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?

Ans:

Categorical variables represent a qualitative method of scoring data (i.e. represents categories or group membership). They can be represented as binary values or multi-way(3 or more). Identified Categorical Variables in the Dataset are, Season, Year, Month, Holiday, Weekday, Workingday, weathersit.

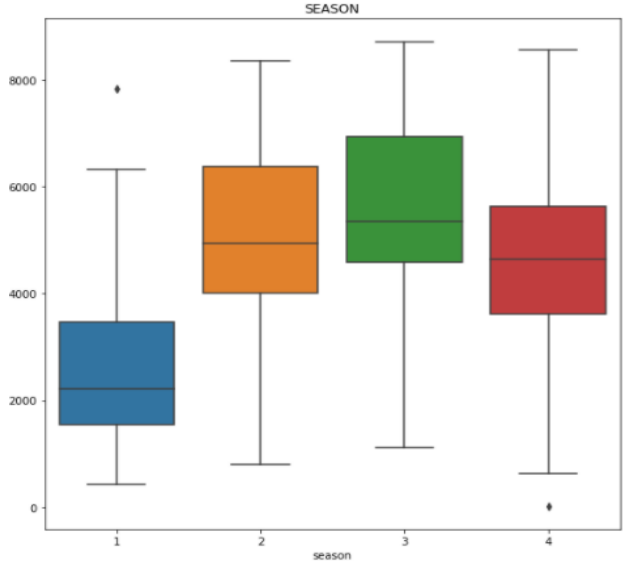
Dependent Variable in The Bike Renting Dataset is cnt(casual and registered are part of target.)

Season(multi-way**)**:

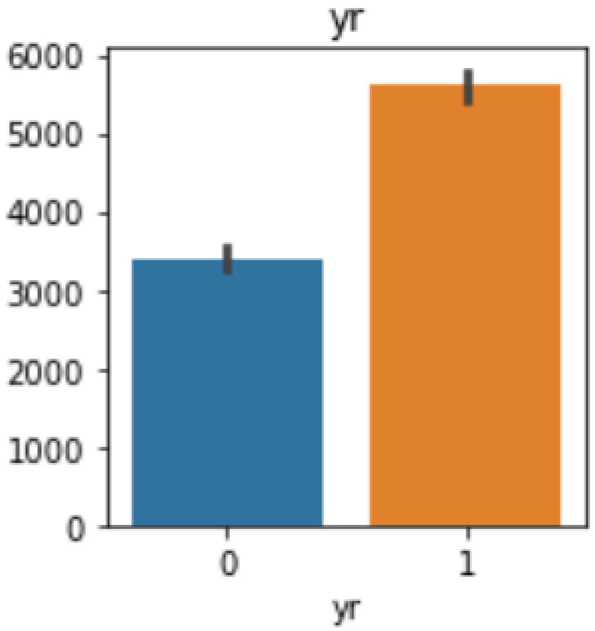
Values represented as ,
1:spring, 2:summer, 3:fall, 4:winter
demand is not as high as summer and fall in spring and winter

Snowfall ,temparature or other reasons specific to season can affect the demand of rentals.

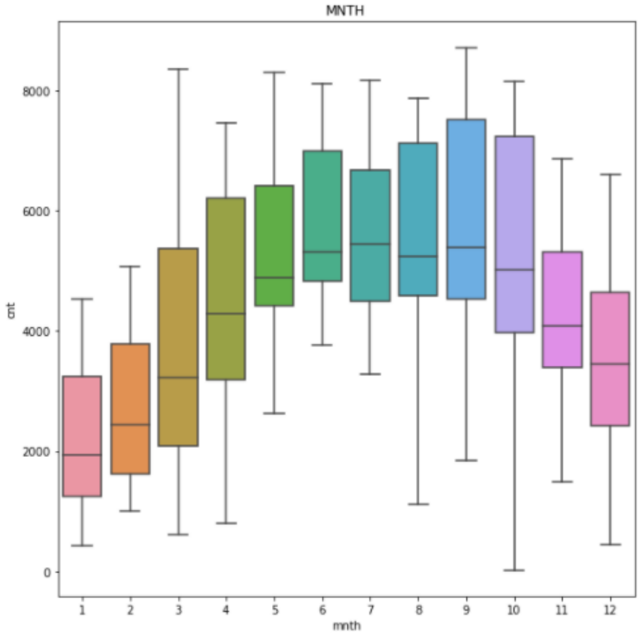
Example : On a snowfall and rain can reduce the demand



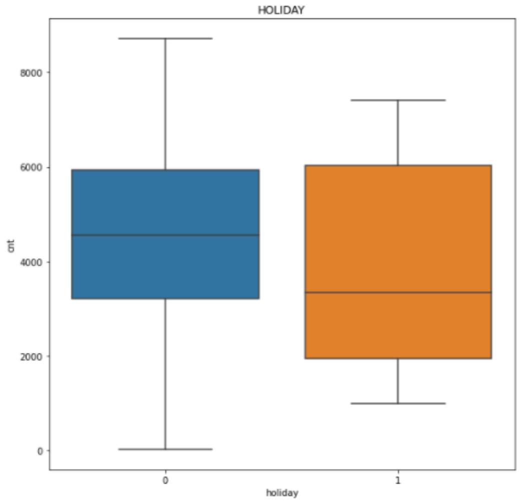
Year(binary) : Maybe demand can be increased or decreased with time. Due to business decisions and marketing. Demand Increased with year

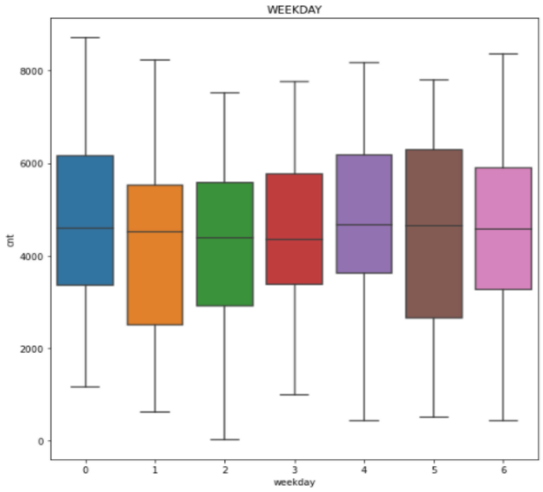


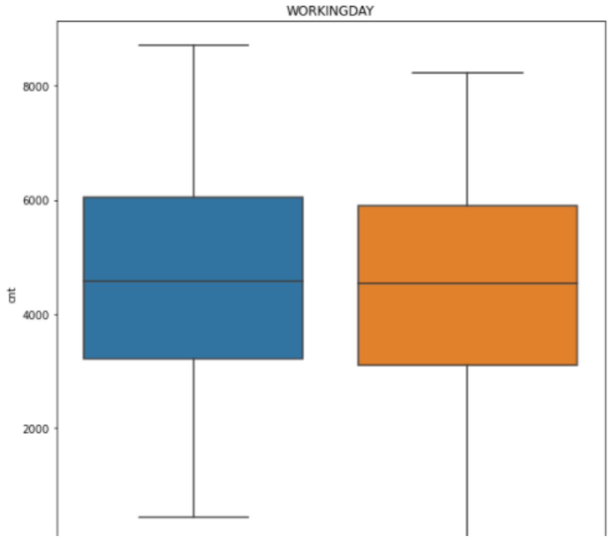
Month: temperature in the month , holidays in the month, or marketing campaigns can affect the demand. In the dataset demand varied with month november, december, jan, feb demand is less



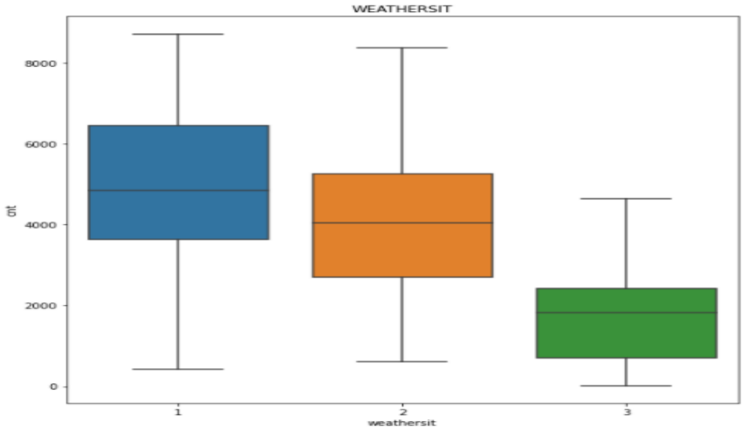
Holiday, Weekday, Workingday: Variables convey how the demand is on holiday and working day. Demand on Working day is bit higher than holiday







