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)

I have done analysis on categorical columns using the boxplot and bar plot. Below are the few points we can infer from the visualization –

- Fall season seems to have attracted more booking. And, in each season the booking count has increased drastically from 2018 to 2019.
- Most of the bookings has been done during the month of may, june, july, aug, sep and oct. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.
- Clearly, weather attracted more booking which seems obvious.
- Thu, Fir, Sat and Sun have more number of bookings as compared to the start of the week.
- When it's not holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family.
- Booking seemed to be almost equal either on working day or non-working day.
- 2019 attracted more number of booking from the 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.
Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not A and B, then It is obvious C. So we do not need 3rd variable to identify the C.

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)

   I have validated the assumption of Linear Regression Model based on below 5 assumptions -

- Normality of error terms
  - Error terms should be normally distributed
- Multicollinearity check
  - There should be insignificant multicollinearity among variables.
- Linear relationship validation
  - Linearity should be visible among variables
- Homoscedasticity
  - There should be no visible pattern in residual values.
- Independence of residuals
  - No auto-correlation

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

Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –

- temp
- winter
- September month

General Subjective Questions

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

Linear regression is a statistical method used to find the relationship between two variables, typically a dependent variable and one or more independent variables. The goal is to find a linear equation that best predicts the value of the dependent variable based on the values of the independent variables.

The linear regression algorithm works by fitting a straight line to the data points in such a way that the differences between the actual values and the predicted values (residuals) are minimized. This line is represented by the equation:

[y = mx + b]

where:

- (y) is the dependent variable we are trying to predict,
- (x) is the independent variable,
- (m) is the slope of the line, which represents how much (y) changes for a unit change in (x),
- (b) is the y-intercept of the line, which represents the value of (y) when (x) is zero.

The algorithm calculates the values of (m) and (b) that minimize the sum of the squared differences between the actual values of the dependent variable and the values predicted by the linear equation. This process is called "least squares regression."

Once the best-fitting line is found, it can be used to make predictions about the dependent variable based on new values of the independent variable.

In summary, linear regression is a simple and powerful algorithm used to find and quantify the relationship between variables by fitting a straight line to the data points.

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

Anscombe’s quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

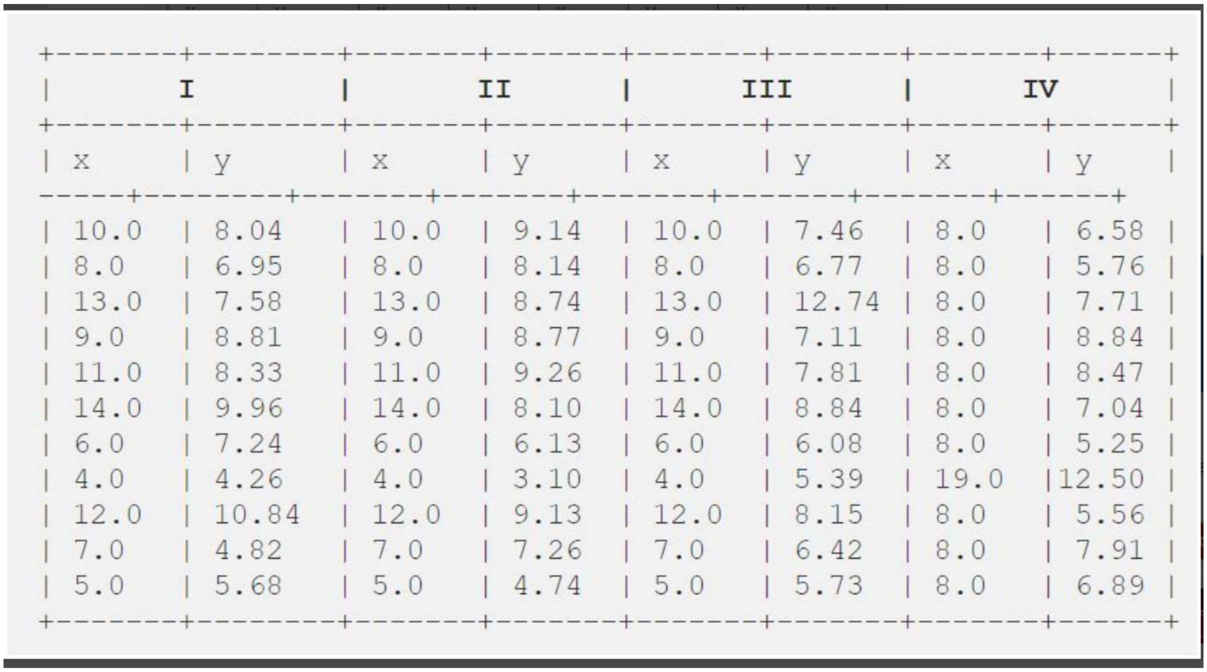
The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The four datasets that make up Anscombe’s quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Purpose of Anscombe’s Quartet

