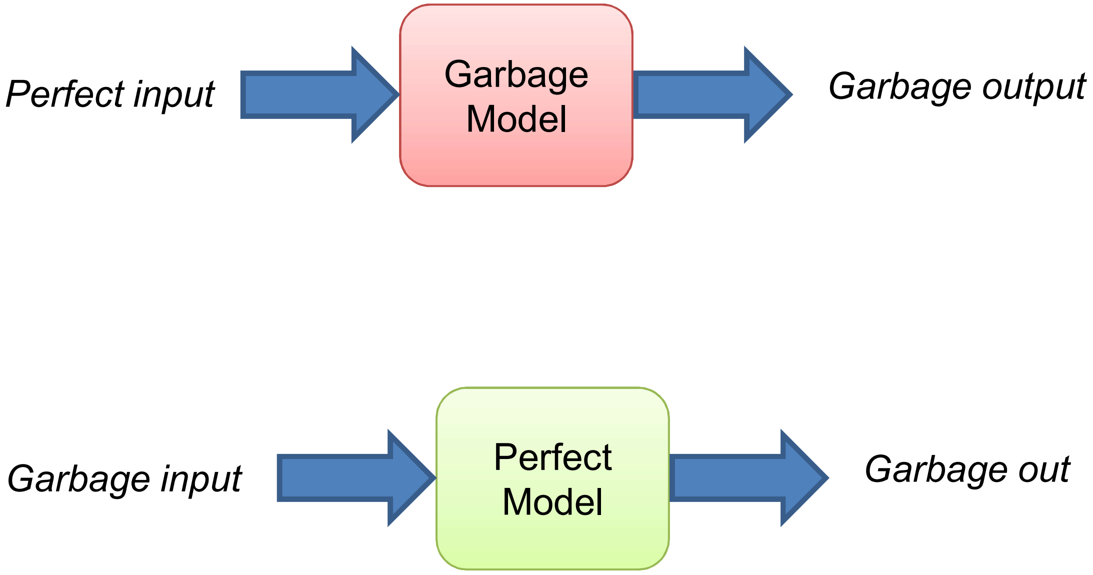
Preparing Data and Feature Engineering

Week 2

Mohammad Esmalifalak

Why Feature Engineering or Data Cleaning?



