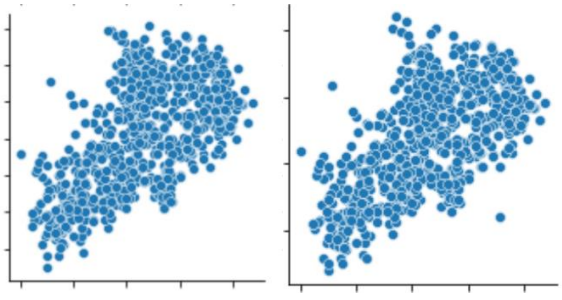
Assignment-based Subjective Questions:

1.From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
Ans: The categorical variables from the dataset are year, seasons, weekdays, weather and
Months. From the year column**,** we can infer that the number of bikes rentals are increased in the year 2019 than 2018, this can be due to many factors, one being the adoption time or people getting accustomed to the new option of transport, second is the increase in population and people preferring not to drive and get stuck in the traffic. From the season's column, we can see that people rent the bikes more in fall, followed by summer, winter, and spring. From the weekday columns, we can infer that except Sunday, all the remaining days, the average number of people renting bikes is almost the same, this might be because of the weekend, more people are getting out of their houses. From the column working day, we can infer that more people are using bikes on working days rather than a holiday. From the categorical column weathersit, the data doesn't have any data of the heavy rain in the data set but we can infer that the bikes are rented more in clear weather sit followed by mist and then light snow.

2. Why is it important to use drop_first=True during dummy variable creation?
Ans: drop_first=True is very important to use, basically we perform dummy encoding for variables more than two levels(assume K), after dummy encoding we can see that the variable can be defined with just k-1 levels, so we use drop_first=True to reduce the extra column and increase the efficiency.

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

Ans: As per the pair-plot among the numerical variables, the highest correlation with the target variable is temp and temp with an equal correlation of 0.63.



a)temp-0.63(corr) **b)atemp-0.63(corr)**

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

Ans: - There is a linear relationship between X and Y

- Errors terms are normally distributed

- Error terms are independent of each other

- Error terms have constant variance (homoscedasticity)

After building the model on the training set, I've checked for the error terms whether they are normally distributed or not and the error terms are normally distributed.

Checked for constant variance and found them to be in homoscedasticity.

The training and testing accuracy are nearly equal hence there is no Overfit/Underfit situation.

The predicted values have a linear relationship with the actual values.

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

Ans: Based on the final model , the top three features contributing significantly towards explaining the demand of shared bikes are temperature, light snow and windspeed.

Temperature: 0.3953, it is obvious that the temperature plays a major role for people renting the bikes.

Light Snow: -0.2742, it is observed that when it's snowing, people tend to avoid bikes.

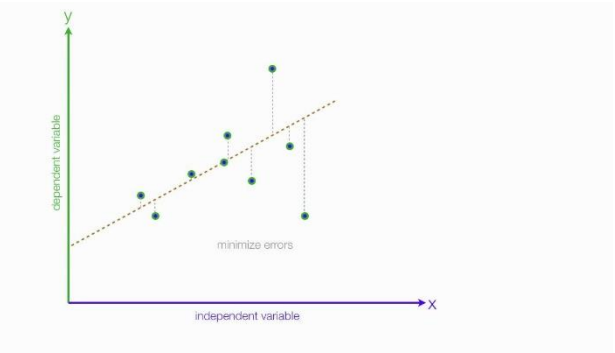
windspeed: -0.1516, it is observed that when the windspeed is increasing, people tend to avoid bikes.

General Subjective Questions:

1. **Explain the linear regression algorithm in detail?**

Ans: **Definition**: Linear Regression is one of the easiest algorithms in machine learning, It is a statistical model that attempts to show the relationship between two variables with the linear equation.
They are two types of regressions:
1. Simple linear regression
2. Multiple linear regression

1. Simple Linear Regression: Simple linear regression is the most elementary type of regression model which explains the relationship between a dependent and an independent variable using a straight line.
Examples: Predicting the sales of a particular item on an E-commerce website, predicting the scores of the prices of a commodity with the past year's data, etc.
A straight line is plotted on a scatter plot having the data points plotted which is called the regression line.
We use linear regression for many things such as evaluating the trends and sales estimates, Analyzing the impact and price changes, assessing risk in the financial services & insurance domain, or assessment of results in medical equipment, etc.
The equation used in linear regression is "Y=mx+c" where m- is the slope, x- is the independent variable, Y- is the dependent variable, and c – is the intercept.



The diagram on the left shows the example of the linear relationship between the dependent and independent variable, The dotted line in the image shows the best-fitted line, this line can be modeled by the linear equation '**Y=mx+c"**

The main target of the linear regression is to find the values of m and c.

The best-fitted line is obtained by the method called the "**Ordinary least squares**" method.

The cost function is the calculation of error between the actual and predicted values, the cost function of linear regression is mean squared error or Root mean square error.

