Discovering Knowledge in Data\_ An Introduction to Data Mining (2nd ed.)

دوره مقدماتی آموزش داده کاوی

* جلسه اول

1. معرفی دوره و چگونگی تدریس و آموزش
2. معرفی افراد
3. شروع مقدمات با طرح موضوع مفهوم داده کاوی و مراحل داده کاوی

**- معرفی دوره :**

* شما توی دوره مقدماتی تحلیل گر شو (مثبت دیتا) ثبت نام کردین، این دوره کلا آموزشش با خودم هست و در این دوره سعی میشه پیش نیازهایی که احتیاج دارید و یکسری مفاهیم پایه ای رو براتون بگیم.
* اول از همه اینکه شک نکنید این دوره بسیار دوره شیرینی خواهد بود و اصلا حوصلتون سر نمیره. هدف من در درجه اول یادگیری شماست و اینکه هزینه ای که میکنید در نهایت در آینده برای شما کسب هزینه کنه.
* دوم اینکه کسانی که بتونن توی این دوره و دوره پیشرفته جزو سه نفر اول دوره بشن، در پروژه های واقعی بکارگیری میشن و آموزش داده میشن. از همین اول بهتون بگم تلاشتون رو زیاد کنید و برای هزینه ای که کردین احترام قائل باشید. نفر اول این دوره که بعدا معیارهای نفر اول شدن رو هم براتون توضیح میدم به صورت رایگان توی دوره پیشرفته ما شرکت میکنه.
* در کنار آموزش مفاهیم سعی میشه زبان پایتون که یکی از قدرتمندترین زبان ها برای داده کاوی هست رو هم آموزش بدیم.
* چرا پایتون و چرا بقیه نرم افزارها نه؟؟؟ چون پایتون رایج ترین نرم افزار هست و برای شما یک مهارت خوب هست
* بچه ها تفکر من این نیست که شما رو مثه دانشگاه درس بدم، مفاهیم رو باید بفهمید، باید درک کنید چی به چیه، اینجور بدردرتون میخوره و بهتون قول میدم در آینده نه چندان دور برای همتون پول ساز خواهد بود.

**- روند کلاس:**

* کلاس ها در روزهای یکشنبه و چهارشنبه ساعت 19 الی 22 (بین کلاس ها 20 دقیقه استراحت داریم)
* گروه بندی کلاس (5 تا گروه)
* رسم نقشه ذهنی برای مفاهیم درسی
* کوییز مفهومی
* کارگاه در صورت نیاز
* anydesk

**-داده کاوی چیست؟** داده کاوی فرایند کشف الگوها و روندهای مفید در مجموعه داده های بزرگ است.

میدونین تفاوت اختراع و کشف چیه؟ اختراع یک حرکت نو آورانه است که قبلا نبوده و ابتکار یک نفر هست ولی کشف چیزی هست که بوده ولی به چشم کسی نیومده... یعنی چیزی که وجود داره ولی نمیبینیش... دقیقا داده کاوی هم همین کار رو میخواد بکنه. شما یکسری عدد دارید، نگاهشون که میکنید هیچ چیزی دستگیرتون نمیشه و اگر هم بشه چیز خاصی نیست... ولی وقتی وارد داده ها میشین اگر از متدهای درست استفاده کنید به یه کشفیاتی میرسین که از چشم همه حتی کسایی مستقیما با اون داده ها سر و کار داشتن پنهون مونده.. پس با داده کاوی قراره چیزهای ببینیم که وجود دارن ولی از دید پنهون شدند.

**مراحل داده کاوی:** مثل همه فرآیندهای علمی داده کاوی هم دارای یه فرآیند هست که طبق اون اگر جلو بریم میتونیم بگیم که تونستیم به یه نتیجه درست برسیم. بذارید یه مثال ملموس براتون بزنم. من قبلا یعنی قبل دانشگاه تابستونا میرفتم نقاشی ساختمون کار میکردم. این بنده خدا اوستای ما همیشه یه حرف خوبی بهم میزد... میگفت یه نقطه رو شروع قرار بده و یه نقطه رو پایان ، از وسط کار شروع نکن، اینجور می فهمی چکار کردی و کجا رو رنگ کردی... خلاصه یعنی توی کار گم نشید. این فرآیند داده کاوی هم دقیقا همونه.. یعنی باید یاد بگیریم که از یک جا شروع کنیم و یک جا تموم کنیم. توی کار نیفتیم. حالا این فرآیند داده کاوی مراحل زیر رو داره:

1. **فهم حوزه ‌ی منبع داده ها:** این حوزه هدفش چیه، میخواد به کجا برسه، درآمدش از کجاس، بازار فروشش کجاست و کلا تمام روابط کاری این حوزه رو باید درک کنید.
2. **فهم داده ها:** داده های ذخیره شده از چه نوعی هستند؟ نقاط ضعف و قوت داده هاشون چیه؟ حجم داده ها چقدره؟ چندتا مشخصه داره؟ کدوم مشخصه ها برای مدیران اون حوزه اهمیتش بیشتره؟ کدوم از دیتاها بد جمع شده و باید خیلی مورد بررسی قرارش داد و کلا کیفیت داده ها مورد بررسی قرار بگیره
3. **پیش پردازش داده‌ها:** داده ها حتما و حتما دچار مشکلاتی هستند، مثل گمشده بودن داده ها، همبستگی داده ها با هم، بی نظم بودن در صورتی که از چند دیتابیس مختلف جمع شده باشند، هم مقیاس نبودن و ... برای رسیدن به یک پردازش خوب و نتیجه قابل اطمینان داده ها قبل از استفاده باید مورد بررسی دقیق قرار گیرند. این مرحله یکی از مهمترین مراحل داده کاوی محسوب میشه.
4. **مدل کردن داده‌ها:** در این قسمت بسته به نوع دیتا و خروجی مد نظر خود یک مدل را برای داده ها انتخاب کرده و آن را برازش میدهیم. این نکته بسیار حائز اهمیت است که انتخاب متد درست منجر به خروجی اطلاعات و مفاهیم درست می گردد.
5. **فاز ارزیابی :** در این قسمت بر روی داده ها مدل برازش داده شده و خروجی هم حاصل شده، حال باید این سوال را از خودمان بپرسیم که مدل برازش داده شده مدل درستی است؟ خطای حاصل از این مدل چقدر است؟ چه پیشنهادی برای کم شدن خطا داریم؟
6. **فاز توسعه :** در این فاز باید این امر بررسی شود که آیا مدل برازش شده برای جامعه دیگر نیز قابل توسعه است؟ و یا فقط در چارچوب همین مسئله جواب میدهد؟

