Joanna Riascos

Dr. Jaume Sylvian

Assignment Five

Capstone Big Data Analytics

February 17, 2017

**Algorithm Overview and Implementation in Zeppelin**

What we intend to create is an algorithm that expresses what we would like to accomplish with our parking dataset. The algorithm consists of running a linear regression and a time series analysis. What is linear regression? Linear regression is a methodology for modeling many relationships between a scalar dependent variable y and one or more variables denoted by x. If it’s only one explanatory variable then we called it simple linear regression. If it’s more than one explanatory variable then it’s called multiple linear regression. Linear regression is a supervised machine-learning algorithm. It is the foundation of many complex and long algorithms. In linear regression the relationships are modeled by using and implementing predictor functions where the unknown model parameters reside in the data. These types of models are linear models.

Likewise, linear regression has many useful and practical uses. Linear regression can be used to fit a predictive model to an observed dataset of x and y values. Afterwards, when developing this type of model, an additional value x is then given without its accompanying value of y, which is the fitted value. If we get a variable, y and a number (x1,..xp) it might be related to y. Linear regression analysis can be implemented to count the relationship between y and x1 which may result in not having a relationship at all and to analyze which subsets of the x1 contain useful information about y. Similarly, linear regression implements a statistical model that even though some relationships between independent and dependent variables are almost linear they will still show optimal results and observations. Linear regression models, most of the time are fitted using the least squares method. The least squares technique is the easiest and most common used form of linear regression and delivers an optimal solution to the issue of having to find the best fitting straight line through a set of points. The data is then transformed so that the resulting line is straight. Linear regression models may be fitted in many other ways however; they may also be fitted by minimizing the “lack of fit” or by minimizing a penalized version of the least square function as in a ridge function.

**Assumptions**

The main assumptions of the linear regression equation are the following:

* Linearity – the mean of the response variable is a linear combination of the parameters and the predictor variables. This is very much less restrictive than it may at first seem.
* Homoscedasticity (constant variance) – means that the different response variables have the same variance in their errors regardless of the predictor variables.
* Independence of errors – this assumes that the errors of the response variables are uncorrelated with each other. No correlation between consecutive errors in such a case of time series data.
* Multicollinearity - happens when the independent variables are not independent from each other. Also, the error of the mean has to be independent from the independent variables.
* Normality – when we draw a histogram of the residuals and then examine the normality of the residuals. If we see that the residuals are not skewed, then that means that the assumption is satisfied.
* Equality of variance – if we create a scatter plot and the residuals do not depict in a triangular manner then that means that the equal variance assumption is met.

**Least Square Estimation and Related Techniques**

Ordinary Least Square (OLS)– is a method for estimating the unknown parameters in a linear regression model. It is theoretically simple and computationally straightforward. It is used to establish a line of best fit by minimizing the sum of squares created by the mathematical function.

**Generalized Least Squares**

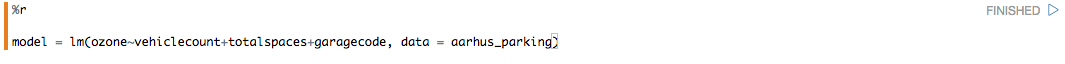
Is an extension of the OLS method that allows efficient estimation of when either heteroscedasticity or correlations or both are present among the terms of the model as long as the form of heteroscedasticity and or correlations or both are present among the error terms of the model as long as the of heteroscedastically and correlation is known independently of the data. GLS can be implemented to perform linear regression when there exists a particular degree of correlation between the residuals in a regression model.

The reason for implementing linear regression is to identify the impact of the parking management system by using different and various tool such as: Zeppelin and Tableau. By running a linear regression we will read in the data and analyze the dataset. We will take as a dependent variable the “Ozone” field and the other fields vehicle count, total spaces, and garage code as the independent variables. What we have done here is that we have merged the two dataset files, the parking and the pollution dataset. We believed that it was a good idea of adding the “Ozone” field from the pollution dataset to our parking dataset since our idea was to investigate how smart parking can help the environment. The reason as to why we chose the “Ozone” column is because car emissions are one of the major contributors of air pollution.

First of all, we will load the parking dataset “aahrus\_parking.csv” file on Zeppelin. Next, we will choose our dependent and independent variables to run the linear regression analysis. Subsequently, we added the following import packages for the linear regression and created the Spark environment.

