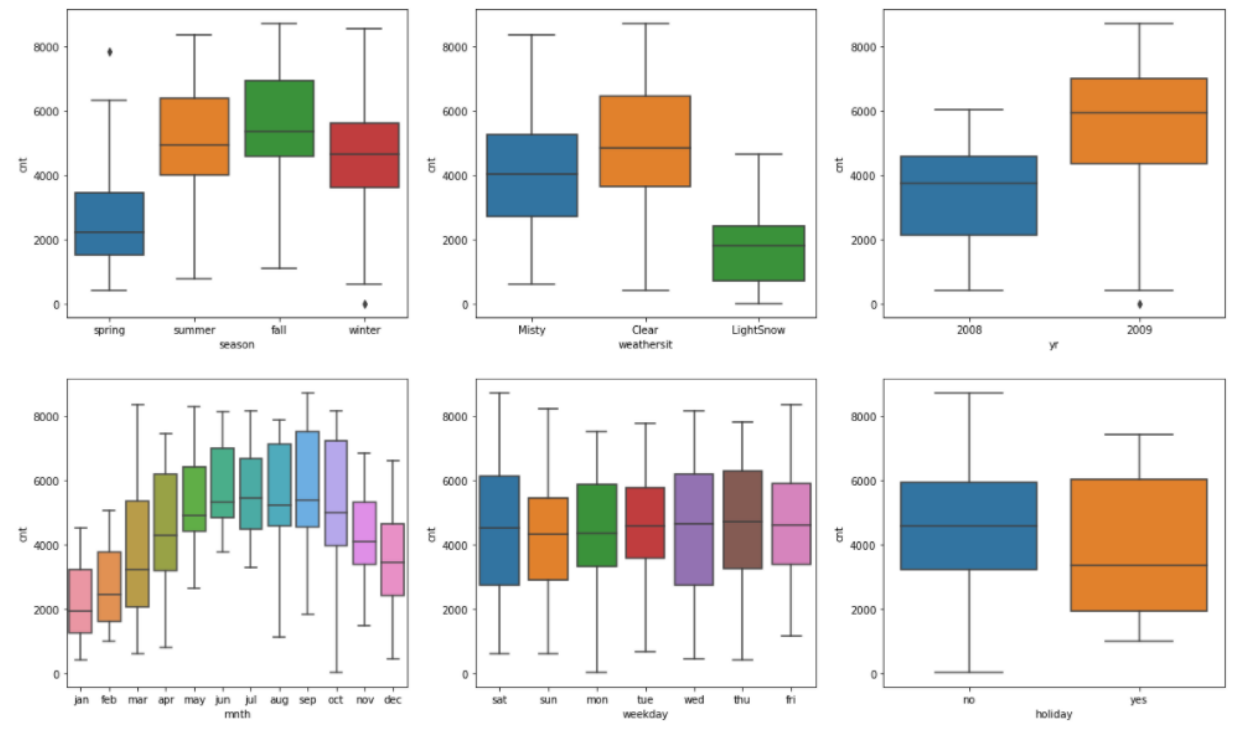
Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
Ans:

Several inferences can be got from analysing the categorical variables. Few are listed below:

a) More people tend to ride during fall season followed by summer season and then winter season
b) More people tend to ride when the weather is clear followed by misty weather.
c) On an average, more rides fall on workdays than on holidays.
d) May to October registers higher ride count
e) Average rides remain more or less the same on all days of the week.
f) There is a very good rise in demand from 2008 to 2009.