In linear regression, the model tries to get the best-fitted line to predict the values of y based on the given input x. after the model calculates the cost function, it tries to minimize the cost function. To minimize the cost function, the model needs to have the best value of x and y, initially, the model selects the x and y randomly and then iterates the value to minimize the cost function until it reached the minimum. Finally, when the model minimizes the cost function, we will have the best x and y values, and then when we update these values of the linear equation, the model predicts the value of x in the best and accurate manner.

---

2. Explain Anscombe's quartet in detail?
Ans: Anscombe's quartet can be defined as a group of data sets that look identical in simple descriptive statistics, but there is some distinctness in the dataset, they produce different distributions and look very different when plotted on a scatter plot.
This is constructed or explained to understand the importance of visualization before analyzing or building any model.

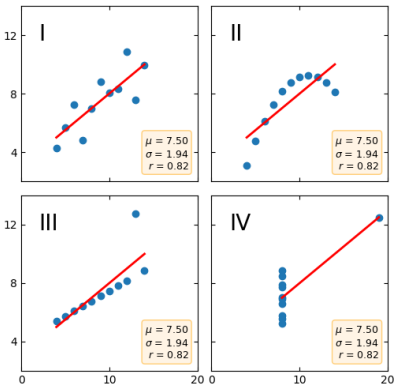
To explain Anscombe's quartet in detail, we took a dataset that has some observations,
Looking at the dataset and calculating the statistical details such as mean, variance, and standard deviation, we found out that there is no difference in the statistical information.

Continued…

This is the dataset we have for understanding and we can see that the summary statistics are the same for all the observations.

When we plotted a scatter plot with these data points, all the plots generated are different and are not interpretable.

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	



From the four scatter plots on the left side, we can witness that only the first data set is suitable for linear regression as the line is fitting the model, all the remaining three models don't look interpretable and no linear model can be built with those data sets.

Hence, this is **Anscombe's quartet,** explaining the importance to plot the data before building any model.

---

3. What is Pearson's R?

Ans: The relation between two variables is measured by the correlation coefficients, there are many types of correlation coefficients ,out of which the most widely used is Pearson's correlation which is also called as pearson's-R.it is used for linear regression. The full name of Pearson-r is pearson product moment correlation.

Pearson's – R draws a line through the data points of variables and shows us the relationship between them and the relation between them is given by pearson coefficient. The relation between the variables can be either positive or negative also..

Examples:

Positive Linear relationship: The cost of the empty land always increases with time.

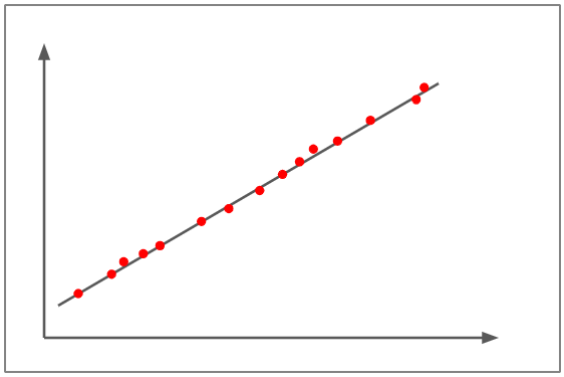
Negative Linear relationship: The cost of a car always decreases with respect to the time, as a depreciation.

The pearson coefficient follows a formula which is :

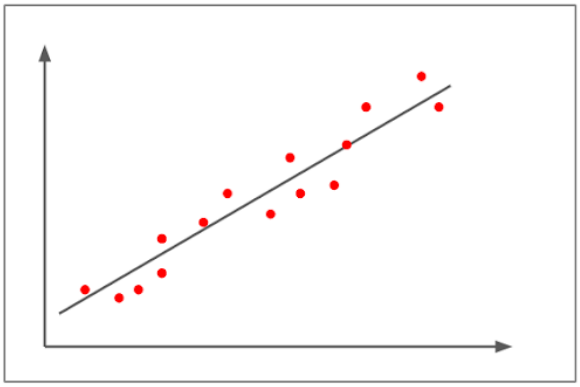
$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\,n\Sigma x^2 - (\Sigma x)^2\,]\,[\,n\Sigma y^2 - (\Sigma y)^2\,]}}$$

Where n = number of pairs
∑xy = the sum of the product of pairs
∑x=sum of x scores
∑y=sum of y scores
∑x^2=sum of squared x scores
∑y^2=sum of squared y scores

1.Positive correlation(r^2=1):

