**CSCI 59000**

**Big Data Management Project Report**

**Analysis on Traffic Violations in the US**

By

**Keyur Mehta, Love Modi, Darsh Sanghavi**

*Department of Computer Information and Science*

*Indiana University Purdue University Indianapolis, US*

**TABLE OF CONTENTS**

[1 INTRODUCTION 3](#_Toc501206301)

[2 RELATED WORK 3](#_Toc501206302)

[3 PROPOSED APPROACH 3](#_Toc501206303)

[3.1 Requirement Analysis and System Parameter Decisions 3](#_Toc501206304)

[3.2 Preprocessing Data 4](#_Toc501206305)

[3.3 Map and Reduce 4](#_Toc501206306)

[3.4 D3 Visualizations 4](#_Toc501206307)

[4 RESULTS 6](#_Toc501206308)

[4.1 State wise Count 6](#_Toc501206309)

[4.2 Violation count every minute 7](#_Toc501206310)

[4.3 Gender wise Analysis 7](#_Toc501206311)

[4.4 Top Reason of Violation 8](#_Toc501206312)

[4.5 Analysis visualized on Amazon QuickSightTM 9](#_Toc501206313)

[4.6 Tabular Analysis 11](#_Toc501206314)

[5 INDIVIDUAL CONTRIBUTIONS 12](#_Toc501206315)

[6 CONCLUSION 12](#_Toc501206316)

[7 REFERENCES 13](#_Toc501206317)

**List of Figures**

[Figure 1 Project Approach 4](#_Toc501134947)

[Figure 2 Traffic Analysis website 6](#_Toc501134948)

[Figure 3 US State Map showing tickets count 6](#_Toc501134949)

[Figure 4 Violation count per minute 7](#_Toc501134950)

[Figure 5 Gender wise tickets count 7](#_Toc501134951)

[Figure 6 Top Reason of Violation 8](#_Toc501134952)

**List of Tables**

[Table 1: Analysis and Visualization tools 5](#_Toc501204996)

Table 2: QuickSight Analysis 8

*Table 3: Analysis of the Results……………………………………………………………………………………………………………………………… 11*

[Table 4: Member's Contribution 12](#_Toc501204998)

# **INTRODUCTION**

Our objective with this project is to synthesize the traffic violation data maintained by the US government into small bits of useful information. The dataset that we have used has more than 35 parameters spread over 1.5 million rows. With such a large volume of data, it is difficult to extract important information. Big data management tools proved to be an ideal instrument to handle such large amount of data. Our analysis will prove immensely useful for any organization who wants to work on traffic rule implementation. It acts as a guiding parameter for the agencies to target the right community, gender, states etc. to spread awareness, traffic rule education as well stricter traffic violation penalties. These implications of our project were a great motivator for us to select this project in the first place.

We were able to filter out states which were the highest violators of traffic rules. Even the race, gender, vehicle type etc. distributions were computed using the big data skills learned. We also analyzed complex multi-dimensional data collating information from several parameters at one go. Once we were successful in extracting all the required information, we exposed ourselves to many unknown technologies for data visualization. We had to learn technologies like D3, jQuery, Amazon QuickSightTM to visualize the data extracted in an effective, concise and easily interpretable way. We were able to collate all our data and represent it via a website representation hosted over Amazon S3.

# **RELATED WORK**

During our requirement analysis and system design phase, we thought if easier tools (like using an excel sheet with data filters) would be able to do all the work. However, on closer evaluation we found that MS excel is only able to support around 1 million rows. The data filters sometimes don’t work that accurately with enormous amount of data such as this. Also, it is impossible to extract the kind of multidimensional data we were working on. Big data tools proved to be an ideal tool for the kind of data extraction that we wanted to do.

We started our work from scratch albeit the dataset that we collected from a government website [1]. We had to do considerable research while trying to formulate the scripts. Pig Latin documentation [2] helped us with this task. For visualizations we had to look onto a lot of tutorials related to D3.js [3] and read technical documentations for the different technologies that we have used [4].

# **PROPOSED APPROACH**

Our approach has been summed up by Figure 1. Our project is divided into four phases or milestones, each step acting as a base for the next step.

