Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

Answer: Some effect of categorical variable on the dependent variable are.

a. Demand of bike falls in spring.
b. Demand of bike has increased from 2018 to 2019
c. Demand of bike increases from June to September
d. when weather is clear demand of bike is more.
e. There is no major changes in demand on weekdays.

2. **Why is it important to use drop_first=True during dummy variable creation?**

Answer: drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Example: In case of relationship variables has three levels single,married,in a relationship, if we drop a level 'single' we will be still able to explain it as if 'married' and 'in a relationship' is 0 it would mean status is 'single'.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

Answer: The 'temp' and 'atemp' variable has highest correlation with target variable 'cnt'.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

Answer:

I have done the following validation

Error terms are normally distributed

Error term do not follow any pattern

Multicollinearity check using VIF(s).

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

Answer: The top 3 feature contributing are:

1. light snow and rain has negative corelation,
2. temp(temperature) has positive corelation,
3. yr(Year) has positive corelation

General Subjective Questions

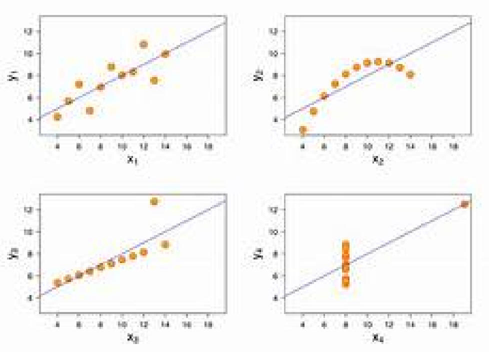
1. Explain the linear regression algorithm in detail. (4 marks)

Answer: Linear Regression Algorithm is a machine learning algorithm based on supervised learning,where the output variable is continous.It is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. It is used to model the relationship between two variables by fitting a linear equation to the observed data. The goal of linear regression is to find the best fit line that can predict the value of the dependent variable based on the value of the independent variables. There are two types of Linear regression model, simple linear regression which has one independent variable and Multiple Linear Regression Model which has two or more independent variables.

In simple linear regression, the equation of the line is given by y = mx + b, where y is the dependent variable, x is the independent variable, m is the slope of the line, and b is the y-intercept 3. In multiple linear regression, the equation of the line is given by y = b0 + b1x1 + b2x2 + ... + bnxn, where y is the dependent variable, x1, x2, …, xn are the independent variables, and b0, b1, b2, …, bn are the coefficients
An important aspect of linear regression are the assumptions of linear regression.
There is a linear relationship between X and Y:
Error terms are normally distributed with mean zero(not X, Y):
Error terms are independent of each other:
Error terms have constant variance (homoscedasticity):

Some other aspects in case of multiple linear regression are Overfitting,Multicollinearity,Feature selection

2. Explain the Anscombe's quartet in detail.



Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough". The quartet is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties. Each dataset in the quartet has a unique pattern when plotted, which highlights the importance of visualizing data before drawing conclusions .

3.**What is Pearson’s R?**

Answer: The Pearson correlation coefficient is a statistical measure that quantifies the strength and direction of a linear relationship between two variables. It is a number between -1 and 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation 1.

The formula for calculating Pearson’s r is:
r=∑i=1n(xi−x¯)2∑i=1n(yi−y¯)2∑i=1n(xi−x¯)(yi−y¯)
where x and y are the two variables, n is the number of data points, and x_i and y_i are the i-th data points of x and y, respectively. The bar over x and y represents the mean of x and y, respectively 1.

The Pearson correlation coefficient is widely used in many fields, including finance, economics, psychology, and biology, to name a few. It is used to determine whether two variables are related and how strong that relationship is 1.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer: Scaling is the process to normalize the data within a particular range. It is performed to bring multiple variables in different ranges to a single range. The two most discussed scaling methods are Normalization and Standardization.The difference between normalized scaling and standardized scaling is:
Normalizing can either mean applying a transformation so that you transformed data is roughly normally distributed, but it can also simply mean putting different variables on a common scale.
Normalization=X-Xmin/Xmax-Xmin

Standardizing means subtracting the mean and dividing by the standard deviation.
Standardization= X –mu/sigma

