Madhuri Gaikwad

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 conducted a box plot analysis to visualize the relationship between categorical variables and the target variable. The findings indicate the following effects on the target variable:

- Among the seasons, fall (Season 3) exhibits the highest demand for rental bikes.
- There is a noticeable growth in demand for the upcoming year.
- Demand shows a consistent increase each month until June, with September having the highest demand. Subsequently, demand starts to decrease.
- Demand decreases during holidays.
- Weekdays do not provide a clear picture of demand.
- The highest demand is observed during clear weather conditions.

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

Answer:

1. The parameter `drop_first=True` is important to use when creating dummy variables. Its purpose is to eliminate the extra column generated during the dummy variable creation process.

2. By using `drop_first=True`, the correlations between the dummy variables are reduced. This is beneficial because highly correlated dummy variables can introduce multicollinearity issues in the data.

3. If we do not drop one of the dummy variables created from a categorical variable, it becomes redundant in the dataset. This redundancy is due to the presence of a constant variable (intercept) in the model, which can lead to multicollinearity problems.

To summarize, using `drop_first=True` is important in handling categorical variables to avoid multicollinearity and ensure the model's effectiveness.

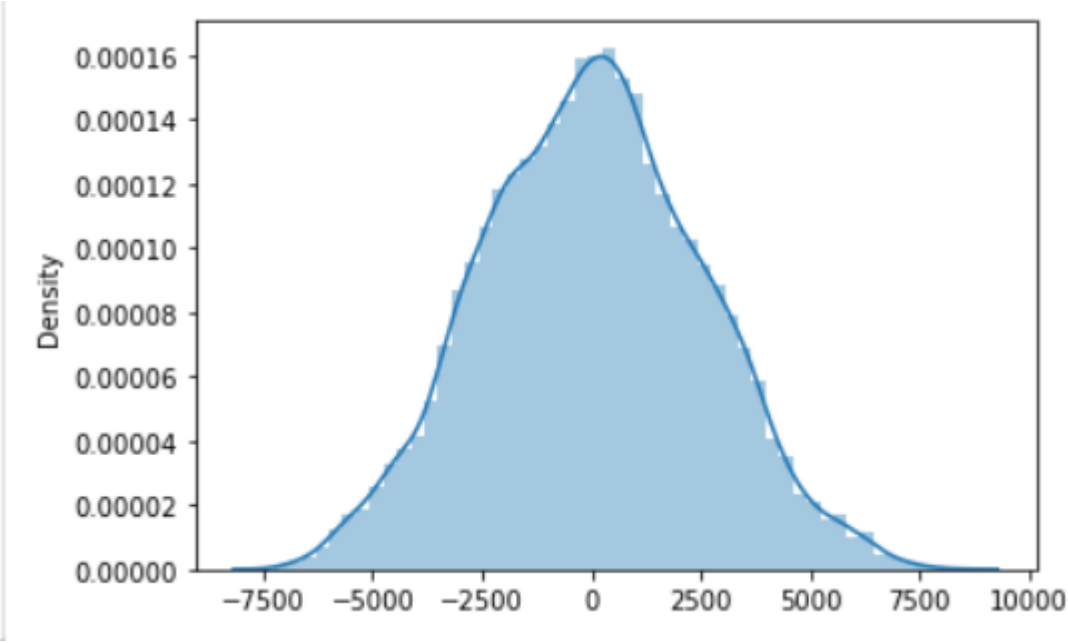
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer: By looking at the pair-plot among numerical variables, **temp and atemp** are the two variables which are highly correlated.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer:

- Checked for VIF values for the independent variables to ensure minimal collinearity.
- The residuals follow normal or close to normal distribution for y_pred and y_test data.



- The residual errors are homoscedastic as in the variance does not increase towards either side of the distribution substantially.

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:
According to my prediction of the model following are the top 3 features significantly contributing towards the demand of the shared bikes:

- **Temp**
- **Weathersit (Good&Clear)**
- **Year**

General Subjective Questions:

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

Answer:

Linear regression is a statistical regression method used for predictive analysis and observe the relationship between the continuous variables. It shows the relationship between the **independent variable** (X) also known as predictor variable and the **dependent variable** (Y-axis) also called target variable.

Depending on the number of input variables or predictor **variables**, there are two types:

Simple linear regression: When the number of independent variables is 1

Multiple linear regression: When the number of independent variables is more than 1

Linear regression estimates the relationship between a dependent variable and an independent variable using the **line equation**:

y = mx+c where **Y** is the **target variable**, **x** is the **input variable**, **m** and **c** are **slope** and **y intercept** respectively

The above equation could be also written as **y = b1x+b0**, which implies that for every rise/fall in the value of x, the y is rising or falling b1 times provided b0 is kept constant.

A regression line can be a **Positive** Linear Relationship or a **Negative** Linear Relationship based on the coefficient values.

The main aim of the linear regression algorithm is to get the best values for b0 and b1 to find the best fit line for the model. The best fit line should have the least error means the error between predicted values and actual values should be minimized.it can be done using Cost function.

