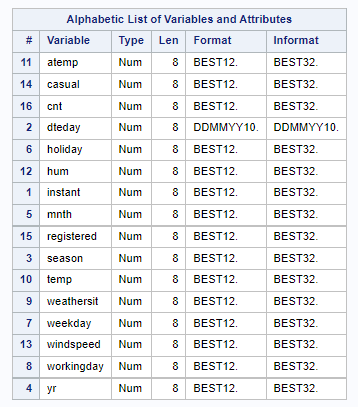
**Statistics for Data Sciences**

(Bike Sharing Systems)

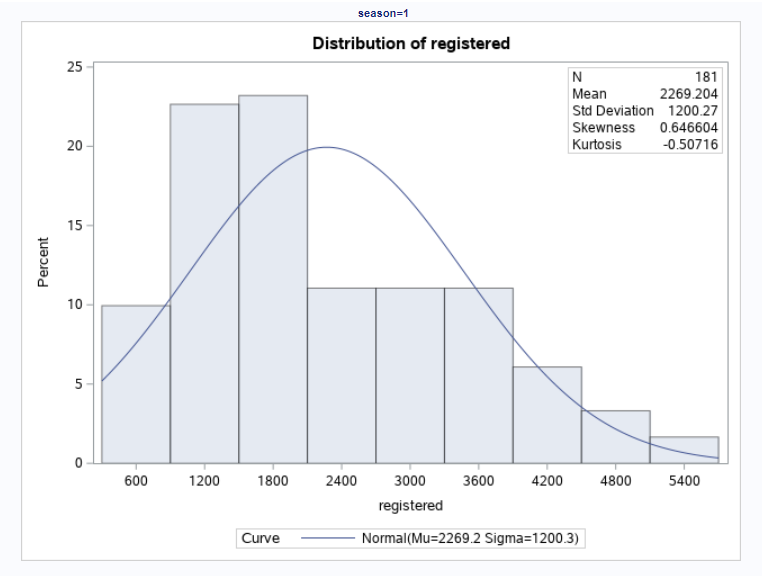
The data file for this assignment is called daily.csv and contains daily counts of bike rentals for 2011 and 2012, derived from Capital Bikeshare trip history data, with additional weather and seasonal information. The data was downloaded from the UCI Machine Learning Repository. Variables in that file are:

