Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

Dependent variable is changing as per changes in Categorical variables.

Their effects are as follows –

Month – Dependent variable is changes depending on January, July and September month.

Season – Dependent variable is changes depending on Winter and Summer season.

Weather Situation – Dependent variable is changes depending on Weather Situation 3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) and Weather Situation 2 (Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist).

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

During converting Categorical variable's unique value into dummy variables; we can create 'n-1' dummy variables to represent 'n' unique Categorical values.

We achieve this setting 'drop_first=True' in 'get_dummies()' method, this removes one of Dummy variable during conversion of Categorical variable into Dummy variables.

Due to this we eliminate chances of multi-collinearity issue during model building.

It also ensures that model interprets coefficient correctly. It also simplifies the model.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

'Temperature (temp)' numeric variable has highest correlation with target variable.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 4 goes below this line> (Do not edit)

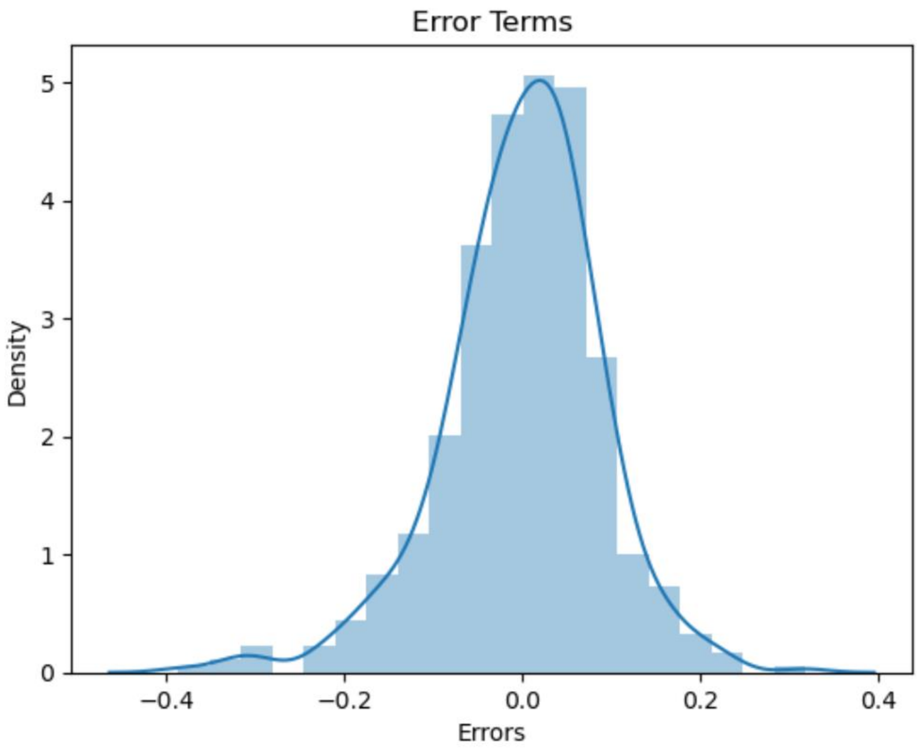
We check p-value and VIF for each independent variable.

p-value should be less than 0.05 for each independent variable.

VIF value should be less than 5 for each independent variable.

We also perform Residual Analysis; where we plot Histogram of 'Y Train Actual' vs 'Y Train Predicted'.

Histogram should show graph of Normal Distribution with mean of zero.



Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

Temperature (temp)
Year (yr)
Weathersit

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)
Linear regression algorithm is Supervised machine learning model.
It is used for Continuous dependent variable.

The objective of Linear regression is to find line or hyperplane that minimizes difference between actual target values and predicted target values.
This is achieved by minimizing cost function Residual Sum of Squares (RSS) using either Differentiation method or Gradient Descend method.

Strength of linear regression model is mainly explained by R^2
Where $R^2 = 1 – (RSS/TSS)$
RSS – Residual Sum of Squares
TSS – Total sum of squares

