NEVaulate

Zhi Zhang (18210054)
Han Lin (18206786)

A report submitted in part fulfilment of COMP 30780

Data Science in Practice

Demonstrator: Rian Dolphin

UCD School of Computer Science
University College Dublin

May 12, 2022

Table of Contents

Abstract

The future direction of global automotive development is new energy, which has become the consensus of global countries and automobile enterprises. People intend to buy the new energy vehicles more than ever, such as fully electric, hybrid, hydrogen energy vehicles and so on. There has been considerable interest in the rapid spread of new energy vehicles, and the comparison between new energy and traditional vehicles especially in some aspects of mobility and safety. The media, car commentators and automotive practitioners have come up with their insights and suggestions. In this project, we will look at whether new energy vehicles will dominate the future and how they differ from traditional vehicles in terms of mobility, performance, and safety from a data science perspective. We present the findings of large-scale data analysis to assist in answering some of the most prevalent concerns, such as how to find the safest vehicle in both new energy and traditional vehicle? We achieve this by combining data from more than 0.8 million real vehicle information to investigate from diverse perspectives.

Chapter 1: **Introduction**

At present, we're in the middle of the automotive industry going electric. While most of the new energy vehicle market is mainly driven by supply and policy, the fluctuating new energy vehicle market over the past few years has gradually taken on a demand-oriented character as 2022 approaches.

Nowadays, consumers are more interested in sustainable, environmentally friendly lifestyles. As well as changing consumer attitudes, regulators are also changing their stance. China, for example, has pledged to be carbon neutral by 2060[1], and the electrification of its vehicles will help achieve that goal. Such changes in environmental policy can be seen in other parts of the world. In Europe, countries have already set a ban date for the sale of gas-powered cars[2]. In the United States, the government is expected to set environmental priorities. The federal government will buy new energy vehicles and deploy half a million new public charging stations across the country [3]. Under the parallel development of various automobile manufacturers and countries, the new energy vehicle market began to burst out great vitality.

Our interest was first led by a series of news reports about the recent surge in gasoline prices[4], and the electric vehicle sales are also increasing year on year. We are wondering if new energy vehicles will overtake traditional gas-powered cars to dominate the future and what advantages they have over traditional vehicles.

In the next chapter I discuss why new energy vehicle is suitable as an interesting data science topic, and outline the key research questions to be explored. This is followed by a brief summary of related work that has been conducted on NEV. Chapter 4 summarises the datasets used in this study, including data we retrieved from existing dataset and a self-collected dataset of vehicle safety from *National Highway Traffic Safety Administration* API interface. In Chapter 5 we describe how we have answered each of the key research questions and present our findings in detail. Chapter 6 concludes with a summary of the key findings, a discussion of the limitations of the work as presented, and several possibilities for future work.

Chapter 2: **Project Objectives**

New energy vehicles are vehicles that use non-fossil energy as a source of power, integrate advanced technologies in vehicle power control and drive system. The rise of new energy vehicles has been helped by rising gasoline prices and people's awareness of sustainable development. A growing number of people are paying attention to this field in recent years.

For instance, Figure 2.1 presents one version of the the level of interest in NEV, based on Google searches. It shows a gradual increase to the peak since 2020 in NEV interest.

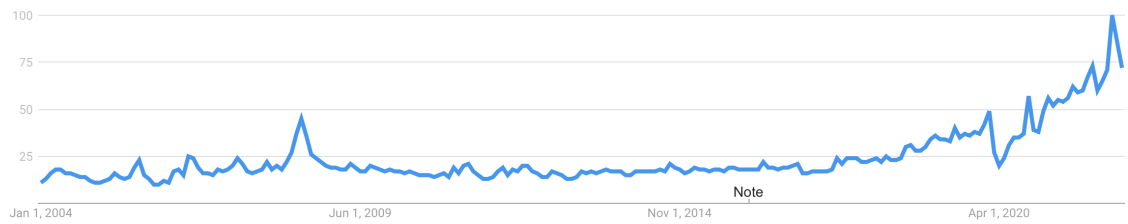

Figure 2.1: Google Trends shows the rapid rise in NEV popularity at the April 2020

New energy vehicles not only mean pure electric vehicles, they also refer to plug-in electric vehicles (PHEV), fuel cell vehicles (FCEV) and hydrogen engine vehicles(HEV), etc. We will focus on all of these types in our research.

This project could help drivers who want to switch from gasoline to new energy vehicles or even for people who are willing to drive in the future.

2.1 Objectives & Motivations

A suitable data science topic generally comes down to at least two key points: (1) Is there a interesting set of research questions to ask and (2) Is there enough data to answer them?

The new energy vehicle project can be studied from the following aspects: social and environmental factors, as well as the performance and safety of vehicles are the most concerned topics. In addition, we can find data on the official government website to ensure there is enough data and its authenticity and accuracy.

At the end of the study, we hope to figure out what influences the development of new energy vehicles, how they develop and what their advantages are compared to traditional vehicles. And it should be possible to provide some useful suggestions for drivers.

2.2 Research Questions

For this project then, we will use two sources of data: Obtain dataset from portals & collection. Use existing dataset by the United States official government website and vehicle safety data on *National Highway Traffic Safety Administration* [5] through API. Both of these datasets are described in detail in Chapter 4 and we will use them to explore the following research questions.

RQ1: How social environment influences the development of NEV in the United States?

We will analyse how it affects the sales of new energy vehicles in the United States from four aspects:

- GDP
- Income per capita
- Tax policy
- Climate

Analyze the correlation between them to find out which factor has the most influence on the NEV developments.

RQ2: What are the advantages of new energy vehicles compared to traditional vehicles?

We will find out new energy vehicles' advantages by compared with traditional vehicles from a few different perspectives:

- Compare of number/distribution of energy supply stations
- Performance of vehicles
  - Quality-price ratio (engine size vs. price)
  - Gas emission
  - Value retention

RQ3: How to find the safest vehicle in both new energy and traditional vehicle?

Analyze the safety of new energy vehicles from the following three aspects:

- Complaints – Judge on the number of broken components by different fuel types and whether they were caused by crashes.
- Recalls – Judge by the date between the release of the recall and the model year and the recall frequency.
- Ratings – Judge by the collision rating results for different makers and models.

Chapter 3: Related Work

In fact, the field of new energy vehicles has a long history, even as it has become popular in recent years. In the 20th century, electric vehicles with high cost, low speed and short battery range, have been in a negative position in the market compared with internal combustion engine cars. Since 2010, global sales of all-electric automobiles and utility vans have surpassed 1 million units delivered, with 4.8 million electric cars in operation by the end of 2019.[6]

There have been some interesting discussions and analyse about electric vehicle market and future forecasts, such as how to predict the price of electric cars. For example, A predictive model that provides an effective decision-making reference for similar product markets.[7]

In addition, there are also studies on people's willingness to buy electric cars. They utilize data from a national poll in which participants were asked to pick their preferred gasoline vehicle and choose two electric versions of their favourite gasoline car. The data were then used to estimate the willingness to pay for five electric car qualities using a latent class random utility model.[8]

Instead of electric vehicles, we extend the scope to all new energy vehicles in our project. There are many factors affecting the development of new energy vehicles. We want to find out the most influential factor in the social environment. Based on the existing research results, we know that people may consider more the performance of new energy vehicles, the use of price and safety. We will study these aspects in more depth in our project.

Chapter 4: Data Considerations

4.1 Data Collection

Most of our datasets are from official government websites which can ensure that the data is real and accurate. For this study we will use two different sets of data:

- Obtained datasets from portals & collection.
- Write crawlers to fetch data from the API interface.

In this chapter we describe how both of these datasets were collected and we summarise the methods utilized to analyze them in order to answer the research questions in the next chapter.

4.1.1 Portals Data

For Research Question 1:

- GDP & Income datasets collected from '*Bureau of Economic Analysis*'[9]
- Population datasets collected from '*Bureau of Economic Analysis*'[9]
- Climate datasets collected from '*National Oceanic And Atmospheric Administration*' [10].
- Tax policy datasets collected from '*U.S. Department of Energy*' [11].
- EV sales datasets collected from '*GitHub existing datasets*' [12]
- GeoJSON data for the United States from '*eric*'[13]

For Research Question 2:

- Alternative Fueling Station Counts by State datasets collected from '*U.S. Department of Energy*' [14]
- Gasoline Station Counts by State datasets collected from '*U.S. Department of Energy*' [15]
- Vehicle dataset contains price and performance collected from '*DVM-CAR*' [16]
- Vehicle dataset contains annual fuel cost collected from '*U.S. Department of Energy*' [17]

4.1.2 API Data

For Research Question 3:

Write crawlers to fetch the data from: *National Highway Traffic Safety Administration*.[5] For our research question 3, we will focus on these aspects:

1. Complaints - The complaint database is used to identify safety issues that warrant investigation and to determine if a safety-related defect trend exists.

2. Recalls - The manufacturer recalls a product or piece of original equipment if it finds a safety defect or does not meet federal safety standards.

   There are four steps to fetch all vehicle complaints / recalls data.

   - Step 1: Get all Model Years Request a list of available Model Years for a given product type (Vehicle). Response is a list of available Model Years in the repository. Use a Model Year from this list to use in the next step.

   - Step 2: Get all Makes for the Model Year Request a list of vehicle Makes by providing a specific vehicle Model Year. Response is a list of vehicle Makes for that Model Year. Use the Model Year and a Make from this list to use in the next step.

   - Step 3: Get all Models for the Make and Model Year Request a list of vehicle Models by providing the vehicle Model Year and Make. Response is a list of vehicle model for that make and each model Year. Use the Model Year, make and a model from this list to use in the next step.

   - Step 4: Get all complaints / recalls for the selected Model Year, Make, Model Now that you have all three key pieces of information (Model Year, Make, Model), make the request to get the complaints for the required combination. Response is a list of complaints / recalls for the given Model Year, Make and Model.

3. Ratings - NHTSA's New Car Assessment Program rates vehicles to determine crashworthiness and rollover safety.

   There are five steps to fetch all vehicle ratings data.

   - Step 1: Get all Model Years.

   - Step 2: Get all Makes for the Model Year.

   - Step 3: Get all Model for the Make and Model Year.

   - Step 4: Get available vehicle variants for a selected Model Year, Make and Model. Make the request to get the crash tested vehicle variants for the Model Year, Maker and Model combination.

   - Step 5: Get the Safety Ratings for the selected variant. Now that we have the VehicleId of the precise model which we want look up, make the request to get the Safety Ratings for the required vehicle variant by passing in the VehicleId.

In total, we obtained the following amount of data from *NHTSA* API.

API Data List		
Data Name	Json File Number	DataFrame Size
Complaints	6,540	750,742 rows $\times$ 12 columns
Recalls	32,146	63,599 rows $\times$ 14 columns
Rating	7,314	7,314 rows $\times$ 34 columns

4.2 Data Preparation

The data cleaning process consists of the following steps:

- Load dataset into jupyter notebook.
- Check that all state names are correct and that all data sets are in the same format.
- Handel with null values.
- Convert date data into time series.
- Conversion of strings to numeric types of data
- Aggravating multiple files into one data frame if it is needed.

For example:

In tax policy dataset, state names are abbreviations, and them need to be converted to full names for manipulation with other datasets

State	Type
NV	Utility/Private Incentives
NV	Utility/Private Incentives
WI	State Incentives

Replacing abbreviations with full name:

State	Type
Nevada	Utility/Private Incentives
Nevada	Utility/Private Incentives
Wisconsin	State Incentives

Grouped according to the name of the state and the type of policy:

	Laws and Regulations	State Incentives	Utility/Private Incentives	Total
State				
Alabama	3	2	3	8.0
Alaska	2	0	4	6.0
Arizona	9	5	10	24.0
Arkansas	4	1	1	6.0
California	47	40	46	133.0

For Combining multiple datasets to a master dataset:

- Each month's temperature data for each state is a separate data set. Annual average temperatures are calculated and then combined with data from other states.
- Income, GDP, number of Stations data should be per capita. Therefore they are all combined with demographic data to form new data.

Chapter 5: **Results**

In this chapter we present the outcomes of our analysis for each research question. The outcomes of our study for each research topic are presented in this chapter. Each case includes a description of the datasets utilized and the approach taken, as well as a discussion of the findings and their implications.

5.1 RQ1: How social environment influences the development of NEV in the United States?

In this research question, we will focus on the relationship between GDP, income, tax policy, climate and the sales of new energy vehicles in each state of the United States. Also, find the most influential social environment for NEV development.

5.1.1 Data & Method

In this research question, three areas will be analyzed: economic, climate and tax policy.