یکسری کج فهمی ها در خصوص داده کاوی هست که اینجا باید براتون توضیح بدم :

* اول اینکه داده کاوی یک فرآیند هست و به صورت اتوماتیک انجام نمیشه، یعنی یه مرحله رو باید انجام بدی ببینی چه خبره، اگر همه چیز اوکی بود بعد بری مرحله بعد، حتی بعضی اوقات باید از مرحله های جلوتر برگردی مرحله های عقب تر بعضی از نقوص رو برطرف کنی
* داده کاوی به نظارت انسان احتیاج دارد زیرا که هر مرحله باید مورد بررسی قرار بگیره وگرنه خروجی اشتباه هست.
* عدم تجزیه و تحلیل از تجزیه و تحلیل اشتباه بهتر است. اگر شما در یک سیستم بخواهید با استفاده از داده کاوی مشاوره بدهید به مدیریت دستگاه، اگر مشاوره ندهید بهتر است تا مشاوره اشتباه دهید.

دوره مقدماتی آموزش داده کاوی

* جلسه دوم (مفاهیم پایه پایتون)

-در این قسمت قرار به توضیح کدهای پایتون نیست و تنها برخی مفاهیم پایه پایتون رو مورد بررسی قرار می دهیم، دوستانی که اصلا با پایتون آشنایی ندارند، میتونن از سایت W3school و یا سایت های مرتبط آموزش ببینند. اگر هم ظرفیتتون به 20 نفر برسه، کارگاه یک روزه براتون برگزار میکنیم. ولی خب کدهایی که اینجا داریم باهاشون کار میکنیم، کدهای مربوط به کار با دیتا اونم درسطح ابتدایی هست.

* کامنت

کامنت قطعه ای از کد است که اجرا نمیشود. در نوشتن کامنت نباید زیاد وسواسی باشیم، از طرفی نداشتن کامنت در کدها، باعث سردرگمی شما در مراجعات بعدی خود به کدها میشه. از طرفی استفاده افراطی از کامنت در کدها، بیشتر باعث سردرگمی و بی نظمی کدها می شود.

# This is a comment

* وارد کردن پکیج ها در پایتون

در حالی که بسیاری از کارها را می‌توان در پایتون **"**خارج از جعبه" (out of the box - یعنی پس از نصب پایتون در اختیار هست و احتیاجی به نصب ماژول و یا کتابخانه ای نداریم) انجام داد، بیشتر کارهایی که می‌خواهیم انجام دهیم نیازمند وارد کردن بسته‌ها هستند. بسته ها حاوی کدهای طراحی شده ویژه ای هستند که ما را قادر می سازد تا کارهای پیچیده علم داده را بدون نوشتن کد خود انجام دهیم. به عنوان مثال، در فصل 6، باید یک مدل طبقه بندی و درخت رگرسیون **(**classification and regression tree) بسازیم. به جای اینکه بفهمیم چگونه یک مدل CART را از ابتدا بسازیم، بسته ای را وارد می کنیم که حاوی آن کد است. هنگامی که بسته وارد شد، می توانیم کد را برای ایجاد یک مدل CART اجرا کنیم.

دو مورد از مهمترین پکیج ها (ماژول و یا همان بسته) در پایتون که جز پرکاربردترین پکیج ها در علوم داده هستند، pandas و numpy هستند. برای وارد کردن این دو پکیج در پایتون باید دوخط زیر تایپ و اجرا کنیم:

import pandas as pd  
import numpy as np

توجه داشته باشید که ما بسته ها را با استفاده از دستور import وارد می کنیم. کد as با استفاده از نامی که ما می توانیم تعیین کنیم، بسته را تغییر نام می دهد. ما بسته ها را تغییر نام می دهیم تا کار با آنها آسان تر شود.

برای استفاده از دستورات موجود در بسته‌های panda و numpy، باید نام بسته‌ها را قبل از نام فرمان بیان کنیم. به عنوان مثال برای استفاده از دستور read\_csv()، باید

()pandas.read\_csv

را تایپ کنیم. همچنین می‌توانیم یک نام مستعار به بسته بدهیم. در مورد بالا، بسته pandas را با استفاده از as pd به pd تغییر می دهیم و با استفاده از as np بسته numpy را به np تغییر می دهیم. پس دستور بالا به شکل زیر بازنویسی می‌گردد:

pd.read\_csv()

همیچنین ما میتوانیم قسمتی از یک پکیج را در پایتون بازخوانی کنیم:

from “module name” import Func1,Func2,…

برای مثال از ماژول sklearn.tree توابع DecisionTreeClassifier **و** export\_graphviz را فراخوانی میکنیم:

from sklearn.tree import DecisionTreeClassifier, export\_  
graphviz

* وارد کردن داده ها در پایتون

اکنون در مورد نحوه وارد کردن یک مجموعه داده به پایتون بحث خواهیم کرد. در این متن از دستور read\_csv() با استفاده از ساختار زیر استفاده می کنیم:

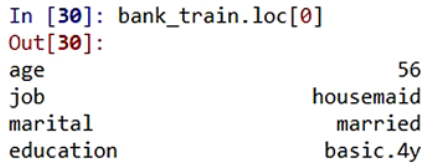
your\_name\_for\_the\_data\_set = pd.read\_csv ("the\_path\_to\_ the\_file")

برای مثال جهت خواندن داده ها در فایل (bank\_marketing) از کد زیر استفاده میکنیم:

bank\_train=pd.read\_csv("C:/Users/Data\_Science/Data/bank\_marketing”)

* دسترسی به رکوردها و متغیرها در پایتون

در بررسی های متفاوت ممکن است بخواهید یک رکورد خاص را بررسی کنید. به عنوان مثال، چگونه به یک رکورد در مجموعه داده bank\_train دسترسی پیدا کنیم؟ ما از متد loc که در بسته Pandas فراهم شده برای انجام این کار استفاده خواهیم کرد.