Perfect input

Garbage Model

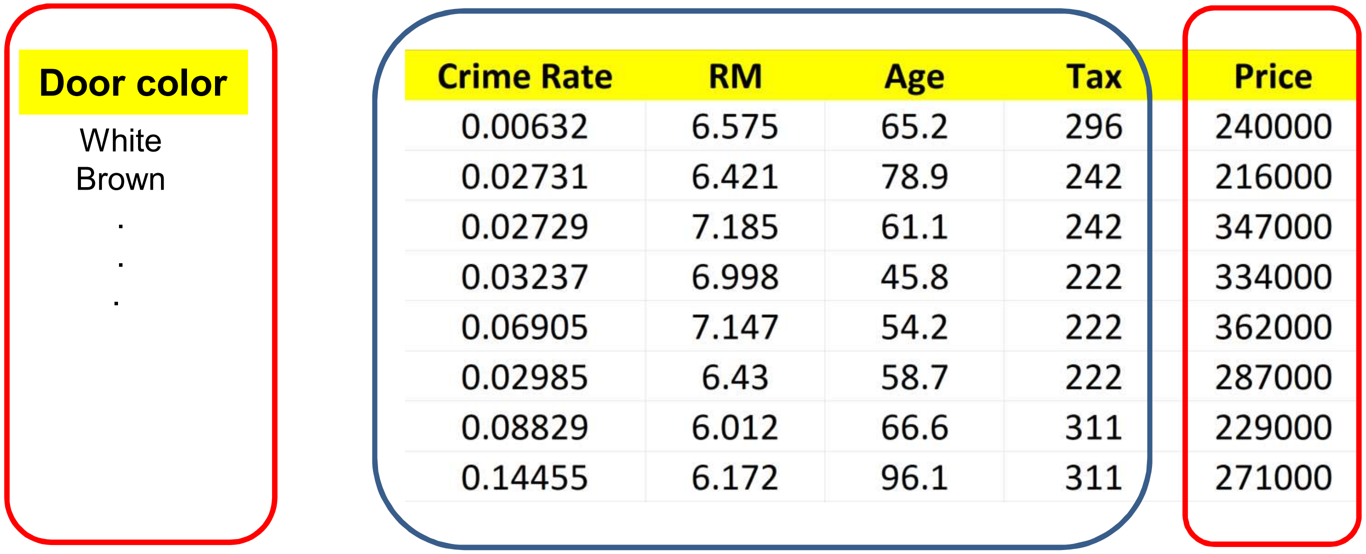
Garbage output

Garbage input

Perfect Model

Garbage out

Feature Engineering

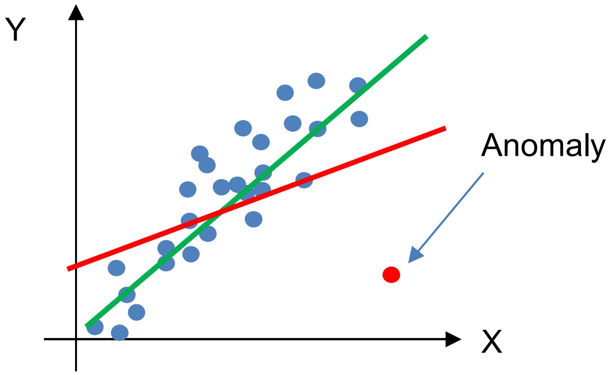


Door color
White
Brown
.
.
.

Crime Rate	RM	Age	Tax	Price
0.00632	6.575	65.2	296	240000
0.02731	6.421	78.9	242	216000
0.02729	7.185	61.1	242	347000
0.03237	6.998	45.8	222	334000
0.06905	7.147	54.2	222	362000
0.02985	6.43	58.7	222	287000
0.08829	6.012	66.6	311	229000
0.14455	6.172	96.1	311	271000

- **Feature selection:** Deciding which data to collect (Domain Knowledge)
- **Feature creation:** Combinations of features

Data Cleaning



Removing Anomalies

Feature Normalization

RM	Age	Tax	Price
6.575	65.2	296	240000
6.421	78.9	242	216000
7.185	61.1	242	347000
6.998	45.8	222	334000
7.147	54.2	Nan	362000
‘6.43’	58.7	222	287000
6.012	66.6	311	229000
6.172	96.1	311	271000

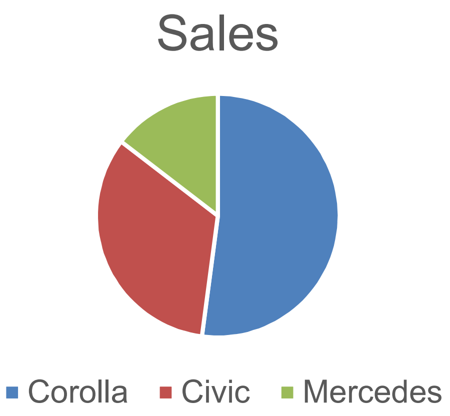
Strings (Convert to int. or Float)

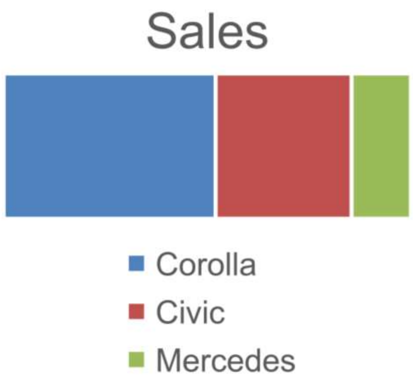
Missing values

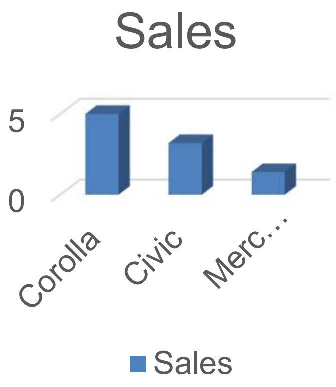
Types of Data-Level of Measurement

Categorical (Nominal, Qualitative):

