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)

Ans :-

- People prefer to use bikes in the season of fall the most and spring the least.
- It has a high change for a person to rent bike on a clear day as compared to rainy day also we do not have data on heavy rain day.
- It is seen that people are usually renting bikes on a working day more compared to holiday so they must be using it for travelling to offices.
- We have observed that sales have increased almost 1.5 times in year 2019 as compared to 2018.
- Weekday data is almost similar and Monday has least rental of all

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

Ans:- As if we will leave drop first then there will be for example 2 features for 2 categories but we can also use 1 and have exactly same output. As 0 and 1 can signify both the features correctly.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans: - Temperature is seen to have the highest correlation.

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

Ans: -

- The residuals are normally distributed.
- Variance of error is low.
- VIF is low.
- P value is low
- Probability of F score is near zero

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans:-

- atemp
- Light rain
- Year

Based on the coefficient

General Subjective Questions

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

Ans: - In linear regression we tend to find the coefficient of the equation.

Bn Xn + Bn-1Xn-1 + . . . + B1X1 +B0 where X1 to Xn are the features and based on the RMSE we tend to find the coefficient with lowest error. The thought behind the same is to find the n dimensional line/hyperplane that cuts all the points in such a way that the RMSE is lowest.

2. Explain the Anscombe's quartet in detail. (3 marks)

Ans: - It is basically different datasets that can fool linear regression as linear regression solely aims at reducing the RMSE value which is wrong as there could be datasets with similar RMSE but different data variance so it helps in understanding that visualising the data is always important.

3. What is Pearson's R? (3 marks)

Ans: - It is a way to analyse the correlation between 2 variables. Its value lies between -1 to 1 where 1 signifies perfect positive correlation, -1 signifies perfect negative correlation. From 0-1 it is positive correlation and from 0 to -1 it is negative corelation.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans:- Scaling means to put variables on a same scale and could be done In many ways either by varying the values from 0-1 to by standardising the standard deviation by not affecting the integrity of information that variable shares.

- Standardized scaling means modifying the variables such that there mean turns out to be 0 and standard deviation as 1
- Normalized scaling means modifying variable such that the min value is modified to 0 and max to 1 and accordingly rest values are adjusted.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans: - This happens when a feature is perfectly corelated to another feature and due to this the R2 comes 1 and thus computing VIF to infinity (Inf).

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans:- A q plot is plotting data from two quantile against one another. This graphical method aims to determine if the two samples of data come from the same distribution or not. If they do not the skewness and kurtosis will increase.