# Chapter -1:

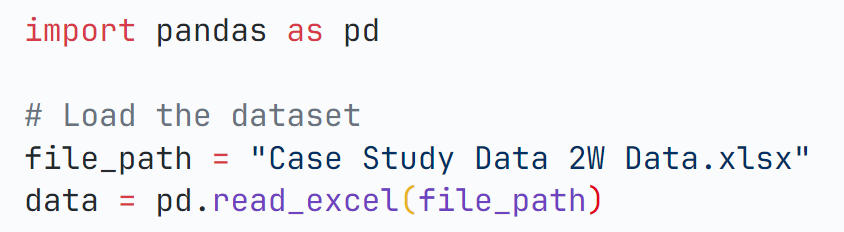
# Implementation

1.1Problem Formulation:

The primary objective of this case study is to develop a comprehensive analysis and provide next steps for a predictive model that accurately forecasts the residual value (RV) of used 2W vehicles. The case study will involve creating and testing hypotheses, conducting statistical analysis, and performing exploratory data analysis (EDA) using the provided dataset.

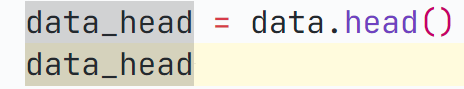
1.2Data Loading:

The `Case Study Data 2W Data.xlsx` dataset was loaded into a pandas Data Frame using read\_excel method.

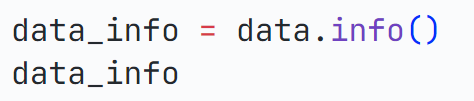


1.3Summary:

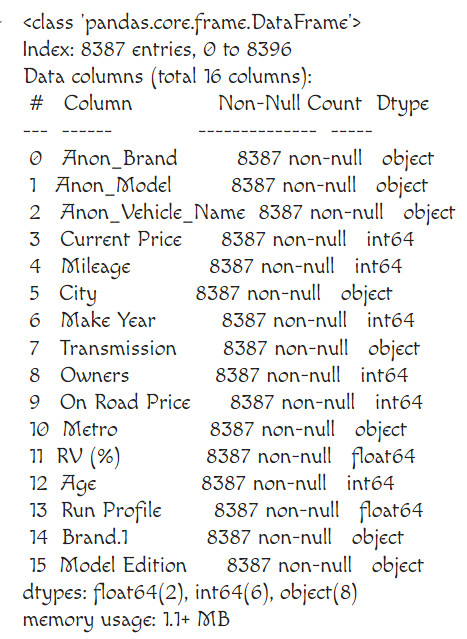
First 5 records have been printed using head method.



Summary of number of rows, Name of the columns, non-null values in each column as well as data type in each column is provided.

****

We have found out that total number of records is 8397 and the data type of each column is given in the code.



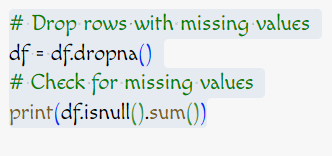
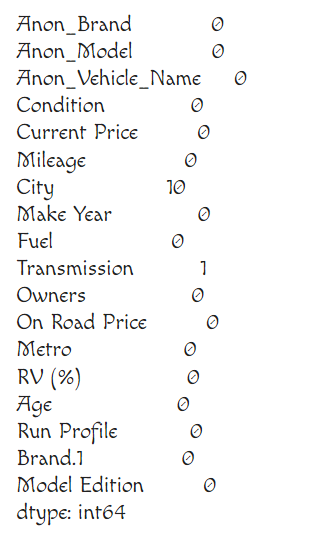
**1.4 Statistics**:

Number of non-null values, mean, standard deviation, minimum value, maximum value has been provided.

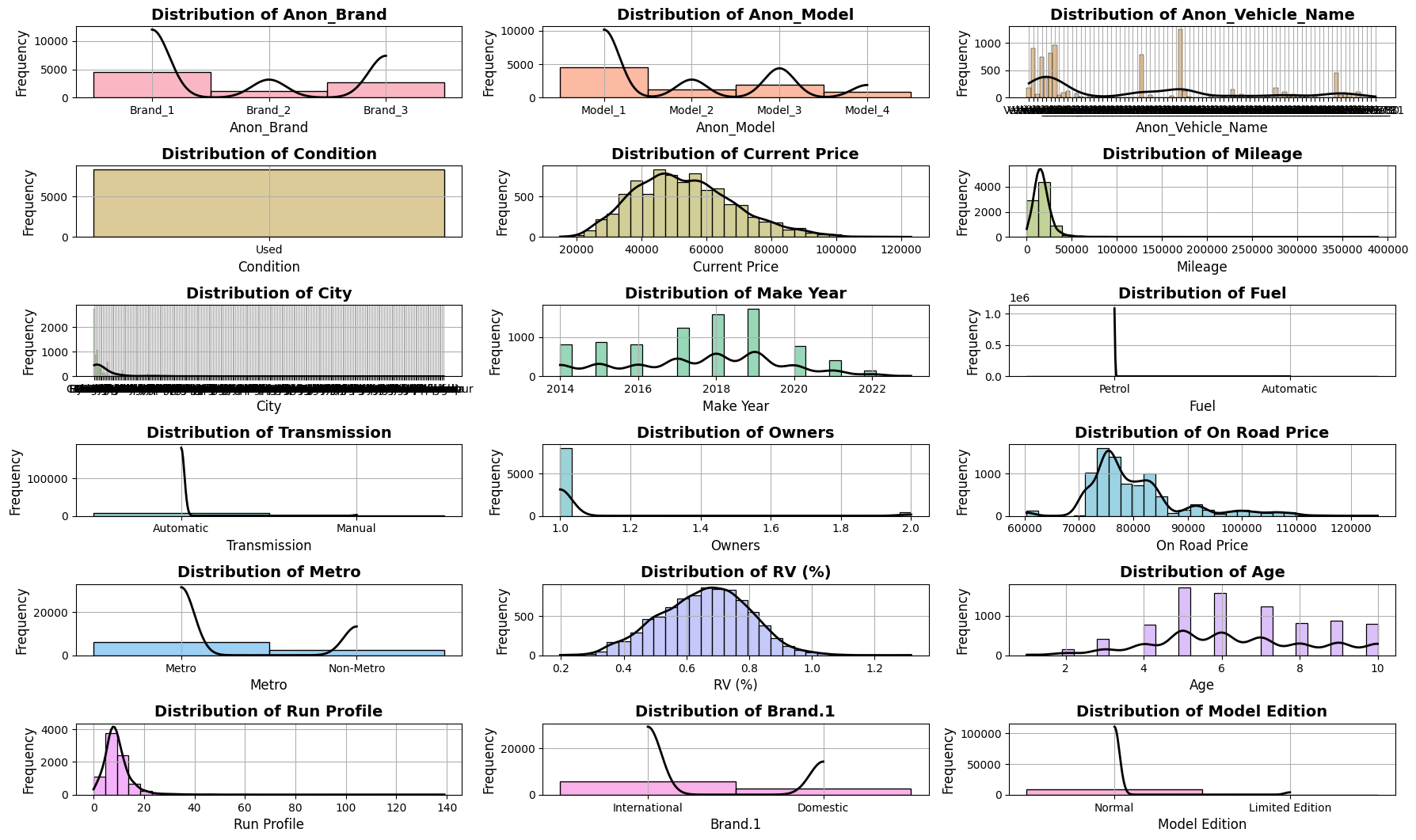
****

**1.5 Missing Values:**

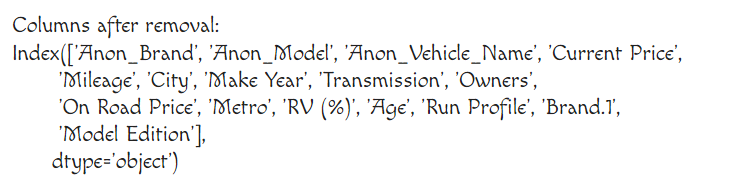
* Here Transmission column as 1 missing values and city column has 10 missing values.
* In the next step we have dropped the records with missing values as it is not that important

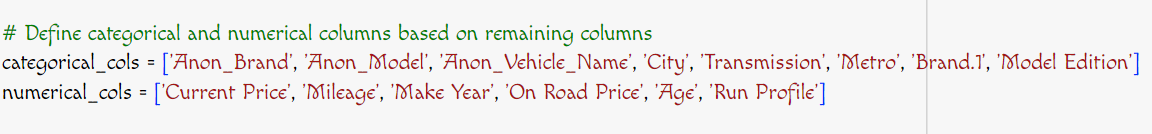
** 1.6 Distribution of features**:

Below given curves depicts the distribution of various features and the frequency of occurrences.