## **Requirement Analysis and System Parameter Decisions**

We started out with building our backend first and used the backend output to develop HTML frontend visualizations. Our first step was to generate requirements out of the dataset. After brain storming on the dataset, we were able to come up with a mix of one and multi-dimensional queries helpful in improving traffic conditions. The requirements document was a great help to keep us on track and in tracking progress. Next, we wrote the required scripts and executed them using Amazon AWS Elastic Map-Reduce (EMR).

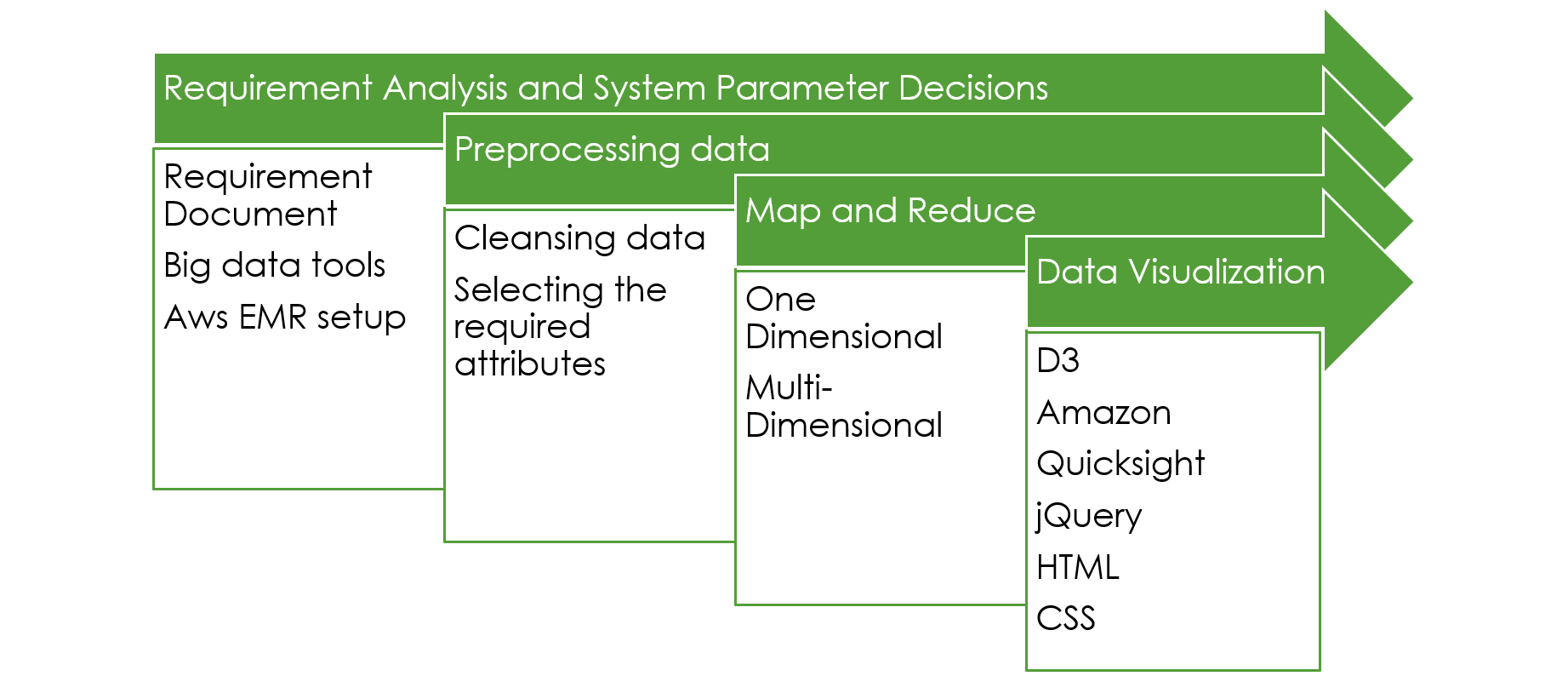


Figure 1: Project Approach

## **Preprocessing Data**

We have a very large dataset maintained by the US government [1]. It has more than 1.5 million rows and around 35 attributes. After our requirements were finalized, we pruned through our data to select the required attributes for our scripts. This was a kind of optimization or innovation as this helped in improving the time of operation. We could cut down on considerable data and hence the query operations were faster.

