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:

a. 'fall' has highest demand for rental bikes.
b. Demand has grown for next year (2019) than previous year (2018).
c. The demand experiences a consistent growth every month until June. The month of September stands out with the highest demand. However, after September, the demand starts to decline.
d. Demand is decreasing in 'holiday'
e. The demand is not clearly reflected by the weekday.
f. Booking seems to be almost equal either on working day or non-working day.
g. The good ('Clear, Few clouds, Partly cloudy, Partly cloudy') weathershit has highest demand.
h. Demand could be declining after September due to weather condition.

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

Answer:

a. Utilizing drop_first = True is essential as it help in decreasing the surplus column produced while generating dummy variables. This, in result, lessens the correlations formed among the dummy variables.
b. **drop_first: bool**, default is **False**, which implies to get n-1 dummy variables out of n categorical variable levels by removing the first level.
c. Assuming we have a Categorical column with three distinct values, we aim to generate a dummy variable for this column. If a variable is neither A nor B, it can be inferred that it is C. Therefore, there is no necessity for a third variable to identify C.

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

Answer:

a. The variables "temp" and "atemp" exhibit the strongest correlation with the target variable "cnt".
b. Note: since “temp” and “atemp” exhibits Multicollinearity, we have to remove “atemp” from dataframe.

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

Answer:

a. Multicollinearity
    i. No significant multicollinearity among variables.
b. Linear relationship

i. There should be linear relationship among variables.

c. Normal distribution of error terms
    i. Residual errors should be normally distributed.
d. Homoscedasticity
    i. The residual values must not show any visible pattern.
e. Residuals to be independent
    i. There should not a relationship between the residuals and the variable.

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

Answer:

a. temp
b. winter
c. september

General Subjective Questions

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

Answer

Linear regression is a statistical model that examines the linear association between a dependent variable and a given set of independent variables. This model indicates that when the value of one or more independent variables changes (either increases or decreases), the value of the dependent variable will also change correspondingly (either increase or decrease).

Mathematical formula : $\mathbf{Y = \beta_0 + \beta_1 X}$

where

Y is the dependent variable.
X is the independent variable.
β1 is the slope of the regression line which represents the effect X has on Y
$\mathbf{\beta_0}$ is a constant, known as the Y-intercept. If X = 0, Y would be equal to $\mathbf{\beta_0}$

There are two types of Linear Regression

a. Simple Linear Regression
b. Multiple Linear Regression

Linear Regression Model has following assumptions.

Multi-collinearity

Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.

Autocorrelation

The Linear regression model also assumes that there is minimal or no autocorrelation present in the data. Autocorrelation refers to the existence of a relationship between residual errors.

Relationship between variables

There should be linear relationship among variables.

Normality of error terms

Error terms should be normally distributed.

Homoscedasticity

There should be no visible pattern in residual values.

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

Answer

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet they differ significantly when graphically represented and analysed. This quartet was created by the statistician Francis Anscombe in 1973 to emphasize the importance of visualizing data and not relying solely on summary statistics.

The four datasets in Anscombe's quartet share the same mean, variance, correlation coefficient, and linear regression line, but they exhibit distinct patterns when graphed. Each dataset consists of 11 (x, y) pairs. Let's take a closer look at each of the four sets:

	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

Dataset I:

- **Description:** A simple linear relationship.
- **Graphical Representation:** A scatter plot with a clear linear trend.
- **Summary Statistics:**
  - Mean of x: 9.0
  - Mean of y: 7.5
  - Variance of x: 11.0
  - Variance of y: 4.12
  - Correlation coefficient: 0.82
  - Linear regression: y = 3.0 + 0.5 * x

Dataset II:

- **Description:** A non-linear relationship.
- **Graphical Representation:** A scatter plot with a curve; not a simple linear relationship.
- **Summary Statistics:**
  - Mean of x: 9.0

- Mean of y: 7.5
- Variance of x: 11.0
- Variance of y: 4.12
- Correlation coefficient: 0.82
- Linear regression: y = 3.0 + 0.5 * x

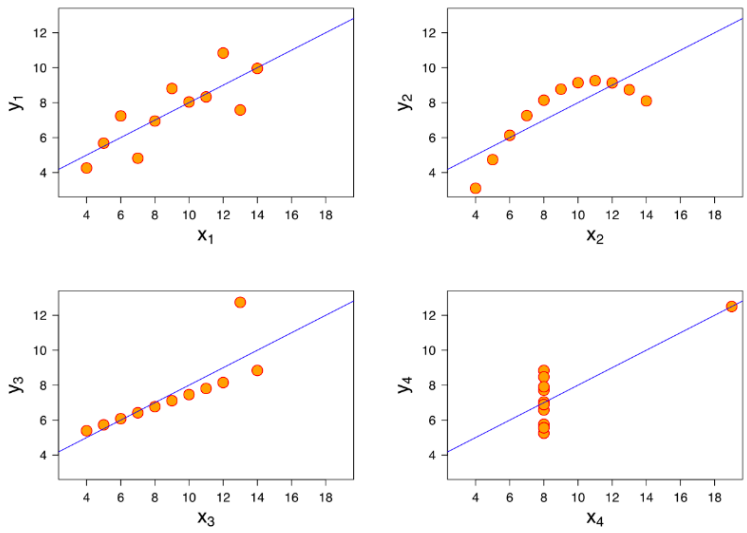
Dataset III:

- **Description:** A situation with an outlier.
- **Graphical Representation:** A scatter plot with a clear linear trend, except for one outlier.
- **Summary Statistics:**
  - Mean of x: 9.0
  - Mean of y: 7.5
  - Variance of x: 11.0
  - Variance of y: 4.12
  - Correlation coefficient: 0.82
  - Linear regression: y = 3.0 + 0.5 * x

Dataset IV:

- **Description:** Two distinct clusters.
- **Graphical Representation:** Two separate clusters with different linear relationships.
- **Summary Statistics:**
  - Mean of x: 9.0
  - Mean of y: 7.5
  - Variance of x: 11.0
  - Variance of y: 4.12
  - Correlation coefficient: 0.82
  - Linear regression: y = 3.0 + 0.5 * x

When we graph these four datasets on an x/y coordinate plane, it becomes evident that they exhibit identical regression lines. However, each dataset conveys a distinct narrative.



The quartet illustrates the importance of visually inspecting data through plots and graphs, as relying solely on summary statistics can be misleading. Even though these datasets share similar basic statistics, their underlying structures are vastly different.

This emphasizes the need for exploratory data analysis and visualization to gain a more comprehensive understanding of the data.

3. **What is Pearson's R?**

Answer

Pearson's correlation coefficient, often denoted as "r" or "Pearson's r," is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It was developed by Karl Pearson, a pioneer in the field of statistics.

The Pearson correlation coefficient is a dimensionless value that ranges from -1 to 1, where:

- 1 indicates a perfect positive linear relationship: as one variable increases, the other variable increases proportionally.
- -1 indicates a perfect negative linear relationship: as one variable increases, the other variable decreases proportionally.
- 0 indicates no linear correlation: the variables are not linearly related.

The formula for Pearson's correlation coefficient (r) is given by:

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

where

X_i and Y_i are the individual data points for the two variables X and Y

$\overline{X}$ and $\overline{Y}$ are the means of the variables X and Y, respectively.

The numerator in the formula represents the covariance between X and Y, while the denominator represents the product of the standard deviations of X and Y. The division normalizes the covariance, resulting in a correlation coefficient that is independent of the scales of the variables.

Pearson's correlation coefficient is widely used in various fields, including statistics, economics, psychology, and the natural sciences, to assess the degree and direction of linear relationships between two variables. It's important to note that Pearson's r specifically measures linear correlations and may not capture non-linear relationships between variables.

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

Answer

Feature scaling is the process of transforming the numerical features of a dataset into a standard scale. It involves adjusting the range or distribution of values of different features to make them comparable and prevent certain features from dominating the analysis due to differences in scale. Feature scaling is commonly performed in machine

learning and statistical modelling to ensure that algorithms operate effectively and provide meaningful results.

Reasons for Feature Scaling

1. **Magnitude Adjustment:** Features with larger magnitudes may dominate those with smaller magnitudes in certain algorithms and analyses.
2. **Algorithm Sensitivity:** Some machine learning algorithms, particularly those based on distances or gradients, are sensitive to the scale of features. Scaling helps prevent bias towards features with larger scales.

Difference between Normalized Scaling and Standardized Scaling:

Normalized Scaling

- **Objective:** Scaling the features to a specific range, often [0, 1].
- **Formula:** $X\text{normalized} = (X - X\text{min}) / (X\text{max} - X\text{min})$
- **Properties:** The resulting values are between 0 and 1, where 0 corresponds to the minimum value of the feature, and 1 corresponds to the maximum value.

Standardized Scaling (Z-score normalization or Standardization):

- **Objective:** Scaling the features to have a mean of 0 and a standard deviation of 1.
- **Formula**: $X\text{standardized} = (X - \mu)/ \sigma$
- Properties: The resulting values have a mean of 0 and a standard deviation of 1.

Key Differences:

Range: Normalized scaling confines values to the [0, 1] range, while standardized scaling centers values around 0 with a standard deviation of 1.

Sensitivity to Outliers: Standardized scaling is more robust against outliers since it uses the mean and standard deviation, while normalized scaling might be affected by extreme values.

Interpretability: Normalized scaling retains the original interpretability of the feature within the [0, 1] range, while standardized scaling transforms the feature into z-scores, making interpretation in the original units less straightforward.

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

   Answer:
   a. The most common scenario leading to infinite VIF is when a variable is a constant multiple of another variable or a combination of other variables. This scenario results in a situation where the denominator in the VIF formula becomes zero (1 – R^2 =0), leading to an infinite VIF value.
   b. When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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, which stands for Quantile-Quantile plot, is a graphical plot used to assess whether a dataset follows a normal distribution. In the context of linear regression, Q-Q plots are often used to check the assumption of normality in the residuals.

User of Q-Q plot As Below

Normality of Residuals:

- A key assumption of linear regression is that the residuals (the differences between observed and predicted values) are normally distributed.
- The Q-Q plot helps visualize whether the residuals follow a normal distribution. If the points deviate significantly from a straight line, it indicates potential departures from normality.

Identifying Outliers

- Outliers or extreme values in the residuals may cause deviations in the Q-Q plot. Identifying these deviations can help detect influential points in the analysis.

Diagnostic Tool

- Q-Q plots are part of a set of diagnostic tools used to assess the assumptions of linear regression. Other tools include residual plots, leverage plots, and Cook's distance.

Model Validity

- A linear regression model with normally distributed residuals is more valid, and its estimates are more reliable.
- Departures from normality might suggest that the model does not adequately capture the underlying structure of the data.