**MONTH – 3**

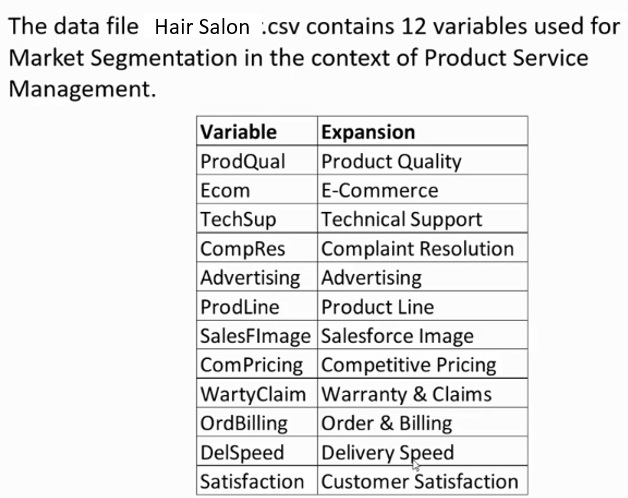
**Project advance stats**

**By : - chetan sharma**

Problem Statement 1: The ‘Hair Salon.csv’ dataset contains various variables used for the context of Market Segmentation. This particular case study is based on various parameters of a

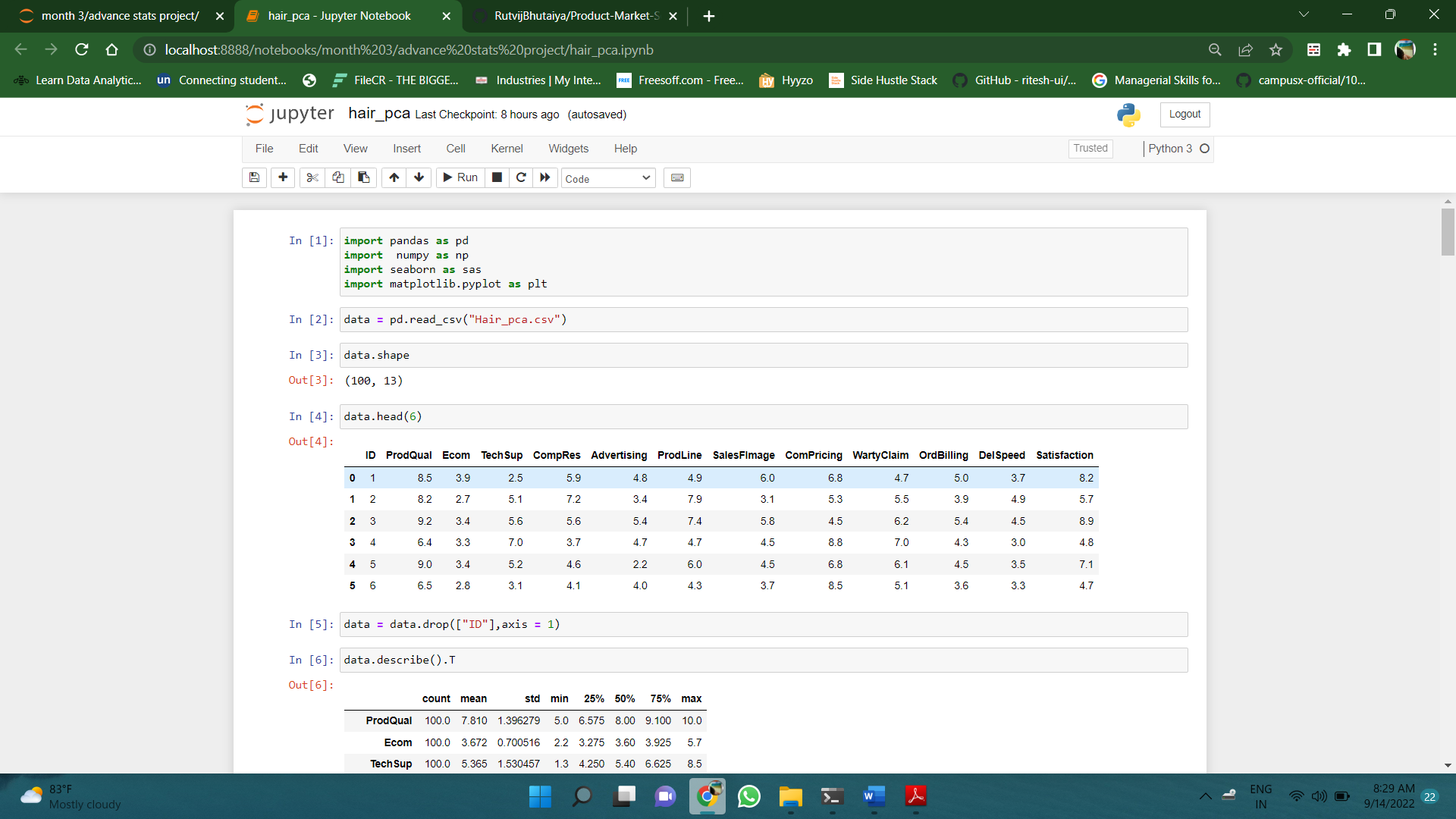
salon chain of hair products. You are expected to do Principal Component Analysis for this case study according to the instructions given in the following rubric.

**Data dictionary**



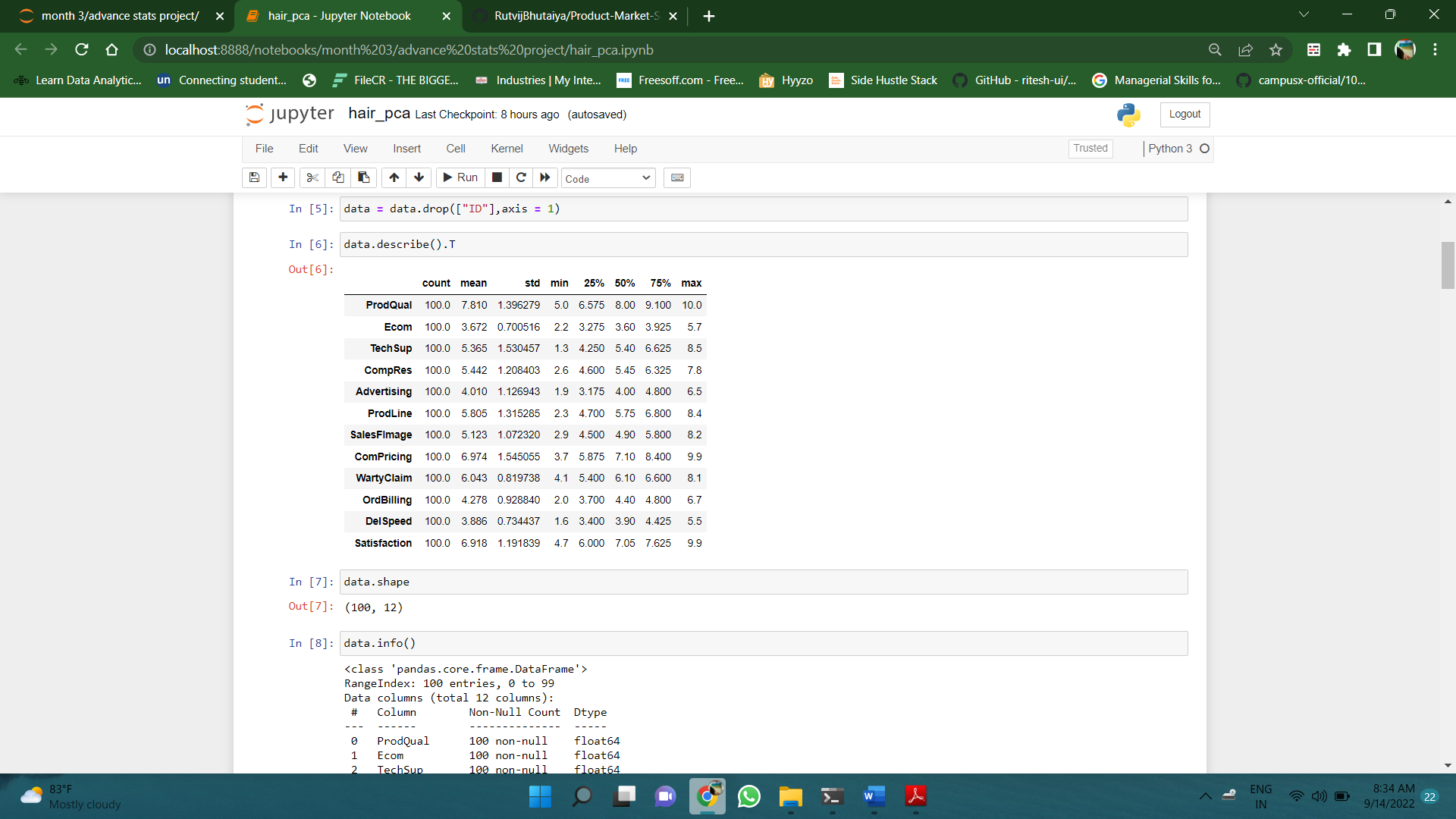
**1) Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. The inferences drawn from this should be properly documented**

**Solution : -**

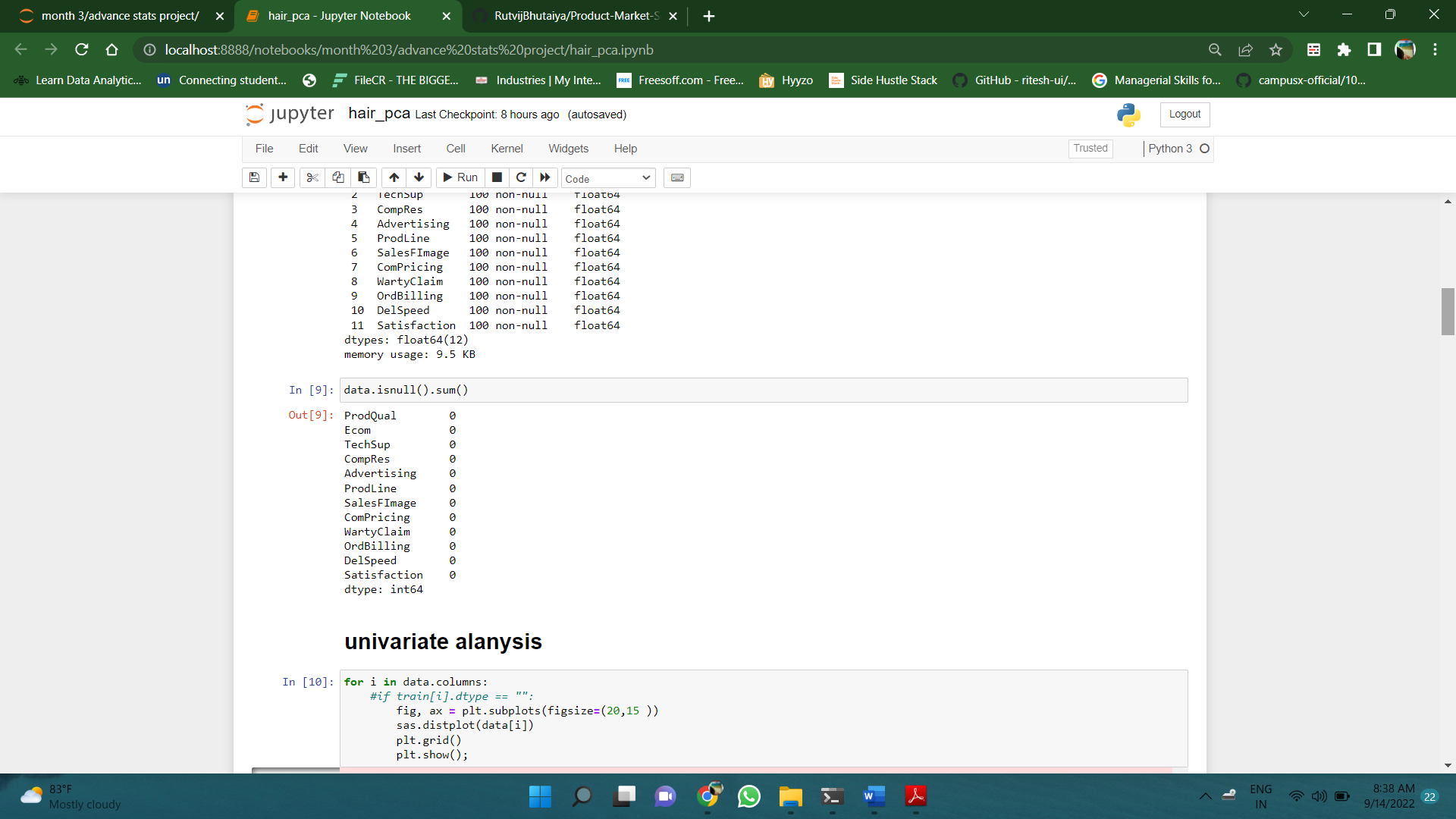


* Out of 13 variables,” satisfaction” is dependent variable and following are independent variable.
* In hari.csv file, categorical data represents customers’ votes ranging between 0 and 10. In the date ‘0’ represents least score and ‘10’ represents highest score. In data set, Satisfaction variable is dependent on ProdQual, Ecom, TechSup, CompRes, Advertising, ProdLine, SalesFImage, ComPricing, WartyClaim, OrdBilling and DelSpeed.