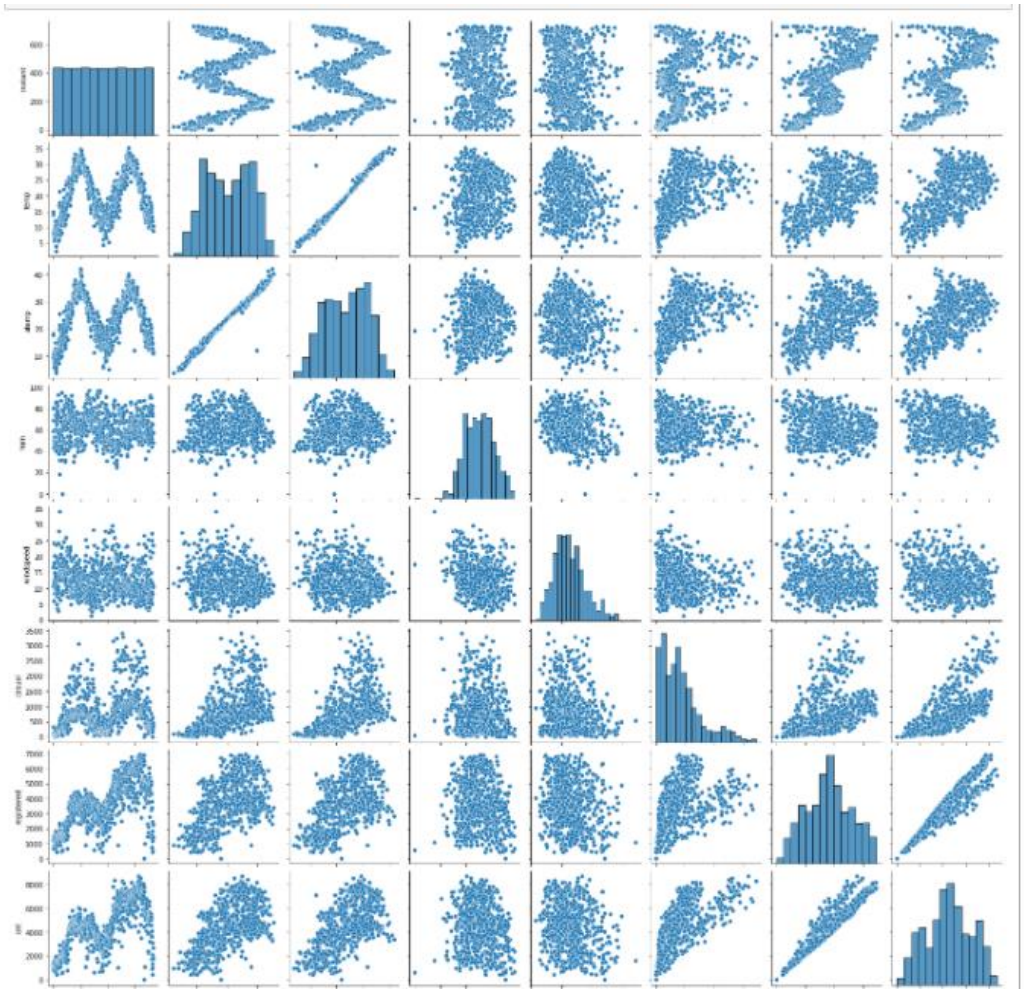
Q2. Why is it important to use drop_first=True during dummy variable creation?
Ans:

Using drop_first=True will remove the first level dummy variable while creating dummy variables from categorical variables.

If we want to remove any specific dummy variable, we can use drop() method instead.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Ans:

Eliminating variables ‘registered’ and ‘casual’, the highest correlation with target variable ‘cnt’ in pair-plot is ‘temp’, closely traced by ‘atemp’.



Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans:

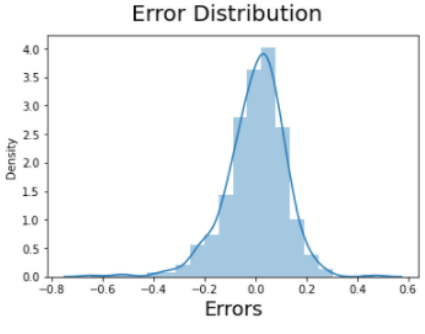
Validation is done by carrying out residual analysis. Histogram / Distribution of Error (between the actual dependant variables and the predicted dependant variables) is a normal distribution with mean at '0'.

```
In [594]: y_train_cnt = lm.predict(X_train_lm)

In [595]: import matplotlib.pyplot as plt
          import seaborn as sns
          %matplotlib inline

In [615]: # Plot the histogram of the error terms
          fig = plt.figure()
          sns.distplot((y_train - y_train_cnt), bins = 20)
          fig.suptitle('Error Distribution', fontsize = 20)                  # Plot heading
          plt.xlabel('Errors', fontsize = 18)                            # X-label

Out[615]: Text(0.5, 0, 'Errors')
```



Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans:

Top 3 features:

Including impact of yearly demand increase i.e. yearly increase in cnt

1. Year
2. Windspeed
3. Spring (season)

(Note: For the model where yearly demand variation is not considered, top 3 features are

1. Windspeed
2. Spring(season)
3. Sep (month))

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Ans:

Linear regression is a statistical method used to get the best estimate of an output variable 'y' given a set of independent input variables 'X', each of which has a linear relationship with y.

