Assignment-based Subjective Questions – Anjana Mohandas

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?** (3 marks)

    - Fall season seems to have attracted a greater number of bookings. And, the booking count has increased drastically from year 2018 to 2019.
    - Most of the bookings has been done during the months of May, June, July, August, September and October.
    - Trend increased from the beginning of the year till mid and then it started decreasing as we approached the end of year.
    - Clear weather attracted more bookings which seems obvious.
    - Thursday, Friday, Saturday and Sunday have a greater number of bookings as compared to other days of the week.
    - During weekdays and working days, booking seems to be less in number which seems reasonable.
    - Year 2019 attracted a greater number of bookings compared to previous year, which shows good progress in terms of business.

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

    "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.

    Syntax – "drop_first= bool, default False", which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

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

    "temp" variable has the highest correlation with the target variable.

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

    Residual Analysis:
    Calculate the residuals (the differences between the predicted and actual values) for the training set.
    Plot a histogram of the residuals to check if they are approximately normally distributed. Normality of residuals is a key assumption of linear regression.
    Residual vs. Fitted Values Plot:
    Create a scatterplot of the residuals against the predicted values (fitted values). Look for patterns in the residuals. Ideally, there should be no clear pattern; residuals should be randomly scattered around zero.

Residuals vs. Predictor Variables:
Create scatterplots of residuals against each predictor variable to check for any systematic patterns or heteroscedasticity (variance of residuals changing with predictor values). A cone-shaped pattern in these plots may indicate heteroscedasticity.
Multicollinearity:
Assess multicollinearity among predictor variables by calculating correlation coefficients or variance inflation factors (VIFs). High multicollinearity can affect the model's stability and interpretation.
Homoscedasticity Test:
Perform formal tests for homoscedasticity, such as the Breusch-Pagan or White test. These tests help determine if the variance of the residuals is constant across all levels of the predictors.
Linearity of Relationships:
Check whether the relationships between the predictor variables and the dependent variable are linear. You can use scatterplots and residual plots to assess linearity.
Cross-Validation:
Perform cross-validation on the model to assess its predictive performance on unseen data. Cross-validation can help you detect overfitting or underfitting.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2marks)**

1. Temp
2. Winter
3. September

General Subjective Questions

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

Linear regression is a supervised machine learning algorithm used for modeling the relationship between a dependent variable (target) and one or more independent variables (features or predictors) by fitting a linear equation to the observed data. The goal is to find the best-fitting line (or hyperplane in multiple dimensions) that minimizes the sum of squared differences between the predicted and actual values. This line represents the linear relationship between the variables. Here's a detailed explanation of the linear regression algorithm:

1. Model Representation:

Linear regression assumes a linear relationship between the dependent variable (Y) and the independent variable(s) (X). For simple linear regression (one independent variable), the model can be represented as:

$Y = \beta0 + \beta1 * X + \varepsilon$

Y is the dependent variable (the target you want to predict).

X is the independent variable (the feature used for prediction).

β0 is the intercept (the point where the line intersects the Y-axis).

β_1 is the slope (the change in Y for a unit change in X).

ε is the error term (the part of Y that cannot be explained by the linear relationship with X).

2. Objective:

The objective of linear regression is to find the values of β_0 and β_1 that minimize the sum of squared errors (SSE) or the mean squared error (MSE) between the predicted values (Y_{pred}) and the actual values (Y_{true}).

3. Training:

During the training phase, the algorithm learns the optimal values of β_0 and β_1 from the training data.

This is typically done using a mathematical optimization technique like the method of least squares.

4. Least Squares Method:

In simple linear regression, the least squares method finds β_0 and β_1 that minimize the sum of squared residuals (the vertical distance between data points and the regression line).

Mathematically, the coefficients are calculated as:

$$\beta_1 = (\Sigma((X_i - \bar{X})(Y_i - \bar{Y}))) / \Sigma((X_i - \bar{X})^2)$$

$$\beta_0 = \bar{Y} - \beta_1 * \bar{X}$$

Σ represents summation over all data points.