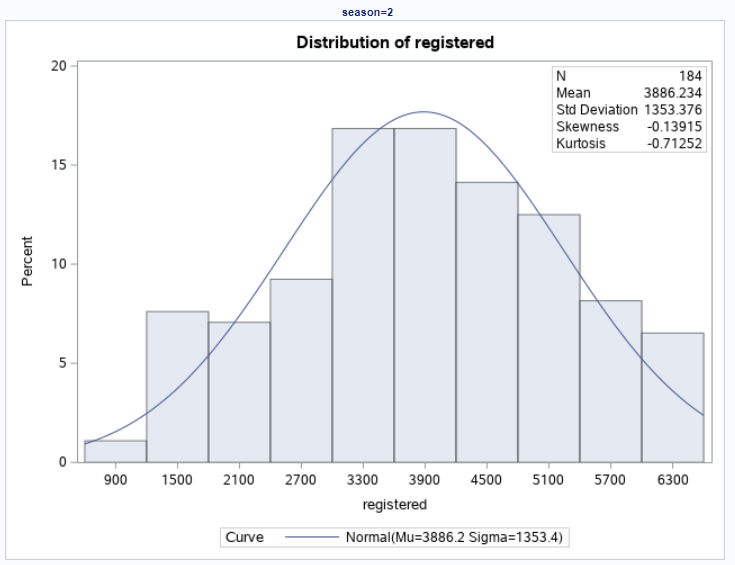


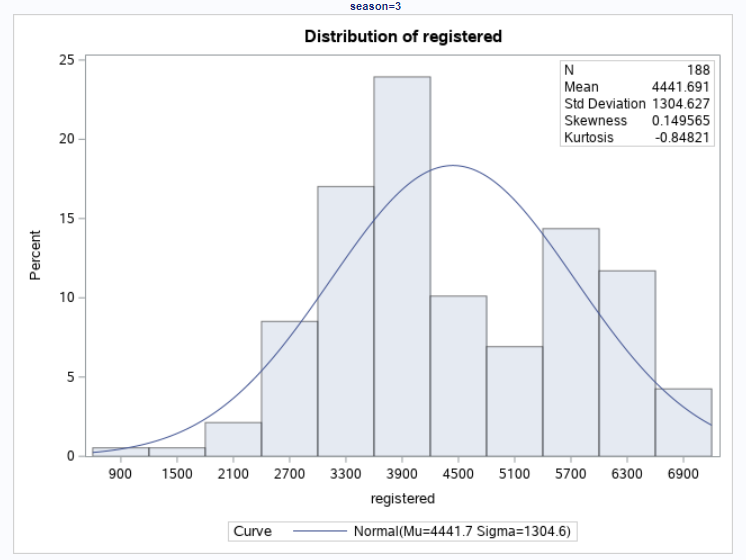
**Assignment Tasks:**

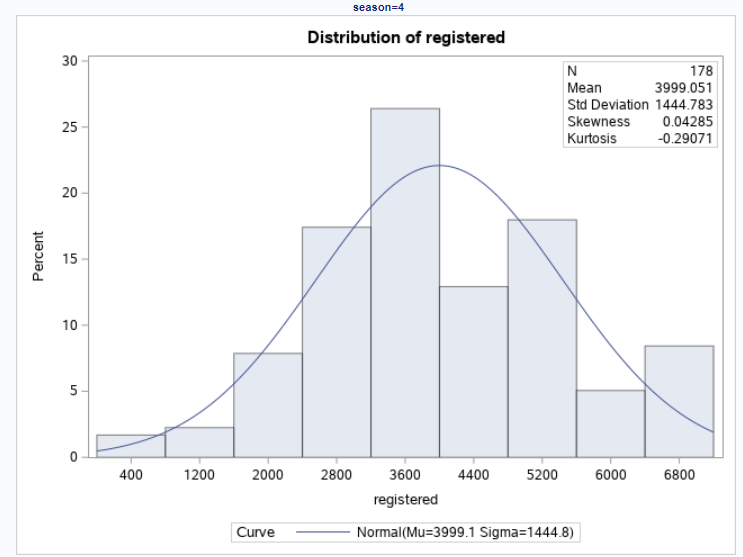
Answer 1:

(a) Distribution of the number of registered users per day (registered) by season.

****

****





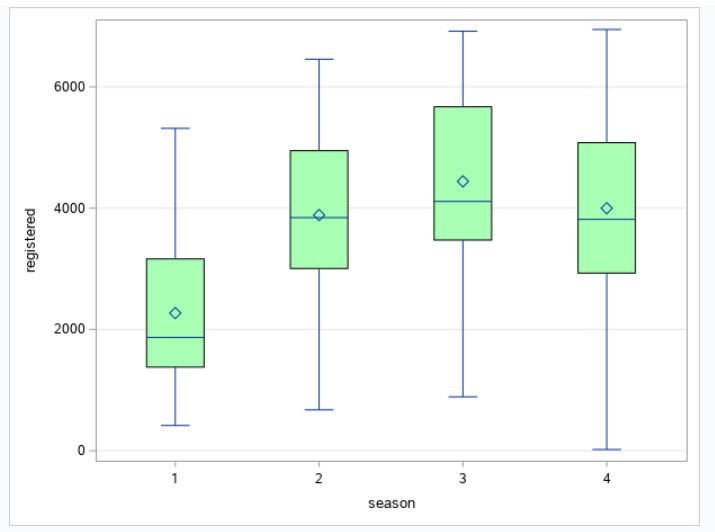
All the above graphs display the distribution of registered count groups by season (**1**:winter, **2**:spring, **3**:summer, **4**:fall). In the case of season 1, we can observe the distribution of the registered group by winter season is positive skewed distribution. As we see that the skewness statistics is 0.65, which indicates that the distribution is positive skewed because when the skewness value is between 0.5 & 1, then the data are slightly skewed or positive skewed. Additionally, we can conclude, in this way, the registered count in the winter season is nearly 2,200.

In the case of season 2, we can observe that the distribution of the registered group by spring season is a nearly symmetrical distribution. As we see that the skewness statistics is -0.14, which indicates that the distribution is nearly symmetrical because the skewness value is between -0.5 & 0.5, then the data are symmetrical. Additionally, we can conclude, in this way, the registered count in the spring season is nearly 3,900.

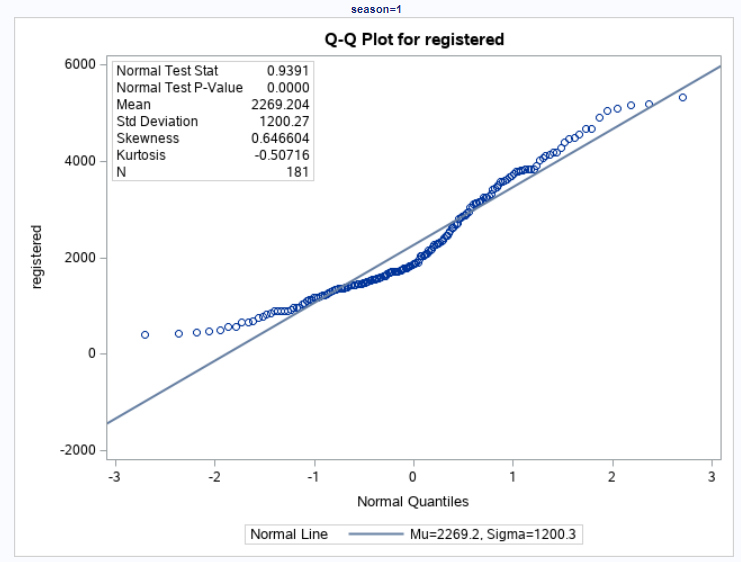
In the case of season 3, we can observe that the distribution of the registered group by summer season is a nearly symmetrical distribution. As we see that the skewness statistics is 0.14, which indicates that the distribution is nearly symmetrical because the skewness value is between -0.5 & 0.5, then the data are symmetrical. Additionally, we can conclude, in this way, the registered count in the summer season is nearly 4,400.

