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? (3 marks)

Answer – The dataset has the following categorical columns

- Season
- Holiday
- Workingday
- Weathersit
- Month
- Weekday
- Year

Following are my observation of categorical columns impacting the dependent variables

a. Season - People seems to be booking more ride in summer and fall and avoiding the rides in spring and winter.
b. Holiday, Weekday & Workingday- There is low variation of rides on holidays, working days, weekdays and non-holidays.
c. Weathersit – People prefer to take more rides when the sky is clear or partly cloudy.
d. Month and Year – There is an increasing trend of people preferring rides as the years pass and the months are May till October seem to be observing higher rides compared to the remaining months.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Answer – This is done to avoid multicollinearity. When there are N categories we should think of having N-1 variables while performing the encoding.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer – Registered user seems to have the highest correlation with the target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer – I will perform the following steps to validate the assumption of Linear Regression

- The residual(i.e y_train – Y_train_pred) should be normally distributed and centred around zero.
- The model r2 score should be high maybe more than 60 or 70% to ensure the selected features explains the variation of the target.
- RMSE can also be a good metric to confirm the quality of the model build.

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

Answer – following are the top features impacting the rides –

- weather
- working day
- year

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Answer – Linear regression algorithm explains the relationship between dependent variables and one or more independent variables. When this relation is linear in nature or if they are highly correlated with the target variable then we call the dataset a good candidate for linear regression modelling. The equation of a straight line is given by y =mx + c. If the target is dependent on multiple indepent variable then the equation is given by y = m1x + m2x + m3X…..+ c + e

The following are the assumption of simple linear regression –

- There is a linear relation between x and y.
- Error terms are normally distributed.
- Error terms are independent of each other.
- Error terms have constant variance.

The strength of the linear regression model can be assessed using 2 metrics –

- R2 – R2 is a number which explains what portion of the given data variation is explained by the developed model. It always takes a value between 0 & 1.
- RMSE - In statistics, it is defined as the total sum of errors across the whole sample. It is the measure of the difference between the expected and the actual output. A small value indicates a tight fit of the model to the data.

2. Explain the Anscombe's quartet in detail. (3 marks)

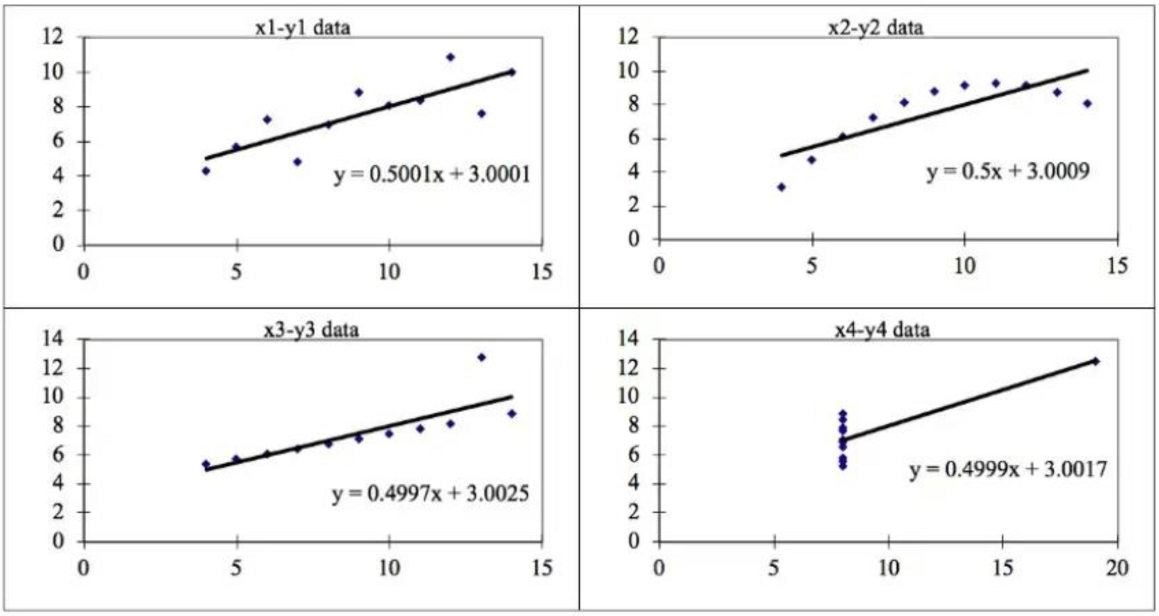
Answer – Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them.

