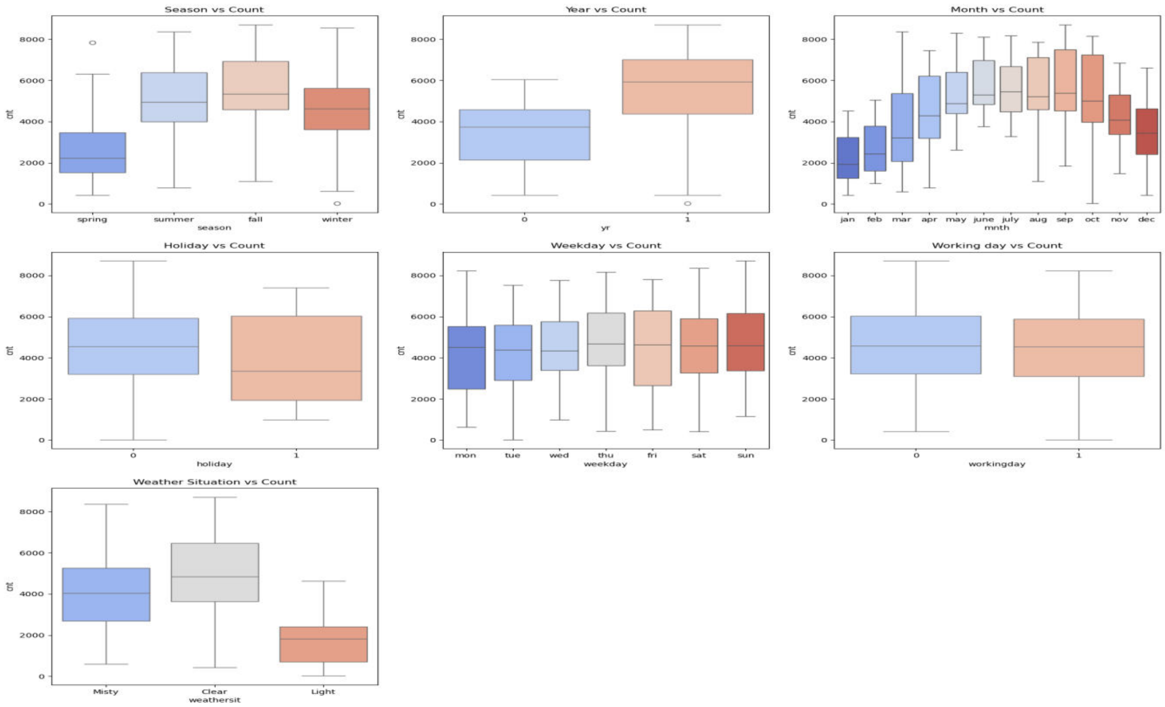
Assignment-based Subjective Questions

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

Answer:



The box plot generated illustrates the analysis of categorical variables on the dependent variable. The key observations and inferences are as follows:

- Seasonal Impact: Fall season shows the highest number of bookings, with a significant increase in booking counts from 2018 to 2019 across all seasons.
- Monthly Trend: The majority of bookings occur between May and October, showing a rising trend from the start of the year until mid-year, followed by a decline towards the year-end.
- Weather Influence: Clear weather conditions attract more bookings, which is expected.
- Weekly Pattern: Bookings are higher on Thursdays, Fridays, Saturdays, and Sundays compared to the beginning of the week.
- Holiday Effect: Fewer bookings are made on non-holiday days, which is reasonable as people may prefer staying at home and spending time with family on holidays.
- Workday Comparison: Bookings are almost equal on working days and non-working days.
- Annual Growth: There is an increase in bookings in 2019 compared to the previous year, indicating good business progress.

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

Answer:

The purpose of creating dummy variables is to handle categorical variables with 'n' levels by generating 'n-1' new columns. Each new column uses zeros and ones to indicate the presence of a specific level. By setting drop_first=True, we ensure that the resulting dummy variables represent 'n-1' levels, effectively reducing correlation among the dummy variables and avoiding multicollinearity.

In our case: We have a categorical variable "season" with four categories: spring, summer, fall, and winter. Without drop_first=True, we would create four dummy variables. With drop_first=True, only three are created, implicitly representing the fourth category. The dropped category can be inferred from zeros in the remaining dummies.

	season_spring	season_summer	season_winter
0	0	0	0
1	1	0	0
2	0	1	0
3	0	0	1

---

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

Answer:

The 'temp' and 'atemp' variables exhibit the highest correlation with the target variable 'cnt' compared to other variables.

---

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

Answer:

I validated the robustness and reliability of the Linear Regression Model based on following:

- Normality of Error Terms:
  - The error terms should be normally distributed.
- Multicollinearity Check:
  - There should be insignificant multicollinearity among the variables.
- Linear Relationship Validation:
  - Linearity should be visible among the variables.
- Homoscedasticity:
  - There should be no visible pattern in the residual values, indicating equal variance of the errors.
- Independence of Residuals:
  - The residuals should be independent, with 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?**

Answer:

The top three features significantly contributing to the demand for shared bikes are:

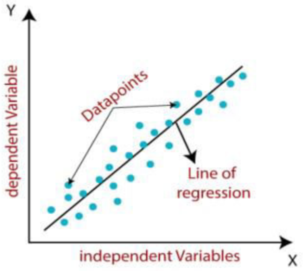
- Temperature (temp)
- Year (yr)
- September (mnth_sep)

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Answer:

Linear regression is a fundamental and widely used algorithm in statistics and machine learning for modeling the relationship between a dependent variable and one or more independent variables. Linear regression aims to establish a linear relationship between the dependent variable (target) and independent variables (predictors) [shown in graph].



Mathematically, we can represent a linear regression as:

$$y= \beta_0+\beta_1X_1+ \varepsilon$$

Here,
Y = Dependent Variable (Target Variable)
X_1 = Independent Variable (predictor Variable)
β_0 = intercept of the line (Gives an additional degree of freedom)
β_1 = Linear regression coefficient (scale factor to each input value).
ε = random error
The values for x and y variables are training datasets for Linear Regression model representation.

Linear regression can be further divided into **two types of the algorithm**:

- **Simple Linear Regression:** If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.
  Formula for the Simple Linear Regression: $Y=\beta_0+\beta_1X_1 +\epsilon$

- **Multiple Linear regression:** If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

  Formula for the Multiple Linear Regression: $Y= \beta_0+ \beta_1X_1+ \beta_2X_2+\ldots+ \beta_pX_p+\epsilon$

Steps in Linear Regression:

a) **Data Preparation**: Collect and preprocess the data, ensuring it is clean and suitable for modeling.
b) **Model Selection**: Choose the linear regression model (simple or multiple).
c) **Model Fitting**: Use methods like Ordinary Least Squares (OLS) to estimate the coefficients (β/beta). OLS minimizes the sum of squared residuals (difference between observed and predicted values).
d) **Model Evaluation**: Evaluate the model's performance using metrics such as R^2 (coefficient of determination), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).
e) **Prediction**: Use the fitted model to make predictions on new data.

Assumptions of Linear Regression:

- **Linearity**: The relationship between the dependent and independent variables should be linear.
- **Independence**: The residuals (errors) should be independent.
- **Homoscedasticity**: Constant variance of residuals across all levels of the independent variables.
- **Normality**: The residuals should be normally distributed.
- **No Multicollinearity**: Independent variables should not be highly correlated with each other.

Benefits and Limitations:

Benefits:

- Easy to implement and interpret.
- Works well with linearly separable data.
- Efficient for large datasets.

Limitations:

- Assumes a linear relationship.
- Sensitive to outliers.
- Requires careful handling of assumptions.

Linear regression is a robust and easily interpretable algorithm for predictive modeling. To achieve accurate and reliable predictions, it is crucial to validate its assumptions and thoroughly evaluate the model.

2. Explain the Anscombe’s quartet in detail.

Answer:

Anscombe's quartet is a set of four datasets that have nearly identical simple statistical properties, such as mean, variance, correlation, and linear regression lines, yet they appear very different when graphed. This quartet was constructed by the statistician Francis Anscombe in 1973 to emphasize the importance of graphing data before analyzing it and to show how statistical calculations alone can be misleading.

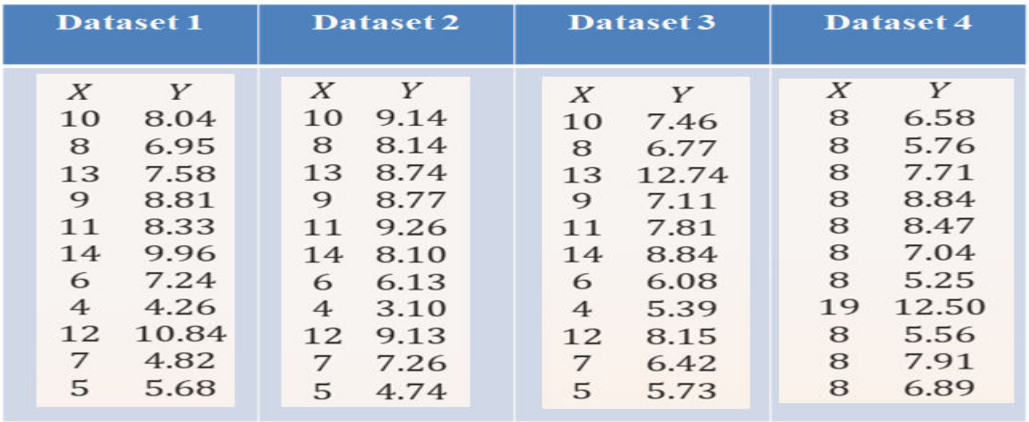
Properties:

All four datasets in Anscombe’s quartet have:

1. Same Mean of X and Y
2. Same Variance of X and Y
3. Same Correlation between X and Y
4. Same Linear Regression Line: y = 3.00 + 0.5x

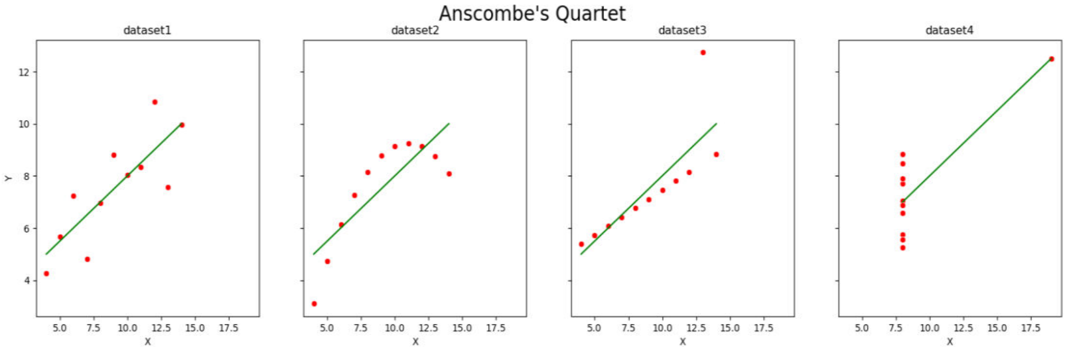
The Four Datasets

- Dataset 1: A typical linear relationship between X and Y.
- Dataset 2: Data that shows a clear curvilinear relationship.
- Dataset 3: An outlier heavily influences the statistics.
- Dataset 4: A vertical outlier with a nearly perfect linear relationship.



Dataset 1		Dataset 2		Dataset 3		Dataset 4	
X	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Visualization: The power of Anscombe's quartet comes from visualizing the datasets:



Key Lessons from Anscombe's Quartet

- Importance of Visualization: Simply relying on summary statistics can be misleading. Visualizations can reveal patterns, relationships, and anomalies not evident from statistics alone.
- Outliers and Influential Points: Outliers or specific influential points can distort statistical summaries and affect the interpretation of the data.
- Context Matters: Understanding the context and characteristics of the data is crucial for accurate analysis and interpretation.

Anscombe’s quartet powerfully demonstrates why it is essential to visualize data and not to rely solely on summary statistics. Graphing data helps to uncover the underlying structures and nuances that numbers alone might hide.