**DESCRIPTIVE STATS**



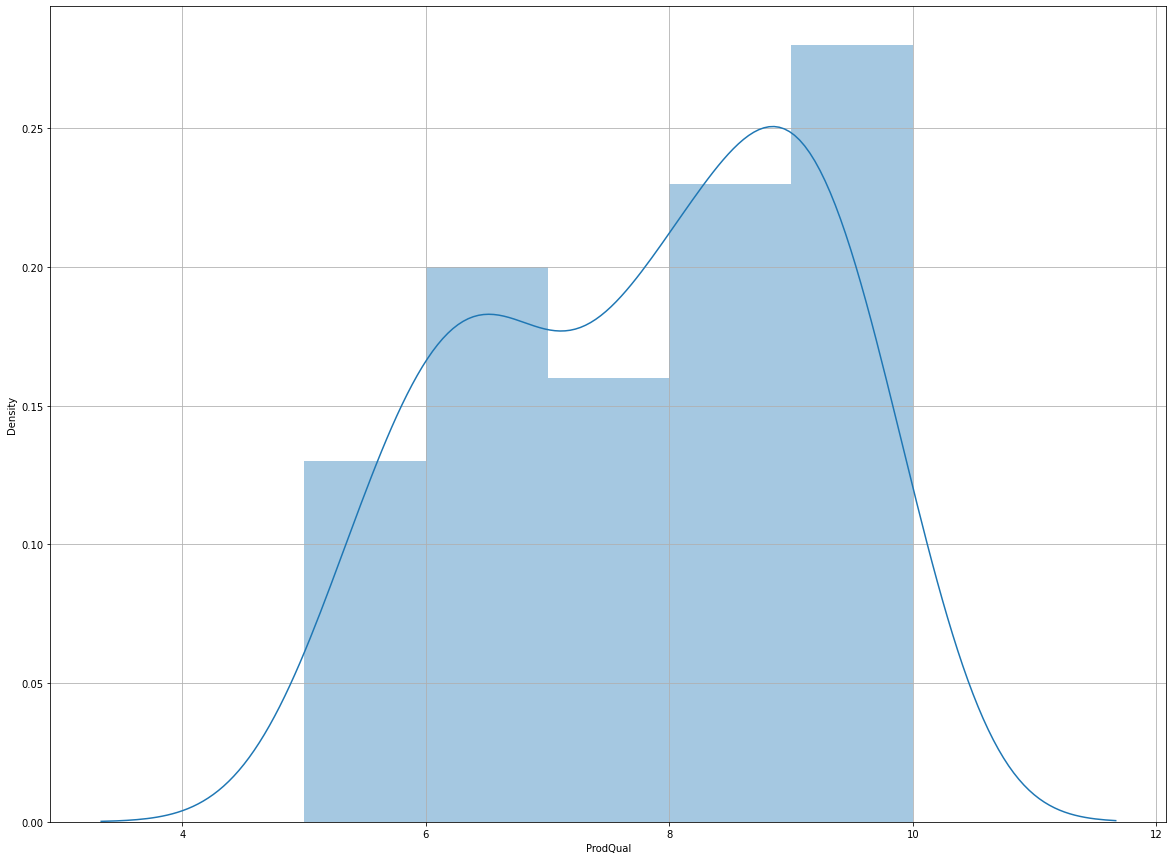
* This data set contain only numeric columns and the value in the data set are in between 0 to 10.
* This data set doesn’t carry any missing value.



1. **Univariate analysis**

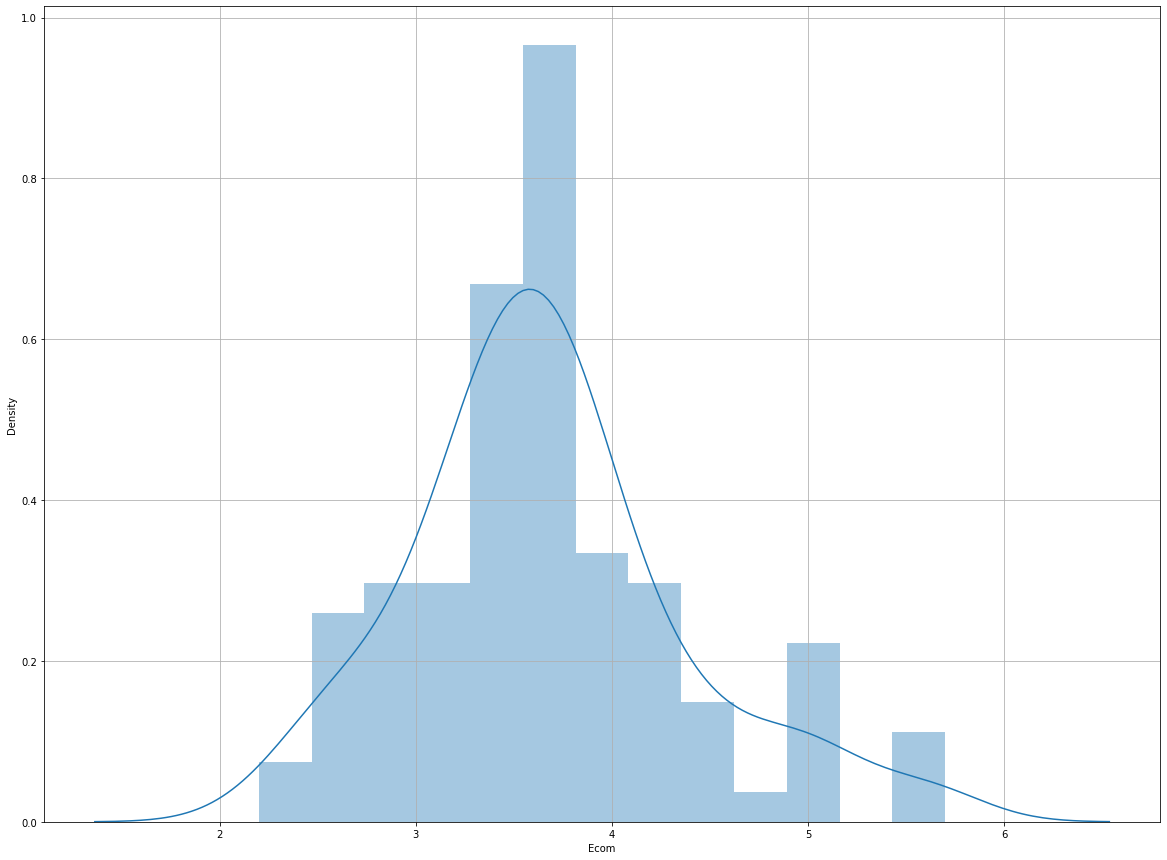
**Distribution plot of all independent variable**

1. **Density vs prodQual**

****

* 1. This graph shows the distribution of PRODUCT\_QUAL rating of FMCG products .
  2. As we see this graph is a kind of normally distributed but slightly right skew.
  3. The minimum value any product get is 5 and the max value a product gets is 10.
  4. A large variety of product gets rating between 8 to 10

