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: -

Few Categorical variables in the assignment are Weekday, Month, Season, WeatherSit, holiday and working day

With the analysis of these variables using box and bar plot we can infer that

1. There is considerable about the increase in demand from 2018 to 2019
2. There is a pattern seen in spread of data across months, with highest peaks from May to Oct I.e., >4000 bookings from May to Oct with Sep the highest month
3. More bookings are in season fall followed by summer, winter and spring
4. We also see that maximum bookings are in clear weather condition. Also, with this data set, we don't see any bookings in weathersit (4) = heavy rain.
5. With Weekday graph analysis, we can see that there is very minimal variance on weekend and weekday bookings, so this data is not enough to conclude in the very beginning, since this variable may or may not have influence in predicting a good model.

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

Ans:

It is important to use drop_first = True, because it helps in reducing the extra column when the same result can be achieved with n-1 columns. For e.g. In this assignment, we have four seasons in season column. Spring, summer, fall and winter.

So, when converting them to numeric variables, we can safely drop one column say first one spring. So, in the result set, there will be only 3 columns summer, fall and winter. But when all the values are 0, it clearly implies that the season is spring. We don't need an additional column to conclude the season. Hence it is safe to use drop_first with an added advantage on extra maintenance.

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

Ans: - atemp and temp 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)

1. F-Statistics - With F-Statistics of 254.0, (> 1) and p-values = ~0.00, we can prove that overall, it is a significant model.
2. Const interpretation - with the const coefficient of 0.075325, it indicates that even in absence of all predictor variables, the bike rental can still increase by 0.075 units.

3. Absence of multicollinearity - all predictors have VIF less than 5 and also, with above heatmap, we can consider that there is insignificant collinearity between the 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)

After the final model, top 3 features contributing significantly are

1. temp with coeffecient value of '0.549936' indicates that it has high influence on the bike renting
2. snow(weathersit = 5) with coeffcient value of -0.288021 indicates that it brings down the booking count of bike
3. yr with coeffecient value 0.233056 proves that booking has improved over years

General Subjective Questions

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

Linear Regression model is the supervised machine learning algorithm where a dependent variable y is predicted on given independent variable x.

There are 2 types of LR models

1. Simple Linear Regression where only one independent variable x is used to predict y
2. Multiple Linear Regression where multiple independent variables x1, x2, x3 etc. are used to predict the dependent variable y

The simple y = mx+c formula is used as basic analysis approach. C is the intercept and m is the coeffecient of x.

There are multiple approaches to build the model to predict the y value. Once we have m and c, which is considered as beta0 and beta1 in machine learning language, we will get the best fit line.

The main aim on predicting the best fit line is to predict the y value in such a way that the error I.e difference between the predicted value and actual value is very minimal.

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

The Anscombe's quartet is 4 different datasets with very simple statistics which look very much similar to each other but look very differently when plotted in graph.

This explains the importance of analyze the data using different graphs and why basic statistic properties are not enough for explaining the relationship between variables.

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

Pearson's R is nothing but the correlation coefficient which explains the linear correlation between 2 sets of data.

The result always varies from –1 to 1. There is a specific formula to calculate the R2 in statistics which is dependent on y (dependent variable) and x (independent variable). We can get the same values using df.corr() function in python.

Scaling is the technique to standardize the independent features available in the dataset in a fixed range.

Scaling is performed to ease the data handling and analysis within the given fixed range otherwise and give equal weightage to all the data be it high or low during analysis. If scaling is not done, then ML algorithm tend to weigh greater values high and less values low irrespective of their unit of measurement.

Standardization scaling or Z-score scaling is transformation of data by subtracting from mean and dividing by standard deviation. Not range bounded.

Normalization or Min-Max Scaling is dependent on min and max values of variables. I.e. (x-xmin)/(Xmax-xmin). This is used when data is of different scales. Ranges from 0 to 1 and sometimes –1 to 1

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

In case of perfect correlation of 1, then VIF will be infinity. High correlation means high VIF. VIF > 10 is considered as highly correlated.

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

Q-Q plot is used to analyze the distribution of data and check if it follows any theoretical pattern like linear, exponential or uniform distribution.

This helps in linear regression when we have both train and test data sets received separately. We can plot the Q-Q plot to understand and conclude whether both the data sets come from populations with similar distribution.