Mentors:

Shivam Pandey, Nivin Vinod

**Final Documentation**

CUSTOMER SEGMENTATION AND

CREDIT RISK MODELLING

By

Finance and Analytics club,

Science and Technology council, IIT Kanpur

Summer Project of

Acknowledgement:

Mentors: Shivam Pandey,Nivin Vinod

We are grateful to both of our mentors , Finance and Analytics Club, SnT for offering us an opportunity to learn through this project.The project was very insightful and practical.It introduced us to the amazing world of finance and machine-learning.We learnt about the intricacies of credit-risk modeling and some central concepts of ML like K-means algo, pipeline, test/train splitting and k-fold-cross-validation.On the least, We enjoyed learning through this project and definitely plan to venture further in this field…

**CONTENTS**

1. **Customer-Segmentation**
   1. What is Customer-Segmentation?
   2. Significance of Customer-Segmentation
2. **Dataset**
   1. Brief introduction of the Dataset
   2. Expected Outcome of the Dataset
3. **PCA**
   1. What is PCA
   2. Basics of PCA
   3. Why PCA
   4. Implementation of PCA
   5. Results obtained after applying PCA
4. **Clustering**
   1. Introduction to K-Means Clustering
   2. Elbow Method
   3. K-Means Algorithm
5. Credit-risk modelling

a- Train test split

b. Chi-squared test

c. ANOVA test

1. Weight of Evidence (WoE) and Information Value (IV)
2. Understanding binnig
3. Introduction to WoE
4. Introduction to IV
5. How to implement WoE and IV
6. Making a PD model
7. Steps involved in PD model
8. K-fold cross validation
9. ROC curve
10. Confusion matrix
11. Threshold value
12. ROC
13. AUC
14. Credit Score
15. What is credit score
16. Youden’s J-statistic
17. Setting loan approval cutoffs

* About Customer Segmentation

2. **Definition:**

* The customer segmentation process refers to audience segmentation and targeting. It is a way to group similar customers together in order to do reporting and analysis on the health of database.
* The whole concept is about the skilful categorization of the target audience and loyal customers. By categorization what is meant is ***dividing customers into various segments and creating targeted marketing campaigns.***

2. **How to achieve this segmentation**

* You can [segment your customer base](https://www.newbreedrevenue.com/blog/database-segmentation) by their industry, product tier or usage level, revenue, employee size, geographical locations, date they became a customer, channel their contact originated from, buyer persona and ideal customer profile.
* Grouping by these data points can help you identify your [ideal customer profiles](https://www.newbreedrevenue.com/blog/what-is-an-ideal-customer-profile) and determine which types of prospects you should avoid selling to.

3. **Benefits of Customer Segmentation**

* **More Customer Retention**

§ A conventional marketer will always prioritize its existing customers more than the new customers.

§ **About 75% of satisfied customers** are more likely to remain with that organization that regularly meets up their needs.

§ A loyal customer always is an advocate among the vast crowd. Hence, retaining valued customers will make business stand stronger. Segmenting customers based on their value will help in retaining them easily with the targeted strategies and campaigns.

* **Improved Communication with customer**

§ Customer segmentation enables you to identify patterns in the way you’ve acquired customers and the way different segments have engaged with your company.

§ This allows you to gain a better [understanding of what resonates](https://www.newbreedrevenue.com/blog/why-personas-matter-to-marketing-sales-and-service) with different types of prospects during the marketing and sales process and what kinds of communications they prefer to receive during service delivery

* **Enhanced competitiveness**

§ The more customers retained, the more revenue generation. And all of this will enhance competitiveness in the market.

§ Segmenting the customers helps targeting a specific audience for the campaign consequently, saving from bugging all the customers. As a result customers will be more inclined towards the marketing efforts.

§ This will eventually give marketing campaigns a better response than expected and enhance competitive spirit in the market as well.

### **Leads To Price Optimization**

§ Segmenting the customers on the basis of their expenditure will help in analyzing their spending limits. As a result, companies can price their products in a way that fits customers’ budgets.

§ On a bigger scale, financial analysis saves companies from losing customers because the products are too pricey.

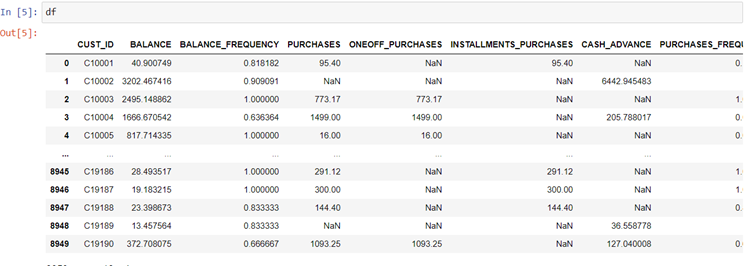
* About the dataset

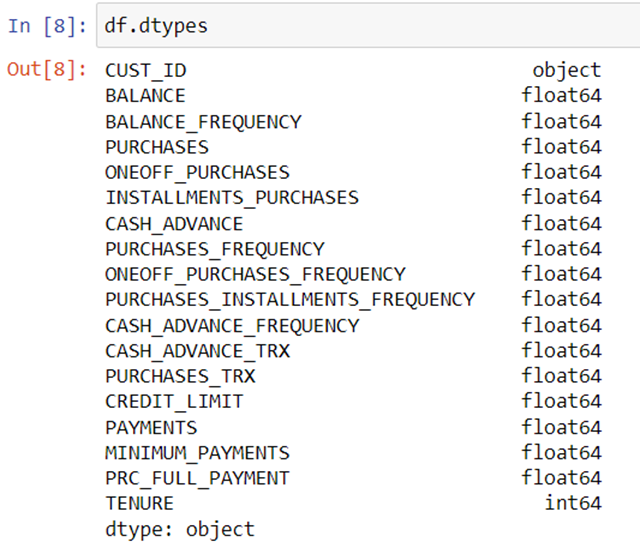
§ The dataset used for the customer segmentation part comprises the different types of attributes related to credit cards of customers.

§ There is data of 8949 customers and for each customer their data have been collected under 17 different columns pertaining to different attributes.

§ Data corresponding to each customer are under the following heads:

* **BALANCE**
* **BALANCE FREQUENCY**
* **PURCHASES**
* **ONE OFF PURCHASES**
* **INSTALLMENTS PURCHASES**
* **CASH ADVANCE**
* **PURCHASES FREQUENCY**
* **ONEOFF PURCHASES FREQUENCY**
* **PURCHASES INSTALLMENTS FREQUENCY**
* **CASH ADVANCE FREQUENCY**
* **CASH ADVANCE TRX**
* **PURCHASES TRX**
* **CREDIT LIMIT**
* **PAYMENTS**
* **MINIMUM PAYMENTS**



****

**DATATYPE OF DIFFERENT COLUMNS**

* **Data Cleaning**

**1. Definition**

**·** Data cleaning, or data cleansing, is the important process of correcting or removing incorrect, incomplete, or duplicate data within a dataset.

· Data cleaning should be the first step in your workflow. When working with large datasets and combining various data sources, there’s a strong possibility you may duplicate or mislabel data. If you have inaccurate or incorrect data, it will lose its quality, and your algorithms and outcomes become unreliable.

**2. What is a cleaned dataset**

**Consistency:** Data should be consistent across the database

**Accuracy:** Data should be close to the true values

**Completeness:** Data should include all the required information

**Validity:** Data should correspond with business rules and/or restrictions

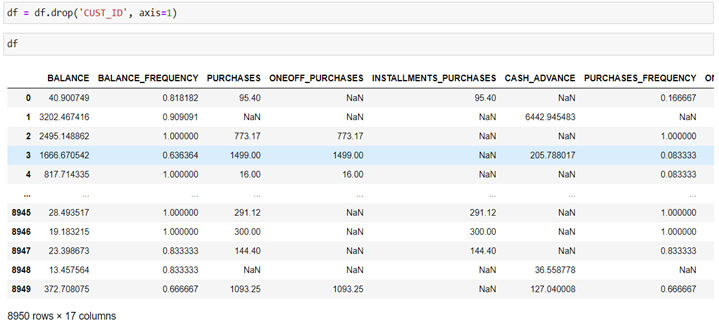
**Uniformity:** Data should be specified using consistent units of measurement

**3. Steps followed for data cleaning**

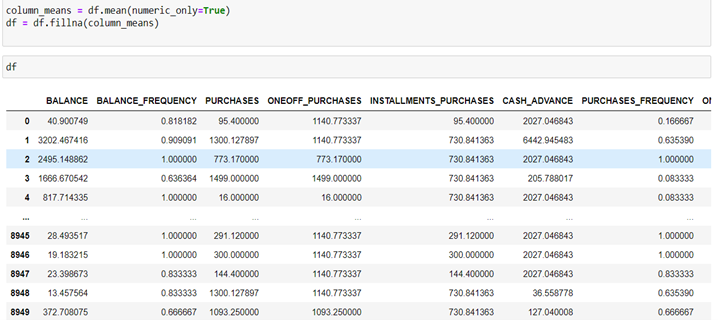
* Remove the columns which do not have data type as int or float as computations cannot be done on such columns.
* Check for the percentage of NaN values in a column, if the percentage exceeds 70 % then remove the column from the dataset.
* For the remaining columns try to fill the NaN the values using any of the following methods

§ fill it by the mean of the remaining values of that column

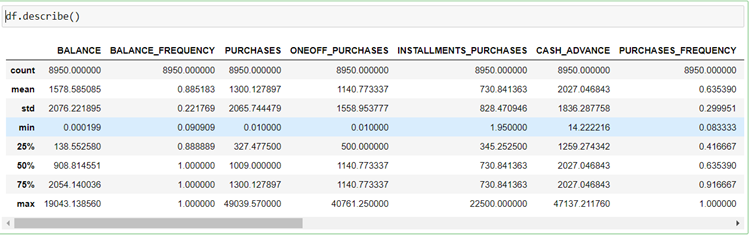
§ fill it by the median of the remaining values of that column

****

**Dropping the column with object as data type**

****

**Filling the NaN values with the mean**

****

**Description of data frame after cleaning**

**Skewness of Data**

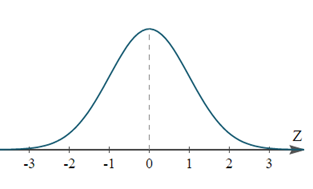
1. Definition

§ Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or [normal distribution](https://www.investopedia.com/terms/n/normaldistribution.asp), in a set of data. If the curve is shifted to the left or to the right, it is said to be skewed.

§ Skewness can be quantified as a representation of the extent to which a given distribution varies from a normal distribution.

# 2. Normal Distribution

§ In a normal distribution, the graph appears as a classical, symmetrical “bell-shaped curve.” The mean, or average, and the mode, or maximum point on the curve, are equal.

****

**Types of Skewness:**

In a symmetric bell curve, the mean, median, and mode are all the same value. But in a skewed distribution, the mean, median, and mode are all different values.

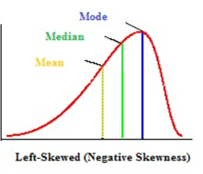
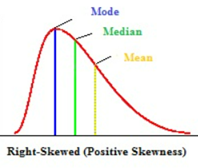
A skewed data distribution or bell curve can be either positive or negative.

### Positively Skewed Distribution

A positively skewed distribution means that the extreme data results are larger. This skews the data in that it brings the mean (average) up. The mean will be larger than the median in a Positively skewed distribution.

### Negatively Skewed Distribution

A negatively skewed distribution means the opposite: that the extreme data results are smaller. This means that the mean is brought down, and the median is larger than the mean in a negatively skewed distribution.\



# **4 Reducing skewness**

A data transformation may be used to reduce skewness. A distribution that is symmetric or nearly so is often easier to handle and interpret than a skewed distribution. More specifically, a normal or Gaussian distribution is often regarded as ideal as it is assumed by many statistical methods.

### **§ Square root:**

The square root, x to x^ (1/2) = sqrt(x), is a transformation with a moderate effect on distribution shape: it is weaker than the logarithm and the cube root. It is also used for reducing right skewness, and also has the advantage that it can be applied to zero values. Note that the square root of an area has the units of a length. It is commonly applied to counted data, especially if the values are mostly rather small.

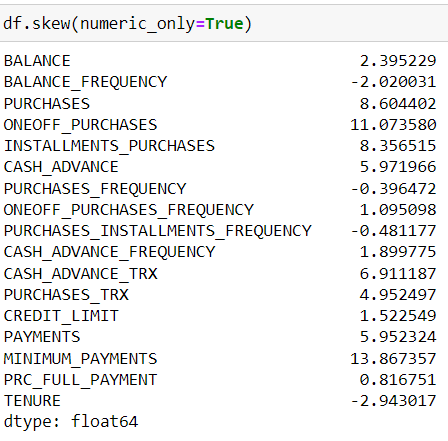
### § **Cube root**

The cube root, x to x^ (1/3). This is a fairly strong transformation with a substantial effect on distribution shape: it is weaker than the logarithm. It is also used for reducing right skewness, and has the advantage that it can be applied to zero and negative values. Note that the cube root of a volume has the units of a length. It is commonly applied to rainfall data.

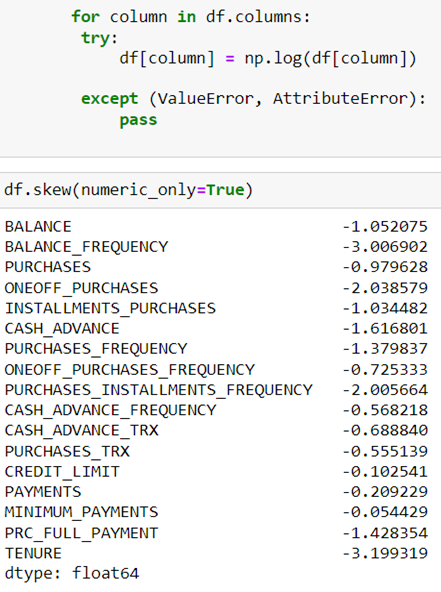
### § **Logarithms**

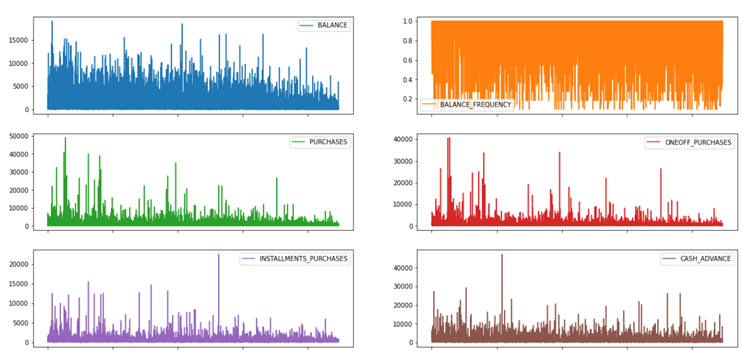
The logarithm, x to log base 10 of x, or x to log base e of x (ln x), or x to log base 2 of x, is a strong transformation with a major effect on distribution shape.

**Skewness index without transforms:**

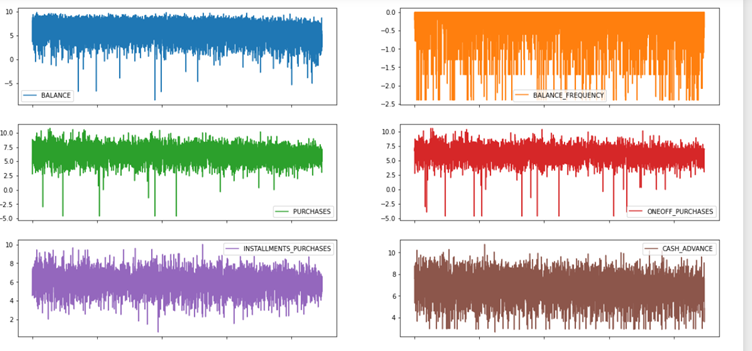
****

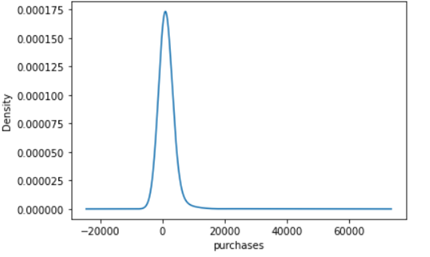
* **Skewness index after transforms:**

****

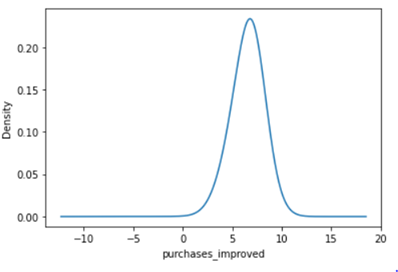
****

**(Subplot of columns before transformation)**

**(Subplot of columns after transformation)**

****

**KDE plot of purchases before transformation**

****

**KDE plot of purchases after transformation**

* **Standardization of data**

**1. Definition**

* Data standardization is the process used to ensure that internal data is consistent, each data type needs to have the same content and format for it to be considered standardized data, making them easier to track and analyze.
* Data standardization helps improve the quality of your data by transforming and standardizing it. Think of it like a uniform for your databases. By taking this step, you are formatting your records in a way that creates consistency across your systems and makes it easy for businesses to use.

