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?

The categorical variables used in the dataset were season, year, holiday, weekday and weathersit and mnth.

Season – The boxplot shows that fall has the maximum demand for bikes and sprint has the least
Weathersit – There were no demand for bikes during heavy rain indicating unfavorable conditions and demand was high during clear and partly cloudy conditions.
Year – The demand was high in 2019 when compared to 2018
Holiday – Demand for bikes reduced during holiday
Month – Demand for bikes was highest during September and least during December.

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

During dummy variable creation drop_first=True helps in reducing the extra column creation. It reduces the correlation created among dummy variables. If we have all dummy variables it leads to multicollinearity between dummy variables so , If we have n categorical variables we need n-1 columns to represent dummy variables.

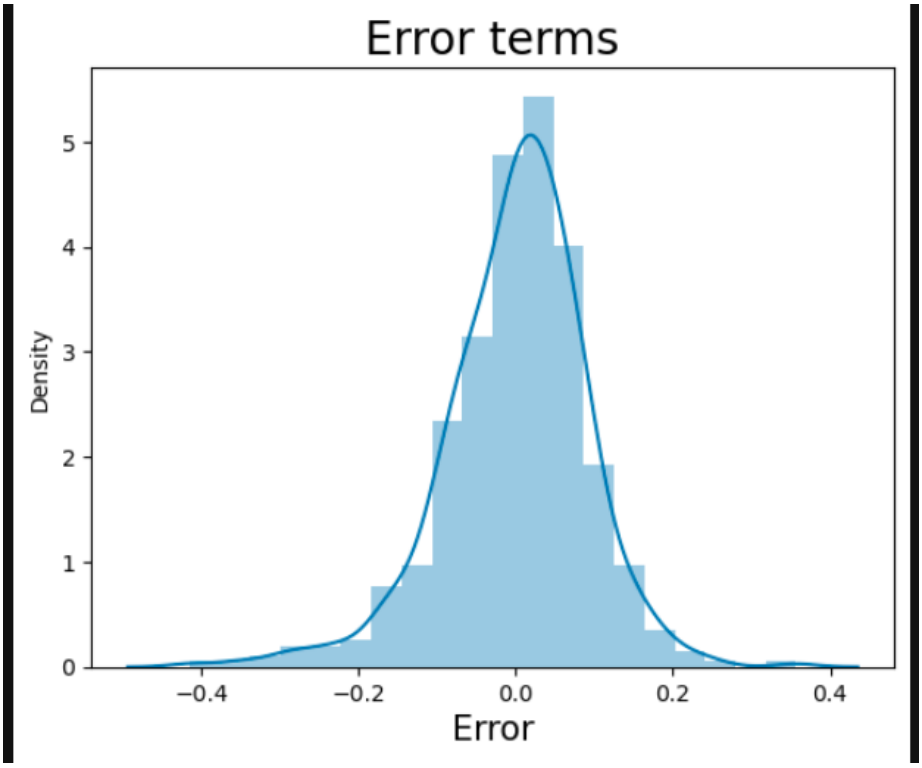
Eg: column for gender that contains 4 variables- "Male", "Female", "Other", "Unknown". So a person is either "Male", or "Female", or "Other". If they are not either of these 3, their gender is "Unknown".

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

Temp and atemp are the numerical variables which have the highest correlation with the target variable cnt.

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

We validate the normal distribution of error terms and mean is around 0 and we validate that by plotting a distplot of residuals are following normal distribution. The below graph shows the residuals are distributed at mean 0.



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

The top 3 features contributing significantly towards the demand of shared bikes are

Temp - 0.49

Year - 0.23

Weathersit - -0.28

General Subjective Questions

Q) Explain linear regression algorithm in detail?

Linear regression is a method used to predict a continuous outcome based on one or more input features. It finds the best-fitting straight line through the data points, such that the distance between data points and line is minimized and it is called the "regression line." The equation of this line is:

$y = \beta 0 + \beta 1x1 + \beta 2x2 + ... + \beta nxn + \varepsilon$

where:

- **y** is the dependent variable (the outcome we are trying to predict).
- **x1, x2, ..., xn** are the independent variables (the input features).
- **β0** is the y-intercept (the value of y when all x values are 0).
- **β1, β2, ..., βn** are the coefficients for the independent variables (they represent the change in y for a one-unit change in x).
- **ε** is the error term (the difference between the actual and predicted values).

It is mostly used in finding out relationship between variables and forecasting.

There are two types of linear regression:

1. **Simple Linear Regression**: Involves only one independent variable.
2. **Multiple Linear Regression**: Involves two or more independent variables.

To apply linear regression, certain assumptions must be met:

- **Linearity**: The relationship between the dependent and independent variables is linear.
- **Independence**: The residuals (errors) are independent.
- **Homoscedasticity**: The residuals have constant variance at every level of x.
- **Normality**: The residuals of the model are normally distributed.