For economic analysis: data on income per capita, GDP, and sales per capita were needed. Before starting this research question, we had the data for each state. In this research question they are aggregated according to state name, as below(first 5 rows of dataframe):

	income	GDP	EV sale
Alabama	43157	40648.99	0.000201
Alaska	61639	72584.15	0.000195
Arizona	47242	45809.18	0.001074
Arkansas	45037	38875.25	0.000179
California	64333	69266.30	0.003667

GDP, income and sales per capita are compared after being put together. GDP and income are then compared separately with the effect of sales per capita

For climate analysis: data on annual temperature and sales per capita were needed. The available cleaned dataset contains annual average temperatures by state from 1990-2021, and sales per capita from 2016-2019. The year 2019 was selected for analysis. Regions were divided according to temperature and, to observe the distribution of EV sales as below(first 5 rows of dataframe):

	temp	EV sale
Massachusetts	48.5	0.001120
North Carolina	61.3	0.000520
Arkansas	61.2	0.000179
North Dakota	38.0	0.000147
Kansas	53.8	0.000289

For tax policy analysis: we used the total number of tax policies related to new energy vehicles in each state and combined them with the average EV sales data from 2016-2019 and US geographic regions and party information into a new data frame, first 5 rows as follows:

	Tax Policy Number	EV sale	Geo Region	Party
Alabama	8.0	0.000130	South	Republican
Alaska	6.0	0.000161	West	Republican
Arizona	24.0	0.000714	West	Democratic
Arkansas	6.0	0.000108	South	Republican
California	133.0	0.002959	West	Democratic

5.1.2 Results

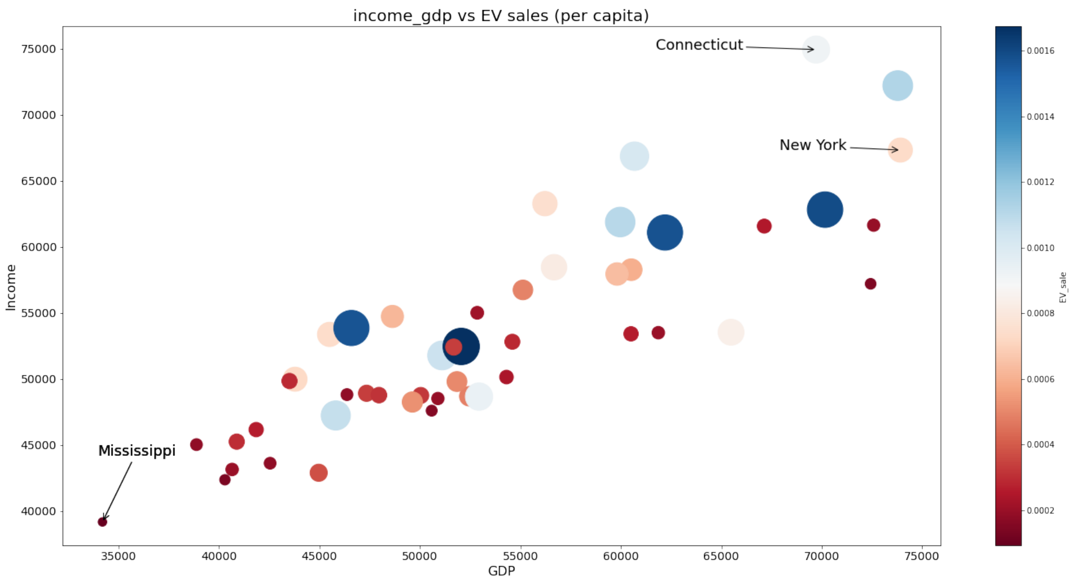

Figure 5.1: The impact of GDP and income on the NEV sales (per capita).

We start with an overall comparison of GDP, income and sales per capita. A scatter plot is created using GDP and income as axes. Each scatter point corresponds to a state and is given a size based on the volume of EV sales. To make the results more visible, a colour bar was added, and after removing data that had a particularly large impact on the results (e.g. California's sales far exceeded those of all the states) this graph was obtained. The graph shows that GDP and income are proportional, and that EV sales change from red at the beginning to blue as GDP and income increase.

To verify the idea, we conducted a search on the internet. Based on the data looking at the top 5 countries with the highest share of electric vehicles versus the bottom 5 countries, it is true that higher GDP corresponds to a higher market share for electric vehicles. [18]

The states with the highest as well as the lowest GDP and income are labeled. The interesting point is that the lowest GDP and income states do correspond to very low sales. But the highest income states correspond to higher sales volumes than the highest GDP states. Does this mean that income has a greater impact on EV sales than GDP?

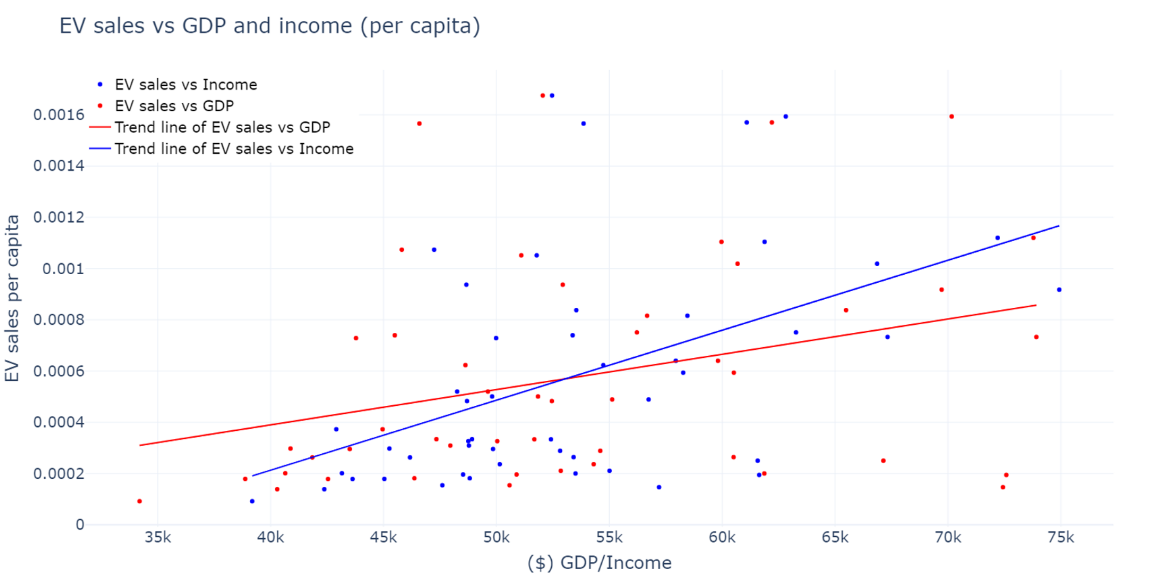

Figure 5.2: The impact of GDP and income on the NEV sales (per capita).

Figure 5.2 presents an analysis of whether income has a greater impact on sales than GDP. Using the values of GDP and income as the x-axis (in US dollars) and EV sales per capita as the y-axis. Separate colours are used to indicate GDP and income, and a linear regression model is built based on the distribution of points. According to the model, the slope corresponding to income is indeed greater than that of GDP, which validates the idea that income has a greater impact.

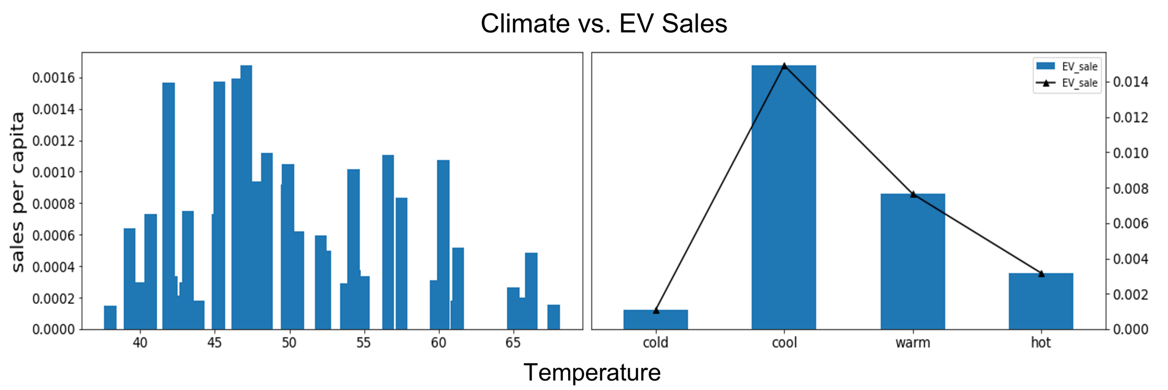

Figure 5.3: The relationship between climate and NEV sales.

We note that temperature can have an impact on the performance of electric vehicles: *'At –15 C, EVs drop to 54 per cent of their rated range, meaning a car that rated for 402 km will only get on average 217 km'* [19]

The left part of the graph in Figure 5.3 is made up of the annual average temperature (the sum of the average temperatures for each month divided by 12) on the y-axis and EV sales on the y-axis, with each bar representing a state, divided into four regions based on temperature to make the graph on the right. (e.g. annual temperature lower than 40 °F consider as cold place)

The graph shows that the majority of EV sales are in areas where the weather is suitable. When the climate is too cold or hot, EV sales drop significantly, to the detriment of EVs.

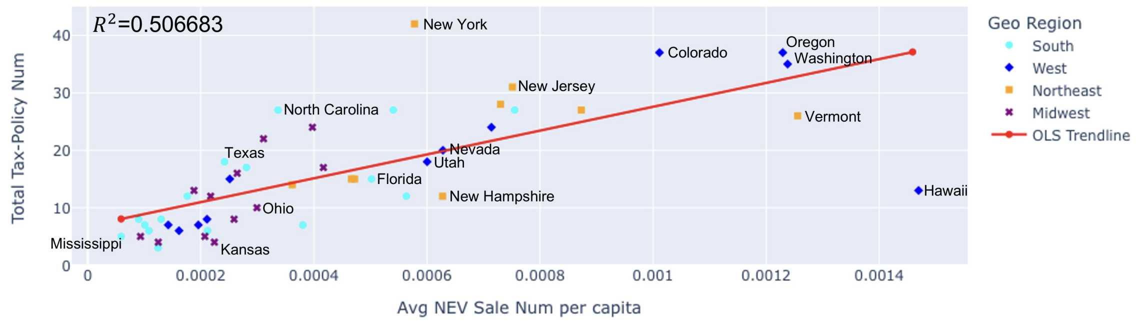

Figure 5.4: The relationship between tax policy and NEV sales.

We also build a linear regression model for the tax policy which refer to the new energy vehicle. It's clearly show that most southern and mid-western states are in the lower left corner, which means they have fewer tax policies and lower car sales records. Some western states, like Washington, are in the upper right corner, but the highest tax policies don't lead to the highest car sales. Like New York. '*New York has ambitious electrification goals, but it lags behind other major cities in the United States significantly.*'[20] From the report of *The New York Times*, we can see that the most important reason is the dense urban environments and shortage of new energy vehicle infrastructure construction in New York. And we'll talk more about that in our next research question.

In addition, we also try to figure out whether political party will affect development of NEV or not. The reason why we study this aspect is that in the United States, the topic of new energy vehicles is actually related to politics. Different parties may have different attitudes towards new energy vehicles to guide the market and policy. In 2012, for example, "*Mitt Romney, then a presidential candidate, was seen as taking the party line when he called Tesla a "loser" and criticized Department of Energy grants under President Obama.*"[21]

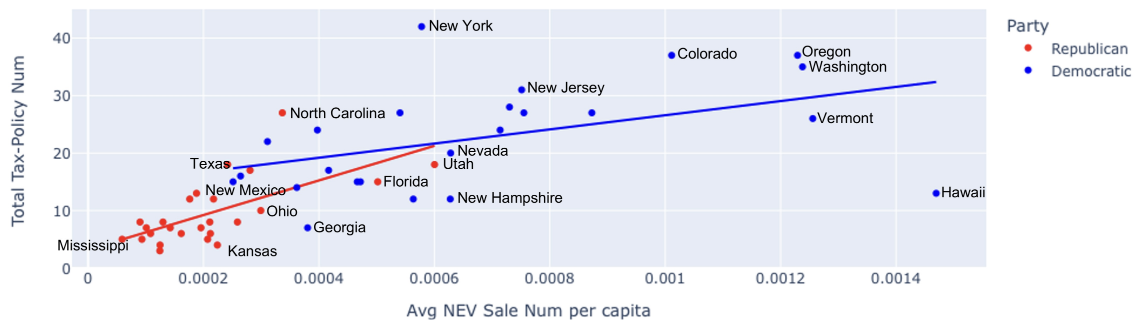

Figure 5.5: The relationship between tax policy and NEV sales in political party perspective.

Our regression model told us that most **democratic** states have more NEV than **republican** states. North Carolina and Utah are the most prominent red states, with the most tax policies and the most new energy vehicle sales, surpassing most blue states, respectively. In contrast, New Mexico and Georgia had the lowest tax and sales among the blue states, respectively.

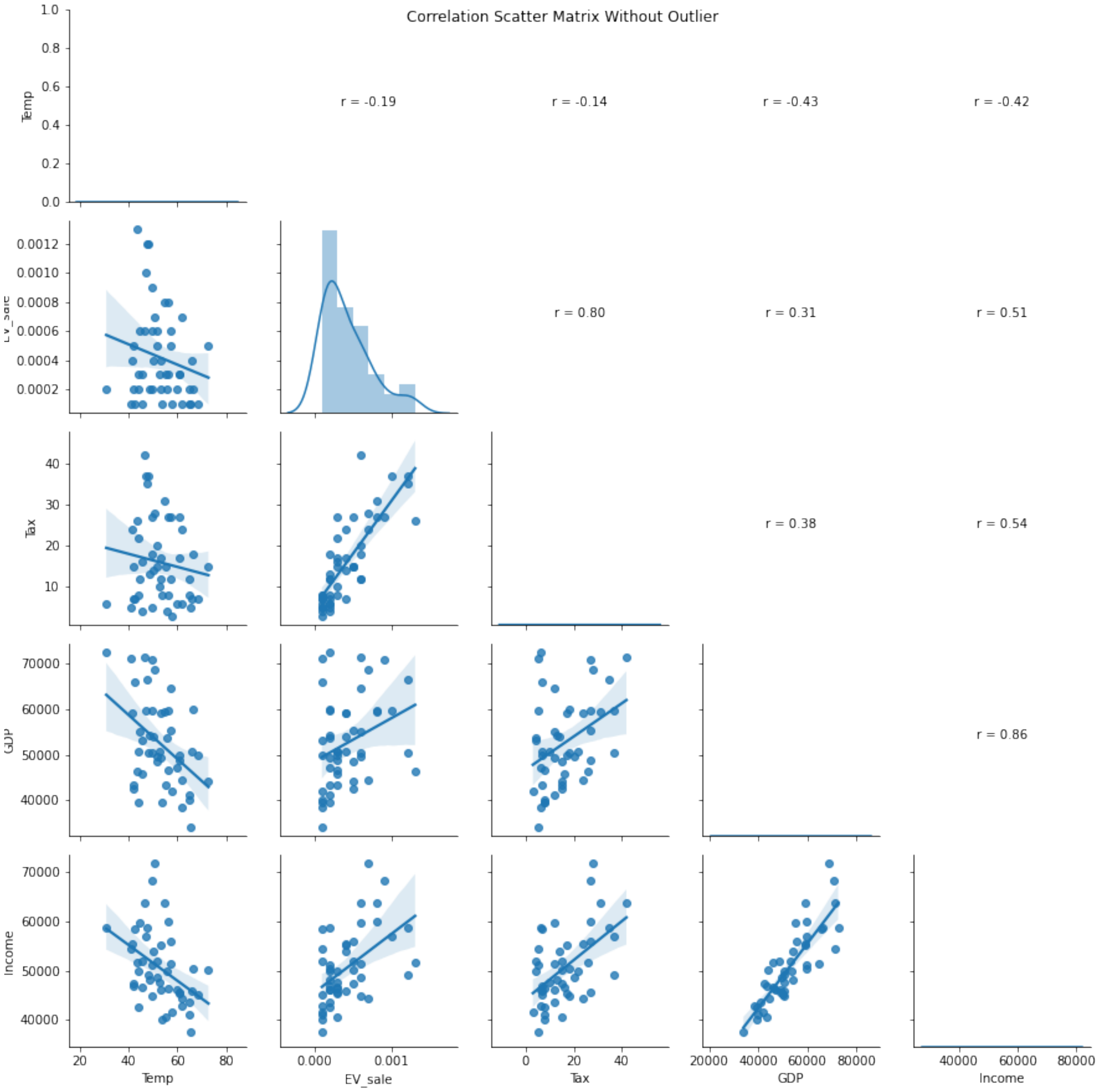

Figure 5.6: Correlations Between Social Environment (Without Outlier).

