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: Count is the dependent variable here. The categories are weathersit, season, weekday and month. As from the boxplots in the assignment, we can see the following effect on the bike usage or demand:

- Weathersit: If the weather is clear, obviously there will be more demand for the bikes. There will not be humidity or snow in the same weather.
- Season: Fall and summer season has the highest median, so it is considered as most favourable seasons to ride the bikes or the usage will be maximum
- Weekday: We can see on Sunday and Monday, the median is less apart from the other weekdays. So bike usage is low on Sunday as it is a holiday
- Month: There is more demand in July for the bikes

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

Ans: The drop first is more usable in our scenario as it will not create the unnecessary category and the variable. So in weather we have 3 categories .ie. CloudY+Mist, Clear and LightSnow. So once we apply dummies with drop first, it will only create the values for the two categories i.e. CloudMisty and Light Snow so it will not create the third category i.e. Clear.

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

Ans: In Our case study, we have target variable I,e, count. Highest correlated numericals are from Temp and Year in the pair plot.

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

Ans: There are following assumptions from which we can validate that it is following linear regressions on the training sets:

- As we can see in the chart that temperature, casual, registered, instant variables are mostly correlated in training sets.
- Error terms should follow normal distribution (Homoscedasticity)
- When we plot the pyplot, it is following multivariate normal distribution

- Also we can observer linear relationship between target variables and other variables in the pyplot (temp with count)
- The line should follow Y=mx +c equation

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:

1. US bike-sharing provider BoomBikes can focus more on Temperature as the 0.4918 coefficient plays important role in the increase of the bike usage

2. We can see the high number of revenues with the bike usage in 2019 compared to 2018. But company needs to observe the Corona Pandemic situation to make further decisions

3. Also, BoomBikes can focus more on the inventory availability on Summer, Winter seasons, August and September month, Weekends and Working days as the coefficients can highly increase the usage and revenue.

General Subjective Questions

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

Ans: Linear Regression is a ML 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.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output).

Hypothesis Function for Linear Regression is : Y= c+ mx

Where x is input data and Y is the value to be predicted

C is constant/intercept

M is coefficient of x

If you see our equation which we derived, it is matching the same pattern

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

   Ans: Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.
   It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.
   Dataset 1: this fits the linear regression model pretty well.
   Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
   Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
   Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

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

   Ans: In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.
   Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.
   Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.
   Pearson's Correlation Coefficient is named after Karl Pearson. He formulated the correlation coefficient from a related idea by Francis Galton in the 1880s.
   Using the formula proposed by Karl Pearson, we can calculate a linear relationship between the two given variables. For example, a child's height increases with his increasing age (different factors affect this biological change). So, we can calculate the relationship between these two variables by obtaining the value of Pearson's Correlation Coefficient r.
   There are certain requirements for Pearson's Correlation Coefficient:

   - Scale of measurement should be interval or ratio
   - Variables should be approximately normally distributed
   - The association should be linear
   - There should be no outliers in the data

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

Ans: Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).If we have higher range of values in our datasets, then our values will not be corrected in the regression and building the models. It is always a best practice to scale the values using above techniques to predict the correct values.

Normalisation	Standardisation
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
It is useful when we don’t know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

(Normalisation = (x-xmin)/(xmax-xmin))

(Standardisation= (x-mu)/ sigma)

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

(3 marks)

Ans: When VIF = Infinite, then 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: 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. Also, it helps to determine if two data sets come from populations with a common 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.