horizontal line

**PGP-DSE ONLINE JULY-2022**

**TEAM-06**

**Mini Project of EDA and Statistics**

# ****EDA and Statistical Analysis of Credit Card Dataset****

**Submitted By:-**

Sanamala Dasaratha Prem Kumar

Shreyas sachin Deshpande

Shekhar Singh

Reshma Begum

Ayush Kaushik

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Topic** | **Page No** |
| 1 | Abstract | 1 |
| 2 | Introduction, specification, | 1,2 |
| 3 | Need of fraudulent detection | 2 |
| 4 | Data dictionary | 2 |
| 5 | Part A | 3 |
| 6 | Part B | 10 |

## Abstract:

## 

A **credit card** is a [payment card](https://en.wikipedia.org/wiki/Payment_card) issued to users (cardholders) to enable the cardholder to pay a [merchant](https://en.wikipedia.org/wiki/Merchant) for [goods and services](https://en.wikipedia.org/wiki/Goods_and_services) based on the cardholder's accrued debt i.e., promise to the [card issuer](https://en.wikipedia.org/wiki/Credit_card_issuer) to pay them for the amounts plus the other agreed charges. The card issuer creates a [revolving account](https://en.wikipedia.org/wiki/Revolving_account) and grants a [line of credit](https://en.wikipedia.org/wiki/Line_of_credit) to the cardholder, from which the cardholder can borrow money for payment to a merchant or as a cash advance.

But there are cases were many fraudulent transactions happening through this credit card and we hereby studying the data set and finding insights regarding the fraudulent transaction and provide suggestions to avoid.

## Introduction:

## It is important that credit card companies are able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase.

## We will be suggesting the methods to avoid the same in transactions and will find how the fraudulent transaction happening and areas, amount, classes and places where this transaction happening.

**Specifications:**

**Part -A** is concept based of descriptive statistics, probability distributions and inferential statistics including confidence intervals and hypothesis testing.

**Part -B** on the other hand is dataset based and explore various data cleaning options, data analysis options and using EDA to derive deep and meaningful insights for the business

## Need for Fraudulent Detection:

Credit Card Fraudulent is a form of identity theft where criminals make purchases or obtain cash advances in your name. This can be with an existing account, via theft of your physical credit card or your account numbers and PINs, or by opening new credit card accounts in your name. Once they're in, thieves run up charges and stick you and your credit card company with the bill. Being defrauded is certainly a major inconvenience, but it's unlikely to cause lasting financial harm. Unfortunately, dealing with credit card fraud [has other unexpected costs](https://www.experian.com/blogs/ask-experian/what-are-unexpected-costs-of-identity-theft/), including time and aggravation. It can take months for credit card issuers to investigate fraud, and lingering issues can result in damage to your credit, which also takes time to remedy.

**Data Dictionary:**

a) It contains only numeric input variables. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, .. V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'.

b) Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset.

c) The feature 'Amount' is the transaction Amount,

d) Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

**Part A:**

The following are the ages of CEOs of 42 Top Fortune 500 Companies when they took over the post of CEO. Use this data for answering following questions where relevant.

57 61 57 57 58 57 61 54 68 51 49 64 50 48 65 52 56 46 54 49 50 47 55 55 54 42 51 56 55 54 51 60 62 43 55 56 61 52 69 64 46 54

**Compute the mean, median, mode, range, variance and standard deviation**

Mean of the data is 54.904761904761905

Median of the data is 55.0

Mode of the data is 0 54

Range of the data is 27

Variance of the data is 39.4541231126597

Standard Deviation of the data is 6.2812517154353635

Mean Deviation of data: 4.814058956916099

**Calculate the Pearson coefficient of skewness**

Pearson coefficient of skewness is -0.04548683903435661

Interpretation: The coefficient is negative but near zero. Thus, the data is said to be near symmetric.

**Count the number of data values that fall within two standard deviations of the mean**

Number of data values that fall within two standard deviations: 39

Percentage of data values that fall within two standard deviations: 92.85714285714286

According to Chebyshev's theorem more than 90% data fall within 2 standard deviation

**Find min, max and inter-quartile range**

Minimum: 40.875

Q1: 51.0

Q2: 55.0

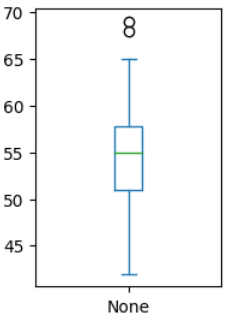
Q3: 57.75

Maximum: 67.875

IQR: 6.75

**Find outliers and plot**

Outliers - [68, 69]

****

**Find the percentile rank of the data point 50.**

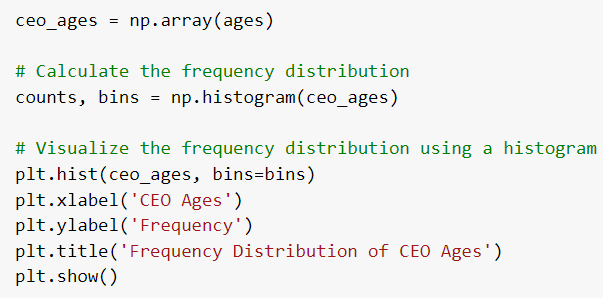
Number of 50 present in data: 2

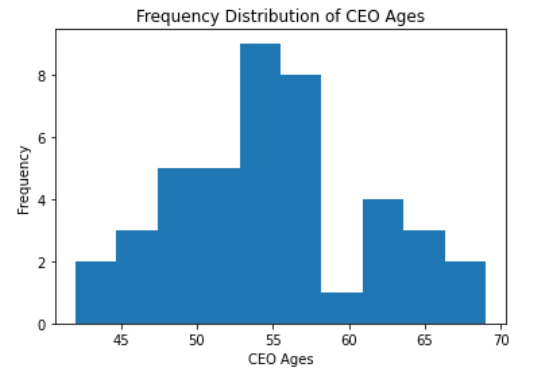
Percentile rank of 50: 4.761904761904762

**Find probability that a person becoming a CEO is below 50 years of age**

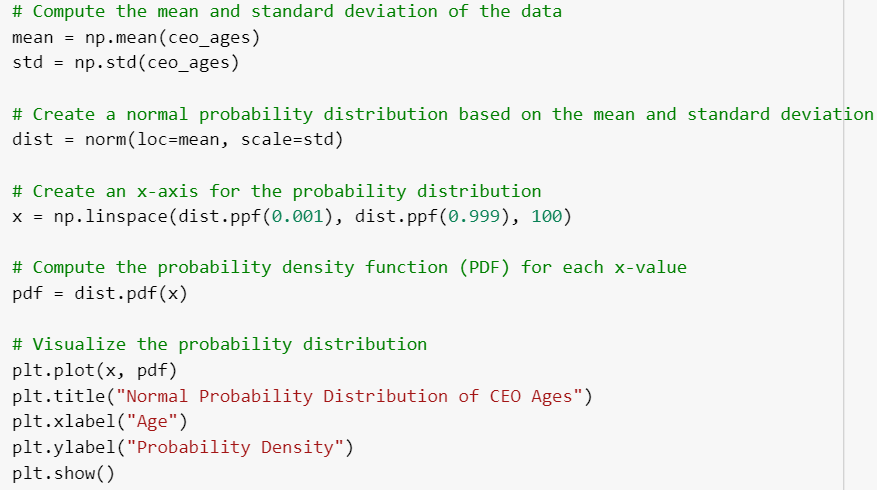
The probability of a person becoming a CEO below 50 years of age is: 0.2146696702826586