We also needed to clean the data so that Pig Latin does reads our data as we expect it to. For example, our comma separated version of the original dataset had a lot other comma in the strings as well as data points. This led to incorrect reading of data by Pig Latin (it reads comma and considers the next statement as the value for the next attribute). We had to remove the extra commas from the data. We used Microsoft Excel’s find and replace feature to do so. Once the data was clean, Pig Latin was easily able to read the data as per requirement.

## **Map and Reduce**

This phase basically related to development of the one dimensional and multidimensional scripts. During this exercise, we were able to learn lots of new details (keywords and functions) regarding Pig Latin. The Outcomes section aptly describes the results we obtained during this phase.

## **D3 Visualizations**

Once the scripting part was successfully completed, we worked on developing visualizations using D3, HTML, CSS, jQuery, Amazon QuickSightTM. We developed interactive graphs showing country maps with data numbers along with animated maps using D3 technology. We used Amazon QuickSightTM to visualize our multi-dimensional data as it has a lot of convenient tools. We collated all the maps in a single webpage and hosted it over internet using Amazon S3. With this approach, we were able to display our project effectively during the project demonstration. Below table gives the brief overview of the analysis and its representation techniques:

| **Sr #** | **Analysis (Scripts)** | **Visualization Technique** | **Tool Used** |
| --- | --- | --- | --- |
| 1 | State wise Count - Alcohol, Belt, Commercial License, Personal Injury, Property Damage, Fatal | US State Map | D3 |
| 2 | Violation count every minute | Line graph | D3 |
| 3 | Gender Wise Analysis - condition HAZMAT, Belt, Alcohol, Commercial License | Pie chart | D3 |
| 4 | Top Reason of Violation | Bubble chart | D3 |
| 5 | Vehicle type Alcohol yes - count Alcohol No - count | Clustered bar combo chart | Amazon QuickSight |
| 6 | Count of arrest type | Vertical Bar graph | Amazon QuickSight |
| 7 | Vehicle color wise ticket count | Bar graph | Amazon QuickSight |
| 8 | Race wise ticket count | Horizontal bar graph | Amazon QuickSight |
| 9 | Hourly ticket count | Pie chart | Amazon QuickSight |
| 10 | Vehicle type ticket count | Bar graph | Amazon QuickSight |
| 11 | Violation type count | Pie chart | Amazon QuickSight |
| 12 | Violation type (4 types) Gender | Multi column Bar | Amazon QuickSight |
| 13 | Month-Year ticket count | Area Line graph | Amazon QuickSight |
| 14 | Average daily ticket count | Line Graph | D3 |
| 15 | Gender wise count | Line graph | Amazon QuickSight |
| 16 | Time of Day - vehicle type | Line graph | Amazon QuickSight |

Table 1: Analysis and Visualization tools

# **RESULTS**

Once we obtained the analytical results using Elastic Map & Reduce, these results were required to be visualized in user understandable form. For this we have used 2 different methods for visualization. We have created a webpage to display all the analysis in an interactive form. This webpage is hosted on the Amazon S3 Cloud platform and can be access using below link:

<http://iupui-big-data-project-final.s3-website.us-east-2.amazonaws.com/>

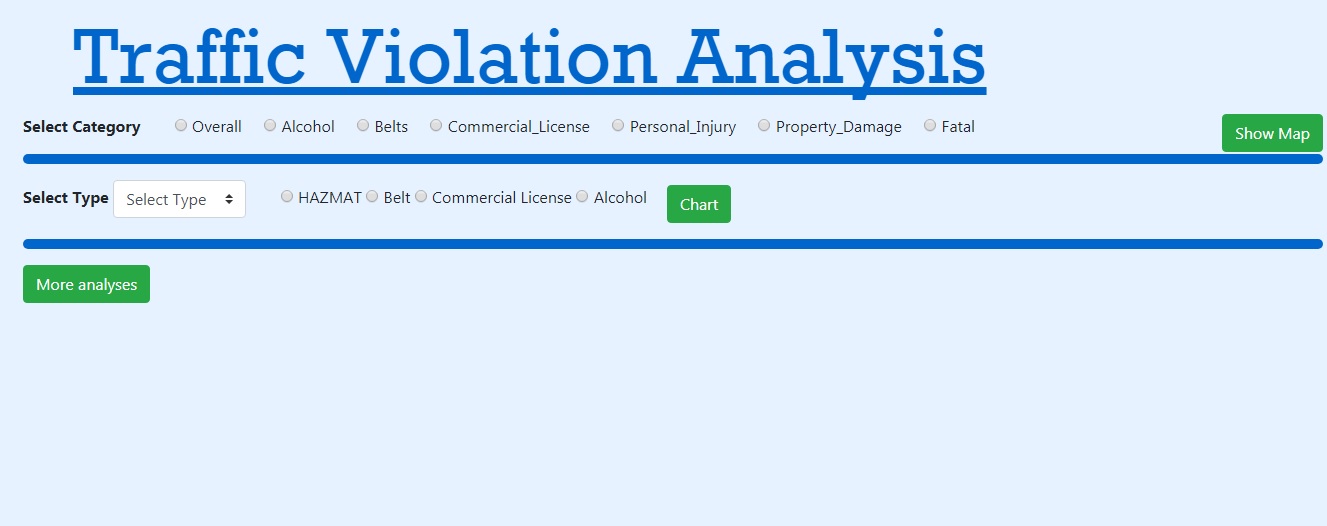


Figure 2: Traffic Analysis website

## **State wise Count**

In this visualization, we have displayed the analysis result in the US state map, where different color coding scheme is used in order to display the count of the ticket issued to the people. The darker the color, more number of tickets issued in that particular state. We have provided interactive form to select the analysis based on the different categories. Based on the selected categories, the map will display the count of tickets issues state wise. We have also provided the feature to get the actual count of the ticket by hovering on the state through mouse pointer. The categories are Overall, Alcohol, Belts, Commercial License, Personal Injury, Property Damage, and Fatal. This visualization is created used D3.js and HTML. Figure 3 shows the result.

