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?**

   Solution:

After doing the analysis on the categorical columns by plotting boxplots and bar plots, the following observations are made.

- The bike rentals were high during a non-holiday time, which is quite acceptable as most of the people prefer to rest and spend time with family on a holiday.
- The bike rentals are relatively stable on both working and a non-working day.
- The bike rentals were high during the Fall season, where people would mostly prefer to go on a bike trip during to enjoy the weather and then the summer season also, where most of the people would get holidays and like to roam.
- The bike rentals were more from Thursday(being the high) till Sunday, which is quite obvious that people choose to put a leave from Thursday in order to get a long weekend.
- The bike rentals were more when the weather is Clear, which is more common.
- The bike rentals were more during the month of September and around that time.
- Overall, the bike rentals were most during the year 2019 compared to the previous year, which showed a positive growth in the business.

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

   Solution:

As we know for a categorical column with N categories, we represent these categories with N-1 dummy variables. While creating the dummies we give drop_first=True which ensures to drop the first column during dummy variable creation. This is to avoid creation of excess columns unnecessarily which shows a variation in the VIF and another evaluation metrics.

For example, if we have a column called “status” with 3 categories as “yes”, “no”, “unknown”, without using this “drop_first=True” it would result in 3 dummy columns. But by using this it would drop the first column (“yes”), and we can represent this information by using the other 2 dummy columns. This approach simplifies the model by reducing the correlations created among dummy variables.

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

Solution:

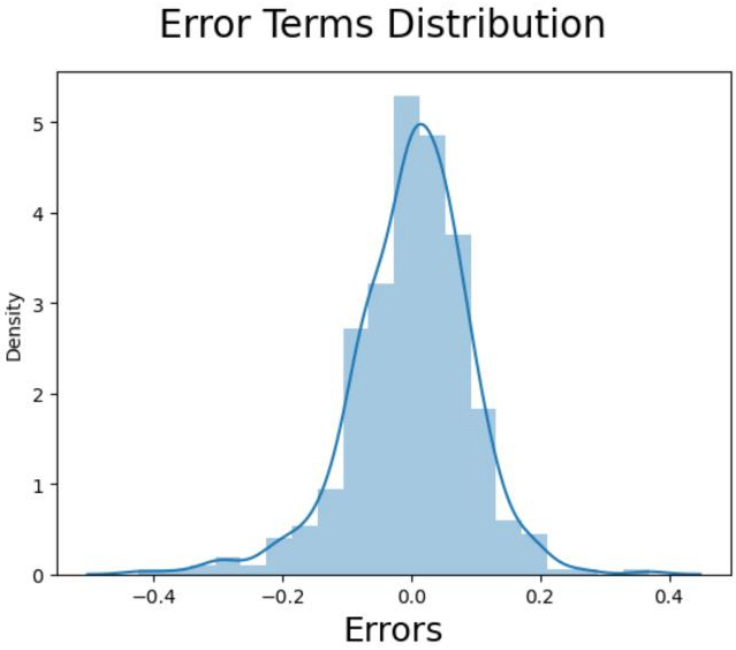
‘temp’ variable has the highest correlation with the target variable.

