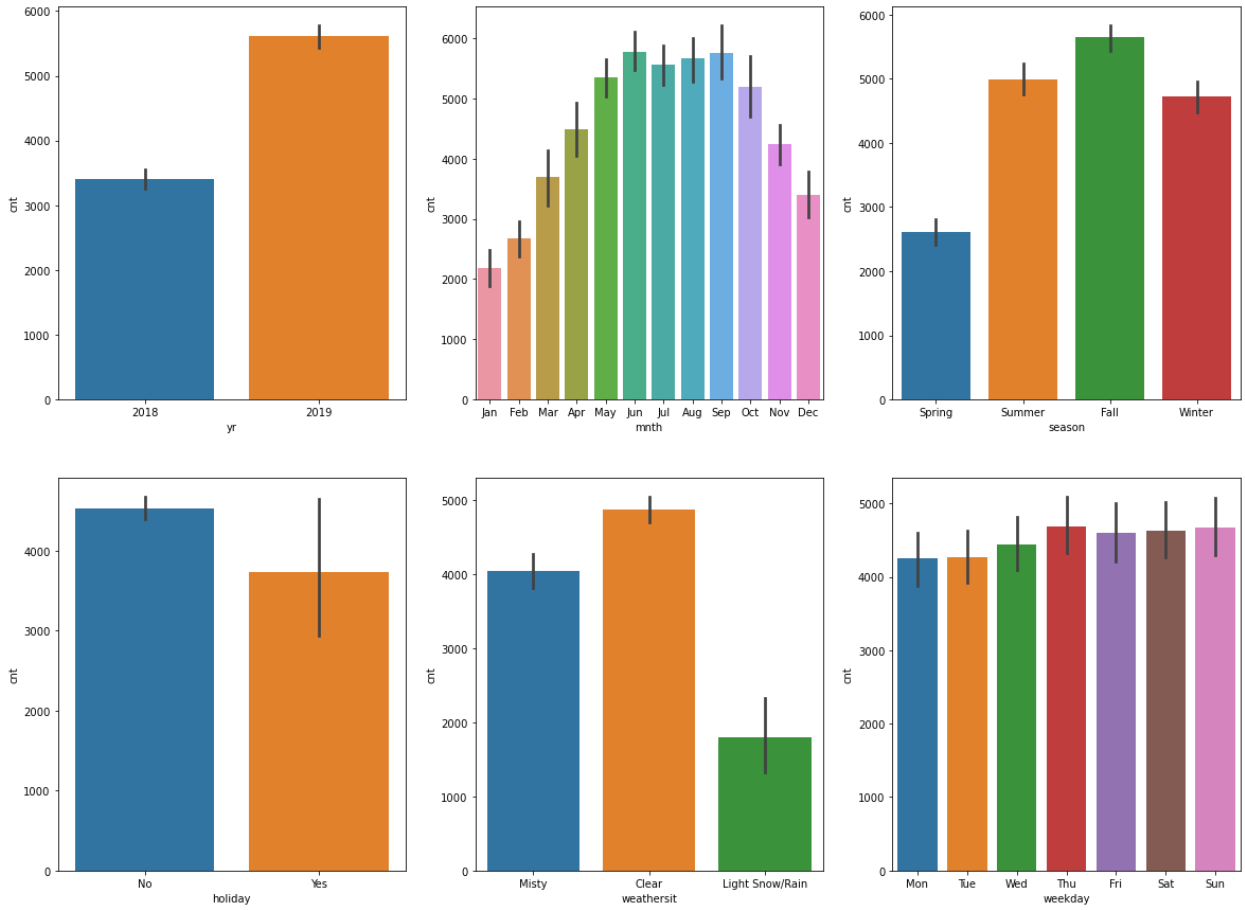
