# BIKE RENTAL DEMAND ASSIGNMENT

## 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. The important categorical variables that explain the variation in dependent variable(cnt) with significant coefficient are as follows :

a) Year
b) Holiday
c) Light_snow_rain
d) September

- The effects of these categorical variable on the dependent variable could be explained in terms of following :
- In reference to the base year that 2018 (given by the constant) where in increase in demand was by 0.07, the bike rental demand increased by the 0.232 in year 2019.
- The coefficient of holiday is -0.10 which indicates that demand for rental bikes fall by -0.10 during holidays when compared to the other days.
- So far as the weathersit is concerned the rental bike demand fall by 0.22 during light rain and snow, whereas during clear, pleasant and cloudy weather, the demand increased by 0.07 .
- In the month of September, the rental bike demand improved by 0.08 which is almost same to the reference month i.e April where in the coefficient is 0.07.

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

2. In regression analysis the dependent variable, is not only influenced by the ratio scale variables such as income, output, prices, costsbut also by variables that are essentially qualitative, or nominal scale, in nature, such as sex, race, color. However, we need to be very careful while incorporating the dummy variables. If we add dummy for each category of and also an intercept, then there will be a case of perfect collinearity, that is, exact linear relationships among the variables. In such case, the estimation of model is impossible. Therefore , it is important use drop_first=True during dummy variable creation.

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

3. In our case study the target variable is 'cnt' i.e number of total bike rentals. The pairplot clearly indicated high correlation between temp and the target variable. The results are confirmed by looking at the heatmap

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

4. After building the model on the training set, the assumptions are checked by calculating the VIF values and doing residual analysis.

- The speed with which variances and covariances increase can be seen with the variance-inflating factor (VIF). VIF shows how the variance of an estimator is inflated by the presence of multicollinearity. In our case, we have removed those variables where the value of VIF was greater than 5 and later updated the model.
- Residual analysis shows whether the error terms follow normal distribution or not.
- Residual plot helps us to check whether the relationship between the outcomes and the predictors is linear.

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

5. The top three features that explains the maximum variance in the dependent variable (significant at 1% level ) are as follows :

- Temperature : The coefficient is positive i.e 0.50 which indicates that with the increase in temperature the bike rentals demand increases. Therefore, the company should try to target more customers during high temperature.
- Year : the coefficient is 0.23, which means that in 2019 the demand for bike rentals rises when compared to 2018. Therefore, the company should focus on the trends that took place in 2019 which elevated the bike rental demand.
- Light_snow_rain : In this case the coefficient is -0.22, which means the demand for bikes fall by 0.22 units during light snow and rainy season.

## General subjective questions

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

1. Linear regression is a form of predictive modelling technique (supervised) that helps to explain the variation in dependent (target variable) owing to independent variables (predictors).

The linear regression algorithm is as follows:

1. Reading and Understanding the data by doing data quality check and handing missing values.

2. Understand the data frame by visualization: Here we can interpret the meaningful insights from the data by doing univariate and multivariate analysis. Further, we can use pair plot and heatmap to look at the correlation amongst the given variables . It helps to detect the relationship between dependent and independent variables along with the associations between independent variables.

3. Data preparation : It is important to map the categorial variables and convert them into dummies to run the linear regression model. It is important "quantify" nominal variables by constructing artificial variables that take on values of 1 or 0, 1 indicating the presence (or possession) of that attribute and 0 indicating the absence of that attribute. Such variables are called dummy variables.

4. Splitting the data into training and testing sets : In order to check whether the prediction of our model is true or not , we need to split out data and evaluate the model on testing set.

5. Building a linear regression model that covers the following steps :

Feature selection using RFE

Updating the model by checking p value of coefficients and VIF.

6. Residual Analysis of the train data : Further, it is important to check if the error terms are also normally distributed (which is infact, one of the major assumptions of linear regression) or not.

7.Making predictions using the final model :

Here we need to calculate the coefficients and make prediction about the dependent variable on the testing dataset.

8. Model Evaluation by plotting the error terms

9. The final step is to calculate the r square and mean square error

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

2. Anscombe's Quartet was developed by statistician Francis Anscombe. It constitutes four datasets, containing (x,y) pairs. Although these datasets shared the same statistics, but results completely changed when we visualize the data.

