Exploratory Data Analysis for audit risk dataset

Sector score

An S-Score is a numerical value that shows how consumers and investors feel about a company, stock, exchange-traded fund (ETF), sector, or index as expressed over social media. S-Scores are created with data gathered by social media monitoring engines to help investors make trades and to help companies with market analysis and decision-making.

Inherent Risk:

Inherent Risk represents the susceptibility of a financial statement assertion to material misstatement before considering the effect of internal controls. It is an assessment of the risk factors inherent in an organization's operations, industry, and environment. Inherent Risk can be calculated using the following formula:

Inherent Risk = Risk of Material Misstatement (RMM) without considering controls

Control Risk:

Control Risk refers to the risk that a material misstatement could occur in a financial statement assertion but would not be prevented or detected by the entity's internal controls. It assesses the effectiveness of an organization's internal controls in preventing or detecting errors or fraud. Control Risk can be calculated using the following formula:

Control Risk = Risk of Material Misstatement (RMM) due to ineffective controls

Detection Risk:

Detection Risk represents the risk that the auditor fails to detect a material misstatement in the financial statements. It is the complement of the auditor's assurance of detecting a misstatement if it exists. Detection Risk can be calculated using the following formula:

Detection Risk = Risk that the auditor fails to detect a misstatement

These three components are multiplied together to calculate the Audit Risk, as mentioned in the previous response:

Audit Risk = Inherent Risk \* Control Risk \* Detection Risk

By multiplying these three factors, the Audit Risk quantifies the overall risk that the auditor will issue an incorrect opinion on the financial statements.

It's important to note that these calculations are subjective and require professional judgment from auditors to assess the risks accurately. The formulas provided serve as a general framework for understanding the concepts and relationships between Inherent Risk, Control Risk, and Detection Risk.

Audit Risk is the risk that an auditor will issue an incorrect opinion on financial statements. It is calculated by multiplying three components: Inherent Risk, Control Risk, and Detection Risk.

Audit Risk = Inherent Risk \* Control Risk \* Detection Risk

In the table, we have the values for Inherent Risk, Control Risk, and Detection Risk for each entry. Let's take the first entry as an example:

Inherent Risk = 8.574

Control Risk = 0.4

Detection Risk = 0.5

Using the formula mentioned above, we can calculate the Audit Risk as follows:

Audit Risk = 8.574 \* 0.4 \* 0.5

= 1.7148

So, the Audit Risk for the first entry is 1.7148.

Similarly, we can calculate the Audit Risk for each entry in the table using the respective values for Inherent Risk, Control Risk, and Detection Risk. The calculated Audit Risk represents the risk associated with each specific case.

It's important to note that the values of Inherent Risk, Control Risk, Detection Risk, and Audit Risk may vary depending on the context and specific calculations used in a particular audit or risk assessment. The provided values are examples for illustrative purposes.

The Audit risk dataset is used to advice auditors of fraudulent and non fraudulent firms using the risk factor of the companies based on present and historical data.

Now in this dataset there are 27 variables and 20952 records i.e., these records are from 776 companies from 14 different industry sectors.

Out of the 27 variables 12 variables is the data that is already recorded while collecting the sample data by two auditors who prepared the planned expenditure of inspection and summary reports A and B.

EDA on the dataset of audit risk

In the Exploratory Data Analysis of Audit Risk Dataset, we begin with identifying all the data types of the variables that are in the dataset. After finding all the datatypes of the variables, we must understand all the data in the dataset i.e., from when and where the data is collected, which methods are used to collect the data etc. Up next is representing the data in a graphical format so that anybody can understand the data by just referring to the graphs.

In the exploratory data analysis of this particular dataset, first we have identified all the datatypes there are, and then we have written a code for identifying the five highest values and five of the lowest values in the dataset.

Now based on the above steps, we have arrived at a conclusion that Location\_ID and TOTAL are slightly insignificant while calculating risk when compared to other variables. After removing insignificant variables we have described the data. Describing the data basically means we see what are the number of records present, mean, standard deviation, least value and highest values in each column of the dataset.

