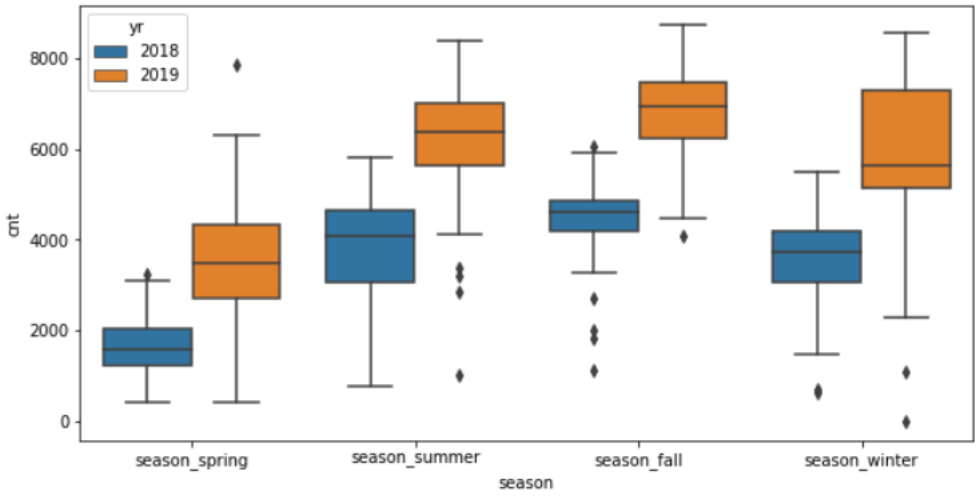
Assignment-based Subjective Questions

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

   The categorical data present in the dataset are season, year, month, weekday, working day and weather situation.

   Based upon Data visualisation we can infer the below:

   - *Year 2019 has more count.*
   - *Seasons summer and fall followed by winter has more counts. In spring season, the count is less.*
   - *Months May to October has more count compared to other months.*
   - *Working day and weekday doesn't have much impact on the count.*
   - *Clear and misty weather has more count.*



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

   drop_first=*True reduces the correlations created among dummy variables*.

   drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation.

   Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then it is obviously unfurnished. We require only two columns to represent 3 categories.

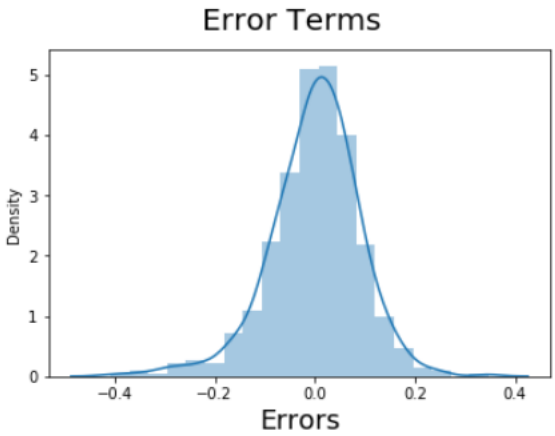
   Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

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

   *The columns **temp(0.64)** and **atemp** has same highest correlation among numerical variables.*

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

   *The Linear Regression Assumptions are validated by residual analysis of train data. Plotting the **error terms**:*



   ***Multicollinearity** check: There should be insignificant multicollinearity among variables.*
   ***Linear relationship validation**: Linearity should be visible among variables*
   ***Homoscedasticity**: There should be no visible pattern in residual values.*
   ***Independence of residuals**: No auto-correlation*

