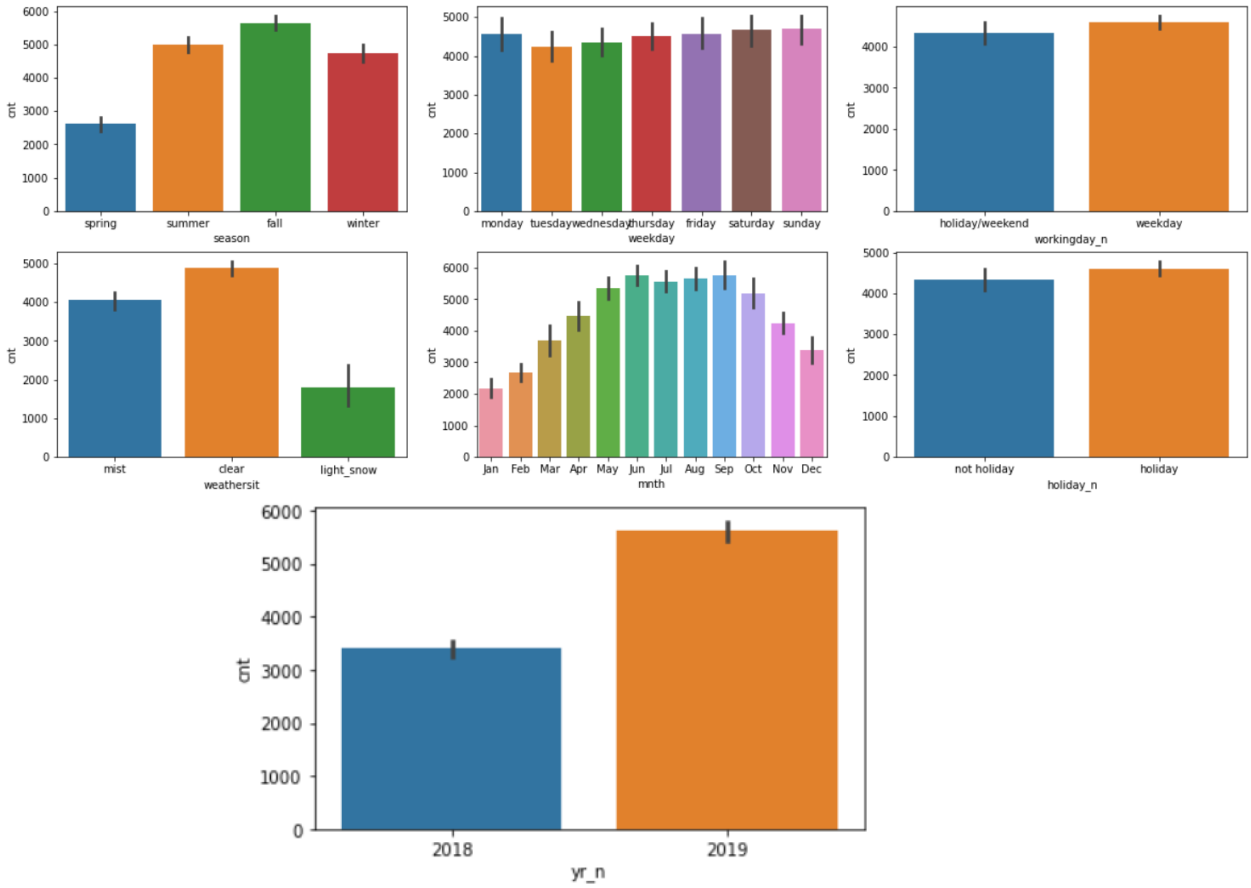
Assignment-based Subjective Questions

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

Answer:



- Most of the people are taking shared bikes when the weather is clear .
- Usage of shared bikes is high in months of June,July,Aug and September.
- Demand of shared bikes has increased in 2019 when compared to 2018.
- Demand for bikes is high during fall.
- Demand of shared bikes is slightly high on weekday,but there is no significant difference and same is the case with holiday/not holiday. Demand of shared bikes is slightly high on holiday but no significant difference.

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

Answer: Because when a dummy variable is created for feature/column with n levels,n-1 features/columns are sufficient to represent n levels.But during creation of dummy variables,as we create n features for n levels,we use drop_first=true to remove the extra column.Thus reducing the correlation among dummy variables.

For example,lets say there is a column named Marital status with values married,single and divorced.
To represent these values we can create only two dummy variables and that can be used to represent all 3 variables information.