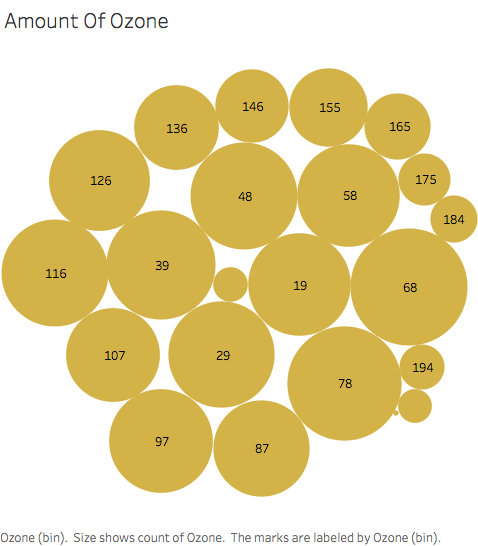
* **import org.apache.spark.sql.functions.\_**
* **import org.joda.time.format.DateTimeFormat**

****

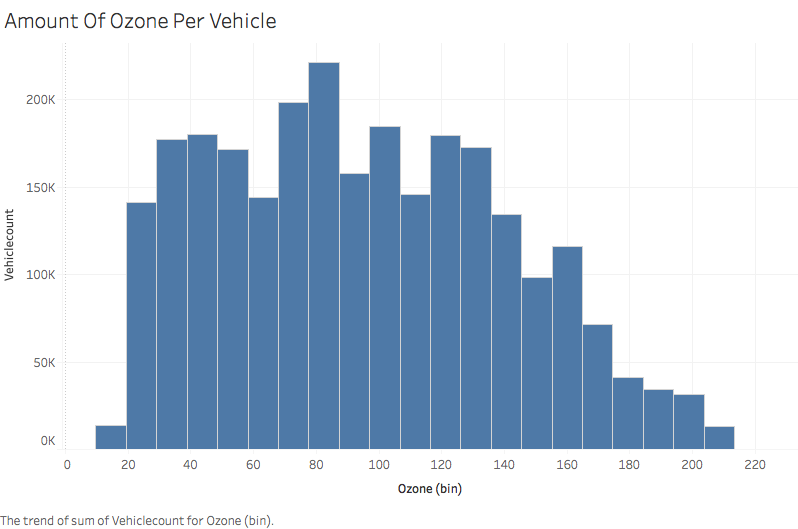
****

We then added the %r in the script to run our R code in each sections of the code. After running the code we can see that the significant coefficient is vehicle count. The vehicle count is for how many cars there is, the total spaces is for the amount of spaces in the parking lot or garage, the garage code is the name of the garage space and ozone is the amount of ozone.

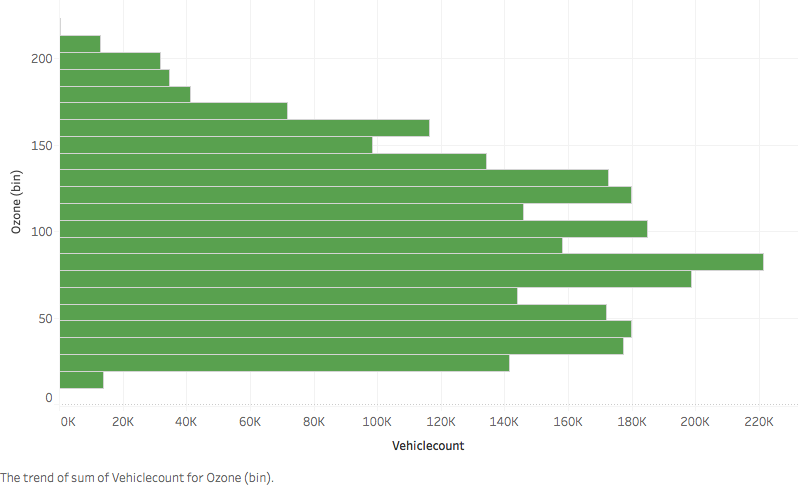
Afterwards, we uploaded our parking and pollution csv file on Tableau and put the vehicle count, total spaces, garage code, and ozone columns in the x and y axis. We chose Tableau to create the charts because it is great for creating and implementing data visualizations. The first chart that we created with Tableau was the Bubble chart, which consists of the total amount of ozone.



Then we went ahead and created a bar chart with the vehicle count and with the total amount of ozone. With this we can see that there is a big amount of ozone in regards to the vehicle count.

