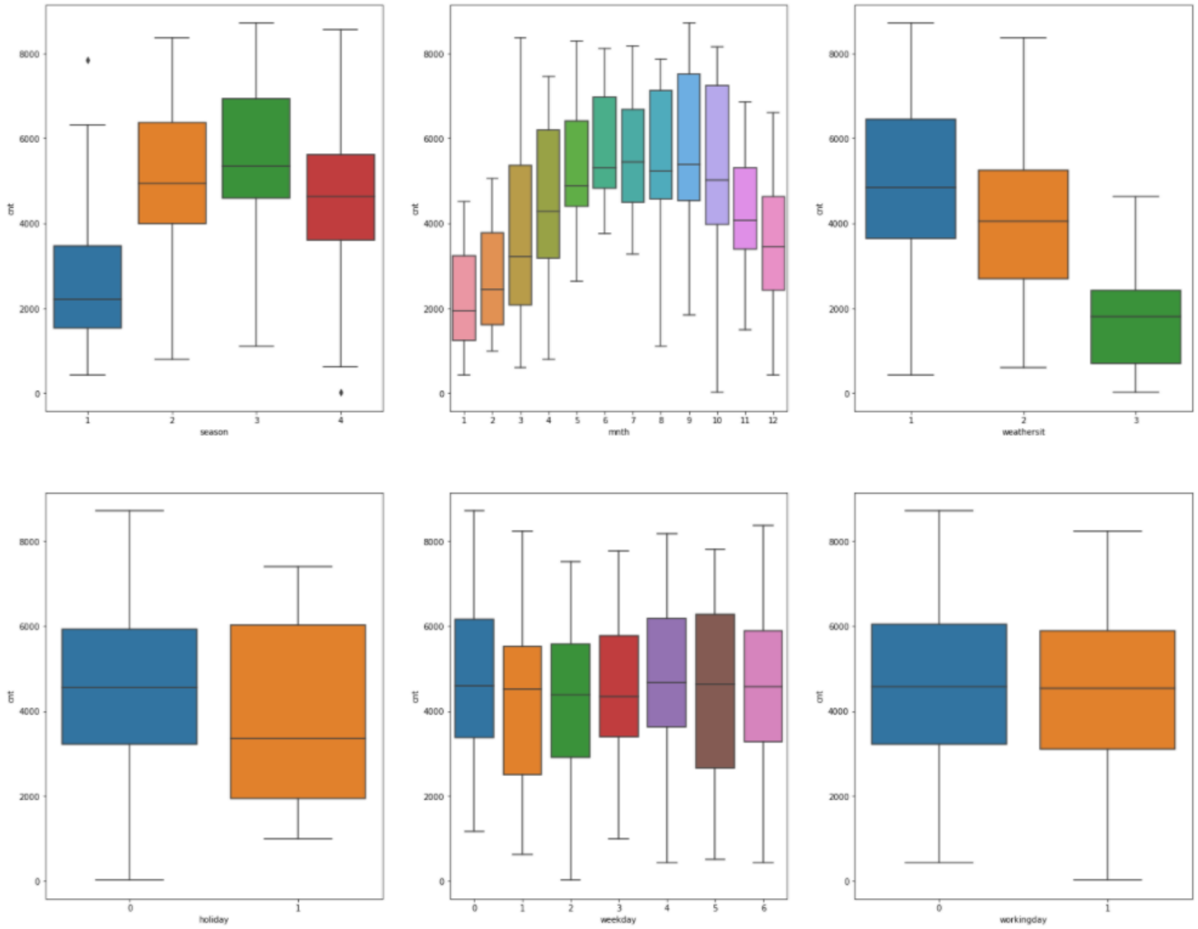
Assignment-based Subjective Questions

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

Ans : The categorical variables in the dataset that helps us decide about the demand count of bikes are - season, mnth, weekday, weathersit, year, holiday, workingday

All these categorical variable are initially studied as boxplot in the python notebook and the results convey the effect of these variables on the dependent variable.



From the box plot, we can easily conclude that:

1. Bike sharing count is more in season of summer and fall compared to winter and spring. Indicates:- season can be a good predictor for the dependent variable

2. Bike sharing count is more in month of June to October than other months.

3. Bike sharing count is more when weathersit is Clear, Few clouds, partly cloudy, Partly

cloudy.