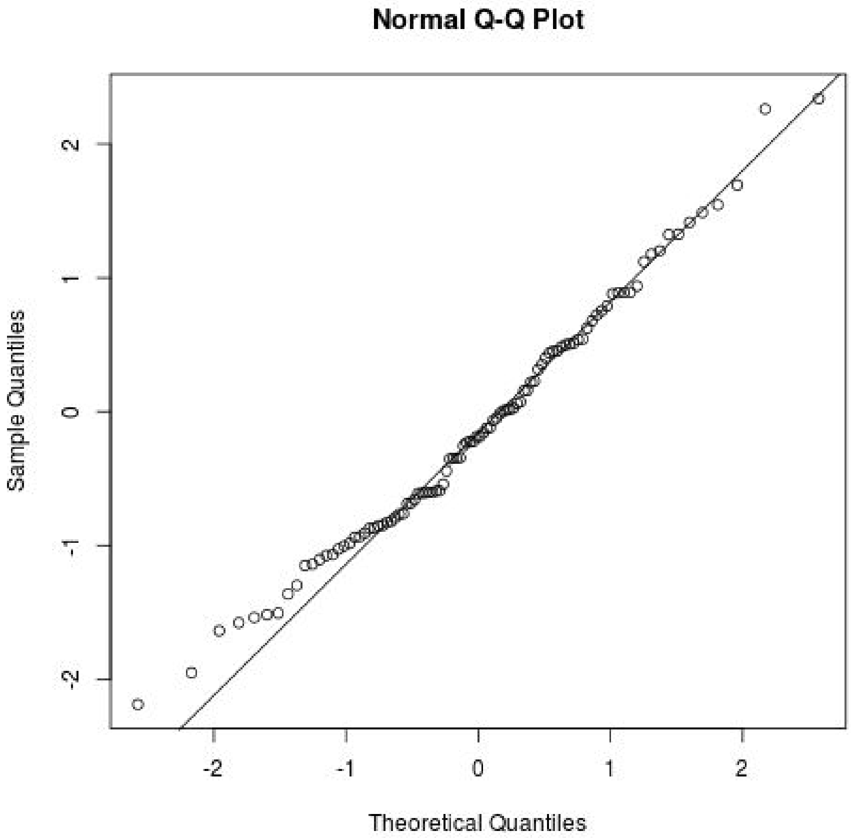
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

An infinite value of VIF for a given independent variable indicates that it can be perfectly predicted by other variables in the model. The greater the VIF, the higher the degree of multicollinearity. In the limit, when multicollinearity is perfect, the VIF tends to infinity.

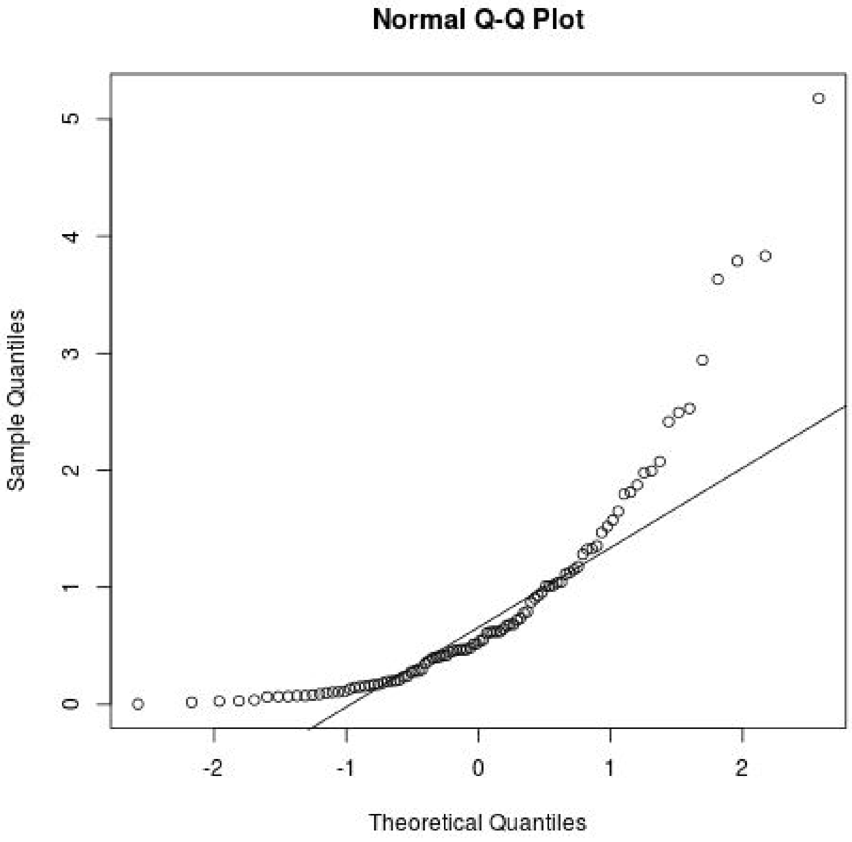
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A Q-Q plot, short for quantile-quantile plot, is a type of plot that we can use to determine whether or not the residuals of a model follow a normal distribution. If the points on the plot roughly form a straight diagonal line, then the normality assumption is met.

The following Q-Q plot shows an example of residuals that roughly follow a normal distribution:



However, the Q-Q plot below shows an example of when the residuals clearly depart from a straight diagonal line, which indicates that they do not follow normal distribution:


Normal Q-Q Plot
Sample Quantiles
0
1
2
3
4
5
-2
-1
0
1
2
Theoretical Quantiles