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots.. + \beta nXn + €$$

Basic assumptions are error terms (Y_pred – Y_actual) are normally distributed with zero mean, independent and have constant variance. In multiple linear regressions, care must be taken to avoid multi-collinearity and overfitting, where feature fitting becomes critical.

Algorithm of linear regression is as follows:

- Step 1: Reading and Understanding Data

- Step 2: Visualizing the Data

Steps 1 and 2 involve general understanding of the data.

- Step 3: Data Preparation

In this step, the data is modified so that categorical variables with N distinct values are converted to N-1 distinct dummy variables. All strings are converter to numerical variables. And all the variables are normalized using standardization or minmax techniques.

- Step 4: Splitting Data into Training and Test Data sets

A given dataset is generally split into 70:30 ratio where 70% is used for training the model and 30% is used for testing the model arrived.

- Step 5: Building Linear Models and analysing statistics

Linear models are built by using different algorithms by studying the correlation features, p-values, R2/adjusted R2 values, etc.

- Step 6: Residual Analysis of Train dataset

Validate the assumptions of the model by plotting and ensuring the error values are normally distributed.

- Step 7: Making Predictions Using the finalized model

Make predictions using the finalized model on the test data set.

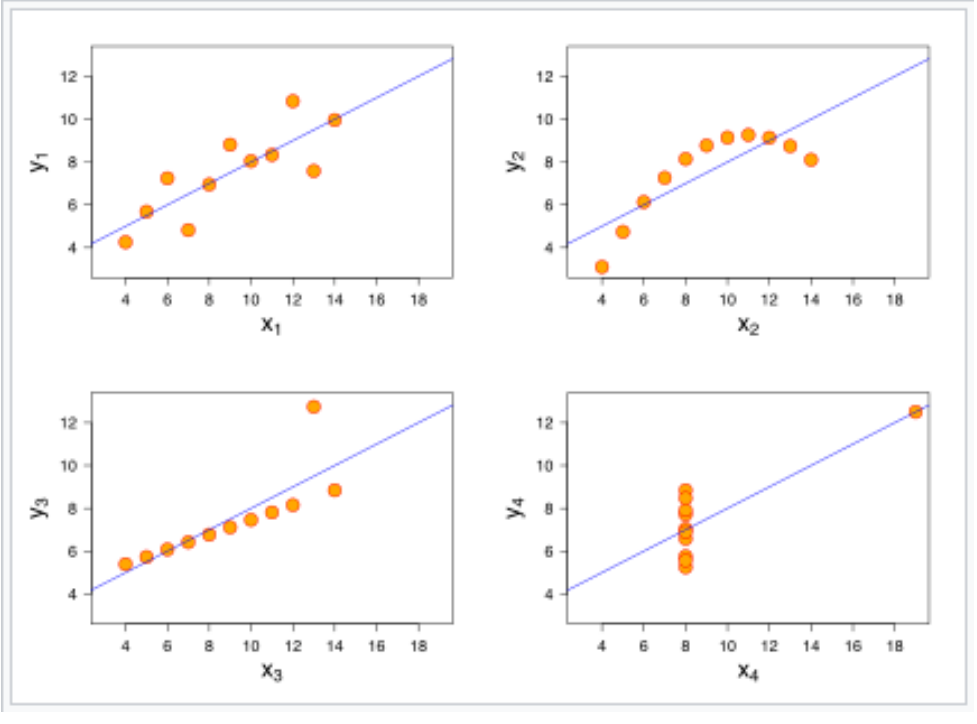
- Step 8: Model Evaluation

Evaluate by comparing predicted data with actual test data. A very high R2 value indicate that the model is good.

