

# 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 : Top variables that are seen effecting and benefitting the Bike Rental count are as follows:

- Spring season: -0.6657 : In Spring Season bike hires are reduced by .665 units
- Feeling Temperature: 0.3768 : Feeling Temperature Indicates that a unit increase in feeling temperature increases bike hire by .3768 units.
- Year : 1.0415 : In 2019 there were 1.0415 times more bike hires than 2018.
- Light Snow : -1.3836 , Mist : -0.3647 : Light Snow and Mist have inverse reaction to bike hires.
- Sun: 0.2749 : On Sunday we have a increase in bike hires by .2749 units.
- working\_day: 0.2610: Working day have a impact of .2610 on bike hires.

- 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. Hence 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: Temp and Atemp have highest correlation with target variable.

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

Ans: Concluded based upon the Linear Regression –

- The error terms are normally distributed.
- The training and testing accuracy are nearly equal hence there is no Overfit/Underfit situation.
- The predicted values have linear relationship with the actual values.
- Also by using the in-build function of scikitlearn called R2

`r2_score(y_test, y_test_pred)`

which predicted that model has accuracy of 81%.

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

**Ans :**

1. Bike Rentals are more during the Fall(Monsoon) season
2. Bikes seem to be rented more in Partly cloudy weather.
3. Bikes seem to be rented more on working days.

## General Subjective Questions

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

**Answer-**Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on –the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

$$y = mx + c$$

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

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

**Answer-**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.

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.

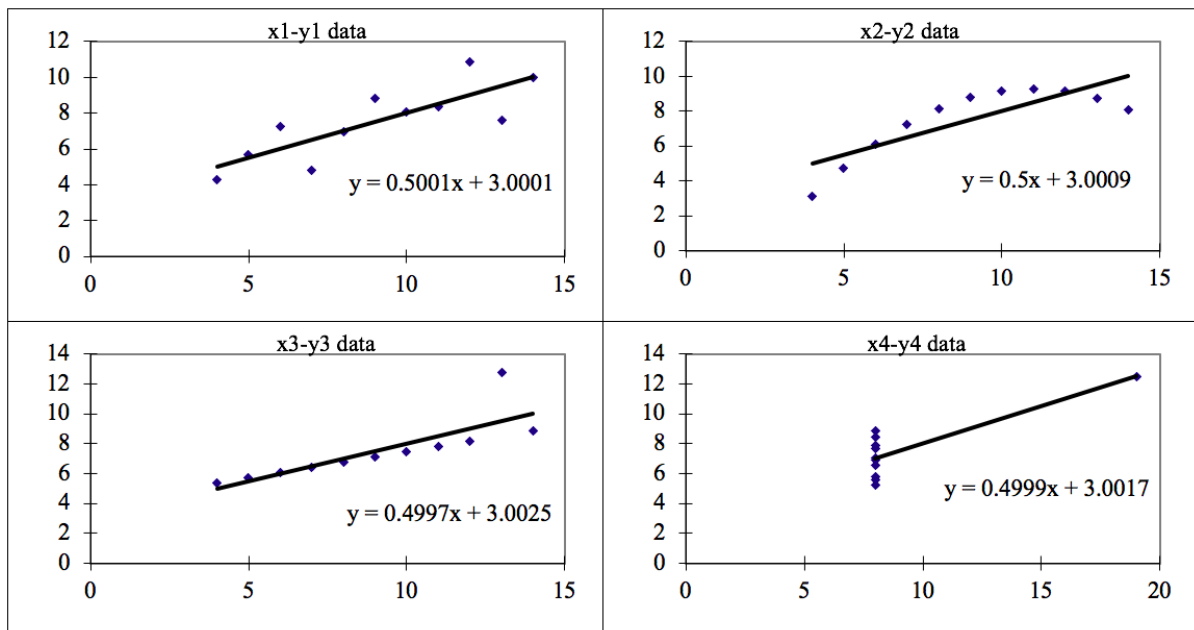
This tells us about the importance of visualizing the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be consider eda fit for the data with linear relationship and is incapable of handling any other kind of datasets.

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 all these four datasets are approximately similar and can be computed 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	

When these models are plotted on a scatter plot, all datasets generates 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:

1. **Dataset 1:** this **fits** the linear regression model pretty well.
2. **Dataset 2:** this **could not fit** linear regression model on the data quite well as the data is non-linear.
3. **Dataset 3:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model
4. **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 visualisation and how any regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model.

### 3. What is Pearson's R?

**Answer :** 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.

Pearson correlation coefficient ( $r$ )	Correlation type	Interpretation	Example
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the <b>same direction</b> .	Baby length & weight:  The longer the baby, the heavier their weight.
0	No correlation	There is <b>no relationship</b> between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and $-1$	Negative correlation	When one variable changes, the other variable changes in the <b>opposite direction</b> .	Elevation & air pressure: The higher the elevation, the lower the air pressure.

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

**Answer :**

#### What is 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 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. It is important to note that **scaling just affects the coefficients** and none of the other parameters like **t-statistic, F-statistic, p-values, R-squared**, etc.

### 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.

$$\text{MinMax Scaling: } x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

### Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean ( $\mu$ ) zero and standard deviation one ( $\sigma$ ).

$$\text{Standardisation: } x = \frac{x - \text{mean}(x)}{\text{sd}(x)}$$

- `sklearn.preprocessing.scale` helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.

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

**Answer:** If there is perfect correlation, then  $VIF = \infty$ . This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get  $R^2 = 1$ , which leads to  $1/(1-R^2) = \infty$ . To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

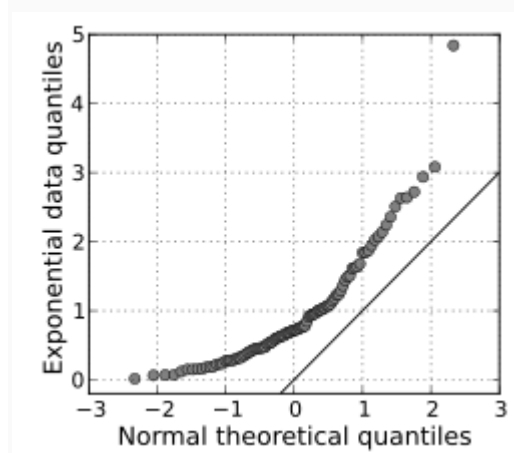
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

**Answer :**

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.