**2. Benefits of standardization**

* [**Cost Reduction**](https://winpure.com/blog/cost-of-poor-data-quality/)**:**

Currently, many organizations collect data in free-format which usually leads to all kinds of errors. These include inconsistent use of abbreviations, spelling errors, duplicates, and so on. By using Word Manager you can make significant gains in cost, time and manpower.

**Improved productivity**

Word Manager will save your staff’s time by providing them clean records. For example, your sales team will not have to waste their time trying to contact customers using inaccurate or obsolete information.

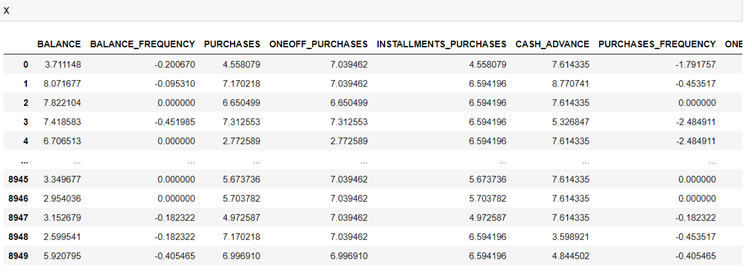
**Improved Customer Acquisition**

Accurate data, cleansed, and standardized, can help in the marketing process. It brings better returns on your email or postal campaigns.

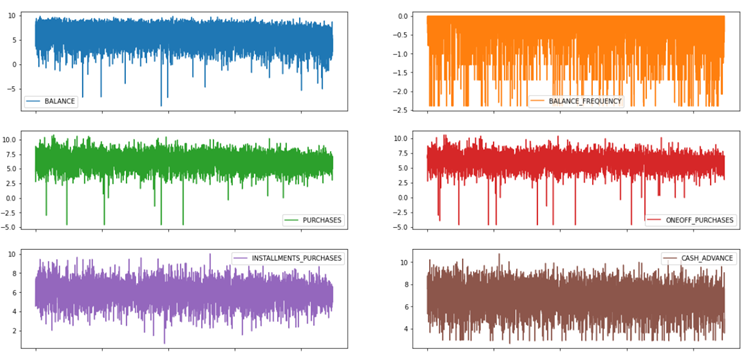
**3 How to accomplish standardization:**

****

**(Standardization using scikit learn)**

****

**Data frame after standardization of data**

****

**Subplot after data standardization**

**Co-variance**

**1. Definition**

* In mathematics and [statistics](https://corporatefinanceinstitute.com/resources/knowledge/basic-statistics-concepts/), covariance is a measure of the relationship between two random variables. The metric evaluates how much – to what extent – the variables change together.
* In other words, it is essentially a measure of the variance between two variables. However, the metric does not assess the dependency between variables.

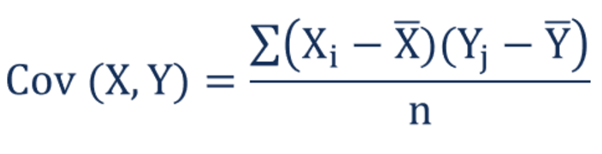
**2. Significance**

The variance can take any positive or negative values. The values are interpreted as follows:

* Positive covariance: Indicates that two variables tend to move in the same direction.
* Negative covariance: Reveals that two variables tend to move in inverse directions.

**7.3 Formula to calculate covariance**

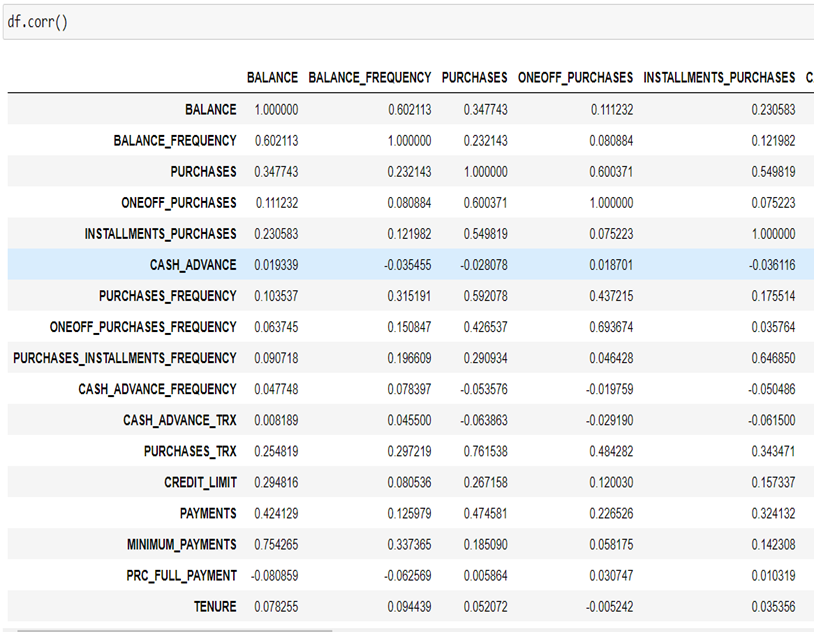
The covariance between two random variables X and Y can be calculated using the following formula:



* **Xi** – the values of the X-variable
* **Yj**– the values of the Y-variable
* **X̄** – the mean (average) of the X-variable
* **Ȳ** – the mean (average) of the Y-variable
* **n** – the number of data points

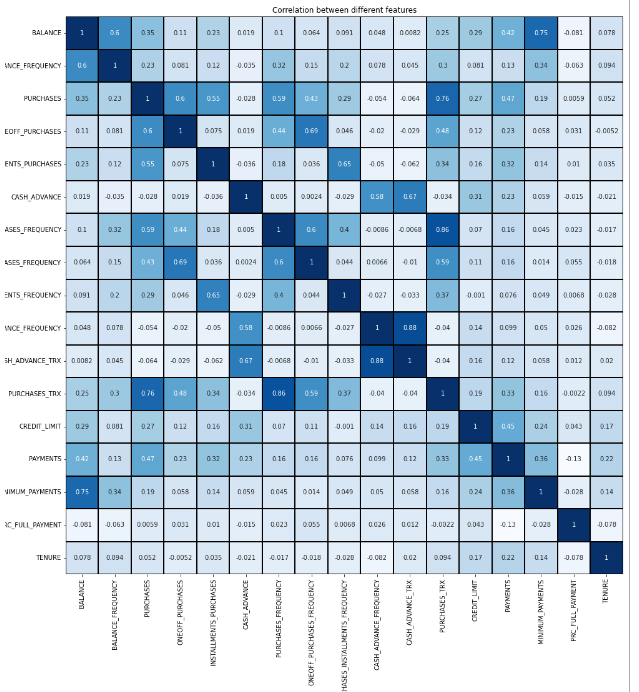
**4. Covariance Matrix**

* In probability theory and statistics, a covariance matrix (also known as auto-covariance matrix, dispersion matrix, variance matrix, or variance–covariance matrix) is **a square matrix giving the covariance between each pair of elements of a given random vector**.

**(Covariance Matrix)**

* **Plotting heat map to show this covariance matrix using seaborn**

**(Code Snippet)**

**Heat map using seaborn**

* **Features of the heat map:**
* It represents a symmetric matrix (17 x 17 in this case)
* Magnitude of covariance is determined by the darkness of color i.e. darker the color more is the covariance.
* The value of covariance along the main diagonal will always be 1 as it represents covariance between the same random variables.

**What is PCA?**

Principal component analysis (PCA) is a standard tool in modern data analysis. Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity. Because smaller data sets are easier to explore and visualize and make analyzing data much easier and faster for machine learning algorithms without extraneous variables to process.

So to sum up, the idea of PCA is simple — reduce the number of variables of a data set, while preserving as much information as possible.

# **Basics of PCA-(Principal component Analysis)?**

The Principal Component Analysis (PCA) is an unsupervised statistical technique that can be used for dimension reduction, feature extraction, and data visualization. PCA can analyze the data to identify patterns in order to reduce the dimensions of the dataset with minimal loss of information.

## **Why PCA?**

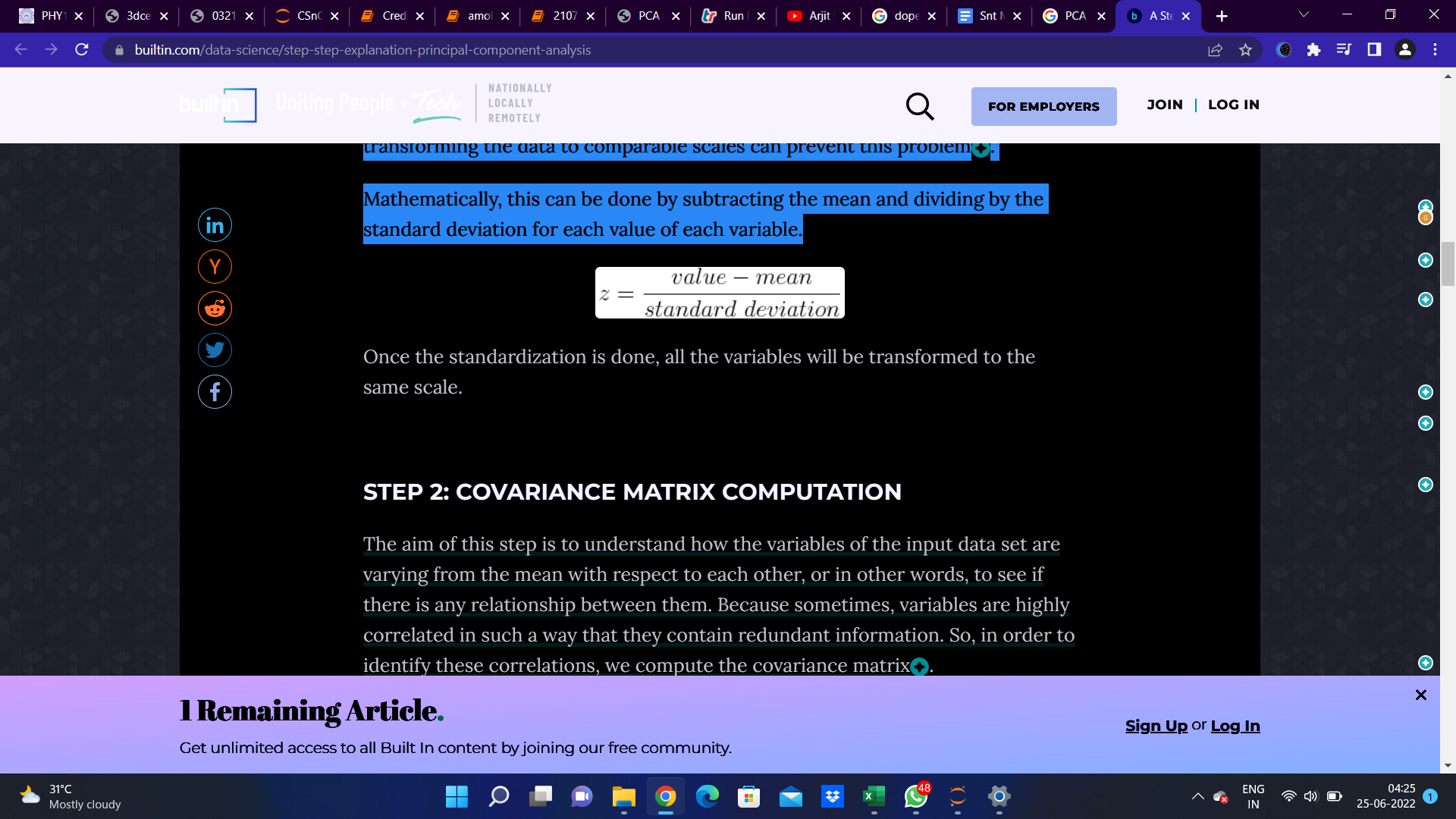
* High dimensionality means that the dataset has a large number of features, which demands loads of computational power and unnecessary waste of time and energy.
* By reducing the dimensions of datasets, PCA provides an effective and efficient method for data description and visualization.
* PCA produces a low-dimensional representation of the feature space by finding a sequence of linear combinations of the features that have maximal variance and are mutually uncorrelated. The proportion of variance explained by each principal component is a measure of the strength of each component.

**Assumptions for PCA:**

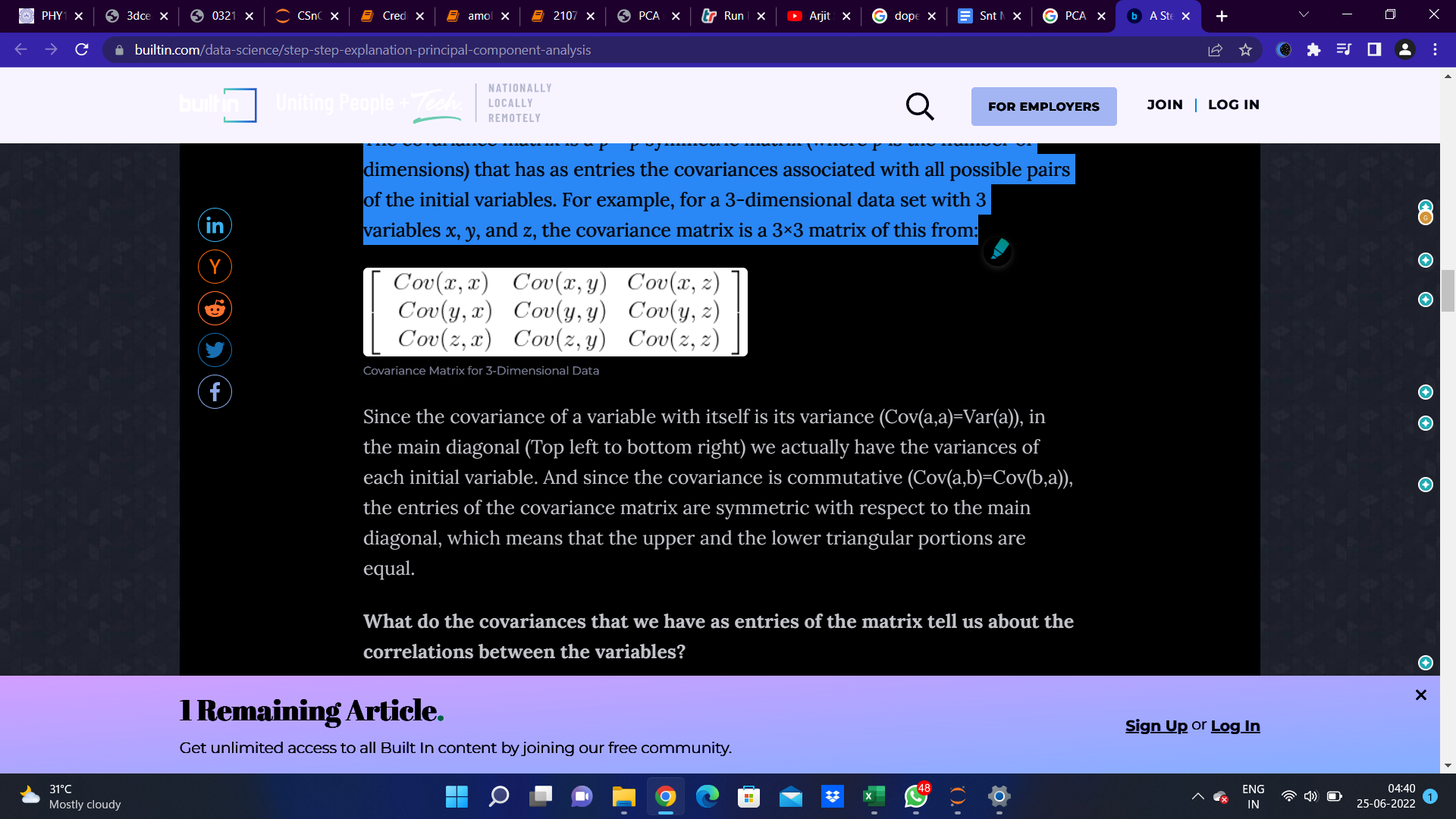
* Linearity Linearity frames the problem as a change of basis. Several areas of research have explored how extending these notions to nonlinear regimes.
* Large variances have important structure. This assumption also encompasses the belief that the data has a high SNR. Hence, principal components with larger associated variances represent interesting structure, while those with lower variances represent noise. Note that this is a strong, and sometimes, incorrect assumption
* The principal components are orthogonal. This assumption provides an intuitive simplification that makes PCA soluble with linear algebra decomposition techniqu

**Implementation of PCA**

**Step 1:** **Standardization of the data**

* The aim of this step is to standardize the range of the continuous initial variables so that each one of them contributes equally to the analysis.
* More specifically, the reason why it is critical to perform standardization prior to PCA, is that the latter is quite sensitive regarding the variances of the initial variables. That is, if there are large differences between the ranges of initial variables, those variables with larger ranges will dominate over those with small ranges (For example, a variable that ranges between 0 and 100 will dominate over a variable that ranges between 0 and 1), which will lead to biased results. So, transforming the data to comparable scales can prevent this problem
* Mathematically, this can be done by subtracting the mean and dividing by the standard deviation for each value of each variable.
* Once the standardization is done, all the variables will be transformed to the same scale.

