Exploratory Data Analysis

The main goals of EDA are:

1. Provide summary level insight into a data set
2. Uncover underlying patterns and structure in the data
3. Identify outliers, missing data and class balance issues
4. Carry out quality control checks

The principal steps in the process of EDA are:

1. **Summarize the data** - Generally done using dataframes in R or Python
2. **Tell the Story** - Summarize the details of what connects the dataset to the business opportunity
3. **Deal with missing data** - Identify the strategy for dealing with missing data
4. **Investigate** - Using data visualization and hypothesis testing delve into the relationship between the dataset and the business opportunity
5. **Communicate** - Communicate the findings from the above steps

Data visualization

- Jupyter notebooks in combination with pandas and simple plots are the basis for modern EDA when using Python as a principal language.

Advantages of Jupyter notebooks:

- They are portable: They can be used locally on private servers, public cloud, and as part of IBM Watson Studio
- They work with dozens of languages
- They mix markdown with executable code in a way that works naturally with storytelling and investigation

- matplotlib and its child libraries like seaborn are the core of the Python data visualization landscape
- pandas and specifically the dataframe class works naturally with Jupyter, matplotlib and downstream modeling frameworks like sk-learn

EDA and Data Visualization best practices

1. The majority of code for any data science project should be contained within text files. This is a software engineering best practice that ensures re-usability, allows for unit testing and works naturally with version control. In Python the text files can be executable scripts, modules, a full Python package or some combination of these.
2. Keep a record of plots and visualization code that you create. It is difficult to remember all of the details of how visualizations were created. Extracting the visualization code to a specific place will ensure that similar plots for future projects will be quick to create.

3. Use you plots as a quality assurance tool. Given what you know about the data it can be useful to make an educated guess before you execute the cell or run the script. This habit is surprisingly useful for quality assurance of both data and code.

Missing values

- Dealing with missing data sits at the intersection of EDA and data ingestion in the AI enterprise workflow
- Ignoring missing data may have unintended consequences in terms of model performance that may not be easy to detect
- Removing either complete rows or columns in a feature matrix that contain missing values is called **complete case analysis**
- Complete case analysis, although commonly used, can lead to undesirable results. The category or categories of missingness present in the data can significantly impact the quality of these results.

The categories of missingness are:

- **Missing completely at random or MCAR**
- **Missing at random or MAR**
- **Missing not at random or MNAR**

- The best case scenario is that the data are MCAR. It should be noted that imputing values under the other two types of missingness can result in an increase in bias.
- In statistics the process of replacing missing data with substituted values is known as **imputation**
- It is a common practice to perform multiple imputations
- The practice of imputing missing values introduces uncertainty into the results of a data science project
- One way to deal with that additional uncertainty is to try a range of different values for imputation and measure how the results vary between each set of imputations. This technique is known as **multiple imputation**

CASE STUDY: Data visualization

It can be easy to get lost in the details of the findings when communicating the finding from EDA to business stakeholders. Project planning and milestones are important so remember to talk about what you:

1. Have done
2. Are doing
3. And plan to do

- Remember that deliverables are generally a presentation or a report and they should use a portable format (e.g. PDF or HTML)
- Deliverables should be concise and clear. Appendices are useful as supplemental materials to a deliverable and they help keep them free of unnecessary items
- Visual summaries are a key component of EDA deliverables
- There is no single right way to communicate EDA, but a minimum bar is that the data summaries, key findings, investigative process, conclusions are made clear