Weathersit: Weather has a significant effect on demand. May be demand can be lower on bad weather condition example heavy snowfall



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

Let’s Take values in the Categorical column **Season** and create a dummy variable for this column.

	spring	summer	fall	winter
spring	1	0	0	0
summer	0	1	0	0
fall	0	0	1	0
winter	0	0	0	1

spring represented 1000, summer as 0100, fall as 0010, winter as 0001.
This same message we can convey using 3 variables.

	summer	fall	winter
spring	0	0	0
summer	1	0	0
fall	0	1	0
winter	0	0	1

We can represent columns with n categories with help of (n-1) dummy variables.

In the above scenario we removed the spring column and represented it as 000.
The same result can be achieved by deleting any of the columns. It need not be a first column.

	spring	summer	winter
spring	1	0	0
summer	0	1	0
fall	0	0	0
winter	0	0	1

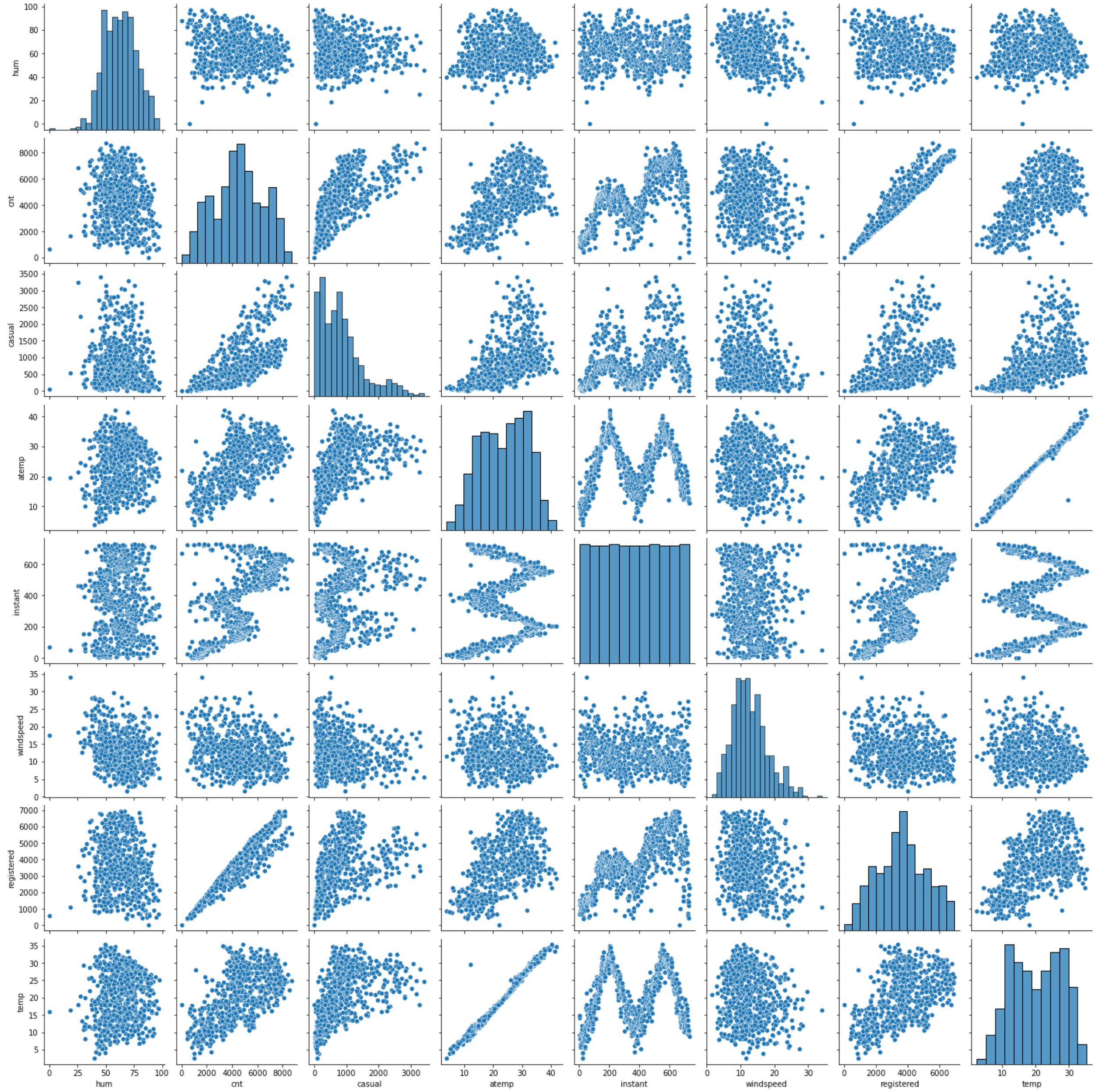
So, To represent n categories we need (n-1) dummies. But ‘n’ number of dummies gets created when we use pd._get_dummies() to create dummies. So, we need to drop any one of the Dummies from the data set to reduce correlation among dummy variables. For Simplicity we are removing the first dummy using drop_first=True.

Hence if we have categorical variables with n-levels, then we need to use n-1 columns to represent the dummy variables.

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

Ans:

Temp: Temperature has highest correlation with target variable(cnt). Correlation coefficient is 0.63

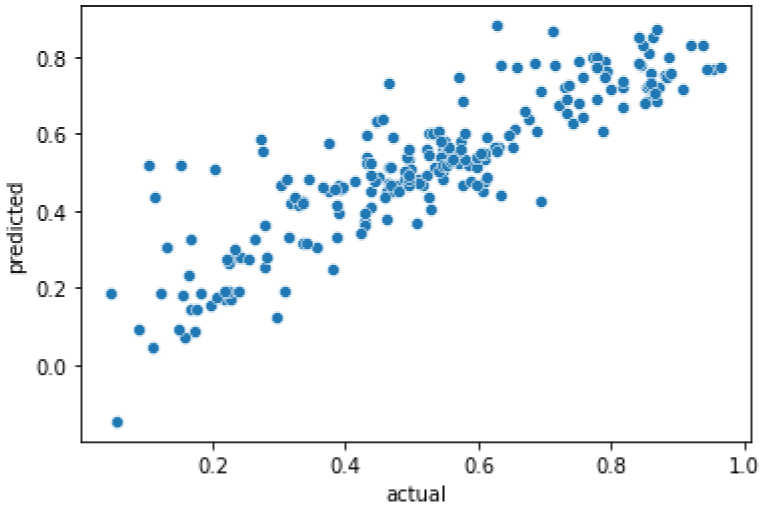


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans:

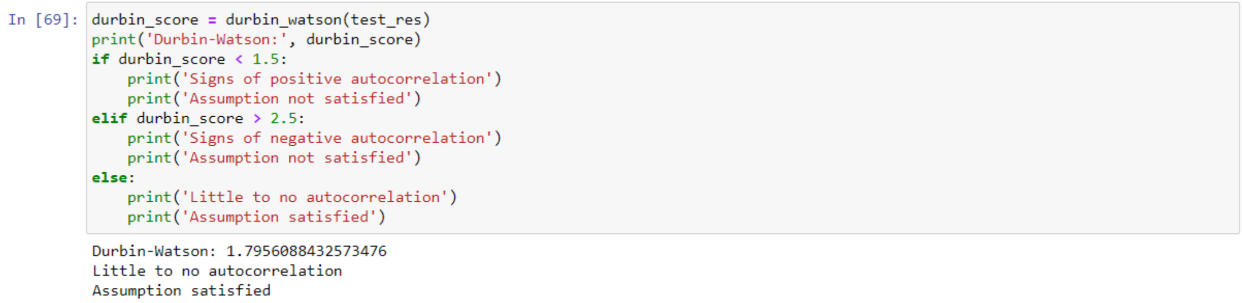
1. Linear relationship: There exists a linear relationship between the independent variable(x), and the dependent variable(y).

The Plot Between actual vs Predict is showing a diagonal line. Which indicate the Linear Relation



2. Independence: The residuals are independent. In simple terms, there is no correlation between consecutive residuals.

Durbin Score is 1.795 which indicating little or no correlation

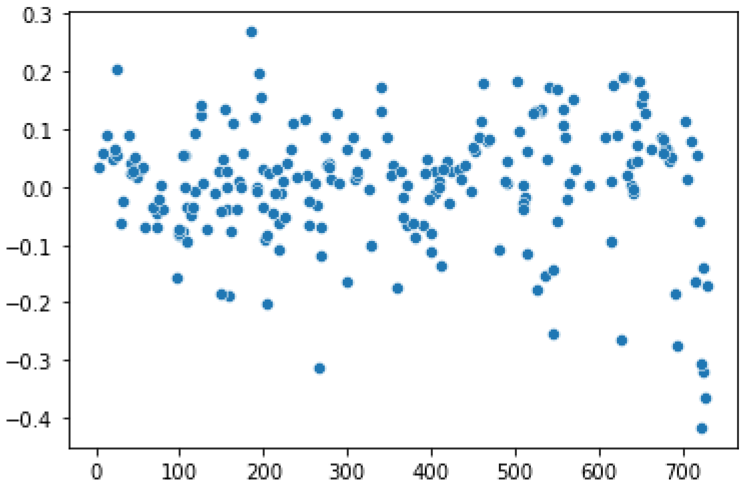


```
In [69]: durbin_score = durbin_watson(test_res)
         print('Durbin-Watson:', durbin_score)
         if durbin_score < 1.5:
             print('Signs of positive autocorrelation')
             print('Assumption not satisfied')
         elif durbin_score > 2.5:
             print('Signs of negative autocorrelation')
             print('Assumption not satisfied')
         else:
             print('Little to no autocorrelation')
             print('Assumption satisfied')
```

```
Durbin-Watson: 1.7956088432573476
Little to no autocorrelation
Assumption satisfied
```

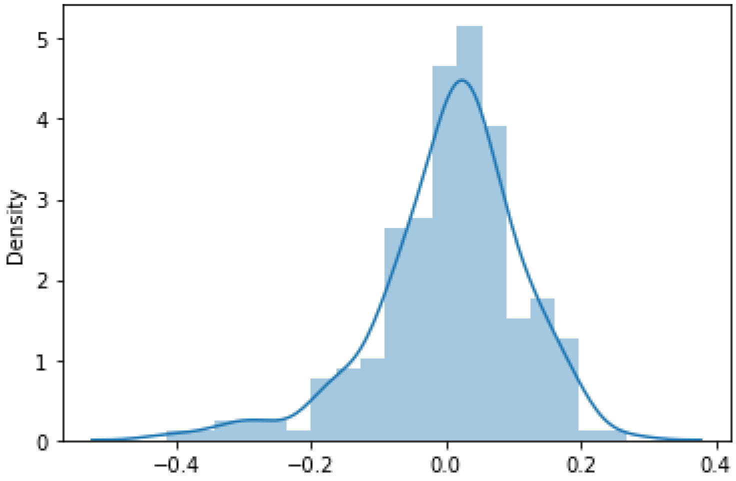
3. Homoscedasticity: The residuals have constant variance at every level of x.

Residuals are not have uniform or any pattern



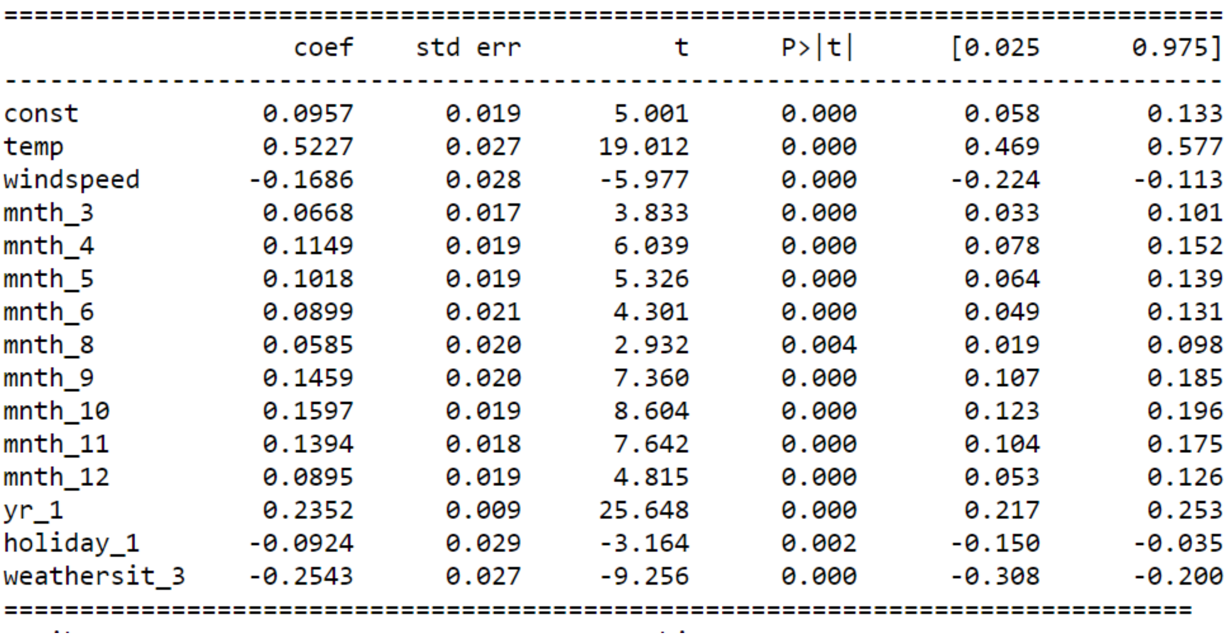
4. Normality: The residuals of the model are normally distributed

