Assignment-based Subjective Questions

From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

- Season 3 (Fall) shows the highest rentals followed by Season 2 (Summer) and 4 (Winter). The lowest rentals are in Season 1 (Spring).
- 2019 showed more rentals when compared with 2018.
- Most of the bookings were done during May, June, Jul, Aug and Sept.
- 97% of the rentals occur when there's no Holiday, proving that the data is skewed/ biased and not reliable for prediction.
- The weekday variable shows a very close trend of between 13.59% to 14.83% of total bookings. This variable can have some or no influence towards the predictor.
- About 68.49% of the bike bookings were happening on a ‘working day'. This indicates the working day can be a good predictor for the dependent variable.
- About 68.61% of the bike bookings were happening on a 'Clear' day, followed by 30.26% on a 'Misty' day. This indicates that weather can be a good predictor for the dependent variable.
- In the ‘month’ variable, the months between May to September show a very close trend of between 10.07% to 10.67%. The next adjacent months, April and October show a very close trend between 8.17% and 7.74% respectively.

Why is it important to use drop_first=True during dummy variable creation?

- The drop_first=True argument in dummy variable creation is important to avoid a statistical issue called multicollinearity.
- When you create dummy variables for a categorical feature with multiple levels, you end up with one binary variable for each level.
- By setting drop_first=True, we essentially remove one of the dummy variables. This doesn't lose information because the missing variable's value can be inferred from the others.
- With multicollinearity, the model's coefficients for the dummy variables become difficult to interpret. Dropping one variable helps avoid this confusion.

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

- The ‘temp’ and ‘atemp’ variable has the highest correlation with the target variable, followed by ‘yr’ variable.

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

My analysis confirms that the data meets the following assumptions for linear regression:

- **Normally Distributed Errors:** The errors (differences between predicted and actual values) follow a normal distribution.
- **Low Multicollinearity:** The independent variables exhibit little to no correlation with each other.
- **Linear Relationship:** There appears to be a linear trend between the independent and dependent variables.
- **Homoscedasticity:** The spread of the residuals (errors) is consistent across the range of the independent variables. There's no pattern like increasing or decreasing variance.
- **Independent Residuals:** The errors are independent of each other, meaning no correlation exists between errors at different data points.

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

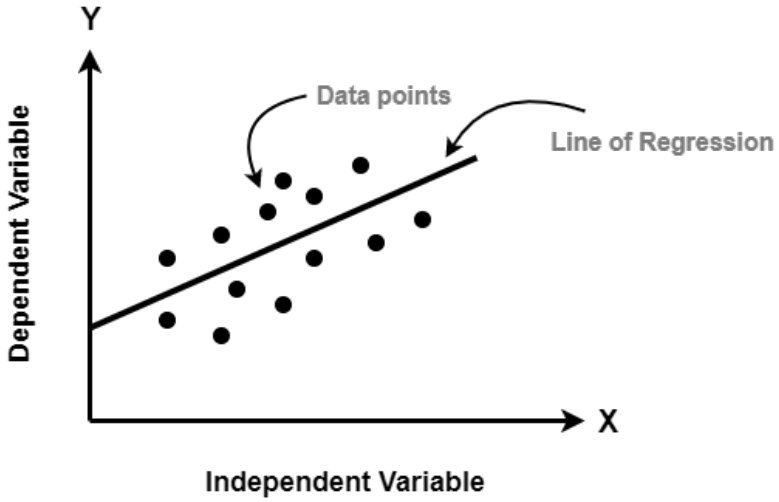
Top 3 Features:

- Temperature (positive influencer)
- Year (positive influencer)
- Light Snow/Rain weather (negative influencer)

General Subjective Questions

Explain the linear regression algorithm in detail.

Linear regression tackles prediction using a straight line. Imagine predicting a dependent variable (Y, like house price) based on an independent variable (X, like house size). The algorithm finds the best-fit line through your data points.