Furthermore, we did correlation analysis and generated heat maps with regression scattering matrix of these social and environmental factors. There is no doubt that income and GDP are strongly correlated, but in the sales of NEV, tax policy is the most important social environment factor.

5.1.3 Discussion

- GDP and income do affect the sales of new energy vehicles, but income is more important.
- A decent climate is important, more driver prefer to buy NEV in cool places rather than in cold and hot places.

- Tax policy is the most influential factor. More tax policies related to NEV could lead to more car sales, but the higher number of policies does not mean more sales.
- Political parties affect people's willingness to purchase new energy vehicles, and it seems that Democrats are more supportive of new energy vehicles than Republicans.

5.2 RQ2: What are the advantages of new energy vehicles compared to traditional vehicles?

We will compare traditional cars from several different perspectives to find out the advantages of new energy vehicles: mobility(energy supply station) and vehicle performance as well as quality-price ratio, gas emission and value retention.

5.2.1 Data & Method

For vehicle mobility, we have traditional gas station data from 1996 to 2012 and new energy station data from 2007 to 2021. Due to the limitations of the data itself, most our research can only analyze the data based on their intersection: from 2008 to 2012.

To achieve this, we intercepted the range of data in two data frames for different fuel types. In order to get the increment value, we calculate delta values between each two years and add them together. Then combine them together with the geographic information by state name to obtain the following data.

	NEV	Gas	geometry	coords
Alabama	-48.0	100.0	MULTIPOLYGON (((-88.12466 30.283...	(-86.7330, 26.61320)
Alaska	187.0	-205.0	MULTIPOLYGON (((-166.10574 53.98...	(-152.9134, 62.87804)
Arizona	-13.0	27.0	POLYGON ((-112.53859 37.00067, -112...	(-111.6685, 34.16846)
Arkansas	633.0	-647.0	POLYGON ((-94.04296 33.01922, -94...	(-92.4985, 34.75095)
California	3399.0	130.0	MULTIPOLYGON (((-122.42144 37.86...	(-119.991, 37.27239)
...	...	...	...	...

To analyse trend for the development of all NEV supply type. We group all NEV station data by 'Year' and calculate the sum of each year. We can get:

Year	Biodiesel	CNG	E85	Electric	Hydrogen	LNG	Propane	Total
2007	633	771	1699	440	51	38	2110	5742
2008	633	771	1699	440	51	38	2110	5742
...	...	...	...	...	...	...	...	...
2020	712	1549	3946	106814	63	106	2956	116146
2021	730	1510	4331	128474	67	103	2805	138020

In order to analyze the influence of geographical location and political party on the development of supply stations. We chose the 2012 data for both the new energy and traditional station, because both are closest data. then we Concatenate them with our 'Geo&Political' information together.

	Gas	New_Energy	Region	Geo Region	Party
Alabama	6300.0	12.0	Region4	South	Republican
Alaska	361.0	198.0	Region10	West	Republican
Arizona	2027.0	122.0	Region9	West	Democratic
Arkansas	2107.0	682.0	Region6	South	Republican
California	10100.0	4261.0	Region9	West	Democratic

To compare the performance of new energy vehicles, the following aspects will be compared: gas emission, value for money of the vehicle purchase, fuel cost, and value retention. Therefore, emissions, the price of the car in different years, annual fuel consumption and performance data are required.

For purchasing value for money, we have the following data. Comparison based on maximum power and engine size corresponding to price. Whether a new energy vehicle will buy a better performing car for the same price.

name	year	selling_price	fuel	engine	max_power
Maruti Swift Dzire VDI	2014	450000	Diesel	1248 CC	74 bhp
Skoda Rapid 1.5 TDI Ambition	2014	370000	Diesel	1498 CC	103.52 bhp
Honda City 2017-2020 EXi	2006	158000	Petrol	1497 CC	78 bhp
Hyundai i20 Sportz Diesel	2010	225000	Diesel	1396 CC	90 bhp
Maruti Swift VXI BSIII	2007	130000	Petrol	1298 CC	88.2 bhp

For emissions as well as retention of value, we have the following data. An analysis of cars of different fuel types according to year, price, and emissions.

Maker	Genmodel	Year	Price	Gas_emission	Fuel_type	Engine_size
Abarth	124 spider	2016	29365	148	Petrol	1368
Abarth	124 spider	2016	31365	153	Petrol	1368
Abarth	124 spider	2017	29365	148	Petrol	1368
Abarth	124 spider	2017	31365	153	Petrol	1368
Abarth	124 spider	2017	26665	148	Petrol	1368

For annual cost of vehicle, We have official statistics for the period 2014-2022 on annual fuel cost. After cleaning and consolidation, the following data will be analysed: the change in the annual spend on fuel for different fuel types of vehicles and a comparison.

Year	cost_of_gas	cost_of_EV	cost_of_PHEV	cost_of_FCV	Avg_NE
2022.0	1859.905660	698.125000	1318.604651	1300.0	1105.576550
2021.0	2095.423729	695.098039	1492.000000	1320.0	1169.032680
2020.0	2076.817793	646.052632	1445.555556	1350.0	1147.202729
2019.0	1904.559748	628.571429	1258.333333	1350.0	1078.968254
2018.0	1778.682171	618.750000	1213.235294	1250.0	1027.328431

5.2.2 Results

In vehicle mobility analysis, we focus on energy supply station first. While most electric vehicle charging takes place at home, having access to public charging, whether at a retail establishment, a public garage, or other location, may make owning a battery-powered vehicle significantly more feasible. In this project, we will only consider all public new energy vehicle energy supply stations, excluding any home charging places.

We generate this distribution map of new energy stations per vehicle per state in 2020 which can tell us how convenient are new energy charging stations in the United States.

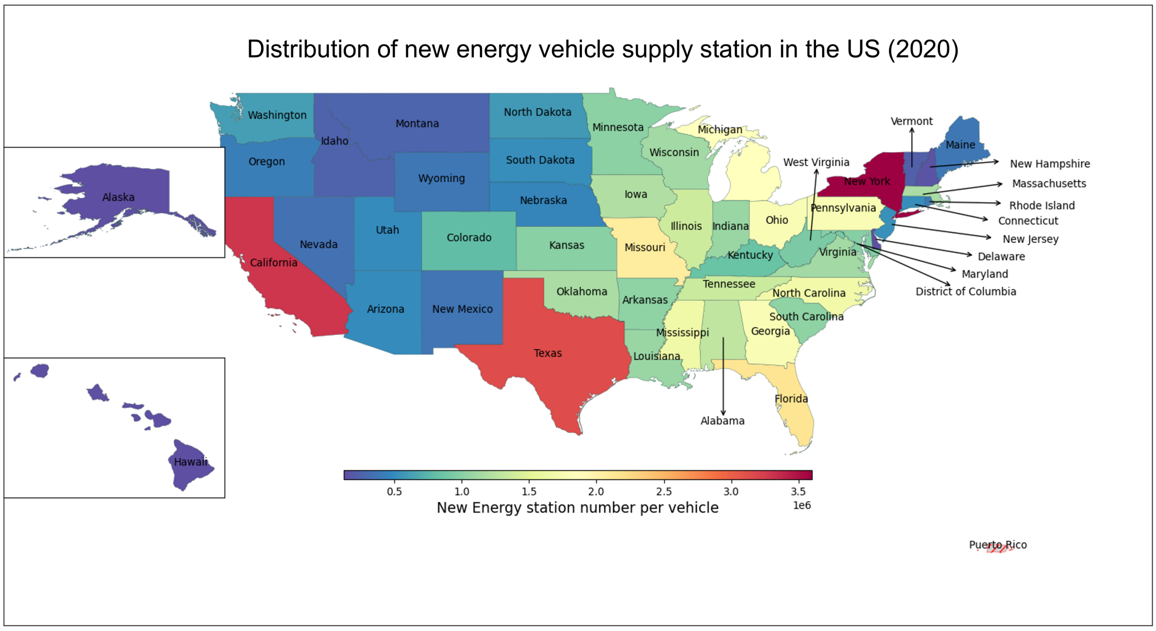

Figure 5.7: NEV Station per vehicle in 2020

From the map above, it is no surprise to see that New York is the most convenient state for NEV drivers, followed by California and Texas. Almost every car can have more than three stations to choose from. Certainly, it because the number of NEVs in New York is relatively low. In addition, from east to west, the average number of supply station per car is decreasing.

This is for new energy vehicle stations only, if we also include traditional gasoline stations, and calculate some of the differences between 2007 and 2012, we can get the following map (Figure 5.8).

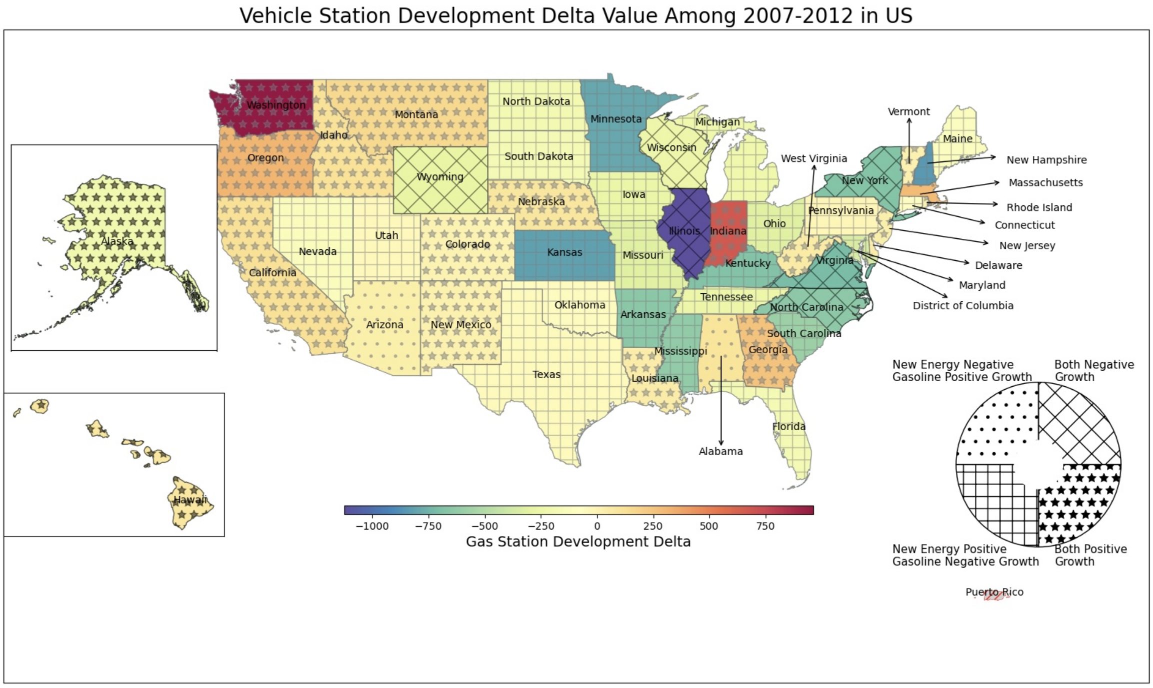

Figure 5.8: Vehicle Supply Station Development Delta Value Among 2007-2012 in US

The above map have two major variables.

- The color of state refers to the gas station development delta value from 2007 to 2012.
- The pattern of state refers to the four types of growth for new energy stations and gasoline stations.

Pattern	Explanation
*	NEV & Gas station both positive growth
x	NEV & Gas station both negative growth
+	NEV station positive growth, Gas station negative growth
.	NEV station negative growth, Gas station positive growth

Surprisingly, New York, Virginia and North Carolina have seen negative growth in both new energy and traditional stations in these years, while most western states have seen positive growth. In the south-central region, there are more alternative fuel stations and fewer traditional gas stations. Only Arizona and Alabama are the completely opposite. This may because of the number of traditional cars is higher than that of new energy vehicles.

