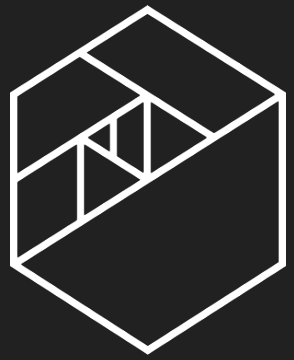


The Data Science Workflow:
Data Cleaning + Exploratory Data Analysis

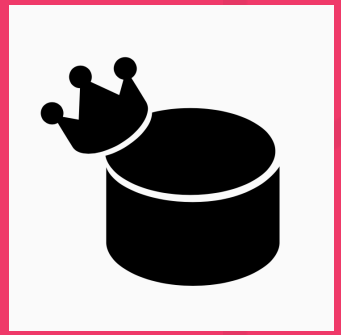
DATA SCIENCE WORKFLOW

DATA SCIENCE WORKFLOW

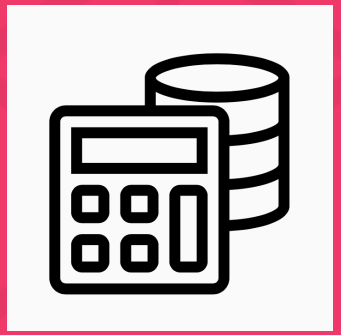
Data Cleaning

METIS

WHY IS DATA CLEANING SO IMPORTANT?



“Data is king”



“Data >> Features >> Algorithms”

“Garbage in, garbage out”

HOW CAN DATA BE MESSY?

1. Duplicate or unnecessary data
2. Inconsistent text and typos
3. Missing data
4. Outliers

… and more!

1. DUPLICATE OR UNNECESSARY DATA

- Let's say I'd like to do some analysis on Metis students

Name	Applied Date	Status	Campus
Alice	Jan 2018	Enrolled	Chicago
Bob	Feb 2018	Enrolled	Chicago
Charlie	June 2017	Rejected	NYC
Charlie	Jan 2018	Enrolled	NYC
Eve	Jan 2018	Enrolled	NYC
Frank	Feb 2018	Deferred	Seattle
Grace	Dec 2017	Enrolled	Seattle
Henry	Jan 2018	Enrolled	SF

1. DUPLICATE OR UNNECESSARY DATA

- Let's say I'd like to do some analysis on Metis students

Name	Applied Date	Status	Campus
Alice	Jan 2018	Enrolled	Chicago
Bob	Feb 2018	Enrolled	Chicago
Charlie	June 2017	Rejected	NYC
Charlie	Jan 2018	Enrolled	NYC
Eve	Jan 2018	Enrolled	NYC
Frank	Feb 2018	Deferred	Seattle
Grace	Dec 2017	Enrolled	Seattle
Henry	Jan 2018	Enrolled	SF

1. DUPLICATE OR UNNECESSARY DATA

- Keep an eye out for duplicate values and **dig into why** there are multiple values

- It’s a good idea to look at the features you’re bringing in and **filter down the data** as necessary (although be careful not to filter too much if you may use the features at a later point)

2. INCONSISTENT TEXT AND TYPOS

- Let’s say I’d like to do some analysis on Metis students

Name	Applied Date	Status	Campus
Alice	Jan 2018	Enrolled	Chicago
Bob	Feb 2008	Enrolled	Chicago
Charlie	June 2017	Rejected	NYC
Charlie	Jan 2018	Enrolled	NYC
Eve	Jan 2018	Enrolled	new york city
Frank	Feb 2018	Deferred	Seattle
Grace	Dec 2017	Enrolled	seattle
Henry	Jan 2018	Enrolled	SF

2. INCONSISTENT TEXT AND TYPOS

- Let’s say I’d like to do some analysis on Metis students

Name	Applied Date	Status	Campus
Alice	Jan 2018	Enrolled	Chicago
Bob	Feb 2008	Enrolled	Chicago
Charlie	June 2017	Rejected	NYC
Charlie	Jan 2018	Enrolled	NYC
Eve	Jan 2018	Enrolled	new york city
Frank	Feb 2018	Deferred	Seattle
Grace	Dec 2017	Enrolled	seattle
Henry	Jan 2018	Enrolled	SF

2. INCONSISTENT TEXT AND TYPOS

- Look at some summary statistics for each column
    - For numerical fields, what are the **minimum and maximum values** - do they make sense?
    - For categorical fields, what are the **unique values** - can some values be grouped together?