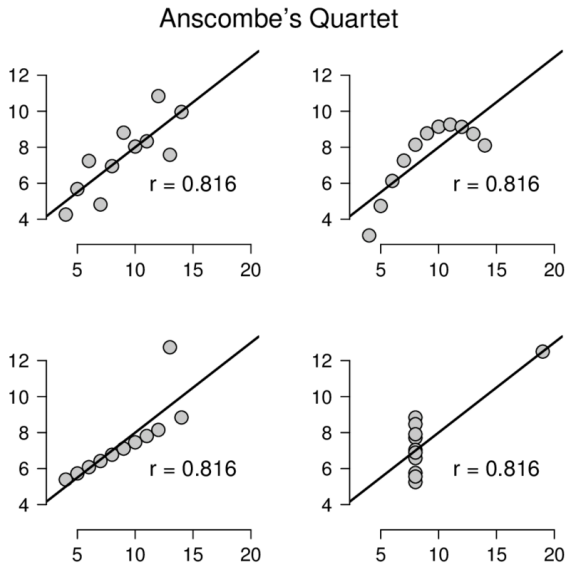
- Model: We use the equation $Y = \beta_0 + \beta_1 X + \varepsilon$. Here, β_0 is the y-intercept, β_1 is the slope, and ε represents the error (difference between actual and predicted Y).
- Training: We want to minimize the error (ε) for all data points. The method of least squares comes in. It calculates the sum of squared errors (ε^2) and aims to minimize it. This ensures the best line fit.
- Optimization: The sum of squared errors acts like a cost function. We use techniques like gradient descent to adjust β_0 and β_1 iteratively. With each step, the algorithm tweaks these values to minimize the cost function, essentially finding the best slope and intercept for the line.
- Evaluation: Once we have the optimal β_0 and β_1, we can use the equation to predict Y for new data points. But how good is our model?
  - R-squared: This value (between 0 and 1) tells us how well the model explains the variation in Y using X. A higher value indicates a better fit.
  - P-value: This helps determine if the X-Y relationship is statistically significant. A low p-value suggests a strong correlation.
- Types:
  - Simple: Uses one X to predict Y (e.g., size predicts price).
  - Multiple: Uses two or more Xs (e.g., size, location, bedrooms predict price).

Linear regression is powerful for its simplicity and interpretability. However, it assumes a linear relationship between variables. If the relationship is more complex, the model won't perform well. It's also sensitive to outliers in the data.

Explain the Anscombe’s quartet in detail.

Anscombe's quartet is a famous example in statistics created by Francis Anscombe in 1973 to highlight the importance of data visualization. It consists of four sets of data, each containing 11 data points (X, Y).

- Deception by Statistics: All four datasets have nearly identical statistical properties - mean, variance, correlation coefficient, and even the least squares regression line. This might lead you to believe the data is similar.
- Visual Reality Unveils the Truth: When you plot these datasets, you see a completely different story. Each dataset showcases a very distinct pattern. One has a clear linear relationship, another has a curved pattern, one has a single outlier, and the last has a cluster of outliers.
- Why It Matters: Anscombe's quartet demonstrates that relying solely on summary statistics can be misleading. Visualization is crucial to uncover the underlying trends, outliers, and non-linear relationships that statistical measures might not capture.

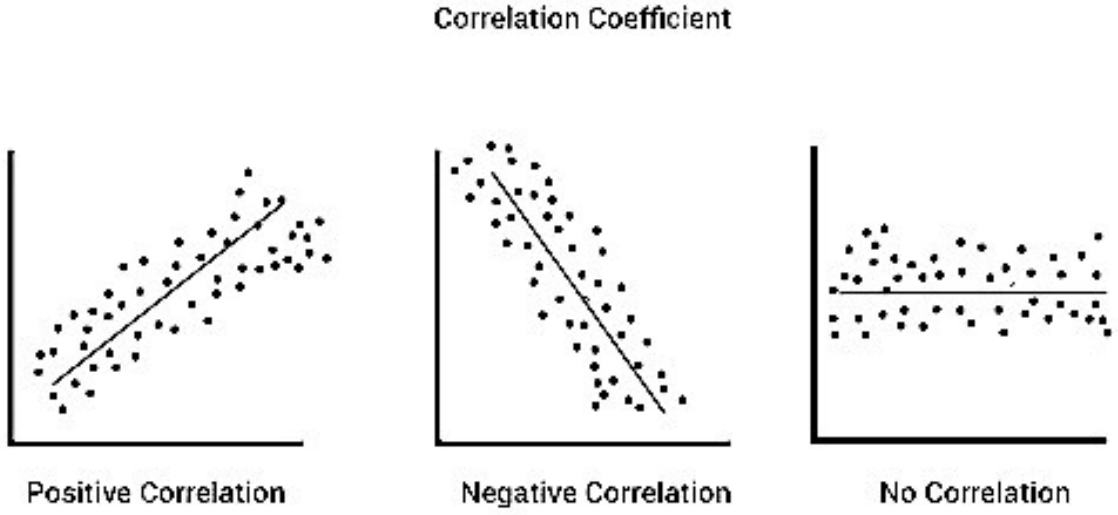


In essence, it emphasizes that data can be deceptive. Even if numbers seem to suggest similar characteristics, the true nature of the data can only be revealed through visualization. This is a cautionary tale for data analysts to not solely rely on statistical summaries and to always incorporate data visualization techniques for a more comprehensive understanding.

What is Pearson’s R?

Pearson's R, also called the Pearson correlation coefficient, is a statistical measure that reflects the strength and direction of a linear relationship between two continuous variables. It's denoted by the symbol "r" and falls between -1 and +1.

