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: -

Inferences after analysing the categorical variables with the help of boxplot are: -

1. Fall season has the highest cnt i.e. highest demand for bikes followed by summer season.

2. The demand for bike increases till September and September month has the highest number of rental bikes. After September the demand decreases.

3. The no. of rental bikes remains approximately the same throughout the week with slight peak on Friday.

4. The highest demand for the rental bikes was during the clear weather followed by cloudy weather.

5. The demand for rental bikes varies on holiday.

6. The demand for rental bikes in 2019 is much higher as compared to year 2018.

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

Answer: - drop_first allows us whether to keep or remove the reference (whether to keep k or k-1 dummies out of k categorical levels). drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

Answer: - 'temp' and 'atemp' has the highest correlation with the target variable.

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

Answer: -

Assumptions are as follows: -

1. Linear relationship validated by plotting a scatterplot and observing the plot, linearity visible.

2. Small or no multicollinearity validated with the help of VIF (variance inflation factor), the value of VIF for variables is under limit.
3. Homoscedasticity validated with the help of regplot(). No pattern in residual values.
4. Normal distribution of error terms validated by plotting a distribution plot, error terms are normally distributed.
5. No autocorrelation of errors validated with the help of Durbin Watson test, O/P value is 2.

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

Answer: - The top 3 features are

1. Temp (temperature)
2. Month_sept(September)
3. Season_winter

General Subjective Questions

1. Explain the linear regression algorithm in detail.

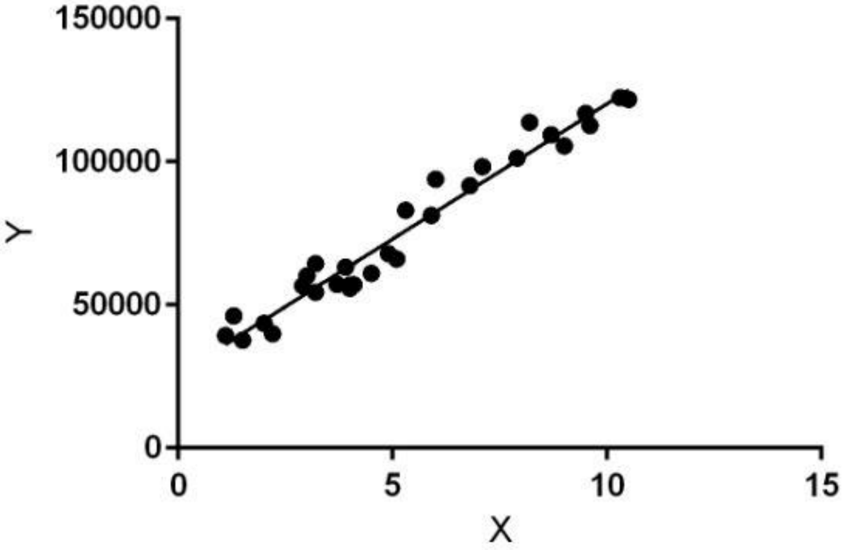
Answer: - It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.

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.

Types of Linear Regression: -

1. Simple Linear Regression
2. Multiple Linear Regression

The linear regression model provides a sloped straight line representing the relationship between the variables.



The mathematical equation can be given as:

Y = β0 + β1*x

Where

1. Y is the response or the target variable
2. x is the independent feature
3. β1 is the coefficient of x
4. β0 is the intercept

Assumptions are as follows: -

1. Linear relationship between the features and target: Linear regression assumes the linear relationship between the dependent and independent variables.
2. Small or no multicollinearity between the features: Multicollinearity means high-correlation between the independent variables. Due to multicollinearity, it may difficult to find the true relationship between the predictors and target variables.
3. Homoscedasticity Assumption: Homoscedasticity is a situation when the error term is the same for all the values of independent variables. With homoscedasticity, there should be no clear pattern distribution of data in the scatter plot.
4. Normal distribution of error terms: Linear regression assumes that the error term should follow the normal distribution pattern. If error terms are not normally distributed, then confidence intervals will become either too wide or too narrow, which may cause difficulties