3. MISSING DATA

- Let’s say I’d like to do some analysis on Metis students

Name	Applied Date	Status	Campus
Alice	Jan 2018	Enrolled	Chicago
Bob	—-	Enrolled	Chicago
Charlie	June 2017	Rejected	NYC
Charlie	Jan 2018	Enrolled	NYC
Eve	Jan 2018	Enrolled	new york city
Frank	Feb 2018	Deferred	Seattle
Grace	Dec 2017	Enrolled	seattle
Henry	—-	—-	—-

3. MISSING DATA

- Let’s say I’d like to do some analysis on Metis students

Name	Applied Date	Status	Campus
Alice	Jan 2018	Enrolled	Chicago
Bob	—-	Enrolled	Chicago
Charlie	June 2017	Rejected	NYC
Charlie	Jan 2018	Enrolled	NYC
Eve	Jan 2018	Enrolled	new york city
Frank	Feb 2018	Deferred	Seattle
Grace	Dec 2017	Enrolled	seattle
Henry	—-	—-	—-

3. MISSING DATA

- Things to do about missing data
  - **Remove** the row(s) entirely
  - **Impute** the data = replace with substituted values
    - Fill in the missing data with the most common value, the average value, etc.

- What are the pros and cons of each of these approaches?

4. OUTLIERS

- An outlier is an observation in data that is **distant** from most other observations
- Typically, these observations are aberrations and **do not accurately represent** the phenomenon we are trying to explain through the model
- If we do not identify and deal with outliers, they can have **a significant impact** on the model

HOW TO FIND OUTLIERS

1. Plots

Histogram

Density Plot

Box Plot

2. Statistics

Interquartile Range

Standard Deviation (normally distributed data)

3. Residuals

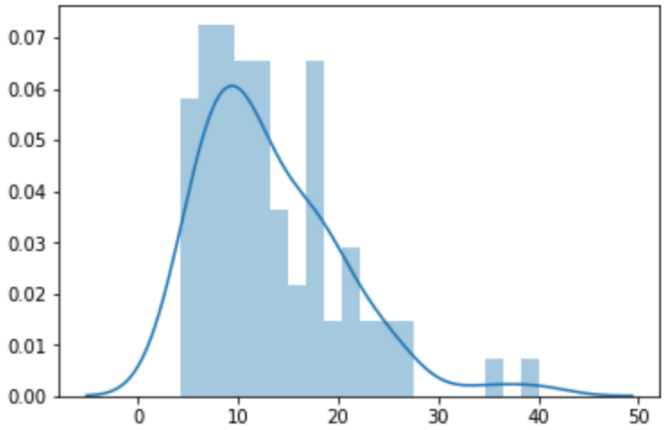
Studentized Residual

Deleted Residual

(for regression problems)

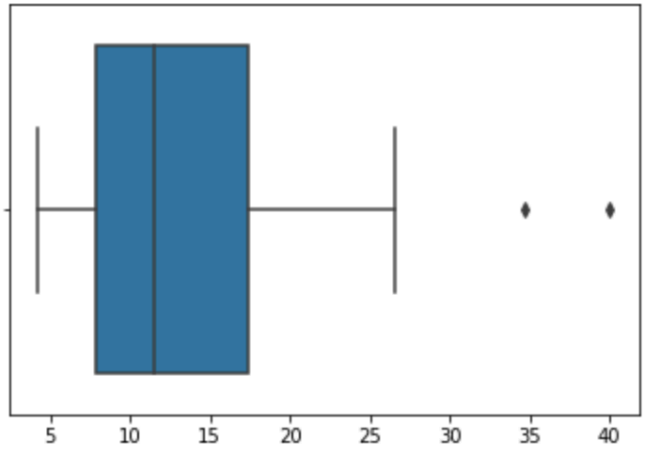
1. PLOTS

```
# plot a histogram and density plot
sns.distplot(data, bins=20);
```



HISTOGRAM

```
# plot a boxplot
sns.boxplot(data);
```



BOX PLOT

2. STATISTICS

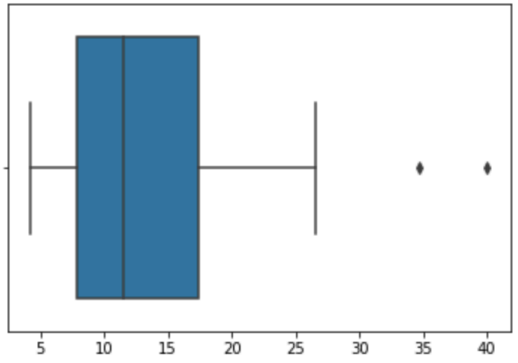
INTERQUARTILE RANGE

```
import numpy as np

# calculate the interquartile range
q25, q50, q75 = np.percentile(data, [25, 50, 75])
iqr = q75 - q25

# calculate the min / max limits to be considered an outlier
min = q25 - 1.5*(iqr)
max = q75 + 1.5*(iqr)

print(min, q25, q50, q75, max)
```

```
-6.6 7.8 11.5 17.4 31.8
```

```
# identify the points
[x for x in data['Unemployment'] if x > max]
```

```
[40.0, 34.700000000000003]
```



