1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans. The demand of bike is comparatively less in the month of spring

- The demand of bike increased in the year 2019 as compared with year 2018.
- Month Jun to Sep is the period when bike demand is high. The Month Jan is the lowest demand month.
- Bike demand is lesser on holidays.
- The demand of bike is almost similar throughout the weekdays.
- The bike demand is high when weather is clear and few clouds however demand is less in case of lightsnow and light rainfall. We do not have any data for Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog , so we can not derive any conclusion. Maybe the company is not operating on those days or there is no demand of bike.

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. By doing so, 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?

Ans. Looking at the pair plot, the highest correlation with the number of bikes sold seems to be with the temperature in Celsius (and feeing temperature)

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

Ans. The assumption is that the error terms are normally distributed with mean zero. We prove this by plotting a histogram of y_train – y_train_pred.

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

Ans. Temperature (temp) - A coefficient value of ‘0.5636’ indicates that a unit increase in temp variable increases the number of hired bikes by 0.5636 units. Which is a very high correlation.

Weather Situation 3 (weathersit_3) - A coefficient value of ‘-0.3070’ indicated that, a unit increase in Weathersit3 (Light Snow, Light Rain, Thunderstorm etc) decreases the bike hire numbers by 0.3070 units.

Year (yr) - A coefficient value of '0.2308' indicated that a every year the number of hired bikes increases by 0.2308 units.

These three variables seem to be contributing the most towards the demand for shared bikes.

General Subjective questions

1. Explain the linear regression algorithm in detail.

Ans. Linear Regression is a machine learning algorithm. It is based on supervised learning. It performs a regression task. Regression essentially models a target value we need to predict based on mutually independent variables. It is mostly used for forecasting and for making business decisions regarding which variable contributes to what extent in the final outcome.

Linear regression is used to forecast a dependent variable value (y) based on a given independent variable (x). So, this creates a linear relationship between x (input) and y(output). Hence, it is called Linear Regression.

2. Explain the Anscombe's quartet in detail.

Ans. Anscombe's quartet comprises four datasets that have nearly identical statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

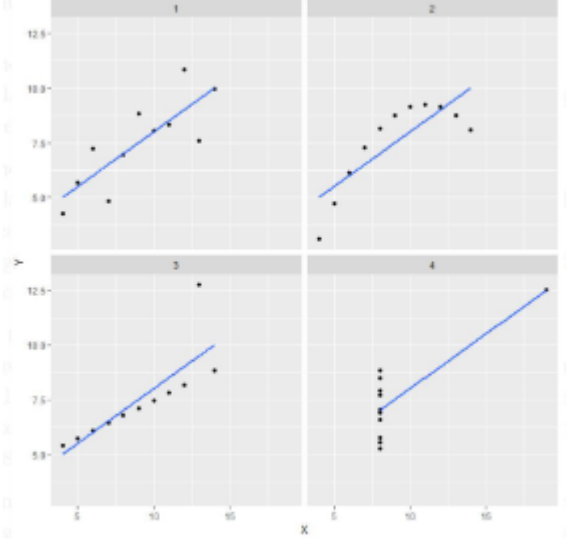
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Shown below are their statistical properties

Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

But when plotted they look very different



This is used as an example to emphasize the importance of graphing an data/ visualizing it instead of just computing the statistical properties, because they don't give us enough information.

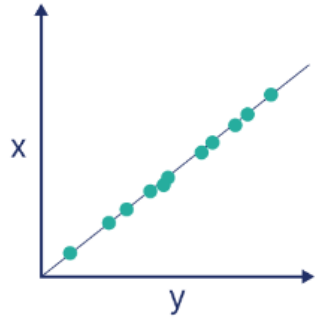
3. What is Pearson's R?

Ans. The Pearson correlation coefficient (*r*) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

You can look at r as measure of how close the observations are to a line of best fit.