(دستورات در ژوپیتر ذخیره شده است)

* خروجی گرافیکی در پایتون: قبل از اینکه بخش کدنویسی پایتون را ترک کنیم، باید به یک چیز دیگر نیز بپردازیم: نحوه به دست آوردن و تغییر خروجی گرافیکی در پایتون.(دستورات در ژوپیتر ذخیره شده است)

دوره مقدماتی آموزش داده کاوی

* جلسه سوم (پیش پردازش دیتا)

1. چرا به پیش پردازش دیتا احتیاج داریم؟
2. مدیریت داده های از دست رفته

**1- چرا به پیش پردازش دیتا احتیاج داریم**

بسیاری از داده های خام موجود در پایگاه های داده ، پردازش نشده ، ناقص و نویزی هستند، به عنوان مثال ، پایگاه های داده ممکن است حاوی :

* اطلاعات منسوخ شده باشد (برای مثال فرض کنید یک خط تولید از یک کارخانه کاملا از روند تولید خارج شده باشد در این صورت، دیتای مربوط به آن منسوخ شده است)
* مقادیر از دست رفته (Missing value)
* داده های پرت (Outlier)
* داده های خارج از فرم و نامناسب برای ورود به مدل (داده هایی که باید با استفاده از عملیات هایی آن ها را تبدیل به داده های خوش فرم کنیم)
* داده هایی که عقلانی نیستند.

اغلب داده هایی که با آن ها سر و کار داریم با این مشکلات مواجه هستند، برای اینکه درست برنامه ریزی نشده اند و اپراتور (چه ماشین و چه انسان) دارای خطای ثبت اطلاعات است. حال چرا باید پیش پردازش داده ها انجام شود؟ جواب یک اصطلاح است : GIGO (garbage in garbage out)

اگر به مدل خود آشغال وارد کنید، حتما مدل به شما آشغال تحویل میدهد. پس باید دیتا به صورت تمیز و پیش پردازش شده به مدل داده شود، تا به خروجی درست و قابل اطمینان دست پیدا کنیم. باید حواسمان باشد که گاهی اوقات تا 60 درصد فرآیند داده کاوی در همین مرحله خلاصه میشود و این مرحله مهمترین مرحله داده کاوی است.

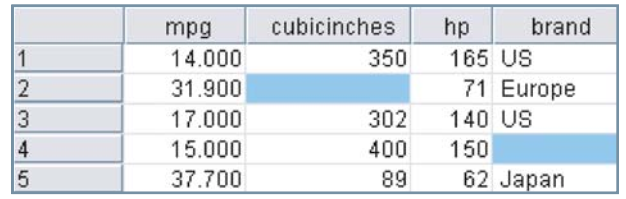
برای مثال به جدول زیر که یک قسمت از دیتابیس یک فروشگاه اینترنتی هست دقت کنید تا مشکلاتش رو با هم بررسی کنیم:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Customer ID | Zip | Gender | Income | Age | Marital Status | Transaction Amount |
| 1001 | 10048 | M | 78,000 | C | M | 5000 |
| 1002 | J2S7K7 | F | -40,000 | 40 | W | 4000 |
| 1003 | 90210 | …. | 10,000,000 | 45 | s | 7000 |
| 1004 | 6269 | M | 50,000 | 0 | S | 1000 |
| 1005 | 55101 | F | 99,999 | 30 | D | 3000 |

* متغیر Customer id کلید اصلی جدول و نشان دهنده شماره مشتری است که مشکل خاصی از نظر پیش پردازش دیتا در آن وجود ندارد
* متغیر Zip متغیر کد پستی مشتریان است، در این جا فرم های متفاوت برای این متغیر میتواند مشکل ساز باشد، برای مثال مشتری 1002 از کانادا و مشتری 1004 از انگلیس هستند و بقیه از آمریکا، اگر پایه ریزی این متغیر براساس کدهای 5 رقمی آمریکا باشد، آنگاه این متغیر بد فرم است.
* متغیر جنسیت یا Gender دارای مقدار گمشده برای مشتری شماره 1003 است.
* متغیر درآمد از چند بعد دچار ایراده، اول اینکه مشتری شماره 1002 درآمد منفی
* .براش ثبت شده که این میتونه خطای برنامه نویسی پایگاه داده باشه که اجازه ثبت مقدار منفی در این فیلد رو میده، دوم اینکه داشتن درآمد 10 میلیون دلار در سال برای فردی (مشتری 1003)که در منطقه هیلس هست مقداری دور از انتظاره، سوم اینکه داشتن درآمد 99999 برای مشتری 1005 غیرعادیه، برای اینکه مگه میشه کسی دقیقا این درآمد رو داشته باشه و از طرفی این عدد برخی اوقات شیطنت مدیران پایگاه داده برای خالی نموندن فیلدها هم به حساب میاد. مشکل آخر و مهمترین مشکل اینکه درآمد افراد در این فیلد بر مبنای یک واحد پولی ثابت نیست، مشتری 1002 بر اساس دلار کانادا، مشتری 1004 براساس پوند انگلیس و بقیه مشتری ها براساس دلار آمریکا در آمد خود رو اعلام کردند.
* متغیر Age و یا همان سن مشتری دارای دو مشکل است، اول اینکه سن افراد به دو صورت گروه بندی شده و عددی وارد شده است و دیگر اینکه عدد غیرممکن صفر برای مشتری شماره 1004 ثبت شده است.
* متغیر Marital status و یا وضیعت تاهل هم دچار مشکل کدگذاری است و برای مثال ما نمیدانم کد s نشان دهنده مجرد (single) و یا جدا شده (separated) است.
* متغیر میزان تراکنش یا Transaction amount با شرط اینکه مقدار این تراکنش ها با یک واحد پولی باشد، مشکلی ندارد.

