UNIVERSITY OF CALIFORNIA, RIVERSIDE
Department of Computer Science and Engineering

Final Report

Project title: **Car Owning Quality Evaluation Using Big Data Analysis**
Course: **CS 226 Big Data Management**
Project Members: Yuanjie Zhu
Chunhan Zhang
Tanish Arora
Qicheng Hu

1. Introduction

Today, transportation is necessary in order to be able to live in most places in the U.S. However, on top of the price of an automobile, there are a lot of hidden costs that come with owning a vehicle. These costs and inconveniences vary from state to state, and an analysis of these costs per state would be good for people that need to decide whether they are suitable for owning a vehicle.
Our project's focus is to highlight the hidden costs that come with owning a vehicle. These costs and inconveniences vary from state to state, also depending upon their safety information such as fatality reports in different states.

2. Related Work

Based on our understanding of car purchases and normal expenditure considerations, we initially searched for information in related fields and inquired a lot of relevant information[1-10], whether it was car sales or car insurance, car safety and the like. For our feature selection, we have a general direction, and then we search for detailed data sets and gradually optimize the selected features to establish an evaluation system.
In the subsequent data set search and project implementation, based on the conclusions of our analysis from different papers in our Literature Survey, data will be extracted in a targeted manner, and finally a model of satisfaction with car use will be delivered.

Model: We apply equal weights to each feature we extracted from the dataset of costs and features of traffic safety. We believe that each cost should be an individual parameter, and the safety issues the same. After that, even with the same weight of each feature, the total weights for each individual feature are different. Therefore, we multiplied the ratio towards each total weight, and then added them up to the final results. Details for each part of features, and how we handle different datasets with different models is given in Section 3, and the feature extraction details in Section 4.

Tools: Pyspark, Basemap, NumPy

Milestones:
Our plan goes quite well with our progress, except at week 8 where we extended a week to verify the results and visualize the features.
The following table contains the recorded milestones of our project.

Step Numbers	Milestones	Time Frame
1.	Data collection and preprocessing	Week 4 - 5
2.	Features weight decision and evaluation algorithm creation	Week 6
3.	Data analysis	Week 6 - 7
4.	Quality evaluation model formation and visualization	Week 7 - 9
5.	Final report	Week 10 - 11

3. Data Processing

3.1 Cost of Owning a Car

The analysis of Edmunds Inc., named True Cost to Own a Car (TCO), shows the components to consider when evaluating the true value of a vehicle. TCO calculates the factor of depreciation, interest of financing, taxes, insurances, fuel, maintenance, repairs and etc.
We divide these features into the following main components: financing of car, insurance of car and operation cost of car. Financing of a car includes the original price, taxes and interest of financing if applicable; insurance of car includes the cost of car insurance counted by year; the operation cost of car includes the fuel, maintenance and inspection, which are related to distance driven.

3.2 Traffic Safety Facts and Car Accidents

The data sources we collected related to this topic are: State Traffic Safety Information (STSI), Traffic Safety Facts Annual Report Tables, NHTSA's Fatality Analysis Reporting System (FARS), FastFARS (FF), and Monthly Fatality Counts[11]. Open history datasets from the United States Department of Transportation. They show the data for traffic safety such as fatality, injuries with different types of vehicles, fatal accidents from state to state as well as different driving factors related accidents and fatalities.
After consideration for data integrity and data time rigor, we found from datasets and defined these crucial features: fatalities count for last 10 years in each state, fatality rate per 100 million VMT(vehicle miles traveled), the proportion of fatal car accidents in each state to the population of that state, statistics on the number of injuries and deaths in car accidents based on

different types of cars, car accidents ranking for states related to human error, car accidents ranking for states related to atmosphere reasons.

4. Feature Extraction

4.1 Cost of Owning a Car

As it indicates in section 3.1, the expenditure of owning a car can be divided into several components. Here, for the convenience of calculating cost in common scenarios, we exclude the original price of car, which varies due to variant makes and models, from the financing of car while keeping the rest factors. The two other main features taken into consideration are: insurance cost and operation cost. The data of insurance cost of cars categorized by states and the average cost of fuel, maintenance are available on different data sources, and can be integrated.

4.2 Traffic Safety Facts and Car Accidents

As it's shown in section 3.2, data related to automobile safety can be divided into several features. For the convenience of analysis, we use Apache Spark to read related datasets mentioned in section 3.2. Then extract those crucial features based on a given time period range from 2010 to 2018 of the original datasets and do data preprocessing to calculate the average or total number of accidents and fatalities with different concerns.

4.3 Features Integration

As above two sections mentioned, we found features ranging from lots of concerns and aspects. In order to get a better understanding and comprehension of the ranking for state. We define four colors utilized in each feature. The principle is that based on the ranking percentage of a feature of the state compared to other states with the same feature, we color such feature in Blue or Green or Orange or Red. Each color has its own weight, then calculate the ultimate color for the combination of features with different colors and weight considered.

4.4 Summary Graph Creation

As the above section mentioned, after determining the ranking of each state and giving them a color based upon their ranking and other factors. It's important to be able to visualize the given ranks based upon colors given to each state and not just output into a simple text file. Being able to visualize the ranks based on their colors is more user friendly. Here the different colors and states are taken from the resulting file after feature integration and combined with the basemap library with the US state structure files to output the states. However, the basemap library for some reason still does not support Michigan and Wisconsin as states.

5. Experiments & Results

5.1 Cost of Owning a Car Features

The average cost of licenses, taxes and registration fees of having a car is 738 dollars per year

across the United States and the cost of finance interest is 683 dollars per year. Therefore, the first component of owning a car is approximately 1466 dollars per year on average. The cost of insurance varies from state to state. The average fuel cost of a car is averagely 10.26 U.S. cents per mile and the maintenance cost is 7.91 cents per mile as the data from Statia shows. The overall yearly cost of operation for a car is 2725.5 dollars assuming that the car is driven for 15000 miles per year.

The overall cost of owning a car is computed as the aggregation of the above features. The most expensive states to own a car are (by order of descending): Michigan (18394 dollars per year), Rhode Island, Nevada, Florida, New Jersey, California, Mississippi, Minnesota, Tennessee(13791 dollars per year). The least expensive states to own a car are (by order of descending): South Dakota (10971 dollars per year), North Carolina, Georgia, Alabama, Wyoming, Hawaii, Colorado, Delaware, Oregon, Montana, New Hampshire, Alaska (5185 dollars per year).

5.2 Traffic Safety Facts and Car Accidents Features and Integration

Based on each state, we output its colors which are the ranking for different features to be considered. We give color Blue with weight four, Green with weight three, Orange with weight two and Red with weight one. Each color represents a different percentage of such features compared to other states. Blue(0~25%), Green(25~50%), Orange(50~75%), Red(75~100%). We can see that for each feature, less percentage/higher weight which means less car accidents probability and fatalities.

With all such features calculated, we compute the ultimate color for each state with evenly consideration, which means each feature counts the same importance as others. In such way, we get ultimate color for each state with different type of vehicles considered shown in the following tables(only part of states shown in tables):

For QOL(quality of life) of owning a vehicle, Blue better than Green, Green better than Orange, Orange better than Red.(The full results can be reachable under deliverables folder/final_color/)

Alabama	Orange	2.57142857
Alaska	Green	3
Arizona	Orange	2.28571429
Arkansas	Orange	2.14285714
California	Orange	2.14285714
Colorado	Green	3
Connecticut	Orange	2.85714286
Delaware	Green	3.14285714
Florida	Red	1.42857143

Georgia	Orange	2.57142857
Hawaii	Green	3.28571429
Idaho	Orange	2.57142857
Illinois	Orange	3
Indiana	Orange	2.42857143
Iowa	Green	3
Kansas	Orange	2.42857143
Kentucky	Red	1.57142857
Louisiana	Orange	2.14285714
Maine	Green	3
Maryland	Orange	2.85714286

Table 1: States Ultimate Color For Passenger Cars

Alabama	Orange	2.857142857
Alaska	Green	3.285714286
Arizona	Orange	2.571428571
Arkansas	Orange	2.428571429
California	Orange	2.428571429
Colorado	Green	3.285714286
Connecticut	Green	3.142857143
Delaware	Green	3.428571429
Florida	Red	1.714285714
Georgia	Orange	2.857142857
Hawaii	Green	3.571428571
Idaho	Orange	2.857142857
Illinois	Green	3.285714286
Indiana	Orange	2.714285714
Iowa	Green	3.285714286
Kansas	Orange	2.714285714
Kentucky	Red	1.857142857
Louisiana	Orange	2.428571429

Maine	Green	3.285714286
Maryland	Green	3.142857143

Table 2: States Ultimate Color For Light Trucks

Alabama	Green	3.142857143
Alaska	Green	3.571428571
Arizona	Orange	2.857142857
Arkansas	Orange	2.714285714
California	Orange	2.714285714
Colorado	Green	3.571428571
Connecticut	Green	3.428571429
Delaware	Green	3.714285714
Florida	Red	2
Georgia	Green	3.142857143
Hawaii	Green	3.857142857
Idaho	Green	3.142857143
Illinois	Green	3.571428571
Indiana	Orange	3
Iowa	Green	3.571428571
Kansas	Green	3
Kentucky	Orange	2.142857143
Louisiana	Orange	2.714285714
Maine	Green	3.571428571
Maryland	Green	3.428571429

Table 3: States Ultimate Color For Buses

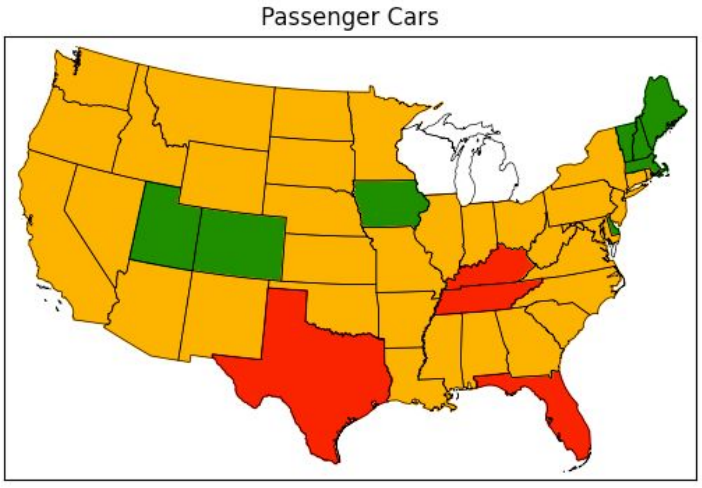
Alabama	Green	3.142857143
Alaska	Green	3.571428571
Arizona	Orange	2.857142857
Arkansas	Orange	2.714285714
California	Orange	2.714285714

Colorado	Green	3.571428571
Connecticut	Green	3.428571429
Delaware	Green	3.714285714
Florida	Red	2
Georgia	Green	3.142857143
Hawaii	Green	3.857142857
Idaho	Green	3.142857143
Illinois	Green	3.571428571
Indiana	Orange	3
Iowa	Green	3.571428571
Kansas	Green	3
Kentucky	Orange	2.142857143
Louisiana	Orange	2.714285714
Maine	Green	3.571428571
Maryland	Green	3.428571429

