

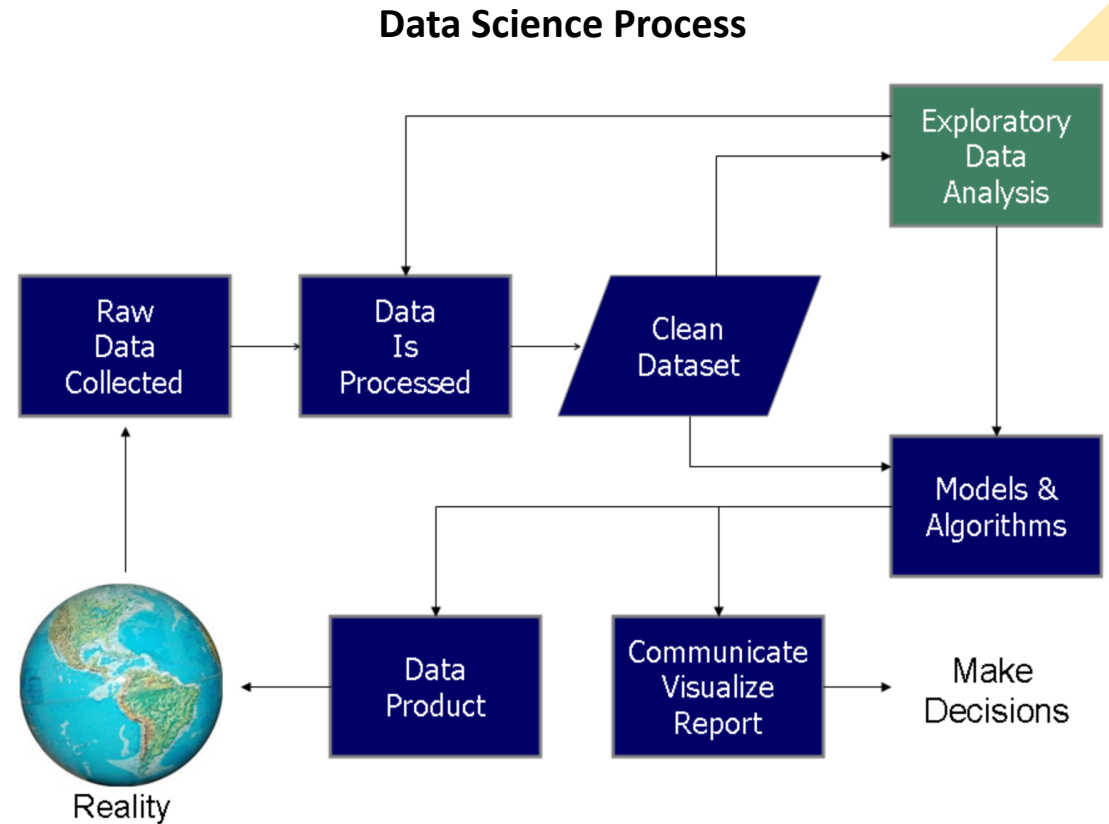
Exploratory Data Analysis (EDA) using Python