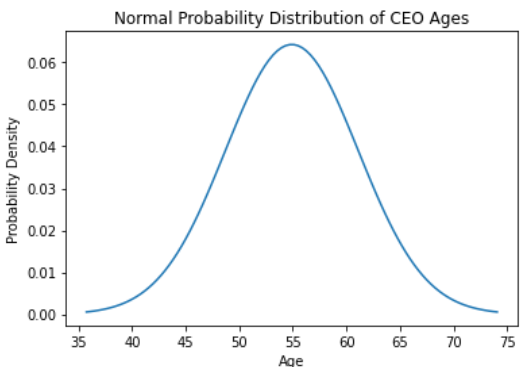
Frequency distribution for the data and visualize it appropriately



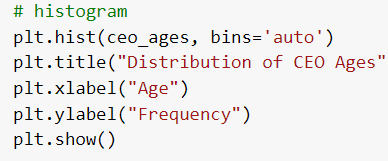


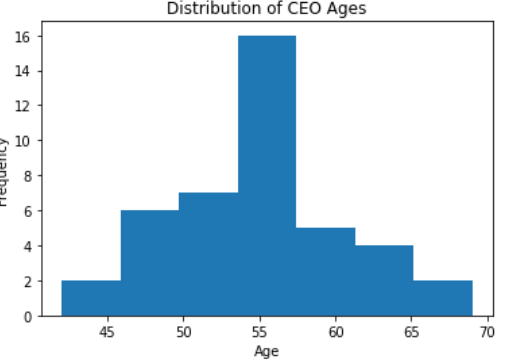
Create a probability distribution of the data and visualize it appropriately

