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)

Ans:- From just visualizing the data we found out that :

- most of the bookings were done in season 3(fall) and 2(summer)
- More bookings done in yr 1(2019)
- More bookings in the middle months
- more bookings when its not a holiday
- more booking on workingday
- bookings spread evenly between all weekdays
- more bookings when weathersit was 1(Clear, Few clouds, Partly cloudy, Partly cloudy) or 2(Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist)

However after creating the model we get our best fit line which was -

cnt = 0.1945 + (0.2292 * yr)+ (-0.0558 * holiday)+ (0.0444 * workingday)+ (0.5301 * temp)+ (-0.1692 * hum)+ (-0.1857 * windspeed)+ (-0.0582 * 2_Mist) + (-0.2486 * 3_rain_snow)+ (0.0529 * sun)+ (0.056 * aug)+ (0.1255 * sep)+ (0.0411 * oct)+ (0.1039 * summer)+ (0.1348 * winter)

And below are the insights from the model for the categorical variable-

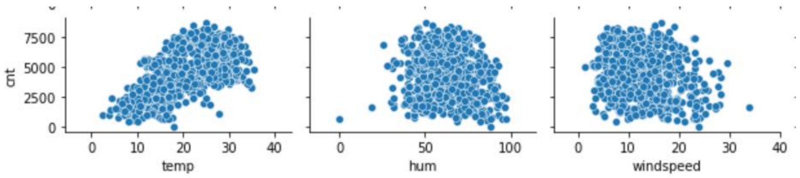
- winter - when effects of season is taken into account, winter will increase the cnt by 0.13 times than the reference group (spring)
- sep - when effects of months is taken into account, sep will increase the cnt by 0.12 times than the reference group (january)
- summer - when effects of season is taken into account, summer will increase the cnt by 0.10 times than the reference group (spring)
- aug - when effects of months is into account, aug will increase the cnt by 0.05 times than the reference group (january)
- sun - when effects of weekday is into account, sun will increase the cnt by 0.05 times than the reference group (monday)
- workingday - the cnt will increase by 0.04 times when it’s a workingday
- oct - when effects of months is taken into account, oct will increase the cnt by 0.04 times than the reference group (january)
- holiday - the cnt will decrease by 0.05 times when it’s a holiday
- weathersit(2) - when effects of weather status is taken into account, cnt will decrease by 0.05 times than the reference group (Clear, Few clouds, Partly cloudy, Partly cloudy) when weather is Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
- weathersit(3) - when effects of weather status is taken into account, cnt will decrease by 0.24 times than the reference group (Clear, Few clouds, Partly cloudy, Partly cloudy) when weather is Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

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

Ans:- When we are creating dummy variables for “n” category , then if we already know n-1 categories then we can easily predict the nth category, so having an extra variable is redundant. That’s why we drop the first variable whose value is implicitly explained by the others

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

Ans:-We have the below pair plot of cnt with the numerical variables -

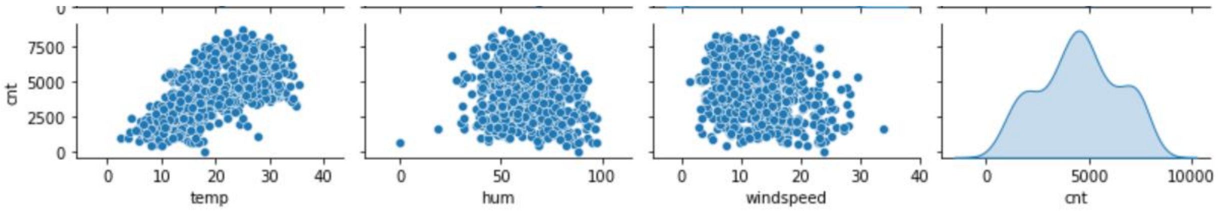


Looking at the pair plot , its safe to say that “**temp**” has the highest correlation with the target variable

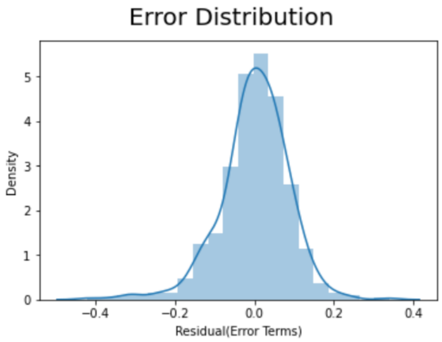
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans:- We validated our assumptions of Linear Regression by doing the Residual Analysis . Following are the proofs of the assumptions-