- Don’t have order (e.g. Sex, Preferred type of car, Color)
- Can be summarized by frequency of observation for each category
- Not possible to calculate mean







Slide 5

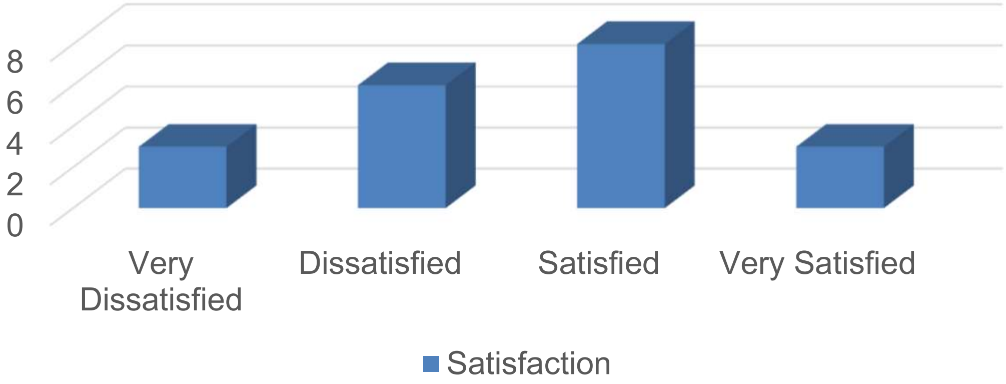
7 really glad you included this
-Noelle Ibrahim
, 5/8/2018

2 Thanks
Haitham Alhajj, 5/8/2018

Types of Data-Level of Measurement

Ordinal:

- Have meaningful order (e.g. Rank, Satisfaction, Fanciness)
- Can be summarized by frequency of observation
- Usually not good idea to get mean



Slide 6

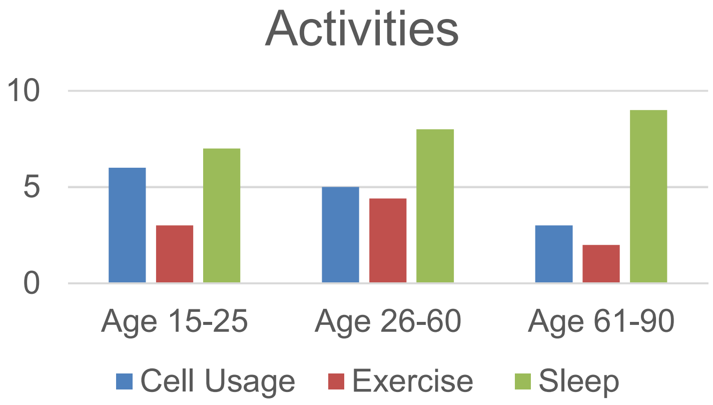
7 really glad you included this
-Noelle Ibrahim
, 5/8/2018

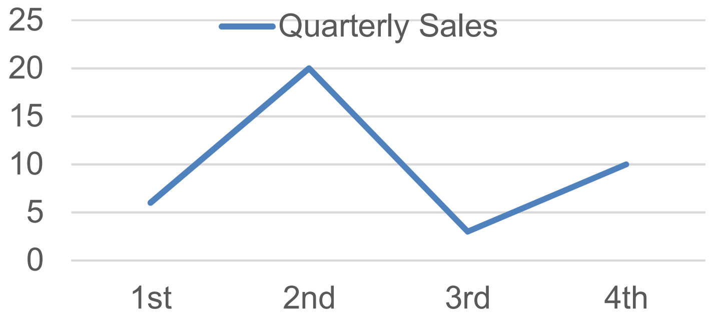
2 Thanks
Haitham Alhajj, 5/8/2018

Types of Data-Level of Measurement

Interval/Ratio (Scale, Quantitative, Parametric):

