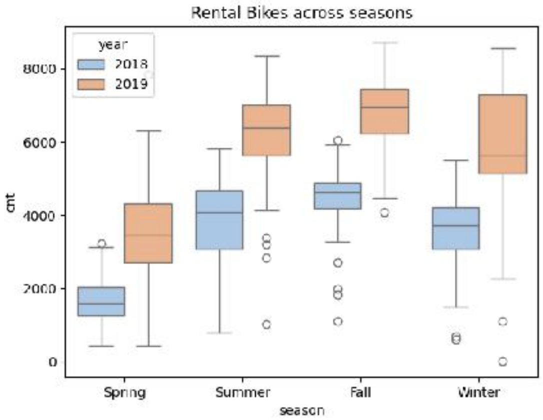
I Assignment Based Subjective Questions

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

Below are the inferences on the case study w.r.t categorical variables

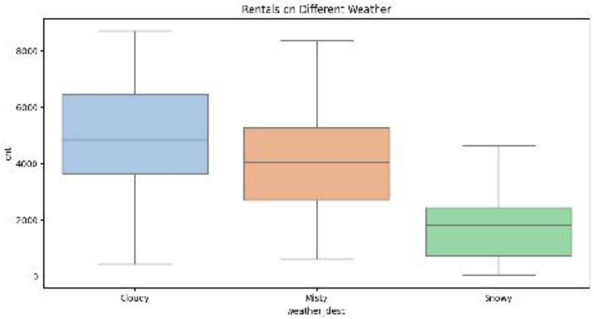
1. Seasons:

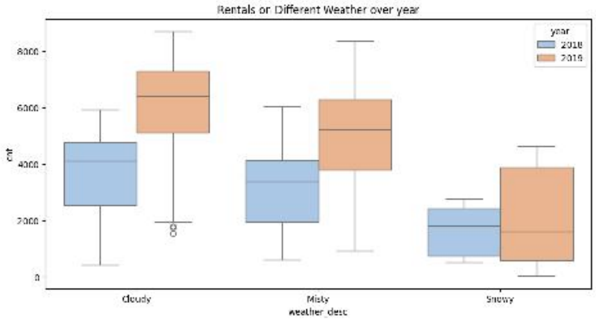
Out of the 4 seasons in the dataset , ‘Spring’ has less number of rental bikes compared to the other seasons both in 2018 and 2019. The Summer and Fall is where the count of rental bikes are high. The winter has performed slightly better in upper quantile areas compared to that of spring. .



2. Weather:

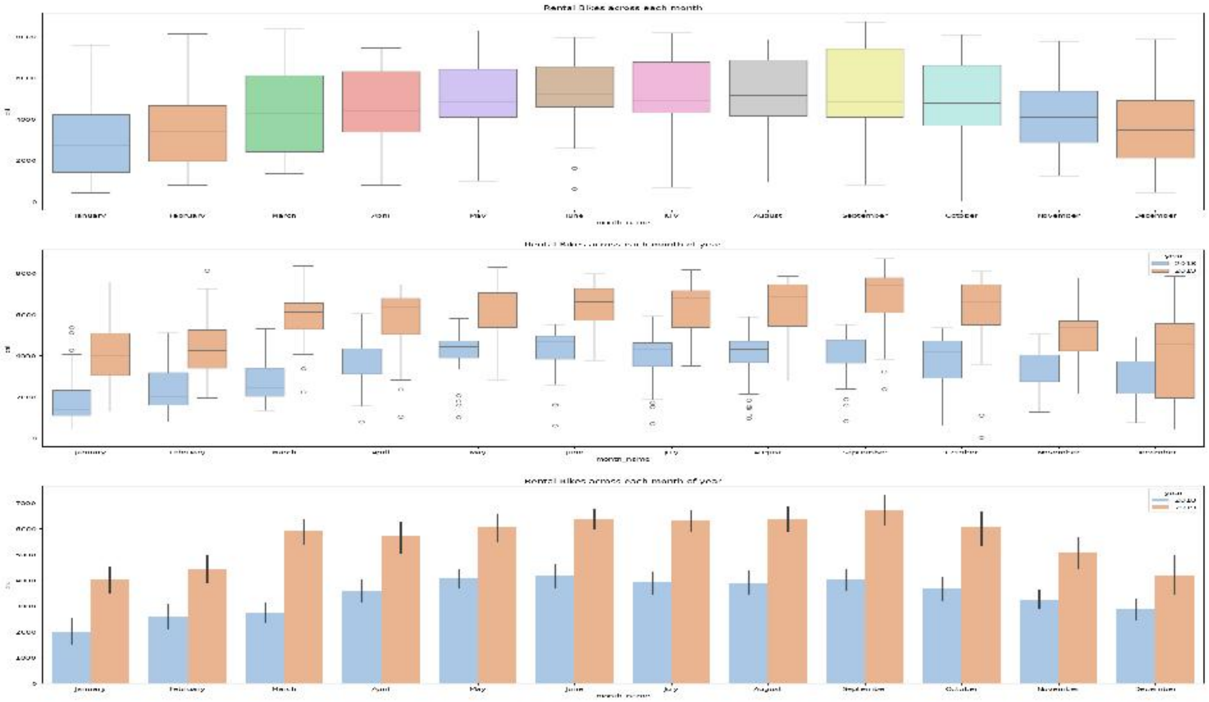
Cloudy and Misty weather seem to be aiding the increase in the number of Rental Bikes which again is coherent with Season ‘Summer’ and ‘Fall’ seasons





