ITM 891

Large Scale Data Analysis for MSBA

Spring 2020

Project Report
Highway Selection for Electric Vehicle Charging Station

Made by:

Karan Pranav Dalal dalalkar@msu.edu

MICHIGAN STATE UNIVERSITY

Broad College of Business

Executive Summary

Provide an analytical solution to an Electric Vehicle Charging establishment company to decide which interstate highway between any two major US cities are best to set up an electric vehicle charging station.

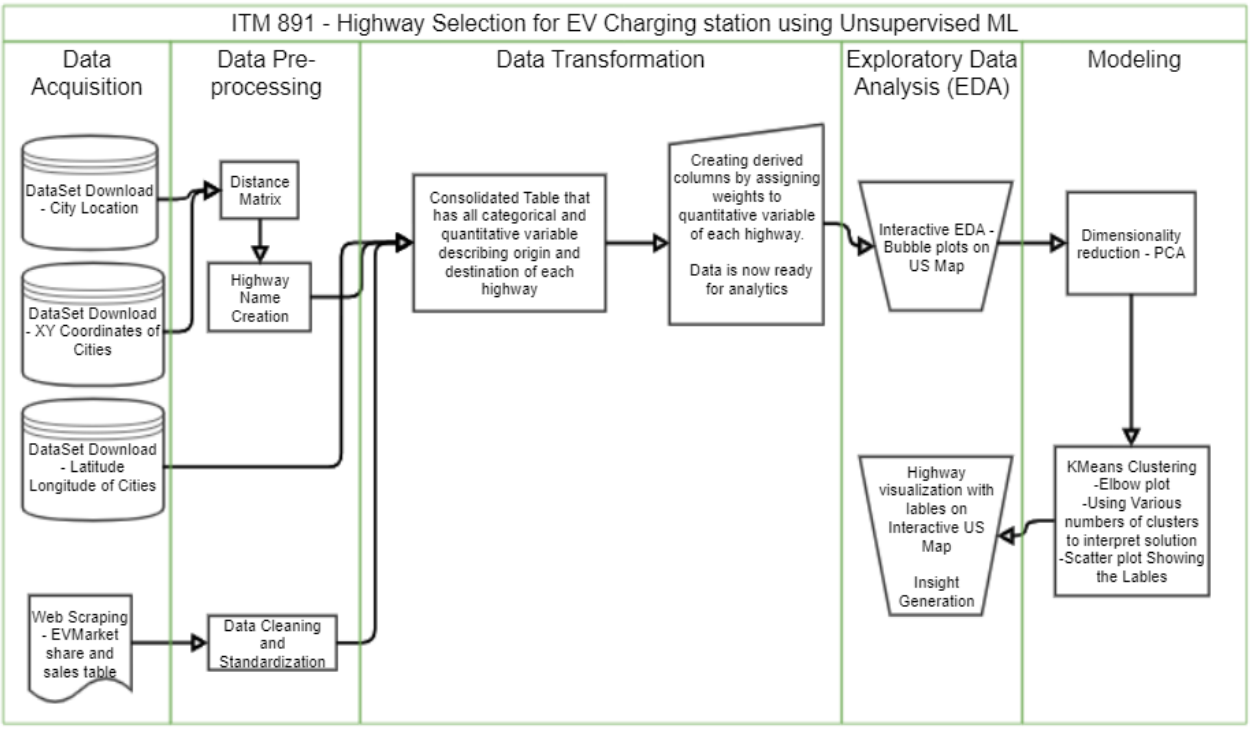
Business problem: With the budget on hand, A startup company is looking to open not more 20 charging stations. They want to make an analytics drive decision.

For carrying out the analysis, 48 capitals of the US have been chosen. Highways have been chosen assuming that each state capital has direct road connectivity with 2 nearest state capitals. Each highway represents a potential site for setting up the EV charging station. The project aims to identify a subset of highways out of the 96 highways identified among 48 US State capitals. The parameters/variables used to find the subset of highways is:

- **Distance:** Distance between two cities
- **EV_Sale_2018:** EV Sales in $ for that State in 2018
- **EV_Sale_2017**: EV Sales in $ for that State in 2017
- **EV_2018_Percent_Marketshare:** Market share of EV vs Non-EV for that State in 2018
- **EV_2017_Percent_Marketshare:** Market share of EV vs Non-EV for that State in 2017

To get the right subset of highways, KMeans Clustering algorithm has been used, as this is an unsupervised learning problem. Before applying clustering, the cleaned data is reduced to two dimensions using PCA.

