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?

Ans: Analysis of categorical variable done through box plot. Below the point of infer from the visualization are follows

- Highest demand for bike during fall.
- Count increased by next year 2019 hence demand will continuously grow in next year.
- Demand is growing till Jun and gradually decreasing upcoming months.
- Demand is higher for Thursday, Friday, Saturday and Sunday.
- Demand comparatively lesser than start of the week.
- Clear weather sit has highest demand.
- Due to weather conditions demand is getting decreased.
- Booking seems same as weekday and weekend.

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

Ans: 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.

Syntax: By default, drop_first=False which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

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

Ans: Temp has 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?

Ans:

- The Two Variables Should be in a Linear Relationship- The first assumption of simple linear regression is that the two variables in question should have a linear relationship.
- All the Variables Should be Multivariate Normal Distribution-the linear combination of the random variables should have a normal distribution.
- There Should be No Multicollinearity in the Data-multiple linear regression is that there should not be much multicollinearity in the data. Such a situation can arise when the independent variables are too highly correlated with each other.
- There Should be No Autocorrelation in the Data-One of the critical assumptions of multiple linear regression is that there should be no autocorrelation in the data. When the residuals are dependent on each other, there is autocorrelation.
- There Should be Homoscedasticity Among the Data-The data is said to homoscedastic when the residuals are equal across the line of regression. In other words, the variance is equal.

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

Ans: Below are the 3 features contributing significantly
Temp, Season winter, Month September

General Subjective Questions

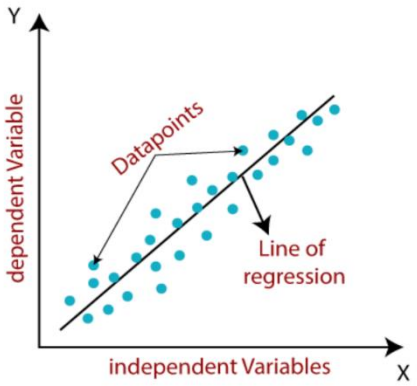
1.Explain the linear regression algorithm in detail.

Ans. 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.

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:

$y= a_0+a_1x+ \varepsilon$



Y - Dependent Variable (Target Variable)
X - Independent Variable (predictor Variable)
a0 - Intercept of the line (Gives an additional degree of freedom)
a1 - Linear regression coefficient (scale factor to each input value).
ε - random error

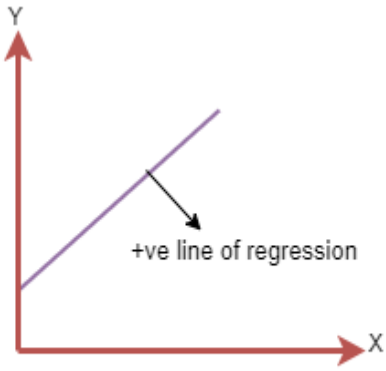
The values for x and y variables are training datasets for Linear Regression model representation.

Linear regression can be further divided into two types of the algorithm:

- **Simple Linear Regression:**
  If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.
- **Multiple Linear regression:**
  If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Linear Regression Line

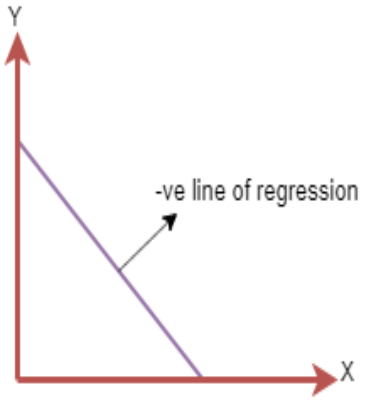
A linear line showing the relationship between the dependent and independent variables is called a **regression line**. A regression line can show two types of relationship:



The line equation will be: $Y= a_0+a_1x$

- **Positive Linear Relationship:**

If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.



The line of equation will be: $Y= -a_0+a_1x$

- **Negative Linear Relationship:**

If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a Negative linear relationship.

2.Explain the Anscombe's quartet in detail.