STANDARD DEVIATION

```
import numpy as np

# calculate the mean and standard deviation
mean = float(np.mean(data))
sd = float(np.std(data))

# identify cases where points are 2-3 standard deviations away
[x for x in data['Unemployment'] if (x < mean - 2*sd) or (x > mean + 2*sd)]
```

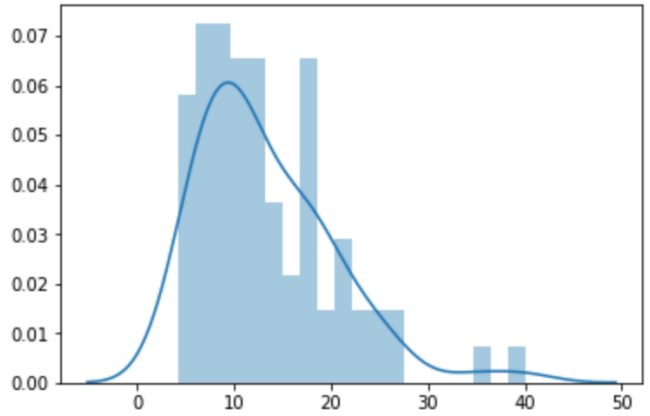
```
[40.0, 34.700000000000003]
```

HOW TO DEAL WITH OUTLIERS

- Remove them
- Assign the mean or median value
- K-nearest neighbors
- Use regression to try and predict what the value should be
- Transform the variable

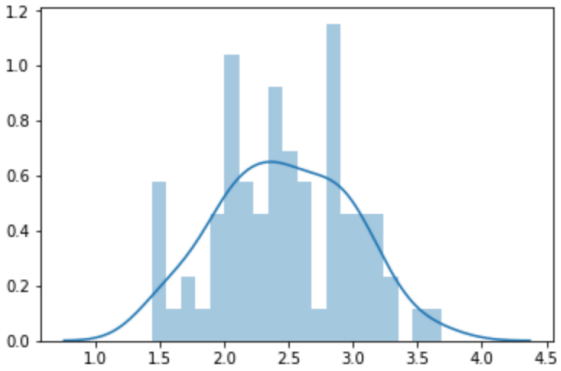
THE POWER OF TRANSFORMATIONS


```
# plot a histogram and density plot
sns.distplot(data, bins=20);
```



RIGHT SKEWED

```
import math
log_data = [math.log(d) for d in data['Unemployment']]

# plot transformed plots
sns.distplot(log_data, bins=20);
```



NORMAL!

HOW CAN DATA BE MESSY?

1. Duplicate or unnecessary data
2. Inconsistent text and typos
3. Missing data
4. Outliers

… and more!