Analysis Pipeline



Methodology

i. Data Acquisition

Four data sources are involved in this process:

A. Data: Electric Vehicle market share within state - Source: Website
The data is web scrapped from a table residing on a website. This table has details of EV Market Share and Sales by US State. I have assumed that the 50 state capital follows the same trend (proportionally) as it state's number. Library used: Beautifulsoup
https://evadoption.com/ev-market-share/ev-market-share-state/

B. Data: The name and state of each capital - Source: USCAP dataset
The USCAP dataset has Names of each state and their capital cities in a txt file format. The file was downloaded and changed to tab delimiter for proper loading into a python dataframe.
https://people.sc.fsu.edu/~jburkardt/datasets/cities/uscap_name.txt

C. Data: The (X,Y) coordinates of each capital - Source: USCAP dataset
The USCAP dataset has (X,Y) coordinates of each capital cities in a txt file format. The coordinates were calculated using a cylindrical projection. The file was downloaded and changed to tab delimiter for proper loading into a python dataframe.
https://people.sc.fsu.edu/~jburkardt/datasets/cities/uscap_xy.txt

D. Data: Latitude and longitude of each capital - Source: USCAP dataset
The USCAP dataset has Latitude and longitude of each capital cities in a txt file format. The coordinates were calculated using a cylindrical projection. The file was downloaded and changed to tab delimiter for proper loading into a python dataframe.
https://people.sc.fsu.edu/~jburkardt/datasets/cities/uscap_ll.txt

ii. Data Pre-processing

Data cleaning was carried out on the captured data, along with standardizing data types, renaming columns and removing rows not suitable for analysis.
Also, Distance matrix was calculated showing distance of each city select with one another. Using this, highways were selected for each state capital.

[19]:

	Montgomery	Phoenix	Little Rock	Sacramento	Denver	Hartford	Dover	Tallahassee	Atlanta	Boise	...
Montgomery	0.000000	1783.885334	449.200825	2468.842413	1389.335431	1142.651646	879.060131	191.390397	162.403666	2211.197020	...
Phoenix	1783.885334	0.000000	1367.034403	738.873153	654.903530	2782.170637	2555.917103	1932.060682	1912.954737	759.015951	...
Little Rock	449.200825	1367.034403	0.000000	2030.520388	940.135677	1442.295418	1200.705829	630.595840	552.825830	1762.056270	...
Sacramento	2468.842413	738.873153	2030.520388	0.000000	1141.957372	3378.605004	3174.681299	2630.381930	2583.332133	502.791251	...
Denver	1389.335431	654.903530	940.135677	1141.957372	0.000000	2236.678176	2035.770234	1568.360718	1481.727742	822.363861	...
Hartford	1142.651646	2782.170637	1442.295418	3378.605004	2236.678176	0.000000	266.780324	1119.468412	980.346523	3012.555759	...
Dover	879.060131	2555.917103	1200.705829	3174.681299	2035.770234	266.780324	0.000000	852.873930	717.191246	2829.714872	...
Tallahassee	191.390397	1932.060682	630.595840	2630.381930	1568.360718	1119.468412	852.873930	0.000000	228.725753	2388.542561	...
Atlanta	162.403666	1912.954737	552.825830	2583.332133	1481.727742	980.346523	717.191246	228.725753	0.000000	2303.456159	...
Boise	2211.197020	759.015951	1762.056270	502.791251	822.363861	3012.555759	2829.714872	2388.542561	2303.456159	0.000000	...
Springfield	563.234235	1610.005587	394.880711	2200.159864	1059.503075	1180.769220	976.824642	744.165681	552.553955	1856.033552	

Fig 1: Snapshot of distance matrix

[20]:

	HighwayName	From	To	Distance
0	MontgomeryAtlanta	Montgomery	Atlanta	162.403666
1	MontgomeryTallahassee	Montgomery	Tallahassee	191.390397
2	PhoenixSanta Fe	Phoenix	Santa Fe	449.101232
3	PhoenixSalt Lake City	Phoenix	Salt Lake City	504.984988
4	Little RockJackson	Little Rock	Jackson	222.279732
...	...	...	...	...
91	CharlestonFrankfort	Charleston	Frankfort	223.414006
92	MadisonSpringfield	Madison	Springfield	228.167510
93	MadisonSaint Paul	Madison	Saint Paul	287.204287
94	CheyenneDenver	Cheyenne	Denver	97.986275
95	CheyennePierre	Cheyenne	Pierre	380.504080

96 rows × 4 columns

Fig 2: Highway Identification (Highway name = City1City2)

iii. **<u>Data Transformation</u>**

The following were performed for make a consolidated table having all the highway related information:

A. Joined Selected highways with quantitative variable obtained through web scraping.

B. Joining with latitude/longitude table for obtaining that information for each highway

```
2 merged_inner.dtypes

[38]: HighwayName                                    object
From                                           object
To                                             object
State_From                                     object
State_To                                       object
Distance                                      float64
EV_Sale_2018_To                               float64
YoY_Sale_Percent_Increase_2018_2017_To        float64
EV_2018_Percent_Marketshare_To                float64
YoY_Share_Percent_Increase_2018_2017_To       float64
EV_Sale_2017_To                               float64
EV_2017_Percent_Marketshare_To                float64
EV_Sale_2018_From                             float64
YoY_Sale_Percent_Increase_2018_2017_From      float64
EV_2018_Percent_Marketshare_From              float64
YoY_Share_Percent_Increase_2018_2017_From     float64
EV_Sale_2017_From                             float64
EV_2017_Percent_Marketshare_From              float64
lat_From                                      float64
lon_From                                      float64
lat_To                                        float64
lon_To                                        float64
Lables                                          int32
text                                           object
dtype: object
```

Fig 3: Summary columns in Consolidated table

This, consolidated table is used to create a subset table which has only those variables used for doing analytics. So, 5 quantitative variables are selected, and highway names are kept as index.

Out[32]:

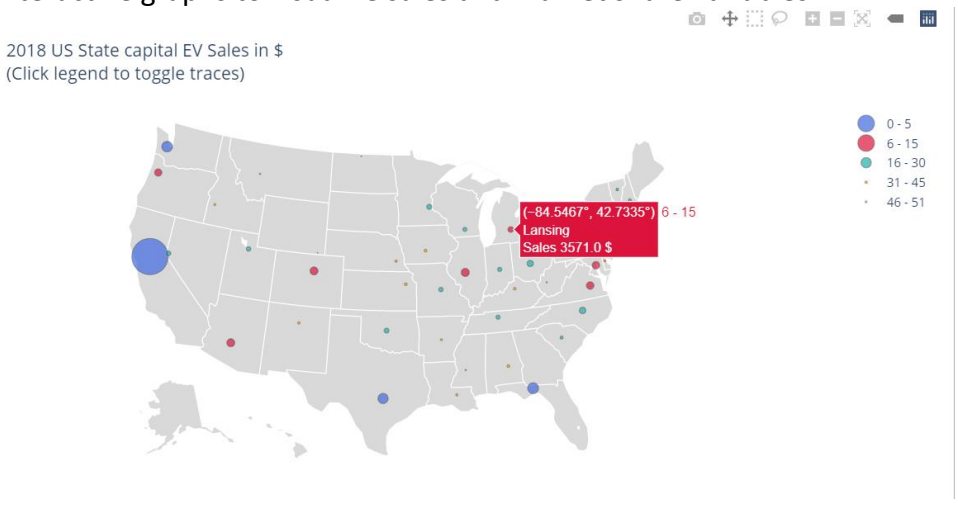
HighwayName	Distance	EV_Sale_2018	EV_Sale_2017	EV_2018_Percent_Marketshare	EV_2017_Percent_Marketshare
MontgomeryAtlanta	162.403666	6870.0	2808.0	1.59	0.72
MontgomeryTallahassee	191.390397	14571.0	6954.0	1.44	0.71
PhoenixSanta Fe	449.101232	7791.0	3345.0	2.65	1.37
PhoenixSalt Lake City	504.984988	9381.0	4139.0	3.44	1.84
Little RockJackson	222.279732	666.0	315.0	0.57	0.26
...	...	...	...	...	...
CharlestonFrankfort	223.414006	1005.0	473.0	0.80	0.42
MadisonSpringfield	228.167510	9313.0	5388.0	1.99	1.29
MadisonSaint Paul	287.204287	4809.0	2974.0	1.93	1.28
CheyenneDenver	97.986275	7143.0	4207.0	2.96	1.82
CheyennePierre	380.504080	227.0	130.0	0.70	0.47