4. Holiday- There is not much of a difference in the level of demand here. Except that all of the datapoints are located a little higher for the nonholiday day, and for holiday, the datapoints are a bit spread, and the median is lower than that of a non- holiday

5. Yr - Bike sharing count is more in year 2019

6. Weekday - Bike Sharing count have no major impact of working day and weekday, which makes us come to an inference that people create demand on working days as well as weekend almost equally. This is also a similar case in working day variable

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

Ans: Using drop_first=True is more common in statistics and often referred to as "dummy encoding". When we convert the categorical variables to dummies, indirectly we are giving importance to each value in a categorical column by making each value as a column. If we don't drop the first column then your dummy variables will be correlated. This may affect some models adversely and the effect is stronger when the cardinality is smaller.

In the process of creating dummy variables, if the number of levels is n, then n-1 dummy variables is sufficient for representing them.

For example, consider the number of levels is 3

And, the variables are A, B and C

While making the dummy values for all the 3 variables, we get

A = 1 0 0 (dummification can be thought of as a sensor signalling whenever the

variable is A, and is silent for the other two, hence we get this)

B = 0 1 0 (Similarly, firing for B)

C = 0 0 1 (for C)

Now, we have use 3 variables for the dummy variable creation

But, there's a possibility of using 2 variables

Converting all the values to 0 0 can also be thought of as A, and the other two

variable value remains the same

A = 0 0 , B = 1 0 , C = 0 1

The solution to the dummy variable trap is to drop one of the categorical variables if there are n number of categories, we have to use n-1 in the model and hence we are using drop_frist=True while creating dummy variables.

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

Ans: Pair-Plot tells us that there is a LINEAR RELATION between 'temp/atemp' and 'cnt' variables has good relationship. We can say that 'temp' and 'cnt' variables we can consider while performing analysis.

Note:- Registered and Casual were highly correlated with cnt, hence they needs to be ignored in analysis as they directly relates to cnt.

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

Ans:

• Linear relationship between X and Y variables.

• No multicollinearity for this we have used VIF method in the analysis to overcome

multicollinearity and removed those features whose VIF>=5.

• Error terms are normally distributed or not – We have plot distplot by plotting residuals

which is difference between y_train – y_train_pred. We can also use Q-Q plot for the

same

• Error terms are independent and have constant variance(homoscedasticity) – Validated

this assumption based on the scatter plot taking y_train_pred on X axis and residuals on

y axis. If there is no pattern as such it shows that they have constant variance and

independent to each other.

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

Ans: Based on the final model and the coefficient of the variables given by building the model, the top 3 features actually contributing significantly and explaining the demand for the shared bikes are :

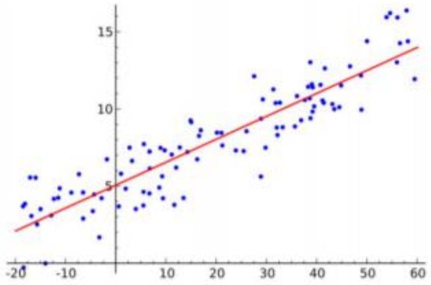
- 1st one is temp the count of bikers will increase.
- 2nd one is weathersit category 3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) during these weather situation demands of bikes are reduced.
- 3rd one is a year, every year we can see the growth of demand in bikes.

General Subjective

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

   Ans: Linear Regression algorithm is a machine learning algorithm based on supervised learning method. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

*Linear Equation => y = B1*X1 + B0*

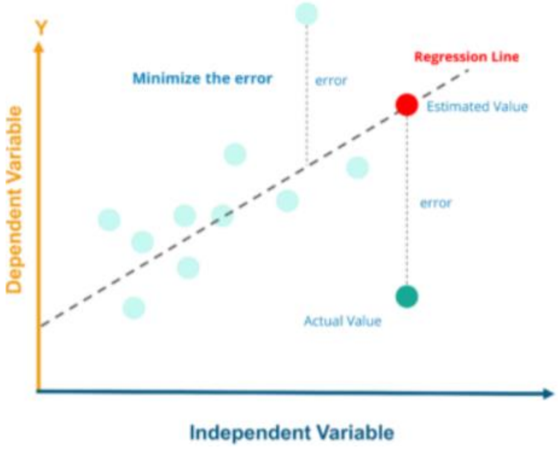


Linear regression quantifies the relationship between one or more predictor variable(s) and one outcome variable. For example, it can be used to quantify the relative impacts of age, gender, and diet (the predictor variables) on height (the outcome variable).