Method of Least Squares

The most common method to determine the best-fitting line is the least squares method. This approach minimizes the sum of the squared differences between the observed values and the values predicted by the line.

Q2) Explain the Anscombe's quartet in detail.

Anscombe's Quartet was developed by the statistician ***Francis Anscombe.*** Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet appear very different when represented in graphs. It was developed to emphasize the importance of graphing data before analyzing. It shows the effect of outliers on statistical properties.

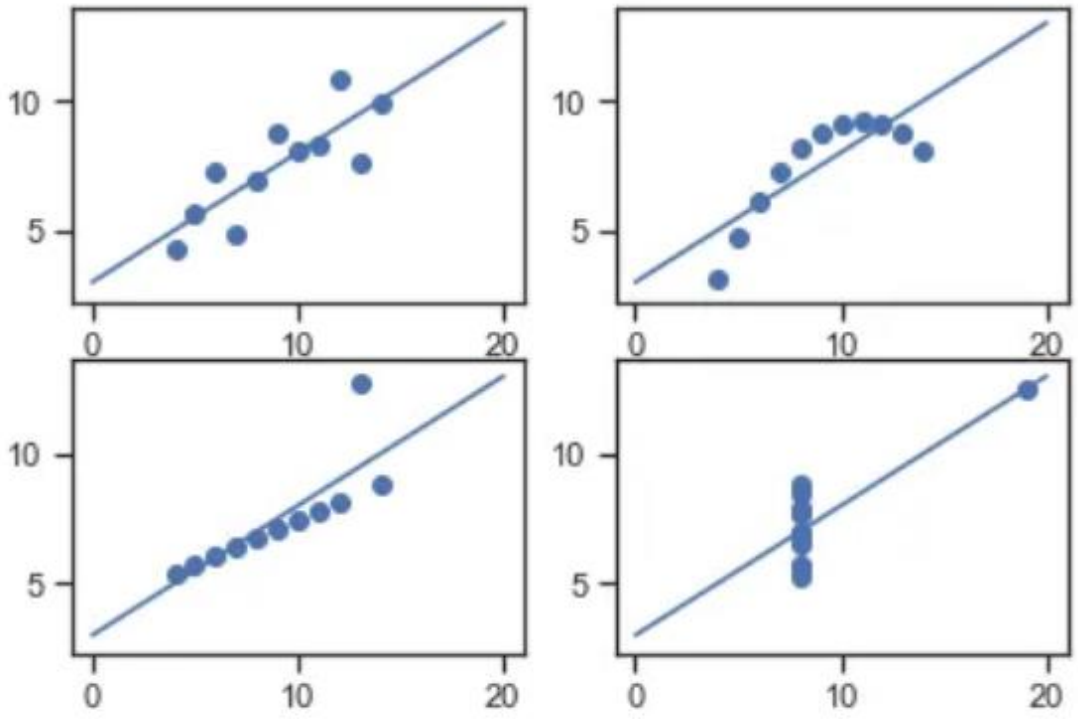
x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Four Data-sets

Apply the statistical formula on the above data-set,

Average Value of x = 9
Average Value of y = 7.50
Variance of x = 11
Variance of y =4.12
Correlation Coefficient = 0.816
Linear Regression Equation : y = 0.5 x + 3

Statistical data of these sets look similar, but when graphs are plotted each graph shows different behavior,



Graphical Representation of Anscombe's Quartet

The first graph shows linear relationship
The second graph is non linear relationship
The third graph looks like a tight linear relationship between *x* and *y*, except for one large outlier.
The last graph looks like the value of *x* remains constant, except for one outlier as well.