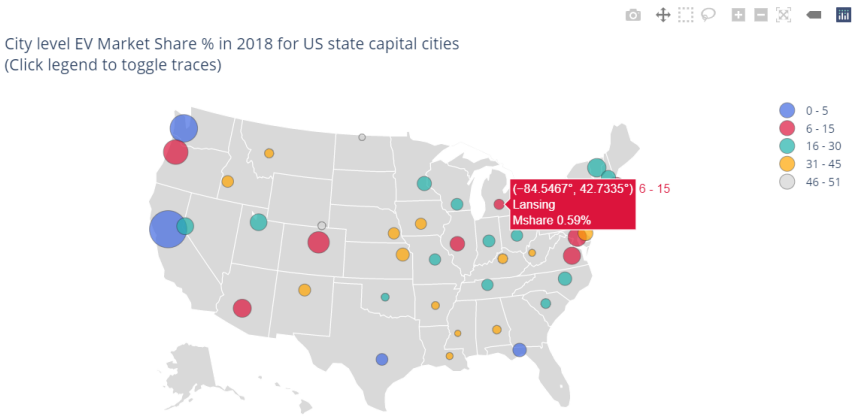
96 rows × 5 columns

Fig 4: Final table Ready for Analytics and EDA

iv. Exploratory Data Analysis

Used Interactive graphs to visualize sales and market share variables.





v. Modeling

Step 1: Dimensionality reduction Using Principal Component Analysis

	0	1
0	-7935.588337	683.238920
1	794.040224	1220.628402
2	-6869.789355	716.582337
3	-5100.090519	880.935333
4	-14521.658804	-468.199438
..	...	...
91	-14150.299242	-423.722378
92	-4499.462987	-220.076809
93	-9599.336700	-542.486492
94	-6966.059391	-362.427758
95	-14992.181237	-540.175104

[96 rows x 2 columns]

Fig 5: Reduced 5D to 2D variables using PCA

Step 2: KMeans clustering (with various K clusters)

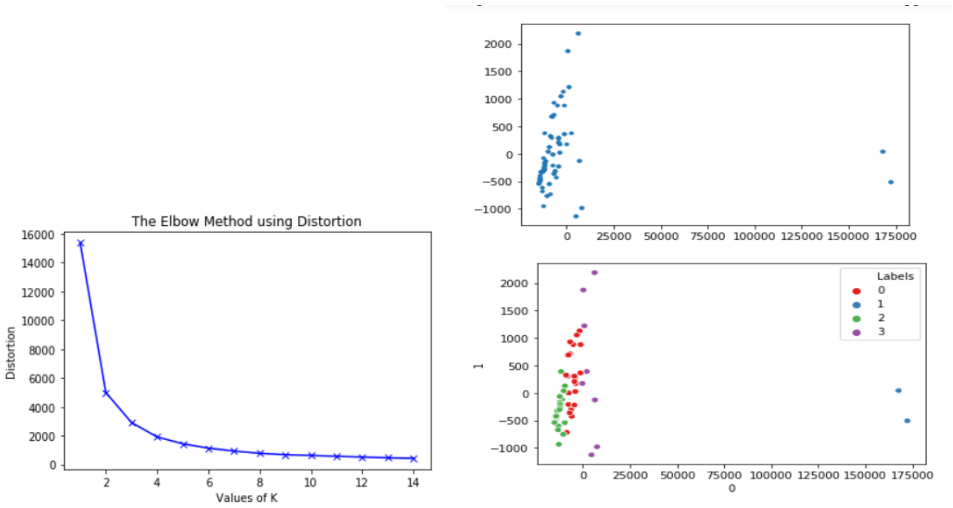


Fig 6: KMeans Clustering when K=4

vi. Insights

As the analytics is in a unsupervised learning form, we can have multiple interpretations for various number of cluster.