### **STEP 2: Covariance Matrix Computation**

* The aim of this step is to understand how the variables of the input data set are varying from the mean with respect to each other, or in other words, to see if there is any relationship between them. Because sometimes, variables are highly correlated in such a way that they contain redundant information. So, in order to identify these correlations, we compute the covariance matrix.
* The covariance matrix is a *p* × *p*symmetric matrix (where *p* is the number of dimensions) that has as entries the covariances associated with all possible pairs of the initial variables. For example, for a 3-dimensional data set with 3 variables *x*, *y*, and *z*, the covariance matrix is a 3×3 matrix of this from:
* Since the covariance of a variable with itself is its variance (Cov(a,a)=Var(a)), in the main diagonal (Top left to bottom right) we actually have the variances of each initial variable. And since the covariance is commutative (Cov(a,b)=Cov(b,a)), the entries of the covariance matrix are symmetric with respect to the main diagonal, which means that the upper and the lower triangular portions are equal.

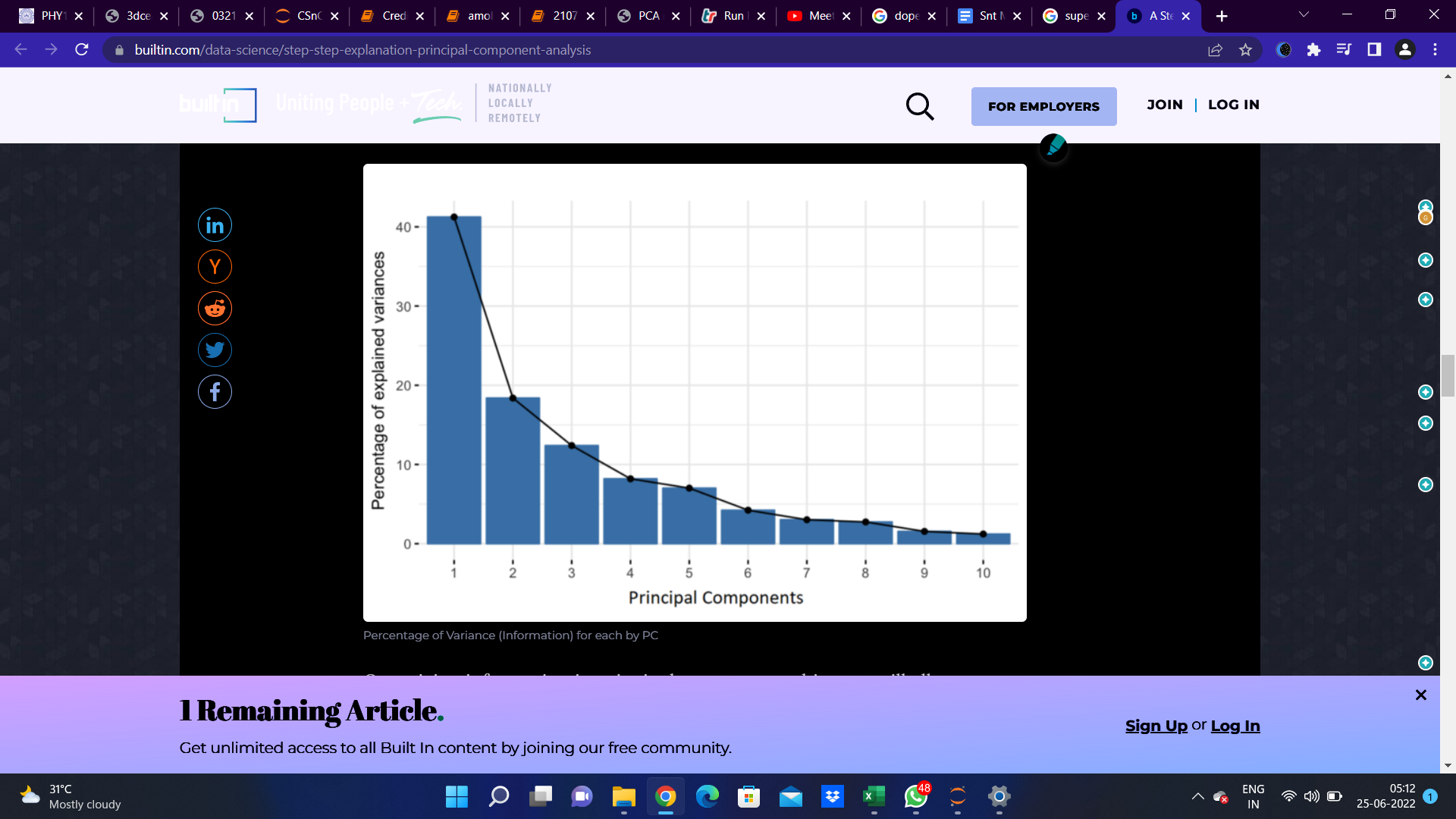
**#If the sign of the co-variance value is positive, the two variables Increase or decrease together.**

**#If the sign of the co-variance value is negative, it suggests that as One variable increases, other decreases, and vice-versa.**

### **STEP 3: Compute the Eigenvectors and Eigenvalues of the Covariance Matrix to identify the PRINCIPAL COMPONENTS**

Eigenvectors and eigenvalues are the linear algebra concepts that we need to compute from the covariance matrix in order to determine the ***principal components*** of the data. Before getting to the explanation of these concepts, let’s first understand what do we mean by principal components.

Principal components are new variables that are constructed as linear combinations or mixtures of the initial variables. These combinations are done in such a way that the new variables (i.e., principal components) are uncorrelated and most of the information within the initial variables is squeezed or compressed into the first components. So, the idea is 10-dimensional data gives you 10 principal components, but PCA tries to put maximum possible information in the first component, then maximum remaining information in the second and so on, until having something like shown in the scree plot below.



Organizing information in principal components this way, will allow you to reduce dimensionality without losing much information, and this by discarding the components with low information and considering the remaining components as your new variables.

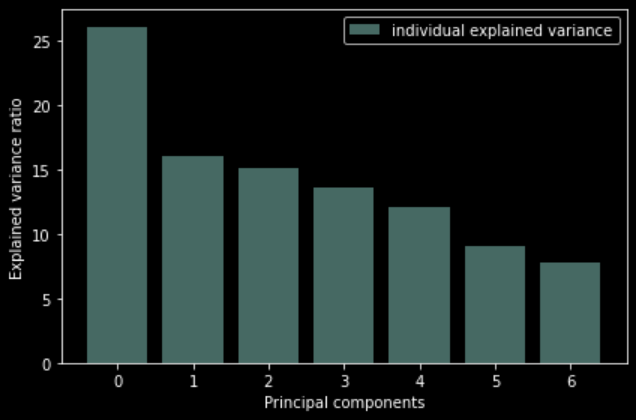
An important thing to realize here is that, the principal components are less interpretable and don’t have any real meaning since they are constructed as linear combinations of the initial variables

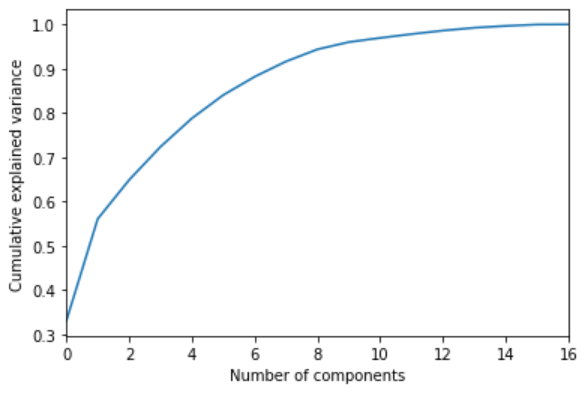
Geometrically speaking, principal components represent the directions of the data that explain a **maximal amount of variance**, that is to say, the lines that capture most information of the data. The relationship between variance and information here, is that, the larger the variance carried by a line, the larger the dispersion of the data points along it, and the larger the dispersion along a line, the more the information it has. To put all this simply, just think of principal components as new axes that provide the best angle to see and evaluate the data, so that the differences between the observations are better visible.

**Step 4:** **Select the columns with most amount of variance.**

As we saw in the previous step, computing the eigenvectors and ordering them by their eigenvalues in descending order, allow us to find the principal components in order of significance. In this step, what we do is, to choose whether to keep all these components or discard those of lesser significance (of low eigenvalues)

* Get the covariance matrix from the given data-set.
* Obtain corresponding eigen-values and eigen-vectors from the covariance matrix.
* Arrange the eigen-values in a descending order. Magnitude of each eigen-value would be proportional to its contribution in the total variance.
* Obtain the no. of eigen-values that cover most of the variance , let’s say 90%.
* Construct a matrix of the eigen-vectors corresponding to these eigen-values.
* Project the dataset onto a new space using this matrix.





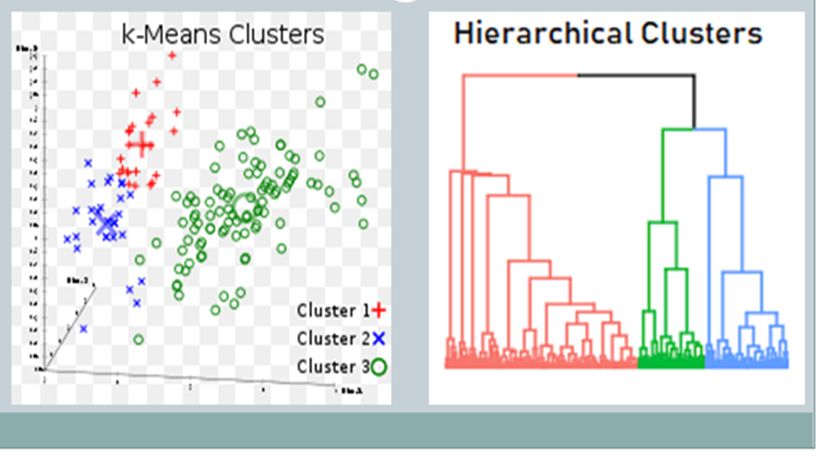
**Clustering of data**:

* **Clustering** is a classic data mining technique based on machine learning that divides groups of abstract objects into classes of similar objects. Clustering helps to split data into several subsets.
* Each of these clusters consists of data objects with high inter-similarity and low intra-similarity.
* It is an unsupervised learning method.

**Clustering Methods:**

1. **K-means Clustering**
2. **Hierarchical Clustering**

(In our project we have focussed on K-means Clustering.)



**K means clustering** :

* **K-means** clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups).
* The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K.
* In K-means clustering, each cluster is represented by its center (i.e., centroid) which corresponds to the mean of the observation values assigned to the cluster.

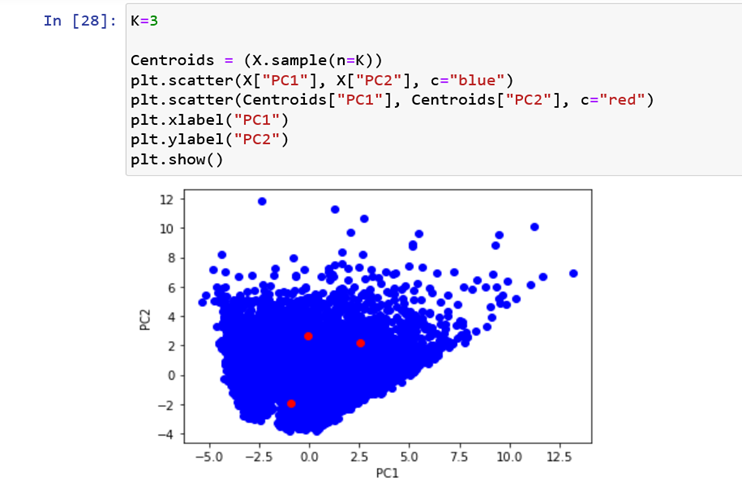
**Why K means** :

* K-means tries to classify observations into mutually exclusive groups (or clusters), such that observations within the same cluster are as similar as possible, whereas observations from different clusters are as dissimilar as possible.
* The basic idea behind k-means clustering is constructing clusters so that the total within-cluster variation is minimized.

**Centroid :**

A centroid of a given cluster is a point that is the mean of all the points of that cluster.

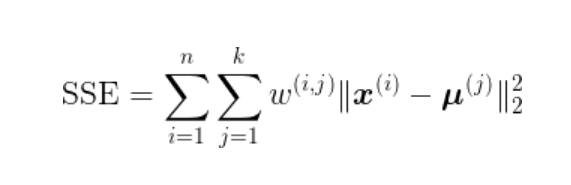
Initializing centroids to our dataset in python(randomly):

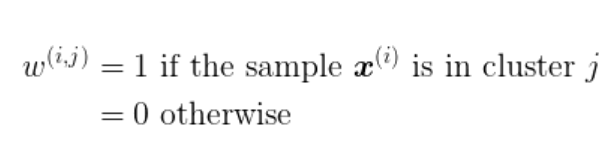


**Elbow Method**

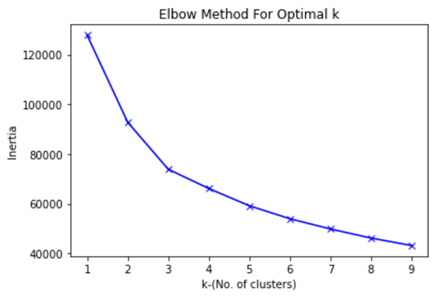
**(For finding optimal number of clusters)**

* Here we plot a graph between Inertia and Number of clusters.
* Inertia: It is a quantity also called Within-Cluster-Sum-of-Squared-Errors(SSE) mathematically represented by:



* 
* 

**Final plot:**



**OPTIMAL NUMBER OF CLUSTERS:**

In the graph we can see that after K=3, the value of inertia almost remains the same.

Hencein our project, we have used K=3 as the optimal number of

clusters.

**#K(no. of cluster) is very low:** WRONG APPROACH as observations within the same cluster can be vastly different.

**#K(no. of cluster) is very high:** WRONG APPROACH as comprehending each cluster would be difficult due to loss in simplicity.

**K-means algorithm**

– **Outline of algorithm** :

Assuming the total no. of the required cluster to be K. We should follow the following procedure on our dataset to cluster it.

1. Pick K points as the initial centroids from the dataset, either randomly or the first K.

2. Find the Euclidean distance of each point in the dataset with the identified K points (cluster centroids).

3. Assign each data point to the closest centroid using the distance found in the previous step. 4. Find the new centroid by taking the average of the points in each cluster group.

5. Repeat 2 to 4 for a fixed number of iterations or till the centroids do not change.

– **How to find K?**

In order to implement the K-means clustering in our project, first we choose the optimal number of clusters (K) for which one of the most popular methods is “Elbow method”.

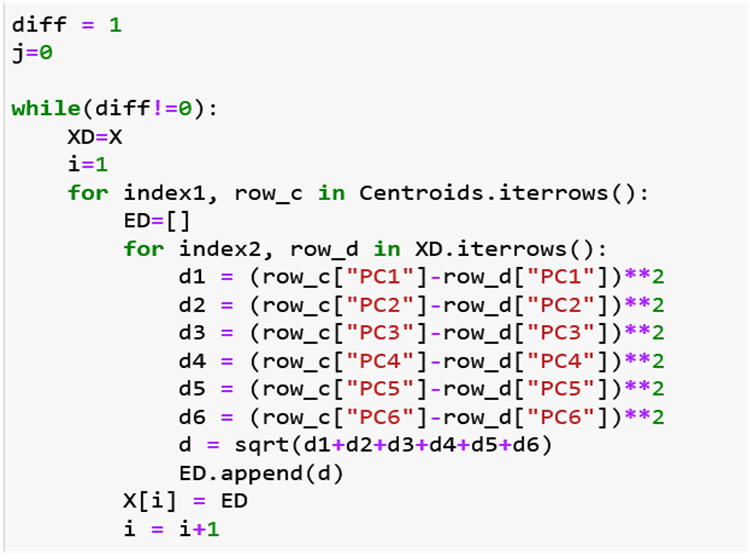
In the Elbow method, it uses the within cluster sums of squares by looking at the total within-cluster sum of squares as a function of the number of clusters. The location of a knee or elbow in the plot is usually considered as an indicator of the appropriate number of clusters.

Some other methods which could be use for finding K are the average silhouette, and the Gap statistic methods.