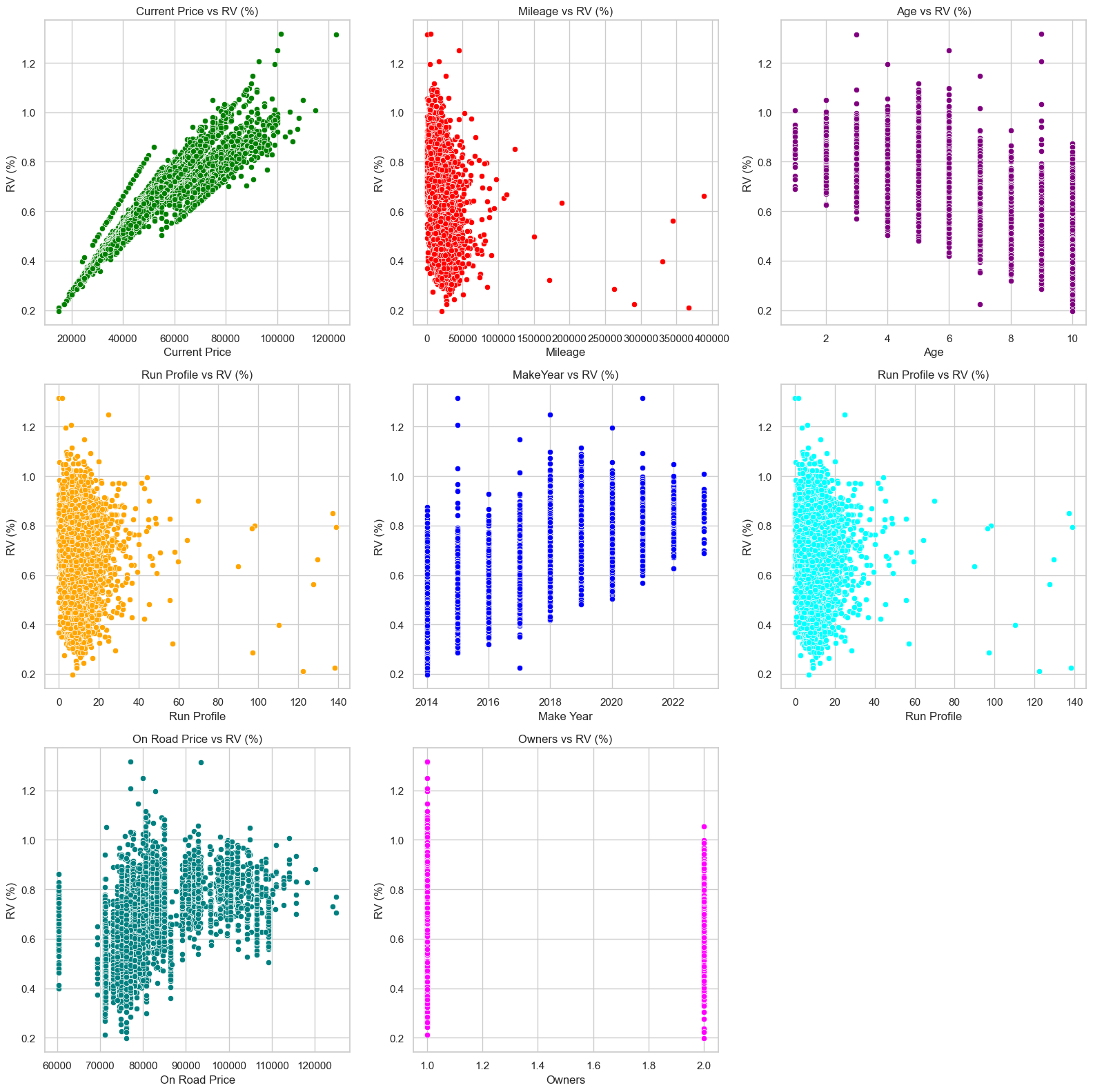


As all vehicles have been used so condition column can be dropped. Also, in the fuel column as number of vehicles with automatic fueling is 3 it can be dropped.

****

****

In the above step categorical and numerical columns have been segregated for future processing.

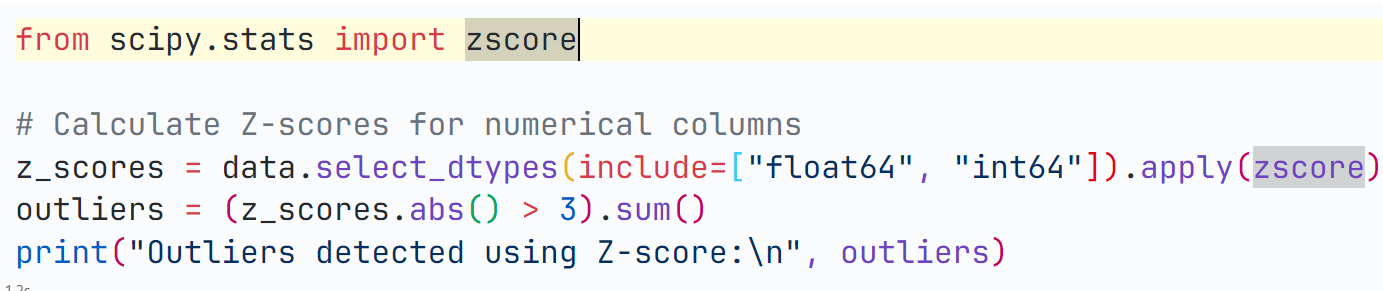
**1.7** Plot of Features with RV:

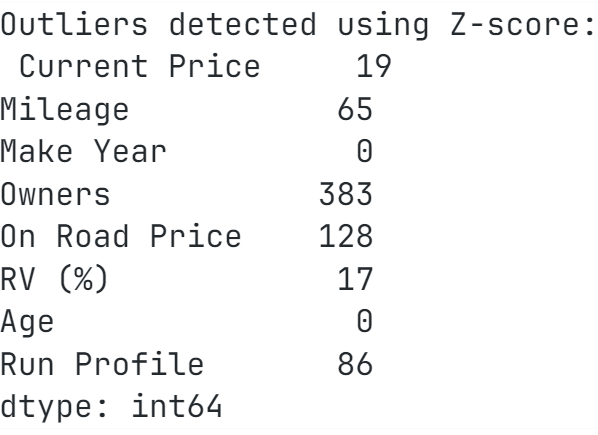
* In the above figure plots of RV with various features have been plotted. We can see Current Price showing a strong positive correlation with RV.
* Mileage, Run Profile vs RV (%): Shows a weak negative correlation, suggesting that higher mileage and Run Profile might be associated with slightly lower RV (%).
* Age shows somehow negative correlation indicating older vehicles generally have lower RV values.
* Make Year shows somehow positive correlation indicating newer vehicles might have slightly higher RV values.

