Subject:
Year. Month. Date. ( )

تدریس یار داده کاوی جلسه اول ۹۲/۶/۲۴ (غایب)

یک فایلی داریم به نام Data mining_haeri time ، این فایل داده کاوی با نرم افزار SPSS Clemen است. از صفحه ۴ به بعد این فایل، اطلاعاتی مربوط به تیم داده کاوی شروع میشود، آشنایی با نرم افزار داده کاوی است.

نرم افزارهای متعددی در رابطه با داده کاوی وجود دارد که یکی از آنها SAS Enter prisemaner است که یک نرم افزار جداگانه است. نرم افزار Veka هست، نرم افزار Tanagra است، نرم افزار SPSS clementine است و.... یک فایل تئوری هم هست بنام Slid 1 که طریقه نصب نرم افزار را آموزش می دهد. چرا cleme Tine ؟ ما میگوییم کسی که میخواهد داده کاوی بکند به حاطر پایگاه داده های بزرگ است، یعنی از زمانی که داده های پایگاه داده ی بزرگ به وجود آوردند نرم افزارهای متناسب با آن هم بوده، در حقیقت داده کاوی شروع شده، حالا ما در رشته های مختلف مثلاً آمار و سایر رشته های مربوط به آن ممکن است نرم افزارهای که تا به امروز داشتیم، محدودیت تعداد مشاهده داشته باشد، به عنوان مثال، مثلاً ما ممکن است نرم افزار Exel داشته باشیم و با آن بتوانیم همه کار بکنیم، اما پایگاه داده آن باز هم محدود و تعداد شیت های آن کم است ولی در SPSS به این صورت نیست و میلیون ها رکورد میتوان ذخیره کرد. در نرم افزار Exel برای مثال ما اگر در هر shit که اگر 2003 باشد 64 هزار مشاهده داریم، 2007 یا 2010 ما نیم یک میلیون و دویست هزار داده داریم، حال اگر ما ده میلیون داده داشته باشیم و بخواهیم

Subject:
Year. Month. Date. ( )

یک تفسیر انجام دهیم، حالا در سنسور 2007 یا 2010 ما باید حدود 8 تا 10 shit داشته باشیم و داده ها

در shit های مختلف برایمان بیاید، نرم افزارهای آماری به ویژه spss تا 14 میلیون رکورد استاد

داشتن، آن هم spss معمولی بود، یعنی spss clementin هم نبود که ما بخواهیم بگیم این قابلیت را

دارد که برای پایگاه داده ای بزرگ باشد، الان اغلب نرم افزارها به ویژه هر چه که جلوتر می رویم، می توانند

چندین میلیون رکورد را از فایل های relational، بتوانیم فایل flat ای را استخراج کنیم، حالا وقتی

که ما می آییم و این کار را انجام میدهیم، یکبار داده ها را از پایگاه داده های مختلف فراخوانیم در یک فایل و

فایل تخت

به صورت flat در بیاوریمش و از آن جا به بعد فعالیت های آماری مان را انجام می دهیم. یعنی نمایی

که ما در نرم افزار Exel می بینیم و فایلی که در نرم افزار spss می بینیم، نمایی است که switch نداریم

جای رکورد، ترجیحاً این گونه است آن ما فایلی را در spss باز میکنیم و متصل به هیچ پایگاه داده ای رکوردی

نیست. اگر قرار باشد DataBase ای در حقیقت link باشد به DataBase ای که الان داریم، آن

قبلش هم باید فکرش را کرده باشیم که یک فایل تخت که جامع باشد را بیاوریم. این واقعیت برای نرم افزار

spss clementine وجود دارد که اگر ما بخواهیم فایلی را فراخوانیم، اطلاعات را می رود از source اولیه

فرا می خواند و هر کاری که شما بگیری انجام بده، ممکن است که انجام بده، حالا ممکن است DataBase

اضافه شده باشد یا خیر. اگر اضافه نشده باشد که همان اطلاعات قبلی استفاده میکند و اگر اضافه

Subject:
Year. Month. Date. ( )

شده بسته، اطلاعات تون را بروز میکند و در اختیارتان قرار می‌دهد، ممکن است در اینجا بگوییم که ما داریم

از همان spss استفاده میکنیم اما spss که این قابلیت را نداشت.

یکی از کارهایی که در داده‌کاوی انجام میدهیم، اسمش روبش هست، مانند صاحب معدن طلا. در

داده‌کاوی ما اکتشاف می‌کنیم، استخراج میکنیم، فرضیه‌سازی میکنیم. در آمار یک فرضی را به ما می‌دهند و ما

آزمون می‌کنیم! اما در اینجا ما می‌خواهیم فرضیه‌سازی کنیم. از دیدگاه آمار، میگوییم الگوریتم‌ها و ابزار و روش‌هایی

داریم که میتوانند این کار را برای ما انجام دهند یکی از این الگوریتم‌ها clustering است، یکی دیگر از الگوریتم‌ها فاکتور

آنالیز است، یکی دیگر سری‌های زمانی است که کاملش پیش‌بینی است، یکی دیگر از این الگوریتم‌ها یا مدل‌ها،

پیش‌تدریس است ( )، یکی دیگر ماشین‌لرنینگ معمولی است که به روش‌های مختلف باشد،

یکی دیگر cub است، online(olap) و جدول‌های چندبعدی بکشیم، یعنی داده‌ها را با حضور

و اطلاع از سه متغیر ها تدریسمش بکنیم. حالا ما میتوانیم با استفاده از cub، سه متغیر یا جدول سه

بعدی که یکی از آنها جنسیت (gender)، یکی از آنها تحصیلات (e) و یکی از بعدهای دیگر را میتوان وضعیت تاهل (m) در نظر گرفت.

پس ما میتوانیم یک جدول ۳ بعدی داشته باشیم، که در هر بعد ما میتوانیم وضعیت تاهل و تحصیلات را ببینیم و در یک

بعد دیگر تحصیلات و جنسیت را میتوان ببینیم و میتوانیم در عین حال هر ۳ تا را ببینیم، پس کاری که ما انجام میدهیم

این است که میتوانیم با تدریس همین able های معمول خودمان، ولی اینبار به جای اینکه در

یکی ۲ تا کتاب آموزشی گام به گام برای نرم‌افزار: ۱- خانم سمیه علیزاده ۲- دکتر نصیری || چند کتاب مفید: بانک داده تجمعی (کیان رایانه) (راجع به داده انبار گفتم)

Subject:

Year Month Date ( )

Datawar house

۲ جهت باشد، در ۳ جهت باشد، در چندین جهت بگیریم، حالا هدف دیگر این است که سریع پاسخ دهیم.

پس ما به دنبال روش‌هایی هستیم که سریع هم جواب برسد (در زمان کم). این سریع جواب دادن در حقیقت ما را به سوی Data mining می‌آورد و نرم‌افزارهایی که در آن استفاده می‌شود.

پی‌دی‌اف slide

اسلاید ۱) مراحل نصب clementin 12 spss است. در داده‌کاوی یک فرایند خیلی مهمی که باید از پیش ساده بگذریم، بررسی کیفیت داده‌ها است، اغلب الگوریتم‌هایی که ما از آنها نام بردیم، باید این را بدانیم که این الگوریتم‌ها لازمشان این است که اگر Data‌هایمان از کیفیت مناسب برخوردار باشد، ما می‌توانیم به نتایجی که آن الگوریتم‌ها به دست می‌آورند، اعتماد داشته باشیم. برای نصب نرم‌افزارهایی مانند همه‌ی نرم‌افزارها فایل setup را اجرا می‌کنیم، پنجره‌ای که باز می‌شود، نصب clementin را انتخاب می‌کنیم (Install clementine)، در پنجره بعدی گزینه Next را انتخاب می‌کنیم، ممکن است در سیستم‌ها بعد از اینکه گزینه Next را زدیم، پنجره‌ای ظاهر شود که مربوط است به این که آیا ما قبلاً نسخه‌ای را نصب داشته‌ایم یا همین نسخه قبلاً نصب شده باشد، این پنجره، می‌گوید Remove کنید یا Repair. نرم‌افزار در حال نصب کردن است. پس از اینکه کار نصب نرم‌افزار تمام شد، قبل از هر کاری، داخل سی‌دی مربوط رفته داخل فولدر crack چند تا فایل وجود دارد، این اطلاعات مربوط را هم را copy کرده و در مسیری که نرم‌افزار clementine را نصب کردیم، paste می‌کنیم. بعد از اینکه این عملیات را انجام دادیم، حالا می‌توانیم نرم‌افزار را اجرا کنیم.

Subject:
Year. Month. Date. ( )

در نهایت از قسمت start، آیکن clementine12 را انتخاب میکنیم و اجرا میکنیم، همان طور که ملاحظه می کنیم، برای اولین بار وقتی خیلی از نرم افزارها را باز میکنیم، یک پنجره این می آید (صر ۹)، اگر بخواهیم که آن page باز نشود، نرم in future Don't show this dialog را انتخاب میکنیم. بعد از بسته شدن این پنجره، محیط نرم افزار قابل رویت است، که نه shit این دارد، نه ظاهرش شبیه excel و نه حتی شبیه خود spss است. این نرم افزار از چند قسمت تشکیل شده است. به آن صفحه ای که صفحه pdf10 است و شماره های 1، 2، 3، 4 در آن قرار دارند، صفحه ی جریان میگویند. به قسمت بالا سمت راست قسمت مدیریت نرم افزار می گویند، که ما میتوانیم فایل هایمان را میتوانیم در اینجا ببینیم که ۳ تا گره دارد: streams، outputs، Models. قسمت مربوط به مدیریت پروژه های داده کاوی که از چه روشی یا پرچم اساسی، قسمت سازی را ما انجام داریم، براساس مدل crisp_DM قسمت سازی داده کاوی را انجام میدهیم یا براساس قسمت سازی class که در قسمت پایین سمت راست قرار گرفته. و قسمت پایین سمت چپ، در حقیقت میانبرهای است که گره هایی که ما و شما برای کار به آن ها نیاز داریم، این گره ها را به ما خواهد داد. اگر ما بخواهیم یک فایلی را باز کنیم و یک خروجی معمولی ترسیم کنیم باید بتوانیم آن فایل را اول باز کنیم. برای این کار، باید گره ای مربوط به فایل را به تسم مصوفی کنیم و از طریق آن گره که دریم است برای لود کردن فایل به نرم افزار، از آن استفاده بکنیم. اگر ما نیز ها را ملاحظه کنید

Subject:
Year. Month. Date.

یک گره‌ای داریم بنام source ، در ورژن 12 نرم افزار، 9 تا منبع وجود دارد که با سیستم‌های مختلفی هستند، میتوانند دریچه‌های یا گره‌هایی برای معرفی و ورود Data به نرم افزار باشند. در گره‌ی source ، اولین Enterprise view است، در بین Database هست و سومی var file می‌باشد که برای فایل‌های notpad است و بعد fix file را داریم و بعد فایل‌های spass را داریم و بعد فایل‌های Dimensions و بعد SAS Files و بعد فایل‌های Excel را میخواند در انتها User Input که یوزر خودش میتواند فایل‌هایش را وارد کند با انواع فایل‌ها را میتواند با سیستم خودش برای ما در نرم افزار لود کند به عنوان مثال: اگر ما فایلی داریم که قبلاً در محیط نرم افزار Excel مثل شماره ۳ داخل صفحه را بخواهیم در نرم افزار لود کنیم این اجازه را نیستم به ما خواهد داد. فقط یادمان باشد که فایلی را که به عنوان فایل Eexcel وارد میکنیم، شرایطی را باید داشته باشد، که بعداً اگر ما از سیستم دستوری گرفتیم، به مشکل برنخورد. این گره فقط مثل در حقیقت import اگر در Data از نرم افزارهای مختلف به clementine است. ما در نرم افزار Excel در یک این گره Data split داریم، هم نمودار داریم، هم یک قسمت جدول نوشتیم و کارهای متعددی را داریم انجام میدهیم، ما چون فقط میخواهیم Data را بخوانیم، اینها گره‌هایی هستند که از جمله گره‌های هستند که ما می‌توانیم برای خواندن داده‌ها از آنها استفاده بکنیم. ما میخواهیم Data ای را خوانده‌ایم این Data ی

Subject:
Year. Month. Date. ( )

قبلاً در محیط نرم افزار مثلاً Not pad وارد سیستم شدیم، ما برای اینکه بتوانیم این Data را وارد سیستم بکنیم، باید یکی از از گزینه Var و گزینه Fix، از این دو تا میتوانیم استفاده کنیم و به همین ترتیب برای معرفی Data هایی که در محیط Excel نوشته شده، به سیستم از گزینه Excel استفاده میکنیم (در قسمت Source).

pdf 1.1 صفحه A صفحه جریان داده ها است، قسمت B مربوط میشود به مدیریت داده ها، ما ممکن است که چند فایل را همزمان open را باز کنیم، در اینجا دیگر نیاز نیست که چندین بار spass clementine اجرا شود، در stream ما میتوانیم چندین stream را باز نگه داریم، میتوانیم stream ای را تغییر نام بدهیم، میتوانیم stream ای را حذف کنیم، کپی کنیم، ما کاری که انجام میدهیم در این قسمت، در گزینه stream، ایجاد، تغییر نام جریان ها است. گزینه output، ما در نرم افزارهای آماری هر پنجره، فقط یک کار انجام میدهد. Data فقط مربوط به Data هاست، output ها و عسل خروجی و نتایج حاصل از اجرای گره ها را، مربوط به جدول باشد، نمودار باشد، در این قسمت نمایش داده میشود. گزینه مدل هر مدل آماری که ما اجرا بکنیم، که انواع و اقسام مدل های آماری و غیر آماری در این گزینه ی مدل قرار دارد، هر مدلی که ما در اینجا انتخاب کنیم، نتیجه ی حاصل اش در این گزینه می نشیند، پس اگر ما بخواهیم stream، یا صفحه جریانی داشته باشیم، باید نودها یا گره هایی را از قسمت D انتخاب کنیم، و منتقل اش کنیم به صفحه A، تا اینها را به هم وصل نکنیم جریانی ایجاد نمی شود، قسمت C مربوط به نحوه ی یا مدلی که

Subject:
Year. Month. Date. ( )

ما در فرایند داده کاوی ما از آن استفاده میکنیم، فرایند داده کاوی ما میتواند با استاندارد های CRISP باشد یا میتواند با استاندارد یا چرخه ی classes باشد. درحقیقت در CRISP ما باید مثل EZO؛ یا مانند یک الگوریتم step by step فرایند کارمان را انجام که می دهیم، مشخص سازی هم انجام بدهیم. classes کارش دسته بندی است ولی در classes آن فرایندی کاری که ما داریم انجام میدهیم فعالیتی که داریم انجام میدهیم، regresiun ای که داریم انجام میدهیم، آن گره و گره های مشابه را برایمان دسته بندی می کند، ولی CRISP فرایند کار را به ما می گوید. ما اولین کاری که باید در داده کاوی انجام بدهیم باید ببینیم متغیرهایمان چی هستش؟ برای کی می خواهیم کار بکنیم؟ چی میخواهیم؟ جنس داده ما چیست؟ در اینجا گزارش تهیه می کنیم. این قسمت C، که میتواند فایل Excel باشد، pdf باشد و هر چی باشد، یک قسمت سازی است مثل یک زدم که جداگانه ای است که ما در آن اطلاعاتمان را قرار می دهیم، مگر این نرم افزار این قابلیت را دارد که ما بیایم فایل pdf بسازیم و...؟ این کارش این است که از ما shortcut میگیرد، یعنی وقتی که ما میخواهیم توضیح بدهیم، یک فایل word ای را بهش معرفی می کنیم، و قسمتش را برایمان نگه میدارد. درقسمت D گره ها را می بینیم، گره source اگر انتخاب شود، منابع و گره هایی که دارای ورودی استفاده میشود در شکل اش هم دایره است، یعنی منابع ورودی ما گره ای دایره ای شکل است.