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

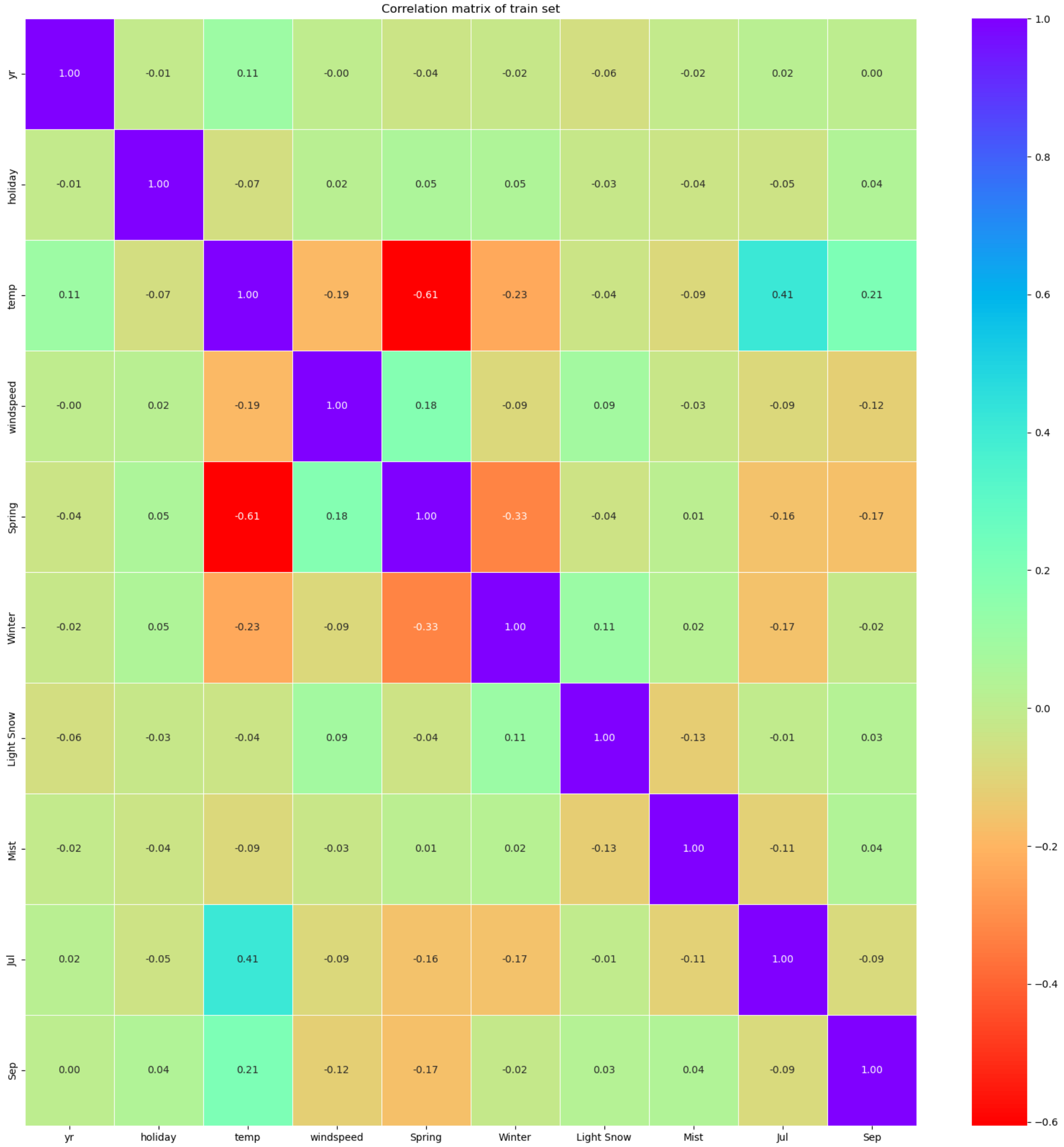
Solution:

I have validated the model by considering 5 assumptions:

1. Normality of error terms - It states that all the error terms should be normally distributed at mean is equal to zero.



