**Traffic Violation Data Analysis**

**Python**

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Author** | **Reviewer** |
| Initial Project description | 06/07/2020 | Dinesh Arockiasamy | Marc Lowenthal |
| Project overview | 06/23/2020 | Dinesh Arockiasamy |  |

Table of Contents

[Project Direction 3](#_Toc2318)

[Introduction: 3](#_Toc16929)

[Functional Overview 3](#_Toc30111)

[Tech stack 3](#_Toc21938)

[How to run the program : 3](#_Toc16831)

[How is this useful 4](#_Toc9465)

[#1 Project Idea: 4](#_Toc24731)

[#2 Learning & experience : 4](#_Toc11126)

[Scope of the project : 4](#_Toc2195)

[Input Sample : 5](#_Toc21348)

[Output of the project : 5](#_Toc15079)

[Plans for future enhancements : 7](#_Toc8493)

[Summary : 7](#_Toc31195)

# Project Direction

## Introduction:

I would like to design and develop a program which does exploratory data analysis on traffic violation data from various states of America. The project will leverage Python language and its various frame works along with the power data structures of Python.

## Functional Overview

This project, on a high level will try to extract necessary information from the traffic violation data and will make sense out of the raw data that we have got from internet. This program will bring out a few use full informative analysis such as search by state, search by gender, timing and seasonal details for certain violations etc.

Along with that, this project also does generate an output text file which contains the data that was extracted from csv file.

# **Tech stack**

I have used below tech stack to build this project

Python language

Python modules - Numpy, Pandas, os, random

A data set that is downloaded from kaggle

PyCharm as IDE for developing program

# **How to run the program :**

1. Un zip the folder which contains the code and input files. Please keep the “main.py” as the start up page and run the program to get the results.
2. import “pandas”, “Numpy”, install if not already installed.
3. The output was written into a text file in the same location where the code resides.

# **How is this useful**

To explain how would this be useful, I would like to take two aspects of it.

## #1 Project Idea:

Traffic violation data set contains raw information about any violation or accident like location, vehicle and kind of violation etc., By doing this project, the raw data is cleaned, grouped, filtered and modeled to help insurance companies, Automobile companies and legal departments to make informed decisions about accidents and traffic violations. This data modeling also can help predict a result for similar situations. This project can convert the raw data to information.

