Assignment Based Subjective Question :

1) from your analysis of the categorical variable from the dataset what could you infer about their effect on the dependent variable?
Answer: Different categorical variable have different effect on the dependent variable for example take holiday

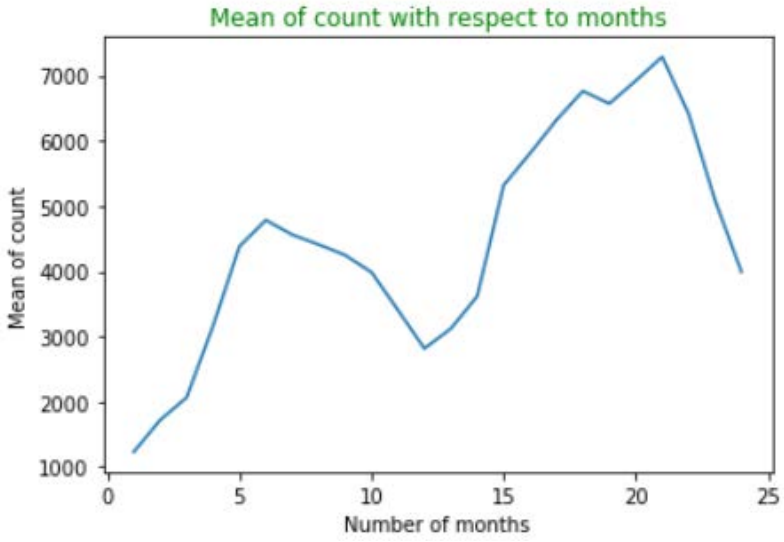
```
the mean of the bike data count 4508.006849315068
the standard deviation of the data count 1936.0116473612595
```

holiday	cnt
0	4530.90268
1	3735.00000

here we can see that mean for the count 4508 and 1936 is standard deviation
As you can see here there are two means when there is holiday and not when there is not a holiday and a significant drop in the mean.

categorical variable for the year and month

yr	mnth	cnt
0	1	1231.903226
	2	1721.964286
	3	2065.967742
	4	3162.333333
	5	4381.322581
	6	4783.733333
	7	4559.387097
	8	4409.387097
	9	4247.266667
	10	3984.225806
	11	3405.566667
	12	2816.870968
1	1	3120.774194
	2	3617.964286
	3	5318.548387
	4	5807.466667
	5	6318.225806
	6	6761.000000
	7	6567.967742
	8	6919.451613
	9	7285.766667
	10	6414.225806
	11	5088.800000
	12	3990.741935

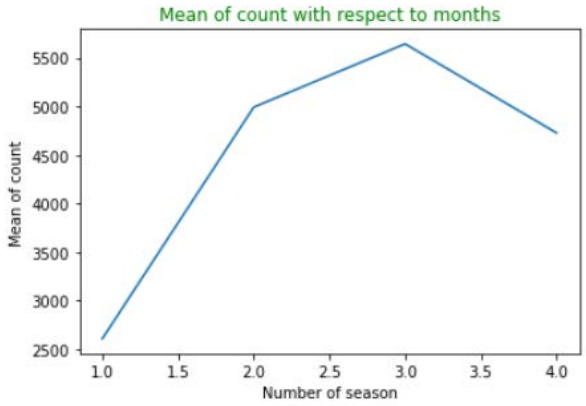


here year and months segmented analysis is done if you look closely you will find the difference between year1 and year2 will be 2000 and year 1 and year2 follow the same cyclic pattern month- month wise however it will be better fit than season but we will have to perform sine and cosine function but since it is still not in the purview of the course **so we will just use yr variable**.

Season Variable:

here we can see there is significant difference of the mean of the count according to season and we will be able to use it as a dependent variable

season	cnt
1	2608.411111
2	4992.331522
3	5644.303191
4	4728.162921



There is clearly a very hughe variance across the 4 seaosn

1:spring --> lowest
2:summer --> higher
3:fall --> highest
4:winter --> 3rd but not significantly less

There are many such categorical variables will not write about everything here like weekdays , working days, weathersit will however attach the result

weathersit	cnt
1	4876.786177
2	4044.813008
3	1803.285714

workingday	cnt
0	4330.168831
1	4590.332665

2) why is it important to use **drop_first=True** during dummy variable creation?