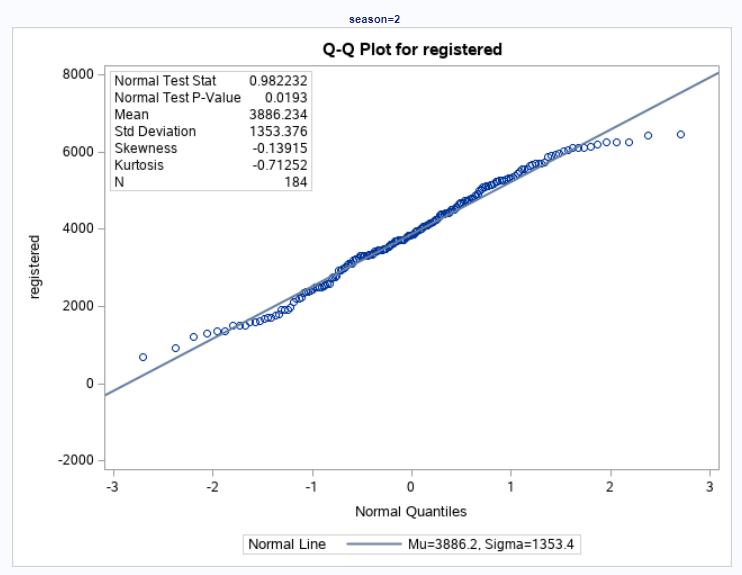
In the case of season 4, we can observe that the distribution of the registered group by fall season is a nearly symmetrical distribution. As we see that the skewness statistics is 0.04, which indicates that the distribution is symmetrical because the skewness value is between close to zero, then the data is symmetrical. Additionally, we can conclude, in this way, the registered count in the fall season is nearly 4,000.