**2) Density vs Ecom**

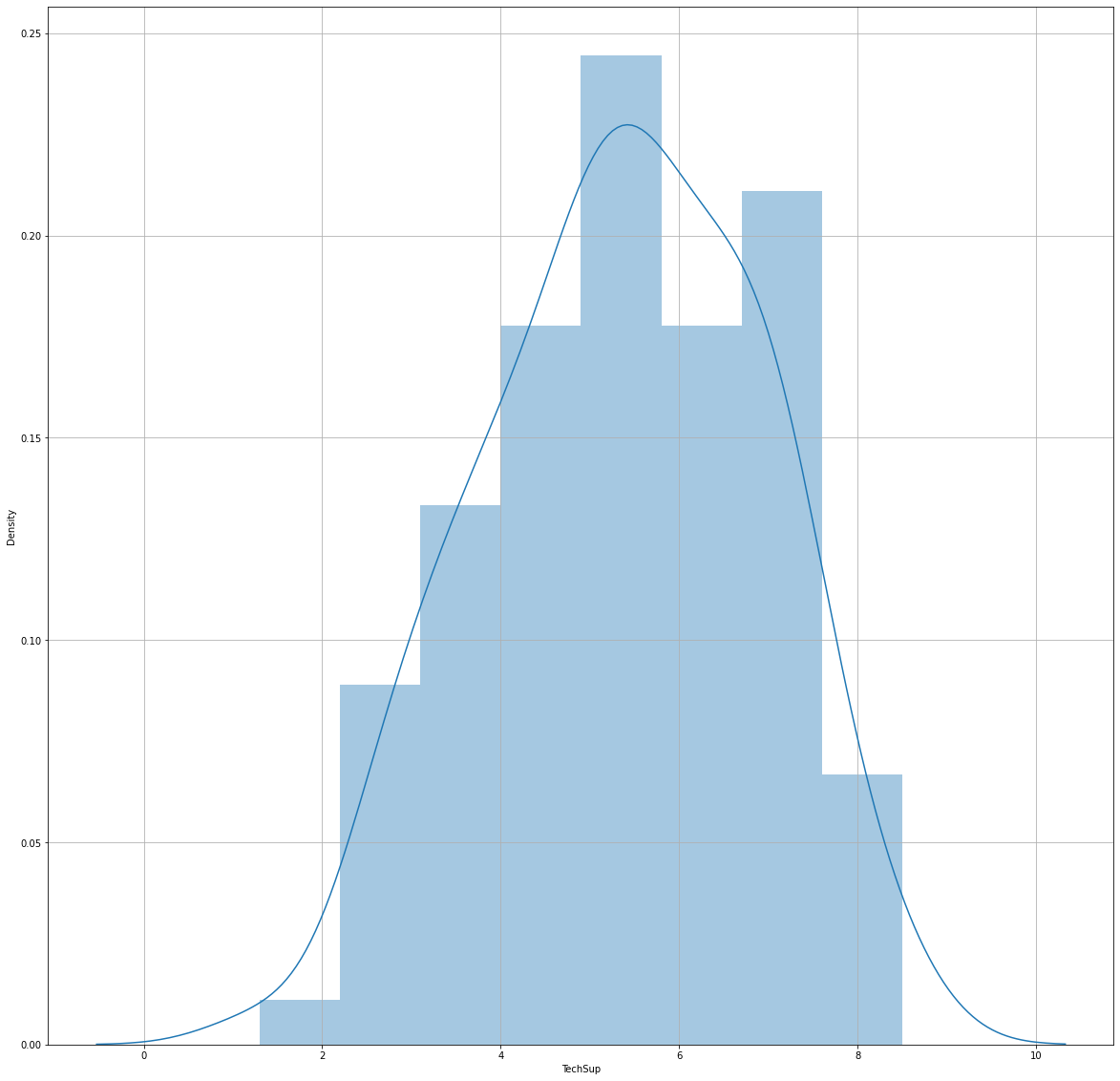
****

**2.1)** This graph shows the distribution of ECOM rating of FMCG products.

**2.2)** As we see this graph is a properly normally distributed.

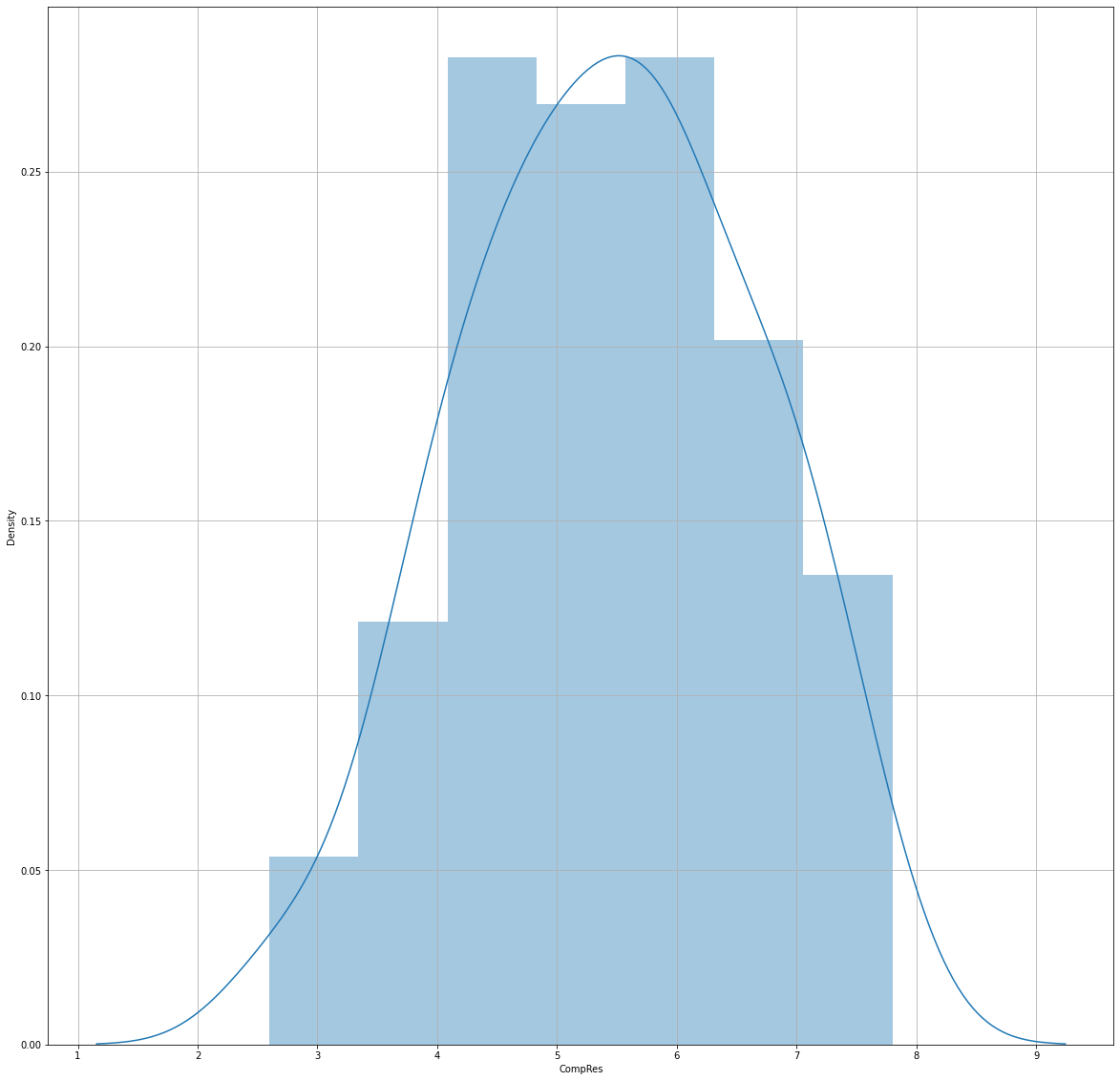
* 1. The minimum value in ECOM is 2.2 and the max value a product gets is 5.7.
  2. A large variety of product gets rating between 3 to 4

3 ) **DENSITY vs TECHSUP**



* 1. This graph shows the distribution of TechSup rating .
  2. As we see this graph is a kind of normally distributed but slightly right skew.
  3. The minimum value any product get is 1.3 and the max value a product gets is 8.5 and mean of 5.365
  4. A large variety of product gets rating between 5 to 7.25

4) **DENSITY vs COMPRES**



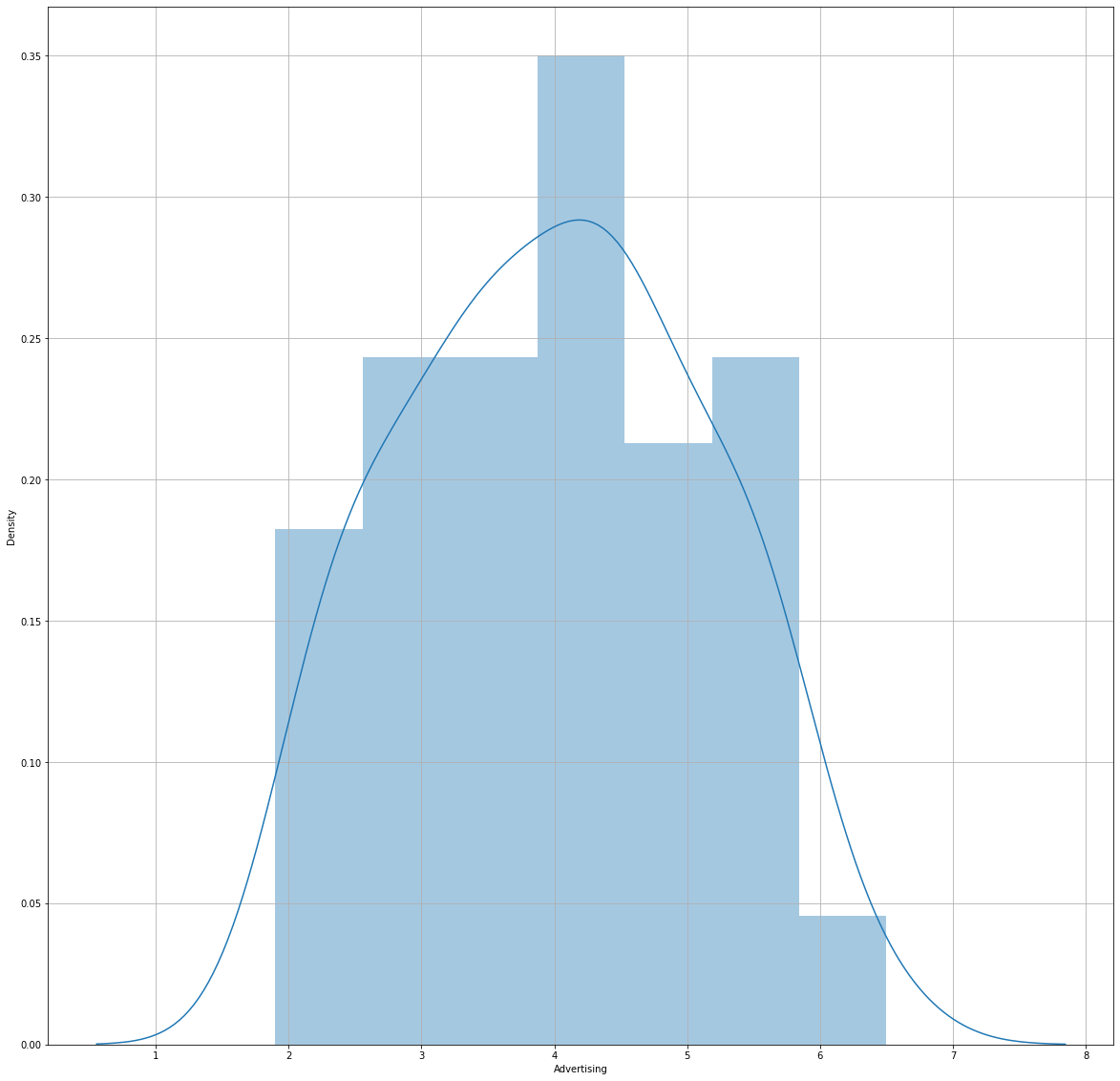
4.1) This graph shows the distribution of Compres rating.

4.2) As we see this graph is a kind of normally distributed but slightly right skew.

4.3) The minimum value any product get is 1.3 and the max value a product gets is 8.5 and mean of 5.365

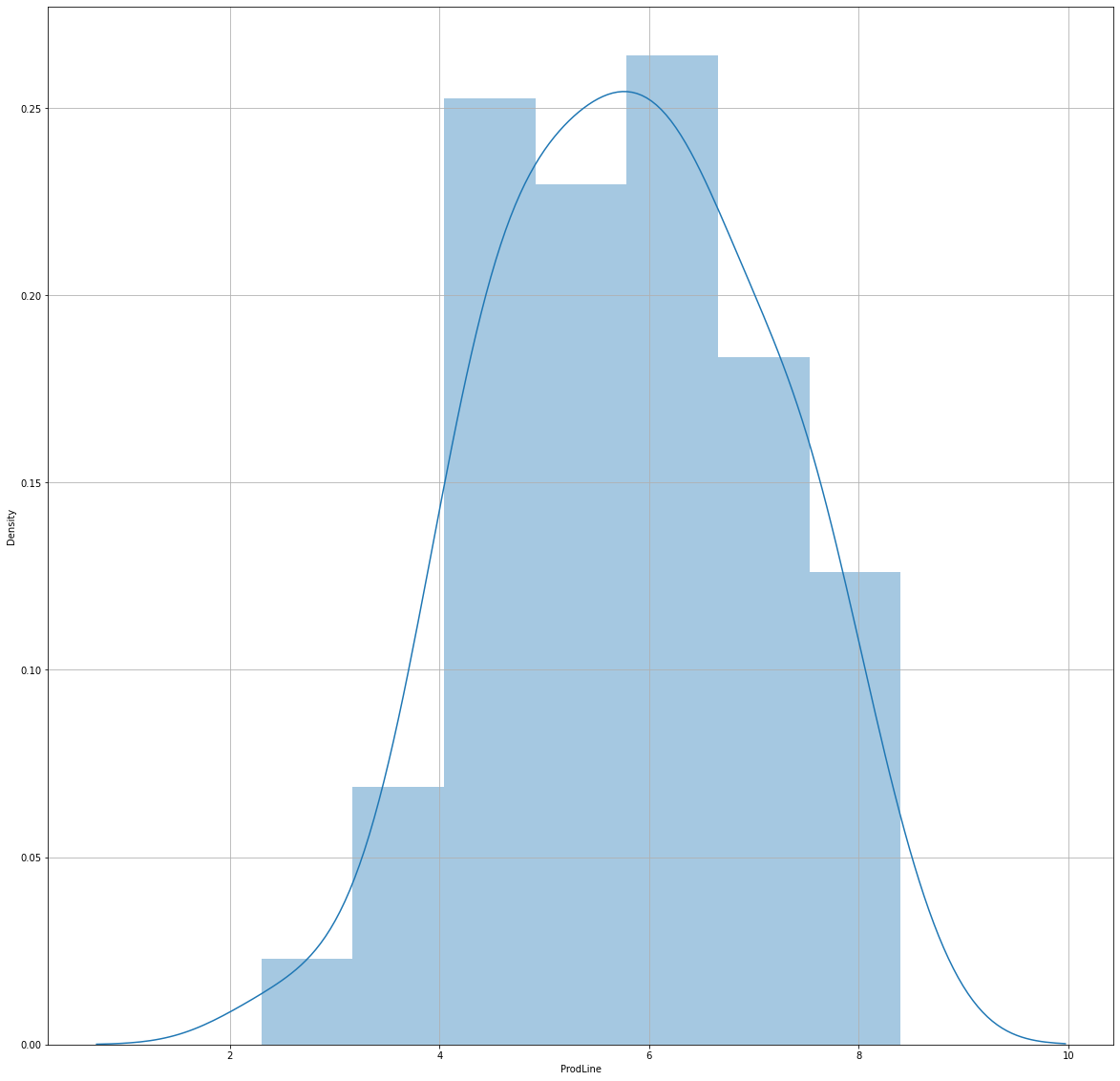
4.4) A large variety of product gets rating between 5 to 7.25

