Name: Harish Kumar

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?**

In case of Bike Sharing Problem, there are many Categorial variables effect the value of dependent variable, for example from our analysis we could clearly see that the Bike sharing numbers has been increased drastically from 1,243,103 in Year-2018 to 2,047,742 Year-2019, so could inter that year on year Bike sharing numbers are increasing.

With respect to weather sit, when it is Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds people don't prepare to ride a bike so we could inter that people don't prepare to ride a bike when the weather is Light Snow, Light Rain + Thunderstorm + Scattered clouds or Light Rain + Scattered.

If we look at the Months, as per the analysis in the month 9(Sept) has high positive correlation compare to other month, so we could infer that bike sharing is more in the 9^{th} month.

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

We pass this parameter while creating the dummy variable for categorical column. In general, the pd.get_dummies function creates new number of columns which is same as number of unique items of that column, and if that item column is applicable then it will be indicate as 1 else 0.

In the below example there are 3 unique items, so the get_dummies function 3 columns. If the product size Small, the Small column will be indicated as 1 else 0. But actually using 2 columns itself we could find the Product's size. So could use the drop_first=True to drop the first column.

By dropping the first column, we could reduce the size of dataset, improve the efficiency of the model so the adjusted R-Squared value and if we are note dropping the column there will be high VIF associated with that column

For example:

Product' Size Categories – Small, Medium and Big

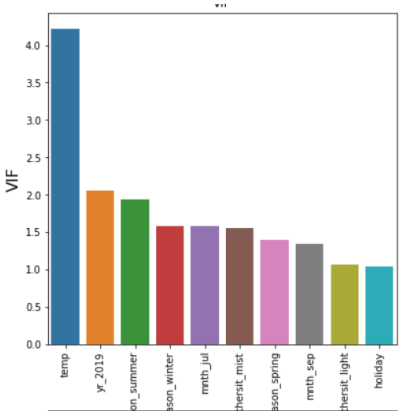
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

The “Registered” numerical variable is having high correlation of 0.95%, followed by “Casual -0.67%” and “Temp-0.63%”.

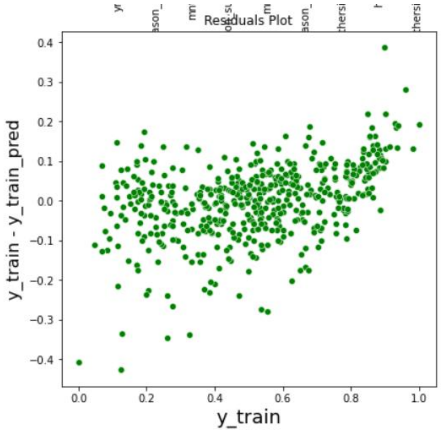
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

I have verified the following assumptions of Linear Regression after building my model on the training set

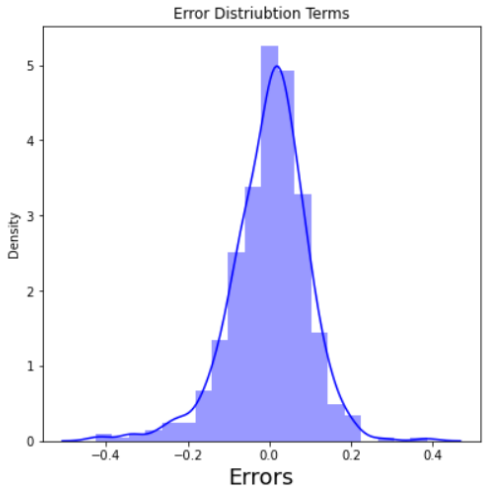
- There should be a Linear relationship between Independent and dependent variables. The co-efficiency value of the independent variables indicates if there is any linear relationship among the independent and dependent variable
- There should be no correlation between the residual(error) terms. We could use the P-value find if there is any re
- The independent variables should not be correlated, the absence of this phenomenon is known as multicollinearity. We could use VIF method to find if there is a multicollinearity. The VIF value anything which more than 5 indicates high multicollinearity



- The error terms must have constant variance. This phenomenon is known as homoskedasticity. If the non-constant variance is referred to as heteroskedasticity.