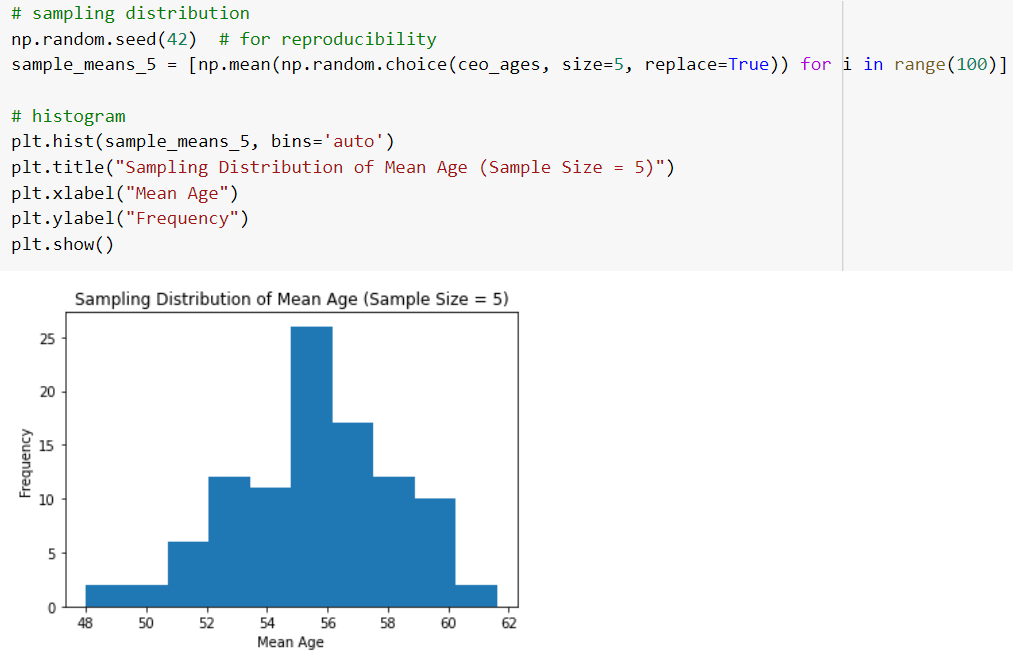


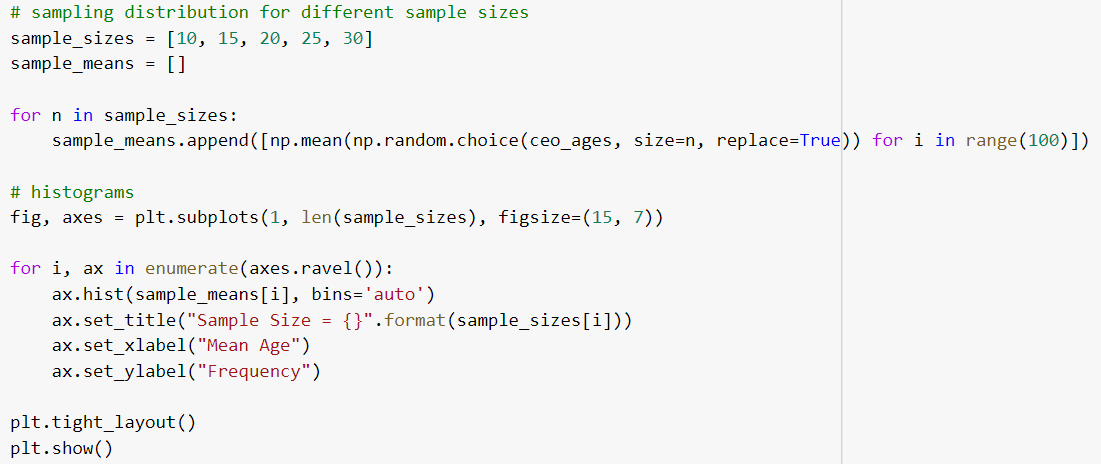


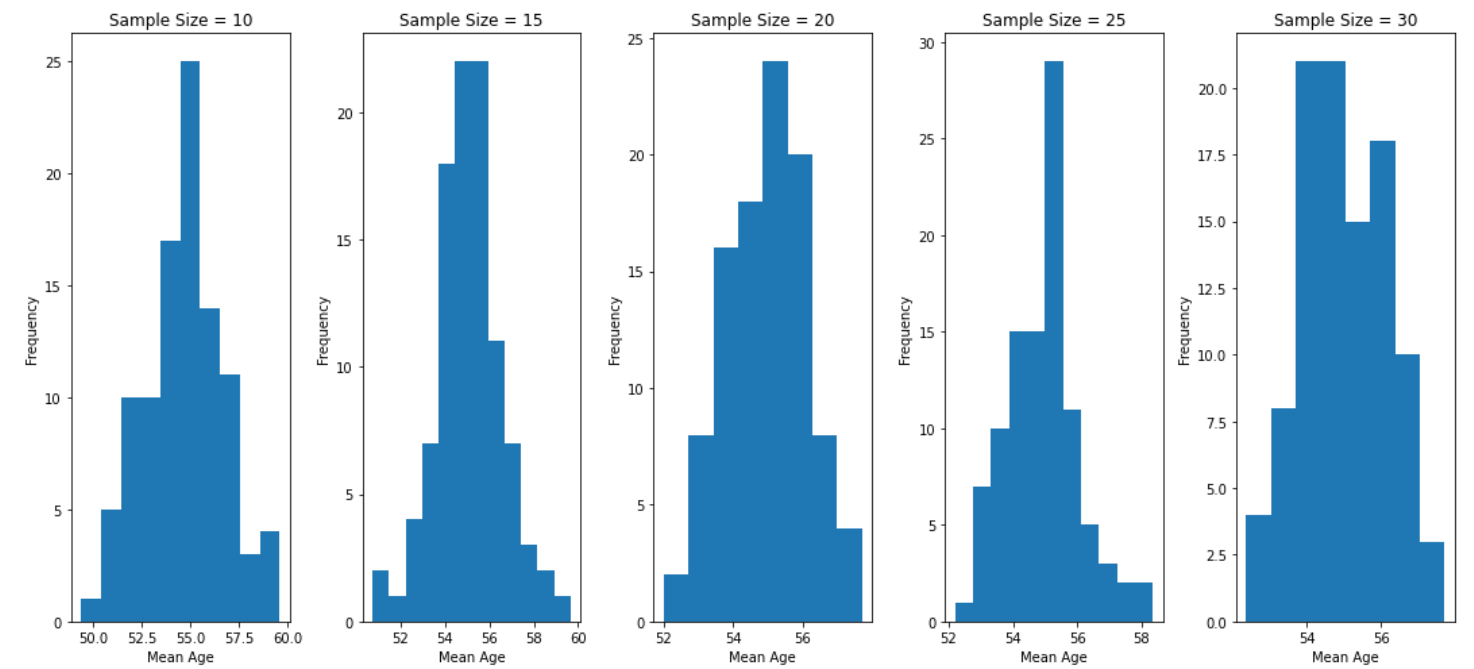
What is the shape of the distribution of this dataset? Create an appropriate graph to determine that. Take 100 random samples with replacement from this dataset of size 5 each. Create a sampling distribution ofthe mean age of customers. Compare with other sampling distributions of sample size 10, 15, 20, 25, 30. State your observations. Does it corroborate the Central Limit Theorem?



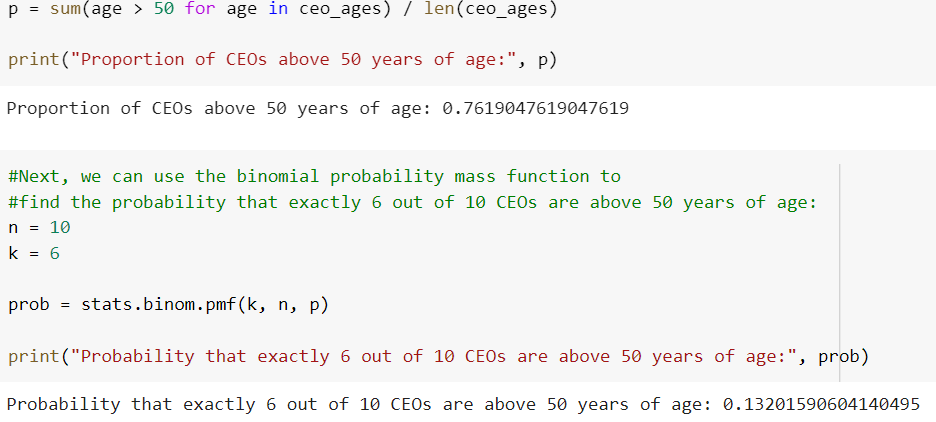




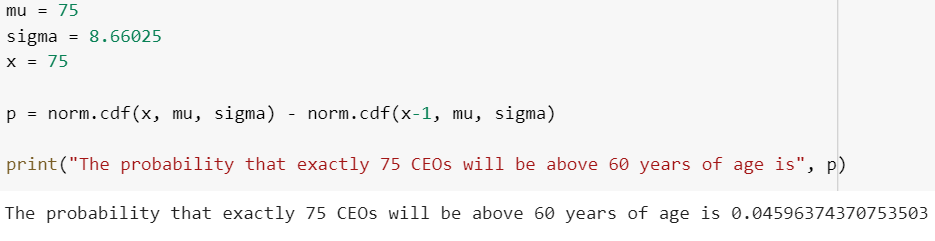




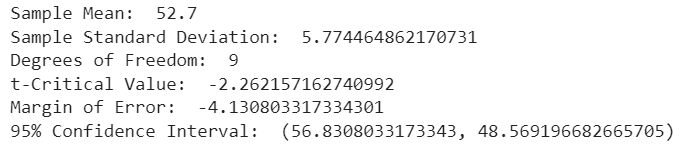
Treat this dataset as a binomial distribution where p is the probability that a person becomes a CEO above 50 years age. What is the probability that out of a random sample of 10 CEOs of Fortune 500 companies exactly 6are above 50 years of age?



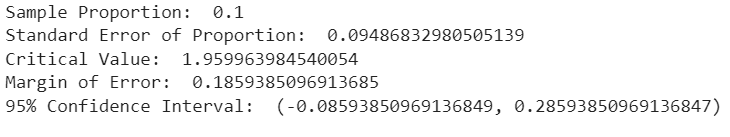
A study claims that 25% of all Fortune 500 companies becoming a CEOare above 60 years of age. Using the Normal approximation of a Binomial distribution, find the probability that in a random sample of 300 Fortune 500 companies exactly 75 CEOs will be above 50 years of age.



Compute a 95% Confidence Interval for the true mean age of the population of CEOs for the given dataset using appropriate distribution.( Statereasons as to why did you use a z or t distribution)



A data scientist wants to estimate with 95% confidence the proportion ofCEOs of Fortune 500 companies are above 60 years in the population. Another recent study showed that 25% of CEOs interviewed were above 60. The data scientist wants to be accurate within 2% of the true proportion. Find the minimum sample size necessary.



**Part B:**

## Project Statement:

The key objective is to able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase. We also find methods to analyses the transactions and insights to avoid the fraudulent transaction.

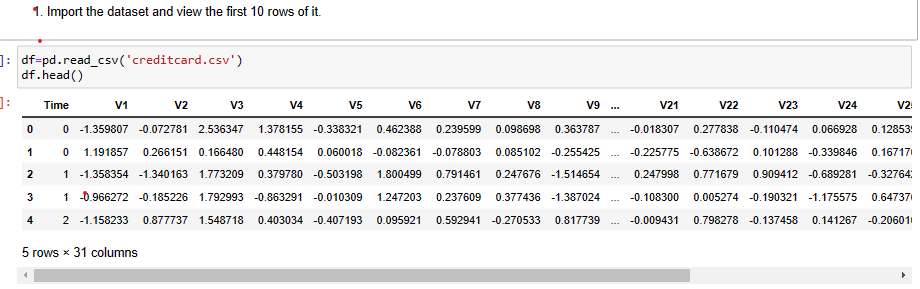
## Complexity involved:

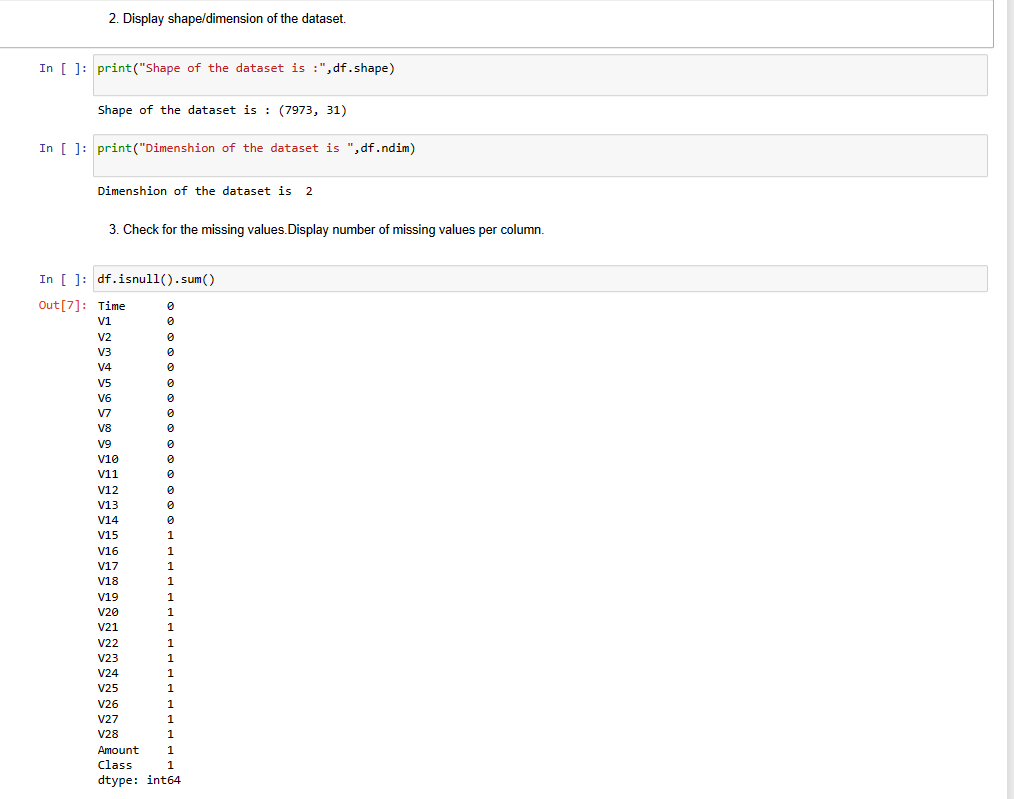
* From the dataset we understood that most of the variables in the dataset is not normally distributed and so that we need to transform them to normal distribution.
* The explanation about the variables and its name is not specified properly
* Categorization of the variables are quite difficult
* The number of variables in the dataset is high

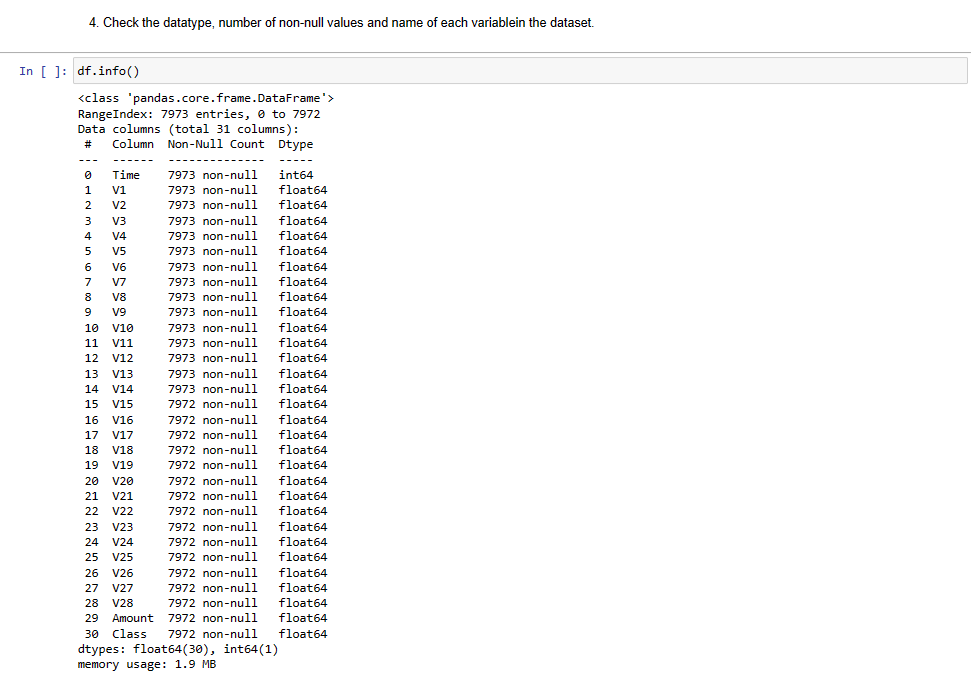
## Project outcome:

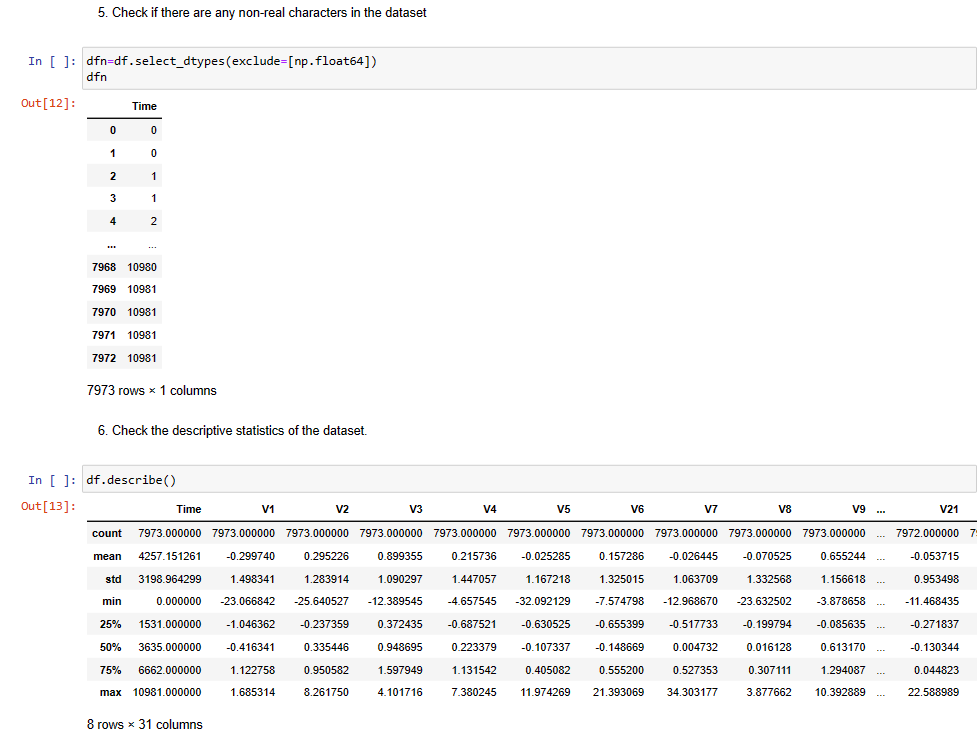
* + The outcome of the project is to find the fraudulent transaction happening in the credit card transactions and to find the insights in overall transactions and to study the details of how ,when and under what circumstances is the fraudulent transaction happening
  + Suggesting methods to reduce the fraudulent transactions , making all the transaction safer
  + Getting details on how the systematic transactions happening and what are the factors under which the whole process is running.