**5)DENSITY vs Advertising**



* 1. This graph shows the distribution of Advertising rating .
  2. As we see this graph is a kind of normally distributed .
  3. The minimum value any product get is 1.9 and the max value a product gets is 6.5 and mean of 4.010
  4. A large variety of product gets rating between 2 to 6

**6) DENSITY vs PROD line**



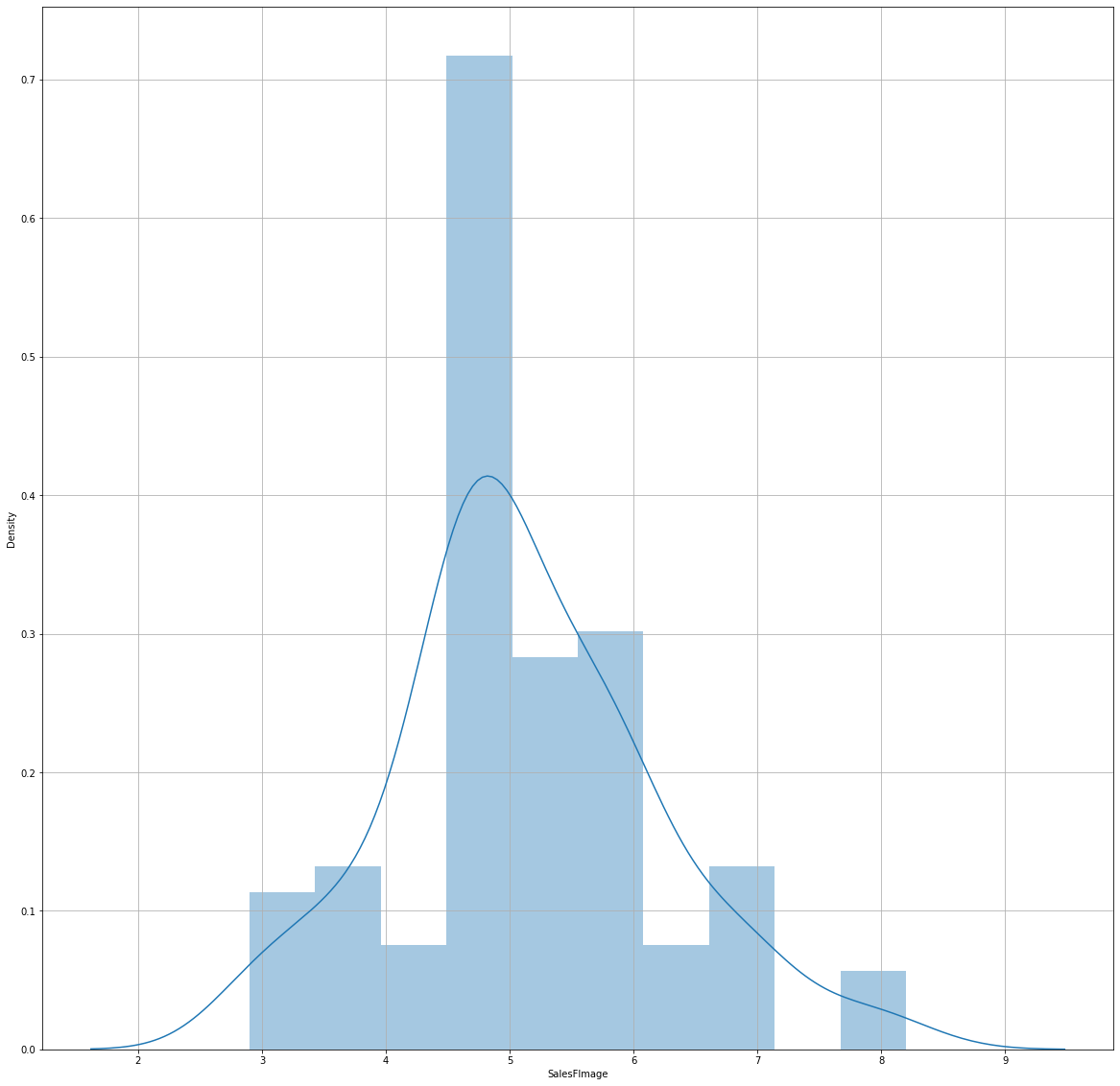
* 1. This graph shows the distribution of ProdLine rating

6.2) As we see this graph is a kind of normally distributed .

6.3) The minimum value any product get is 2.3 and the max value a product gets is 8.4 and mean of 5.805

6.4) A large variety of product gets rating between 4 to 6.25

**7) DENSITY vs SalesFimage**



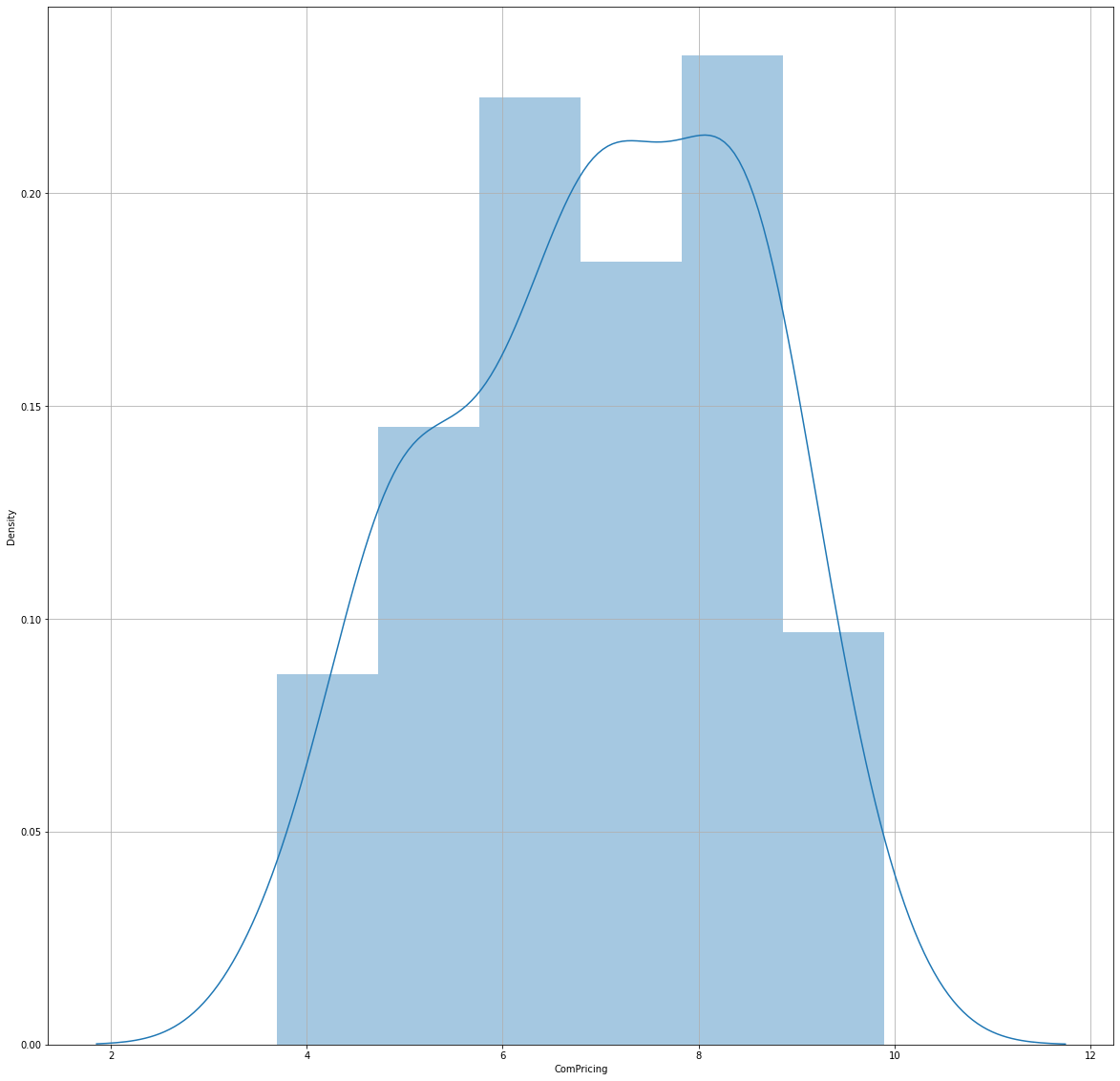
7.1) This graph shows the distribution of SalesFimage rating.

7.2) As we see this graph is a kind of normally distributed but slightly left skew.

7.3) The minimum value any product get is 1.3 and the max value a product gets is 8.5 and mean of 5.365

7.4) A large variety of product gets rating between 5 to 7.25

**8) DENSITY vs ComPricing**



8.1) This graph shows the distribution of ComPricing rating.

8.2) As we see this graph is a kind of normally distributed but slightly right skew.

8.3) The minimum value any product get is 3.7 and the max value a product gets is 8.5 and mean of 9.9