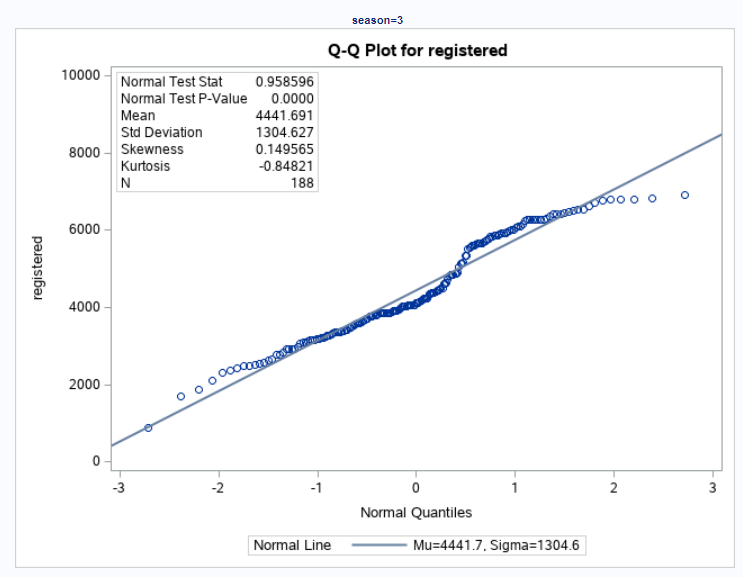
**Box-Plot:**

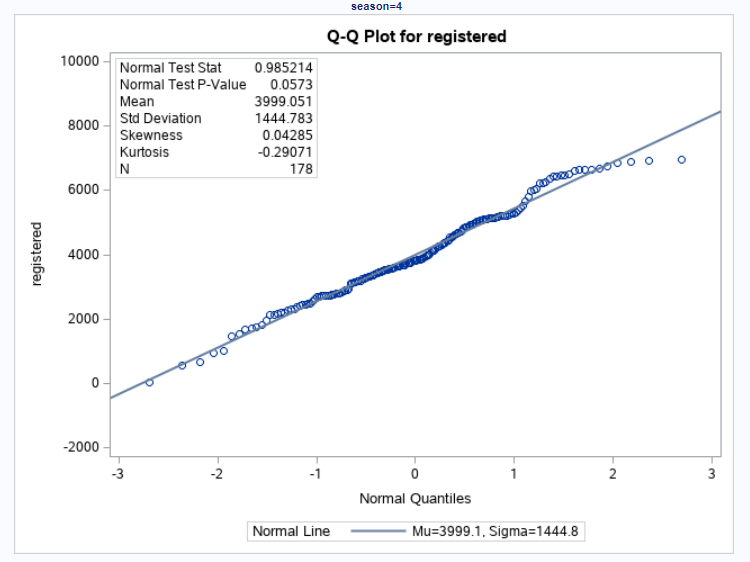
****

**Quantile-Quantile Plot:**

****

****

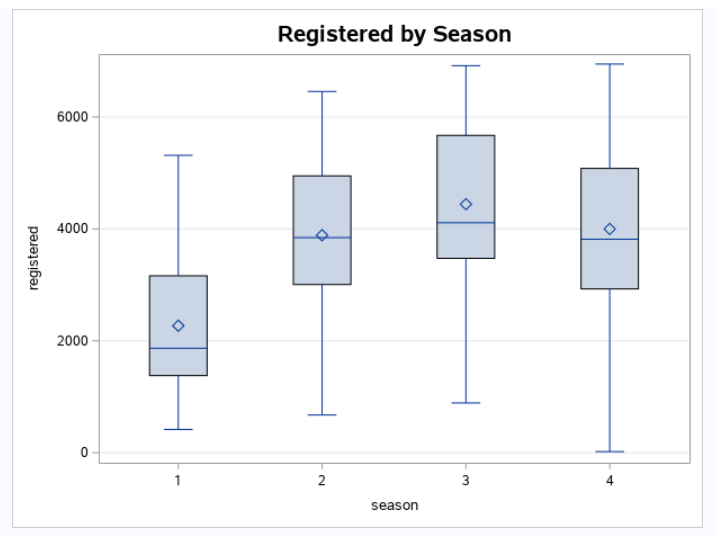
****

****

If the bottom end of the Q-Q plot deviates from the straight line but the upper end is not, then we can clearly say that the distribution has a longer tail to its left or simply it is negatively skewed but when we see the upper end of the Q-Q plot to deviate from the straight line and the lower and follows a straight line then the curve has a longer till to its right and it is positively skewed.

Normal Goodness-of-t test: The test rejects the hypothesis of normality for Q-Q plot 1, 2 and 3 because the p-value is less than, it implies failing the normality test allows you to state with 95% confidence the data does not fit the normal distribution.

(b) Box Plot:

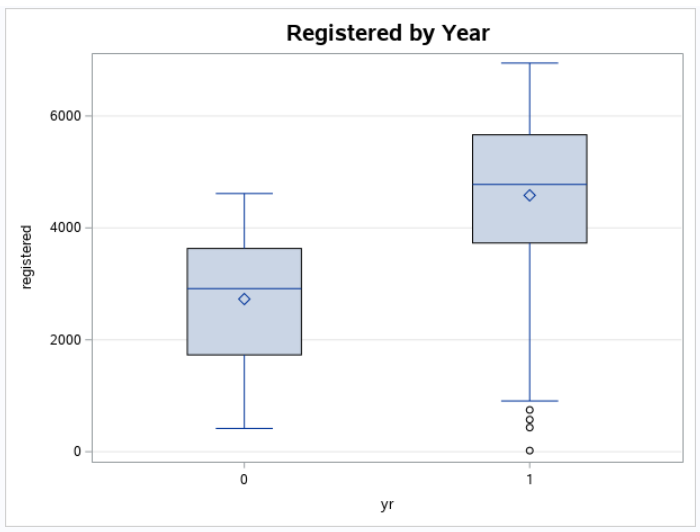


In the above box plot, we can see that the “middle” (or median) line of box 2, 3, and 4 lies above box 1 which indicates that there is likely to be a difference among the groups.

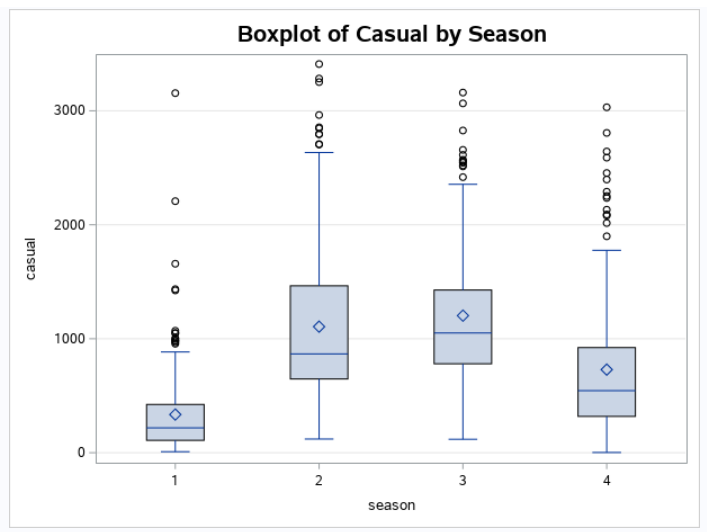
Also, the longer the box the more dispersed the data whereas the smaller the less dispersed the data. There is no outlier in the registered count users.

Interestingly, in season 1 (winter), the daily count of registered users (i.e., bike rentals) of above +2,000. The bike rental count of registered users is below 4000 in season 2 and 4. But daily variation in range is more in season 4 as compared to 2. In season 3, the daily bike rental count of registered users is above +4000 and the daily variation is around 1500 to 7500.

We can easily conclude that registered users are more likely to use daily rental bikes in summers as compared to other seasons.