Linear Regression is the process of finding a line that best fits the data points available on the plot, so that we can use it to predict output values for inputs that are not present in the data set we have, with the belief that those outputs would fall on the line

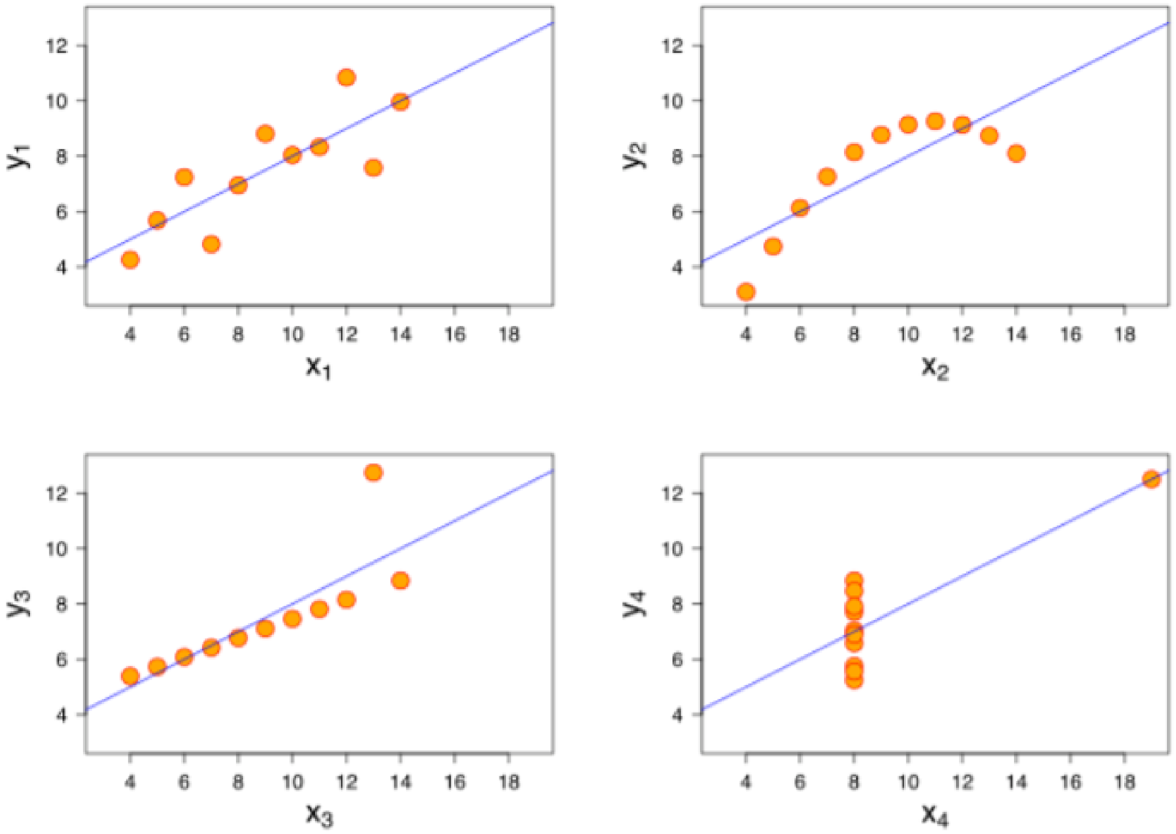
Linear regression is a way to model the relationship between two variables. ... The equation has the form Y= a + bX, where Y is the dependent variable (that's the variable that goes on the Y axis), X is the independent variable (i.e. it is plotted on the X axis), b is the slope of the line and a is the y-intercept.

Linear regression attempts to model the relationship between two variables by fitting a linear equation (= a straight line) to the observed data. One variable is considered to be an explanatory variable (e.g. your income), and the other is considered to be a dependent variable (e.g. your expenses).



2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's Quartet was developed by statistician named Francis Anscombe. It comprises four datasets, each containing eleven (x,y) pairs. The important thing to note here about these datasets is that they have a similar descriptive statistics. But when they are visualized, each graph tells a new visual story irrespective of their identical summary statistics. Visualization may not be as precise as statistics, but it provides a unique view onto data that can make it much easier to discover interesting structures than numerical methods. Visualization also provides the context necessary to make better choices and to be more careful when fitting models or in the preprocessing stage. Anscombe’s Quartet is a case in point, showing that four datasets that have identical statistical properties can indeed be very different.



