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)

Key Findings based on categorical variables:

- **Seasonal Trends:** The fall season experienced a significant increase in bookings, with a notable uptick in bookings observed across all seasons from 2018 to 2019.
- **Monthly Patterns:** The months of May, June, July, August, September, and October consistently had the highest booking volumes. Booking activity demonstrated a gradual increase from the beginning of the year, peaking in mid-year before declining towards the year-end.
- **Weather Influence:** Favorable weather conditions were correlated with higher booking numbers, as expected.
- **Weekday Patterns:** Thursday, Friday, Saturday consistently recorded more bookings than the beginning of the week.
- **Holiday Impact:** Bookings were generally lower on non-holiday days, suggesting a preference for staying home and spending time with family during holidays.
- **Year-over-Year Growth:** 2019 witnessed a substantial increase in bookings compared to the previous year, indicating positive business growth.

Overall, we observe a strong correlation between favorable weather conditions, specific months, and particular days of the week on booking activity. The year-over-year increase in bookings highlights the company's growth and success.

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

When creating dummy variables from a categorical variable with **k** categories, we typically create **k-1** dummy variables. This is to avoid the **dummy variable trap**, which occurs when all dummy variables are perfectly correlated. For example, if we have a categorical variable with three categories (A, B, and C), we can create two dummy variables:

- **is_A** (1 if the category is A, 0 otherwise)
- **is_B** (1 if the category is B, 0 otherwise)

If we include all three dummy variables, we can perfectly predict one dummy variable from the others. For instance, if **is_A** and **is_B** are both 0, then we know that **is_C** must be 1. This redundancy can lead to numerical instability and incorrect model results.

By dropping the first category, we effectively encode the remaining categories relative to the first one. In the example above, if we drop **is_A**, the model will interpret **is_B** as indicating whether the category is B or C, relative to A.

So, using **drop_first=True** when creating dummy variables helps to avoid the dummy variable trap and ensure the stability and interpretability of our statistical models.

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

'**temp**' and '**atemp**' variable have 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)

To ensure the validity of the linear regression model, the following assumptions were verified:

- **Normality of Residuals:** The residuals were assessed for normality using statistical tests and visual techniques like Q-Q plots.
- **Multicollinearity:** Variance Inflation Factors (VIF) were calculated to identify and address any significant multicollinearity among the independent variables.
- **Linearity:** The relationship between the dependent and independent variables was visually inspected through scatter plots to confirm a linear association.
- **Homoscedasticity:** Residual plots were examined to check for any discernible patterns, indicating constant variance across different levels of the independent variables.
- **Independence of Residuals:** Autocorrelation tests were conducted to ensure that the residuals were not correlated with each other, suggesting the independence of observations.

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

Top 3 features contributing significantly towards explaining the demand of the shared bikes:

- temp
- winter
- sep

General Subjective Questions

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

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables and aims to find the best-fitting line (or hyperplane in higher dimensions) that represents this relationship.

The Model

The linear regression model is expressed as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon$$

where:

- **y:** is the dependent variable (also known as the response variable)
- $\mathbf{x_1}$, $\mathbf{x_2}$, ..., $\mathbf{x_p}$**:** are the independent variables (also known as predictor variables)
- $\boldsymbol{\beta_0}$, $\boldsymbol{\beta_1}$, $\boldsymbol{\beta_2}$, ..., $\boldsymbol{\beta_p}$**:** are the coefficients representing the relationship between the independent and dependent variables
- **ε:** is the error term (or residual), which represents the difference between the actual value of y and the predicted value from the model

The Goal

The goal of linear regression is to estimate the coefficients (β_0, β_1, β_2, ..., β_p) that minimize the sum of squared errors (SSE). This is achieved using a technique called **ordinary least squares (OLS)**.

Ordinary Least Squares (OLS)

OLS finds the coefficients that minimize the following equation:

$$SSE = \Sigma(y_i - \hat{y}_i)^2$$

where:

- $\mathbf{y_i}$**:** is the actual value of the dependent variable for the i-th observation
- $\hat{\mathbf{y}}_\mathbf{i}$**:** is the predicted value of the dependent variable for the i-th observation

Assumptions

Linear regression makes several assumptions:

- **Linearity:** There is a linear relationship between the dependent variable and the independent variables.
- **Independence:** The observations are independent of each other.
- **Homoscedasticity:** The variance of the error term is constant for all values of the independent variables.
- **Normality:** The error term is normally distributed.