Answer: Two reasons drop_first=True is a good idea:

-> reduces an extra column without missing information and reduces the column which will better our adjusted R-Squared .

-> decreases the redundancy and multicollinearity in the ML model.

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

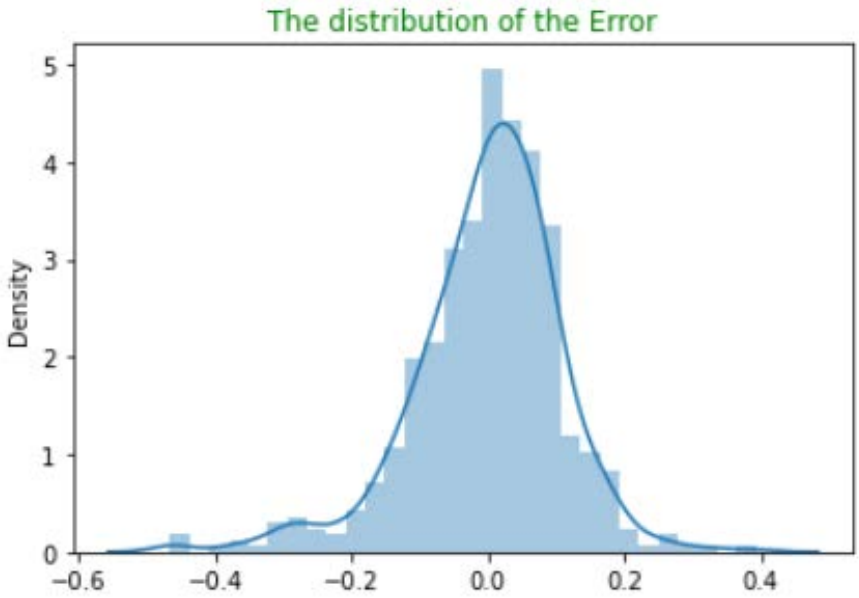
Answer: The highest correlation is **Registered.** but registered + casual gives cnt hence it is expected.

The answer without registered and casual would be **temp** and **atemp**.

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

Answer: Two ways:

1) error is normally distributed and mean of the error is zero
2) error is equally distributed no pattern in the error



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

Answer: The Interpretation id done after min max scaling which i did :

1) Temp/atemp
2) Year
3) constant (but if you are strictly on given independent variables then weathersit_rainy

Dep. Variable:	cnt	**R-squared:**	0.840
Model:	OLS	**Adj. R-squared:**	0.831
Method:	Least Squares	**F-statistic:**	89.99
Date:	Wed, 09 Feb 2022	**Prob (F-statistic):**	1.01e-50
Time:	03:23:59	**Log-Likelihood:**	130.66
No. Observations:	146	**AIC:**	-243.3
Df Residuals:	137	**BIC:**	-216.5
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1466	0.062	2.368	0.019	0.024	0.269
yr	0.2674	0.017	15.534	0.000	0.233	0.301
atemp	0.4770	0.070	6.827	0.000	0.339	0.615
windspeed	-0.1265	0.042	-3.034	0.003	-0.209	-0.044
season_spring	-0.1133	0.043	-2.605	0.010	-0.199	-0.027
season_summer	0.0442	0.030	1.469	0.144	-0.015	0.104
season_winter	0.0609	0.037	1.640	0.103	-0.013	0.134
weathersit_fine	0.0767	0.019	4.128	0.000	0.040	0.113
weathersit_rainy	-0.1765	0.055	-3.220	0.002	-0.285	-0.068

Here is the screen shot of the final model summary that i got .

