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:

- **Season**: Demand of rental bike is highest during 'fall' season followed by 'summer', then 'winter' and least during 'spring'
- **Year**: There is an increase in demand, on year-on-year basis (from 2018 to 2019)
- **Weather Situation**: People tend to rent bike in a good weather situation
- **Month**: Demand increases from January to June, then it remains a bit steady or there is a slow increase, September been the highest, followed by October. Demand decreases drastically from October to November and December
- **Weekday**: Not getting much insights from weekday column as the medians for all the days are not deviating much, although Wednesdays and Saturdays has a wider range from low to high demand.
- **Holiday**: Demand is mostly on the higher side during non-holidays, but highest demand remains the same for holidays and no-holidays.
- **Workingday**: Not getting much insights from workingday column as the medians are around the same range, with mostly higher demand during working day.

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

Ans:

drop_first=True, prevents from creating a redundant categorical variable which can actually be represented with the help other 2 variables. It reduces the number of variables, which helps in avoiding extra multicollinearity issues.

For any categorical variable with n level, n-1 variables are enough to represent all n levels as 0 in each variable means it is actually the last available level for which we don't have the variable.

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

Ans:

'temp' has the highest correlation with 'cnt' (target variable). 'atemp' also has the same corelation with 'cnt', but 'temp' and 'atemp' are highly corelated, so 'atemp' can be ignored.

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

Ans:

Performed following steps for validating the assumptions:

- Normal Distribution: Created a distplot of error terms, to validate that error terms are normally distributed with mean 0. As per the residual plot of error term it can be observed that error terms are independent of each other, there is no pattern observed
- Independent: Created residplot of error terms, to validate that error terms are independent of each other, i.e. they do not follow any pattern.
- Error terms has constant Variance: Created regplot on 'y_train' and 'y_train_pred' data, Heteroscedasticity is not observed in the error terms so we can say that it is truly homoscedastic

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:

Top 3 significant features are as follows:

- '**temp**' (Temperature): It has the highest coefficient value '0.5489', so it has a significant impact on demand of shared bikes.
- '**yr**' (Year): It has the coefficient value '0.2385' which is second highest in the list of variables.
- '**season_winter**' (Season): Winter season also have significant impact on demand, as its coefficient value is '0.1165' which is third highest in the list of variables.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans:

- Linear Regression is a machine learning algorithm based on supervised learning
- It is mostly used for finding out the relationship between variables and forecasting.
- Different regression models differ based on –
  I. the kind of relationship between dependent and independent variables considered
  II. the number of independent variables getting used
- Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x)
- Dependent variables are also called as an outcome variable, criterion variable, endogenous variable, or regressand.
- Independent variables are also called as an exogenous variable, predictor variables, or regressors

2. Explain the Anscombe's quartet in detail.

Ans:

- Anscombe's quartet is a set of four datasets that have nearly identical summary statistics, but very different visual patterns.
- 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.
- It also 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.
- Anscombe's quartet was constructed by statistician Francis Anscombe in 1973

3. What is Pearson's R?

Ans:

- The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation.
- It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.
- The Pearson correlation coefficient is a descriptive statistic, which summarises the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.
- The Pearson correlation coefficient is also an inferential statistic, which can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.
- Another way to think of the Pearson correlation coefficient (r) is as a measure of how close the observations are to a line of best fit.
- The Pearson correlation coefficient also tells you whether the slope of the line of best fit is negative or positive. When the slope is negative, r is negative. When the slope is positive, r is positive

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

Ans:

- What? Scaling is a data Pre-Processing step which is applied to independent variables to normalise the data within a particular range. It also helps in speeding up the calculations in an algorithm.
- Why? Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence result in incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. 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-squared, etc.
- Normalisation/Min-Max Scaling brings all of the data in the range of 0 and 1
- sklearn.preprocessing.MinMaxScaler helps to implement normalisation in python.
- Standardisation Scaling replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).
- sklearn.preprocessing.scale helps to implement standardisation in python.
- One disadvantage of normalisation over standardisation is that it loses some information in the data, especially about outliers.

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

Ans:

- If there is perfect correlation, then VIF = infinity.
- Perfect correlation leads to R2 = 1, then it will lead to infinity value of VIF as VIF = 1/1-R2
- A large value of VIF indicates that there is a correlation between the variables.
- A general rule of thumb is that if VIF > 10 then there is multicollinearity

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

Ans:

- Q-Q plots are also known as Quantile-Quantile plots.
- It plots the quantiles of a sample distribution against quantiles of a theoretical distribution.
- Doing this helps to determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential