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 : As per my final model, the predictor variables are

- **Temperature (temp)** - A coefficient value of '0.548' indicated that a unit increase in temp variable increases the bike hire numbers by 0.548 units.
- **Weather Situation 3 (weathersit_3)** - A coefficient value of '-0.0784' indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.0784 units.
- **Year (yr)** - A coefficient value of '0.2329' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2329 unit
- **windspeed:** - A coefficient value of '-0.1598' indicated that, a unit increase in windspeed variable decreases the bike hire numbers by 0.1598 units.
- **season1:** spring has coeff
- **season4:** winter
- **Holiday** : Holiday has negative Coeff -0.0987 indicating that during holidays, there might be a reduction in Count. However, very less holidays will be there per year. It is almost not so significant unless it is holiday.
- The details of weathersit_1 & weathersit_3
    - **weathersit_1:** Clear, Few clouds, Partly cloudy, Partly cloudy
    - **weathersit_3:** Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

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

ANS : As it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among 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: Highest Correlation is with Temp,Atemp

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

ANS: Based on VIF and P values

- As the R2 & adjusted R2 is 0.845 in the initial model, we can say that 80% variance of the data is explained by the features selected for the Linear regression
- The higher P values indicate that the beta coeffecients are not significant and we can drop those variables
- Higher VIF indicates that there is multiple collinearity in the model and hence always VIF is supposed to be 5 or lesser than 5. Higher VIF should be eliminated
- I also checked the business significance being explainable with the model

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 : YEAR, TEMPERATURE, WEATHER SITUATION

General Subjective Questions

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

ANS : If we have a target variable which needs to be predicted based on PAST DATA with LABELS, where we have DATA as continuous variable (And not categorical), we can predict the data based on LINEAR REGRESSION ALGORITHM. There should be one or more independent variables which can be identified as monotonously increasing or decreasing and can be represented by a straight line(indicating a linear relationship) with Y = C + M1X1 + M2X2 + M3X3 + M4X4+….MNXN.. Here, M1,M2,M3, ..are the coefficients of how the target Variable "Y" is varying. Using this model. We can predict the Future data or future Y.

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Here, we train a model to predict the behaviour of your data based on some variables based on linear regression model achieved through LR algorithm

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

ANS : 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.

In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.

In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.

In the third one(bottom left) you 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.

Finally, the fourth one(bottom right) 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 : In statistics, the Pearson correlation coefficient (PCC, pronounced /ˈpɪərsən/), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation,[1] is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation.

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

ANS : It is a step of Data Pre Processing which is applied to independent variables or features of data. It basically helps to normalise the data within a particular range. Scaling makes all the data to be increasing or decreasing relatively irrespective of its actual magnitude.

It is also said that, scaling helps in speeding up the calculations in an algorithm.

There are 2 types of scaling 1. Minmax scaling(Normalized scaling)

2. Standardized Scaling

MinMax Scaling = (X-Xmin)/(Xmax-Xmin)

Standarized Scaling = X-MEAN/Standard Deviation

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

Value of VIF can be infinite. As VIF = 1 /(1-R^2), When the R^2 is very high approx 1 or equal to 1. Then this situation might arise.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well). We have to drop these variables as this becomes 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. Also, it helps to determine if two data sets come from populations with a common distribution.

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.

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. A 45 degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.