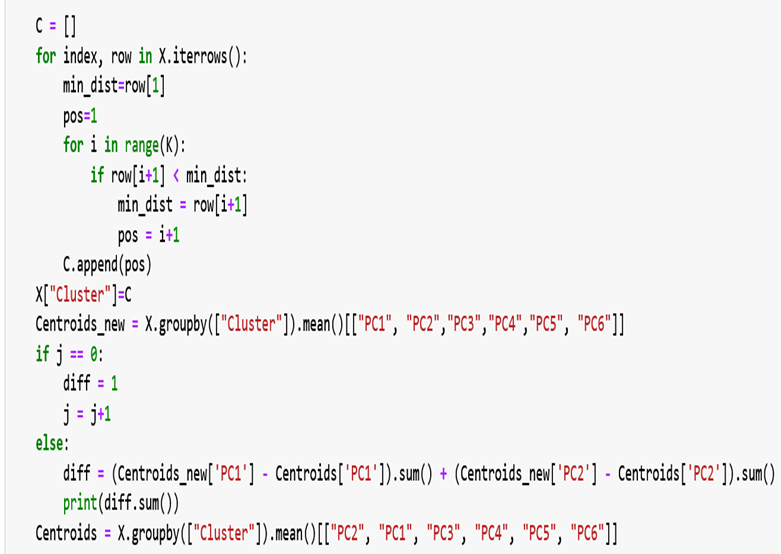
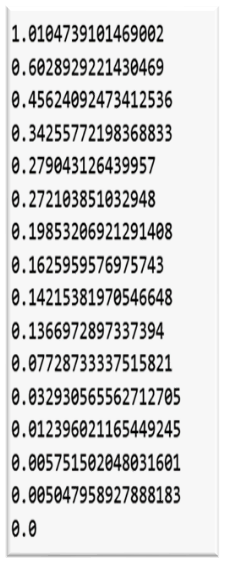
– **Add code and images** :

**Code Snippet 1-**

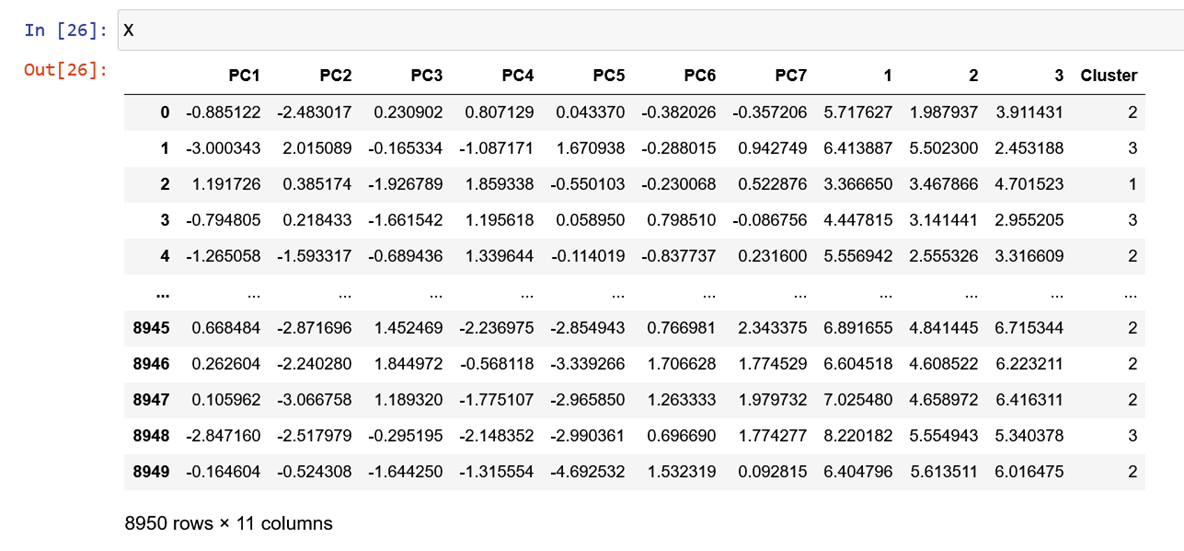
* Finding Euclidean distance of points from the centroid.

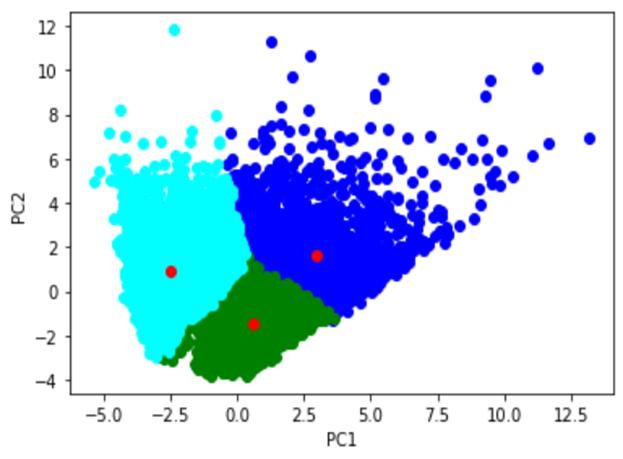


**Code Snippet 2-**

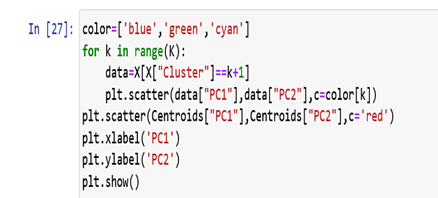
* Assigning each point to a centroid and forming three primary clusters.
* Taking the average of all points in a cluster to generate new centroids.
* Iterating the above procedure until nre centroids generated are identical to the previous ones.
* Here we are done with clustering.

**Final Data after clustering:**





PC1 Vs PC2 plot:-



Code of the above plot:-

**CREDIT RISK MODELING**

* **Environmental setup:**

Starting off with the credit risk modeling, we import the following libraries and modules which we will be needing :

Libraries: Pandas, numpy, seaborn and pyplot.

Modules:

train\_test\_split, RepeatedStratifiedKFold, cross\_val\_score, LogisticRegression, roc\_curve, roc\_auc\_score,confusion\_matrix, precision\_recall\_curve, auc, f\_classif, Pipeline, BaseEstimator, TransformerMixin, chi2\_contingency

**Train-test split**

In machine learning, it is a common practice to split your data into two different sets. These two sets are the training set and the testing set. As the name suggests, the training set is used for training the model and the testing set is used for testing the accuracy of the model.

* *Why do we need to split data into training and testing sets?*

While training a machine learning model we are trying to find a pattern that best represents all the data points with minimum error. While doing so, two common errors come up. These are overfitting and underfitting.

* *Underfitting*:

Underfitting is when the model is not even able to represent the data points in the training dataset. In the case of under-fitting, you will get a low accuracy even when testing on the training dataset.

Underfitting usually means that your model is too simple to capture the complexities of the dataset.

* *Overfitting*:

Overfitting is the case when your model represents the training dataset a little too accurately. This means that your model fits too closely. In the case of overfitting, your model will not be able to perform well on new unseen data. Overfitting is usually a sign of a model being too complex.

Both overfitting and under-fitting are undesirable.

* *How to split training and testing data sets in Python?*

The most common split ratio is 80:20.

That is 80% of the dataset goes into the training set and 20% of the dataset goes into the testing set.

Before splitting the data, make sure that the dataset is large enough. Train/Test split works well with large datasets.

* **To split using x-train**:

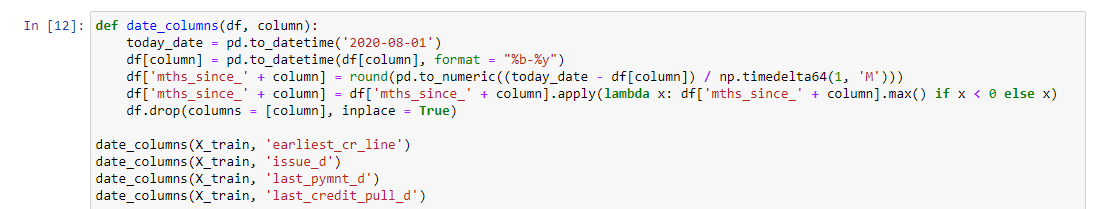
1. Import the entire dataset

2. Split the data using sklearn

To split the data we will be using train\_test\_split from sklearn.

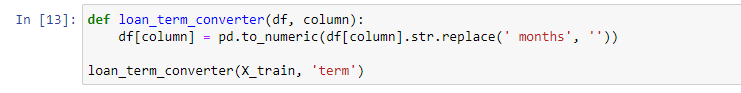
train\_test\_split randomly distributes your data into training and testing set according to the ratio provided.

In the dataframe, we have many columns containing non numerical data. For example, dates, months, etc. ‘Months’ column contains the string type data. We cannot use this data in further algorithms. Hence, we got to convert these data into numerical form.



The function given above converts the datetime format of data into numerical data, specifically the difference between today and the given date.

Similarly, the code below converts the string type data in the column of ‘month’ into numeric type data.



After separating the categorical and numerical type data, we are set to apply:

Chi-squared statistic for the categorical type data

Anova statistic for numerical type data

* **Chi squared test of independence:**

A chi-squared test (symbolically represented as χ2) is basically a data analysis on the basis of observations of a random set of variables. Usually, it is a comparison of two statistical data sets. This test was introduced by Karl Pearson in 1900 for categorical data analysis and distribution. So it was mentioned as Pearson’s chi-squared test.

The chi-square test is used to estimate how likely the observations that are made would be, by considering the assumption of the null hypothesis as true.

A hypothesis is a consideration that a given condition or statement might be true, which we can test afterwards. Chi-squared tests are usually created from a sum of squared falsities or errors over the sample variance.

Null hypothesis: this hypothesis assumes that there is no relation between the two variables.

Alternate hypothesis: this hypothesis assumes that there is a relation between the two variables.

*Finding P-Value:*

P stands for probability here. To calculate the p-value, the chi-square test is used in statistics. The different values of p indicates the different hypothesis interpretation, are given below:

P≤ 0.05; Hypothesis rejected

P>.05; Hypothesis Accepted

*Properties;*

The following are the important properties of the chi-square test:

-Two times the number of degrees of freedom is equal to the variance.

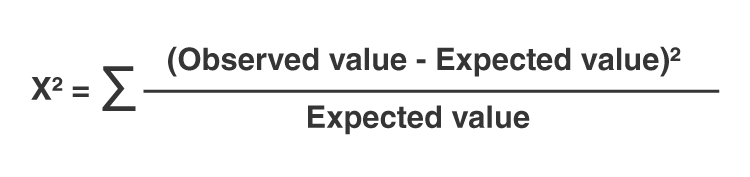
-The number of degree of freedom is equal to the mean distribution

-The chi-square distribution curve approaches the normal distribution when the degree of

freedom increases.

*Formula*

The chi-squared test is done to check if there is any difference between the observed value and expected value. The formula for chi-square can be written as;



or

χ2 = ∑(Oi – Ei)2/Ei

where Oi is the observed value and Ei is the expected value.

* *When to use a chi-square test?*

A Pearson’s chi-square test may be an appropriate option for your data if all of the following are true:

You want to test a hypothesis about one or more categorical variables. If one or more of your variables is quantitative, you should use a different statistical test. Alternatively, you could convert the quantitative variable into a categorical variable by separating the observations into intervals.

* *How to perform a chi-square test?*

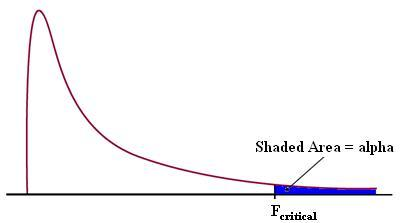
The exact procedure for performing a Pearson’s chi-square test depends on which test you’re using, but it generally follows these steps:

Create a table of the observed and expected frequencies. This can sometimes be the most difficult step because you will need to carefully consider which expected values are most appropriate for your null hypothesis.

Calculate the chi-square value from your observed and expected frequencies using the chi-square formula.

Find the critical chi-square value in a chi-square critical value table or using statistical software.

Compare the chi-square value to the critical value to determine which is larger.

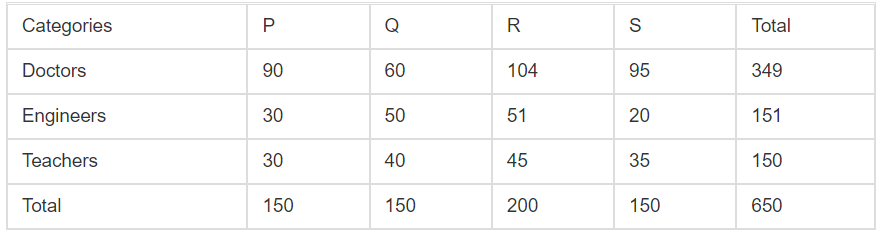


* if the F-value lies within the rejected area, the null hypothesis is rejected.

That is, You should reject the null hypothesis if the chi-square value is greater than the critical value. If you reject the null hypothesis, you can conclude that your data are significantly different from what you expected.

* **Example:-**

Let us take an example of categorical data where there is a society of 1000 residents with four neighborhoods, P, Q, R and S. A random sample of 650 residents of the society is taken whose occupations are doctors, engineers and teachers. The null hypothesis is that each person’s neighborhood of residency is independent of the person’s professional division. The data are categorized as:

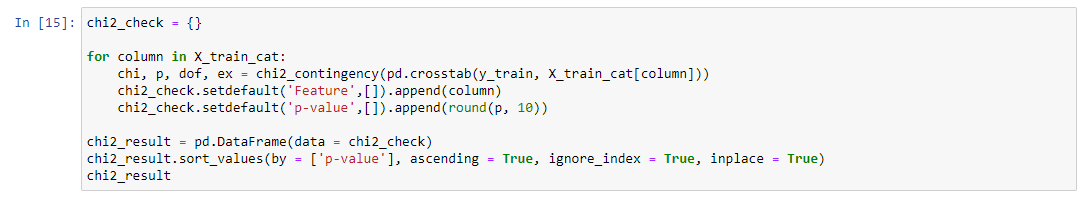


Assume the sample living in neighborhood P, 150, to estimate what proportion of the whole 1,000 people live in neighborhood P. In the same way, we take 349/650 to calculate what ratio of the 1,000 are doctors. By the supposition of independence under the hypothesis, we should “expect” the number of doctors in neighborhood P is;

150 x 349/650 ≈ 80.54

So by the chi-square test formula for that particular cell in the table, we get;

(Observed – Expected)2/Expected Value = (90-80.54)2/80.54 ≈ 1.11

* *Code snippet for chi-squared*

**ANOVA**

**(ANALYSIS OF VARIANCE)**

* **What Is Analysis of Variance (ANOVA)?**

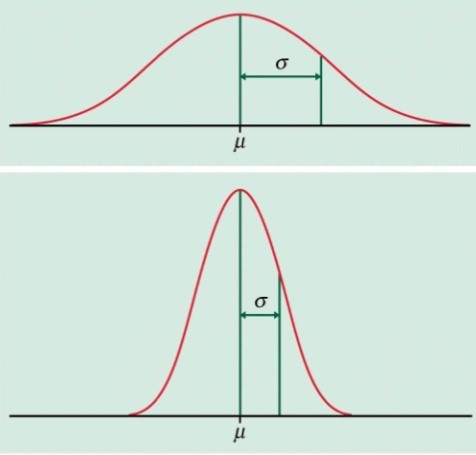
Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study. Here is how we calculate the F-ratio:

**F = Z/A**

Where, Z = variance betweenthe set of means (m1,m2,m3,m4)

A = variance of all the Readings (a1,a2….d4,d5)

* How does the variation look within group:



The plot above is the nature of the graph which has more variance (relative to the lower one). It is more scattered around its mean (u). This sample will lead to a bigger ‘A’.

The plot below is the nature of the graph having lesser variance (relative to the upper one). It is more clustered around its mean (u). This sample will lead to smaller ‘A’.

The ANOVA test allows a comparison of more than two groups at the same time to determine whether a relationship exists between them. The result of the ANOVA formula, the F statistic (also called the F-ratio), allows for the analysis of multiple groups of data to determine the variability between samples and within samples.

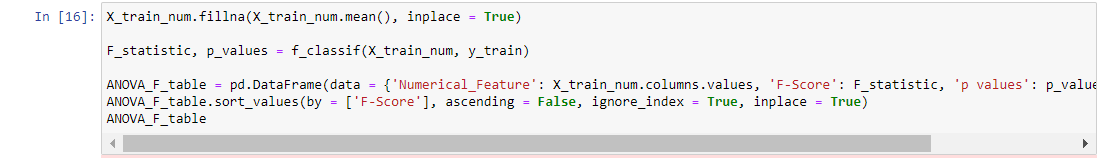
If no real difference exists between the tested groups, which is called the null hypothesis, the result of the ANOVA's F-ratio statistic will be close to 1. The distribution of all possible values of the F statistic is the F-distribution. This is actually a group of distribution functions, with two characteristic numbers, called the numerator degrees of freedom and the denominator degrees of freedom.