**1.8 Outlier Detection**:

Outlier detection is a crucial step in data preprocessing, aimed at identifying data points that deviate significantly from the overall distribution of the data

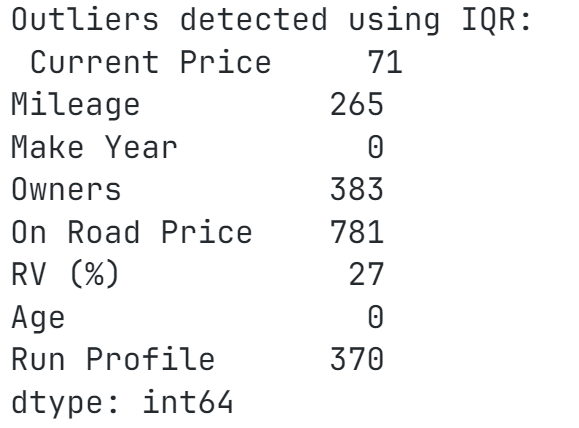
**Z-Score:** Measures how far a data point is from the mean in terms of standard deviations. Data points with a Z-score greater than a certain threshold are considered outliers.



****

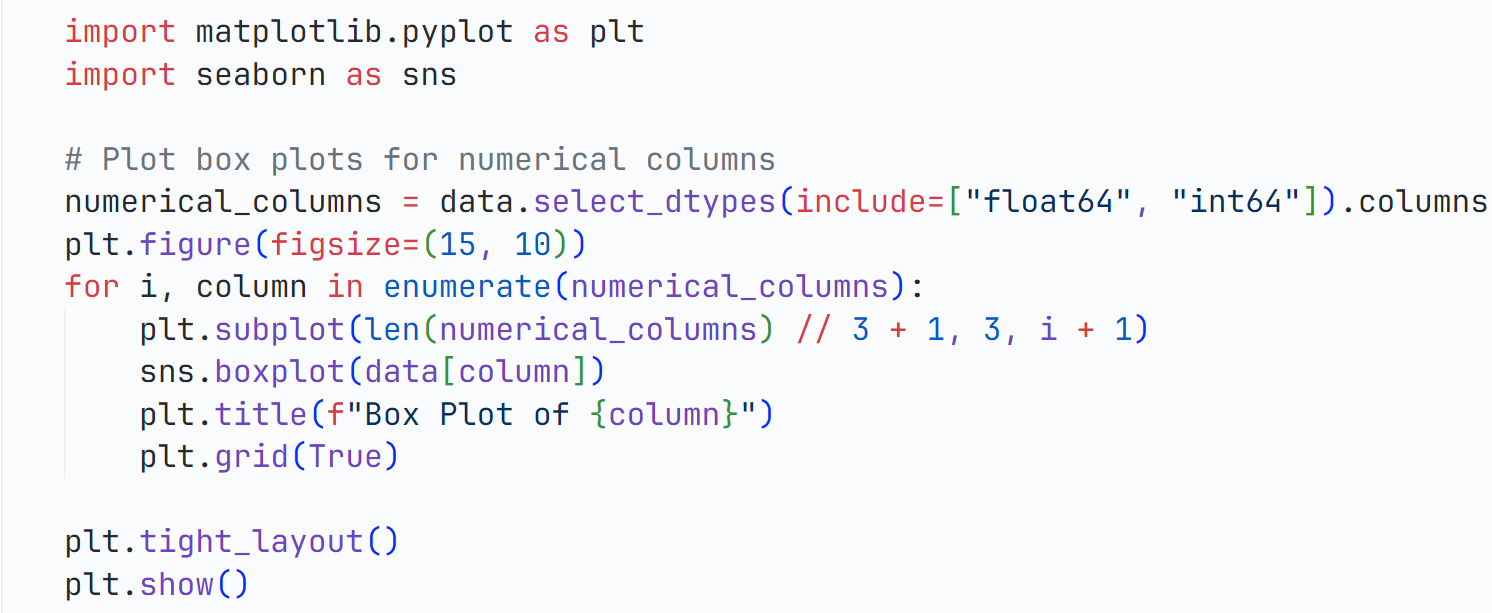
IQR (Interquartile Range):

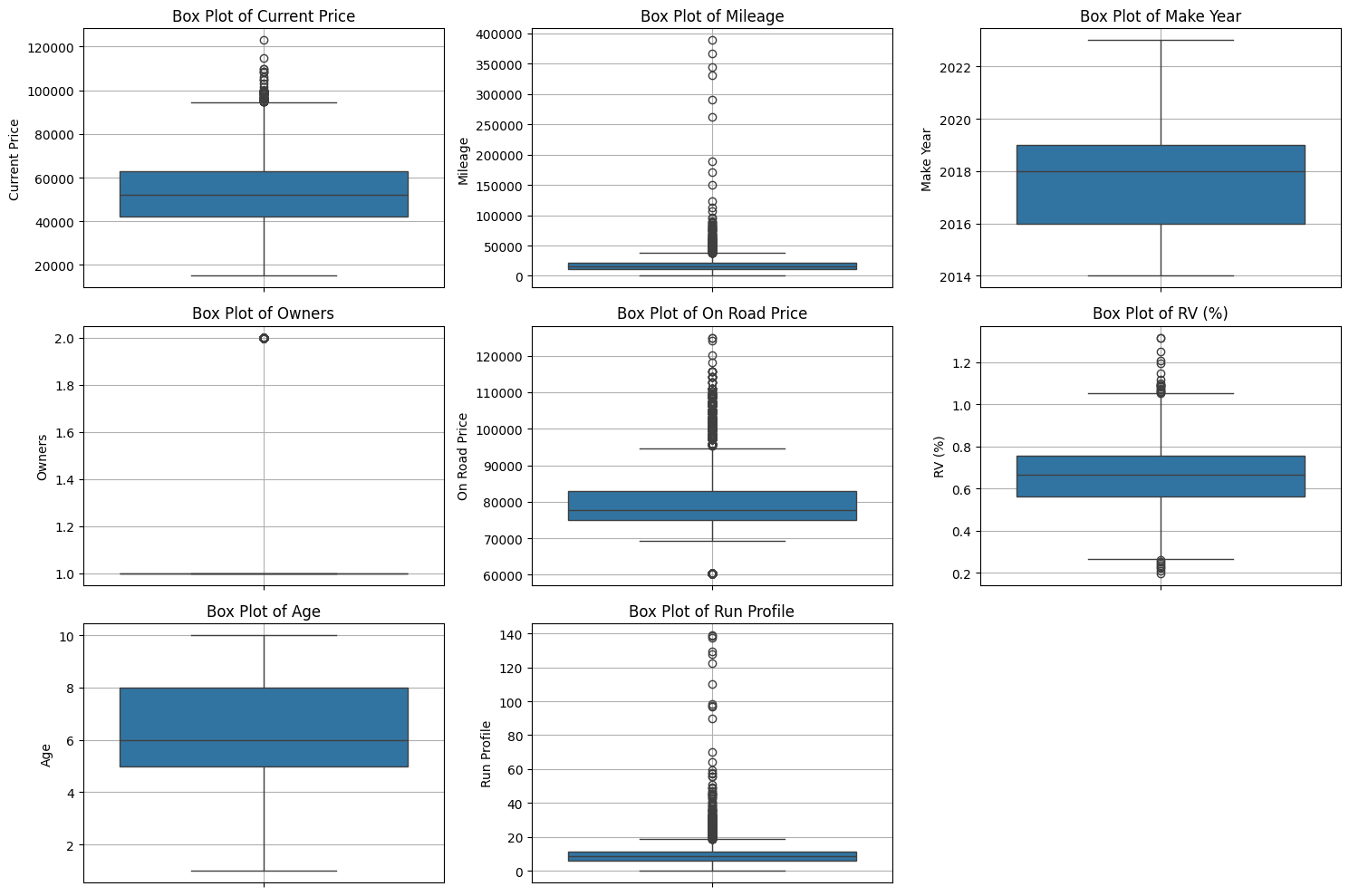
Measures the spread of the middle 50% of the data. Outliers are those that lie outside the range defined by Q1 - 1.5\*IQR and Q3 + 1.5\*IQR.



Box Plot:

A box plot (or box-and-whisker plot) is a great visualization tool for identifying outliers and understanding the distribution of numerical data.

