Graphical user interface, text, application

Description automatically generatedGraphical user interface, application

Description automatically generated

Student Name

***Sanatkumar Rajmogali Ippalpalli***

Title of Project Report

***Guided Project 3 – Exploratory Factor Analysis***

Table of Contents

[Introduction 4](#_Toc82371169)

[Eckovation theme & Question 5](#_Toc82371170)

[Prerequisites before starting coding 6](#_Toc82371171)

[program DEVELOPMENT steps 8](#_Toc82371172)

[Dataset requirements 8](#_Toc82371173)

[Technique – EFA 8](#_Toc82371174)

[PROGRAM / CODE DEVELOPMENT 9](#_Toc82371175)

[Analysis 30](#_Toc82371176)

[CONCLUSION 31](#_Toc82371177)

[Figure 1 Import libraries and datasets/modules 9](#_Toc82371147)

[Figure 2 Load Datasets and converting Pandas Dataframe 9](#_Toc82371148)

[Figure 3 Training Dataset Information 10](#_Toc82371149)

[Figure 4 Dataset Information – Understanding with statistics 11](#_Toc82371150)

[Figure 5 Dataset Information – Removal of NA, Duplicate & Unique Information 13](#_Toc82371151)

[Figure 4 Dataset Preparation -Code for drop & encoding parameters 19](#_Toc82371152)

[Figure 5 results after dataset preparation 20](#_Toc82371153)

[Figure 6 Dataset Balanced/Imbalanced Check 21](#_Toc82371154)

[Figure 6 Dataset check for categorical variables 21](#_Toc82371155)

[Figure 7 Univariate Analysis using Seaborn Catplot 22](#_Toc82371156)

[Figure 8 Univariate Analysis using Seaborn Catplot 22](#_Toc82371157)

[Figure 8 Univariate Analysis using Seaborn Catplot 23](#_Toc82371158)

[Figure 9 Univariate Analysis using Seaborn Catplot 24](#_Toc82371159)

[Figure 9 Univariate Analysis using Seaborn Catplot 24](#_Toc82371160)

[Figure 9 Univariate Analysis using Seaborn Catplot 25](#_Toc82371161)

[Figure 9 Univariate Analysis using Seaborn Catplot 25](#_Toc82371162)

[Figure 9 Univariate Analysis using Seaborn Catplot 26](#_Toc82371163)

[Figure 9 Univariate Analysis using Seaborn Catplot 26](#_Toc82371164)

[Figure 9 Univariate Analysis using Seaborn Catplot 26](#_Toc82371165)

[Figure 10 Univariate Analysis using Matplotlib & Seaborn heatmap 27](#_Toc82371166)

[Figure 11 Removing Outliers from dataset 28](#_Toc82371167)

[Figure 12 Correlation among Features 29](#_Toc82371168)

***EXECUTIVE SUMMARY***

It is very important for any business dataset, to ensure the results produced by any kind of machine algorithm / technique are valid and applicable to any desired business optimization outcomes and goals.

Science and technology improved many technologies and has guided numerous innovative features which advanced the digitization with respect to data (text format) in the charts that we use for quick understanding / judgement.

Exploratory Factor Analysis (EFA) can help answer questions about standard deviations, categorical variables, and confidence intervals. Once EFA is complete and insights are drawn, its features can then be used for more sophisticated data analysis or modeling, including [machine learning](https://www.ibm.com/cloud/learn/machine-learning).

As part of guided project, to find by using dataset which explore / extract top features interact were coded and implemented to capture the details as mentioned in the question.

# Introduction

The objective is to showcase a few of the basics of data science (data cleaning, encoding, feature engineering, and model training), all while attempting to solve a problem that is common among businesses: ***Customer Satisfaction***.