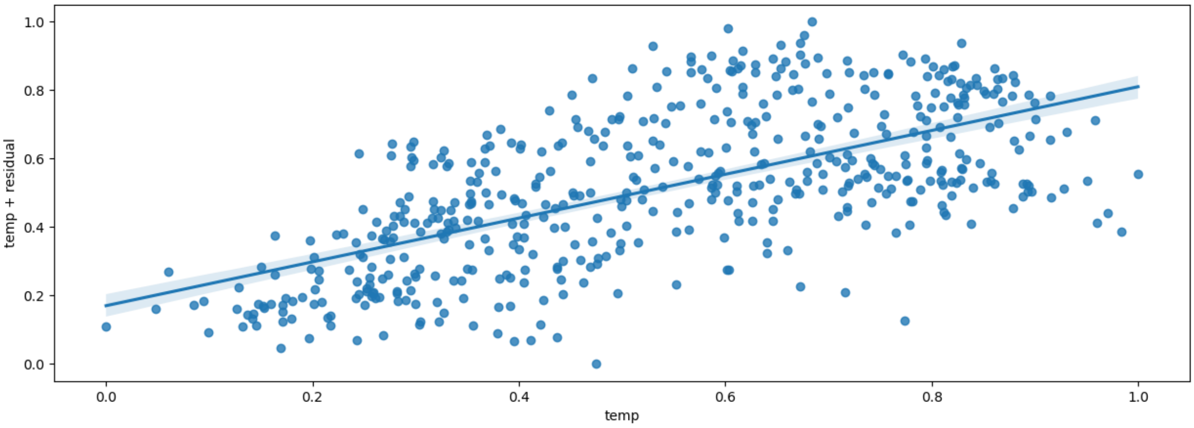
2. Multicollinearity check– There should be no collinearity among variables.

