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)**
   Answer:
   I have analyzed categorical columns using box plots and bar charts. Here are the key insights:

   - Fall season attracts more bookings, with significant increases from 2018 to 2019 across all seasons.
   - The peak booking months are May to October, with a trend increasing until mid-year and declining thereafter.
   - Clear weather clearly correlates with higher booking rates.
   - Thu, Fri, Sat, and Sun have higher booking rates compared to weekdays.
   - Bookings tend to decrease on non-holiday days, which is expected as people prefer to spend holidays with family.
   - Bookings are almost evenly distributed between working and non-working days.
   - In 2019, bookings increased significantly compared to the previous year, indicating positive business growth.
2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
   Answer:

- **Avoiding Dummy Variable Trap**: Including all levels (k) of a categorical variable as dummy variables can lead to multicollinearity issues in linear regression. To mitigate this, we use k-1 dummy variables, with one level omitted as a reference handled by the intercept.
-
- **Handling Nominal Categories**: Nominal categorical variables like colors (e.g., Red, Green, Blue) lack inherent order. Encoding them numerically (e.g., 1, 2, 3) may introduce unintended order biases (e.g., Red < Green < Blue). Dummy encoding preserves the categorical distinctions with binary indicators (0 or 1) for each category, preventing such biases and ensuring accurate model interpretation.
-
- **Numeric Conversion for Model Understanding:** Regression models require numerical inputs; hence, string or text data, such as categorical variables, must be converted into numeric form through methods like dummy encoding to facilitate model comprehension and analysis.

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

Answer:

I have assessed the assumptions of the Linear Regression Model based on the following criteria:

- Normality of error terms: Error terms should follow a normal distribution.
- Multicollinearity check: There should be insignificant multicollinearity among variables.
- Validation of linear relationships: Variables should exhibit linear relationships.
- Homoscedasticity: Residual values should display no discernible pattern.
- Independence of residuals: There should be no autocorrelation among residuals.

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

Answer:

Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –

- atemp (coef: .4714)
- year (coef: .2355)
- sep (coef: .0873)

General Subjective Questions

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

Answer:

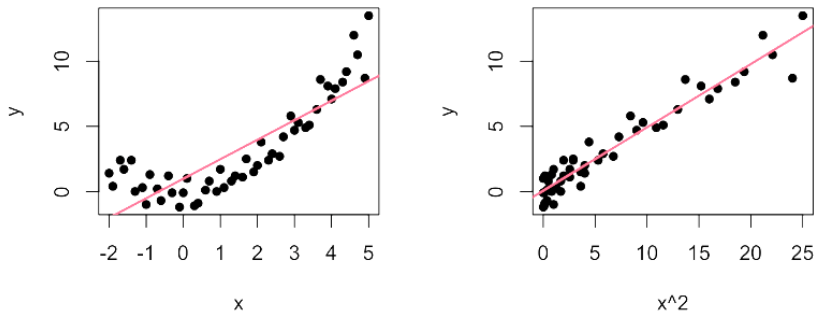
Linear Regression finds the best linear relationship between the independent and dependent variables.

It is a method of finding the best straight-line fitting to the given data.

In technical terms, linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables. It is mostly done by the Sum of Squared Residuals Method.

The assumptions of linear regression are:

a. **The assumption about the form of the model**: It is assumed that there is a linear relationship between the dependent and independent variables.