*Customer satisfaction* (often abbreviated as CSAT) is a measure of how products and services supplied by a company meet or surpass [customer](https://en.wikipedia.org/wiki/Customer) expectation. Customer satisfaction is defined as "the number of customers, or percentage of total customers, whose reported experience with a firm, its products, or its services (ratings) exceeds specified [satisfaction](https://en.wikipedia.org/wiki/Contentment) goals”. Customers play an important role and are essential in keeping a product or service relevant; it is, therefore, in the best interest of the business to ensure customer satisfaction and build customer trustworthiness. Measuring customer satisfaction provides an indication of how successful the organization is at providing products and/or services to the marketplace.

Science and technology improved many technologies and has guided numerous innovative features which advanced measurement of multiple attributes (features) indicate satisfaction on the basis of rankings (ratings) provided.

Hence Eckovation includes this guided project on Exploratory Factor Analysis in the courseware for students to understand, implementation / execute the code themselves on the airline passengers satisfaction dataset available on the Kaggle website. This guided project objective is to showcase a few of the basics of data science (data cleaning/transform, Normalization, encoding, feature engineering, and model training), all while attempting to solve a problem that is common among businesses: Customer Satisfaction.

This report includes the 5W1H about the theme of development of code and running the code with database available over the internet of airline passenger satisfaction estimation. At the end of the report, the conclusions share the features extracted and useful for next course of activities to gain advantages in the development.

# Eckovation theme & Question

**Theme : Exploratory Factor Analysis**

**Exploratory factor analysis or EFA** is a statistical technique used to reduce data to a smaller set of summary variables and to explore the underlying structure of a relatively large set of variables. It is used to identify the underlying relationships between measured variables. Each observed variable is considered as a potential measure of every factor, and the goal is to determine the strongest relationships.

**Question:**

Factor analysis is a useful technique to find latent factors that can potentially describe multiple attributes, which is sometimes very useful for dimensionality reduction. Use the **Airline Passenger Satisfaction dataset** to perform factor analysis. (Use only the columns that represent the ratings given by the passengers, only 14 columns). Choose the best features possible that helps in dimensionality reduction, without much loss in information.

**Airline Passenger Satisfaction dataset**

Dataset Link: <https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction>

# Prerequisites before starting coding

1. Who - Software needed?
2. What - Version / Release of software?
3. Any Prerequisites
4. How - to install the software
5. Which -libraries are needed to execute the problem statement
6. Where – dataset requirements, path location to include in the code
7. When – to use the above feature extraction
8. Who – Software neeed?

Python

1. What- Version / Release of software?

Python version 3.6 (latest version of python)

1. Any Prerequisites

RAM space availability & hard disk space availability

Admin rights to install the software

1. How - to install the software
2. The following url [https://www.python.org/downloads/](https://www.python.org/downloads/%20) can be referred to download python.
3. Second and easier option is to download anaconda and use its anaconda prompt to run the commands. To install anaconda check this url <https://www.anaconda.com/download/>
4. Which -libraries are needed to execute the problem statement
5. Sklearn (scikit-learn) ( pip install -U scikit-learn)
6. Numpy (pip install numpy)
7. Matplotlib (pip install matplotlib)
8. Scipy (pip install scipy)
9. Seaborn (pip install seaborn)
10. Where – dataset requirements, path location to include in the code
11. Once you have python downloaded and installed, you will need to setup PATH variables (if you want to run python program directly, detail instructions are below in how to run software section). To do that check this: [https://www.pythoncentral.io/add-python-to-path-python-is-not- recognized-as-an-internal-or-external-](https://www.pythoncentral.io/add-python-to-path-python-is-not-recognized-as-an-internal-or-external-command/) [command/](https://www.pythoncentral.io/add-python-to-path-python-is-not-recognized-as-an-internal-or-external-command/).
12. Setting up PATH variable is optional as you can also run program without it and more instruction are given below on this topic.
13. When – to use the above feature extraction
14. When – to use the above feature exploratory factor analysis

# program DEVELOPMENT steps

* Dataset requirement
* Technique selections
* Program / code development
* Analysis

### Dataset requirements

The data source used for this project has been downloaded from the Kaggle website.

**Airline Passenger Satisfaction dataset**

Dataset Link: <https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction>

### Technique – EFA

Exploratory factor analysis (EFA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

EFA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate.