## We will be suggesting the methods to avoid the same in transactions and will find how the fraudulent transaction happening and areas, amount, classes and places where this transaction happening.



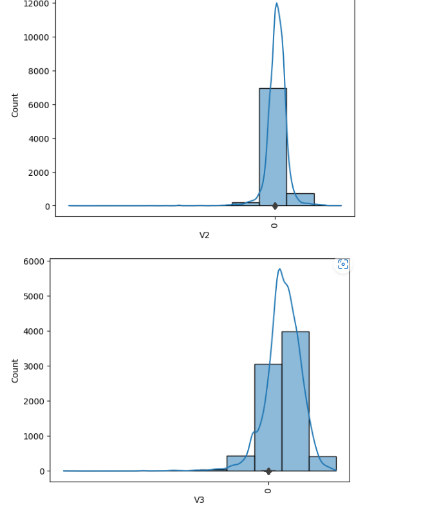
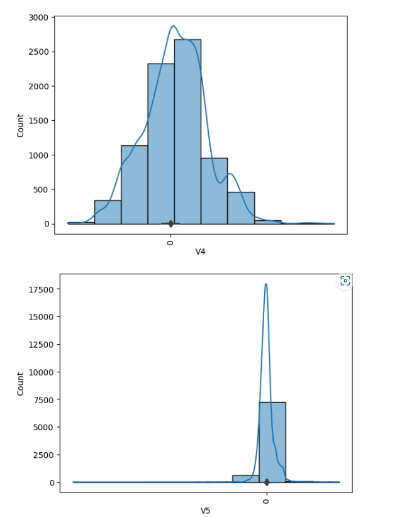
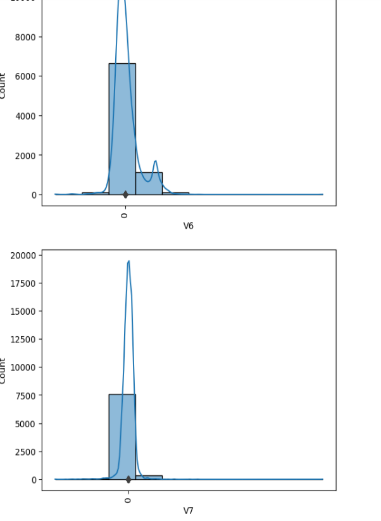


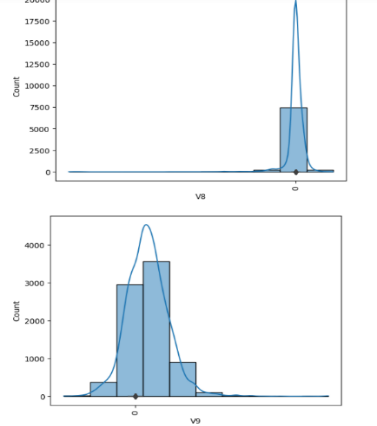
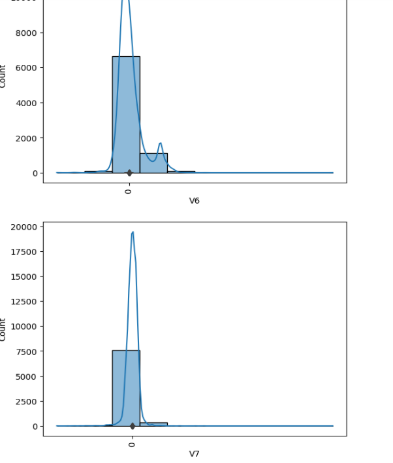


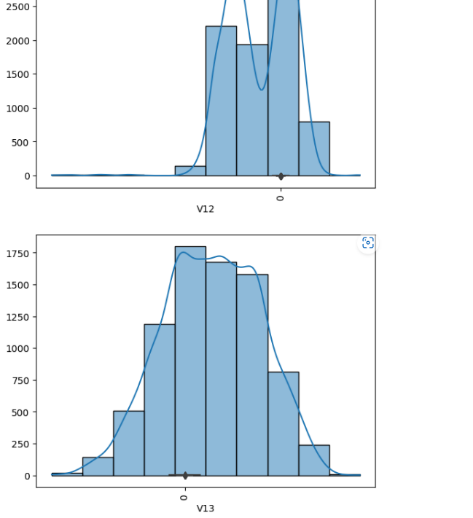
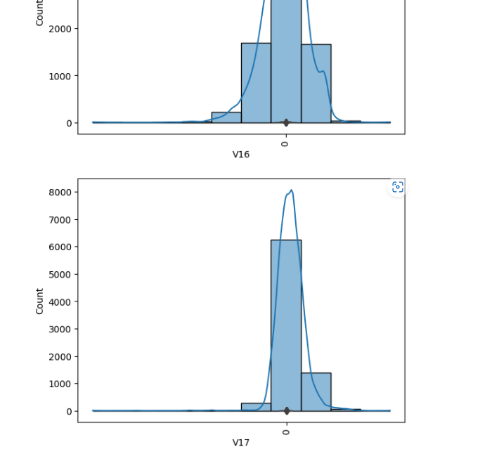


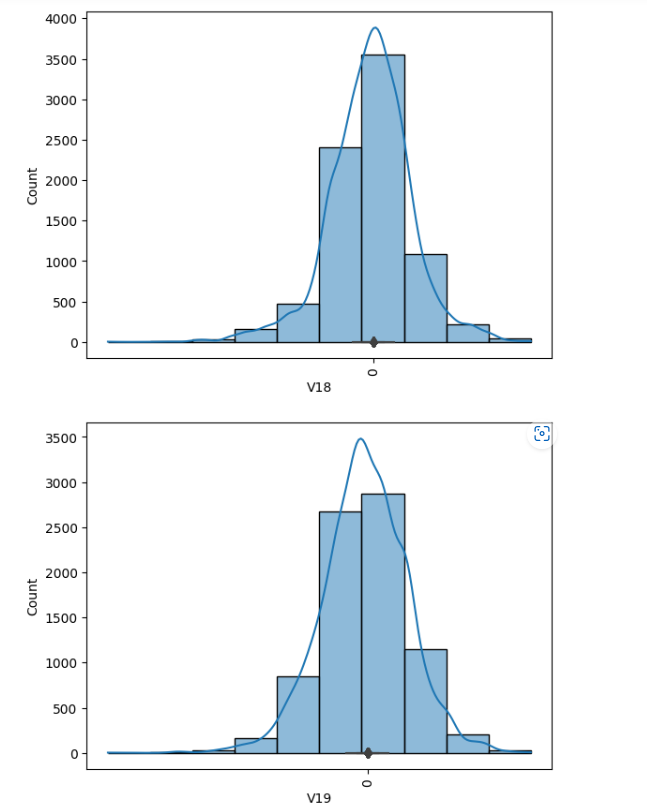
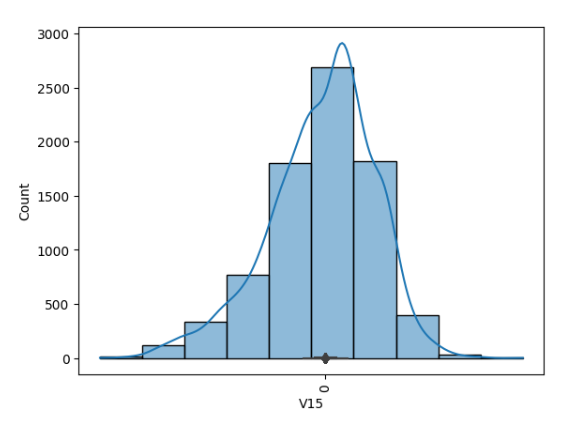
**Finding the Distribution of All Variables**

