

## 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 demand of bike is less in the month of spring when compared with other seasons
- The demand bike increased in the year 2019 when 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 less in holidays in comparison to not being holiday.
- The demand of bike is almost similar throughout the weekdays.

**2. Why is it important to use drop\_first=True during dummy variable creation?**

**Ans:** It helps in reducing the extra column created during dummy variable creation

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

**Ans:** temp and atemp has the 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:** 1. R2 value for predictions on test data (0.815) is almost same as R2 value of train data (0.818). This is a good R-squared value; hence we can see our model is performing good even on unseen data (test data)

2. Errors are normally distributed here with mean 0

3. Variance of the residuals (error terms) is constant across predictions.

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

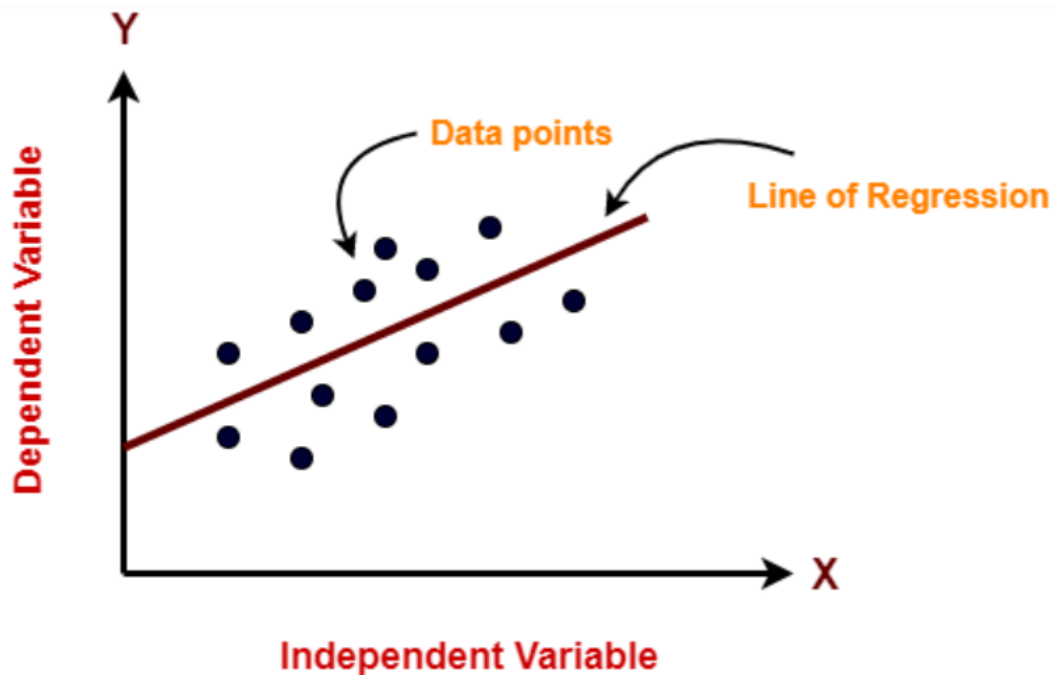
**Ans:** The top 3 features contributing significantly towards the demands of share bikes are:

- Weathersit\_Light\_Snow (negative correlation)
- yr\_2019(positive correlation)
- temp (positive correlation)

## General Subjective Questions

### Explain the linear regression algorithm in detail.

**Ans:** Linear Regression is an ML algorithm used for supervised learning. Linear regression performs the task to predict a dependent variable(target) based on the given independent variable(s). So, this regression technique finds out a linear relationship between a dependent variable and the other given independent variables. Hence, the name of this algorithm is Linear Regression.



In the figure above, on X-axis is the independent variable and on Y-axis is the output. The regression line is the best fit line for a model. And our main objective in this algorithm is to find this best fit line.

#### Pros:

- Linear Regression is simple to implement.
- Less complexity compared to other algorithms.
- Linear Regression may lead to over-fitting, but it can be avoided using some dimensionality reduction techniques, regularization techniques, and cross-validation.

#### Cons:

- Outliers affect this algorithm badly.

- It over-simplifies real-world problems by assuming a linear relationship among the variables, hence not recommended for practical use-cases.

### **Explain the Anscombe's quartet in detail.**

Anscombe's Quartet can be defined as a group of four data sets which are **nearly identical in simple descriptive statistics**, but there are some peculiarities in the dataset that **fools the regression model** if built. They have very different distributions and **appear differently** when plotted on scatter plots.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other **observations on statistical properties**. There are these four data set plots which have nearly **same statistical observations**, which provides same statistical information that involves **variance**, and **mean** of all x, y points in all four datasets.

