|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Field** | **Field Description** | **Data Type** | **Modeling Type** | **Changes Required, If Any** | **Comments/Reason for Change** |
| Id | Record\_ID | Numeric | Continuous | No |  |
| Model | Model Description | Character | Nominal | No |  |
| Price | Offer Price in EUROs | Numeric | Continuous | No |  |
| Age\_08\_04 | Age in months as in August 2004 | Numeric | Continuous | No |  |
| Mfg\_Month | Manufacturing month (1-12) | Numeric | Nominal | No | Nominal - Because it takes finite number of values. |
| Mfg\_Year | Manufacturing Year | Numeric | Nominal | No | Nominal - Because it takes finite number of values. |
| KM | Accumulated Kilometers on odometer | Numeric | Continuous | No |  |
| Fuel\_Type | Fuel Type (Petrol, Diesel, CNG) | Numeric | Nominal | Yes | Petrol, Diesel and CNG are dummy variables converted to  numeric but they are basically categorical variables with  just two values. |
| HP | Horse Power | Numeric | Continuous | No |  |
| Met\_Color | Metallic Color? (Yes=1, No=0) | Character | Nominal | No |  |
| Color | Color (Blue, Red, Grey, Silver, Black, etc.) | Character | Nominal | No |  |
| Automatic | Automatic ( (Yes=1, No=0) | Character | Nominal | No |  |
| CC | Cylinder Volume in cubic centimeters | Numeric | Continuous | No |  |
| Doors | Number of doors | Numeric | Nominal | No | Nominal - Because it takes finite number of values. |
| Cylinders | Number of cylinders | Numeric | Nominal | No | Nominal - Because it takes finite number of values. |
| Gears | Number of gear positions | Numeric | Nominal | No | Nominal - Because it takes finite number of values. |
| Quarterly\_Tax | Quarterly road tax in EUROs | Numeric | Continuous | No |  |
| Weight | Weight in Kilograms | Numeric | Continuous | No |  |
| Mfr\_Guarantee | Within Manufacturer's Guarantee period (Yes=1, No=0) | Numeric | Nominal | No | Nominal - Because it takes finite number of values. |
| BOVAG\_Guarantee | BOVAG (Dutch dealer network) Guarantee (Yes=1, No=0) | Numeric | Nominal | No | Nominal - Because it takes finite number of values. |
| Guarantee\_Period | Guarantee period in months | Numeric | Nominal | No | Nominal - Because it takes finite number of values. |
| ABS | Anti-Lock Brake System (Yes=1, No=0) | Numeric | Nominal | No | Nominal - Because it takes finite number of values. |
| Airbag\_1 | Driver Airbag (Yes=1, No=0) | Numeric | Nominal | No | Nominal - Because it takes finite number of values. |
| Airbag\_2 | Passenger Airbag (Yes=1, No=0) | Numeric | Nominal | No | Nominal - Because it takes finite number of values. |
| Airco | Airconditioning (Yes=1, No=0) | Numeric | Nominal | No | Nominal - Because it takes finite number of values. |
| Automatic\_airco | Automatic Airconditioning (Yes=1, No=0) | Numeric | Nominal | No | Nominal - Because it takes finite number of values. |
| Boardcomputer | Board computer (Yes=1, No=0) | Numeric | Nominal | No | Nominal - Because it takes finite number of values. |
| CD\_Player | CD Player (Yes=1, No=0) | Numeric | Nominal | No | Nominal - Because it takes finite number of values. |
| Central\_Lock | Central Lock (Yes=1, No=0) | Numeric | Nominal | No | Nominal - Because it takes finite number of values. |
| Powered\_Windows | Powered Windows (Yes=1, No=0) | Numeric | Nominal | No | Nominal - Because it takes finite number of values. |
| Power\_Steering | Power Steering (Yes=1, No=0) | Numeric | Nominal | No | Nominal - Because it takes finite number of values. |
| Radio | Radio (Yes=1, No=0) | Numeric | Nominal | No | Nominal - Because it takes finite number of values. |
| Mistlamps | Mistlamps (Yes=1, No=0) | Numeric | Nominal | No | Nominal - Because it takes finite number of values. |
| Sport\_Model | Sport Model (Yes=1, No=0) | Numeric | Nominal | No | Nominal - Because it takes finite number of values. |
| Backseat\_Divider | Backseat Divider (Yes=1, No=0) | Numeric | Nominal | No | Nominal - Because it takes finite number of values. |
| Metallic\_Rim | Metallic Rim (Yes=1, No=0) | Numeric | Nominal | No | Nominal - Because it takes finite number of values. |
| Radio\_cassette | Radio Cassette (Yes=1, No=0) | Numeric | Nominal | No | Nominal - Because it takes finite number of values. |
| Parking\_Assistant | Parking assistance system (Yes=1, No=0) | Numeric | Nominal | No | Nominal - Because it takes finite number of values. |
| Tow\_Bar | Tow Bar (Yes=1, No=0) | Numeric | Nominal | No | Nominal - Because it takes finite number of values. |