ممکن است یکسری داده ها را از Excel فرا بخوانیم، ممکن است نیازی داشته باشد که ما فیلدها را به تغییرات خاصی را بر روی آن بدهیم، مثلاً بیاییم دو تا متغیرمان را با هم ترکیب کنیم و از آن یک متغیر جدیدی

ایمیل استاد: amargostaran@yahoo.com

Subject:
Year. Month. Date. ( )

سیبا زیم، باید بردیم در گوگل، record options، ما داده‌هایی را که بخواهیم باهم ریکورد ترکیب کنیم، نودهای
از این قسمت استفاده میکنیم، ممکن است گاهی اوقات بخواهیم از روی چند تا متغیر، متغیر جدیدی را تعریف کنیم،
یا اینکه بخواهیم بخشی از اطلاعاتی را برای یک متغیر خاص، نمایش دهیم، Record option شش ضلعی است.
اگر بخواهیم نموداری را ترسیم کنیم، فیلدهای مثلثی شکل خواهد بود.

جلسه دوم تدریس یادداده‌کاوی 92,6,31

در چرخه CRISP یادگیری داده‌کاوی به روش CRISP، ما مراحلی که در هر ویژه داده‌کاوی دارد، که در اینجا حدود
۶ مرحله است، مطرح میکنیم، یک مرحله Business Understanding داریم، یکی Data Understanding
است، یک Data Preparation داریم، یک Modeling هست، یک Evaluation هست و یک
Deployment است. کسی که بخواهد یک فرایند داده‌کاوی را انجام بدهد، ابتدا باید در آن صنعتی که میخواهد فعالیت بکند،
شناختی از آن صنعت باید داشته باشد، و مدل‌ها پیش را خوب بشناسد، فرایند کار را باید در حالت معمولی با حاشیه
آشنا باشد، ما اگر بخواهیم از این چرخه استفاده بکنیم، این به درد من تازه‌کاری میخورد که میخواهم بروم
متندات خودمان را مکن است در یک فایل صوتی، در یک فایل word یا Text، مراحل شناخت خودمان
را از آن صنعت، ثبت کرده و میخواهیم داشته باشیم، اینجا در حقیقت مسیری است که یک short cut
برای ما ایجاد میکند و از ما میخواهد که آدرس بدهیم. مثلاً ما در DeskTop یک فایل را

Subject:
Year. Month. Date. ( )

ذخیره کردیم که در رابطه با Business Understanding است، ما میتوانیم آن فایل را کلیک کنیم و Drag کنیم و در آن پوشه قرار دهیم، کاری که این پوشه میکند، در حقیقت shortcut مسیر فایل مان را برای ما ذخیره میکند، حالا ممکن است علاوه بر اینکه ما آن صنعت را شناختیم، نیاز است که یک شناختی از Data داشته باشیم، Data Understanding، اینجا میگیم که ما رفتیم اطلاعات را دیدیم و تمام اطلاعاتش را میتوانیم در این فولدر قرار دهیم، ممکن است که ما بخواهیم تغییری یا تبدیلی بر روی داده هایمان بزنیم، یک پیش پردازش بر روی داده هایمان داشته باشیم، میتوانیم Data preparation ذخیره کنیم. اطلاعات مربوطه را در این محیط

ممکن است کسی بخواهد بر روی Data اش مدل خاصی را بزند، وقتی که ما این کار را بخواهیم انجام دهیم، هم میتوانیم مستندات و هم نتایجی که در قسمت بالا در نظر نیم یا گرفته Model داریم، میتوانیم مدل مربوطه را کلیک کرده و Drag کنیم در این پوشه Modling، shoratcut اش را بگیریم ایجاد شود. بعد از اینکه ما مدل بندی میکنیم باید نتایج را ارزیابی کنیم، آیا این مدل مناسبی بود؟ آیا چند مدل داشته باشیم بین این مدل ها کدام مدل بهتر است؟

پس در این فولدر ما هر چیزی میتواند قرار بگیرد، در واقع یک زدم کم است، وقتی ما داریم یک پروژه ای را انجام میدهیم، یک زدم کلی را برمی داریم و عرض میکنیم که آیا ما میخواهیم Data ای را آماده کنیم، می رویم یک copy میگیریم از Data متغیر هایی که بهم دارند، و بعد یک CD اطلاعات

Subject:
Year. Month. Date. ( )

رو هم گرفته اند، آن را هم داخل این روم کن قرار می دهیم و ... . در حقیقت در اینجا صرف چیزی با صر format ای را میتوانیم قرار بدهیم. در حقیقت مراحل مستندسازی ما را در اینجا نشان داده میشود، در نهایت فولدر Development ما است که وقتی که ما پروژه داده کاوی را اجرا میکنیم، در انتها صدف حال این است که در کارهای بزرگ ما بیاییم اتوماسیون بکنیم و این ها را اجرا بکنیم. در حقیقت آن Document ای، چیزی مراحلی برای خودکارسازی این اتوماسیون انجام دادید، میتوانید در اینجا مستندسازی را قرار دهیم که بعداً یا هر موقع که خواستیم برگردیم و این پروژه را مرور کنیم، بتوانیم تشخیص بدهیم که چه مراحلی، چه اقداماتی، چه کارهایی انجام شده و این اطلاعات کجا ذخیره شده، که این قابلیت در نرم افزارهای رایگان وجود ندارد، ممکن است افرادی بیایند و از تیم classes استفاده کنند، این تیم در حقیقت زمانی ما ازش ممکن است استفاده بکنیم که ما گروه classes، stream ها، نودها، مدل هایی که تولید کردیم، مبادل دیتافرم ها، در حقیقت به جای اینکه بیاییم براساس فرایندکاری، در سری قبل بشه برای افراد جنتی یا افرادی که میخواهند مثلی یک فرایند نظم خاصی کارشان را پیش ببرند استفاده میشه (CRISP_DM)، اینجا ما میخواهیم بگوییم که چنتا stream (classes) استفاده کردیم، یعنی چند تا فایل داده باز کردیم، از چه گره هایی استفاده کردیم، چیزی که برایمان مهم است در حقیقت داریم مراحل یا آزمون ها را ذخیره میکنیم و از آن ها استفاده میکنیم، کسی که بخواهد از این فرایند استفاده کند، میتواند برود اطلاعاتش را برحسب ای

Subject:
Year. Month. Date. ( )

سیستم classes پیش بیرد و کارش را انجام بدهد، در قسمت پایین، قسمت چهارم، قسمت D، اینجا در حقیقت shortcut های آیتم ها، مدل ها و سورس هایی که ما میخواهیم استفاده بکنیم، میتوانیم از این قسمت نشان دهیم. ما در قسمت مربوط به shortcut ها، گروه source را داریم، یعنی الان Favorites ما، بخشی از همه ی shortcut های ما را دارد نشان میدهد، مثلاً از گزینه source دو تا از نودها، از گزینه record operation ۱ تا ۳ از نودها، از گزینه Field ops ۱ تا ۳، از برنامه ها ۱ تا ۳، از مدل ها ۱ تا ۴، از output و Export ها، که به صورت default نشان داده شده است.

ما بر این اینجا بتوانیم از نودهای بیشتری استفاده بکنیم، و نودهایی که مورد نیاز ما باشد و مناسب با آن کارمان باشد استفاده بکنیم، میآییم هر کدام از آنها را میبریم در گروه اختصاصی خودشان. اگر وارد گروه source شویم اولاً این گروه به صورت دایره ای شکل است، این گروه شامل گره هایی است که در حقیقت ورودی های سیستم ما است، که میتواند یکی از فایل مربوط به Notpad، spss، Access، Execl و سایر فایل ها از سیستم های دیگر را میتواند برایمان بخواند، ترجیحاً در عمل کارمان جوری است که داده مان معمولاً در فرمت های دیگری در سیستم های دیگر ذخیره شده اند، یا از معمولاً Exed یا از معمولاً sql server یا از جاهای مختلف دیگری، یا هر سیستمی که سازگار با نرم افزار Microsoft باشد، میتوانیم اطلاعات را به صورت رو خط بخوانیم، ممکن است کسی بخواهد بخشی از Data هایش را select کند، ممکن است بخواهد Data پیش

Subject:
Year. Month. Date. ( )

برای پس یک فرایند خاصی انتخاب بکند، ممکن است بخواهد Data ش را Sort بکند، ممکن است بخواهد

داده هایش را از روی چند تا فایل دیگری بیاید اطلاعاتش را Merg کند، میتواند از گزینه هایی که در Record

options است استفاده کند که صورت اصلی نمایش داده شده. کرکره بعدی Field ops است، آن

کسی بخواهد تبدیل یا تغییری روی داده هایش بدهد، یا مثلاً بخواهیم یک تغییر جدیدی ایجاد بکنیم، از این کرکره استفاده

میکنیم. اگر بخواهیم برای داده هایمان نمودار ترسیم کنیم، نمودار ۲ بعدی، اگر بخواهیم برای ارزیابی مدل هایمان،

بررسی کنیم که کدام یک از مدل های ما مناسب تر است که از نود avaliation استفاده کنیم، باید وارد کرکره مدل شویم.

صفحه هشتم و اسلایدهای Data mining ۲۶ و ۲۷ اگر کسی بخواهد مدلی را بر روی داده هایش اجرا کند، به عنوان مثال ما

میخواهیم یک درخت تصمیم الگوریتم C5.0 را برای داده هایمان اجرا کنیم، از درخت تصمیم استفاده کنیم، ممکن

است بخواهیم از neural Network استفاده کنیم، اینها Item هایی هستند یا نودهایی هستند که در مدل بندی،

کرکره هستند و ما می توانیم از اینها استفاده کنیم. ممکن است ما خروجی خاصی را برای داده هایمان خواسته باشیم، قبلاً

عرض کردیم که اگر بخواهیم Data هایمان را فرا بخوانیم، باید یک نودی را معرفی کنیم که شروع کار با آن نود است،

ما میخواهیم مثل Excel، Data هایمان را ببینیم، برای این منظور باید output بگیریم، output مان

را باید در یک Table جداگانه ای بگیریم، یا داده هایمان را در یک جدول در بعدی به فرم میانگین یا آمار خلاصه

نشان بدهد، هر output ای که بالا نیاز هستیم باید بیاییم در کرکره output ها و بخواهیم که داده ها کرکره

نمایان خروجی دهد.

Subject:
Year. Month. Date. ( )

ممکن است بخواهیم Data هایمان را Export کنیم در نرم افزار دیگر، این امکان برایش وجود دارد که چه را

اگر Data یا Item هایی که ما تولید کردیم از جنس Data باشد یا مثلاً از جنس اطلاعاتی باشد که

امکان Export به sql server یا سایر Database ها و ... باشد که سازگار با نرم افزار

ما باشد، در کرکره Export این امکانش وجود دارد. پس میتوانیم از این shortcut ها برای کار

ازش استفاده کنیم. خوب کنیم میخواهیم Data ای را فراخوانیم در صفحه نرم افزار، برای این کار

باید وارد کرکره source شویم و بگوییم که آن Data ای را که آماده کردیم، با چه فرمتی ذخیره شده

است، نوعی که در اینجا مثلاً داریم (شماره 1) یک فایل SPSS ای است، ما به ۳ طریق میتوانیم نودها

را فعال کنیم به نرم افزارمان: ① روی نود مربوطه دابل کلیک کنیم مثلاً در کرکره source یکی از نودها را

انتخاب و فراخوانی انجام بدهیم، ② روی نود مورد نظر یک بار کلیک کرده و بیاییم در تخته جریان (صفحه جریان) دوباره کلیک کنیم در

صفحه ③ کلیک کنیم، Drag کنیم. آیا میتوان چند تا جریان متفاوت هم زمان وارد صفحه جریان کرد؟

بله میتوان. برای اینکه بتوان Data هایمان را load کنیم، روی نودی که آوردیم دو بار کلیک می‌کنیم،

پنجره ای باز میشود، و از ما آدرس یا مسیری که فایل مورد نظرمان (در اینجا spss) در آن جا ذخیره

شده را load میکنیم و از اینجا پیدا خواهد کرد. فعلاً ما چیزی اینجا نکردیم و تنها یک فراخوانی انجام دادیم.

فرمت داده های spss در هر نسخه ای که باشد، نام فایل .sav خواهد بود.

Subject:
Year. Month. Date. ( )

پس ما در این کاری که انجام دادیم، تنها کاری که انجام دادیم این بود که فایل مورد نظرمان را باز کردیم، هیچ چیز خاصی نبینیم، فقط به ما می‌گوید که یک دریچه‌ای باز خواهد شد، شرایطی برای ورود داده داریم، مثلاً نقطه و فاصله و...

آن شرایط چیه؟ وقتی که ما Data‌هایمان را در محیط spss ذخیره کرده باشیم، محیط spss داده‌هایش شبیه محیط Excel است، هر متغیر در spss حتماً در یک ستون است، اطلاعات هر case، هر فرد در یک ردیف است، قاعدتاً نه از نقطه، نه از سینی کاوش، نه از فاصله برای جداسازی داده‌ها استفاده نشده است. اگر کسی دستور دهی بزند Excel، اگر Data‌ای را بخواهیم از محیط Excel فرا بخوانیم، در سلول‌های جزایی، Data‌هایمان آمده باشد، معمولاً ما در یک سلول Excel نمی‌آییم بنویسیم که مثلاً عدد 5 فاصله عدد 7 فاصله عدد 9 فاصله و... هر فرمتی را که بتیم به عنوان ورودی بشناسد، امکانش هست که ما برای ورود از آن استفاده بکنیم. (صفحه 4 اسلاید 8). یک علامت اساسی هم است که با A نمایش داده شده، اگر کسی روی آن کلیک کند، در حقیقت یکی از فایل‌های از قبل ذخیره شده یا سمپل‌های را دارد که ما می‌توانیم به این sample‌ها مراجعه کنیم، با پسوند spss و Data مربوط‌مان را load کنیم.