$\bar{X}$ is the mean of the independent variable.

$\bar{Y}$ is the mean of the dependent variable.

5. Making Predictions:

After training, the model can make predictions by plugging new or unseen values of X into the linear equation:

$$Y_{pred} = \beta_0 + \beta_1 * X$$

6. Model Evaluation:

The model's performance is evaluated using various metrics, such as the coefficient of determination (R-squared), mean squared error (MSE), or root mean squared error (RMSE), to assess how well it fits the data.

7. Assumptions:

Linear regression assumes that the relationship between the variables is linear.

It assumes that the errors (ε) are normally distributed and have constant variance (homoscedasticity).

It assumes that the errors are independent of each other (independence of residuals).

8. Extensions:

Linear regression can be extended to multiple independent variables, creating multiple linear regression.

When dealing with categorical predictors, dummy variables can be used.

Various regularization techniques, such as Ridge and Lasso regression, can be applied to prevent overfitting.

Linear regression is a fundamental and widely used technique in statistics and machine learning for tasks like prediction, inference, and understanding relationships between variables.

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

Anscombe's quartet is a set of four small datasets that have nearly identical simple descriptive statistics, yet they exhibit significantly different characteristics when graphically analyzed. This quartet was created by the statistician Francis Anscombe in 1973 to emphasize the importance of data visualization and the limitations of relying solely on summary statistics.

The Datasets:
Anscombe's quartet consists of four distinct datasets, each containing 11 data points (x, y pairs):

Dataset I: This dataset has a clear linear relationship between X and Y, following the equation Y = 3X + 2. It is a typical example of linear regression.

Dataset II: Unlike Dataset I, Dataset II also shows a linear relationship but with an outlier. One data point has an unusually high Y value, affecting the regression line's slope and intercept.

Dataset III: Dataset III consists of points that form a non-linear relationship. It's a clear example of how linear regression may not be appropriate for all datasets. A linear model here would not fit the data well.

Dataset IV: This dataset has no apparent relationship between X and Y, as the points are dispersed without any clear pattern. It illustrates that even when the correlation coefficient is close to zero, other relationships may exist.

Summary Statistics:
Despite their differences in data distribution and relationships, all four datasets have nearly identical summary statistics:

Mean of X: ~9.0
Variance of X: ~11.0
Mean of Y: ~7.5
Variance of Y: ~4.12
Correlation coefficient (r): ~0.816

In summary, Anscombe's quartet serves as a reminder of the importance of data visualization and the potential pitfalls of relying solely on summary statistics when analyzing datasets.

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

Pearson's correlation coefficient, often denoted as "r," is a statistical measure of the strength and direction of the linear relationship between two continuous variables. It quantifies how well the relationship between these two variables can be described by a straight-line (linear) equation. Pearson's R ranges from -1 to 1, where:

r = 1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other also increases linearly.
r = -1 indicates a perfect negative linear relationship, meaning that as one variable increases, the other decreases linearly.
r = 0 indicates no linear relationship between the variables. They are not correlated.
Key points about Pearson's correlation coefficient:

Formula:
Pearson's R is calculated using the following formula:

$$r = \frac{\sum(x_i - x)\sum(y_i - y)}{\sum(x_\iota - x)^2\sum(y_\iota - y)^2}$$

where: *xi* and *yi* are individual data points.
x and *y* are the means of the X and Y variables, respectively.

Range:
As mentioned earlier, the range of Pearson's R is from -1 to 1, where values closer to -1 or 1 indicate stronger linear relationships, and values closer to 0 indicate weaker or no linear relationship.

Interpretation:
A positive value of r indicates a positive correlation, meaning that as one variable increases, the other tends to increase.
A negative value of r indicates a negative correlation, meaning that as one variable increases, the other tends to decrease.
The magnitude (absolute value) of r reflects the strength of the correlation. The closer it is to 1 (positive or negative), the stronger the linear relationship.
Correlation measures only the strength and direction of linear relationships. It may miss important associations that are not linear.
The presence of outliers can strongly influence the value of r.