Let us hop to the inscribing carving!

### PROGRAM / CODE DEVELOPMENT

As explained step by step during the lecture by mentor, we would approach steps and understand the basics with brief explanation as needed.

#### Step 1: Import the relevant libraries and applicable datasets/modules

Graphical user interface, text, application

Description automatically generated

Figure Import libraries and datasets/modules

#### Step 2: Load dataset and convert Pandas Dataframe

Download the csv dataset and copy into the disk as Pandas Dataframe. Estimate shape of

training and testing dataset to understand available features in the dataset.

Text

Description automatically generated

Figure Load Datasets and converting Pandas Dataframe

Understand, Clean and Transform Dataset

The main purpose of EFA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

The Training dataset has 25 features and we need to check if all features are numerical (integer) format before proceeding to machine learning model. This is called cleaning / prepare the dataset for transformation.

Table

Description automatically generated

Figure Training Dataset Information

Based on the figure above, there are multiple variables which needs to be analysed before setting for dataset cleaning / preparation. Each parameter uniqueness to be studied beforehand using describe function in pandas dataframe.

Graphical user interface, text, application

Description automatically generated

Figure Dataset Information – Understanding with statistics

For the Age column, we can see that the youngest passenger is 7 years old and the oldest passenger is 85 years old. Average age is 39. When we evaluate the quarters, we can observe that the age is evenly distributed.

For the Flight Distance column, we see that the minimum value is 31 and the maximum value is 4983. The average distance of flight is 1180.4. When we look at the quarters, we can say that there are outliers because there is too much difference between the 3rd quarter and the maximum value.

For the Departure Delay in Minutes and Arrival Delay in Minutes columns, the minimum value is 0 (which corresponds to no delay in that flights) and the maximum value is around 1592/1584. When we examine the 3rd quarter and maximum values, we can see that there are too many outlier values.

There are many categorical features evaluated in 0-1-2-3-4-5 degrees. If we look at the averages of these ratings, the highest level of satisfaction is the Inflight service category with an average of 3.64, while the lowest is the Inflight wifi service category with an average of 2.72 points.

Dataset to be checked if there are any missing values in our dataset or not. It can occur when no information is provided for one or more items or for a whole unit.  issing Data is a very big problem in real-life scenarios. Missing Data can also refer to as NA(Not Available) values in pandas. There are several useful functions for detecting, removing, and replacing null values in Pandas DataFrame :

* [isnull()](https://www.geeksforgeeks.org/python-pandas-isnull-and-notnull/)
* [notnull()](https://www.geeksforgeeks.org/python-pandas-isnull-and-notnull/)
* [dropna()](https://www.geeksforgeeks.org/python-pandas-dataframe-dropna/)
* [fillna()](https://www.geeksforgeeks.org/python-pandas-dataframe-fillna-to-replace-null-values-in-dataframe/)
* [replace()](https://www.geeksforgeeks.org/python-pandas-dataframe-replace/)
* [interpolate()](https://www.geeksforgeeks.org/python-pandas-dataframe-interpolate/)
* isna()
* nunique()
* duplicated()

Table

Description automatically generated

Graphical user interface, text

Description automatically generated

Chart, pie chart

Description automatically generated

Figure Dataset Information – Removal of NA, Duplicate & Unique Information

There are four primary types of EFA:

* Univariate non-graphical. This is simplest form of data analysis, where the data being analyzed consists of just one variable. Since it’s a single variable, it doesn’t deal with causes or relationships. The main purpose of univariate analysis is to describe the data and find patterns that exist within it.
* Univariate graphical. Non-graphical methods don’t provide a full picture of the data. Graphical methods are therefore required. Common types of univariate graphics include:
  + Stem-and-leaf plots, which show all data values and the shape of the distribution.
  + Histograms, a bar plot in which each bar represents the frequency (count) or proportion (count/total count) of cases for a range of values.
  + Box plots, which graphically depict the five-number summary of minimum, first quartile, median, third quartile, and maximum.
* Multivariate nongraphical: Multivariate data arises from more than one variable. Multivariate non-graphical EDA techniques generally show the relationship between two or more variables of the data through cross-tabulation or statistics.
* Multivariate graphical: Multivariate data uses graphics to display relationships between two or more sets of data. The most used graphic is a grouped bar plot or bar chart with each group representing one level of one of the variables and each bar within a group representing the levels of the other variable.