در اسلایدهای بعدی، اگر کسی بخواهد از چند تا فایل Data بخواند، اگر کسی بخواهد از چند تا فیلد، فیلد جدیدی ایجاد بکند، فیلد جدیدی را حذف کند، یعنی روی فیلدهایش کار کند، اگر کسی بخواهد گرافی، نمودار رسم کند، اگر کسی بخواهد مدلی را ارزیابی کند، آزمون کند، خروجی رگرسیونی، آنالیز تم خاصی به هر، اگر کسی

Subject:
Year. Month. Date. ( )

بخواهد خروجی مناسبی بگیرد، اگر بخواهد اطلاعاتش را Export کند در صفحه دیگری که مدنظرش باشد،

همه‌ی اینها را رعایت پایین که قبلاً توضیح داده انجام میشود. حالا همه‌ی نرم‌افزارهای آماری، یا به تعبیری

همه‌ی نرم‌افزارها، یک ویژگی خوبی که دارند، help های قوی دارند، وقتی که برانیم دنبال مثلاً آزمون

TTest هستم، توصیه میشود وقتی help Topic را بزنیم، میتوانیم در کلمه‌ی index این فیلتر TTest، و آزمون‌های

TTest ای که دارد را پیشنهاد میکند و آن نود مربوطه را استخراج میکنیم. گزینه‌ای که در اینجا select شده

و قرمز رنگ شده CRISP-Data mining help است که در حقیقت این متدی که ما برای پروژه‌های داده‌کاوی بیان را

گفتیم با چه سیستمی میخواهیم پیاده‌سازی کنیم. اگر کسی روی CRISP-DM help کلیک کند، در حقیقت می‌آید

مراجع به این الگوریتم، مراجع به این متودولوژی که دارد، توضیحاتی را برایش گذاشته که اگر ملاحظه کنید،

آن 7 مرحله مربوط به آن را گذاشته، در حقیقت خواسته کار را برای ما توضیح داده، حالا در قسمت Index

ما میتوانیم Item یا نود یا مدل یا Table ای که مدنظرمان است را نامش را آن جا تایپ کنیم، به عنوان

مثال، ما C & R Tree تایپ کردیم، خواستیم که برایمان توضیح دهد، پنجره Topics found

باز میشود، ما هر کدام از Item ها را بخواهیم، میتوانیم روی‌اش کلیک کنیم و Display را بزنیم، پنجره مربوط

به آنها با تمام جزئیات مبسوط و جامع نشان داده میشود. در اسلاید بعدی، انتهای اسلاید 1 (slid 1)

صفحه ۱۰

همان که می‌بینید، در صفحه stream، چهار نود آورده شده، پس امکان نودهای آموزشی

Subject:
Year. Month. Date. ( )

مختلف هست به چه روشی؟ به همان ۳ روش که قبلاً توضیح داده شد. File Fix و ویژگی‌ای که واقعه‌ای است

که تغییرهای ما یا Data های ما با فرمت fix وارد شده‌اند، چه طوری وارد شده‌اند؟ مثلاً شماره‌هایی که یک

عدد ۱۰ رقمی است، مثلاً فرضاً یک ملی 0989123 دارم، داده‌هایی که بدون هیچ معدودیتی

به هم چسبیده باشند fix شده باشند، این قبیل داده‌ها را که معمولاً در محیطی Notpad یا word

یا TxT وارد می‌کنیم می‌توانیم از طریق نود var.fix فراخوانیم. ممکن است که بعد آخوندان چطوری این

Data ها را می‌خوانند یعنی همین ۱۰ رقم می‌خوانند؟! فعلاً ۱۰ رقم می‌خوانند، خوب ما که ۱۰ رقمش هم در دهان نمی‌خورد،

می‌گیم به چه صورت به درد می‌خورد؟ می‌گوییم از ما خواسته شده که ۳ رقم اول این عدد را در یک ستون بزنید، ۲ رقم

دیگرش را در یک ستون بزنید، ۱ رقم دیگه در یک ستون بزنید، ۲ رقم دیگه در ستون دیگه و ۱ رقم دیگه ستون دیگه که

هر کدام از این‌ها شناسه‌ی یک چیزی برایمان هستند. یعنی ما می‌خواهیم این عدد ۱۰ رقمی را در ۶ ستون ذخیره

کنیم، پس ما می‌توانیم اینو از آن نود استفاده کنیم. حال ممکن است که ما در یک Text یا Notpad یا

داده‌هایمان را با یک separator ای از هم جدا کرده باشیم، مثل اینجا که داده‌هایمان را با slash جدا کردیم،

ممکن است با یک simicolon یا dash یا ... جدا کنیم، در این حالت این امکان برای ما وجود دارد که بیائیم

داده‌هایمان را با آن separator ای که داریم، جدا کنیم که برای این کار در var.file روی کلیک

عکس ۶

Subject:
Year. Month. Date. ( )

جلسه سوم 92/07/7

ما یک فایل Notpad داریم، یک داده‌ی را داخل فایل Notpad وارد کردیم، الان میخواهیم آن Data ها را در نرم افزار spss clementine، load کنیم. پس ما فرمت Data ای را که در حقیقت در نرم افزار داریم وارد میکنیم، ممکن است با فرمتی که ما مدنظر داشتیم، نخوانند، چون سیستم ما به طور خودکار می رود از روی متغیر فرمتی که خودش تشخیص بدهد را برای فیلد ها در نظر میگیرد، ما میتوانیم به سیستم بگوییم که متغیر بده یا خودمان باید در سیستم این امکان را ایجاد بکنیم که نام متغیر ها را، نوع متغیر ها را، یا هر چی دیگر که بخواهیم براساس چیزی که مدنظر ما است، تغییر دهد البته یک جوری هم نباشد که مثلا ما به اشتباه یک فرمت داده‌ای را که کسی است بخواهیم برایش معرفی کنیم. پس ما میتوانیم از این مسیر استفاده کنیم و حتی نوع متغیر هایی که مدنظرمان است را انتخاب کنیم و مدنظر داشته باشیم، الان میخواهیم فایلی را باز کنیم با فرمت var.file یعنی از گروه source ⇐ فایل های را با فرمت variable تعریف بکنیم، برای این کار نود مربوط به var.file را از گروه source انتخاب میکنیم، وقتی این گره را انتخاب بکنیم امکان load کردن، import کردن فایل هایی که ما مدنظرمان است را برایمان ایجاد میکند، var.file هم فایل های از جنس TXT است، Notpad ایی است، و بعد از انتخاب Double click کردن پنجره ای باز میشود، حالا از قسمت جلوی file یک گزینه ای است بنام Browse که داخل ۳ تا نقطه قرار دارد که به منظور Browse کردن استفاده

Subject:
Year. Month. Date. ( )

میشود، اگر مسیر خاصی مدنظر باشد در این قسمت open میکنیم، ممکن است که فایلی در دسترس نداشته
(slid 2) صفحه 6
باشیم، خود این نرم افزار در کنار Demo آدرس اصلی کلیک کنیم، در اینجا میتوانیم روی این گزینه
کلیک کنیم تا پنجره مربوط به فایل های Demo ای که روی خود نرم افزار است برایمان باز شده، پس
ما میتوانیم از فایل های Demo نیز استفاده کنیم، روی Demo که کلیک کردیم، بسته به فرمت File،
فایل هایی را به ما نشان میدهد که در این help وجود دارد، الان فایل که به فرمت TxT یا notpad ای
باشد میتوانیم در اینجا خیلی بخوانیم، ما فایل ORUG1 را در اینجا ما روش کلیک میکنیم و بعد open را میزنیم، فایلی که
ما کلیک کردیم و open کردیم، به این فرم که در اسلاید 7 می بینید است، اگر ملاحظه فرمایید Data های
ما با کاما از هم جدا شده اند، در حالت قبل Data های ما به هم چسبیده بودند، هیچ جداکننده ای بین
آن ها نبود، به همین دلیل ما فرقی از Var.File استفاده میکنیم، داده هایمان این قالب را خواهند
داشت، پس seprator ما اینجا کاما شده، اگر با این قسمت Delimiters (صفحه 7 اسلاید ۲)
بروید، space ما میتوانیم، 2 تا عدد را با استفاده از space از یکدیگر جدا کنیم، میتوانیم هم زمان، یا
به تنهایی از کاما استفاده کنیم، میتوانیم از کلید Tap استفاده کنیم، میتوانیم Enter کنیم و داده هایمان در
سطر جدید به عنوان seprator عمل کنند، میتوانیم از کاراکتر دیگر هم استفاده کنیم ولی اینجا باید تعریفش
کنیم، other ممکن است ما بگوییم slash، back slash، ...، * و ...
میتوانیم آن را به عنوان seprator در نظر بگیریم، caracter Non-prenting، میتوانیم از فاصله مجازی هم

Subject:
Year. Month. Date. ( )

آدرس Data ها نباشد، از اون هم میتوانید استفاده کنید، و یا میتوانید از چند تا blanck یا space

به عنوان معدد کننده استفاده کنید. پس ما در فایلمان میدانیم هم زبان هم سیمی کالون داشته باشیم، هم

کاما داشته باشیم، هم Tab داشته باشیم، هم newline داشته باشیم، هم space داشته باشیم؟ بلی

پس با این فرق هایی که در قسمت Dilimiters دیدیم، این امکان برایش وجود دارد که بتواند تشخیص

دهد داده ها با چه فرقی در کنار هم قرار گرفته اند. در این پنجره به ملاحظه میکنید، ممکن است که ما در حالت

اول، وقتی قبل از اینکه Data بخوانیم، در این تیک که وجود دارد، نوشته Read field names from file

که ردیف اول من بیان کن نام متغیر های من است، آنچه چیزی که داریم در ردیف داده ها داریم می بینیم،

ردیف اولی که در این پنجره ما وجود دارد، پاسخ های ما است، نام متغیر ما نیست. اگر نام متغیر بود این

تیک میخورد اگر نبود ما اجازه نداریم به عنوان نام متغیر استفاده بکنیم. (slid2, safe7).

این مورد را وقتی که ما میخواستیم، فایلی را در محیط نرم افزار Excel و یا محیط دیگر نیز فرا بخوانیم،

این امکان برایمان وجود داشت، وقتی از محیط Excel فایلی را ذخیره کرده باشیم و بخواهیم در یک محیط

دیگر باز کنیم، یا برعکس، Data ای داشته باشیم با فرمت دیگر و بخواهیم در محیط Excel باز کنیم، از ما

می پرسد که اسم فیلد هایت در ردیف اول است یا خیر. اگر اسم فیلد ها در ردیف اول باشد، این تیکی

که پایین داده ها ملاحظه می کنید، را باید فعال کنید تا آن اسم ها را برای ما load کند و بخواند.

Zarrin Barg

Subject:
Year. Month. Date. ( )

ممکن است که ما بخواهیم در فایل مان چند ردیف را حذف کنیم، ما بگوییم از ردیف ۵ به بعد میتوانیم از skip

header characters را انتخاب بکنیم، اگر این گزینه را بزنیم، از اطلاعات چند ردیف صرف نظر می کند مثلاً

از ردیف 5 6 به بعد بیاییم Data هایمان را load کنیم و ادامه بدهیم تا بتوانیم از شان استفاده نماییم.

ممکن است که ما مثلاً بخواهیم داده هایی که داریم، در این داده ها فقط قسمتی از آن فیلد را فرا بخوانیم، این

امکان برایمان وجود دارد که ما بتوانیم بخشی از این Data ها یا فیلد هایی که داریم را فرا بخوانیم. الان همه ی

Data ها بوسیله ی کاما از هم جدا شده اند و ما اینجا خاصیتش را داریم می بینیم. اگر ما وارد گزینه ی Data

شویم، الان در قسمت Feild ها یک داده فیلد های ما را لیست میکند و به عنوان جدا کننده به ما نشان

می دهد، اگر ما وارد گزینه Data شویم، در این گزینه انتظار داریم که هر کدام از این ستون ها که با کاما

جدا شده اند، به ترتیب بیان کننده ی از این فیلدهای فرد باشد. شروع ~~[illegible]~~ : خب ما Data

pdf 3.1 مرا

هایمان را در این فرمت هایی که داریم در نظر میگیریم، در قسمت شماره ۱، آن اطلاعات مان را داریم ملاحظه

میکنیم، وقتی read file names زدیم، الان که بالای تصویر می بینید، Age, sex, EP, ...

این چند تا متغیر را داریم در اینجا می بینیم، قسمت ۲، الان وقتی read Field names From File زدیم

را بزنیم، ردیف اول را به سر ستون هایمان در نظر میگیرد، و میتوانیم به عنوان متغیر مان از آن استفاده کنیم،

در قسمت شماره ۳ اطلاعات هر ردیف را برای یک نفر در نظر می گیرد، ممکن است که ما بیاییم

Subject:
Year. Month. Date. ( )

برای اینکه محیط Notpad مان، سعه هایش زیاد نشود، هر دو نفر را اطلاعاتشان را در یک ستون در نظر بگیریم. پس اتوماتیک داریم، جیاد میشلند و اگر ۱۲ تا متغیر ما در ستون مان داشته باشیم، وقتی اینجا تعریف کنیم که در هر ردیف اطلاعات مربوط به ۲ نفر است، مثلاً ۲ تا ۶ تا در نظر می‌برد، ۶ تا اول برای نفر اول، ۷ تا ۱۲ برای نفر دوم (ردیف دوم)، پس هر ردیف را نتوانیم بنویسیم اطلاعات مربوط به چند نفر در فایل Notpad وجود دارد، تا بتواند آنها را load نماید. می‌توانیم از سیستم بخواهیم چند تا از اطلاعات‌مان را حذف نماید (شماره ۴). در قسمت ۵ می‌توانیم بگوییم جدا کننده های ما کدام اند. آیا داده‌های من به وسیله‌ی کاما sem، ... . پس در کرکره اول که کرکره مربوط به File است، اطلاعات مربوط به متغیر را داریم می‌بینیم. منتظر هم وارد کرکره دوم، کرکره Data شویم، وقتی ما کرکره Date را می‌بینیم، آن سرستون های ما، چون آن تیک فعال بود، به عنوان نام فیلد های ما داریم ملاحظه می‌کنیم. (صفحه ۲ فایل 3.1) الان کدام کرکره ما فعال است، کرکره Type و در Data هم فقط سرستون را داریم می‌بینیم که در این تصویر نمی‌بینیم و نوع متغیر را مثل حالت قبل. الان ما وارد ستون Type شویم، کرکره Type قرار بگیرد تا بگیرد چه کند! نام متغیرها در فیلد بود، در Data بود، اگر می‌خواهیم تغییر بدهیم فقط در کرکره Data این امکان برای ما وجود دارد که بیاییم به جای age، Type کنیم sex. پس امکان تغییر فرمت متغیرها، امکان تغییر نام متغیرها نه در فایل هست و نه در Type. پس در کرکره Data و فیلد امکان اینکه

Subject:
Year. Month. Date. ( )

اگر ما بخواهیم نوع متغیرمان که الان کردیم اش باز است، اسم متغیر که را که ما اینجا داریم لحاظ میکنیم، اگر بخواهیم این
اسامی را تغییر دهیم، در این دو تا گزینه قبلی برایمان امکان پذیر است. مثل: هشتم وارد گزینه Type شدیم،
متغیر sex ما به صورت range معرفی شده، ما چند جور متغیر داریم؟ متغیرهایمان یا بصورت Rang
هستند یعنی کمی هستند، یا از جنس Discrate هستند یعنی متغیرهای کیفی از هم جدا یعنی حالت کمی ندارد،
گسسته است، یا نوع میتواند Flag باشد، یعنی در حالتی یعنی از جنس True, False، این هم نوعی
گسسته است ولی گسسته ای که باینری است. set هم Discrate مانند است ولی با این قابلیت که میتواند
بیش از Discrate (متغیرهای گسسته) هم عدد داشته باشد، هم حرف داشته باشد و... پس Discrate
بیشتر به خاطر گسسته بودن یا حرفی بودن یا رشته ای بودن مورد استفاده قرار می گیرد. در set ما تنها امکان این
برایمان وجود دارد که به صورت حرفی داشته ای عددی و... برایمان وجود خواهد داشت. Ordinal set
یعنی از جنس کیفی ترتیبی هستند، ما میتوانیم متغیرهایمان را مثلاً به فرم رضایت مندی، مقطع تحصیلی، سلسله مراتب
متغیرها
مدیریتی و... اینها مواردی هستند که داده های ما از جنس ordinal هستند، (در مواردی ما در حقیقت درجه
عضویت داریم ولی وقتی که متغیرها از جنس ordinal هست، اینجا صحبت از رضایت مندی خیلی کم، کم و...
آن ها مثلاً درجه عضویت بد عدد میشه
است، یعنی در عین حال که از جنس set هستند ولی ترتیب در آن ها مهم است.) ! ممکن است
که نوع داده ها Typeless باشند مثل کدملی، شماره دانشجویی، کد پستی و... که اینها نه ترتیبی

Subject:
Year. Month. Date. ( )

در آنها مهم است، نه چیزی دیگر و در واقع یک Unic است که در واقع در محاسبات استفاده نمی‌شوند.