Distribution plot of residuals indicating normal distribution



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

Ans: Temp, Yr, Weather



```
===============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
const          0.0957      0.019      5.001      0.000       0.058       0.133
temp           0.5227      0.027     19.012      0.000       0.469       0.577
windspeed     -0.1686      0.028     -5.977      0.000      -0.224      -0.113
mnth_3         0.0668      0.017      3.833      0.000       0.033       0.101
mnth_4         0.1149      0.019      6.039      0.000       0.078       0.152
mnth_5         0.1018      0.019      5.326      0.000       0.064       0.139
mnth_6         0.0899      0.021      4.301      0.000       0.049       0.131
mnth_8         0.0585      0.020      2.932      0.004       0.019       0.098
mnth_9         0.1459      0.020      7.360      0.000       0.107       0.185
mnth_10        0.1597      0.019      8.604      0.000       0.123       0.196
mnth_11        0.1394      0.018      7.642      0.000       0.104       0.175
mnth_12        0.0895      0.019      4.815      0.000       0.053       0.126
yr_1           0.2352      0.009     25.648      0.000       0.217       0.253
holiday_1     -0.0924      0.029     -3.164      0.002      -0.150      -0.035
weathersit_3  -0.2543      0.027     -9.256      0.000      -0.308      -0.200
=============================================================================
```

General Subjective Questions

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

Ans:

Linear Regression is one of the basic machine learning algorithms based on supervised learning.It predicts a dependent variable value(y/target variable). This **regression** technique finds out a **linear** relationship between x (input) and y(output).

Here, Dependent variable is Continuous in nature. Independent variables can be either continuous or categorical variables.

- It is widely used for Forecasting and prediction
- It Guarantees Interpolation of the data, not extrapolation
- It implies correlation, not causation
- It is parametric regression

If the number of Independent variables are 1, then is it a Simple Linear Regression.

It can be represented as Y = m*x + b

Here , Y is the output or the prediction.

m is the slope or the "weight" given to the variable x.

x is the input you provide based on what you know.

b is the intercept. Essentially given 0 for your input, how much of Y do we start off with.

If the number independent variable > 1(assumes there is no relationship between x1 and x2). It is represented as Y = b0+b1X1+b2X2+...bpXp

Once The best fitting line is found(with least error score), we can validate our model using r2 error and we can perform residual analysis(Mean Squared Error or cost function) to find whether the model is biased or anything is missed.

Cost function:

$$\text{MSE} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}}_{\text{Mean}} \underbrace{(\overset{\substack{\text{Actual}\\\text{output}\\\downarrow}}{Y_i} - \overset{\substack{\text{Predicted}\\\text{output}\\\downarrow}}{\hat{Y}_i})^2}_{\text{Squares of the errors}}$$

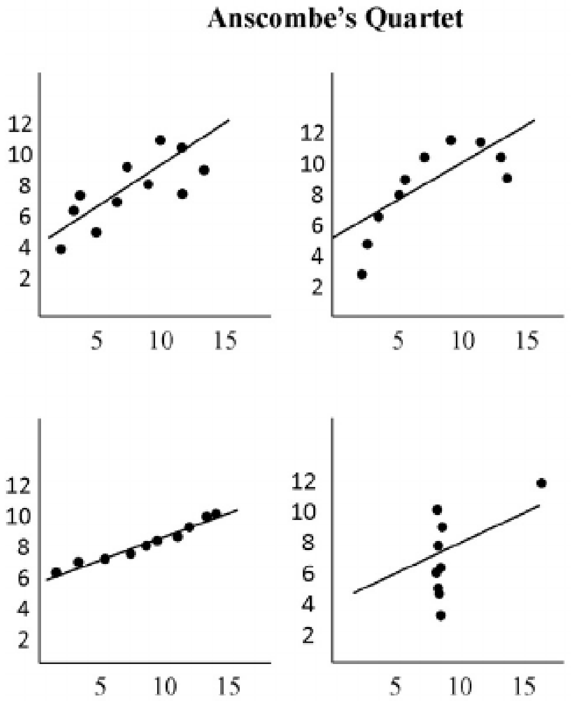
Assumptions of Linear regression:

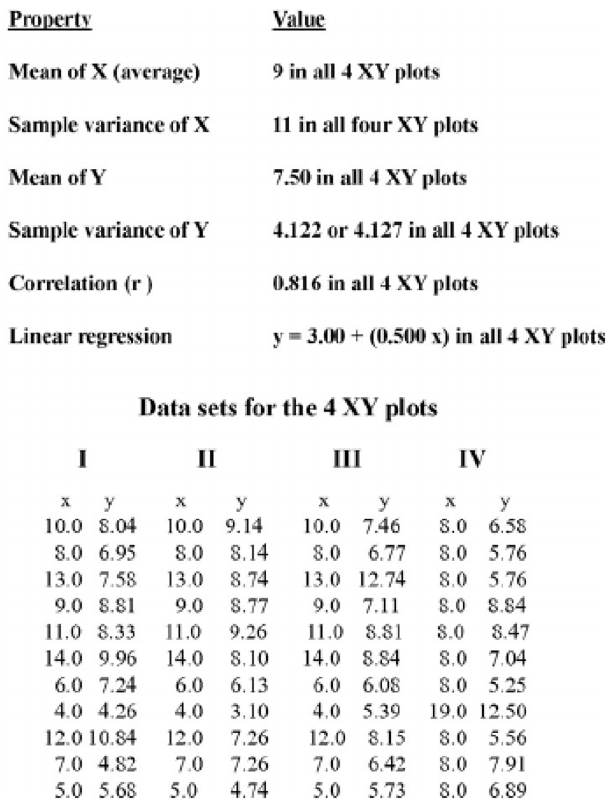
1. Linear relationship: There exists a linear relationship between the independent variable(x), and the dependent variable(y).
2. Independence: The residuals are independent. In simple terms, there is no correlation between consecutive residuals.
3. Homoscedasticity: The residuals have constant variance at every level of x.
4. Normality: The residuals of the model are normally distributed

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

Ans:

In 1973 statistician Frances Anscombe published a paper that contained 4 fictitious XY datasets plotted as below





Property	Value
Mean of X (average)	9 in all 4 XY plots
Sample variance of X	11 in all four XY plots
Mean of Y	7.50 in all 4 XY plots
Sample variance of Y	4.122 or 4.127 in all 4 XY plots
Correlation (r)	0.816 in all 4 XY plots
Linear regression	y = 3.00 + (0.500 x) in all 4 XY plots

Data sets for the 4 XY plots

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	5.76
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	8.81	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	7.26	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

He uses these datasets to make an important point that becomes clear if we blindly go about doing parameter estimation. If we observe parameters of 4 datasets, mean, sample variance, residuals are the same for 4 datasets, but they don't follow the same distribution.

Plot 01, 11 clearly don't have a linear relationship. Doing Linear Regression on this dataset is not ideal.

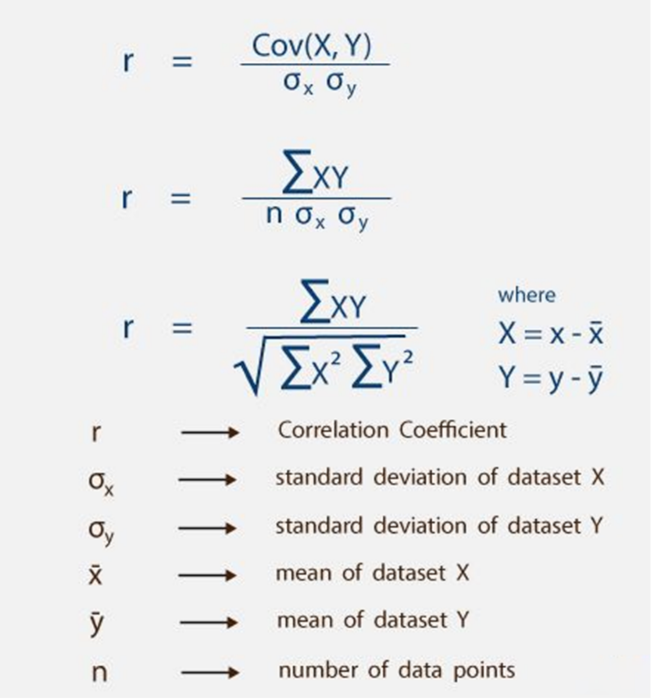
The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistical properties for describing realistic datasets.

