**Assignment**: Bike Sharing Assignment Subjective Questions
**Student Name**: Kajal Kaspate
**Batch ID**: C61

## 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? | 1. Categorical variables based on season and month have impact on the target variable. These are important variables from business point of view. For users to rent the bike, weather, temperature and season should be conducive. We see 'cnt' is highly correlated with 'Summer'.<br>2. Also, we see that target variable is correlated with the months. This is understandable as temperature patterns can be generally predicted by months.<br>3. We also see 'weather' variable affecting bike renting. We see that there were no bikes rented when the weather was 'thunderstorms or hailstorms'. We do see that when the weather is 'wet' or 'misty', it affects 'cnt' negatively.<br>4. In the final model, we see that coefficient for 'yr' is 0.23 which means 'cnt' will increase by 0.23 units for every unit increase in 'yr'<br>5. Coefficient for weather 'wet' is -0.29 that means, this causes the most reduction to 'cnt' per unit increase in 'wet' |
| 2 | Why is it important to use **drop_first=True** during dummy variable creation? | 1. When we use Pandas function to get the dummies, for N variables, we do get N number of dummies.<br>2. However, because dummy variables use 0s and 1s to store the information in the columns, it is easier to use toggle method and know if the value of a variable is 0, one of the other variables must be 1.<br>3. In case of a variable which takes 2 values, for example 'Educated' and 'Not Educated', when the column 'Educated' is 0, 'Not Educated' is 1 and vice versa. If we keep one column as 'Educated', 0 will mean that 'Not educated' is 1 without having a need for storing this second column.<br>4. By extrapolating this to N variables, we can store the same amount of information using N-1 variables.<br>5. Having one less column to store the same amount of information helps in memory management during runtime.<br>6. Also, during the model building, potential for multicollinearity is reduced slightly. |
| 3 | Looking at the pair-plot among | 1. We see that 'cnt' is highly correlated to 'temp', 'atemp', 'casual' and 'registered' variables. |

| | | |
|---|---|---|
| | the numerical variables, which one has the highest correlation with the target variable? | 2. There is a mathematical relation between ‘cnt’ and ‘registered’ and ‘casual’. (cnt = casual+registered). Hence these two columns are redundant.<br>3. Also, there is almost perfect relation between ‘temp’ and ‘atemp’ hence we can use one of them to predict ‘cnt’ |
| 4 | How did you validate the assumptions of Linear Regression after building the model on the training set? | 1. I plotted histogram for error terms and looked if their error term mean is near 0.<br>2. Also, I plotted the error terms on to a graph to see if there is no predictable pattern and if they are completely random.<br>3. This also ensured that the variance is consistent across error terms which is referred to as homoscedastic. |
| 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, I see below features contribute the most towards explaining demand of the shared bike<br>1. Temperature – Coefficient for the temperature is 0.45 and this is the max for all the predictors combined.  Correlation coefficient for this is 0.63 based on corr function. This can explain the increase in bike rentals as the higher temperatures could drive people to go out much and enjoy the bike rides or also rides to parks etc on holidays.<br>2. Weather – Second most important predictor is weather. If the weather is bad with thunderstorms, no bikes are rented as we see from the data set. Also, from data visualization we saw that ‘Dry’ weather accounted for more bike rentals. In the final model, we have negative 0.29 as coefficient for ‘Wet’ weather. This is the lowest coefficient we have amongst all predictors. This could also mean, when there is light rain/snow – count of rented bikes falls. So, when the weather is Dry, we should have increase in the bikes rented.<br>3. Year – year has a coefficient as 0.23. This is significant in saying that demand for bike sharing is going up from the time it was introduced. |

# General Subjective questions:

## 1. Explain the linear regression algorithm in detail.

*Overview*:

Regression is a statistical method of creating a model (or mathematical equation) to predict target variable from set of independent variables using a data set at hand. This regression model then can be used to predict future values for the target variable given the values for the predictor variables.

Linear Regression is a type of regression model which establishes linear relation between target variable and the predictor variables. Linear relation is where all the data points can be plotted on a straight line.

*How it works:*

A linear relation described by a straight line can be represented by a mathematical equation of the format like below.

$$y=mx+c$$

Here, y is the predictor variable. 'm' is the slope of the equation and 'c' is the constant. Depending on the various values of x, we can calculate the values of y.