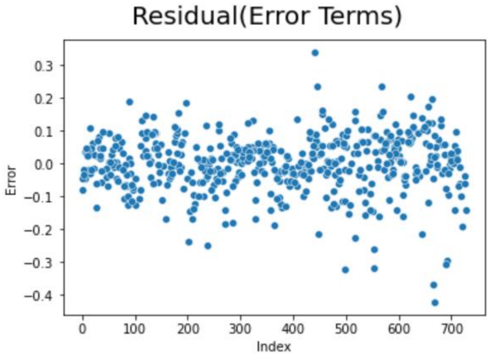
- There is a linear relationship between X and Y (which we can see from the below plot that we plotted)



- Error terms are normally distributed with mean zero



- Error terms are independent of each other and have a constant variance



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

Ans:- The top 3 predictors were

- **temp** - with unit increase in temp the cnt increases by .53 times
- **yr** - as yr increases by a unit, the cnt increases by 0.22 times
- **winter** - when effects of season is taken into account, winter will increase the cnt by 0.13 times than the reference group (spring)

General Subjective Questions

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

Ans:- Linear regression is a machine learning algorithm which estimates how a model is following a linear relationship between one target variable and one or more predictors.

There are 2 types of linear regression:

- Simple Linear Regression
- Multiple Linear Regression

Simple Linear Regression: It is a type of linear regression model where there is only one predictor variable.

Multiple Linear Regression: It is similar to simple linear regression but here we have more than one predictor variable

Steps followed in Linear Regression Algorithm:

- Reading and understanding the data
  - Importing required libraries like pandas & numpy for data analysis and seaborn & matplotlib etc for data visualization
  - Cleaning and manipulating data to make it up to the standards by treating null values if any, changing data types if needed, removing unwanted rows or columns etc.

- Visualizing the data (Exploratory Data Analysis)
    - Visualizing numerical variables using scatter or pairplots to find visual patterns
    - Visualizing categorical variables using barplots or boxplots to find some patterns
- Data Preparation
    - Converting categorical variables with varying degrees of levels into dummy variables so that these variables can be represented as numbers during model building in order to contribute to the best fitted line for the purpose of better prediction
- Splitting the data into training and test sets
    - Splitting the data into two sections so that one subset which is the known (Train dataset generally 70%) is used to generate the best fit line and using this we will try to estimate the target variable from the unknown(test dataset generally 30%)
    - Rescaling the features: It is a method used to normalize the range of numerical variables with varying degrees of magnitude so that the contribution of each variable with varying magnitude has similar impact
- Building a linear model
    - Forward Selection: We start with null model and add variables one by one. These variables are selected on the basis of high correlation with target variable. First we select the one, which has highest correlation and then we move on to the second highest and so on
    - Backward Selection: We add all the variables at once and then eliminate variables based on high multicollinearity (VIF>5) or insignificance (high p-values)
    - RFE or Recursive Feature Elimination: Its like an automated version of feature selection technique where we select “m” variables out of “n” variables and then machine provides a list of features with importance level given in terms of support and rankings. A rank 1 means that feature is important for the model, while a rank 4 implies that we are better off. Similarly support will be True or False implying whether it is important or not for model building
- Residual analysis of the train data
    - It tells us how much the errors (y_actual — y_pred) are distributed across the model. Generally there are 4 assumptions that we need to verify if those are holding true once the model is created
        - There is a linear relationship between X and Y (which we can see from the below plot that we plotted)
        - Error terms are normally distributed with mean zero
        - Error terms are independent of each other
        - Error terms have a constant variance
- Making predictions using the final model and evaluation
    - We will predict the values by applying our model on the test dataset
    - Compare how much accurate is our predicted values to the original values in the test data set

- Calculate the R2 and Adjusted R2 for the test data set and see the difference with the R2 and Adjusted R2 for train dataset .A difference of 3-5 % is generally acceptable to conclude that the model predicted correctly