- Can be measured (rather than classified or ordered).
- Can be discrete (#of costumers, age) or continuous (distance, weight)
- Common summary measures are, mean, median, standard deviation





Slide 7

7 really glad you included this
-Noelle Ibrahim
, 5/8/2018

2 Thanks
Haitham Alhajj, 5/8/2018

Dealing with Missing Values - Methods

Replace with mean, median, interpolation, remove from dataset : each of these choices is associated with various trade-offs.

Method	Strengths	Weaknesses	Python

Mean	Averages of numerical data are used in many computations that will need to be done with numerical data, replacing with average value will not distort	Skewed by outliers
		Only meaningful for rational data types (i.e. float, int) and possibly interval, but not nominal (categorical) or ordinal
			mean_age = df.Age.mean()
			df.Age = df.Age.fillna(mean_age)

Slide 8

3 This slide and the next one are very nice and efficient. But again, there are too much information per slide. O the other hand, students are very good in absorbing concepts when we show them plots and figures rather than only words and sentences. So, I suggest to put each method in a separate slide and show two plots of two data sets for which one of them the method is working and another one themethod is not.
-hossein taghinejad
, 5/8/2018

Dealing with Missing Values - Methods

Replace with mean, median, interpolation, remove from dataset : each of these choices is associated with various trade-offs.

Method	Strengths	Weaknesses	Python
Median	Robust to outliers	May not be appropriate for datasets with “skewed’ distributions (i.e. poisson) in certain applications	med_age = df.Age.median()

Slide 9

4 This slide and the next one are very nice and efficient. But again, there are too much information per slide. O the other hand, students are very good in absorbing concepts when we show them plots and figures rather than only words and sentences. So, I suggest to put each method in a separate slide and show two plots of two data sets for which one of them the method is working and another one themethod is not.

-hossein taghinejad

, 5/8/2018

Dealing with Missing Values - Methods

Replace with mean, median, interpolation, remove from dataset : each of these choices is associated with various trade-offs.

Method	Strengths	Weaknesses	Python
Mode	Using the most frequent value in a large dataset will not usually distort the average or other values too greatly	May not be appropriate for datasets with kurtosis (i.e. stable distributions) in certain applications	mod_age = df.Age.mode()[0] *Mode returns a series

Slide 10

5 This slide and the next one are very nice and efficient. But again, there are too much information per slide. O the other hand, students are very good in absorbing concepts when we show them plots and figures rather than only words and sentences. So, I suggest to put each method in a separate slide and show two plots of two data sets for which one of them the method is working and another one themethod is not.

-hossein taghinejad

, 5/8/2018

Dealing with Missing Values - Methods

Replace with mean, median, interpolation, remove from dataset : each of these choices is associated with various trade-offs.

Method	Strengths	Weaknesses	Python

Remove	Does not introduce any bias if missing values are randomly distributed	Selection bias may occur if missing values are concentrated among population subgroups (i.e. mostly older or mostly younger patients in a medical database)	A = df.dropna(how='all')
			df.dropna(how='any')
			df.dropna(thresh=2)