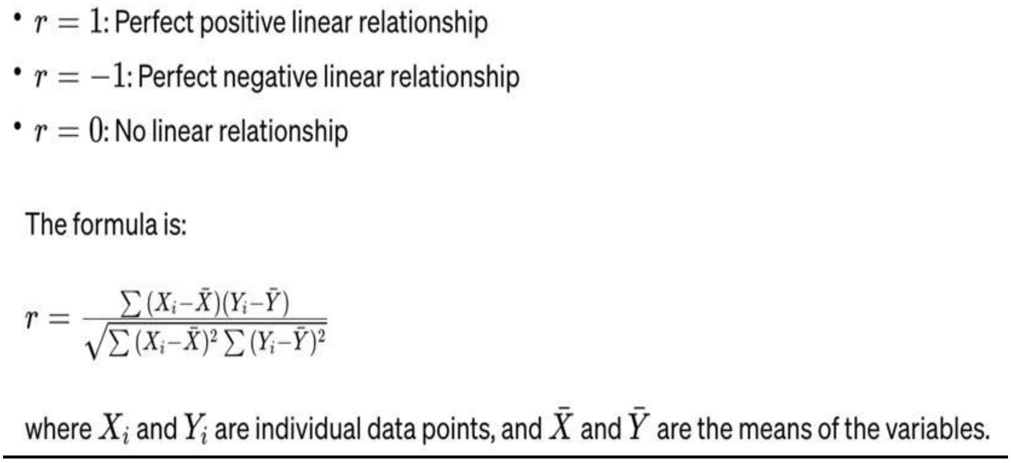
---

3. What is Pearson’s R?

Answer:

Pearson's R, also known as Pearson’s correlation coefficient, is a measure of the linear relationship between two variables. It quantifies the strength and direction of this relationship, providing insights into how one variable changes in relation to the other.

Key Points:



- $r = 1$: Perfect positive linear relationship
- $r = -1$: Perfect negative linear relationship
- $r = 0$: No linear relationship

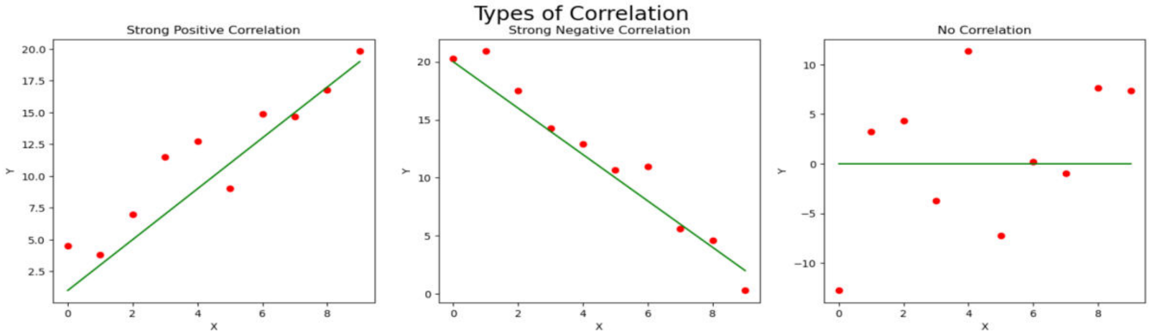
The formula is:

$$r = \frac{\sum (X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum (X_i-\bar{X})^2 \sum (Y_i-\bar{Y})^2}}$$

where X_i and Y_i are individual data points, and $\bar{X}$ and $\bar{Y}$ are the means of the variables.

Interpretation:

- **Positive Correlation**: As one variable increases, the other also increases. For example, height and weight typically show a positive correlation.
- **Negative Correlation**: As one variable increases, the other decreases. For example, the number of hours spent studying and the number of errors made on a test might show a negative correlation.
- **No Correlation:** No consistent relationship between the variables. For example, shoe size and intelligence score might have no correlation.



Usage:

Pearson's R is widely used in:

- **Statistics**: To understand relationships between variables.
- **Data Science**: For feature selection and data analysis.
- **Research**: To analyze experimental and observational data.

---

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

Answer:

Scaling is the process of transforming the features of your dataset so that they are on a similar scale. This is crucial in many machine learning algorithms, especially those that rely on distance calculations or assume features are normally distributed.