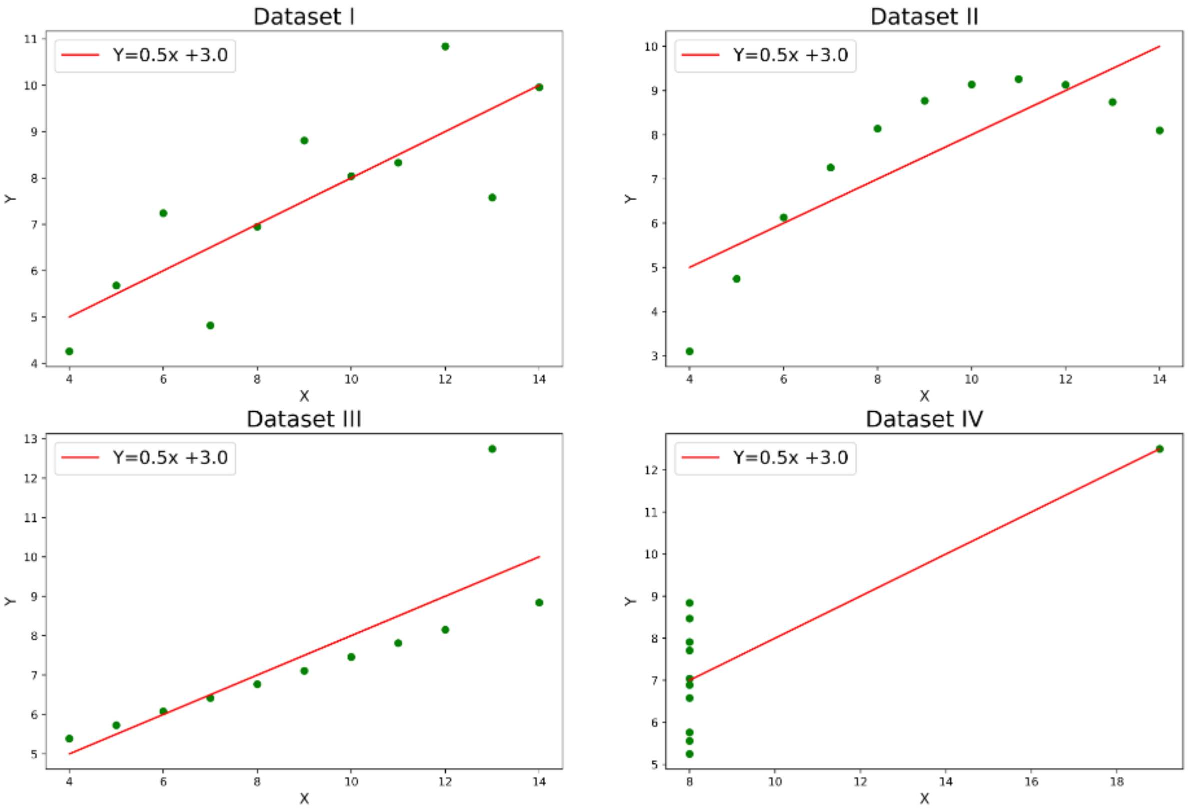
R-square values lies between 0 and 1; where 1 mean all variance in data is explained by model and 0 means none of the variance in data is explained by model.
Generally model having R-square value of more than 0.7 is considered good model; but it can change based on business requirements.

<Your answer for Question 6 goes here>

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Anscombe's quartet consists of four datasets having identical descriptive Statistical properties like R-squared, means, co-relation, variance but having different representations when we plot scatter plot graph.

This dataset is created by Francis Anscombe in 1973 to show the importance of visualizing the data and to show that summary statistic alone can be misleading. Each dataset includes 11 x-y pairs of data when plotted; they will show different graph for each dataset but will have identical statistical properties.



It highlights importance of combining Statistical analysis with Graphical exploration for accurate data interpretation.

<Your answer for Question 7 goes here>

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Pearson's R, also known as Pearson's Correlation Coefficient, is statistical measure that gives strength and direction of Linear relationship between two continuous variables.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2]\,[n\Sigma y^2 - (\Sigma y)^2]}}$$

Range –
Its value 'r' ranges from -1 to 1
r = 1 means perfect positive linear relationship
r = -1 means perfect negative linear relationship
r = 0 means no linear relationship

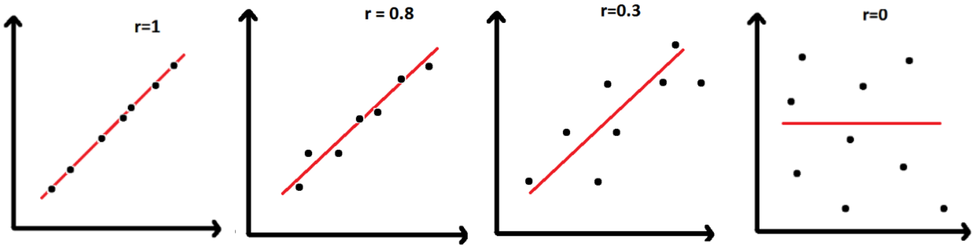
Direction –
Positive r indicates as one variable increases other also increase
Negative r indicates as one variable increases other variable decreases.