The 2008 global financial crisis may also have contributed to the decline in energy supply stations in some states. The financial crisis and the subsequent Great Recession had a significant negative impact on the oil and gas industry, resulting in a sharp drop in oil and gas prices and a contraction in credit. Oil and gas businesses' income have decreased as prices have fallen. Due to restrictive credit circumstances caused by the financial crisis, many explorers and producers had to pay high interest rates while obtaining cash, limiting potential revenues.[22]

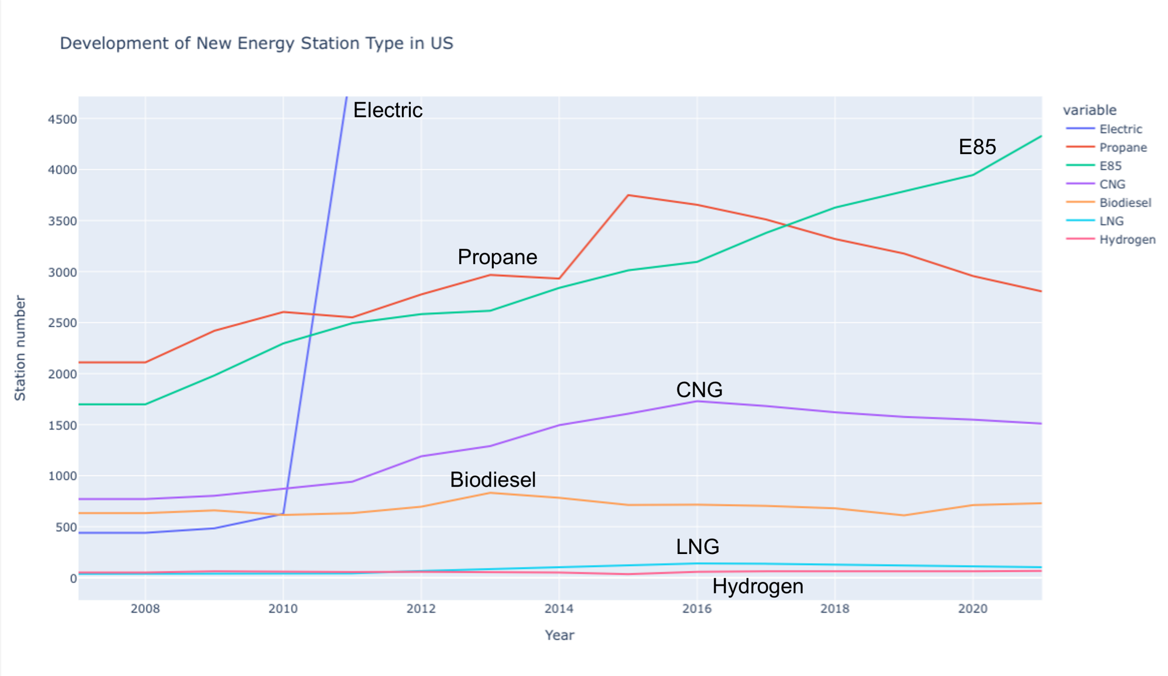

Figure 5.9: Development of New Energy Station Type in US

After analyzing the gas station, we start looking at the development of new energy stations. The results clearly show that the **electric** station is the most popular type with exponential growth since 2010. **E85** is an abbreviation typically referring to an ethanol fuel blend of 85% ethanol fuel and 15% gasoline or other hydrocarbons by volume. It has continued to increase over the past decade. That was alongside by propane fuel, which led ethanol gasoline until 2017, But it was surpassed by ethanol fuel. It is interesting that most of the new energy fuels are generally on the rise, only **propane** and **CNG** start decrease after peaking. We noticed that because the most famously "polar vortex" which saw a sharp drop in temperatures across eastern Canada and much of the United States. The cost of heating fuel has skyrocketed. Prices at major propane hubs have dropped to their lowest levels since 2002 in the two years following the polar vortex.[23]

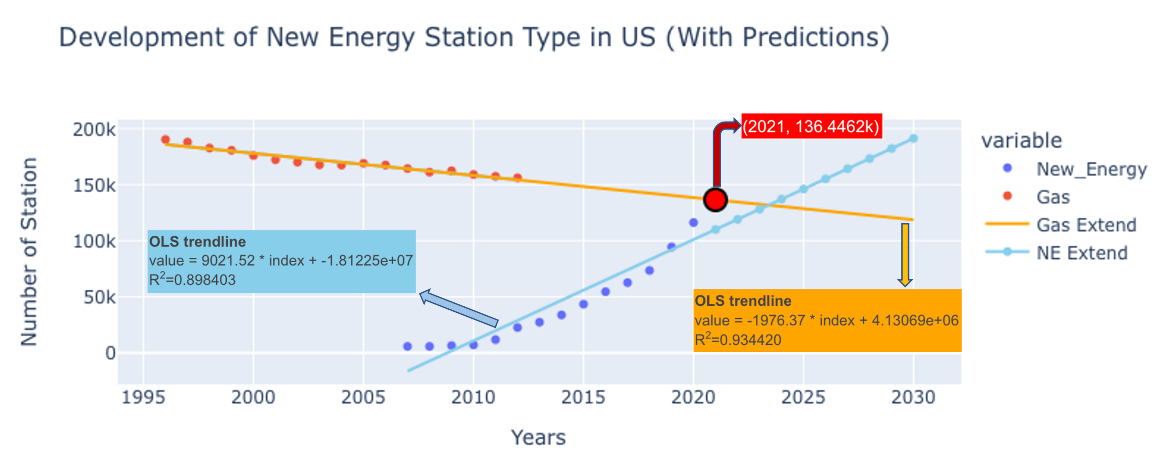

Figure 5.10: Development of New Energy Station Type in US (With Predictions)

Due to the limitations of the year of the data set, we could not actually directly compare gas stations and new energy stations from 1995 to 2021. Therefore, we established a linear regression model for the prediction of unknown and future data through existing data.

Based on our model show in Figure 5.10 , there was a intersected point for stations in 2021, and new energy stations will continue to increase, while gas stations will decrease.

For performance comparison. Gas emissions will be compared first and the value for money.

Figure 5.11: Gas emission (left) and performance vs. price (right))

For the graphic on the left, gas emission will compared. Outliers arise due to the particularly high emissions of some luxury sports cars, such as Lamborghini, Ferrari and Rolls-Royce of the traditional car brands, and Bentley of the new energy hybrids. However, on average and in general, emissions from new energy vehicles are smaller, which suggests that new energy vehicles do cause less pollution. And When people buy a car, value for money is certainly a very important point. Does a new energy car offer better performance for the same price?

For the retention price graphic on the right. A scatter plot was created to compare their value for money. The graph on the right is made by dividing the available engine size and maximum power data by the corresponding price. Thus, a car with better performance and with a lower price corresponds to a scatter point closer to the top right corner of the graph.

Correspondingly, a linear regression model is built to show the difference in overall value for money. The model for new energy vehicles is positioned above the model for conventional vehicles, which means that new energy vehicles are more cost effective overall. For the same price point, new energy vehicles offer better performance, or in other words, they are cheaper for the same performance.

After comparing the advantages considered during the purchase process, it is also important to consider how the new energy vehicle compares to a conventional vehicle after the purchase.

This following chart combines a chart of average vehicle prices in the market with a chart of annual fuel costs. The linear model of price comparison shows that the average price of new energy vehicles in the market is higher than that of conventional vehicles, but their annual fuel cost is only half that of conventional vehicles.

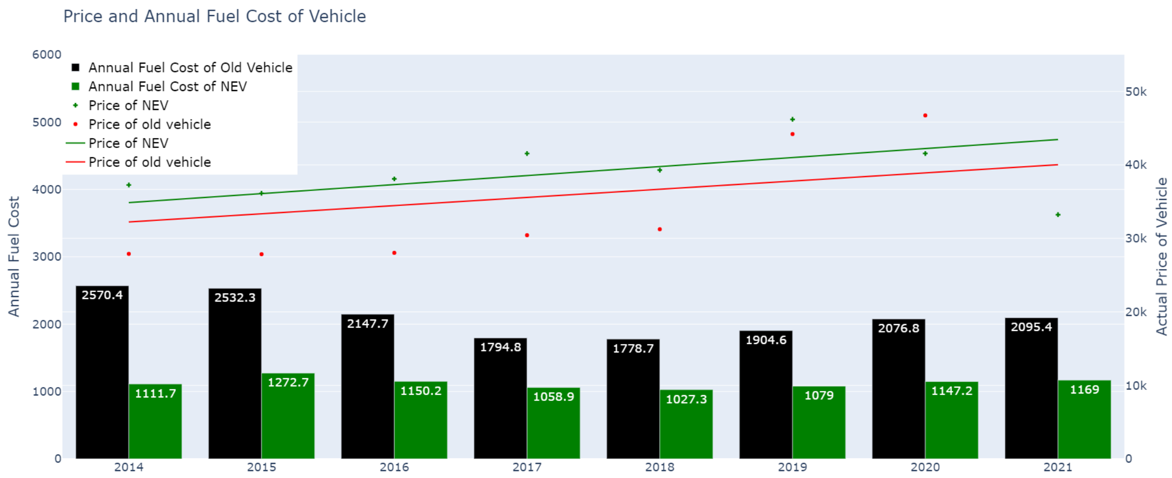

Figure 5.12: Price and annual fuel cost

They do save money in future driving, but compared to the extra money spent on a new car, we have done a rough calculation: from the linear regression model, the average price of a new car is about 5k higher than a conventional car, and the new car saves 1k per year. This also means that it takes more than five years of driving time for the new energy vehicle to start saving money.

But how much lower will the value of the car be in five years' time? A Value Retention graphic was created to explore this:

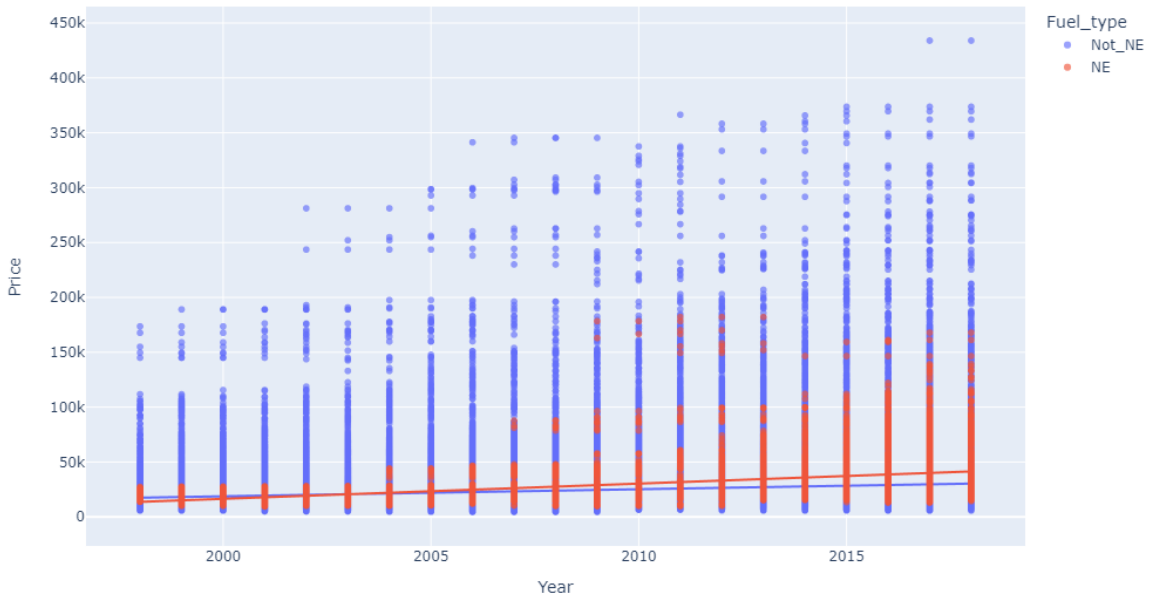

Figure 5.13: Value Retention - Price comparison of new energy and traditional vehicles)

In Figure 5.13, The year is the x-axis and the price of the vehicle is the y-axis. We have over 330,000 vehicles of different fuel types corresponding to the price as well as the year, and we compare the overall retention rate by building a regression model on the whole. It can be seen that the slope of the red part representing new energy vehicles is larger, which indicates that the prices of new energy vehicles are more variable, that new energy vehicles have increased in value to a greater extent than traditional vehicles, and that they retain their value at a higher rate than traditional vehicles. Combined with the result of fuel cost savings, the longer the driving time, the more worthwhile the purchase of a new energy vehicle.

5.2.3 Discussion

The purpose of this research question was to answer one of the most common questions about NEV: Why we need new energy vehicle, what is the advantages? We first looked at the energy supply station of two different fuel type vehicles. As expected, the supply stations of new energy vehicles will completely surpass the construction number of traditional gas stations in the foreseeable future and new energy owners will be more convenient to drive their vehicles.

Some western states, such as California, have been more aggressive in the development of new energy stations due to the major bases of NEV companies such as Tesla. From the perspective of new energy, electric power is still the most popular type in modern society, followed by ethanol, propane and compressed natural gas(CNG).

As the above results show, new energy vehicles have lower emissions and are more cost effective. We believe that those who are conscious of environmental protection and value for money are better suited to purchasing new energy vehicles. For those who want to save money, they need to consider how often they replace their cars. And you are sure you will not be buying a new car anytime soon, then a new energy car would be a good choice.

5.3 RQ3: How to find the safest vehicle in both new energy and traditional vehicle?

We will use the datasets we collect from API of user complaints, recalls, and ratings to compare and analyse traditional and new energy vehicles. Find out which component have most complaint and recall and correlation between them. Also find the brand who have highest grade in safety rating in both fuel type fields.

5.3.1 Data & Method

There is a limitation of our complaint, recall and rating datasets is that it does not include vehicle fuel type, it is an aggregate data. To distinguish a car powered by new energy or gasoline, we analyze it by its model name. We check if the string contain any key words, for example, 'HYBRID', 'ELEC', 'HEV', 'HV', 'EV', 'ENERGI', 'ECO', 'PLUG-IN', 'TESLA', 'PLUGIN', etc. This works for most cases, but there are some specific brands that can be misdetected, like the 'CHEVROLET'. Because of 'CHEVROLET' inclusion of 'EV' in its name, and we reworked our keyword list after discovering this problem. We add a space before 'EV' (' EV')and after 'EV' ('EV ') to avoid this kind of situation. This method also applies to other keywords.

For the component analysis, since a cell may contain multiple components, we split them by comma and count the frequency, then we can get following data:

Component	Frequency
ELECTRICAL SYSTEM	102386
POWER TRAIN	93831
SERVICE BRAKES	86440
AIR BAGS	82026
STEERING	68543
...	...

For rating analysis, our dataset describe how each vehicle was rated in the collision experiment. These grades are divided into sections and scored individually based on the components of the vehicle. e.g. Front Crash Driver side Rating, Front Crash Passenger side Rating, etc. We combine them to get a score ratio for each car. There are 14 aspects in total, and each of them have 5 max grade. Therefore, the grade range is [0,70]. To calculate grade ratio, just add 14 aspects subgrade together then divide by 70.

Here is part of the final rating dataframe:

	Make	Model	ModelYear	VehicleDescription	gradeRatio
0	MITSUBISHI	ECLIPSE	1992	1992 Mitsubishi Eclipse 2-DR.	0.114286
1	BMW	6 SERIES	2017	2017 BMW 6 Series C RWD	0.071429
2	BMW	X5 HYBRID	2021	2021 BMW X5 Hybrid SUV AWD	0.214286
3	LEXUS	LS 500 2021	2021	Lexus LS 500 4 DR AWD	0.214286
...	...	...	...	...	...
7313	BMW	7 SERIES	2002	2002 BMW 7 Series w/SAB	0.071429

5.3.2 Results

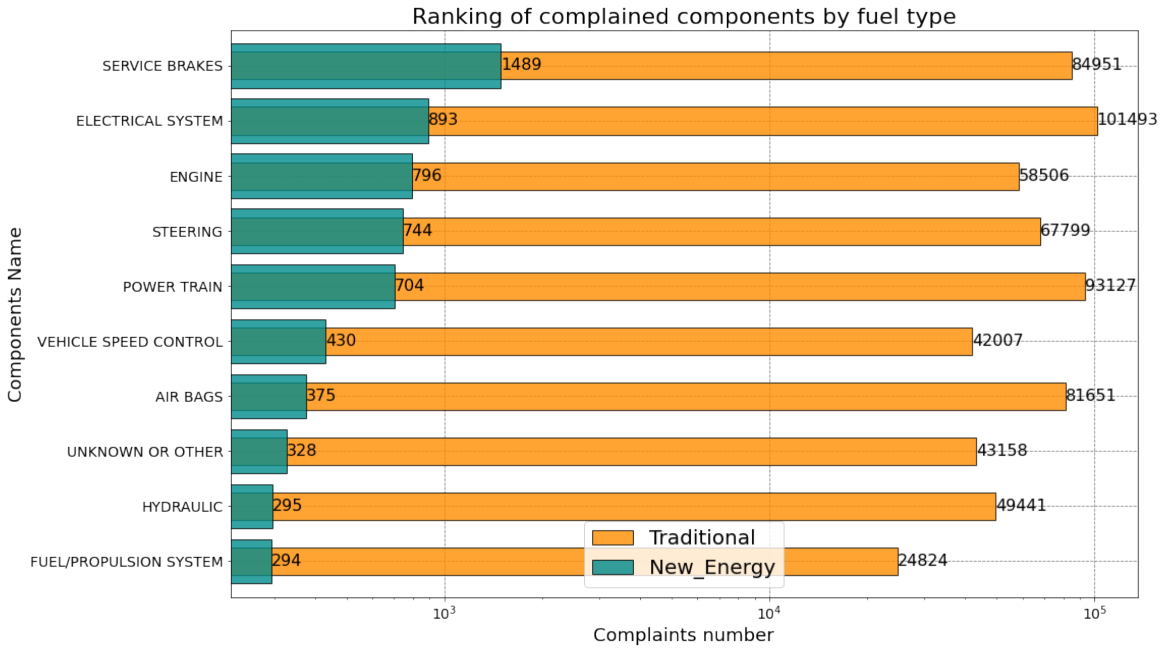

Figure 5.14: Complaint components rank by fuel type

It is interesting to know which component of the vehicle are most vulnerable to complaints. We are realized it is true that most of the components are complained in new energy and traditional vehicle are different. The number of complaints received by the different parts of the vehicle was made into Figure 5.14, and the top 10 components received the most complaints were ranked.

As can be seen from the histogram above, the most common complaint is the electrical system of traditional cars, while the most common complaint of new energy vehicles is the service brakes. It is clear that there are different components to consider when purchasing a vehicle, depending on the type of car.

We also looked at the ranking of the number of complaints received in daily driving, and the ranking of the number of complaints received in crashes. This was used to examine what kind of vehicle would be safer, both in normal life and in the event of an unexpected event (e.g. a collision).

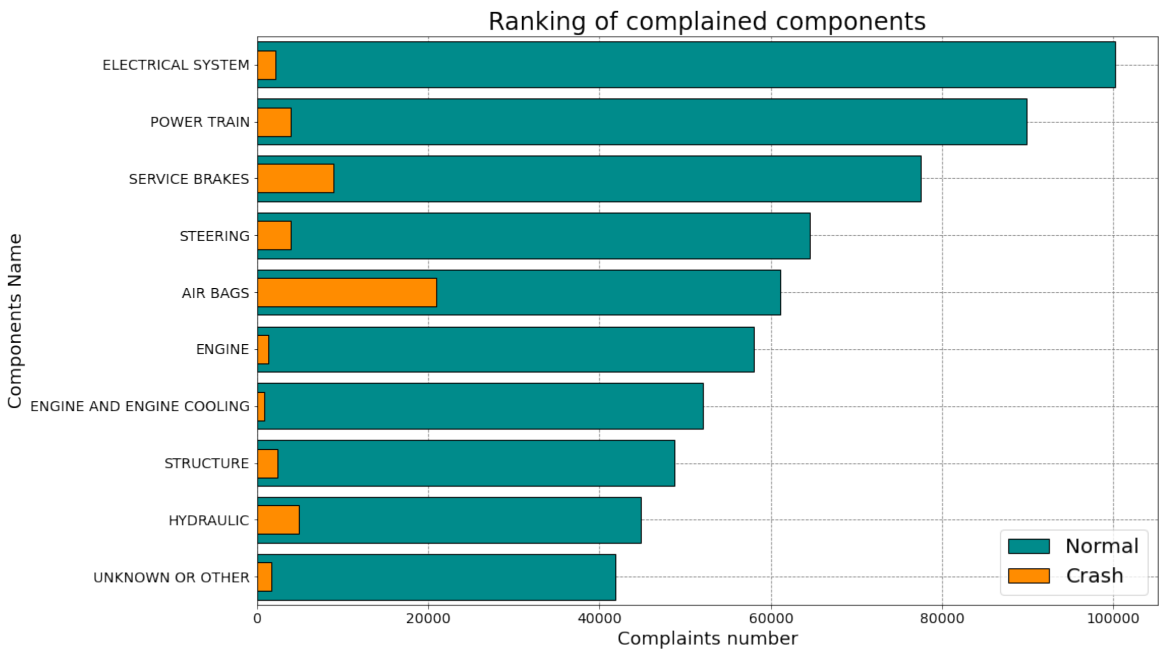

Figure 5.15: Complaint components rank crash

Figure 5.15 shows the top 10 components that received the most complaints in normal life as well as in crash. The results show that the electric system is the most complained about in everyday life and that complaints about air bags are much higher in collisions than in other components.

When purchasing a vehicle, people need to know more about the electric system and power train, especially about air bags, which are very important for safety.

After analyzing the ranking of components, we believe that in daily maintenance, many components will be damaged at the same time, and it may be interesting to study the correlation between them.

From the Figure 5.16 below, except for the obvious components like electric and electrical system and air with air bags. We can also see that hydraulic have some connection with service brakes. It seems chest clip always crash with buckle and harness, same as base, shell and carry handle.

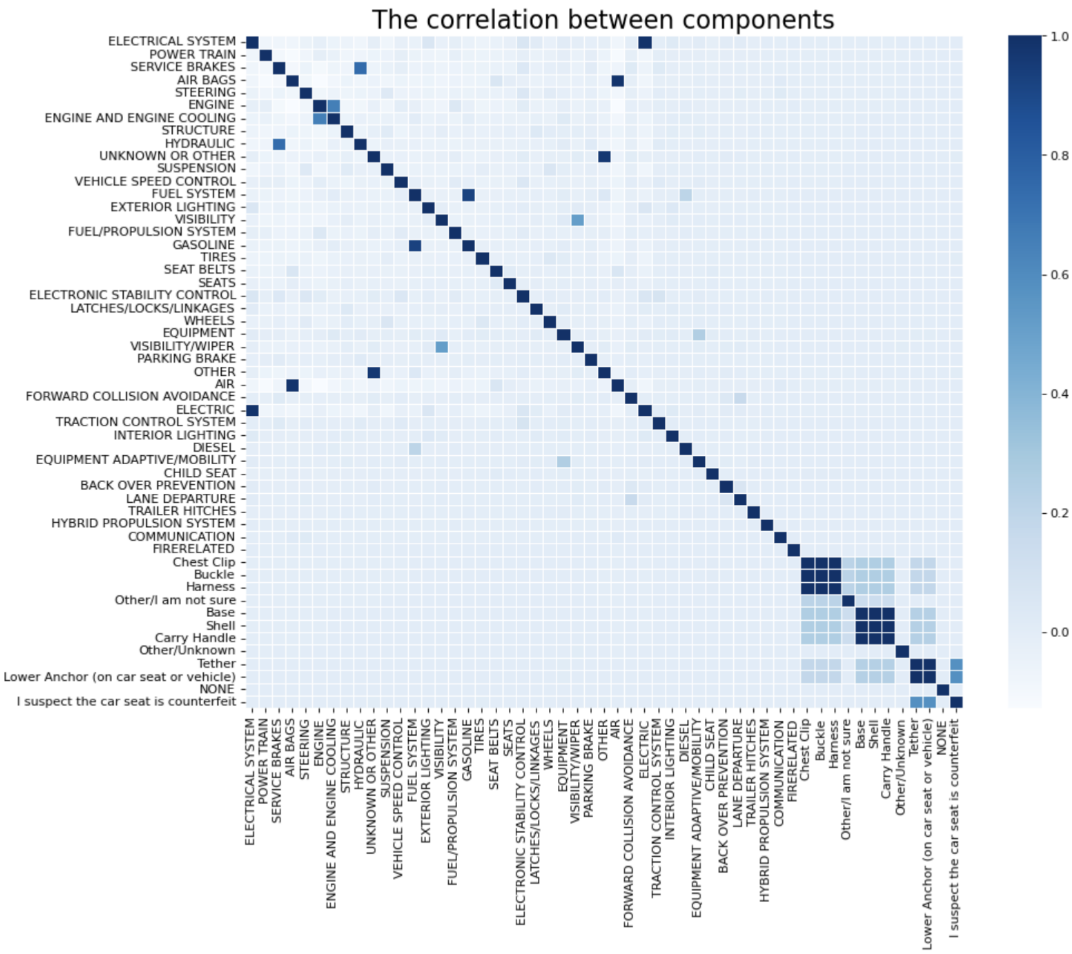

Figure 5.16: Complaint components correlation

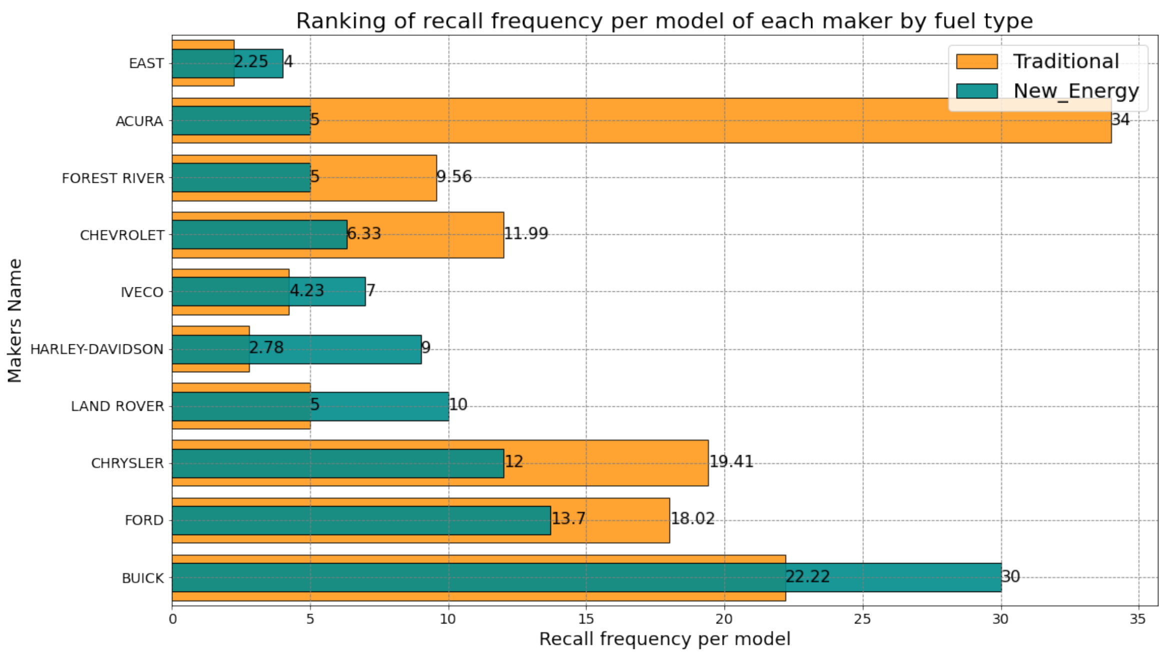

Figure 5.17: Complaint ratio per model by makes & fuel type

In addition, we also ranking of recall frequency per model of each maker by fuel types. From figure 5.17, it is clearly show that *Buick* had the most new energy vehicles recalls, with an average of 30 recalls per model. It also made a number of recalls of its traditional cars, taking second place by 11.78 behind the *Acura* (34 recalls). *Ford* had the second largest number of new energy vehicles recalls which is 13.7 times and the fourth largest number of traditional vehicles (18.02 times).

