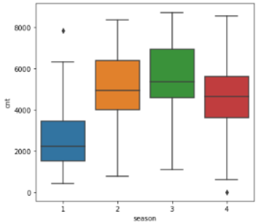
Linear Regression – Assignment

Assignment-based Subjective Questions

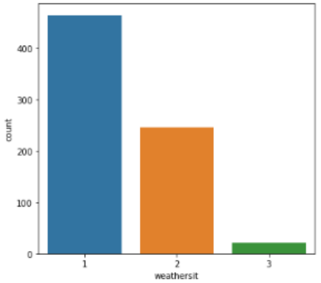
Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans. There are six categorical variables as below:

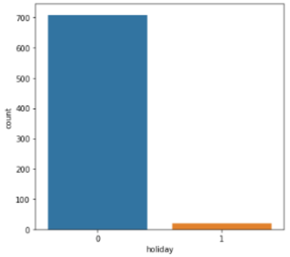
- **Season** – In this variable four season are there as follows
  - 1:spring, 2:summer, 3:fall, 4:winter
  - *We can observe from the graph that usage of cycle increase in summer and fall season*



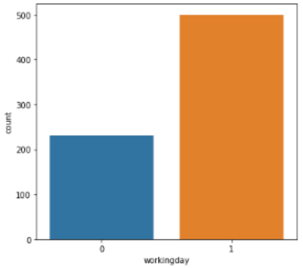
- **Weather situation**: In this variable four season are there as follows
  - *1: Clear, Few clouds, Partly cloudy, Partly cloudy*
  - *2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist*
  - *3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds*
  - *4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog*
  - *In this variable if wheatear is clear people are tend to use more cycle*



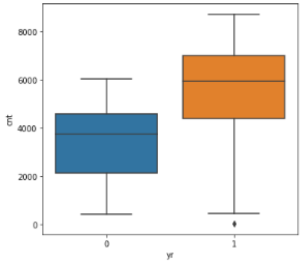
- **Holiday-** it is binary variable 0- no holiday and 1- holiday
  - *We have observe that there when there is no holiday people prefer to use more cycle*



- **Working day**
  - *People use more cycle on their working day*



- **Yr-(year)** : Year has two value 0 shows 2018 and 1 shows 2019 and *we can observe as company is growing old usage of cycle is increasing*



- **Weekday** : day of the week
  - *There is no pattern in the in this variable all the day looks*

Q2. Why is it important to use *drop_first=True* during dummy variable creation?

Ans. To explain the variance of the target variable, we need to create a dummy variable because the model/machine doesn't understand categorical variables. If we don't remove one of the dummy variables, the dataset will become multicollinear, and we don't need to create unnecessary columns if the machine can understand the data with fewer columns. As an example, If there is only one column of gender with the two values male and female, if we create two columns, they will be collinear, either a person being male or female. Therefore, we only create one column of female with the values 0 and 1, 0 indicating that the person is male and 1 indicating that the person is female.

Hence , we always need to use drop_first-=true while creating dummy variable

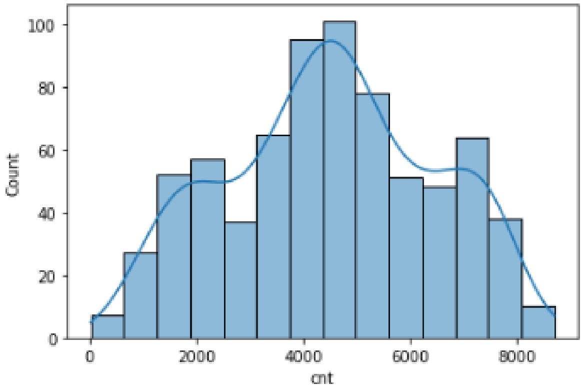
Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans. If I look the pair plot the most corelated variable is ***temp*** and ***atemp*** with target variable, (*I have not considered the casual and registered variable as they are the function of target variable*)