After describing the data, we are performing handling of missing data. Handling of missing data means we find if there are any missing values in any of the variables/attributes. Then we treat these missing values by filling up the missing value cells with the mean of that respective column.

The missing values that are replaced are helpful while plotting graphs and charts to better understand the data. We have tried to represent a few aspects of the data in the format of graphs and charts after handling missing values.

The first variable that we have considered to represent in the graphical format is “Money\_Value”. This variable shows the amount of money that is recorded in the misstatements of the companies. The reason for considering this variable is that, a company having higher value is more likely to be fraudulent and whichever has comparatively lower money value might be a non-fraudulent company. The following image shows the line graph of the variable money value.

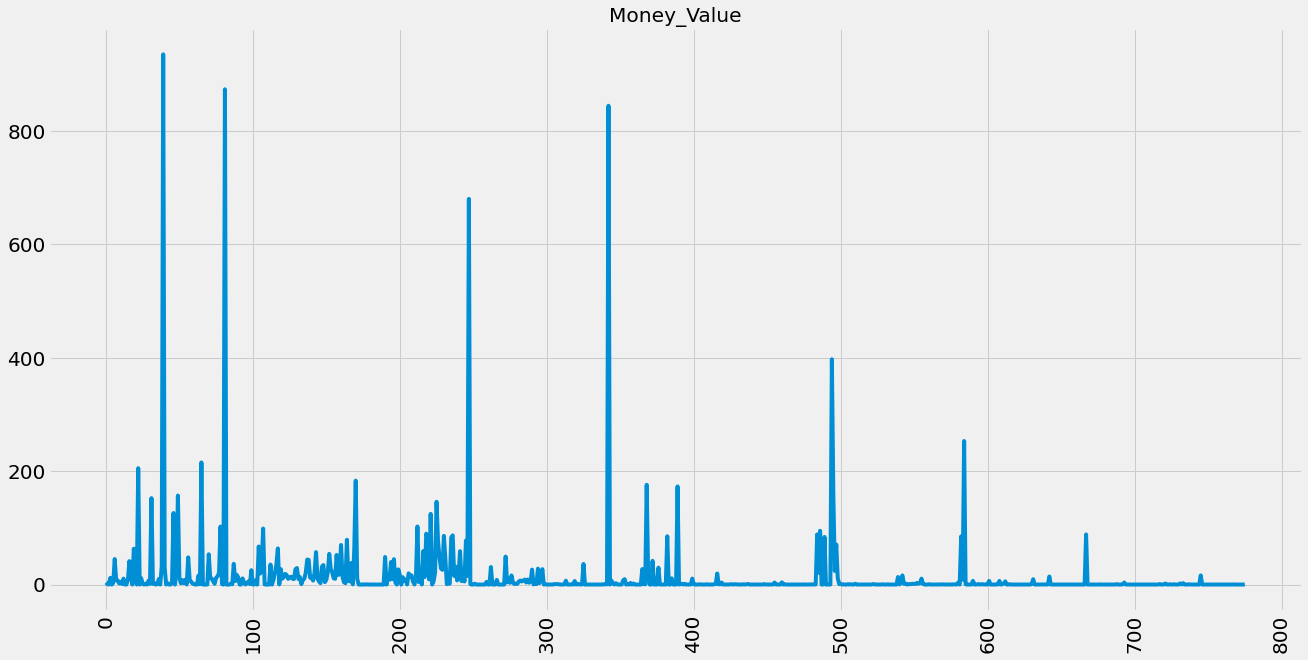


Figure 1: Line Graph of the variable “Money\_value”

Since having to look at one variable graph is not enough to understand the data as a whole, we have taken two variables against one another to see if we can better understand the data. The variables we choose to represent in the graph are “Score” and “Risk”. The following graph show a density graph on score against risk.