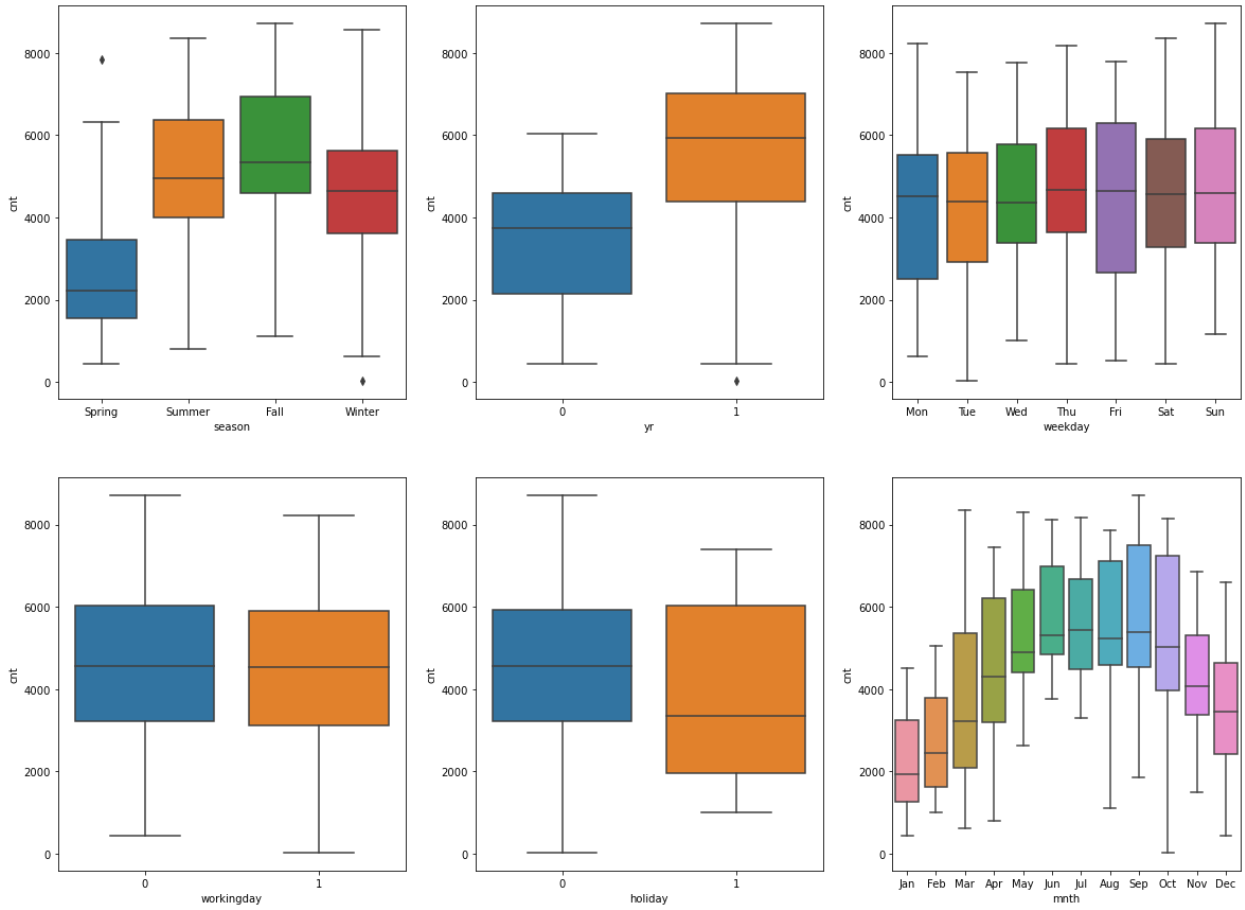
Subjective Questions