Text

Description automatically generated

Chart

Description automatically generated

Chart, bar chart

Description automatically generated

Graphical user interface, chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart, waterfall chart

Description automatically generated

Thus, EFA is primarily used to see what data can reveal beyond the formal modelling or hypothesis testing task and provides a provides a better understanding of data set variables and the relationships between them. The above plots represent Univariate visualization of each field in the raw dataset, with summary statistics. Also, we shall apply Multivariate visualizations, for mapping and understanding interactions between different fields in the data as we proceed.

#### Step 3: Dataset Preparation

There are quite a few parameters which need to be cleaned in order to prepare the data for use in a machine learning model.

There are several categorical variables that need to be converted (encoded), including our target variable 'Satisfaction'.

There are also a couple of columns that are unnecessary, such as 'Unnamed:0' and 'id' which shall be dropped. The functions below will be used to perform the dataset preparation.

Loop over dataset variables like Gender, customer type, travel type, class, satisfaction to be encoded as shown below code.

Text, timeline

Description automatically generated

Figure Dataset Preparation -Code for drop & encoding parameters

Graphical user interface

Description automatically generated with medium confidence

Table

Description automatically generated

Figure results after dataset preparation

#### Step 4: Balanced /Imbalanced Dataset

Apply the train\_test\_split module from sklearn.model\_selection and split the data (X-features and y-labels)into training data and testing data, with 10% of the data used for testing and the remaining 90% to train the model.

Chart, bar chart, box and whisker chart

Description automatically generated

Figure Dataset Balanced/Imbalanced Check

Text

Description automatically generated with low confidence

Figure 6 Dataset check for categorical variables

#### Step 5: Exploratory Factor Analysis

Data Visualization is the process of analyzing data in the form of graphs or maps, making it a lot easier to understand the trends or patterns in the data. There are various types of visualizations –

* Univariate analysis: This type of data consists of only one variable. The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it.
* Bi-Variate analysis: This type of data involves two different variables. The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the two variables.
* Multi-Variate analysis: When the data involves three or more variables, it is categorized under multivariate.

We will use ***Matplotlib and Seaborn*** library for the data visualization.

#### Analysis Using Seaborn

**Customer Type**: Loyal passengers are very high in number. Even among loyal passengers, the ratio of satisfied and dissatidfied ones are almost *49:51*.

**Gender**: It is observed that gender-wise distribution of dissatisfied and satisfied customers are quite same. For both male and female passengers, no. of dissatisfied customers are on the higher side compared to no. of satisfied customers.