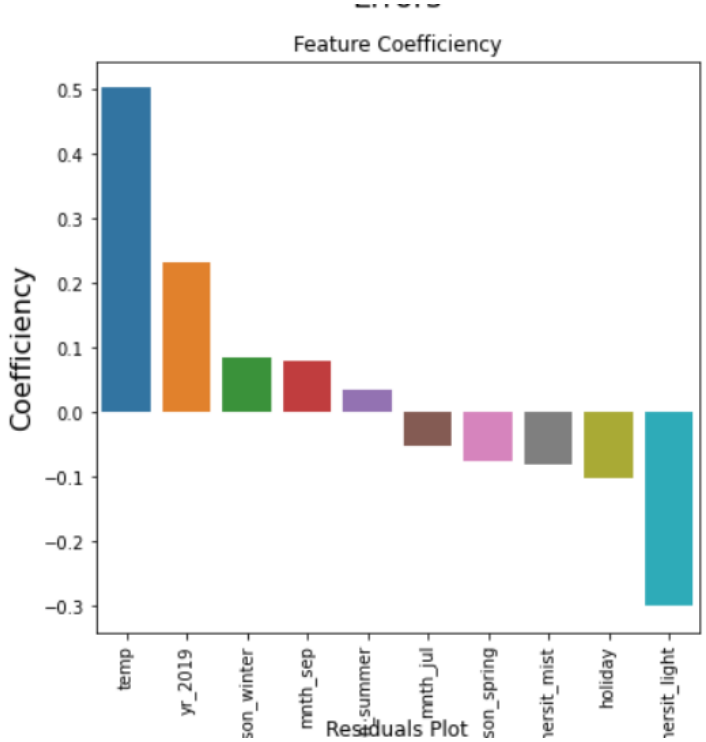
- The error terms must be normally distributed. We can plot distribution plot for residual(Y_train – Y_train_pred) to find the same.



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

Based on my Analysis Temperature(temp) and Year (Yr-2019) having high positive co-relation and weather sit - Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered is having high negative co-relation. So these 3 features contributing significantly towards explain the demand of the shared bikes.

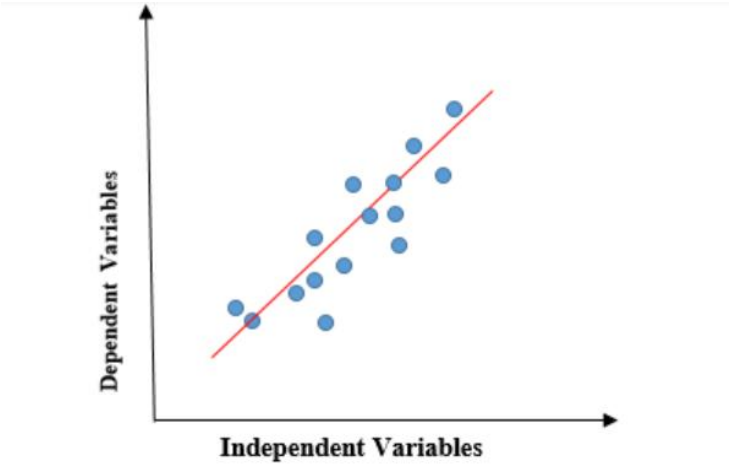
The below table and plot show the significance of other features.



General Subjective Questions

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

   The linear regression algorithm can be used to predict the behavior of linearly correlated data point. For example, the sales promotion and sales numbers are linearly correlated, which mean when the promotion budget increases, the sales figures are also expected to increase. Since there is a correlation among the promotion and sales, we can train the model using the linear regression algorithm and then can predict what could be the sales value if we know the promotion budget.



   The linear regression algorithm is simplest and most widely used algorithm. It is supervised algorithm, which means it needs to have to target variable to train the model.

   Linear Regression equation: y = a + bx

   A is slope
   b is intercept
   x is independent variable
   y is dependent variable or target

   If there are more than one independent variable, then we can use the **multiple linear regression algorithm** to predict the dependent variable

   Use case of Linear Regression:
   - Predict the Price of a any product based on the features of that product. For example, we can find the Price of the house based on the area, no. of bedroom, and price of similar own in that area.
   - Risk Management – Using regression, we can analyze risk in the financial and insurance sector

2. Explain the Anscombe's quartet in detail.

Anscombe's Quarter can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. The Statistics summary for all the 4 set is same, as shown in the below table.

Anscombe's Data									
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

When these models are plotted on a scatter plot, all the datasets generate a different kind of plot that is not interpretable by any regression algorithm. So it shows the importance of data visualization. Hence all the important features in the dataset must be visualized before implementing any machine learning algorithm.