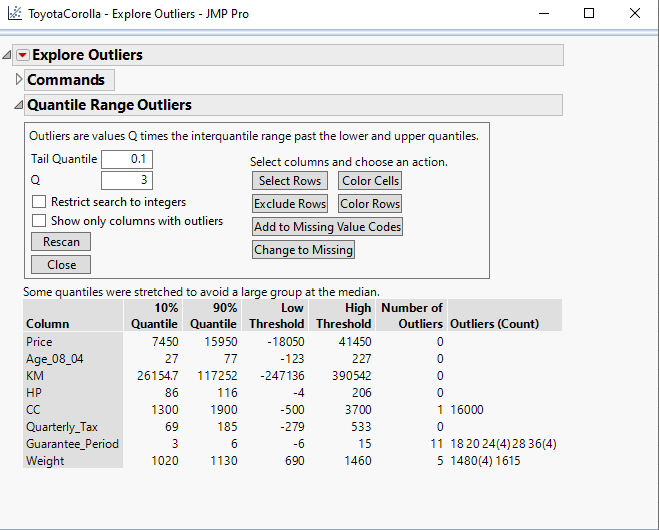
**1.  Review the variable types for each column.  Make sure that each type is correct.  Document what you have changed and why.**

The fuel types – Petrol, Diesel and CNG (dummy variables) needs to be changed from continuous to nominal as they are indicator variables.

**2.  Exclude any outliers and explain the justification for exclusion.**

**Steps to check for Outliers:**

1. Analyze -> Screening -> Explore Outliers -> Quantile Range Outliers

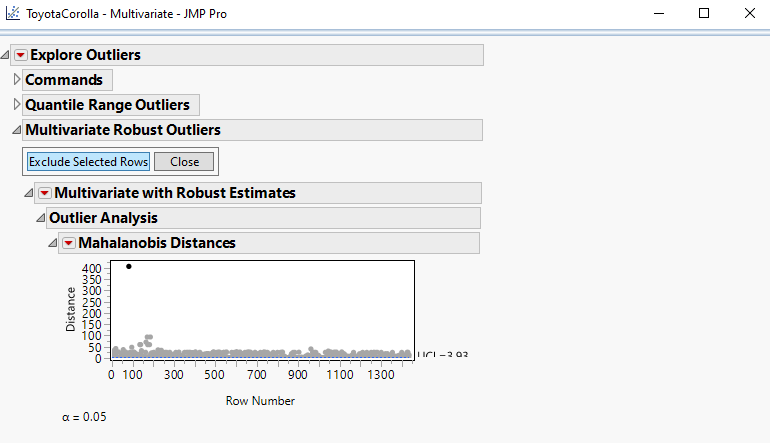


CC -> 16000 is listed as an outlier

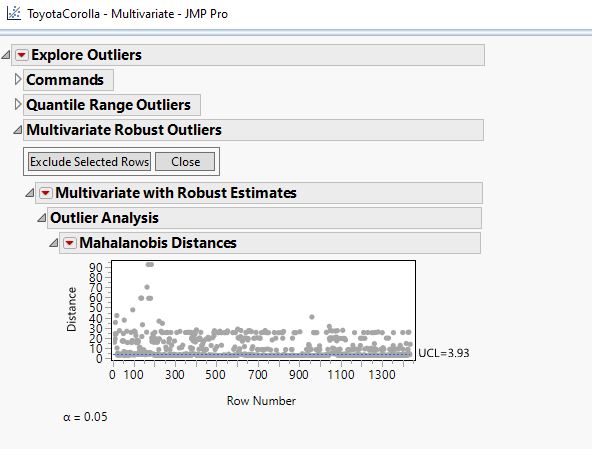
Guarantee Period -> 18,20,24,28,36 is listed as an outlier.

Weight - > 1480, 1615 is listed as an outlier.