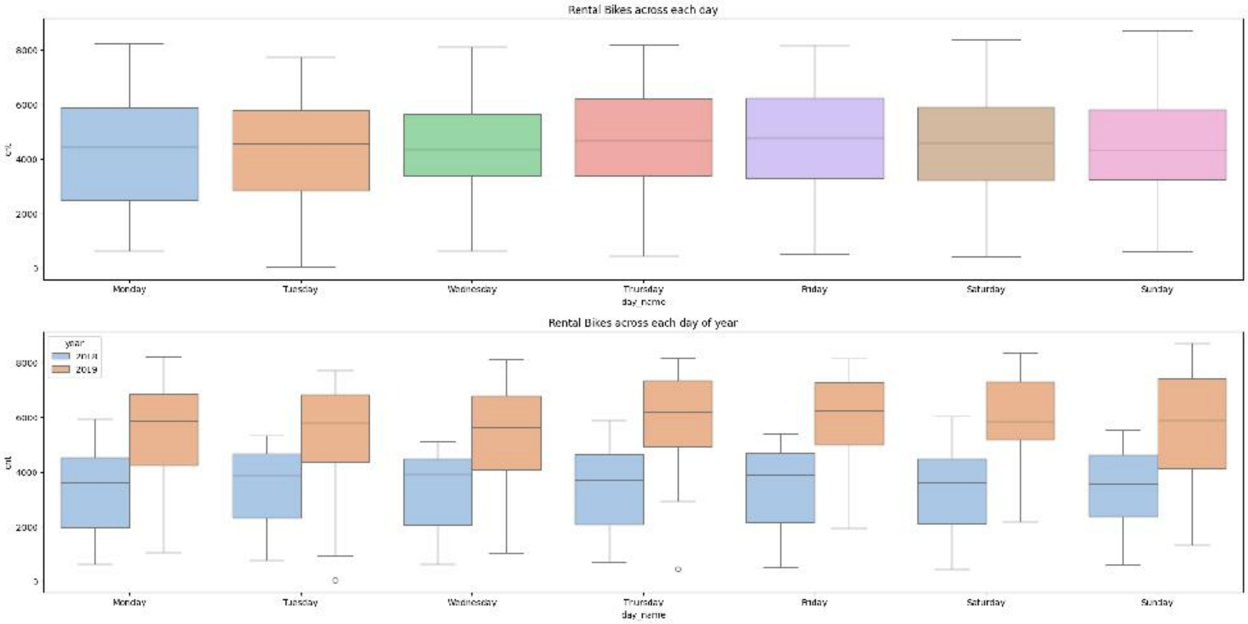
3. Month:

Mid months starting from June to September seem to have increased the count of Rental Bikes which seems coherent with the season data as Summer and Fall is observed during those months. December sees a decline which seems to be a holiday season and year end.



