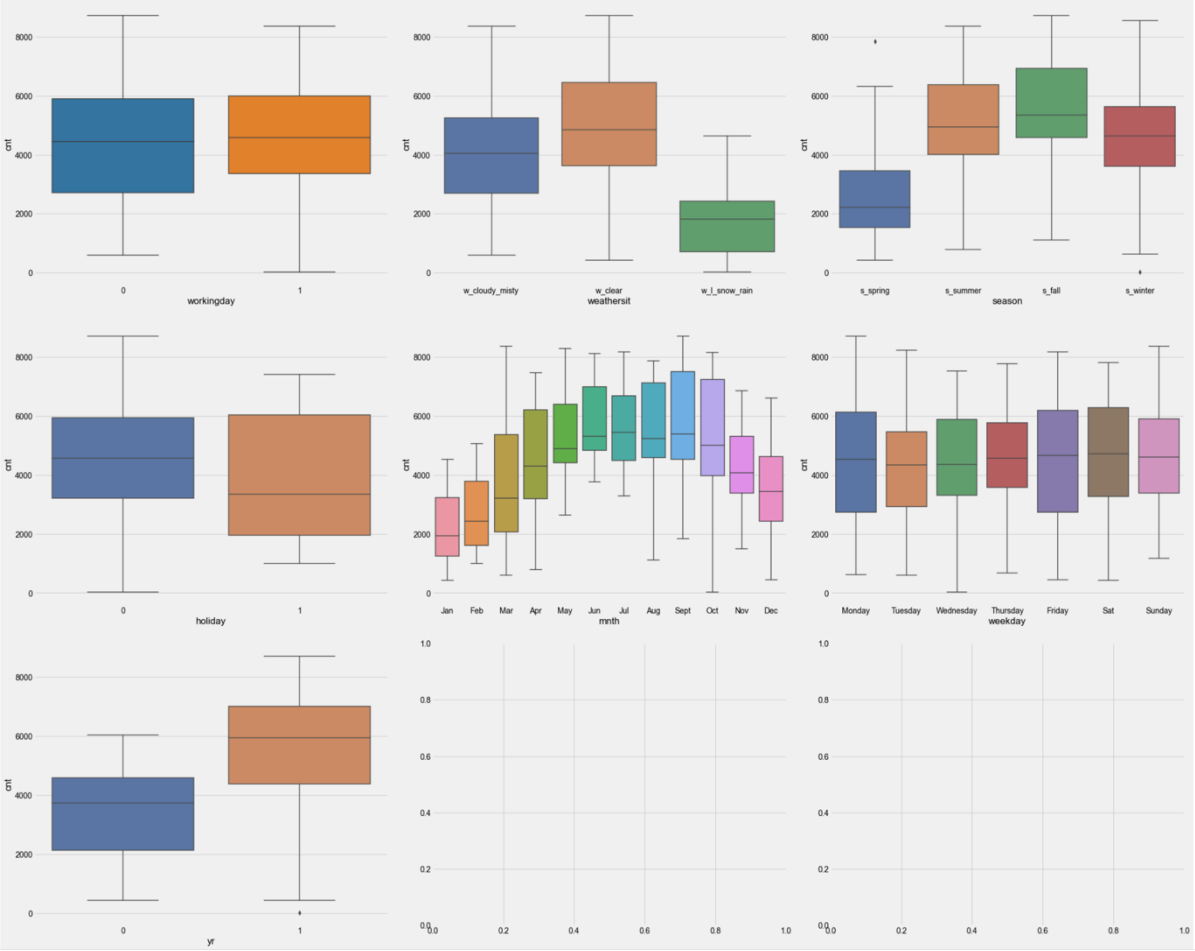
Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

   The categorical variables in the dataset are, working_Day, weather, season, holiday, month, weekday, yr.

   Below is the plot of the Categorical Variables, with the Dependant variable ie count.