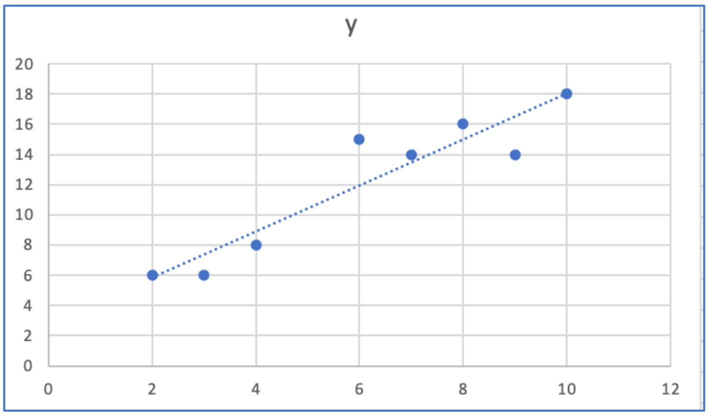
Let's take example of below 8 data points and assume, that the linear equation between x and y is represented by y=mx+c

| x | y |
|---|---|
| 3 | 6 |
| 2 | 6 |
| 4 | 8 |
| 9 | 14 |
| 6 | 15 |
| 7 | 14 |
| 8 | 16 |
| 10 | 18 |

These 8 data points can be calculated as below –

| Y - actual |
|---|
| $y_1 = mx_1 + c$ |
| $y_2 = mx_2 + c$ |
| $y_3 = mx_3 + c$ |
| $y_4 = mx_4 + c$ |
| $y_5 = mx_5 + c$ |
| $y_6 = mx_6 + c$ |
| $y_7= mx_7 + c$ |
| $y_8 = mx_8 + c$ |

Below is the scatter graph for the same data points. Here, I have tried to find a linear relation between x and y by plotting a straight line. Not all data points lie on the line, however, the error introduced in predicting the values of y is minimised.



Error terms are difference between actual value of y and the predicted value of the y because we tried to fit them on the line.

| Y – actual | Error term for the predicted y for the best fitted line |
|---|---|
| $y_1 = mx_1 + c$ | $y_1 - y_{pred1}$ |
| $y_2 = mx_2 + c$ | $y_2 - y_{pred2}$ |
| $y_3 = mx_3 + c$ | $y_3 - y_{pred3}$ |
| $y_4 = mx_4 + c$ | $y_4 - y_{pred4}$ |
| $y_5 = mx_5 + c$ | $y_5 - y_{pred5}$ |
| $y_6 = mx_6 + c$ | $y_6 - y_{pred6}$ |
| $y_7= mx_7 + c$ | $y_7 - y_{pred7}$ |
| $y_8 = mx_8 + c$ | $y_8 - y_{pred8}$ |

With linear regression algorithm, we simply try to find a best-fit line where the error terms are minimised. Also, we want to measure the magnitude of the error terms and not the sign, hence we square these differences. The value that we want to minimise is also called as 'Residual Sum of Squares (RSS)'. This can be represented as below.

$\sum_{i=1}^{n} RSS = y_i - y_{\text{pred } i}$

## *Types of Linear Regression:*

Above example only had one single predictor which is 'x'. Such linear regression models are called 'Simple Linear Regression'(SLR) model. Where there more than one predictor variables present, it is called as 'Multiple Linear Regression' (MLR) models.

## *Constraints for linear regression model:*

1. To find the best fit line, we have to make sure that error terms are normally distributed with mean zero.
2. Error terms are independent of each other and there is no visible pattern in them.
3. Error terms have constant variance which is also referred to as 'homoscedasticity)

## *Examples of where this can be applied:*

Linear regression is used when the predictor variable is continuous type. For example -

a. predicting the sales based on the marketing budget,

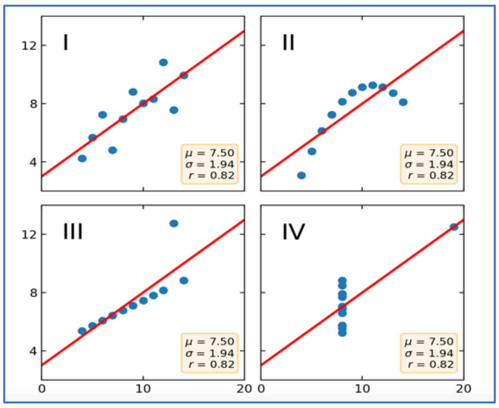
b. predicting weight based on height and age.

## 2. Explain the Anscombe’s quartet in detail.

Anscombe quartet is a group of datasets which have identical statistics like same mean, standard deviation, regression line but they are qualitatively different as shown in below graph. Each dataset consists of 11 points, and they were constructed by statistician Anscombe to demonstrate importance of graphing the data and the effect of outliers on the statistical properties.

Datasets are as below –

| Data set I | | Data set II | | Data set III | | Data set IV | |
|---|---|---|---|---|---|---|---|
| *x* | *y* | *x* | *y* | *x* | *y* | *x* | *y* |
| 10 | 8.04 | 10 | 9.14 | 10 | 7.46 | 8 | 6.58 |
| 8 | 6.95 | 8 | 8.14 | 8 | 6.77 | 8 | 5.76 |
| 13 | 7.58 | 13 | 8.74 | 13 | 12.74 | 8 | 7.71 |
| 9 | 8.81 | 9 | 8.77 | 9 | 7.11 | 8 | 8.84 |
| 11 | 8.33 | 11 | 9.26 | 11 | 7.81 | 8 | 8.47 |
| 14 | 9.96 | 14 | 8.1 | 14 | 8.84 | 8 | 7.04 |
| 6 | 7.24 | 6 | 6.13 | 6 | 6.08 | 8 | 5.25 |
| 4 | 4.26 | 4 | 3.1 | 4 | 5.39 | 19 | 12.5 |
| 12 | 10.84 | 12 | 9.13 | 12 | 8.15 | 8 | 5.56 |
| 7 | 4.82 | 7 | 7.26 | 7 | 6.42 | 8 | 7.91 |
| 5 | 5.68 | 5 | 4.74 | 5 | 5.73 | 8 | 6.89 |