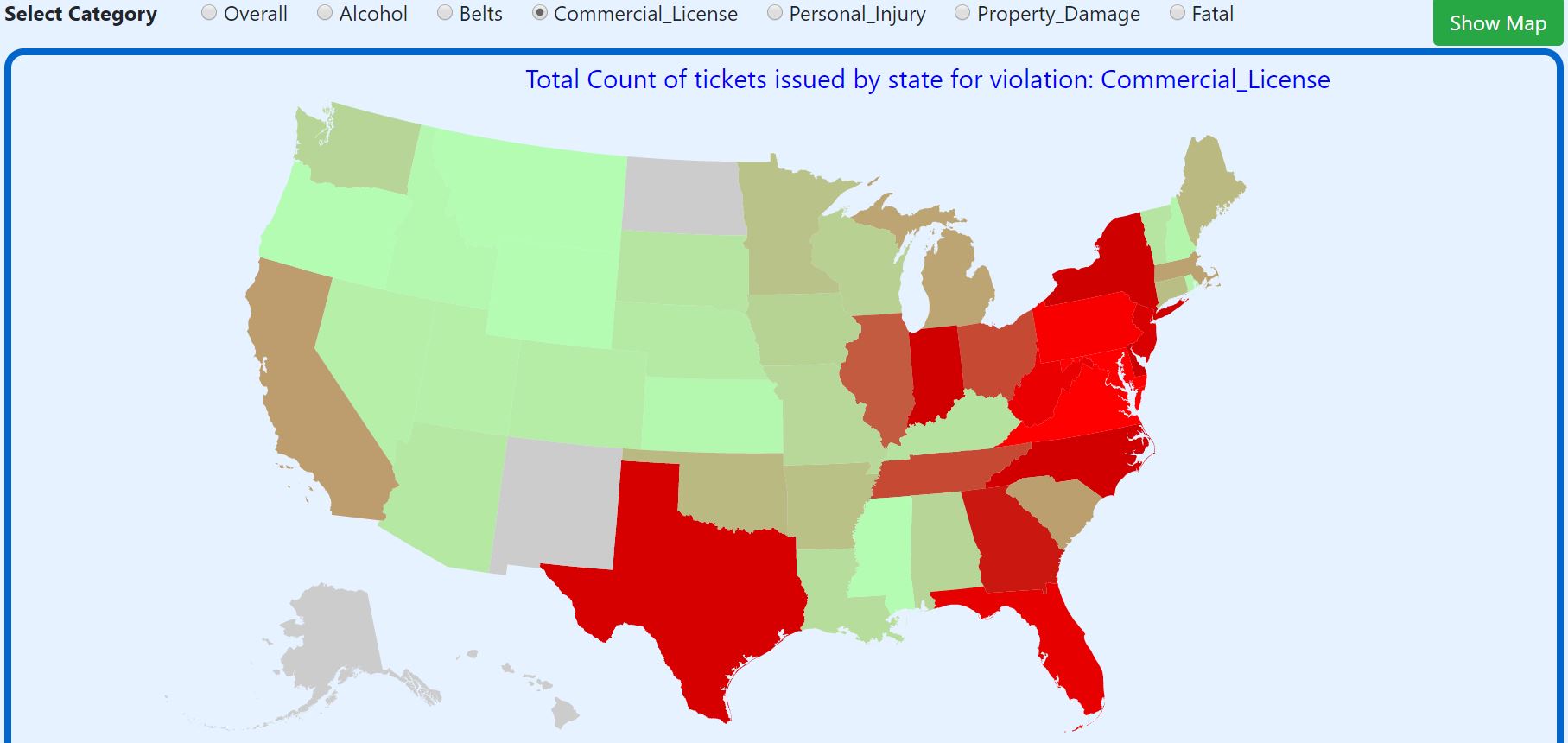


Figure 3: US State Map showing tickets count

## **Violation count every minute**

This is another D3.js visualization which represents the average number of tickets issued to the citizens per minute. This analysis is represented in form of ‘line Graph’ where we can clearly see the trend of tickets issued in the particular point of the day. We can analyze from this visualization that the ticket count is higher from 9.00 PM to 12.00 PM

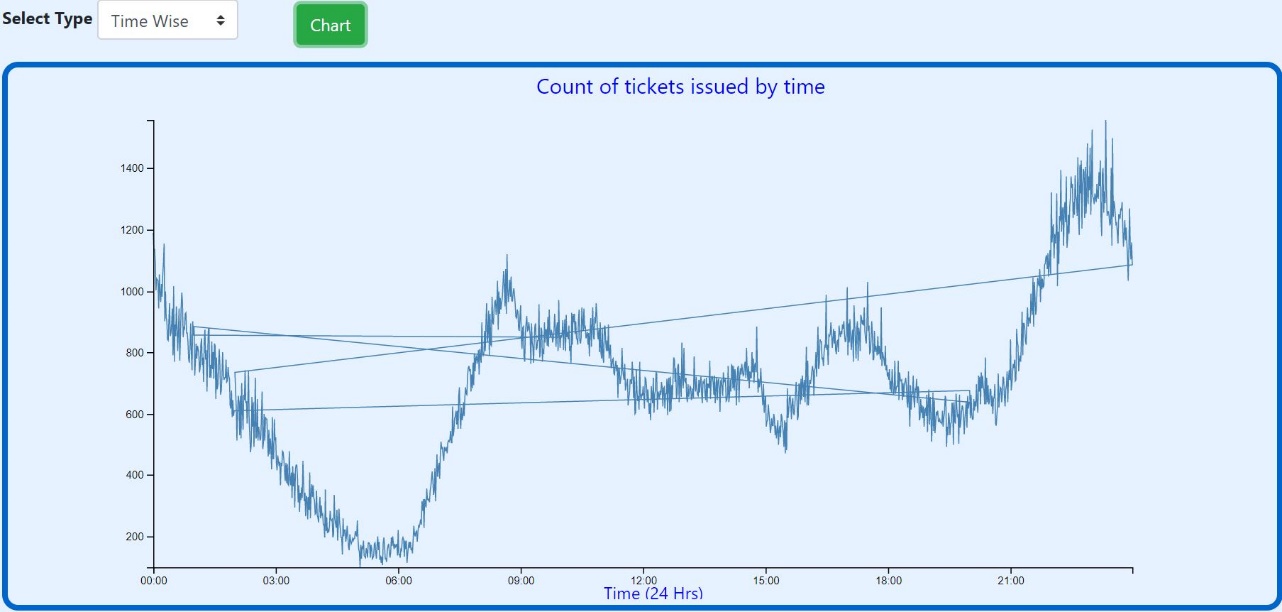


Figure 4: Violation count per minute

## **Gender wise Analysis**

In this analysis we have provided interactive pie chart analysis which displays the gender wise tickets issued based on the different conditions. Here, user will have option to select the condition and the pie chart based on that specific condition will be generated. We have provided the option of HAZMAT, Belt, Commercial License, and Alcohol Consumption.