Chart, bar chart

Description automatically generated

Figure Univariate Analysis using Seaborn Catplot

Chart, bar chart

Description automatically generated

Figure Univariate Analysis using Seaborn Catplot

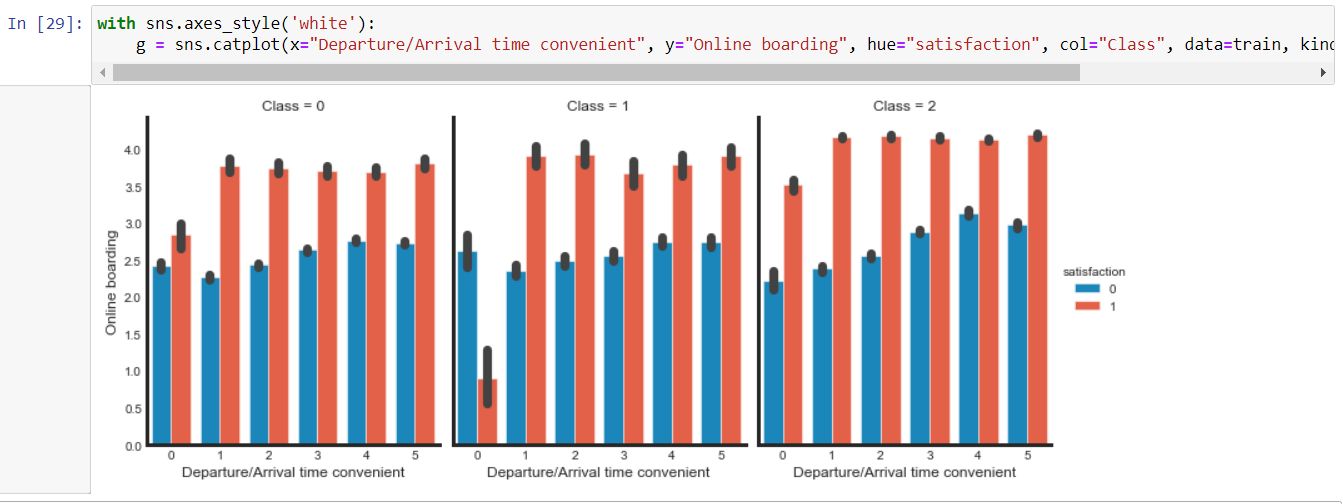


Figure 8 Univariate Analysis using Seaborn Catplot

with sns.axes\_style('white'):

g = sns.catplot(x="Departure/Arrival time convenient", y="Online boarding", hue="satisfaction", col="Class", data=train, kind="bar", height=4.5, aspect=.8)

**Online Boarding, Departure/Arrival Time Convenience grouped by Class:**

For Eco Plus class, very inconvenient Departure/Arrival time (Departure/Arrival\_time\_convenient = 0) has really high no. of dissatisfied passengers, even when online boarding is done very well. For other combinations, no. of satisfied passengers are on the higher side compared to no. of dissatisfied passengers.

with sns.axes\_style('white'):

g = sns.catplot(x="Class", y="Departure Delay in Minutes", hue="satisfaction", col="Type of Travel", data=train, kind="bar", height=4.5, aspect=.8)

g = sns.catplot(x="Class", y="Arrival Delay in Minutes", hue="satisfaction", col="Type of Travel", data=train, kind="bar", height=4.5, aspect=.8)

Chart, bar chart

Description automatically generated

Figure Univariate Analysis using Seaborn Catplot

Departure Delay, Arrival Delay grouped by Type of Travel: For personal travel (specially Eco Plus and Eco), the no. of dissatisfied passengers are really high when arrival delay in minutes is high. Now, this is quite obvious. By minute comparison, all combinations have higher no. of dissatisfied passengers compared to no. of satisfied passengers.

with sns.axes\_style('white'):

g = sns.catplot(x="Gate location", y="Baggage handling", hue="satisfaction", col="Class", data=train, kind="box", height=4.5, aspect=.8)

Chart, bar chart

Description automatically generated

Figure 9 Univariate Analysis using Seaborn Catplot

**Baggage Handling, Gate Location grouped by Class:**

For business class, it is observed that all gate locations have higher no. of dissatisfied passengers when baggage handling is not done perfectly well (rating <= 4). For Eco Plus, when the gate location is 1 and for Eco, when the gate location is 2, even when the baggage are handled in a mediocre way (rating in range 2.0 - 4.0), passengers remained dissatisfied.

with sns.axes\_style('white'):

g = sns.catplot(x="Inflight wifi service", y="Inflight entertainment", hue="satisfaction", col="Class", data=train, kind="box", height=4.5, aspect=.8)

Chart, box and whisker chart

Description automatically generated

Figure 9 Univariate Analysis using Seaborn Catplot

**Inflight Entertainment, Inflight wi-fi Service grouped by Class:**

It is interesting to find that Eco Plus passengers are mostly satisfied without in-flight wi-fi service (rating 0) and medium level of in-flight entertainment (rating 2 - 4). For Business class passengers, only highest level of in-flight entertainment (rating 5) can make them satisfied. For Eco passengers, high level of in-flight entertainment (rating 3 - 5) and very high wi-fi service availability (rating 5) can make them satisfied.