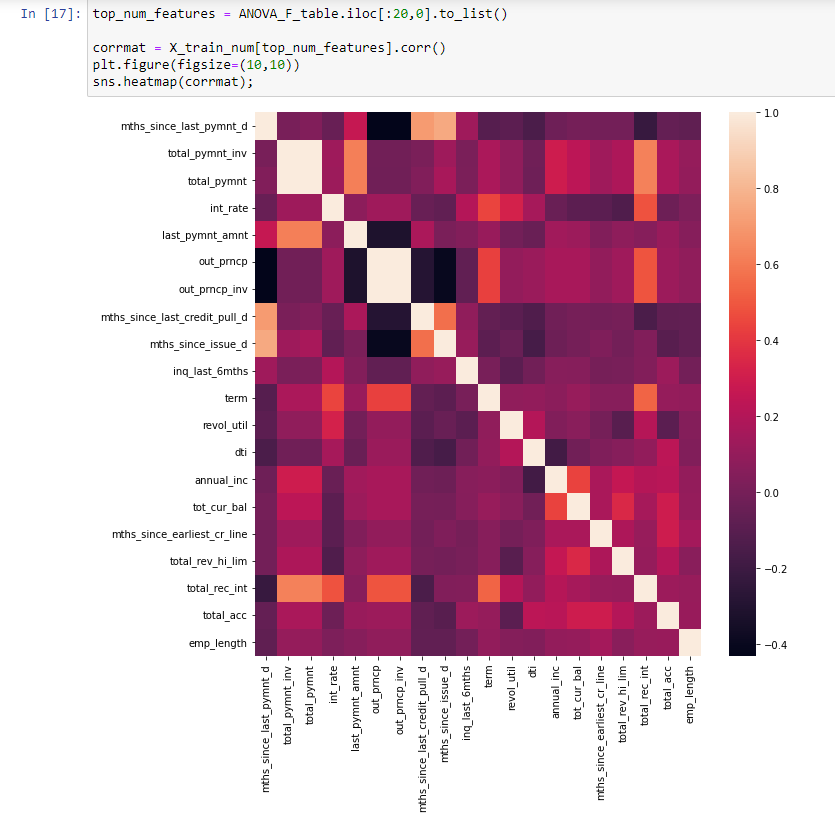
* *One-Way ANOVA Versus Two-Way ANOVA:*

There are two main types of ANOVA: one-way (or unidirectional) and two-way. One-way or two-way refers to the number of independent variables in your analysis of variance tests. A one-way ANOVA evaluates the impact of a sole factor on a sole response variable. It determines whether all the samples are the same. The one-way ANOVA is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups.

A two-way ANOVA is an extension of the one-way ANOVA. With a one-way, you have one independent variable affecting a dependent variable. With a two-way ANOVA, there are two independents. For example, a two-way ANOVA allows a company to compare worker productivity based on two independent variables, such as salary and skill set. It is utilized to observe the interaction between the two factors and tests the effect of two factors at the same time.

* *Code snippet for ANOVA applied to each pair of column:*



* *Heatmap of thus obtained relations between the columns:*

WoE (Weight of Evidence) & IV (Information Value)

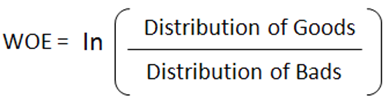
* Understanding Binning:

Data binning is a data pre-processing method used to minimize the effects of small observation errors. The original data values are divided into small intervals known as bins and then they are replaced by a general value calculated for that bin.

* Introduction to Weight of Evidence:

**Definition:** The purpose of the Weight of Evidence (WoE) is to serve as a flexible tool to recode the values in continuous and categorical predictor variables into discrete categories automatically, and to assign a unique Weight-of-Evidence value to each category.

* + The Weight of Evidence or WoE value is a widely used measure of the strength of a grouping for separating good and bad risk. Basically, it helps to transform a continuous independent variable into a set of groups or bins based on similarity of dependent variable distribution i.e. number of events and non-events.
  + **Formula**:



1. **Distribution of Good -** % of Good Customers in a particular group
2. **Distribution of Bad -** % of Bad Customers in a particular group
3. **ln -** Natural Log

* **Observation:**

Positive WOE means Distribution of Goods > Distribution of Bad.

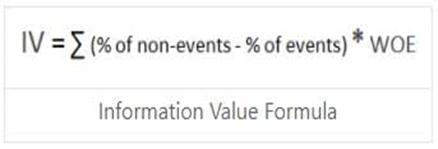
Negative WOE means Distribution of Goods < Distribution of Bad.

The value of WoE will be 0 if the value of (Distribution Goods / Distribution Bad) is equal to 1.

Introduction to Information Value:

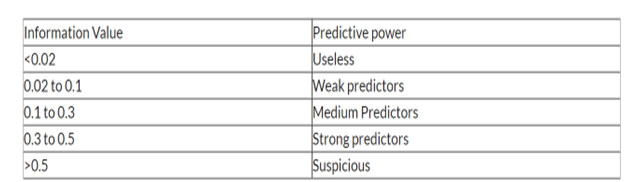
**Definition**: Information value is one of the most useful techniques to select important variables in a predictive model. After calculation of Information value (IV) we can see if it useful for further calculation or not this can be refer from given formula and the given range:

* **Formula**:



Here we get WOE from the formula stated previously.

* **Optimal range for IV**:

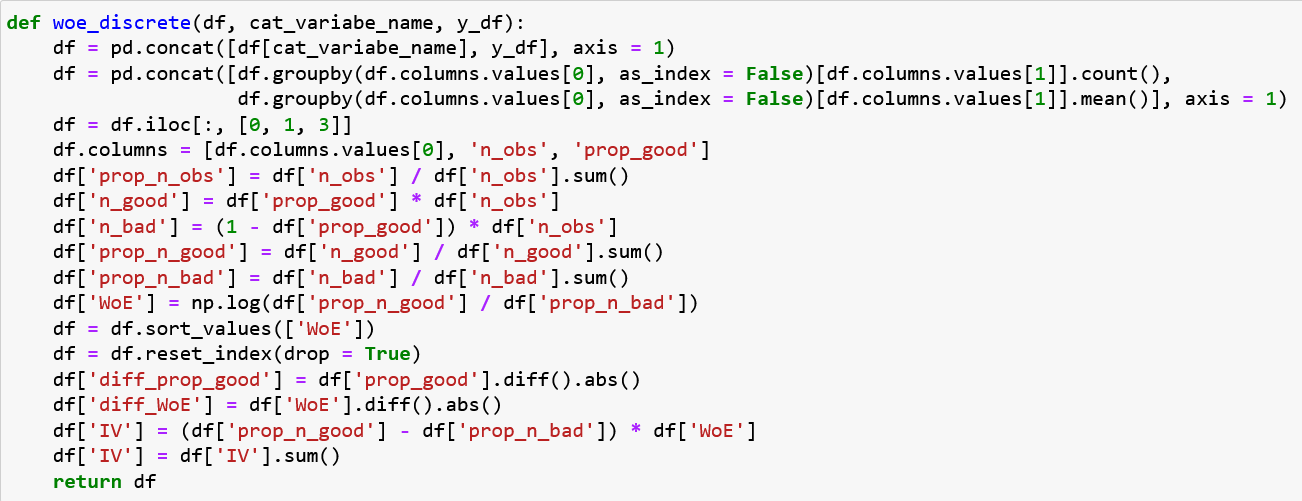
The table below gives you a fixed rule to help select the best features for our model:

Thus we handled our dataset and binned it using WOE and IV for further processing of our credit risk management task.

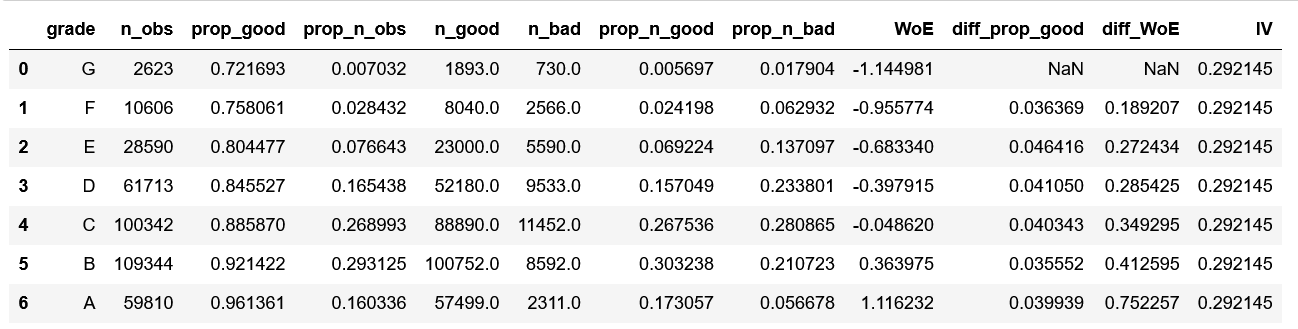
* How to implement WoE and IV?

We will analyze both categorical and continuous variables one by one from our dataset.

• **For categorical independent variables (Discrete) :** Combine categories with similar WOE and then create new categories of an independent variable with continuous WOE values.

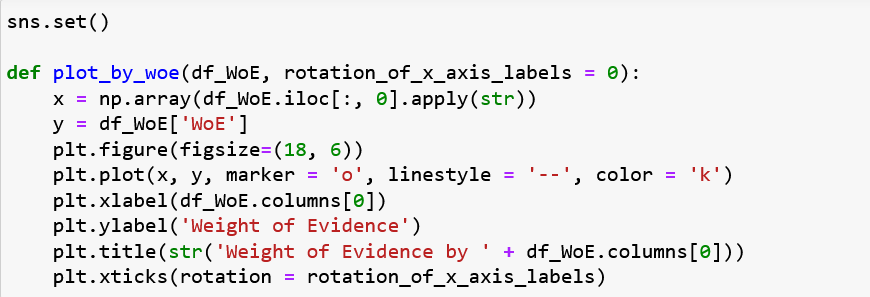
**Code for which is Given below:**

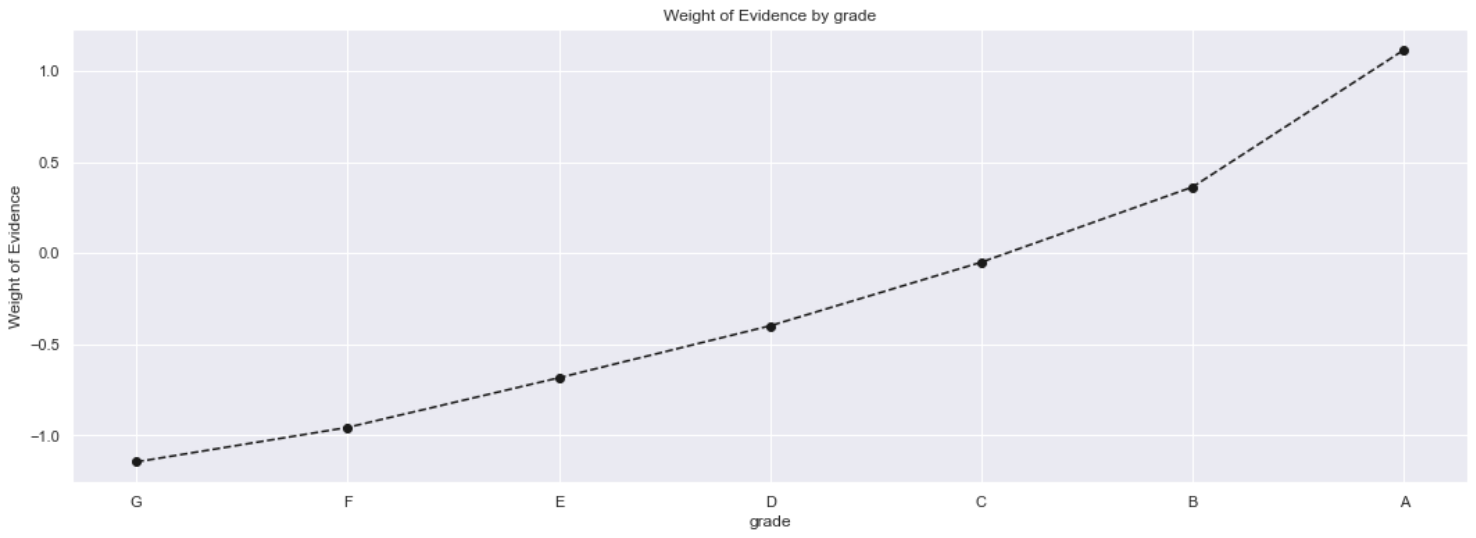
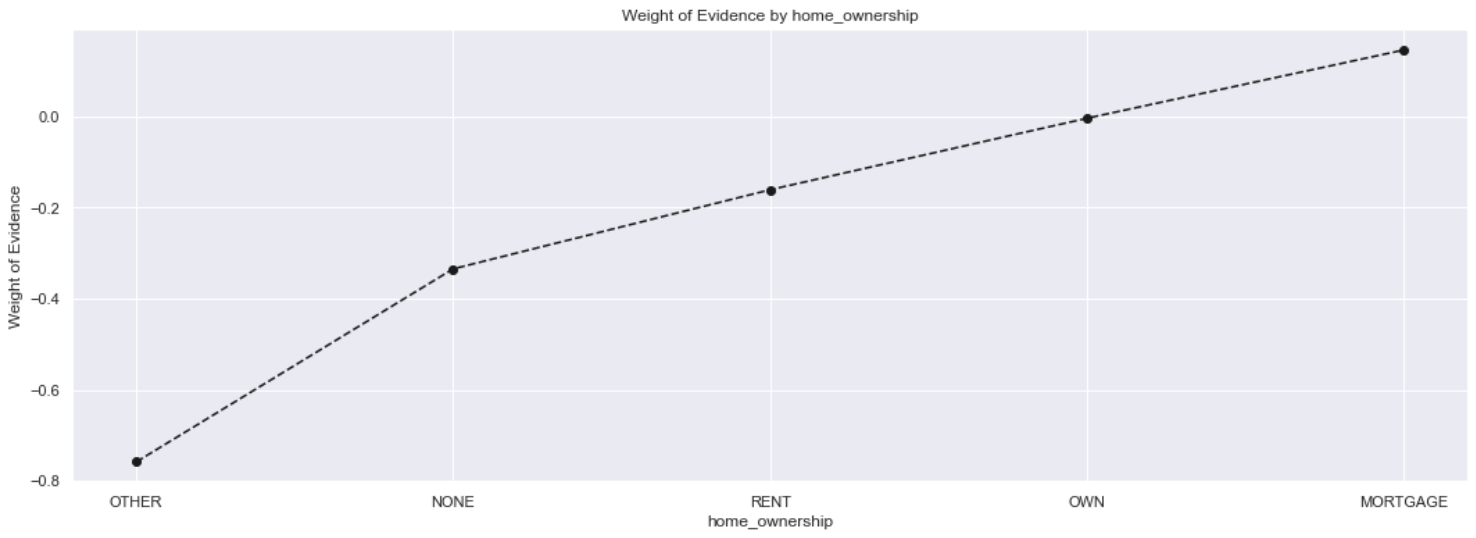
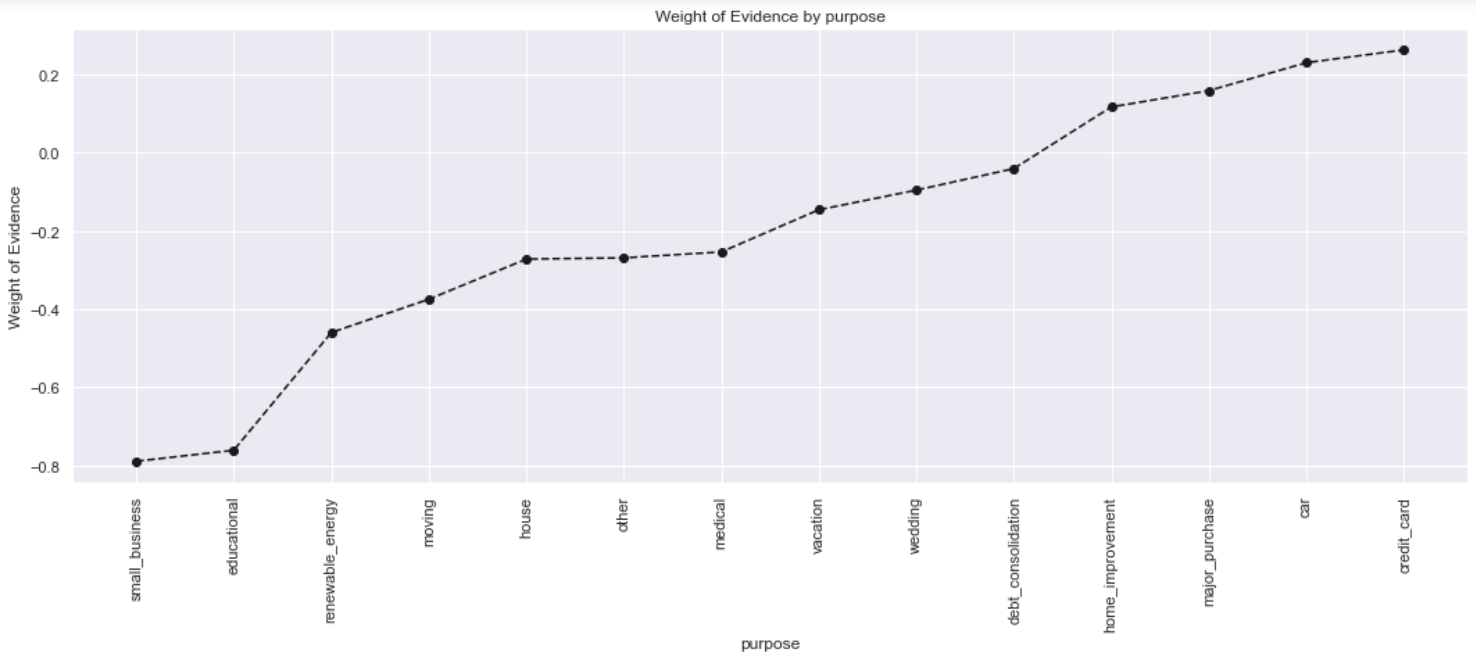
And the resultant Data Frame will look like this:



The same process will be repeated for all the four categorical variables that we have identified and then plot corresponding curves.