General Subjective Questions

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

Answer: The linear regression algorithm is the art of fitting a line (simple linear regression) or fitting a plane (multiple linear regression) so that we can predict a certain continuous value (desired value) there are certain assumptions too and also the evaluations way but first let us understand these concepts in detail:

Understanding : fitting a line and line will be of the from :

$$Y=mX+C$$:

here: Y --> is the dependent variable
m--> is the slope of the line.
C --> intercept of the line

so our goal is to find the best m and C such that the error between actual Y and predicted Y is as low as possible which we call it as minimizing the cost function and the process of minimizing the cost function is called Gradient Descent.

In Multiple linear regression the formula changes to

$$y=b_0+b_1(x_1)+b_2(x_2)+.....+b_{(n-1)}(x_{(n-1)})+b_n(x_n)$$

here we have to find the best b_0, b_1 and so on so the cost function is minimum

Assumptions: 1) The variable used to determine is independent of each other
2) There is a linearity between independent and dependent variable
3) Homoscedasticity all the independent variables
4) error are normally distributed and the mean is at 0
5) error are uniformly spread out (means no pattern)

Evaluation: we evaluate the model using R - Squared value

Formula for R-Squared

$$R^2 = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}}$$

Unexplained Variation: RSS- residual sum squared
Total Variation -MSS - mean sum squared

Formula for RSS:

$$RSS = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2$$

2) Explain the Anscombe's quartet in detail.

Answer: Anscombe's data produced by Francis Anscombe to show how the descriptive data can be deceptive and why data visualization is a very important part of the data science

Francis produced four different data which had similar mean, standard deviation and variance and if you put all these four data into the Linear Regression algorithm will produce almost identical results but the underlying distribution of the data is very different. As depicted in the picture below

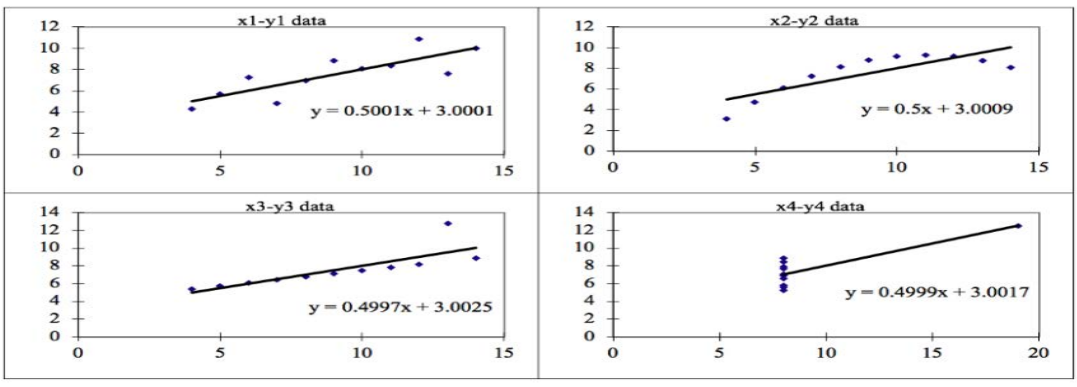


Image by Author

3) What is Pearson's R?

Answer: first let me define correlation

Correlation: it is a statistic that measures the relationship between two variables and it lies between -1.0 to 1.0

Pearson's r or bivariate correlation. It is a statistic that measures the linear correlation between two variables. it does not capture no linear relation between the variable

now to the formula of Pearson's R

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

Where,

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

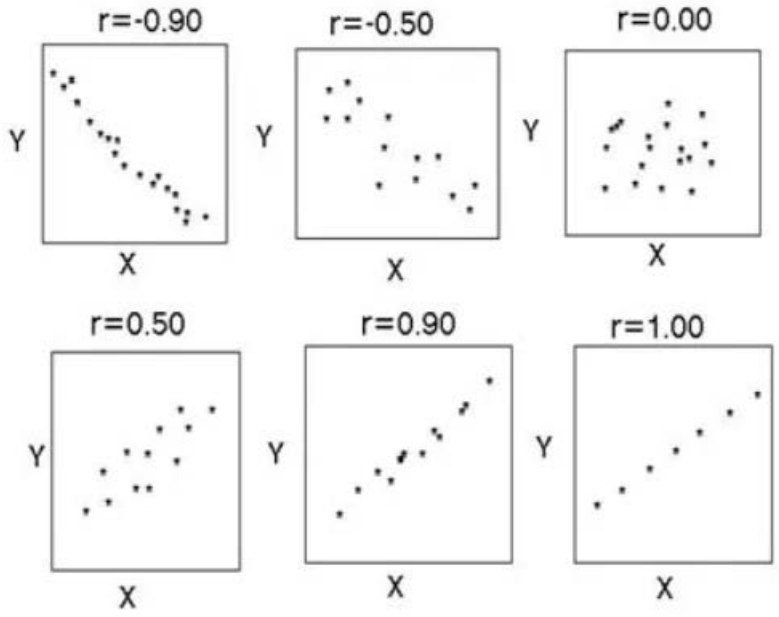
Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores

Effect of r : The Higher the mod(R) the more closely related are the variables linearly. here the sign of the number represents the slope not the magnitude if the r is very high negative number means when one variable the other decreases and positive sign means when one variable increase the other also increases. The magnitude or the numerical part describes the strength of th relation.
note: pearson's r assumes the data is normally distributed.



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

Answer: you either multiply or divide a number by a specific constant to convert it into a desirable value so that you can perform mathematics on it quite easily or you can understand the data quite easily
for instance: you are going over a sales chart of a company and its values is in millions of dollars if you divide the number by 1 million then you can understand the sales very easily rather if the whole number is just printed out

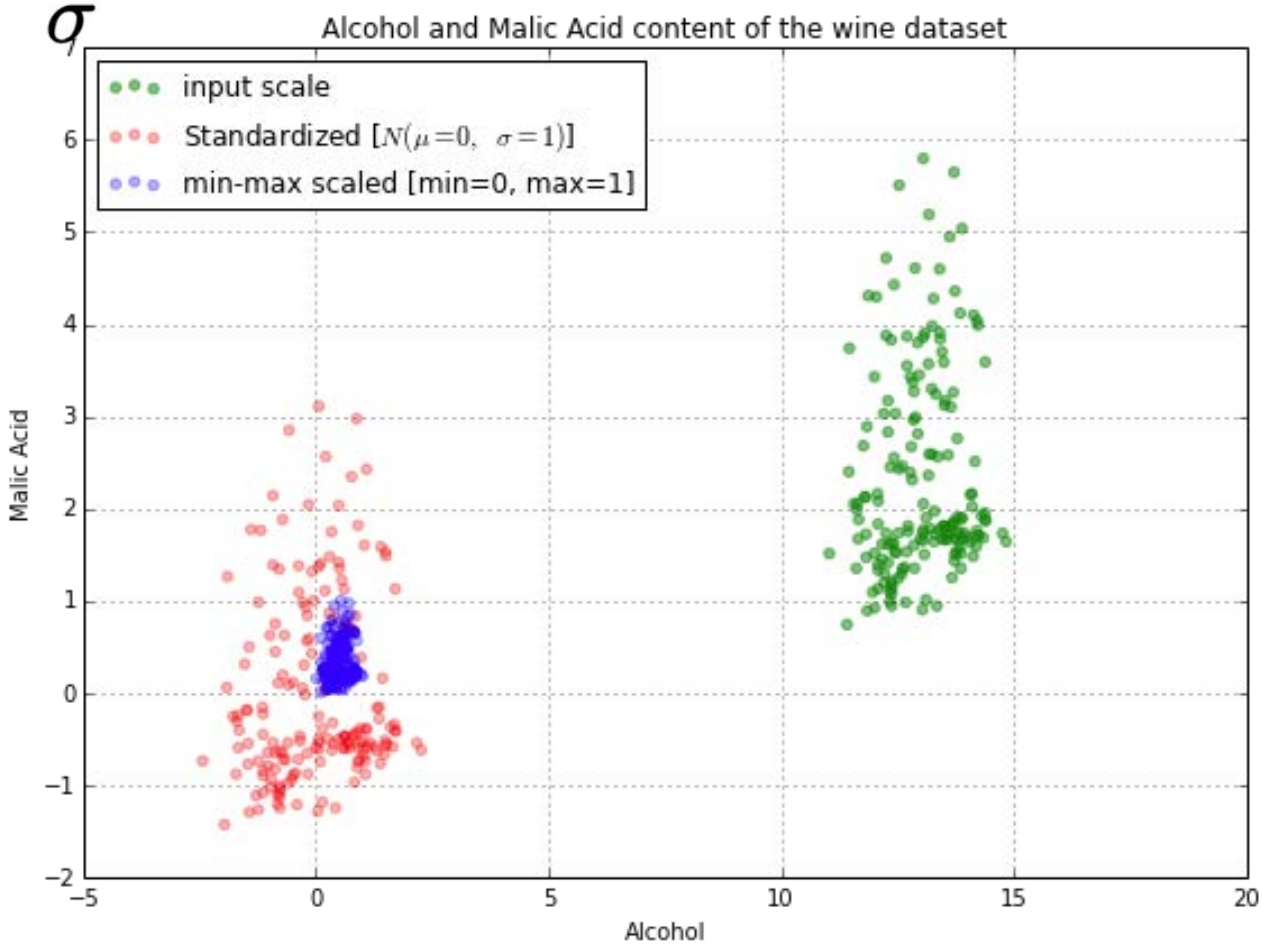
Scaling is performed for a better interpretation of the data and to understand which variable is contributing most to the target variable the main variable in the independent variables.

