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- **Because dummy coding compares the mean of the dependent variable for each level of the categorical variable to the mean of the dependent variable at for the reference group, it makes sense with a nominal variable.**

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. 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- **Registered.**

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

Ans**- Pair-wise scatterplots may be helpful in validating the linearity assumption as it is easy to visualize a linear relationship on a plot.**

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

   Ans- Sun ,working_day and Temperature.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

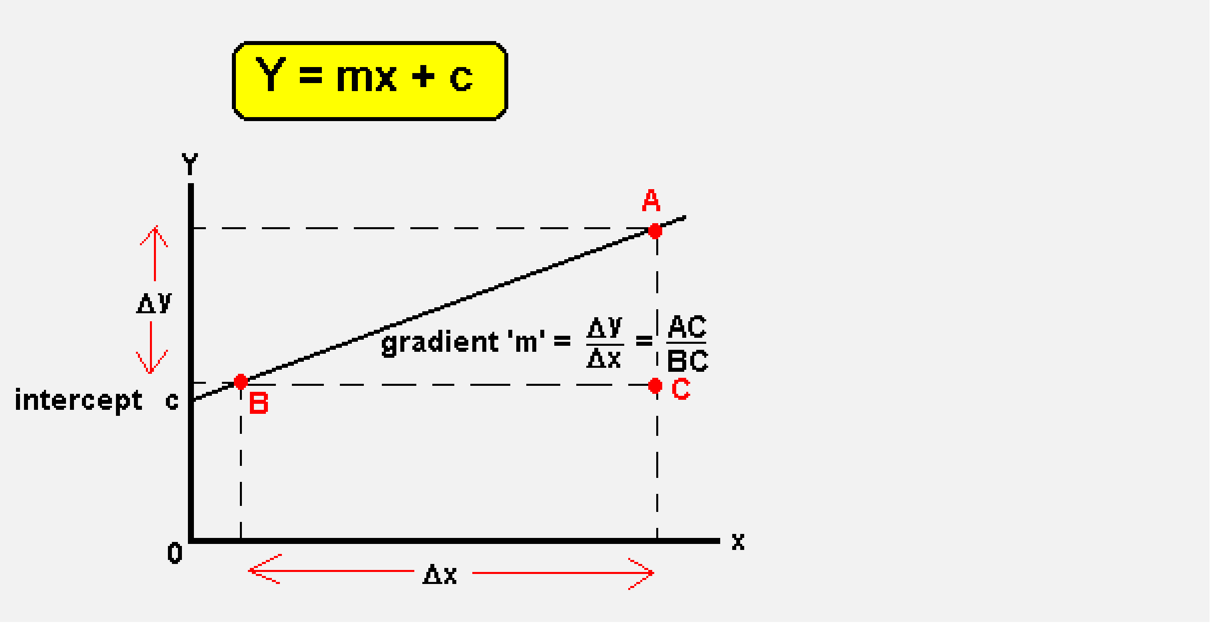
Ans- Linear Regression is a linear approach to modelling the relationship between a scalar response (or dependent variable or y) and one or more explanatory variables (or independent variables or x).

Simple Linear Regression:-The case of one explanatory variable is called Simple Linear Regression (i.e it is a linear relationship between x and y).

Mathematically speaking, everyone must have heard about the straight-line equation:

y = mx + c

where y is the dependent variable, x is the independent variable, m is the slope of the line and c is y-intercept(value *of y when x=0*)



2. Explain the Anscombe's quartet in detail.

Ans- 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 analysing 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 visualising 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 be 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

Image by Author

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	

Image by Author

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:

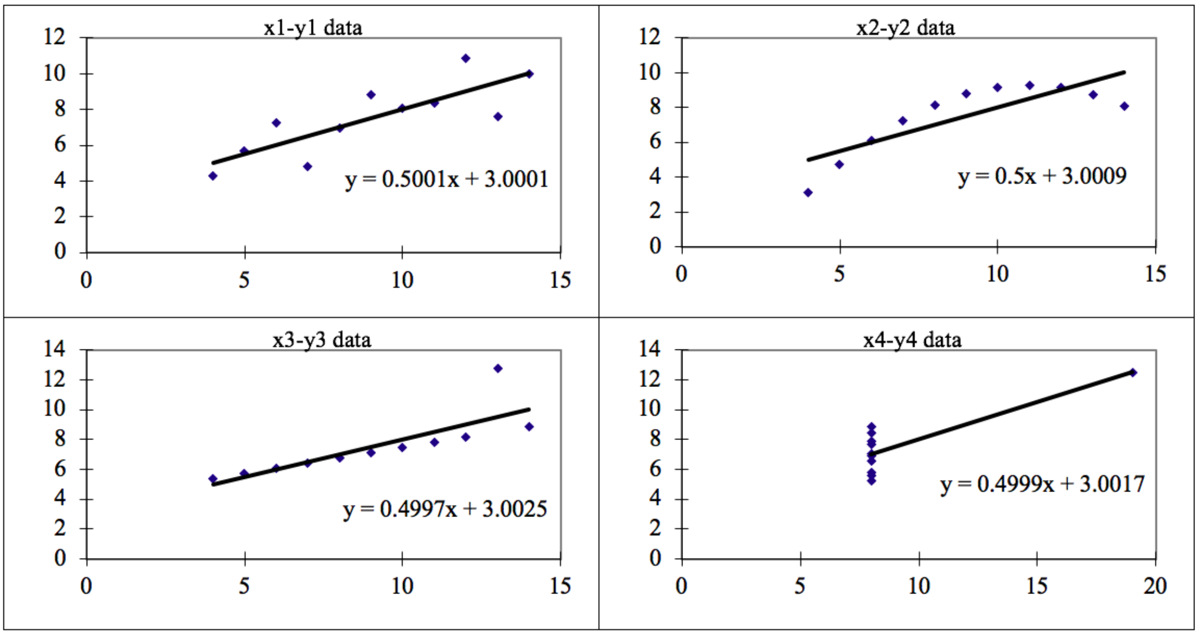


Image by Author

The four datasets can be described as:

1. **Dataset 1: this fits the linear regression model 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.**

3. What is Pearson's R?

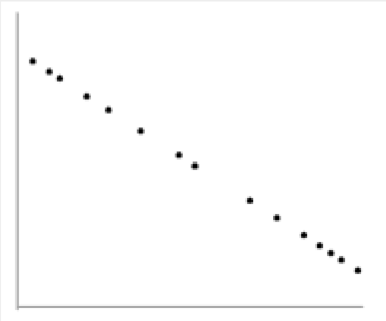
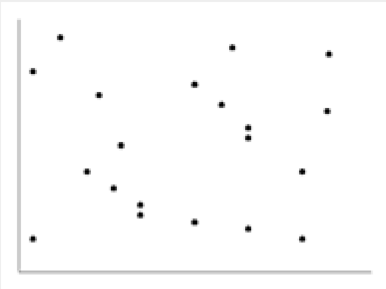
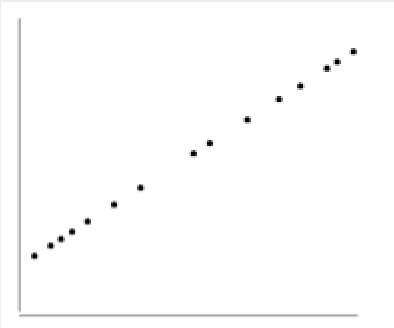
Ans- Correlation is a technique for investigating the relationship between two quantitative, continuous variables, for example, age and blood pressure. Pearson's correlation coefficient (r) is a measure of the strength of the association between the two variables.

The first step in studying the relationship between two continuous variables is to draw a scatter plot of the variables to check for linearity. The correlation coefficient should not be calculated if the relationship is not linear. For correlation only purposes, it does not really matter on which axis the variables are plotted. However, conventionally, the independent (or explanatory) variable is plotted on the x-axis (horizontally) and the dependent (or response) variable is plotted on the y-axis (vertically).

The nearer the scatter of points is to a straight line, the higher the strength of association between the variables. Also, it does not matter what measurement units are used.

Values of Pearson's correlation coefficient

Pearson's correlation coefficient (r) for continuous (interval level) data ranges from -1 to +1:

r = -1		**data lie on a perfect straight line with a negative slope**
r = 0		**no linear relationship between the variables**
r = +1		**data lie on a perfect straight line with a positive slope**

Positive correlation indicates that both variables increase or decrease together, whereas negative correlation indicates that as one variable increases, so the other decreases, and vice versa.

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 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.**

It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

The terms normalization and standardization are sometimes used interchangeably, but they usually refer to different things. Normalization usually means to scale a variable to have a value between 0 and 1, while standardization transforms data to have a mean of zero and a standard deviation of 1.

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

Ans- 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).

