Term Project – Targeted Advertising for New Vehicle Sales

Michelle Helfman

Bellevue University

DSC650-T301 Big Data (2245-1)

Nasheb Ismaily

May 31, 2024

Targeted Advertising for New Vehicle Sales

Introduction and Problem Statement

Selling automobiles, including trucks, is a business with many different parts: inventory, advertising, sales commissions, and the dealership itself. After deducting these costs, the profit margin after selling a new vehicle is very low. Some costs, such as the facility and commissions, are fixed, but inventory and advertising can be fine-tuned to create additional profits.

Advertising is one such area. By analyzing all the vehicle sales for the region, not just one dealership or brand, zip codes are used to target advertising for a particular type of vehicle. This targeted approach ensures that money is spent in locations where a given vehicle type is already sold, areas where new customers can be reached, and less in areas that would not attract business for that vehicle type but would be prime targets for another type of vehicle, minimizing wasted spending.

In the days before cable/streaming television, websites, and mailers, managers would examine sales data for their dealership and direct competitors for a short period by hand, then use their gut and advertising services to decide where to place their ads. However, dealerships are generational, and adopting concepts different from how Dad did it is difficult even though what worked in years past may not be suitable for the business now.

Utilizing big data in advertising eliminates guesswork, allowing for a comprehensive analysis of all regional sales over a longer period. This information, combined with demographics and viewing trends, tailors advertising to only when the targeted group is watching, maximizing the impact of the ads.

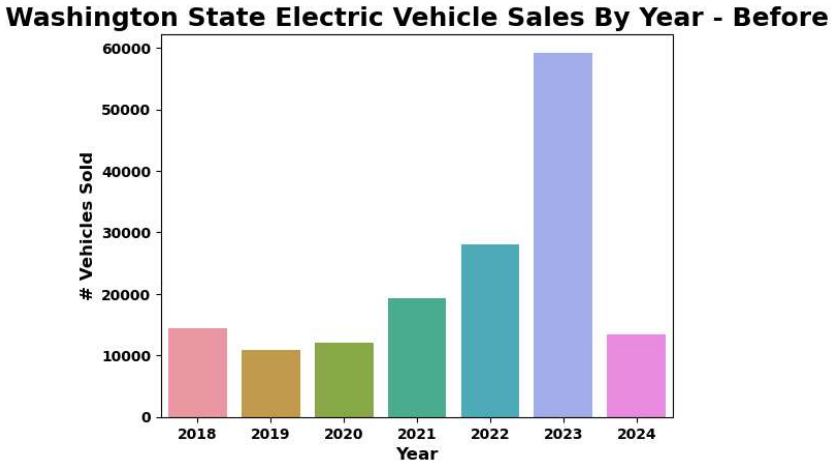
Data and Technology

Everything starts with the data. Electric vehicle sales in Washington state give a complete picture of a vehicle type while keeping the size manageable for the available resources. Using Jupyter Notebook, the original data consists of 157240 rows and 13 columns for 2018 through 2024, with information about make, model, city, zip code, and other electric vehicle-related information. The majority of base prices are 0 and the vin is extra information.

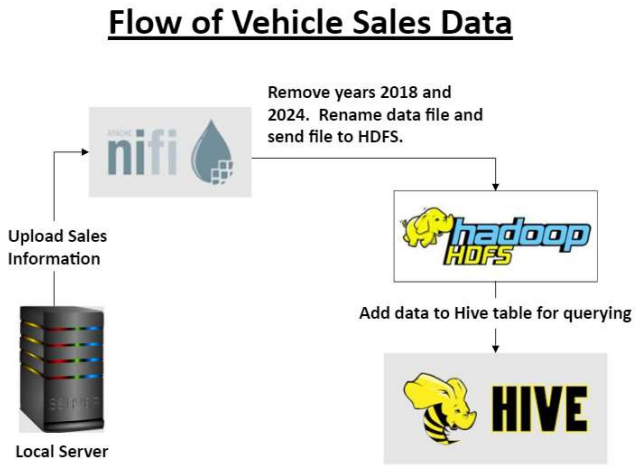
```
RangeIndex: 157240 entries, 0 to 157239
Data columns (total 13 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   vin                  157240 non-null  object
 1   county               157240 non-null  object
 2   city                 157240 non-null  object
 3   state                157240 non-null  object
 4   zipcode              157240 non-null  int64
 5   year                 157240 non-null  int64
 6   make                 157240 non-null  object
 7   model                157240 non-null  object
 8   ev_type              157240 non-null  object
 9   cafv_eligibility     157240 non-null  object
 10  electric_range       157240 non-null  int64
 11  base_msrp            157240 non-null  int64
 12  legislative_district 157240 non-null  int64
dtypes: int64(5), object(8)
```