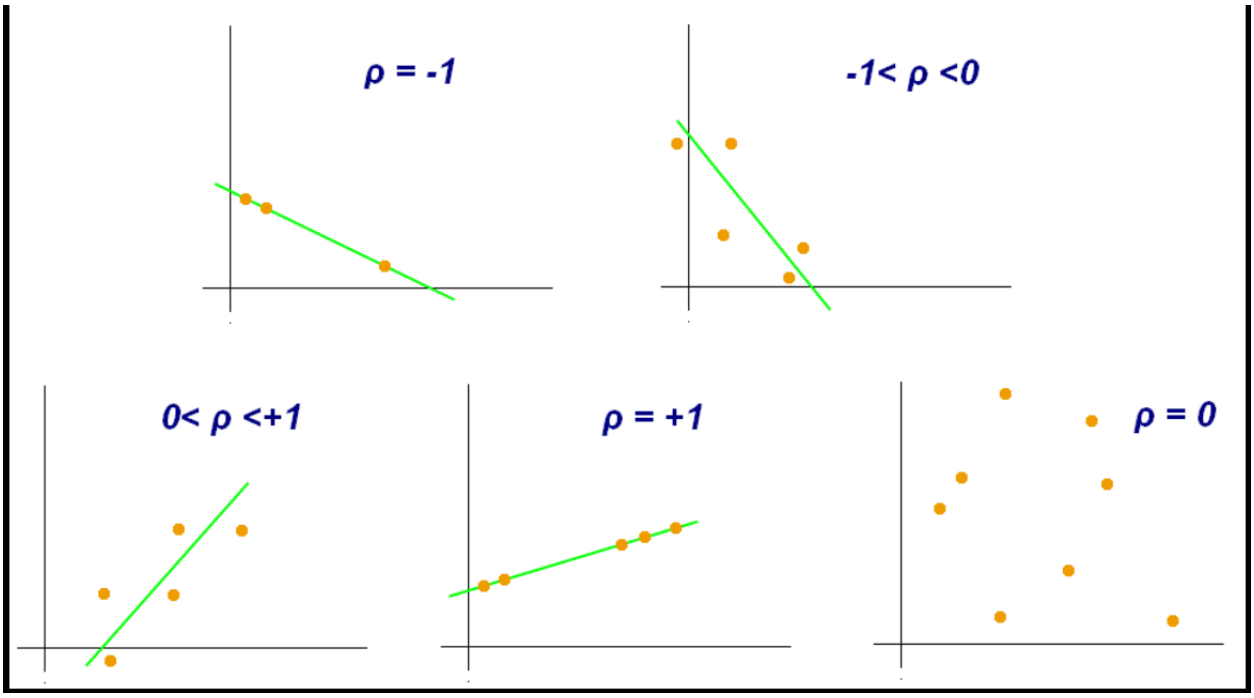
Anscombe's quartet visualizes data to uncover underlying patterns and anomalies that may not be apparent through summary statistics alone.

Q3) What is Pearson's R?

The pearson's correlation coefficient is the measure of the linear correlation between two variables. It ranges between -1 and 1 and indicates the strength and direction of the linear relationship.

Range of the value:
r = 1 perfect positive linear correlation
r = -1 perfect negative linear correlation
r = 0 No linear correlation

The below graph represent the same.



Calculation

$$r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}}$$

where:

- x_i and y_i are the individual data points.
- $\overline{x}$ and $\overline{y}$ are the means of the x and y variables, respectively.

Properties

1. **Range**: Pearson's R ranges from -1 to 1.
    - **r = 1**: Perfect positive linear correlation.
    - **r = -1**: Perfect negative linear correlation.
    - **r = 0**: No linear correlation.
2. **Symmetry**: Pearson's R is symmetric, meaning the correlation between x and y is the same as the correlation between y and x.

3. **Unit-Free**: Pearson's R is a unit-free measure, meaning it does not depend on the units of the variables being correlated.
4. **Linear Relationship**: Pearson's R measures the strength and direction of the linear relationship between two variables.

5. **Sensitivity to Outliers**: Pearson's R is sensitive to outliers, which can significantly affect the correlation coefficient.

Q4) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is a method to standardize or normalize data. It is performed during the data preprocessing stage to deal with data in different values. This is essential in many machine learning algorithm that rely on distance between data points. Without scaling, features with larger range may dominate and lead to biased model.

Normalized scaling
Also known as Min-Max scaling, this technique scales data to a range of 0 to 1. It's useful when the data's distribution is unknown or not normal. Normalization is more likely to be affected by outliers because of its restricted range.
Standardized scaling
This technique scales data to have a mean of 0 and a standard deviation of 1. It's useful when the data follows a Gaussian distribution. Standardization is less likely to be affected by outliers because it doesn't have a restricted range.

Q5) You might have observed that sometimes the value of VIF is infinite. Why does this happen?

VIF (variance inflation factor) is a measure to detect the severity of multicollinearity in regression analysis .If the VIF is perfectly correlated, then VIF is infinity. It gives the quantitative idea about how much the variables are correlated with each other. It is vital parameter to test our linear regression model.

$$VIF = \frac{1}{1 - R^2}$$

When R-1 is the R squared value of the independent variable, which we want to check how well it is explained by other variables. If it is well explained its perfect correlation and its R2 values will be 1. So VIF = 1/1-1 which results in infinity. This indicates independent variable can be perfectly explained by leading to exact linear dependency.

Interpreting VIF
1 = not correlated
Between 1 and 5 = moderately correlated
Above 5 = Highly correlated

Q6) What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A **Q-Q plot** (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a datasets, such as the normal distribution. It plots the quantiles of the dataset against the quantiles of the another data set. If Q-Q plot lie approximately along a straight line, it indicates that the dataset came from same distribution.

Use and Importance in Linear Regression:

1. **Assessing Normality**: In linear regression, one of the assumptions is that the residuals (errors) are normally distributed. A Q-Q plot helps to visually assess whether this assumption holds true.
2. **Identifying Outliers**: Q-Q plots can help identify outliers or deviations from the expected distribution, which can affect the performance of the regression model.
3. **Model Validation**: By comparing the residuals to a normal distribution, a Q-Q plot helps validate the linear regression model and ensures that the assumptions are not violated.

In summary, a Q-Q plot is an essential tool in linear regression for assessing the normality of residuals, identifying outliers, and validating the model.