*Reference: https://matplotlib.org/stable/gallery/specialty_plots/anscombe.html*

Below are the statistical values for data sets.

| Property | Value |
|---|---|
| Mean of x | 9 |
| Sample variance of x | 11 |
| mean of y | 7.5 |
| Sample variance of y | 4.125 |
| Correlation beween x and y | 0.816 |
| Linear regression line | *y* = 3.00 + 0.500*x* |
| R-squared | 0.67 |

## 3. What is Pearson's R?

Pearson's R is a coefficient that is used to measure correlation between two variables. Variance for one variable is a measure of how much that variable varies away from its mean. Extending this idea to two variables makes it co-variance and it tells how two variables change together. Also, co-variance reflects the directional relationship between random variables, but not the magnitude of the relationship. However, covariance changes if the units of the variables are changed. To address this, Covariance coefficient is introduced.

If we have data set for two variables - x and y,
Covariance for them will be –

$$\sigma_{xy} = \frac{\sum(x_i - \mu_x)(y_i - \mu_y)}{N}$$

Because covariance cannot address the problem with the scaling of the variables, below is the correlation coefficient.

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\frac{\sum(x_i - \mu_x)(y_i - \mu_y)}{N}}{\sigma_x \sigma_y}$$

Pearson's coefficient can be used if one of the following conditions are true –

1. Both the variables are quantitative
2. Variables are normally distributed or close to normally distributed.
3. There are no outliers as presence of outliers may significantly skew data and the resulting Pearson coefficient will not accurately reflect the correlation of the two variables.
4. This can be used when the relation between two variables is linear. This is best for data sets that with a reasonably straight trend line.

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

### *What is scaling*

Scaling is a data transformation technique where we transform a feature to fit in a particular range of values. e.g., between 0 and 1 or a range where mean is 0 for all the features. If the range is used as 0 and 1, the min and max is changed to 0 and 1 and all the values are fitted in that range.

### *Why is scaling performed*

If your data set contains predictors with different scales, after creating a model, coefficients of these predictors will not be comparable. E.g., in the housing data set, if the area of the house is in sq. fts and the number of bedrooms is one digit number, after creating the model, co-efficient ranges will vary a lot and it may not be possible to explain business how predictors are affecting the target variable.
Some algorithms that use Gradient Descent as an optimisation method require that data is scaled. Step size in the optimisation function will vary for all the features if they use different scales and this will make algorithm inefficient.

### *Difference between standardized scaling and normalized scaling*

Normalized scaling is where all the features are scaled to a specific range e.g. 0-1 or 0-100.
Normalization equation is as below -

$$\frac{\mathrm{x - min}}{\mathrm{max - min}}$$

Standardization is rescaling data to have mean as 0 and standard deviation is 1.
Equation for standardisation is –

$$\frac{\text{x - mean}}{\text{standard deviation}}$$

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

Below is the formula for VIF

$$\mathrm{VIF_i} = \frac{1}{1 - \mathrm{R_i^2}}$$

$\mathrm{R_i^2}$ will reach 1 when there is perfect correlation of *i' th* variable with all other predictor variables excluding the target variable. Mathematically, VIF fraction will be infinity when such relation exists. This is the case of severe multicollinearity, and we will have to review the model to drop the variables that are causing this.

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

### *What is it*

Q-Q plot stands for the Quantile-Quantile plot. If we have two data sets, quantiles from one data set are plotted against quantiles from another data set. It is a form of a scatter plot. Quantiles are nothing but the percentiles and can be calculated at regular intervals (e.g 0.2, 0.4,0.6,0.8,0.9) and QQ plot is formed by plotting quantiles from sorted dataset 1 on vertical axis and plotting quantiles from sorted dataset 2 on horizontal axis.

### *Use and importance of QQ plot*

After plotting the quantiles of the two data sets on the scatter plot, if points seem to fall on the straight line, then it can be said that the data sets follow the same distribution. It is not a foolproof test; hence it can be subjective. However, this can be used as a tool to see the possibility of both the data sets following same kind of distribution e.g., normal or exponential.

QQ plot can be used to answer below questions –

1. Do datasets have similar distribution like shape and type?
2. Do datasets have similar tail behavior?
3. Do both the datasets come from population sets that have same distribution?

### *Importance of QQ plot in linear regression*

1. QQ Plot can be used in linear regression to check if the error terms are normally distributed.
2. We can also use QQ plot to see if the error terms have constant variance (homoscedasticity)