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

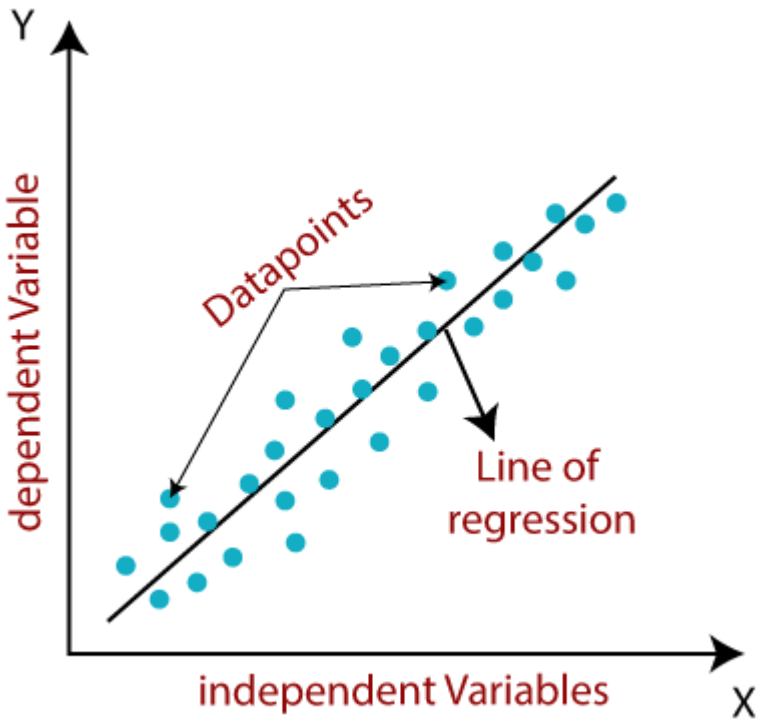
   *Based on the final model the top 3 features contributing significantly towards explaining the demand of shared bikes are: **Temperature, year and weather situation** (negatively correlated).*

   However year feature may be related to other condition of how the pandemic situation and people conditions are in general.

General Subjective Questions

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

Linear Regression is a machine learning algorithm based on supervised learning. Linear regression is understanding the relationship between one dependent variable and several independent variables. It is used to predict the value of a variable based on value of one or more variables.



The assumptions of simple linear regression are:
1. There must be Linear relationship between X and Y
2. Error terms are normally distributed (not X, Y)
3. Error terms are independent of each other
4. Error terms have constant variance (homoscedasticity)

Types of Linear Regression:
1. *Simple Linear Regression- Model with only one independent variable*
2. *Multiple Linear Regression- Model with more than 1 independent variable.*

Equation of straight line: $y = mX+c$*, where c is constant, Y is predictor variable, m is slope, X is independent variable.*

Every liner regression model has cost function which has to be optimized.

Cost Function(J): By achieving the best fit line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So it is

important to reach the best value that minimize the error between predicted y value and the true value.

Cost function of linear regression is Root Mean Squared Error (RMSE) between predicted y value and true y.

There are two types of optimizations: Constrained and Unconstrained.

Unconstrained optimization is further divided into differentiation and gradient descent(iterative)

Gradient descent start with random b1 and b2 values and then iteratively updating values, reaching the minimum cost.

After fitting straight line on the data, we have to do hypothesis testing on it to find if the straight line is significant or not.

Null hypothesis: H0: B1 = 0

Alternate hypothesis: H1: B1 not equal to 0.

The parameters used to assess a model are:

1. *t statistics*
2. *F statistics*
3. *R squared*

R-squared is a statistical method the determines the goodness of fit. It measures the strength of the relationship between the dependent and independent variable on scale of 0-100%

