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? (3 marks)
Answer:
Observations from the bar plots for categorical variables:

- Yearly distribution of Total number of Users indicates that more bikes are rent during 2019.
- Seasonal distribution of Total number of Users indicates that more bikes are rent during fall season.
- Distribution of Total number of Users across various seasonal changes indicates that more bikes are rent when weather is clear for both the years.
- More number of bikes were rented on Working days.
- More bikes are rent for the year 2018 in the month of June.
- More bikes are rent for the year 2019 in the September month.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
Answer:
drop_first = True helps in reducing the extra column created during dummy variable creation.
Hence it reduces the correlations created among dummy variables.
If we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.
Ex:
We have 2 types of values(Holiday or Working-day) in Categorical column and we want to create dummy variable for that column. If one variable is not Holiday, then it is obvious Working-day. So we do not need 2nd variable to identify the Holiday.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
Answer:
Variable ‘temp’ (temperature) has the highest co-relation with the target variable ‘cnt’.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
Answer:

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
Answer:
Following are the top 3 features contributing significantly towards explaining the demand of the shared bikes based upon their co-relation with target variable:

- temp
- Yr 2019
- Weather Light Snow

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
Answer:
Linear Regression is the statistical technique to understand the relationship between the dependent variable and the independent variable/variables.
It is an Machine learning algorithm based on supervised learning.
The best fit relation can be represented with the help of following equation:
Y = mx + c
Where,
x: is the independent variable
c: intercept
m: coefficient of x.

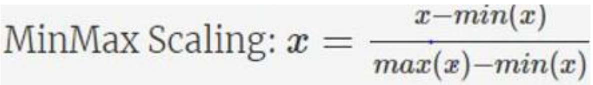
2. Explain the Anscombe’s quartet in detail. (3 marks)
Answer:
Anscombe’s quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

It is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. What is Pearson’s R? (3 marks)
Answer:

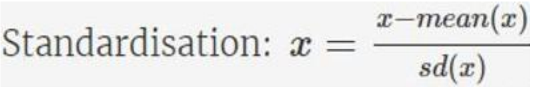
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
Answer:
Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:
The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.



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

Standardization Scaling:
The variables are scaled in such a way that their mean is zero and standard deviation is one



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

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
Answer:
If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
Answer:
A Q - Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

It is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.

Importance of Q-Q plot: When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.