with sns.axes\_style(style='ticks'):

g = sns.catplot("satisfaction", col="Ease of Online booking", col\_wrap=6, data=train, kind="count", height=2.5, aspect=.9)

Chart, bar chart

Description automatically generated

Figure 9 Univariate Analysis using Seaborn Catplot

with sns.axes\_style(style='ticks'):

g = sns.catplot("satisfaction", col="Seat comfort", col\_wrap=6, data=train, kind="count", height=2.5, aspect=.8)

Chart, bar chart

Description automatically generated

Figure 9 Univariate Analysis using Seaborn Catplot

with sns.axes\_style(style='ticks'):

g = sns.catplot("satisfaction", col="Cleanliness", col\_wrap=6, data=train, kind="count", height=2.5, aspect=.8)

Chart, bar chart

Description automatically generated

Figure 9 Univariate Analysis using Seaborn Catplot

with sns.axes\_style(style='ticks'):

g = sns.catplot("satisfaction", col="Food and drink", col\_wrap=6, data=train, kind="count", height=2.5, aspect=.8)

Chart, bar chart

Description automatically generated

Figure 9 Univariate Analysis using Seaborn Catplot

**Ease of Online Booking, Seat Comfort, Cleanliness, Food and Drink:**

For all of these features, maximum no. of satisfied passengers belongs to the category of 4 and 5 rating givers. Below rating 4, passengers are mostly dissatisfied.

#### Step 6: Analysis using Matplotlib

After completion of data analysis using Seaborn, now proceed forward using Matplotlib for data analysis.

Text

Description automatically generated

Chart, treemap chart

Description automatically generated

Figure Univariate Analysis using Matplotlib & Seaborn heatmap

**Checkin Service, Inflight Service, On-board Service, Leg-room Service:**

For checkin service, 0-2 rating givers are predominantly dissatisfied. For other three services, only 4 and 5 rating givers belong to satisfied passengers category.

#### Step 7: Label Encoding of Categorical Variables

Text

Description automatically generated with medium confidence

Graphical user interface, table

Description automatically generated

Figure Removing Outliers from dataset

Graphical user interface, text, application

Description automatically generated

Chart

Description automatically generated

Figure Correlation among Features

Graphical user interface, application

Description automatically generated

Text

Description automatically generated

Index(['Type of Travel', 'Class', 'Flight Distance', 'Inflight wifi service',

'Online boarding', 'Seat comfort', 'Inflight entertainment',

'On-board service', 'Leg room service', 'Cleanliness'],

dtype='object')

These are top 10 features impacting on passenger satisfaction. We will check feature importance with other methods as well.

### Analysis

Though we had got top 10 features impacting on passenger satisfaction in the first attempt but improvement in the code while changing methods like Random Forest, Logistic Regression, Naïve Bayes Classifier, K-nearest neighbour Classifier, Decision Tree Classifier, Neural Network (Multilayer Perceptron), Extreme Gradient Boosting, Adaptive Gradient Boosting as illustrated / commented in the program result in better prediction of top features.

As we conclude, Random Forest with less time betters all other models results.

This is first attempt, the improvements in the code with time with multiple attempts may be checked and justified in future.

# CONCLUSION

In this guided project, we built a Exploratory Factor Analysis for the passenger satisfaction airline dataset using seaborn and matplotlib technique before application of machine learning models. The seaborn customized plot types were used to corelate interactions of the variables (univariate / multivariate). This is followed by using analysis by matplotlib for univariate interactions to understand if any differences.

The multivariate analysis is done using Seaborn HeatMap using correlations estimated after label encoding, removal of outliers.

On further analysis with multiple methods of machine learning models, it’s deduced in the first attempt, Random Forest technique is faster.

As it’s first attempt, the improvements in the code with time with multiple attempts may be checked and justified for the conclusions.

This entire program runs within few minutes.

references:

1. <https://www.ibm.com/cloud/learn/exploratory-data-analysis>
2. <https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/>