   From the above plot we can infer that :

   - Snowy and Rainy weather situations adversely impact the count of Riders for Boom Bikes
   - Workingday has higher or holiday
   - The ridership in Spring is less as compared to summer fall Summer, Fall and Winter. That could mean, that Bike Stations are predominantly in Urban areas, near Office Commute Location.
   - The Ridership steadily decreases from October to January. This is a decreasing trend, and then rises from February to September.
   - The ridership on all the days of the week is steady around 15-16% and day of the week is not a significant predictor

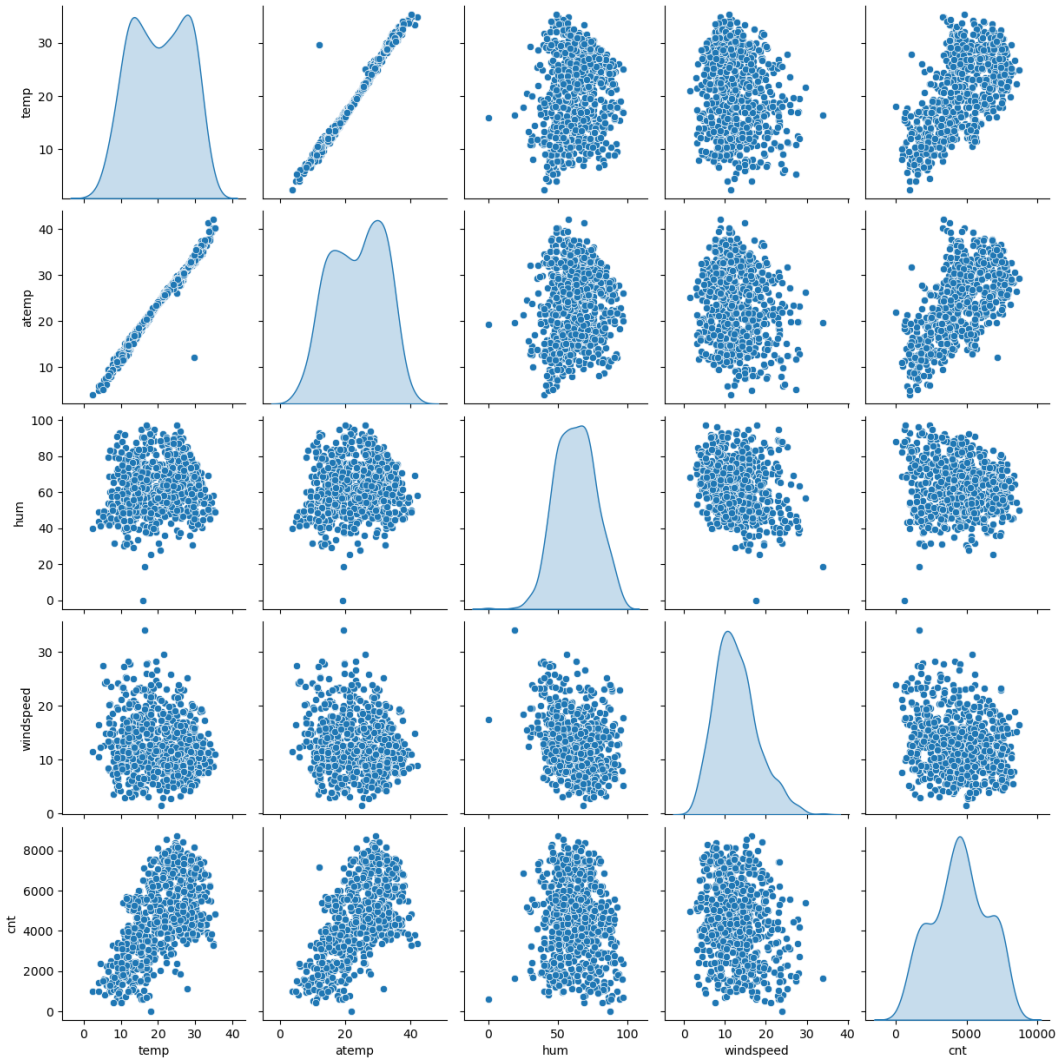
- Ridership is slowly increasing across 2018 to 2019, and this increasing trends are initial. But we can infer that business is growing.