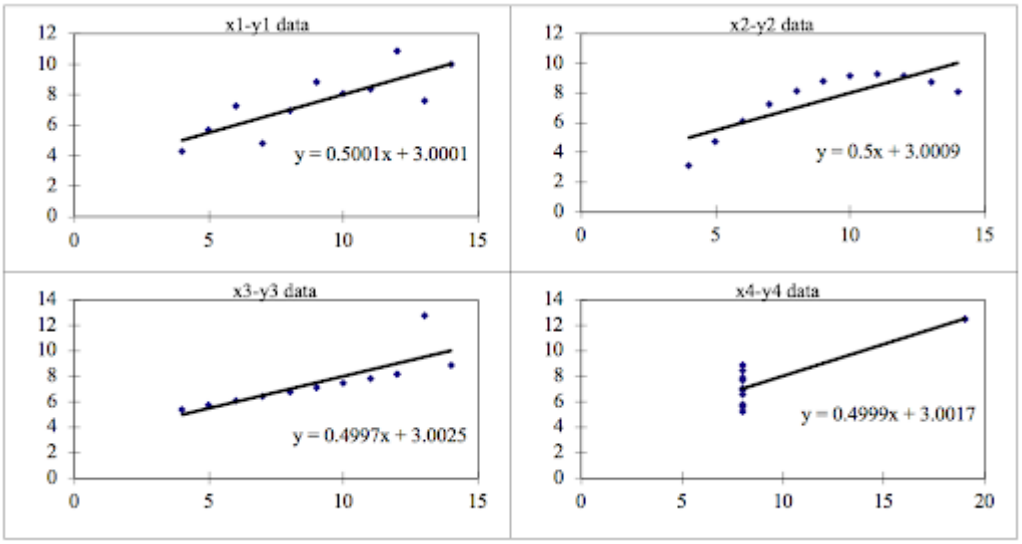
Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x, y) points.

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.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

The statistical information for these four dataset are approximately similar. We can compute them as follows:

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	

However, 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, as you can see below:



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.

As you can see, Anscombe's quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine

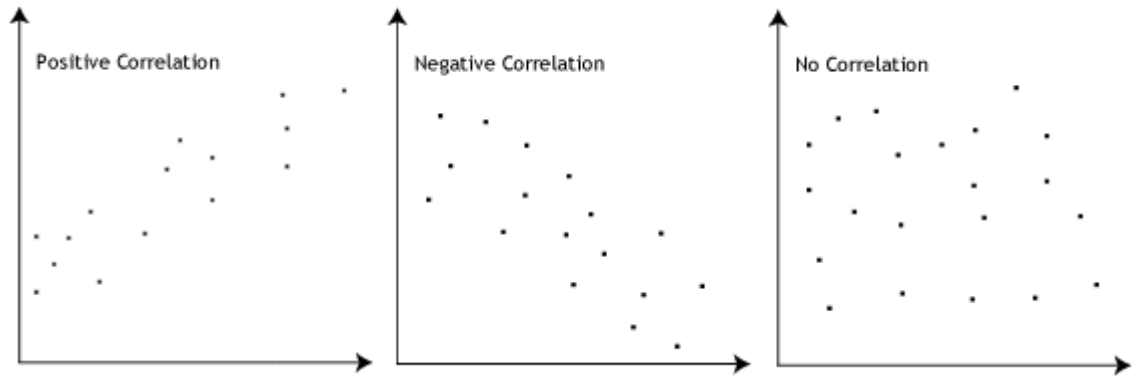
learning algorithm, we first need to visualize the data set in order to help build a well-fit model

3. What is Pearson's R?

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations. it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association



Pearson's Correlation Coefficient formula is as follows,

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Where,

- r = Pearson Coefficient

- n= number of the pairs of the stock
- $\sum xy$ = sum of products of the paired stocks
- $\sum x$ = sum of the x scores
- $\sum y$= sum of the y scores
- $\sum x^2$ = sum of the squared x scores
- $\sum y^2$ = sum of the squared y scores

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

Ans: 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.

Why is scaling performed:

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. Scaling only affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Difference between Normalization and Standardization

S.NO.	Normalization	Standardization
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.

S.NO.	Normalization	Standardization
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.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

When there is a perfect relationship then VIF=infinity where as if all independent variables are orthogonal than to each other then VIF=1.0. Means if a variable is expressed exactly by a linear combination of other variable then it said that VIF is infinite.

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

Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution and quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.

How Q-Q plots can help us identify the distribution types?

The power of Q-Q plots lies in their ability to summarize any distribution visually.

QQ plots is very useful to determine

- If two populations are of the same distribution
- If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this.
- Skewness of distribution

In Q-Q plots, we plot the theoretical Quantile values with the sample Quantile values. Quantiles are obtained by sorting the data. It determines how many values in a distribution are above or below a certain limit. If the datasets we are comparing are of the same type of distribution type, we would get a roughly straight line. Here is an example of normal distribution.