## Assignment-based Subjective Questions

**Question 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 :** We have following categorical in our final model
**1. Season** : Winter help to increase the deamand while Spring impactively negatively
**2. Month**: In September demand increases while in Jan, July, Nov and Dec demand decreases
**3. Year :** Demand in 2018 is less than in 2019. Which is showing a increased demand. So we company should increase bikes stocks for next year so that customer demand of bikes can be fullfilled
**4. Weathersit:** Whether situation Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds **and** Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist decreasing my sales.

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

**Answer :** It helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

For Ex: We creating dummy variable for season column. Now as per data set we will have four types of dummy varriables(Columns in Data Set), 1. Spring, 2. Summer, 3. Winter and 4. Fall. Using these dummy variable we show season such as for Winter(0010) and Summer(0100). Each character denote columns. But we can show this status using three variables also such as

Spring 000
Summer 100
Winter 010
Fall 001

Thats why we do not required lets consider Summer column but can denote theri value.

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

**Answer :** After seeing the pair plot of all numeric variables(atemp, windspeed, hum, cnt), **atemp** have the highest correlation with **cnt(target variable).**

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

**Answer :** In last we can compare our final model with our initial pair plot and boxplot. We can see that **in pair plot atemp have highest correlation with target(cnt) variable**. When we see **boxplot season, yr, whethersit have a good correaltion with target(cnt) variable**.

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

**Answer : Positively Significant:** atemp(coffe: 0.4109) , yr_2019 (coffe: 0.2361), X Light Snow(coffe: -0.2881)

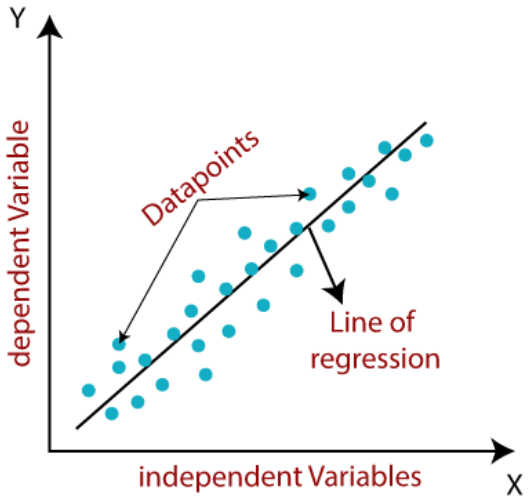
## General Subjective Questions

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

**Answer :** Linear regression is one of the statistical method used for predictive analysis. Linear regression make predictions for categorical(continous/real) or numeric variables such **season, wethersituation, temprature, month etc.**

Linear regression algorithm shows a linear relationship between a dependent variable(y) and one or more dependent variables(X), hence called linear regression which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

We represent linerar regression using $y=B_0+B_1x+B_2x+.....+B_ix+\varepsilon,$ below is the referenced image to understand.




Here,

y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
$B_0$= intercept of the line (Gives an additional degree of freedom)
$B_i$ = Linear regression coefficient (scale factor to each input value).
ε = random error

The values for x and y variables are training datasets for Linear Regression model representation.

If there are only one indepent variable then its called **Simple Linear Regression($y=B_0+B_1x+\varepsilon$)**, but independent variables are more than 1 then its called **Multiple Linear Regression($y=B_0+B_1x+B_2x+.....+B_ix+\varepsilon$).**

When working with linear regression, our main goal is to find the best fit line that means errors between predicted and actual value should be minimized. Our best fit line should have least error. The best-fit line is obtained by minimising a quantity called **Residual Sum of Squares (RSS)**

RSS = $e1^2+e2^2+e3^2+e4^2+e5^2+.....++eN^2$

Where $\mathbf{eN=y_{actual}-Y_{predicted}}$

$\mathbf{Y_{predicted}=B_0+B_1x}$

$\mathbf{RSS=(y_1-B_0-B_1x)^2+(y_2-B_0-B_2x)^2+(y_3-B_0-B_3x)^2+.....+(y_N-B_0-B_Nx)^2}$

**RSS=**

**Residuals:** The distance between the actual value and predicted values is called residual. If the observed points are far from the regression line, then the residual will be high, and so cost function will high. If the scatter points are close to the regression line, then the residual will be small and hence the cost function.

**Gradient Function:** Gradient descent is used to minimize the MSE by calculating the gradient of the cost function. It is done by a random selection of values of coefficient and then iteratively update the values to reach the minimum cost function.

**Model Performance:** The process of finding the best model out of various models is called optimization. It can be achieved by below method:

1**. Prob(F-statistics):** It denotes the overall fitness of model. It should be less then 0.05(5 Significance level)

**2. Adjusted R-Square:** It denotes the goodness of fit of overall models. Its beasically calculated using R-Square using following formula $1/(1-R^2)$ .
**3. P-Value :** P-Value of all independent variables should be less then 0.05. If its higher than drop that variable and do modeling again than check OLS summery again.
**4. VIF(Variance Inflaction Factor):** Check VIF value for all variables and for this assignment I take number which <=5 for good fit. VIF is basically show the multicollinearity between variables.

## Assumptions of Linear Regression:

**1. No multicollinearity between the features:** Multicollinearity means high-correlation between the independent variables. Due to multicollinearity, it may difficult to find the true relationship between the predictors and target variables. It should not be >5 (or >10 in some business cases). If its greater than 5 for any variable then our model become overfit.

**2. Homoscedasticity Assumption:** Homoscedasticity is a situation when the error term is the same for all the values of independent variables. With homoscedasticity, there should be no clear pattern distribution of data in the scatter plot.