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

- When we create dummy variables we are actually increasing the number of variables.
- For eg – A House can either be Semi-Furnished, Furnished, Unfurnished. If a house is neither Semi-Furnished, nor Unfurnished then its obvious that its Furnished. So 2 dummy variables will suffice. Alternatively we can create 3 dummy variables, also
- If we do so, then we are creating correlated variables which may have a strong adverse impact on the models converging. This is especially true if the cardinality is less, and high multicollinearity means unpredictable sign changes

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

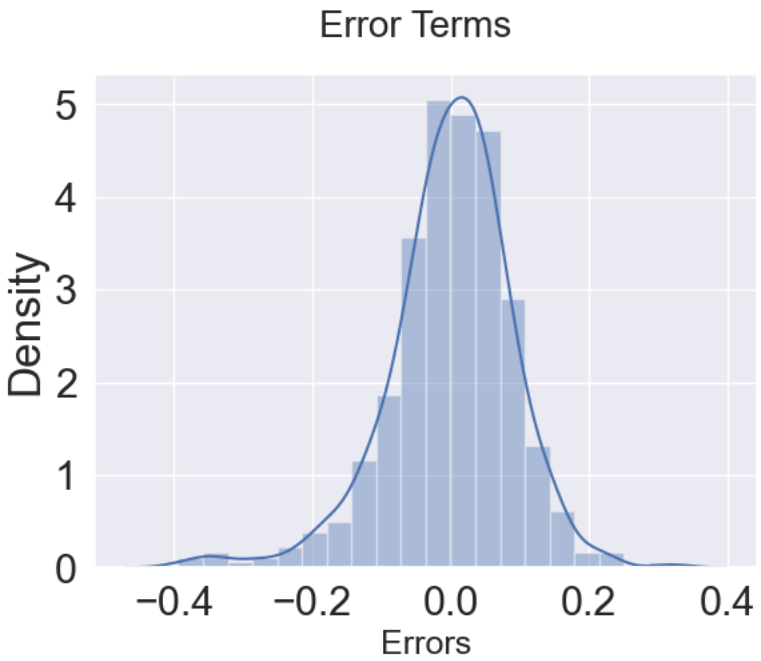
- The highest corelation is between temp and atemp based on the pair plot below



-

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

   The Linear Regression Model can be validated using Residual Analysis, as shown below



   - We plot a distribution of the Errors(which is the difference between y_train, and y_train_cnt, and validate the below two points
     - The residuals should be normally distributed
     - The residuals should have a mean of zero
   - For our model as well these points hold true and are validated.

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

   Below are the coefficient values of the final model.

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1698	0.029	5.937	0.000	0.114	0.226
yr	0.2294	0.008	28.272	0.000	0.213	0.245
workingday	0.0536	0.011	4.875	0.000	0.032	0.075
temp	0.5709	0.020	28.559	0.000	0.532	0.610
hum	-0.1613	0.038	-4.295	0.000	-0.235	-0.087
windspeed	-0.1861	0.026	-7.259	0.000	-0.236	-0.136
w_cloudy_misty	-0.0553	0.011	-5.255	0.000	-0.076	-0.035
w_l_snow_rain	-0.2439	0.026	-9.250	0.000	-0.296	-0.192
s_summer	0.0899	0.010	8.859	0.000	0.070	0.110
s_winter	0.1407	0.010	13.575	0.000	0.120	0.161
Sept	0.1025	0.016	6.610	0.000	0.072	0.133
Monday	0.0630	0.014	4.450	0.000	0.035	0.091

Based on the above we can conclude that

- Temperature impacts the Traffic count positively (0.53)
- Weather with Light Snow and Rain impacts the Traffic negatively(-0.24)
- Windspeed being high also impacts the Traffic count negatively (-0.18)

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Linear Regression is a supervised Machine Learning Algorithm that is used for the prediction of numeric values. Linear Regression is the most basic form of regression analysis

Regression is most commonly used predictive analysis model. Linear Regression is based on linear equation : y = mx +c. It assumes a linear relationship between dependant variable y and independent variable x.

Regression is performed when the dependent variable is of continuous data type and Predictors or independent variables could be of any data type like continuous, nominal/categorical etc. Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error.

In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term.
Regression is broadly divided into simple linear regression and multiple linear regression.

1. Simple Linear Regression : SLR is used when the dependent variable is predicted using only **one** independent variable.

