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.

- There were 6 categorical variables that has some effect on the dependent variable ('cnt'). Their effect is as follows:
  - **season** : Most of the bike booking were happening in season3 with a median of over 5000 booking.
  - **mnth** : A trend can be observed with number of booking increasing from January till September with some downward trend after it towards the end of the year.
  - **weekday** : Not a lot of significant difference can be observed in each level but some components were part of the final equation with a positive impact/relationship with the target
  - **holiday** : Most of the bookings happened when it was not a holiday.
  - **weathersit** : Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds had a negative impact on the overall sales of bikes.
  - **workingday** : Not a lot of significant difference was observed in the levels.

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 all the levels of a categorical variable are not required to extract the meaning out of that variable.
- Eg : If we have 3 levels in a categorical variable, let's say – furnished, semi furnished and unfurnished, then if it's not furnished or semi furnished then it automatically means that it must be unfurnished and hence we can drop this variable as it contains no information of value and will only increase the correlation and multicollinearity amongst variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
Ans. Looking at the pair-plot among the numerical variables, 'temp' variable has the highest correlation with the target variable as it is increasing with the increase in target variable and the positive linear relationship is quite visible even 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.

- I checked the distribution of the error terms which was normally distributed with mean around 0.
- I checked the linearity among the dependent and independent variables and found linear relationship among some of the independent variables with the dependent one.
- I checked for multicollinearity among variables and made sure they were not multicollinear with each and were in permissible limits.
- I checked if there was any autocorrelation among the error terms and there was no autocorrelation present.

- I checked variance of the error terms and found it to be constant throughout, that is homoscedastic in nature.

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.

- The top 3 feature contributing significantly towards explaining the demand of the shared bikes are as follows :
  - Temperature (temp) - A coefficient value of '0.5136' indicated that a unit increase in temp variable increases the bike hire numbers by 0.5136 units.
  - Weather Situation 3 (weathersit_3) - A coefficient value of '-0.2897' indicated that w.r.t to Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.2897 units.
  - Year (yr) - A coefficient value of '0.2343' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2343 units.

General Subjective Questions

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

- Ans. Linear Regression is a supervised machine learning algorithm. It predicts the value within a continuous range of numbers. Linear Regression is a method of modelling a target value based on independent predictors.
- Linear Regression can be of 2 types in our case, one is simple linear regression and the other is multiple linear regression. Simple linear regression has one dependent variable and one independent variables. Multiple linear regression has one dependent variable and two or more independent variables.
- There are a few assumptions we make when using linear regression:
  - The relationship between the dependent and independent variables should be almost linear.
  - The data should be homoscedastic in nature that is there should be constant variance of the error terms.
  - The results obtained from observation should not be influenced by the results obtained from the previous observation that is no autocorrelation.
  - The residuals should be normally distributed. This assumption means that the probability density function of the residual values is normally distributed at each independent value.
  - There should be no multicollinearity present between the independent variables.
- Some of the most popular applications of Linear regression algorithm are in financial portfolio prediction, salary forecasting, real estate predictions and in traffic in arriving at ETAs.

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

- Ans. Anscombe’s quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyse it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another.

- It tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.
- Even though the statistical summary of each of those 4 datasets is similar to one another, it is actually not possible to fit a linear regression on all of them due to presence of non-linearity or outliers in 3 of the sets out of 4.
- The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

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

- There are many situations in our daily life where we know from experience, the direct association between certain variables but we can't put a certain measure to it. Eg : You know that the chances of you going out to watch a newly released movie is directly associated with the number of friends who go with you because the more the better.
- But there are many other factors too, like your interest in that movie, your budget etc. Thus to analyse the situation in detail, we need to note down our similar past experiences and form a sort of distribution from that data. It is at this point that you require a Correlation Coefficient, which will now provide us with a value, based on which you can calculate the possibility of you not going for the movie this time if our friends don't turn up! Pearson's Coefficient of Correlation is one such type of parameter denoted with R.
- The value of this correlation ranges from -1 to +1. A value greater than 0 indicates a positive association i.e. as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association i.e. as the value of one variable increases, the value of the other variable decreases. 0 signifies no association among variables.

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

- Ans. Feature scaling is one of the most important data pre-processing step in machine learning. Algorithms that compute the distance between the features are biased towards numerically larger values if the data is not scaled. Hence, we scale the data so that it does not get biased towards the larger or smaller values.
- For example : In a linear regression, the coefficient for each variable will have a significant difference in it's value when compared before and after scaling and we will get a true sense of contribution of each variable towards predicting the target since they will be on same scale.
- Difference between normalized scaling and standardized scaling is as follows:
- Normalized scaling : Minimum and maximum value of features are used for scaling. It is used when features are of different scales and scales values lie between [0, 1] or [-1, 1]. It is really affected by outliers.
- Standardized scaling : Mean and standard deviation is used for scaling. It is used when we want to ensure zero mean and unit standard deviation and it is not bounded to a certain range. It is much less affected by outliers.

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

- VIF value can be infinite if there is a perfect correlation among the independent variables. That means, value of one of the variable is completely explained by either a single or a combination of other variables.
- The formula for VIF is 1/(1-r^2) and if the r^2 value is 1, then by calculation 1/(1-1) = Infinite. In this scenario, the variance or relationship in our variable is entirely explained with the help of other variables.
- When such scenario arrives, we should drop one of the variables causing this perfect correlation and then observe the VIF values.
- I had encountered such scenario in my linear regression case study for which I've shared my approach in the python notebook.

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

- 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. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.
- A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.
- In linear regression, it is often used to determine whether the residuals or error terms follow a normal distribution by plotting it against the quantiles to check whether they are normally distributed or not. Thus, we can validate one of the assumptions of linear regression that the error terms should be normally distributed.