* **Code for plotting the curve:**

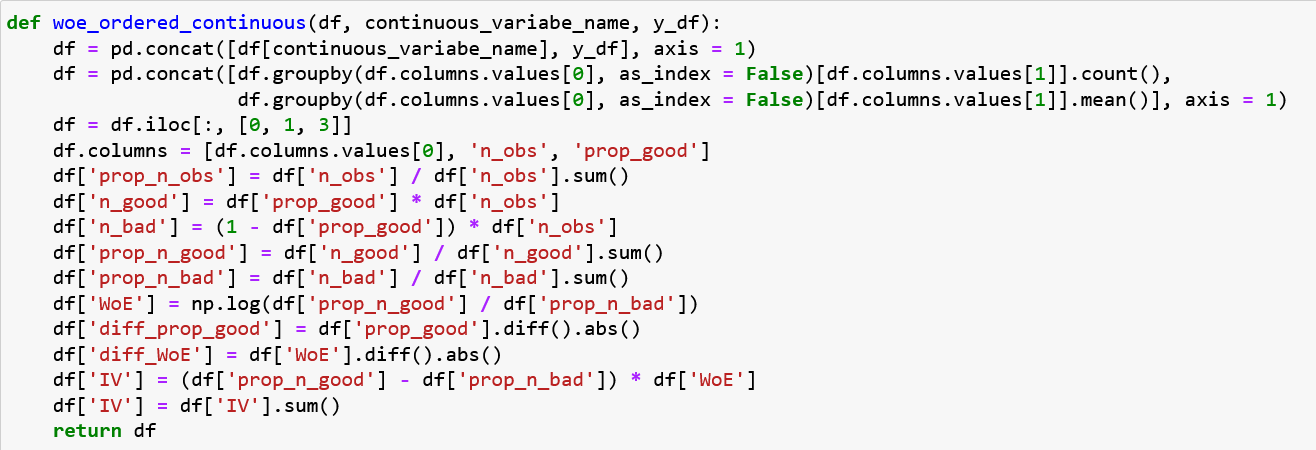


* **The plot we get for corresponding categorical variable looks as follow:**
* ****
* 
* 
* 

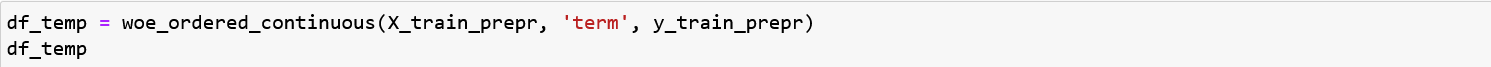
**Note:** Now we observe graphs of WOE. If there is a continuous increase in WoE across the different categories then we do not need to combine any features together and should leave all these categories as they are.

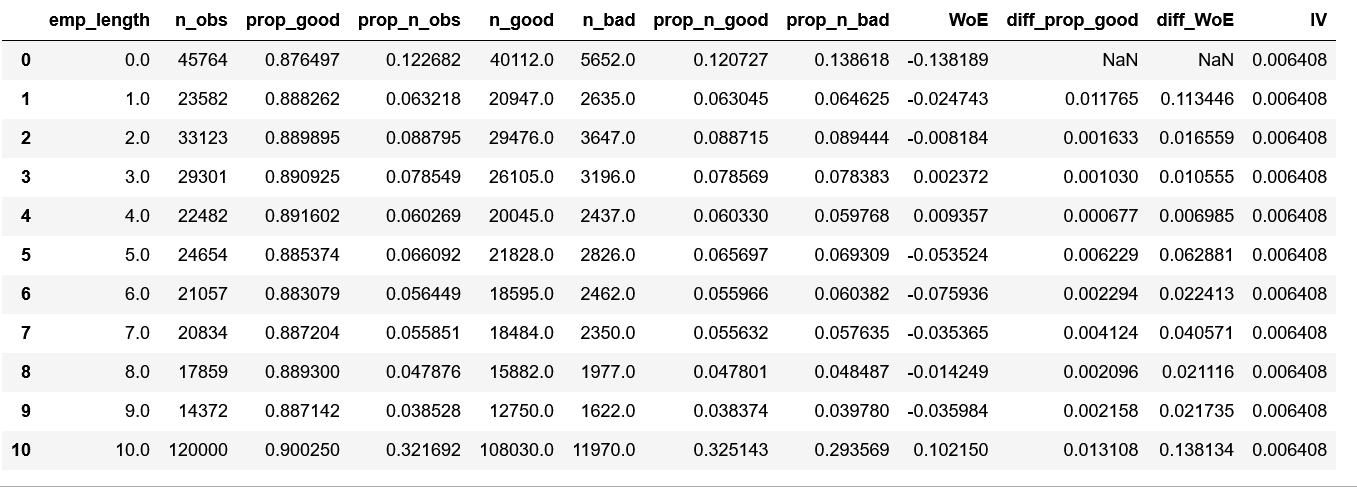
* **For continuous independent variables:** First, create bins (categories / groups) for a continuous independent variable and then combine categories with similar WOE values and replace categories with WOE values. Similar code as for discrete variables with minor changes is applied.

**Code for which is Given below:**

****

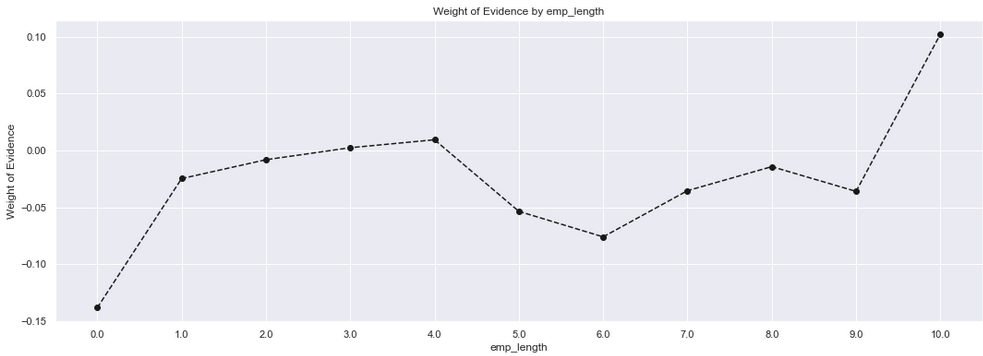
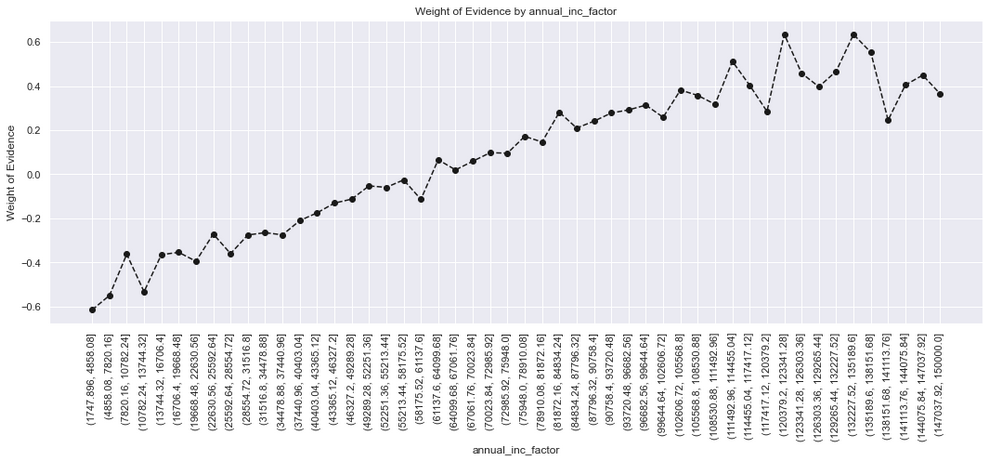
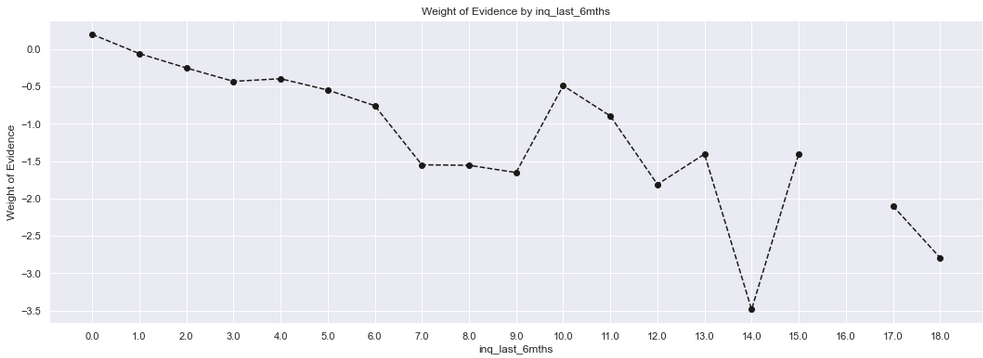
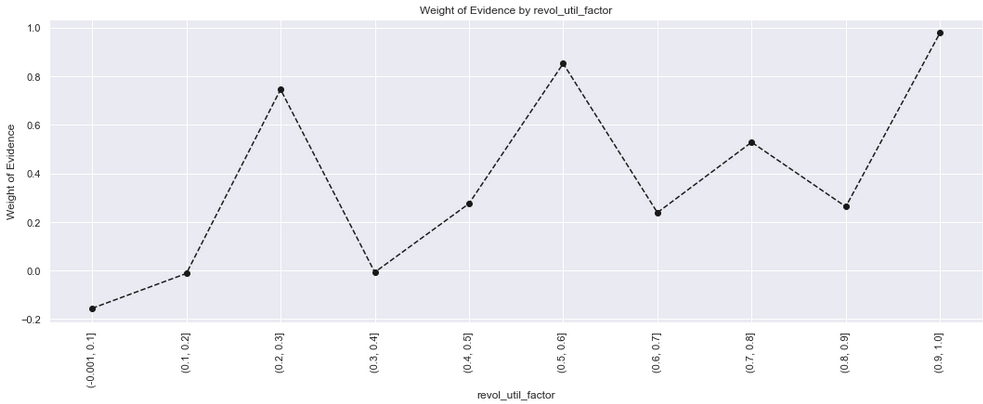
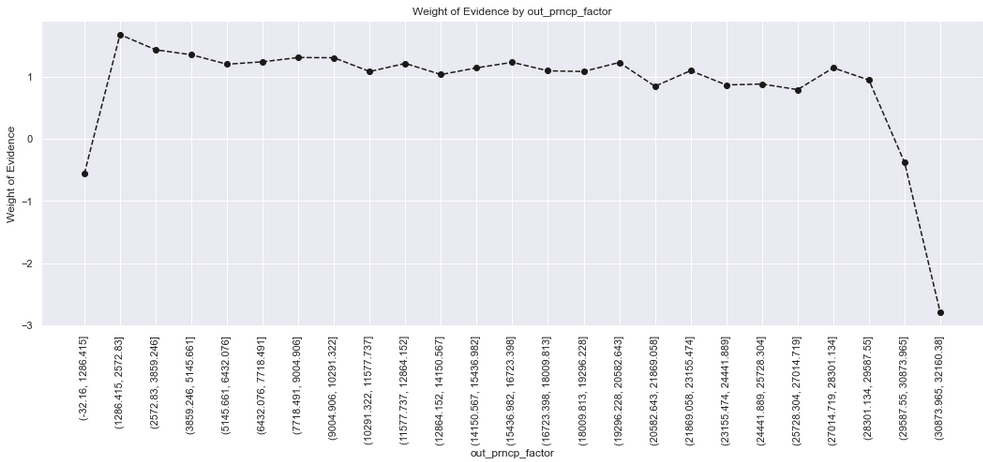
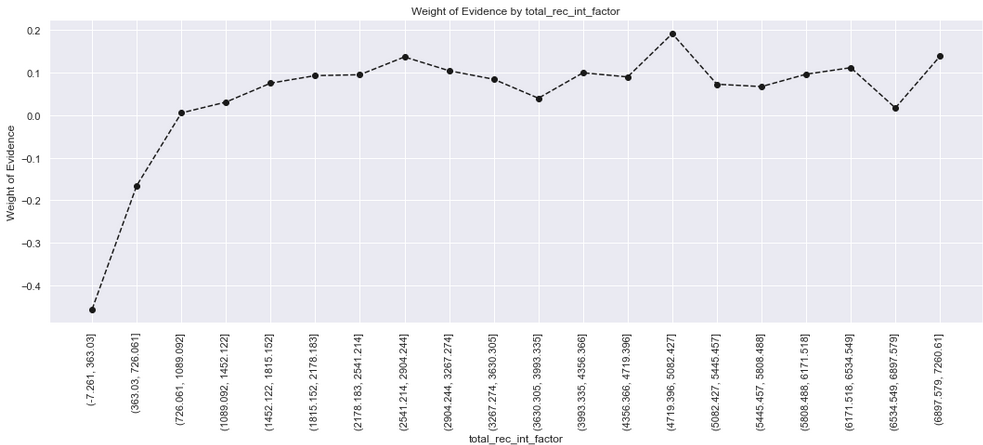
And the resultant Data Frame will look like this:





The same process will be repeated for all the continuous variables that we have identified and then plot corresponding curves.

**Some of the plots looks as follow:**

* 
* 
* 
* 
* 
* 
* 
* Analysis from obtained graph and values:

Now for further processing of data we will not consider following features:

1. Feature with very low IV: As it will become very less significant hence, useless.

2. Features which have very high IV: As it would be suspicious and too correct to be a real scenario.

3. WoE ranges between a very small range, implying low power of differentiating between good and bad loans

Note: When the IV is borderline close to the minimum or maximum ideal threshold, we proceed without ignoring that feature.

* Some rules of using WoE:

1. Each bin should have at least 5% of the observations otherwise it would be insignificant.

2. Each bin should be non-zero for both non-events and events.

Similar groups should be aggregated.

3. The WOE should be monotonic i.e. either growing or decreasing with the groupings.

4. Missing values are binned separately.

* Benefits of using WoE and IV:

WOE can handle missing values as missing values can be binned separately. WOE transformation helps us to build strict linear relationships with log odds. Otherwise, it is not easy to accomplish a linear relationship using other transformation methods such as log, square-root etc.

The overall calculations are nearly accurate processing through this method and thus we handled our dataset and binned it using WOE and IV for further processing of our credit risk management task.

* Custom Classes**:**

The custom class creates new categorical dummy features based on the cut-off points that we manually identified based on the WOE plots and IV earlier.

Now, we create a custom scikit-learn class to take care of all binning transformations on any given data set. This custom class will help us in performing k fold cross validation.

This custom class will create new categorical dummy features based on the cut-off points that we manually identified based on the WoE plots and IV above

Now we will use it as a part of a scikit-learn Pipeline.

**Making A PD Model**

The Probability of Default (PD) is the likelihood that a borrower willfail to pay back a debt. It enables us to gauge credit risk in quantitative terms.

**Steps involved in a PD-Model**

STEP-1: Define a Modeling Pipeline-A machine learning pipeline is used to help automate machine learning workflows. It involves a series of transformations on the given data that work as a single atomic unit.



Pipeline-Steps in a PD-model:

1. WOE-Binning.

2. Logistic Regression: - Logistic regression estimates the probability of an event occurring, such as default/no default, based on a given dataset of independent variables.

Step-2: Define Cross-Validation Criteria:

1. The goal of cross-validation is to test the model's ability to predict new data that was not used in estimating it.

2. Here use “Repeated-Stratified-K-fold”. This involves simply repeating the cross-validation procedure multiple times and reporting the mean result across all folds from all runs.

K-Fold Cross Validation:

1. Shuffle the dataset randomly.

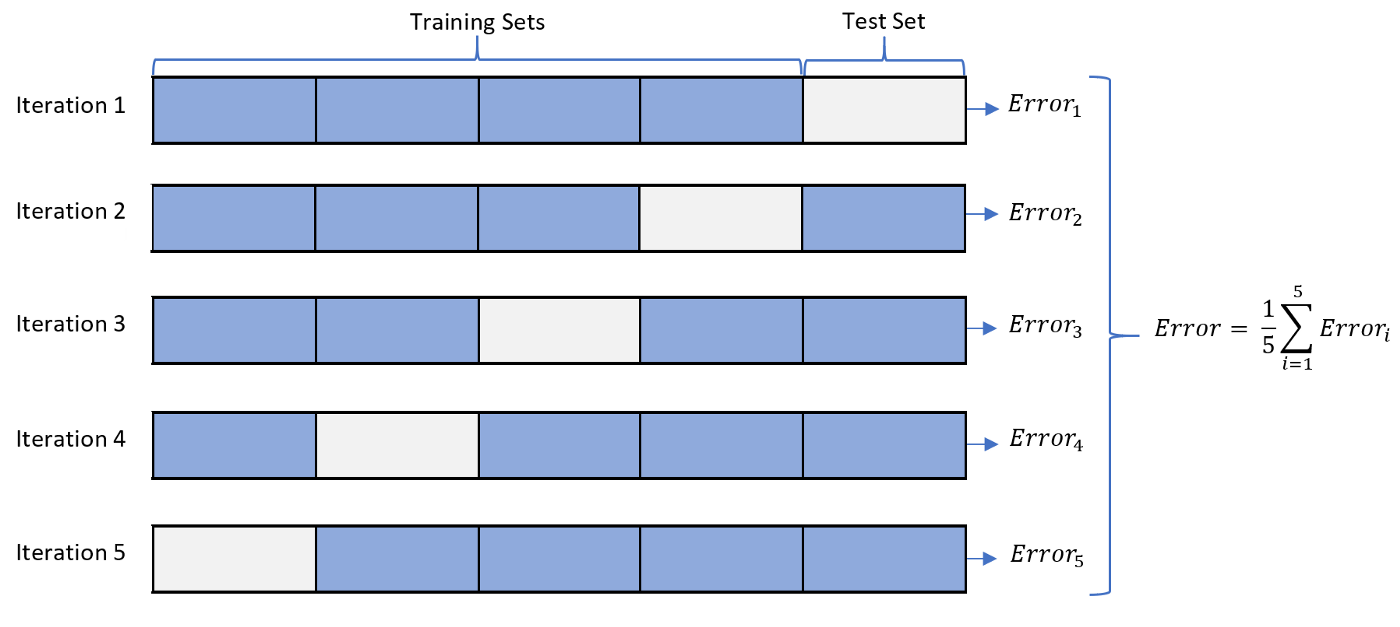
2. Split the dataset into k groups

3. For each unique group:

i. Take the group as a hold out or test data set

ii. Take the remaining groups as a training data set

iii. Fit a model on the training set and evaluate it on the test set.



