DATA ANALYSIS FOR DATA SCIENCE PROJECT

Project team members: Sai Karthik Mantri, Syed Naseer Rizwan, Harshith Sai Kesa

Executive Summary:

In recent years, electric vehicles (EVs) have grown in popularity as a viable way to cut greenhouse gas emissions and slow down climate change. Understanding EV adoption and use trends is crucial as the number of EVs on the road rises. This project intends to study and display the population of electric vehicles in the US state of Washington between two well-known car manufacturers and determine why one brand is more popular than the other based on several data-provided parameters. In conclusion, this study will offer a thorough examination of the demographic statistics for electric vehicles to offer insights on the adoption and usage of EVs as of the time indicated by the data released by the Washington State Department of Licensing. EXCEL, PSPP, and R are utilized as analysis software. Data has been prepared, cleaned, and geographically mapped before analysis.

Introduction:

The initiative will gather information on EV sales, registrations, and the construction of charging stations using publicly accessible data sources, including EV manufacturers and government transportation organizations. The data will be examined statistically to estimate the geographic distribution of EVs in the state of Washington, detect trends in EV usage regarding car brands, and determine why one brand is more popular than another.
This dataset shows the Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) that are currently registered through Washington State Department of Licensing (DOL).

By using Excel, we understood the data like what quantitative data are available which can be used for analysis and mapping. We also gathered information on trends of Electric Vehicles utilized in the state of Washington depending on car brands. Secondly. We utilized PSPP software for statistical analysis and descriptive statistics, for better understanding we have run a sample test by structuring the data. Thirdly, we used R Programming to plot Geographical usage of electrical vehicle by the owners in the state of Washington.

Background and Questions:

The major goal of data analysis is to try to understand why one brand in the electric car market is so popular compared to others. This research aims to provide an answer to the issue of why one brand is more well-known than others given the information presented in the data. A sizable dataset with 17 columns and more than 13,000 rows is the subject of the discussion. To make the data's overall analysis and comprehension simpler, we trimmed it to 100 records.

Data Selection:

The dataset was gathered via the data.gov website, which is run by the government. The dataset may be found online at https://catalog.data.gov/dataset/electric-

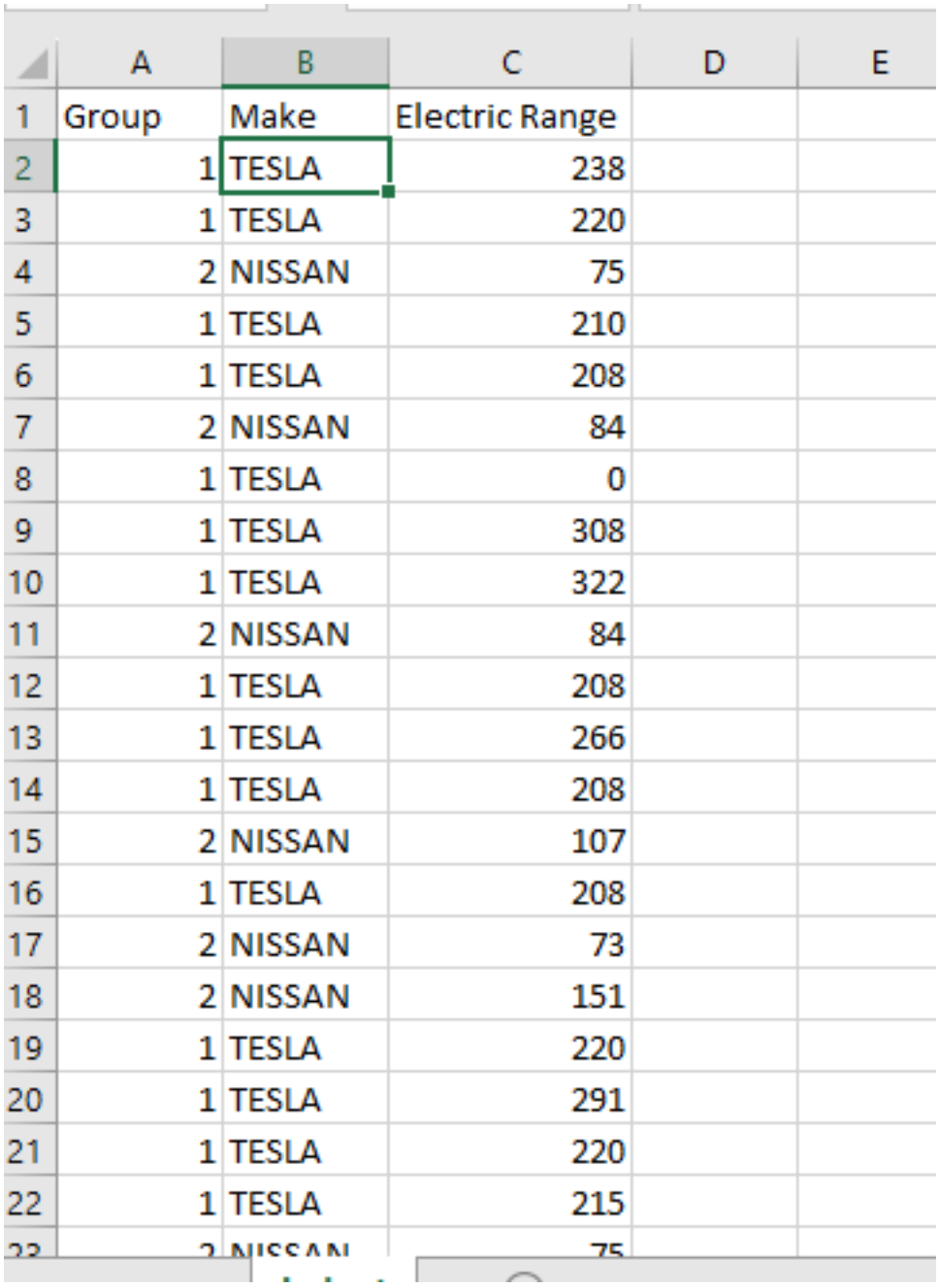
vehicle-population-data. The dataset contains data on car serial numbers, counties they are from, states they are from, cities they are in, electrical range, or range of electrical vehicles, which is another way of saying miles, and other similar data on electrical vehicles. The manufacturer of the automobiles included in the dataset's make column is one of the crucial pieces of information used for this investigation.

Data Structures:

We initially imported the data as comma separated values (csv), then we trimmed the records from more than 10,000 to 100 for simpler comprehension and analysis using Excel. The dataset contains information on the car's number, county, city, state, postal code, model year, manufacturer, and kind of electric vehicle, as well as information about CAFV, electric range, base MSRP, and others. We grouped the data for analysis and utilized an independent two-sample t-test using PSPP software. The vehicle position is supplied in a different format for R programming, namely POINT (-120.56754 46.98765). By dividing and removing the undesired sections, we were able to derive longitude and latitude.

Data Cleaning:

To get the appropriate output for computing popular automobile brands, we first shortened the records and formatted the data in Excel. To do an independent two sample test, we must arrange the data according to the [Make] and [Electric Range] Columns, as shown in the image below. Secondly, we utilized PSPP to extract descriptive and statistical analysis reports.



	A	B	C	D	E
1	Group	Make	Electric Range		
2	1	TESLA	238		
3	1	TESLA	220		
4	2	NISSAN	75		
5	1	TESLA	210		
6	1	TESLA	208		
7	2	NISSAN	84		
8	1	TESLA	0		
9	1	TESLA	308		
10	1	TESLA	322		
11	2	NISSAN	84		
12	1	TESLA	208		
13	1	TESLA	266		
14	1	TESLA	208		
15	2	NISSAN	107		
16	1	TESLA	208		
17	2	NISSAN	73		
18	2	NISSAN	151		
19	1	TESLA	220		
20	1	TESLA	291		
21	1	TESLA	220		
22	1	TESLA	215		
23	2	NISSAN	75		

Grouping of data.
For Mapping the data geographically, we need to format the data to use in R programming, We had longitude and latitude given as a single string format later we have split the single string into 2 separate columns and feed the data into R and plotted the map.

Data Analysis:

We calculated the most popular vehicle brand from the dataset column name [make]. We used =countif(G2:G100,”TESLA”) formula on a separate column to get the count of each vehicle brand. Then according to the calculated information we concluded that tesla EV brand is the most popular one. To find its competitor, we have checked for the maximum value from the new column with count of number of

vehicles with respect to brands excluding the most popular one. We used =max(U2:U35) formula to get second most popular brand i.e. “NISSAN”. We also want to determine if all the vehicles of the top two brands have how many number of Battery Electric Vehicles(BEV) and Plug-in Hybrid Electric vehicles(PHEV), we achieved It using the formula =countif(I:I,” Battery Electric Vehicles(BEV)”,G:G,”Tesla”) for tesla vehicles and the same for Nissan but using nissan as string in place tesla in the above formula. Next, we want to find the average electric range for both the brands. So, we used the formula =averageifs(K:K,G:G,”tesla”) and similarly for nissan. We also calculated the number of tesla and nissan cars by the model year of the car by using =countifs(G:G,”tesla”,F:F,W3) and similarly for Nissan. Electric range by brand by year is calculated to better understand rise of electric range in both the brands. We used =averageifs(K2:K100,G2:G100,”TESLA”,F2:F100,W3) and similarly for Nissan.

cars	count	max two
TESLA	41	41
HONDA	2	18
NISSAN	18	1
FORD	7	
AUDI	1	
KIA	6	
CHEVROLET	13	
SMART	1	
BMW	3	
TOYOTA	2	
JEEP	1	
FIAT	1	
VOLVO	3	

	TESLA	NISSAN
Battery Electric Vehicle (BEV)	41	18
Plug-in Hybrid Electric Vehicle (PHEV)	0	0

ELECTRIC RANGE	
231.268293	TESLA
100.666667	NISSAN

TESLA COUNT BY MODEL YEAR	model year	NISSAN COUNT BY MODEL YEAR	ER BY BRAND	
9	2018	2	231.268293	TESLA
0	2021	0		
6	2019	1		
4	2013	6		
4	2017	2		
2	2016	0		
0	2023	0	100.666667	NISSAN
1	2014	1		
1	2015	2		
2	2022	0		
12	2020	2		
	2012			

ER BY BRAND BY YEAR		
	TESLA	NISSAN
	222.666667	151
	#DIV/0!	#DIV/0!
	220	150
	208	75
	207.5	107
	205	#DIV/0!
	#DIV/0!	#DIV/0!
	208	84
	208	84
	0	#DIV/0!
	305.833333	149
	#DIV/0!	73

#DIV/0! Represents that average of selected cells is not possible or infinite. In our case, as number of 2021 car model for Nissan and tesla are both zero the average was not possible.

PSPP: We Imported the formatted data as shown in figure below.

Case	Group	Make	Electric_Range
1	1	TESLA	238
2	1	TESLA	220
3	2	NISSAN	75
4	1	TESLA	210
5	1	TESLA	208
6	2	NISSAN	84
7	1	TESLA	0
8	1	TESLA	308
9	1	TESLA	322
10	2	NISSAN	84
11	1	TESLA	208
12	1	TESLA	266
13	1	TESLA	208
14	2	NISSAN	107
15	1	TESLA	208

We performed Independent Sample T test in PSPP for the above data. Where group 1 represents Tesla and group 2 represents Nissan. The results of the test determined a huge difference in average electrical range as shown below.

```
T-TEST /VARIABLES= Electric_Range
        /GROUPS=Group(1,2)       /MISSING=ANALYSIS
        /CRITERIA=CI(0.95).
```

Group Statistics

	Group	N	Mean	Std. Deviation	S.E. Mean
Electric_Range	1	41	231.27	68.68	10.73
	2	18	100.67	33.05	7.79

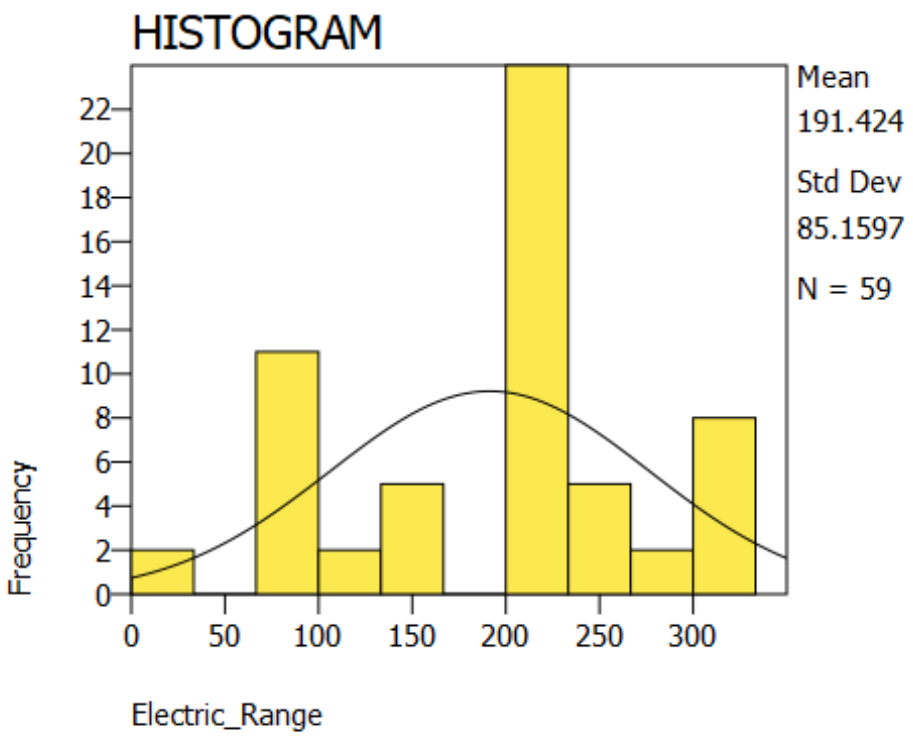
Descriptive and Statistical analysis reports are pasted below.

```
DESCRIPTIVES
    /VARIABLES= Electric_Range
    /STATISTICS=DEFAULT VARIANCE KURTOSIS SKEWNESS.
```

Descriptive Statistics

	N	Mean	Std Dev	Variance	Kurtosis	S.E. Kurt	Skewness	S.E. Skew	Minimum	Maximum
Electric_Range	59	191.42	85.16	7252.18	-.56	.61	-.34	.31	0	330
Valid N (listwise)	59									
Missing N (listwise)	0									

```
GRAPH /HISTOGRAM (NORMAL) = Electric_Range.
```



As the histogram and Skewness values suggest the data is normally distributed.

The results of the Independent Sample T test are as follows.

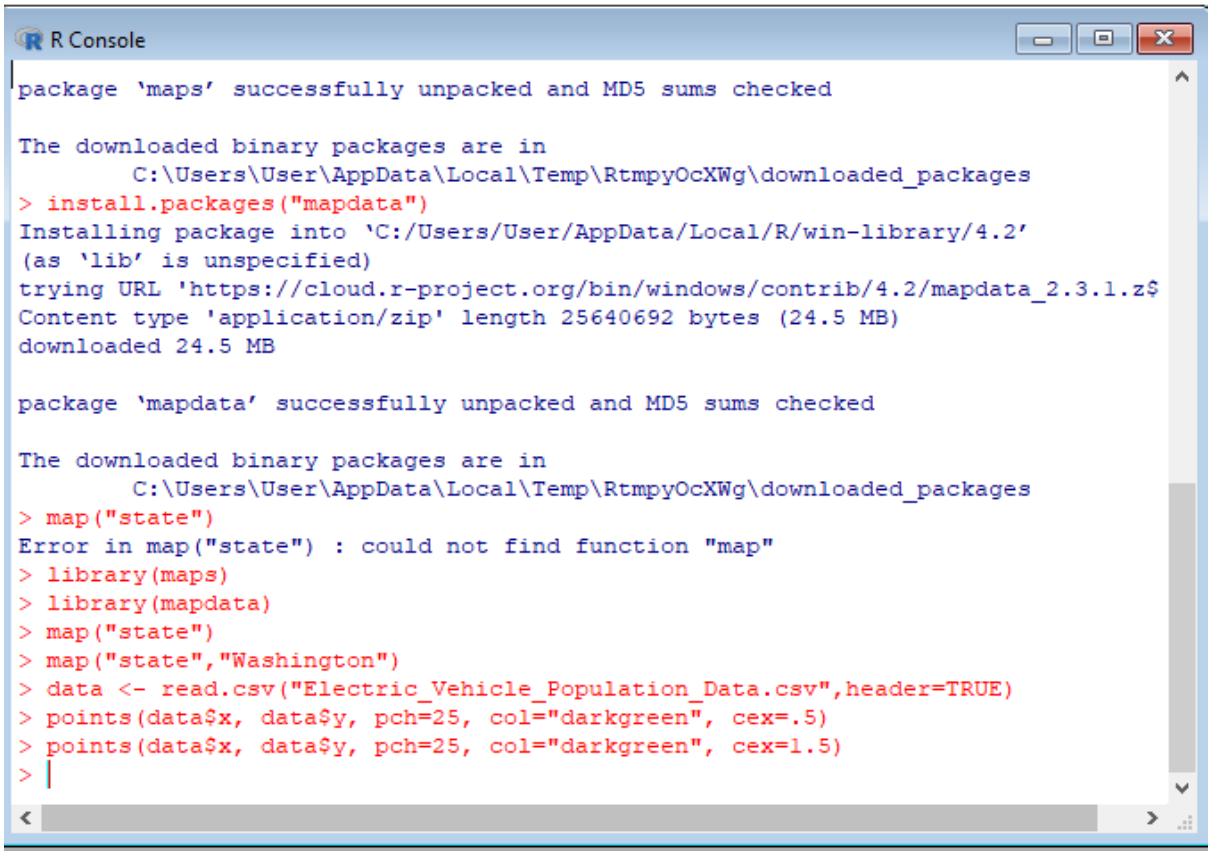
Independent Samples Test

		Levene's Test for Equality of Variances		T-Test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Electric_Range	Equal variances assumed	1.61	.209	7.66	57.00	.000	130.60	17.05	96.46	164.74
	Equal variances not assumed			9.85	56.40	.000	130.60	13.26	104.05	157.15

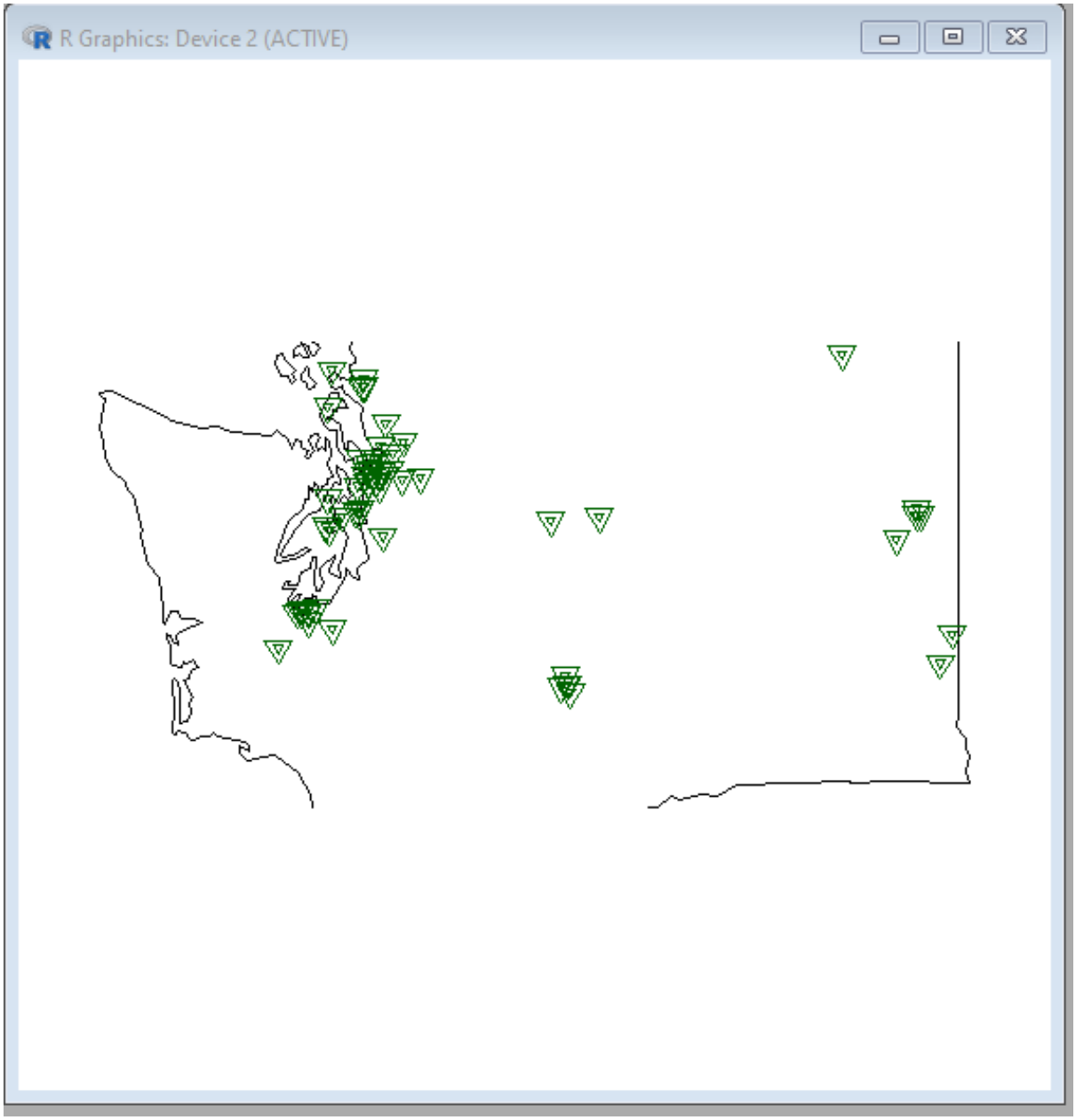
Null Hypothesis: Mean of Group 1 i.e. mean of electric charge of Tesla is equal to Mean of group 2 i.e. mean of electric charge if Nissan.
Alternate Hypothesis: Mean of Group 1 i.e. mean of electric charge of Tesla is Greater than or not equal to Mean of group 2 i.e. mean of electric charge if Nissan
Null Hypothesis is accepted when sig value is <0.05 but in over case it is >0.05 hence Null hypothesis is rejected Where Null hypothesis is mean of group 1 is equal to mean of group 2 Here the means are not equal.
In this case, we are accepting the Alternate Hypothesis which indicate that mean of electric charge of tesla is greater than mean of Electric charge of Nissan. Which also indicates the popularity of Tesla.