Normalized Scaling: in this type of scaling the variable is scaled to the range of [0 1] and it is done by subtracting the minimum from the variable and then dividing it by the difference of maximum and minimum
the problem with this type of scaling the data also loses variance and some information due to it

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

Standardized Scaling: in this type of scaling the mean is 0 and the standard deviation is 1. in here we subtract the variable with the mean and divide it by standard deviation
the problem with standardization is after standardization still the data could be uninterpretable

$$x' = \frac{x - \bar{x}}{\sigma}$$



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

Answer: VIF (Variance inflation Factor) is the metric to test the multicollinearity in the data means whether the data is highly correlated or the one or more variable together can totally explain the data
and its formula is

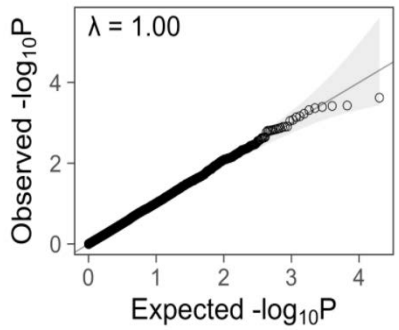
$$VIF = \frac{1}{1 - R_i^2}$$

Here the r is the r-squared value that represents total variation in a variable explained by other variables

If the VIF is equal to infinity then that means the variable in question can be totally explained by one or more variables hence is redundant and should not be used data modeling.

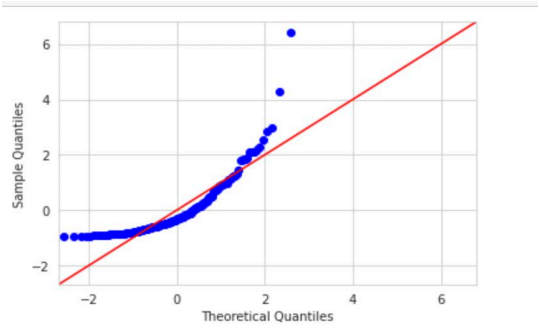
6) What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression?

Answer: The Q-Q plot is used to determine whether a variable belongs to a particular distribution or not. First you take certain amount of values at different quantiles from your variable and you take the same value of quantile from the distribution and plot them against each other if they fit on a straight line then they belong to the same distribution and if they don't fit on the straight line then they do not belong to the same plot



As you can see the plot is fitting very well on a straight line this means the variable does belong to the theoretical distribution

y axis --> Observed value (variable)
x axis --> Expected value (distribution value)



This here indicates the underlying variable is not distributed as per the expectation.

We use Q-Q Plot in linear regression for two purposes:

1) to check whether all the independent variables are normally distributed as per the linear regression assumption
2) to check whether the variables are homoscedasticity and not hetroscedasiticity.