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:

   The Bar plot was used to analyze categorical features. Here are the conclusions drawn.

   - During the autumn, which normally starts in September, most men and women who have rented motorcycles have done so. Therefore, people prefer to travel when the weather is clear.
   - The majority of people have leased bikes during the workday. They probably prefer riding a motorbike to work.
   - The busiest days for bike rentals were Friday and Saturday.
   - The number of bike rentals has significantly increased in 2019. Probably as a result of the beneficial effects it had on individuals the year before, which is a positive trend.
   - Myriad of people rented Bike during National Holiday.

2. **Why is it important to use drop_first=True during dummy variable creation?**
   Answer:
   drop_first= true is crucial since it minimises the extra column that is formed when a dummy variable is created.
   As a result, it lessens the connections that dummy variables cause.
   For instance, we want to build a dummy variable for a categorical column that has three different types of data. One variable is obviously unfurnished if it is neither furnished nor semi-furnished. Therefore, there is no need for a third variable to identify unfurnished.

Furnishing status	Furnished	Semi-Furnished
Furnished	1	0
Semi-Furnished	0	1
Un-Furnished	0	0

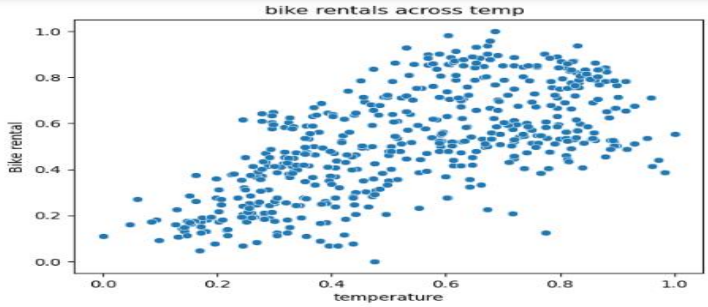
   Therefore, if we have a categorical variable with n levels, we must use n-1 columns to represent the dummy variables.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   Answer:
   Temp has the highest correlation with the target variable (i.e 0.63)

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   Answer:
   Assumptions of Linear regression was validated based on below 4 assumptions:

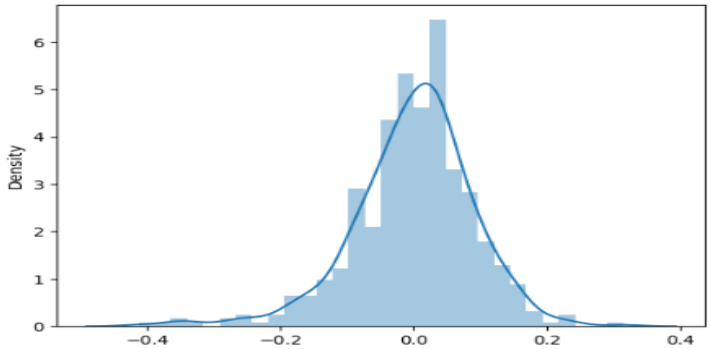
   - Linear relationship between X and Y

linear relationship should be present between feature and the target. The scatter plot is used to visualise linear relationships.



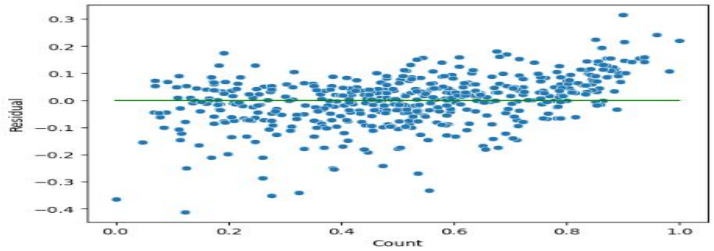
Based on the above analysis, sales of rental bike increase as the temperature rises.

- Error terms are normally distributed.
  Distplot is used to check the uniform distribution error term



Error term is normally distributed, there is no skewness seen.

- Error terms have constant variance (homoscedasticity)
  Error term is same across all the independent variables. A scatter plot of residual values vs predicted values is used to ascertain the homoscedasticity.



the graph shows no specific pattern. there exists common variance among the residuals,

- Error terms are independent of each other
  Error terms obtained were not correlated with each other.