**2-مدیریت داده های از دست رفته**

اول از همه باید گفت که هر چه اطلاعات بیشتر باشد، با فرض انتخاب روش درست تحلیل، خروجی تحلیل قابل اعتمادتر می باشد. مشکل داده های گمشده تقریبا در همه ی پایگاه های داده ای هست و برای حل مشکل اون باید کاری بکنیم.



یک روش متداول برای "کنترل" مقادیر از دست رفته صرفاً حذف رکوردها (سطر دیتایی) با مقادیر گمشده از تجزیه و تحلیل است. با این حال، این ممکن است خطرناک باشد، زیرا الگوی مقادیر از دست رفته ممکن است در واقع سیستماتیک باشد، و صرفاً حذف رکوردهای دارای مقادیر از دست رفته منجر به یک زیرمجموعه مغرضانه از داده ها می شود. علاوه بر این، صرفاً به این دلیل که یک مقدار از یک ویژگی از بین رفته است، حذف اطلاعات در تمام ویژگی ها دیگر هدر می‌رود (یک سطر دیتا از دست میدهیم).

در واقع اگر تنها 5٪ از مقادیر داده از مجموعه داده 30 متغیری وجود نداشته باشد (یعنی هر ویژگی یا صفت 5 درصد دیتای گمشده داشته باشه) و مقادیر از دست رفته به طور مساوی در سراسر داده ها پخش شوند، تقریباً 80٪ از رکوردها حداقل یک مقدار گمشده دارند (یعنی در یک سطر دیتا یک مقدار گمشده است). بنابراین، تحلیل‌گران داده‌ به روش‌هایی روی آورده‌اند که مقادیر گمشده را با مقداری جایگزین می‌کنند که بر اساس معیارهای مختلف است.

برخی از معیارهای رایج برای انتخاب مقادیر جایگزین برای داده های از دست رفته به شرح زیر است:

1. مقدار از دست رفته را با مقداری ثابت که توسط تحلیلگر مشخص شده است جایگزین کنید.

2. مقدار گمشده را با میانگین آن ویژگی (برای متغیرهای عددی) یا نما (برای متغیرهای طبقه‌بندی) جایگزین کنید.

3. مقادیر از دست رفته را با یک مقدار ایجاد شده به صورت تصادفی از توزیع مشاهده شده متغیر جایگزین کنید.

4. مقادیر از دست رفته را بر اساس سایر ویژگی های رکورد جایگزین کنید. (یعنی برای مثال از یک مدل رگرسیونی استفاده کنیم)

|  |
| --- |
|  |
| Replacing missing field values with user-defined constants. |
|  |
| Replacing missing field values with means or modes |

|  |
| --- |
|  |
| Replacing missing field values with random draws from the distribution of the variable. |

|  |  |
| --- | --- |
| متغیرها | |
| میزان مصرف بنزین بر حسب مایل | mpg |
| حجم موتور برحسب اینچ مکعب | cubicinches |
| قدرت موتور بر حسب اسب بخار | hp |
| برند خودرو | brand |

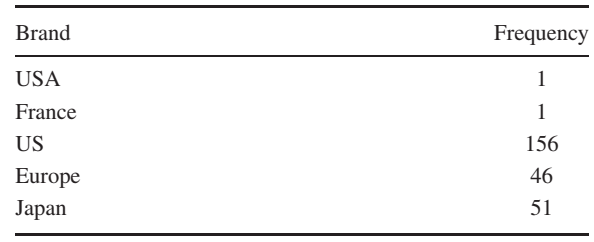
مشکلات این روش ها:

* ممکن است داده های حاصل منطقی و درست نباشند. مثلا در بالا هیچ خودروی ژاپنی در دیتاها موجود نیست که حجم موتورش 400 اینچ مکعب باشد.
* مسئله دیگه اینه که معیارهای مرکزی به مقادیر پرت حساس هستن، اگر نقطه پرتی باشه، برآورد ما رو برای مقدار گمشده داغون می کنه

استفاده از روش های مدل کردن داده ها باید انجام بگیره تا این مشکلات حل بشه.

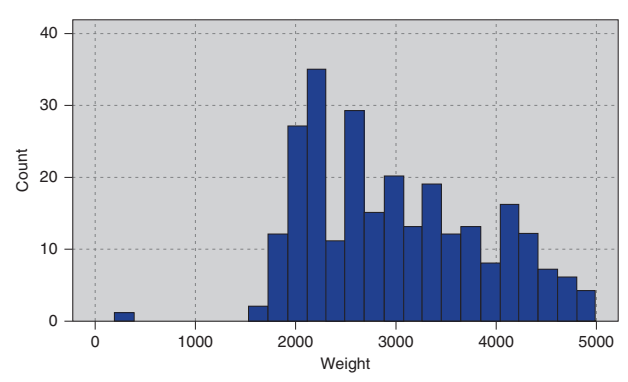
**3-شناسایی طبقه بندی های نادرست**

اجازه دهید نمونه ای از بررسی برچسب های طبقه بندی روی متغیرهای طبقه بندی را بررسی کنیم تا مطمئن شویم که همه آنها معتبر و سازگار هستند. جدول زیر توزیع فراوانی پنج کلاس، ایالات متحده آمریکا، فرانسه، ایالات متحده، اروپا و ژاپن را نشان می دهد. با این حال، دو کلاس، ایالات متحده آمریکا و فرانسه، هر کدام تنها یک خودرو دارند. آنچه به وضوح در اینجا اتفاق می افتد این است که دو مورد از رکوردها به طور متناقض با توجه به مبدأ ساخت طبقه بندی شده اند. برای حفظ سازگاری با بقیه مجموعه داده‌ها، رکورد با مبدأ ایالات متحده US باید برچسب ایالات متحده USA و رکورد با مبدا فرانسه باید دارای برچسب اروپا باشد.

