## What you'll learn

- Explore how to find and sculpt stories
- Discuss how PACE applies to telling stories using data
- Perform exploratory data analysis on a dataset in Python
- Data sources, data types, data structuring, and data cleaning
- Missing values, outliers, and categorial data
- Ethics of exploring and cleaning raw data
- Communicating your questions and findings
- Tableau
- Visuals and presentations
- Six practices of exploratory data analysis
- How PACE fits into the process of exploratory data analysis
- Importance of data visualization in the data exploration process

## Exploratory data analysis (EDA)

The process of investigating, organizing, and analyzing datasets and summarizing their main characteristics, often employing data wrangling and visualization methods

## Practices of EDA

- Discovering
- Structuring
- Cleaning
- Joining
- Validating
- Presenting

## Discovering

Data professionals familiarize themselves with the data so they can start conceptualizing how to use it

## Structuring

The process of taking raw data and organizing or transforming it to be more easily visualized, explained, or modeled

## 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

## Joining

The process of augmenting or adjusting data by adding values from other datasets

## Validating

The process of verifying that the data is consistent and high quality

## Presenting

Making your cleaned dataset or data visualizations available to others for analysis or further modeling

## Data visualization

The graph, chart, diagram, or dashboard that is created as a representation of information

## Data visualization Python Packages

- Matplotlib
- Seaborn
- Plotly

## Review

- Practices of EDA
- PACE workflow
- Ethics of working with data
- Data visualizations

## What you’ll learn

- Data types and data sources
- Discovering
- Structuring
- Formatting
- Workplace skills and PACE

## Understanding raw data

- Data sources
- Data formats
- Data types
- Python functions

## Data source

The location where data originates

## Examples of data source

- Report from a computer system
- Selection from a large online database

- Data table that has been manually entered

## Data formats

- Tabular files
- XML files
- CSV files
- Excel files, rows - objects, columns - aspects
- DB files
- JSON files

pandas.read_csv()

Import json
pandas.read_json
pandas.df.to_json()

## Types of data

- First-party data
- Second-party data
- 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

## Other types of data

- Geographic
- Demographic
- Numeric
- Time-based
- Financial
- Qualitative

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?

## Info()

Gives the total number and data types of individual entries. Keep in mind that data types are called Dtypes in pandas.

## Int64

Standard integer somewhere between negative nine quintillion and positive nine quintillion

## Strings

A sequence of characters or integers that are unchangeable

## Helpful Python methods for ‘discovering’

- .describe()

- .sample()
- .size()
- .shape()

## Hypothesis

A theory or an explanation, based on evidence, that is not yet proved true

## 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?

## Questions to ask

- Which months have the most passenger traffic?
- Which weeks, dates, or known holidays have the highest number of passengers?
- When are tickets typically purchased?

## Hypothesis

Analyze the data to understand whether the airlines would attract more customers by lowering prices on those specific fights

## Organize or alter data

- Regroup entries into months/years or age
- Group customer ages into age ranges
- Combine or split data columns

- Change data formats or time zones

## Sorting

The process of arranging data into meaningful order for analysis

## Extracting

The process of retrieving data 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

## Slicing

A method for breaking information down into smaller parts to facilitate efficient examination and analysis from different viewpoints

## Grouping

Aggregating individual observations of a variable into groups

## Merging

Method to combine two different data frames along a specified starting column

## Box plot

Data visualization that depicts the locality, spread and skew of groups of values within quartiles

## Merging

Combining two different data sources into one

## Review (Data structuring techniques)

- Data sources
- Data types
- Data formats
- How to use Python to uncover big picture understandings
- Date and time transformations in Python
- Sorting
- Extracting
- Filtering
- Slicing
- Joining
- Merging
- Grouping

## Review (Workplace skills)

- Timing for communicating updates
- Posting questions to project stakeholders, managers, and subject matter experts

## Review (Techniques for gathering data-driven insights)

- Making and testing hypotheses on your datasets

## What you'll learn

- Missing values
- Outliers
- Transforming categorical into numerical data

- Input validation

## Workplace skills

- When to communicate with stakeholders and engineers about missing or outliers values
- Ethical implications you must consider when dealing with missing and outlier data values

## Missing data are often encoded as

- N/A
- NaN (‘Not a number’)
- [blank]

## Missing data

A value that is not stored for a variable in a set of data

A zero (0) could be considered a missing value, but in other datasets could be a legitimate data point.