8.4) A large variety of product gets rating between 3.75 to 7.50

**9) DENSITY vs WartyClaim**

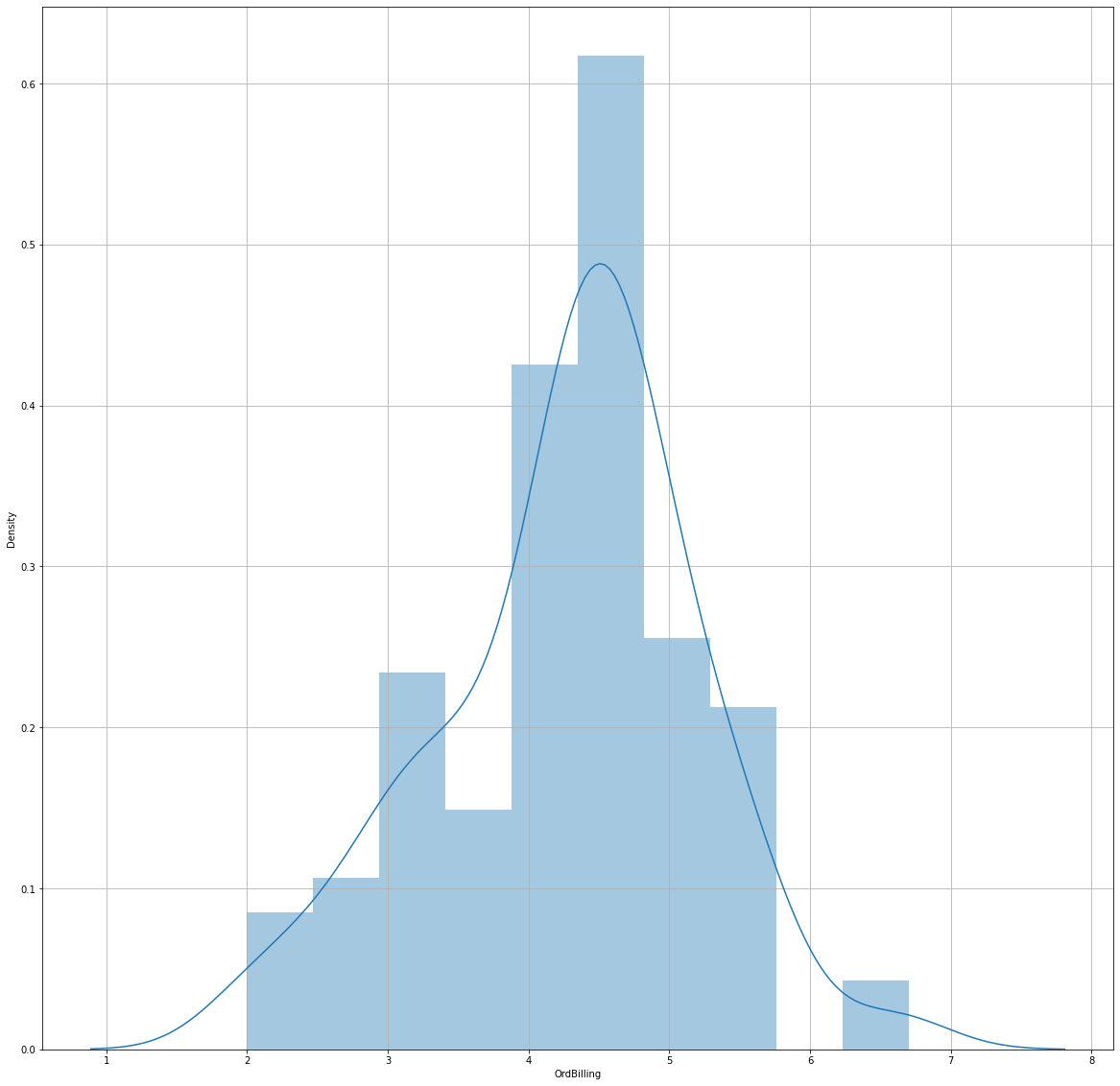
9.1) This graph shows the distribution of WartyClaim rating.

9.2) As we see this graph is a kind of normally distributed.

9.3) The minimum value any product get is 4.1 and the max value a product gets is 8.1 and mean of 6.043

9.4) A large variety of product gets rating between 5 to 7.25

**10) DENSITY vs OrdBilling**



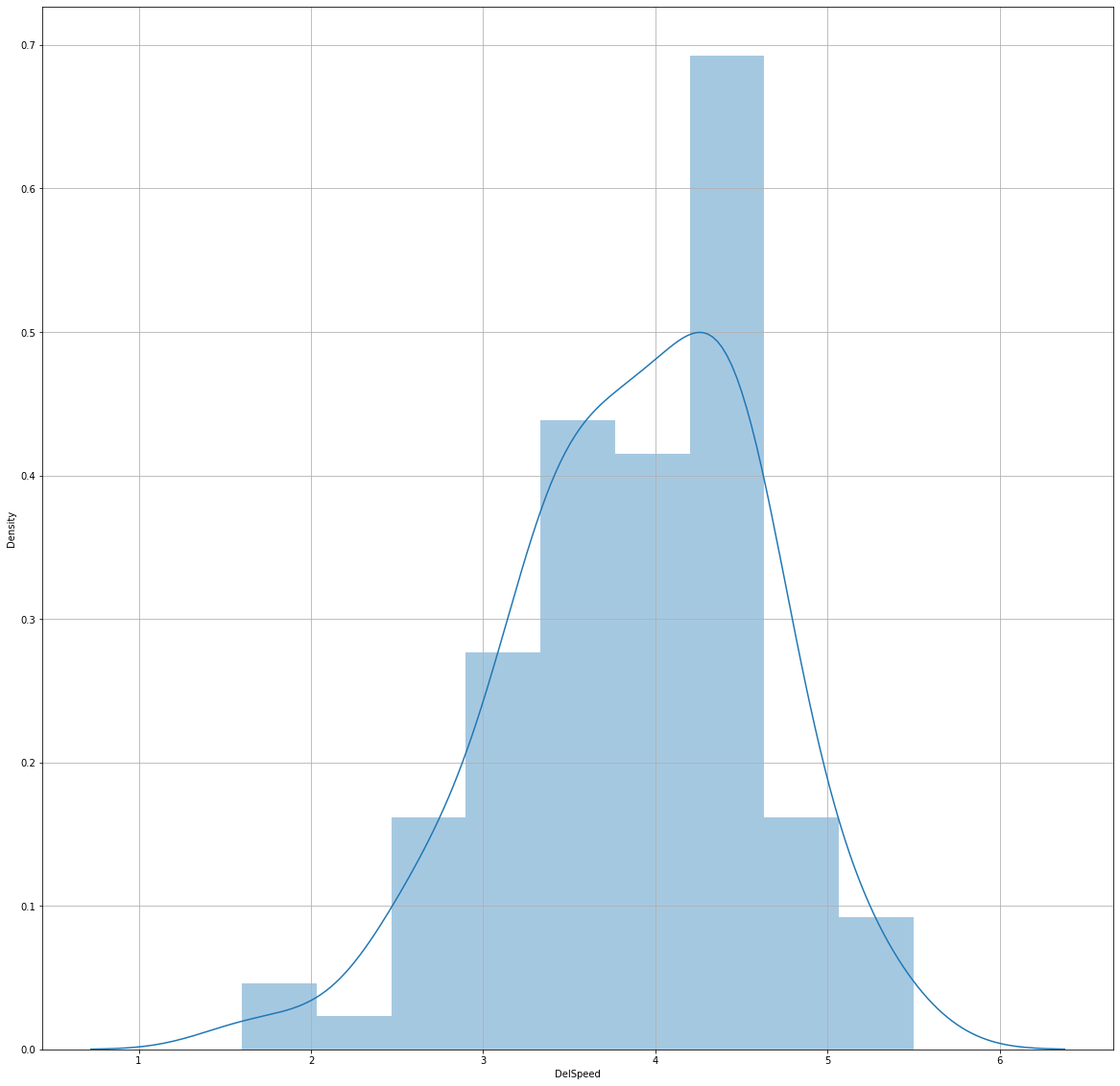
10.1) This graph shows the distribution of OrdBilling rating.

10.2) As we see this graph is a kind of normally distributed but slightly right skew.

10.3) The minimum value any product get is 2.0 and the max value a product gets is 6.7 and mean of 4.278

10.4) A large variety of product gets rating between 4 to 4.80

**11) DENSITY vs DelSpeed**



11.1) This graph shows the distribution of DelSpeed rating.

11.2) As we see this graph is a kind of normally distributed but slightly right skew.

11.3) The minimum value any product is 1.6 and the max value a product gets is 5.5 and mean of 3.886

11.4) A large variety of product gets rating between 3.25 to 4.75

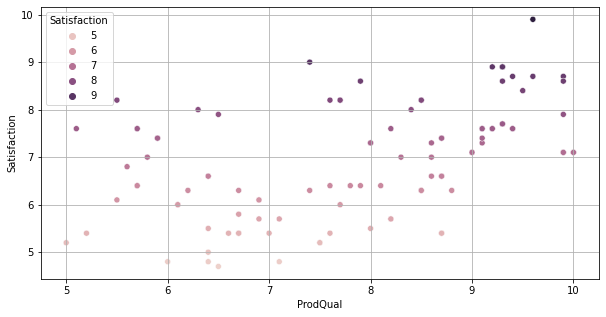
**Conclusion on univariate analysis**

* After analysis of distribution plot of all the feature in the data sets, we can say that the distribution of rating is unbiased. And there are no outliers present in the dataset.
* ,variables’ w-statistics is more than 95% [except ProdQual variable]. For data normality, higher the w-statistics means normal data. However, we also observed that p-value for ProdQual, Ecom, SalesFImage, ComPricing and OrdBilling are less than 0.05. That concludes deviation in normality for selected variables

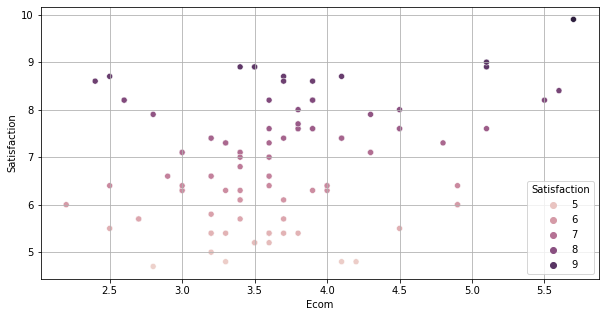
**Bivariate analysis**

As we know that scatter plot is a best way to analysis bivariate analysis because it show the clear picture about the relation ship between two variable