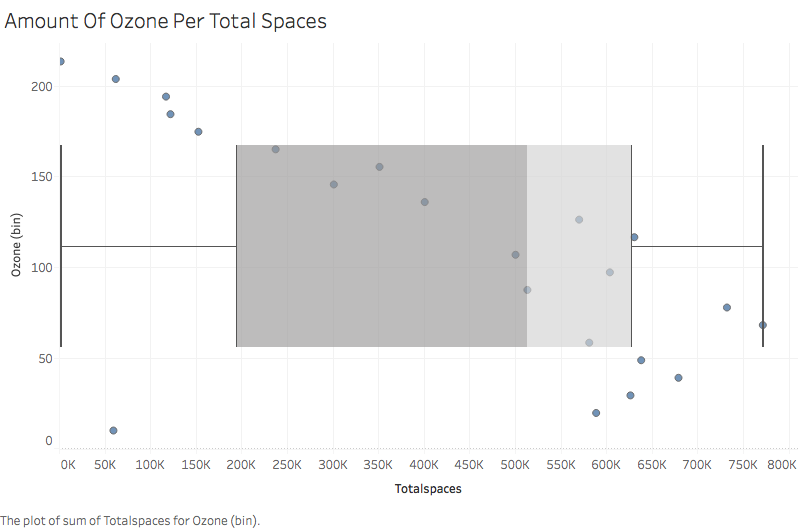


The following graph contains the same information; the only thing that changed was the field that we chose for the x-axis. This time, we chose the ozone field as the y-axis. There is a large amount of the vehicle count and the amount of ozone is large as well.

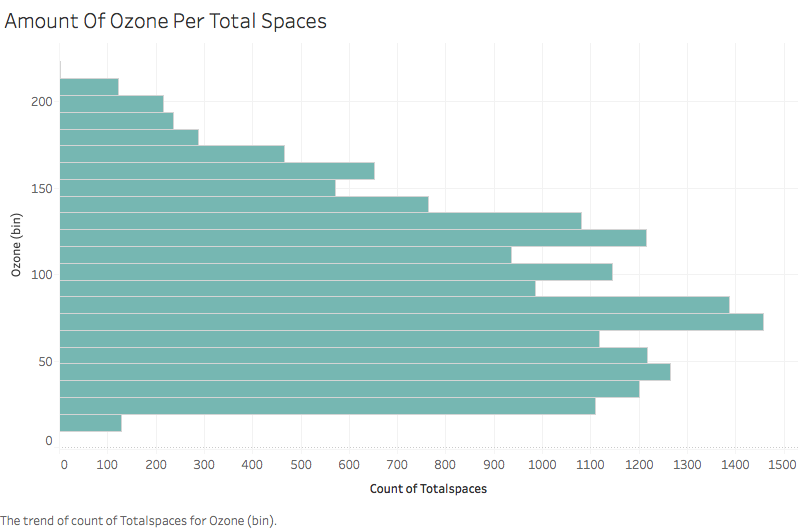


Formerly, we created a box plot for the amount of ozone and the total parking spaces.

As indicated in the box plot, the amount of total spaces is very much high. The box plot is skewed left; the most observations are on the high end of the scale.



The following chart depicts the amount of ozone per total car parking spaces. The more that there is total parking spaces the more there is the amount of ozone. It means that most of the cars that are not parked in the parking lots are outside. Since there are more cars outside, there are more car emissions causing the increase of ozone in the atmosphere.



**Time Series**

Is a collection of observations xt, each of these are recorded at time t. (Time can be discret, t t=1, 2, 3.. or cotinous t>0)

**Time Series Analysis**

Time series analysis is composed of a series of methods for analyzing time series data that extracts meaningful statistics and other characteristics and parts of data. Time series is an ordered order of values of a variable at equally spaced time intervals. With the help of time series we can acquire an understanding of the structure that produced a certain observed data. Time series forecasting is the use of a model to predict future values based upon previously observed values. Time series analysis is used for many applications such as: economic forecasting, sales forecasting, budgetary analysis, stock market analysis, yield projections, process and quality control, inventory studies, workload projections utility studies, and census analysis.

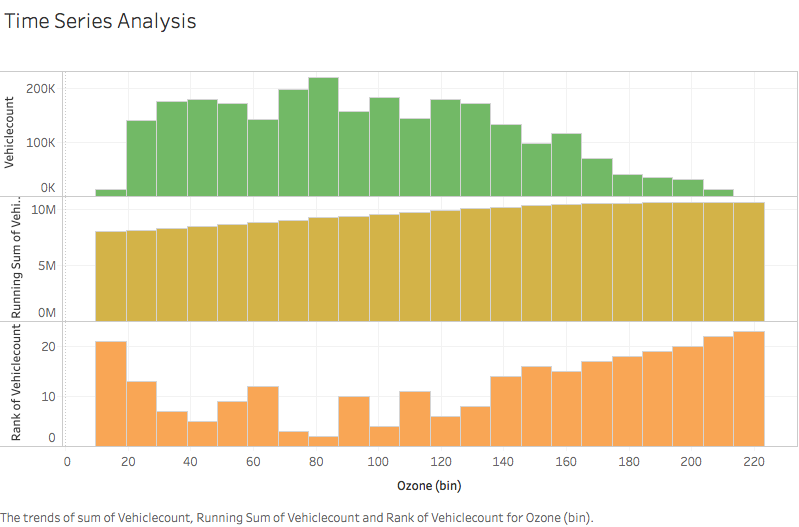
**Objective of Time Series Analysis**

* Data compression (provides compact description of the data)
* Descriptive (identify patterns in correlated data)
* Explanatory (understanding and modeling the data)
* Forecasting: (prediction of future trends from old trends)
* Intervention analysis (how a single event changes within time)
* Prediction (use of the model to predict future values of time series)

