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?

- Bookings are more during fall season.
- Bookings are more during the months of May, June, July, Aug, Sep and Oct
- Bookings are more during clear weathersit.
- Bookings are less during holidays
- Overall Bookings are more in 2019
- There is a linear relationship between temp and atemp variables

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

- drop_first=True drops the first column during dummy variable creation.
- Dummy variables will create as many variables as the categories of the feature. For instance, if a feature has p categories, it creates p number of variables assigning 0 and 1.
- Since we have only 0s and 1s, we can drop one variable since we need only p-1 variables for analysis

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

- Temp has the highest correlation

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

- Linear relationship validation
    - There should be a linearity between features and dependent variable
- Multicollinearity between the features
    - There should be insignificant multicollinearity among variables. Temp and atemp has multicollinearity, which were dropped during analysis
- Homoscedasticity:
    - No specific pattern was found during residual analysis
- Normal distribution of error terms
    - Error terms should be normally distributed
- Independence of residuals:
    - The value of Durbin-Watson test is 2.019 quite close to 2. This confirms that, residuals are not correlated

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

- Season - spring
- Weathersit - Light_snowrain
- Year

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a fundamental statistical method used for modeling the relationship between a dependent variable and one or more independent variables. It's called "linear" because the relationship between the dependent variable (often denoted as y) and the independent variable(s) (often denoted as x) is assumed to be linear, meaning it can be represented by a straight line. Here's a detailed explanation of how linear regression works:

- Model Representation:
  In its simplest form, linear regression with one independent variable can be represented as:
  $y = c + mX + \epsilon$
  - y is the dependent variable.
  - x is the independent variable.
  - c is the intercept (the value of y when x=0).
  - m is the slope (the change in y for a unit change in x).
  - ϵ represents the error term, which captures the difference between the actual value of y and the value predicted by the model. It includes all other factors not included in the model that affect y.
- Objective:
  The goal of linear regression is to find the best-fitting line through the data points. "Best-fitting" typically means minimizing the sum of the squared differences between the observed dependent variable values and the values predicted by the linear model. This method is called the method of least squares.
- Estimating Parameters:
  To find the best-fitting line, we need to estimate the parameters c and m. This is usually done using the ordinary least squares (OLS) method. In OLS, the parameters are chosen to minimize the sum of the squared residuals (the differences between the observed and predicted values).

- Fitting the Model:
  Once the parameters are estimated, we have the equation of the best-fitting line. We can use this equation to make predictions for new values of x.
- Assessing the Fit:
  It's essential to assess how well the linear regression model fits the data. Common metrics for this include the coefficient of determination (R^2), which measures the proportion of the variance in the dependent variable that is predictable from the independent variable(s), and visual inspection of residual plots.
- Assumptions:
  Linear regression relies on several assumptions, including:
  - Linearity: The relationship between the dependent and independent variables is linear.
  - Independence: Observations are independent of each other.
  - Homoscedasticity: The variance of the errors is constant across all levels of the independent variables.
  - Normality: The errors are normally distributed with mean zero.
- Extensions:
  Linear regression can be extended to handle more complex relationships by including multiple independent variables (multivariate regression) or by incorporating polynomial terms, interaction terms, or other transformations of the variables.

Overall, linear regression is a powerful and widely used technique for understanding and predicting relationships between variables, but it's crucial to ensure that its assumptions are met and to interpret its results carefully.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a famous example in statistics illustrating the importance of visualizing data and the potential pitfalls of relying solely on summary statistics. It consists of four datasets that have nearly identical descriptive statistics (e.g., mean, variance, correlation) but exhibit vastly different relationships when plotted. This quartet was introduced by the statistician Francis Anscombe in 1973 to emphasize the necessity of graphical exploration in data analysis. Here's a detailed explanation of each dataset within Anscombe's quartet:

Dataset I:

Characteristics:

- The relationship between x and y is linear.
- There's a strong positive correlation.
- The data points are relatively closely clustered around the line of best fit.

Implications:

- Summary statistics alone (e.g., mean, variance, correlation coefficient) might suggest a straightforward linear relationship, which could lead to the mistaken assumption that linear regression is appropriate for modeling the data.

Dataset II:

Characteristics:

- The relationship between x and y is also linear but with a different slope.
- There's a strong positive correlation.
- One outlier significantly affects the line of best fit.

Implications:

- The presence of outliers can have a substantial impact on regression analysis, potentially leading to misleading conclusions if not properly addressed.

Dataset III:

Characteristics:

- The relationship between x and y is non-linear.
- There's a perfect linear relationship between the two variables except for one outlier.
- The outlier skews the summary statistics but not the graphical representation.

Implications:

- A single outlier can distort summary statistics, such as the correlation coefficient, but examining the data visually can reveal the true nature of the relationship.