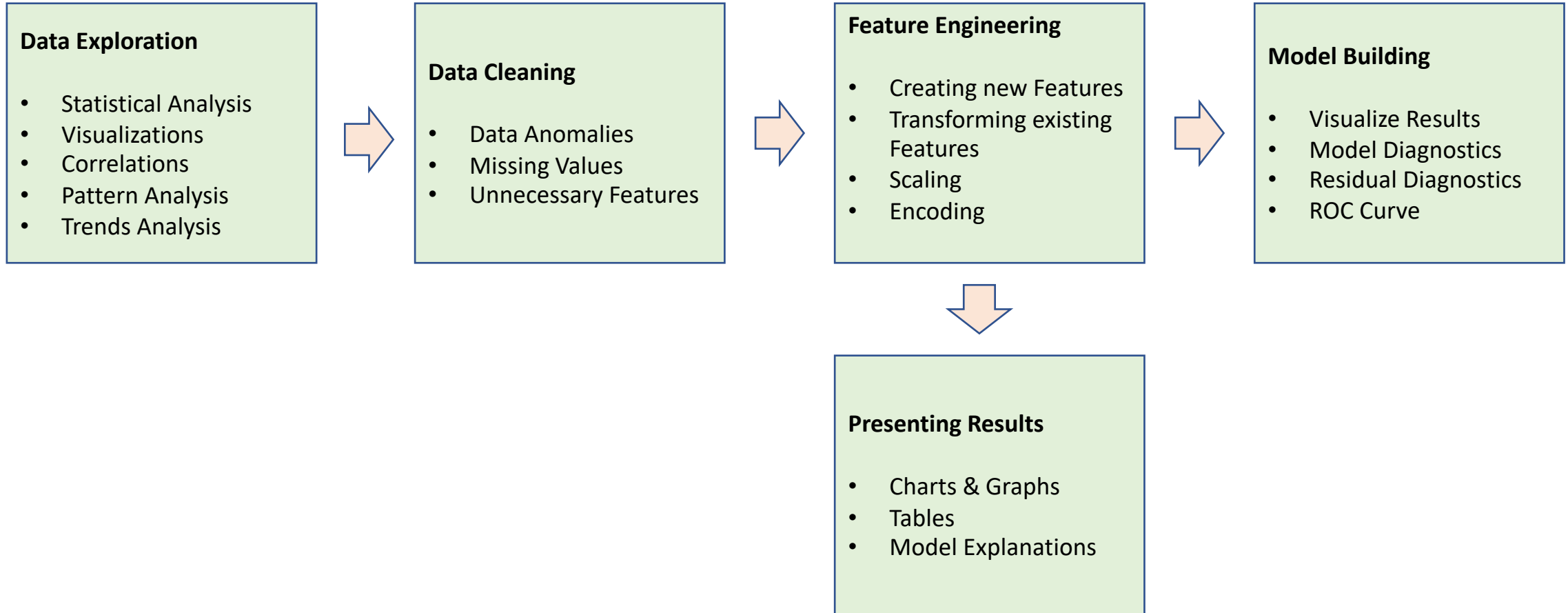
Introduction to EDA

Exploratory Data Analysis (EDA) is used by data scientists to **analyse 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.

EDA is primarily used to see **what data can reveal beyond the formal modelling and 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.



Use of EDA in various Phases of DS/ML



Goals of EDA

Analytics

Analysis and Finding Patterns, Trends, Relationships and Insights

Reports and Visualizations

Machine Learning Modelling

Pre-Processing Data for Machine Learning

Establishing Assumptions about Data. E.g., Normality.

Types of EDA

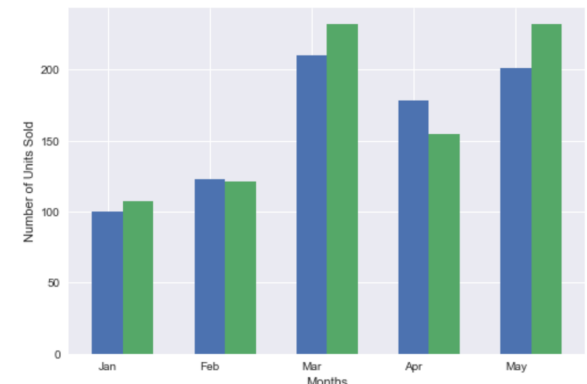
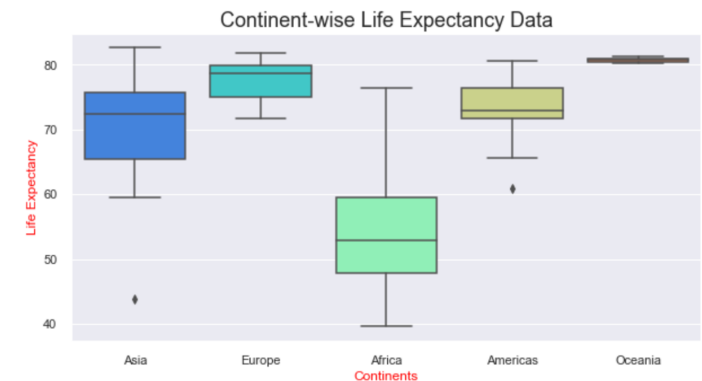
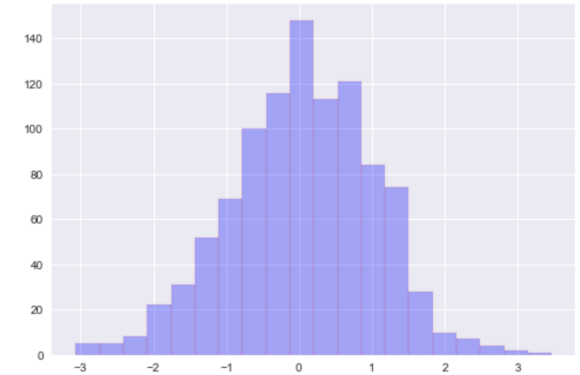
Univariate Non-Graphical: This is simplest form of data analysis, where the data being analysed consists of just one variable. Since it's a single variable, it doesn't deal with causes or relationships. The main purpose of univariate analysis is to describe the data and find patterns that exist within 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:

- 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.

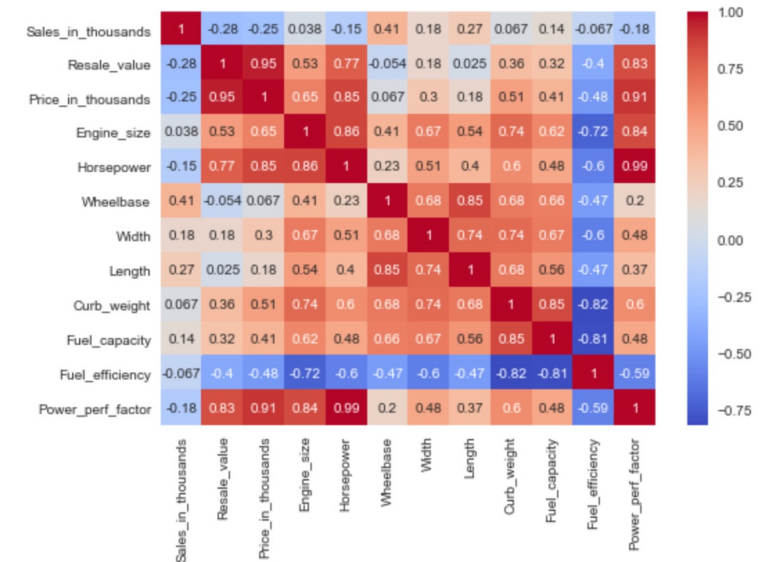
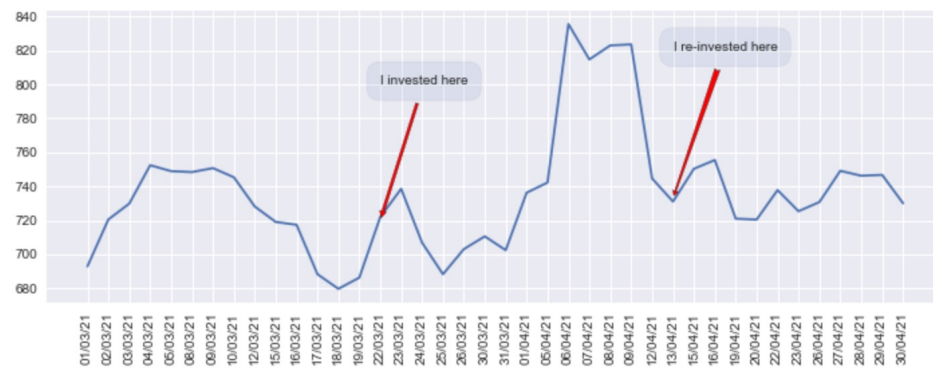
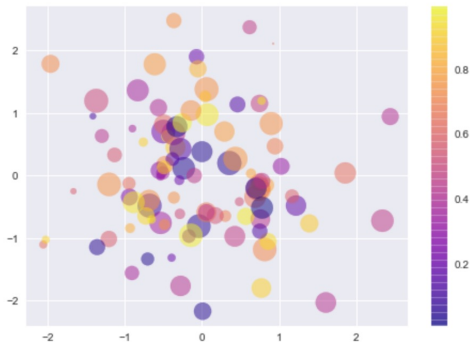
Multivariate Non-Graphical: Multivariate data arises from more than one variable. Multivariate non-graphical EDA techniques generally show the relationship between two or more variables of the data through cross-tabulation or statistics.

