# 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)**

There were 6 categorical variables in the dataset. The inference that we could derive were:

* 32% of the booking were happening in season3. In season2 and season4 with 27% and 25% .So season can be a good predictor for the dependent variable
* 10% booking were happening in the months 5,6,7,8,9.So mnth can be a good predictor for the dependent variable ‘cnt’
* 67% booking were happening during weathersit1 and 30% in weahersit2.So weathersit is also a good predictor for the dependent variable
* 69% of the booking were happening in workingday .So it is a good predictor.
* The variable weekday shows 13% of the booking. So this may not have close linear relationship between dependent variable.

1. **Why is it important to use drop\_first = True during dummy variable creation? (2 mark)** 
   1. We create a dummy variable by passing drop\_first=True since to avoid multicollinearity
   2. If we not provide drop\_first=True then n dummy variables will be created and these n variables are themselves correlated which is known as multicollinearity.
2. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

temp and atemp

1. **How did you validate the assumptions of Linear Regression after building the model on the**

**training set? (3 marks)**

* 1. To validate assumptions we need to check errors are normally distributed with mean zero. We need to perform Residual Analysis of Training Data.
  2. Linearity of the data
  3. X are Independent and observed with Negligible Error
  4. Residual Errors are independent from each other and predictors.

1. **Based on the final model, which are the top 3 features contributing significantly towards**

**explaining the demand of the shared bikes? (2 marks)**

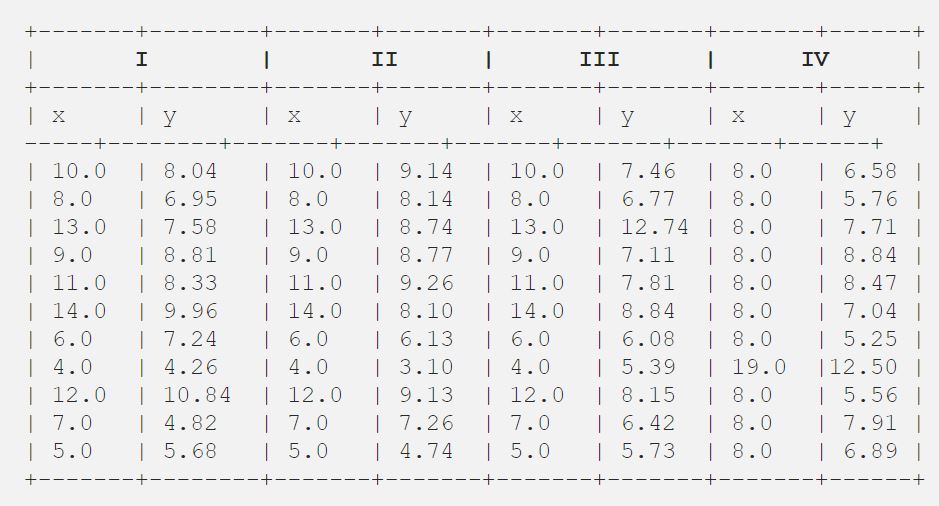
As per the final model the top 3 features contributing significantly towards explaining the demand of the shared bikes are

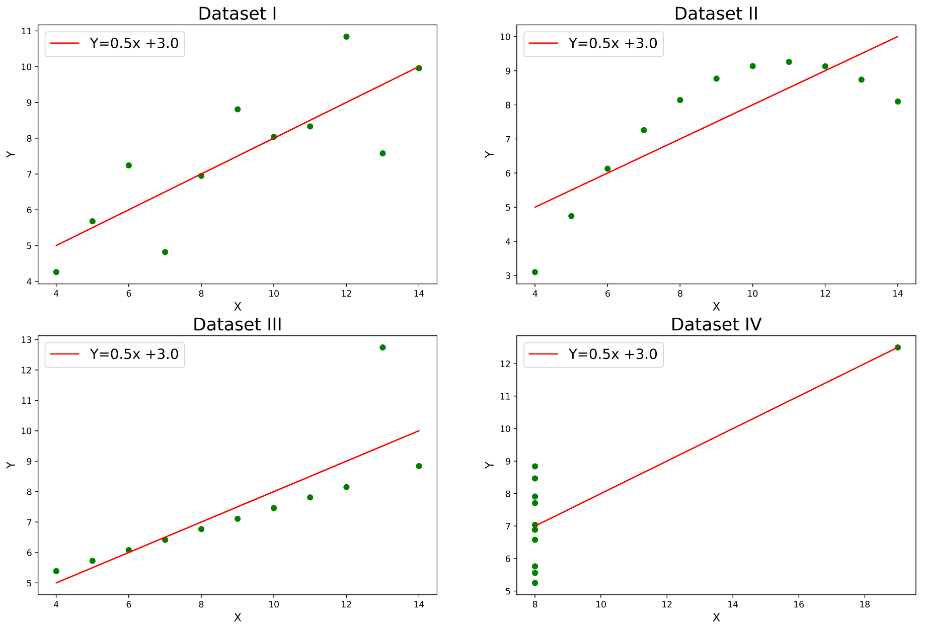
1. Temperature(temp): Increase in the temperature variable increases the bike hire numbers
2. Weather Situation 3(weathersit\_3): A unit increase in weathersit\_3 variable decreases the bike hire numbers.
3. Year(yr): A unit increase in year variable also increases the bike hire numbers.

