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: We saw that all the categorical variable has effect on the dependent variable. Below are some of the variable which has affected the demand:

a. **yr - 2019 has more demand of these bikes than 2018. So demand will increase if situation become normal**
b. **spring(season) - Demand has decreased in spring season. So company can focus to improve demand in spring**
c. **Mist(weathersit) - Demand has decreased in Mist weather.**
d. **Winter(season) - Demand has increased in winter season**
e. **Jul(mnth) - Demand has decreased in Jul month.**
f. **Sep(mnth) - Demand has increased in sept month**
g. **Sun(weekday) - On Sundays we see less demand for bikes**
h. **LightSnow(weathersit) - Demand has decreased in LightSnow/Light Rain weather.**

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

Ans: Because n variables can be represented by n-1 variable. So we can drop first column of the dummy variables

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

Ans: Temp and atemp 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? (3 marks)

Ans: assumptions of Linear Regression were validated as below:

a. **There was a linear relationship between Independent variables and Dependent Variable which we verified using pairplot of the dataframe.**
b. **Residual vs Fitted values plot does not show Heteroscedasticity.**
c. **All the predictors has VIF<5, which indicates there was not Perfect Multicollinearity.**
d. **Using Distribution plot on the residuals we verified that Residuals are normally distributed.**

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: temp, yr and weathersit are the top 3 features contributing significantly towards explaining the demand of the shared bikes

General Subjective Questions

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

Ans: Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Linear regression performs the task to predict a dependent variable value (y) based on a given set of independent variables (x). So, this regression technique finds out a linear

relationship between x (input) and y(output). We have given dataset with independent and target variables. We clean the data, check the correlation and then divide the data into Train and test sets. We perform scaling of numerical features and encode categorical features. We then train the model using train data set and then predict using test data.

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

Ans: Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties(mean, std deviation etc.), 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 analysing it and the effect of outliers on statistical properties.

Below are the explanations for these data points:

1. **In the first oneif you look at the scatter plot you will see that there seems to be a linear relationship between x and y.**
2. **In the second oneif you look at this figure you can conclude that there is a non-linear relationship between x and y.**
3. **In the third oneyou can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.**
4. **Finally, the fourth one shows an example when one high-leverage point is enough to produce a high correlation coefficient.**

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

Ans: The Pearson's R also known as Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Pearson correlation coefficient (r):

a. **Between 0 and 1 – It shows Positive correlation. Which means when one variable changes, the other variable changes in the same direction.**
b. **Zero (0) - No correlation. Which means there is no relationship between the variables.**
c. **Between 0 and –1 - It shows Negative correlation. Which means when one variable changes, the other variable changes in the opposite direction.**

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

Ans: Scaling is part of the pre-processing step which is applied on independent variables (particularly numerical variables) to normalize the data within a particular range.

Most of the times the data in the data sets have different units and some have very high magnitudes and some has low magnitudes. If scaling is not done on these data sets 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.

There are basically two types of scaling one can perform. Below are the types and differences:

a. **Normalization/Min-Max Scaling: It brings all of the data in the range of 0 and 1. It is really affected by outliers as High Max value can cause small values having very small scaled value. Below is the formula for the scaling:**
   X_new = (X - X_min)/(X_max - X_min)
b. **Standardization Scaling: Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ). It is much less affected by outliers.**
   X_new = (X - mean)/Std

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

(3 marks)

Ans: If there is perfect correlation, then VIF is 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)

Ans: Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.