نکته‌ای که هست ممکن است که ما بخواهیم دوست نداریم متغیر خاصی را در ادامه فایلمان از آن استفاده کنیم و

ببینیم، می‌توانیم وارد کرکره Filter شویم و از سیستم بخواهیم که این را به ما نشان ندهد؛ اگر بخواهیم متغیری را بخواهیم

در فرایند تغییرات Data ست را load نشده، می‌توانیم از Filter روی آن فلشی که از قسمت Field

خارج شده، روی آن کلیک کنیم و علامت x بخورد که در این صورت آن Data را نشان نخواهد داد.

اگر ما متغیری که در اینجا فرمتش را تغییر می‌دهیم، باید Read value با این کلیک را فعال کنیم،

Run کنیم، تا در قسمت ستون ما Values ها بیاید، ما مشاهداتمان را باید Read value

بزنیم. اگر Read values را نزنیم هنوز Data ها را نخوانده، ممکن است که نتواند خروجی به ما بدهد،

چون Data را نتوانسته بخواند. پس ما باید Read value کنیم تا فرمت خاص داده‌ها را بتوانیم ببینیم.

اگر Read values را بزنیم، ملاحظه می‌کنیم که در ستون Value ها، Data های را که منتظر

ما هست را سیستم نشان می‌دهد (صفحه ۱۴ اسلاید) کلیه داده‌های را که ما در سیستم‌مان معرفی کردیم

برای سن در گروه گذاشته، از مینیمم 15 تا ماکزیمم 74، جنسیت به شکل Flag هست

یعنی دو حالت male , Female. فشار خون به صورت set نشان داده است. در

سیستم ممکن است Data ای را که داریم، بخواهیم که این داده‌ها را سیستم برایمان فراخواند، در این

Subject:
Year. Month. Date. ( )

در سیم آلارم values را کلیک کنید، این کادری را که ملاحظه میکنید (ص ۴) ظاهر میشود. این کادر وظیفه‌اش این است که نوع Data ای را که ممکن است ما در فرایند بخواهیم استفاده کنیم، به چه صورت باهاش برخورد کنیم، به عنوان مثال اولینش current است در شبیه‌ها، اگر داده‌هایمان حالت current باشد، و ما read value را بزنیم، دیگه داده‌های ما بعداً در فایل اصلی‌مان تغییری نکند، اینجا در حین اینکه ما داریم داده‌هایمان را load می‌کنیم و فرآیندش می‌خوانیم، داده‌هایمان دیگر تغییر نخواهند کرد. همین داده‌های جاری را به عنوان، داده‌های اصلی ما تلقی می‌کنند و از آن استفاده می‌کنند، بعداً اگر تغییر کرد، از آن استفاده نمی‌کند، فقط چیزی که الان لود شده را استفاده می‌کند مگر اینکه دوباره بروم، فرمتش را تغییر دهم. اگر کسی گزینه <Read> را انتخاب کند، به این معنا است که هر بار که سیستم می‌خواهد مدلی را Run بکند، فرایندی را از اول بخواند، نیاز هست که دوباره شما باید داده‌هایتان را read value را بزنید تا داده‌های جدید را مجدداً برایت فرا بخواند، اگر گزینه <Read +> را بزنید، ممکن است که فایل اصلی ما داده‌هایش کم و زیاد شده باشد، اگر داده‌ها کم و زیاد شده باشد و + Read را بزنید، این امکان وجود دارد که تغییرات Data را در فرایندی که ما مدل را Run می‌کنیم از اول چون می‌رود از source داده، Data را خودش فرا می‌خواند و می‌آید وقتی می‌رسد به + Read تغییرات احتمالی که ممکن است روی پایه‌های داده‌هایمان داشته باشیم، اعمال خواهد کرد. اگر گزینه <pass> را بزنید، وقتی که سیستم به این قسمت می‌رسد ما value اش

Subject:
Year. Month. Date. ( )

را pass زدیم، اصلاً مقادیر مربوط به این متغیر را برایمان ملاحظه نمی‌کند یعنی در نظر نمی‌گیرد. اگر ما مقادیری از value ها را به اشتباه وارد کرده باشیم، امکان پاک کردن آنها توسط clear value وجود دارد. clear value می‌آید این تغییرات که الان ما در نظر گرفتیم برای value ها، آن ها را پاک میکند و دوباره باید Read value بزنیم. ولی اگر clear All value را بزنیم، کاری که برای ما انجام می‌دهد این است که، ممکن است ما تایپ رو هم عوض کرده باشیم، اگر این گزینه زده بشه، نه تنها values را برایمان Reset میکند، بلکه Type را هم Reset میکند. ممکن است که ما در value گزینه Specify را انتخاب بکنیم، امکان این را که بتوانند برای داده هایی که در نظرش است مثلاً برای بعضی پاسخ‌هایش توضیحات خاص خودش را اضافه بکند، مقادیری را برایش پیشنهاد بکند، این امکان ما وجود دارد که بتواند از طریق specify تغییرات لازم را برای خودش انجام بدهد. یک ستون دیگه داریم بنام Direction ما در فرایند داده کاوی، چیزی که برایمان خیلی مهم است این است که آیا متغیر مورد نظر چه نقشی برایمان بازی میکند؟ یعنی نقش input داره یا output یا هر دو نقش را دارد؟! پس ما میتوانیم از طریق جهت آن متغیرمان تعیین کنیم که آیا متغیرهایمان چه نقشی را دارند که این موضوع در الگوریتم‌های داده کاوی مهم است.

حال فرض می‌کنیم که ما داده هایمان را Read values زدیم، فراخواندیم، متغیرهایمان از نوع input است و حالا میخواهیم یک خروجی بگیریم، چه خروجی میخواهیم بگیریم؟! مثلاً خواستیم که یک نمودار میله ای برای داده هایمان

Subject:
Year. Month. Date. ( )

رسم کنیم، وقتی که میخواهیم نمودار ترسیم کنیم باید وارد کرنر graph شویم، و تعیین کنیم نوع نمودار را در آن جا،

آیا نمودار ما میخواهیم برحسب متغیری که داریم، یعنی کیفی است یا کمی، در مورد ما الان Distribution

را انتخاب کرده ایم، سومین نمودار که ویژه ای متغیرهای کیفی است، این متغیرهای کیفی یا متغیرهایی که از جنس

Flag هستند یا متغیرهایی که از جنس Discrate هستند یا متغیرهایی که از جنس set هستند! متغیرهایی

که از نوع ordinal set هستند، ما میتوانیم برای این نوع متغیرهایمان، نمودار بکشیم با به ازای این کرنر استفاده

کنیم. می توانیم این کرنر را هم مثل کرنر source که به سه طریق میتوانستیم وارد صفحه بکنیم، وارد کنیم تا زمانی که

جریان source دارد، ایمان را به این نمودار وصل نکنیم، امکان اینکه این نمودار بتواند متغیرهایی که

در فایل داده های ما هست را تشخیص بدهد وجود ندارد. چگونه اتصال را برقرار کنیم؟ میتوانیم از طریق

غلتک وسط موس بیاییم روی نود منبع داده هایمان قرار بگیریم و غلتک را کلیک کنیم و بکشیم تا روی آن کرنر

که امکان قرار بگیرد تا جریان برقرار شود. یک راه دیگرش این است که ما بیاییم روی کره صدا source های

کلیک کنیم، کلید F2 را بزنیم، وقتی کلید F2 را بزنیم یک علامت شش ضلعی با یک نامش ایجاد میشود و

میاییم روی کرنر مقصدمان کلیک میکنیم، اتصال از نقطه شروع به آن مقصدی که در نظرمان است اگر

اجازه برقرار کردن اتصال وجود داشته باشد را به ما خواهد داد. ما میتوانیم به صورت خودکار و این که

همان وقت که در پایین کلیک کردیم، اتصال اتوماتیک داده برقرار شود، این امکان نیز وجود دارد.

Subject:
Year. Month. Date. ( )

چطوره؟ قبل از اینکه ما بریم کلیک و Drag کنیم، بریم روی گزینه مبدأمان کلیک کنیم وقتی که کلیک
کردیم از گزینه گراف، وقتی دابل کلیک کنیم بر روی گزینه دلخواهمان، اتوماتیک وارد لینک بین
گره مبدأ و گره مقصد ما برقرار میشود. ما فقط اتصال را برقرار کردیم و هنوز متغیری را نتوانسته ایم وارد
بخوانیم، حالا که اتصال برقرار شد، روی مثلث که با شماره 1 نشان داده شده، دابل کلیک کرده تا پنجره ایی باز
شود، همان طور که ملاحظه میکنید (ص ۵) روی پنجره باز شده، علامت ؟ است، علامت ؟ نشان
دهنده این است که هنوز متغیری را load نکردیم. در این پنجره در کنار Feild روی آن گزینه اش
کلیک میکنیم و نام متغیری را که میخواهیم برایش گراف را ترسیم کنیم می آییم انتخاب میکنیم. به عنوان مثال
وقتی روی این گزینه کلیک کردیم، فیلدهایی را به من نشان خواهد داد که ترجیحاً کنی باشد یعنی Flag
set، ordinal set، discrate. بعد از انتخاب متغیر مورد نظر، اگر بخواهیم صرفاً نمودار
مربوط به این متغیر را ببینیم، باید Execute را بزنیم، و اگر فقط ok کنیم هیچ اتفاقی نمی افتد و
فقط داده دستور گرفته خواهد بود. انتظار داریم که خروجی مورد نظر بعد از Execute کردن در گزینه
output آن بالا سمت راست صفحه نرم افزار، قرار بگیرد. ممکن است که ما بخواهیم ۲ تا نمودار داشته باشیم،
ممکن است که ما بخواهیم دو تا متغیر را در مقابل هم بگیریم، یعنی scater pliot را بخواهیم رسم کنیم، از
گزینه گراف Pliot را انتخاب میکنیم، همان طور که ملاحظه میکنید ص 6 از یک منبع داده ها مانند فایل دیتا ست
نیز کشیده ایم.

Subject:
Year. Month. Date. ( )

هر گزینه‌ای این امکان برایش وجود ندارد، که در رسم plot، پنجره‌ای که باز میشه با دابل کلیک، به صورت ?T.? است چون که قبلش برای آن مشخص شده است. پس باید برویم برایش x و y را مشخص کنیم، در این پنجره‌ای که باز شده، هنوز اسمی نیامده، که یعنی باید مشخص کنیم کدام متغیر در مقابل کدام متغیر دیگر میخواهیم قرار بگیرد. برای انتخاب متغیرها، روی کرکره‌های موجود، کلیک کرده و لیست متغیرهایمان می‌آید و از بین آن‌ها، ما را انتخاب می‌کنیم و بعد از اینکه متغیرها انتخاب شد، اگر ok کنیم، هیچ اتفاقی نمی‌افتد و اگر Apply نیز کنیم، باز هم اتفاقی نخواهد افتاد، یعنی عمل میکند اما خروجی را به ما نشان نخواهد داد تا زمانی که Execute را اجرا کنیم، یعنی به ما خروجی نخواهد داد. که در ص 7 رابطه دو تا متغیر Na و k برای مثال نشان داده شده است. چه زمانی از Distribution استفاده میکردیم؟ الان Na و k مقادیر کمی هستند، Distribution ها متغیر کیفی است، قرار نیست بین دو تا گزینه Distribution ها و Plot ها کاری انجام شود، این عمل این است که ما بخواهیم از یک منبع داده‌ها، یک جدول توزیع داده‌ها را بدهد، یک بار برایمان Scater Plot را بکشد، یک بار برایمان ..... در ص 8 همان طور که ملاحظه می‌کنید، با این تابع مربوط به کار می‌توانیم ببینیم که وقتی است که Run یا Execute می‌کنیم این اتفاق خواهد افتاد، و در نهایت در کرکره Output بالا سمت راست، نمودارها و جداول خروجی قرار می‌گیرد و باید برویم بر روی Double click کنیم تا نشان ما را بدهد. ممکن است که ما بخواهیم یک محاسباتی را هم برایمان انجام دهد، همان طور که

Subject:
Year. Month. Date. ( )

داریم ملاحظه میکنید ما از سیستم میخواهیم که Na را تقسیم بر K کند. یعنی یک متغیر جدید برایمان بسازد.
که از تقسیم این دو متغیر ایجاد شده باشد، قرار شد که ما یک فیلد جدید را بسازیم. برای این کار ابتدا باید
از گروه Field option آن گزینه Drive یا مشتق شده را بیاوریم، که همان طور که ملاحظه می‌کنید
در گروه Drive، shortcut اش یک علامت + است یعنی چند تا فایلش داریم یک فایلش اضافه
شده، یعنی هر فایلش ما بیان گر یک متغیر است، پس سیستم این امکان را دارد که بیاید و یک متغیر جدید
را برای ما اجرا کند. پس از دار دکمه گروه 2 صفحه، اتصالش را مانند حالت قبل با Data برقرار
می‌کنیم و بر روی آن Double click کرده تا پنجره مربوطه اش باز شود. یک قسمت داریم
نام Formol، روی گزینه ی expression اش تیک میزنیم وارد پنجره ای میشویم و ما فرمول را
مینویسیم و سپس چک می‌کنیم و این متغیر را وقتی ok می‌کنیم، از Drive 2 بعد سیستم ما یک متغیری
دارد به نام Drive1 اضافه شده 2 فایل داده ها که میتوانیم در محاسباتمان از آن استفاده
کنیم. (اسلاید 9)

جلسه ۴ 92/07/14

اسلاید 8 » در قسمت Drive as چند حالت وجود دارد که آیا پیش شرط ما Formolu هست
ما میتوانیم از سیستم بخواهیم که یک فرمولی را برای ما در نظر بگیرید، محاسبات بر مبنای آن فرمول

Subject:
Year. Month. Date. ( )

انجام دهد، و مقادیرش را در فایل جدیدمان ذخیره کند، پشت صفحه یک علامت ماشین حساب وجود هست

فرمولا هست، اگر ماوس روی آن کلیک کنیم، پنجره ۴ باز میشود. در این Drive as گزینه Field

Type هست، که میگوید نوع متغیری را که میخواهد حساب کنم، میخواهیم چی در مقابلش بگذاریم که در اینجا

نوع متغیرمان را براساس Data ای که داریم، اگر داده ما کلی باشد قاعدتاً در اینجا هم نوع کلی

را خواهیم داشت یعنی Rang خواهد بود. میتوانیم این را خودمان از قبل تعیین کنیم، یا میتوانیم از سیستم

بخواهیم بصورت پیش فرض محاسبه ما هر چه بود براساس آن کارهایمان را انجام دهد، در اینجا ما Default

گذاشتیم که همان پیش فرض است. حال روی پنجره ماشین حساب کلیک کرده، تا پنجره ۴ باز شود، هدف

ما این است که یک متغیر جدیدی ایجاد بکنیم که از تقسیم متغیر Na بر روی k ایجاد شده. از قسمت

پایین که لیست متغیرهای ما است، متغیر Na را انتخاب میکنیم و با فلش منتقل اش میکنیم بالای صفحه.

عملگرهای جبری، ریاضی و... هر کدام که نیاز است را انتخاب میکنیم و در نهایت متغیر بعدی را به شکل اول انتخاب

میکنیم. میخواهم بدانم که این فرمولی که نوشتم ایراد دارد یا خیر؟! میتوانیم از گزینه check استفاده کنیم.

و اگر فرمول ما دارای ایراد میکنم در محاسبه اش، مشکلی وجود داشته باشد، نرم افزار هم این جا به من خطا

خواهد داد. در قسمت سمت چپ، لیستی از توابع مختلفی که سیستم میتواند برمبنای آن ها، محاسباتی را برای

ما انجام دهد وجود دارد. اگر محاسبه مشکلی نداشت ok را میزنیم تا وارد صفحه بعدی شویم. همان طور

Subject:
Year. Month. Date. ( )

که ملاحظه میکنید، فرمولی که ما خواستیم برایمان محاسبه کند، در قسمت Formula آمده، ما میتوانیم Drive

یک را انتخاب کنیم و بر مبنای آن از سیستم بخواهیم که این محاسبه را برایمان انجام دهد. ممکن است که مثلاً

ما بخواهیم دو تا فرمول با این سیستم بدهیم، در هر بار مسیری را که در یک قفسه داریم در نظر میگیریم، هر بار که قفسه را

ما داریم میاوریم، امکان محاسبه یک فرمول بیشتر برای ما نیست، حتماً از ای هر بار قفسه‌ای که میخواهیم بیاریم،

باید بیاییم یکبار این گره را حتماً اجرا کنیم تا محاسبات را برایمان انجام دهد. حال میخواهیم هیستوگرام

مربوط به این قفسه جدید را ببینیم، بعد از این گره جدید، ما از فدی Graph، نود histogram را به صفحه

فعال میکنیم، و جرایش را به قفسه جدید مان وصل میکنیم. فعلاً تا این جا کار هیچ خروجی نداریم چون

دستور execute را اجرا نکردیم. (اسلاید 12)

میخواهیم برایم داده‌ایی را که وارد کردیم تا به حال، درست وارد سیستم شده است یا خیر، برای این کار از گره output

Table را انتخاب میکنیم (اسلاید 13) و از گره منبع داده‌ها یا از گره DRUG، جریان داده‌ها را

از منبع داده‌ها، میخواهیم یک Table ایی، یک Data Flat ی ایی را برایمان نمایش دهد.

کارش چیست؟ امکان تغییر در این Table وجود ندارد، فقط کاربرد این است که به ما نمایش

می‌دهد در آید این Data به این صورت هستند. همچنین از سیستم میخواهیم که یکسری آمارهای

خلاصه مثلاً شامل نمودارهای آمارهای مرکزی، آزمون‌های آماری را برای ما اجرا کند، ما وقتی

Subject:
Year. Month. Date. ( )

از بین این کار را میخواهیم، میتوانیم وارد همان گزینه output شویم و از آنجا بخواهیم نود Data Audit

را برای ما بیاورد، این نود نه تنها آمارهای توصیفی را به ما می‌دهد امکان اینکه بیاید مجموع داده‌ها، چولگی

پیچیدگی، انحراف استاندارد، مال مربوطه را به ما بدهد این امکان وجود دارد.

ما هدر گره یا نودی را که خواندیم، اتصال آن را به جریان داده‌ها وصل میکنیم، باید روی آن Double click کنیم تا

پنجره مربوطه باز شود، ~~حالا~~ الان این پنجره Default، همه‌ی متغیرها را آمار خلاصه‌ی آن را به ما می‌دهد، (صفحه ۱۳)

ممکن است که ما بخواهیم که ~~[illegible]~~ تعداد خاصی از متغیرها را بخواهیم، الان فایل ما چند تا متغیر دارد؟ در قسمت ۳

نوشته Field 8، یعنی متغیر که ۸ تا ستون در مجموعه‌ی داده‌های DRUG1n ما ۸ تا الان وجود دارد.

الان در پنجره setting هستیم، میخواهیم آمارها و اطلاعات بیشتری را به ما بدهد. اگر روی گزینه‌ی

quality کلیک کنیم، قبل از اینکه این پنجره را ok کنیم بسته شود، یک اطلاعات دیگری را به ما ارائه می‌کند.

یعنی وقتی ما از setting بروی quality کلیک کنیم، قسمت اول مقادیر بی‌پاسخ را به ما می‌دهد، ردیف

اول تعداد مشاهدات، تعداد رکوردها را در حقیقت براساس داده‌های معتبر! valid valuse برای ما در نظر

می‌گیرد و اگر داده یا مقادیر نامعتبری برسد out میکند و میگوید آن مقادیر بی‌پاسخ را از تحلیل کنار میگذارد.

مقادیر بی‌پاسخ به دلایلی یا جواب ندادند یا ما خودمان به سیستم معرفی کنیم که اینها بی‌اثر هستند نمی‌توانند در تحلیل

مورد محاسبه قرار بگیرند. در پاسخ سوم و قسمت بعدی که ما داریم ملاحظه میکنیم، اطلاعاتی که وجود دارد

Subject:
Year. Month. Date. ( )

با توجه به این که آیا مشاهداتی که ما داریم در نظر میگیریم آیا این مشاهدات، مشاهدات outlier است؟! آیا مقدارهایی که ما داریم در نظر میگیریم، مقدار صحیح هستند. ما مشاهداتمان میتوانند به دو دسته تقسیم شود، مشاهداتی که outliers هستند و مشاهداتی که Extreme هستند، Extreme ها یعنی خیلی دور از مرکز داده هایمان فاصله دارند، الان در همین بکس ۳ که ملاحظه میکنید ص ۱۴ Detectio Method چطوری ما تشخیص دهیم که آیا داده های ما outliers هستند یا خیر؟! در گزینه اول نوشته standard devation from mean میشه اگر داده های ما با استفاده از اطلاعات انحراف معیار، (پس کی به درد می خورد؟ وقتی که داده های ما کمی هستند) اگر داده های ما به اندازه ۳ انحراف معیار از سمت راست و به اندازه ۳ انحراف معیار از سمت چپ یعنی همان 3σ فاصله داشته باشند، داده هایمان معمولی هستند، اگر داده های ما بین ۳ تا ۵ انحراف معیار از میانگین ها فاصله داشته باشند، به عنوان داده های outliers شناسایی میشوند. اگر مشاهده ای بیرون از فضای ۵ انحراف معیار یعنی میانگین + ۵ انحراف معیار، داده های من بزرگتر از این مقادیر یا کوچکتر از میانگین - ۵ انحراف معیار باشند، به این داده ها میگوییم داده های صحیح، یعنی بهتر است که این داده ها را حذف کنیم. ممکن است که ما بخواهیم داده هایمان را به جای اینکه براساس انحراف معیار از میانگین مورد مطالعه قرار دهیم، ممکن است داده های ما کمی نباشند که ما بخواهیم میانگین و انحراف معیارشان را محاسبه کنیم.

Subject:
Year. Month. Date. ( )

چطور تشخیص دهیم که آیا داده هایمان پرت هست یا خیر؟ ما میتوانیم از ترم Interquartile ranges from استفاده کنیم. میتوانیم بگوییم اگر داده های ما، یعنی داده های میان چارکی، یعنی فاصله بین چارک اول تا چارک سوم از چاه عددی چند است، به اندازه 1.5 برابر این انحراف معیار، اگر داده ها در فاصله 1.5 برابر کمتر بیشتر از این چارک اول و چارک سوم باشند، ما میگوییم با اینها داده مشکلی ندارند، اگر بیشتر از این 1.5 برابر دامنه میان چارکی فاصله داشته باشند، یعنی از چارک اول کوچکتر و از چارک سوم بزرگتر، اگر این فاصله بین 1.5 تا 3 برابر دامنه میان چارکی باشد، میگوییم داده ها، outliers هستند. ما به طور اتوماتیک داده ها را محاسبه میکنیم، خودش میره چارک اول، دوم، سوم را محاسبه میکند، همان طور که نوشتم تم انحراف معیار و میانگین را خود تم محاسبه میکند، خود چارک ها را هم خود تم محاسبه میکند و مقادیر بین 1.5 تا 3 انحراف معیار را outliers و بیشتر از 3 دامنه میان چارکی را پرت در نظر میگیرد که توصیه میکنیم باید از تحلیل خارج شود. گزینه output یک گزینه در این جا داریم، به نام گزینه output، اگر کلیک کنیم، گزینه دوست دارید خروجی مربوط به این نود با چه نامی، در چه مسیری، در چه فرمتی تایپ شود، میتوانیم روی این گزینه کلیک کرده، در قسمت output name یک قسمت داریم Auto داریم، یک قسمت custom، در قسمت اتوماتیکش به صورت خودکار، یک نامی روی خروجی مان میگذارد، اگر خودمان بخواهیم اسم دهیم، میتوانیم در custom اسم مورد دلخواه مان را بدهیم، اگر بخواهیم Data مان را در مسیر خاصی برایمان save کند

Subject:
Year. Month. Date. ( )

در گزینه File names ، Brows میکنیم و با فرمتی که مدنظرمان است و درست داریم Data های مربوطه را با آن فرمت مربوطه می‌آید، ذخیره میکنید به شرطی اینکه ما از آن خواسته باشیم، output مان را در یک فایل دیگر ذخیره کند، اما ذخیره‌ای این فایل در محیط Excel، در محیط Notpat در محیط HTML و در محیط خود clementin دارد. اگر ما بیاییم Apply و ok کنیم، برمیگردیم در صفحه جراحی‌مان. ما هنوز اطلاعاتی که در این صفحه است را اجرا نکردیم فقط ازش خواستیم که این فعالیت‌ها، این کارها را برایمان انجام دهد. اگر ما بیاییم از اینستا بخواهیم که وقتی میخواهد برایمان کار انجام دهد، از این گزینه Paly استفاده میشود که تمام گرافها، محاسباتی که در گره Drag بوده Audit ایی که در قسمت Data Audit است، Table ایی که ما خواستیم در اینجا رسم کنیم همه‌شان هم‌زمان با هم اجرا میشود و نتایج این‌ها در قسمت مدیریت OutPut ذخیره میشود و ظاهر میشود. البته این خروجی هنوزمان باید نتایجش که میاید، پنجره‌های مربوطه خروجی‌ها هم باز میشود. ما اگر پنجره‌ها را نیز close کنیم، اطلاعاتمان shortcut اش در قسمت مدیریت output است و هر وقت روی shortcut کلیک کنیم، پنجره مربوطه برای خودمان ظاهر میشود. حالا ما درست نداریم که همه‌ی اطلاعاتمان هنوزمان با هم Run شدند، برای این منظور کنار علامت Play یک شکل منفی است که گوشه‌اش علامت play هست، این زمانی ازش استفاده میکنیم که مثلاً

Subject:
Year. Month. Date. ( )

یا چند گره که مدنظرمان است را select کنیم و از بین گزینه‌ها چندتایی را برای ما اجرا کند که ما خودمان select اش کردیم.

ص 16

فایل بعدی (pdf 3.2) در این فایل ما آمدیم روی قسمت #2 Histogram of Na_to_K کلیک کرده، این به ما اطلاعاتی را به ما می‌دهد، ما خواسته بودیم هیستوگرام را جدایی از این مراحل برای ما ترسیم کند، البته این نمودار که میدهد بر حسب جنسیت male و Female است، به تفکیک نوع دارو، دارد A و B و C و ...

و در عین حال هیستوگرام مربوط به Na_to_k را برای هرکدام از اینها کشیده است. اگر ملاحظه کنید در صفحه 17 به اندازه پاسخ‌ها، وقتی داروی ما از y خصوصاً برای زن و مرد مشخص شده است.

همچنین اگر از مدیریت output روی تب نمودار ما کلیک کنیم، آن را برایش خواهد داد.

زمانی که ما گزینه Distribution را برای داده‌هایمان کشیدیم، این گزینه نه تنها نمودار میله‌ای را برای داروها می‌دهد بلکه جدول توزیع فراوانی اش را در قسمت جلوی نمودارها نیز می‌دهد ص 18. ما تا حالا نمودار داده‌هایمان را دیدیم بودیم، ما وقتی که Table را Run کردیم، اطلاعات مربوط به آن، در پنجره ص 19 فقط نمی‌توانیم ببینیم، اگر mouse را روی هر متغیر نگه داریم، ملاحظه می‌کنید که اطلاعات مربوط به بر حسب متغیرمان در آن جا ظاهر داده می‌شود. اگر از Plot استفاده کنیم ص ۲۰ که می‌توانیم داده‌ها را بخوانی پراکندگی شان را در صفحه ببینید، محور تا نیم اطلاعاتش از 0.2 تا 0.8 و سمت Na از 0 تا 40 است.

Subject:
Year. Month. Date. ( )

تا این جا اطلاعات خاصی هیچ یک از متغیرها نداریم، ولی وقتی که ما $\frac{k}{Na}$ کردیم، اطلاع مناسب تری را میتوان از آن استنباط کرد، در حقیقت این محاسبات را کسی که نسبت به آن Business شناخت دارد به ما می گوید، Business understanding یکی از ویژگی هایش این است که حداقل ما در آن تیم کاری که داریم، میتوانیم از افراد بخواهیم که کمک بخواهیم که چه فیلد ها، چه متغیرهایی را ما می توانیم داشته باشیم.

ما وقتی که Data Audit را اجرا کردیم، این خود پیغام ایی را که به ما می دهد، در ص 21 نشان داده شده است، ستون اولش نام فیلدها را نشان می دهد، ستون دوم نوع متغیرها را نشان می دهد.

A الفابتیک را نشان می دهد که در کنار نوع متغیرها قرار میگیرد، مشاهده میکنید که کنار نوع متغیر $\frac{Na}{k}$ یک علامت ؟ قرار گرفته، یعنی وقتی که ما می خواستیم متغیرمان را بسازیم، گفتیم پیش شرط بزار، هنوز هیچ وقت عددی به سیستم در نظر نگرفتیم ولی وقتی که به نمودارش نگاه میکنیم، می فهمیدیم که هر جا که داده های ما کمی بوده، این نمودارها ستون هایشان به هم چسبیده، و هر جا اسمی بوده ستون ها فاصله دارند...

ستون دیگری که ملاحظه می کنید، ستون Type است، نوع متغیرها (ستون سوم) متغیرسن که نوع اش عددی بود، نوع اش Range است، یعنی داده های ما از جنس فاصله ای و نسبتی است قیاسش، جنسیت اینجا Discarate در نظر گرفته، اگر زنانی که سیستم جنسیت را فرا می خوانند ما جنسیت را در حالتی به سیستم تعریف میکردیم، (Flag) اینجا به حالت Flag به ما نشان می دادند...