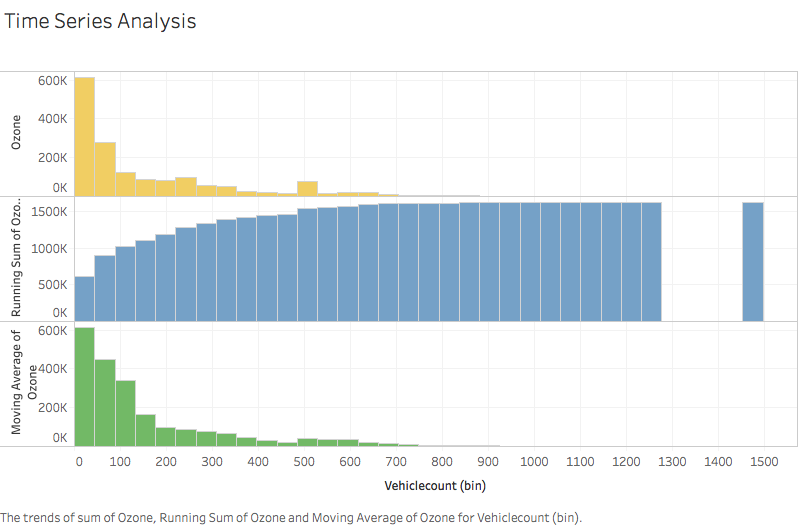
**Advantages Of Time Series Analysis**

* Time series models are quite useful models when you have serially correlated data.
* The best way to deal with temporal effects is by running the time series analysis.
* Spectral analysis (frequency domain)- aims to isolate periodic or cyclical components in a time series.
* Intervention analysis – it is used to establish if an event can lead to a variation in the time series
* Explanative analysis – studies the cross correlation or relationship between two series and the dependence of one on another.

We then ran a time series analysis using the vehicle count and the amount of ozone. First, we took the count of the vehicle count, the sum of the vehicle count, and the rank of the vehicle count. We chose the rank as the data calculation because we wanted to visualize the amount of ozone.

.

Furthermore, the following time series analysis chart is composed of the count of the amount of ozone, the running sum of the ozone, and the moving average of the ozone.



Next, we ran a linear regression analysis by using Zeppelin. As we mentioned previously, we will take the “ozone” field as the dependent variable and the vehicle count, total spaces, and garage code as the independent variables. There are a total of 55264 observations and we are taking four variables. Moreover, in the next section of the paper we will provide the pseudo steps for the linear regression analysis.

**Pseudo Code For Linear Regression**

1. Import the parking dataset in Zeppelin

2. Import the pollution dataset in Zeppelin

3. Merge the parking and pollution datasets together.

4. Call the merge function to call the two datasets

5. Make the Ozone field count the dependent variable and the vehicle count, total spaces, and garage code fields as independent variables.

6. Run the linear regression analysis using the lm function.

7. Check the P-values for each of the parameters.

8. Recognize the parameters that affect the ozone count in the model

9. Plot the graphs.

**Flowchart of Parking System**

Macintosh HD:Users:joannariascos:Downloads:Untitled Diagram (12).png

The flow chart consists of the admin and the user. The admin is in charge of updating and deleting the parking slots. The smart parking system checks whenever a car is parked or not. It checks when a car leaves the garage and when it enters the garage. As soon as an available parking spot is available it notifies the user that needs to make a parking reservation. On the other hand, the user opens the smart parking application and checks which parking spots are available. It also sends the user information of the parking spot details. All of this data gets stored in the database.

In conclusion, our algorithm consists of the linear regression and the time series analysis. As mentioned previously, all of the analyses will run on Zeppelin and there we will import the parking and pollution datasets. Afterwards, we will merge these two files and then we will take the ozone field count as the dependent variable and the vehicle count, total spaces, and garage code fields as independent variables. Then we will call the lm function and run our analysis. Finally, we will plot these values to look at the residuals vs fitted, and qq-plots. By doing this we will get a better understanding of our data values and be able to perform more accurate analysis.