00-Divorced
10-Married
01-Single

Married	Single
0	0
1	0
0	1

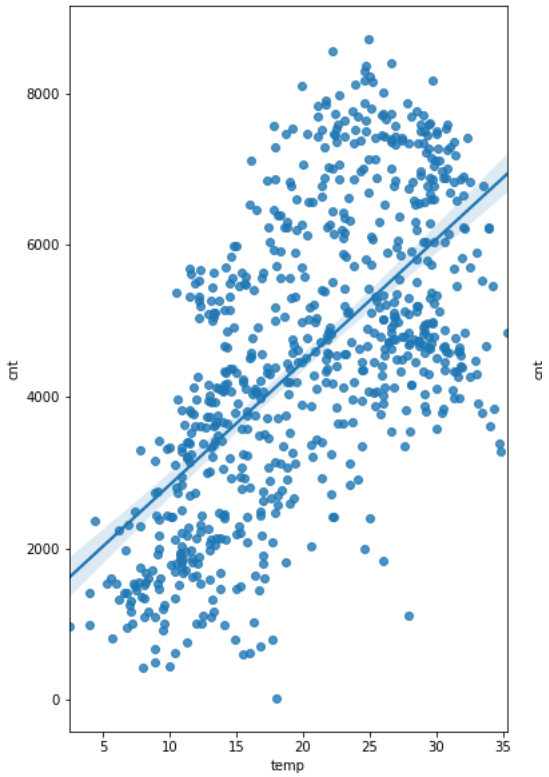
3.Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer:Both atemp and temp are having high correlation with target variable.As temp and atemp are also having high correlation,atemp was dropped.

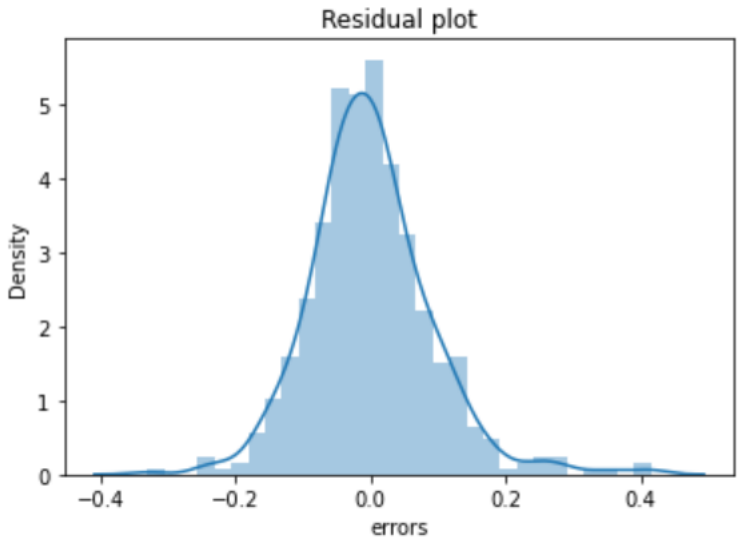
4.How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer:Following are assumptions of linear regression.