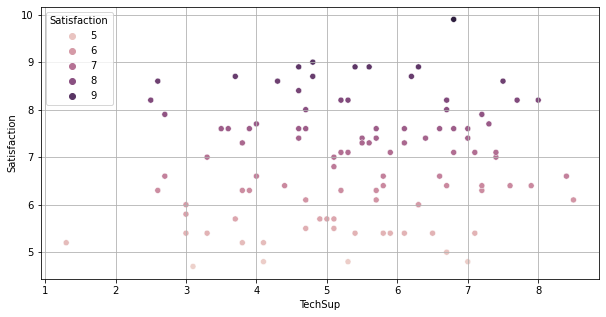
1. ProdQual vs satisfaction



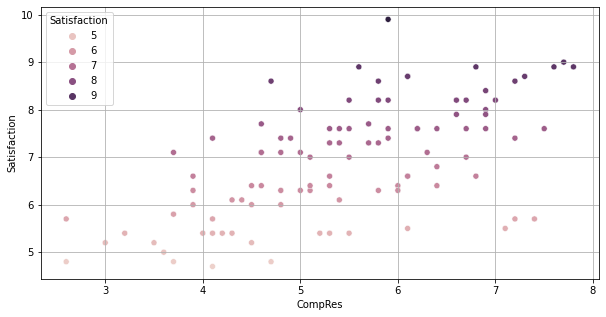
1. Ecom vs satisfaction



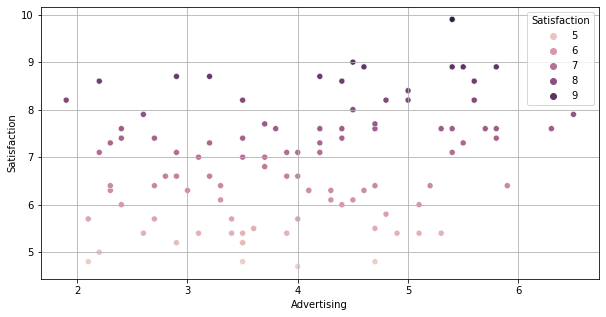
1. TechSup vs satisfaction



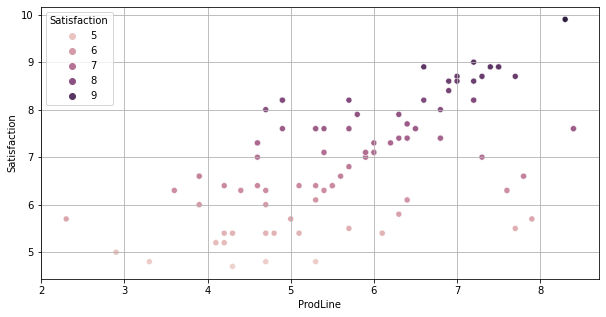
1. compRes vs satisfaction



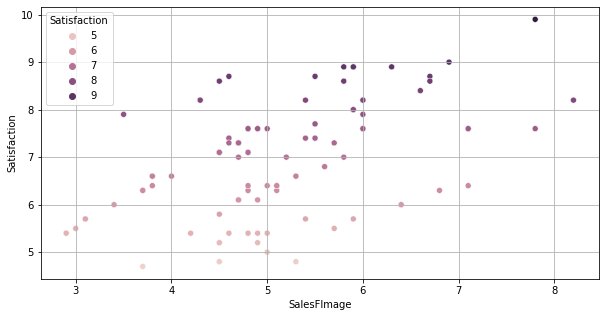
1. Advertising vs satisfaction



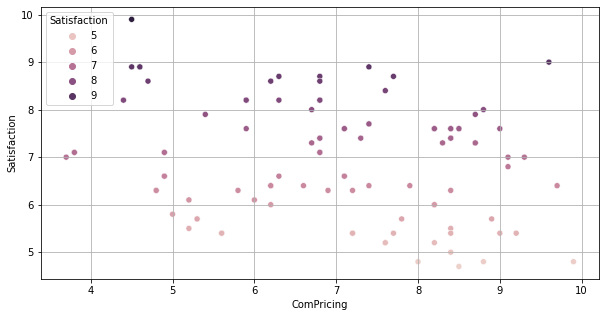
1. ProdLine vs satisfaction



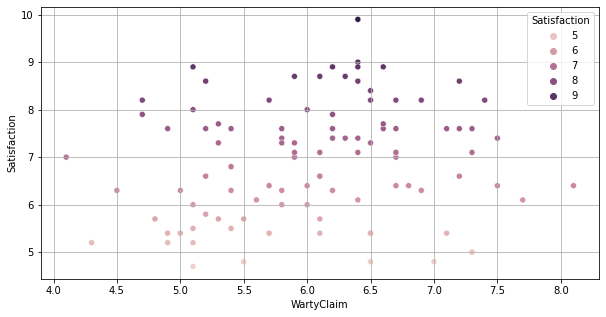
1. SalesFlmage vs satisfaction



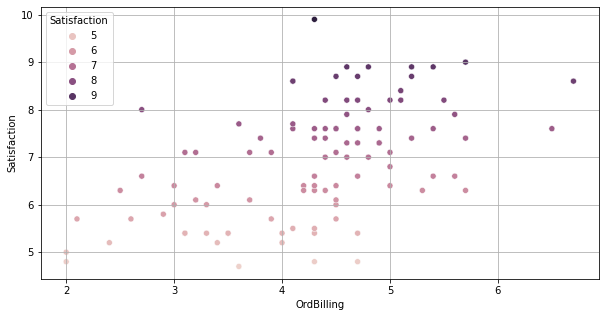
1. ComPricing vs satisfaction



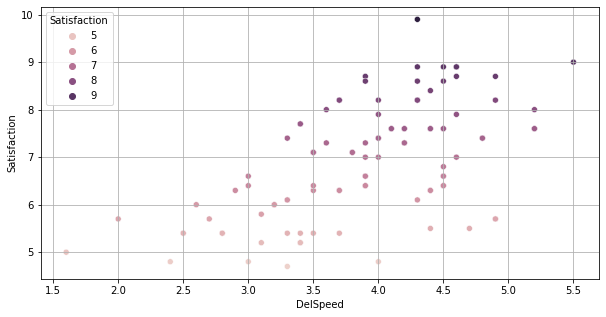
1. Wartyclam vs satisfaction



1. Orderbilling vs satisfaction



1. Delspeed vs satisfaction



**Conclusion on bivariate analysis**

* As all the graphs state that there are not a proper relation between variable and dependent variable. There are mix response in between variable.
* Some with high rating have low satisfaction level and some with low rating have high satisfaction level .
* We cant predict the clear relation among them

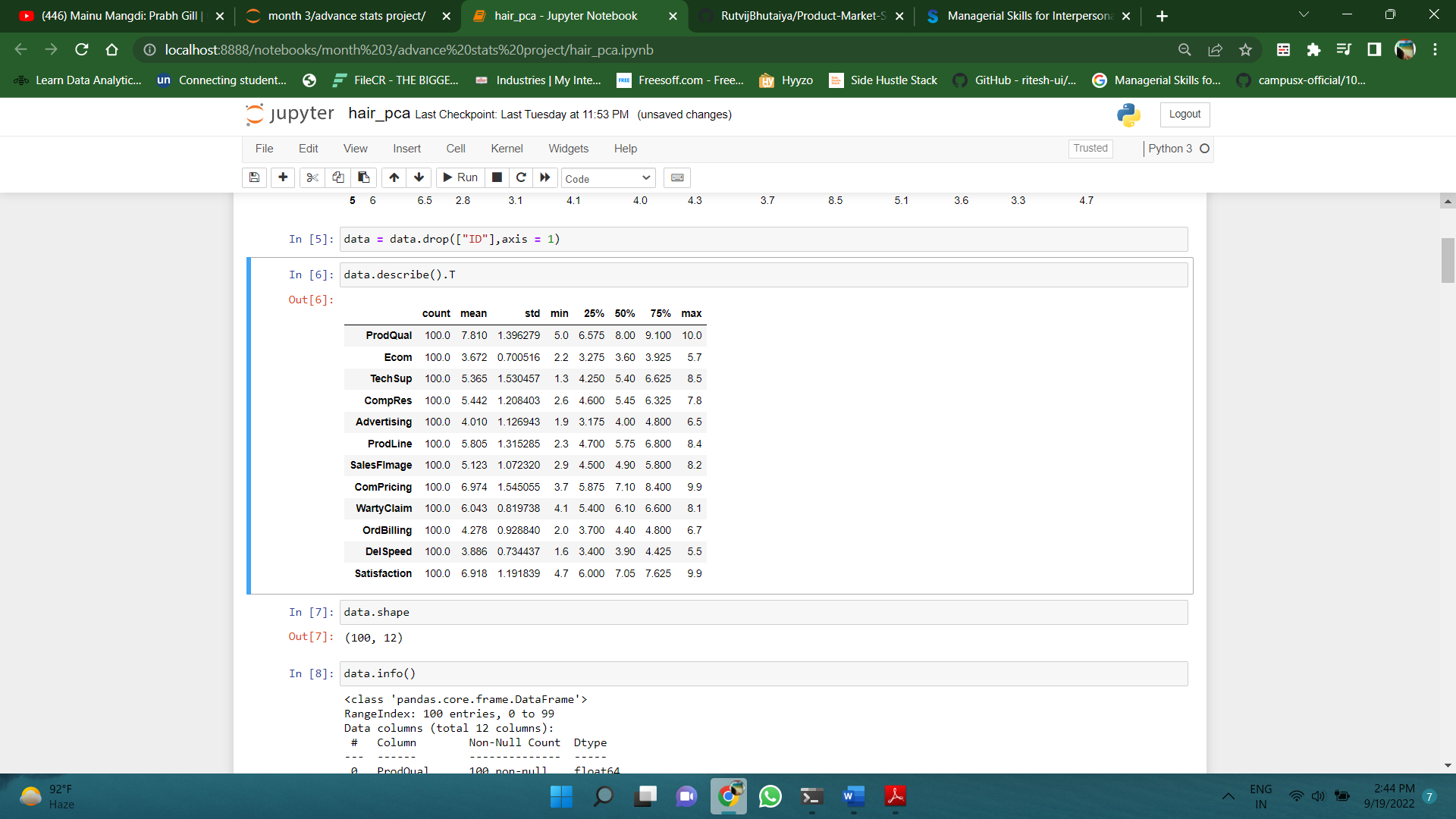
**Q2)** Scale the variables and write the inference for using the type of scaling function for this case study.

* For scaling we use z-score scaling or standard scaler (both are same).
* Scaling is usually done to give all variables an even contribution to the variation in the data. Scaling enables comparing variables that once had different value ranges. The scaling done was standardizing the values of all variables. This way, all variables will have a mean of 0 and a standard deviation of 1.
* The standard score of a sample x is calculated as:

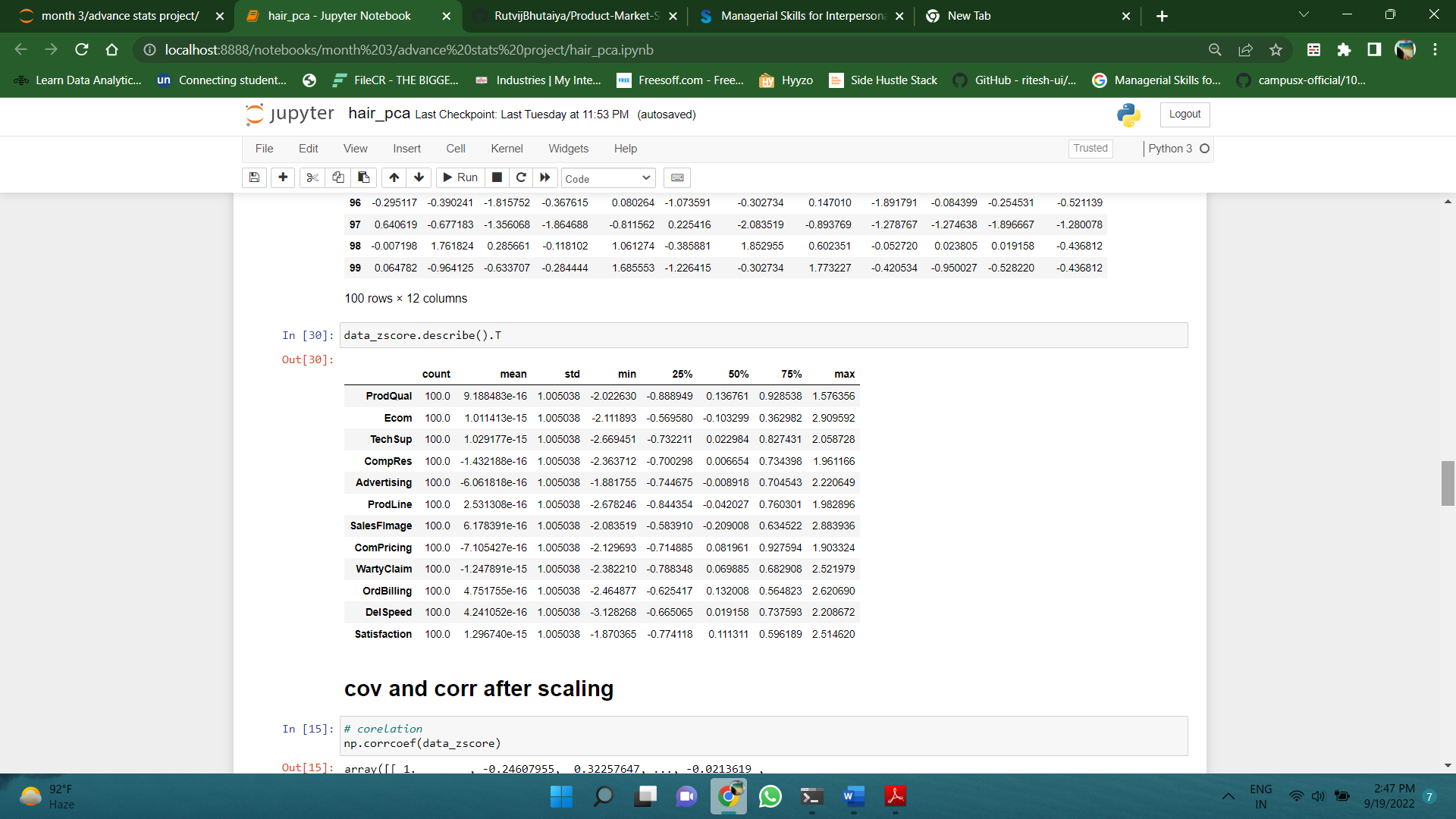
z = (x - u) / s

where u is the mean of the training samples or zero if with\_mean=False, and s is the standard deviation of the training samples or one if with\_std=False.