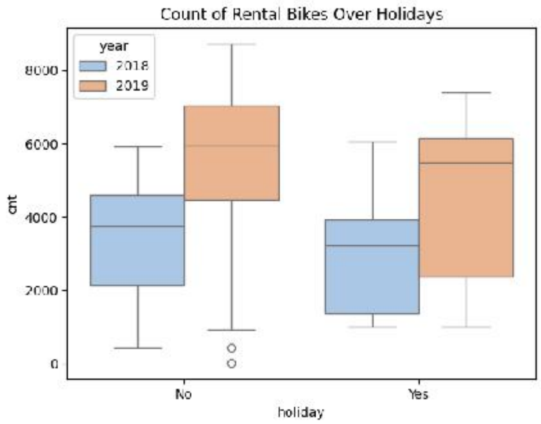
4. Week:

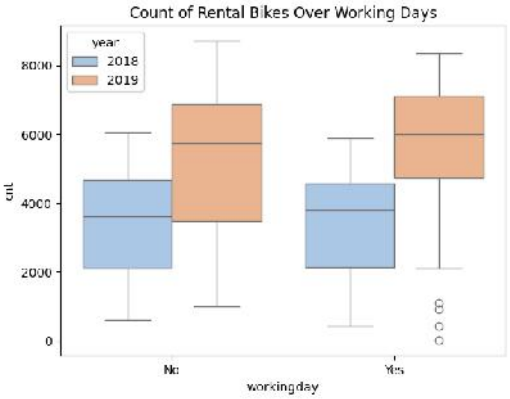
There is not much of a difference w.r.t to the days of week in 2018. The spread is equal across days of week. In 2019 Thursdays ,Fridays and Saturdays seem to show a slight uptick. The Sundays of 2019 perform poor in the lower quartile region.



5. Working day:

The working day shows higher momentum in rental bikes compared to that of weekends. This could mean rentals are being used for office commute.



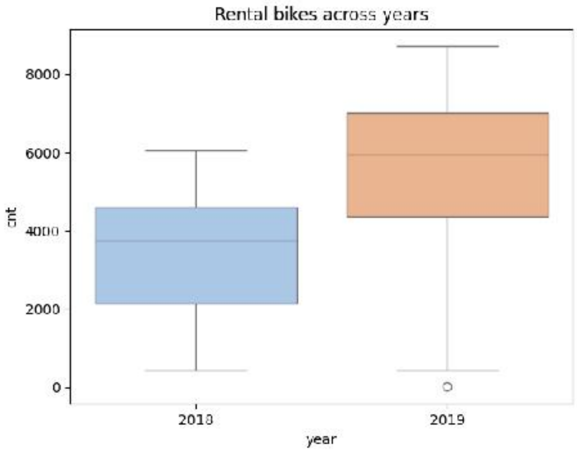


6. Holiday:

Holidays are having lesser traction compared to that of Weekdays , which proves the above positive correlation with ‘working day’.

7. Yr:

The year 2019 shows significant improvement in the rental bikes compared to that of 2018.



The Linear model derived at the end of the case study is explained by below categorical variables along with the coefficients

$$5348.48 * atemp + 2024.18 * yr + 785.48 * Cloudy + 634.88 * September + 706.05 * Summer$$

$$+ 1039.87 * Winter - 589.49 * holiday + 268.78 * Saturday$$

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

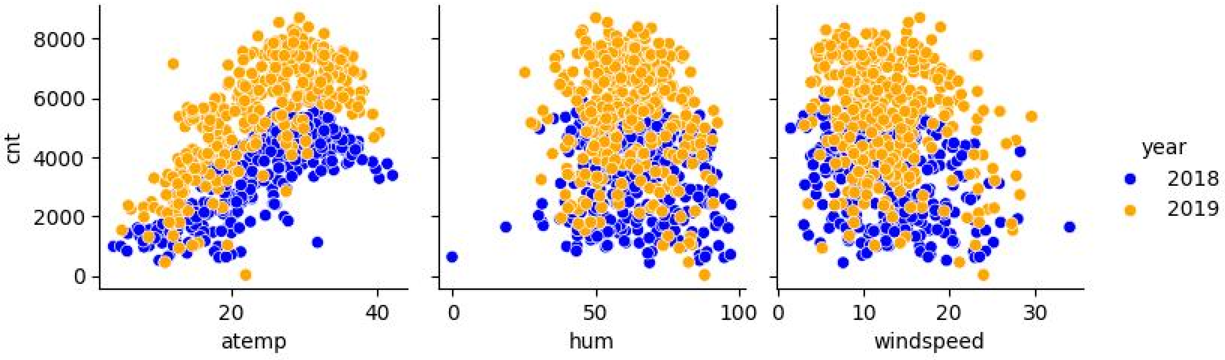
(Dummy variable creation or) One hot encoding is a way of encoding categorical variables into multiple dummy variables represented in a binary forms (0 or 1). For a column represented by k categorical variables, k-1 dummy variables are sufficient to explain the model. The convention is to drop the first variable from the k different variables . The python library pandas has get_dummies which has a parameter ‘drop_first’ which enables it to drop the first dummy variable from the list of generated ones.