# General Subjective Questions

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

* Linear Regression is a statistical model. It analyses the linear relationship between a dependent variable with given set of independent variables.
* Linear relationship between variables means when the value of one or more independent variables will change then the value of dependent variable will also change accordingly.
* Linear relation between dependent and independent variables is given by
  + Y = mX + b
  + Y is the dependent variable.
  + X is the independent variable.
  + m is the slop of the regression line which effect X has on Y
  + b is a constant and also known as intercept. If X=0 , Y would be equal to b
* A linear relationship will be called positive if both variables increases.
* A linear relationship will be called negative if independent increases and dependent variable decreases.
* Linear regression model assumes that there is very little or no multi-collinearity in the data.

1. **Explain the Anscombe’s quartet in detail. (3 marks)** 
   1. Anscombe’s Quartet is the model to demonstrate the importance of data visualization was developed by the statistician Francis Anscombe in 1973 to signify the importance of plotting data before analysing it with statistical properties.
   2. It includes four data-set and each data-set consists of eleven (x,y) points.
   3. They share all the same descriptive statistics but different graphical representation.
   4. Data-sets which are identical over a number of statistical properties but produce dissimilar graphs.
   5. The four data set of Anscombe’s quartet as shown below
   6. The scatter plot of each data-set as shown below



* 1. Even for similar statistical properties they appear different very different when they graphed

1. **What is Pearson’s R?(3 marks)**

* The Pearson correlation coefficient ( R ) is a 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.
* If the Pearson correlation coefficient is between 0 and 1 then it is of type Positive correlation. So when one variable changes the other variable changes in the same direction.
* If the Pearson correlation coefficient is 0 then it is considered as no correlation. There is no relationship between the variables.
* If the Pearson correlation coefficient is between 0 and -1 then it is considered as Negative correlation. So when one variable changes the other variable changes in the opposite direction
* It is descriptive statistics. It summarizes the characteristics of a dataset.
* The relationship strength also vary between based on the coefficient value
* The fallowing table shows strength of the relationship based on Pearson correlation coefficient ( R ) value

| Pearson correlation coefficient (r) value | Strength | Direction |
| --- | --- | --- |
| Greater than .5 | Strong | Positive |
| Between .3 and .5 | Moderate | Positive |
| Between 0 and .3 | Weak | Positive |
| 0 | None | None |
| Between 0 and –.3 | Weak | Negative |
| Between –.3 and –.5 | Moderate | Negative |
| Less than –.5 | Strong | Negative |

1. **4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**
   1. **WHAT** 
      1. Scaling is a technique to standardize the independent features present in the data in a fixed range.
      2. It is performed during the data pre-processing to handle highly varying magnitude or values
   2. **WHY**
      1. The machine learning models results in greater values in case if we miss scaling.
      2. Scaling guarantees that all features are on a comparable scale and have comparable ranges. This is called normalization
      3. When the features are scaled several machine learning methods perform better or converge very quickly.
      4. Numerical instability can be prevented by avoiding significant scale disparities between features.
   3. **NORMALIZED SCALING**
      1. Normalization is a data pre-processing technique used to adjust the values of features in a dataset to a common scale.
      2. This is done to facilitate data analysis and modelling to reduce the impact of different scales on the accuracy of machine learning models.
      3. Here values are shifted and rescaled so that they end up ranging between 0 and 1.
   4. **STANDARDIZED SCALING**
      1. Standardization is another scaling method where the values are centred around the mean with a unit standard deviation.
      2. This means that the mean of the attribute become zero and the resultant distribution has a unit standard deviation

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

* If there is perfect correlation then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.
* VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity.
* To determine VIF we fit a regression model between the independent variables.
* If the VIF is 4 ,this means that the variance of the model coefficient is inflated by a factor 4 due to the presence of multicollinearity.

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

* Q-Q plot is the way to test the distribution of continuous variables graphically.
* It is a graphical tool to help us assess if a set of data came from some theoretical distribution such as a Normal, exponential or Uniform distribution.
* Q-Q plot or Quantile-Quantile plot is a scatter plot created by plotting 2 different quantiles against each other.
* A Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
* The first quantile is that of the variable you are testing the hypothesis and the second one is the actual distribution of against.
* **STEPS TO GENERATE Q-Q plot**
  + Take your variable of interest and sort it from smallest to largest value.
  + Take a normal curve and divide it into 20 equal segments.
  + Compute z score for each of their points.
  + Plot the z-score obtained against the sorted variables.
  + Observe if data points align closely in a straight 45-degree line.