Anscombe’s quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

The four datasets of Anscombe’s Quartet:



```
+-------+--------+-------+-------+-------+-------+-------+------+
|       I        |       II      |      III      |      IV      |
+-------+--------+-------+-------+-------+-------+-------+------+
| x     | y      | x     | y     | x     | y     | x     | y    |
-----+--------+-------+-------+-------+-------+-------+------+
| 10.0  | 8.04   | 10.0  | 9.14  | 10.0  | 7.46  | 8.0   | 6.58 |
| 8.0   | 6.95   | 8.0   | 8.14  | 8.0   | 6.77  | 8.0   | 5.76 |
| 13.0  | 7.58   | 13.0  | 8.74  | 13.0  | 12.74 | 8.0   | 7.71 |
| 9.0   | 8.81   | 9.0   | 8.77  | 9.0   | 7.11  | 8.0   | 8.84 |
| 11.0  | 8.33   | 11.0  | 9.26  | 11.0  | 7.81  | 8.0   | 8.47 |
| 14.0  | 9.96   | 14.0  | 8.10  | 14.0  | 8.84  | 8.0   | 7.04 |
| 6.0   | 7.24   | 6.0   | 6.13  | 6.0   | 6.08  | 8.0   | 5.25 |
| 4.0   | 4.26   | 4.0   | 3.10  | 4.0   | 5.39  | 19.0  |12.50 |
| 12.0  | 10.84  | 12.0  | 9.13  | 12.0  | 8.15  | 8.0   | 5.56 |
| 7.0   | 4.82   | 7.0   | 7.26  | 7.0   | 6.42  | 8.0   | 7.91 |
| 5.0   | 5.68   | 5.0   | 4.74  | 5.0   | 5.73  | 8.0   | 6.89 |
+-------+--------+-------+-------+-------+-------+-------+------+
```

3. What is Pearson’s R? (3 marks)

Pearson's correlation coefficient, often denoted as (r), is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1, where:

- A value of 1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable also increases proportionally.
- A value of -1 indicates a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases proportionally.
- A value of 0 indicates no linear relationship between the variables.

Pearson's (r) is calculated by dividing the covariance of the two variables by the product of their standard deviations. It is commonly used to assess the association between variables and to determine the strength and direction of the relationship.

Pearson's correlation coefficient is a widely used statistic in various fields such as psychology, economics, biology, and more, to measure the degree of association between two variables. It provides valuable insights into how changes in one variable are associated with changes in another variable.

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

➢ What is scaling?

Scaling is a preprocessing technique used in data analysis and machine learning to change the range of feature values. It's the process of transforming the data into a specific scale, typically to ensure that all features contribute equally to the analysis or to improve the performance of machine learning algorithms.

➢ Why is scaling performed?

Scaling is performed for several reasons:
a) Feature uniformity: It brings all features to a similar scale, preventing features with larger numeric ranges from dominating those with smaller ranges.
b) Algorithm performance: Many machine learning algorithms perform better or converge faster when features are on a similar scale.
c) Gradient descent optimization: For algorithms that use gradient descent, scaling can help reach convergence faster.
d) Distance-based algorithms: For algorithms that use distances between samples (like K-Nearest Neighbors), scaling ensures all features contribute equally to the distance calculation.
e) Improved interpretability: In some cases, it makes it easier to interpret the importance of features.