In the year of 2011 (coded as 0), the yearly bike rental count of registered users is on an average 3500 and the range of the count is 500 (minimum) to 4500 (maximum). On the other hand,in the year 2012 (coded as 1), the yearly bike rental count of registered users is on an average 5000, but there are few outliers in the data set. Also, the minimum count of 2012 is 1500 and the maximum count is more than 7000.



In the above box plot, we can see that the “middle” (or median) line of box 2, 3, and 4 lies above box 1 which indicates that there is likely to be a difference among the groups.

Also, the longer the box the more dispersed the data whereas the smaller the less dispersed the data. There are many outliers in the casual count of users for all the seasons.

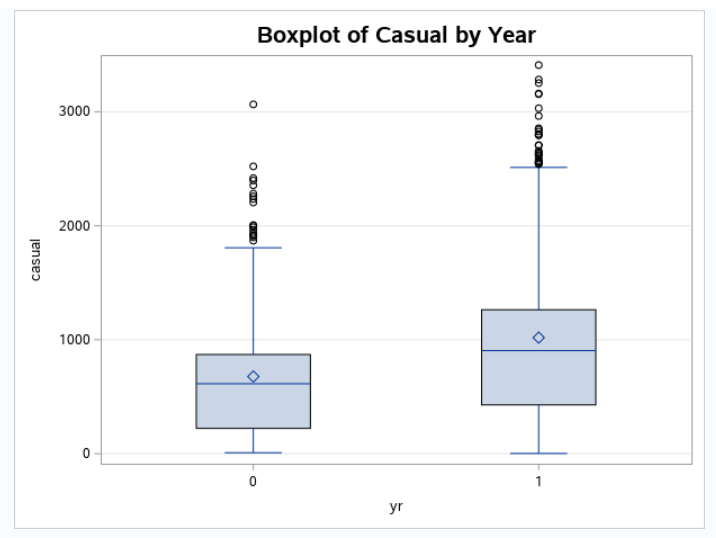
The daily casual count of users in winter on an average is less than 500.

The daily casual count of users in spring on an average is +500.

The daily casual count of users in summer on an average is +1000.

The daily casual count of users in fall is on an average 500.

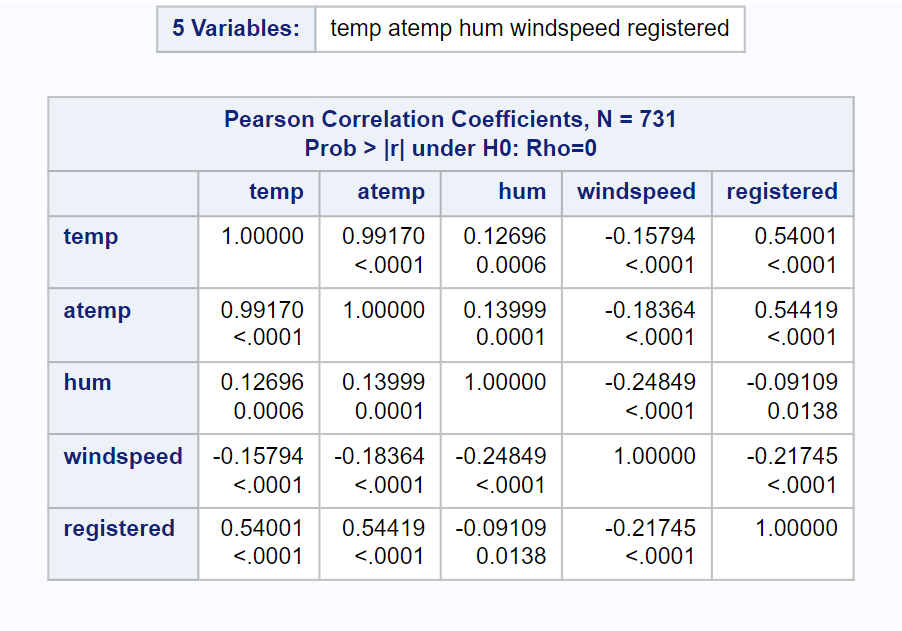
We can easily conclude that casual users are more likely to use rental bikes in summers as compared to other seasons. The reason may be that it is a vacation time.