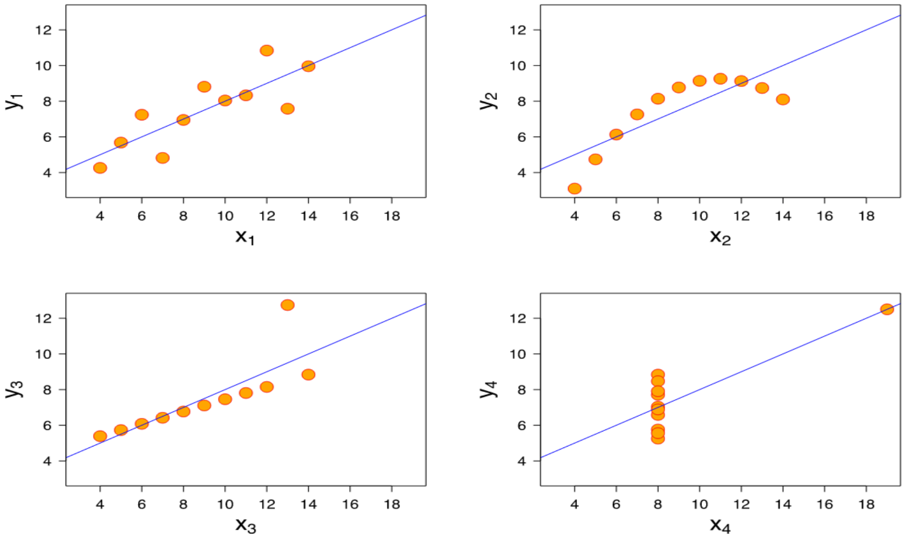
2. Explain the Anscombe’s quartet in detail. (3 marks)

Ans- It’s a group of four datasets that appear to be similar when using typical summary statistics yet tell four different stories when graphed. Each dataset consists of eleven (x,y) pairs as follows

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

All the summary statistics that are generally computed are close to identical:

- The average x value is 9 for each dataset
- The average y value is 7.50 for each dataset
- The variance for x is 11 and the variance for y is 4.12
- The correlation between x and y is 0.816 for each dataset
- A linear regression (line of best fit) for each dataset follows the equation y = 0.5x + 3
- So far these four datasets appear to be pretty similar. But when we plot these four data sets on an x/y coordinate plane, we get the following results:

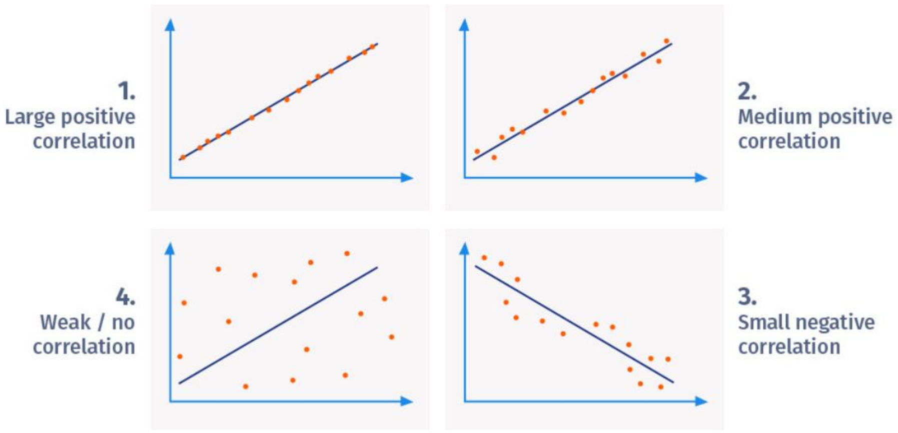


Now we see the real relationships in the datasets start to emerge. Dataset I consists of a set of points that appear to follow a rough linear relationship with some variance. Dataset II fits a neat curve but doesn't follow a linear relationship .Dataset III looks like a tight linear relationship between x and y, except for one large outlier. Dataset IV looks like x remains constant, except for one outlier as well. Computing summary statistics or staring at the data wouldn't have told us any of these stories. Instead, it's important to visualize the data to get a clear picture of what's going on.

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

Ans- Pearson correlation coefficient or Pearson's r is defined in statistics as the measurement of the strength of the relationship between two variables and their association with each other.

In simple words, Pearson's correlation coefficient calculates the effect of change in one variable when the other variable changes. 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 i.e. as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association i.e. as the value of one variable increases, the value of the other variable decreases.



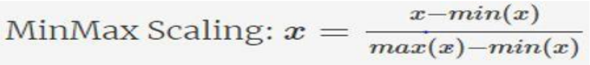
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans- It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

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.

