Linear Regression Subjective Questions

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)

season: Bike reservations are highest during the Summer and Fall season and least during the Spring season. There are more rentals in winter as compared to spring. Season can be a good predictor for the dependent variable.
weathersit: Higher bike rental observed when weather is more clear and sunny.There is no instance of rental during a HeavySnow/Rain condition .Weather can be a good predictor for the dependent variable.
mnth: Demand varies across months with lowest demand on january. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.
holiday: Bike booking are less when it is a holiday.This indicates, holiday may not be a good predictor for the dependent variable.
weekday: On weekdays median is between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. We will let the model decide if this needs to be added or not.
workingday: Majority of bike rental are happening on workingdays. However median of rentals on Working day or Non-working day are closer to 5000. This indicates, workingday can be a good predictor for the dependent variable
yr: Median bike rentals has increased in the year 2019, compared to 2018

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

Using drop_first=True during dummy variable creation is important.

- This avoid multicollinearity in regression models and improve the interpretability of the model.
- It automatically drops one of the dummy variables for each categorical variable. Dropping one dummy variable ensures that the number of dummy variables (k - 1) is equal to the number of categories (k) minus one, which is necessary for model estimation.
- The dropped dummy variable becomes the reference category for the categorical variable, making it easier to interpret the coefficients of the remaining dummy variables. The coefficients represent the difference in the effect of each category compared to the reference category.

In the dataset that some of the variables like 'weathersit' and 'season' have values as 1, 2, 3, 4 which have specific labels associated with them

	season_fall	season_spring	season_summer	season_winter
0	0	1	0	0
1	0	1	0	0
2	0	1	0	0

We can drop first column, as the type of season as spring can be identified with just the last three columns. if all 3 columns has value as zero, then it is spring season

	weather_Mist	weather_clear	weather_light_snow_rain
0	1	0	0
1	1	0	0
2	0	1	0

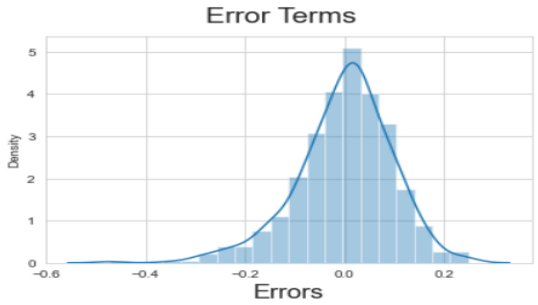
We can drop first column, as the type of whether as clear(Clear, Few clouds, Partly cloudy, Partly cloudy) can be identified with just the last two columns. if all 2 columns has value as zero, then it is clear weather.

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

This plots indicate a strong positive correlation of count with temperature – temp,atemp.
'Temp' is the temperature in Celsius and 'atemp' is the feeling temperature in Celsius

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

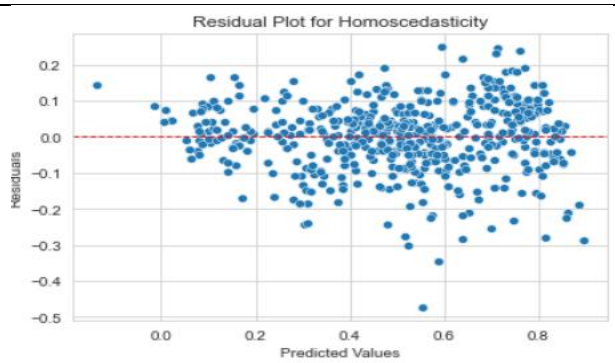
Plotted distplot of residuals to see error terms are normally distributed with mean zero. Below diagram shows it is normally distributed



Used the Durbin-Watson test to check for autocorrelation in the residuals. A value close to 2 indicates no significant autocorrelation.

```
Durbin-Watson Statistic: 1.9857584576363367
No significant autocorrelation detected.
```

Residuals vs. Fitted Values Plot has been used to check Homoscedasticity or consistent spread of residuals across all predicted values. Heteroscedasticity is indicated by a funnel shape.