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? (3 marks)

   Created few plots to check relation between Categorical variables





- Bike Demand increased in 2019
- Bike demand is high on Non holidays
- Bike Demand is highest in fall then summer winter less in spring.
- Month wise high demand is from Jul to Nov , this shows Month and season may have correlation.
- Bike demand is high when weather is clear, it is least in when it Rains and Snow. This variable is also related to Season
- Day wise avg demand is almost same. its show higher/lower demands on some days but not much difference.

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

   When we create dummy variables for categorical variables with N level there is no need to create N variables as we can identify the N variable from other dummy variables created. Let's say we have season variable with 3 Levels Summer,Winter,Rainy and we create 3 variables out of it.

Summer	Winter	Rainy
1	0	0
0	1	0
0	0	1

- When Values in 1,0,0 Its Summer
- When 0,1,0 Then its winter
- When 0,0,1 Then its rainy season

But we don't need last variable because when it's not summer (0) and winter (0) it is indication of Rainy season. So only creating two column is sufficient

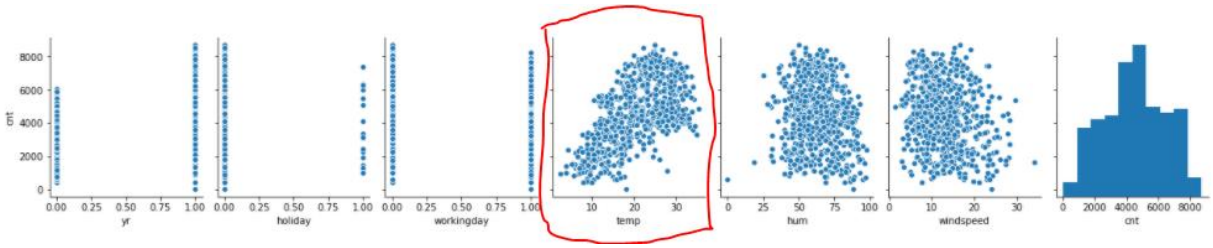
Summer	Winter
1	0
0	1
0	0

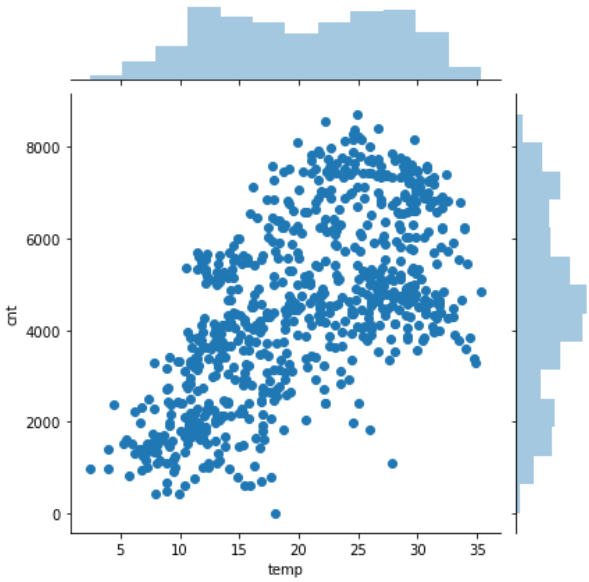
This helps in model building as we will have less number of columns to choose and build model from.

This saves lot of effort to choose correct features for model building.

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

   By looking at plot **temp** is showing highest correlation with target variable as it is showing fairly strong positive linear relation in scatter plot.

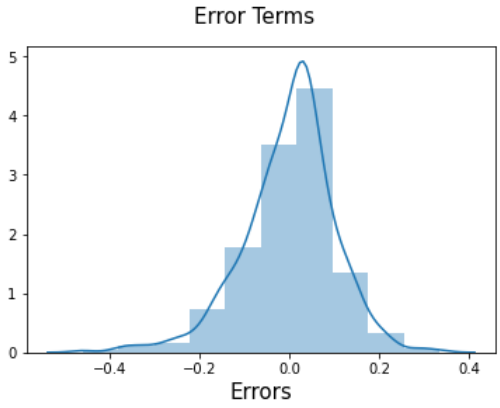




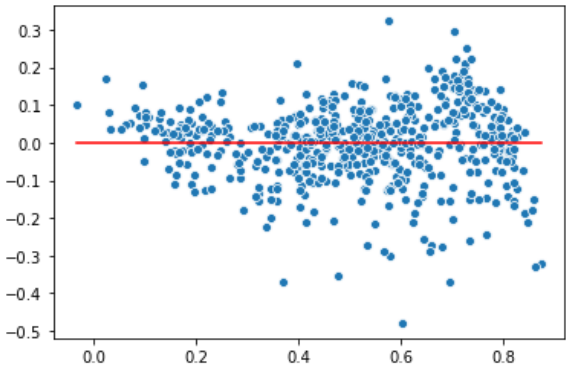
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