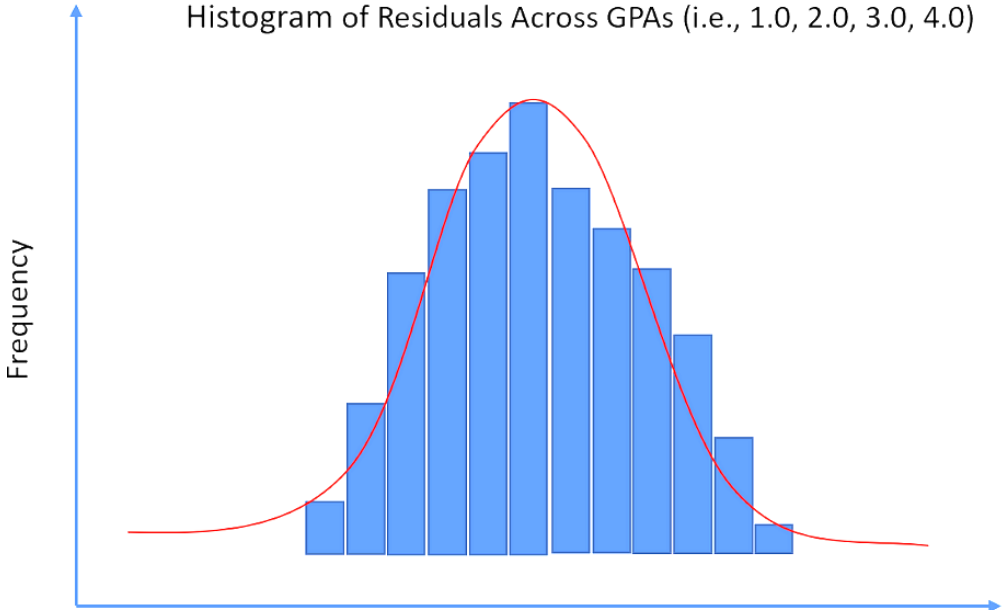
b. **Assumptions about the residuals**:

1) *Normality assumption*: It is assumed that the error terms, ε(i), are normally distributed.

2) *Zero mean assumption:* It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.

3) *Constant variance assumption*: It is assumed that the residual terms have the same (but unknown) variance, sigma square. This assumption is also known as the assumption of homogeneity or homoscedasticity.

4) *Independent error assumption*: It is assumed that the residual terms are independent of each other, i.e., their pair-wise covariance is zero.



c. Assumptions about the estimators:

1) The independent variables are measured without error.

2) The independent variables are linearly independent of each other, i.e., there is no multicollinearity in the data.

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

Answer:

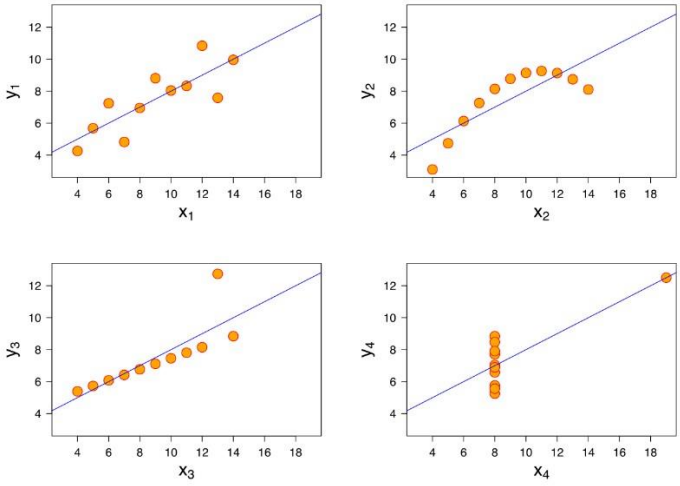
Anscombe's Quartet, devised by statistician Francis Anscombe, consists of four datasets, each containing eleven pairs of (x, y) values. Despite having identical descriptive statistics, these datasets reveal entirely distinct relationships when graphed. Each graphical representation tells a unique and sometimes contrasting story, highlighting the critical importance of visualizing data to understand its true nature beyond summary statistics. This underscores the potential pitfalls of relying solely on summary statistics without examining the actual data distribution and patterns visually.

	I		II		III		IV	
	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	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	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
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

The summary statistics show that the means and the variances were identical for x and y across the groups:

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
- The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

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

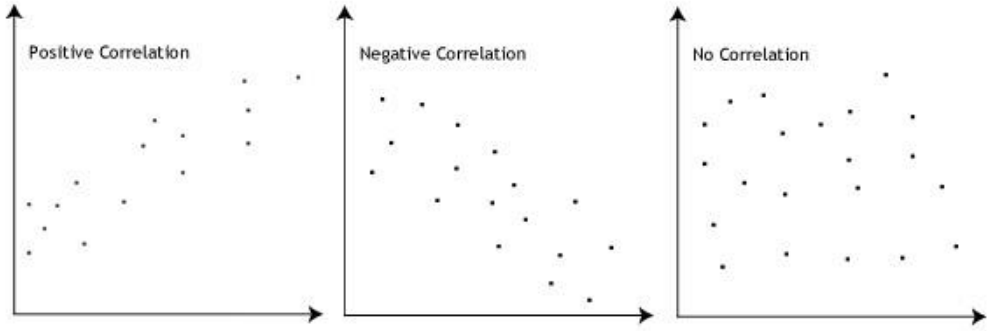
Answer: Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will

be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative

The coefficient varies between -1 and +1 where:

r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
r = 0 means there is no linear association
r > 0 < 5 means there is a weak association
r > 5 < 8 means there is a moderate association
r > 8 means there is a strong association

The figure below shows some data sets and their correlation coefficients.