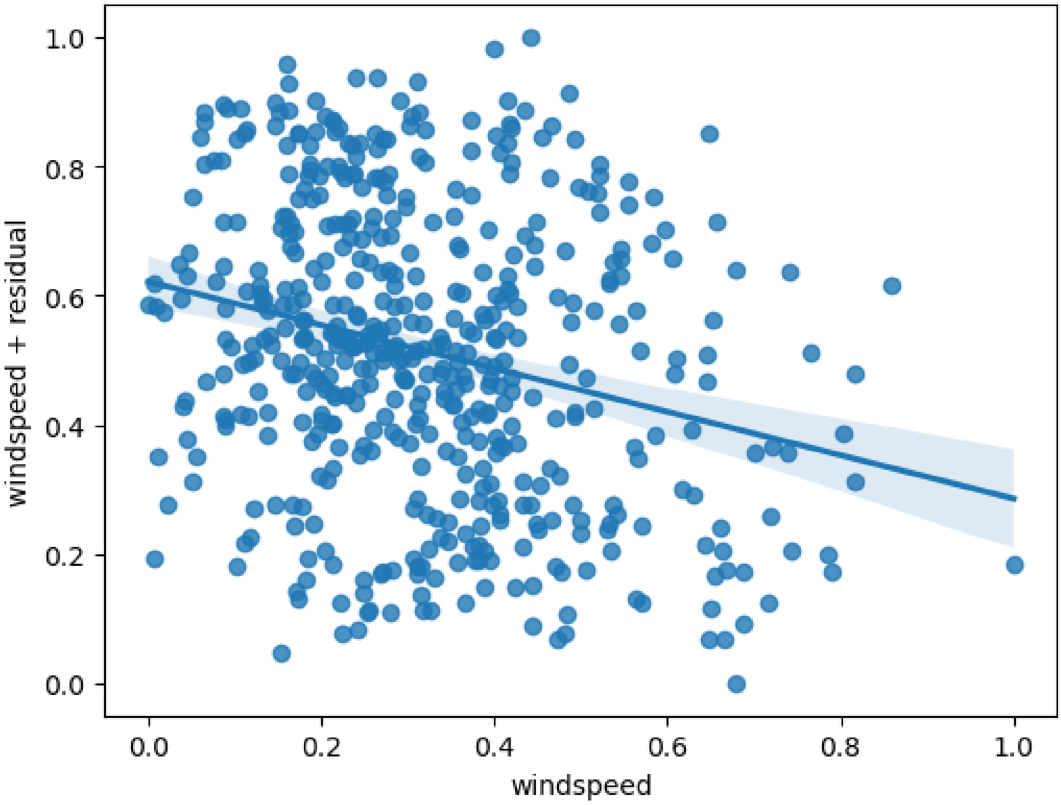


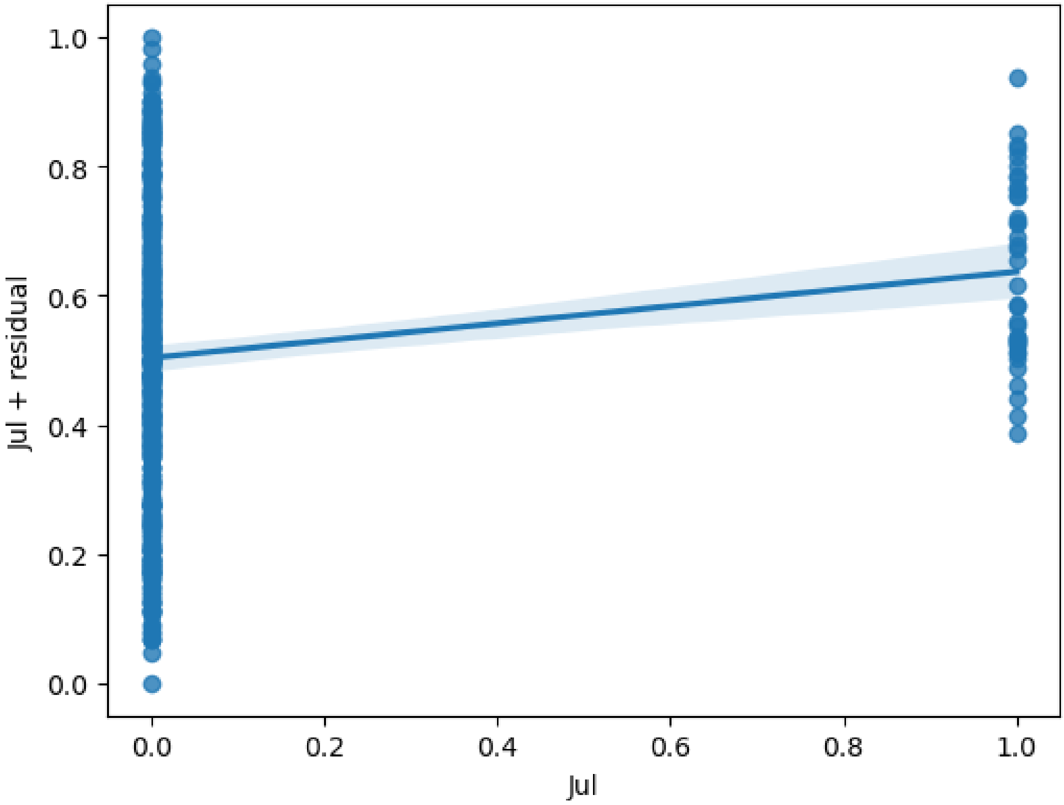
Correlation matrix of train set

	yr	holiday	temp	windspeed	Spring	Winter	Light Snow	Mist	Jul	Sep
yr	1.00	-0.01	0.11	-0.00	-0.04	-0.02	-0.06	-0.02	0.02	0.00
holiday	-0.01	1.00	-0.07	0.02	0.05	0.05	-0.03	-0.04	-0.05	0.04
temp	0.11	-0.07	1.00	-0.19	-0.61	-0.23	-0.04	-0.09	0.41	0.21
windspeed	-0.00	0.02	-0.19	1.00	0.18	-0.09	0.09	-0.03	-0.09	-0.12
Spring	-0.04	0.05	-0.61	0.18	1.00	-0.33	-0.04	0.01	-0.16	-0.17
Winter	-0.02	0.05	-0.23	-0.09	-0.33	1.00	0.11	0.02	-0.17	-0.02
Light Snow	-0.06	-0.03	-0.04	0.09	-0.04	0.11	1.00	-0.13	-0.01	0.03
Mist	-0.02	-0.04	-0.09	-0.03	0.01	0.02	-0.13	1.00	-0.11	0.04
Jul	0.02	-0.05	0.41	-0.09	-0.16	-0.17	-0.01	-0.11	1.00	-0.09
Sep	0.00	0.04	0.21	-0.12	-0.17	-0.02	0.03	0.04	-0.09	1.00