1. There is a linear relationship between X and Y.

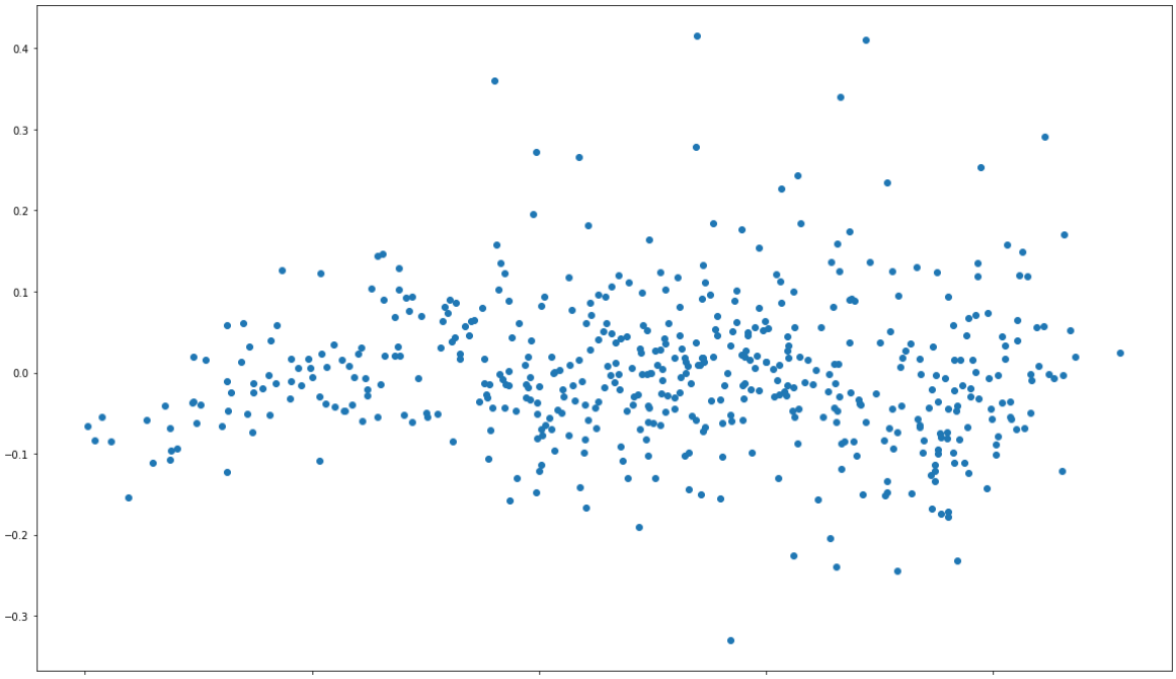


2.Error terms are normally distributed with mean zero.

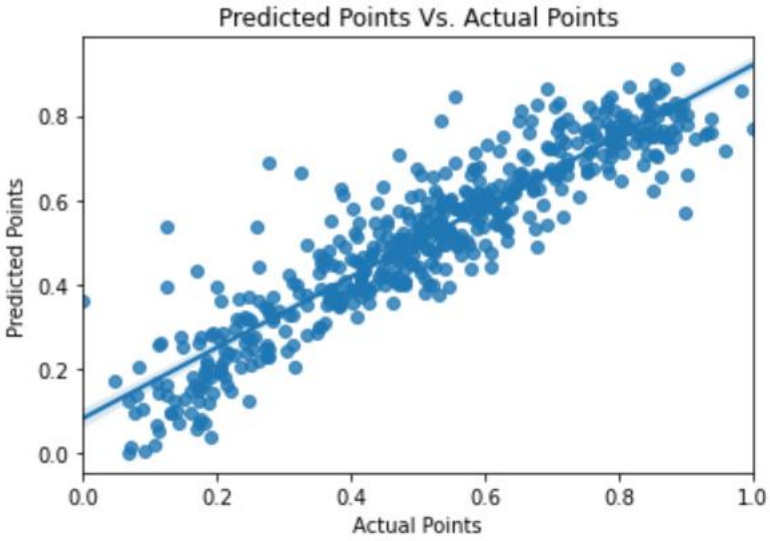


If error terms are not normally distributed then the p-values obtained become unreliable.

3.Error terms are independent of each other.
There should be no visible pattern.



4.Error terms should have constant variance(homoscedasticicty)
The variance should not increase or decrease as the error values change.



5.There should be no multi-collinearity between predictor variables.

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

Answer: Top 3 features contributing significantly towards explaining demand of shared bikes are:

1) Temperature
2) Year
3) Winter season

General subjective questions:

1. Explain the linear regression algorithm in detail:

Answer: Linear regression is a supervised learning algorithm,which is a form of predictive modelling technique that tells us the relationship between the dependent(target) variable and independent(predictor) variables.

There are two types of linear regressions:

1. Simple linear regression
2. Multiple linear regression

Simple Linear regression

Simple linear regression explains how the value of dependent variable varies based on value of a single independent variable using straight line equation:
y=mX+c

where y=dependent/target variable to predict
X=independent variable/predictor variable used to make predictions
m = slope of the line which shows how y varies based on X
c=constant,if x=0 then y=c

If the value of m is positive ,it means there is positive relationship between X and y i.e if value of X increase then y increases

If the value of m is negative,it means there is negative relationship between X and y i.e if value of X increase then y decrease and vice-versa.

The best fit line is found by minimising the expression of Residual sum of squares(RSS). The strength of linear regression model can be assessed using R-square or coefficient of determination.

Assumptions of simple linear regression:
1.Linear relationship between X and Y.
2.Error terms are normally distributed.
3.Error terms are independent of each other.
4.Error terms have constant variance(homocedasticity)

Multiple linear regression:
Multiple linear regression is a statistical technique to understand relationship between one dependent variable and several independent variables.

Multiple linear regression equation can be simply represented as:
y=m0+m1X1+
m2X2+…….+mnXn

where y is the dependent variable and x0,x1…xn are independent variables.
Following are new considerations we need to make while moving from simple to multiple linear regression:
1.Overfitting:When more and more variables are added,model might end up memorising all data points in training set and may not generalise while making predictions on test dataset
2.Multi-collinearity:There should be no correlation between predictor variables.

2.Explain the Anscombe's quartet in detail.

Answer:Anscombe's quartet is the modal example to demonstrate the importance of data visualisation.It's intended to counter the impression "numerical calculations are exact,but graphs are rough".

It is developed by statistician Francis Anscombe in 1973 to signify importance of plotting data before analysing it with statistical properties.