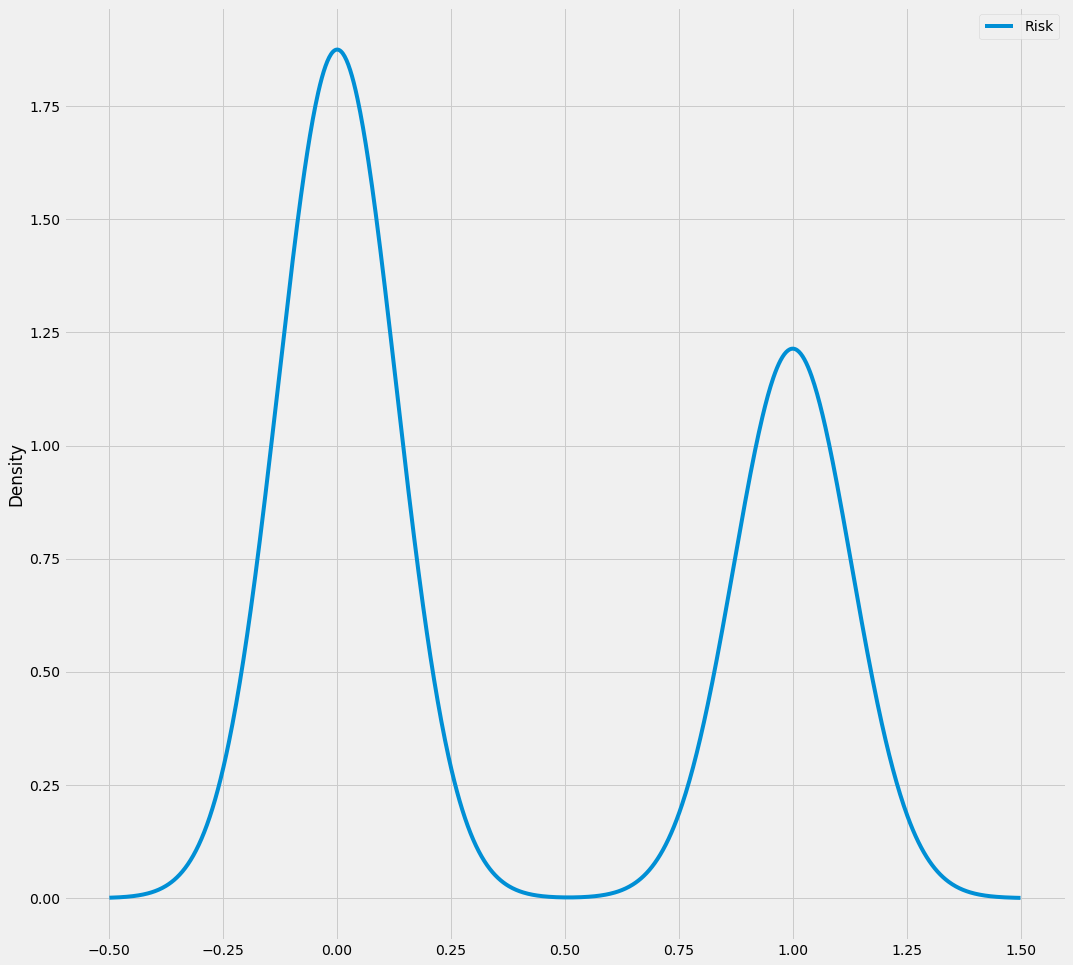


Figure 2: Density Graph on Score against Risk

Next we have Location\_ID and Money\_Value comparison using a barplot in python. The comparison of these two variables would let us understand the data in another perspective. That is we can see how many companies and in what locations have more money value and which companies and at what location have less money value.