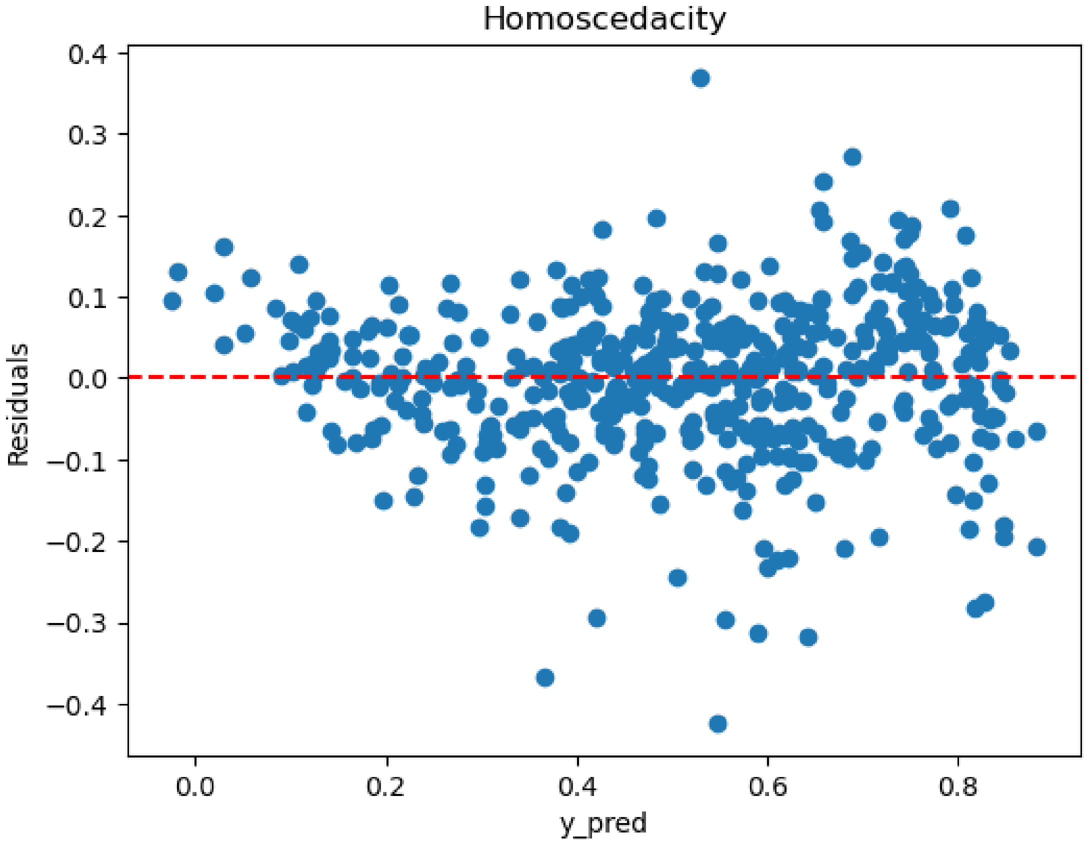
3. Linear relationship between X and Y







4. Homoscedacity – No patterns should be visible in the residual values



5. Independence of residuals – The residuals are independent of each other, indicating no correlation or autocorrelation. Auto correlation is also evaluated on the correlation matrix.

I verified the validity and dependability of the linear regression model and provided appropriate results interpretations by analysing these presumptions.

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

Solution:

Three key characteristics primarily affect the demand for shared bikes:

Temperature, the wintertime, and September

The variations in demand for shared bikes have been found to be mostly explained by these three variables.

General Subjective Questions

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

Solution:

Linear regression is a machine learning algorithm used to predict continuous numeric values. It establishes a linear relationship between input features (independent variables) and the target variable (dependent variable) by fitting a straight line that best represents the data. The goal is to find the line that minimizes the differences between predicted and actual values.

