# 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)

   **Ans:**

   From boxplot analysis in ipynb notebook

   - Bike demand was more in summer and fall season and less in spring and winter
   - Bike demand was more in sept oct and less in Jan dec
   - Bike demand was almost same over weekdays except slight dip in sunday
   - Bike demand was less during snow thunderstrom environment and high during clear less cloudy

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

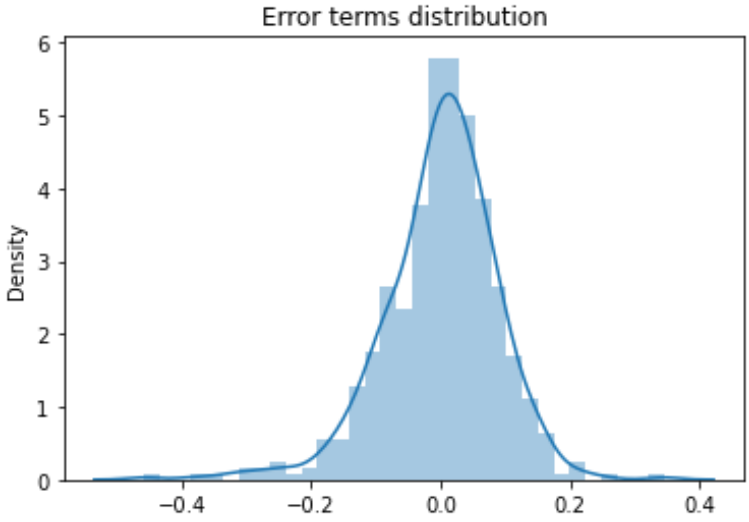
   **Ans:**
   If we don't use drop first, example if we have 3 levels in categorical variables, 3 dummy variables will be created in which one will be redundant. This will increase multicollinearity as one of them will be highly correlated with other.

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

   **Ans:** Variable temp has high correlation with cnt .

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

   **Ans:** Error terms should be normally distributed , We plot sns.distplot of residuals to see if error is normally distributed



5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?** (2 marks)

**Ans:** Temperature, September month, winter season

## General Subjective Questions

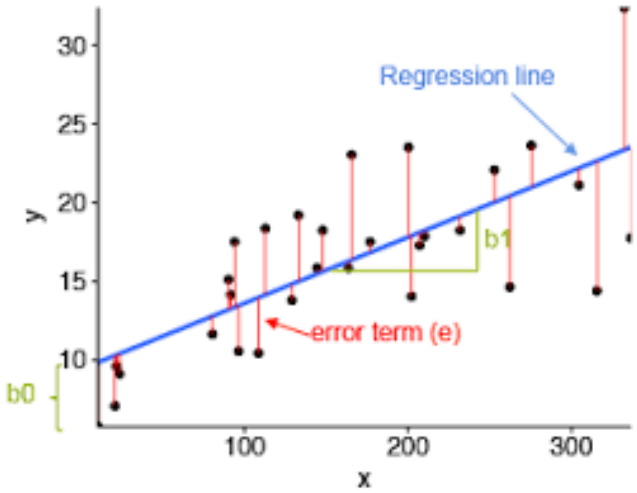
1. **Explain the linear regression algorithm in detail.** (4 marks)

**Ans:** Linear regression is type of supervised machine learning model.

**Simple linear regression:** It represents relation between dependent variable y and independent variable x. The relation is represented in form of best fit straight line.

$Y=mx+c$, **y is dependent variable and x is independent variable.**

M is slope which indicates effect of x on y, c is the intercept




**Multiple linear regression:** It represents relation between dependent variable y and multiple independent variables

$Y=\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_pX_p$

**Assumptions of linear regression:**

1. X and Y have linear relationship
2. Error terms are *normally distributed* with mean zero
3. Error terms are *independent* of each other
4. Error terms have *constant variance* (homoscedasticity)

Best fit straight line is calculated by minimizing the least square errors.

