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?**

Ans: - There are Six Categorical variables in the dataset namely season, weathersit, mnth, holiday, weekday, working day.

Inference about the categorical variables:

- **Season: -** Compared to 2018, 2019 has more demand for bike in all the season. Season 3 has the highest number of booking with a median of over 4000 in 2018 and with a median of over 6000 in 2019. This was followed by season 2 and season 4 for both the years. This says that season a pretty good predictor for dependent variable.
- **Weathersit:** - Here as well there is an increasing trend in demand for bike and it clearly seen in the boxplot. The highest bookings were done in weathersit 1 followed by weathersit 2 and weathersit 3. Weathersit 1 has a median of over 6000 for 2019 while it was over 4000 for 2018 so clearly that is an increase in bookings of bike. Hence weathersit is a good predictor for dependent variable.
- **Mnth: -** Here a we can see a rise in bookings between month 3 to month 10 with a median of over 5000 in 2019. Hence, we can say that mnth is good predictor for dependent variable.
- **Holiday: -** It shows that most of the bookings were done when it was not a holiday. From this we can infer that data might be biased but if we see the trend between 2018 and 2019, it shows an increasing trend of bookings being made not on holidays. We will let model decide to keep it or not.
- **Working Day:** - Working day shows a close trend having medians around 6000 for both working and non-working day. We will let model decide about this variable.
- **Week Day:** - Week day as well has close trend between all the week days has medians are around 6000 for 2019 and we might say for now that it has little to no impact on depend variable. We will let model decide.

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

Ans: - When we have a categorical variable with, say, 'n' levels, the idea of dummy variable creation is to build 'n-1' variables, indicating the levels. So, what happens is in dummy variable, where there is zero for the variable it becomes the base state (reference state) and while where there is 1 it becomes effect on top of the base state. So, in order for good interpretation the level with all the zero is removed to check how does it affect (how much extra effect it has) other levels i.e., effect of state (‘1’) on the base level (‘0’). And so, drop_first is used to remove that level and it is importance for good interpretation also lesser the columns better the analysis.

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

Ans: - From looking at the pair-plot temp, atemp has the highest correlation with target variable cnt.

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

Ans: -

- Error terms are normally distributed with mean 0 as can be seen from Residual analysis done with the help of distplot.
- As seen from the pair plot there is a linear relationship between temp and atemp. And also, both them have high corelation with target variable.
- There is no multicollinearity between the predictor variables as the VIF of all of them are within the permissible range of 5.
- Constant Variance assumption is validated by plotting a scatter plot between y_pred and residual and most of the points lie around 0.0 or in between -0.1 to 0.1 so we can say that the model has constant variance.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

Ans: - As per the model the top 3 features are: -

- **Temperature(temp): -** A coefficient value of '0.5308' indicated that a unit increase in temp variable increases the bike hire numbers by '0.5308' units.
- **weathersit_3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds): -** A coefficient value of '-0.2401' indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by -0.2401 units.
- **Year(yr):** - A coefficient value of '0.2288' indicated that a unit increase in year variable increases the bike hire numbers by '0.2288' units.

These three features should be given most importance for getting maximum bookings.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: - Linear Regression is a machine learning algorithm. It comes under the supervised learning methods. Here the output variable predicted is a continuous variable. Linear regression is the simplest form of method in machine learning as it uses the slope and intercept to get the best fit line. As the name suggest, it has a linear relationship between input variables (features) (X) and single output variable (Target variable) (y). It is one of the simplest & easiest methods in machine learning and it uses statistical method for predictive analysis.

In linear regression, it assumes: -

- There is linear relationship between X and y.
- There error terms are normally distributed.
- Error terms should have constant variance (Homoscedasticity).
- Observations are independent of each other.

There are two type of linear regression, namely

- SLR (Simple Linear Regression).
- MLR (Multiple Linear Regression).

SLR (Simple Linear Regression): In SLR, there is only one independent variable. If only one independent variable is used to predict the dependent/ Target variable, then regression is known as Simple Linear Regression.

MLR (Multiple Linear Regression): - In MLR, there multiple independent variables. If multiple independent variables are use for predicting the dependent/Target variable, then regression is known as Multiple Linear Regression.

The Standard Notation in Regression is: -

$y = mX + c$ ⟹ $y = \beta_1 X + \beta_0$

Where y is Dependent variable (Target Variable), X is independent variable, β_1 is the slope, β_0 is the intercept

For getting the best fit line the error between the actual and the predicted values should be minimum. The best fit line has the minimum error. And so different value of slope and intercept gives different line of regression. So, for getting optimal values of slope and intercept cost function is used.

Cost function is used to get optimal slope and intercept for the best fit line.

$$J(m,c) = \frac{1}{\mathbb{N}} \sum_{i=1}^{n} (y_i - y_P)^2$$

Where $y_p = mx_i + c$, N = total number of observations, y_i = actual value.

The distance between the actual value and predicted values is called residual. If the scatter points are far from the regression line, then the residual will be high, and so cost function will high. If the scatter points are close to the regression line, then the residual will be small and hence the cost function.

2. Explain the Anscombe's quartet in detail.

Ans: - Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the 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. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

Hence, it says that before applying any algorithm it is a good practice to visualize the data which narrow downs the selection of algorithm to be applied and also by visualizing many insights can be gathered which might help in building the model and predictive analysis.

3. What is Pearson's R?

Ans: - Pearson's r is a numerical summary of the strength of the linear relationship 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.

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.

The main result of a correlation is called the correlation coefficient (or "r"). It ranges from -1.0 to +1.0. The closer r is to +1 or -1, the more closely the two variables are related.

If r is close to 0, it means there is no relationship between the variables. If r is positive, it means that as one variable gets larger the other gets larger. If r is negative, it means that as one gets larger, the other gets smaller.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

Ans: - Scaling is a step which is done before building the model. It is applied on the independent variables to normalize the variables within a particular range of 0 to 1. It also helps in speeding up the calculation of algorithm.

Scaling is done because most of the times some independent variables have values of high magnitude and ranges. Scaling brings all the independent variables into same range and level of magnitude. If scaling is not done the algorithm may work on the values but the prediction made will be useless because it will not be accurate.

Normalized Scaling brings all the values in the range of 0 and 1 while standardization replaces all the values by their z-scores and so basically it brings the data into standard normal distribution with mean (μ) 0 and standard deviation one (σ).

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

Ans: - If predictors are in perfect corelation the VIF of the corresponding variables will be infinity. These happens because there is strong multi collinearity between the independent variables. And so Infinite VIF indicates that the corresponding maybe represented exactly by the Model because of some other variables having the same multicollinearity.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**

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. A 45-degree reference line is also plotted.

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.

Following things can be interpreted from Q-Q plot: -

- Distribution: If all points of the quantiles lie on the line or near the straight line which is at 45 degree from x-axis.
- If x-values>y-values
- If y-values>x-values.