Moreover, we also calculate the automobile recall time cycle difference for each maker by following criteria.

$$\textit{Time difference}(delta - T) = Year\ of\ recall\ announced - Year\ of\ vehicle\ released \quad (5.1)$$

Then, we separate them into two groups by fuel type and showed their median values via boxplot which showing below:

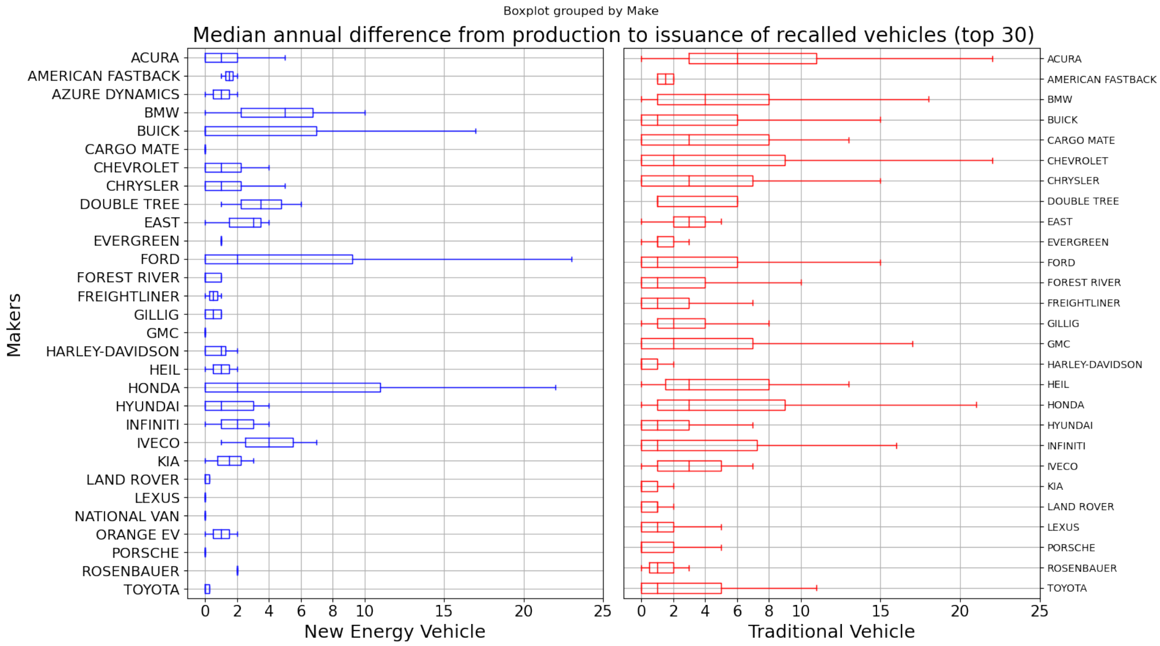

Figure 5.18: Median annual difference from production to issuance of recalled vehicles (top 30)

It is clearly show that the recall delta-T (time differ) of new energy vehicles is generally shorter than that of traditional fuel vehicles. Most of them are concentrated in 0-2.5 years slot, while traditional fuel vehicles are concentrated in the range of 2.5 to 5 years, which also shows that new energy vehicles technology is not mature, at least not to shake the status of traditional vehicles.

In the ranking of new energy vehicles, BMW and IVECO's value is higher than others that means they have high recall cycles and that they are better than others. However, some brand like Buick, Ford and Honda, even they have higher records, but their median recall period still lower than others, although this is perhaps due to the large number of their models.

In traditional gasoline vehicle, Acura have the higest recall time differ, median 6 years followed by BMW with 4 years. It's seems BMW is good at both type of vehicles.

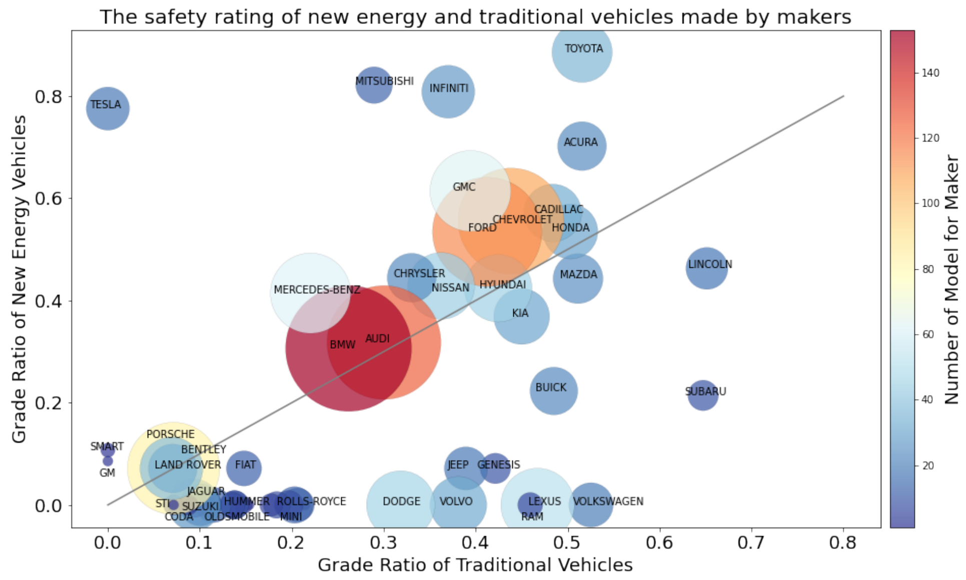

Figure 5.19: Collision rating grades by makes with model number

After calculating the total grade ratio obtained for each car in the crash test. We grouped the data by vehicle type(Traditional and new energy) and maker to make Figure 5.19. Using grade ratio of traditional vehicles as $x-axis$, and new energy vehicles as $y-axis$. Each scatter point in the graph corresponds to the grade ratio of each manufacturer in the safety factor for new energy vehicles versus conventional vehicles. Also, the sizes were given to scatters by their numbers of model.

To make it easier to see whether they balance out in terms of safety for new or conventional cars, we have added an equation with $y = x$. The closer to the line of this equation, the more balanced the cars produced by the manufacturer.

From the result, the more model the carmaker have more balance on the overall safety rating they have, for example BMW and Audi. The Toyota and Lincoln are safer than others on NEV and traditional vehicles, respectively.

5.3.3 Discussion

In this final research section we considered whether some maker were more safe than others by three dimension, complaints, recalls and ratings. The results of the data tell us that different components indeed receive variate complaints and recalls on different fuel type vehicles. Service break is the most vulnerable component to complaints for new energy vehicle, but the largest number of complaints about gasoline vehicles are electrical systems. BMW performs best in car recalls, with both new energy and gasoline vehicles having the very good no-recall retention time. It is also a balance maker in terms of safety in different fuel types. The more model the carmaker have more balance on the overall safety rating they have. Hyundai is the most balanced maker even it have lower amount model than Audi and Ford.

Chapter 6: **Conclusions**

In the stage of global automotive development toward Eco-friendly, more and more people begin to choose to buy new energy vehicles.

This report describes an analysis of new energy vehicles on the future automotive industry from three aspects and whether they will lead the future direction. We analyze the impact of social environment on the development of new energy vehicles, as well as the comparison between new energy vehicles and traditional vehicles, especially in terms of performance and price. In addition, we compared the safety of new energy vehicles and traditional gasoline vehicles from three dimensions, complaints received, vehicle recalls and lab simulated collision test rating.

In order to support our research topic, we not only found the existing datasets through the government website but also wrote the crawler to obtain and integrate some data through API. Based on the analysis of the data set after data cleaning, we observed the following findings:

- A region with the right climate, a high GDP and income per capita, and the government that offers more incentives or more tax policies may conducive to the development of new energy vehicles.

- More new energy supply stations mean that the mobility of new energy vehicles is steadily increasing and surpassing that of conventional vehicles.

- In comparison, new energy vehicles are more environmentally friendly and more cost effective. In terms of saving money, if a person wants to save money on fuel, a new energy vehicle is indeed a good choice, but the costs that need to be considered are: whether the savings on fuel in terms of driving time outweigh the cost of purchasing the vehicle, and the need to understand that the longer the driving time the more worthwhile the purchase of a new energy vehicle.

- It is worth to consider the most complained and recalled component when people wish to buy a new energy vehicle, such as service break and electrical system. And manufacturers with more models will have a better balance between the safety of new energy vehicles and traditional vehicles, indicating their technology is mature.

Finally, there is one undeniable point: there are some limitations to this project. Most comparisons between new and traditional vehicles are based on overall performance, and it cannot be ruled out that some brands of conventional cars perform better than other new energy cars. In addition, the data set does not include all car models on the market. The aim of this project is to identify the advantages of new energy vehicles in terms of general trends, but consumers will need to make their choice based on the performance of the manufacturer in the purchasing process.

Acknowledgements

We would like to express our deepest thanks to Professor Barry Smyth, for giving us inspiration at the beginning of the project, guiding the project through the process and weekly advice during our project.

We would also like to thank our teaching assistant: Ciara Feely who gave a lot of recommendations for the direction of research.

And we would like to give special thanks to our demonstrator: Rian Dolphin. Rian has helped us with many questions, gave advice and provided many useful resources.

Bibliography

1. Mallapaty, S. How China could be carbon neutral by mid-century. *Nature* **586,** 482–484 (2020).
2. Plötz, P., Axsen, J., Funke, S. A. & Gnann, T. Designing car bans for sustainable transportation. *Nature Sustainability* **2,** 534–536 (2019).
3. Friedman, L. Biden Administration Begins $3BillionPlanforElectricCarBatteries.$ *The New York Time.* https://www.nytimes.com/2022/05/02/climate/biden-electric-car-batteries.html (May 2022).
4. Faisal, I. Ukraine conflict: Petrol at fresh record as oil and gas prices soar. *BBC.* https://www.bbc.com/news/business-60642786 (Mar. 2022).
5. *NHTSA datasets and apis* https://www.nhtsa.gov/nhtsa-datasets-and-apis.
6. Awan, A. English. in *Global EV Outlook 2020* 185–219 (International Energy Agency, June 2020).
7. Wan, J. P., Xie, L. Q. & Hu, X. F. Study on the electric vehicle sales forecast with TEI@I methodology. *International Journal of Knowledge Engineering and Data Mining* **7,** 1–38 (2021).
8. Hidrue, M. K., Parsons, G. R., Kempton, W. & Gardner, M. P. Willingness to pay for electric vehicles and their attributes. *Resource and energy economics* **33,** 686–705 (2011).
9. *GDP and Personal Income* https://apps.bea.gov/iTable/iTable.cfm?reqid=70&step=1&acrdn=2.
10. NCEI.Monitoring.info@noaa.gov. *Climate* https://www.ncdc.noaa.gov/cag/statewide/time-series/48/tavg/ann/12/1990-2022?base_prd=true&begbaseyear=1901&endbaseyear=2000.
11. *Federal and state laws and Incentives* https://afdc.energy.gov/laws/.
12. Abdoul84. *EVsales* https://github.com/Abdoul84/Project-One-.
13. Eric. *Geojson and KML data for the United States* https://eric.clst.org/tech/usgeojson/.
14. *Alternative fueling station counts by State* https://afdc.energy.gov/stations/states.
15. *Gas station counts by State* https://afdc.energy.gov/files/u/data/data_source/10333/10333_gasoline_stations_year.xlsx.
16. Huang, J., Chen, B., Luo, L., Yue, S. & Ounis, I. *DVM Car Dataset: A large-scale dataset for Automotive Applications* Aug. 2021. https://deepvisualmarketing.github.io/.
17. *Download Fuel Economy Data* https://www.fueleconomy.gov/feg/download.shtml.
18. *Electric cars: Lower-income countries fall behind, with uptake linked to GDP per capita* Aug. 2021. https://www.acea.auto/press-release/electric-cars-lower-income-countries-fall-behind-with-uptake-linked-to-gdp-per-capita/.
19. Briscoe, N. *How much range will your electric vehicle lose in cold weather?* Jan. 2022. https://www.irishtimes.com/life-and-style/motors/how-much-range-will-your-electric-vehicle-lose-in-cold-weather-1.4757241#:~:text=%E2%80%9CAt%20optimal%20temperatures%2C%20EVs%20are,loss%20of%20range%20is%20apparent..

20. Hu, W., Gavrielov, N. & Ewing, J. *Here's how slowly New York City is moving on Electric Vehicles* Apr. 2022. https://www.nytimes.com/2022/04/05/nyregion/nyc-electric-vehicle-evs.html.

21. *Romney slams Tesla amp; Fisker again in Final presidential debate* Oct. 2012. https://www.cbsnews.com/pittsburgh/news/romney-slams-tesla-fisker-again-in-final-presidential-debate/.

22. Investopedia, T. *How did the 2008 financial crisis affect the oil and gas sector?* Sept. 2021. https://www.investopedia.com/ask/answers/052715/how-did-financial-crisis-affect-oil-and-gas-sector.asp.

23. Regulator, C. E. & of Canada, G. May 2016. https://www.cer-rec.gc.ca/en/data-analysis/energy-commodities/natural-gas-liquids/report/2016/propane-market-review-2016-update-energy-briefing-note.html?=undefined&wbdisable=true.

Appendices

This appendix contains information relevant to our project as it is similar to the content in the main report and is included in the appendix for brevity.

Appendix A – Individual Contributions

- Zhi Zhang:
  - API data collection
  - Data Cleaning
  - RQ1 - NEV sales, Tax Policy & Correlations.
  - RQ2 - Mobility, Energy supply station.
  - RQ3 - Complaints, Recalls & Ratings.
  - Conclusion.
- Han Lin:
  - Portals data collection
  - Data Cleaning
  - RQ1 - Income, GDP & Climate.
  - RQ2 - Vehicle performance & Retention price.
  - RQ3 - Complaints, Recalls & Ratings.
  - Conclusion.

Appendix B – Datasets & Notebooks

Here we list the main datasets and notebooks used in our project, along with a brief description of each file name.

Datasets Used

1. *Ev_sales_per_capita.csv* – Electric vehicle sales per capita data in each state of the United States.
2. *incomes.csv* - Income per capita in each state of the United States.
3. *gdp.csv* - GDP in each state of the United States.
4. *climate_state.csv* - Climate in each state of the United States.