   To validate assumption I did residual analysis, first created variable of residual from y train and y train predicted values then first assumption

   - Error terms are normally distributed : - Created distribution plot residuals to check if they are normally distributed and has mean = 0
     This Plot show error are normally distributed and has mean 0



   - Error terms are independent of each other
     All error values are independed and there is no visual pattern present



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

Top 3 Feature significantly impact are

Feature	Coef
Temp	0.4647
year	0.2371
Light Snow/Rain	-0.2618

Based on final model bike demand is influenced by below features.

Temp has 0.4647coef shows when temp increase by 1 unit bike hire increases by 0.3857 Times, As **Year** with coef 0.2371 indicates when unit increase in year bike hire numbers increase by 0.2371 time.
Whereas When It there is **light rain or snow** bike demand will decrease by 0.2433 times

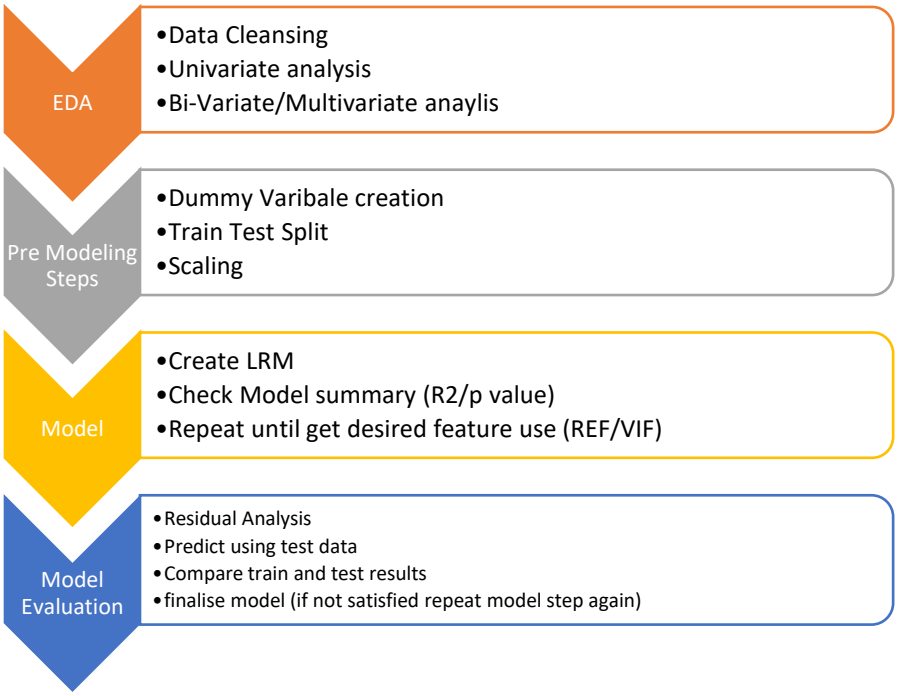
The linear equation for this model.

Bike Demand = 0.2772 + (0.2371 X Year) + (0.4647 X Temp) + (0.0562 X season Winter) + (-0.0936 X holiday) + (-0.1315 X Windspeed) + (-0.1137 X Spring Season) +(-0.2618 X Light Snow/Rain)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Regression model is type supervised machine learning method where output variable is continuous variable like sales, marks etc. We understand the relation between target variable and other predictor variable/s and create a formula which can best determine the values of target variable.



   EDA: - Explanatory data analysis is first step performed to get insights from the data

   - Data Cleansing: - Data is wrangled using multiple methods. Changing column values to upper case/lowercase, changing date format etc are some types. Cheking missing values is main task where if data is missing dropping that row/column or imputing that data using Mean/Mode/Median is done in this step
   - Univariate analysis: - Checking individual variables in data checking there distribution using histogram.
   - Bivariate/Multivariate analysis: Understanding the relationship between two and more variables. This can be done using scatter plot, bar graphs, heat maps etc.
     This is important in linear regression model as it shows if we have any linear relation in our data.
   - Derive Variables:- Derive new variable as per requirement

Pre Modeling steps: After we identify linear relationship we move to next pre model steps

- Create Dummy Variable: - In LMR all values should be numerical so important categorical variables are converted into dummy variables example as below.

