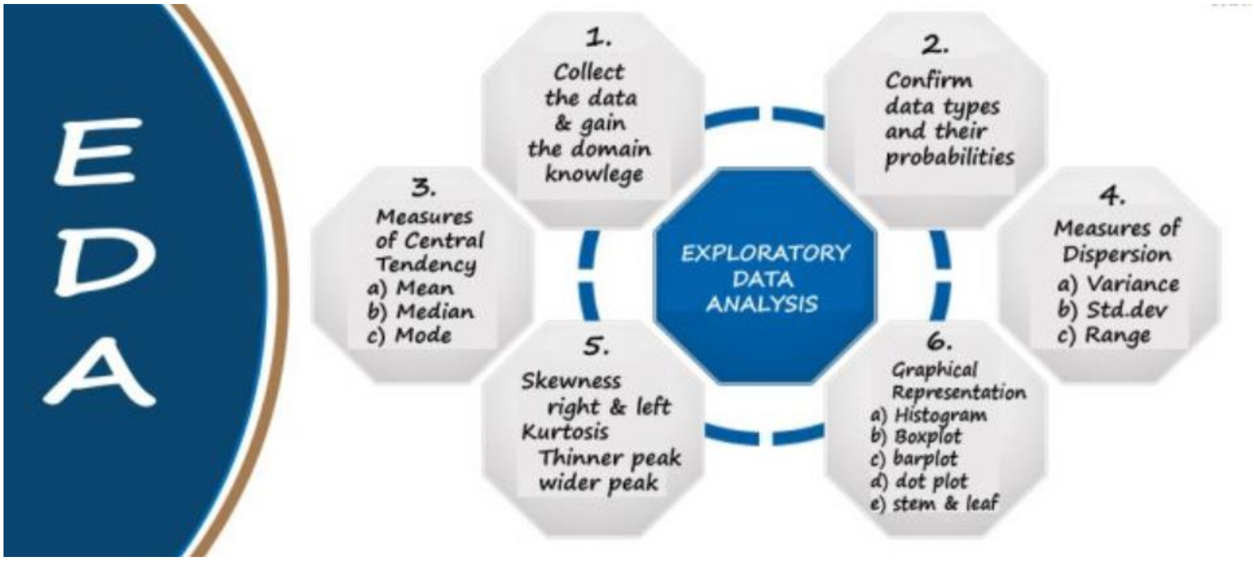
Chapter 2: Exploratory Data Analysis (Analysis)

Introduction

In the previous chapter, we learned about the aspects related to Exploratory Data Analysis. We were analyzing the datasets by making some investigations that were focusing on the structure, quality, content, and quantity of that data. Now in this chapter, we will be focusing on the more advanced structure of EDA. We will be practicing the concepts through exercises.

Exploratory data analysis is a data exploration approach that encourages statisticians to explore their data and possibly formulate hypotheses that might cause new data collection and experiments. EDA is a robust and efficient approach for checking that your data contains all the required information for model fitting and hypothesis testing. It also provides you with an understanding of how to handle missing values and make transformations of variables as needed.

It provides a solid base to build a robust understanding of the data, and issues associated with either the info or process. EDA is an innovative way to solve business problems by researching, planning, and creating solutions.


EDA
EXPLORATORY DATA ANALYSIS
1. Collect the data & gain the domain knowlege
2. Confirm data types and their probabilities
3. Measures of Central Tendency a) Mean b) Median c) Mode
4. Measures of Dispersion a) Variance b) Std.dev c) Range
5. Skewness right & left Kurtosis Thinner peak wider peak
6. Graphical Representation a) Histogram b) Boxplot c) barplot d) dot plot e) stem & leaf

Exploratory data analysis is cross-classified in two different ways where each method is either graphical or non-graphical. And then, each method is either **univariate, bivariate or multivariate.** These three methods will be covered in this chapter.

The data set that will be used in this chapter contains the following features:

- The name, state, district #, party, room #, phone #, and committee assignments of the US House of Representatives.