**3. Normal distribution of error terms:** If error terms are not normally distributed, then confidence intervals will become either too wide or too narrow, which may cause difficulties in finding coefficients.

**4. No autocorrelations:** The linear regression model assumes no autocorrelation in error terms. If there will be any correlation in the error term, then it will drastically reduce the accuracy of the model.

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

**Answer: Anscombe's quartet** a statistician of great repute always give a importance for visualizations in favor of summary statistics. But People ignore that and show interest in only statistical information like mean, standar deviation etc.

So **Anscombe's quartet** comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

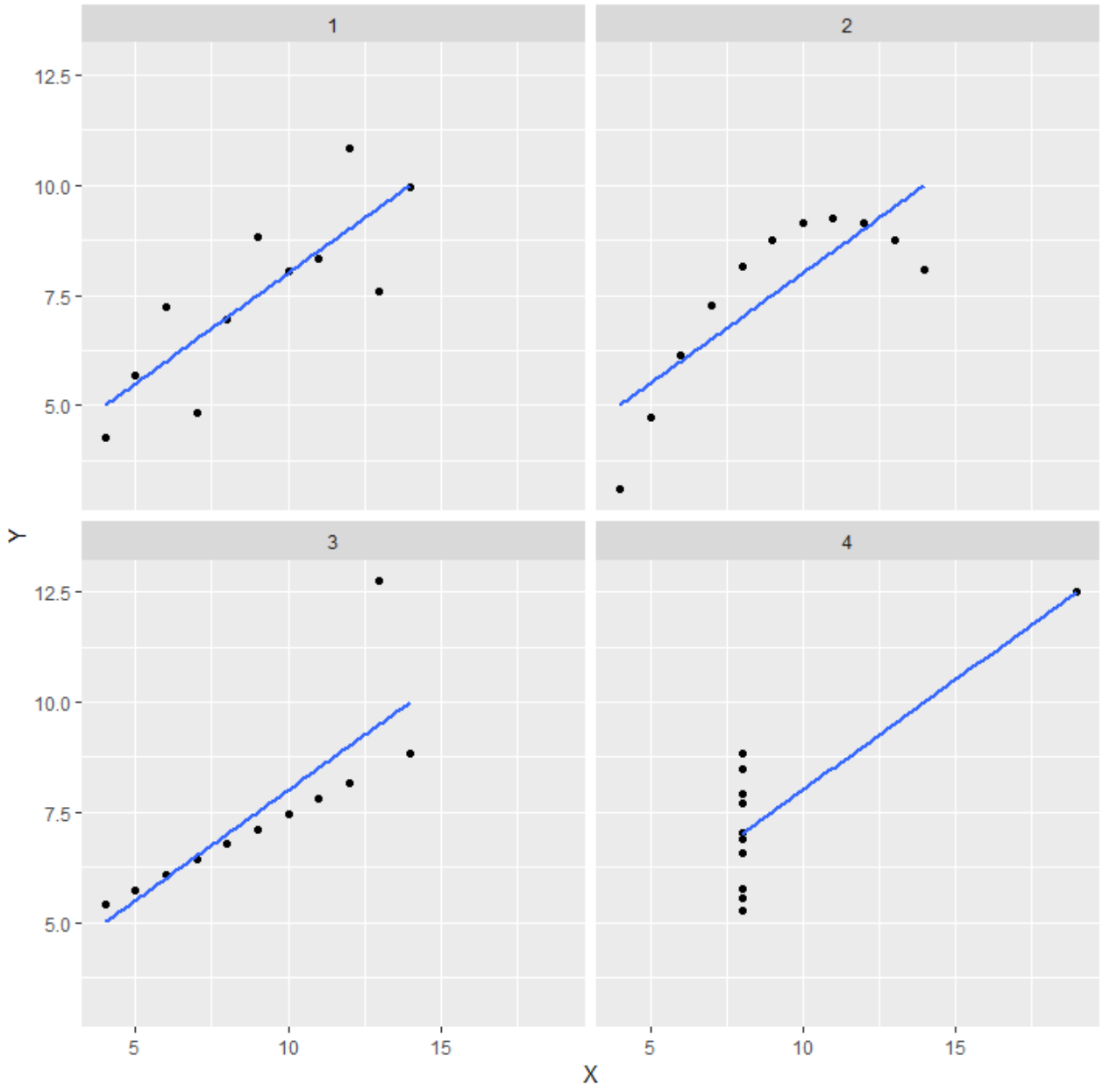
| I | | II | | III | | IV | |
|---|---|---|---|---|---|---|---|
| x | y | x | y | x | y | x | y |
| 10.0 | 8.04 | 10.0 | 9.14 | 10.0 | 7.46 | 8.0 | 6.58 |
| 8.0 | 6.95 | 8.0 | 8.14 | 8.0 | 6.77 | 8.0 | 5.76 |
| 13.0 | 7.58 | 13.0 | 8.74 | 13.0 | 12.74 | 8.0 | 7.71 |
| 9.0 | 8.81 | 9.0 | 8.77 | 9.0 | 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 |

## Statistical Summery of above dataset

**Summary**

| Set | mean(X) | sd(X) | mean(Y) | sd(Y) | cor(X,Y) |
|---|---|---|---|---|---|
| 1 | 9 | 3.32 | 7.5 | 2.03 | 0.816 |
| 2 | 9 | 3.32 | 7.5 | 2.03 | 0.816 |
| 3 | 9 | 3.32 | 7.5 | 2.03 | 0.816 |
| 4 | 9 | 3.32 | 7.5 | 2.03 | 0.817 |

## Grahpical Representation of data



Note: Graphs were completely different even though the summary was exactly similar

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

**Answer :**

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

**Answer :**

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

**Answer :**

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

**Answer :**