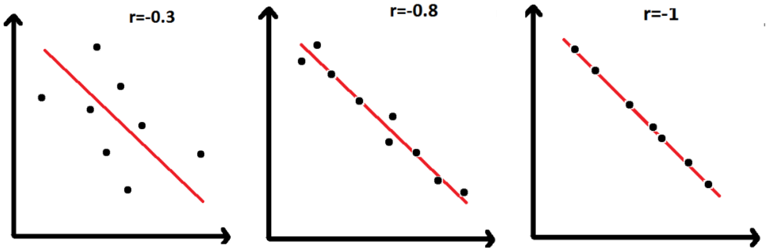
Strength –
r value closer to -1 or 1 indicates strong relationship between two variables
r value closer to 0 indicates weak or no relationship

Assumptions –

1) Both variables are continuous and normally distributed
2) The relationship between variables is linear.
3) Data does not contain significant outliers as they can affect 'r' value.

Pearson's R (Correlation Coefficient) is used for determining strength of association between two continuous variables.





<Your answer for Question 8 goes here>

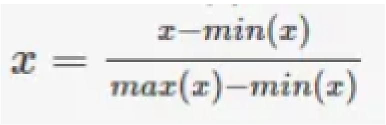
Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)
Scaling is process of transforming values of features in dataset to common scale.

Scaling is performed to –

1) Make data easier to visualize before modelling
2) Prevent some features from dominating the model, which can lead to poor performance
3) Help optimize algorithm to reach minima faster

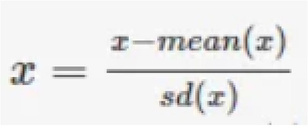
The difference between normalized scaling and standardized scaling as below –

1) Normalized scaling, scales data to range (usually 0 to 1) based on minimum and maximum value; while Standardized scaling, centers data around mean 0 and scales it by standard deviation 1.
2) Normalization can help in adjusting outlier if used correctly depending on technique; where as Standardized is more consistent approach to fix outlier problem.
3) Formula of Normalized scaling –



$$x = \frac{x - min(x)}{max(x) - min(x)}$$

Formula for Standardized scaling –



$$x = \frac{x - mean(x)}{sd(x)}$$

<Your answer for Question 9 goes here>

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this

happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

VIF stands for Variance Inflation Factor. It is statistical concept that measures how much multi-collinearity is present in regression model.
It is calculated by the formula below.

$$VIF_i = \frac{1}{1 - R_i^2}$$

Where ‘i’ is i^{th} variable

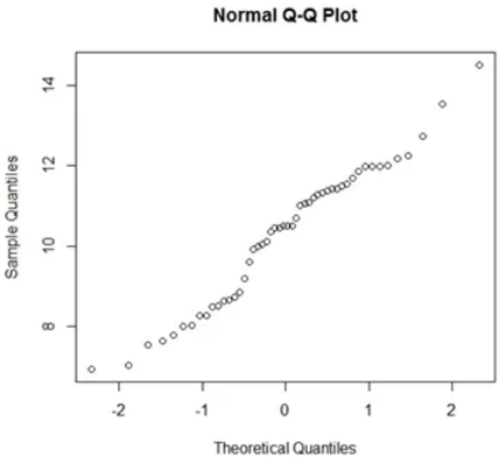
As you can see, when R_i^2 becomes 1 than denominator becomes zero; which means VIF becomes infinity.
It denotes perfect correlation in variables; essentially meaning one variable can be perfectly predicted by another.

<Your answer for Question 10 goes here>

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Q-Q plot (quantile-quantile plot) is graphical technique to check if two datasets come from population with common distribution.

Q-Q plot is scatterplot of two sets of quantiles plotted against each other. If both sets of quantiles come from same distribution than plot will show that all points are forming a line that is roughly straight.



Use of Q-Q plot in linear regression –
After building Linear Regression model, it is used to check if 'y actual' and 'y predicted' points lie approximately on the line. If they are not that means our residuals are not Gaussian (Normal) and hence our errors are also not Gaussian.

Importance of Q-Q plot –

1) Multiple distribution aspects can be simultaneously tested using Q-Q plot. i.e. presence of outliers, shifts in location, shifts in scale, etc.
2) Sample size for Q-Q plot do not need to be equal.
3) It provide more insights into nature of difference than analytical method.

<Your answer for Question 11 goes here>