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

Scaling is a data preprocessing technique used in statistics and machine learning to transform the values of variables so that they all fall within a similar numerical range.

Scaling is performed to ensure that variables with different units or magnitudes do not disproportionately influence the outcome of certain algorithms or analyses. It helps bring consistency to the data and can improve the performance and convergence of various machine learning algorithms.

Two common scaling methods are **normalized scaling and standardized scaling**.

Normalized Scaling:

- Normalization, also known as min-max scaling, scales the data to a specific range, typically [0, 1].
- Normalization preserves the relative differences between data points but can be sensitive to outliers.

Standardized Scaling (Z-score normalization or standardization):

- Standardization scales the data to have a mean of 0 and a standard deviation of 1.
- Standardization centres the data around zero and scales it by the standard deviation. It is less affected by outliers compared to normalization.

Differences between Normalized Scaling and Standardized Scaling:

Normalized scaling	Standardized scaling
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.
It is really affected by outliers.	It is much less affected by outliers.
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

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

A Variance Inflation Factor (VIF) is a measure used in regression analysis to assess the extent of multicollinearity among predictor variables. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, making it difficult to determine their individual effects on the dependent variable. VIF quantifies how much the variance of the estimated regression coefficients is inflated due to multicollinearity.

In some cases, you may observe that the VIF is calculated as infinite (or very large) for a particular predictor variable. This happens for the following reasons:

1. **Perfect Multicollinearity**: The most common reason for infinite VIF is perfect multicollinearity. Perfect multicollinearity occurs when one or more predictor variables in the regression model can be exactly predicted from other variables in the model. In other words, there is a linear dependency among the predictor variables.
    - For example, if you have two predictor variables, X1 and X2, and X2 is a linear combination of X1 (e.g., X2 = 2 * X1), then the VIF for X2 would be infinite because X2 can be perfectly predicted from X1.
2. **Linear Dependence**: Even if the multicollinearity is not perfect, but there is a very high degree of linear dependence among predictor variables, it can lead to extremely high VIF values. This indicates that the variance of the estimated coefficients for these variables is greatly inflated due to multicollinearity.

3. **Data Issues**: In some cases, data issues like duplicate records or coding errors can lead to inflated VIF values. If you have duplicate or nearly identical records for a particular variable, it can cause problems in VIF calculations.
4. **Algorithmic Limitations**: Some software or libraries used for calculating VIF may have limitations or may not handle certain scenarios well, leading to unusual or infinite VIF values. This is less common but possible.

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

(3 marks)

A Q-Q plot, short for Quantile-Quantile plot, is a graphical tool used in statistics to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. It compares the quantiles (percentiles) of the observed data to the quantiles of the specified theoretical distribution**.**

Use and Importance of Q-Q Plots in Linear Regression:

1. **Normality Assumption:**
   Q-Q plots are used to visually check whether the residuals follow a normal distribution. If the points in the Q-Q plot closely follow a diagonal line, it suggests that the residuals are normally distributed.
2. **Detecting Departures from Normality:**
   In a Q-Q plot, if the points deviate significantly from the diagonal line, it indicates a departure from normality. Departures can include skewness, heavy tails, or other non-normal characteristics.
3. **Outlier Detection:**
   Q-Q plots can also help identify outliers in the data. Outliers may appear as points that deviate substantially from the expected quantiles.
4. **Model Assessment:**
   Q-Q plots can be used not only to check the normality of residuals but also to assess the appropriateness of the chosen regression model. Deviations from the expected diagonal line may suggest model misspecification.
5. **Model Improvement:**
   If a Q-Q plot reveals non-normality in the residuals, it may prompt the need for data transformation or consideration of alternative modelling techniques that can better accommodate non-normal errors.
6. **Communicating Results:**
   Q-Q plots are a valuable tool for visualizing and communicating the distributional characteristics of residuals to stakeholders. They provide an intuitive way to convey information about the quality and reliability of a linear regression model.