Zarrin Barg

Subject:
Year. Month. Date. ( )

ستون بعدی مینیمم را به ما نشان می‌دهد، صحبتی که مقادیر ما عددی است، کمترین مقدار به عنوان مینیمم و هر جا که عددی نیست، قابل محاسبه نبوده و اطلاعی به ما نمی‌دهد. اطلاعات بعدی، ستون بعدی Max - Min یعنی Rang دامنه تغییرات،

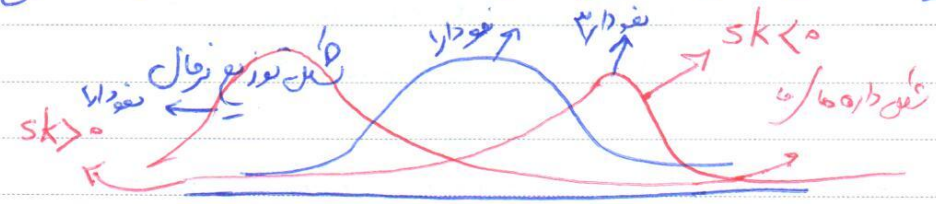
ماکزیمم را نشان می‌دهد، بعدی sum مجموع، بعدی Mean میانگین، انحراف معیار میانگین‌ها یعنی انحراف معیار استاندارد میانگین‌ها.

ستون بعدی، ستون مربوط به انحراف معیار است، ستون بعدی واریانس است، ستون بعدی چولگی است.

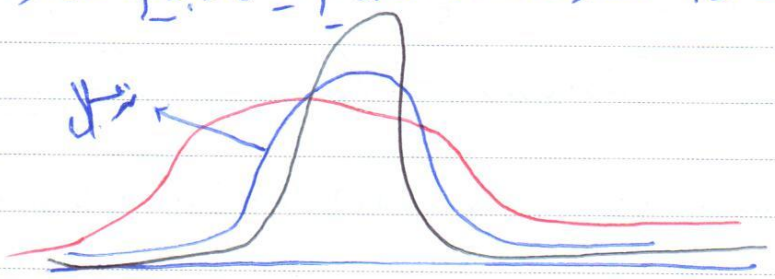
ستون بعدی انحراف استاندارد چولگی است، همچنین اطلاعات مربوط به کشیدگی و انحراف معیارش و همینطور تعداد داده‌ها را در اختیار ما خواهد داد. یادآوری کنم که چولگی و کشیدگی زمانی معتبر خواهد بود که ما صدتایی مقایسه داده‌ها با داده‌های توزیع نرمال باشد. لوزی‌های کنار فیلد‌ها بیان‌گر این هستند که متغیرهای ما عددی هستند. ص ۲۳: ادامه جدول قبل است که kurtosis و Unique و Valid بودن را نشان می‌دهد، ما وقتی که کشیدگی را مطرح می‌کنیم، این kurtosis فقط در مقام مقایسه با داده‌های نرمال مناسب است. مابقیه جاها به درد ما نمی‌خورد. در قسمت Unique وقتی که ما متغیر کیفی داریم، متغیرهای از جنس عددی نیستند، به ما می‌گوید که چند تا کد در مجموع داده‌های ما وجود دارد، مثلاً در قسمت شماره خونه ۳ تا کد وجود دارد: ۱ بالا، پایین، متوسط، در قسمت جنسیت ۲ تا کد وجود دارد: زن و مرد. در قسمت Valid یعنی چند تا Data ی معتبر در مجموعه فایل‌های ما وجود دارد، در اینجا ۲۰۰ تا مشاهده در این فایل وجود دارد. چولگی یعنی کجی (skewiness)، kurtosis یعنی کشیدگی. اگر ما شکل

Subject:
Year. Month. Date. ( )

زنگوله‌ای توزیع نرمال را در نظر بگیریم. اگر توزیع داده‌های ما نسبت به شکل زنگوله‌ای، انحراف داشته باشد، می‌گوییم داده‌هایمان یا انحراف به چپ دارد یا انحراف به راست دارد. شکل زیر شکل توزیع نرمال است:



فرض کنیم که داده‌هایمان فرمش نرمال است، ما توزیع داده‌هایمان ممکن است به این صورت باشد، یعنی خروجی اکثر داده‌هایمان به صورت نرمال باشد. در این نمودار که ما کشیدیم نسبت به نمودار اول کج شده است به سمت چپ، اینجا می‌گوییم چولگی اتفاق افتاده است (چولگی مثبت). ممکن است توزیع داده‌های ما به صورت نمودار ۳ باشد، یعنی اغلب داده‌های ما در یک سمت توزیع باشند، ما برای این حالت می‌گوییم sk<0 (چولگی منفی) یعنی نسبت به فرم نرمال اصلی‌مان، انحراف داریم، انحراف‌مان به سمت چپ، انحراف‌مان به سمت راست. حالا اگر بخواهیم کشیدگی را بگوییم. نمودارهای زیر را در نظر می‌گیریم.



کشیدگی

نمودار قرمز: شبیه نرمال است ولی پایین‌تر از توزیع نرمال است می‌گوییم k<0 (داده‌ها پست‌تر از توزیع نرمال هستند)

Zarrin Barg

Subject:
Year. Month. Date. ( )

یعنی درصد کمتر حول و حوش میانگین نسبت به توزیع نرمال جمع شده اند، ولی ممکن است که داده ها کاج

خم نمودار خمشی باشند یعنی درصد خیلی بیشتر از داده ها حول و حوش مثلاً توزیع نرمال جمع شده باشند، اگر درصد

بیشتر از داده ها در این رنج باشند، یعنی بیشتر از 68% در این قسمت، ما میگوییم که توزیع دارای کشیده تر

است یعنی $k>0$ ولی k توزیع نرمال چند است $k=0$ است، چولگی توزیع نرمال هم صفر است

$sk=0$، پس چولگی و کشیدگی فقط وقتی که ما داریم در مقام توزیع نرمال داریم کار میکنیم میتوانیم؟ ما کمک

کنند، ما در این جدول ص 23 چولگی و کشیدگی و ... را میتوانیم ببینیم، در نمودار بعدی ما میتوانیم در قسمت

quality هر چیزی که در کنار quality مدنی کرده بودیم، اینجا هم میتوانیم ببینیم، ما نام متغیرهایی

را داریم، Type آنشان را داریم، در رابطه با outlier ها، آیا ما مشاهدات outlier داریم که در اینجا

تعدادش صفر است، آیا مشاهده Extreme داریم، که در اینجا صفر است، مشاهده outlier و

Extrem کجاها و برای چه نوع متغیرهایی ما میتوانیم بگوییم مشاهده outlier و Extrem است؟

جایی که متغیرهای ما کمی یا حداقل کیفی ترتیبی باشند، پس ما برای متغیرهای اسمی نمی توانیم این دو را

مشخص کنیم. اگر outlier و Extrem داشتیم، اینجا در مقدار Action ما میگفتیم که چکار باید

با این نوع داده ها بکنیم؟ Delete کنیم یا صفر شک کنیم یا جایش ... . اگر قرار است جایش کنیم از چه روشی استفاده

کنند، چطوری جایش کنند؟! ما در قسمت missing imputation میتوانیم نحوه ای برخورد با مشاهدات بی پاسخ

را بگوییم.

haeri.stat@gmail.com

Subject:
Year. Month. Date. ( )

درقسمت مدل از چه روشی برای برخورد با این بی پاسخی ها استفاده میکند، وقتی که ما مدل خاصی را در نظر میگیریم،
و میخواهیم سیستم برایمان معرفی کند، به ویژه وقتی که به فرم شرطی دستوری را به سیستم می دهیم، اینجا میتوانیم
مدلش را تعیین کنیم، که حالا مدل ما میتواند Fix باشد، میتواند Random باشد، میتواند براساس یک
الگوریتم خاصی باشد، و... . ستون بعدی: چند درصد داده های ما کامل است، یعنی نه Missing است نه
بهتر است، که در اینجا 100٪ داده های ما کامل است. چند تا مشاهده داشتیم در ستون بعدی گفته که در
اینجا 200 جمله است. ممکن است که بخواهیم اطلاعاتمان را save کنیم، ما اصلاً خروجی را نمی خواهیم، این
stearm یا اجرایی که در این صفحه است را میخواهیم save کنیم، میدریم گزینه File ⇐ save stream
ممکن است که بخواهیم همه خروجی ها، output ها و نتایج و هر کاری که کردیم را بخواهیم save کنیم،
درقسمت project باید بیاییم ⇐ save project. اگر بخواهیم پروژه جدیدی را سیستم برایمان در نظر
بگیرد New project را می زنیم. فایل class_4: همان طور که قبلاً گفتیم ممکن است وقتی
(3.1)
که ما بایستی Data کار می کنیم، داده ها واقعی باشند. در صفحه 5 فایل قبلی ما آوردیم در یک
Drive و از سیستم خواستیم که از روی قسمت Nack یک فرمول جدید را برای ما ارائه بدهد. درقسمت
فرمول آن نوشته، ما گزینه ی تیک دار به نام Flag، گزینه ی تیک دار به نام set، گزینه ی
تیک دار به نام state ..... در کاری که ما انجام می دهیم اگر نیاز به فرمول نویسی داشته باشیم

Subject:
Year. Month. Date. ( )

بر مبنای آن فرمول را در قسمت Formula می‌نویسیم. ممکن است کسی Flag را انتخاب کند. اگر کسی
در قسمت Drive as وقتی کلیک میکند Flag را بزند، جواب خروجی فرمول باید درست یا غلط باشد،
یعنی ما باید یک شرطی را ایجاد کنیم، یعنی در قسمت فرمولا اگر Flag را بزنیم انتظار داریم جوابمون یا درست بشه یا
غلط. مثلاً اگر $\frac{a+c}{k}$ ما کوچکتر از 5 باشد ok کنیم در این حالت کوچکتر از 5 ها را True میکند
والا False میکند. ممکن است از سیستم بخواهیم به جای Flag، set را برایمان تعریف کند، ما میتوانیم؟
سیستم بگوییم دسته‌بندی کن اگر سن ما به عنوان مثال کوچکتر از ۲۰ بود، سیستم بیاید در کاری که میخواهد
انجام دهد، کد ابجدی، در set file اش میزنیم کد ۱، دوباره شرط بزاریم که اگر ۲۰ < سن یا مساوی ۲۰
بود و مجدداً ۳۰ > سن بیا label ای که میخواهی برایش قرار بدهی، یا کدی که میخواهی برایش قرار بدی کد ۲
بده و در ردیف بعدی به عنوان مثال اگر بگوییم ۳۰ =< سن بود، کد ۳ بزن، خروجی که سیستم برای ما
استخراج خواهد کرد، خروجی خواهد بود که شامل یک متغیر جدید خواهد بود که در فایل ما خواهد آمد و این خروجی
متغیر جدید شامل دهرقمی که دارد سه تا سطر دارد، مثلاً کد ۱، کد ۲، کد ۳، کدهایی هستند که برایش
تعریف کردیم، پس ما میتوانیم به نوعی از سیستم بخواهیم داده‌هایمان را دسته‌بندی کند. یک حالت دیگر
هم داریم به نام stats که در ۱۰ حالت دارد و آنها را از state استفاده کنیم، متغیر جدیدی که
سیستم به ما خواهد داد این است که به عنوان مثال: اگر متغیر ۷۰ =< سن بود، در قسمت on از این
شرط استفاده کرده و در قسمت off اش را رها می‌کنیم. پس نتیجه‌ای که در state به ما می‌دهد یک متغیر

Subject: http://www.red-gate.com/products/sql-development/sql-data-generator ⇐ (با این نرم افزار میتوانیم Data تولید کنیم)

Year. Month. Date. ( )

در حالتی است که هر کدام از این حالتها یک شرط خاصی را خواهد داشت. یک حالت دیگر داریم در قسمت Drive as ، به عنوان count ، کارش این است که می آید وقتی ما متغیرها را بهش معرفی می کنیم مجموع متغیرها یا متغیری که ما به آن می دهیم، مثلاً می گوید اگر کوچکتر یا بزرگتر از یک مقدار خاص باشد، کارش این است که آن مقادیر بزرگ یا کوچکتر را صرفاً برای ما می شمارد.

جلسه ۵ 92,07,21

آخرین مطالبی که جلسه قبل داشتیم مربوط میشد به فایل class4_1 آخرین موضوعی که بحث کرده بودیم، یک فایلی بود بنام cars ، این فایل در sample های spss معمولی، وجود دارد. در فایل های spss معمولی ما یک فلدری داریم در مسیر sample اش که میتوانیم این فایل cars را load کنیم و بر مبنای آن، کارهای را که داریم انجام می دهیم را پیش ببریم. بجز از اینکه ما نود مربوط به statics File را فعال کردیم، فایل مربوط یک فایلی هست بنام cars که آن را لود کردیم، اطلاعات مربوط به متغیرهایمان را در گرگره Data می بینیم، در گرگره Filter اگر متغیری را می خواستیم حذفش کنیم این کار را انجام میدادیم، در گرگره Type نوع متغیرهایمان را در صورتی که نیاز باشد تغییر می دهیم و بر مبنای آن اگر متغیر یا متغیرهایی را که ما فراخواندیم، بخواهیم مقادیرش را ببینیم و در گام های بعدی از آن استفاده کنیم، باید Read value را بزنیم تا در این ستون values ها اطلاعات مربوط به متغیرها

Subject:
Year. Month. Date. ( )

مربوطه لود شود، ما ممکن است که در داده هایمان بی پاسخی داشته باشیم. به عنوان مثال: فرض کنید که ما برای متغیر MPG تعدادی از وسایل نقلیه که اطلاعاتشان را مورد مطالعه قرار دادیم، متغیر MPG اطلاعاتش نباشد، اسم فایل بالای پنجره ما آمده، Cars.sav، در sample، clementin و روی ۱۴ فریم.