	vin	county	city	state	zipcode	year	make	model	ev_type	cafv_eligibility	electric_range	base_msrp	legislative_district
0	WBY8P6C58K	King	Seattle	WA	98115	2019	BMW	I3	BEV	CAFV Elegible	153	0	43
1	5YJSA1E26J	King	Kent	WA	98042	2018	TESLA	MODEL S	BEV	CAFV Elegible	249	0	47
2	5YJXCDE23J	King	Bellevue	WA	98004	2018	TESLA	MODEL X	BEV	CAFV Elegible	238	0	41
3	WBY33AW0XP	King	Seattle	WA	98109	2023	BMW	I4	BEV	Unknown/Not Researched	0	0	36
4	5YJ3E1EB5L	King	Bothell	WA	98011	2020	TESLA	MODEL 3	BEV	CAFV Elegible	322	0	1

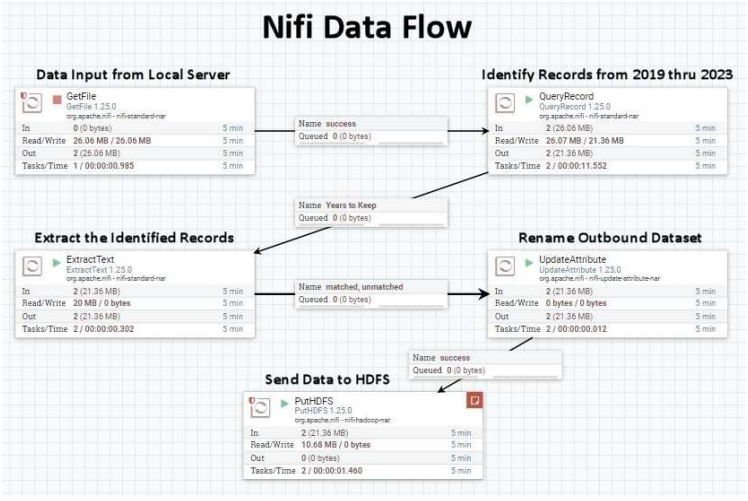
This project will examine the relationship between year, city, zip code, make, and vehicle sales.



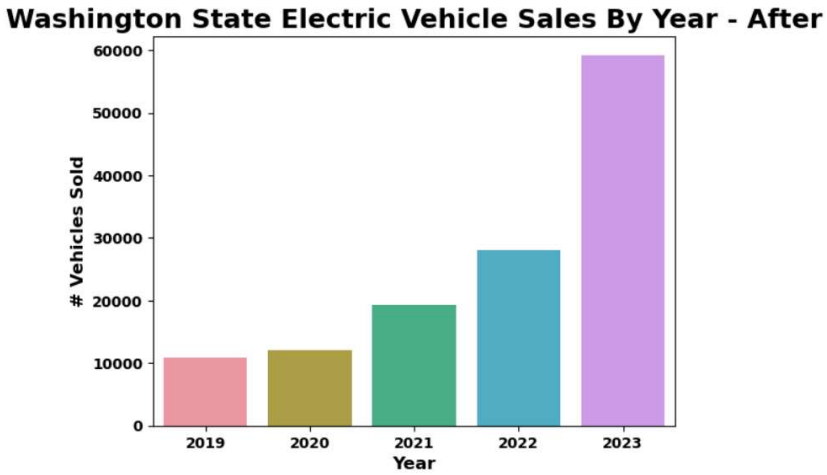
The information spans over six years and shows electric vehicle sales are rising annually. Thirty-seven different vehicle brands have some form of electric vehicle, with Tesla selling the majority. Seattle is the largest city in Washington state, with 35 zip codes. For this experiment, we will use Tesla and the city of Seattle for the ending queries, and the years 2018 and 2024 will be discarded during the following steps, giving a complete five-year sales history.