Calculated Variance Inflation Factor (VIF) values for each independent variable. High VIF values (usually > 5 or 10) suggest 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)

As per our final Model, the top 3 predictor variables that influences the bike booking are:

- **Temperature (temp)** - A coefficient value of '0.3683' indicated that a unit increase in temp variable increases the bike hire numbers by 0.3683 units.
- **Year (yr)** - A coefficient value of '0.2463' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2463 units.
- **season_spring**- A coefficient value of '-0.1987' indicated that, w.r.t season, a unit increase in season_spring variable decreases the bike hire numbers by 0. 1987 units.

  Next top feature is weather

- **weather_light_snow_rain (weathersit_3)** - A coefficient value of '-0.1743' indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.1743 units.

General Subjective Questions

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

Linear regression is a supervised machine learning algorithm that finds the best linear-fit relationship on any given data, between a dependent variable (target) and one or more independent variables, by fitting a linear equation. It's widely used for tasks like prediction, forecasting, and understanding the relationships between variables. Equation is given below

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

Linear regression makes several assumptions, including:

- Linearity: The relationship between features and the target is assumed to be linear.
- Independence: The errors (residuals) are assumed to be independent.
- Homoscedasticity: The variance of the errors is assumed to be constant across all levels of the independent variables.
- No or Little Multicollinearity: The independent variables should not be highly correlated with each other.

There are different variants of linear regression, including:

- Simple Linear Regression: When there's only one independent variable.
- Multiple Linear Regression: When there are multiple independent variables.
- Ridge Regression and Lasso Regression: Regularized linear regression techniques to address multicollinearity and overfitting.
- Polynomial Regression: Allows modeling non-linear relationships by introducing polynomial terms.

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

Anscombe's Quartet is a famous statistical paradox that highlights the importance of visualizing data and the limitations of relying solely on summary statistics to understand a dataset. It consists of four small datasets that have nearly identical simple descriptive statistics but exhibit very different distributions when graphed. This paradox was created by the statistician Francis Anscombe in 1973 to emphasize the need for data exploration and visualization in addition to numerical summaries.

Anscombe's Quartet consists of four distinct datasets, each containing 11 data points (X, Y pairs):

Dataset I: This dataset represents a simple linear relationship. When graphed, it resembles a linear regression line with some random noise.

Dataset II: This dataset also represents a linear relationship but has an outlier that significantly affects the regression line.

Dataset III: This dataset shows a non-linear relationship between X and Y, forming a parabolic curve when graphed.

Dataset IV: This dataset contains an entirely different relationship where all X values are the same, but the corresponding Y values vary widely. It demonstrates the importance of examining the data distribution beyond just the summary statistics.

Anscombe's Quartet serves as a powerful reminder of several important concepts in data analysis:

Data Visualization: It underscores the importance of graphing data to visually explore its distribution and relationships. A summary statistic alone may not reveal the true nature of the data.
Outliers: Dataset II, with its outlier, illustrates how a single data point can heavily influence summary statistics and regression results. Identifying and handling outliers is crucial in data analysis.
Diversity of Patterns: Despite similar summary statistics, the datasets have entirely different patterns and relationships. This highlights that summary statistics can be deceiving and that examining data visually is essential for a comprehensive understanding.

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

Pearson's correlation coefficient (Pearson's R) is a widely used statistic for quantifying the strength and direction of the linear relationship between two continuous variables. It provides valuable insights into how variables are related and is a fundamental tool in statistical analysis and data exploration.

The sign of Pearson's R indicates the direction of the linear relationship:

Positive R suggests a positive linear relationship, meaning that as one variable increases, the other tends to increase.
Negative R suggests a negative linear relationship, meaning that as one variable increases, the other tends to decrease.
The magnitude of R indicates the strength of the linear relationship:

If R is close to 1 or -1, it indicates a strong linear relationship.
If R is close to 0, it suggests a weak or no linear relationship.
Pearson's R is sensitive to outliers. Outliers can have a significant impact on the value of R, potentially inflating or deflating it. Therefore, it's important to examine data for outliers and consider their influence when interpreting R.

Pearson's R is calculated using the following formula:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

Where:

- r is Pearson's correlation coefficient.
- X_i and Y_i are data points from the two variables.
- $\bar{X}$ and $\bar{Y}$ are the means (averages) of the respective variables.