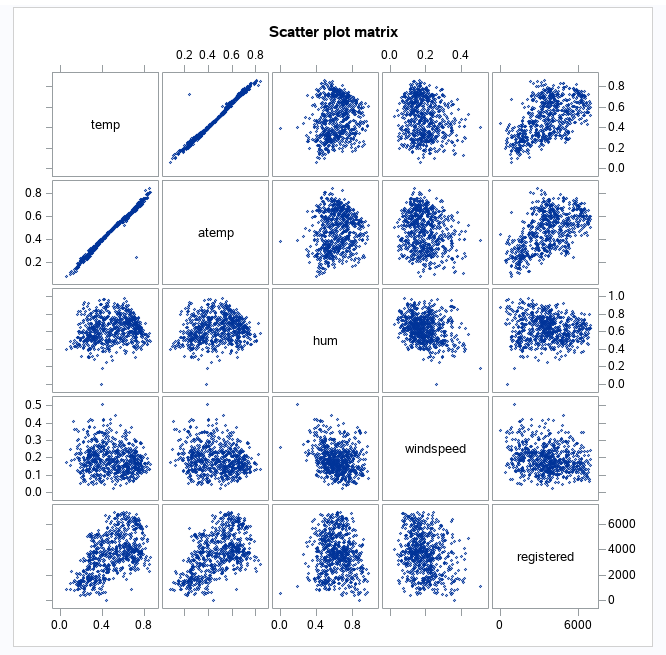


In 2011 and 2012, casual users had shown outliers in the boxplot. As we observe that the daily casual count of users in 2011 and 2012 on an average is +500 and +800, respectively. But casual users count variation more in 2012.

Answer 2:

(a) Pearson Correlation Matrix and Scatter Plot Matrix

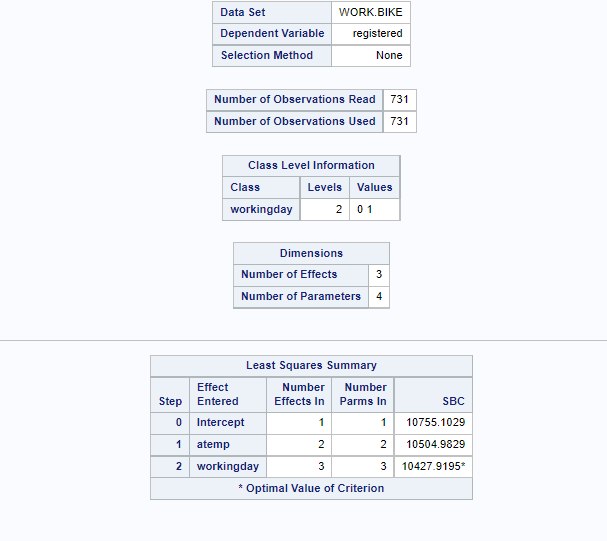


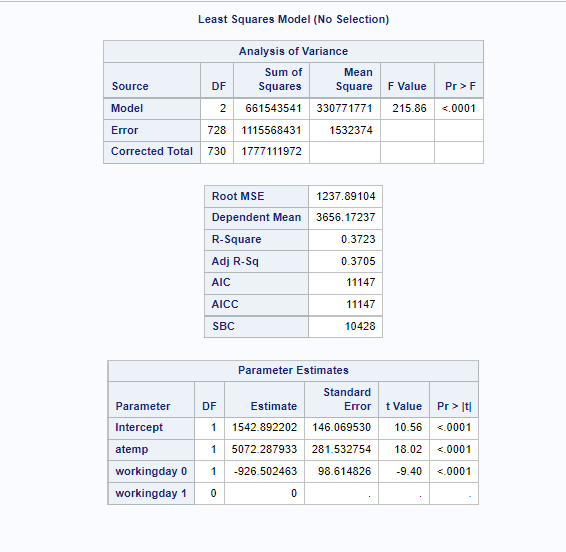


The correlation matrix tells the exact correlation value between the variables, similarly, scatter plot tells that type of data visualization that shows the relationship between different variables.

So, the Pearson correlation or scatter matrix plot, indicates that the temp and atemp has a very strong (perfectly positive) correlation i.e., r = 0.99. On the other hand, there is a positive correlation between the temp and registered, and the atemp and registered, r = 0.5. Also, Registered and hum has no or zero correlation.

(b) Regression Results:





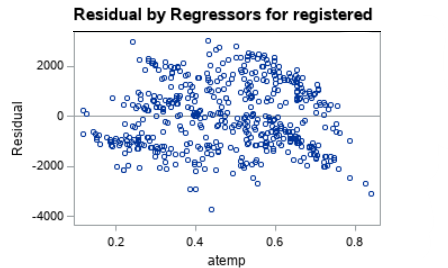
In this model, the dependent variable is registered and the independent variables are atemp and workingday which is set =1. It is a dummy variable and can take 1 (Working day) and 0 (weekend and public holiday). We run a regression using SAS and get the results above. The regression equation is