5. *policy_state.csv* - Number of tax policies related to new energy vehicles in each state of the United States.
6. *population.csv* - Population in each state of the United States.
7. *US_Map_5m.json* - US GeoJSON shape file.
8. *station.csv* - The number of new energy stations in each state each year.
9. *gas_station.csv* - The number of traditional gasoline stations in each state each year.
10. *States_Info.csv* - Contain some geographic region and political party information for each state.
11. *car_reg_2015_and_2020_per_capita.csv* - Number of all types vehicle registrations per capita by state in 2015 and 2020.
12. *EV_reg_2020.csv* - Number of electric vehicle registrations by state in 2020.
13. *train-data.csv* and autos.csv - Vehicle's performance data.
14. *Car details v3.csv* - Price and max power of vehicles.
15. *Vehicle_Dataset.csv* - Vehicle's gas emission data.
16. *Vehicle_annual_cost.csv* - Annual cost of fuel by vehicle fuel types.
17. *Complaints.pkl* - Vehicle complaints aggregation dataset.
18. *Recalls.pkl* - A aggregated dataset on recalls by automakers.
19. *Ratings.pkl* - A aggregated dataset of vehicle collision simulation scores.

Notebooks Used

Crawl API & Data Clean [Zhi]

- *0101_Fetch_Complaints_Data.ipynb* - Crawl all complaints data from API and write to JSON files.
- *0102_Clean_Complaints_Data.ipynb* - Aggregate all complaint files and clean the data.
- *0201_Fetch_Recalls_Data.ipynb* - Crawl all recalls data from API and write to JSON files.
- *0202_Clean_Recalls_Data.ipynb* - Aggregate all recall files and clean the data.
- *0301_Fetch_Rating_Data.ipynb* - Crawl all ratings data from API and write to JSON files.
- *0302_Clean_Rating_Data.ipynb* - Aggregate all rating files and clean the data.

Data Clean

- *1000_Clean_Vehicle_Dataset.ipynb* [Han] - Clean up data including manufacture, model, year and gas emission, in preparation for the RQ2 gas emissions comparison.
- *2000_population_clean.ipynb* [Han] - Clean up the data that includes population. Make it easier to combine with EV sales, GDP and income to calculate per capita values to ensure comparability between states

- *2100_income_cleaning.ipynb* [Han] - Clean up the dataset of income for different states, it will be used to analysis in RQ1
- *2200_GDP_cleaning.ipynb* [Han] - Clean up the dataset of GDP and use 'population' dataset to get GDP per capita for different states.
- *2300_policy_clean.ipynb* [Han] - Clean up the dataset of policy to get the number of laws or incentives for different states.
- *2400_climate_clean.ipynb* [Han] - Clean up the datasets of temperature in different states, calculate annual temperature and combine each file into one master datasets.
- *3000_EV_clean.ipynb* [Han] - Clean up the dataset of EV sales for different states, combine with population dataset to get per capita value.
- *3100_car_reg_clean.ipynb* [Zhi] - Clean up the dataset of total number of registered vehicles per state in 2015 & 2020, including combine with population dataset to calculate per capita value.
- *4000_Station.ipynb* [Zhi] - Clean up the dataset contains the number of all new energy stations in different sates including simplify operations such as simplify column names.
- *4200_gas_cleaning.ipynb* [Han] - Clean up the dataset of number of gas stations for different states, including check the name of states are correct.
- *5000_vehicle_cost_cleaning.ipynb* [Zhi & Han] - Read annual fuel cost data from multiple files and combined according to fuel type.

Data Analysis

- *1000_US_EV_Sales_Map.ipynb* [Zhi] - Visualizing data by presenting it on a map.
- *2000_Tax_and_Sales.ipynb* [Zhi] - Explore the relationship between tax policy and sales of new energy vehicles.
- *3000_Economic.ipynb* [Han] - Describe how economic aspect affect the sales of new energy vehicles.
- *4000_Climate.ipynb* [Han] - Study on the relationship between climate and new energy vehicle sales.
- *5000_Correlation.ipynb* [Zhi] - Find the correlation between these factors.
- *1000_US_Station_Map.ipynb* [Zhi] - Analyze the development of energy supply stations and display them on maps.
- *2000_Gas_vs_NEV_Station.ipynb* [Zhi] - To study the future relationship of development of energy supply stations.
- *3000_Performance_comparison.ipynb* [Han] - Compare the new energy and traditional vehicles in three ways to find out their advantages (Quality-price, emission and value retention).
- *4000_Price_Prediction.ipynb* [Han] - Predict the actual price of future new energy vehicles and traditional vehicles in future markets from manufacturing and driving costs/savings.
- *1000_Complaints.ipynb* [Zhi & Han] - Analyze the complaint data to determine the number of component damaged by different fuel types and whether they are caused by accidents.
- *2000_Recalls.ipynb* [Zhi & Han] - Study vehicle recall data and judge the date difference between recall release date and model year and count the recall frequency.
- *3000_Ratings.ipynb* [Zhi & Han] - Analyze the rating of automobile crash test and judge the result of collision rating of different manufacturers and models.