## What to do with missing data

- Request the missing values to be filled in by the owner of the data
- Delete the missing column(s), rows(s), or value(s)
- Create a NaN category
- Derive new representative values(s)

## Derive new representative value(s) strategy

- Forward filling
- Backward filling (backfilling)
- Deriving mean values

- Deriving median values

## Non-null count

The total number of data entries for a data column that are not blank

## Outliers

Observations that are an abnormal distance from other values or an overall pattern in a data population

## 3 types of outliers

- Global outliers
- Contextual outliers
- Collective outliers

## Global outliers

Values that are completely different from the overall data group and have no association with any other outliers

## Contextual outliers

Normal data points under certain conditions but become anomalies under most other conditions

## Collective outliers

A group of abnormal points that follow similar patterns and are isolated from the rest of the population

## Documentation string or Docstring

A line of text following a method or function that is used to explain to others, using your code, what this method or function does. A docstring represents good documentation practice in Python.

## Categorial data

Data that is divided into a limited number of qualitative groups

## Dummy variables

Variables with values of 0 or 1, which indicate the presence or absence of something

## Label encoding

Data transformation technique where each category is assigned a unique number instead of a qualitative value

## Heatmap

A type of data visualization that depicts the magnitude of an instance or set of values based on two colors

## Input validation

The practice of thoroughly analyzing and double-checking to make sure data is complete, error-free, and high-quality

## Why validate data?

- Make more accurate business decisions
- Improve complex model performance
- Prevent future system crashes, coding issues, or wrong predictions

## Questions to ask while validating data

- Are all entries in the same format?
- Are all entries in the same range?
- Are the applicable data entries expressed in the same data type?

## Joining

The process of augmenting data by adding values from other datasets

## Review

- EDA practices
- Missing data and outliers
- Categorical and numerical data
- Input validation
- Workplace skills
- Ethical considerations

## What you'll learn

- Improve your data visualizations skills
- Accurate representation aspects and techniques
- Techniques for making all your data visualizations accessible

## Working with Tableau

- How to create data visualizations in Tableau
- Explain technical concepts to non-technical audiences
- Create dynamic visualizations with interactive and motion elements
- Alter your visualizations based on different audiences needs

## Your analysis criteria

- The locations of rental units where the owners have 40 or more property listings
- Properties with many good reviews
- Price between 90 and 250 Euros

## Tableau—--------------------------------------------------------------

## Dimensions

Qualitative data values, used to categorize and group data to reveal details about it

## Measures

Numeric values that can be aggregated or placed in calculations

## Continuous

A measure or dimension has an infinite and uncountable number of outcomes

## Discrete

A measure or dimension has a finite and countable number of outcomes

## Heatmap

A type of data visualization that depicts the magnitude of an instance or set of values based on two colors

## Box plot

Data visualization that depicts the locality, spread and skew of groups of values within quartiles

## Histogram

A data visualization that depicts an approximate representation of the distribution of values in a dataset

## Bins

A Tableau term that describes the custom segments of data that values can be grouped into

## Basic organizing strategies for a presentation

- Chronological
- Generic-to-specific
- Specific-to-generic

A chronological approach to data visualizations is useful for data that is best understood in a time series.

A Generic-to-specific approach helps an audience consider an issue before describing how it affects them.

A Specific-to-generic approach is useful to highlight impacts the data can have on a broader scale.

## Story

A Tableau term for a group of dashboards or worksheets assembled into a presentation

## Set

A Tableau term for a custom field of data created from a larger dataset based on custom conditions

## Action

A Tableau tool to help a user interact with a visualization or dashboard by allowing control of a selection

## Review (Working with Tableau)

- Use Tableau to tell your data's stories
- Adjust content depending on audience
- Share technical concepts with non-technical audiences
- Use Tableau at a high level
- Applied techniques data professionals use to create
- Ensure visualizations accurately represent data
- Make data visualizations accessible for people who have

## Experiential learning

Understanding through doing

## Questions you might get in interviews

- What is your process for cleaning data?
- What tools do you use for creating data visualizations?
- How and why do data visualizations enhance the stories data tells?
- What considerations are top of mind when sharing data stories with non-technical stakeholders?

## Review

- How data professionals take care of missing data and outliers

- How to change categorical data into numerical data using the label encoding technique
- How to design visualizations and present your data in impactful ways
- Advanced concepts of visualizing
- Started using Tableau

## Workplace skills

- Communicating to different audience
- Importance of ethics
- The need for accessibility
- The importance of following the PACE workflow