1. Analyze -> Screening -> Explore Outliers -> Multivariate Robust Outliers

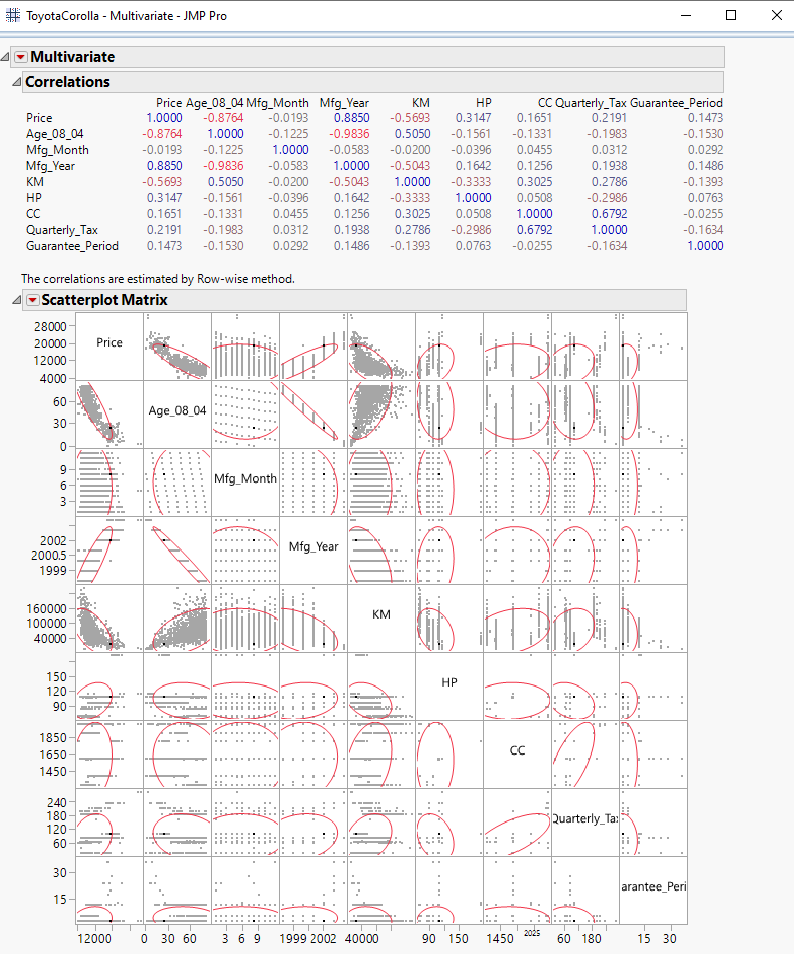


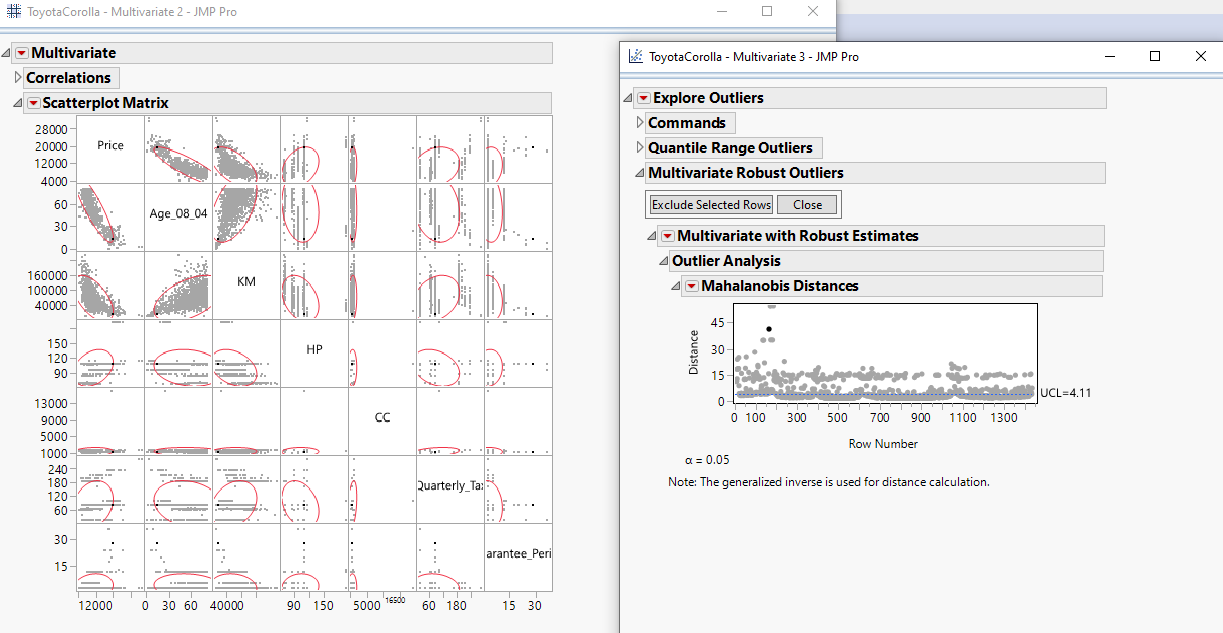
After Excluding rows (The outlier was the row where CC was 16000) and clicking Re- Run,



The plot contains an upper control limit (UCL) of 3.93. The UCL is meant to be a helpful guide to show where potential outliers might be.