The various pieces are integrated and work like a chain, requiring very little human intervention. The data flow starts by uploading a file from the local server/computer to Apache NiFi, where the outlying years are removed from the final output file. The information is then sent to Apache Hadoops' HDFS and loaded into Apache Hive for evaluation.



Apache Nifi is an ETL (Extract, Transform, and Load) tool for automating data flow between systems and locations. In this case, the vehicle sales data is uploaded into NiFi from a Docker instance on a local server/computer. The outlying years, 2018 and 2024, are discarded so that there are five full years of history.



After excluding years 2018 and 2024 from the final dataset, there are still 13 columns but only 129445 rows.

```
Found 3 items
-rw-r--r--   1 mhelfman supergroup   11196868 2024-05-31 22:57 /tmp/Electric_Vehicle_Sales_Data.csv
drwx-wx-wx   - root     supergroup          0 2024-05-31 22:55 /tmp/hive
drwxrwxrwt   - root     root                0 2024-05-31 22:59 /tmp/logs
bash-5.0#
```

The information is downloaded into another local Docker instance containing Hadoop's HDFS to be imported into Apache Hive.

```
    > CREATE TABLE evsales (
    >     vin STRING,
    >     county STRING,
    >     city STRING,
    >     state STRING,
    >     zipcode STRING,
    >     year STRING,
    >     make STRING,
    >     model STRING,
    >     ev_type STRING,
    >     cafv_eligibility STRING,
    >     electric_range int,
    >     base_msrp int,
    >     legislative_district string
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE
    > tblproperties("skip.header.line.count"="1");
OK
Time taken: 1.486 seconds
```

```
hive> describe evsales;
OK
vin                     string
county                  string
city                    string
state                   string
zipcode                 string
year                    string
make                    string
model                   string
ev_type                 string
cafv_eligibility        string
electric_range          int
base_msrp               int
legislative_district    string
Time taken: 0.372 seconds, Fetched: 13 row(s)
```

A table (evsales) is created to hold the final Electric Vehicle Sales Data. The zip code and year columns are now categorical and used for grouping and are converted to strings when the data is loaded.

```
hive> LOAD DATA INPATH '/tmp/Electric_Vehicle_Sales_Data.csv' INTO TABLE evsales;
Loading data to table default.evsales
OK
Time taken: 0.597 seconds
hive> select * from evsales limit 10;
OK
WBY8P6C58K      King    Seattle WA      98115   2019    BMW     I3      BEV     CAFV Elegible   153     0       43
WBY33AW0XP      King    Seattle WA      98109   2023    BMW     I4      BEV     Unknown/Not Researched  0       0       36
5YJ3E1EB5L      King    Bothell WA      98011   2020    TESLA   MODEL 3 BEV     CAFV Elegible   322     0       1
1V2GNPE86P      King    Sammamish       WA      98075   2023    VOLKSWAGEN      ID.4    BEV     Unknown/Not Researched  0       0       41
5YJ3E1EB0M      Yakima  Yakima  WA      98908   2021    TESLA   MODEL 3 BEV     Unknown/Not Researched  0       0       14
1N4BZ1CP3K      Kitsap  Bainbridge Island       WA      98110   2019    NISSAN  LEAF    BEV     CAFV Elegible   150     0       23
KNDCC3LGXK      King    Kirkland        WA      98033   2019    KIA     NIRO    BEV     CAFV Elegible   239     0       45
SADHD2S10L      King    Bellevue        WA      98004   2020    JAGUAR  I-PACE  BEV     CAFV Elegible   234     0       41
5YJ3E1EB9K      King    Seattle WA      98177   2019    TESLA   MODEL 3 BEV     CAFV Elegible   220     0       36
5YJYGAEE8M      Snohomish       Snohomish       WA      98296   2021    TESLA   MODEL Y BEV     Unknown/Not Researched  0       0       1
Time taken: 2.693 seconds, Fetched: 10 row(s)
```

The dataset loaded into the evsales table is the same type of information as that seen through Jupyter Notebook.