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.

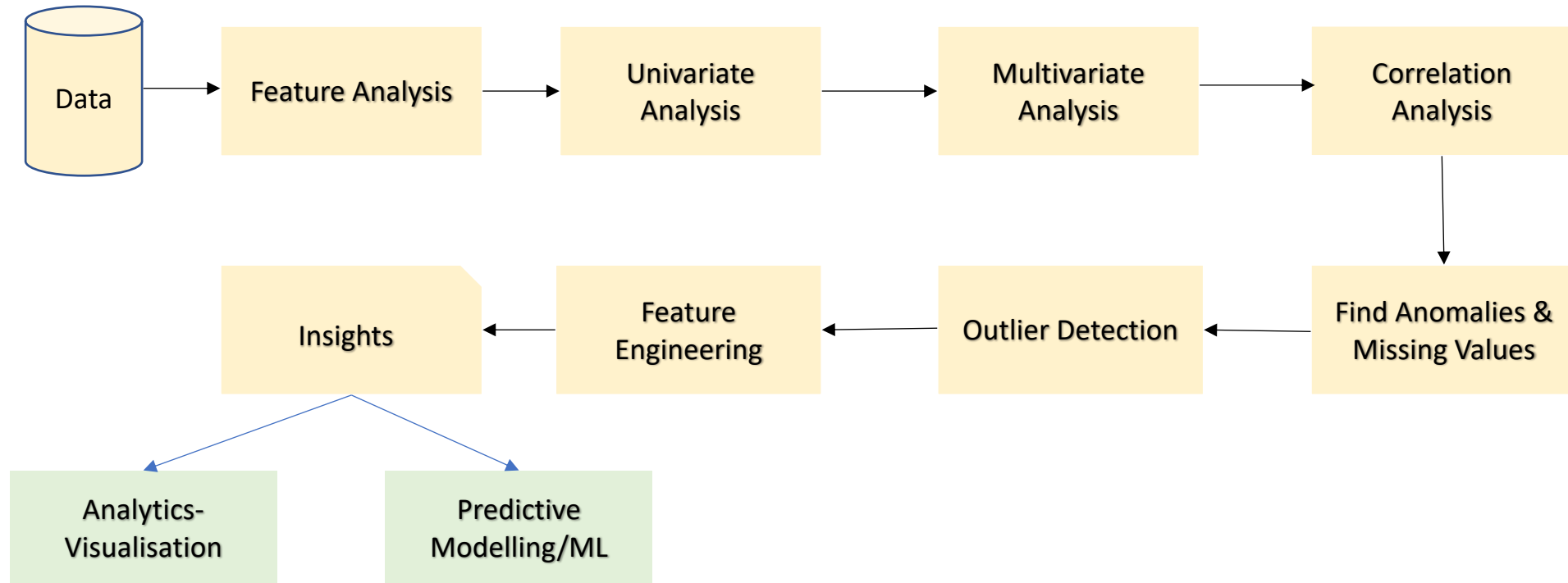


Tools of EDA


- **Scatter plot**, which is used to plot data points on a horizontal and a vertical axis to show how much one variable is affected by another.
- **Run chart**, which is a line graph of data plotted over time.
- **Bubble chart**, which is a data visualization that displays multiple circles (bubbles) in a two-dimensional plot.
- **Heat Map**, which is a graphical representation of data where values (degrees) are depicted by colour.