****

****

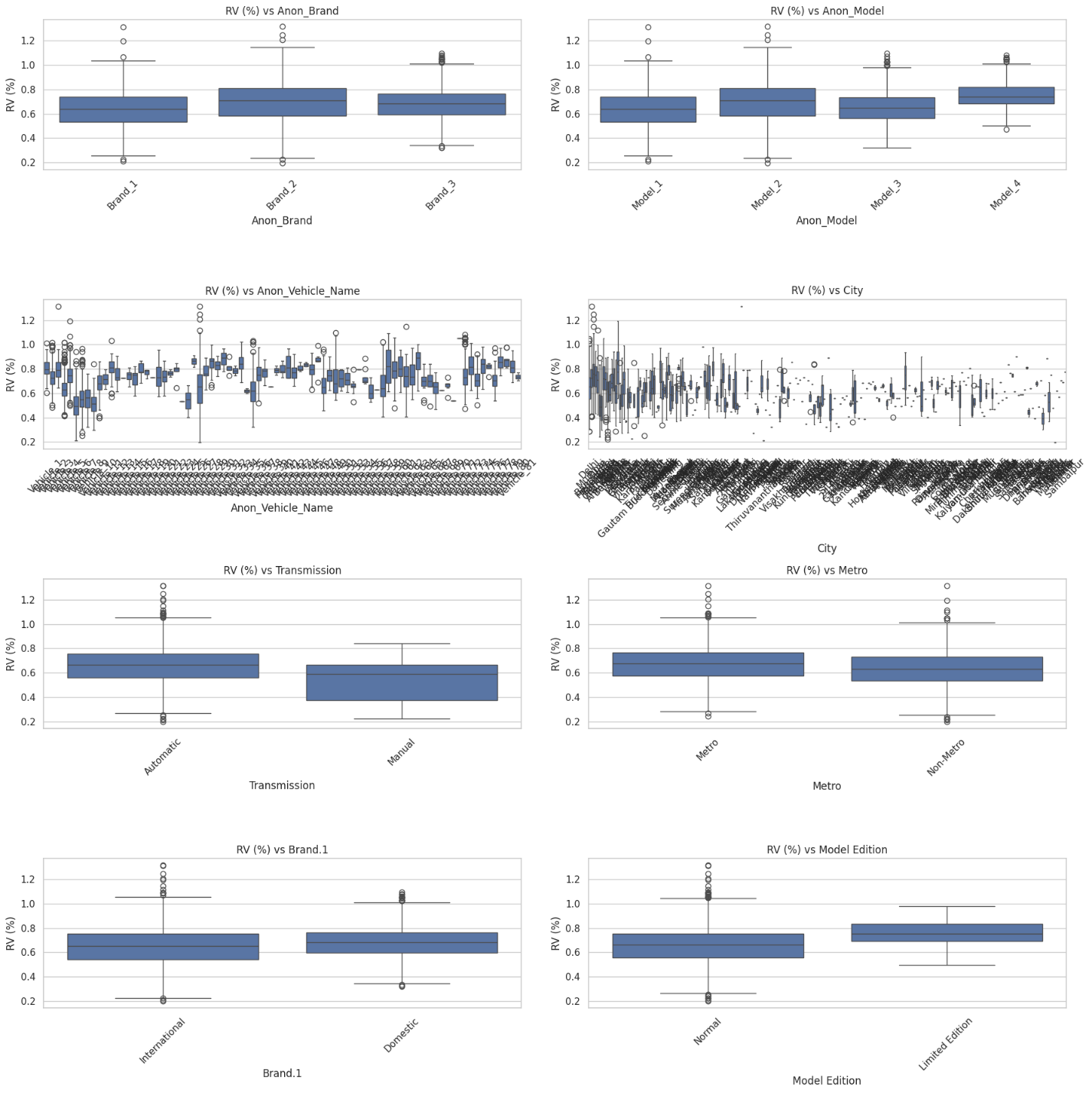
Above plot shows current price, mileage, RV, Run Profile and On road price has outliers but we do not need preprocessing as this values are important.

**1.9 Plot of RV vs Categorical columns:**

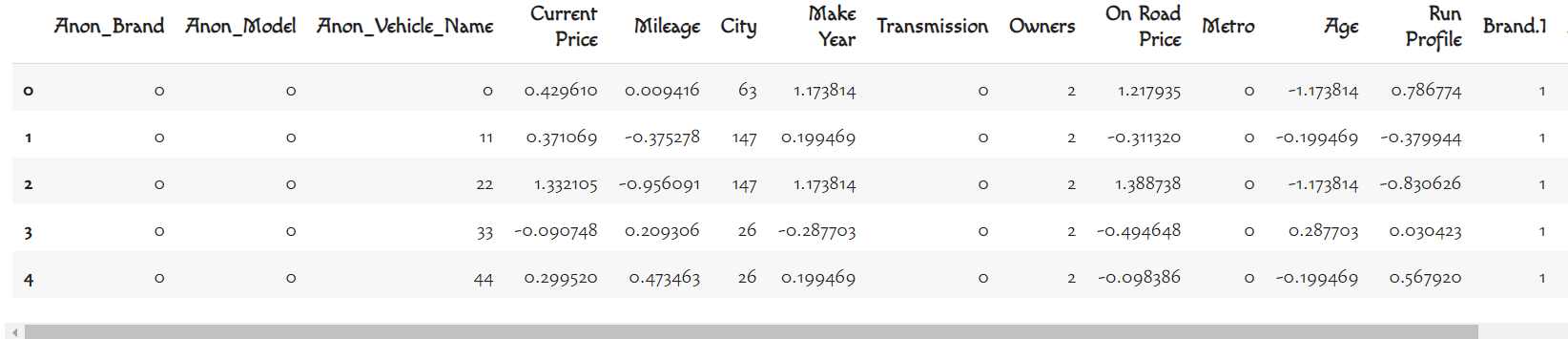
1. For Anon\_Brand: The median RV (%) is similar across different brands, indicating that the typical RV value is roughly the same for most brands.

**Brand 3** seems to have a slightly higher typical RV value and a slightly **wider spread** of data compared to the other brands.

1. **Anon Vehicle Name & City:** The box plots for Anon Vehicle Name as well as city exhibit a wider range of RV (%) values, suggesting that the vehicle name and city name might be a more influential factor.
2. **Transmission:** The box plots for Transmission suggest that automatic transmission vehicles might have slightly higher RV (%) values compared to manual transmission vehicles.
3. **Metro:** The box plots for Metro indicate that vehicles in metropolitan areas might have slightly higher RV (%) values compared to those in non-metropolitan areas.
4. **Model Edition:** The box plots for Model Edition suggest that limited edition models might have slightly higher RV (%) values compared to standard or normal models.



* 1. **Label Encoding & Scaling:**
* **Label encoding** Convert categorical variables into a numerical format that machine learning algorithms can process. This is useful when the categorical variable is **ordinal** (i.e., the order of the categories is meaningful) or when the machine learning model can handle the integer-based categories effectively.
* **Scaling** normalizes numerical features to have a mean of 0 and a standard deviation of 1.