2. Multiple Linear Regression :MLR is used when the dependent variable is predicted using multiple independent variables.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

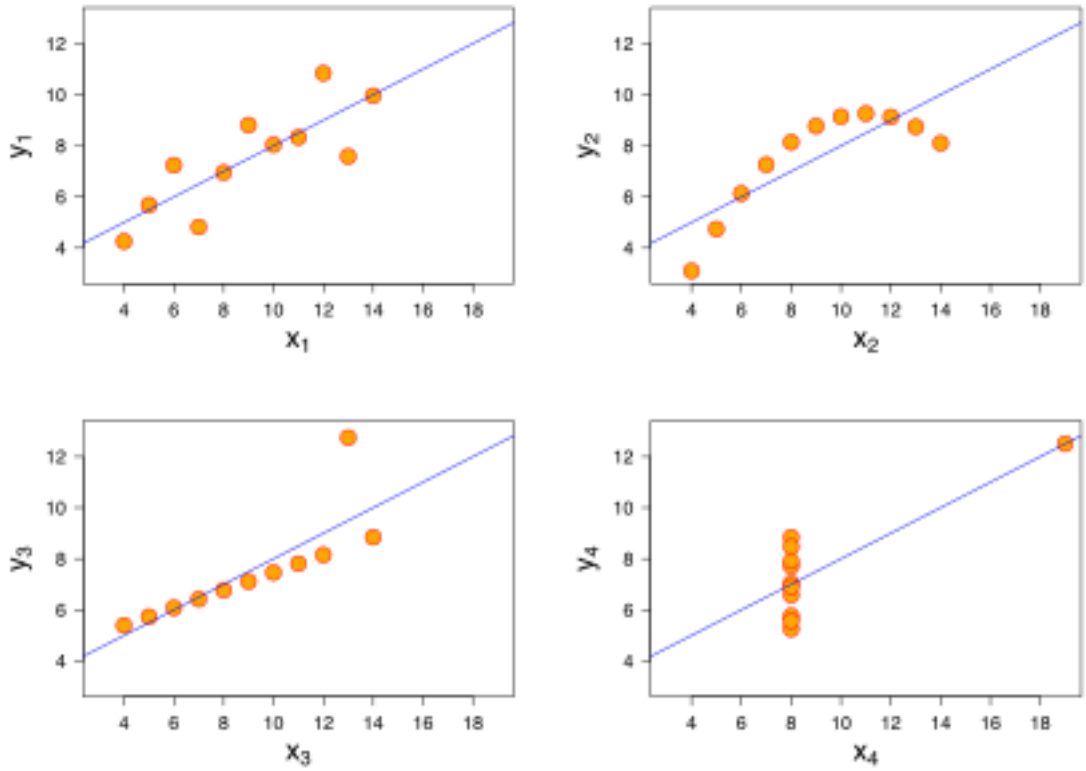
Where B0.. Bn are coeffecients of each of the independent variables.

MLR requires close attention to Overfitting, Multi Colleanarity and Model Selection, which makes it more complicated.
The other assumptions of Linear Regresson are same across both MLR and SLR ie Zero mean, independent, Normally distributed error terms that have constant variance

-

2. **Explain the Anscombe's quartet in detail.**

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough.



- The first scatter plot (top left) appears to be a simple linear relationship.
- The second graph (top right) is not distributed normally; while there is a relation between them,it's not linear.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

3. **What is Pearson's R?**

Pearson's r is a numerical summary of the strength of the linear association between the variables.It value ranges between -1 to +1. It shows the linear relationship between two sets

of data. In simple terms, it tells us can we draw a line graph to represent the data?

- r = 1 means the data is perfectly linear with a positive slope
- r = -1 means the data is perfectly linear with a negative slope
- r = 0 means there is no linear association

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

Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. I
t is performed during the data preprocessing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.

● **Normalization** is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

● **Standardization**, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

$$x = \frac{x - mean(x)}{sd(x)}$$

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

VIF basically helps explaining the relationship of one independent variable with all the other independent variables

$$\mathrm{VIF_i} = \frac{1}{1 - R_i^2}$$

Here Ri stands to SE for regression of Xj on the other covariates

The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. It gives us a measure of how much a variable can be explained by all the other independent variables.