Each graph tells a different story irrespective of their similar summary statistics. This could be explained with the help of following example:

Let say we have following four data sets:

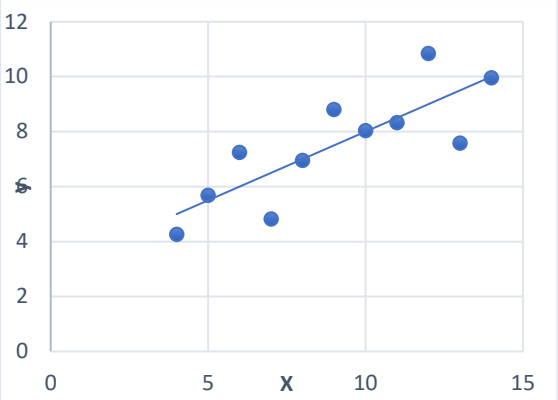
| I | | II | | III | | IV | |
|---|---|---|---|---|---|---|---|
| x | y | x | y | x | y | x | y |
| 10 | 8.04 | 10 | 9.14 | 10 | 7.46 | 8 | 6.58 |
| 8 | 6.95 | 8 | 8.14 | 8 | 6.77 | 8 | 5.76 |
| 13 | 7.58 | 13 | 8.74 | 13 | 12.74 | 8 | 7.71 |
| 9 | 8.81 | 9 | 8.77 | 9 | 7.11 | 8 | 8.84 |
| 11 | 8.33 | 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 |

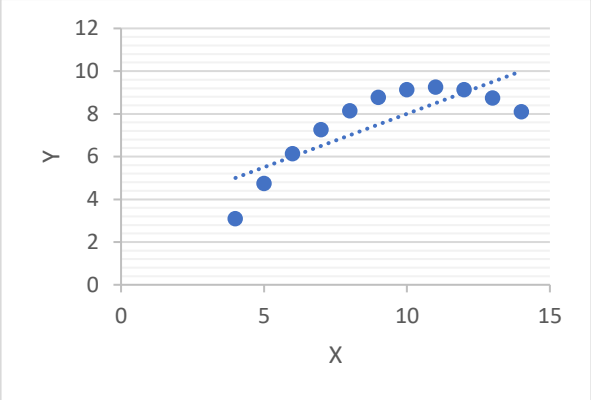
Following is the summary statistics of above metioned datasets:

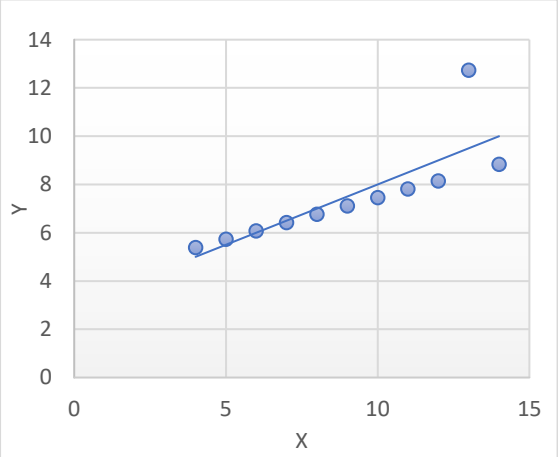
| Set | mean(X) | sd(X) | mean(Y) | sd(Y) | cor(X,Y) |
|---|---|---|---|---|---|
| I | 9 | 3.32 | 7.5 | 2.03 | 0.816 |
| II | 9 | 3.32 | 7.5 | 2.03 | 0.816 |
| III | 9 | 3.32 | 7.5 | 2.03 | 0.816 |
| IV | 9 | 3.32 | 7.5 | 2.03 | 0.817 |