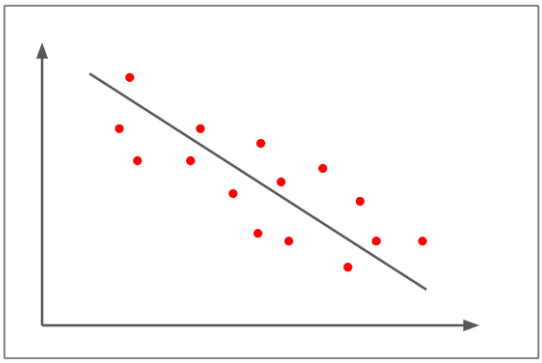


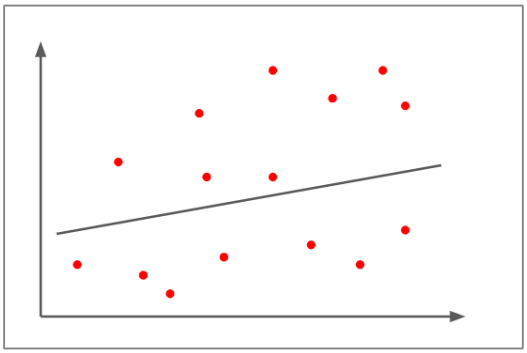
2.medium positive correlation(r^2=0.70):



3.Negative correlation(r^2=-0.70):

4.Weak or no correlation(r^2=0.05):





The above graphs tell us the different correlations as per Pearson's correlation factor.

---

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

Ans: Scaling is a method used in normalizing the range of independent variables of the data. Suppose we have different variables in the dataset such as age, salaries or heights, etc, now they are all on different scales like ages at 1 to 100, salaries at 10,000 to 10,00,000, now while we perform any analysis on this data, we might get a different range of values which will be very hard or difficult to interpret, so we'll do a small modification to this data by changing all the values into the values of range 0 – 1 and bring the data into a standard normal distribution with mean zero and standard deviation one, which will help us interpret the data.

The basic difference between the normalized scaling and standardized scaling is normalization rescales the values into the range of 0-1 whereas standardization rescales the data to have the mean of 0 and a standard deviation of 1.

The formulae in the background used for each of these methods are as given below:

Standardization: x= (x-mean(x))/sd(x)

MinMax scaling:x= (x-min(x))/Max(x)-Min(x)

Some differences between normalized scaling and standardized scaling are :

Normalized scaling	Standardized scaling
Minimum and maximum values of features are used for scaling	Mean and standard deviation is used for scaling
Scaled valued are always between [0,1] or [-1,1]	Scaled values are not bounded to a certain range
It is affected by outliers	It is much less affected by outliers
It is often called as scaling normalization	It is often called as Z-score normalization

---

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

Ans: Basically variance inflation factor or VIF gives a basic quantitative idea about how much the feature variables are correlated with each other.it is an extremely important parameter to test our linear model. The formula for calculating VIF is VIF= $\frac{1}{1-R^2}$…

SO, whenever we have very high or perfect correlation , we get the value of R^2 =1, then the denominator in the formulae becomes 0 which will make the whole VIF infinity. To handle such a situation, we can drop the highly correlated value which is causing this high multicollinearity.

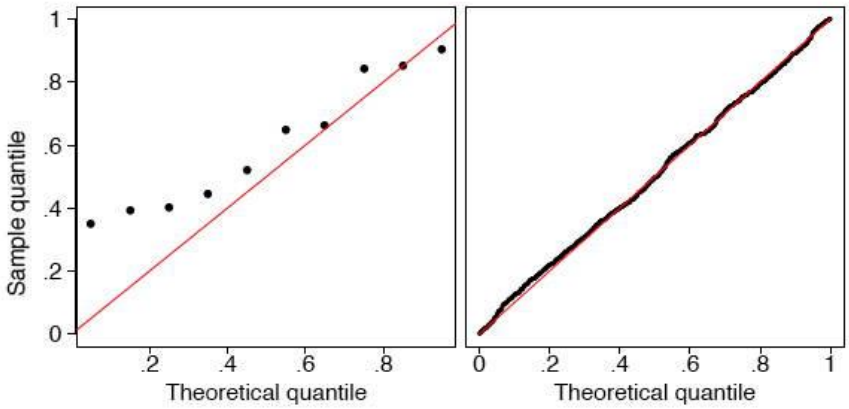
So infinite VIF value means that the corresponding variable may be expressed exactly by a linear combination of other variables which shows infinite VIF.

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

Ans: Technically, “Quantiles are just the lines that divide the data into equally sized groups.”

The median is quantile because it splits the data into groups that contain the same number of data points.

Sometimes, if we have data and we want to know whether the data is normally distributed or not, we will plot a Q-Q plot which will help us to know the answer.



The figure on the left side shows how the data is distributed, after plotting the data points, we’ll see how well the dots fit the straight line, if the data were normally distributed, most of the points would be on the line. In this case, the fit is not awesome.

The figure on the right side shows how the data is distributed, after plotting the data points, we can see that the points are much closer to the line indicating that the data is uniformly distributed and gives a better fit.

A Q-Q plot generally used to distinguish the shapes of the distributions by providing a graphical view of how properties such as scale, skewness are similar or different in the two distributions.