	Summer	Winter
Summer	1	0
Winter	0	1
Rainy	0	0

- **Train Test Split: -** After all variable creation is done we move to split the data into test and train. We divide data either 70% train 30% Test or 80% Train or 20 % test. **This is important step as it gives us data set to evaluate our model**. If use same data for training and testing the model we could lessen the effect of data differences and better understanding of features of the model

- **Feature Scaling: -** in Dataset variables could be at different scales such as Temp could be from -10 to 30 Deg. Rainfall could be from 0 to 50 mm sales 0 to Million dollars. Because of this model could give odd coeff. So it is always better to scale variable at one level for ease of interpretation. There are popular method to do scaling like **Standardizing** where data is scaled such as there mean is 0 and SD is 1, **MinMax** Scaling where data is scaled such that all available values are scaled to fit in between 0 and 1.

Model Building:-

Feature Selection:-

- Start building model we could build model using manual way means choosing independent variable one by and try all combination to arrive at good model. This could be very tedious and time consuming when # of variable are high.
- There is automated way where features are selected using REF (Recursive Feature Elimination) where we can pass number of feature we want to function and it provide list of columns which are important in model.
- We could use mixed approach where first some feature are selected using REF and then next steps are built using manual way.
- To check multicollinearity VIF method is used based on this feature could be selected or removed from model (if VIF is > 5 remove the feature)
- Model building is recursive step where each model is checked for validity, significance and accuracy here concept of R2/adjusted R2 P value , Prod F stat are checked and based on all these final model is selected for evaluation
  Good Model indication
  High R2/adjusted R2
  Low value for Prod (F- stat)
  Low P value for all selected independent variables

Model Evaluation

Residual analysis:-Model is tested for assumption like

- Error terms are normally distributed
- Error terms are independent of each other
- Error terms have constant variance (homoscedasticity)

Predict Test dataset

Test data us predicted using the final model and then test pred and train pred comparison is done.

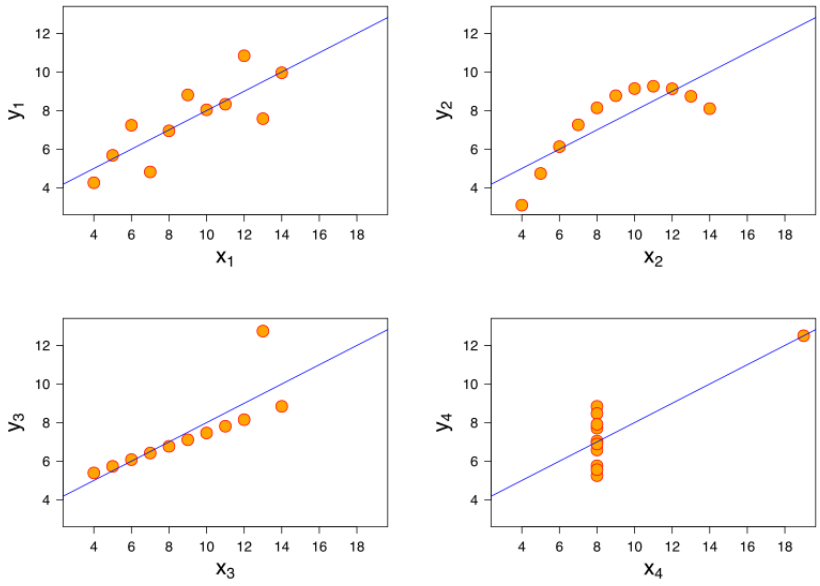
R2 and of test set is calculated and predicted with train set and then final with all these model is finalized.

2. Explain the Anscombe's quartet in detail.

This is data set created by Francis Anscombe in 1973 , it consist of 4 data set which has approximately identical simple statistical properties such as Mean, Sample Variance of x, Linear line etc.