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

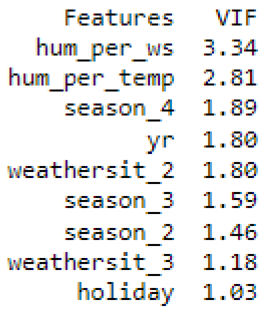
Ans. Below are the validation I have done during training of my model (check this link):

- The target variables was normally distributed , as it is showing bell curve , it shows the normality of the data



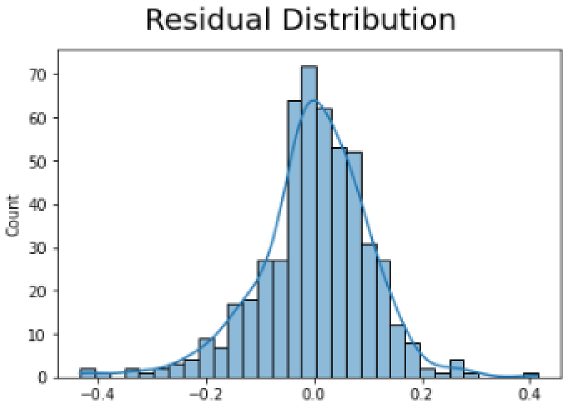
Q3. continue answer of Q3

- There should not multicollinearity in my model variables :

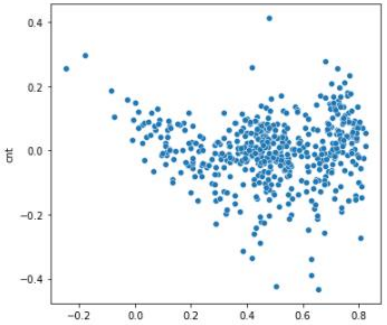


```
     Features   VIF
    hum_per_ws  3.34
  hum_per_temp  2.81
      season_4  1.89
            yr  1.80
  weathersit_2  1.80
      season_3  1.59
      season_2  1.46
  weathersit_3  1.18
       holiday  1.03
```