Cost function adjusts the regression coefficients and measures how a linear regression model is acting. The cost function is used to find the correctness of the mapping function that maps the predictor variable to the target variable. This mapping function is also known as the Hypothesis function.

Following are the steps to be followed in linear regression algorithm:

1. Reading and understanding the data: Importing required libraries like pandas & numpy for data analysis and manipulation and seaborn & matplotlib for data visualization

2. Performing EDA on the data: Visualization of data: Visualizing numerical variables using scatter or pair plots and for categorical variables using bar plots or boxplots in order to interpret the inferences.

3. Data preparation: modification: Converting categorical variables with varying degrees of levels into dummy variables which are numerical in nature

4. Splitting the data into training and test sets either 70-30 or 80-20 : Splitting the data into 2 sections in order to train a subset of given dataset to generate a trained model that will very well generalize how test data will be evaluated

5. Build a linear model with help of library functions: We add all the variables at once and then eliminate variables based on high multicollinearity (VIF>5) or insignificance (high p- values).

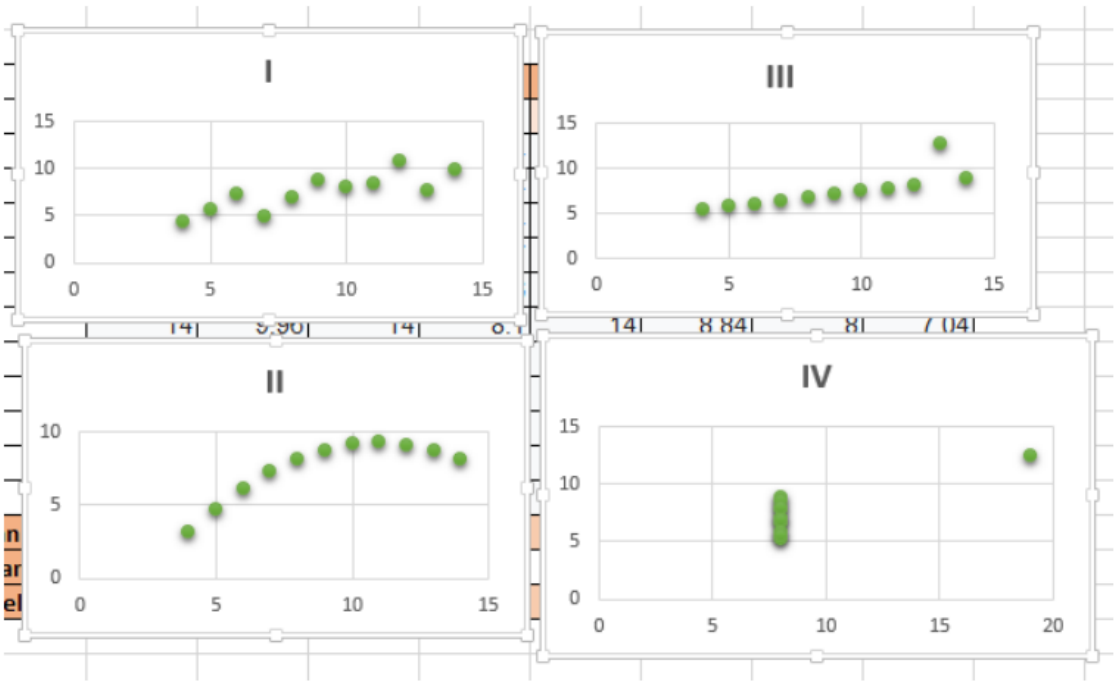
6. Residual analysis of the train data, error terms are analysed: It tells us how much the errors (y_actual — y_pred) are distributed across the model. A good residual analysis will signify that the mean is centred on 0.

7. Making predictions using the final model and evaluation: We will predict the test dataset by transforming it onto the trained dataset

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

Answer: Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. Each dataset consists of eleven (x,y) points.

	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
Mean	9	7.500909	9	7.500909	9	7.5	9	7.500909
Variance	11	4.127269	11	4.127629	11	4.1226 2	11	4.123249
Correlation	0.816421		0.816237		0.816287		0.816521	



The four datasets can be described as:

• **Dataset 1:** this fits the linear regression model pretty well.

• **Dataset** 2: this could not fit linear regression model on the data quite well as the data is non-linear.