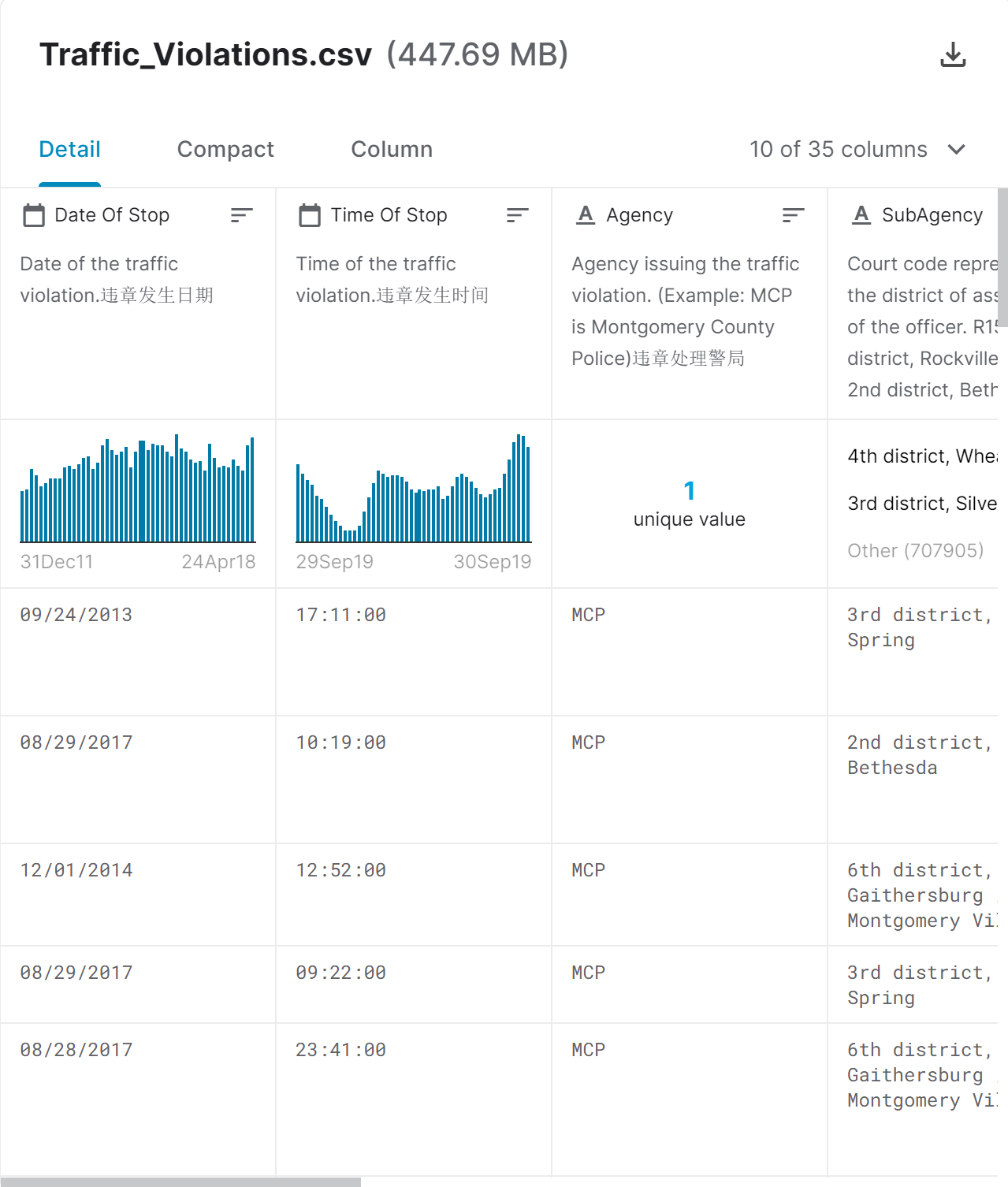
#2 Learning & experience :

As a student, this was a good learning to combine and consolidate the things that was taught through out the course. I could also explore a few more modules like Pandas, Numpy in python which are widely used in Machine Learning. This project would definitely lead my interest towards Machine learning and related stuff in future. This was good set platform to practise almost all the concepts that was learned through materials and homework assignments.

# **Scope of the project :**

Though, I have all the reasons to use Pandas to validate and train the model, I have restricted the scope of the project to few searches from the data set. Training this data set with a selected algorithm and predict the result by keeping the data as input are left for future enhancement of this project.