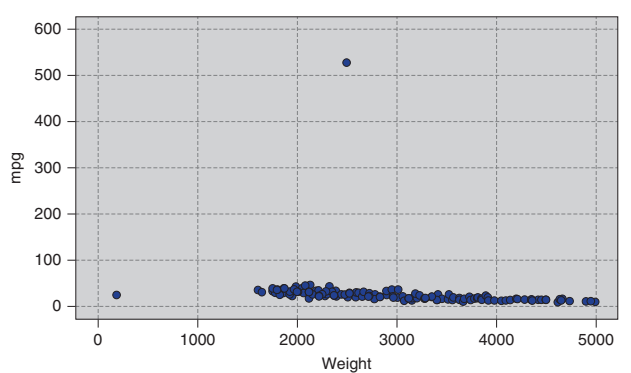


**4-روش‌های گرافیکی برای شناسایی نقاط پرت (**outlier**)**

مقادیر پرت مقادیر شدیدی هستند که برخلاف روند مابقی داده ها هستند. شناسایی نقاط پرت مهم است زیرا ممکن است نشان دهنده خطا در ورود داده ها باشد. همچنین، حتی اگر نقطه پرت یک نقطه داده معتبر باشد و خطا نباشد، روش‌های آماری خاصی به وجود نقاط پرت حساس هستند و ممکن است نتایج غیر قابل اعتمادی ارائه دهند. یک روش گرافیکی برای شناسایی نقاط پرت برای متغیرهای عددی، بررسی هیستوگرام آن هاست.



(هیستوگرام وزن خودرو) از آنجایی که 192.5 پوند برای یک خودرو بسیار سبک است، ما تمایل داریم در صحت این اطلاعات شک کنیم. می توانیم حدس بزنیم که شاید وزن 1925 پوند بوده، با این حال، نمی‌توانیم مطمئن باشیم و تحقیقات بیشتری در مورد منابع داده‌ها مورد نیاز است. گاهی اوقات نمودارهای پراکندگی دوبعدی می توانند به آشکارسازی نقاط پرت در بیش از یک متغیر کمک کنند.



بیشتر نقاط داده در امتداد محور افقی در امتداد هم هستند، به جز دو نقطه پرت. خودروی سمت چپ همان وسیله نقلیه ای است که در شکل قبل شناسایی کرده ایم و تنها 192.5 پوند وزن داشت. نقطه پرت نزدیک به بالا چیز جدیدی است: خودرویی که بیش از 500 مایل به ازای هر گالن سوخت می‌پیماید! واضح است، مگر اینکه این وسیله نقلیه با کریستال های دی لیتیوم کار کند. ما باید به دنبال خطای ورود اطلاعات باشیم.

توجه داشته باشید که وسیله نقلیه 192.5 پوندی از نظر وزن بسیار پرت است اما از نظر مسافت پیموده شده نیست. به طور مشابه، خودرویی که 500 مایل در هر گالن از نظر مسافت میپیماید پرت است اما از نظر وزن نه. بنابراین، یک رکورد ممکن است در یک بعد خاص پرت باشد اما در بعد دیگر نه. ما روش های عددی را برای شناسایی نقاط پرت بررسی خواهیم کرد.

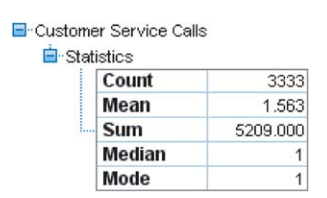
**معیارهای مرکزی و پراکندگی**

فرض کنید که ما علاقه مندیم که مرکز یک متغیر خاص را که با یکی از معیارهای عددی مرکزی اندازه گیری می شود (رایج ترین آنها میانگین، میانه و مد است)، محاسبه کنیم. اندازه‌های مرکز (measures of center) یک مورد خاص از اندازه‌گیری‌های مکان (measures of location) هستند، خلاصه‌های عددی که نشان می‌دهند مشخصه خاصی از متغیر در کجای یک خط عددی قرار دارد. نمونه هایی از معیارهای مکان صدک و چندک هستند.

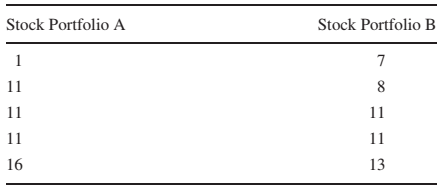
معیار میانگین ، مجموع تمام مقادیر یک متغیر تقسیم بر تعداد آن هاست.



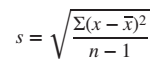
از آنجایی که معیار میانگین، نسبت به مقادیر پرت حساسیت زیادی دارد، استفاده از معیار میانه در وجود نقاط پرت میتواند معیار بهتری برای نشان دادن مرکز داده ها باشد(در اصطلاح اگر توزیع داده ها چوله باشد، استفاده از میانه به جای میانگین پیشنهاد می گردد) توجه داشته باشید که معیارهای مرکز همیشه در مورد اینکه مرکز مجموعه داده در کجا قرار دارد صدق نمیکنند. در جدول زیر (معیارهای مرکزی برای داده های تماس مشتریان) میانه 1 است، به این معنی که نیمی از مشتریان حداقل یک تماس با خدمات مشتری برقرار کردند. مد نیز 1 است، به این معنی که بیشترین تعداد تماس های خدمات مشتری 1 بوده است. میانه و مد موافق هستند. با این حال، میانگین 1.563 است که 56.3 درصد بیشتر از سایر معیارها است. این به دلیل حساسیت میانگین به راست چولگی داده ها است.



اندازه گیری مکان برای خلاصه کردن یک متغیر به طور موثر کافی نیست. در واقع، دو متغیر ممکن است مقادیر بسیار یکسانی برای میانگین، میانه و حالت داشته باشند و در عین حال ماهیت های متفاوتی داشته باشند. به عنوان مثال، فرض کنید که پرتفوی سهام A و پرتفوی سهام B هر کدام شامل 5 سهام هستند.(با نسبت قیمت/درآمدP/E ) همانطور که در جدول زیر نشان داده شده است. پورتفولیوها به طور مشخص از نظر نسبت P/E متفاوت هستند. پرتفوی A شامل یک سهام با نسبت P/E بسیار کوچک و دیگری با نسبت P/E نسبتاً بزرگ است. با این حال، نسبت‌های P/E پورتفولیو B به‌شدت حول میانگین خوشه‌بندی می‌شوند. با این حال، با وجود این تفاوت ها، میانگین، میانه و نسبت P/E و مد پرتفولیوها دقیقاً یکسان است: میانگین نسبت P/E 10 ، میانه 11 و مد برای هر پورتفولیو 11 است.