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

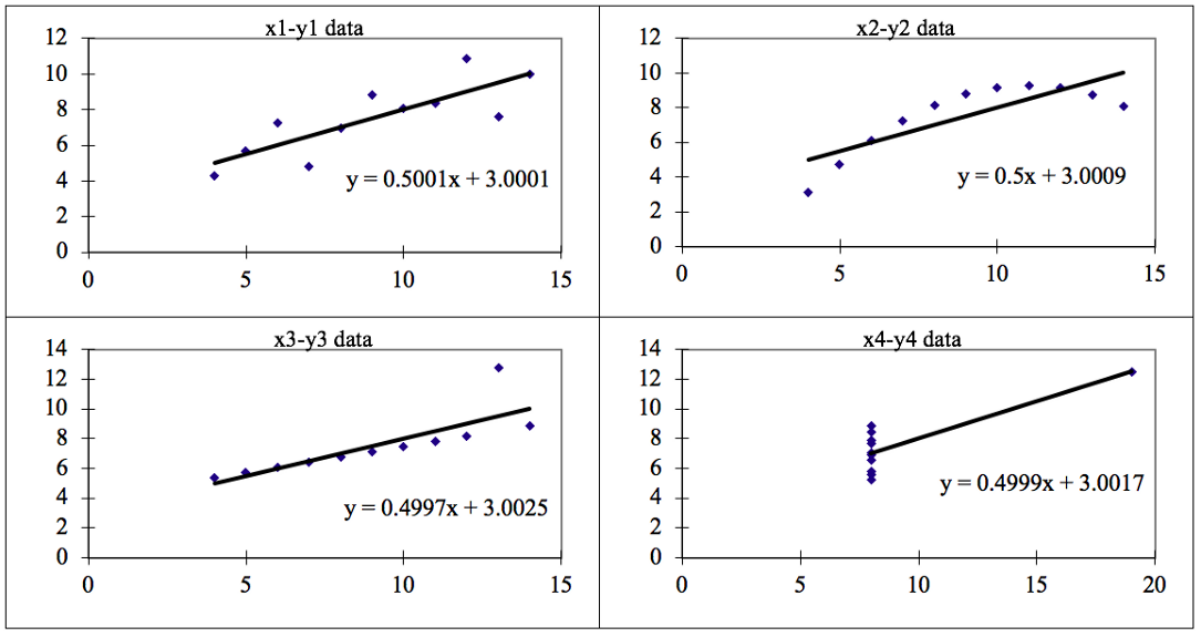
Ans: Anscombe's quartet explains importance of data visualisation and how any regression algorithm can be fooled by the same.

It was stated by Francis Anscombe to illustrate the importance of plotting the graphs which has to be done before analyzing and model building, and the effect of other observations on statistical properties.

Anscombe's Quartet can be defined as a group of four data sets which are identical in descriptive statistics but have very different distributions which is proved clearly when plotted on scatter plots.

| 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 | |

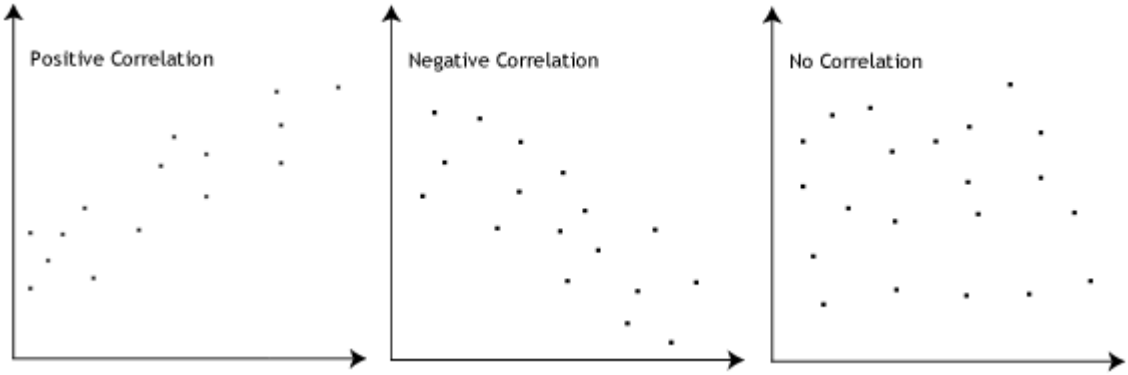
The statistical information for all these four datasets are approximately similar. When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm.



3. **What is Pearson's R?** **(3 marks)**

**Ans:** Pearson's R is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations.
The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive
- r = -1 means the data is perfectly linear with a negative slope
- r = 0 means there is no linear association




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

**Ans:** In raw dataset obtained data can be of wide range. ML algorithms wont work effectively with different scale of data. Therefore, Scaling is a method used to generalize the range of independent variables It is performed after test train split as test data should be hidden from train data. Second advantage of scaling is that gradient descent converges much faster with feature scaling than without it.

**Methods for Scaling**

1. **Normalization**

It is known as min max scaling , it is the simplest method and consists of rescaling the range of features to scale the range in [0, 1]. The general formula is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Here, max(x) and min(x) are the maximum and the minimum values of the feature respectively.

2. **Standardization:**

In standardization scaling the range of rata is converted to data with 0 mean and unit standard deviation. The formula is given as:

$$x' = \frac{x - \bar{x}}{\sigma}$$

Here, σ is the standard deviation of the feature vector, and x̄ is the average of the feature vector.

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

**Ans:** VIF is given by **1/1-R^2**

VIF happens to be infinite when R2=1, R2 is 1 when a variable is perfectly correlated with other.

So VIF = 1/(1-1) which is infinite.

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

Ans:

A Q-Q plot is a plot of the percentiles of the first data set against the percentile of the second data set..

There will be a 45-degree reference line plotted.

If points in plot fall along this reference line then 2 datasets come from same population.

If points deviate from reference line then we conclude that the two data sets have come from populations with different distributions.