Exploratory and Confirmatory Data Analysis using python

Key Takeaways

Task 1

Title: Introduction

- EDA or Exploratory Data Analysis is one of the data analysis methods where we use different statistical summaries and graphical representations to perform initial investigations on the data to discover interesting patterns, spot anomalies, and overall, for a better understanding of our data.
- EDA is used to see how our data can be useful.

Task 2

Title: Exploratory Data Analysis - Where to start?

- The first step of Data Exploration is to check what kinds of data types we are working with.
- Create a road map for your data exploration based on the different data types you have in your dataset.
- Having a list of different information types (Time, Place, Product, Sales, etc.) that are in your dataset always helps.

Task 3

Title: Data Exploration - Time and Costumer Information Aspect

- If you have datetime column in your data frame, make sure it has the datetime64 data type.
- To start your data exploration always check the time span of your data.
- If you have datetime column in your data frame, you can explore your data based on different granularity levels (Year, Month, Day, Hour, Minute and Second). For example, you can aggregate the profit gained based on different Years, Months and Days.
- Data aggregation is one of the key required knowledge of data exploration.
- Line charts are the most common visualization techniques used while working with time series data

Task 4

Title: Data Exploration - Geo Information

- Choropleth maps are the most common visualization techniques used, for exploring Geo Data.

Task 5

Title: Exploratory Data Analysis - Hierarchical Information about the products

- Sunburst Diagram and Treemap Diagram are two most common data visualization techniques that are used to explore hierarchical data.
- Exploring hierarchical data always can be very insightful. Try to find hierarchical information in your data.
- Time information is also a hierarchical information. You can use Treemap and sunburst diagrams to explore your data based on different hierarchical level (granularity level) such as year, month, day, hour, minute and even second.

Task 6

Title: Data Exploration - Distributional analysis of sales information columns

- You can apply distribution analysis to any numerical value column in your data.
- You can use statistical summaries to see if there are any outliers in your column.
- Histograms and Box plots are two visualization techniques used for distributional analysis.
- Always pay attention to the skewness of your histogram.
- Right-skewed histogram is telling you there are outliers in the right side of your data range. You can see the tail on the right side of your histogram
- Left-skewed histogram is telling you there are outliers in the left side of your data range. You can see the tail on the left side of your histogram

Task 7

Title: What is Confirmatory Data Analysis (CDA)?

- Confirmatory Data Analysis is the process of using statistical summary and graphical representations to evaluate the validity of an assumption about the data at hand.
- One of the popular data analysis methods is CDA. Where you make some assumptions about your data, and you start to validate them.