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x	11	exact
Mean of y	7.5	to 2 decimal places
Sample variance of y	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	y = 3.00 + 0.500x	to 2 and 3 decimal places, respectively

Although these data set have approx. same properties when they plotted on graph shows very different. Plot No shows nonlinear relationship, plot 4 shows one huge value (outlier) can twist the summary statistics



This was done to prove importance of Visual analysis and EDA and not to just blindly rely on descriptive statistics

3. What is Pearson's R?

Pearson's r measures the strength of linear relationship between two variables.it has value -1 to 1. Where +1 means perfect positive relation and -1 means perfect negative relation. And 0 means no linear relation.

Formula

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

Where:

- N = number of pairs of scores
- Σxy = sum of the products of paired scores
- Σx = sum of x scores
- Σy = sum of y scores
- Σx^2 = sum of squared x scores
- Σy^2 = sum of squared y scores

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

What

In Dataset variables could be at different scales and magnitude such as Temp could be from -10 to 30 Deg. Rainfall could be from 0 to 50 mm sales 0 to Million dollars. Because of this model could give odd coeff.

Why ?

- Ease of interpretation of coeff
- Algorithm like gradient decent converges much faster with feature scaling than without it

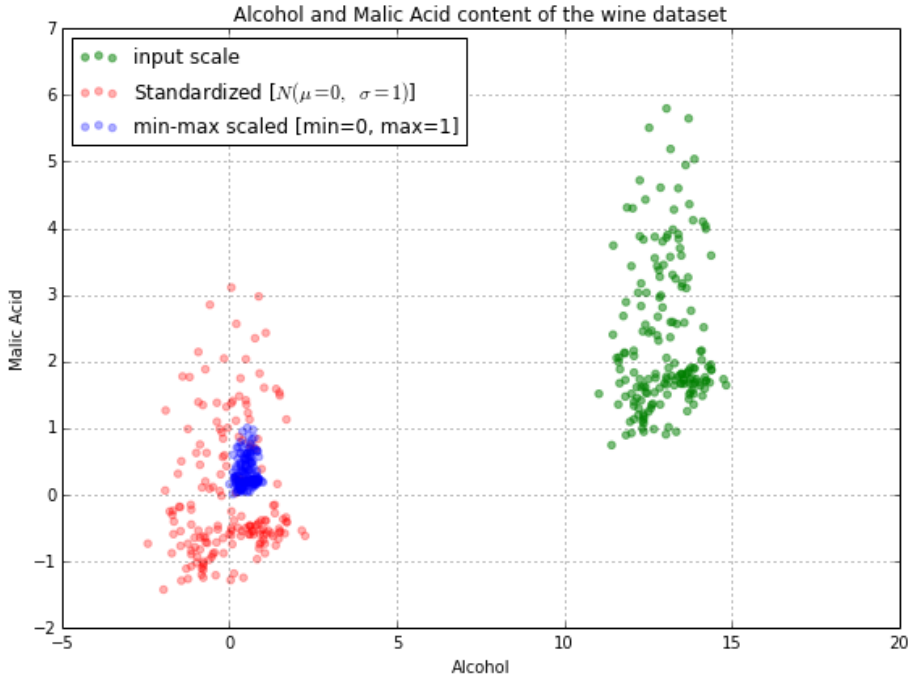
Difference between normalized scaling and standardized scaling

Normalized: In this values are rescaled to fit between 0 to 1. e.g. Min max method

Standardized:- in this Values are scaled in such way that their mean is zero and standard deviation is one

This below diagram shows all three patterns

Green :- Data at actual scale / Red :- Standardized scaled data / blues:- Normalized scaled data



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

   If dependent variables are perfectly correlated then VIF becomes infinite.
   In this case R2 becomes 1 and as per VIF formula 1/1-R2 result to infinite value.

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

Quantile-Quantile (Q-Q) plot are used to assess if your residuals are normally distributed.

Q-Q Plots are scatter plot created by plotting two sets of quantiles against each other. If quantiles came from same distribution then forms straight line shows residuals are normally distributed if it is forming and curve then it is skewed.