Registered = b0 + b1atemp + b2\*workingday

Registered = 1542.89 + 5072.29\*atemp - 926.50\*workingday

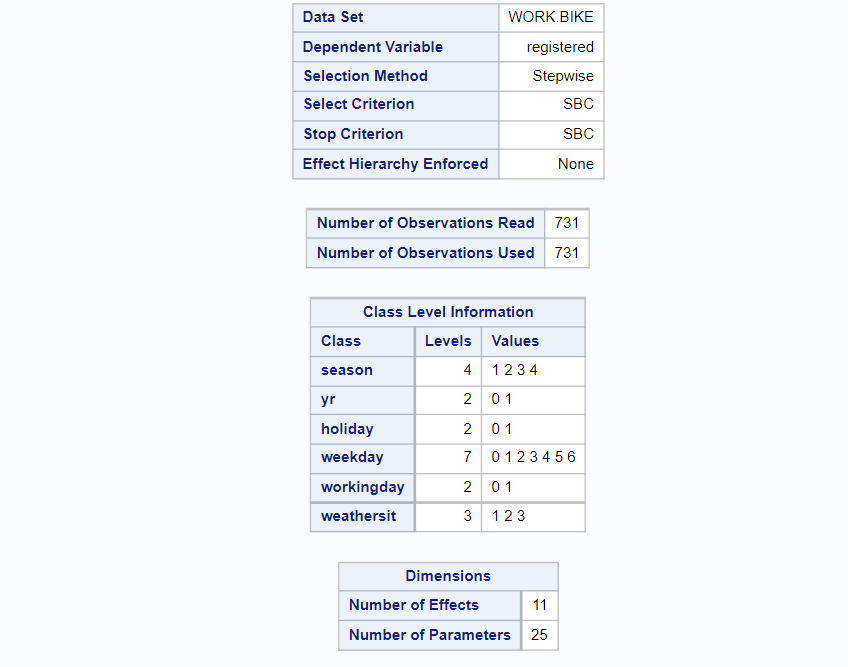
As the p-value of the variables is less than 0.000 which means that the variables are statistically significant in the model.

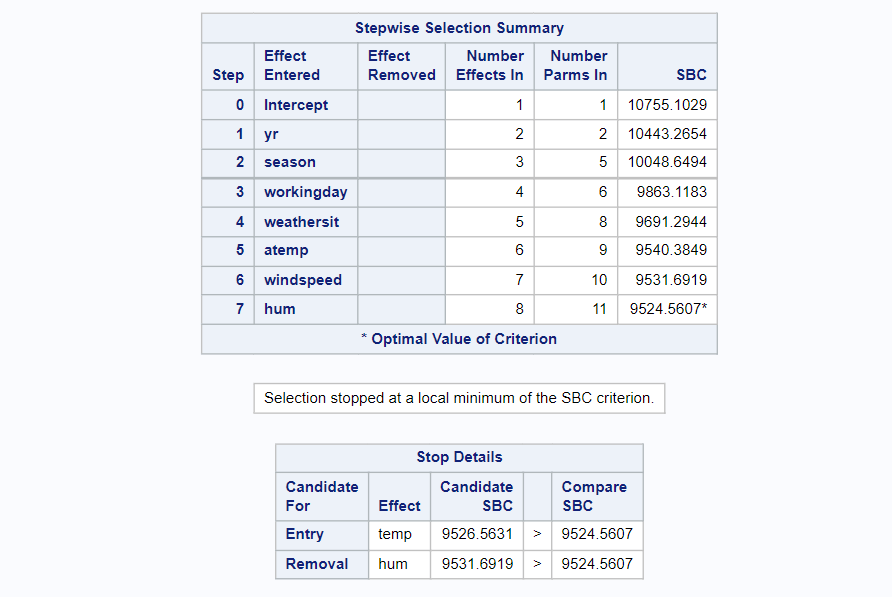
On the other hand, the r-square of the model is 37% which implies that the 37% of the variation in the dependent variable (registered) can be explained by the variation in the independent variables. We can observe that the r-square of the model is low and more than 50% of the variability in the outcome data cannot be explained by the model.

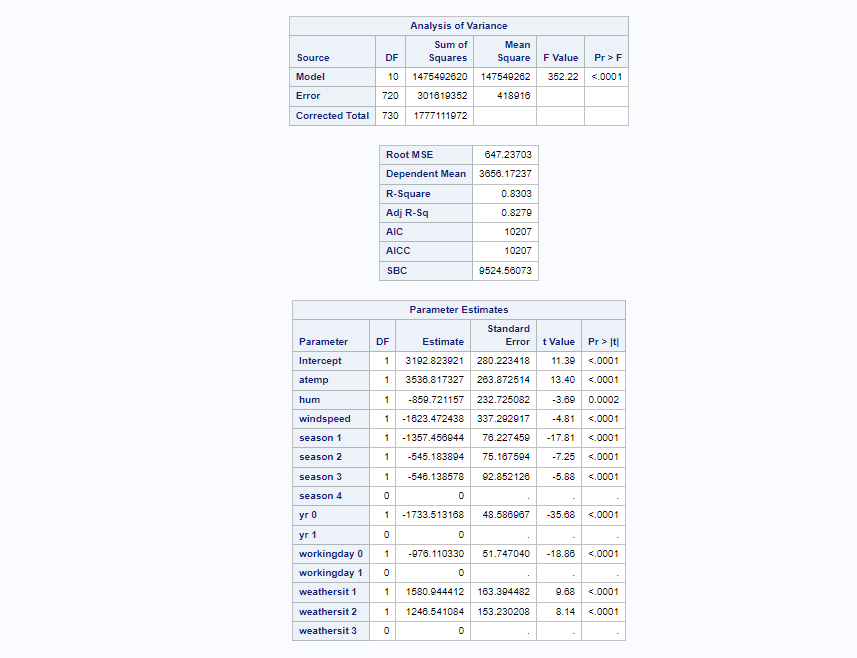


The above chart displays the residual plot and we can see that the residual = 0 line in the residual plot. The registered points scattered randomly away from the residual = 0. So based on the residual pattern, we can say that the linear model is not too appropriate.