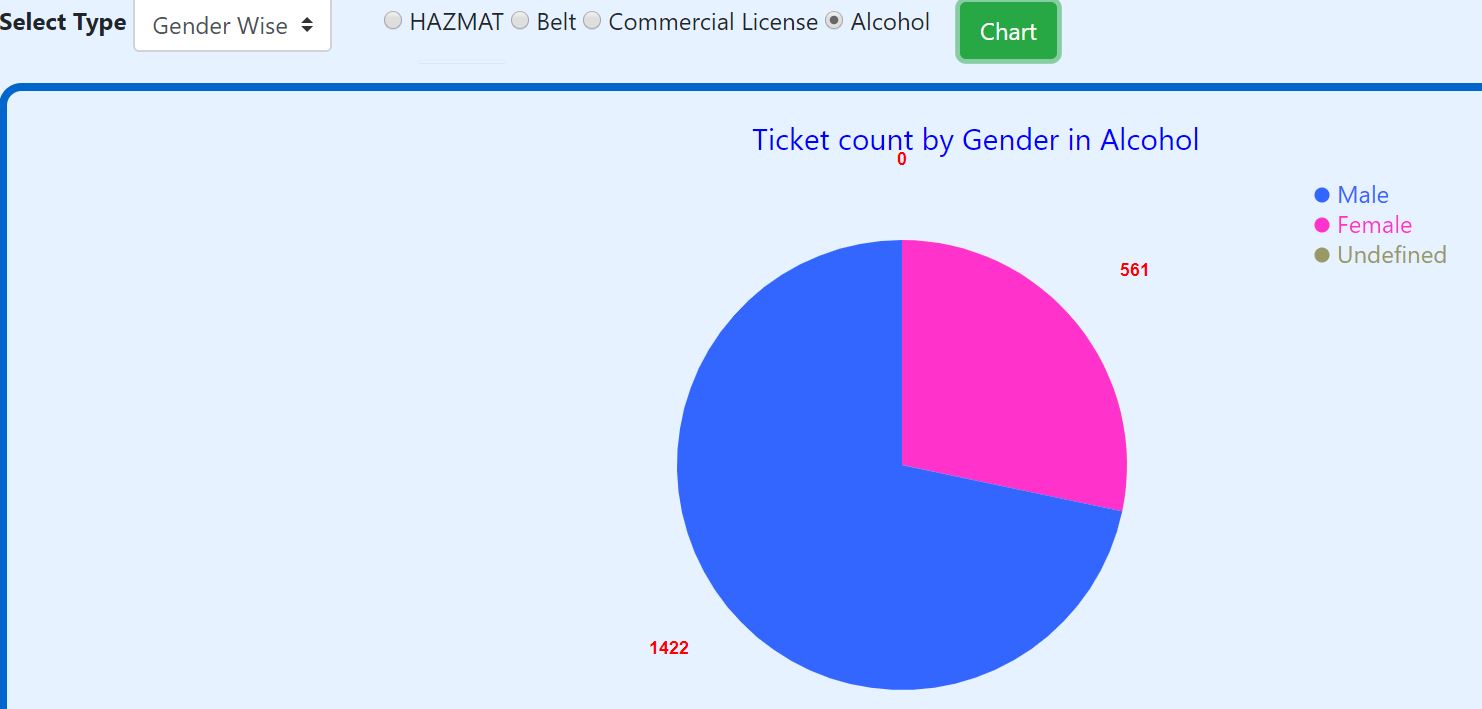


Figure 5: Gender wise tickets count

## **Top Reason of Violation**

In this analysis, we represents the top 10 reason of observed by for violations. It is represented by interactive bubble chart, where radius of the bubble and its color will changes as per the counts of the tickets issued for that particular reason. Higher the count, darker the color and bigger the radius of the bubble.