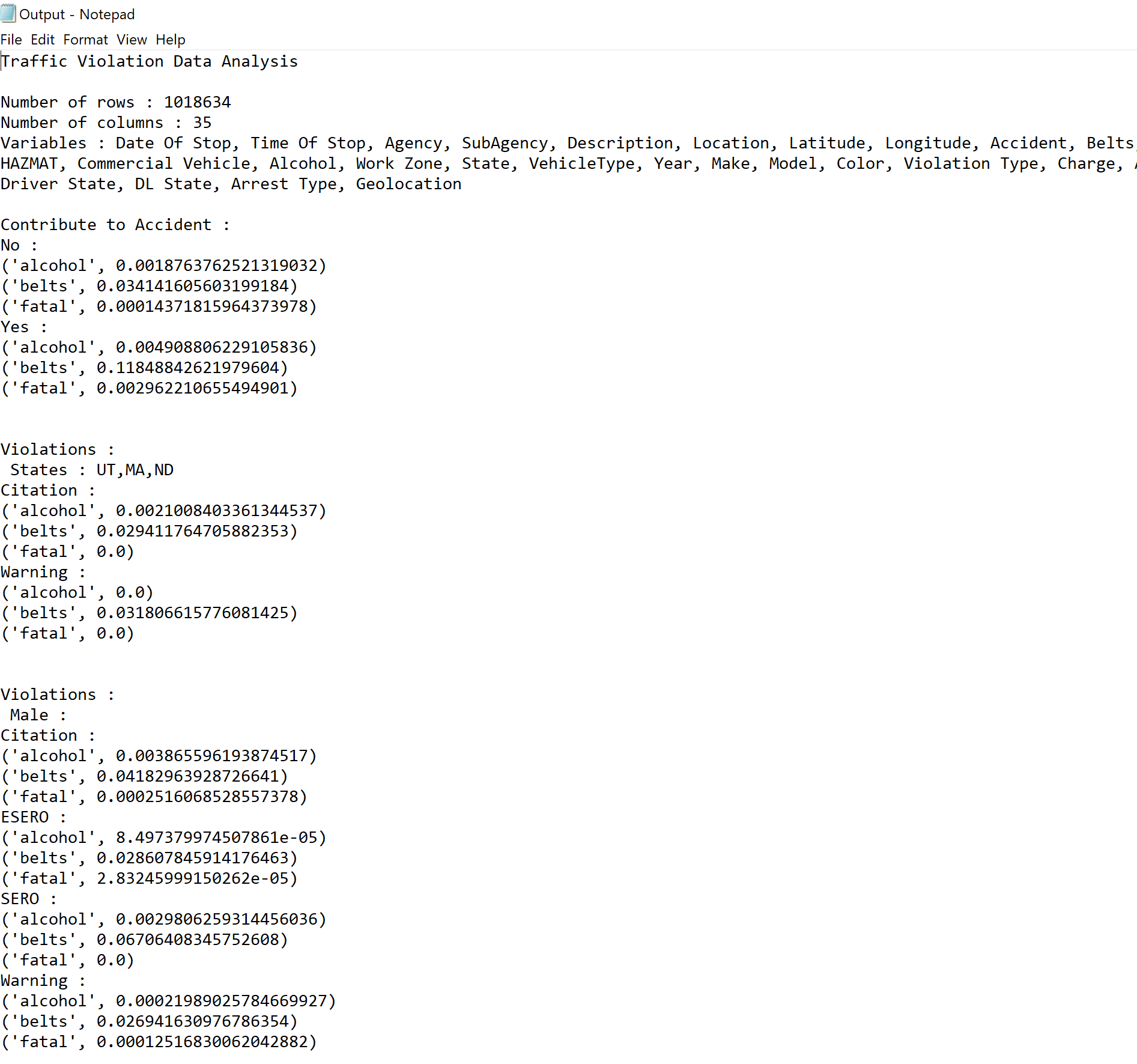
# **Input Sample :**



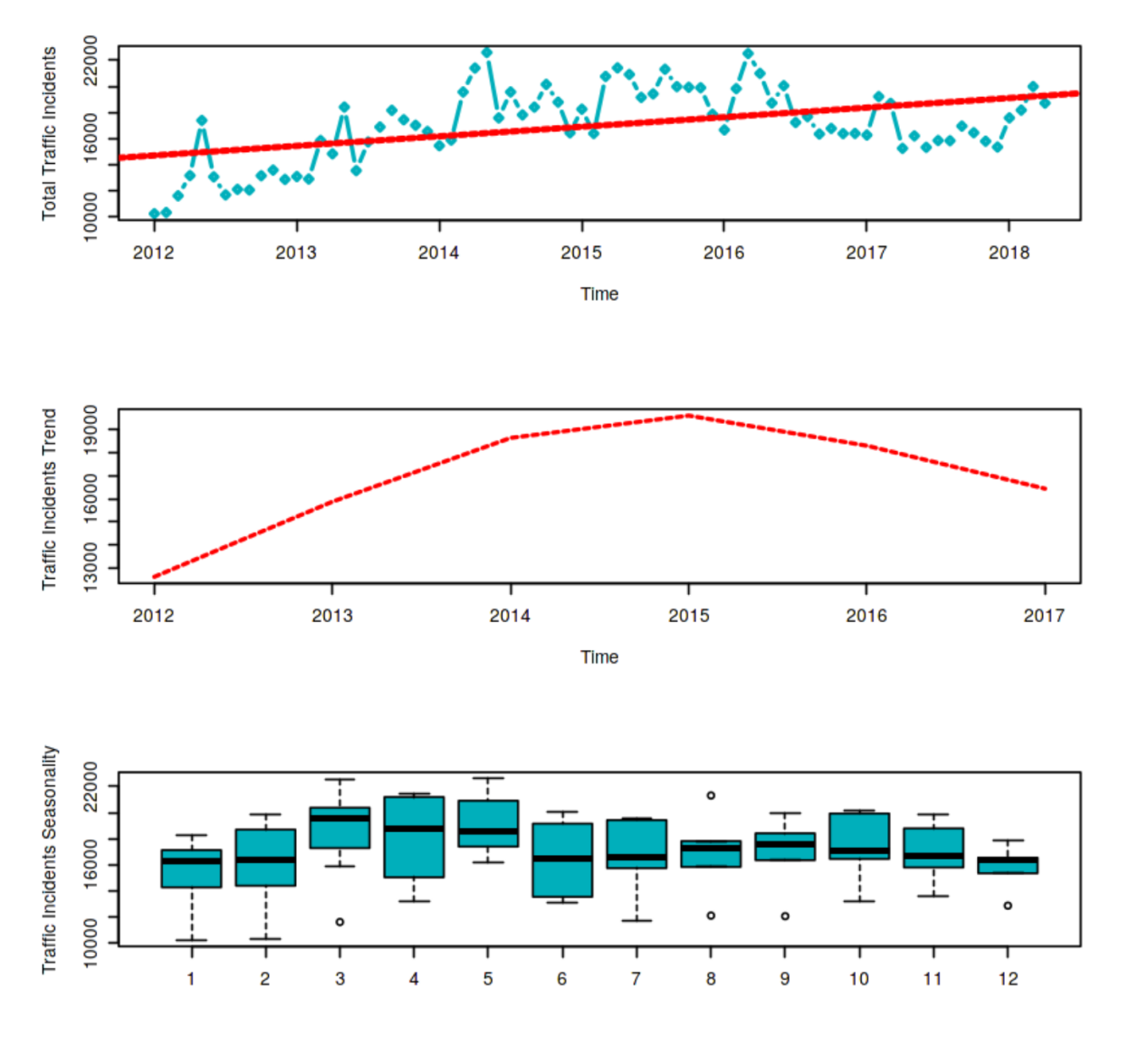
# **Output of the project :**

As I have written the search results with in a text file, I have de-scoped the visualization of data from this project. Plotting would also be a good candidate for enhancement of this project.

Output file :



I was thinking to demonstrate the data using some useful plots like below.



The plots were possible to print it using matplotlib module. But as I was writing the output to a text file, I didn’t include it in the output.

# **Plans for future enhancements :**

1. Implement Data frame slicing and subsetting for training and validation
2. Validate the groups of k-cross validations
3. Pick and choose appropriate algorithm for model training
4. Predict results for similar incidents

# **Summary :**

Overall it was a good learning and a very good first step towards a new phase of my career. I have around 14 years of IT experience and so far it was primarily on coding applications using programming languages like C#, Java, Javascript etc. I hope the graduation with Data analytics concentration along with python/R etc., would bring me into the would of data and data analysis. Thank you for all the help!