All four of these data sets have the same variance in x, variance in y, mean of x, mean of y, and linear regression. But, as you can clearly tell, they are all quite different from one another. Anscombe's Quartet shows that multiple data sets with many similar statistical properties can still be vastly different from one another when seen through visual plots. Additionally, Anscombe's Quartet warns of the dangers of outliers in data sets. Think about it: if the bottom two graphs didn't have that one point that strayed so far from all the other points, their statistical properties would no longer be identical to the two top graphs. In fact, their statistical properties would more accurately resemble the lines that the graphs seem to depict. It is obvious that the top right graph really shouldn't be analyzed with a linear regression at all because it's a curvature. Conversely, the top left graph probably must be analyzed using a linear regression because it's a scatter plot that moves in a linearly manner. These observations demonstrate the value in graphing your data before analyzing it. Anscombe's Quartet is a important demonstration which shows us that visualization is a pre-processing step which must be included in the process before model building to throw in some accurate results from the model.

3. What is Pearson’s R?

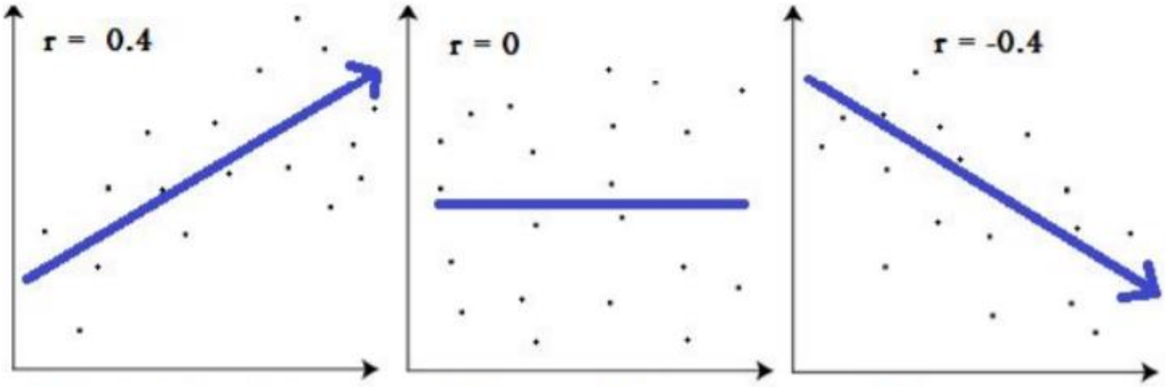
Ans: s Correlation coefficients are used in statistics to measure how strong a relationship is between two variables. There are several types of the correlation coefficient, but the most popular is Pearson’s. Pearson’s correlation (also called Pearson’s R) is a correlation coefficient commonly used in linear regression. Correlation coefficient formulas are used to find how strong a relationship is between data.

The formulas return a value between -1 and 1, where:

● 1 indicates a strong positive relationship.

● -1 indicates a strong negative relationship.

● A result of zero indicates no relationship at all.



Positive Correlation No correlation Negative

One of the most commonly used formulas in stats is Pearson's correlation coefficient formula:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

The assumptions that Pearson's R coefficient –

- For the Pearson's R correlation, both variables should be **normally distributed**.
- There should be no significant **outliers**. We all know what outliers are but we don't know the effect of outliers on Pearson's correlation coefficient, R. This coefficient is very sensitive to outliers, can have a very large effect on the line of best fit. This also means including outliers in your analysis dataset, can lead to misleading results.
- Each variable should be **continuous** i.e. interval or ratios for example time, height, age etc. must be equal.
- Scatter plots will help you tell whether the variables have a linear relationship. If the data points have a straight line, then the data satisfies the linearity assumption. If the data you have is not linearly related you might have to run a non-parametric .
- The observations are **paired observations**. That is, for every observation of the independent variable, there must be a corresponding observation of the Assignment 02 8 dependent variable.
- **Homoscedascity** - Homoscedascity simply refers to 'equal variances' of errors as well as dependent variable. A scatter-plot makes it easy to check for this. If the points lie equally

on both sides of the line of best fit, then the data is homoscedastic. As a bonus — the opposite of homoscedascity is heteroscedascity which refers to refers to the circumstance in which the variability of a variable is unequal across the range of values of a second variable that predicts it.

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

Ans: Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

It is important to note that **scaling just affects the coefficients** and none of the other parameters involved in the modelling.

Normalization/Min-Max Scaling: Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

- It brings all of the data in the range of 0 and 1. **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python.
- Normalization can also be used to scale categorical variables as their values are always 0 and 1. Normalization can be safely used on all the variables overall.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling: Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

- **sklearn.preprocessing.scale** helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it loses some information in the data, especially about **outliers**.