High Level EDA Process



Lending Club EDA Case Study



BORROW ▲

INVEST ▼

ABOUT US

HELP

Personal Loans >

Borrow up to \$40,000 and get a low, fixed rate.

Business Loans >

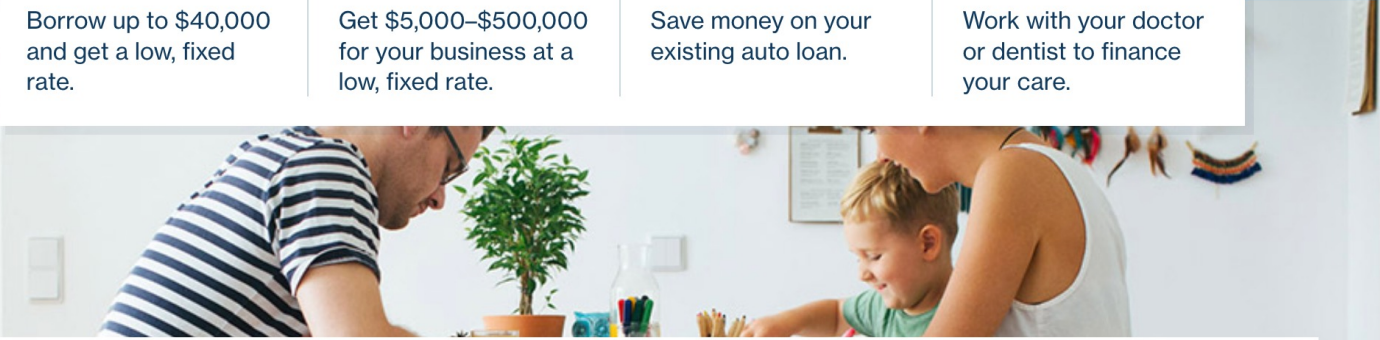
Get \$5,000–\$500,000 for your business at a low, fixed rate.

Auto Refinancing >

Save money on your existing auto loan.


Patient Solutions >

Work with your doctor or dentist to finance your care.



Personal Loans Up to \$40,000


Check your rate. It won't impact your credit score.

 Privacy & security
PROTECTION

\$ How much do you need?

What's the money for? ▼

Check Your Rate

 Respond to Mail Offer

Lending Club is an US based Consumer Finance Company. They have shared their Loan Transaction data in public for the purpose of Data Science research.

The most popular usage of Lending Club Data has become the EDA Case Study.

Lending Club Project Problem Statement

When Lending Club receives a loan application from an individual, the company has to make a decision for loan approval based on the Applicant's profile and Credit History. This is called **Credit Decision**. A correct Credit Decision is important to the company as a rejected Credit is loss of business, at the same time, credit given to an un-credit-worthy person leads to risk of default.

To make the Credit Decision process efficient and correct, Lending Club wants to undertake an exercise to analyse past loan data with the incidences of '**default's**' and '**non-default's**'.

Through this analysis, **Lending Club wants to find what factors in the Applicants Profile and the Loan Product must they focus on to ensure Credit is given to the right applicants and reduce chance of Credit Decision errors.**

EDA Process

Statistical Analysis and Summarisation of the Data



```
graph TD; A[Statistical Analysis and Summarisation of the Data] --> B[Univariate Analysis]; B --> C[Bivariate Analysis]; C --> D[Conclusions];
```

The diagram illustrates the EDA process as a four-step flowchart. The steps are represented by colored rectangular boxes arranged in a descending staircase pattern from top-left to bottom-right. The first box is blue, the second is teal, the third is green, and the fourth is olive green. Downward-pointing arrows connect each box to the next one below it.

Univariate Analysis

Bivariate Analysis

Conclusions

As part of the Analysis, we will use a number EDA techniques including abundant amount of Visualisations.

The final goal will be to draw conclusions from the data that would help Lending Club in better profiling of customer for better Credit Decisions.

Visualizations