1. Analyze -> Multivariate Methods -> Multivariate

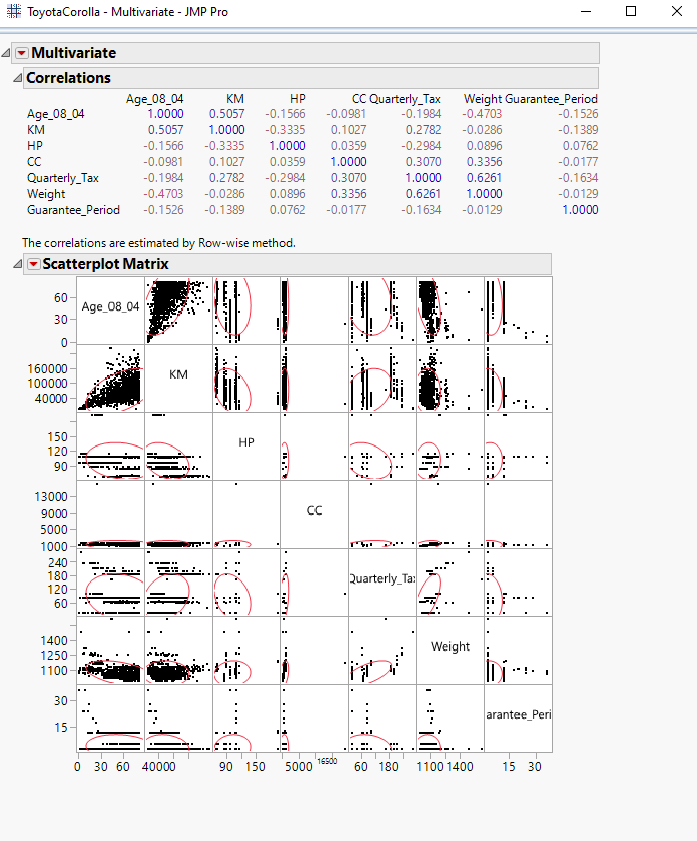




|  |  |  |  |
| --- | --- | --- | --- |
| **Column** | **Outlier Value** | **Outlier Count** | **Exclusion (Yes/No)** |
| CC | 16000 | 1 | Yes. |
| Weight | 1615  1480 | 1  4 | No.  The data points fall inside the ellipses. (Any point outside the ellipses is a potential outlier) |
| Guarantee Period | 18  20  24  28  36 | 1  1  4  1  4 | No.  It is a categorical variable and we cannot differentiate it as good/bad values. |

On comparing the key plots between Multivariate Robust Outliers and Multivariate Screening we conclude that only 16000 is an obvious outlier and can be excluded from the analysis.

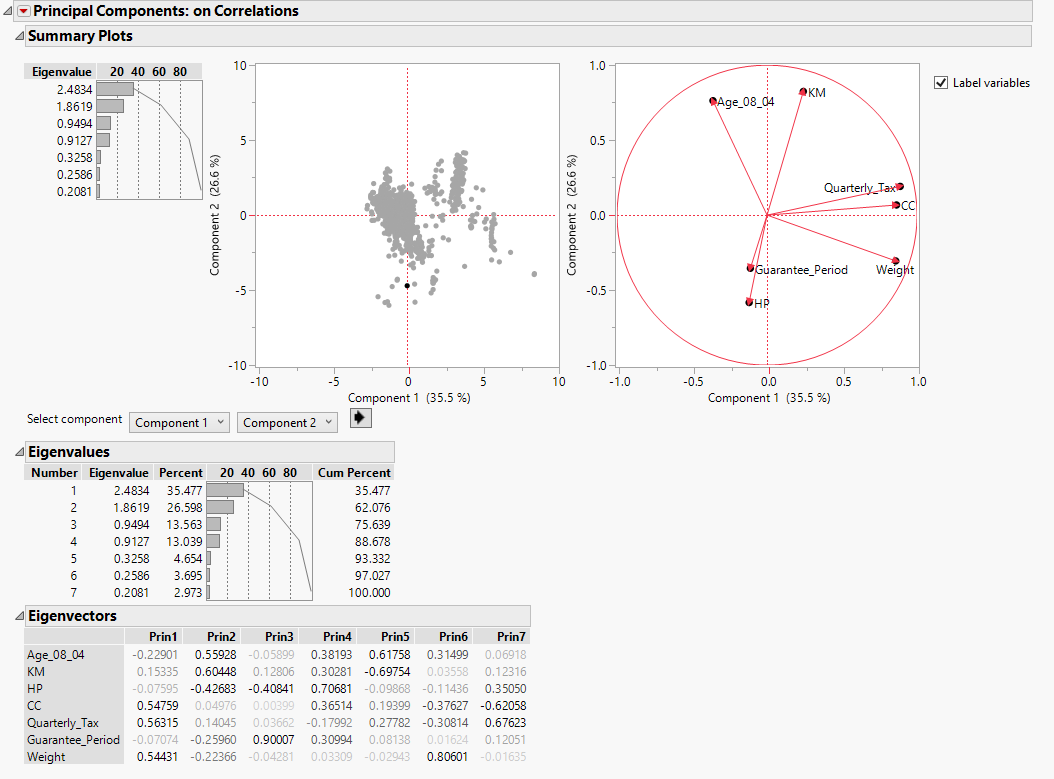
**3.  Decide which continuous variables you would use to predict Price.  For each continuous variable list why or why not you will use it.**