It comprises of 4 datasets and each dataset contains 11 (x,y) points.All the four datasets share same descriptive statistics(means,variance,standard deviation ect) but different graphical representation.

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Average value of x = 9

Average value of y =7.5

Variance of x=11

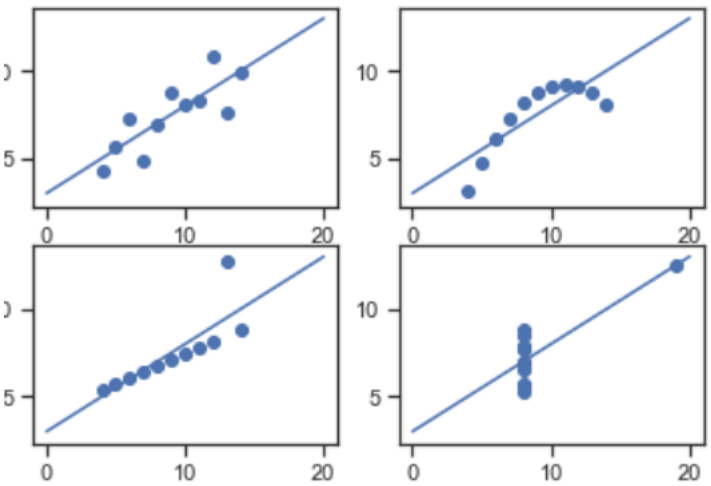
Variance of y=4.12

Correlation coefficient=0.816

Linear regression equation : y=0.5x+3

The above values are same for all 4 datsets.

But when we plot these values we get the following pictorial view:



Dataset1 – shows linear relationship between x and y with some variance.

Dataset2 – shows curve shape and there is no linear relationship

Dataset3 – shows a linear relationship but with few outliers

Dataset4:Value of x remains constant except for one outlier.

3. What is Pearson's R?

Answer:Pearson correlation coefficient (r) is the most coomon way of measuring linear correlation between two variables.The value ranges between -1 and 1 and measures strength and direction of relationship.

Positive correlation:r value ranges between 0 and 1.When one variable changes the other variable changes in same direction.

Negative correlation: r value ranges between -1 and 0.When one variable changes the other variable changes in opposite direction.

No correlation:r value is 0.There is no relationship between 2 variables.

If the coefficient value lies between +/-0.5 and +/-1 then it is strong correlation.

If the coefficient value lies between +/-0.3 and +/-49 then it is medium correlation.

If the coefficient value lies between +0.29 then it is small correlation.

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

Answer:When we have lot of independent variables in model and each of them are in different scales,it will result in incorrect modelling,giving more weightage to features having high magnitudes and ignoring other parameters.

We need to scale features for two reasons

1.Ease of interpretation

2.Faster convergence for gradient descent methods.

Two popular methods used for scaling features are:

Standardised scaling

Normalised scaling

Standardised scaling	Normalised scaling
x=(x-mean(x))/sd(x)	x = (x-min(x))/(max(x)-min(x))
Mean and standard deviation are used	Minimum and maximum values are used
Used to ensure zero mean and unit standard deviation	Used when features are of different scales
Scales values not bounded in certain range	Scales values between (-1,1) or (0,1)
Not affected by outliers	Affected by outliers
Transformer called StandardScaler is used for scaling features	Transformer called MinMaxScaler is used for scaling features
Also called z-score normalisation.	Also called scaling normalistion
Useful when feature distribution is normal	Useful when we don't know about distribution
Used when algorithms make assumptions about data	Used when algorithms do not make assumptions about data

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

Answer:VIF is variance inflation factor which is used to identify multi-collinearity between independent variables.If VIF value is >10 it means there is high correlation and value > 5 should be inspected.

Formula for calculation of VIF= 1/(1-R2)

If value of R2=1 then VIF becomes infinity which means there is perfect correlation between two variables.To solve this we need to drop one of the variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression ?
   Q-Q plot is a graphical technique for determining if two datasets come from populations with a common distribution.

   Uses and importance:
   Q-Q plot is used to find type of distribution for random variable whether it be a gaussian distribution,uniform,exponential or even pareto distribution etc.In Q-Q plot we plot quantiles against quantiles.Whenever we are interpreting Q-Q plot we concentrate on y=x line.
   Many distributional aspects like shifts in location,shifts in scale,changes in symmetry and presence of outliers can be detected using this plot.

   Interpretations of two datasets:
   1)Similar distribution : IF all points of quantiles lies on or close to straightline at 45 degree angle from x axis
   2)Y-values < X-values : If y-quantiles are lower than X-quantiles
   3) X-values < Y-values : If x-quantiles are lower than y-quantiles
   4)Different distribution:It all point of quantiles are away from straight line at 45 degree from x axis