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: As per our analysis form our Box plots we see below observation
   1) Summer and Fall seasons had high number of bikes rented
   2) Year 2019 has more bikes compared to 2018
   3) More number of bikes were rented on working day compared to holiday or weekend
   4) People rented bikes more on clear sky day

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

Ans: When we create dummy variables using pd.getdummies() , For each unique categorical value in the column one separate column is created

- drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

   Ans: Temp as highest corelation with target variable cnt (0.84) which we also were able to see in our heatmap

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

Ans:

- Verified the Linear relationship between target variable cnt and independent variable Temp, Atemp,season etc
- Verified that error terms are normally distributed with a distplot
- Verified that the homoscedastic between error terms with scatter plot as they were randomly distributed

- Verified Durbin-Watson (DW) statistic test. The values should fall between 0-4. If DW=2, no auto-correlation; if DW lies between 0 and 2, it means that there exists a positive correlation
- Verified No Multicollinearity - (Variance Inflation Factor). Where VIF<=4 implies no multicollinearity, whereas VIF>=10 implies serious multicollinearity

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 Features are

- Temp
- Yr_2019
- Month_sept

General Subjective Questions

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

Ans:
Linear Regression is machine learning algorithm based on supervised learning. It uses regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), which is called linear regression.

Linear regression uses a traditional formula $y = mx + b$ or $y = a0 + a1x$
y – Dependent variable
x – Independent variable
m – Slope of line
b – Intercept of the line
The cost function helps to figure out the best possible values for a0 and a1, which provides the best fit line for the data points.

In Linear Regression, Mean Squared Error (MSE) cost function is used, which is the average of squared error that occurred between the predicted values and actual values.

When we have more than one input we can use Ordinary Least Squares to estimate the values of the coefficients. The Ordinary Least Squares procedure seeks to minimize the sum of the squared residuals.

Gradient descent is a method of updating a0 and a1 to minimize the cost function (MSE). A regression model uses gradient descent to update the coefficients of the line (a0, a1 => xi, b) by reducing the cost function by a random selection of coefficient values and then iteratively update the values to reach the minimum cost function.

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

Ans:

Anscombe proved the importance of the graphs with some sample datasets

The basic thing to analyse about these data-sets is that they all share the same descriptive statistics like mean, standard deviation, variance etc but represented with different graphical representation.

Each graph plot shows the different behaviour compared to statistical analysis.

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties yet appear very different when graphed. Each dataset consists of eleven (x,y) points.

Francis Anscombe demonstrated the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

3. What is Pearson's R? (3 Marks)

Ans:

Correlation measures the strength of association between two variables as well as the direction. There are mainly three types of correlation that are measured. One significant type is Pearson's correlation coefficient. This type of correlation is used to measure the relationship between two continuous variables.

The Pearson's correlation coefficient varies between -1 and +1

Pearson r Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Requirements for Pearson's Correlation Coefficient:

- Scale of measurement should be interval or ratio
- Variables should be approximately normally distributed
- The association should be linear
- There should be no outliers in the data

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

Ans:

The 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.

In most of the times if scaling is not done, 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 so it will lead to incorrect modelling.

Normalisation	**Standardisation**
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
Scikit-Learn provides MinMaxScaler for Normalization.	Scikit-Learn provides StandardScaler for standardization

It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
Outlier are affected more	Outlier are affected less

Any data before passing as an parameter it should undergo a process of Standardising or Normalizing.

Standardizing :
Its used most widely in most of the models.
Its spreads data points between -3 to 4 or like that

Normalising/Scaling:
Its used less like in IMAGE PROCESSING for PIXEL intensities to be fix in 0 and 1
Scaling removes outliers by default as its getting placed between 0 to 1.

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

Ans :

When there is a perfect corelation between independent variables then the VIF = INFINITY. If multiple or more number of variables are corelated then the VIF will be more. VIF of more than 5 is considered redundant and should be dropped.

In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.

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

Ans:

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.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.