Group Project on Electric Vehicles in New York State

Group Members:

Jia Ying Lv (jl5727@columba.edu)

Taotao Jiang (tj2441@columbia.edu)

Wanying Li (wl2722@columbia.edu)

Yunxuan Chen (yc3682@columbia.edu)

Using New York State as an example, our project aims to analyze the historical trend and current status of the EV market, and predict the future image of EV industry. Specifically, our project consists of 3 parts: first, we analyze the historical trend of the EV industry in New York State, together with the contributing factors. Second, we map the current distribution of EV registration and charging stations in New York State at county level. Finally, we employ sentiment analysis and LDA model analysis to explore customer's needs and expectation about EV.

Part 1-graph 1: Original EV Registration in New York State (2011-2019)

The first graph we created shows the historical trend of electric vehicle registration in New York State from 2011 to 2019. The data was downloaded from New York State Energy Research and Development Authority (NYSERDA). Since the original dataset separates “Year” and “Month Name” variables in two columns, we combined them together with a new variable “Time”. Then, an interactive area chart was created via highcharter package with year on the x axis and cumulative sum of EV registration on the y axis.

Part 1-graph 2: New York State Cumulative Population

The second graph is about the cumulative population growth in the New York State from 2010 - 2019. The data was collected from U.S. Census Bureau. The original dataset contained demographic info of New York State at different levels, including info about gender and race distribution. We select only the growth number of total populations from the dataset. The graph was created with ggplot2, with year on the x axis, and cumulative population on the y axis.

Part 1-graph 3: Transportation Sector Greenhouse Emissions Inventory, 1990-2016

The third graph focused on the air pollution and policy incents for EV. We found the data regarding the carbon emission in transportation sector from NYSERDA. We transformed the data about air pollution from the PDF to csv, and use ggplot2 to make this graph. The x axis is Year, and y axis is emission inventory. Also, we marked the different type of pollution sources with different colors.

Part 1-graph 4: EV make-models in New York State by popularity

The fourth graph displays the popularity of EV make-models in New York State by its registration numbers updated in 2021. The graph was created using the same data source as in graph 1. The purpose is to show the great variety of EV models available in the current U.S. market. Since the original dataset separates "Make" and "Model" variables in two columns, we combined them together with a new variable "EV_make" to display both carmakers and model names. Then an interactive treemap was created via highcharter package. A treemap was chosen because it can rank dozens of variables in the same graph by color density. Since we have 66 observations in the original dataset, we found a treemap most effective for visualization.

Part 1-graph 5: Features of EV by Make-models

The fifth graph plots the features of EVs by make-models in the current U.S. market. We collected the information of EV's features from Visual Capitalist. The dataset included information about EV's range, MSRP, brand, and type name. We use a scatter point plot to show the pros and cons of different types of EV, and the relative advantage of OEMs. As some type names overlapped with each other, we use ggploty to show the graph, so we could zoom in at certain region to get a clearer view of the type names.

Part 2-map 1 for EV charging stations in New York State

In this part, we mapped the current distribution of EV registration and charging stations in New York State, to examine whether charging stations would affect EV sales in the NY State. The assumption is that the number of charging stations is positively associated with EV sales. Both EV charging station and registration data are downloaded from New York State Energy Research and Development Authority (NYSERDA), and were updated in 2021. The data has already included longitude and latitude of each charging station, so there is no need for geocoding.

Utilizing the dplyr and leaflet packages, we created an interactive map with each circle representing one charging station. We made the popup text to display the important features (city

name, zipcode, EV connector type, access days & time) of each EV charging station. To make it easier to read, we added marker clustering so that zooming in will reveal the individual locations but the zoomed-out map only shows the clusters. This gives the audience a direct understanding of the distributions of EV charging stations in major NY districts.

Part 2-map 2 for EV charging stations in New York City

In the second map we were interested in exploring the distribution of EV charging stations in New York City. The shapefiles of borough boundaries in NYC is downloaded from NYC Department of City Planning. Then, we filtered the NY state charging station data into NYC data by borough name. We grouped by borough names in NYC, counted the number of charging stations in each borough, and store the values into the shapefile data. Finally, we used tm package to create a heatmap with color density to represent the number of charging stations in NYC.

Part 2-map 3 for original EV registration in New York State

The third map displays the distribution of EV registration in New York State by zipcode. It serves as a comparison of the first map to show the relationship between EV registration and charging stations. Firstly, the NY zipcode data was downloaded from Opendatasoft. The data contains longitude and latitude of each zipcode in NY state. Next, EV registration data is accessed from the same source as in part 1 graph 1. We joined the two datasets by zipcode and assign any NA values in terms of EV registration numbers into zero. Finally, we used leaflet to create an interactive map with popup content. The popup content displays key features of each zipcode: number of PHEV/EREV, number of BEV, and total number of EVs within that specific zipcode area. We further added marker clustering to make it coherent with the setting of the first map.

Part 3- LDA Analysis

Data Wrangling: In this part, our team decides to analyze tweets about electric vehicles on Twitter. Using Python Tweepy, we scaped all tweets containing words “electric car” from 2020-04-06 to 2020-04-14 (scraping date is limited by Twitter API). In total, we get 11,706 tweets.

Graph 1: Interactive Table for Electric Tweets: The interaction data table is provided to give users a straightforward impression on what our original dataset looks like and they can search for any texts based on key words.

Graph 2: Visualization of LDA results: Our first attempt is to have an overall knowledge on what these tweets are generally talking about. So, we manage to categorize all EV-related tweets into three topics and visualize these topics with LDAvis package. The major takeaways from the graph are the prevalence levels for different topics and the most common words within each topic.

Part 3 – Twitter Sentiment Analysis

We retrieved EV related twitter posts from April 6th - April 14th, 2021 to conduct a comprehensive sentiment analysis. The goal of this section is to highlight the following three patterns:

- Identify consumer needs and preferences: The LDA model and word cloud illustrated consumers have placed elevated focus on charging infrastructure, battery usage and the EV giant – Tesla.
- Uncover a positive sentiment pattern with regards to the 7-day window tweets in April 2021
- Gain forward-looking perspective based on emotion classification

Data Wrangling: Some basic text cleaning is performed by removing punctuations, lower casing letters and eliminating repeated key words such as "electric", "car" and "vehicle." Next, the text data is converted into a corpus and then vectorized into term document matrix. And finally, we created a dataframe containing the individual words and its respective frequencies.

Graph 1 - word cloud: Using the above tokenized dataframe, we applied the word cloud package to create a visualization of the most frequently mentioned terms with the larger sized words indicating a higher frequency.

Graph 2 – Frequency bar graph: for easier readability, we decided to highlight the frequencies of the top 17 words in a descending order.

Graph 3 – Total Tweets by Sentiment: Using the *plotly* package, we would like to examine if there is a positive sentiment or negative sentiment pattern in the midst of the pandemic. To achieve this end goal, we converted our stemmed text into *quanteda* corpus using the *quanteda* package. Then, the *Hu & Deng Dictionary* is utilized to create a function of sentiment calculations and a new column is added to the corpus dataframe to indicate the tone of each tweet text. Finally, we used the *ifelse* statement to categorize positive, neutral and negative sentiment and visualized the frequency distribution for each tweet.

Graph 4 – In-depth Sentiment Text Breakdown: Using the *Bing lexicon* and *get_sentiments* function from the *tidytext package*, we evaluated the emotion prevalent in the tokenized word list created in Graph 1. Then, the dataframe is divided into a set of negative and positive words: within each set, the top 20 words are identified and turned into a bar graph visualization. The goal of this

chart is to highlight what individuals are actually talking about on Twitter whether it's positive or negative comments related to the EV industry.

Graph 5 – Emotion Classification: To gain a forward-looking perspective on electric mobility, we utilized the *NRC Emotion lexicon* to the sentiment of each word and created a visual representation of the top emotions present in the twitter data set. The most frequent emotion helped to demonstrate a relatively optimistic view towards the future of EV business, exemplified by emotions such as "anticipation" and "trust."