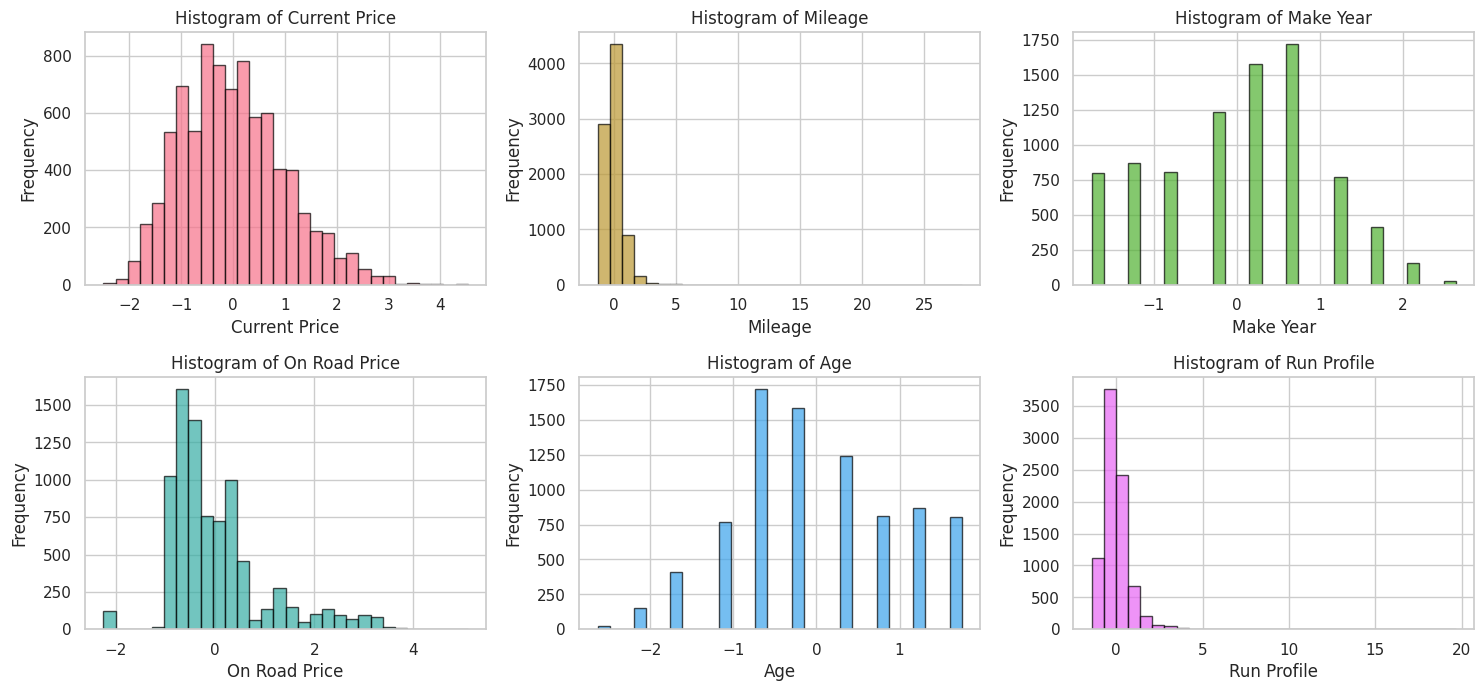
****

**1.11 EDA:**

Below given table is the statistics of given dataset after label encoding and scaling the dataset.

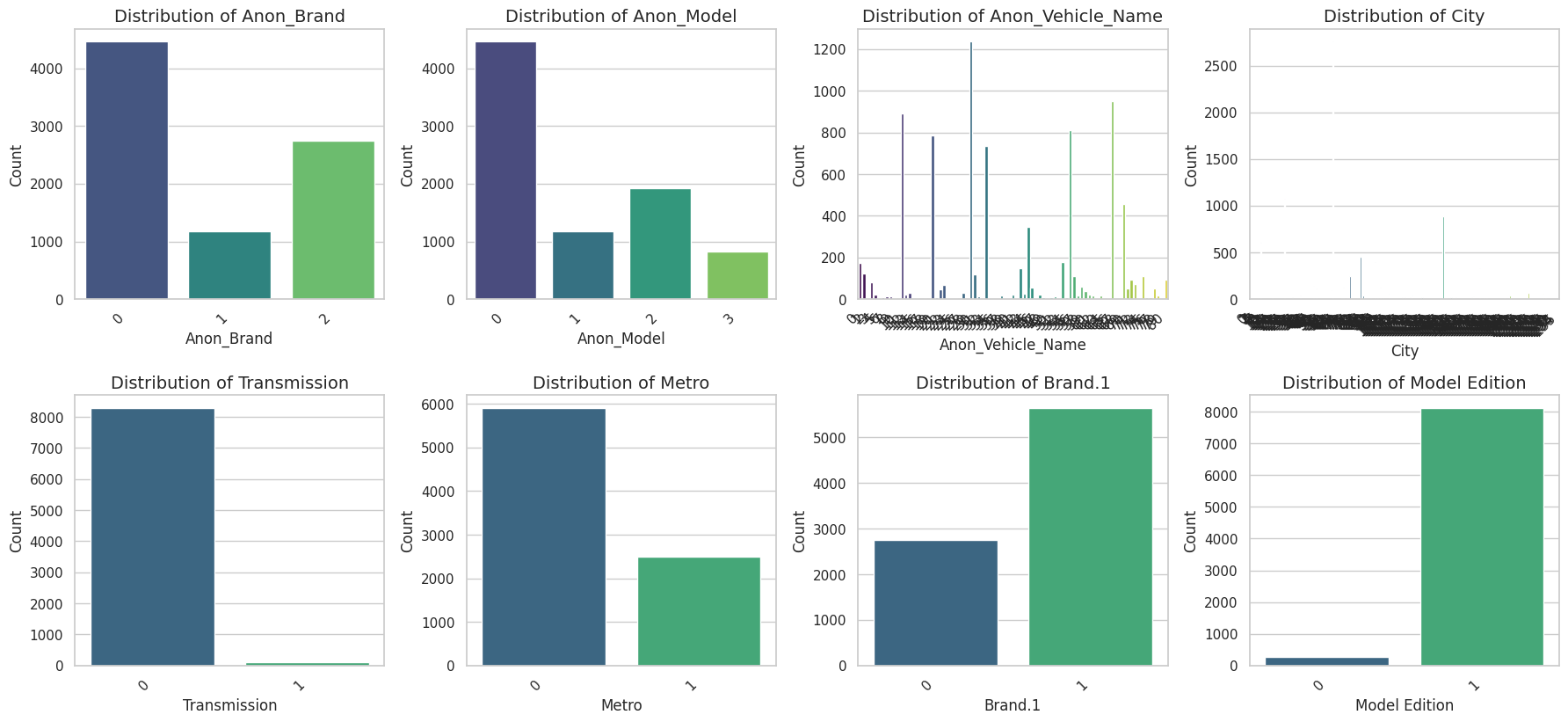


**1.12 Histogram plot of Features:**



Above plot depicts the frequency of occurrence of each features in a histogram plot.

**1.13 Plot of Categorical Features:**



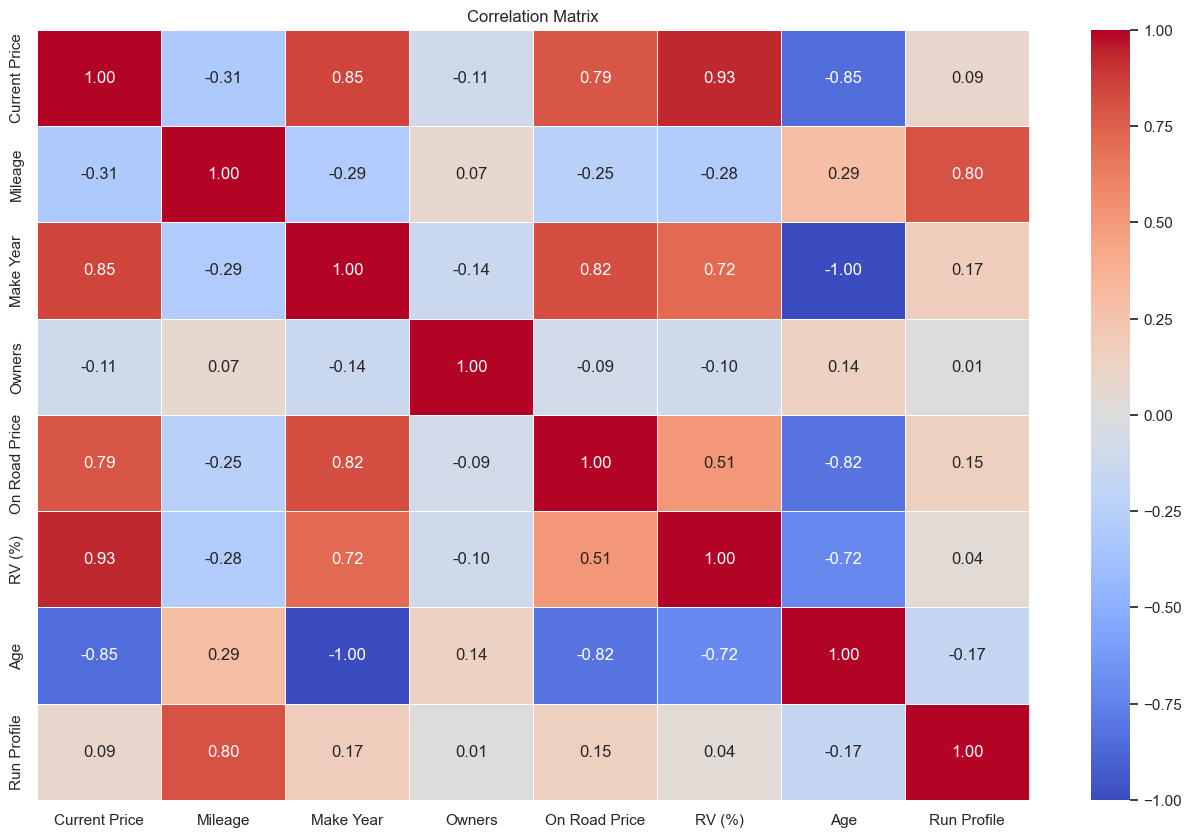
**1.14 Correlation Matrix**:

A correlation matrix is a table that shows the correlation coefficients between multiple variables in a dataset. Each cell in the matrix represents the correlation between two variables. The values range from -1 to 1, where:

* +1 indicates a perfect positive correlation: as one variable increases, the other increases.
* -1 indicates a perfect negative correlation: as one variable increases, the other decreases.
* 0 indicates no correlation: the variables do not have a linear relationship.



From the above table we can calculate the correlation between various input feature but as our target is correlation between each feature and Residual Vehicle (RV). We can select the features that we want based upon the correlation and we can observe that for RV and owners correlation is not much so we can ignore owner column as well as Run Profile feature.

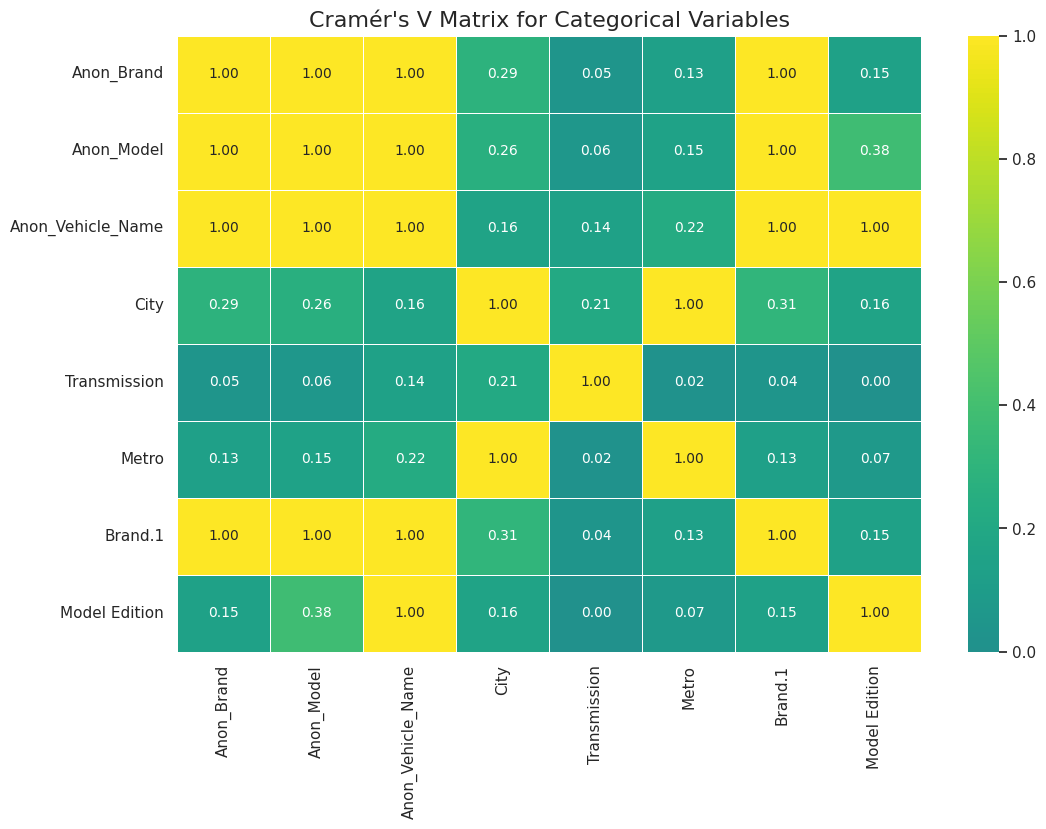
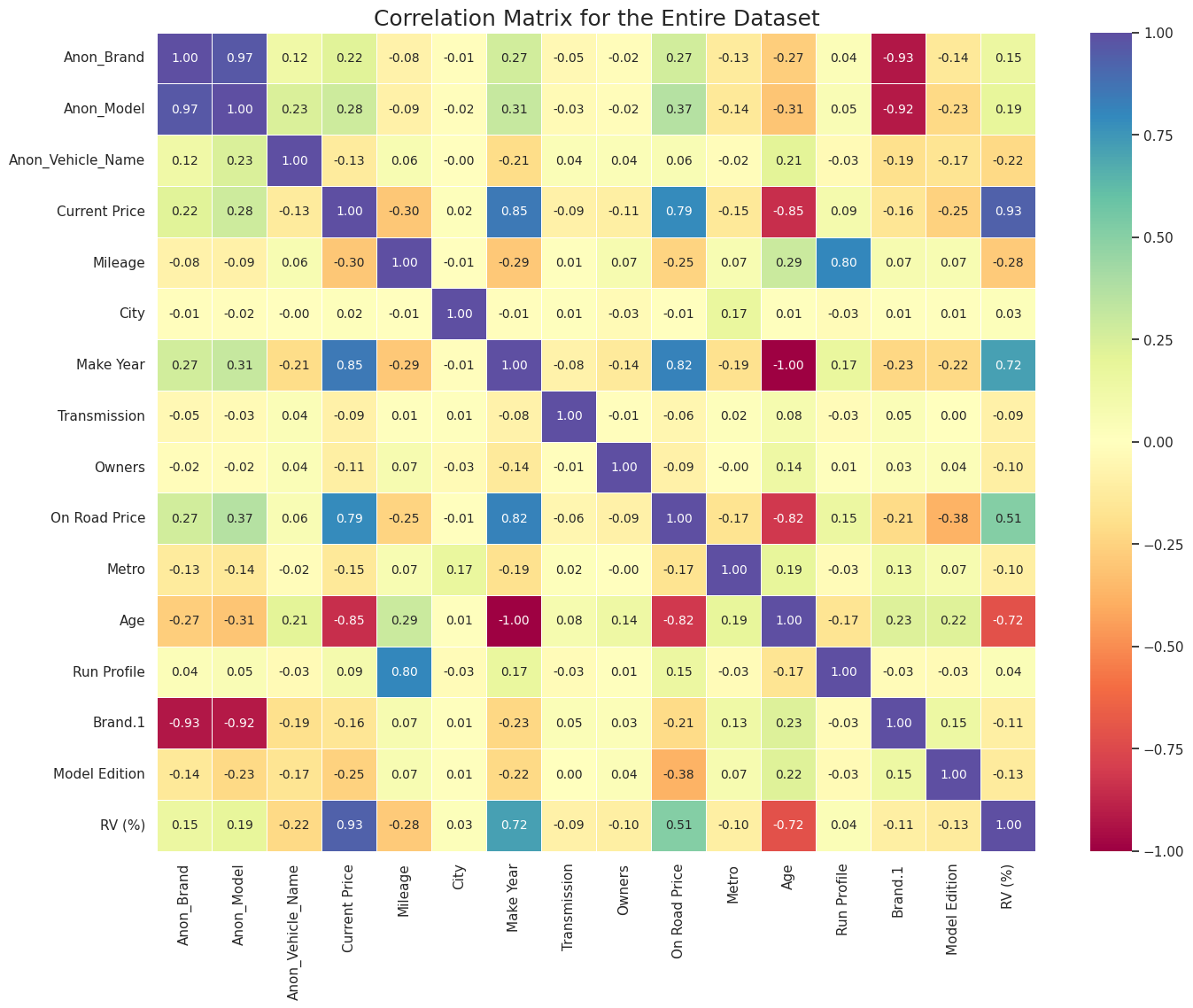


1.15 Crammer’s Matrix & Entire Correlation Matrix:

Cramér's V is a measure of association between two nominal (categorical) variables. It is a way to quantify the strength of the relationship between variables in a contingency table (cross-tabulation).