- **In VIF, each feature is regression against all other features. If R2 is more which means this feature is correlated with other features. [0]**
    - **VIF = 1 / (1 – R2)**
    - **When R2 reaches 1, VIF reaches infinity**
- **We try to remove features for which VIF > 5**
- **Once we identify high VIF for features we need to reduce it**
    - **We can do it by eliminating some features**
    - **How to identify which feature to remove?**
        - **Check the correlated features for feature having high VIF**
        - **In the example at [1] weight and BSA were correlated**
        - **Practically it is easy to measure weight so we kept it**
            - **So such decision depends on the practical implication**
        - **There can be the case that one feature is correlated with many others and we might want to remove**

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-12.87	2.56	-5.03	0.000	
Age	0.7033	0.0496	14.18	0.000	1.76
Weight	0.9699	0.0631	15.37	0.000	8.42
BSA	3.78	1.58	2.39	0.033	5.33
Dur	0.0684	0.0484	1.41	0.182	1.24
Pulse	-0.0845	0.0516	-1.64	0.126	4.41
Stress	0.00557	0.00341	1.63	0.126	1.83

it

Correlation: BP, Age, Weight, BSA, Dur, Pulse, Stress

	BP	Age	Weight	BSA	Dur	Pulse
Age	0.659					
Weight	0.950	0.407				
BSA	0.866	0.378	0.875			
Dur	0.293	0.344	0.201	0.131		
Pulse	0.721	0.619	0.659	0.465	0.402	
Stress	0.164	0.368	0.034	0.018	0.312	0.506

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

Ans- **The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set**.

Quantile-Quantile (Q-Q) plot is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Few advantages:

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behaviour

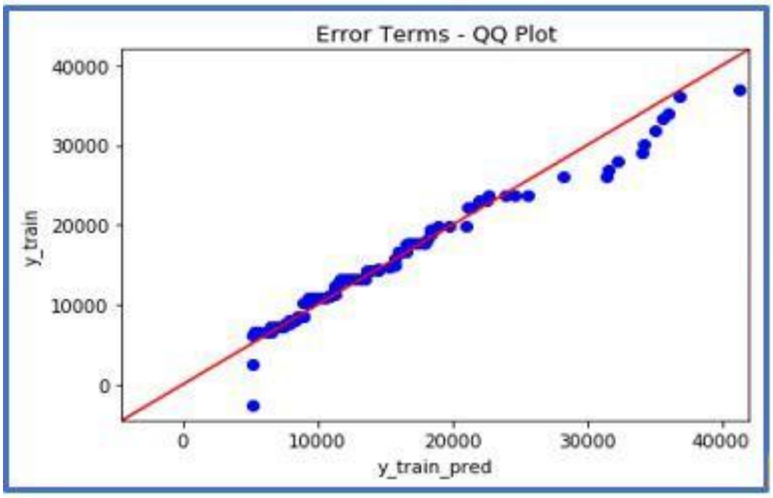
Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

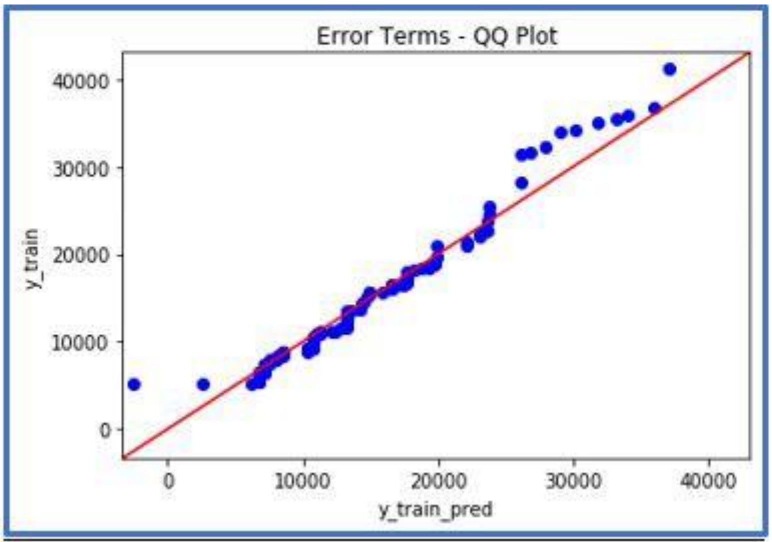
Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis

Python:

statsmodels.api provide qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets, respectively.