Below is a sample usage from case study

```
dummy_seasons=pd.get_dummies(data['season'],dtype=int,drop_first=True)
```

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

Looking at the scatterplot and heatmap of the numerical variables in the case study , the feeling temperature , the column (‘atemp’) shows the highest correlation with 0.63 as correlation coefficient. The regplot also shows an observable linear slope between ‘atemp’ and ‘cnt’ columns with a pearson coefficient of 0.63



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

The Assumptions of Linear Regressions are as belows

- The variables analyzed should show linear pattern with the target variable

As seen in the scatterplot (in the pairplot) of numeric variables

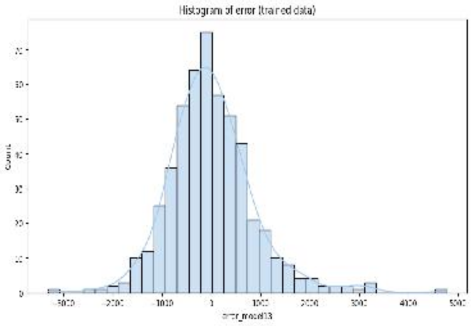
1. The below variables show linear upward trend
   - ➢ Season
   - ➢ Year
   - ➢ Mnth
   - ➢ Weekday

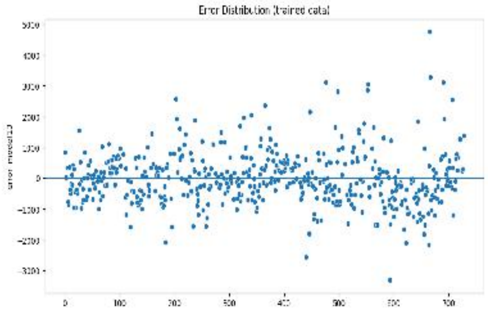
- ➢ atemp
- ➢ temp

2. The variables hum, windspeed shows a slightly downward trend
3. The working day, Weathersit, holiday variables differs in their spread with their values

- The error distribution follows a normal distribution with mean around 0.

  As seen in the residual analysis, the histogram of errors y-y_predicted (for model 13) shows a normal distribution with mean centered at 0.





- The error distribution does not follow any pattern
- Homoscedasticity (Constant variances)
  - ○ As seen in the residual analysis , the scatter plot does not follow any pattern
  - ○ Also the variances are fixed around the mean and does not show any increasing trend.

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

As seen from the model

- The feeling temperature (atemp)
- Summer season (Summer)
- Clear / Cloudy / Partly Cloudy weather (weathersit)

seem to explain the demand very well.

Boombikes should plan ahead procuring more bikes as the demand would shoot up during the summer season in mid months (June to September) in the upcoming calendar year.

II General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression:

Linear regression explains the relationship between dependent (target) variable and the independent variable in a dataset identified by a best fitting line with a minimal cost function.

y=βx+c explains the relationship between x (the independent variable) and y(the target variable) where β is the coefficient or slope and c is the intercept.

Incase of more than one independent variable , the best fit line becomes a hyperplane and the equation becomes

*y=β1*x1 + β2*x2 + β3* x3 + β4*x4 + ...βm*xn + c*

$y=\Sigma\ \beta jxi + c$ where i=1,2..n and j=1,2..m

where xi, x2.., xm defines m independent features describing the dependent variable y.
β1, β2 , β3 are the coefficients of each feature c is the intercept.