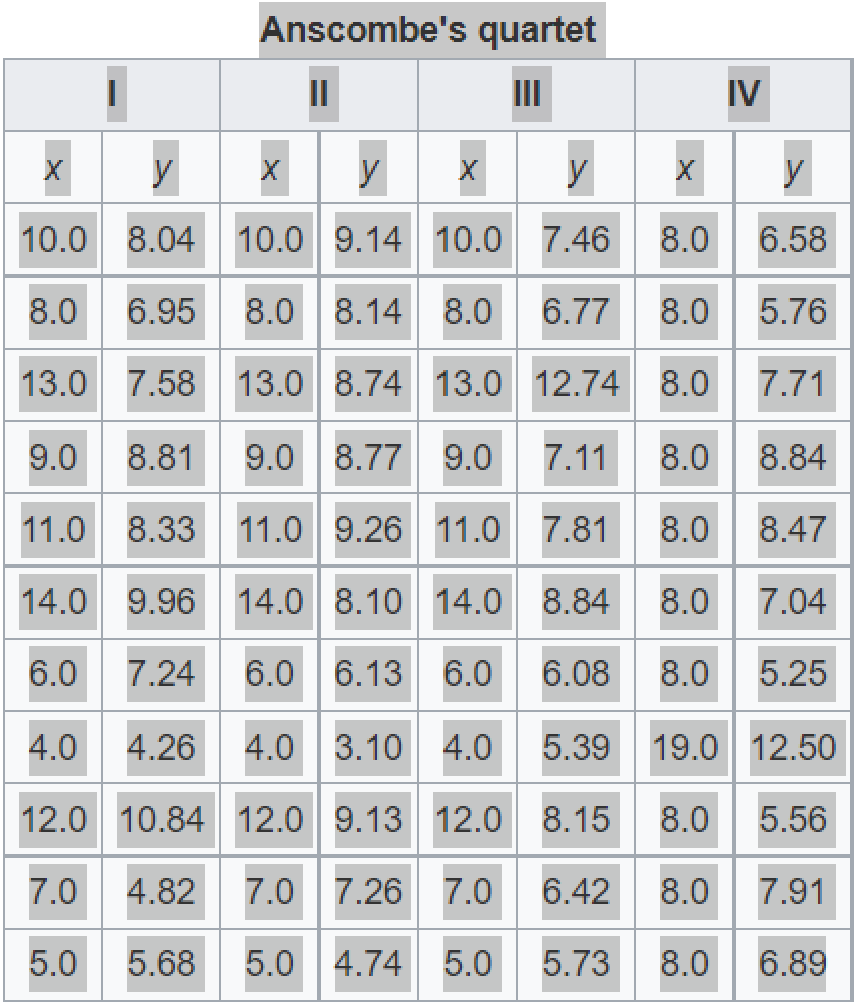
Data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.

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.

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 datasets are as follows. The x values are the same for the first three datasets.



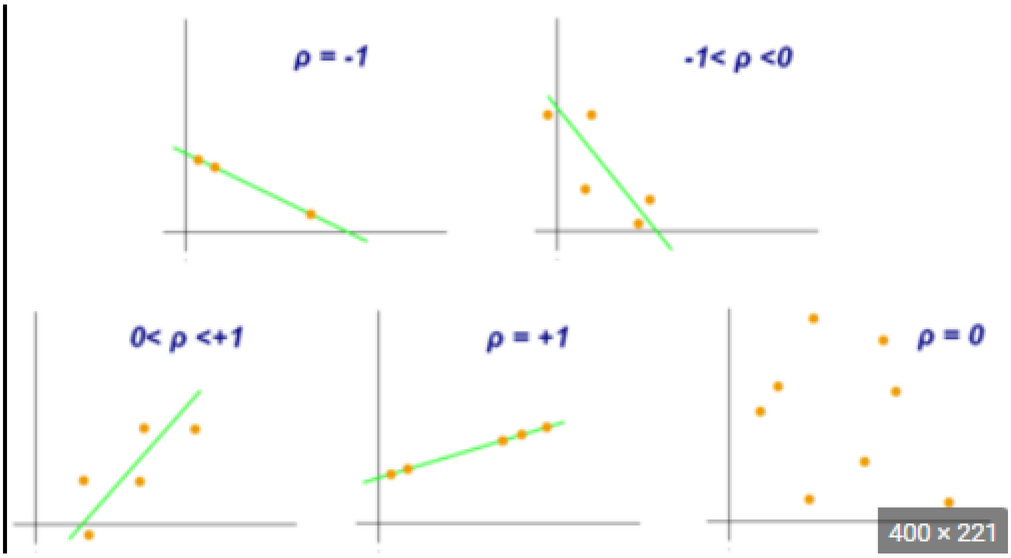
Anscombe's quartet

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

3. What is Pearson’s R? (3 marks)

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. When one variable changes, the other variable changes in the same direction.

The below diagram explains this beautifully



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer – Scaling is the process of normalizing the data between 0 to 1. Scaling is performed to ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model. Or in other words, the model tends to get biased with higher values and the coefficient for this variable would be much higher than the others.

