1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

    - There are many categorical variables in dataset. Some of variables are highly corelated to bike rental.
    - To see correlation , build boxplot of categorical variables.
    - From pairplot, we can notice that,
        - bike rental is more in 2019 year compared to 2018
        - bike rental is more on holidays
        - bike rental is more on Sat, Wed, Thu and Working day
        - bike rental is more in clear weather in Summer and Fall season

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

    - drop_first=True helps in reducing the extra column created during dummy variable creation and 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?

    - Bike rental correlated to temp

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

    - By doing residual analysis and making the final predictions on training dataset

    Result Comparison between Train model and Test:

    - Train R^2 : 0.804
    - Train Adjusted R^2 : 0.800
    - Test R^2: 0.781
    - Test Adjusted R^2: 0.754
    - Difference in R^2 between train and test: 1.029%
    - Difference in adjusted R^2 between Train and test: 1.035% which is less than 5%

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

    - Based on the final model , 3 variables contributing to demand of the shared bike are
        1. Temp - A coefficient value for 0.5377 indicated that a unit increase in temp variable increases the bike hire numbers by 0.5377 units.
        2. Year - A coefficient value for 0.2331 indicated that a unit increase in year variable increases the bike hire numbers by 0.2331 units.
        3. Saturday - A coefficient value for 0.0119 indicated that a unit increase in year variable increases the bike hire numbers by 0.0119 units.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

- Linear regression is a process of estimating the relationship among variables. The focus here is to establish the relationship between a dependent variable and one or more independent variable(s)

Linear regression performs the task to predict a dependent variable(target) based on the given independent variable(s)

Mathematically Linear regression can be mentioned as

$y = mx + c$

Where m is slope , c is intercept value on y axis

y is dependent variable and x is independent variable

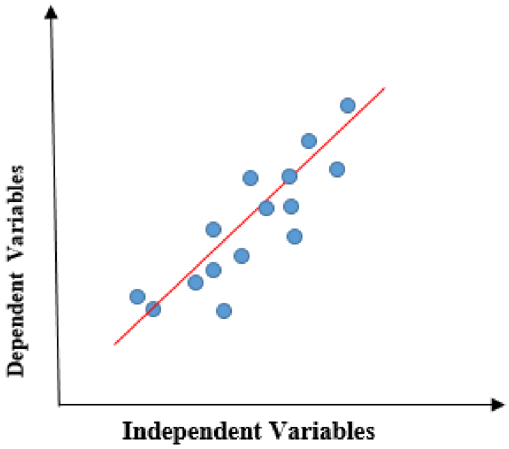


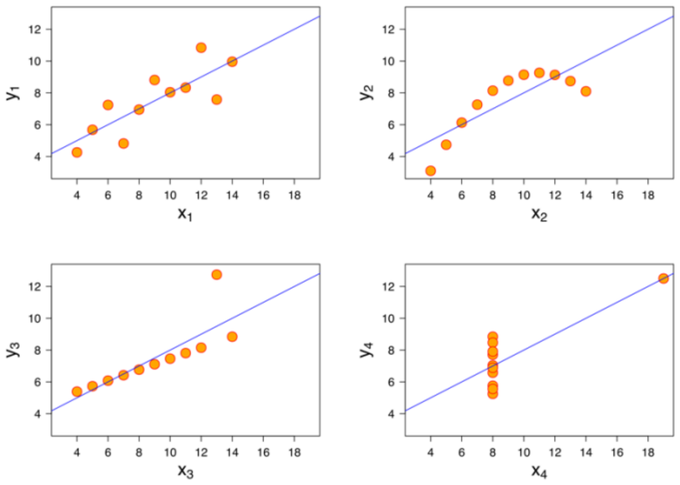
Figure 1

Figure 1 denotes the linear relationship between the dependent variable and independent variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. As the value of x (independent variable) increase, the value of y (dependent variable) increase. The red line is referred to as the best fit straight line. Based on the given data points, try to plot a line that fits the points the best.

If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression.

2. Explain the Anscombe’s quartet in detail.

    - 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. 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 when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough."



- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.

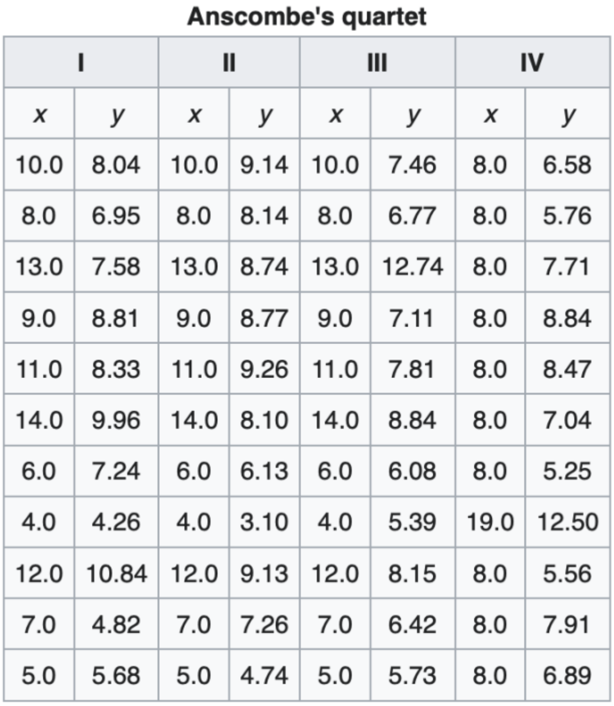
- The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.

- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

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

Reference: https://en.wikipedia.org/wiki/Anscombe%27s_quartet

3. What is Pearson's R?

    - Pearson's *r* is correlation coefficient, it is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

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

    - Scaling is done when there are many independent variables in a model, each variable are on different scales. This may lead a model with very weird coefficients, that might be difficult to interpret. So we need to scale features because of two reasons:

    A. Ease of interpretation

    B. Faster convergence for gradient descent methods

Two type of scaling are:

1. Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.
2. Min-Max Scaling or Normalized Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

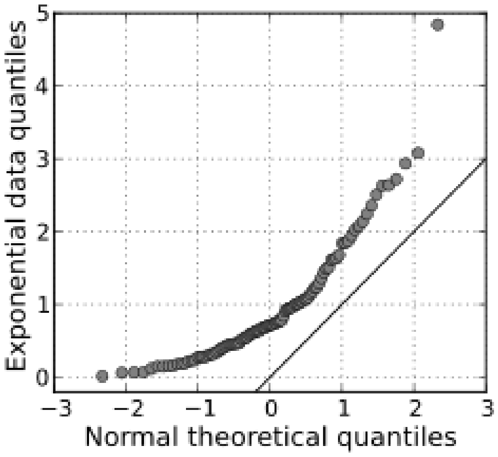
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

    - An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.
      If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. In order to solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

    Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.[1] First, the set of intervals for the quantiles is chosen. A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). Thus the line is a parametric curve with the parameter which is the number of the interval for the quantile.

    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.

Reference: https://en.wikipedia.org/wiki/Q–Q_plot