Data Report for Autolib Electric Car-Sharing Service Company: Blue Cars

Summary of the Data Set.

The data presented shows the electric car usage in a certain area. Autolib electric car-sharing service company is interested in investigating a certain claim about the blue cars from the provided Autolib dataset.cars. The data scientist is tasked with the duty of formulating a hypothesis around the data given observation during exploratory data analysis. She then has to test the hypothesis and determine whether to accept the null hypothesis or to reject the null hypothesis.
The data can be found here [http://bit.ly/DSCoreAutolibDataset and http://bit.ly/DSCoreAutolibDatasetGlossary

Understanding the Data

Importing Libraries

She began by importing the necessary libraries and loading the data set in order to have a feel of the data.

Data Cleaning

The data set is large and is arranged in dates. The data scientist began by cleaning the data through removing any null values and checking the data types. Upon checking the data types, she realized that the date types were objects while the postal code were integers. She changed the date to date type and the postal code to string. She changed the postal code to string since the numerical value of the postal code do not have an actual bearing. Upon making relevant changes, she checked the data types again just to be sure. She then changed the values of the rows in the column ‘Day of week’ from numeral 0-6 to actual days of the week. In this case, she assumed that 0 is Monday while 6 will be Sunday.

Dealing With Anomalies

She then did a box plot for the blue cars returned as well as the blue cars taken in order to see if there were any anomalies. She discovered that she had some anomalies but since the data was fairly evenly distributed she did not remove any anomaly.

Exploratory Data Analysis

She then did some univariate and bivariate exploratory data analysis in order to see what kind of data she was dealing with. She noticed that there seemed to be a correlation between the number of blue cars taken and the number of blue cars returned with the day of the week. She noticed that there seemed to be great activity in the weekdays more than the weekends. She decided to explore this further and noticed that while the number of cars on weekends was lower than weekdays, on particular days, the number of cars used over the weekend was far much higher. She also noted that the number of blue cars taken and that of blue cars returned was fairly equal throughout the week, This could mean that almost all the cars taken were eventually returned.

Hypothesis Testing

She came up with the hypothesis that the mean number of blue cars taken on saturday and sunday were equal and to test this hypothesis.
Null Hypothesis: The mean rate of ordering blue cars on Saturday is equal to Sunday

Alternative Hypothesis: The mean rate of ordering blue cars on Saturday is not equal to Sunday.

She chose a sample of 35 from the data for both Saturday and Sunday for the blue cars taken. To choose her sample, she used simple random sampling as there seemed to be no other correlation between the variables.
She used the z score since the sample size is larger than thirty. She calculated the means of each of the samples as well as the variance. The variance was quite large. She assumed a standard deviation of 221 for both samples as they were not that different from one another. She then calculated the probability using the second score. This was a two tailed test as the mean could either be higher or lower. Using a 95% confidence interval, we determine that 47% is within the confidence level and thus we do not reject the null hypothesis. We indeed agree that the means for saturday and sunday for blue cars taken is fairly equal.

Recommendations and Conclusions

While this was not explored in depth, there is evidence that a lot of cars are used over the weekend. The company should therefore pay more attention to the weekends and ensure that none of their customers miss cars on the weekends. Further, in the weekend, none of the two days is of more consequence than the other. This means that the company should pay attention to both days with equal measure to get optimum results. The data has a large standard deviation meaning that a lot of figures are away from the means. This could mean that there is a surge somewhere that needs to be investigated further.