* The function of scaling is to arrange or reduce the spread of data.
* Data before scaling



* Here we can see that the spread for all variable is between 0 to 10. This spread creates problem related to the distance between the elements.
* Data after scaling



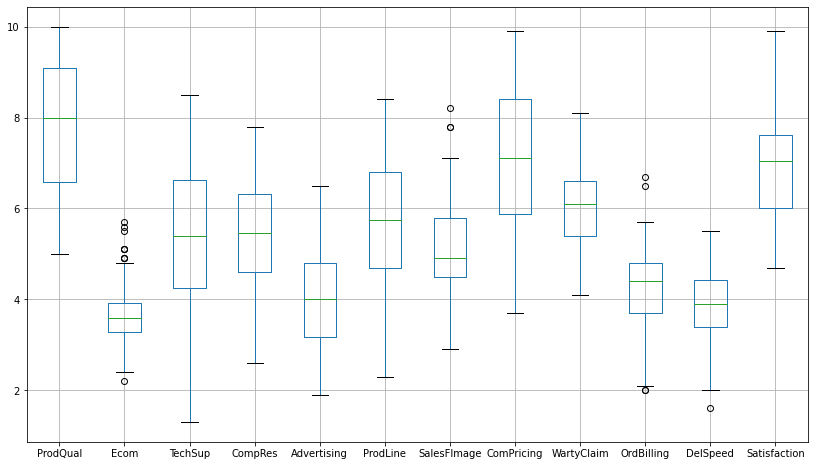
* In this table we can see that the distance between two variable is between -3 to 3

Q- 3 ) Comment on the comparison between covariance and the correlation matrix after scaling.

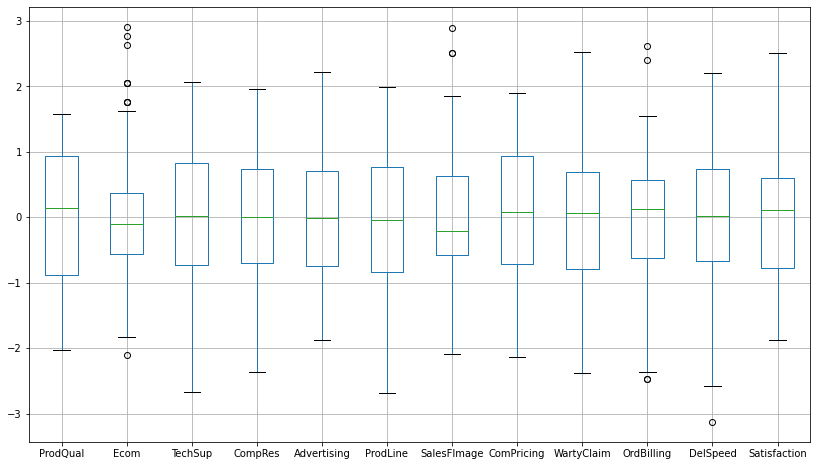
* Covariance and correlation are two terms that are opposed and are both used in statistics and regression analysis. Covariance shows you how the two variables differ, whereas correlation shows you how the two variables are related.
* The diagonal values in correlation matrix is 1 while diagonal value in covariance matrix is1.010101
* After scaling, the covariance and correlation matrices become identical. Furthermore, the correlation matrix before scaling is similar to the covariance and correlation matrices after scaling. The reason those variables became similar after scaling is that all variables have a standard deviation of 1. Correlation is computed as:
* *ρ*=*stdev*(*X*)∗*stdev*(*Y*)*covariance*(*X*,*Y*)​  
  As aforementioned, after scaling, all variables had a standard deviation of 1, which means the denominator becomes 1 (you multiplied 1 and 1). Hence, the correlation of two standardized variables will be equal to their covariance.
* After scaling we saw a noticeable change in covariance matrix while there is no noticeable change in correlation matrix.

4) Check the dataset for outliers before and after scaling. Draw your inferences from this exercise.

Box plot before scaling



Boxplot after scaling

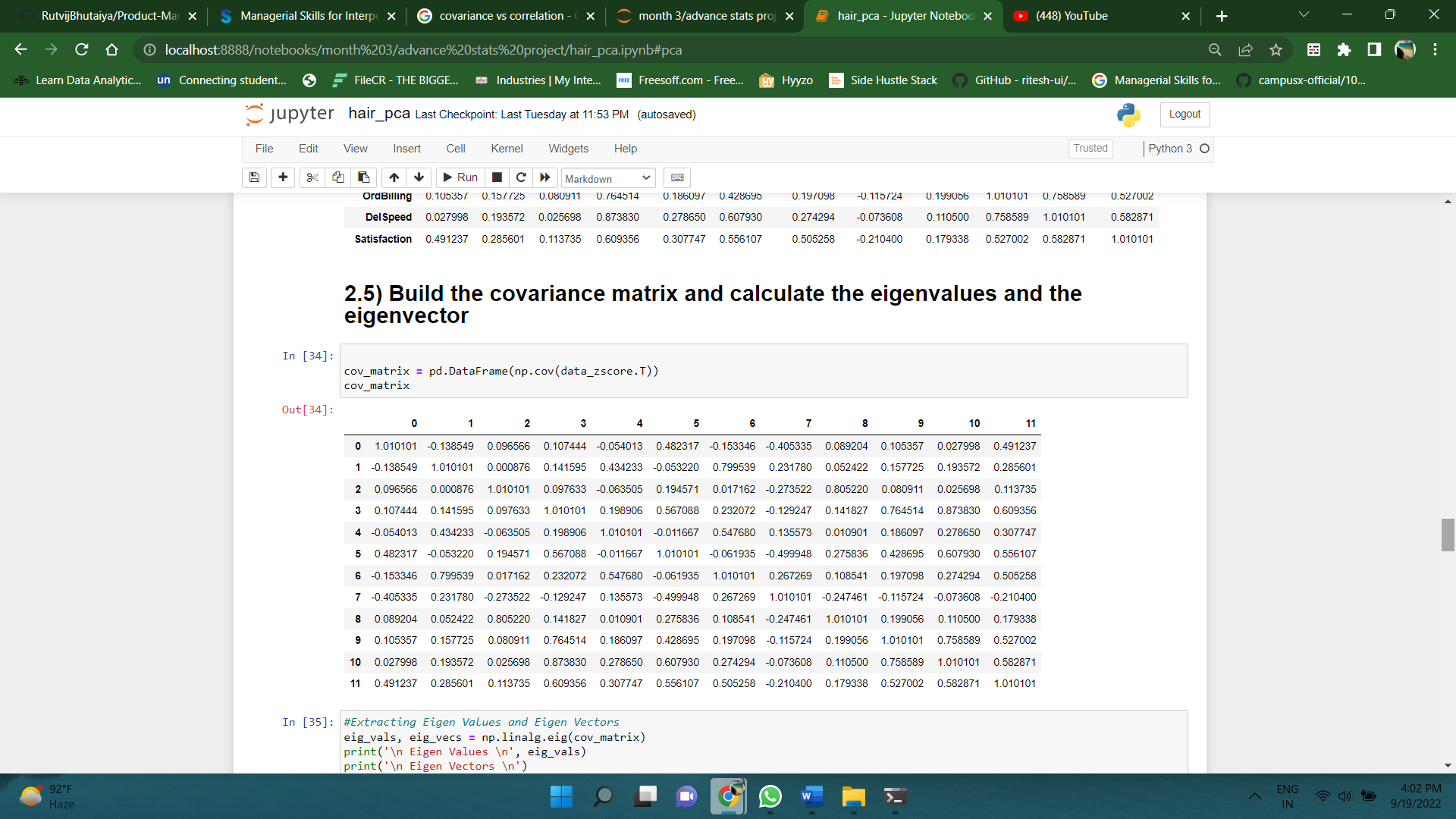


After analyzing both the boxplot plot

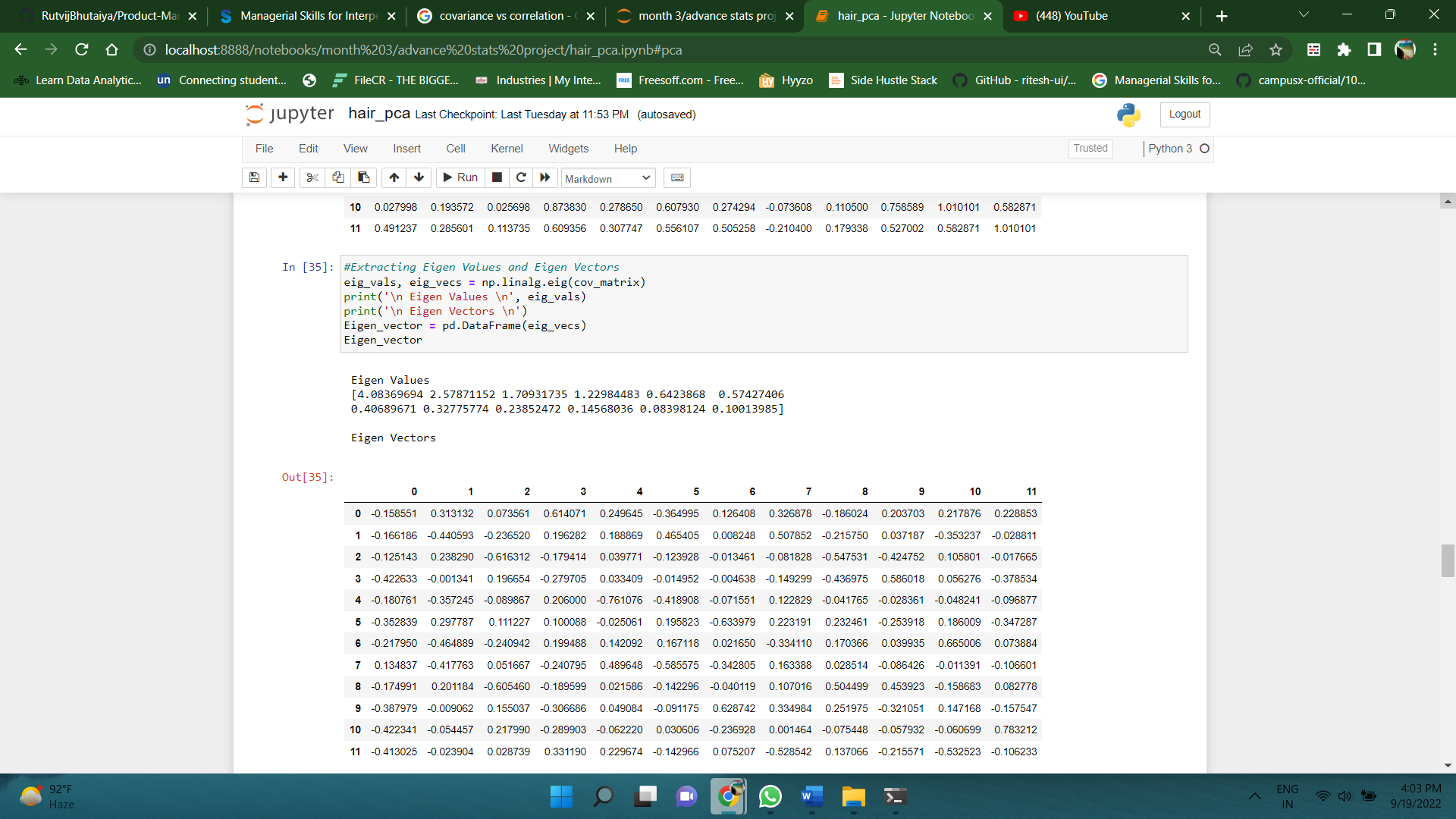
* In second boxplot chart we can see that the shift to variable to same scale. deviation is less
* Scaling didn’t work on outliers
* All the outliers in both the boxplot are remain same.
* In second boxplot we can see that all the variables are placed around 0 . spread is in controlled manners .

5) Build the covariance matrix, eigenvalues and eigenvector.