- Error term should be normally distributed

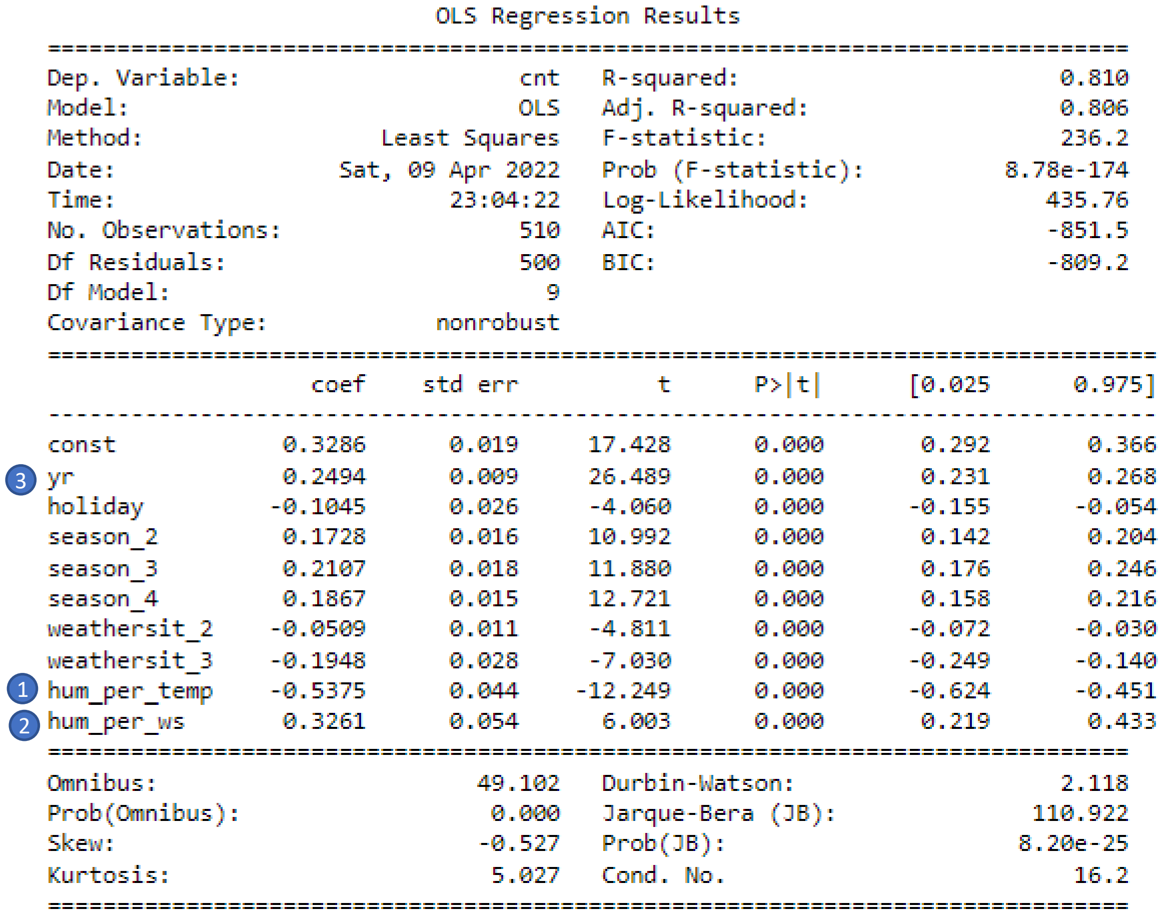


- the residuals have constant variance at every level of x



Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans. There are three variables which are contributing significantly towards explaining the demand of the shared bikes