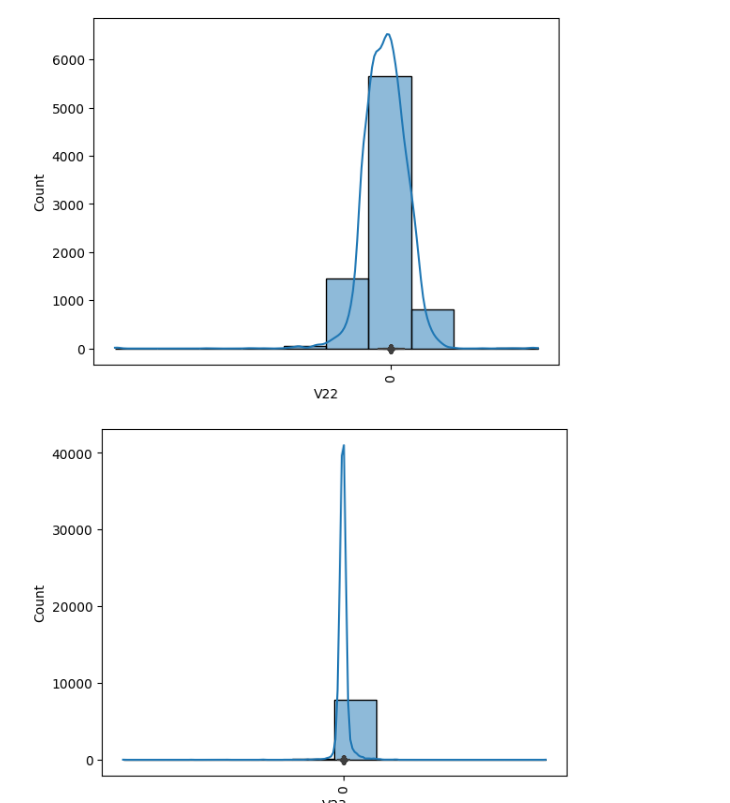
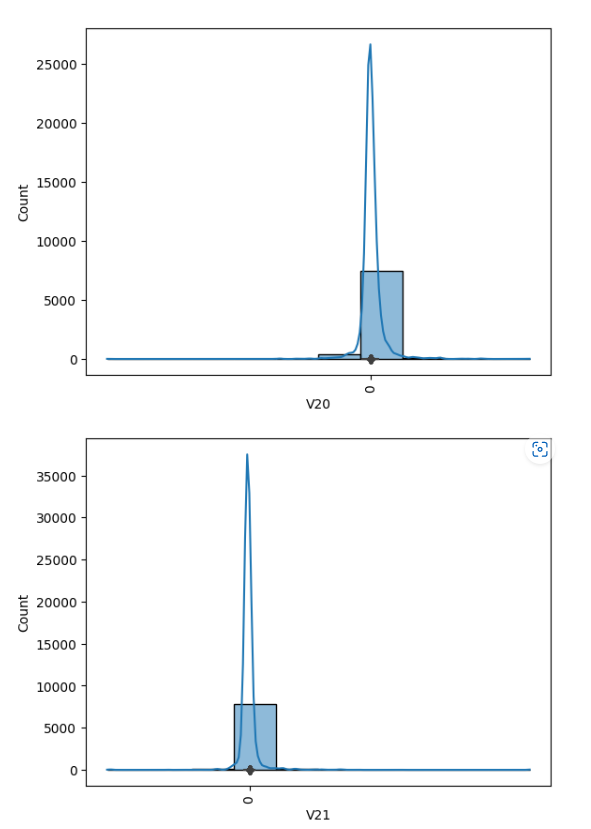
**Here we are plotting the distribution of the varibles in the data set to check whether it is normally distributed or not.**

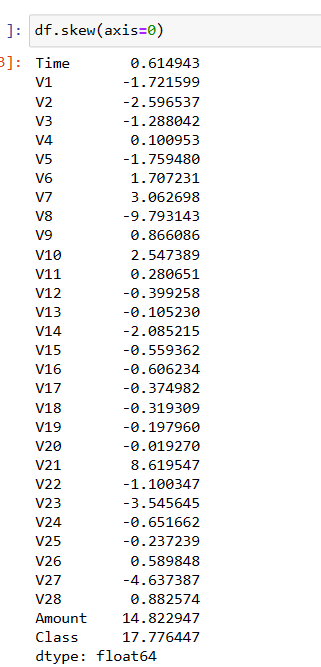




From this distribution we can see that Time,V1,V2,V3,V4,V24,V25,V26,V27,V28,Amount and Class are not normally distributed.

**Finding the skewness of the variable:**

From this dataset we can see positive skewed and negatively skewed variables so that we hereby finding out the skewness of the variables in the dataset