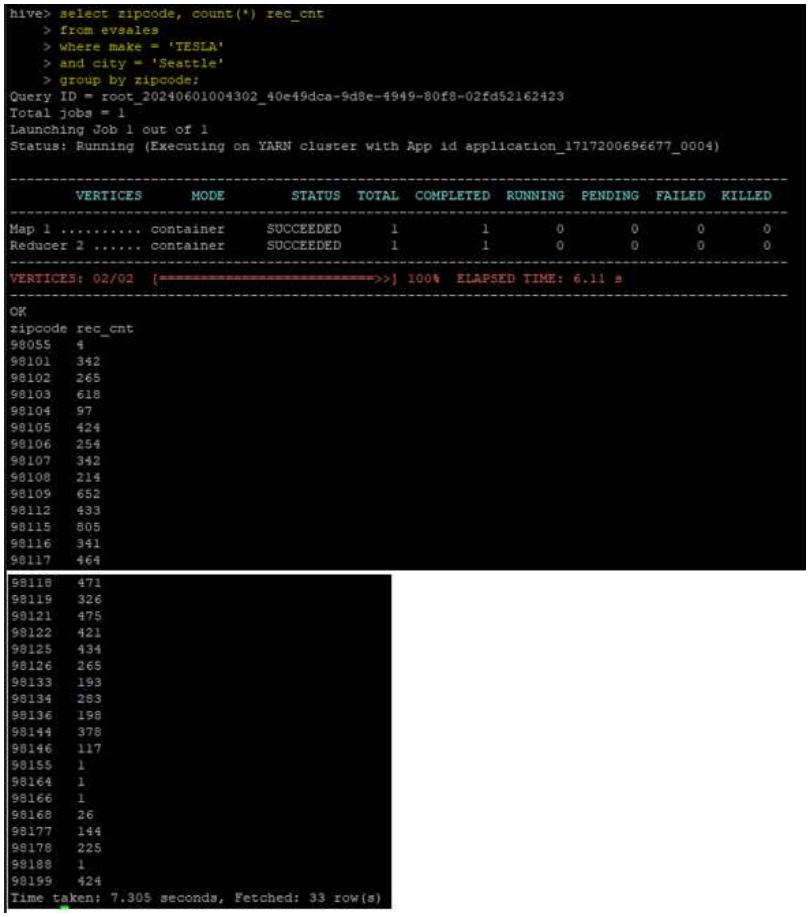
```
hive> select year, count(*) as rec_cnt from evsales
    > group by year;
Query ID = root_20240601002341_a6df032e-26ec-4c3e-9c8c-db03fcc87abf
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1717200696677_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 5.82 s
--------------------------------------------------------------------------------
OK
year    rec_cnt
2019    10904
2020    11990
2021    19254
2022    28067
2023    59230
Time taken: 6.784 seconds, Fetched: 5 row(s)
```

The total count (129445) and the vehicle sales by year also match the “After” results from the Jupyter Notebook, showing no data loss. The information shows an upward trend in electric vehicle popularity by year.