واضح است که معیارهای مرکزی تصویر کاملی را در اختیار ما قرار نمی دهد. آنچه وجود ندارد، معیارهای گسترش یا معیارهای تغییرپذیری است که میزان پراکندگی مقادیر داده را توصیف می کند. نسبت‌های P/E پورتفولیو A نسبت به پورتفولیو B پراکنده‌تر است، بنابراین معیارهای تغییرپذیری برای پورتفولیو A باید بزرگ‌تر از شاخص‌های B باشد. معیارهای معمول تغییرپذیری شامل محدوده (Range) (حداکثر - حداقل)، انحراف استاندارد (SD) میانگین انحراف مطلق، و محدوده بین چارکی (IQR) است. SD نمونه‌ای شاید رایج ترین معیار تغییرپذیری باشد :



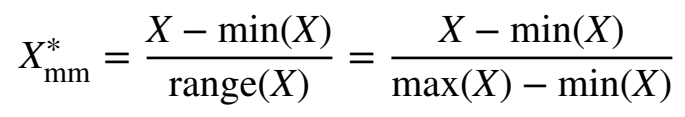
به دلیل مجذور شدن صورت، SD به وجود نقاط پرت حساس است، که باعث می شود تحلیلگران دیگر معیارهای پراکندگی، مانند میانگین انحراف مطلق، را در شرایطی که داده ها مقادیر پرت را شامل می شود، ترجیح دهند.

**5- تبدیل داده ها**

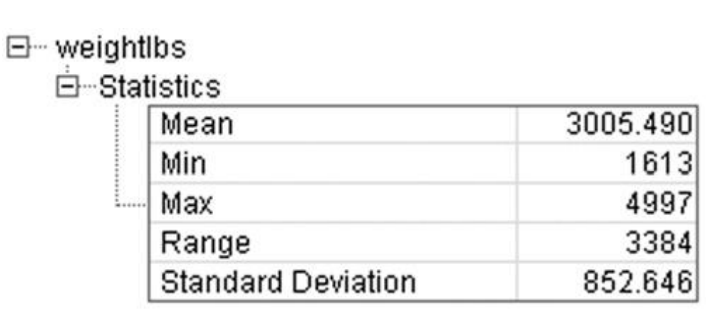
متغیرها معمولاً محدوده هایی دارند که با یکدیگر تفاوت زیادی دارند. به عنوان مثال، اگر به لیگ برتر بیسبال علاقه مند باشیم، میانگین ضربه زدن بازیکنان از صفر تا کمتر از 0.400 متغیر خواهد بود، در حالی که تعداد ران های خانگی در یک فصل از صفر تا حدود 70 متغیر خواهد بود. برای برخی از الگوریتم های داده کاوی، تفاوت در دامنه ها منجر به تأثیر نامناسب بر نتایج می شود. به این معنی که تنوع بیشتر در هوم ران ها بر تنوع کمتر در میانگین ضربه زدن غالب خواهد بود. بنابراین، متخصصان باید متغیرهای عددی خود را نرمال کنند تا مقیاس تأثیر هر متغیر بر نتایج را استاندارد کنند. شبکه‌های عصبی و الگوریتم‌هایی که از معیارهای فاصله استفاده می‌کنند، مثل الگوریتم k نزدیک‌ترین همسایه، از نرمال‌سازی استفاده می‌کنند. چندین تکنیک برای عادی سازی وجود دارد، و ما دو تا از روش های رایج تر را بررسی خواهیم کرد. اجازه دهید X به مقدار متغیر اصلی ما اشاره کند و X\* به مقدار متغیر نرمال شده اشاره دارد.

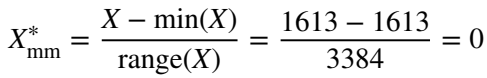
* نرمال سازی مینی ماکس (Mini-Max)

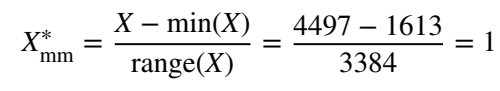
نرمال سازی مینی ماکس به صورت زیر عمل می کند:



اگر مقدار مینیمم به این نرمال سازی وارد شود، مقدار صفر و اگر مقدار ماکسیمم وارد شود مقدار یک برگرداننده می‌شود. همچنین برای مقدار midrange مقدار حاصل از این نرمال سازی برابر با 5. می‌شود. پس نرمال سازی مینی ماکس داده ها را به فضای 0 تا 1 نگاشت می ‌کند.

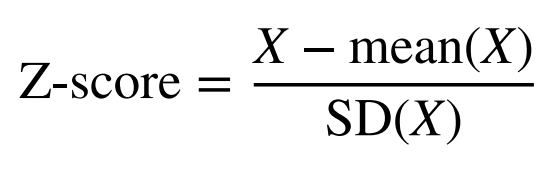




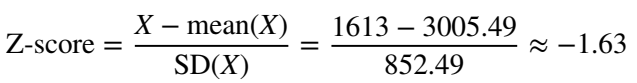


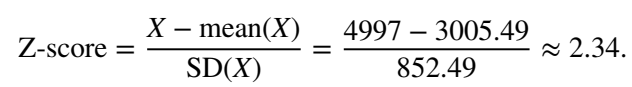
* استاندارد سازی Z-score

استانداردسازی Z-score که در دنیای تجزیه و تحلیل آماری بسیار گسترده است، با در نظر گرفتن تفاوت بین مقدار اصلی و مقدار میانگین، تقسیم بر انحراف استاندارد مقادیر عمل می کند که به صورت زیر تعریف می‌شود :



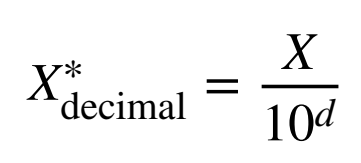
این استاندارد سازی داده ها را تقریبا بین 3 تا 3- نگاشت می‌کند.