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

Scaling in the context of data preprocessing refers to the process of transforming the values of variables (features) in a dataset to a common scale, typically within a specific range or distribution. Scaling is performed to address issues related to the magnitude of variables and to ensure that they have a similar influence on the analysis or machine learning models. There are two common types of scaling: normalized scaling and standardized scaling.

Scaling is performed for below reasons:

Magnitude Differences: Variables in a dataset may have different units or scales. When performing mathematical operations or using algorithms that rely on distances or magnitudes (e.g., gradient descent in machine learning), variables with larger scales can dominate the analysis or model training.

Algorithms that Use Distances: Many machine learning algorithms, such as k-means clustering or support vector machines, rely on measuring distances between data points. Scaling ensures that all variables contribute equally to these distance calculations.

Convergence and Stability: Scaling can help algorithms converge faster and be more stable during training. Gradient-based optimization algorithms, for example, tend to work better with scaled data.

Difference Between Normalized Scaling and Standardized Scaling is given below:

Range: Normalized scaling transforms data into a specific range (e.g., 0 to 1), while standardized scaling centers the data around 0 with a standard deviation of 1.

Preservation of Relative Relationships: Normalized scaling preserves the relative relationships and order of data points, while standardized scaling also preserves relative relationships but centers the data around zero.

Impact on Outliers: Normalized scaling can be influenced by outliers, as it uses the minimum and maximum values, which can be heavily affected by extreme values. Standardized scaling is less sensitive to outliers because it uses the mean and standard deviation.

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

A Variance Inflation Factor (VIF) can become infinite when there is perfect multicollinearity among the independent variables in a regression model. Perfect multicollinearity occurs when two or more independent variables in the model are perfectly linearly related, meaning that one variable can be expressed as an exact linear combination of the others. In such cases, it's impossible to estimate the unique contribution of each variable to the model because they are redundant.

$$VIF = \frac{1}{1 - R^2}$$

In the case of perfect multicollinearity, one of the independent variables can be expressed as a perfect linear combination of the others. For example, if you have two variables, X1 and X2, where X2 = a * X1 for some constant 'a,' then there is perfect multicollinearity.
When the regression model attempts to calculate the VIF for a variable involved in perfect multicollinearity, it finds that R squared is equal to 1.

Plugging $R^2 = 1$ into the VIF formula:

$$VIF = \frac{1}{1-1} = \frac{1}{0}$$

Since you cannot divide by zero, the VIF becomes infinite in the presence of perfect multicollinearity.

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

Q-Q plots are valuable tools in linear regression for checking the normality assumption of residuals. They help identify deviations from normality, outliers, and potential issues with the model's assumptions. Addressing non-normality in residuals can lead to more reliable and accurate regression analysis results.

Use of Q-Q plot:

Distribution Assessment: Q-Q plots are primarily used to visually assess the distribution of a dataset. They help you determine if the data follows a known theoretical distribution, like the normal distribution.

Identification of Deviations: Q-Q plots can reveal deviations from the expected distribution. If the data points deviate significantly from a straight line in the plot, it suggests that the data does not follow the expected distribution.

Importance in Linear Regression:

In the context of linear regression, Q-Q plots are important for below reasons:

Normality Assumption: Linear regression models often assume that the residuals (the differences between observed and predicted values) are normally distributed. This assumption is important for the validity of statistical tests and confidence intervals associated with the regression coefficients.

Residual Analysis: After fitting a linear regression model, you can create a Q-Q plot of the residuals to check if they follow a normal distribution. If the residuals are normally distributed, the Q-Q plot will closely resemble a straight line.

Detecting Non-Normality: A departure from normality in the residuals can be detected using the Q-Q plot. Deviations from the straight line in the Q-Q plot indicate non-normality in the residuals. Non-normality may suggest that the model assumptions are not met, and it can affect the reliability of the model's results and predictions.

Outlier Detection: Q-Q plots can also help in identifying outliers. Outliers can cause deviations from the expected distribution, which would be evident in the Q-Q plot as points far away from the line.

Model Improvements: If the Q-Q plot reveals non-normality in the residuals, it may prompt you to explore model improvements. You can consider transformations on the dependent variable or applying robust regression techniques to address non-normality.