Applications

Linear regression is widely used in various fields, including:

- **Economics:** Predicting economic variables like GDP, inflation, and unemployment.
- **Finance:** Forecasting stock prices, bond yields, and risk.
- **Marketing:** Analyzing the relationship between marketing expenditures and sales.
- **Social sciences:** Studying relationships between social variables like education, income, and crime rates.
- **Engineering:** Modeling the relationship between inputs and outputs in manufacturing processes.

Limitations

While linear regression is a powerful tool, it has limitations:

- **Non-linear relationships:** If the relationship between the variables is non-linear, linear regression may not provide accurate predictions.
- **Outliers:** Outliers can significantly affect the results of linear regression.
- **Multicollinearity:** If the independent variables are highly correlated, it can make it difficult to interpret the model's coefficients.

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

Anscombe's Quartet is a fascinating collection of four datasets that illustrate the importance of data visualization in statistical analysis. Created by statistician Francis Anscombe in 1973, the quartet serves as a powerful reminder that summary statistics alone can be misleading.

Each of the four datasets in Anscombe's Quartet consists of 11 pairs of (x, y) values. Despite having nearly identical statistical properties—such as means, variances, and correlation coefficients—the datasets exhibit very different distributions and relationships when graphed.
This stark contrast emphasizes that relying solely on numerical summaries can obscure the true nature of the data.

Key Statistics for All Datasets

For all four datasets, the following statistics are identical:

- Mean of x: 9
- Mean of y: 7.50
- Variance of x: 11
- Variance of y: 4.125
- Correlation between x and y: 0.816
- Linear regression line: y = 3 + 0.5x

The Four Datasets

Dataset I
This dataset shows a clear linear relationship, where as x increases, y also increases. When plotted, it appears to fit the linear regression model well.

Dataset II
In this dataset, there is a strong relationship between x and y, but it is not linear; instead, it follows a quadratic pattern. This dataset demonstrates how a single outlier can significantly affect the regression line, leading to potentially misleading conclusions if one only looks at summary statistics.

Dataset III
This dataset also appears to have a linear relationship but contains one significant outlier. The presence of this outlier skews the results, highlighting how outliers can distort statistical analyses and lead to incorrect interpretations.

Dataset IV
In this final dataset, most x values are identical except for one extreme value. This creates a situation where the linear regression model is not appropriate, as the majority of data points do not follow any discernible trend.

Importance of Visualization

The primary lesson from Anscombe's Quartet is that visualizing data is crucial before drawing conclusions or applying statistical models. Graphs can reveal patterns, relationships, and anomalies that summary statistics might hide.

For instance:

- **Outliers:** As seen in Datasets II and III, outliers can significantly influence regression results.

- **Non-linearity:** Dataset II illustrates that not all relationships are linear; visualizing data helps identify these complexities.

- **Distribution Patterns:** Dataset IV shows how similar summary statistics can mask entirely different underlying distributions.

Conclusion

Anscombe's Quartet serves as a compelling reminder that data analysis should involve both numerical summaries and visual exploration. By examining data visually, analysts can uncover insights that would otherwise remain hidden and avoid falling into the trap of

oversimplified conclusions based solely on statistics. This approach encourages a more nuanced understanding of data and its complexities, ultimately leading to better decision-making in various fields such as science, business, and social research.

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

Pearson's correlation coefficient (r) is a statistical measure that quantifies the linear relationship between two variables. It ranges from -1 to 1:

- **r = 1:** Indicates a perfect positive correlation, meaning the two variables increase or decrease together perfectly.
- **r = -1:** Indicates a perfect negative correlation, meaning one variable increases as the other decreases perfectly.
- **r = 0:** Indicates no correlation between the variables.

Formula for Pearson's correlation coefficient:

$$r = (\Sigma(x_i - \bar{x})(y_i - \bar{y})) / \sqrt{(\Sigma(x_i - \bar{x})^2)} * \sqrt{(\Sigma(y_i - \bar{y})^2)}$$

where:

- $\mathbf{x_i}$ and $\mathbf{y_i}$ are the individual data points for the two variables.
- $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are the means of the respective variables.