Normalization/Min-Max Scaling:

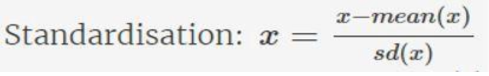
- It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.



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

Standardization Scaling:

- 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).



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

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

Ans- VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity.

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which leads to 1/(1-R2) which is equal to infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

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

Ans- A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

QQ plots is very useful to determine

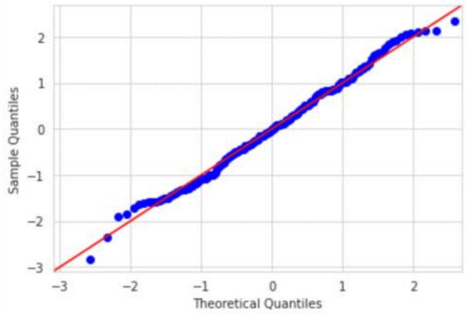
- If two populations are of the same distribution
- If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this.
- Skewness of distribution

In Q-Q plots, we plot the theoretical Quantile values with the sample Quantile values. Quantiles are obtained by sorting the data. It determines how many values in a distribution are above or below a certain limit

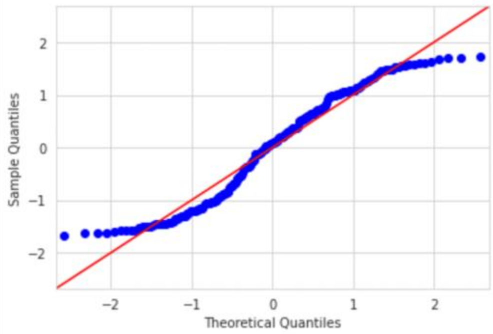
As we build our linear regression model, we can check the distribution of the error terms or prediction error using a Q-Q plot. If there is a significant deviation from the mean, we should check the distribution of feature variable and consider transforming them into a normal shape.

Following are few of the scenarios:

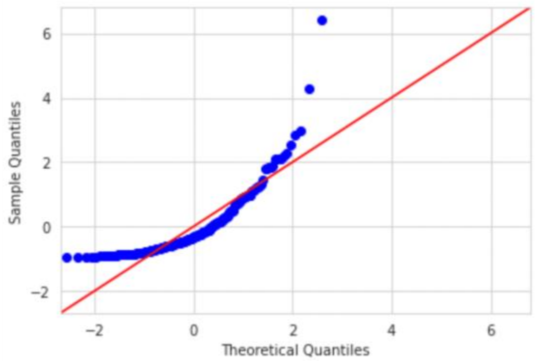
- If the datasets we are comparing are of the same type of distribution type, we would get a roughly straight line.



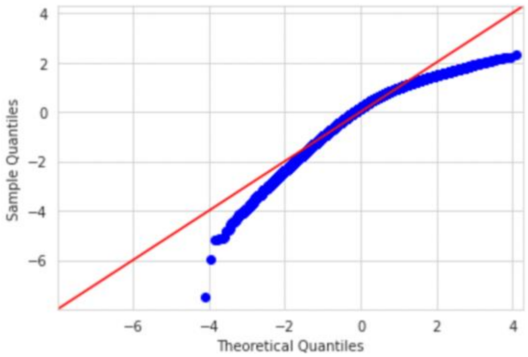
- If the dataset is uniformly distributed , then we will get graph like below



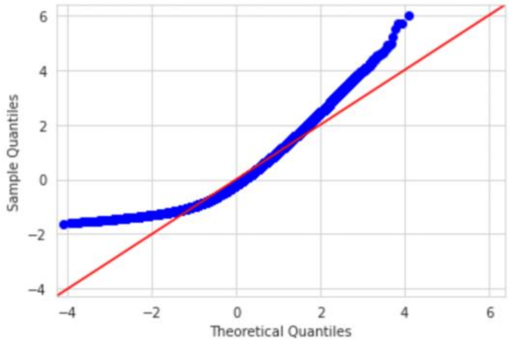
- If we plot a variable with exponential distribution with theoretical normal distribution, the graph will look like below



- Q-Q plots can be used to determine skewness as well. If we see the left side of the plot deviating from the line, it is left-skewed



- Similarly, a right-skewed distribution would look like below. We can observe deviation on the right side


Sample Quantiles
Theoretical Quantiles