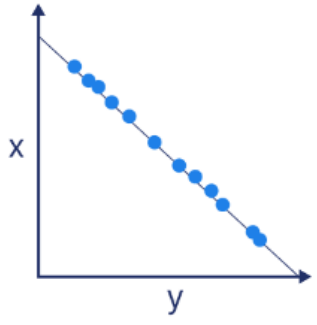
Perfect positive correlation

$r = 1$



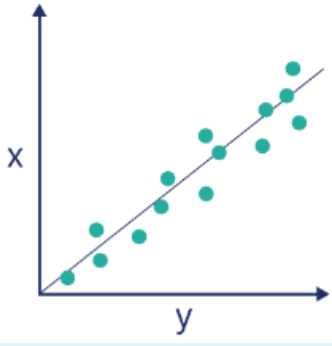
Perfect negative correlation

$r = -1$



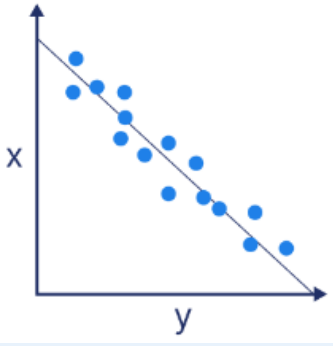
Strong positive correlation

$r > .5$



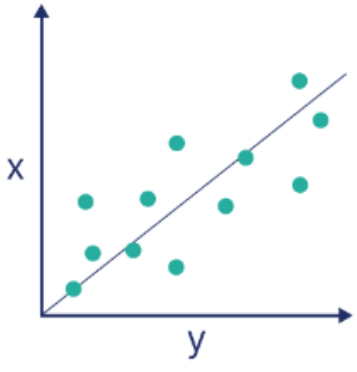
Strong negative correlation

$r < -.5$



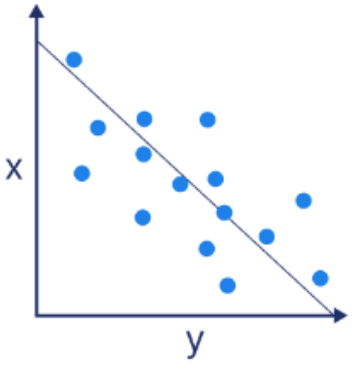
Weak positive correlation

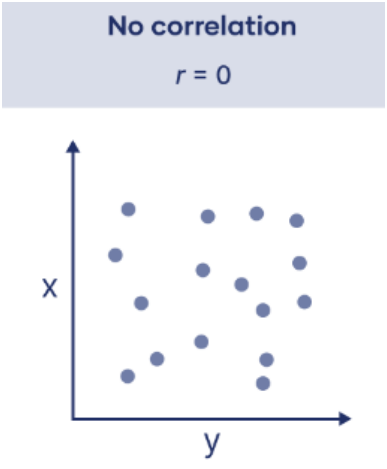
$.3 > r > 0$



Weak negative correlation

$0 > r > -.3$





Pearson's coefficient is most useful when both variables are quantitative, normally distributed and when the relationship is linear. It should not be used if there are outliers.

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

Ans. Feature Scaling is a technique to bring down all the independent features present in the data in a uniform range. It is a step under data pre-processing to handle data with different magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values. So for example if one value is 300 m and another value is 5km, the algorithm will read 300 m as a greater value, which obviously is wrong.
The most common techniques of feature scaling are

a) MinMaxScaling, and
b) Standardization.

Min Max Scaling -

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

This Scaler shrinks the data within the range of -1 to 1 if there are negative values. We can set the range like [0,1] or [0,5] or [-1,1].

Standardization –

$$x_{new} = \frac{x - \mu}{\sigma}$$

The Standard Scaler assumes data is normally distributed within each feature and scales them such that the distribution centered around 0, with a standard deviation of 1.

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

Ans. VIF being infinite means there is perfect correlation between the two variables. In this case we might want to eliminate the duplicate variables/ columns.

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

Ans. Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against.

The slope tells us whether the steps in our data are too big or too small (or just right). This is because each step in the data traverses a fixed and constant percentage.

A steeply sloping section of the QQ plot means that in this part of our data, the observations are more spread out than we would expect them to be if they were normally distributed. A flat QQ plot means that our data is more bunched together than we would expect from a normal distribution.