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?

The following are the points we can infer from the effect on dependent variable using Box plot

1) The demand is high in the summer and Fall season

2) From the month of feb to oct there is higher demand after oct the demand falls back

3) the bike shares increased in 2019 where the maixmum share of 2018 is the median of 2019

4) High sales in clear and cloudy whereas there is no sale in Heavy rain weather

5) Sales is higher in the Clear and cloudy weather situation

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

drop_first=True is mainly used to reduce the extra column created during dummy variable creation.
Which in terms helps to reduce the correlation created among dummy variables

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

After looking at the pair plot we can see that temp variable is highly correlated with the target variable

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

The training set is validated using the following assumptions

1) Normality of Error Terms : Error terms should be normally distributed
2) Multicollinearity : Insignificant multicollinearity among the variables

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

The following are the top 3 features contributing significantly towards explaining the demand of the shared bikes

1. Year
2. Atemp
3. Winter

General Subjective Questions:

1) Explain the linear regression algorithm in detail?

Linear regression is a statistical method that is used for prediction based on the relationship between the continuous variables. In simple words, we can say that linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression

The linear regression model depicts the relationship between the variables as a sloped straight line.
When the value of x (independent variable) increases, the value of y (dependent variable) is likewise increasing. In linear regression what we do is find a best fit straight line.

Mathematically we represent a linear regression as,

$y = a + bx$, for simple linear regression

$y = a+ b1x1 + b2x2 + b3x3 + ...$ for multiple linear regression

where,

a = intercept of the line or bias

$b, b1, b2,...$ = linear regression factor or scale factor or weights

$x, x1, x2, ...$ = independent variables

y = dependent variable

A linear line showing the relationship between the dependent and independent variables is called a **regression line**. On the basis of the relationship between the independent and dependent variables, the regression line can be of two types.

Negative Linear Relationship:

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

$$\mathbf{y = -a + bx}$$

Positive Linear Relationship:

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

$$\mathbf{y = a + bx}$$

There are two type of linear regression

Simple Linear regression :

Simple linear regression is a regression model that estimates the relationship between one independent variable and one dependent variable using a straight line.

Multiple Linear regression

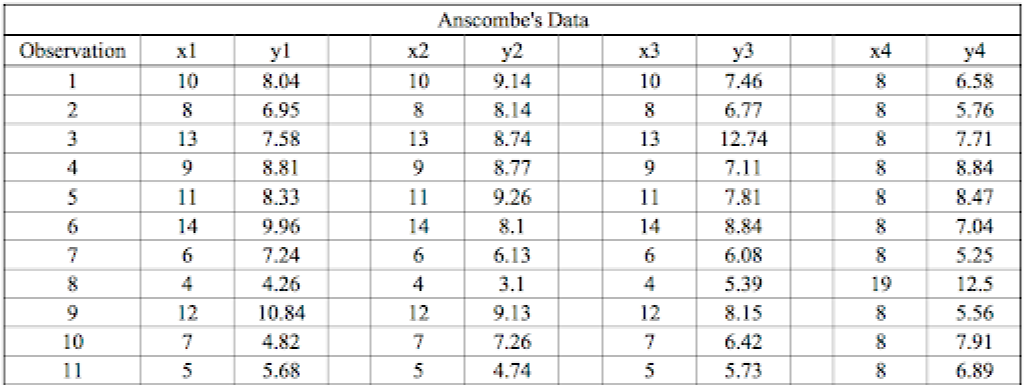
Multiple linear regression is a statistical technique that uses several explanatory variables to predict the outcome of a response variable

2) Explain the Anscombe’s quartet in detail.

Anscombe’s quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set. As you can see, the data sets have very different distributions so they look completely different from one another when you visualize the data on scatter plots

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.

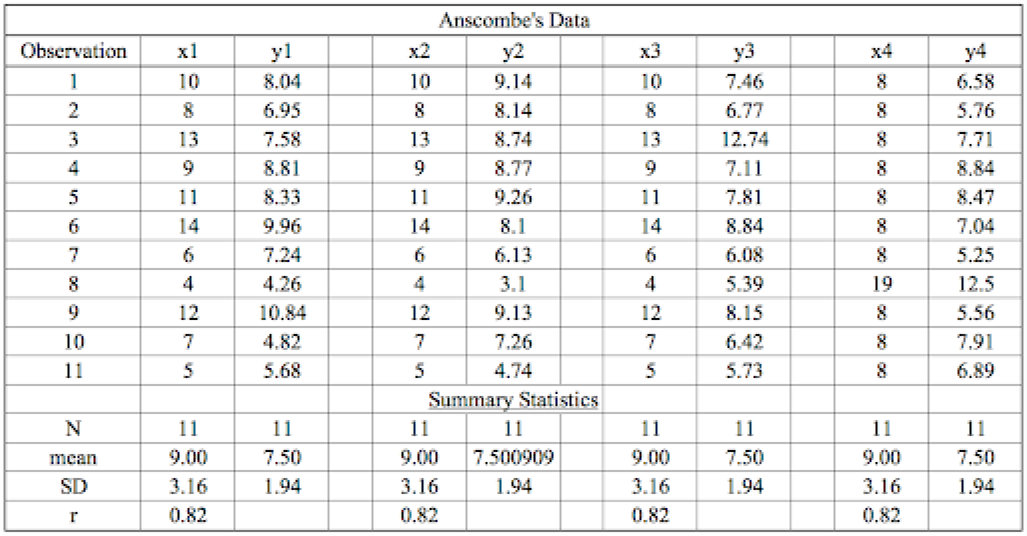
We can define these four plots as follows:



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

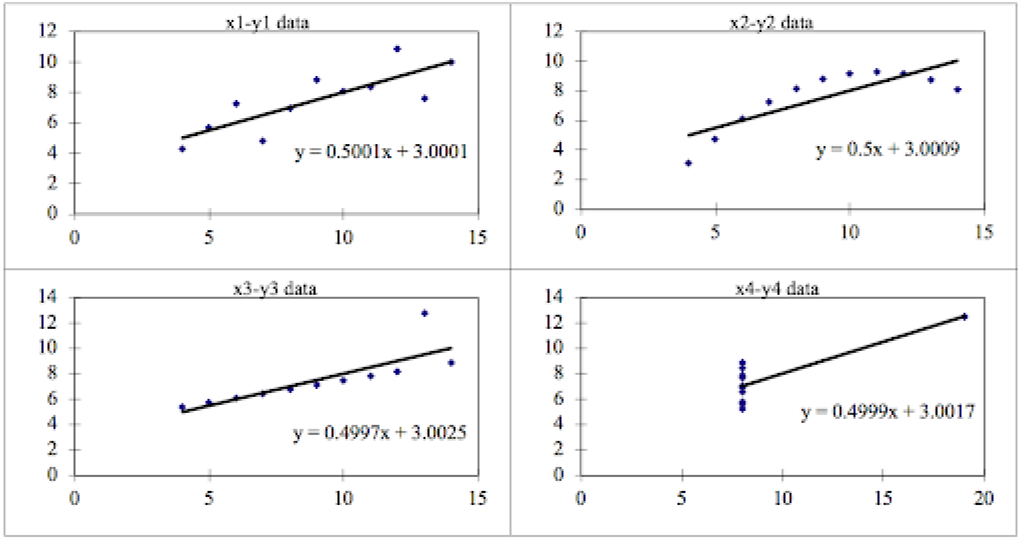
The statistical information for these four data sets are approximately similar. We can compute them as follows:



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:



We can describe the four data sets as:

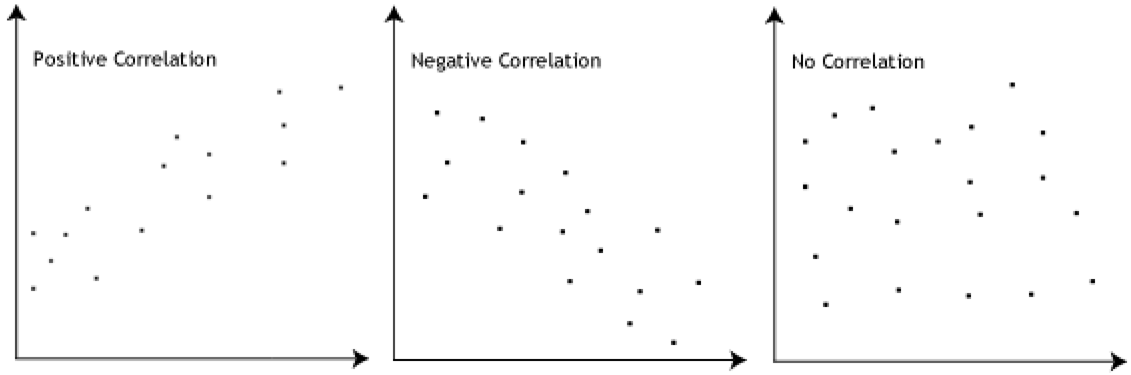
- Data Set 1: fits the linear regression model pretty well
- Data Set 2: cannot fit the linear regression model because the data is non-linear
- Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model
- Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model

3) What is Pearson’s R?

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; thus 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



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

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.

Why 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

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

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**).

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

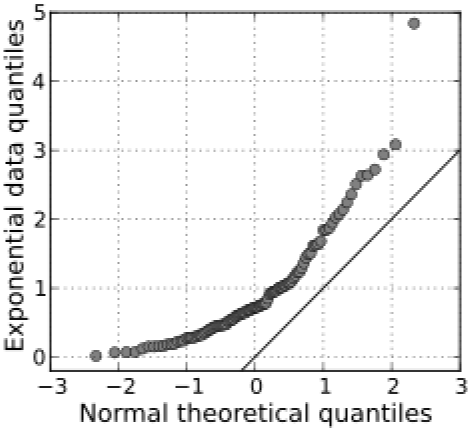
If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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.

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.