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.810
Model:                            OLS   Adj. R-squared:                  0.806
Method:                 Least Squares   F-statistic:                     236.2
Date:                Sat, 09 Apr 2022   Prob (F-statistic):          8.78e-174
Time:                        23:04:22   Log-Likelihood:                 435.76
No. Observations:                 510   AIC:                            -851.5
Df Residuals:                     500   BIC:                            -809.2
Df Model:                           9
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const            0.3286      0.019     17.428      0.000       0.292       0.366
yr               0.2494      0.009     26.489      0.000       0.231       0.268
holiday         -0.1045      0.026     -4.060      0.000      -0.155      -0.054
season_2         0.1728      0.016     10.992      0.000       0.142       0.204
season_3         0.2107      0.018     11.880      0.000       0.176       0.246
season_4         0.1867      0.015     12.721      0.000       0.158       0.216
weathersit_2    -0.0509      0.011     -4.811      0.000      -0.072      -0.030
weathersit_3    -0.1948      0.028     -7.030      0.000      -0.249      -0.140
hum_per_temp    -0.5375      0.044    -12.249      0.000      -0.624      -0.451
hum_per_ws       0.3261      0.054      6.003      0.000       0.219       0.433
================================================================================
Omnibus:                       49.102   Durbin-Watson:                   2.118
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              110.922
Skew:                          -0.527   Prob(JB):                     8.20e-25
Kurtosis:                       5.027   Cond. No.                         16.2
==============================================================================
```

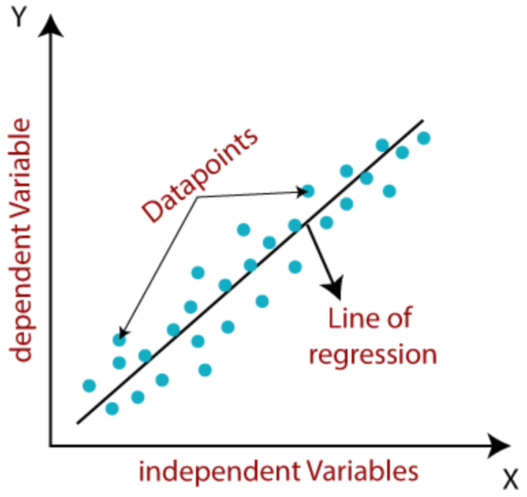
Linear Regression – Assignment

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Ans: One of the most basic and widely used Machine Learning methods is linear regression. It's a statistical technique for performing predictive analysis. Sales, salary, age, product price, and other continuous/real or numeric variables are predicted using linear regression. The linear regression algorithm, as the name implies, reveals a linear connection between a dependent (Y) and one or more independent (X) variables. Because linear regression reveals a linear connection, it determines how the value of the dependent variable changes as the value of the independent variable changes.

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



$$\Upsilon = \alpha_\theta + \alpha_1\chi_1 + \cdots + \alpha_\eta\chi_\eta + \varepsilon$$

Υ= Dependent Variable (Target Variable)
χ = Independent Variable (predictor Variable)
α_θ = intercept of the line (Gives an additional degree of freedom)
α_1 = Linear regression coefficient (scale factor to each input value).
ε = random error

The values for χ and Υ variables are training datasets for Linear Regression model representation

Types of Linear Regression

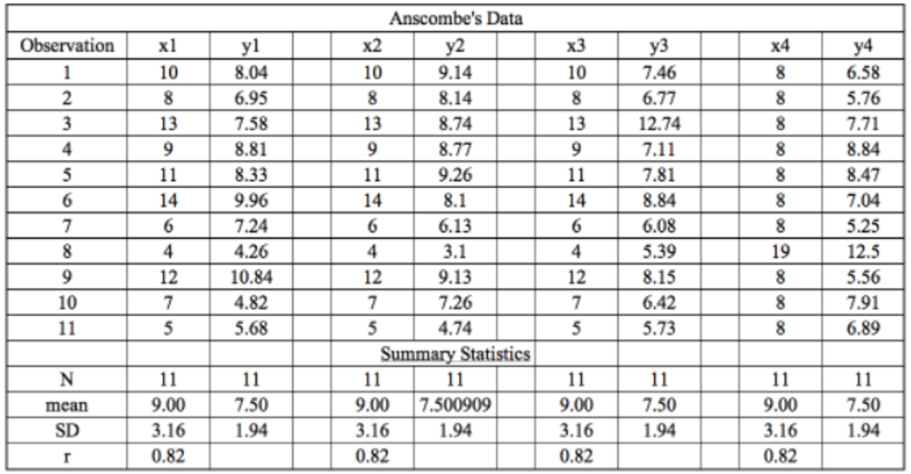
Linear regression can be further divided into two types of the algorithm:

- **Simple Linear Regression:** Simple Linear Regression is a Linear Regression approach that uses a single independent variable to predict the value of a numerical dependent variable.
- **Multiple Linear regression:** Multiple Linear Regression is a Linear Regression approach that uses more than one independent variable to predict the value of a numerical dependent variable.