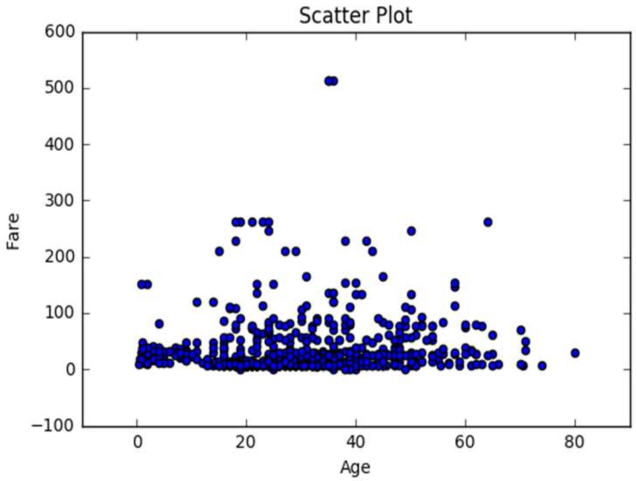
Pandas Data Frames and Useful Commands

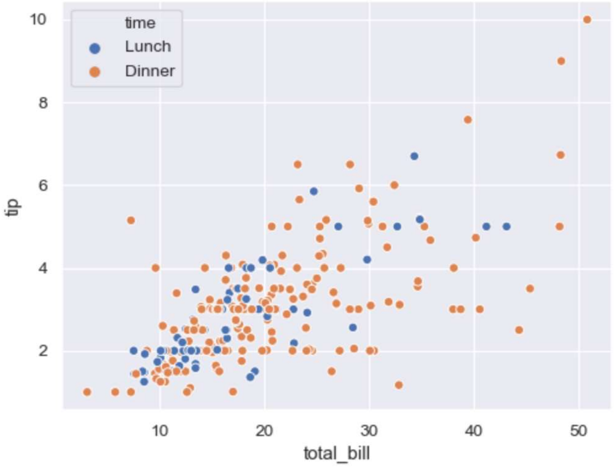
Some commands that can be used to understand your data set. This can help you decide how to best handle missing values and other data cleaning tasks we will discuss in this lecture

- df.head() (method)
- df.tail() (method)
- df.columns (attribute)
- df.shape (attribute)
- df.info()
- df.column.value_counts(dropna=False)
- df.describe()

Visualizing Data with Graphs

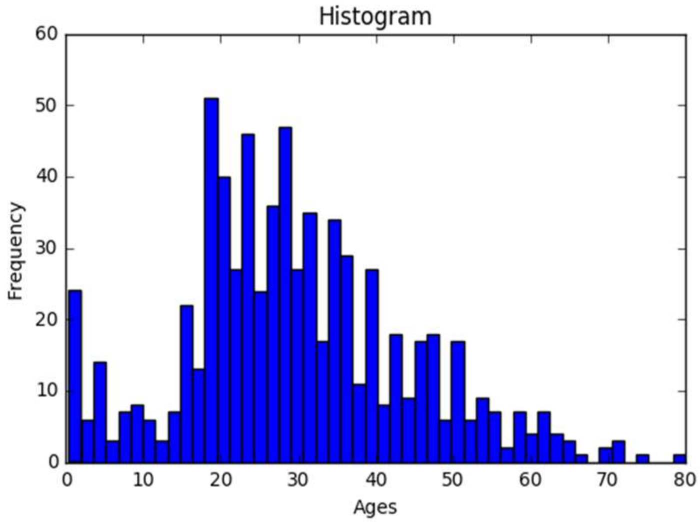
Plot	Strengths	Weaknesses	Python
Scatter	Visualize relationships between variables, see outliers	Useful only for 2 or 3 variables at a time	**# Matplotlib** Import matplotlib.pyplot as plt plt.scatter(x, y) **# Seaborn** Import seaborn as sns Sns.scatterplot(x=…,y=…,hue=…)