Business problem: With the budget on hand, A startup company is looking to open not more 20 charging stations. They want to make an analytics drive decision.

Solution: According to the model created, the following will be the data that be reviewed by a Business Analyst and suggested to the startup.

	Distance		EV_Sale_2018		EV_Sale_2017		EV_2018_Percent_Marketshare		EV_2017_Percent_Marketshare	
	mean	len	mean	len	mean	len	mean	len	mean	len
Labels										
0	246.906104	43.0	2329.023256	43.0	1198.348837	43.0	1.506279	43.0	0.888140	43.0
1	289.593266	4.0	157592.500000	4.0	97401.000000	4.0	10.355000	4.0	6.595000	4.0
2	181.212369	35.0	8773.885714	35.0	4492.571429	35.0	2.556000	35.0	1.450857	35.0
3	213.706245	14.0	16759.428571	14.0	9188.428571	14.0	3.330000	14.0	2.127143	14.0

Fig 7: Highways details (Mean distance, sales numbers) for each labels.

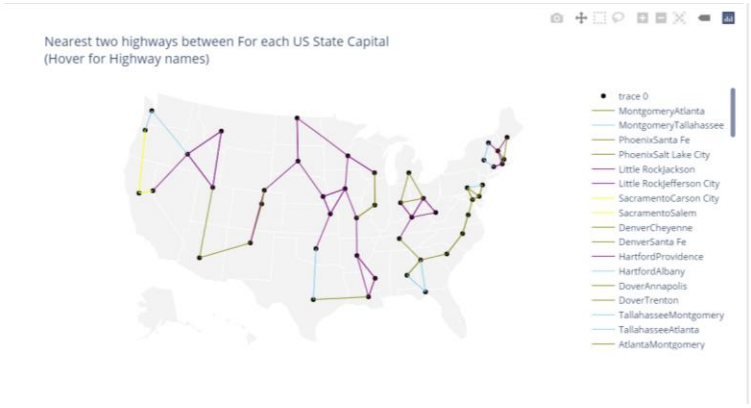


Fig 8: All US Highways(All labels)

Fig 9: Highways with labels ‘1’ and ‘3’ (Suggested by Analysis)

```
]:  1  qnt_analytics_table_1.loc[qnt_analytics_table_1['Labels'].isin([1,3])]
```

:[23]:

HighwayName	Distance	EV_Sale_2018	EV_Sale_2017	EV_2018_Percent_Marketshare	EV_2017_Percent_Marketshare	Labels
MontgomeryTallahassee	191.390397	14571.0	6954.0	1.44	0.71	3
SacramentoCarson City	125.667416	155767.0	95941.0	9.46	5.81	1
SacramentoSalem	453.519115	159418.0	98861.0	11.25	7.38	1
HartfordAlbany	98.123344	19167.0	12394.0	3.58	2.42	3
TallahasseeMontgomery	191.390397	14571.0	6954.0	1.44	0.71	3
TallahasseeAtlanta	228.725753	19709.0	9000.0	2.21	1.05	3
AtlantaTallahassee	228.725753	19709.0	9000.0	2.21	1.05	3
Carson CitySacramento	125.667416	155767.0	95941.0	9.46	5.81	1
TrentonHarrisburg	146.484161	15293.0	8379.0	2.51	1.46	3
AlbanyHartford	98.123344	19167.0	12394.0	3.58	2.42	3
AlbanyMontpelier	138.943878	16576.0	10961.0	3.48	3.16	3
Oklahoma CityAustin	360.684330	14447.0	6110.0	1.13	0.49	3
SalemOlympia	146.185005	18626.0	11056.0	7.69	4.87	3
SalemSacramento	453.519115	159418.0	98861.0	11.25	7.38	1
AustinOklahoma City	360.684330	14447.0	6110.0	1.13	0.49	3
MontpelierAlbany	138.943878	16576.0	10961.0	3.48	3.16	3
OlympiaSalem	146.185005	18626.0	11056.0	7.69	4.87	3
OlympiaBoise	517.297862	13147.0	7309.0	5.05	2.92	3

Fig 10: Final Highway suggestions