ما درستون value وقتی که Read value را می زنیم، مقادیر داده هایمان در این فیلد نمایش داده میشود، حالا ممکن است که مقدارهای از متغیرهایی که ما داریم مورد مطالعه قرار می دهیم، در حقیقت blanck باشد، خالی باشد، یا اصلا ما یک code این داده باشیم به عنوان کدی نامعتبر، اگر این اتفاق ها افتاده باشد، آن وقت ما میتوانیم بیاییم و در ستون بعدی که نوشته Missing این بی پاسخی ها را بررسی کنیم، ممکن است وقتی که ما وارد گزینه missing میشویم برای هر متغیر، یعنی اگر ما می خواهیم بگوییم داده هایمان بی پاسخی دارد، مشکل دارد، و ما بخواهیم این کارها را با آن Data انجام بدهیم، برای هر متغیری که مد نظر ما باشد باید بیاییم تعریف کنیم اگر بی پاسخی داریم، تا بتوانیم از آن اطلاعاتش استفاده کنیم، وقتی روی گزینه مربوطه missing کلیک کنید، اطلاعاتی که به ما خواهد داد: on یعنی میتوانیم missing مان را ارزش کنیم، off یعنی میتوانیم missing مان را فراموش کنیم و specify: اگر label ای برای داده هایمان تعریف کرده ایم بیاییم براساس آن specify که انجام داریم و تغییراتی که برای متغیر انجام داریم، بیاییم بی پاسخی را برای سیستم مان تعریف کنیم. به هر جهت ما کاری که ممکن است برای متغیرهایمان بخواهیم انجام بدهیم طوری باشد که ما فقط در صورتی که داده های متغیری

Subject:
Year. Month. Date. ( )

داشته باشند که این متغیر به عنوان مثال بی پاسخ باشد، اطلاعاتش را نداشته باشیم، میخواهیم جایی هی کنیم به
عنوان مثال در ستون MPG تعدادی از خودروها ممکن است، اطلاعاتش را نداشته باشیم، میتوانیم بیاییم
بزرگترین مربوطه کلیک کنیم و برچسب آن تعریف کنیم که مثلا ما در بیاییم در صفحه ۱ برای MPG، Missing
را on کردیم، سیستم وقتی که به این متغیر برسد امکان برایش وجود دارد که باید در این متغیر بررسی کند که
آیا ما برایش چیزی تعریف کرده ایم، در ستون بعدی check، گفتیم با این بی پاسخی ها چکاری بکند یا خیر؟!
وقتی که ما بیستم هیچ اطلاعی نداریم و ازش نخواستیم هیچ کاری انجام دهد، بیستم می آید با بی پاسخی، هرجایی
که اطلاعاتی داشته باشد، استفاده میکند، و هرجایی Null یا Missing value باشد، پیش شرط کلی هست و هرجا
به آن برسد (به Null) آن را رها می کند. سه حالت وجود دارد که ما میگوییم ممکن است بی پاسخی اتفاق
افتاده باشد: به عنوان مثال ممکن است که در متغیر MPG اطلاعی کسی وارد نکرده (خالی است) یعنی
Null گذاشته، این یک نوع بی پاسخی است، ممکن است به عنوان مثال یک نوع خودروی هست بهش گفتیم
که مصرف بنزینش در هر ماه چقدر است؟! مثلا گفتن: ۲ تا ۷، ما گفتیم به طور متوسط یک عددی را میخواهیم نه
یک دامنه ایی از مقادیر و اعداد، پس این هم یک نوع بی پاسخی است. ممکن هم هست که بگوییم خودروی شما
در هر ماه چقدر سوخت مصرف میکند، یک عدد؟ ما به هیچ که معقول نباشد یا خیلی بزرگ باشد یا خیلی کوچک
باشد، ما به عنوان یک فرد متخصص که داریم با این فایل سروکار میزنیم، متوجه میشویم که یک همچین مقداری

Subject:
Year. Month. Date. ( )

وجود ندارد. حالا اگر در Data هایمان بی پاسخی داشتیم باید با آن ها چه کاری بکنیم؟! اگر در گزینه ی check وارد شویم در ستون بعدی، در ستون check آن جا جاهایی را برای ما تعیین می کند که میتوانیم از آن استفاده کنیم.

ص ۲ در پنجره ی MPG values قسمت Null این تیک خورده، که یعنی آن را به عنوان بی پاسخی می شناسد.

اگر white space باشد، یعنی اگر ما بیایم خودمان بگوییم که آن را اصلا اعلامی در سیستم وجود نداشت، ما خودمان یک کد 99 یک کد خاصی تعریف کرده باشیم، مثلا white space را زدیم، ممکن است ما بگوییم که مقادیر مثلا از ۵۰ تا ۹۰ هستش این ها مقادیر بی پاسخ ما هستند و ما میخواهیم این ها را در مجموعه داده هایمان نداشته باشیم، یا اینکه مثلا ممکن است که ما دون لیست متغیرهایی که داریم، اگر این "define blanks" را بزنیم، کدهای خاصی را برای سیستم مان تعریف کنیم، ما در قسمت blank وقتی که تیکش را فعال می کنیم، میایم در ردیف در قسمت missing value

- جایی که دارد، عدد یا اعداد مربوطه را دونه به دونه وارد می کنیم Enter میکنیم، میگوییم مثلا اگر کد 9 را دیدی این کد 9 یک مقدار بی پاسخ است.

اگر در ستون check کلیک کنیم برای اینکه به ما بگوید بی پاسخی ها چه رفتاری داشته باشیم، گزینه های زیر را می بینیم:

ص ۳ None، Nullify، coerce، Discard، Warn، Abort، ما میگوییم که میتوانیم همین جا که این پنجره را باز کردیم و بی پاسخی ها را برای سیستم مان تعریف کردیم، می توانیم بگوییم نحوه برخورد با بی پاسخی ~~اش به چه صورت باشد. وقتی که ما missing sys off باشد، وقتی که ما از سیستم بخواهیم که~~

Subject:
Year. Month. Date. ( )

هیچ کار خاصی انجام نمی‌دهد، یعنی check value ها Null باشد، وقتی به سیستم به این اعداد می‌رسد،
کار خاصی انجام نمی‌دهد، آن‌هایی که Missing تعریف شدند، سیستم‌ها نشان می‌کند و در هیچ تحلیلی
نشان نشان نمی‌دهد، در صورتی که مثل الان که ما انتخاب کردیم Nullify باشد، نرم‌افزار تمام بی‌پاسخی‌ها
که به صورت کدهای خاصی ممکن است Data (ربوط به) باشد، همه را تبدیل به Null یا Missing سیستم می‌کند و در
حقیقت از محاسبات خارجش می‌کند، حالا ممکن است که ما بخواهیم حالت‌های دیگری را استفاده کنیم،
اینجا ما از گزینه coerce استفاده کنیم، در حقیقت ما زمانی که بخواهیم تبدیل سیستم، وقتی که به بی‌پاسخی برمی‌خورد
چکاری انجام دهد؟! اینجا برای ما مهم است که نوع داده‌هایی که انجام می‌دهیم Rang است، order set است،
یا set است؟! در صورتی که ما داده‌هایمان تایپش Range باشد و بی‌پاسخی در داده‌هایمان وجود
داشته باشد، و سیستم بخواهد از متد coerse استفاده کند، وقتی که به محدوده‌های بی‌پاسخی، یعنی وقتی که
به آن Rang های که ما تعریف کردیم، یعنی وقتی به آن کدهایی که خودمان تعریف کردیم، وقتی که به Null می‌رسد
وقتی که به white space می‌رسد، اگر متغیر ما کلی باشد، سیستم می‌آید میانگین داده‌ها را برای داده‌های
که در رنج محدوده‌های داده‌های ما قرار دارد، قرار می‌دهد. یعنی بی‌پاسخی‌هایمان را می‌آید با میانگین جایگزین
می‌کند، نه این که در فیلد داده‌ها شیت را برای ما پر کند، وقتی سیستم می‌خواهد از این مقادیر استفاده کند،
داده‌های کلی که می‌رسد، می‌آید میانگین می‌گیرد، اگر سیستم به داده‌های کلی برخورد به عنوان مثال

Subject:
Year. Month. Date. ( )

مثلاً ما گفته باشیم داده های ما بین صفر تا ۱۰۰ معتبر است، حالا ما داده ایی داریم با 105 ، 110 وقتی که

سیستم به این ۲ تا می رسد، اینها را با عدد ۱۰۰ جایگزین میکند، اگر یک عدد منفی در ستون MPG ما وارد شده باشد،

یعنی کمتر از عددی که در Rang ما است، یعنی الان ما مقادیر ما در این فایل بین 9 تا 47 است، فرض کنید Data هایی هستند که در فایل ما وجود دارند

که ما به سیستم گفته ایم که داده های ما از ۱۰ تا ۴۰ است، حالا که داده ۹ داریم مقادیر کمتر از ۱۰ را می آید با عدد

۱۰ جایگزین می کند. ممکن است که در حقیقت در سیستم که ما داریم با هاش کار می کنیم، داده های ما را با میانگین مقادیر

به عنوان مثال ما بگوییم ما هیچ Rang ایی برای بی پاسخ تعریف نکردیم، حالا بی پاسخ چه کاری میکند سیستم؟!

سیستم اتوماتیک وارد می آید میانگین داده های کلی را برای بی پاسخ ها قرار می دهد. ما ممکن است داده هایی که به

سیستم داریم معرفی میکنیم، به صورت کد باشد، و سیستم به صورتی عددی آن را تشخیص دهد، ؟! اگر داده های اول

خالی باشند به چه روش میانگین می گیریم؟! به عنوان مثال فرض کنید در کلاس ما ۴۰ نفر هستند، ۱۰ نفر نمی آیند

۳۰ نفر دیگر که در کلاس حضور دارند، از اطلاعات افرادی که در کلاس حضور دارند استفاده میکنیم، میانگین اطلاعاتشان

را حساب میکنیم و می گوییم هر کسی که نیامد، میانگین اطلاعات افراد را برای سایر کلاسی که در کلاس نیامدند، جایگزین

می کنیم، آن ها Data شان خالی است، عددی ندارند ما میانگین را برای آن ها قرار می دهیم. اگر داده های

ما از نوع کمی یا Rang باشد، این کار را انجام می دهیم. حالا ممکن است Data هایی که ما داریم از نوع عنوان

مثال تعداد اعضای خانواده باشد، از یک داریم تا ۱۰، مثلاً اگر تعداد اعضای خانواده ای صفر بخورد

Subject:
Year. Month. Date. ( )

پس صفر را بی پاسخ می‌شناسد. مثلاً آن‌هایی که کمتر از یک هستند را به عنوان کد در نظر گرفتیم، مد نظر ما کافی نیست، اینجا اگر داده‌های ما به صورت عددی باشد، پس صفر را به عنوان بی پاسخی میشناسد و صفر را با نزدیک‌ترین عدد معتبرش جابه‌جا میکند، یعنی صفر را با یک جایش را عوض میکند، این واسه زمانی است که ما در specify ما Missing را فعال کنیم، یعنی بگوییم که داده‌هایی که خارج محدوده است بعنوان بی پاسخ می‌شناسیم، داخل محدوده را با هاش کاری نداریم، آن وقت مقدار کمتر از محدوده را با مینیمم و مقادیر بالاتر از محدوده را با ماکزیمم جایگزین میکنیم. در Coerse اگر داده‌های ما از نوع set بود، پس می‌آید نگاه میکند، قاعدتاً در set، ما متغیرمان تغییر کیفی اسمی است یعنی ترتیب برایش مهم نیست، اگر داده‌های ما set باشند، کوچکترین کد (حالا ممکن است ما با A، B، C و D و ... نشان دهیم در این صورت در صورت بی پاسخی همه را با a جایگزین میکند) جایگزین میشود. حالا اگر در داده‌هایمان عدد هم باشد بی پاسخی‌ها را با کوچکترین عدد جایگزین میکند. اگر داده‌های ما از نوع Flag باشد، یعنی دوحالتی باشد، یعنی به صورت صفر و یک باشد، اگر بی پاسخی در این نوع داده‌ها باشد، باز هم میتوانیم به صورت set لحاظ کنیم، یا یک کوچکترین را قرار می‌دهد که از نوع Flase خواهد بود. ممکن است که نوع داده‌های ما Dis card باشد، وقتی که پس به متغیرها می‌رسد که مقدار بی پاسخ دارد، اطلاعات مربوط به آن را، یعنی کلش را، یعنی کل اطلاعاتی که ممکن است از این ردیف در جای

Subject:
Year. Month. Date. ( )

مربوط به رکورد

دیگر در همین مسیر استفاده بشود، آن فرد را میگذارد کنار (کل اطلاعات را خارج میکند). اگر ما گزینه Warn در Check را انتخاب کنیم، سیستم تنها کاری که برای ما انجام می دهد یک warning ای جلوی کاربر ظاهر می شود یعنی یک خطایی میدهد که یعنی داده ما بی پاسخ دارد و لی اگر فقط Null باشد، سیستم تحت هر شرایطی Null را میگذارد کنار. اگر تعریف نکرده باشیم در warn، Null باشد سیستم گذاشتن کنار. در coerse ما تکلیف Null ها را مشخص کردیم، در Nullify چون Null هست سیستم میگذارتش کنار. در Discard چون Null هست، سیستم کل اطلاعاتش را کنار میگذارد. در Abort اگر ما این گزینه را برای بی پاسخی برای سیستم مشخص بکنیم، سیستم همین که به یک نامعتبر یا بی پاسخ برسد، Error میگیرد و میگوید بیا این را برطرف کن بعد آ یعنی ما تا این را رفع نکنیم اجازه ی اینکه ما فرایند مان را ادامه بدهیم، سیستم به ما نخواهد داد. توصیه میشود برای پاسخ به بی پاسخی ها از coerse استفاده شود از اینجا به بعد اگر سیستم مان بخواهد کاری را انجام دهد، برای بی پاسخی هایش تعیین کردیم که چه کاری انجام دهد. حالا می رسیم به ستون Direction، که حالت های مختلفی دارد، نوع متغیرهای ما میتواند از جنس Input، که در ورژن پایین با In نشان داده و در ورژن بالا با Input نشان داده است، میتواند از جنس Target یا output باشد، یعنی متغیر ما میتواند نقش وابسته یا خروجی را به خودش بگیرد، میتواند نقش Boss را داشته باشد یعنی هم نقش Input را در صورت نیاز بازی کند و هم نقش Target را. همچنین

Subject:
Year. Month. Date. ( )

داده‌های ما می‌توانند بدون نقش باشند مثل: شماره دانشجویی افراد، یعنی این شماره نه به درد میخورد که
جمع اش بزنیم، نه واریانس بگیریم، فقط یک کدی است که برای شناسایی به کار می‌رود. در ورژن جدید Direction
داده‌ها Role است. ممکن است متغیر ما از جنس متغیر بخش نبوی باشد یعنی ما بیاییم متغیری را که
تعریف میکنیم بفرماییم که این متغیر سه قسمت است: Train، Test و avaluation، یک دفعه
که به این متغیر می‌رسد، مثلاً بعضی از الگوریتم‌ها برای مان مهم است مثلاً میگوییم داده‌هایی که Test
و Train، avaluation هستند، طوری استفاده از این کند، وقتی که به متغیری می‌رسد که ما گفتیم
نقش پارتیشن را دارد، در قسمت Train از داده‌های Train استفاده خواهد شد فقط، در قسمت
Test فقط از داده‌های Test استفاده خواهد شد و در قسمت avaluation فقط از داده‌های
split
avaluation استفاده خواهد شد. ممکن است که داده‌های ما یکی از متغیرهایشان نقش script
را داشته باشد، یعنی ما میخواهیم برای مثلاً ۴۸ مایشن را یک متغیر split تعریف میکنیم، وظیفه اش این
متغیر split این است که هر کاری که ما میخواهیم انجام دهیم مثلاً ما سال ۹۱ داریم ۹۲ داریم و...
این متغیر split شده است، هر کاری که ما در این گره انجام دهیم، برای سال ۹۰ جدا خروجی خواهد
گرفت، برای سال ۹۱ جدا و برای سال ۹۲ جدا، چرا؟! چون که متغیر ما مثلاً در اینجا year ما نقش split
بوده. ممکن است که نوع یا نقش متغیر ما از نوع wait یا Frequency باشد.

Subject:
Year. Month. Date. ( )

این متغیری که ما باهاش سروکار داریم به عنوان مثال: ما میخواهیم که از متغیرهایی که در این فایل داریم استفاده میکنیم، استفاده کنیم ولی میدانیم که اینها اطلاعات مربوط به خودروهای متفاوت هستند، هر کدام از این خودروها، حالا یک وزن خاصی، یک ضریب خاصی را، مثلاً در نظر صنایع بهشان داده، میخواهیم آن وزن هم در اینها ضرب شود، یعنی متناسب با سهم وزنی که ممکن است متغیر داشته باشد بخواهیم از آن استفاده بکنیم، اینجا از گزینه‌ی Frequency استفاده میکنیم، یعنی متغیر که می‌رسد، هر ردیف، مثلاً اگر در ستون Frequency ما، در ردیف اولش 1000 باشد، ردیف اول را در فایل ما 1000 بار برای خودش تکرار می‌کند، wait در ردیف دوم اگر 50 باشد، ردیف دوم را 50 هزار بار برای خودش تکرار میکند. متغیر میتواند نقش record ID را برایمان بازی کند، که الان دقیق نمیدانیم این چه نقشی خواهد بود، آیا صرف این است که ما متغیری را به عنوان مثال Record ID اش را میزنیم، آیا ID مشاهدات میخواهد توی آن ثبت شود؟ نمیدانیم!!