If a variable Xj, is perfectly corelated, the then the Ri2 for each Xi =1. So VIF will be 1/0 = infinity

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

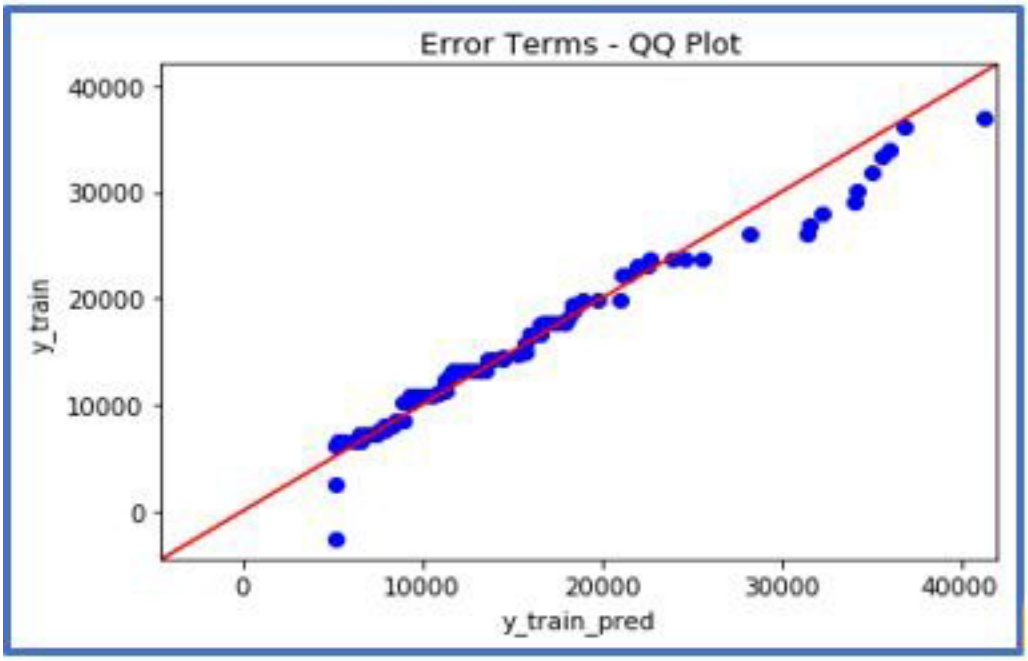
Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

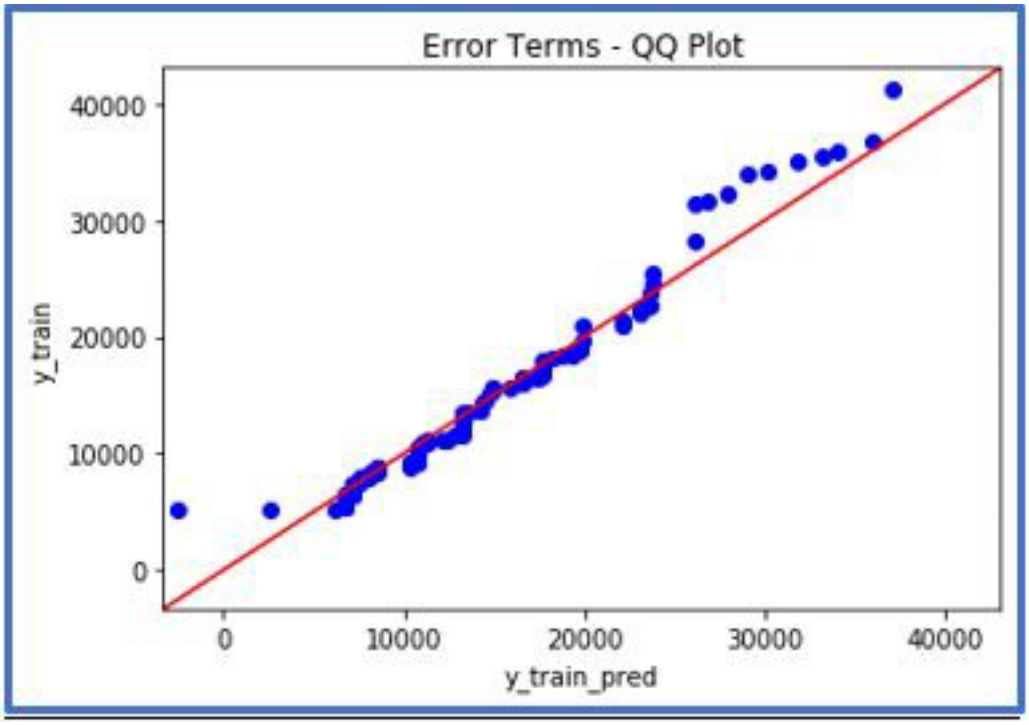
Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis

Following Questions can be answered by a Q-Q Plot

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behavior?