Evaluate the Results

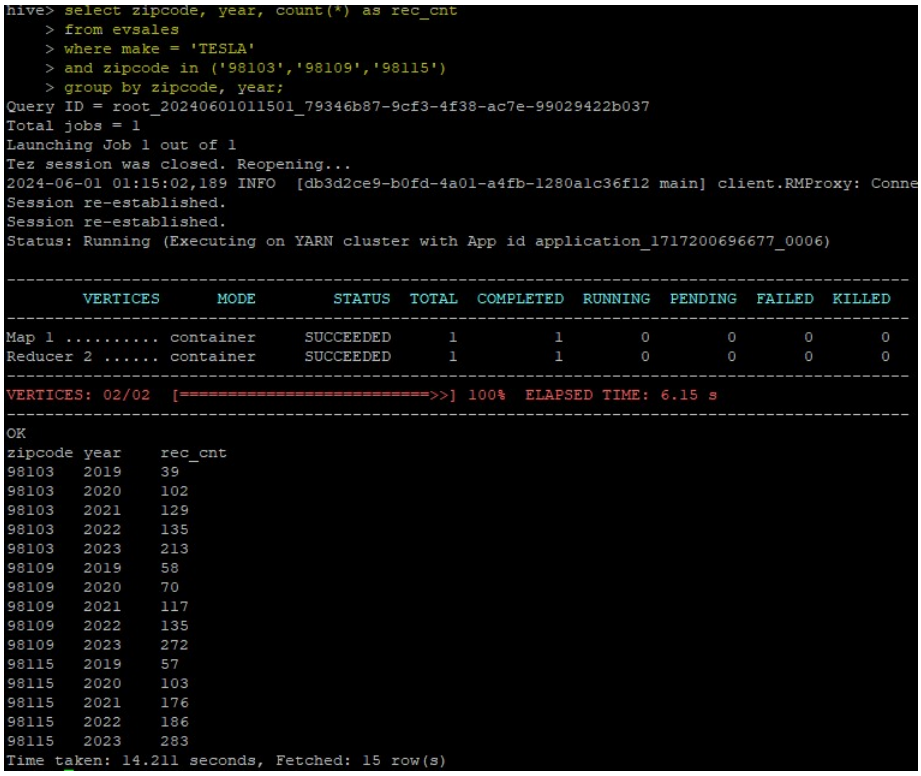
Based on the data exploration in Jupyter Notebook, the city of Seattle and the make of Tesla will be used to narrow down the data sampling. We will begin to examine the relationship between zip code and vehicle sales.



```
hive> select zipcode, count(*) rec_cnt
    > from evsales
    > where make = 'TESLA'
    > and city = 'Seattle'
    > group by zipcode;
Query ID = root_20240601004302_40e49dca-9d8e-4949-80f8-02fd52162423
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1717200696677_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 6.11 s
--------------------------------------------------------------------------------
OK
zipcode rec_cnt
98055   4
98101   342
98102   265
98103   618
98104   97
98105   424
98106   254
98107   342
98108   214
98109   652
98112   433
98115   805
98116   341
98117   464
98118   471
98119   326
98121   475
98122   421
98125   434
98126   265
98133   193
98134   283
98136   198
98144   378
98146   117
98155   1
98164   1
98166   1
98168   26
98177   144
98178   225
98188   1
98199   424
Time taken: 7.305 seconds, Fetched: 33 row(s)
```