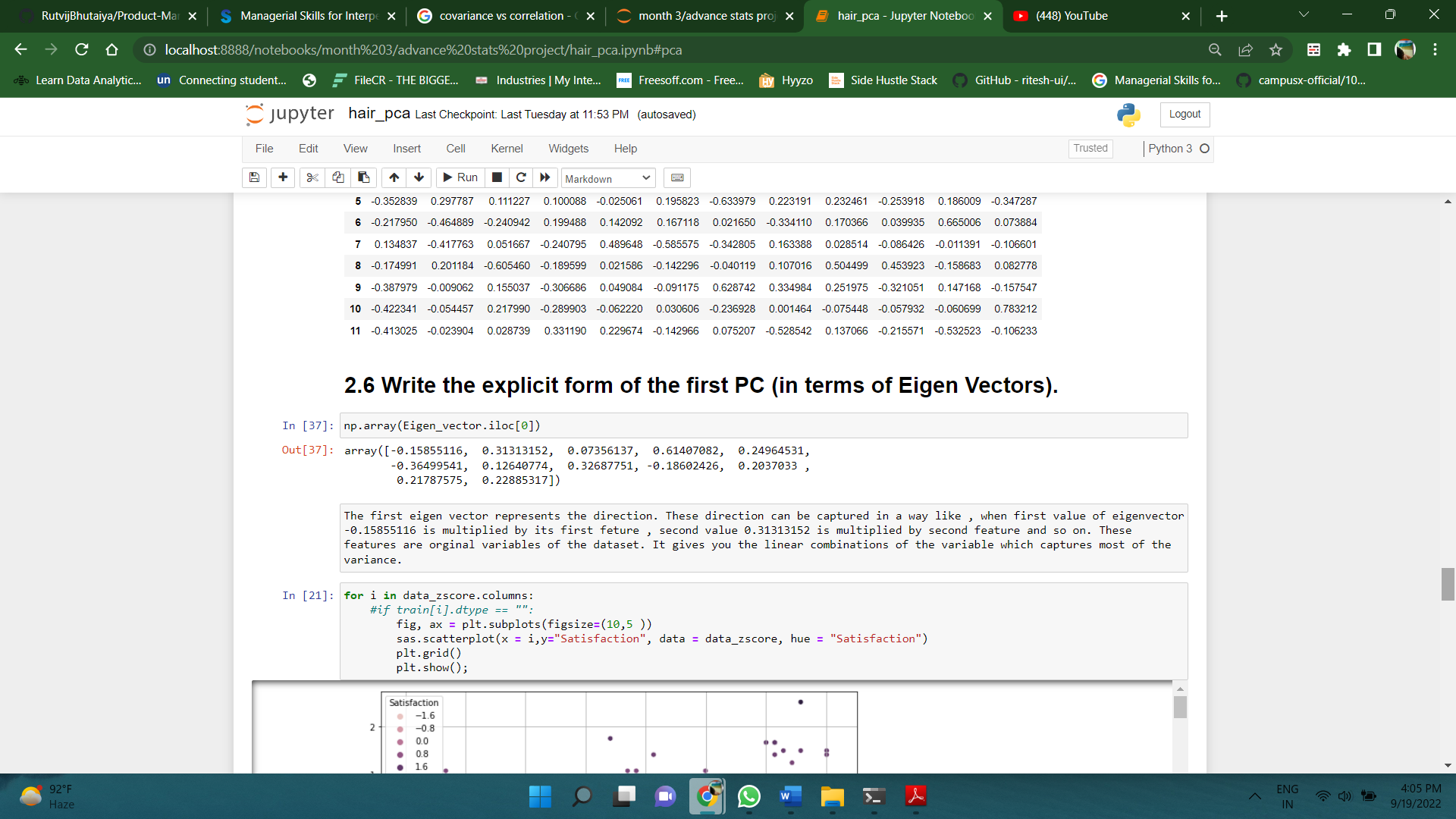
* Covariance matrix



* Eigen value and eigen vector



6) Write the explicit form of the first PC (in terms of Eigen Vectors



The first eigen vector represents the direction. These direction can be captured in a way like , when first value of eigenvector -0.15855116 is multiplied by its first feture , second value 0.31313152 is multiplied by second feature and so on. These features are orginal variables of the dataset. It gives you the linear combinations of the variable which captures most of the variance.

7) Discuss the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate? Perform PCA

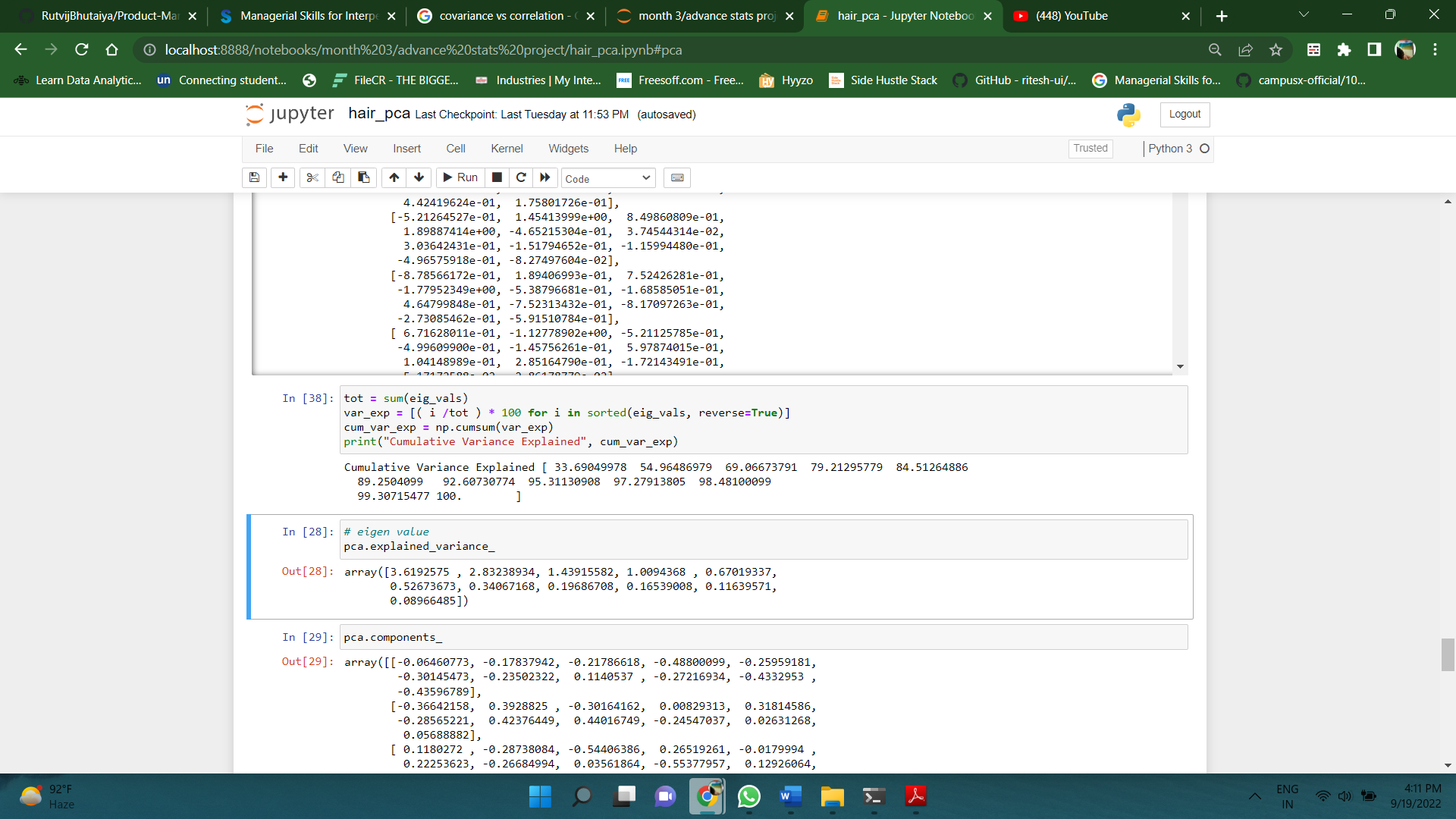
and export the data of the Principal Component scores into a data frame

Eigen values are the variances corresponding to different Principal Components. First eigen value corresponds to first eigen value OR first PC, if that is high proportion of total variance then we are accomplished in reduction of high dimensions using first PC only. Similarly, we can go for as many as PC's are required that gives us maximum variability and helps us to capture most of the variance from the data set.[¶](http://localhost:8888/notebooks/month%203/AS_Nishtha%20Laul_20.12.2020-1.ipynb#Eigen-values-are-the-variances-corresponding-to-different-Principal-Components.-First-eigen-value-corresponds-to-first-eigen-value-OR-first-PC,-if-that-is-high-proportion-of-total-variance-then-we-are-accomplished-in-reduction-of-high-dimensions-using-first-PC-only.-Similarly,-we-can-go-for-as-many-as-PC's-are-required-that-gives-us-maximum-variability-and-helps-us-to-capture-most-of-the-variance-from-the-data-set.)

In the above Cumulative Variance Explained array, we see that the first feature derives 33.69% of the variance within our data set while, the first two gets 54.96% and so on. If we employ 10 features like this, we will be able to capture approximately 98.48% of the variance within the dataset, considering it as diminishing marginal return on total variance explained. Hence for now, we are generating only 11 PCA dimensions .

Eigen vector indicates that

This line of best fit, shows the direction of maximum variance in the dataset. The Eigenvector is the direction of that line,



You can use the eigenvalues and cumulative variance to determine the number of principal components you should retain. Principal components with eigenvalues of 1 are usually retained. PCs are arranged in decreasing order of variance. It means that PC1 will always have the highest proportion of variance covered. Higher proportion of variance that is covered, the better. You can observe that as you go to the last PC, difference in proportion of variance becomes less and less, making it significant. Only retain PCs that have noticeably high proportion of variance. **In this case study, four principal components should be retained.**

### 2.8 Mention the business implication of using the Principal Component Analysis for this case study.

PCA is a statistical tool that is used to reduce multidimensional data which means picking an observation in a high dimensional space and representing in into lower dimension where we can visualise the data more easily and effectively by retaining its accuracy and the information that is present in the data. It captures the variance in the data in a best possible way. It uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables. Principal Component Analysis (PCA) is a well-established mathematical technique for reducing the dimensionality of data, while keeping as much variation as possible. In this case study, we have applied PCA on all the variables except column header “Names”, as its categorical and we have dropped it from the data set. Since, all other variables are continuous in nature on which PCA performs. Its effectiveness depends upon the scales of attributes and that is why standardization of data is the important part through which data gets centred on the origin by subtracting the mean from data and divided by Standard Deviation. Here, we have used z score scaling method as it handles the outliers in a better way. Further, we move on to generate the covariance matrix/ correlation matrix for all the variables that help us to capture the relationship between all the variables of the data. After covariance matrix, we have decomposed it into the coordinate axis and perform the eigen decomposition i.e, to compute the eigen vectors which are basically the Principal Components and the corresponding eigen values which are the magnitudes of variances captured. Eigen values are the variances corresponding to different Principal Components. First eigen value corresponds to first eigen vector OR first PC, if that is high proportion of total variance then we are accomplished in reduction of high dimensions using first PC only. Similarly, we can go for as many as PC's are required that gives us maximum variability and helps us to capture most of the variance from the data set. In the above Cumulative Variance Explained array, we see that the first feature derives 33.48% of the variance within our data set while, the first two gets 54.8% and so on. If we employ 10 features like this, we will be able to capture approximately 98.3% of the variance within the dataset, considering it as diminishing marginal return on total variance explained. Hence for now, we are generating only 10 PCA dimensions. With the help of 10 PC’s, variability of universities are described in a better way in lieu of all the given variables given in the data set.