**WEEK 1:**

1. Exploratory data analysis (EDA): the process of investigating, organizing, and analyzing datasets and summarizing their main characteristics, often employing data wrangling and viz methods.

* 6 EDA practices:
* Discovering
* Structuring
* Cleaning
* Joining
* Validating
* Presenting

1. Discovering: data professionals familiarize themselves with the data so they can start conceptualizing how to use it.

* They review data n ask questions about it
* “what’s the column headers?”, “what do they mean?”, “how many total data points are there?”

Structuring: the process of taking raw data and organizing or transforming it to be more easily visualized, explained, or modeled.

* Refers to categorizing and organizing data columns based on data already in the data set.

Bias (in data structuring): organizing data in groupings, categories, or variables that don’t accurately represent the whole dataset.

Cleaning: the process of removing errors that may distort your data or make it less useful.

* Missing values, misspelling, duplicate entries, extreme outliers,…

Joining: the process of augmenting or adjusting data by adding values from other datasets.

* You might add more value or context to the data by adding more info from other data sources.

Validating: the process of verifying that the data is consistent and high quality.

* Checking misspelling and inconsistent numbers or data formats, checking data cleaning process didn’t introduce more errors.
* Use R, Java script, Python

Presenting: making your cleaned dataset or data viz available to others for analysis or further modeling.

* Is sharing what you’ve learned thro EDA and asking for feedback whether in the form of a clean data set or data viz.

Data viz: a graph, chart, diagram, or dashboard that is created as a representation of info.

* Data viz should be used throughout the EDA, not only for representing part
* Help to understand data and point out trends and insights to others.

**Combine PACE and EDA practices**

1. PACE: Plan, Analyze, Construct, Execute.

* EDA applies to every part of PACE

A white and black text on a white background

Description automatically generatedA diagram with colorful squares

Description automatically generatedA white text on a white background

Description automatically generatedA screenshot of a computer

Description automatically generated

**PACE with data visualizations**

* The manufacturing supervisors
* The executive leadership team
* Data viz are designed to communicate different messages to stakeholders will vested interests in different parts of the project.
* Your data driven storytelling is an opportunity to present facts and viz that are ethical, accessible and representative of the data.

A white text with black text

Description automatically generated

**[ end of WEEK 1…]**

**WEEK 2:**

**Where the data comes from**

1. Data source: the location where data originates

* Know how and when to contact subject matter experts such as engineers or database owners,
* Help to tell story of data and make ethical decisions about its use.
* Understanding how the data is collected
* Report from a computer system
* Selection from a large online database
* Data table that has been manually entered

1. Data formats:

* Tabular files
* XML files
* CSV files
* Excel files
* DB files
* JSON files (JavaScript Object Notation)

**EXCEL & TABULAR:**

**A screenshot of a spreadsheet

Description automatically generated**

* Organized by ROWS and COLUMNS:
* Rows representing the objects
* Columns representing aspects.
* Adv: a clear identification of patterns between variables.

**CSV**

A screenshot of a computer

Description automatically generated

* In a form of text file
* Easy to import or store other software, platforms, and databases.
* Rows of text and numbers separated by commas
* Adv: is a file type which is easily read even in a text editor

Easy to create and manipulate

**DataBase files (DB)**

A screenshot of a computer

Description automatically generated

* Storing data in tables, indexes, or fields
* Great for searching and storing
* SQL knowledge

**JSON files**

A screenshot of a phone number

Description automatically generated

* Data storage files that are saved in JavaScript format
* Provide nested info

1. Data stories in formats of HTML, audio files, photos, email, text images, or text files.
2. Data types: First-, Second-, Third- party data

* First-party data: data that was gathered from inside your own organization.
* second-party data: data that was gathered outside your organization but directly from the original source.
* third-party data: data gathered outside your organization and aggregated.
* Others: Geographic, Demographic, Numeric, Time-based, Financial, Qualitative data.
* 2 IMPORTANT QUESTIONS:
* Given what you know of the data so far, does it align with the PLAN, as defined by your PACE workflow?
* do you have enough data to follow through with the plan in the PACE workflow?

1. **df.head():** will return as many rows of data as you input in the argument field.
2. **info():** gives the total number and data types of individual entries. Keep in mind that data types are called Dtypes in pandas.
3. **Int64:** means INTEGER 64, means standard integer somewhere between negative nine quintillion and positive nine quintillion.

A number and numbers on a white background

Description automatically generated

1. **Strings**: a sequence of characters or integers that are unchangeable.
2. Helpful Python methods for ‘discovering’:

* .describe()
* .sample()
* .size
* .shape

are all useful to learn about dataset.

1. Hypothesis: a theory or an explanation, based on evidence, that is not yet proved true.
2. Questions during the discovering process:

* “how can I break this data into smaller groups so I can understand it better?”
* “how can I prove my hypothesis?”
* “in its current form, can this data give me the answers I need?”

1. Organize or alter data:

* Regroup entries into months/years or age ranges
* Group customer ages into age ranges
* Combine or split data columns
* Change data formats or time zones

1. **.strftime():** are datetime data into new strings.

Code it as %Y-W%V

* % sign is the command which tells the datetime format to use the year data in the string.
* W implies this is a week, and V stands for values.

**Use structuring methods to establish order in your dataset**

1. Sorting: the process of arranging data into meaningful order for analysis.

* Contain characteristics

Extracting: the process of retrieving from a dataset or source for further processing.

Filtering: the process of selecting a smaller part of your dataset based on specified parameters and using it for viewing or analysis.

* selecting rows of a dataset

Slicing: a method for breaking info down into smaller parts to facilitate efficient examination and analysis from different viewpoints.

* Either or both options for columns and rows
* A combination of extraction and filtering

Grouping (bucketizing): aggregating individual observations of a variable into groups.

Merging: method to combine 2 different data frames along a specified starting column.