The above can be represented in a matrix form as

$$\begin{bmatrix} Y1 \\ Y2 \\ Y3 \\ . \\ . \\ YN \end{bmatrix} = \begin{bmatrix} X11 & X21 & . & XM1 \\ X12 & X22 & . & XM2 \\ X13 & X23 & . & . \\ . & . & . & . \\ . & . & . & . \\ X1N & X2N & . & XMN \end{bmatrix} . \begin{bmatrix} \beta 1 \\ \beta 2 \\ \beta 3 \\ . \\ . \\ \beta M \end{bmatrix} + \begin{bmatrix} c1 \\ c2 \\ c3 \\ . \\ . \\ cN \end{bmatrix}$$

The coefficient β *is given by*

$\beta = (X^T X)^{-1} (X^T Y)$

Residuals and R-square

The residuals or errors are represented by y-yi which is the difference between the actual y and y fitted.

Sum of squares of residuals is given by RSS = Σ(*y-yi)*^2 which is also the cost function f(β)

R^2 = 1-(RSS/TSS).

R^2 explains the % of variance explained by the independent variables with the target variable..

This is also called the OLS (order of least square) technique.

Minimizing the cost function

Minimizing the cost function are carried out by

1. Closed Form
    a. Taking the partial derivative of the cost function $f'(\beta)$ and equating to 0
2. Iterative form : Gradient Descent
    a. Taking the partial derivative of the cost function $f'(\beta) = \delta f(\beta)/\delta\beta$
    b. Finding θ_0 such that $\beta_0 = \beta - \alpha * f'(\beta)$ where α is the learning rate.
    c. Repeating the steps a & b till convergence $\beta_{optimal}$.
    d. $\theta_{optimal}$ gives you the best possible coefficients minimizing the cost function

Following are the assumptions for Linear Regression :

1. There should be a linear relationship between the dependent variable y and the independent variable (x1,x2…)
2. The errors should be normally distributed with a mean centered around 0.
3. The errors should not exhibit any pattern
4. The variance of the errors should be constant. (Homoscedasticity)

Factors to consider

1. Overfitting
2. Multicollinearity
3. Feature Selection

Advantages/Disadvantages:

1. Linear Regression can be helpful in prediction,forecasting with the existing dataset.
2. Extrapolation would be a challenge.
3. It is helpful in identifying the driver variables and the correlation , however linear regression cannot explain causation.

2. Explain the Anscombe’s quartet in detail.

Anscombe’s quartet is a group of 4 different dataset which has the same variances , mean , standard deviations and regression line, but qualitatively different. This explains why the summary statistics are not sufficient and why we should use data visualization to understand more on the data.

Below is the Anscombe’s dataset.

Dataset I		Dataset II		Dataset III		Dataset 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

Summary Statistics of above dataset:

1. The mean of the above data set for each x and y lies at 9 and 7.5 Respectively
2. The standard deviation for each of the dataset lies at 3.031 and 2.03 respectively
3. The regression statistics for each of the dataset shows R^2 as 0.667 and Adjusted R^2 as 0.629

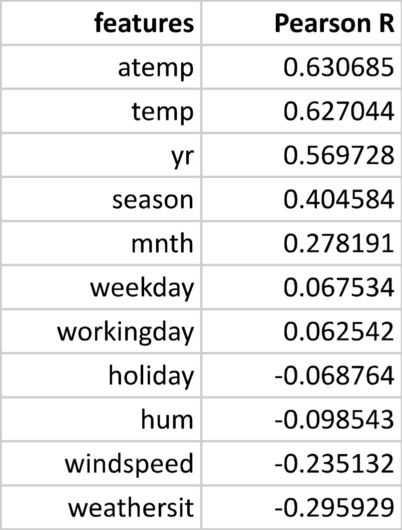
3. What is Pearson’s R?

Pearson R is the correlation coefficient between two variables of analysis. The pearson R is given by the formula

1. $\sum_{i=n}^{i=0} (xi - \hat{x})(yi - \hat{y}) / \sqrt{(xi - \hat{x})^2 (yi - \hat{y})^2}$

2. $(n\sum xy - \sum x \ \sum y \ ) \ / \ (n\sum(x)^2 - (\sum x)^2)(n\sum(y)^2 - (\sum y)^2)$

The Pearson coefficient ranges between -1 to 1 . Negative values indicate negative correlation. Positive value indicates positive correlation. Below are the coefficients from the case study dataset.



features	Pearson R
atemp	0.630685
temp	0.627044
yr	0.569728
season	0.404584
mnth	0.278191
weekday	0.067534
workingday	0.062542
holiday	-0.068764
hum	-0.098543
windspeed	-0.235132
weathersit	-0.295929

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

1. Feature scaling is required to bring the feature(s) in the dataset to the same scale.

2. Feature scaling helps in
   a. Handling the outliers within each feature in comparable scale

b. In faster convergence on optimized β in algorithms such as gradient descent
c. Prevents Numerical instability because of underflows or overflows in numerical values
d. Equal importance is given for each of the feature during the learning process
e. Prevents a single feature from dominating the model as such.