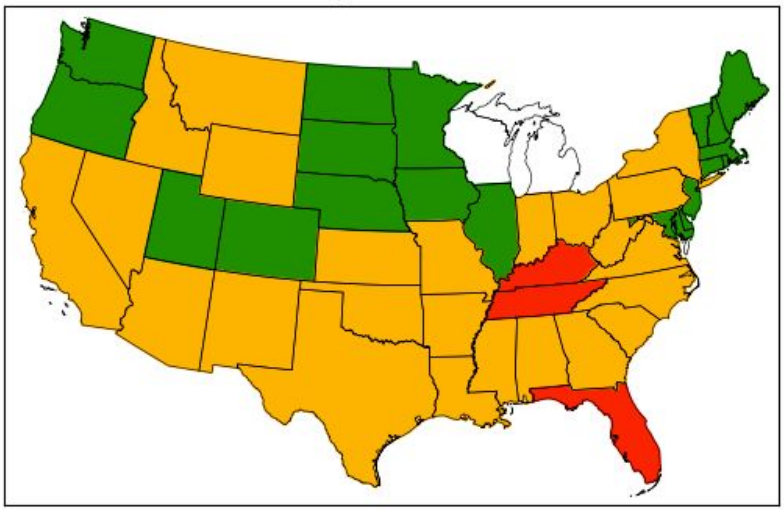
Table 4: States Ultimate Color For Large Trucks

Also, the output colored US maps as follows:

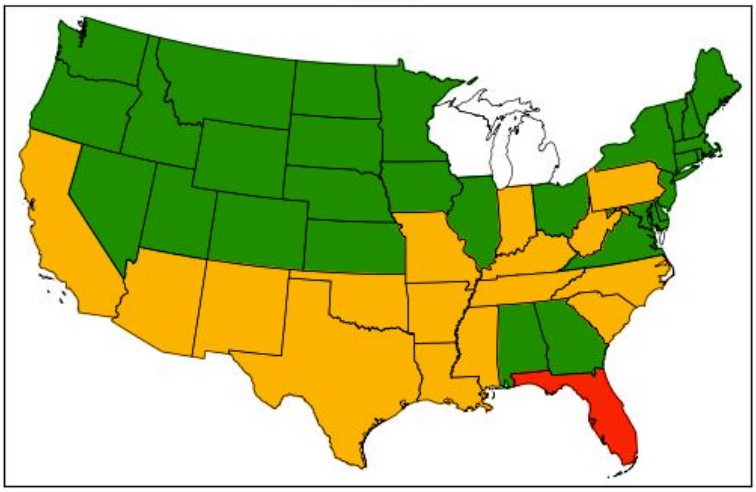
Quality Level: Blue: Excellent, Green: Very Good, Orange: Good, Red: Fair

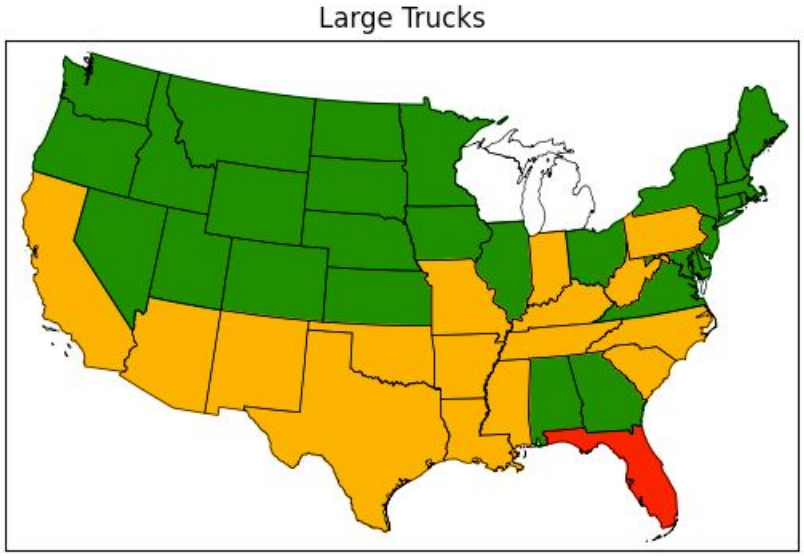


Light Trucks



Buses





5.4 Feature Extraction Output Validation

Because what we are discussing is based on a variety of conditions, consider the rating of owning a vehicle from different perspectives. From the cost of maintaining car use, to the risk factor based on the frequency of car accidents in each state's individual situation. Therefore, it is inappropriate to directly compare with the existing conclusions on the Internet, because the angle and weight of consideration are not consistent, and the years of consideration are not the same. Therefore, based on the results, we hope to provide a different degree of reference for people to consider owning cars. Combined with other ranking data[12, 13], it can be more clear that a certain state ranks under various consideration angles.