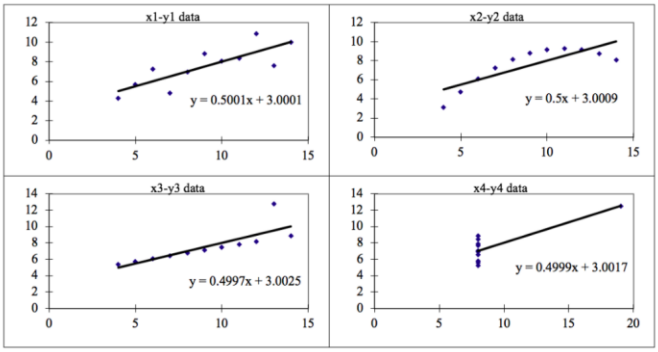


Image by Author

3. What is Pearson's R?

The Pearson's R is also referred as Pearson's correlation coefficient, the Pearson product-moment correlation coefficient(PPMCC), or bivariate correlation. It is a statistic that measure the linear correlation between two variables. Like all the correlation, it also has a numerical value that lies between -1.0 to +1.0

Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient. However, it cannot capture nonlinear relationships between two variables and also cannot differentiate between dependent and independent variables.

The Pearson's R is the covariance of the two variables divided by the product of their standard deviation.

There are certain requirements for pearson's correlation coefficient:

- Scale of measurement should be interval or ratio
- Variables should be approximately normally distributed
- The association should be linear
- There should be no outliers in the data

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

Where,

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores

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

Most of the times, collected data set will have highly varying in magnitudes, units and range. So the scaling is a way of pre-processing the dataset which are having continuous numbers to bring the numbers to same range, scale or magnitudes. For example, if you are predicting house price, the no. of bedrooms(3 bhk) and Area(1200 sqft) are continuous numbers, but numbers difference is very large is it good to bring them to same scale.

Most of algorithm expect the data to be in the same range to predict the values correctly, if scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue we have bring all the variables value to the scale, and it also helps in speeding up the calculations.

It is important to note that scaling just affects the coefficients and none of the other matrics like t-statistic, F-statistics, p-values and R-squred etc.

Two ways of doing scaling are :

- **Normalized scaling:** It is also called min max scaling. The following formula will be applied to calculate or convert the value to normalization. The MinMax function can also be accessed from sklearn.preprocessing.MinMaxScaler library.

  Formula : x – min(x)/ (max(x) – min(x))

  Where,
  X – is the value of feature variable
  Min – is the Minimum value in that feature column
  Max - is the Maximum value in that feature column
  Note: Once the MinMax scaling applied, the column values will range from 0 to 1.
- **Standardization scaling**: Standardization replaces the values by their Z scores. Basically it indicates a data point as number of standard deviation away from the mean value.

  Formula: x – mean(x)/sd(x)
  Where,
  X – is the value of feature variable
  Mean – is the mean value in that feature column
  SD – is the Standard Deviation of that feature column

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

The VIF(Variance Inflation Factor) is a matric used to find the correlation among the independent variables. If there is a perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables.

Formula for VIF : 1/(1-R2)

Where,
R2 – R-Squared

In case of perfect correlation, we get R2 =1, so

VIF = 1/(1-1) = 1/0 = infinite

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

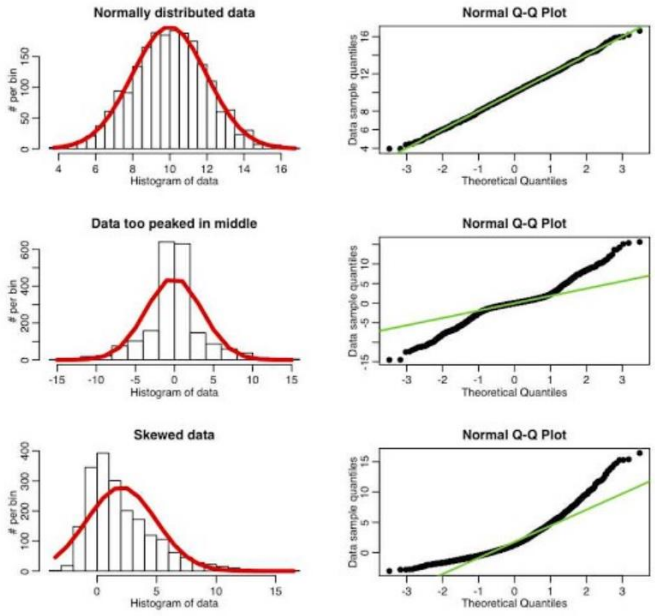
The Q-Q(Quantile-Quantile) plot is a graphical tool to help us assess if a set of data possibly came from same theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training our model with Uniform distribution Train dataset, and using Q-Q plot we can check if the Test Dataset also follows the Uniform distribution, so that our model predicts the target variable value correctly.

Q-Q plot can be used to check the following scenarios if two data sets –

- Comes from population with a common distribution
- Have common location and scale
- Have similar distribution shapes
- Have similar tail behavior

Interpretation – 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.



Source: Sherrytowers Q-Q plot examples