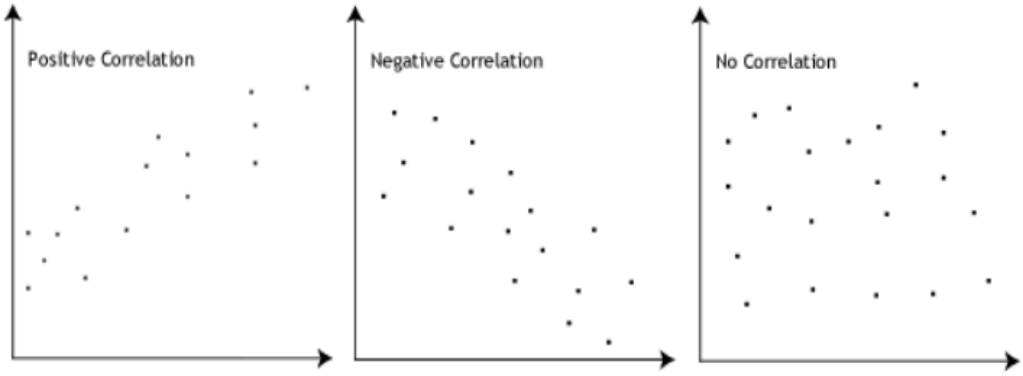
• **Dataset 3:** shows the outliers involved in the dataset which cannot be handled by linear regression model

• **Dataset 4:** shows the outliers involved in the dataset which cannot be handled by linear regression model

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.
- Pearson's r measures the strength of the linear relationship between two variables. Pearson's r always between -1 and 1.
- If data lie on a perfect straight line with negative slope, then r = -1



Positive correlation indicates the both the variable increase and decrease together. Negative correlation indicates the one the variable increase and the other variable decrease and vice versa.

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

Answer:

- **Scaling** is a step of data processing in data analyse using models, it is applied to independent variables to normalize the data within a particular range.
- It also helps in faster calculations in an algorithm.
- The data collected often contains features/variables which are highly varying in magnitudes or units or range.
- If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling.
- To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
- It is important to note that scaling just affects the coefficients and none of the other parameters like **t-statistic**, **F-statistic**, **p-values**, **R-squared**.

❖ The **difference** between **Normalized scaling** and **Standardize scaling** is as follows:

Normalized scaling	Standardize scaling
It brings all of the data in the range of 0 and 1.	It brings all of the data into a standard normal distribution which has mean zero and standard deviation one
Uses **MinMaxScaler** from sklearn sklearn.preprocessing.MinMaxScaler	Uses **scale** from sklearn sklearn.preprocessing.scale
$x = \frac{x - \min(x)}{\max(x) - \min(x)}$	$x = \frac{x - \text{mean}(x)}{sd(x)}$
it loses some information in the data, if there are outliers in the dataset	it retains the information in the data, if there are outliers in the dataset

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

Answer:

- VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity.
- A large value of VIF indicates that there is a correlation between the variables.
- If the VIF is infinite which implies that there is a perfect correlation between two predictor variables.
- It is the case of perfect correlation. In this case we get Rsquared value equal to 1.
- Due to which the term 1/ (1-Rsquared) reached infinity.
- For this we need to identify the feature which is causing this perfect correlation and should be dropped to get a best model
- An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables

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

Answer:

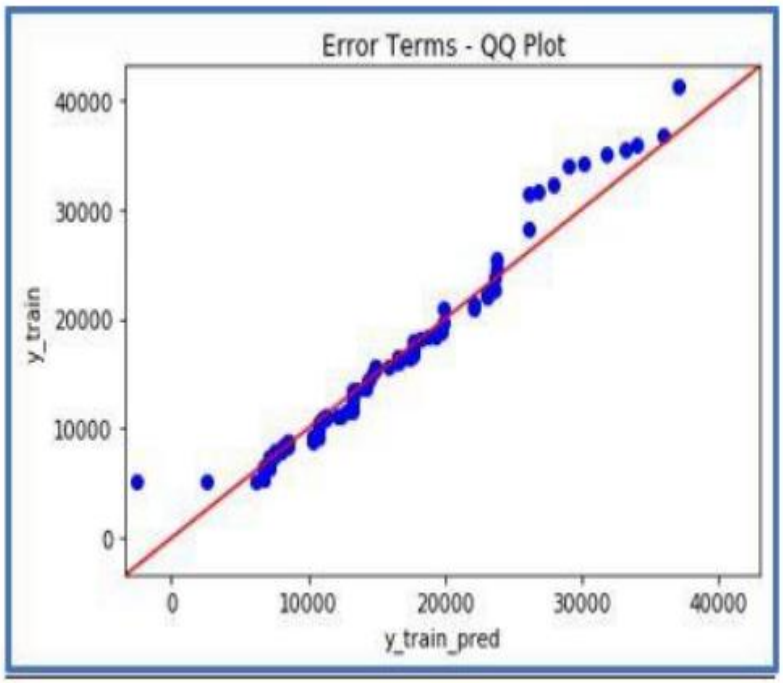
- **Quantile-Quantile (Q-Q) plot**, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.
- Also, it helps to determine if two data sets come from populations with a common distribution
- It is used for determining if two data sets come from populations with a common distribution.
- A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
- Whether the Distributions is Gaussian, Uniform, Exponential or even Pareto distribution, it can be found out.

Few advantages:
a) It can be used with sample sizes also
b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:
If two data sets —

i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behaviour

Interpretation:
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
Below are the possible interpretations for two data sets.
a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degrees from x -axis
b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.
c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degrees from x -axis