Cramér's V ranges from 0 to 1.

* A value of 0 indicates no association between the variables.
* A value of 1 indicates a perfect association.
* Generally, value from 0.7-1 are called as strong association & 0.3-0.7 are called moderate while others are called as weak.

****

1.16 Hypothesis Testing:

Hypothesis testing is a statistical method used to make inferences or draw conclusions about a population based on sample data. It involves formulating a hypothesis, then using statistical techniques to test whether the hypothesis is likely to be true or not.

* **Null Hypothesis (H₀):**

There is no linear relationship between the two variables.

* **Alternative Hypothesis (H₁ or Ha):**

It is the opposite of the null hypothesis.

* **Test Statistic:**

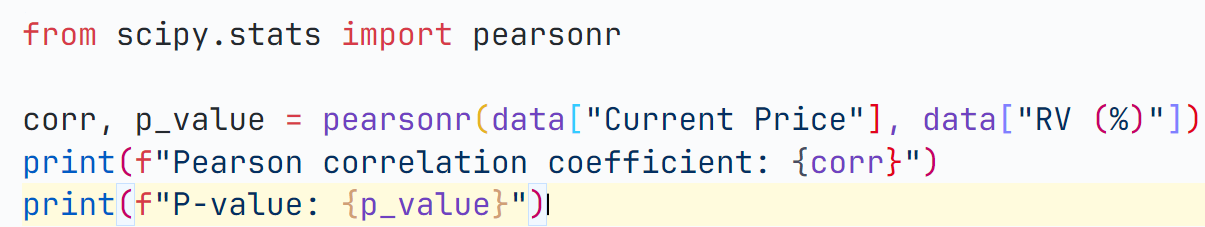
A value calculated from the sample data that is compared against a distribution to determine whether to reject H₀.

Examples include the t-statistic, z-statistic, F-statistic, etc.

**P-Value:**

* The probability of obtaining test results under the assumption that the null hypothesis is true.
* If the p-value is less than α, we can reject the null hypothesis.
* Generally alpha value is 0.05

**Sample Code:**

In the below table Pearson correlation coefficient and p values have been given. **Pearson correlation coefficient** (often denoted as **r**) is a measure of the strength and direction of the linear relationship between two continuous variables analogous to correlation.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Current Price | Mileage | Age | Run Profile | Make Year | On road Price |
| Pearson Coefficient | 0.9307 | -0.2827 | -0.7174 | 0.0408 | 0.7174 | 0.5119 |
| P value | 0.0 | 5.64e-150 | 0.0 | 0.0001 | 0.0 | 0.0 |

**Current Price (Pearson Coefficient: 0.9307, P-Value: 0.0):**

* **Inference:** There is a strong positive linear correlation between Current Price and RV (%). The p-value of 0.0 indicates that this correlation is statistically significant, meaning there is very strong evidence against the null hypothesis of no correlation.

**Mileage (Pearson Coefficient: -0.2838, P-Value: 2.2597):**

* **Inference:** There is a weak negative linear correlation between Mileage and RV (%). The p-value suggests that this result is not statistically significant, so we fail to reject the null hypothesis of no correlation between Mileage and RV (%).

**Age (Pearson Coefficient: -0.7174, P-Value: 0.0):**

* **Inference:** There is a strong negative linear correlation between Age and RV (%). The p-value of 0.0 indicates that this correlation is statistically significant.

**Run Profile (Pearson Coefficient: 0.0408, P-Value: 0.0001):**

* **Inference:** There is a very weak positive linear correlation between Run Profile and RV (%). Despite the small correlation, the p-value indicates that the relationship is statistically significant.

**Make Year (Pearson Coefficient: 0.7174, P-Value: 0.0):**

* **Inference:** There is a strong positive linear correlation between Make Year and RV (%). The p-value of 0.0 indicates a statistically significant relationship.

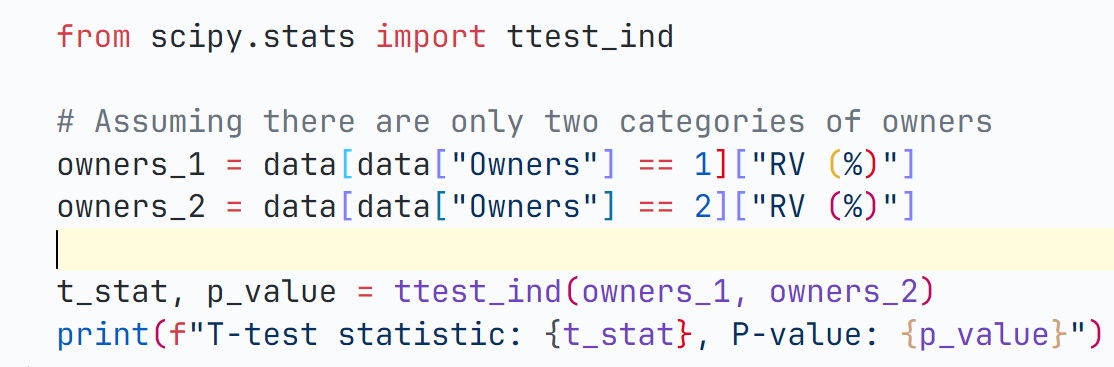
**On Road Price (Pearson Coefficient: 0.5119, P-Value: 0.0):**

* **Inference:** There is a moderate positive linear correlation between On Road Price and RV (%). The p-value indicates that this correlation is statistically significant

**T Test:**

T-tests are statistical tests used to determine if there are significant differences between the means of two groups or if a sample mean significantly differs from a known value.

* **One-Sample T-Test:** Compares the mean of a single group to a known value or population mean.
* **Independent Two-Sample T-Test:** Compares the means of two independent groups to see if they differ significantly.
* **Paired T-Test:** Compares means from the same group at different times (e.g., before and after a treatment) or under two different conditions.

Sample Code:

**ANOVA (Analysis of Variance)** is used to compare the means of three or more groups to determine if there are any statistically significant differences between them. The **F-statistic** is the test statistic used in ANOVA.

**Large F-Statistic:** Indicates that there is more variability between group means compared to within-group variability, suggesting that at least one group mean is significantly different from the others.

**P-Value:** The p-value associated with the F-statistic tells you if the observed F-statistic is significant. A p-value less than the chosen significance level (e.g., 0.05) means you reject the null hypothesis and conclude that there are significant differences between group means.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Brand | Model | Vehicle | Fuel | Transmission | Metro | Brand.1 | Edition |
| T-Test |  |  |  | 4.82 | 7.27 | 9.52 | -10.31 | -16.70 |
| p-value | 8.13 | 1.81 | 0 | 0.04 | 6.49 | 2.47 | 2.19 | 2.47 |
| H-value | 241.07 | 565.8 | 3699.85 |  |  |  |  |  |
| Anova | 131.16 | 197 |  |  |  |  |  |  |

Some model names have been changed as space in the table was shorter.

**T-Test & P-Values:**

A T-test was performed to check the difference in means for these categorical variables against RV (%).

Significant results (low p-values) suggest that the means of RV (%) differ significantly across these categories.

**H-Value (Kruskal-Wallis Test):**

The H-Value is used to compare medians across multiple groups. Higher values indicate a greater difference between the groups.

Significant H-values would suggest that there are differences in RV (%) across the categories.

**ANOVA:**

Used to compare means across more than two groups.

