Assignment-based Subjective

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

•Bike rental is comparative high in Fall and Summer seasons.

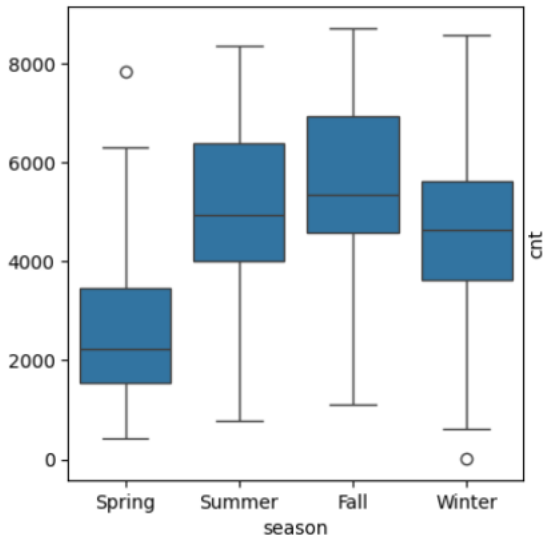
•Bike demand takes a dip in Spring.

•Bike demand in year 2019 is higher as compared to 2018.

•Bike demand is high in the months of September and October.

•Bike demand is high if weather is clear or with mist cloudy while it is low when there is light rain or light snow.

•The demand of bike is almost similar throughout the weekdays.



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

drop_first=True is important to use in order to achieve n-1 dummy variables as it can be used to delete extra column while creating dummy variables.

For Example: We have three variables: Furnished, Semi-furnished and un-furnished. We can only take 2 variables as furnished will be 1-0, semi-furnished will be 0-1, so we don't need unfurnished as we know 0-0 will indicate un-furnished. So we can remove it .

It is also used to reduce the collinearity between dummy variables . Hence if we have n categorical variables then we can use n-1 columns to represent the dummy variables.

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

temp and atemp both have same correlation with target variable cnt of 0.63 which is the highest among all numerical variables.

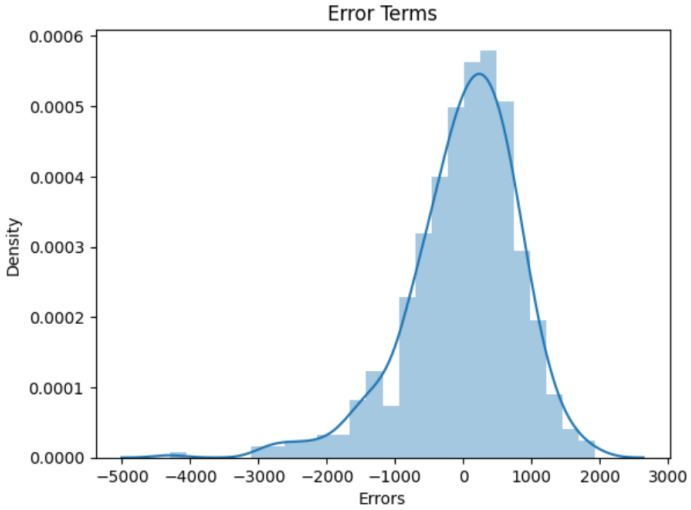
4.How did you validate the assumptions of Linear Regression after building the model on the training set?

Validating the assumptions of linear regression is a crucial step to ensure the reliability of the model. After building the model on the training set, here are the steps I followed to validate the assumptions:

1. Residual Analysis:

- Process: Examine the residuals (the differences between observed and predicted values).

- Check: Residuals should be approximately normally distributed, and there should be no discernible patterns in the residual plot.



2. Homoscedasticity (Constant Variance):

- Process: Plot residuals against predicted values.

- Check: The spread of residuals should be roughly constant across all levels of the predicted values.

3. Linearity:

- Process: Create a scatterplot of observed vs. predicted values.

- Check: The points should fall approximately along a diagonal line, indicating a linear relationship

4. Independence of Residuals: - Process: Examine residuals for autocorrelation. - Check: There should be no discernible pattern in the residuals when plotted against >me or other relevant variables.

5. Multicollinearity: - Process: Calculate Variance Inflation Factors (VIF) for predictor variables. - Check: VIF values should be below a certain threshold (commonly 5 or 10) to ensure no problematic multicollinearity.

6. Cross-Validation: - Process: Validate the model on a test set or through cross-validation. - Check: Assess the model's performance on new data to ensure generalizability and consistency. 7. Check for Overfitting: - Process: Evaluate model performance on a test set. - Check: Ensure that the model generalizes well to new, unseen data without overfitting the training set.

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 directly influencing the count are the features with highest coefficients. These are: Temp, Year (positively influencing) and snowy and rainy weather (negatively influencing).