Dataset IV:

Characteristics:

- There's no apparent relationship between x and y except for one outlier.
- Removing the outlier results in a perfect fit.

Implications:

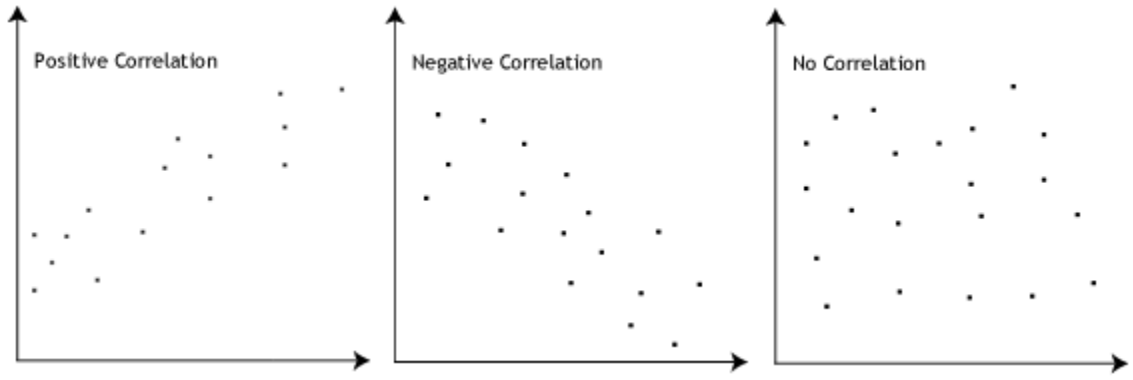
- The presence of influential data points, like outliers, can obscure meaningful patterns in the data and lead to erroneous conclusions if not identified and handled properly.

In summary, Anscombe's quartet demonstrates that summary statistics alone may not provide a comprehensive understanding of the relationships within a dataset. Visualizing the data through plots and graphs is crucial for identifying patterns, outliers, and the appropriateness of statistical models. It serves as a cautionary tale for statisticians and data analysts to exercise diligence and thoroughness in exploratory data analysis.

3. What is Pearson’s R?

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with the high values of the other, the correlation coefficient will be negative.

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



Pearson's R is commonly used in various fields, including statistics, social sciences, economics, and natural sciences, to assess the strength and direction of relationships between variables.
It is often used in exploratory data analysis to identify potential associations between variables and in hypothesis testing to assess the significance of those associations.

Pearson's correlation coefficient is a widely used and valuable tool for understanding relationships between continuous variables, but it's essential to remember that correlation does not imply causation. Additionally, Pearson's R measures only linear relationships and may not capture more complex associations between variables.

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

Feature scaling, also known as data normalization, is a preprocessing technique used in machine learning to standardize the range of independent variables or features of a dataset. The goal of feature scaling is to ensure that all features have the same scale, typically within a certain range or distribution.

This process is essential because many machine learning algorithms perform better or converge faster when the input features are on a similar scale. Feature scaling helps prevent features with larger magnitudes from dominating those with smaller magnitudes and can improve the numerical stability of certain algorithms.

No	Normalization	Standardization
1	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2	Scales values between [0, 1] or [-1, 1].	It is not bound to a certain range.
3	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
4	It is useful when we don’t know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
5	It is often called as Scaling Normalization	It is often called as Z-Score Normalization

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

When the VIF value becomes infinite, it typically indicates perfect multicollinearity. Perfect multicollinearity occurs when one or more independent variables in the regression model can be expressed as a perfect linear combination of the other independent variables. In other words, one or more independent variables become redundant and provide no additional information to the model beyond what is already provided by the other variables.

Here are some common scenarios where perfect multicollinearity might occur:

- Duplicate Variables: If two or more variables in the dataset are exactly the same or are perfect linear combinations of each other, perfect multicollinearity will arise.
- Dummy Variable Trap: In regression models with dummy variables representing categorical variables, perfect multicollinearity can occur when one dummy variable can be predicted perfectly from the others. This often happens when all, but one category is represented by the dummy variables, resulting in a perfect linear relationship among them.
- Data Preparation Issues: Perfect multicollinearity can also occur due to errors or inconsistencies in data preprocessing. For example, if a variable is mistakenly duplicated or transformed into another variable that is perfectly correlated with it, this can lead to infinite VIF values.

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

The quantile-quantile (Q-Q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not. Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution. They are commonly used in statistics, data analysis, and quality control to check assumptions and identify departures from expected distributions.

Use of Q-Q plot:
It is used to check following scenarios:
If two data sets:

1. come from populations with a common distribution
2. have common location and scale
3. have similar distributional shapes
4. have similar tail behavior

Importance of Q-Q plot:

Q-Q plots provide a visual and intuitive way to assess the goodness-of-fit of the model. If the Q-Q plot shows a clear departure from the straight line, it suggests that the model assumptions may not hold, and adjustments to the model may be necessary.