Also referred to as a congressman or congresswoman, each representative is elected to a two-year term serving the people of a specific congressional district. The number of voting representatives in the House is fixed by law at no more than 435, proportionally representing the population of the 50 states. Currently, there are five delegates representing the District of Columbia, the Virgin Islands, Guam, American Samoa, and the Commonwealth of the Northern Mariana Islands. A resident commissioner represents Puerto Rico. Learn more about representatives at The House Explained.

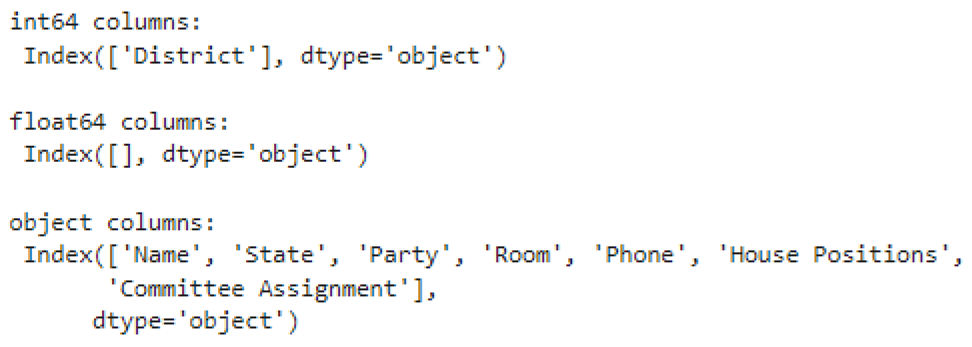
Univariate Analysis

Univariate analysis is a type of descriptive statistics that deal with one variable from a large amount of data. The objective is to derive the data; define and summarize it; and analyze its mean, variance, covariance, or correlation.
In a dataset, it explores each variable separately. It is possible for two kinds of variables:

- Categorical
- Numerical

The most common patterns to be identified when performing a univariate analysis are mean, mode and median. For example, if we were studying the air quality around a city that is experiencing heavy smog levels, we could identify the central tendency of a sample (n = 3) of people's perception of the air quality by looking at their answers to the question: "How clean or dirty did you feel while you were driving?
Similarly, Dispersion (range, variance), Quartiles (interquartile range), and Standard deviation can also be used for this analysis.

Before going into the details of the practice exercise, let us take a look at the data. This data has the following features:

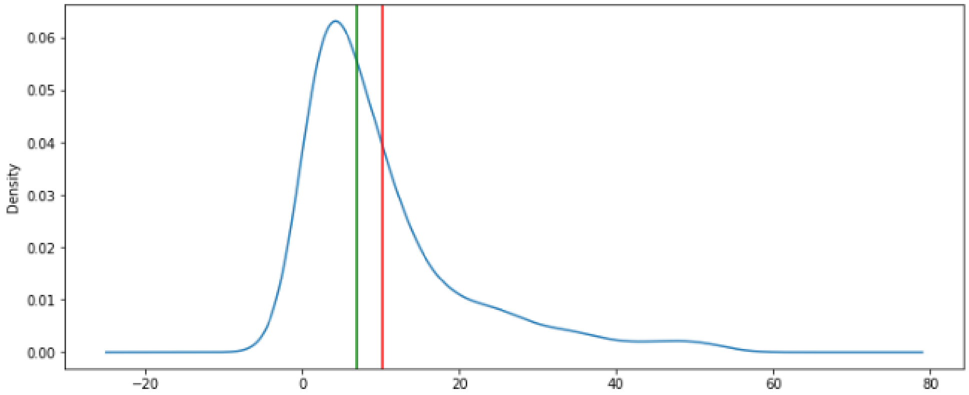


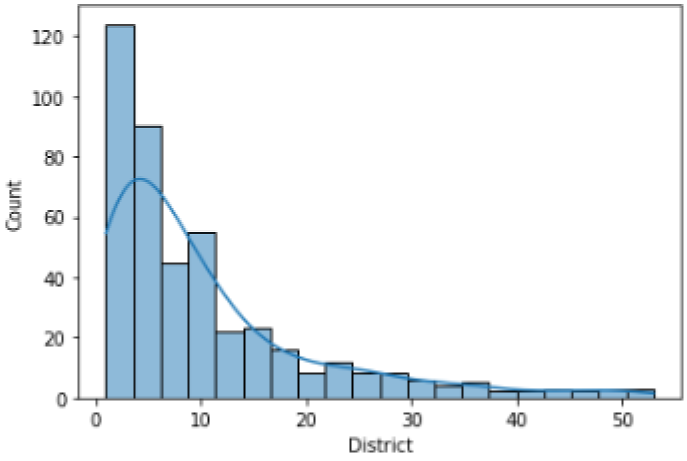
```
int64 columns:
 Index(['District'], dtype='object')

float64 columns:
 Index([], dtype='object')

object columns:
 Index(['Name', 'State', 'Party', 'Room', 'Phone', 'House Positions',
       'Committee Assignment'],
      dtype='object')
```

Histograms

Histograms are similar to bar charts, displaying similar categorical variables against the category of data. Histograms display these categories as bins which indicate the number of data points in a range. They are used to represent a range of values and the frequency of those values within an interval.

In the upcoming exercises, we will be working on the histogram plots of continuous variables. Some outputs have been shown here:



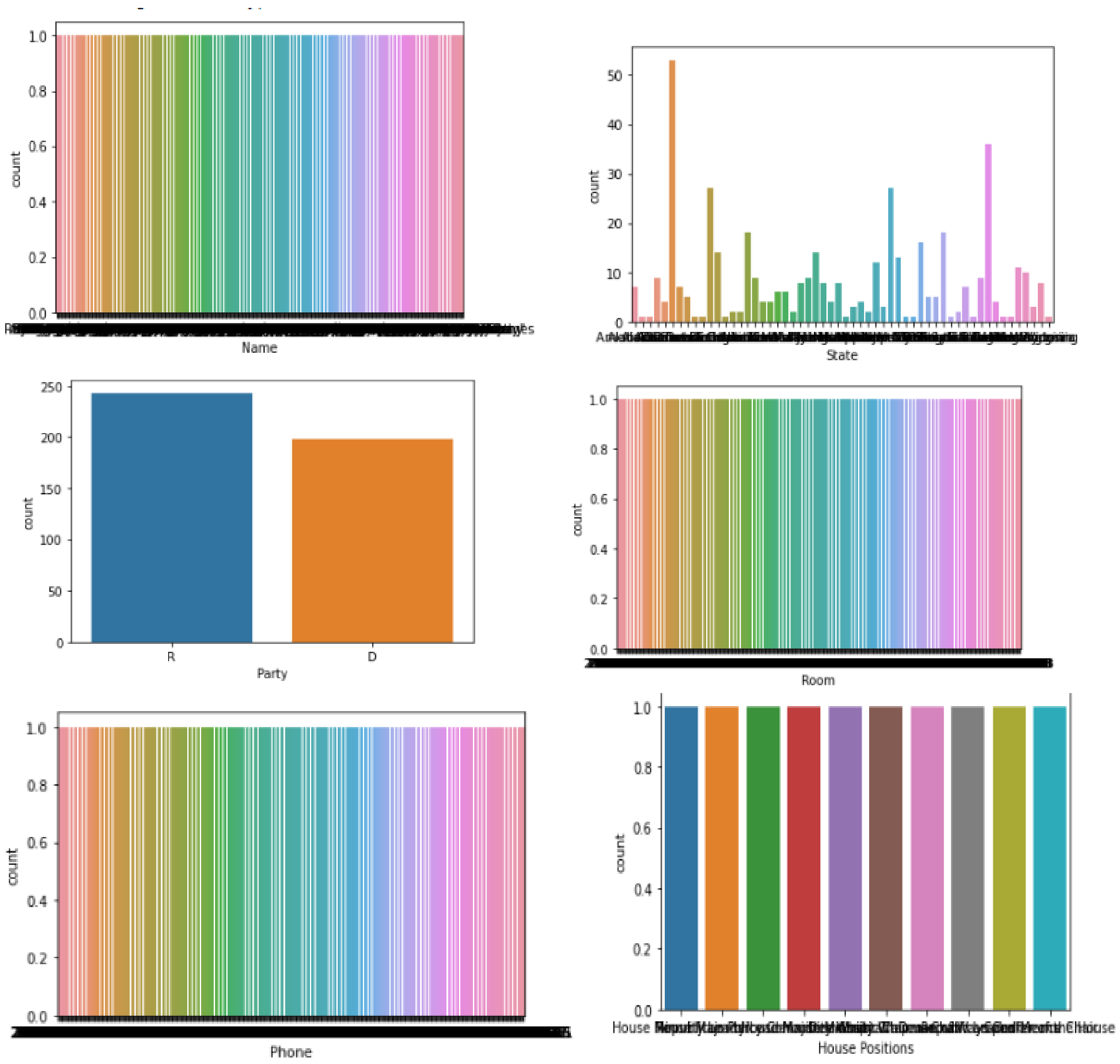


Bar charts/Count Plots

The bar graph is very convenient when comparing categories of data or different groups of data. It helps to track changes over time. It is best for visualizing discrete data. For the bar charts following syntax will be used:

```
sns.countplot(x="Phone", data=df)
df['Phone'].value_counts()
```

It has provided the following results for all variables:



You can learn more about Univariate analysis from here.

Exercise 2.01: EDA – Univariate Analysis

In this exercise, a univariate analysis will be done by plotting histograms and bar charts.

Open the following link to get started (Empty Jupyter Environment) or start with your own local or Hosted Jupyter environment:

https://mybinder.org/v2/gh/fenago/jupyter/HEAD

RESOURCES AND REFERENCES

The code is located here:

datawrangling/Exercise_2_01_UnivariateAnalysis.ipynb at main · fenago/datawrangling (github.com)

The code repository is located here:
datawrangling/Chapter 2 at main · fenago/datawrangling (github.com)