|  |  |  |
| --- | --- | --- |
| **Variable** | **Would you use to predict?** | **Reason** |
| Age\_08\_04 | Yes | There is a strong correlation with Price between these two variables. |
| KM | Depends. | There is a moderate correlation with Price between these two variables. |
| HP | No | There is a weak correlation with Price between these two variables. |
| CC | No | There is a weak correlation with Price between these two variables. |
| Quarterly\_Tax | No | There is a weak correlation with Price between these two variables. |
| Weight | Depends. | There is a moderate correlation with Price between these two variables. |

Therefore, one might use Age\_08\_04, KM and Weight to predict the price.

**4.  Conduct a principal components analysis on the continuous predictor variables used to predict price.  Comment on the results.  Include a screen shot of the Eigenvectors and Eigenvalues (with cumulative percent).  How much variability can you maintain if you just keep 2 principal components? What variables contribute the most information to PC1 and PC2?**

z

Principal Component Analysis is highly helpful for Dimension Reduction. It simplifies the complexity in high-dimensional data while retaining trends and patterns. In the above summary plots, we observe that we get 100% of the information from the 7 variables.

The variability maintained after keeping just two principal components is 2.4834 + 1.8619 = 4.3453

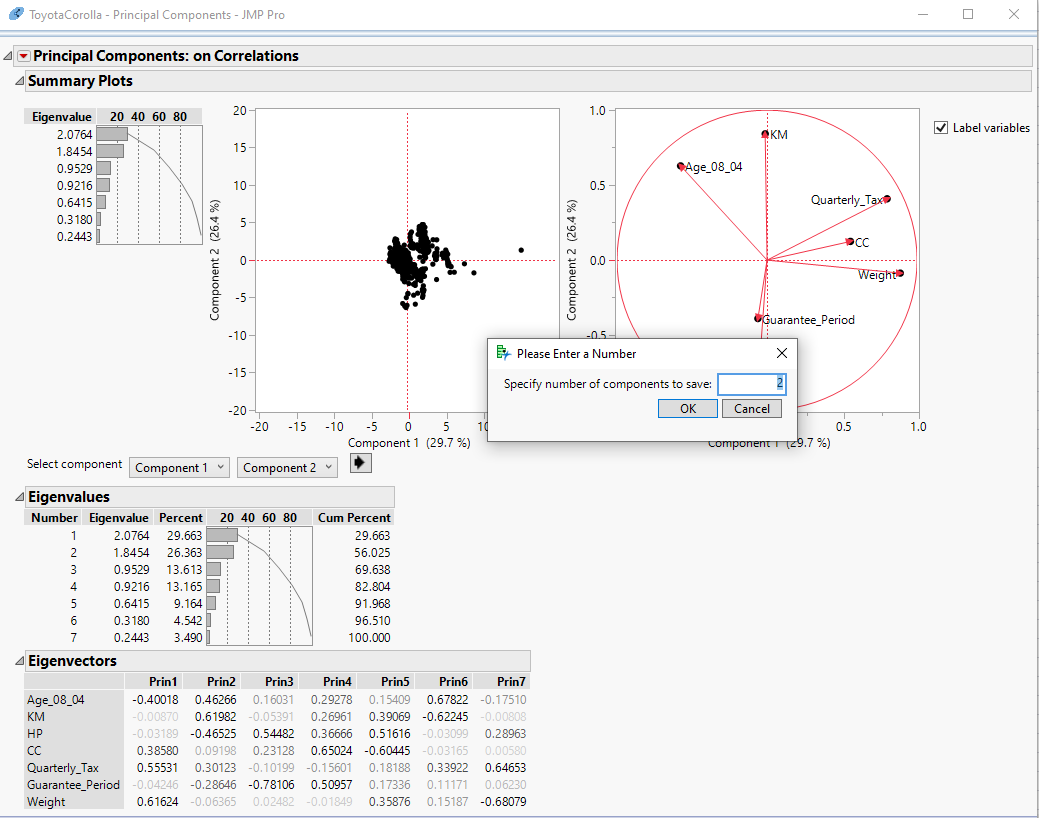
Total cumulative percent with two principal components would be 62.076%.