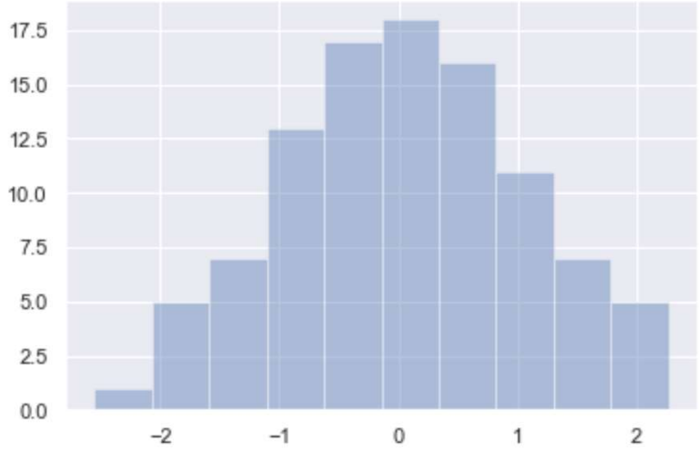




Visualizing Data with Graphs

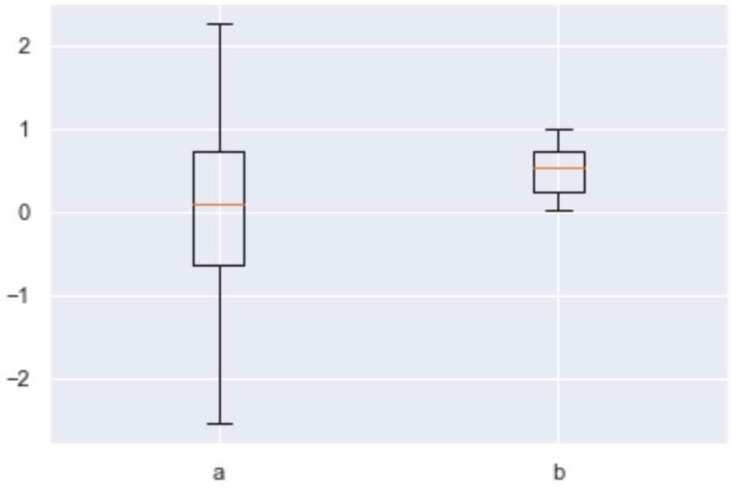
Plot	Strengths	Weaknesses	Python
Histogram	Find distribution of a single variable, see outliers	Selecting Bin size	**# Matplotlib** Import matplotlib.pyplot as plt plt.hist(x, bins=10) **# Seaborn** Import seaborn as sns sns.distplot(x, bins=10, kde=False)