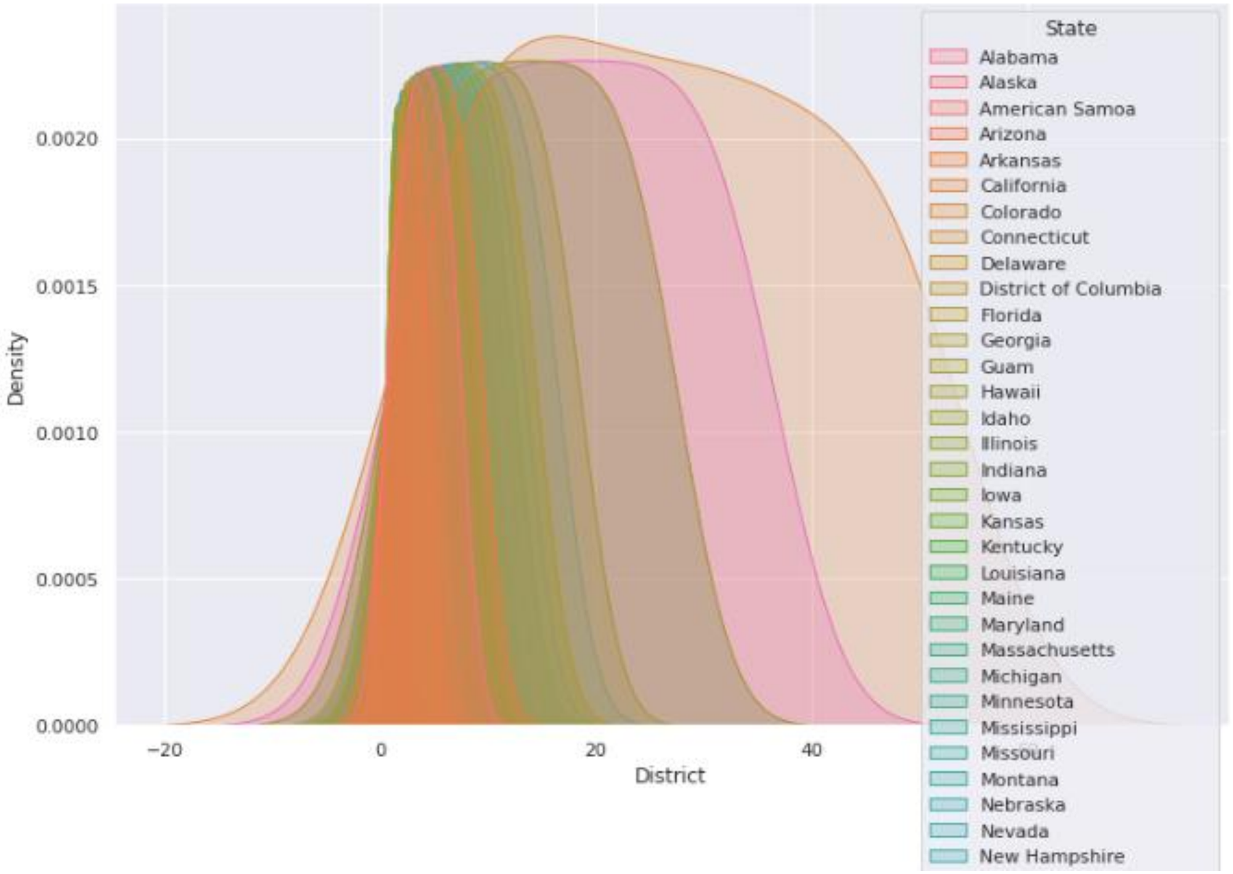
Bivariate Analysis

Bi means two and variate means variable, so here there are two variables. The analysis is related to the cause and the relationship between the two variables.

Types:

- Bivariate Analysis of two numerical variables
- Bivariate Analysis of two categorical variables
- Bivariate Analysis of a numerical variable and a categorical variable

```
# x = <NUMERIC VARIABLE>, hue = <CATEGORICAL VARIABLE>
plt.figure(figsize=(12,8))
sns.kdeplot(data=df,x='District',hue='State',fill=True)
```



Bivariate Correlations

Correlations measure how variables or rank orders are related. Before calculating a correlation coefficient, screen your data for outliers (which can cause misleading results) and evidence of a linear relationship. **Pearson'**s correlation coefficient is a measure of linear association. Two variables can be perfectly related, but if the relationship is not linear, Pearson's correlation coefficient is not an appropriate statistic for measuring their association. In the upcoming exercise, we have used "spearman's method".

You can learn more about bivariate analysis from Here.

We will practice these concepts in the upcoming exercise.

Exercise 2.02: EDA – Bivariate Analysis

In this exercise, the bivariate analysis will be done and the analysis will be done through the plots:

Open the following link to get started (Empty Jupyter Environment) or start with your own local or Hosted Jupyter environment:

https://mybinder.org/v2/gh/fenago/jupyter/HEAD

RESOURCES AND REFERENCES

The code is located here:

datawrangling/Exercise_2_02_BivariateAnalysis.ipynb at main · fenago/datawrangling (github.com)

The code repository is located here:
datawrangling/Chapter 2 at main · fenago/datawrangling (github.com)

Multivariate Analysis

In multivariate analysis, more than 2 variables are analyzed. It is a tremendously hard task for the human brain to visualize a relationship among 4 variables in a graph. Multivariate analysis is used to study more complex sets of data.

Types

- Cluster analysis
- Factor Analysis
- Multiple Regression Analysis
- Principal Component Analysis

Many different ways exist to perform multivariate analysis. But it depends upon the type of data being used.

Cluster Analysis

Cluster analysis is useful for finding groups of related objects in a data set. It tries to classify the objects in such a way that the similarity between two objects from the same group is maximum and minimum otherwise. A well-known example of clustering is in marketing, where customers are grouped into segments based on their buying patterns, demographics, and preferences.