Step-3:Get AUROC and GINI scores:

o AUROC stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC curve.

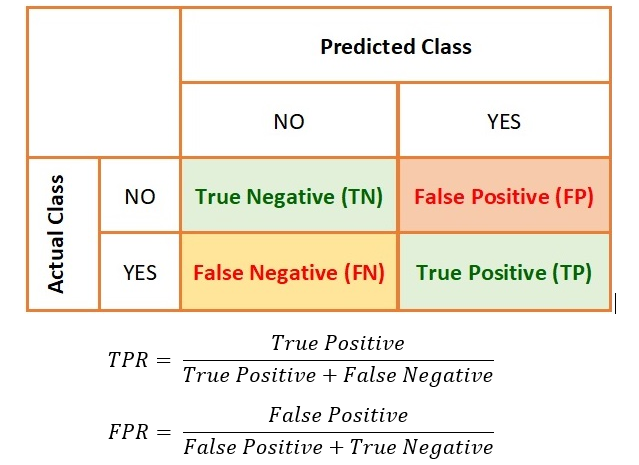
o ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes.

o The Higher the AUC, the better the model is at distinguishing between potential defaulters/non-defaulters.

ROC-Curve:

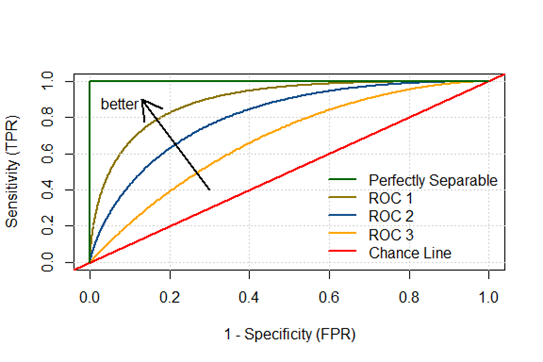
Receiver Operator Characteristic (ROC) curve is a graphical plot used to show the diagnostic ability of binary classifiers. It was first used in signal detection theory but is now used in many other areas such as medicine, radiology, natural hazards, and machine learning.

Creating a ROC curve:

* A ROC curve is constructed by plotting the true positive rate (TPR) against the false positive rate (FPR).
* The true positive rate is the proportion of observations that were correctly predicted to be positive out of all positive observations (TP/(TP + FN)).
* Similarly, the false positive rate is the proportion of observations that are incorrectly predicted to be positive out of all negative observations (FP/(TN + FP)).
* 

**Interpreting the ROC curve:**

* The ROC curve shows the trade-off between sensitivity (or TPR) and specificity (1 – FPR).
* As a baseline, a random classifier is expected to give points lying along the diagonal (FPR = TPR).
* Classifiers that give curves closer to the top-left corner indicate a better performance.
* The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.



How to speculate about the performance of the model?

o An excellent model has AUC near to the 1 which means it has a good measure of separability. A poor model has an AUC near 0.

**Confusion matrix:**

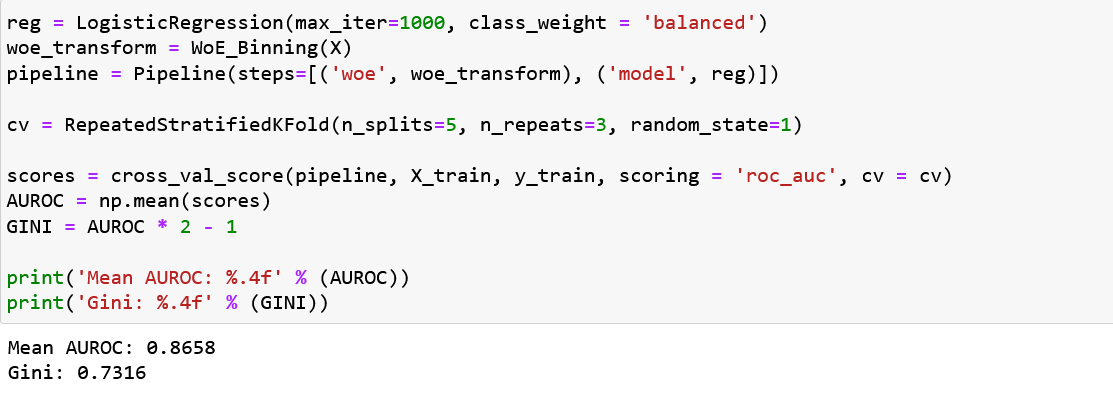
* We predict 0 while the true class is actually 0: this is called a **True Negative**, i.e., we correctly predict that the class is negative (0). For example, an antivirus did not detect a harmless file as a virus.
* We predict 0 while the true class is actually 1: this is called a **False Negative**, i.e., we incorrectly predict that the class is negative (0). For example, an antivirus failed to detect a virus.
* We predict 1 while the true class is actually 0: this is called a **False Positive**, i.e., we incorrectly predict that the class is positive (1). For example, an antivirus considers a harmless file to be a virus.
* We predict 1 while the true class is actually 1: this is called a **True Positive**, i.e., we correctly predict that the class is positive (1). For example, an antivirus rightfully detected a virus.

o  [True positive rate](https://en.wikipedia.org/wiki/Sensitivity_and_specificity#Sensitivity) (**TPR**), aka. sensitivity, [hit rate](https://en.wikipedia.org/wiki/Hit_rate), and [recall](https://en.wikipedia.org/wiki/Information_retrieval#Recall), which is defined as *TP/TP*+*FN*.

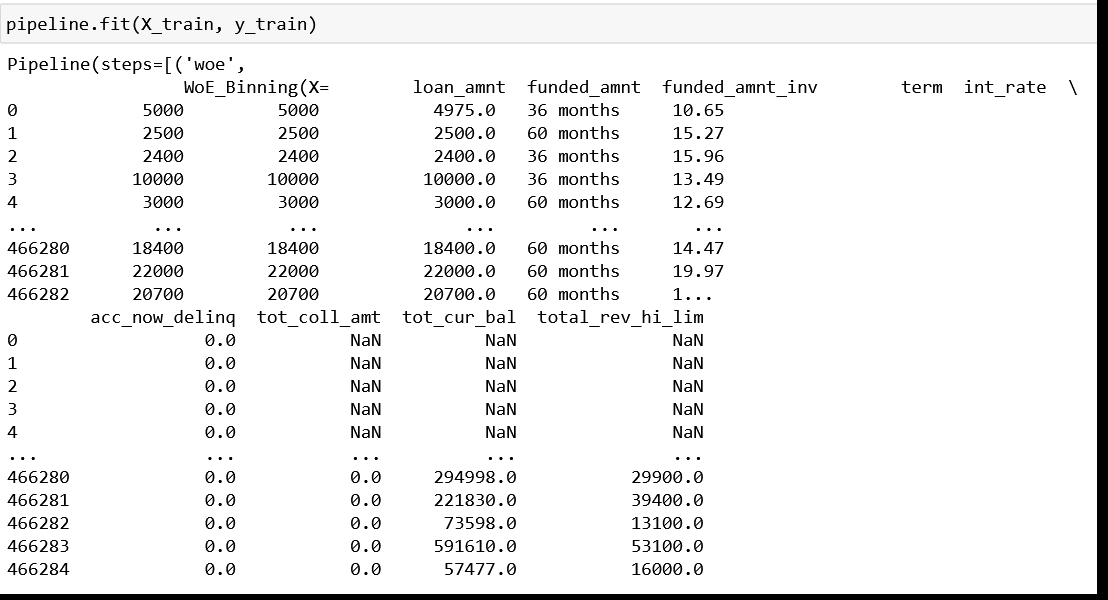
o  [False positive rate](https://en.wikipedia.org/wiki/False_positive_rate) (**FPR**), aka. [fall-out](https://en.wikipedia.org/wiki/Information_retrieval#Fall-out), which is defined as *FP/FP*+*TN*

GINI = AUROC \* 2 – 1

Code Snippet For Steps 1-3:



Next, we fit the pipeline on the training set.



Creating a Summary Table:

o create a transformed training set through our WoE\_Binning custom class.

o store the column names in X\_train as a list

o create a summary table of our logistic regression model (name it summary table).

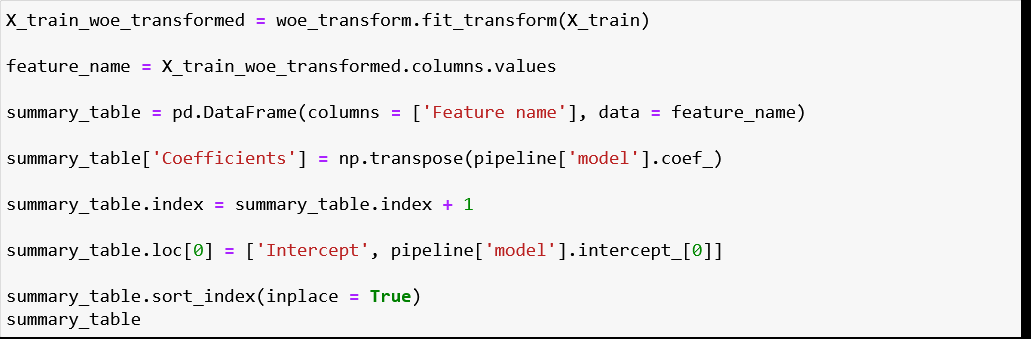
o create a new column in the data frame, called 'Coefficients', with row values the transposed coefficients from the model.

o increase the index of every row of the data frame with 1 to store our model intercept in the 1st row.

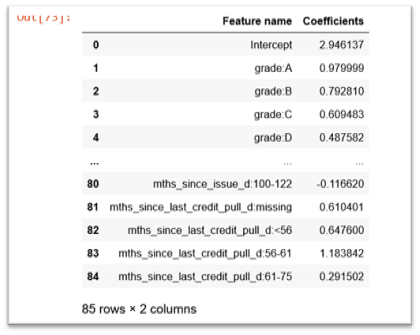
o assign our model intercept to this new row.

o sort the data frame by index.

Input-Cell:



Output Cell:



Prediction Time:

Here we intend to make predictions on our test-set. For that purpose, we use the “pipeline. predict” function from scikit-learn.

1. Make predictions on the test set(X-test) and call it y\_hat\_test:



2. Find out the predicted probabilities and name it y-hat\_test\_proba:

3.Select the probabilities of the positive class (By default-Class 1)

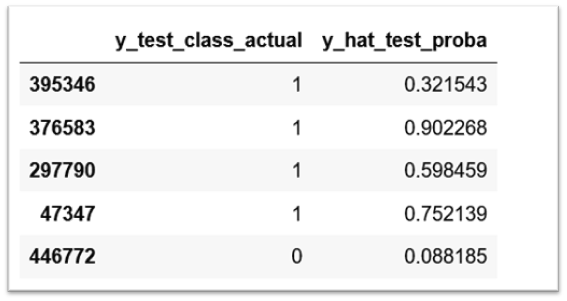
Next, we create a new Data-Frame which contains actual classes and predicted probabilities:

1. Create a new Data-Frame called “y\_test\_temp” which is same as “y\_test”:

2. Reset the index of y\_test\_temp to allow proper concatenation with y\_hat\_test\_proba:

3. Check the shape of y\_test\_temp and make sure that no. of rows are same as that of y\_test:

4. Rename the columns of y\_test\_proba to 'y\_test\_class\_actual' and 'y\_hat\_test\_proba’. Make the index of y\_test\_proba same as that of X\_test:

Finally, this is how y\_test\_proba looks like: 

THRESHOLD VALUE :-

* THE VALUE IS USED TO DIFFERENTIATE GOOD WITH BAD.​
* MANY MACHINE LEARNING ALGORITHMS ARE CAPABLE OF PREDICTING A PROBABILITY OR SCORING OF CLASS MEMBERSHIP, AND THIS MUST BE INTERPRETED BEFORE IT CAN BE MAPPED TO A CRISP CLASS LABEL. THIS IS ACHIEVED BY USING A THRESHOLD, SUCH AS 0.5, WHERE ALL VALUES EQUAL OR GREATER THAN THE THRESHOLD ARE MAPPED TO ONE CLASS AND ALL OTHER VALUES ARE MAPPED TO ANOTHER CLASS.​

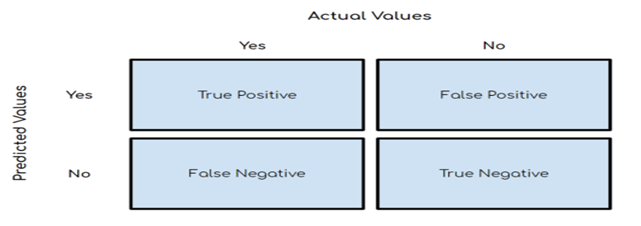
IN SOME CASES, SUCH AS WHEN USING ROC CURVES AND PRECISION-RECALL CURVES, THE BEST OR OPTIMAL THRESHOLD FOR THE CLASSIFIER CAN BE CALCULATED DIRECTLY. IN OTHER CASES, IT IS POSSIBLE TO USE A GRID SEARCH TO TUNE THE THRESHOLD AND LOCATE THE OPTIMAL VALUE.​

CONFUSION MATRIX:-

You’ll evaluate your machine learning model based on some predetermined metrics that you decide to use. When it comes to building classification models, you’ll most likely use a **confusion matrix** and related metrics to evaluate your model. Confusion matrices are not just useful in model evaluation but also model monitoring and model management!

What is a Confusion Matrix?

A confusion matrix, also known as an error matrix, is a summarized table used to assess the performance of a classification model. The number of correct and incorrect predictions are summarized with count values and broken down by each class.



Let us understand the general structure of a confusion matrix as well as the associated key terms:-

**Accuracy**

This is simply equal to the proportion of predictions that the model classified correctly.

ACCURACY = (TP+TN)/(TP+TN+FP+FN)

**Precision**

Precision is also known as **positive predictive value** and is the proportion of relevant instances among the retrieved instances. In other words, it answers the question “What proportion of positive identifications was actually correct?”**Recall**

Recall, also known as the **sensitivity, hit rate**, or the **true positive rate (TPR),** is the proportion of the total amount of relevant instances that were actually retrieved. It answers the question “What proportion of actual positives was identified correctly?”



· **Positive (P)**: Observation is positive.

· **Negative (N)**: Observation is not positive .

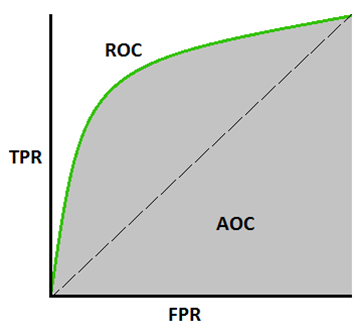
· **True Positive (TP)**: Outcome where the model correctly predicts the positive class.

· **True Negative (TN)**: Outcome where the model correctly predicts the negative class.

· **False Positive (FP)**: Also called a **type 1 error**, an outcome where the model incorrectly predicts the positive class when it is actually negative.

**False Negative (FN)**: Also called a **type 2 error**, an outcome where the model incorrectly predicts the negative class when it is actually positive

**AUC(Area Under the Receiver Operating Characteristic Curve)**

****

AUC represents the probability that a random positive (green) example is positioned to the right of a random negative (red) example. AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0.

* AUC is scale-invariant. It measures how well predictions are ranked, rather than their absolute values.
* AUC is classification-threshold-invariant. It measures the quality of the model's predictions irrespective of what classification threshold is chosen.

**Gini Coefficient**

Gini coefficient is very similar to CAP but it shows proportion (cumulative) of good customers instead of all customers. It shows the extent to which the model has better classification capabilities in comparison to the random model. It is also called the Gini Index. Gini Coefficient can take values between -1 and 1. Negative values correspond to a model with reversed meanings of scores.

If you have concordance and discordance percent, you can compute the Gini Coefficient.

Gini Coefficient = (Concordance percent - Discordance Percent)

* **Concordance percent** refers to the proportion of pairs where defaulters have a higher predicted probability than the good customers.
* **Discordance percent** refers to the proportion of pairs where defaulters have a lower predicted probability than the good customers.