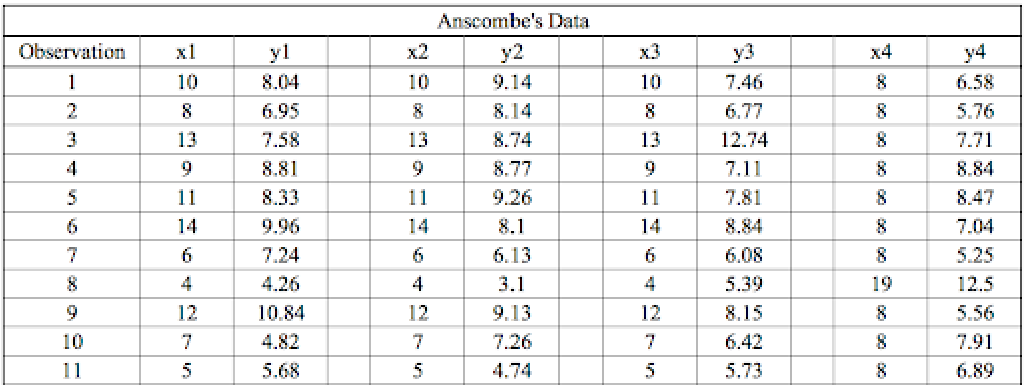
in finding coefficients. It can be checked using the q-q plot. If the plot shows a straight line without any deviation, which means the error is normally distributed.

5. No autocorrelations: The linear regression model assumes no autocorrelation in error terms. If there will be any correlation in the error term, then it will drastically reduce the accuracy of the model. Autocorrelation usually occurs if there is a dependency between residual errors.

2. Explain the Anscombe’s quartet in detail.

Answer: - Anscombe’s quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyse it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another.

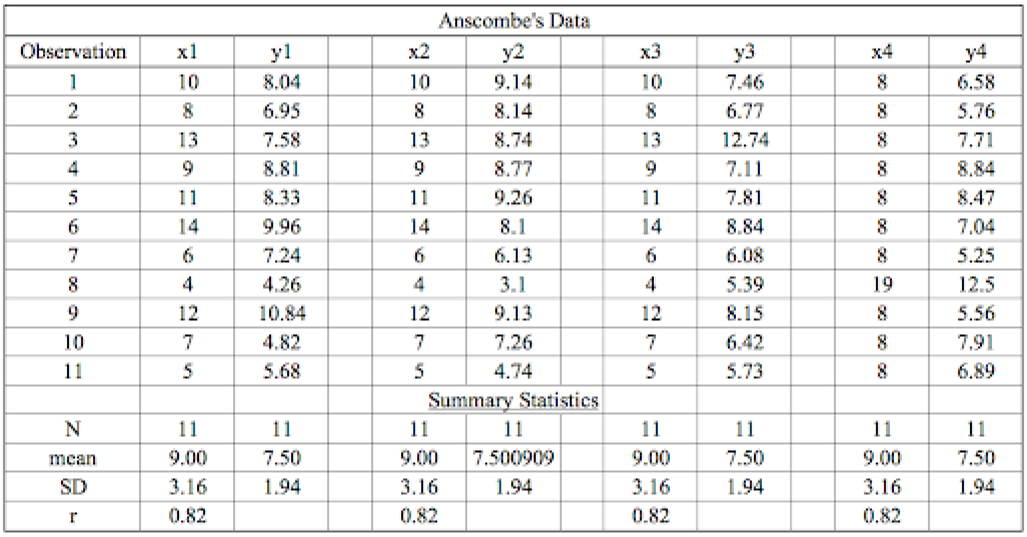
Anscombe’s quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.)