Normalized scaling - Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Here's the formula for normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

Standardized scaling - Standardization is another scaling technique where the values are centred around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Here's the formula for standardization:

$$X' = \frac{X - \mu}{\sigma}$$

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer – VIF infinite means the feature/independent variable is perfectly co-related or explained by other features.

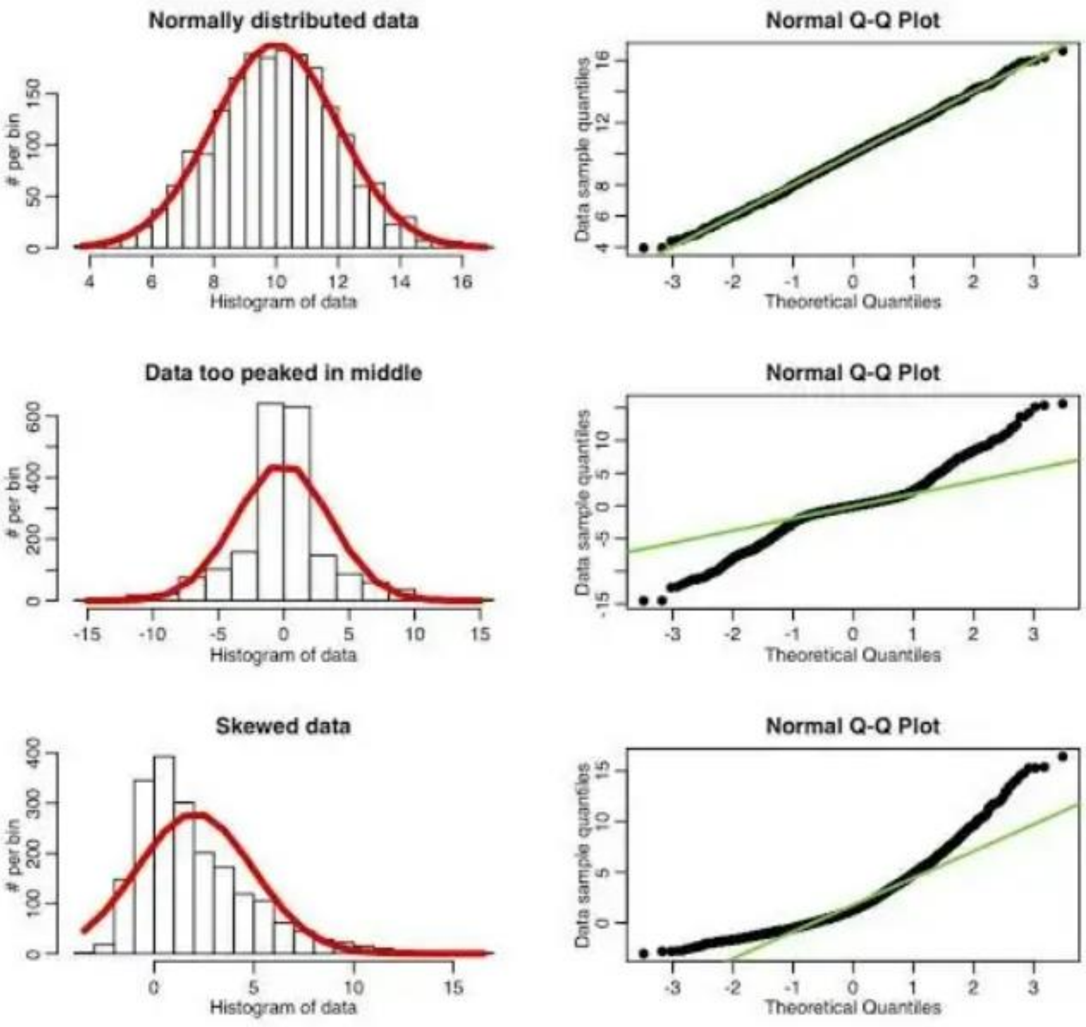
Statistically, the VIF formula is 1 / (1 – R Square), if R square is 1 then VIF is infinite which in turn means perfect co-relation.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer – QQ plot is used to compare two probability distributions by plotting their quantiles against each other. If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line y = x.

Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian Distribution, Uniform Distribution, Exponential Distribution or even Pareto Distribution, etc.

For linear regression they can be drawn to see the distribution of the residuals. Below images explain the QQ plot against the respective distribution


Normally distributed data
# per bin
Histogram of data
Normal Q-Q Plot
Data sample quantiles
Theoretical Quantiles
Data too peaked in middle
# per bin
Histogram of data
Normal Q-Q Plot
Data sample quantiles
Theoretical Quantiles
Skewed data
# per bin
Histogram of data
Normal Q-Q Plot
Data sample quantiles
Theoretical Quantiles