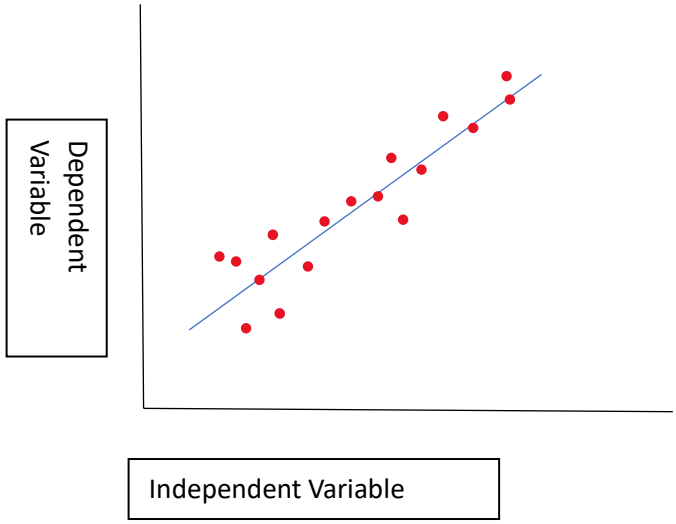
5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   Answer:

Based on the model obtained we can conclude that below three features are the most influential features for Bike Rentals:
- Temperature: with coefficient 0.5499
- Year: with coefficient 0.2331
- Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds: with negative coefficient -0.2880

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**
   Answer:
   The linear regression examines the linear relation between a dependent variable and a set of independent variables. There are two types of Linear Regression
   - Simple Linear Regression:
     It has only one independent feature
   - Multiple Linear Regression:
     It has more than one independent feature

   According to the linear relationship between variables, the value of the dependent variable will vary proportionally (increase or decrease) when the value of one or more independent variables changes.



The dependent variable's linear connection with the independent variables is seen in the graph above. When the value of x, an independent variable, rises, so does the value of y, a dependent variable. The best fit straight line is denoted by the red line.

Mathematically the relationship can be represented with the help of following equation –

$Y = mX + c$

Here, Y is the dependent variable we are trying to predict.

X is the independent variable we are using to make predictions.

m is the slope of the regression line which represents the effect X has on Y.

c is a constant, known as the Y-intercept.

If X = 0, Y would be equal to c.

Additionally, the linear relationship might be either positive or negative in nature.

- Positive Linear Relationship: If both the independent and dependent variables rise, the linear relationship is said to be positive.
- Negative Linear relationship: If the independent variable increases while the dependent variable decreases, then the linear relationship is said to be negative.

Assumption of Linear Regression are:

- Linear relationship between X and Y:
  There should be a linear relationship between feature and target . To show linear relationships, use a scatter plot.
- Error terms are normally distributed:
  The errors in the model are normally distributed. Distplot is used to check the uniform distribution error term
- Error terms have constant variance (homoscedasticity):
  Error term is same across all the independent variables. Model should not exhibit any visible pattern. A scatter plot of residual values vs predicted values is used to ascertain the homoscedasticity.
- Error terms are independent of each other:
  The dataset's observations are not dependent on one another. This means that the value of the dependent variable for one observation does not depend on the value of the dependent variable for another observation.

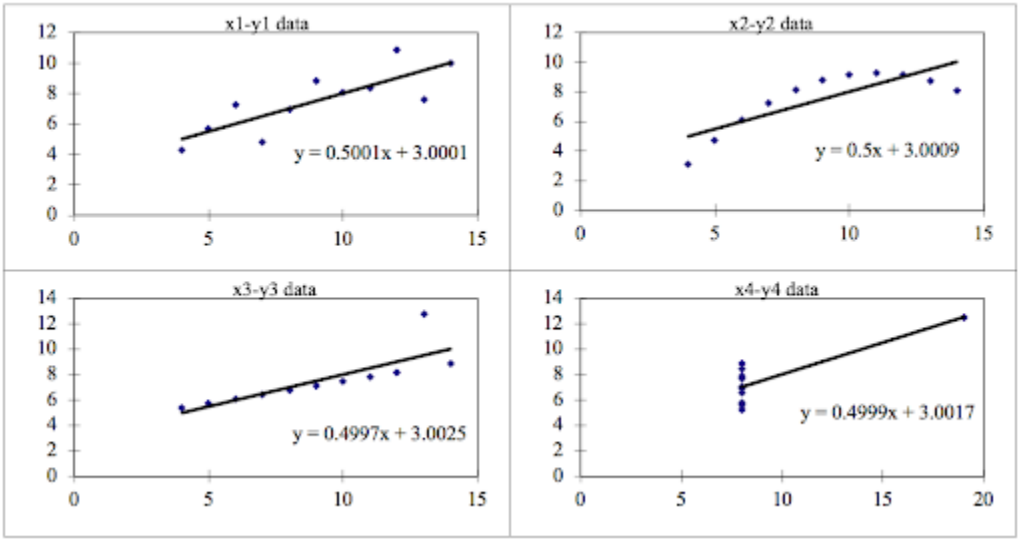
2. Explain the Anscombe's quartet in detail. (3 marks)

Answer:

Anscombe's quartet is a collection of four datasets that, when plotted as a scatter plot on a graph, have different representations despite having similar descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines. The datasets were developed by statistician Francis Anscombe in 1973 . It emphasises the value of data visualisation before applying various algorithms to create models. In order to identify the numerous anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.), it is suggested that the data characteristics be plotted. Additionally, because it cannot handle any other type of data collection, linear regression may only be regarded as a fit for data that have linear connections.