Applications of Pearson's correlation:

- **Statistics:** Analyzing the relationship between variables in various fields.
- **Finance:** Measuring the correlation between stock prices or economic indicators.
- **Psychology:** Studying the relationship between personality traits or cognitive abilities.
- **Social sciences:** Examining the relationship between social variables like income, education, and crime rates.

By understanding Pearson's correlation, we can assess the strength and direction of linear relationships between variables and make informed conclusions based on the data.

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

Scaling is a technique used in data preprocessing to transform data into a standard range or format. It's often necessary to scale data before applying certain machine learning algorithms, especially those that rely on distance calculations or gradient descent.

Why is scaling performed?

- **Normalization and Standardization:** These techniques help ensure that different features contribute equally to the model, preventing features with larger magnitudes from dominating the learning process.
- **Improved Algorithm Performance:** Many algorithms, such as K-nearest neighbors and gradient descent, perform better when features are on a similar scale.
- **Convergence Speed:** Scaling can accelerate the convergence of optimization algorithms.
- **Interpretability:** Scaled data can sometimes make the model's coefficients more interpretable.

<u>Normalized Scaling vs. Standardized Scaling:</u>

Normalized Scaling (Min-Max Scaling)

Transforms data to a specific range: Typically, the range is [0, 1].

Formula: x_scaled = (x - min(x)) / (max(x) - min(x))

- **Advantages:** Preserves the original distribution of the data.
- **Disadvantages:** Sensitive to outliers, as a single outlier can significantly affect the scaling.

Standardized Scaling (Z-score Standardization)

Transforms data to have a mean of 0 and a standard deviation of 1.

Formula: x_scaled = (x - mean(x)) / std(x)

- **Advantages:** Less sensitive to outliers, as the mean and standard deviation are less affected by extreme values.
- **Disadvantages:** May distort the original distribution of the data, especially if the data is skewed.

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

When VIF is infinite, it indicates a perfect multicollinearity between the predictor variables. This means that one predictor can be perfectly predicted from the others, leading to a singular matrix in the regression analysis.

Causes of Infinite VIF:

- **Exact Linear Dependence:** The most common cause is when one predictor is an exact linear combination of the others. For example, if you have two predictors, x1 and x2, and x2 = 2 * x1, then VIF for both predictors will be infinite.

- **Near-Perfect Multicollinearity:** Even if the predictors are not perfectly linearly dependent, but very close to it, the VIF can still become extremely high (approaching infinity). This can happen due to numerical precision limitations in the calculations.

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, or quantile-quantile plot, is a graphical tool used in statistics to compare the distributions of two datasets by plotting their quantiles against each other. This visual representation helps assess whether the two datasets come from the same distribution or to compare a sample dataset against a theoretical distribution, such as the normal distribution.

How Q-Q Plots Work

Construction of a Q-Q Plot

1. **Quantile Calculation:** The quantiles of both datasets are calculated. For a sample dataset, this involves sorting the data and determining the quantiles at specified probabilities (e.g., 0.25, 0.50, 0.75).

2. **Plotting:** The quantiles of one dataset are plotted on the x-axis and the quantiles of the other dataset are plotted on the y-axis.

3. **Reference Line:** A reference line (usually a 45-degree line) is added to the plot. If the points fall along this line, it suggests that the two distributions are similar.

Interpretation

- If the points in the Q-Q plot closely follow the reference line, it indicates that the two distributions being compared are similar in shape.

- Deviations from this line can indicate differences in distribution characteristics such as skewness or kurtosis.

Importance of Q-Q Plots in Linear Regression

Assessing Normality:
In linear regression analysis, one of the key assumptions is that the residuals (the differences between observed and predicted values) should be normally distributed. A Q-Q plot can help visualize this assumption:

Normality Check:
By plotting the residuals against a theoretical normal distribution, analysts can visually assess if they conform to normality. If they do not, it may suggest that a linear regression model is not appropriate.

Identifying Outliers:
Q-Q plots can also help identify outliers:
- Points that fall far from the reference line may indicate outliers in the data that could disproportionately influence regression results.
- Recognizing these outliers allows for further investigation or potential adjustments to improve model accuracy.

Evaluating Model Fit:
When comparing two datasets or assessing how well a model fits:
- A Q-Q plot can visually show if there are systematic deviations from what would be expected under normality.
- This can inform decisions about whether to proceed with linear regression or consider alternative modeling approaches.