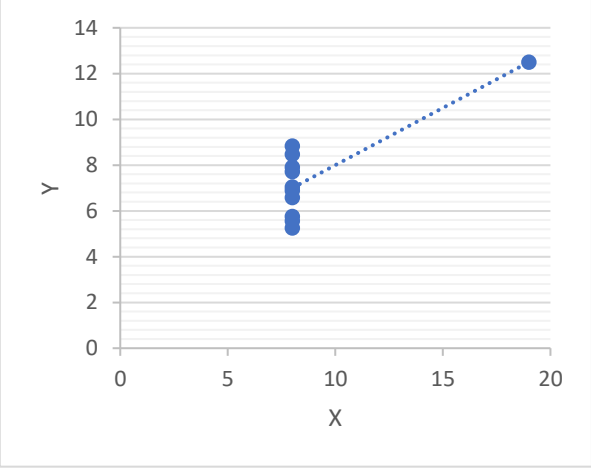
If we look at the summary statistics, all four datasets have same mean, standard deviation and correlation.

Let’s look at these datasets graphically:




- The first scatter plot clearly refers to the simple linear relationship, following the assumption of normality.
- The second graph is not distributed normally. Here the relationship is not linear.
- In the third graph the distribution is linear, but should have a different regression line owing to outlier.
- Finally, the fourth graph shows that one high-leverage point produce a high correlation coefficient, even though the other data points indicates no relationship between the variables.

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

3. Correlation coefficients are used to measure the relationship between two variables. There are different methods to calculate the correlation coefficient such as Pearson, Kendall rank correlation, Spearman correlation, and the Point-Biserial correlation. , but the most popular one is **Pearson's correlation** normally denoted by r. it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship. The value

of r lies between -1 and +1. If r is 1 then tow variables are said yo be perfectly correlated. Positive correlation means that as one variable increases, the other variable tends to also increase and negative correlation means with the increase in one variable, the other variable decreases. It is calculated by the following formula:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

There is certain assumption for Pearson *r* correlation

1. Variables should follow normal distribution
2. the relationship between variables should be linear
3. Homoscedasticity which assumes that data is equally distributed about the regression line

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

4. S**caling** is a method that is used to normalize the range of both dependent and independent variables. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step.

Scaling is required when the features in our dataset vary highly in magnitudes, units, and range. Normalisation should be performed when the scale of a feature is irrelevant or misleading.

There are following methods of scaling:

1. Standardization: In standardization, the values are replaced by z score. This redistributes the features with their mean **μ = 0** and standard deviation **σ =1**

2. Normalization (Min-Max scaling) : This scaling brings the value between 0 and 1.The following formula is used:

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

The advantage of Standardisation over the other is that it doesn't compress the data between a particular range as in Min-Max scaling. This is useful, especially if there is are extreme data point (outlier).

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

5. The VIF is given by the following formula:

$$VIF = 1/(1-r^2_i)$$

Here,'i' refers to the i-th variable, which is being represented as a linear combination of the rest of the independent variables. You will see VIF in action during the Python demonstration on multiple linear regression.

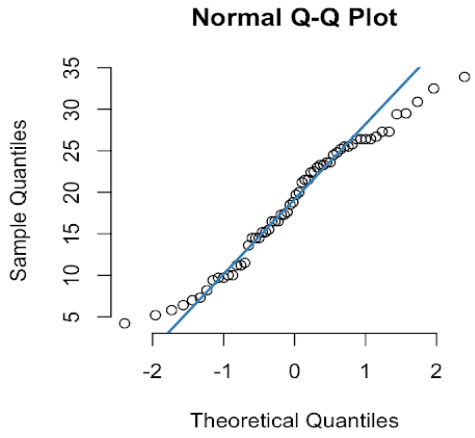
VIF shows how the variance of an estimator is inflated by the presence of multicollinearity. As we look at the formula , as the value of $r^2_i$ approaches 1, the VIF approaches infinity. That is, as the extent of collinearity increases, the variance of an estimator increases, and in the limit it can become infinite. As can be readily seen, if there is no collinearity between variables, VIF will be 1.

Normally, VIF less than 5 is considered to be good, and there is no need to eliminate such variables.

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

6. The Q-Q plot is a visualisation method that helps us to check if the two datasets come from a population with a same distribution. It is plot of quantiles of the first dataset against the quantiles of the second dataset.

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.




This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

The Q-Q plot could be used to check the various assumptions of linear regression:

1. Normal distribution
2. Presence of outlies