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

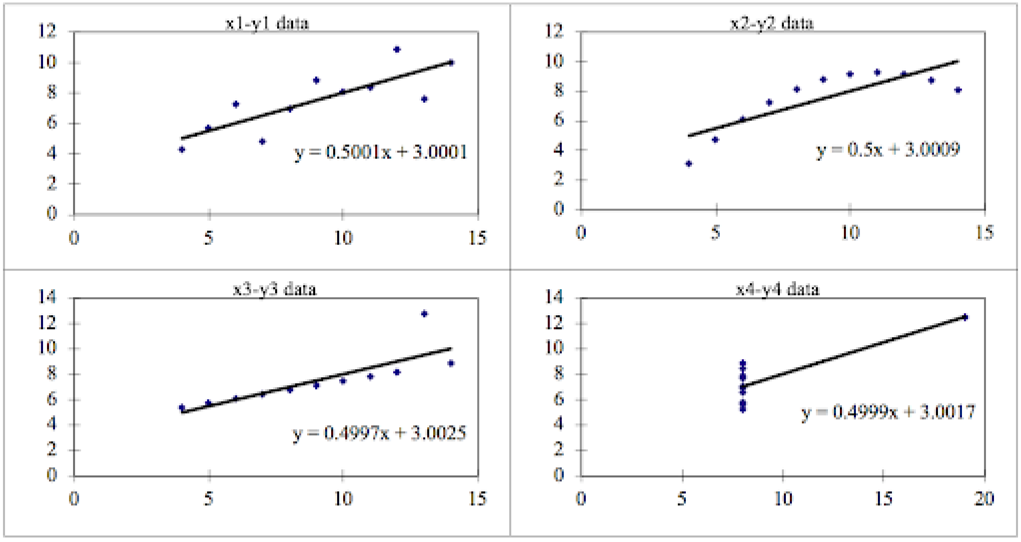
The statistical information for these four data sets is approximately similar



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	

when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm:



ANSCOMBE'S QUARTET FOUR DATASETS

Data Set 1: fits the linear regression model pretty well.

Data Set 2: cannot fit the linear regression model because the data is non-linear.

Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.

Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

3. What is Pearson's R?

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. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The correlation coefficient of -1 means a robust negative relationship. Therefore, it imposes a perfect negative relationship between the variables. If the correlation coefficient is 0, it displays no relationship. If the correlation coefficient is 1, it means a strong positive relationship. Therefore, it implies a perfect positive relationship between the variables.

The Pearson correlation coefficient shows the relationship between the two variables calculated on the same interval or ratio scale. In addition, it estimates the relationship strength between the two continuous variables.

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

Answer: -

i) scaling 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.

ii) Most of the times, collected data set contains features highly varying in magnitudes, units and 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.

Normalization is also known as 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. Formula for Min-Max scaling is {x – min(x)} / {max(x) – min(x)}. It is affected by outliers , used when features are of different scale.

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). sklearn.preprocessing.scale helps to implement standardization in python. Formula for Standardization Scaling is {x – mean(x)} / std(x). It is less affected by outliers , used when we want to ensure 0 mean and 1 standard deviation.

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

Answer: - VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. We know the formula for VIF is 1/1-R2.

High VIF means high correlation between independent variables. For VIF to be infinite the value of R2(r-squared) must be 1 which means perfect multicollinearity. To avoid this, we need to drop the variable.

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

Answer: - The Q-Q plot is a graphical plotting of the quantiles of two distributions with respect to each other. In other words, we can say plot quantiles against quantiles. Whenever we are interpreting a Q-Q plot, we shall concentrate on the 'y = x' line. We also call it the 45-degree line in statistics. It entails that each of our distributions has the same quantiles. In case if we witness a deviation from this line, one of the distributions could be skewed when compared to the other.

Importance of Q-Q plot: -

With the help of Q-Q plot many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected. In addition to this Q-Q plot can be used in scenarios If two data sets, come from populations with a common distribution, have common location and scale, have similar distributional shapes and have similar tail behaviour.