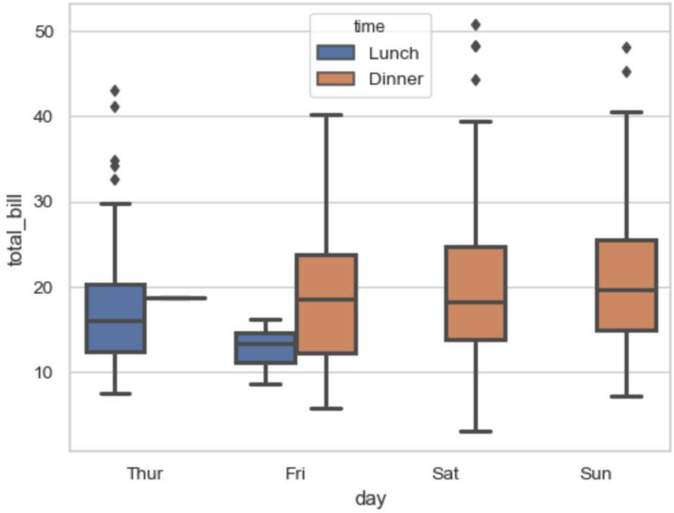




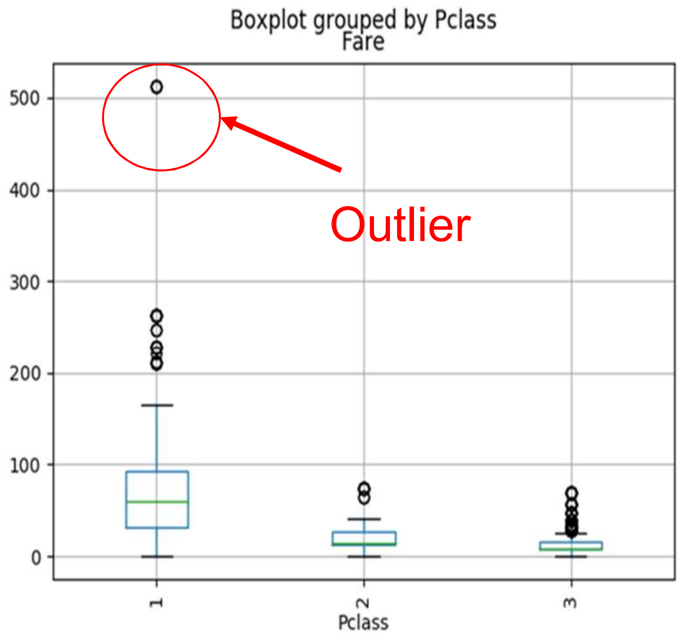
Visualizing Data with Graphs

Plot	Strengths	Weaknesses	Python
Box Plot	Visualize coarse grained distribution of numerical data by category, see outliers Can handle several categorical variables	Distribution is only min, max, quartiles and outliers	**# Matplotlib** Import matplotlib.pyplot as plt plt.boxplot([x ,y],labels=['a','b']) plt.legend **# Seaborn** Import seaborn as sns sns.boxplot(x="day", y="total_bill", hue="time“)



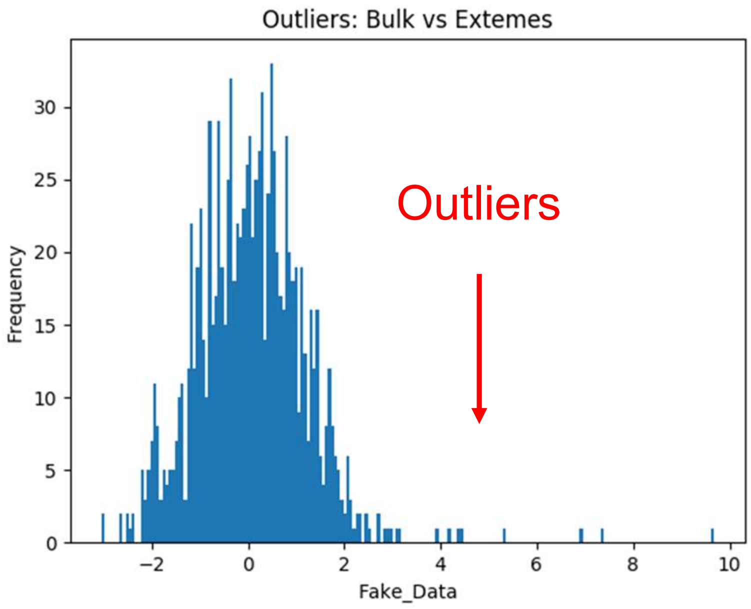


Data cleaning/Outliers



Outlier

May be found using visual inspection



Outliers

Having less than a certain probability of occurring

Categorical Data

Label Encoding:

$$\begin{bmatrix}\text{Toyota}\\\text{Civic}\\\text{Civic}\\\vdots\\\text{F-150}\end{bmatrix}\begin{matrix}\rightarrow\\\rightarrow\\\\\\\rightarrow\end{matrix}\begin{bmatrix}1\\2\\2\\\vdots\\5\end{bmatrix}$$

Five categories → Five Numbers.

Pros: One new column
Cons: Order could confuse algorithm

```
df[“C3"] = df[“C3"].astype('category')
df[“Cn"] = df[“C3"].cat.codes
```

One-Hot Encoding:

$$\begin{bmatrix}\text{Red}\\\text{Blue}\\\text{Blue}\\\vdots\\\text{Green}\end{bmatrix}\begin{matrix}\rightarrow\\\rightarrow\\\\\\\rightarrow\end{matrix}\begin{bmatrix}1&0&0\\0&1&0\\0&1&0\\&\vdots&\\0&0&1\end{bmatrix}$$

Three categories → Three columns

Pros: Doesn’t have any order
Cons: Can create lots cols.+rows

```
pd.get_dummies(df, columns=[‘Sex’,’Color’])
```

McMaster
University
CENTRE FOR
CONTINUING EDUCATION