Q2. Explain the Anscombe's quartet in detail.

Ans:

Anscombe's Quartet is made to emphasize the importance of graphical (visual) study in analysis of a data set. It comprises of four distinct data sets with 11 data points which have similar summary statistic values but differ entirely in relationship between variable, which could be seen through plotting a scatter plot.



All four graphs are different, but have the same mean(x), mean(y), sample variance(Sx2), sample variance(Sy2), Linear Regression line and R2.

Anscombe's quartet

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	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

Source: Wikipedia

Q3. What is Pearson's R?

Ans

Pearson's R, also called as Pearson's correlation coefficient is a measure of linear correlation between two sets of data.

It is the ratio between covariance of two variables and the product of their standard deviation. The result value always lies between -1 and +1.

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where:

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

A value of -1 indicates that the variables are **negatively perfectly correlated.**

A value of +1 indicates that the variables are **positively perfectly correlated.**

A value between -1 and 0 indicate that variables are negatively correlated but are not perfect.

A value between 0 and +1 indicate that variables are positively correlated but are not perfect.

A value of 0 indicate that the variables are not correlated or random in nature.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans:

While modelling with several different independent variables, each variable may be having a different scale. For instance area of a building may range from 600 sqft to 2000 sqft, whereas, number of bedrooms may be between 1 and 4. This different scale range may result in unclear interpretation of coefficients. In such cases, scaling method is used to ensure that all the independent variables have comparable ranges so that coefficient interpretation is easier.

Scaling is performed mainly for:

1. Ease of interpretation
2. Faster convergence for gradient descents method

In Normalized scaling, variables are scaled such that all the values lie between 0 and 1. This is also called as minmax scaling.

Formula is: X = x-min(x) / (max(x) – min(x))

In Standarized scaling, variable are scaled such that their mean is zero and SD is 1.

Formula is: X = x-mean(x)/SD(x)

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans:

VIF is infinite if two variables are perfectly correlated.

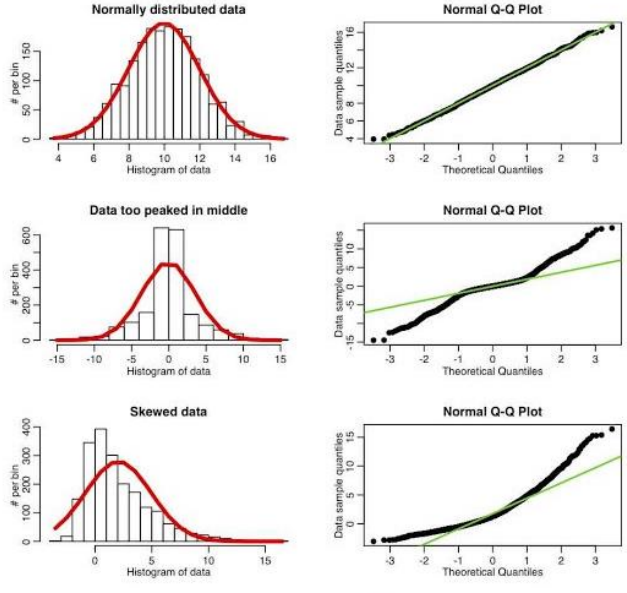
VIF formula is : VIF = 1 / (1-R2)
When two variables are perfectly correlated, R2 = 1
Then the denominator in VIF becomes 0. Anything divided by 0 is infinite.

This happens when two variables considered in the model are related such that one of the variables is redundant. For example, if we consider all three variables, furnished, semi-furnished and unfurnished while building model, then VIF will be infinite for all three variables.
Here, unfurnished = 1 – semi-furnished – furnished. This is the reason, one of the variables is not used in model building.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Ans:

Q-Q plot is a graphical tool used to graphically analyse and compare two probability distributions by plotting their quantiles against each other. If the two distributions under comparison are exactly equal, then the Q-Q plot is a perfect line y = x.

Though it is used for various distributions, lots of practical applications use normal distribution.



Source: Sherrytowers Q-Q plot examples

In linear regression, Q-Q plots can be used to confirm that both the test set and the train set are from the same population having the same distributions.