(c) Regression Result:







In the above model, the dependent variable is registered and the independent variables are atemp, hum, windspeed, season, yr, workingday, and weathersit. We run a regression using SAS and get the results above.

The regression equation is

Registered = b0 + b1\*atemp + b2\*hum + b3\*windspeed + b4\*season1 + b5\*season2 + b6\*season3 + b7\*yr0 + b8\*workingday1+ b9\*weathersit1 + b10\*weathersit2

Registered = 3192.82 + 3536.82\*atemp - 859.72\*hum - 1623.47\*windspeed - 1357.46\*season1 - 545.18\*season2 + 546.14\*season3 - 1733.51\*yr0 + 1580.94\*weathersit1 + 1246.54\*weathersit2

Slopes Interpretations:

The slope of hum is -895.72, which implies that for each increase of 1 normalized humidity in hum, then the value of count of registered users is estimated to decrease by 859.72.

The slope of windspeed is -1623.47, which implies that for each increase of 1 normalized wind speed in hum, then the value of count of registered users is estimated to decrease by 1623.

Similarly, we can interpret for other variables.

For this regression equation, we run a stepwise regression method of regressing multiple variables while simultaneously removing those that aren't important, in other words, a regression model that involves the selection of independent variables to be used in a final model.

Model Diagnosis:

Look at the p-value column, where all the p-value of the variables is less than 0.00 which means that the variables are statistically significant in the model. Furthermore, the p-value of the model is less than 0.05 which means that the model is statistically significant.

On the other hand, the r-square of the model is 83% which implies that the 83% of variation in the dependent variable (registered) can be explained by the variation in the independent variables.

We can observe that the r-square of the model is high and more than 80% of the variability in the outcome data can be explained by the model.



The above plot displays the residual plot and we can see that the residual = 0 line in the residual plot. The registered points scattered randomly around the residual = 0. So based on the residual pattern, we can say that the linear model is appropriate.

(d)



In the above model, the dependent variable is registered and the independent variables are atemp, hum, windspeed, season, yr, workingday, and weathersit. We run a regression using SAS and get the results above.

The regression equation is

Registered = b0 + b1\*atemp + b2\*hum + b3\*windspeed + b4\*season1 + b5\*season2 + b6\*season3 + b7\*yr0 + b8\*workingday0+ b9\*weathersit1 + b10\*weathersit2

Registered = 3192.82 + 3536.82\*atemp - 859.72\*hum - 1623.47\*windspeed - 1357.46\*season1 - 545.18\*season2 + 546.14\*season3 - 1733.51\*yr0 -976.11\*workingday0+ 1580.94\*weathersit1 + 1246.54\*weathersit2

Slope Interpretations:

The slope of hum is -895.72, which implies that for each increase of 1 normalized humidity in hum, then the value of count of registered users is estimated to decrease by 859.72.

The slope of windspeed is -1623.47, which implies that for each increase of 1 normalized wind speed in hum, then the value of count of registered users is estimated to decrease by 1623.

Similarly, we can interpret for other variables.

For this regression equation, we run a stepwise regression method of regressing multiple variables while simultaneously removing those that aren't important, in other words, a regression model that involves the selection of independent variables to be used in a final model.

Model Diagnosis:

Look at the p-value column, where all the p-value of the variables is less than 0.00 which means that the variables are statistically significant in the model. Furthermore, the p-value of the model is less than 0.05 which means that the model is statistically significant.

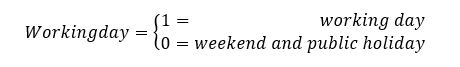
On the other hand, the r-square of the model is 83% which implies that the 83% of variation in the dependent variable (registered) can be explained by the variation in the independent variables.

We can observe that the r-square of the model is high and more than 80% of the variability in the outcome data can be explained by the model.



The above plot displays the residual plot and we can see that the residual = 0 line in the residual plot. The registered points scattered randomly around the residual = 0. So based on the residual pattern, we can say that the linear model is appropriate.

Variable:



If it takes value 0, then this variable is not present in the regression equation and both the model changes but p-value in both the models remains unchanged.

Answer 3: After all the analysis, we came to the conclusion that the count of registered users of rental bikes in the summer season was higher (around +4000) than the other season counts. Similarly, the count of casual users for rental bikes in the summer season was higher (around +1000) than the other season counts.

Furthermore, the count of registered and casual users is more in 2012 as compared to 2011 i.e., +4500 and +800, respectively.

On the other hand, we investigated that there is no or close to zero correlation between registered users and normalized humidity. Alongside, there is a strong or perfect positive correlation between registered users and atemp.

The regression analysis using stepwise regression (part-c) method is the best model as compared to the other models because the p-value of the model and all the variables are statistically significant.