Appendix C – Supplementary Results

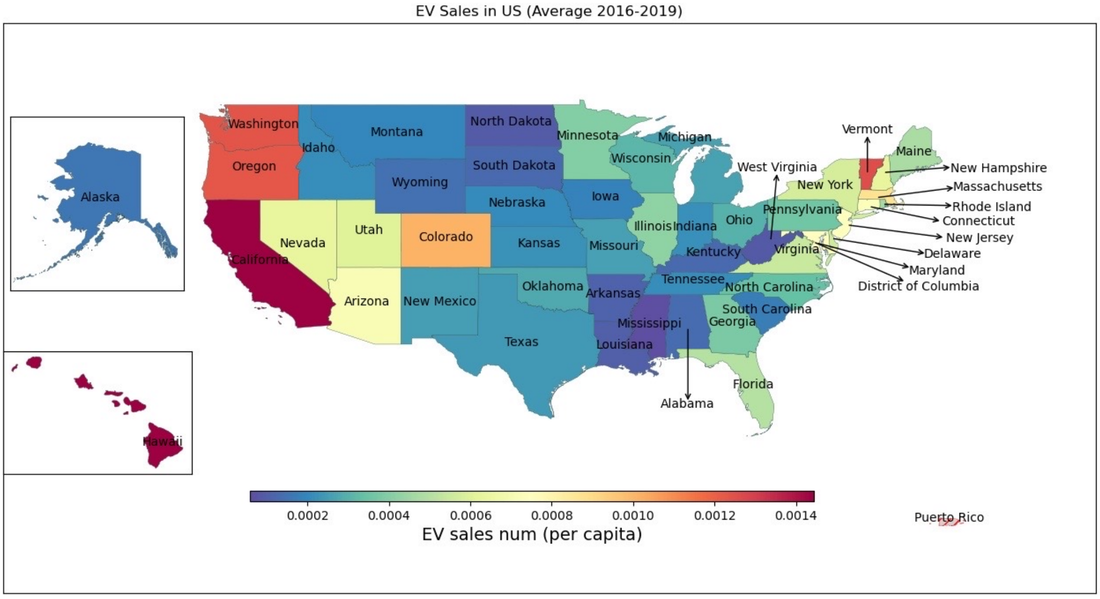

Figure 6.1: US NEV Sales Map

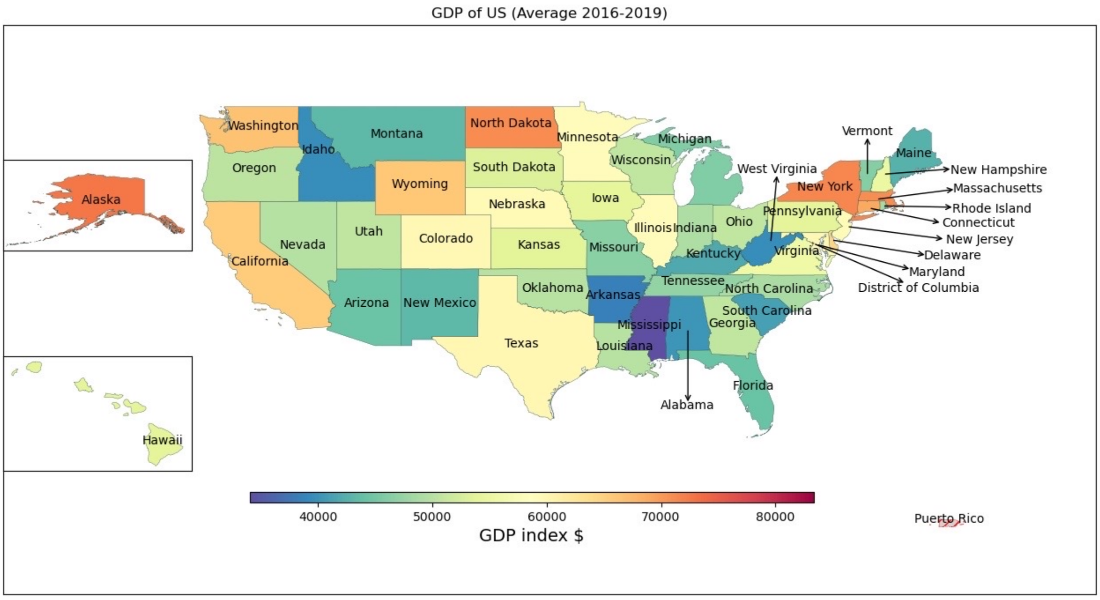

Figure 6.2: US GDP Map

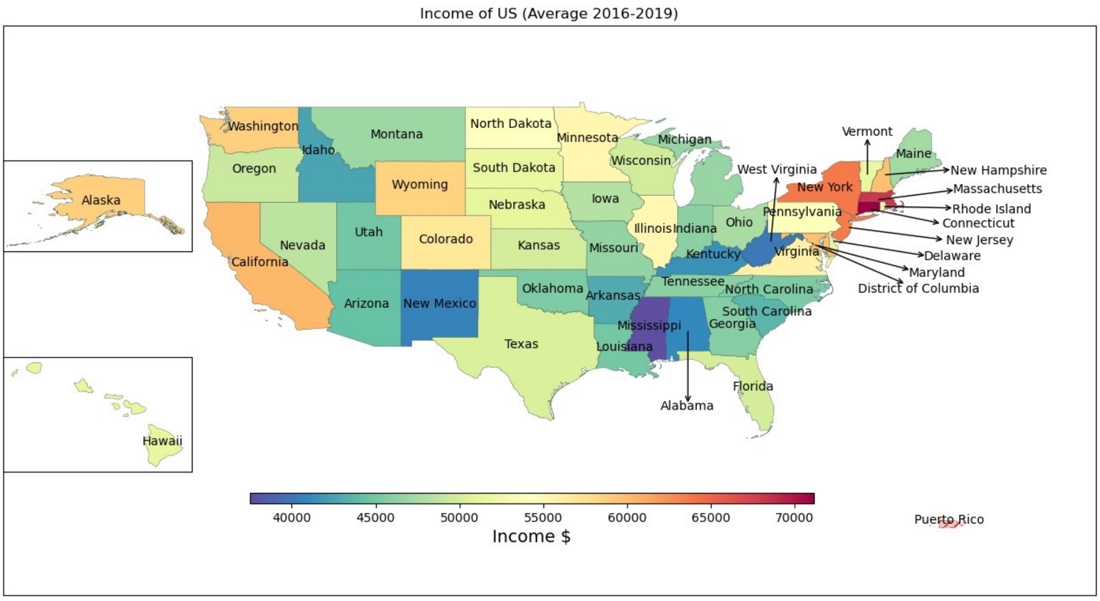

Figure 6.3: US Income Map

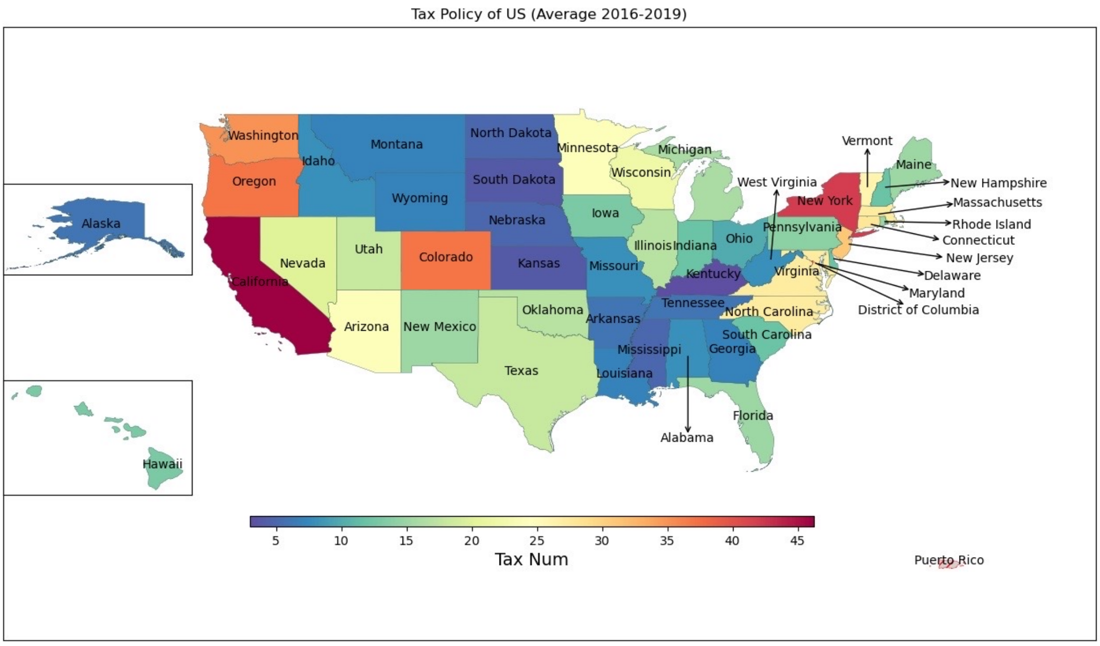

Figure 6.4: US NEV-Related Tax policy Map

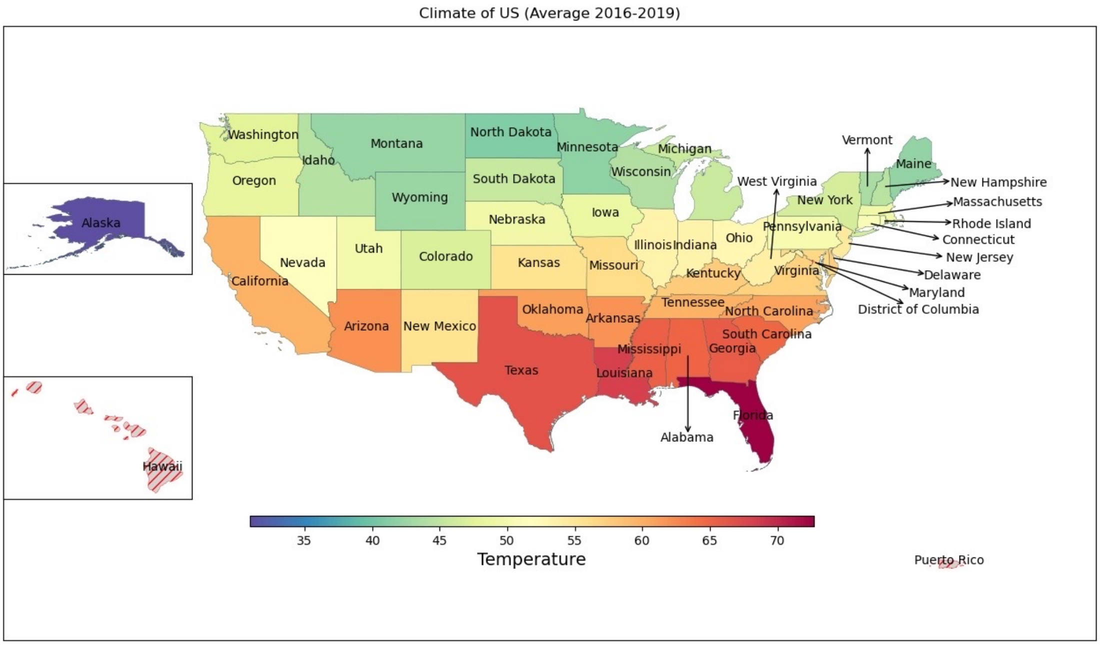

Figure 6.5: US Climate Map

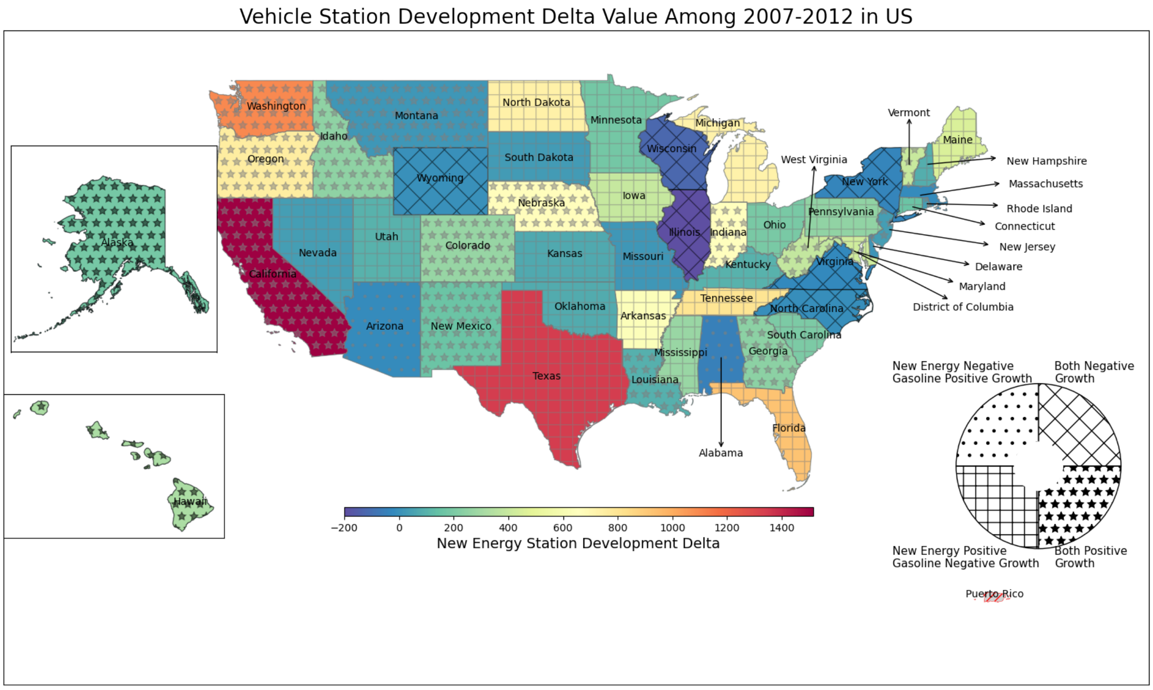

Figure 6.6: Vehicle Supply Station Development Delta Value Among 2007-2012 in US

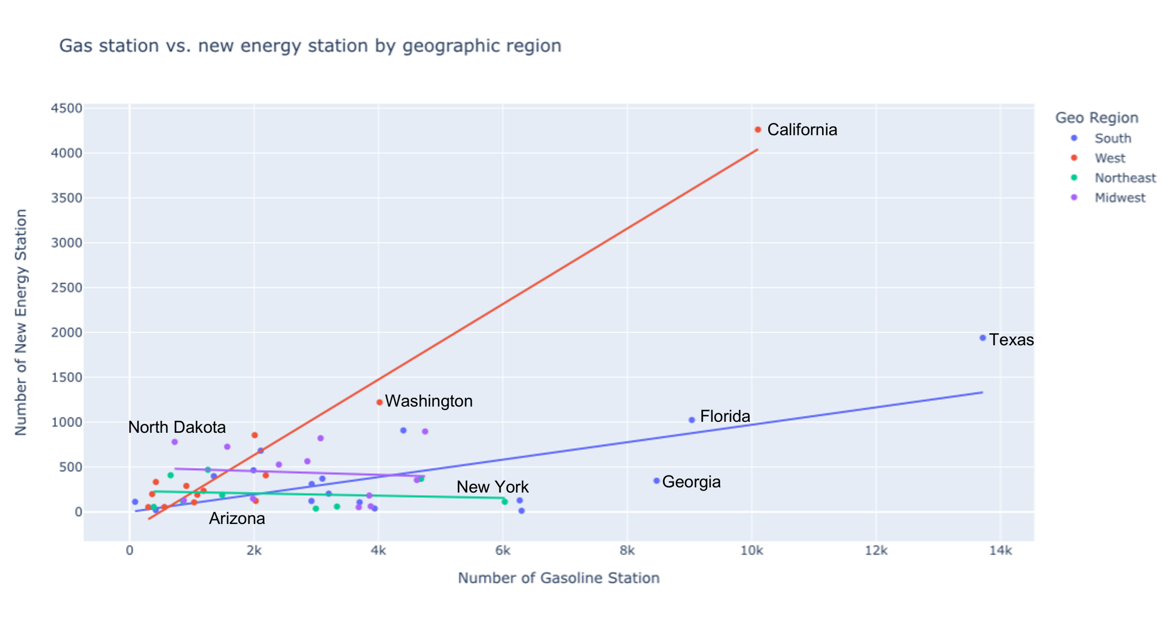

Figure 6.7: Gas station vs. new energy station by US geographic region

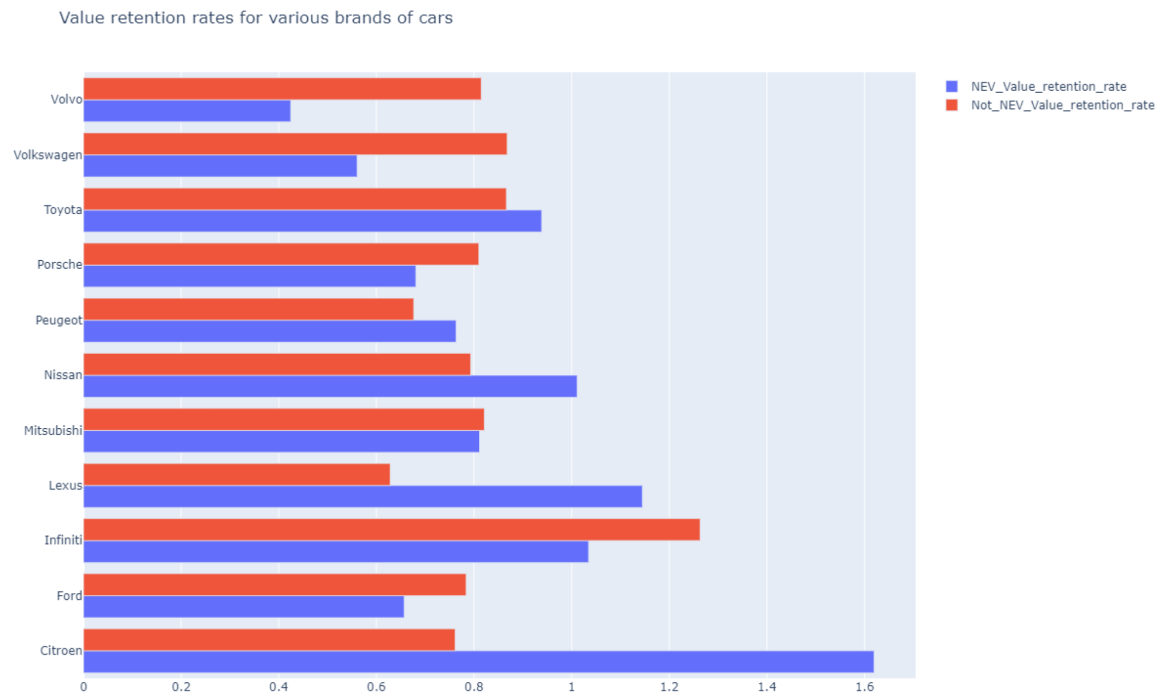

Figure 6.8: Value Retention for different brands

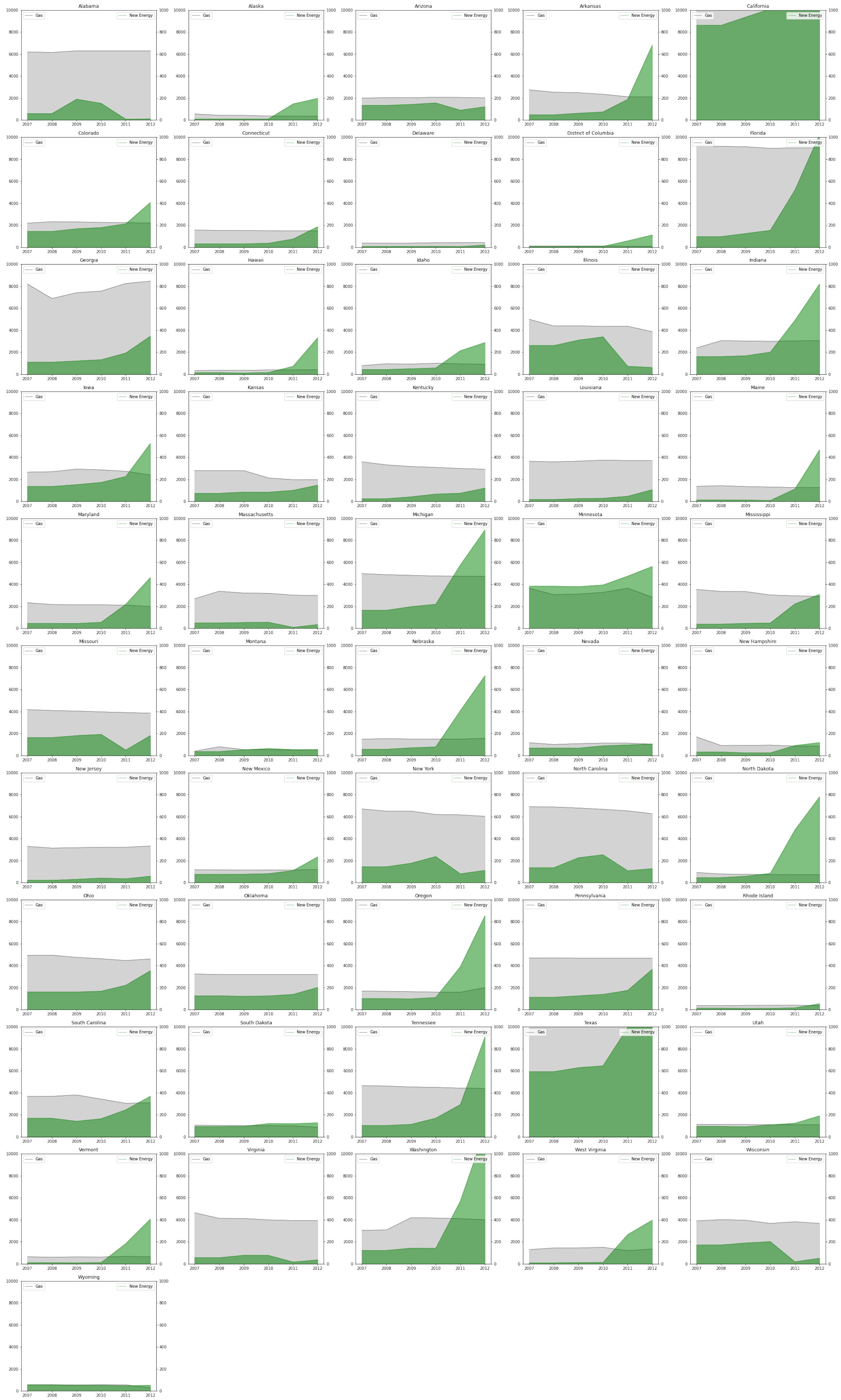

Figure 6.9: The development of all types stations from 2007 to 2012 by states