General subjective questions:

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

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

There are two main types of linear regression:

Simple Linear Regression

This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable. The equation for simple linear regression is:
$y=\beta0+\beta1X$ $y=\beta0+\beta1X$
where:

- Y is the dependent variable
- X is the independent variable
- β0 is the intercept
- β1 is the slope

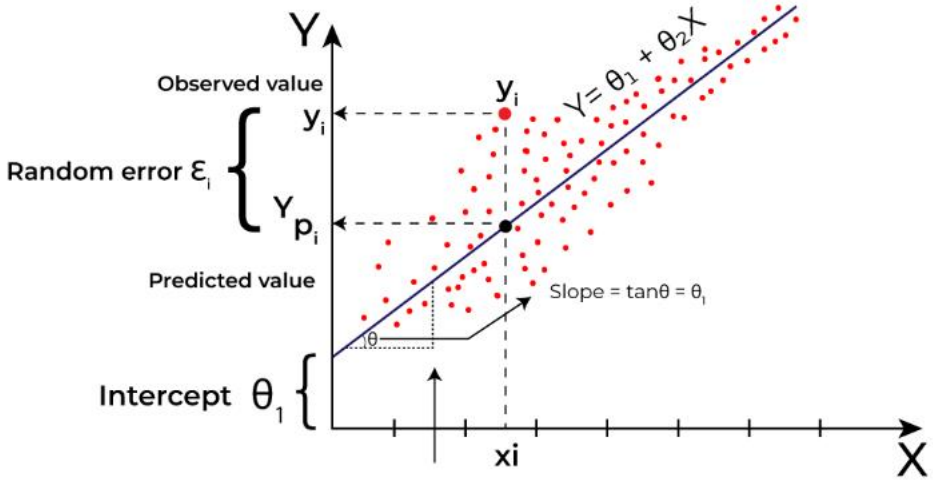
Multiple Linear Regression

This involves more than one independent variable and one dependent variable. The equation for multiple linear regression is:
$y=\beta0+\beta1X1+\beta2X2+\ldots\ldots\beta nXn$ $y=\beta0+\beta1X1+\beta2X2+\ldots\ldots\beta nXn$
where:

- Y is the dependent variable
- X1, X2, …, Xn are the independent variables
- β0 is the intercept
- β1, β2, …, βn are the slopes

The goal of the algorithm is to find the best Fit Line equation that can predict the values based on the independent variables.

In regression set of records are present with X and Y values and these values are used to learn a function so if you want to predict Y from an unknown X this learned function can be used. In regression we have to find the value of Y, So, a function is required that predicts continuous Y in the case of regression given X as independent features.



2. Explain the Anscombe's quartet in detail

Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

Anscombe's Quartet Dataset

The four datasets of **Anscombe's quartet.**

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

3. What is Pearson's R?

Pearson's correlation coefficient, also known as Pearson's R, is a measure of the strength of correlation between two variables. It is commonly used in linear regression. The value of Pearson's R always lie between -1 and +1, the latter indicating a perfectly positive and linear correlation and the former indicating a perfectly linear negative regression. The values in between denotes the relative collinearity of two variables.

Pearson Correlation Coefficient Table

Pearson Correlation Coefficient (*r*) Range	**Type of Correlation**	**Description of Relationship**	**New Illustrative Example**
$0 < r \leq 1$	Positive	An increase in one variable associates with an increase in the other.	**Study Time vs. Test Scores:** More hours spent studying tends to lead to higher test scores.
$r = 0$	None	No discernible relationship between the changes in both variables.	**Shoe Size vs. Reading Skill:** A person's shoe size doesn't predict their ability to read.
$-1 \leq r < 0$	Negative	An increase in one variable associates with a decrease in the other.	**Outdoor Temperature vs. Home Heating Cost:** As the outdoor temperature decreases, heating costs in the home increase.

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

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and 1. **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

- **sklearn.preprocessing.scale** helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.

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

The value of VIF is calculated by the below formula:

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

Where, 'i' refers to the ith variable.

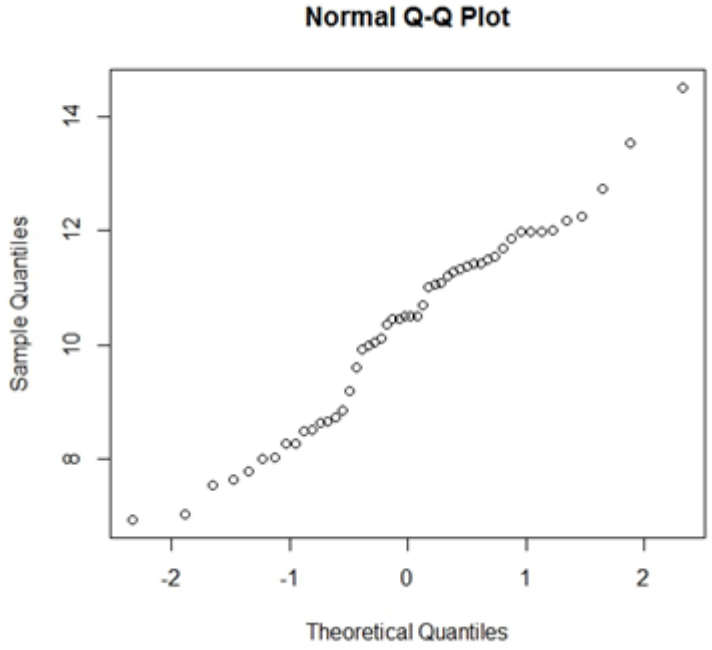
If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

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

The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly

straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.



Use of Q-Q plot in Linear Regression: The Q-Q plot is used to see if the points lie approximately on the line. If they don't, it means, our residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.

Importance of Q-Q plot: Below are the points:

I. The sample sizes do not need to be equal.

II. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.

III. The q-q plot can provide more insight into the nature of the difference than analytical methods.