Correspondence Analysis

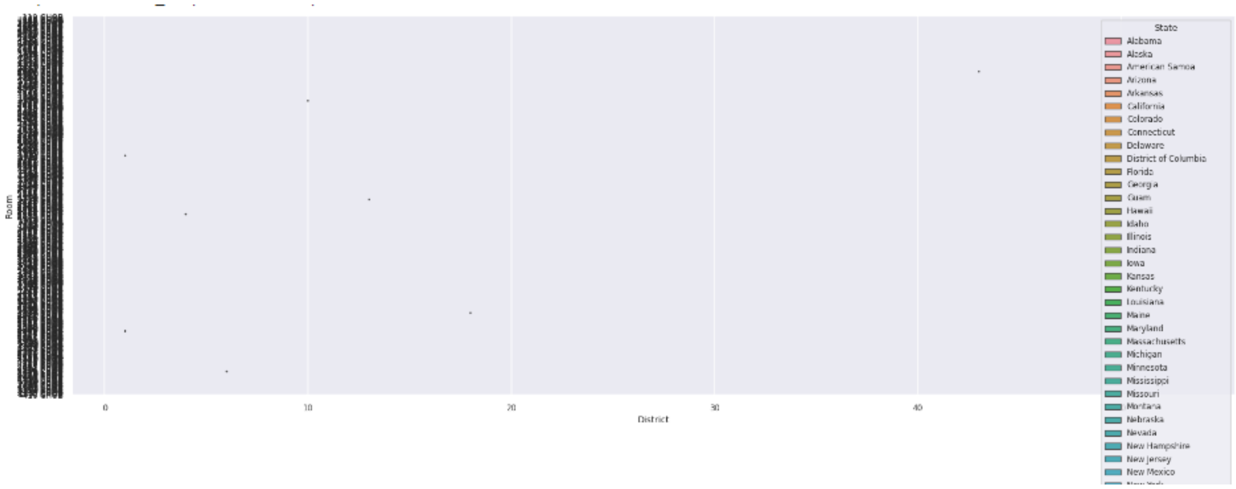
Correspondence Analysis using the data from a contingency table shows relative relationships between and among two different groups of variables. A contingency table is a 2D table with rows and columns as groups of variables.

Principal Component Analysis

Principal Components Analysis (**PCA**) is a linear method for transforming a large dimensional data set into a smaller, standardized dimensionality that is more easily studied and understood. Here we are converting our original variables into a new set of variables called principal components that represent the most important dimensions in a dataset.

Some outputs are the following:

```
sns.boxplot(data=df,x='District',y='Room',hue='State')
```



Single Index:

Categorical value is compared with all continuous values with pivot tables. The following syntax will be followed:

```
# Single Index  –
table = pd.pivot_table(data=df,index=['State'])
table
```

Multiple Index

Multiple values are compared with all continuous values with pivot tables. The following syntax will be followed:

```
# Multiple  values concerning all continuous values in the datas
et
table = pd.pivot_table(df,index=['Room','District'])
table
```

Aggregates

Aggregates on specific features will be explored in the upcoming exercise. It involves the following code snippet:

```
# Aggregate on specific features with values parameter
table = pd.pivot_table(df,index=['Room','State'],dropna=False)
table
```

```
table = df.pivot_table(index=['Room','State'],
               columns='Assignment',
               aggfunc='size',
               fill_value=0,)
table
```

Exercise 2.03: Multivariate analysis

In this exercise, you will practice the concepts related to the relationships between more than two variables.