Q2. Explain the Anscombe’s quartet in detail.

Ans: Anscombe's Quartet is a collection of four data sets that are essentially equal in terms of simple descriptive statistics, but have certain anomalies in the dataset that mislead the regression model if constructed. When displayed on scatter plots, they have extremely distinct distributions and appear differently.

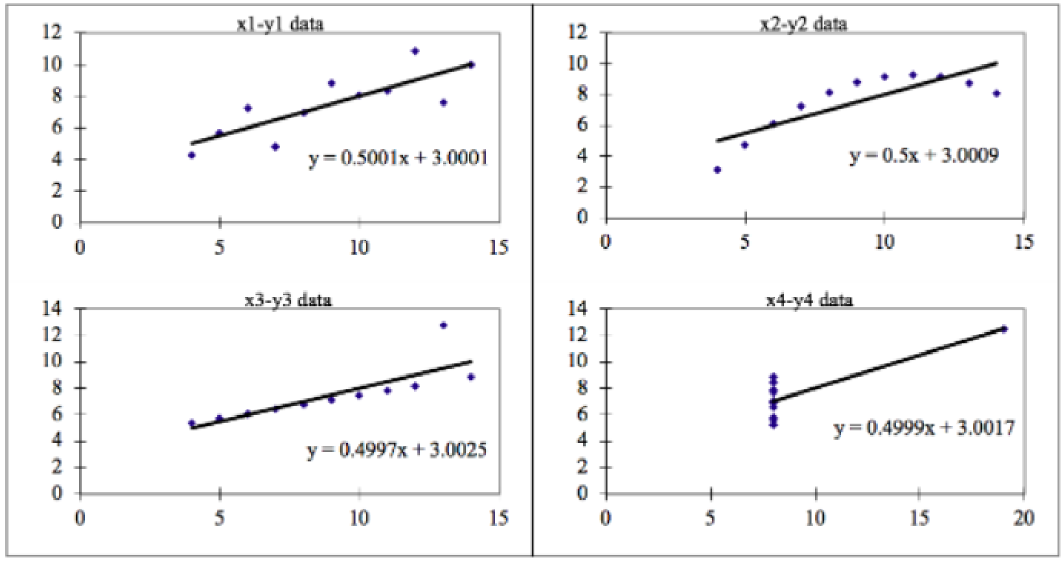
It was built by statistician ***Francis Anscombe in 1973*** to demonstrate the ***significance of plotting graphs before analyzing and modelling***, as well as the impact of additional data on statistical features. There are *four data set plots with virtually identical statistical observations and statistical information*, including variance and mean of all x,y points in each dataset.



					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	

The four datasets are as follows:
Dataset 1: this is a good fit for the linear regression model.
Dataset 2: due to the non-linear nature of the data, a linear regression model could not be fit adequately.
Dataset 3 shows the outliers in the dataset that the linear regression model cannot handle.
Dataset 4 illustrates the outliers in the dataset that the linear regression model cannot handle.

When these models are shown on a scatter plot, all datasets produce a different type of pattern that is unintelligible to any regression programmed misled by their differences, as seen below:



Conclusion:
 We've gone through the four datasets that were intentionally created to demonstrate the importance of data visualization and how it may mislead any regression system. As a result, before applying any machine learning algorithm to the dataset, all of the key characteristics must be shown in order to create a good fit model.

Q3. What is Pearson’s R ?

Ans: Pearson’s R is the method to identify the relationship between two variables , it tell the strength and direction of linear relationships between pairs of continuous variables. The value of this statistics lies between -1 to 1 (***-1 being high negative correlation and 1 means highly positive correlation between two continuous variable***)

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans: Scaling – Scaling is transformation of variables which brings the variables is same scale which helps ML models to converge faster and provide the results

Normalized scaling : when you want to bring the data in 0 to 1 range then we perform normalization of data one of the technique is MinMax scaling $\frac{X_i - Min_i}{Max_i - Min_i}$

Standardized scaling : when you wan to z – score standardization where we subtract mean from the value and divide it by standard deviation $\frac{X_i - \bar{X}}{\sigma}$

Normalization	Standardization
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
It is useful when we don’t know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans: When the variables are explaining all the variance of other variable or we can say the variables are 100% correlated then VIF is infinite as the formula of VIF is

$$\text{VIF} = \frac{\mathbf{1}}{\mathbf{1}-\boldsymbol{R}^{\mathbf{2}}}$$

If we see mathematically if the value of $\boldsymbol{R}^{\mathbf{2}}$ **is 1 or 100%** then the **VIF** value will be **infinity**

Q5. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans: Q-Q plots are used to analyze the distribution of random variables, whether they are Gaussian, Uniform, Exponential, or even Pareto. Simply by glancing at the Q-Q plot, you can identify what sort of distribution it represents.