اگر کسی بارش روی Role متغیرش را تعیین کند، به جز حالت split به جز حالت Frequency، هر جا که لازم شد، از هر الگوریتمی که ما داریم استفاده میکنیم، می‌توانیم از یک متغیر در زمانی که داریم استفاده میکنیم از یک جا هم که نقش آن متغیر را برای ما حفظ کند، ما اولین خروجی که بعد از آمار توصیفی و نمودار میخواهیم بکشیم، میخواهیم داده‌هایمان را با استفاده از الگوریتم k میانگین، باید خوشه‌بندی کنیم. قبل از هر کاری در فایل داده‌های ما، قبل از هر کاری می‌آییم Table اش را ترسیم میکنیم، برای اینکه ببینیم

Subject:
Year. Month. Date. ( )

اصلاً داده های ما به صورت صحیح و مناسب در سیستم ما وارد شده؟!! آیا مشکلی ندارد؟! بعضی وقتها در

Load کردن Data مثلاً ممکن است که ما به ویژه وقتی که داریم به فرمت Notpad یعنی Var.File ، Fix Fil

داریم میخوانیم ، ممکن است اشتباه کرده باشیم ، ممکن است ردیف اول مان ، اسم متغیرها اصلاً نداشته ، میاید

ردیف اول مان را به عنوان نام برای متغیرهایمان در نظر می گیرد ، حالا اصلا ما بریم در split داده ها

و ببینیم اصلا همه چیز چیزی که ما فرا خواندیم ، Data ها هست که ما در نظرمان باشد ، معتبر است؟!

مشکلی ندارد؟! حداقل از لحاظ ظاهری ، بی پاسخی ، Null ای چیزی ندارد؟! پس ما Table را به این

دلیل ترسیم می کنیم که ما ایرادهای احتمالی را ببینیم ، ما میخواهیم الگوریتم k میانگین را برای داده هایمان

ترسیم بکنیم ، الگوریتم k میانگین ، هدفش این است که اگر ما ۱۰۰ نفر به عنوان مثال در کلاس داشته

باشیم ، این تعداد ۱۰۰ نفر را به k دسته تقسیم می کند ، ملاک و مبنایش این است که فاصله مشاهدات

یا افراد را تا یک نقطه ای به عنوان میانگین در نظر میگیرد ، به عنوان مرکز ثقل در نظر میگیرد و ملاک و

معیارش این فواصل است . چرا k میانگین؟ چون ما در روش k میانگین اتفاق اینکه سیستم اتوماتیک

وارد خودش بهترین تعداد خوشه را به ما بدهد ، وجود ندارد ، ما سیستم میگوییم که چند خوشه؟! ما میگوییم

۴ خوشه ، ۵ خوشه و ... پس در الگوریتم k میانگین تنها محدودیتی که میتواند داشته باشد این

است که ما در شرایط کاری به عنوان کاربر خودمان باید بگوییم چند خوشه ، که هر کسی میتواند این

ایراد بگیرد .

Subject:
Year. Month. Date. ( )

ص ۴ آبی روی نود مربوط به k-means ما کلیک کنیم، پنجره‌ای با زیرشاخه‌هایی می‌بینیم، در این پنجره در ستون Field ما با متغیرها یا متغیرهایی که در تحلیل خوشه‌ای می‌خواهیم وارد کنیم را به طور معمول در نظر بگیریم، دو حالت داریم که Use type node settings: اگر این گزینه انتخاب شده باشد، به طور خودکار متغیرهایی که در فایل ما وجود دارد را برای خوشه‌بندی از آن استفاده می‌کند.

Use Custom settings: یعنی ما خودمان می‌خواهیم متغیر یا متغیرهایی که در خوشه‌بندی‌مان نقش داشته باشند، توی سیستم‌مان وارد کنیم که بتوانیم از آن‌ها متغیرها را انتخاب کرد.

یادتان باشد الگوریتم k میانگین تمام متغیرهایش باید از جنس Input باشد، یعنی اگر یکی متغیرش را از جنس output انتخاب کرده باشد یا Frequency یا partition یا ... هر چی به جز Input باشد، در k میانگین محاسبه نخواهد شد، فقط Input ها وارد می‌شود به طور خودکار. یادتان باشد که ما ممکن است در برنامه‌های آماری، وقتی که می‌خواهیم از الگوریتم k میانگین استفاده کنیم، صرفاً متغیرهای کمی را لحاظ کنیم و بگوییم که چون میانگین ویژه متغیرهای کمی است، متغیرهای کیفی که نمی‌توان میانگین را برایشان محاسبه کرد، ولی اینجا ما روش‌هایی را داریم که وقتی که می‌خواهیم فاصله‌ای بین افراد را محاسبه کنیم، فاصله بین متغیر کمی و کیفی، فاصله بین متغیر کیفی کیفی و سایر متغیرها را بخواهیم استفاده کنیم، اینها تعبیه شده است در سیستم. و ما می‌توانیم متغیرهایمان را بر مبنای اینکه کیفی هستند اسمی، کیفی هستند ترتیبی یا کمی هستند فاصله‌شان را محاسبه و با هم جمع کنیم.

بعد از اینکه Fields ها را مشخص کردیم، می‌توانیم بیاییم و از سیستم بخواهیم که ویژگی‌ها و مشخصات مدل‌مان را برایش مشخص کند، اولین ردیف که داریم نوشته Model Name: به طور خودکار سیستم مثلاً می‌گذارد k_means،

Subject:
Year. Month. Date. ( )

ممکن است که شما بخواهید یک اسم خاص خودمان را برعهده بگیریم، می‌توانید بروید گزینه custom را انتخاب کنید،

و در box مربوطه‌اش، توضیحات مربوط به مدل مان را ارائه کنیم. در ردیف دوم نوشته use partitioned data

یعنی ما می‌توانیم از سیستم خواسته که بردار اطلاعات بخش بندی شده استفاده بکن، یعنی سیستم وقتی که می‌خواهد به

ما خروجی بدهد، میاید مثلاً می‌گوید که متغیر مد نظر ما یک متغیری است که نقش مثلاً Partition را دارد، بخش

بندی را دارد، سیستم می‌آید از این متغیر در بخش بندی‌اش یعنی برای داده‌های Test و

Train و evaluation آن داده‌ها را در نظر می‌گیرد، الان می‌خواهد Train کند از داده‌های Train استفاده

استفاده خواهد کرد، برای Test اش از یک بخش دیگر داده‌ها و برای evaluation از قسمت سوم

داده‌ها استفاده خواهد کرد. در k-means ما باید خودمان بگوییم چند خوشه داریم، default این که

خودش برای تعداد خوشه‌ها در نظر گرفته ۵ تا است. آیا می‌خواهید که فیلد مربوط به فاصله‌ی هر مشاهده

را در نظر بگیرید؟ بعضی داده‌ها برای k-means باید input باشد. ولی ما متغیرها متغیرهایی که برای

سیستم مرتبی کردیم، وقتی که سیستم می‌خواست یک مدلی را Run کند، خود k-means فقط

بر اساس داده‌های input ساخته می‌شود، ما می‌گوییم که در داده‌های ما یک Data ای داریم که

نقش بخش بندی را داده برایمان بازی می‌کند. پس در کاری که ما داریم، مدل وقتی که می‌خواهد اجرا شود،

بر مبنای k-means، input ها این کار را انجام می‌دهند ولی یک متغیری داریم که از جنس Partition است

Subject:
Year. Month. Date. ( )

8

proximity

و آن برایمان داده ها را تفکیک میکند. در قسمت show cluster primary این امکان وجود دارد

که ما بخواهیم در shit داده هایی که از این به بعد خواهیم گرفت، اطلاعات مربوط به تشابه خوشه ها را نیز به ما

بدهد، یک گزینه ای دیگر داریم به نام cluster label که میتواند cluster به صورت string باشد یا

Number. خوب بهینه کردن الگوریتم برای اینکه در روش های تکراری بتوان خوب ازش استفاده بکند، میتواند

مهم باشد که ما در این متغیرها یا متغیرهایی که داریم در نظر میگیریم، سیستم باید براساس سرعت دستگاه تنظیماتش

را انجام دهد یا باید براساس حافظه ی نرم افزاری و سخت افزاری اش باید داده ها را با حجمشان سر کار داشته

باشد و کارهایش را انجام دهد. پس در کاری که ما داریم انجام می دهیم برایمان مهم است که از حافظه ی نرم 72 min

افزاری سیستم مان استفاده میکنیم یا از حافظه ی سخت افزاری یا از سرعت کامپیوتر. در قسمت Expert هم

ما نوع خروجی که میخواهیم سیستم بهمان بدهد، برایمان مشخص میشود که ما میخواهیم یک خروجی ساده سیستم بهمان

بدهد که بگوید اینها cluster 2 هستند، 2 هستند یا 3 یا یک میانگین را به ما بدهد یا اینکه نه، میخواهیم که شرط

توقف را و چیزهای دیگر را سیستم بهمان لحاظ کند و برایمان انجام دهد. اگر که گزینه Expert را بزنیم،

شرط توقف را هم در اختیار مان خواهیم داشت، به عنوان مثال ما در کاری که داریم انجام می دهیم اگر گزینه

Expert را فعال کنیم، در Expert (قسمت Mode گزینه Expert فعال شود)، در قسمت

Generate distance Field فاصله ی هر داده را تا میانگین و گره خوشه ای که تعلق گرفته برایمان محاسبه می کند.

Subject:
Year. Month. Date. ( )

ما گفتیم می‌آید داده‌ها را حول و حوش میانگین‌ها قاط می‌کند، اینجا سیستم می‌آید مثلاً می‌گوید: نفر اول فاصله‌اش تا میانگین ۲ است، نفر بعدی ۳٫۲۵ است، نفر بعدی ۱٫۰۲ است و نفر بعدی صفر است و ...

می‌آید یک ستون دیگر به داده‌ها اضافه می‌کند و این اعداد را برایمان در نظر می‌گیرد. در کرکره Expert می‌توانیم شرط توقف را هم به سیستم بدهیم، یعنی آمد Expert را بزنیم، به عنوان مثال این امکان برایمان وجود دارد که ما شرط توقف سال را تعیین بکنیم، بگوییم به طور پیش فرض اش ۲۰ بار که تکرار کرد وایمیستد و اگر تغییراتش مثلاً تا رقم زیر ۱۰ رقم اعشار باشد، توقف می‌کند سیستم، اگر داده‌های ما از نوع set باشند، زمانی که مقدار EN-coding اش ۰٫۷ شود یعنی جذر ۰٫۵ می‌شود تقریباً، به ۰٫۷ برسد، سیستم دیگر متوقف می‌شود و می‌گوید دیگر بیشتر از این نمی‌تواند خوشه‌ها را تعیین کند. حالا اگر excute کنیم، در ورژن پایین‌تر ویژگی model، نوع k_means فعال می‌شود ولی در ورژن بالاتر هم زمان این نود مربوط به k میانگین که نیمه مدل است اتوماتیک وارد وارد صفحه جریان داده‌ها خواهد شد. ما برای اینکه محتوای اطلاعات و نتایج حاصل از k میانگین را ببینیم، باید در ورژن پایین خودمان منتقل اش کنیم در صفحه، ولی در ورژن 14,2 اتوماتیک وارد وارد صفحه می‌شود و حتی لینک اش هم به طور خودکار به source داده‌ها وصل می‌شود. حالا آمدیم روی این نود مربوط به k میانگین، دوبار کلیک کنیم، اطلاعات مربوط به داده‌ها یا نتایج اش در این پنجره‌ای مربوطه

Zarrin Barg

Subject:
Year. Month. Date. ( )

ارائه خواهد شد. در قسمت Model میبینیم حدود 79 رکورد داریم، در کلاستر بعدی 100 تا، در کلاستر بعدی اشاره، کلاستر بعدی 125 تا، کلاستر بعدی 63 اشاره داریم (ص 12). اگر بخواهیم اطلاعات به صورت نمودار یا اطلاعات کامل تر ببینیم، در قسمت میبینیم مدل اگر Expand All را بزنیم، اطلاعات جزئی تر ص 13 در این قسمت ظاهر خواهد شد، اگر به کرکره viewer در همین پنجره بروم، این امکان برایمان وجود خواهد داشت که ما بیاییم در خوشه بندی اطلاعات مربوط به متغیرهایمان را به صورت نمودار ببینیم، به عنوان مثال ما داده ها را به 5 خوشه ص 14 تقسیم شده، متغیر اول ما در خوشه 0 نقش آن چندانی نداشته، بیشتر در خوشه چهارم cluster4، cluster2، cluster1 نقش بالا نشان نوشته بیشتر داشته تا داده های دیگر، متغیر ۲ در کلاستر ۳ تقریبا هیچ نقشی نداشته، حالا درجه اهمیت برایمان گذاشته، بار شن صلحی قرمز، زرد، جریح آبی، قشلاق ملوان نیز، درجه اهمیت بین متغیرها هم برایمان گذارش کرده، ممکن است که این خروجی در نرم افزار clementin ورژن 14 تغییر کرده، مثلا shit ما به ۲ قسمت تقسیم شده، در یک قسمت اطلاعاتی در قسمت Model summery View داریم و cluster، در cluster اول نمودار که در اینجا دیدیم با متغیرات مربوطه اش که چند درصد سهم داشته و به صورت درصدی بیان کرده، در قسمت سمت چپ صفحه، کلاستر اش اطلاعات مربوط به متغیرها را به ما خواهد داد که ما نه تنها میتوانیم فرم درصدی، میتوانیم زیرش فرم نمودار را هم میتوانیم به سیستم بدهیم بهمان بر مبنا.

Subject:
Year. Month. Date. ( )

و در قسمت سمت راست در ماشین لرنینگ میتوانیم اطلاعاتی را که از خوشه ها برای این کار قابل استخراج هست را ارزش استفاده کنیم. در ورژن ۱۴ خروجی مان تغییر کرده، یعنی output ایی که به ما می دهد، یک نمودار جاز ادواره به ما می دهد در قسمت Model summary بنام siluat، نمودار siluat یک نمودار است که مقدارش بین ۱- تا ۱+ را در اینجا نشان می دهد، هر چند مقدارش به یک نزدیک تر باشد، نشان دهنده این است که خوشه هایی را که ما انتخاب کردیم، خوشه های مناسبی است و در تحلیل ارزش استفاده خواهیم کرد. آنها در رنگ خوشه های در دوره بالا این نمودار siluat را نشان کنیم از سه رنگ: مقدار ایی، زرد و قرمز تشکیل شده، حالا آنها بیایم در خوشه ها، قسمتی که به رنگ صفر است ایی شده، یعنی داده خوب بودن خوشه ها را براساس نوع داده هایی که داریم داده نشان می دهد. و در عین حال برای متغیرهای کیفی از نمودار های ستونی و برای متغیرهای کمی از چیزی به هیستوگرام برای تفکیک داده هایش استفاده میکند. (اطلاعات خروجی برای متغیرهای کیفی به صورت میانگین و مقدار انحراف معیار (کیفی اسمی) محاسبه میشود و برای متغیرهای کیفی به صورت درصدی بیان می شود) 20