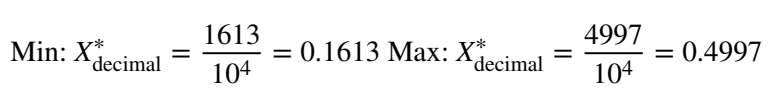


* مقیاس دهی (DECIMAL SCALING)

مقیاس دهی این تضمین را میدهد که تمام مقادیر داده های بین 1 و 1- نرمال سازی می شوند:

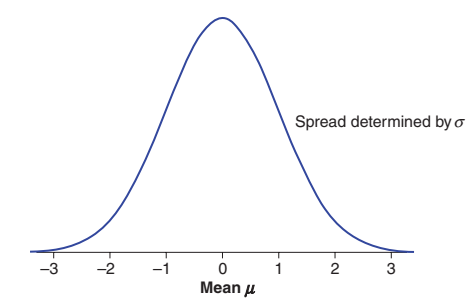


به طوری که d نشان دهنده تعداد ارقام بزرگترین مقدار قدر مطلق اعداد است.

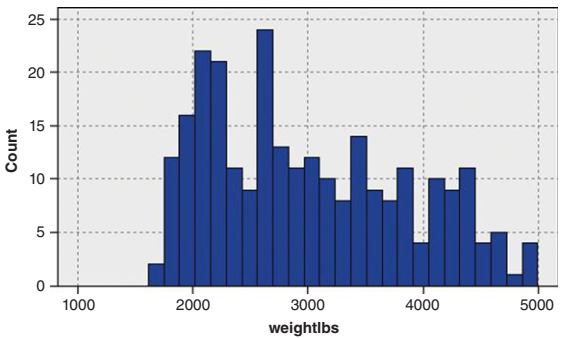


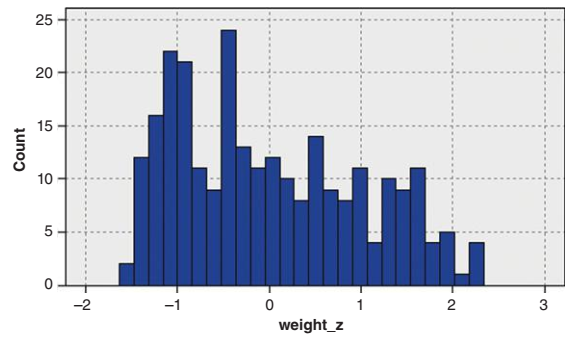
**6-تبدیل داده ها برای رسیدن به داده های نرمال**

برخی از الگوریتم های داده کاوی و روش های آماری مستلزم توزیع نرمال متغیرها هستند. توزیع نرمال یک توزیع احتمال پیوسته است که معمولاً به عنوان منحنی زنگ شناخته می شود که متقارن است. مرکز آن در میانگین µ است و گسترش آن با انحراف معیار σ (سیگما) تعیین می شود. شکل توزیع نرمال را نشان می‌دهد که دارای میانگین 1 و انحراف معیار 0 است که به عنوان توزیع نرمال استاندارد Z شناخته می‌شود.

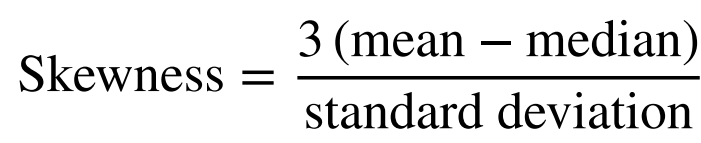


این یک تصور غلط رایج است که متغیرهایی که استانداردسازی نمره Z برای آنها اعمال شده است از توزیع نرمال Z پیروی می کنند. این درست نیست! درست است که داده های استاندارد شده با Z دارای میانگین 0 و انحراف استاندارد 1 خواهند بود اما توزیع ممکن است چوله باشد.





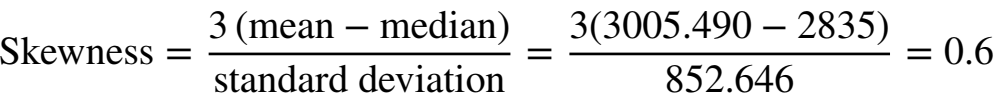
معیار چولگی را با استفاده از عبارت زیر میتوان بدست آورد:

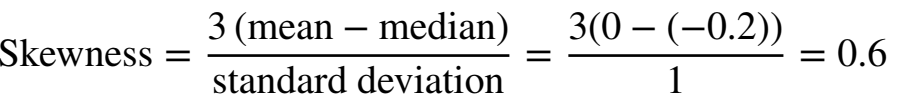


برای داده‌های چوله به راست، میانگین بزرگ‌تر از میانه است و بنابراین چولگی مثبت خواهد بود، در حالی که برای داده‌های چوله چپ، میانگین کوچک‌تر از میانه است و مقادیر منفی برای چولگی ایجاد می‌کند. البته برای داده‌های کاملاً متقارن میانگین، میانه و نما همگی برابر هستند و بنابراین چولگی برابر با صفر است.

|  |  |  |
| --- | --- | --- |
|  | C:\Users\INTERNET\Desktop\Untitled.png |  |

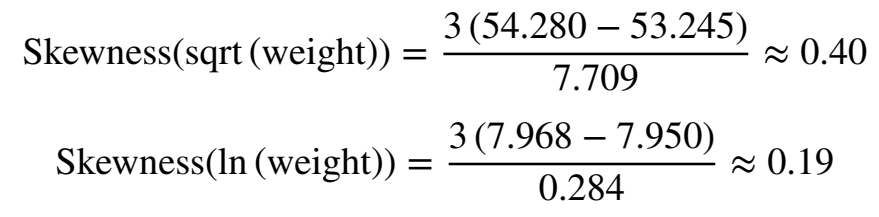
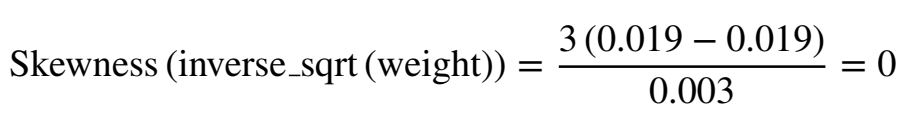
بسیاری از داده های دنیای واقعی، از جمله داده های مالی، دارای انحراف راست هستند. داده‌های چوله چپ چندان رایج نیستند، اما اغلب زمانی رخ می‌دهند که داده‌ها به راست سانسور شوند، مانند نمرات آزمون در یک آزمون آسان، که نمی‌تواند بیشتر از 100 باشد. در اینجا چولگی برای متغیر weight و همچنین متغیر نرمال شده Weight محاسبه شده است :