Here's a simplified explanation of linear regression:

1. Assumptions:

- Linearity: Assumes a straight-line relationship between variables.
- Independence: Assumes observations are unrelated.
- Homoscedasticity: Assumes constant variability in errors.
- Normality: Assumes errors follow a normal distribution.

2. Simple Linear Regression:

Deals with one independent variable and one dependent variable. It predicts the dependent variable based on the independent variable using a linear equation.

3. Multiple Linear Regression:

Extends simple linear regression to multiple independent variables. It predicts the dependent variable based on multiple independent variables using a linear equation.

4. Model Training:

Coefficients (intercept and slopes) are estimated to minimize the difference between predicted and actual values. Techniques like Ordinary Least Squares (OLS) or gradient descent are used.

5. Model Evaluation:

Metrics such as MSE, RMSE, MAE, and R2 score assess the model's performance and how well it fits the data.

6. Making Predictions:

Once trained, the model can make predictions on new data by calculating the predicted value using the learned coefficients.

Linear regression is commonly used for trend analysis, forecasting, and understanding variable relationships.

2. **Explain the Anscombe's quartet in detail.**

Solution:
Anscombe's quartet comprises four datasets that possess identical statistical characteristics but exhibit distinct graphical patterns. Developed by statistician Francis Anscombe in 1973, this quartet serves to emphasize the significance of data visualization and caution against relying solely on summary statistics.

Each dataset consists of eleven pairs of x and y values, sharing the same mean, variance, correlation, and linear regression parameters. However, the distributions and relationships between the variables differ across the datasets. Let's delve into a detailed description of each dataset:

1. Dataset I:
x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
y: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68
Relationship: Approximate linearity

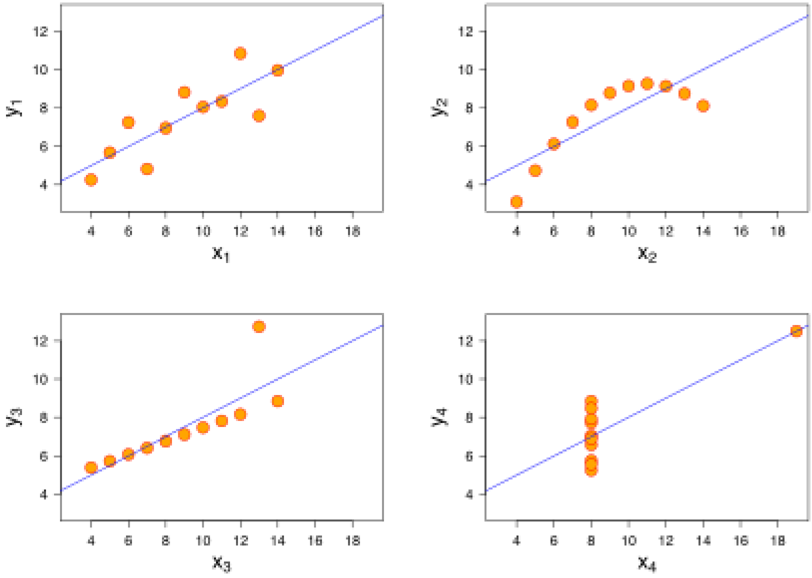
2. Dataset II:
x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
y: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73
Relationship: Non-linear with an apparent quadratic trend

3.Dataset III:

x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
y: 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74
Relationship: Approximate linearity with slight upward curvature

4. Dataset IV:

x: 8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8
y: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89
Relationship: A single outlier significantly impacts the linear regression line



The significance of Anscombe's quartet lies in challenging the notion that summary statistics alone are sufficient for understanding and analysing data. Despite having identical statistical properties, each dataset demonstrates unique patterns when visualized. This highlights the importance of data visualization in gaining insights, identifying outliers, assessing relationships, and verifying assumptions. Anscombe's quartet serves as a reminder that relying solely on summary statistics without considering the data's graphical representation can lead to erroneous conclusions and oversimplification. It underscores the need to explore and visualize data to gain a comprehensive understanding of its characteristics and relationships.