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

Ans:

There are mainly three types of correlation that are measured. One significant type is Pearson's correlation coefficient. This type of correlation is used to measure the relationship between two continuous variables.
Gives the Strength and Direction of the relationship between two continuous variables.
Range lines in (-1, 1)
Useful in feature selection

Formula :



$$r = \frac{Cov(X, Y)}{\sigma_x \sigma_y}$$

$$r = \frac{\sum XY}{n \sigma_x \sigma_y}$$

$$r = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}$$

where
$X = x - \bar{x}$
$Y = y - \bar{y}$

r	→	Correlation Coefficient
σ_x	→	standard deviation of dataset X
σ_y	→	standard deviation of dataset Y
$\bar{x}$	→	mean of dataset X
$\bar{y}$	→	mean of dataset Y
n	→	number of data points

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

Ans:

Scaling: Scaling is bringing all feature values into a similar scale. Ex: age can be in 0 to 150 and salary can be in lakh. Salary values are much higher than age. So there is a chance that model can assume the salary has more weight.

Why We need Scaling: To weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Difference Between Normalized scaling and Standard Scaling:

Min-Max Scaling or Normalized Scaling: `X_new = (X - X_min)/(X_max - X_min)`
Standard Scaling: `X_new = (X - mean)/Std`

Normalized Scaling	**Standard Scaling**
Minimum and Maximum value are used in scaling	Standard Deviation is used in scaling
Used when features values in different scale	Used when we want to ensure mean=0 , standard deviation=1
Scale values between [0, 1] or [-1,1]	No bound in range
Affected by outliers	Less affected by outlier

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

Ans:

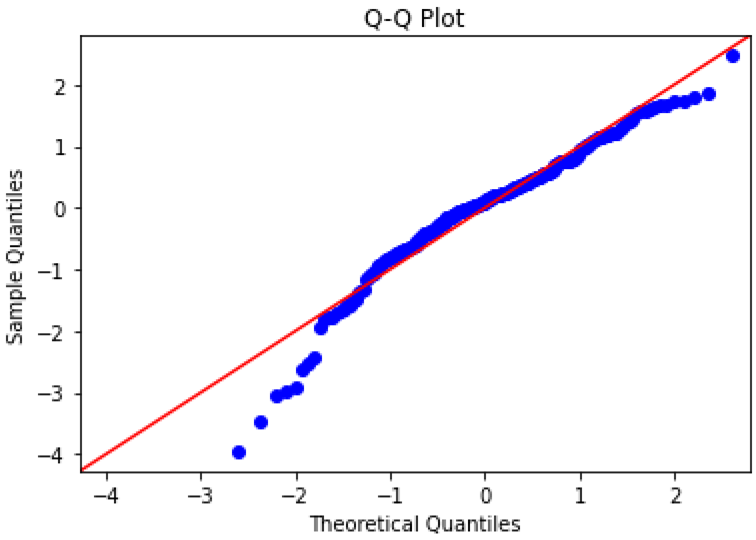
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Ans: Q represents Quantile here. Q- Q plot plots theoretical cumulative probability vs sample cumulative probability.

More abstractly,given two cumulative probability distribution functions *F* and *G*, with associated quantile functions F^{-1} and G^{-1} (the inverse function of the CDF is the quantile function), the Q–Q plot draws the *q*-th quantile of *F* against the *q*-th quantile of *G* for a range of values of *q*. Thus, the Q–Q plot is a parametric curve indexed over [0,1] with values in the real plane $\mathbf{R}^2$

Q_Q Plot of Residuals of Bike rental modal:



Used to check,

- Come from populations with a common distribution
- Have common location and scale
- Have similar distributional shapes
- Have similar tail behavior

Note: Presence of outliers can all be detected from this plot