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: Box plot was used to study effect of different variables on the dependent variable (‘cnt’) .

1. Season: Fall season has comparatively higher number of bookings as compared to other three seasons while spring season has the lowest.
2. Month: Month follows a similar pattern to season as fall months are having more bookings.
3. Weather situation: Light snow condition seems to be a dominating factor affecting drop in bookings.
4. Year: Plot shows considerable increase in bookings from 2018 to 2019; showing good progress in business terms.

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

Answer: For K level categorical variable, (K-1) dummy variables can represent all the information. Hence, by dropping first column we improve the efficiency and reduce 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)

Answer: ‘temp’ variable has the highest correlation with the target variable.

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

Answer:

- Histogram of the error terms to check normality.
- Error terms are independent – no relation
- Homoscedasticity - No visible pattern in residual values.
- Correlation heat map for Multi collinearity checking.
-

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: Top 3 features contributing significantly towards explaining the demand of the shared bikes

- Temp(temp)
- Year(yr)
- Light snow weather situation (highest negative correlation)

General Subjective Questions

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

Answer - Linear regression is a supervised learning algorithm that finds a mathematical relationship between variables and makes predictions for continuous or numeric variables.

The independent variable is also the predictor or explanatory variable that remains unchanged due to the change in other variables. However, the dependent variable changes with fluctuations in the independent variable. The regression model predicts the value of the dependent variable, which is the response or outcome variable being analysed or studied.

The formula for multiple linear regression would look like,

y(x) = p0 + p1x1 + p2x2 + … + p(n)x(n)

The machine-learning model uses the above formula and different weight values to draw lines to fit. Moreover, to determine the line best fits the data, the model evaluates different weight combinations that best fit the data and establishes a strong relationship between the variables

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

Answer: Anscombe’s Quartet is 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.

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.

The four datasets can be described as

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.

Thus, it is important to visualize the data before applying various algorithms out there to build models. Data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.

3. What is Pearson’s R? (3 marks)

Answer: Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

"Tends to" means the association holds "on average", not for any arbitrary pair of observations, as the following scatterplot of weight against height for a sample of older women shows. The correlation coefficient is positive and height and weight tend to go up and down together. Yet, it is easy to find pairs of people where the taller individual weighs less, as the points in the two boxes illustrates.

The Pearson's correlation coefficient varies between -1 and +1 where:

r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)

r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)

r = 0 means there is no linear association

r > 0 < 5 means there is a weak association

r > 5 < 8 means there is a moderate association

r > 8 means there is a strong association

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 fit the data within a particular range.

When you have lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons:

1. Ease of interpretation

2. Faster convergence for gradient descent methods

Features can be scaled using two very popular method:

1. Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.
2. MinMax 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.

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-square, etc

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

Answer: If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

The common heuristic for VIF is that while a VIF greater than 10 is definitely high, a VIF of greater than 5 should also not be ignored and inspected appropriately.

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

Answer: Q Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other.

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.

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

The slope tells us whether the steps in our data are too big or too small. If we have N observations, then each step traverses 1/(N-1) of the data. So we are seeing how the step sizes compare between our data and the normal distribution.

A steeply sloping section of the QQ plot means that in this part of our data, the observations are more spread out than we would expect them to be if they were normally distributed. One example cause of this would be an unusually large number of outliers (like in the QQ plot we drew with our code previously).