3. What is Pearson’s R?

Solution:

Pearson's R, or the Pearson correlation coefficient, is a statistical measure used to quantify the linear relationship between two continuous variables. It is represented by the symbol "r" and can take values between -1 and 1.

The calculation of Pearson's R involves the following formula:
r = (Σ((Xi - X_mean) * (Yi - Y_mean))) / (sqrt(Σ(Xi - X_mean)^2) * sqrt(Σ(Yi - Y_mean)^2))

Here are key characteristics of Pearson's R:

Range: The value of r ranges from -1 to 1. A positive value indicates a positive linear relationship, a negative value indicates a negative linear relationship, and a value of 0 suggests no linear relationship between the variables.

Strength of Relationship: The magnitude of r indicates the strength of the linear relationship. Values closer to -1 or 1 represent a stronger linear association, while values closer to 0 indicate a weaker relationship.

Significance: The statistical significance of Pearson's R can be determined using hypothesis testing. By examining the associated p-value, it is possible to assess whether the observed correlation is statistically significant or occurred by chance. A p-value below a chosen significance level (e.g., 0.05) suggests a significant correlation.

Assumptions: Pearson's R relies on certain assumptions, including the linearity of the relationship between variables, the normal distribution of variables, and homoscedasticity (constant variance of the residuals).

Pearson's R finds wide application in various fields, such as statistics, social sciences, finance, and machine learning. It is used to evaluate the strength and direction of the linear relationship between variables, aiding in prediction, pattern recognition, and identification of associations within the data.

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

Solution:
Scaling is a preprocessing step commonly used in machine learning to transform the features of a dataset into a consistent scale. Its purpose is to ensure that all features have similar ranges, which can be crucial for certain machine learning algorithms and data analysis techniques. Scaling addresses issues related to feature magnitudes and units, preventing features with larger values from dominating the learning process.

The main reasons for performing scaling are as follows:

Comparison of Features: Scaling enables a fair comparison between different features that have diverse units and scales. Without scaling, features with larger values may disproportionately influence the learning process or distance-based algorithms.

Gradient Descent Optimization: Scaling assists in optimizing the performance of gradient descent-based algorithms, which rely on iteratively updating weights. Scaling the features can enhance convergence and expedite the learning process.

Regularization Techniques: Scaling is often necessary when employing regularization techniques like Ridge regression or Lasso regression. These techniques penalize large weights, and without scaling, the regularization term may excessively penalize features with larger values.

There are two commonly used scaling methods:

Normalized Scaling (Min-Max Scaling):

Normalized scaling, also known as Min-Max scaling, rescales the features to a fixed range, typically between 0 and 1.
The formula for normalized scaling is: X_scaled = (X - X_min) / (X_max - X_min)
Here, X represents the original feature value, X_min and X_max are the minimum and maximum values of the feature, respectively.
Normalized scaling preserves the original distribution of the data while scaling it to a specific range.

Standardized Scaling (Z-score normalization):

Standardized scaling, also known as Z-score normalization, transforms the features to have a zero mean and unit variance.
The formula for standardized scaling is: X_scaled = (X - X_mean) / X_std

Here, X denotes the original feature value, X_mean represents the mean of the feature, and X_std is the standard deviation of the feature.
Standardized scaling centers the data around zero and scales it based on the spread of the data.
The primary difference between normalized scaling and standardized scaling lies in the range and distribution of the scaled data. Normalized scaling transforms the data to a fixed range (e.g., 0 to 1), while standardized scaling centers the data around zero and scales it based on the spread of the data. Normalized scaling is suitable when preserving the original distribution and range is important, while standardized scaling is useful when comparing features with different scales and for algorithms that assume normally distributed data.