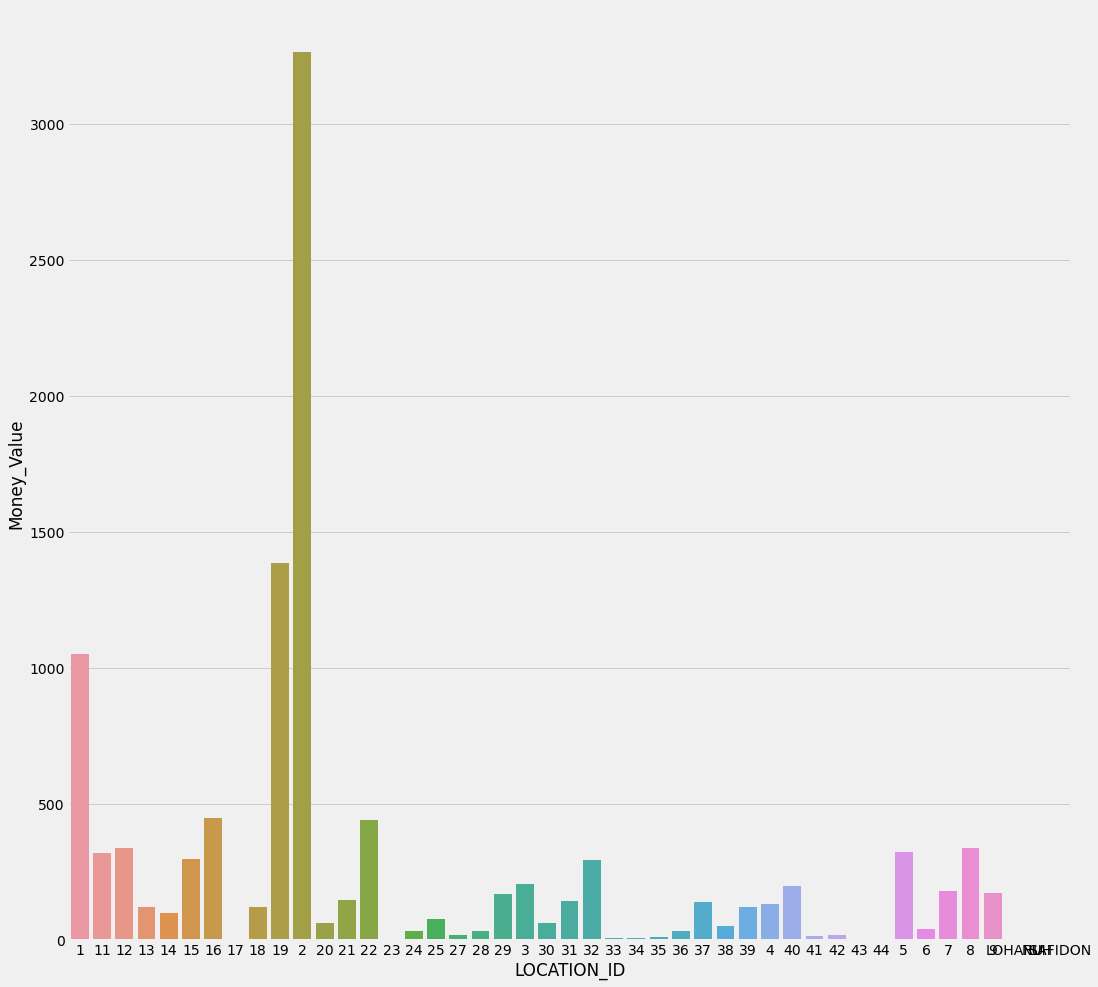


Figure 3: Bar graph showing Location ID against Money value

After just comparing a few variables with each other the next step is testing the relationship between variables with risk variable. For testing this relationship we are first going to take out the risk variable out of the current dataset and make a new dataset with only risk in it. Then we are comparing all the other variables with it to see if those variables have a positive correlation or a negative correlation.