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

Ans: VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables. In VIF, each feature is regression against all other features. If R^2 is more which means this feature is correlated with other features. [0]

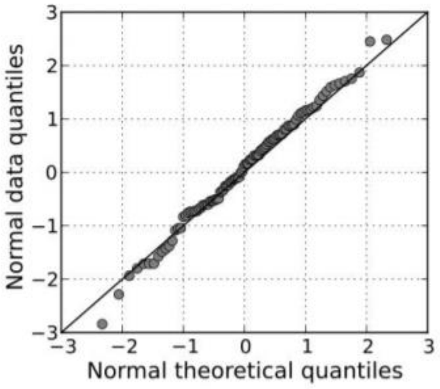
- VIF = 1 / (1 – R^2)
- When R2 reaches 1, VIF reaches infinity.

VIF infinite thus means **perfect correlation** that exists between two independent variable. There might be a possibility of 2 variables being extremely similar in their way of affecting the target variable. For example, one can be a student's CGPA and one variable can be his percentage. This means that these two independent variable essentially tells the same thing, i.e the student's performace, and this is highly collinear and has an infinite VIF. Thus, understanding VIF and how collinear two variables are, hold a very good significance with respect to the model building and its accuracy.

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

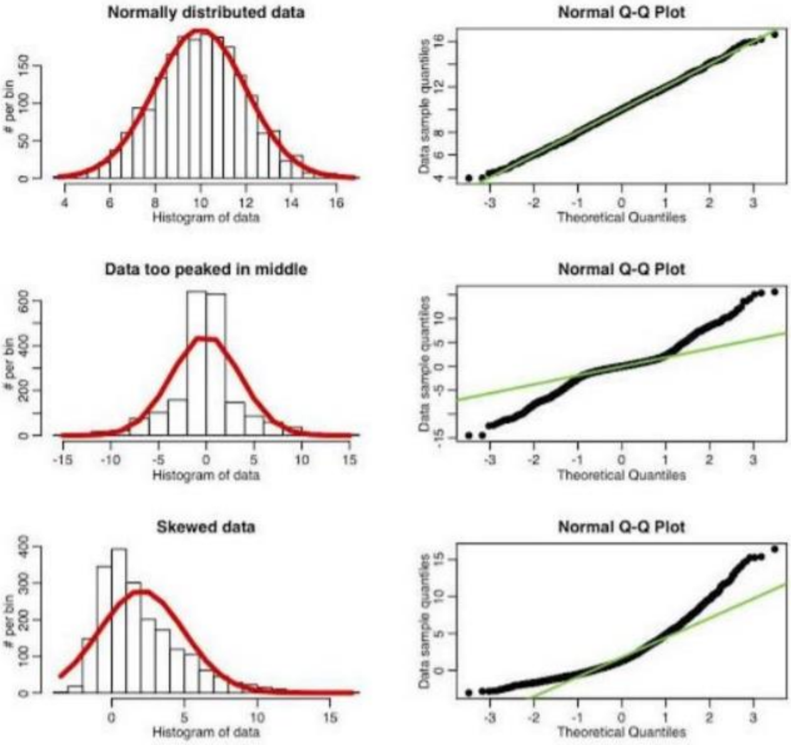
Ans: Q-Q(quantile-quantile) plots play a very vital role to graphically analyze and compare two probability distributions by plotting their quantiles against each other. If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line y = x.

We plot the theoretical quantiles or basically known as the standard normal variate (a normal distribution with mean=0 and standard deviation=1)on the x-axis and the ordered values for the random variable which we want to find whether it is Gaussian distributed or not, on the y-axis. Which gives a very beautiful and a smooth straight line like structure from each point plotted on the graph.



Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian Distribution, Uniform Distribution, Exponential Distribution or even Pareto Distribution, etc. You can tell the type of distribution using the power of the Q-Q plot just by looking at the plot.

Below are the types of output that you can expect.



Normally distributed data
# per bin
0
50
100
150
4
6
8
10
12
14
16
Histogram of data
Normal Q-Q Plot
Data sample quantiles
4
6
8
12
16
-3
-2
-1
0
1
2
3
Theoretical Quantiles
Data too peaked in middle
# per bin
0
200
400
600
-15
-10
-5
0
5
10
15
Histogram of data
Normal Q-Q Plot
Data sample quantiles
-15
-5
0
5
10
-3
-2
-1
0
1
2
3
Theoretical Quantiles
Skewed data
# per bin
0
100
200
300
400
0
5
10
15
Histogram of data
Normal Q-Q Plot
Data sample quantiles
0
5
10
15
-3
-2
-1
0
1
2
3
Theoretical Quantiles