Scaling is performed for several reasons:

– **Equal Weight**: Ensures all features contribute equally to the model, preventing bias from features with larger magnitudes.

– **Improved Convergence**: Enhances the convergence speed in algorithms like gradient descent by aligning feature scales.

– **Enhanced Performance**: Improves algorithms (e.g., k-nearest neighbors, SVM) that rely on Euclidean distance by preventing any single feature from dominating distance calculations.

– **Better Interpretation**: Makes coefficients in linear models more interpretable by standardizing feature scales.

Difference between Normalized Scaling and Standardized Scaling

1. Normalized Scaling (Min-Max Scaling)

– **Purpose**: Transforms the data to fit within a specific range, typically [0, 1].

– **Formula**:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- **Application**: Useful when you want the data to fit within a bounded range.
- **Example**: Scaling pixel values in image processing to be within [0, 1].

2. **Standardized Scaling (Z-score Standardization)**

- **Purpose**: Transforms the data to have a mean of 0 and a standard deviation of 1.
- **Formula**:

$$X' = \frac{X - \mu}{\sigma}$$

- **Application**: Useful when the data is normally distributed and you want to remove the effect of different units and scales.
- **Example**: Standardizing features for Principal Component Analysis (PCA) to ensure that each component explains the same amount of variance.

Normalization	Standardization
Suited for data with varying ranges and units.	Suited for data following a normal distribution.
Keeps the original shape of the data distribution but changes the scale.	Changes the shape of the data distribution to a standard normal distribution.

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

Answer:

An infinite value for the Variance Inflation Factor (VIF) occurs when there is perfect multicollinearity between the independent variables in your regression model. The value of *VIF is calculated by the below formula:*

$$VIF_i = \frac{1}{1 - {R_i}^2}$$

Where, 'i' refers to the i^{th} variable.

Causes of Infinite VIF:

1. Exact Linear Relationship**:** If one variable is an exact linear combination of one or more other variables. For instance, if you have two variables where $X_2 = 2 * X_1$, this will result in perfect multicollinearity.
2. Dummy Variable Trap**:** Including all dummy variables for a categorical feature without dropping one category. This leads to perfect multicollinearity because the dummy variables sum to 1.
3. Redundant Variables**:** Including two highly correlated variables that convey the same information, e.g., 'height in inches' and 'height in centimetres'.

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

Answer:

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If the data follows the specified theoretical distribution, the points will approximately lie on a straight line. Deviations from this line indicate departures from the specified distribution.

How to Create a Q-Q Plot:

1. Sort the Data**:** Arrange the data points in ascending order.
2. Calculate Quantiles**:** Compute the quantiles of the sample data.
3. Theoretical Quantiles**:** Calculate the quantiles from the theoretical distribution (e.g., the normal distribution).
4. Plot**:** Create a scatterplot with the sample quantiles on the y-axis and the theoretical quantiles on the x-axis.

Use and Importance in Linear Regression:

In the context of linear regression, a Q-Q plot is primarily used to check the normality assumption of the residuals (error terms). The assumptions of linear regression include that the residuals should be normally distributed. Here's why it matters:

1. Normality of Residuals**:** The normality of residuals ensures that hypothesis tests and confidence intervals are valid. If the residuals are not normally distributed, the results of these tests may be unreliable.
2. Model Validation**:** A Q-Q plot helps validate the assumption that the residuals follow a normal distribution. If the points lie approximately on the diagonal line, it suggests that the residuals are normally distributed. Significant deviations from the line indicate departures from normality.
3. Detection of Outliers**:** A Q-Q plot can help identify outliers and leverage points that may unduly influence the regression model. These points appear as deviations from the straight line in the plot.
4. Improving Model Fit**:** By identifying non-normality and outliers, the Q-Q plot helps in diagnosing and improving the regression model. Transformations or different modeling techniques might be considered based on the insights from the Q-Q plot.

In conclusion, a Q-Q plot is a crucial diagnostic tool in linear regression for assessing the normality of residuals, detecting outliers, and ensuring the validity of the regression model. By using a Q-Q plot, you can validate your model assumptions and make necessary adjustments to improve the model's reliability and accuracy.