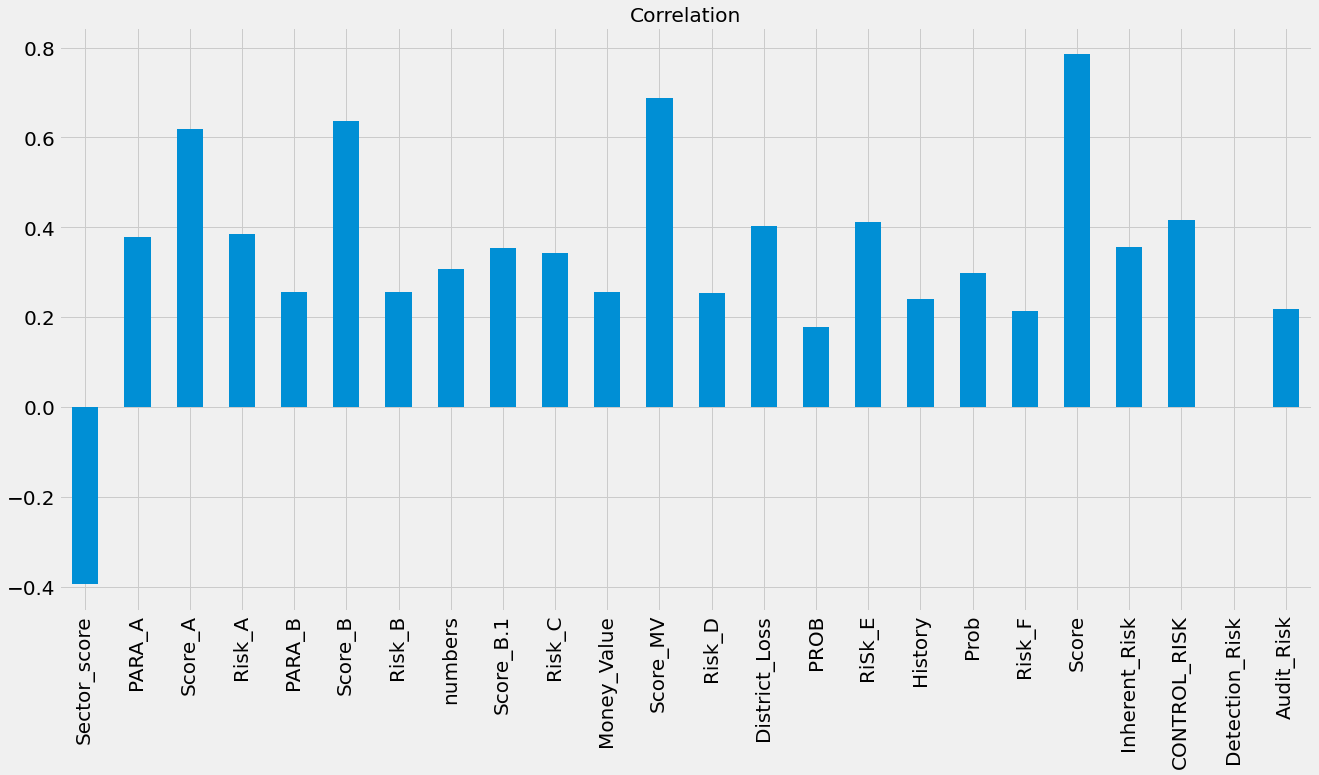


Figure 4: Bar graph showing correlation between risk and all the other variables

And the last step of the exploratory data analysis of the audit risk dataset is showing how many companies are fraudulent and non-fraudulent based on total risk. Now representing the number of the fraudulent and non-fraudulent companies can be done in two types of graphs, one being a bar graph and the other being a pie chart. Now a bar graph shows us how many number of companies are risky or not and/or fraudulent or not. Whereas a pie chart shows us what percentage of companies are risky or nor and/or fraudulent or not out of 776 companies from 14 different sectors.

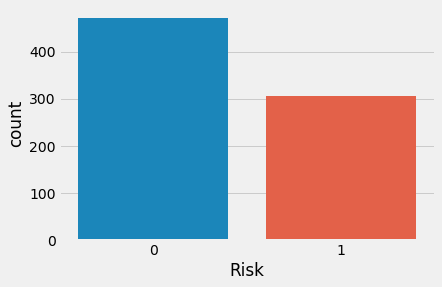


Figure 5: Bar graph showing number of fraudulent and non-fraudulent companies

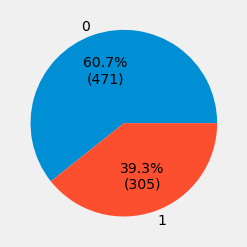


Figure 6: Pie chart showing the percentage of companies fraudulent or non-fraudulent.

This is the complete exploratory data analysis on the Audit Risk dataset which is takes to advice auditors on fraudulent and non-fraudulent firms by calculating the risk of the company based on present and historical data.