R programming is used for geographical mapping of electrical vehicles in the state of Washington. First, the needs the formatted data as shown in the figure below is imported into R.
The formatted data with x and y co-ordinates as shown below:

Electric Range	Base MSRI	Legislative	DOL Vehic	Vehicle Lc	Electric Ut	2020 Cens	x	y
238	0	14	1.41E+08	POINT (-1:	PACIFICOI	5.31E+10	-120.569	46.58514
47	0	23	1.72E+08	POINT (-1:	PUGET SO	5.3E+10	-122.647	47.73689
220	0	36	9426525	POINT (-1:	CITY OF SE	5.3E+10	-122.401	47.65908
75	0	36	2.12E+08	POINT (-1:	CITY OF SE	5.3E+10	-122.368	47.64586
210	0	22	1.86E+08	POINT (-1:	PUGET SO	5.31E+10	-122.754	47.06316
47	0	22	1.54E+08	POINT (-1:	PUGET SO	5.31E+10	-122.877	47.05997
19	0	22	3.48E+08	POINT (-1:	PUGET SO	5.31E+10	-122.892	47.03956
23	0	22	2.27E+08	POINT (-1:	PUGET SO	5.31E+10	-122.754	47.06316
19	0	20	2.3E+08	POINT (-1:	PUGET SO	5.31E+10	-123.087	46.82175
208	69900	26	1.65E+08	POINT (-1:	PUGET SO	5.3E+10	-122.661	47.56573
84	0	22	1.93E+08	POINT (-1:	PUGET SO	5.31E+10	-122.823	47.04437
0	0	14	1.87E+08	POINT (-1:	PACIFICOI	5.31E+10	-120.569	46.58514
308	0	22	2135486	POINT (-1:	PUGET SO	5.31E+10	-122.823	47.04437
322	0	22	1.25E+08	POINT (-1:	PUGET SO	5.31E+10	-122.754	47.06316
20	0	22	1.4E+08	POINT (-1:	PUGET SO	5.31E+10	-122.818	46.98876
26	0	23	1.58E+08	POINT (-1:	PUGET SO	5.3E+10	-122.521	47.62728
84	0	14	1.24E+08	POINT (-1:	PACIFICOI	5.31E+10	-120.523	46.60138
208	69900	22	2.35E+08	POINT (-1:	PUGET SO	5.31E+10	-122.923	47.03779
38	0	21	2.24E+08	POINT (-1:	PUGET SO	5.31E+10	-122.255	47.90456
238	0	12	1.34E+08	POINT (-1:	PUD NO 1	5.3E+10	-120.658	47.5982



Required packages are installed and longitude and latitude data are mapped into R for plotting of the map. After performing the above programming, the geographical map is as follows:



Discussion and Conclusions:

According to our analysis, we concluded that Tesla is the most popular vehicle in Washington since it gives a better electric range than its competitor Nissan. Due to inadequate quantitative data provided in the dataset the analytical possibilities were not extensive enough. Data such as cost per model, lifespan of a car, maintenance cost would further help in better analyses of the data and more evident results. In conclusion, in comparison to other also popular cars tesla outperforms regarding electric charge whereas on the other side of spectrum the least electric charge in Washington state is 3 which evidently tells that tesla is far better in electric charge department with 213 average electric charge.