Gini = 2\*(AUC - 0.5)

Gini = 2\*AUC – 1

**PR CURVE:-**

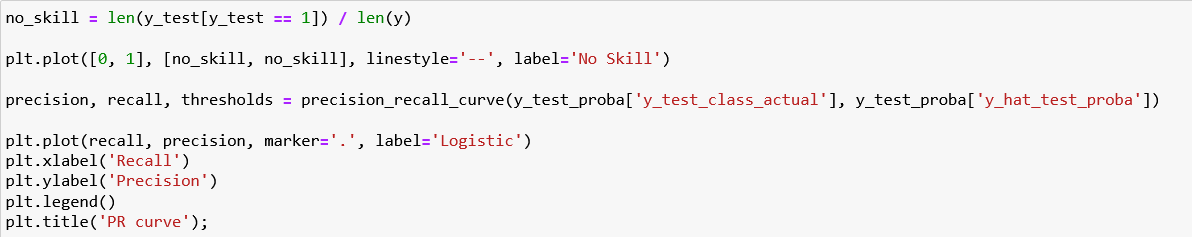
Precision-Recall is a useful measure of success of prediction when the classes are very imbalanced. In information retrieval, precision is a measure of result relevancy, while recall is a measure of how many truly relevant results are returned.

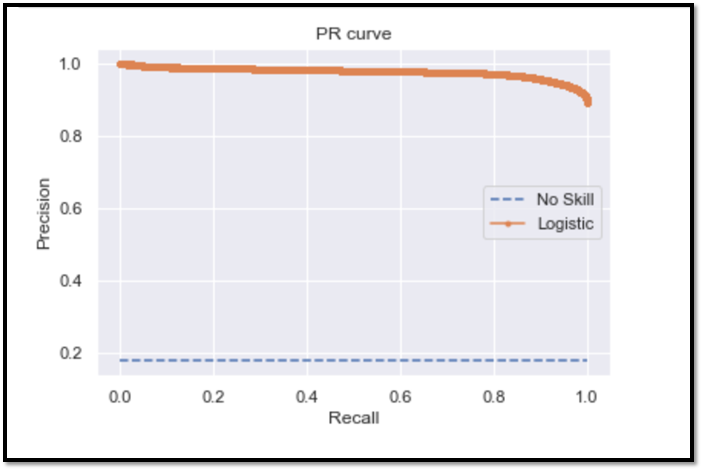
The precision-recall curve shows the tradeoff between precision and recall for different thresholds. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate. High scores for both show that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall).

A system with high recall but low precision returns many results, but most of its predicted labels are incorrect when compared to the training labels. A system with high precision but low recall is just the opposite, returning very few results, but most of its predicted labels are correct when compared to the training labels. An ideal system with high precision and high recall will return many results, with all results labeled correctly.

We plot Recall on the x-axis and Precision on the y-axis to build a Precision-Recall curve once we calculate Precision and Recall across many threshold values.

**Code Snippet**:



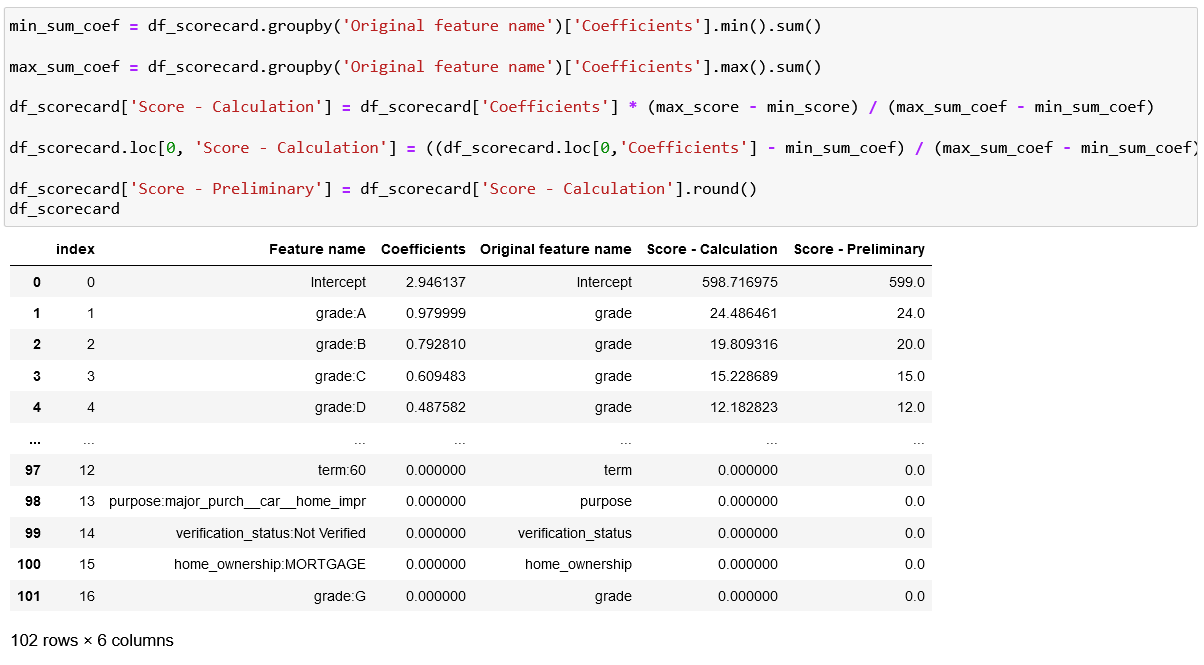


**CONCATENATION OF DATAFRAME:-**

A Data frame is a two-dimensional data structure, i.e., data is aligned in a tabular fashion in rows and columns. We can join, merge, and concat dataframe using different methods.

**In Dataframe df.merge(),df.join(), and df.concat() methods help in joining, merging and concatenate different data frames**.

**Calculating the sum of the minimum and maximum coefficients of each category**

* After concatenating the DataFrame and setting up the minimum and maximum threshold , we calculated the sum of minimum and maximum coefficients .
* Created a new column that has the imputed calculated Score based on the multiplication of the coefficient by the ratio of the differences between maximum & minimum score and maximum & minimum sum of coefficients.

Line 4 of input cell : (df\_scorecard.loc[0, 'Score - Calculation'] = ((df\_scorecard.loc[0,'Coefficients'] - min\_sum\_coef) / (max\_sum\_coef - min\_sum\_coef)) \* (max\_score - min\_score) + min\_score)



**Credit Score**

* **What is a credit score?**

A credit score is a number between 300 and 850 that depicts a consumer's creditworthiness. The higher the score, the better a borrower looks to potential lenders. A credit score is based on [credit history](https://www.investopedia.com/terms/c/credit-history.asp): number of open accounts, total levels of debt, and repayment history, and other factors. Lenders use credit scores to evaluate the probability that an individual will repay loans in a timely manner

* **What does it signify?**

A credit score can significantly affect your financial life. It plays a key role in a lender's decision to offer you credit.

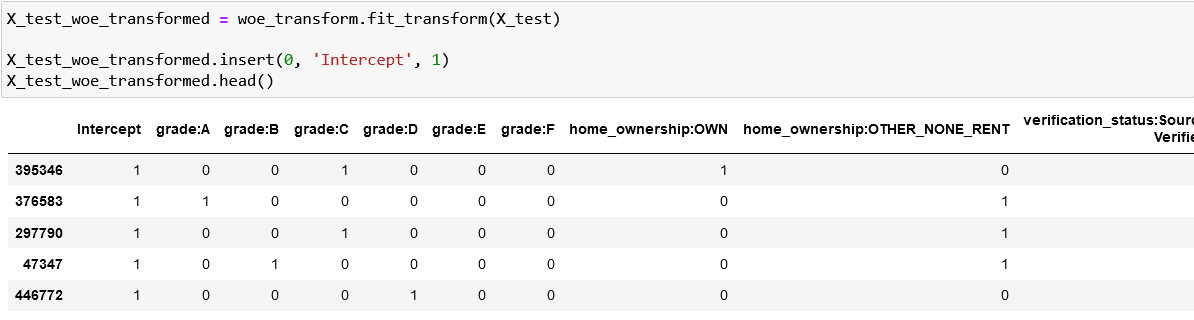
* People with credit scores below 640, for example, are generally considered to be [subprime borrowers](https://www.investopedia.com/terms/s/subprime-borrower.asp). Lending institutions often charge interest on [subprime mortgages](https://www.investopedia.com/terms/s/subprime_mortgage.asp) at a rate higher than a conventional mortgage in order to compensate themselves for carrying more risk. They may also require a shorter repayment term or a co-signer for borrowers with a low credit score.

Conversely, a credit score of 700 or above is generally considered good and may result in a borrower receiving a lower interest rate, which results in their paying less money in interest over the life of the loan. Scores greater than 800 are considered excellent. While every creditor defines its own ranges for credit scores, the average FICO score range is often used

* Excellent: 800 to 850
* Very Good: 740 to 799
* Good: 670 to 739
* Fair: 580 to 669
* Poor: 300 to 579

**Calculating credit scores for all observations in the test data set**

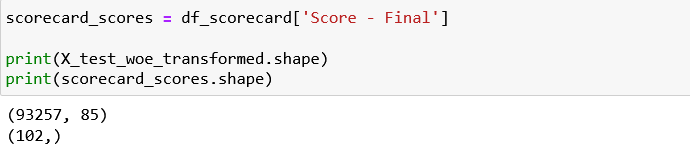
§ create a transformed test set through WoE Binning custom class

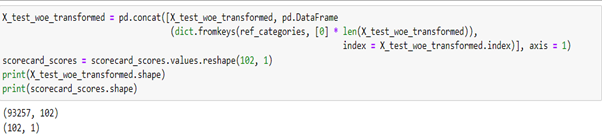
§ insert an Intercept column in its beginning to align with the number of rows in scorecard

§ get the list of final scorecard scores and the shape of test set and scorecard

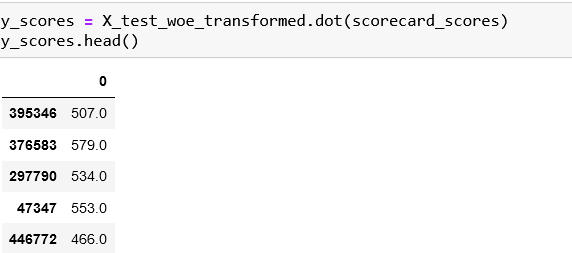
· The number of columns in the test set is lesser (85) than that of rows (102) in the scorecard which does not allow dot multiplication.

· The number of columns in the test set is lesser than that of rows in the scorecard due to reference categories.



§ The reference categories will always be scored as 0 based on the scorecard, so adding these categories to the end of the test set with 0 value.

§ Dot multiplication between test set and scorecard



# 

# **Youden's J statistic**

§ **Definition**

**Youden's J statistic** (also called **Youden's index**) is a single statistic that captures the performance of a [dichotomous](https://en.wikipedia.org/wiki/Dichotomy) diagnostic test. InformedNess is its generalization to the multiclass case and estimates the probability of an informed decision.

The index was suggested by [W.J. Youden](https://en.wikipedia.org/wiki/William_J._Youden) in 1950 as a way of summarising the performance of a diagnostic test. Its value ranges from 0 through 1 (inclusive), and has a zero value when a diagnostic test gives the same proportion of positive results for groups with and without the disease, i.e., the test is useless. A value of 1 indicates that there are no false positives or false negatives, i.e., the test is perfect. The index gives equal weight to false positive and false negative values, so all tests with the same value of the index give the same proportion of total misclassified results.

§ **Calculation and usage**

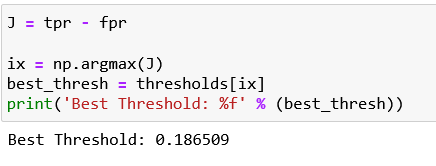
Youden's index is often used in conjunction with [receiver operating characteristic](https://en.wikipedia.org/wiki/Receiver_operating_characteristic) (ROC) analysis. The index is defined for all points of an ROC curve, and the maximum value of the index may be used as a criterion for selecting the optimum cut-off point when a diagnostic test gives a numeric rather than a dichotomous result. The index is represented graphically as the height above the chance line, and it is also equivalent to the area under the curve subtended by a single operating point.

**Setting loan approval cut-offs**

§ Calculate Youden's J-Statistic to identify the best threshold

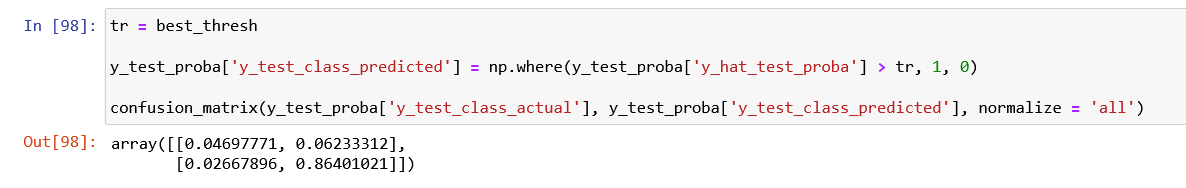
· This means that based on the Youden's J statistic, this is the ideal probability threshold which minimizes the FPR and maximizes the TPR

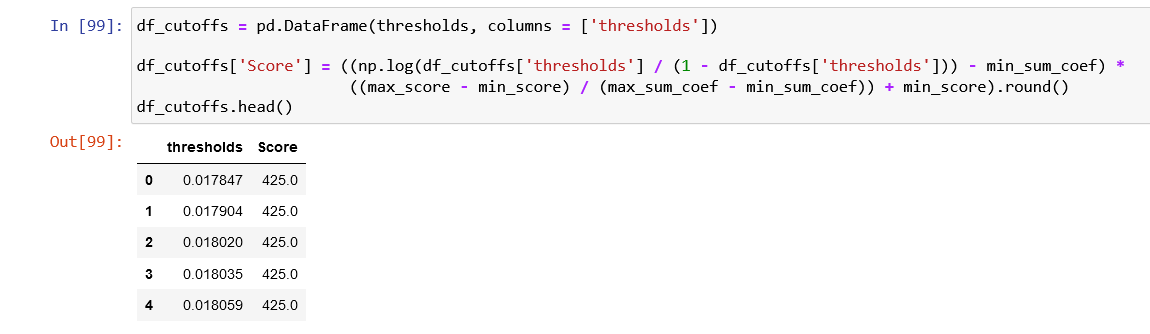
· Which means all samples with a predicted probability higher than this should be classified as in Default and vice versa



§ Update the threshold value, create a new column for the predicted class based on predicted probabilities and threshold and create the confusion matrix

· The updated confusion matrix would show a marked improvement in the TPR at marginal cost of lower TNR but at the same time, FNR has improved drastically with a corresponding marginal increase in FPR.



§ Create a new DF comprising of the thresholds from the ROC output and calculate Score corresponding to each threshold

Create a helper function ‘n\_approved’

· Assigns a value of 1 if a predicted probability is greater than the parameter p which is a threshold and a value of 0 if it is not

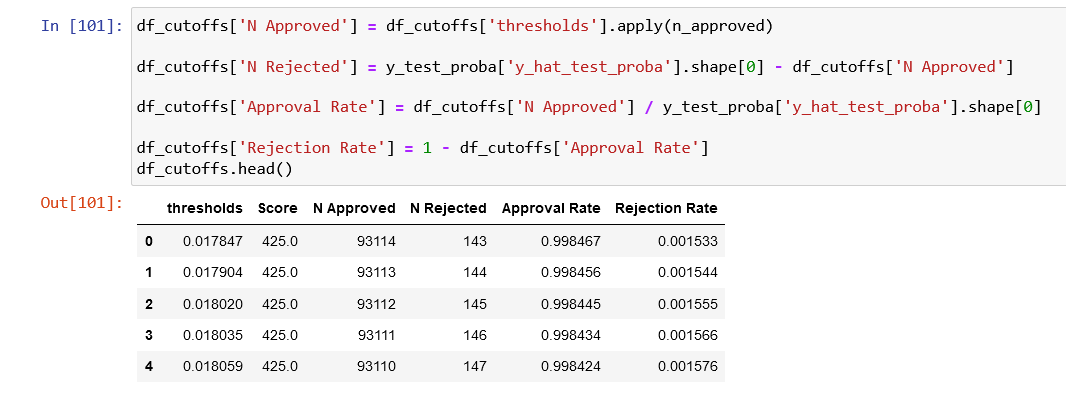
· Then it sums the column.

· for given any percentage values the function will return the number of rows with estimated probabilities greater than the threshold

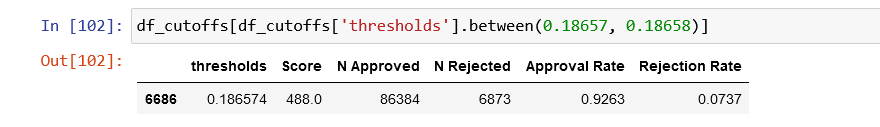


§ Calculate number of approved and rejected applications for each threshold using n\_approved

· approval rate equals the ratio of the approved applications and all applications

· rejection rate equals one minus approval rate

§ Check at the approval and rejection rates at our ideal threshold



§ Compare the above rates with the case of the default 0.5 threshold