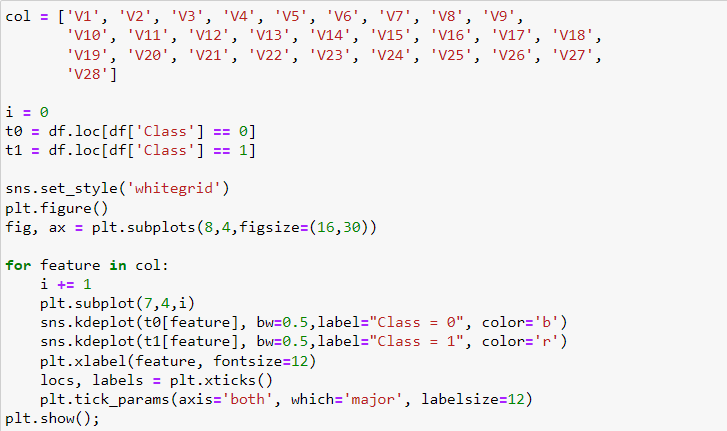


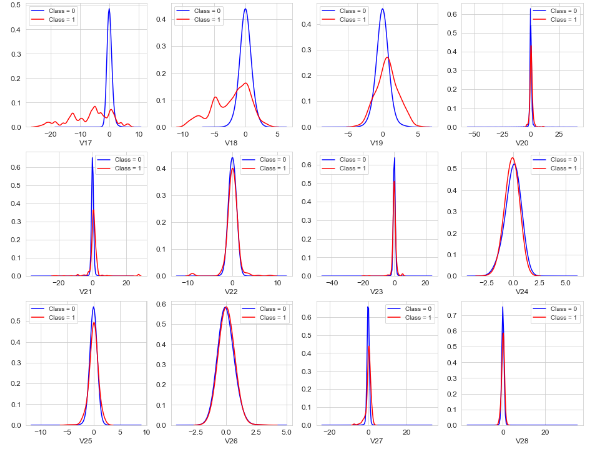
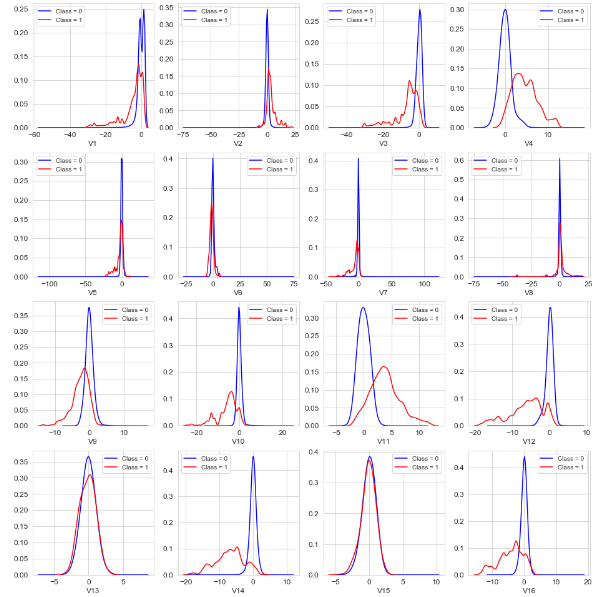
We can see that V8 and V23 are highly negatively skewed and V28 ,class and Amount are highly positively skewed.

**Distribution of all variables with respect to outcome 'Class'**

Plotting the variables with respect to ‘Class’ outcome using **kde plot.**

A kernel density estimate (KDE) plot is a method for visualizing the distribution of observations in a dataset, analogous to a histogram. KDE represents the data using a continuous probability density curve in one or more dimensions. Relative to a histogram, KDE can produce a plot that is less cluttered and more interpretable, especially when drawing multiple distributions. But it has the potential to introduce distortions if the underlying distribution is bounded or not smooth. Like a histogram, the quality of the representation also depends on the selection of good smoothing parameters.





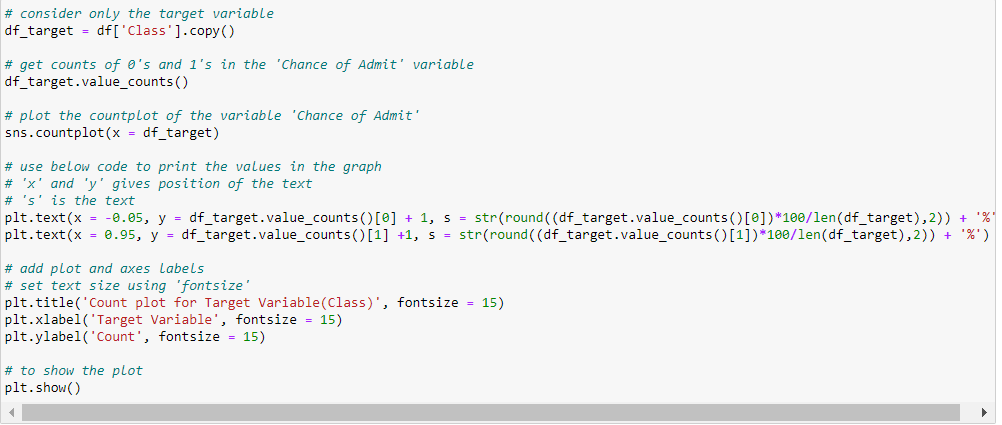
From the above graphs we can see that class outcome is imbalanced.

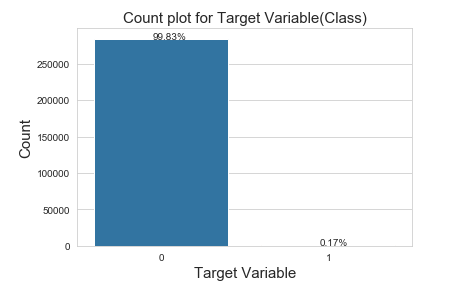
# Count plot for class with percentage

seaborn.**countplot**(*data=None*, *\**, *x=None*, *y=None*, *hue=None*, *order=None*, *hue\_order=None*, *orient=None*, *color=None*, *palette=None*, *saturation=0.75*, *width=0.8*, *dodge=True*, *ax=None*, *\*\*kwargs*)

Show the counts of observations in each categorical bin using bars.

A count plot can be thought of as a histogram across a categorical, instead of quantitative, variable.

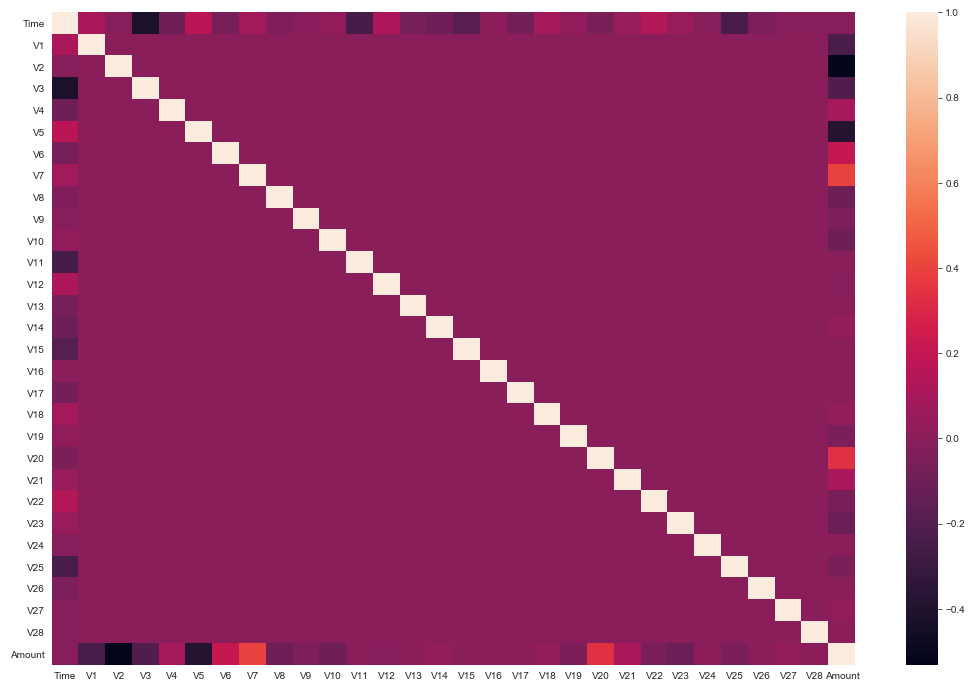


  
Here we can see that class 0 has 99.83% and Class 1 is 0.17%, which shows that class distribution is not balanced.

# Heatmap for Correlation matrix

**Heatmap** is defined as a graphical representation of data using colors to visualize the value of the matrix. In this, to represent more common values or higher activities brighter colors basically reddish colors are used and to represent less common or activity values, darker colors are preferred. Heatmap is also defined by the name of the shading matrix.

The heat-map above indicates that there are no high correlation values among the predictor columns. No predictor column has a high correlation value with the Class column either. However there exists a negative correlation among V2 and Amount as well as a positive correlation among V7 and the Amount feature.