6. Conclusion & Future Work

The main objective of our project was to develop a model that evaluates the quality of life of owning a car in different states all around the country based upon different features and factors. In the project we use tools such as NumPy to work with the non trivial ML algorithm part of the project and Apache Spark, pyspark as our Big data classifiers. We gather our data from various different government websites showing data for traffic collisions, fatalities from state to state as well different human or atmospheric factors which lead to fatal crashes. After gathering all the data from the websites, the data is then passed through a model to determine the ranking of the state and then assign them a different color based upon the quality of owning a vehicle. We then use tools like Basemap to plot out those colors onto the map and output it out to the user. Our future work would be to gather more information about purchasing a car from owners to have a much better idea and to keep updating our data so that our model can be used for interpreting the quality of owning a car respective to each state.

7. Reference

- [1] Luiz Moutinho, Fiona Davies, Bruce Curry, The impact of gender on car buyer satisfaction and loyalty: A neural network analysis, Journal of Retailing and Consumer Services, https://doi.org/10.1016/0969-6989(95)00064-X.
- [2] Annarita Roscino, Alessio Pollice, A statistical analysis of the customer satisfaction with car dealers, https://doi.org/10.1002/asmb.520
- [3] Myers, James H, Measuring Customer Satisfaction: Is Meeting Expectations Enough?
- [4] Dick Ettema, Tommy Gärling, Lars E. Olsson, Margareta Friman, Sjef Moerdijk, The road to happiness: Measuring Dutch car drivers' satisfaction with travel, Transport Policy, https://doi.org/10.1016/j.tranpol.2012.12.006.
- [5] Fuchs K., Abendroth B., Bruder R. (2009) Night Vision - Reduced Driver Distraction, Improved Safety and Satisfaction. In: Harris D. (eds) Engineering Psychology and Cognitive Ergonomics. EPCE 2009. Lecture Notes in Computer Science, vol 5639. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-02728-4_39
- [6] Wardhana A.P., Ishibashi K., Kiyota M., Consideration of Road Management from the ViewPoints of Long- and Short-Distance Road User's Satisfaction, DOI: https://doi.org/10.9744/ced.13.2.90-97
- [7] Taebeum Ryu, Kyunghee Oh, Heecheon You, Myung Hwan Yun, Development of Satisfaction Models for Passenger Car Interior Materials considering Statistical and Engineering Aspects of Design Variables, https://doi.org/10.1177/154193120304700507
- [8] Sajjad Shokouhyar, Sina Shokoohyar, Sepehr Safari, Research on the influence of after-sales service quality factors on customer satisfaction, Journal of Retailing and Consumer Services, https://doi.org/10.1016/j.jretconser.2020.102139.
- [9] Sungha Jang, Ashutosh Prasad, Brian T. Ratchford, Consumer Search of Multiple Information Sources and its Impact on Consumer Price Satisfaction, Journal of Interactive Marketing, https://doi.org/10.1016/j.intmar.2017.06.004.
- [10] S. A. Basheir and A. Z. Karrar, "The Impact of Service Quality on Customer Satisfaction in Location Based Service : (Mishwar and Tirhal as Case Studies)," 2019 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE), Khartoum, Sudan, 2019, pp. 1-6. doi: 10.1109/ICCCEEE46830.2019.9071210
- [11] https://cdan.nhtsa.gov/
- [12] States With the Worst Drivers – 2019 Edition https://smartasset.com/checking-account/states-worst-drivers-2019
- [13] The Best and Worst Drivers by State https://quotewizard.com/news/posts/the-best-and-worst-drivers-by-state