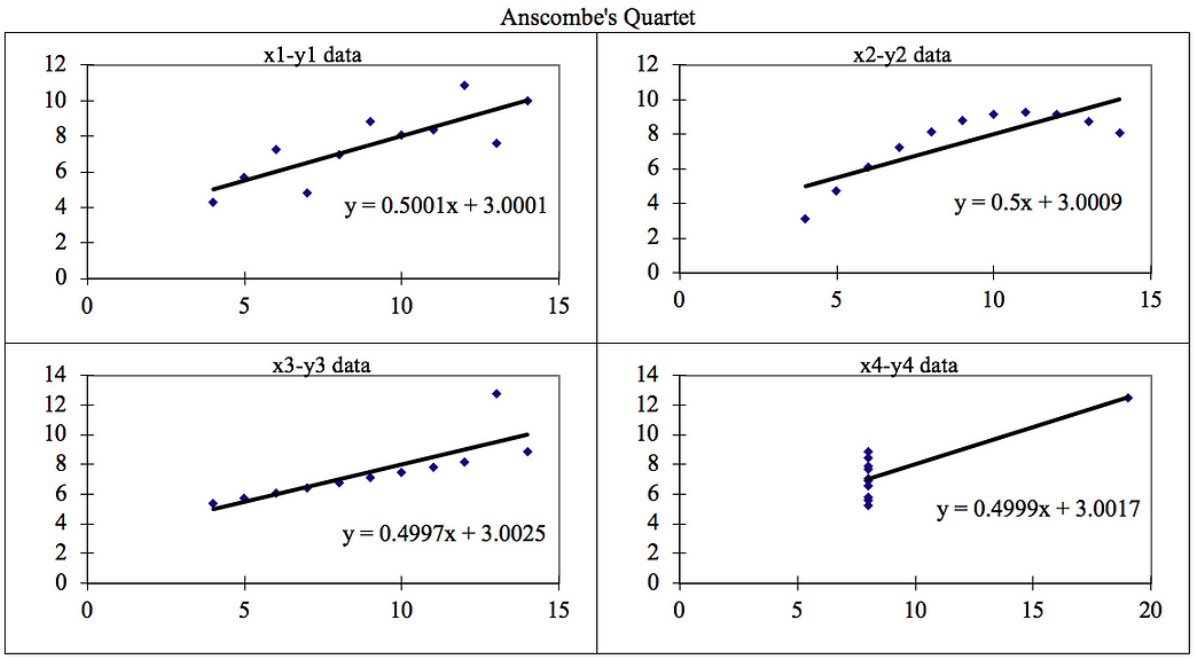
R-Squared = Explained variation/Total variation.

While selecting features in multiple linear regression can you VIF. VIF is Variance Inflation Factor explains the relationship between one independent variable with other independent variables.

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

Anscombe's quartet comprises four datasets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. It was constructed in 1973 by statistician Fancis Anscombe to illustrate importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties. There are these 4 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 in order to see the distribution of samples that can help us identify the various anomalies present in the data like outliers, diversity of data, linear separability of data,etc.

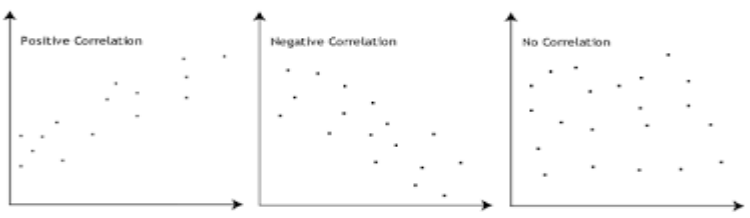


The four datasets which are considered here can be described as:

1. *Dataset 1: this fits the linear regression model pretty well.*
2. *Dataset 2: this could not fit linear regression model on the data quite well as 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? (3 marks)

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative. The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



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

Scaling is bringing different features to same scale.

When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So, we need to scale features because of two reasons:

1. Ease of interpretation

2. Faster convergence for gradient descent methods

Standardized Scaling: The variables are scaled in such a way that their mean is zero and standard deviation is one.

X = x-mean(x)/standard deviation of x

MinMax 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.

X = x-min(x)/max(x)- min(x)

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

If there is perfect correlation between all independent variables then VIF is infinity.

A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity.

To solve VIF infinity we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

In our bike sharing example temp and a temp are highly correlated and may cause VIF to be infinite.

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

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.
***Use of Q-Q plot**: A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data*

fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

***Importance of Q-Q plot**: When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.*