Open the following link to get started (Empty Jupyter Environment) or start with your own local or Hosted Jupyter environment:

https://mybinder.org/v2/gh/fenago/jupyter/HEAD

RESOURCES AND REFERENCES

The code is located here:

datawrangling/Exercise_2_03_MultivariateAnalysis.ipynb at main · fenago/datawrangling (github.com)

The code repository is located here:

datawrangling/Chapter 2 at main · fenago/datawrangling (github.com)

Automated EDA Tooling

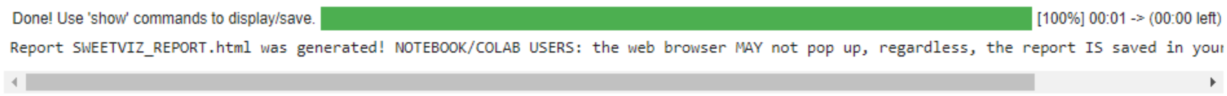
EDA describes the processes of detecting outliers, and missing values, converting categorical variables, and determining the skewness of the dataset. It can be tedious at times to comprehend your data and use it to build models, but with some helpful tools, you can quickly gain an understanding of your database and build better models.

In the upcoming exercise, we will be using “**sweetviz**” for the same purpose.

```
#Installing the library
!pip install dataprep

#Importing
from dataprep.eda import create_report
#Creating report
create_report(df)

!pip install skimpy
from skimpy import skim
skim(df)
```

```
#Installing the library
!pip install sweetviz
#Importing the library
import sweetviz as sv
report = sv.analyze(df)
report.show_html()

# Spliting data set into training and testing set
training_data = df.sample(frac=0.8, random_state=25)
testing_data = df.drop(training_data.index)

#Applying compare function
report2 = sv.compare([training_data,"TRAINING SET"], [testing_data, "TESTI
NG SET"])
report2.show_html()
```

Finally, you will get the following output:



```
Done! Use 'show' commands to display/save. [100%] 00:01 -> (00:00 left)
Report SWEETVIZ_REPORT.html was generated! NOTEBOOK/COLAB USERS: the web browser MAY not pop up, regardless, the report IS saved in your
```

Exercise 2.04: Automated EDA Tooling

In this exercise, you will practice the concepts related to some tools for EDA.

Open the following link to get started (Empty Jupyter Environment) or start with your own local or Hosted Jupyter environment:

https://mybinder.org/v2/gh/fenago/jupyter/HEAD

RESOURCES AND REFERENCES

The code is located here:

datawrangling/Exercise_2_04_EDA_TOOLING.ipynb at main · fenago/datawrangling (github.com)

The code repository is located here:

datawrangling/Chapter 2 at main · fenago/datawrangling (github.com)

Activity 2.01: Loading the "City Land Mass - Miami" Dataset and performing Univariate and Bivariate Analysis on it.

In this activity, we will be checking the working on another dataset. This dataset contains the records of Landmass from Miami. We will be performing Univariate and bivariate analyses on the data to test the skill. This data does not contain many categorical variables.

The City Land Mass Dataset has a layer that shows all land mass within the City of Miami's Official boundary by removing the water bodies. This layer can be used to aesthetically represent the City of Miami in various cartographic products. It shall not be construed as a 'complete' boundary file as it does not show any water bodies about or within the City's boundary.

Open the following link to get started (Empty Jupyter Environment) or start with your own local or Hosted Jupyter environment:

https://mybinder.org/v2/gh/fenago/jupyter/HEAD

RESOURCES AND REFERENCES

The code is located here:

datawrangling/Activity_2.01_Land_Mass_Miami.ipynb at main · fenago/datawrangling (github.com)

The code repository is located here:
datawrangling/Chapter 2 at main · fenago/datawrangling (github.com)

Conclusion

In this chapter, we learned how to perform rigorous analysis on the dataset. Practicing EDA, we have seen Univariate, bivariate, and multivariate analyses of the data. Moreover, we have also learned about automated EDA tooling. Extensive methods were covered in this chapter to analyze the data.

It was observed that different types of relationships exist between the different variables. Moreover, different operations need to be performed between them depending upon their types. Similarly, the graphs were also dependent upon the type of data. It can be a numerical or a categorical variable, but it must be analyzed before starting the process. Data cleaning and Data preprocessing should also be done before starting the procedure.

In the next chapter, we will cover the “**Feature Engineering**” techniques, including feature selection, Feature Importance, Feature Reengineering, and binning.