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]: From analysis of categorical variables, have got below inferences:

I. We found most of bike booking happened in 'Fall' followed by 'Summer' and 'Winter'.

II. We found most of bike booking happened in months May, June, July, August and September.

III. There is also good impact of holiday column. Most of booking happened when there was not holiday as compare to opposite.

IV. For weathersit , we found 'Clear-Few_Partly_cloudy' are having better performance as compare to other.

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

[Answer]: When we use drop_first=True during dummy variable creation, it creates columns for all valid values dropping first valid value. Its important to use because firstly It reduce column creation so we will have less number of columns and it will not impact other activities. Secondly it also reduces correlation.

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

[Answer]: Looking at pair - plot, we see temp variable has highest correlation (0.63) with target variable 'cnt'.

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

[Answer]: After building the model, checked below points for validation:

i. Error terms are normally distributed with mean at zero.

ii. Predicted variables (e.g. temp) are having linear relationship with target variable(cnt).

iii. There is no multicollinearity between predictor variables.

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]: Based on final model, top 3 features which are significantly contributing towards explaining the demand of the shared bikes:

1. Temp (Temperatue) – > Positive correlation
2. Yr (Year) – > Positive correlation
3. season_Winter – > Positive correlation

General Subjective Questions

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

[Answer]: Linear regression algorithm is statistical regression method which we used for predictive analysis for continuous variables. Here we use independent variables to predict dependent variable. Linear regression shows the linear relationship between independent variables and dependent variable. If we are using only one independent variable to predict dependent variable then that model is called simple linear regression. But if we are using two or more independent variables to predict dependent variable then that model is multiple linear regression.

In linear regression, model define a best fitted straight line describing relationship between dependent and independent variables.

Best fitted line is represented by slope – intercept form of straight line.

y = B0 + B1X

where y = dependent variable

X = independent variable

B0 = intercept

B1= Slope / co-efficient of linear regression

For finding best fitted straight line, B0 and B1 should be optimized. Cost function is used to optimize these values.

We use generally two metrices to analyze strength of linear regression:

1. **R2(coefficient of determination)**

R2 indicate how well model able to explain dependent variable using independent variable(s).

Its value always lie between 0 to 1. Overall, for higher R2 value we consider model as a good fit.

2. **Residual Standard Error**

This measures how well a linear regression model fits the data.

We have some pre-assumption while using linear regression:

1. There should be linear relationship between dependent variable and independent variable(s).
2. Homoscedasticity: The variance of residual should be the same for any value of X.
3. Residuals (error term) should be independent. In particular, there should not be correlation between consecutive residuals.
4. The residuals of the model should be normally distributed.

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

[Answer]: Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics but they have very different distributions and appear differently when plotted on scatter plots.

These four data sets were given by statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties. These datasets have nearly same statistical observations, which provides same statistical information i.e. variance, and mean of all x,y points. But when these data sets are getting plotted on a scatter plot, all datasets generate a different kind of plot.

This illustrates the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.

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

[Answer]: It is a statistic measures of the linear correlation between two variables. It is the ratio between the covariance of two variables and the product of their standard deviations. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0. Pearson's Correlation Coefficient is named after Karl Pearson. For measuring correlation in statistics, we generally use Pearson's correlation coefficient. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

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 normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

We do scaling because Most of the times, collected data set contains features highly varying in magnitudes, units and range among each other. So, If scaling is not done 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.

Scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization Scaling: It brings all of the data in the range of 0 and 1. We use below formula for this.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

standardized scaling: It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ). We use below formula for this.

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

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

(3 marks)

[Answer]: When we get VIF = infinity that means there is perfect correlation. 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, 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)

[Answer]: The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.
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. 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.
If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x.
A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.