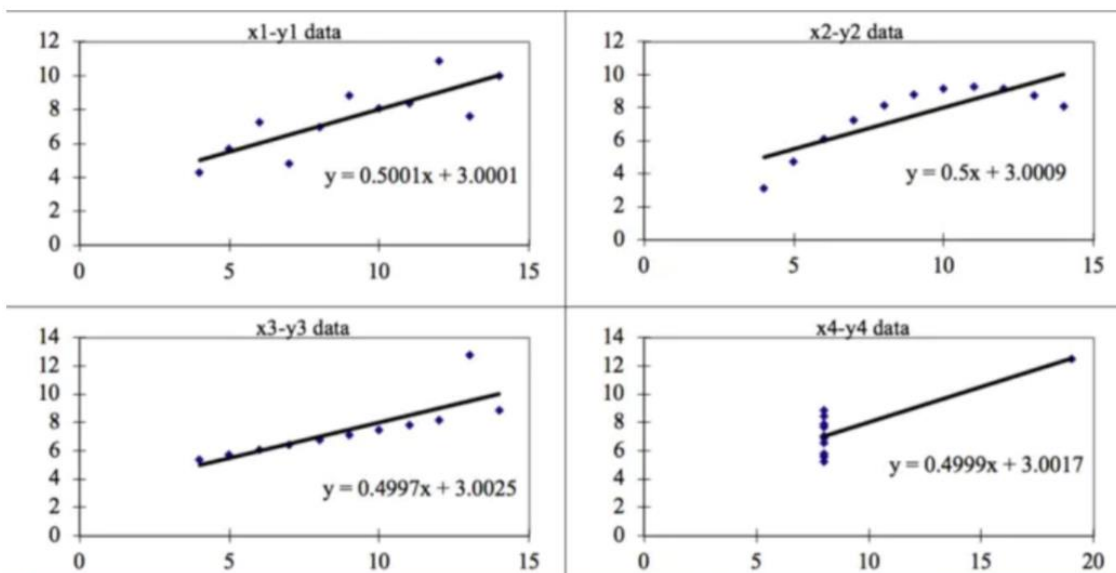
This tells us about the importance of visualizing the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.

Also, the Linear Regression can only be considered a fit for the **data with linear relationships** and is incapable of handling any other kind of datasets.

These four plots can be defined 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

When these models are plotted on a scatter plot, all datasets generate a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:



The four datasets can be described as:

- **Dataset 1:** this **fits** the linear regression model pretty well.
- **Dataset 2:** this **could not fit** linear regression model on the data quite well as the data is non-linear.
- **Dataset 3:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model.
- **Dataset 4:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model.

## Conclusion

- We have described the four datasets that were intentionally created to describe the importance of data visualization and how any regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualized before implementing any machine learning algorithm on them which will help to make a good fit model.

## What is Pearson's R?

The Pearson's Correlation Coefficient is also referred to as **Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation**. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

**It is the covariance of the two variables divided by the product of their standard deviations.**

There are certain requirements for Pearson's Correlation Coefficient:

- Scale of measurement should be interval or ratio
- Variables should be approximately normally distributed
- The association should be linear
- There should be no outliers in the data

It can be calculated using the below formula

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

Where,

$N$  = the number of pairs of scores

$\Sigma xy$  = the sum of the products of paired scores

$\Sigma x$  = the sum of x scores

$\Sigma y$  = the sum of y scores

$\Sigma x^2$  = the sum of squared x scores

$\Sigma y^2$  = the sum of squared y scores

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

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying 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.

Normalized scaling	Standardized scaling
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
It is often called as Scaling Normalization	It is often called as Z-Score Normalization

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

If there is perfect correlation, then  $VIF = \text{infinity}$ . This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get  $R^2 = 1$ ,

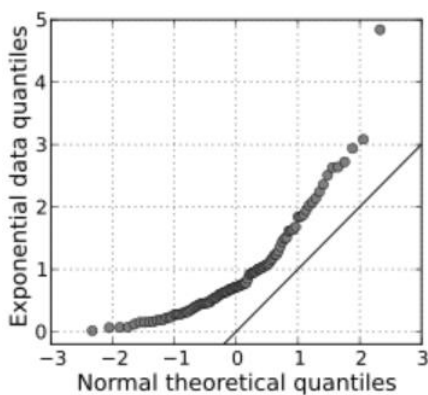
which lead to  $1/(1-R^2)$  infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

### **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.