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

Scaling in the context of machine learning refers to the process of transforming data attributes to a standard range. Here's a breakdown of scaling, its purpose, and the differences between normalized scaling and standardized scaling:

Scaling:

1. **Definition**:
    - **Scaling** adjusts the range of data values to a standard scale, making comparisons between different variables more meaningful.
2. **Purpose**:
    - **Normalization**: Ensures that all features are on a similar scale, typically between 0 and 1, preserving the relative differences in the ranges of the original data.
    - **Standardization**: Centers the feature columns at mean 0 with standard deviation 1, making the data more Gaussian-like.

Why is Scaling Performed?

- **Facilitates Model Convergence**: Many machine learning algorithms perform better or converge faster when features are on a relatively similar scale.
- **Improves Interpretation**: Scaling helps in comparing the importance of different features based on their magnitude.
- **Enhances Regularization**: Regularization techniques like L1 and L2 penalize coefficients based on their scale, so scaling can improve the regularization effect.
- **Mitigates Numerical Instabilities**: Algorithms like gradient descent are sensitive to the scale of features, and scaling can stabilize these algorithms.

S.NO.	Normalized scaling	Standardized scaling
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

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

Your explanation touches on the concepts of Variance Inflation Factor (VIF) and its implications regarding multicollinearity. Here's a refined version for clarity:

When VIF reaches infinity, it indicates a perfect correlation between two independent variables. In such cases, the R-squared (R2) value approaches 1, resulting in a calculation issue where 1 / (1 - R2) becomes infinite. To resolve this, one of the variables causing this perfect multicollinearity should be removed from the dataset. High VIF values, such as 4, suggest significant correlation

among variables, inflating the variance of the model coefficients by a factor of 4 due to multicollinearity. Proper handling of multicollinearity ensures model stability and accurate interpretation of regression coefficients.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.** **(3 marks)**

Answer:

Q-Q Plot (Quantile-Quantile Plot)

A Q-Q plot, short for Quantile-Quantile plot, is a graphical tool used to assess if a dataset follows a particular distribution. It compares the quantiles of the dataset against the quantiles of a theoretical distribution, typically a normal distribution.

Use and Importance in Linear Regression:

1. **Assessing Normality Assumption**:
   - Use: In linear regression, one of the key assumptions is that the error terms (residuals) should be normally distributed. Q-Q plots help in visually inspecting whether the residuals of the model approximate a normal distribution.
   - Importance: If the residuals deviate significantly from the diagonal line (which represents normality), it indicates that the assumption of normality might be violated. This can affect the validity of statistical inference, such as confidence intervals and hypothesis tests.

2. **Identifying Outliers**:
   - Use: Q-Q plots can reveal outliers in the residuals. Outliers are data points that deviate substantially from the expected pattern (in this case, normality).
   - Importance: Outliers can influence the regression model's coefficients and predictions, potentially biasing the results. Q-Q plots help identify these outliers for further investigation.

3. **Model Diagnostics**:
   - Use: Beyond normality, Q-Q plots provide an overall diagnostic tool for the adequacy of the regression model. They can highlight systematic deviations from the expected distributional assumptions.
   - Importance: Detecting and addressing deviations early ensures the model's reliability and helps in making informed decisions about model adjustments or transformations of variables.

Interpreting Q-Q Plot:
- Ideal Scenario: In an ideal case, the points on the Q-Q plot should lie approximately on the diagonal line, indicating that the residuals follow a normal distribution.
- Deviation: If the points deviate from the diagonal line, it suggests that the residuals do not follow a normal distribution. Skewness, heavy tails, or systematic patterns in deviations can indicate issues that need addressing.