➢ Difference between normalized scaling and standardized scaling:

a) Normalized Scaling (Min-Max Scaling):

- Scales features to a fixed range, usually 0 to 1.
- Formula: X_scaled = (X - X_min) / (X_max - X_min)
- Preserves zero values and doesn't center the data.
- Useful when you need values in a bounded interval.

b) Standardized Scaling (Z-score Normalization):

- Transforms data to have a mean of 0 and a standard deviation of 1.
- Formula: X_scaled = (X - μ) / σ, where μ is the mean and σ is the standard deviation.
- Centers the data around zero and gives it unit variance.
- Useful when you need to compare features that have different units or scales.

Key differences:

- Output range: Normalization gives values between 0 and 1, while standardization can give values outside this range.
- Effect of outliers: Normalization is more affected by outliers as it uses the min and max values. Standardization is less affected as it uses mean and standard deviation.
- Data distribution: Standardization assumes your data follows a normal distribution, while normalization doesn't make this assumption.
- Interpretability: Normalized data is often easier to interpret as it's on a fixed scale. Standardized data tells you how many standard deviations a value is from the mean.

The choice between normalization and standardization depends on the specific requirements of your analysis or machine learning algorithm, the nature of your data, and the goals of your project.

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

Variance Inflation Factor (VIF) is a measure used to assess multicollinearity in regression analysis.

High VIF values indicate that a predictor variable is highly correlated with other predictor variables in the model, which can cause issues with the interpretation and stability of the regression coefficients.
When the VIF is calculated to be infinite for a particular predictor variable, it indicates perfect multicollinearity. Perfect multicollinearity occurs when one or more of the predictor variables in the regression model can be exactly predicted by a linear combination of the other predictor variables. In other words, one predictor variable can be expressed as a perfect linear function of the other predictor variables.
This perfect multicollinearity leads to numerical instability in the estimation of regression coefficients, as the model cannot distinguish the unique contribution of the perfectly collinear variables. As a result, the VIF value for the variable involved in perfect multicollinearity becomes infinite, indicating the presence of this issue.
In practical terms, when you encounter a situation where the VIF is calculated to be infinite, it is a signal that there is a serious problem of perfect multicollinearity in the regression model. This issue needs to be addressed by either removing one of the perfectly collinear variables or by using techniques like variable selection or regularization to mitigate the multicollinearity problem and stabilize the regression model.

(3 marks)

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

A Q-Q plot, which stands for quantile-quantile plot, is a graphical tool used to assess whether a dataset follows a particular distribution, such as a normal distribution. In a Q-Q plot, the quantiles of the dataset are plotted against the quantiles of a theoretical distribution. If the points in the plot fall approximately along a straight line, it indicates that the dataset is close to the specified distribution.

In linear regression, Q-Q plots are commonly used to evaluate the assumption of normality of the residuals. The residuals are the differences between the observed values and the predicted values from the regression model. The normality of residuals is an important assumption in linear regression because many statistical tests and confidence intervals rely on the residuals following a normal distribution.

The use and importance of a Q-Q plot in linear regression are as follows:

- **Assessing Normality**: By examining the Q-Q plot of the residuals, you can visually inspect whether the residuals are normally distributed. If the points in the plot deviate significantly from a straight line, it suggests that the assumption of normality may be violated.

- **Identifying Outliers**: Q-Q plots can also help identify outliers in the residuals. Outliers are data points that do not follow the general trend of the dataset and can have a significant impact on the regression model's performance.

- **Model Evaluation**: A well-fitting linear regression model should have residuals that are normally distributed. If the Q-Q plot shows deviations from a straight line, it may indicate that the model needs further refinement or that the data may not meet the assumptions of linear regression.

- **Assumption Checking**: Q-Q plots are an essential tool for checking the assumptions of linear regression, such as linearity, homoscedasticity, and normality of residuals. They provide a visual and intuitive way to assess the validity of these assumptions.

In summary, Q-Q plots play a crucial role in linear regression by helping to evaluate the normality of residuals, identify outliers, assess model fit, and check the underlying assumptions of the regression analysis. They provide valuable insights into the distribution of residuals and the overall quality of the regression model.

(3 marks)