In summary, scaling is performed to ensure consistency in feature scales, facilitate fair feature comparisons, and enhance the performance of machine learning algorithms. Normalized scaling maintains the original distribution and scales the data to a fixed range, while standardized scaling centers the data around zero and scales it based on the spread of the data. The choice between the two methods depends on the specific requirements and characteristics of the dataset and the machine learning algorithm being used.

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

Solution:
Perfect multicollinearity occurs when there is an exact linear relationship between one or more independent variables in a regression model, resulting in unstable parameter estimates and infinite values of Variance Inflation Factor (VIF). Here are some common scenarios that lead to perfect multicollinearity:

Duplicate or Redundant Variables: When two or more variables are identical or perfectly correlated, it introduces redundancy in the information they provide. Including both of these variables in the model would result in perfect multicollinearity.

Data Transformation Issues: Applying inappropriate data transformations can also lead to perfect multicollinearity. For example, if a continuous variable is converted into categorical bins and all the bins are included as independent variables, it can introduce perfect multicollinearity.

Creation of Derived Variables: When new variables are created from existing variables using mathematical operations, it is crucial to avoid inadvertently introducing perfect multicollinearity. If a new variable is created by summing two existing variables that are perfectly correlated, it will result in perfect multicollinearity.

Including Interactions or Polynomial Terms: Interaction terms or higher-order polynomial terms can introduce perfect multicollinearity if they involve variables that are perfectly correlated or linearly related. Including both "age" and "age squared" as independent variables is an example that can lead to perfect multicollinearity.

To address perfect multicollinearity and infinite VIF values, it is necessary to identify and remove the redundant or perfectly correlated variables from the model. This can be done through careful examination of the variables, performing feature selection techniques, or applying dimensionality reduction methods like principal component analysis (PCA).

It's important to note that while infinite VIF values indicate the presence of perfect multicollinearity, high VIF values (but not infinite) suggest strong multicollinearity between variables. In such cases, it is advisable to assess the impact of multicollinearity on the model's performance and consider addressing it by removing or transforming the correlated variables.

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

Solution:
A Q-Q plot, also known as a quantile-quantile plot, is a graphical method used to evaluate the similarity between a given dataset and a theoretical distribution, such as the normal distribution. It displays the quantiles of the observed data against the quantiles of the theoretical distribution, enabling visual examination and identification of deviations from the expected distribution.

The significance and application of a Q-Q plot in linear regression can be summarized as follows:

Assessing the Assumption of Normality:
In linear regression, one of the assumptions is that the residuals (the differences between the observed and predicted values) adhere to a normal distribution. The Q-Q plot assists in evaluating this assumption by comparing the quantiles of the residuals to the quantiles expected from a normal distribution. If the points on the Q-Q plot approximately align along a straight line, it suggests that the residuals follow a normal distribution. Any deviations from the straight line indicate departures from normality.

Detecting Skewness and Outliers:
The Q-Q plot can detect skewness and identify outliers within the data. If the points on the plot systematically deviate from the straight line, it indicates skewness or heavy tails in the distribution. Outliers can be recognized as points that significantly

deviate from the expected line. This information helps in understanding the distributional characteristics of the residuals and identifying influential data points that may impact the linear regression model.

Model Evaluation and Assumption Checking:
The Q-Q plot serves as a diagnostic tool to evaluate the adequacy of the linear regression model. Deviations from the expected line in the Q-Q plot may indicate model misspecification or violations of assumptions. If the Q-Q plot reveals substantial departures from the expected line, it suggests that the linear regression model may not be suitable for the data, and further investigation or refinement of the model might be necessary.

Comparison of Distributions:
In addition to assessing normality assumptions, Q-Q plots can be employed to compare the distributions of two datasets. By plotting the quantiles of two datasets against each other, it becomes easier to compare their distributions, identify differences in skewness, or detect systematic deviations.

To summarize, the Q-Q plot is a valuable tool in linear regression for evaluating the normality assumption, detecting skewness and outliers, assessing model adequacy, and comparing distributions. It provides insights into the distributional characteristics of the residuals and helps identify potential issues that may impact the validity and reliability of the linear regression analysis.