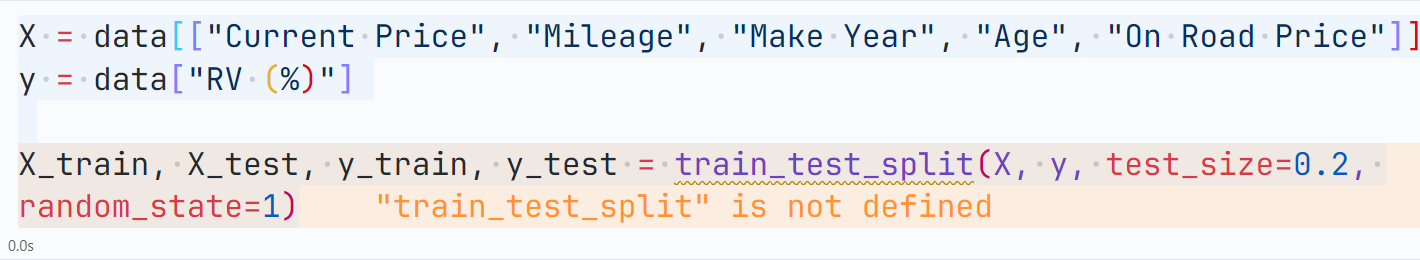
Significant F-statistics suggest that the means of RV (%) differ across the different groups of the categorical variables.

1.17 Model Building:

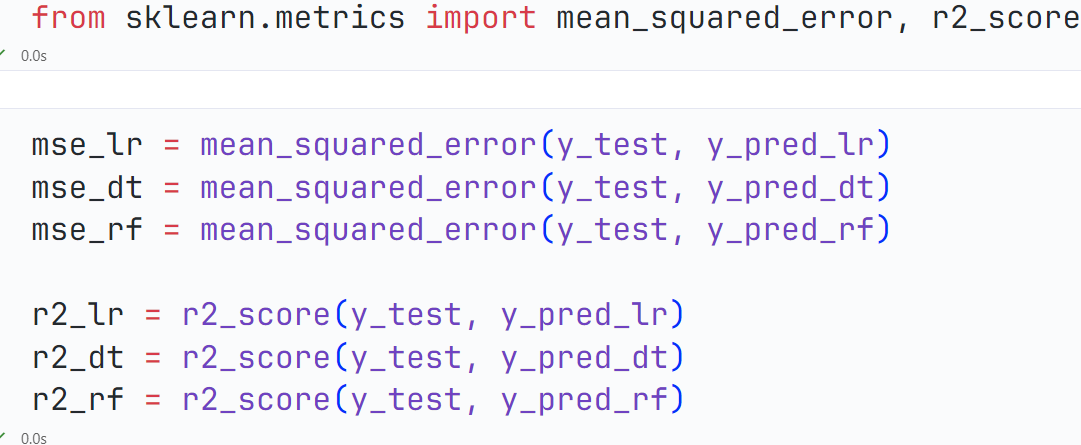
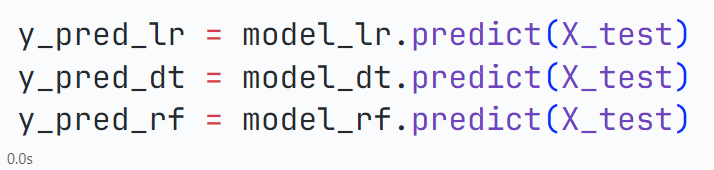
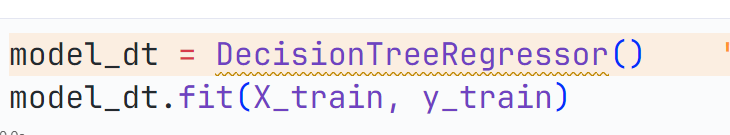
Linear regression, Decision tree and random forest regressor model have been evaluated.

Feature Selection:

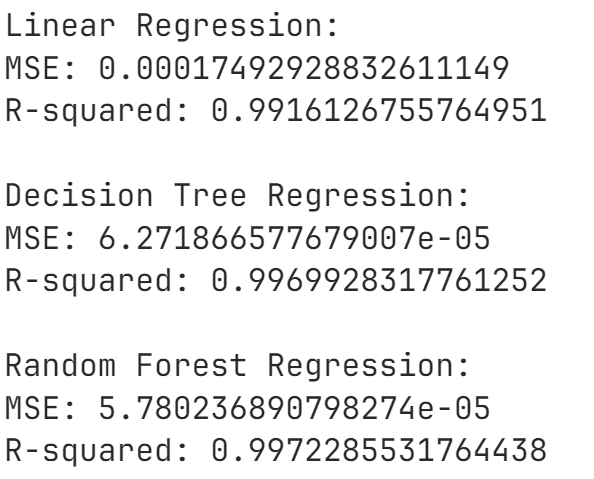
Separate the features (x) and target variable (y), where features are selected on the basis of correlation with RV.

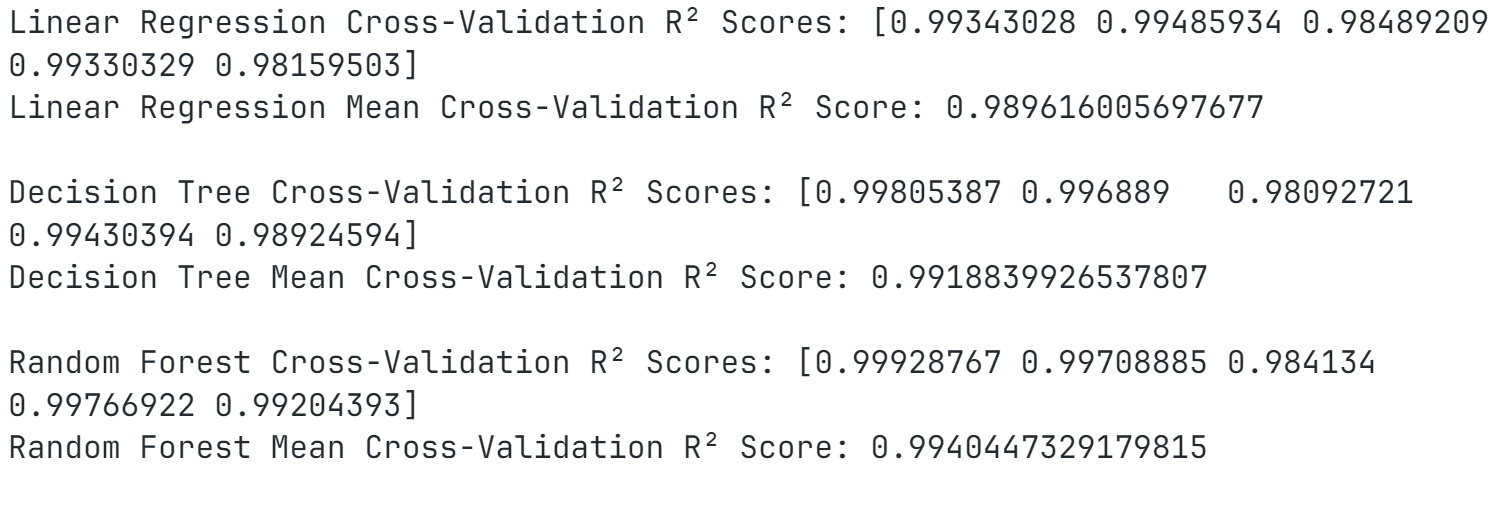
****

5.9Model Initialization and Training:

****Models are trained and prediction is done on test data

**5.10 Regression Metrices:**

****



# CONCLUSION

Now based on the 3 models that we have performed on our dataset H **random forest** is the best model as it has the highest mean cross-validation R² score of 0.9940, indicating that it explains the most variance in the data across different folds. The Random Forest model also has the lowest MSE (5.7802e-05), which means it makes smaller errors in predictions compared to the other models. The Random Forest model has the highest R-squared value on the test data (0.9972), indicating that it has the best fit and predictive power on unseen data