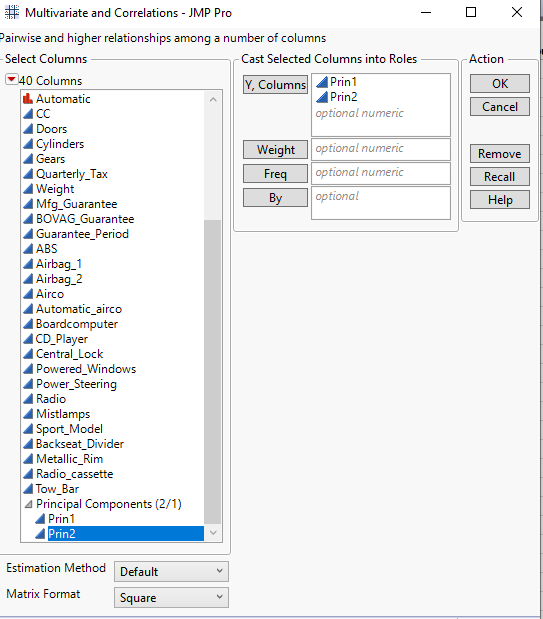
The variables that contribute the most information to

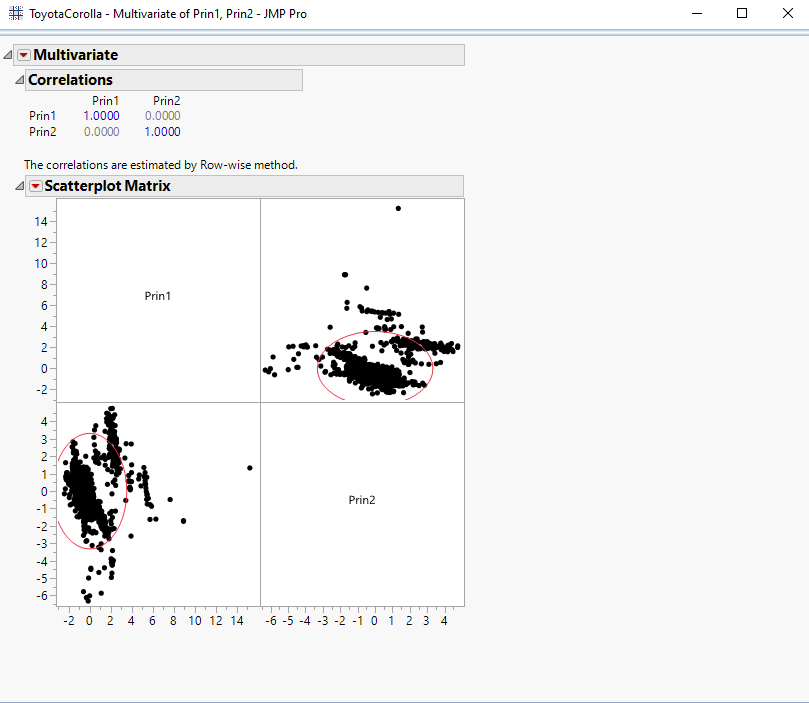
PC1 – Weight, Quarterly\_Tax and Age\_08\_04

PC2 – KM, HP, Age\_08\_04

**5.  Save the first two principal components to the data table.  Create a scatterplot matrix of the two principal components and determine their correlation.  Explain in your own words why the correlations is what it is.**







Thereis no correlation between the 2 principal components as they are perpendicular to each other with no overlap. Hence, the correlation coefficient is zero i.e. **r =0**.