- Strength: The closer the value is to 1 (positive) or -1 (negative), the stronger the correlation.
- Positive R indicates the variables tend to move in the same direction (as one increases, the other increases too, or vice versa).
- Negative R indicates the variables tend to move in opposite directions (as one increases, the other decreases).
- No Correlation: An R of 0 signifies no linear correlation between the variables. The changes in one variable don't have a predictable effect on the other.



Pearson's R only measures linear relationships. It won't capture non-linear connections between variables. It doesn't imply causation. Just because variables are correlated doesn't mean one causes the other.

Pearson's R is a widely used statistic for understanding how two variables are related. It's a helpful tool for initial data exploration but shouldn't be the sole factor in concluding data.

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

In machine learning, scaling refers to the process of transforming the features (variables) in your data to a common range. This is a crucial step in data preprocessing for several reasons:

- Fair Play for Features: Features with larger magnitudes can dominate algorithms that rely on distance-based calculations (like k-nearest neighbours or support vector machines). Scaling puts all features on a similar playing field, ensuring each contributes equally to the learning process.

- Improved Convergence: Gradient descent and other optimization algorithms used to train many machine learning models work more efficiently when features are on a similar scale. Scaling helps these algorithms converge faster to a good solution.

Types of Scaling:

- Normalization: This technique scales the features to a specific range, typically between 0 and 1 (min-max scaling) or -1 and 1. It ensures all features fall within a predefined boundary.
- Standardization: This technique transforms the features by subtracting the mean and then dividing by the standard deviation. This results in features with a mean of 0 and a standard deviation of 1. It emphasizes the relative position of a data point within the distribution of its feature.

The choice between normalization and standardization depends on the specific algorithm you're using. Here's a general guideline:

- Normalization: Preferable for algorithms sensitive to absolute values, like k-nearest neighbours or support vector machines.
- Standardization: Often a good choice for algorithms where the distribution of features matters, such as linear regression or logistic regression.

In essence, both techniques achieve scaling, but they do so in slightly different ways. Normalization focuses on a specific range, while standardization focuses on the relative position within the data distribution.

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

An infinite VIF (Variance Inflation Factor) in linear regression occurs when there's perfect multicollinearity among the independent variables (X variables).

- Multicollinearity: This happens when two or more independent variables are highly correlated, meaning they contain redundant information. In extreme cases, one variable can be perfectly predicted by a linear combination of the others.
- VIF and Multicollinearity: VIF measures how much the variance of an estimated regression coefficient is inflated due to multicollinearity. A VIF of 1 indicates no inflation, while higher values suggest increasing multicollinearity.
- Perfect Collinearity and Infinite VIF: When there's perfect multicollinearity, meaning one variable can be exactly expressed as a linear combination of others, the denominator in the VIF formula becomes zero. Since dividing by zero is undefined, the VIF value goes to infinity. This signifies a major problem in the regression model. It implies,
    - Unreliable Coefficients: With perfect multicollinearity, the coefficients of the individual variables become unreliable. It's difficult to determine the independent effect of each variable on the dependent variable (Y).
    - Unstable Model: The model might be very sensitive to small changes in the data, leading to unpredictable and unreliable predictions.

To fix the same,

- Identify Culprits: Analyze the correlation matrix to pinpoint highly correlated variables.
- Remove Redundant Variables: Consider removing one or more of the collinear variables, but be sure to choose the one that least affects the domain knowledge you're trying to capture in the model.
- Combine Variables: If the variables have a natural interpretation together, you might create a new combined variable.
- Regularization Techniques: Some machine learning algorithms offer regularization techniques that can help reduce the impact of multicollinearity.

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

A Q-Q plot, also known as a quantile-quantile plot, is a graphical tool used in linear regression to assess the normality of the model residuals.

- Quantiles: These are the values that divide data into equal-sized portions. Imagine data points sorted from least to greatest. The median splits the data in half, quartiles split it into fourths, etc.
- Residuals: In linear regression, residuals represent the difference between the actual Y values and the predicted Y values from the model.

Importance of Q-Q Plots in Linear Regression:

- Linear regression often relies on the assumption that the residuals are normally distributed. This assumption is crucial for:
  - Valid hypothesis testing: Many statistical tests used in regression analysis are based on the normality assumption. A Q-Q plot helps you check if these tests are applicable.
  - Confidence interval accuracy: Confidence intervals provide an estimate of the range where the true population parameter (like the slope) might lie. Normality helps ensure the accuracy of these intervals.

Benefits of Using Q-Q Plots:

- Visually informative: Unlike statistical tests, Q-Q plots provide a clear visual representation of how well the residuals align with the expected distribution.
- Flexible: You can use Q-Q plots to assess normality against any theoretical distribution, not just the normal distribution.