# Hypothesis Testing for Fraudulent Transactions

If fraudulent transactions are in general of higher value than normal transactions or not. To check this lets do a hypothesis test. Lets define our Null and Alternative hypothesis.

H0 : Fraudulent transactions are of similar or higher value as normal transactions

HA : Fraudulent transactions are of lower value as normal transactions

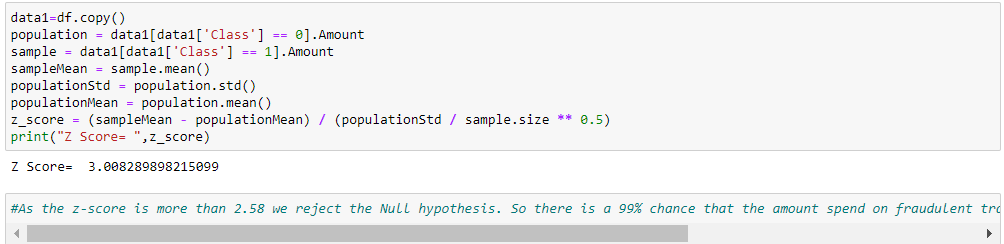
For the hypothesis test I will be performing a Z-test, with the valid transactions acting as the population. Though a T-test can also be performed but given that our sample set (fraudulent transactions) is of size 492 there shouldn't be any difference, as for sample set >= 30 the t distribution and z distribution are nearly the same.

We will be performing the test for 99% significance level, this means that we should get a z-score of atleast 2.58 or higher. formula for z-score, z−score=(x¯−μ)/S.D

Where

x¯ : mean of the sample μ : population mean S.D : Standard deviation of sampling diatribution The standard deviation of sampling distribution in our case is given by the formula : σ/√n , where σ is the Standard deviation of the population and n is the sample size

As the z-score is more than 2.58 we reject the Null hypothesis. So there is a 99% chance that the amount spend on fraudulent transactions are on average significantly lower than normal



**ANOVA test for Statistical Feature Selection**

ANOVA is used for testing two variables, where:

* one is a categorical variable
* another is a numerical variable

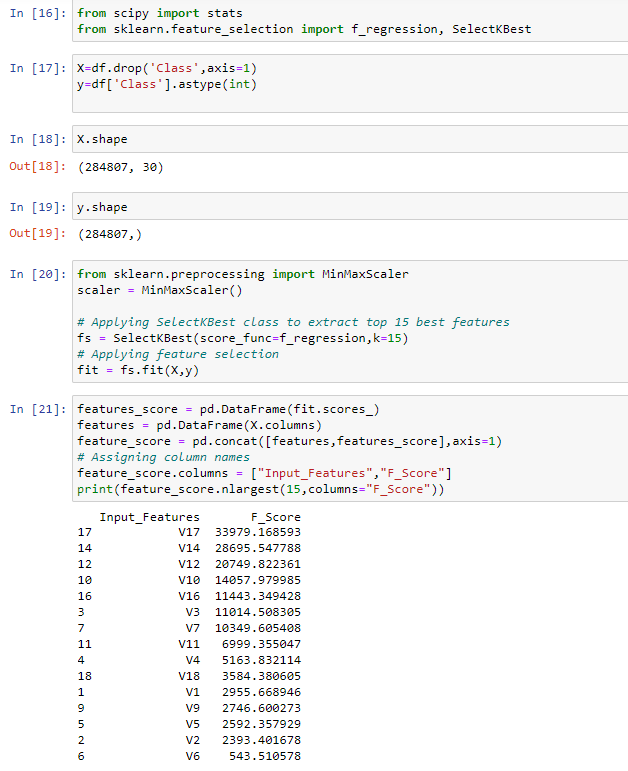
ANOVA is used when the categorical variable has at least 3 groups (i.e three different unique values).

If you want to compare just two groups, use the t-test. I will cover t-test in another article.

ANOVA lets you know if your numerical variable changes according to the level of the categorical variable.

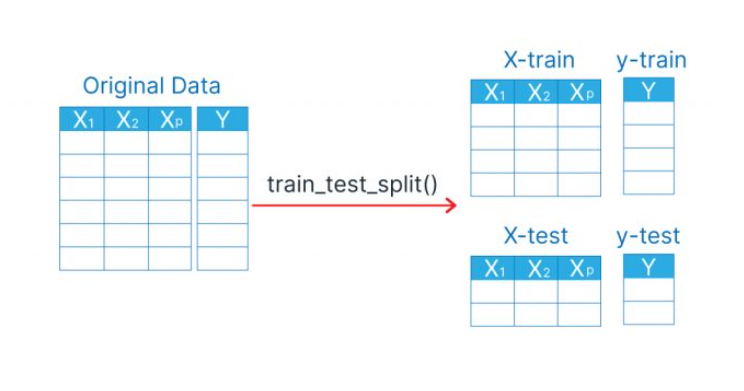
**ANOVA uses the f-tests to statistically test the equality of means**. F-tests are named after its test statistic, F, which was named in honor of **Sir Ronald Fisher**.

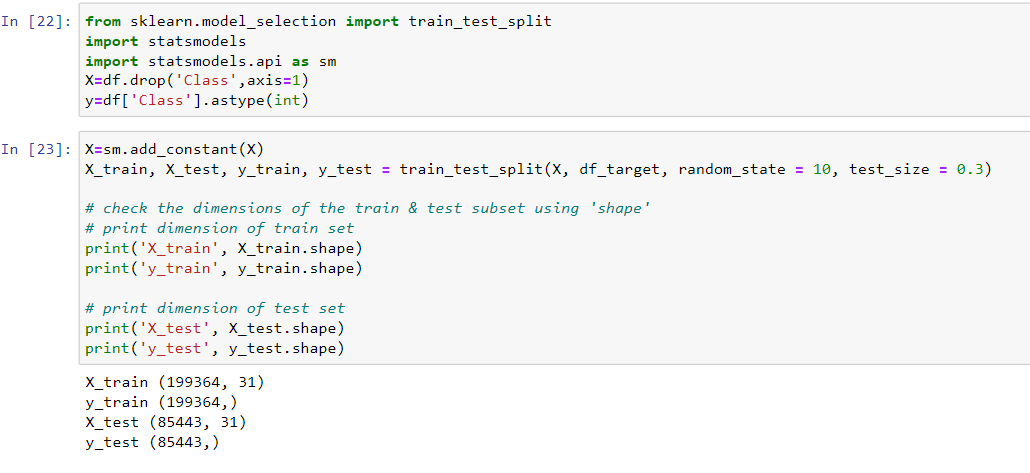
From the ANOVA test performed, we can see that we have listed the features in the descending order of their significance, the least significance can be eliminated.



# Split the dataset randomly into train and test datasets. Use a train -test ratio of 70:30 ratio

A train test split is when you split your data into a training set and a testing set. The training set is used for training the model, and the testing set is used to test your model. This allows you to train your models on the training set, and then test their accuracy on the unseen testing set. There are a few different ways to do a train test split, but the most common is to simply split your data into two sets. For example 80% for training and 20% for testing. This ensures that both sets are representative of the entire dataset, and gives you a good way to measure the accuracy of your models.

****

****