3. Difference between Normalized scaling and Standardized Scaling

Normalization:

Normalization is a method of Feature Scaling where Features are scaled within the range of 0 and 1. Example of Normalized scaling is MinMaxScaling which uses Min and Max of each feature and scales it based on

$$MinMax\,Scaling = (x - x_{min})/(x_{max} - x_{min})$$

Normalization usually handles the outliers as the outliers also falls within the same range of 0 and 1

Standardization:

Standardization is based on the centralized tendencies where the data of each feature is dispersed around $\mu = 0$ with a standard deviation of 1. This is explained by

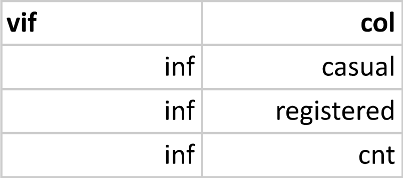
$$Standardized\ Scaling = (x - \mu)/\sigma)$$

Standardization does not affect the distribution as such. The outliers will remain outliers as such.

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

VIF given as $(1/1 - r^2)$ can go to infinity if r^2 is completely explained by the variables analyzed i.e (when $r^2 = 1$). This means that the predictor variable is completely collinear with other regressor variables in the dataset.

For eg) In the given dataset the two predictor variables ‘casual’ and ‘registered’ can completely explain the target variable ‘cnt’ as casual + registered = ‘cnt’. The models r^2 goes to 1 and VIF for each of them shows inf



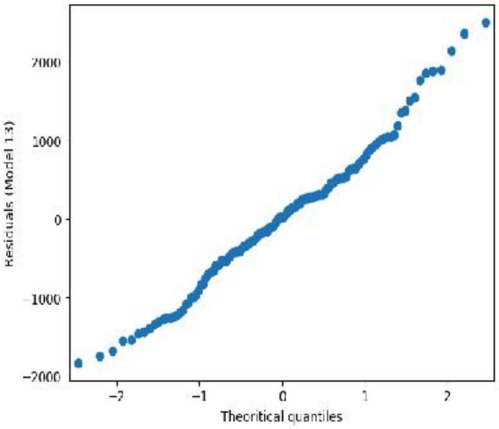
vif	col
inf	casual
inf	registered
inf	cnt

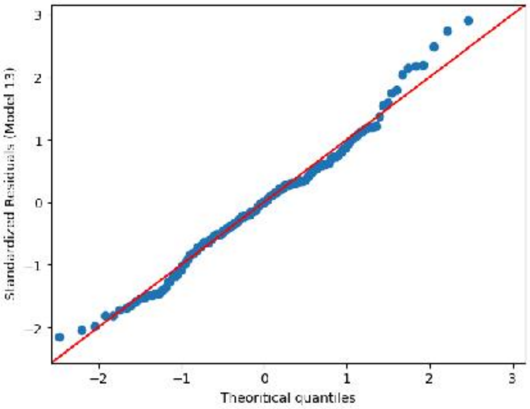
6.What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Q-Q plot plots the quantile distribution of a given dataset with the Theoretical distribution (Normal, exponential,uniform etc..). It gives you an understanding of whether the dataset follows a particular probability distribution. It also helps to identify if the two samples are from the same population and if the x_values are greater than y_values or vice versa.

In Linear Regression, as the assumption states residuals or the prediction error should follow a normal distribution, where Q-Q plot can be used as such. The theoretical distribution (x_axis) becomes the normal distribution and the y_axis becomes the residuals , if the residuals follow normal distribution , it will overlap the theoretical distribution.

In python Statsmodel comes with q-q plot capability. Below is a sample of residuals following normal distribution from the case study. (First one without standardizing the scales , second one with standardized scales.)


Residuals (Model 13)
Theoritical quantiles


Standardized Residuals (Model 13)
Theoritical quantiles