

# **BIA 652**

# **MULTIVARIATE DATA**

# **ANALYSIS**

Spring 2018

Professor: M. Gandomi, PhD

Student: Mukunth Rajendran & Raphael Presberg

## Assignment

So far, we stayed in the first steps of this project which are Business understanding, Data Understanding & Data Preparation. To do so, we made some analysis on the data and run few statistical test.

Concerning the Business Understanding, we figured out for each stakeholder of the project (Drivers, Customers, etc...) which variable we were going to use. This reduce the panda data frame by few and make the system easier.

- Driver

Concerning the driver, with the data we selected, we could predict where the driver has to be at what time to increase his chances to get a client. Also, we could predict the number of client the driver should take if he wants to increase his tip/revenues.

We are selecting the following fields: Trip Time, Fare Amount, Tip Amount, Total Amount, Passenger count Pickup Time, and Pickup coordinates (Pickup Latitude, Pickup Longitude) If we have time, we can add Pick

- Customers

As an international student living in the US, I am not used at all to give tip to the taxi driver. We don't do this in Europe so here, I have no clue of how much should I give to the driver. We could create an algorithm based on the time, the localization, the time trip, if you have luggage or no, to predict how much to give to the driver.

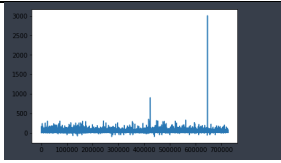
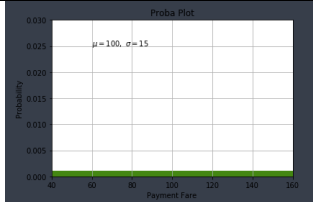
For this we would need this following fields: Trip Time (long? Short?), Pickup Time (night ? day ?), Passenger Count (1? 4?), Surcharge, Tip Amount, and Payment Fare.

- Company
- Environment

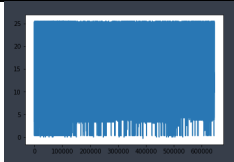
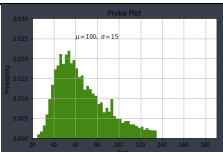
## Project – Extended Abstract

For the Data Understanding and Data Preparation part:

Example for one continuous variable we randomly take: Payment Fare Total

Length	727310
Mean	15
Median	11.15
Variance	173
Standard Variation	13
Range (max – min)	3106
Number of outliers	82703 → 
Distribution	

Let's redo this table with this same field after the data preparation phase, meaning we are taking of outliers.

Length	644707
Mean	10
Median	11.15
Variance	23
Standard Variation	4
Range (max – min)	25
Number of outliers	No more outliers 
Distribution	

We notice that we have to perform this analysis for each variable. By taking of the outliers, we make sure we have a good dataset to use any machine learning technique on.

## Project – Extended Abstract

What do we do now ?

First of all, I think we should reduce our scope. We don't have a lot of time and I want to try several machine learning algorithms. We should focus either on the company, the customer or else.

Then, we still have a lot of data understanding to make by performing statistics tests and data analysis and a lot of data preparation to create a good, workable on dataframe.