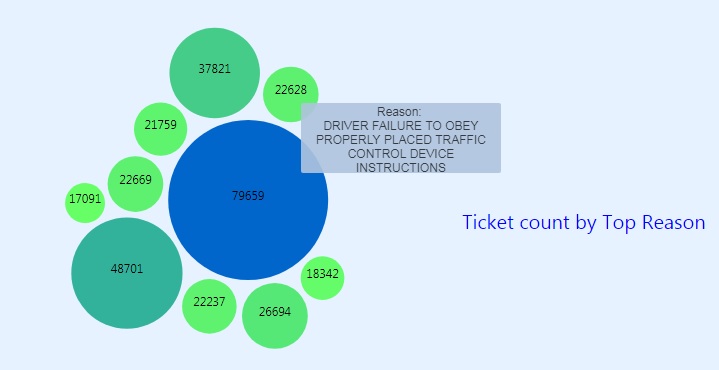


Figure 6: Top Reason of Violation

## **Analysis visualized on Amazon QuickSightTM**

|  |  |  |
| --- | --- | --- |
| **#** | **Analysis** | **Figure** |
| 1 | Count of arrest type | E:\Keyur\Masters\Sem 2\Big Data\Project\Output\Screenshots\graph2.JPG |
| 2 | Vehicle type Alcohol yes - count Alcohol No - count | E:\Keyur\Masters\Sem 2\Big Data\Project\Output\Screenshots\graph2.JPG |
| 3 | Vehicle color wise ticket count | E:\Keyur\Masters\Sem 2\Big Data\Project\Output\Screenshots\graph5.JPG |
| 4 | Race wise ticket count | E:\Keyur\Masters\Sem 2\Big Data\Project\Output\Screenshots\graph6.JPG |
| 5 | Hourly ticket count | E:\Keyur\Masters\Sem 2\Big Data\Project\Output\Screenshots\graph7.JPG |
| 6 | Vehicle type ticket count | E:\Keyur\Masters\Sem 2\Big Data\Project\Output\Screenshots\graph8.JPG |
| 7 | Violation type count | E:\Keyur\Masters\Sem 2\Big Data\Project\Output\Screenshots\graph9.JPG |
| 8 | Violation type (4 types) Gender | E:\Keyur\Masters\Sem 2\Big Data\Project\Output\Screenshots\graph10.JPG |
| 9 | Month-Year ticket count | .E:\Keyur\Masters\Sem 2\Big Data\Project\Output\Screenshots\graph11.JPG |

Table 2: QuickSight Analysis

## **Tabular Analysis**

After studying all the outputs, we have collated important results in a tabular format within Table 3.

|  |  |
| --- | --- |
| **Analyzing Parameters** | **Top Contributor** |
| State with most violators | Maryland |
| Top Reason for Violation | Drive’s failure to obey instructions |
| Vehicles type – Alcohol Violation | Automobile |
| Race wise violation | White |
| Violation Time Range | 9 p.m. – 12 a.m. |
| Violation type | Citation (tickets issued) |
| Month and Year of most violation  Table 3: Result Analysis | March 2013, May 2015 |

# **INDIVIDUAL CONTRIBUTIONS**

The table below highlights the individual contributions of each member throughout this project.

|  |  |
| --- | --- |
| **Team Member** | **Role** |
| Keyur Mehta | * Requirement Analysis * Data Cleansing * Backend Scripting * D3 Visualizations * Project Report |
| Love Modi | * Requirement Analysis * Frontend Design and Development * Backend Scripting * Webpage hosting * Project Report |
| Darsh Sanghavi | * Backend Scripting * Amazon QuickSight Visualization * Webpage Configurations * Requirement Tracking * Project Report |

Table 4: Member's Contribution

# **CONCLUSION**

We have developed a complete big data application from requirement analysis to data extraction followed by its visualization. We hope that this project can be utilized as one of the tools to improve the traffic conditions in the US. Through this project, we got an opportunity to learn various new technologies. We were able to understand and evaluate complex requirements in terms of map-reduce scripts. To visualize the extracted data, gave a real sense of the parameters that need to be focused upon. This project was a very good exposure to real life big data management scenarios.

# **REFERENCES**

|  |  |
| --- | --- |
| [1] | "Maryland Violation Data," [Online]. Available: https://catalog.data.gov/dataset/maryland-statewide-vehicle-crash-data-dictionary. |
| [2] | "Pig Latin Basics," [Online]. Available: http://pig.apache.org/docs/r0.17.0/basic.html. |
| [3] | "D3.js," [Online]. Available: https://d3js.org/. |
| [4] | "Learn to Code," [Online]. Available: https://learn.shayhowe.com/html-css/. |