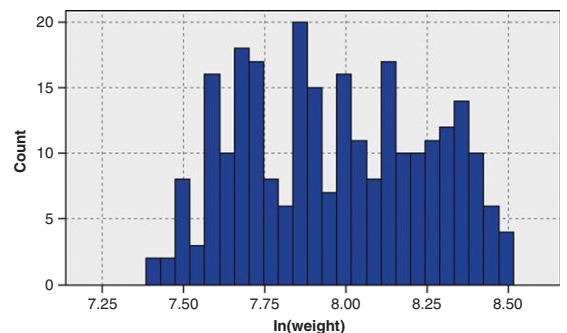


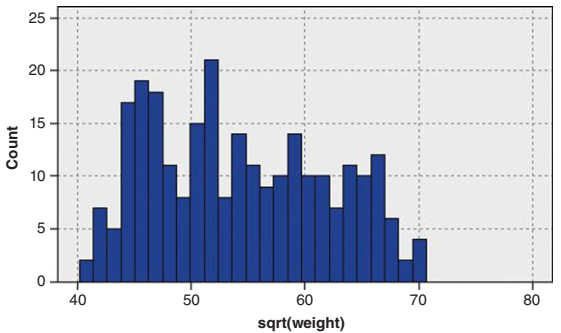


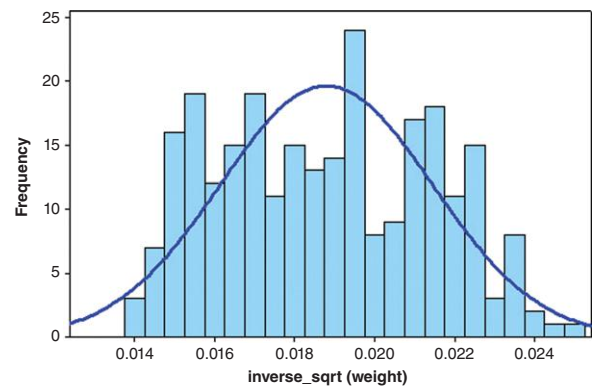
دیده میشود که نرمال سازی داده ها بر روی چولگی آنها تاثیر گذار نیست.

برای اینکه داده‌های خود را «به‌طور نرمال توزیع کنیم»، ابتدا باید آن‌ها را متقارن کنیم، که به معنای حذف چولگی است. برای حذف چولگی، یک تبدیل به داده ها اعمال می کنیم. تبدیل های رایج عبارتند از تبدیل ln(weight)، تبدیل ریشه مربع و تبدیل ریشه مربع معکوس استفاده از تبدیل ریشه مربع تا حدودی چولگی را کاهش می دهد، در حالی که اعمال تبدیل ln(weight) چولگی را حتی بیشتر کاهش می دهد.

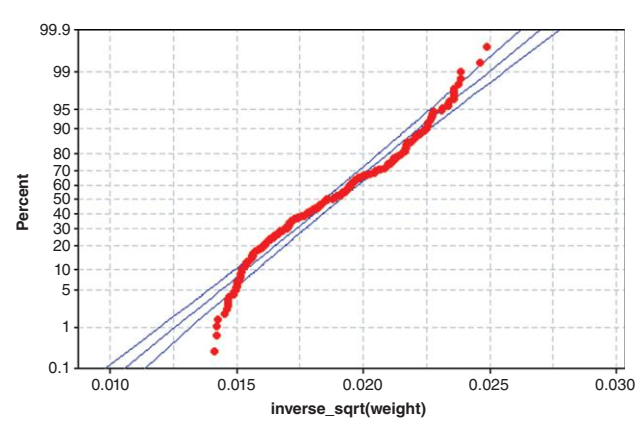




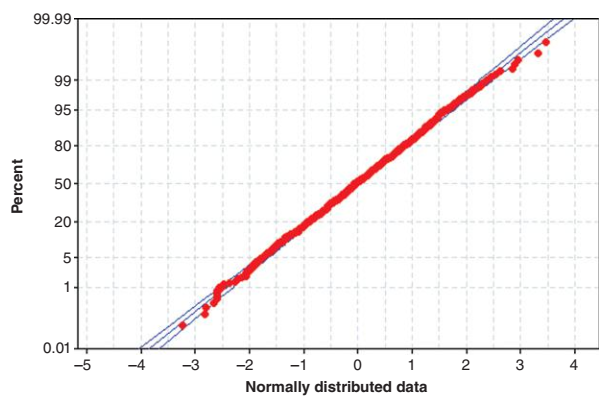


اکنون، هیچ چیز جادویی در مورد تبدیل ریشه مربع معکوس وجود ندارد. اتفاقی برای مقدار چولگی موجود در این متغیر کار می کند. اگرچه به تقارن دست یافته ایم، اما هنوز به حالت نرمال نرسیده ایم. برای بررسی نرمال بودن، یک نمودار احتمال نرمال می سازیم که چندک های یک توزیع خاص را در برابر چندک های توزیع نرمال استاندارد ترسیم می کند.

در نمودار احتمال نرمال، اگر توزیع نرمال باشد، بخش عمده ای از نقاط در نمودار باید روی یک خط مستقیم قرار گیرند. انحرافات سیستماتیک از خطی بودن در این نمودار نشان دهنده غیرعادی بودن است.



آزمایش با تبدیل های بیشتر (نشان داده نشده) نرمال قابل قبولی را برای به همراه نداشت. خوشبختانه، الگوریتم‌هایی که نیاز به نرمال بودن دارند، معمولاً وقتی با داده‌هایی که متقارن و تک‌وجهی هستند، کار می‌کنند.



این نکته قابل بیان است که پس از استفاده از تبدیل، برای تحلیل داده ها، باید دیتا را به مقادیر اصلی خود بازگردانیم.