Anscombe's Data									
Observation	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89
			Summary Statistics					
N	11	11	11	11	11	11	11	11
mean	9.00	7.50	9.00	7.500909	9.00	7.50	9.00	7.50
SD	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94
r	0.82		0.82		0.82		0.82	

According to the summary statistics, x and y's means and variances were the same for all groups for both x and y:
For each dataset, the means of x and y are 9, respectively.
• For each dataset, the variance of x is 11 and the variance of y is 4.13.
• For each dataset, the correlation coefficient (a measure of the strength of a relationship between two variables) between x and y is 0.816.
These four datasets display the identical regression lines when we plot them on an x/y coordinate plane, but each tells a different narrative.



• The linear models in Dataset I seem to be clear and well-fitting.
• Dataset II is not uniformly distributed.
• Although Dataset III's distribution is linear, an outlier causes the estimated regression to be incorrect.
• Dataset IV demonstrates that a high correlation coefficient can be obtained with just one outlier.

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

Answer:

The Pearson correlation coefficient is used to measure the strength and direction of a linear association between two variables. The value ranges between -1 to 1, where the value r = 1 means a perfect positive correlation and the value r = -1 means a perfect negative correlation.

- Positive Correlation(Between 0 and 1):

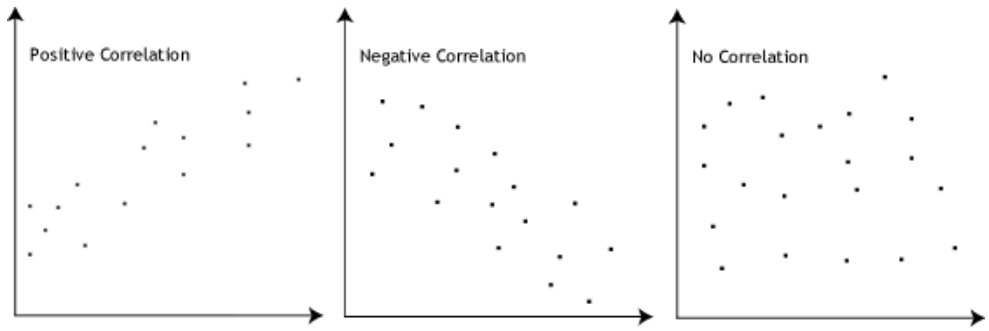
Both variable change in same direction. For instance, increase in salary as the age increases.

- Negative Correlation(Between 0 and -1):

Increase in one variable cause decrease in another variable. For instance, increase in intake of junk food causes decline in one's health

- No correlation(0):

Variables are not related to each other. For example, sales price of the car does not have any relation to the paint applied in the car.



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

Answer:

Scaling is done to normalize the features within a fixed range.It is usually done during the Data preparation stage.

If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give

wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

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:

VIF = infinity if there is perfect correlation. A high VIF score denotes a strong connection between the variables. The presence of multicollinearity causes the variance of the model coefficient to be exaggerated by a factor of 4 if the VIF is 4.

VIF displays a complete correlation between two independent variables when its value is infinite. If the correlation is perfect, we have R-squared (R2) = 1, which results in 1/ (1-R2) infinite. To fix this, we must remove the variable from the dataset that is the exact multicollinearity's cause.

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

Answer:

A graphical method for assessing if two data sets originate from populations with a common distribution is the quantile-quantile (q-q) plot.

Q-Q plot application: A q-q plot compares the quantiles of the first data set to those of the second data set. A quantile is the percentage of points that fall below the specified number. In other words, the 0.3 (or 30%) quantile is the value at which 30% of the data are below it and 70% are above it. Additionally, a 45-degree reference line is plotted. The points should roughly lie along this reference line if the two sets are drawn from a population with the same distibution

Importance of the Q-Q plot: It is frequently desirable to determine whether the presumption of a common distribution is supported when there are two data samples. If so, location and scale

estimators can combine the two sets of data to derive estimates for the shared position and scale. If two samples do differ, it is also helpful to comprehend the variations. More information about the nature of the difference can be gleaned from the q-q plot than from analytical techniques like the chi-square and Kolmogorov-Smirnov 2-sample tests.