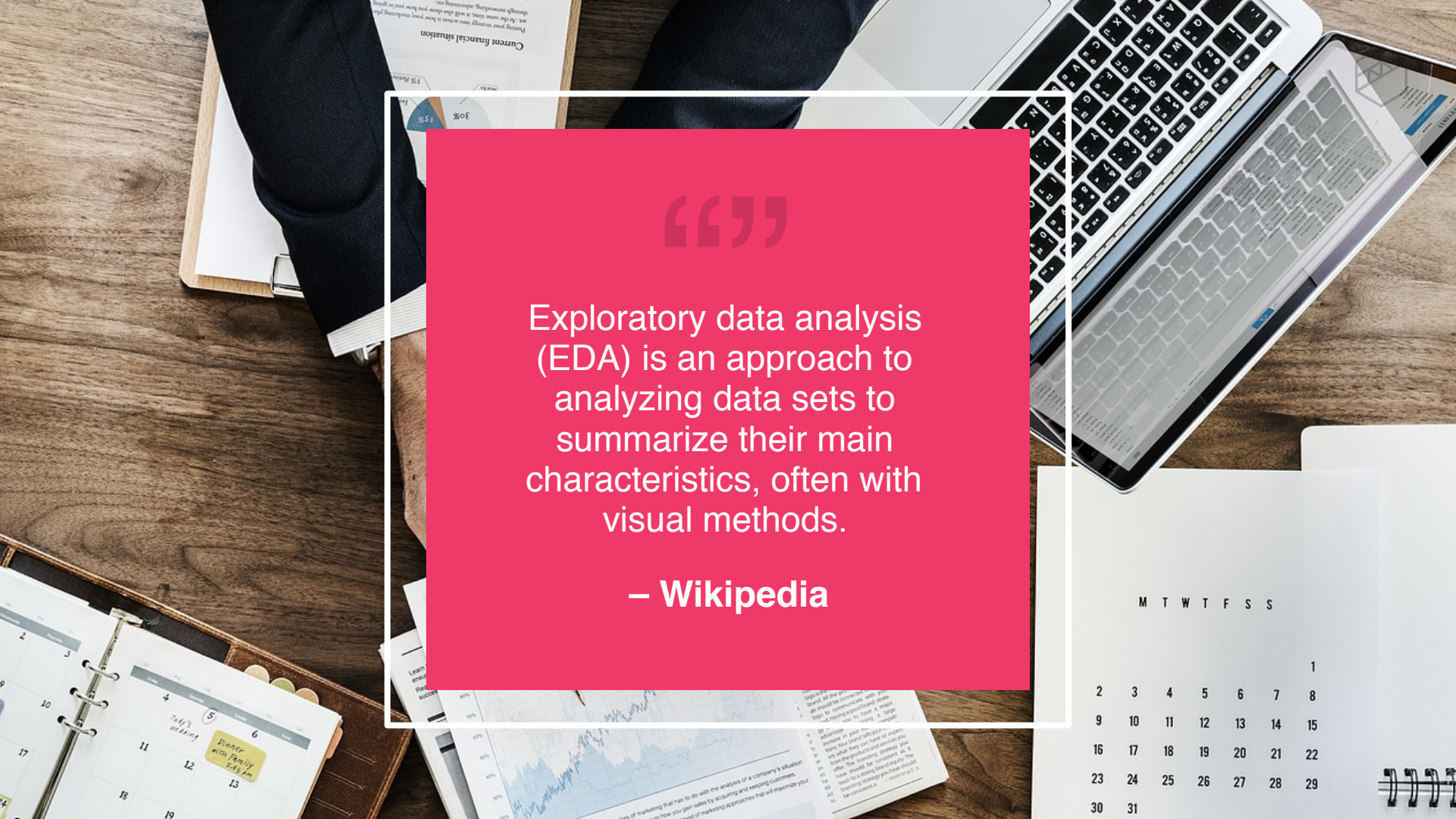
DATA SCIENCE WORKFLOW

DATA SCIENCE WORKFLOW

Exploratory Data Analysis

METIS

WHAT IS EXPLORATORY DATA ANALYSIS?



Exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.

– Wikipedia

WHY IS EDA USEFUL?

- Get an initial feel for the data
- See if the data makes sense and if further cleaning or more data is needed
- Identify patterns and trends in the data - often these can be just as important as your findings from modeling

WHAT ARE SOME TECHNIQUES?

- **Summary Statistics**
  Average, Median, Min, Max, Correlations, etc.

- **Visualizations**
  Histograms, Scatter Plots, Box Plots, etc.

WHAT ARE SOME TOOLS?

- **Data Wrangling**
  - Pandas
- **Data Visualization**
  - Matplotlib
  - Seaborn

OUR QUESTION

- Let’s say I want to do some analysis to see which applicants get accepted into Metis

- As a class, can you brainstorm some ways you can explore this data using (1) statistics and (2) visualizations?

EDA: SUMMARY STATISTICS

- **Average**: I could look at the average of all student interview scores, or perhaps the average of student interview scores by city
- **Max**: I could look at the most common words that accepted vs rejected students use in their application
- **Correlation**: Take a look at the correlation between technical assessment grade and years of Python experience

EDA: VISUALIZATIONS

- **Histogram (numeric)**: Take a look at the distribution of number of years of work experience of our applicants

- **Bar Chart (categorical)**: Create a chart showing the number of applicants with each type of major

- **Scatter Plot**: Create a scatter plot comparing the technical assessment grade and years of Python experience

WHY IS EDA USEFUL?

- Get an initial feel for the data
- See if the data makes sense and if further cleaning or more data is needed
- Identify patterns and trends in the data - often these can be just as important as your findings from modeling

Summary

METIS

SUMMARY

Data Cleaning

Data is king

Lots of time here

May repeat this

EDA

Summary Statistics

Visualizations

Gut check before modeling

ADDITIONAL RESOURCES

- **Book:** Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython by Wes McKinney

- **Website:** Exploratory Data Analysis by DataCamp
  Don’t need to pay for the course, but the course outline shows even more techniques you can use

Up Next...