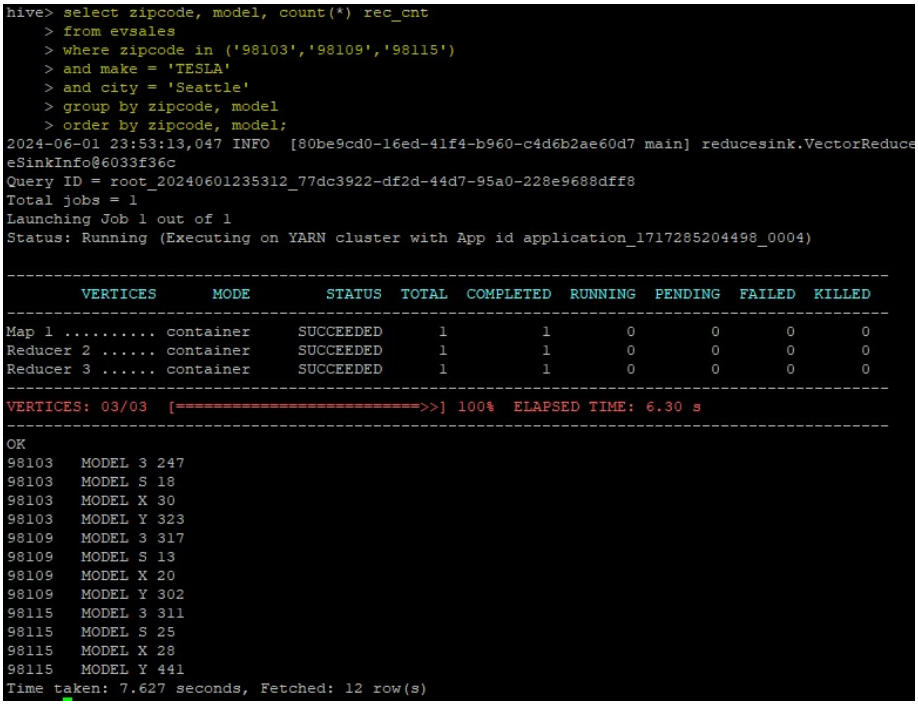
Upon examination, residents of zip codes 98103, 98109, and 98115 purchased the most Teslas during the 5-year period.



```
hive> select zipcode, year, count(*) as rec_cnt
    > from evsales
    > where make = 'TESLA'
    > and zipcode in ('98103','98109','98115')
    > group by zipcode, year;
Query ID = root_20240601011501_79346b87-9cf3-4f38-ac7e-99029422b037
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
2024-06-01 01:15:02,189 INFO  [db3d2ce9-b0fd-4a01-a4fb-1280a1c36f12 main] client.RMProxy: Conne
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1717200696677_0006)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 6.15 s
--------------------------------------------------------------------------------
OK
zipcode year    rec_cnt
98103   2019    39
98103   2020    102
98103   2021    129
98103   2022    135
98103   2023    213
98109   2019    58
98109   2020    70
98109   2021    117
98109   2022    135
98109   2023    272
98115   2019    57
98115   2020    103
98115   2021    176
98115   2022    186
98115   2023    283
Time taken: 14.211 seconds, Fetched: 15 row(s)
```

Looking at Tesla sales by the top three zip codes, the popularity trend continues.



```
hive> select zipcode, model, count(*) rec_cnt
    > from evsales
    > where zipcode in ('98103','98109','98115')
    > and make = 'TESLA'
    > and city = 'Seattle'
    > group by zipcode, model
    > order by zipcode, model;
2024-06-01 23:53:13,047 INFO  [80be9cd0-16ed-41f4-b960-c4d6b2ae60d7 main] reducesink.VectorReduce
eSinkInfo@6033f36c
Query ID = root_20240601235312_77dc3922-df2d-44d7-95a0-228e9688dff8
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1717285204498_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 6.30 s
--------------------------------------------------------------------------------
OK
98103   MODEL 3 247
98103   MODEL S 18
98103   MODEL X 30
98103   MODEL Y 323
98109   MODEL 3 317
98109   MODEL S 13
98109   MODEL X 20
98109   MODEL Y 302
98115   MODEL 3 311
98115   MODEL S 25
98115   MODEL X 28
98115   MODEL Y 441
Time taken: 7.627 seconds, Fetched: 12 row(s)
```

Models 3 and Y are the most popular and least expensive in the sample zip codes, starting at $33,990 and $31,490, respectively. Models S and X have the fewest sales but are the most costly,

with starting prices above $60,000. Using this information, we can examine the sales of comparable vehicles in other zip codes and target advertising to increase sales in those zip codes.

Conclusion

This overview is a sampling of the power of big data. Zip codes with more favorable audiences can also be targeted by including demographics and cable and streaming viewing patterns. Having more details on electric vehicle sales, such as a sale date, can identify trends at a monthly level so advertising dollars are spent during more favorable times of the year for each model. Adding a sale date and color can also increase and decrease the number of vehicles on hand monthly, saving money on inventory.

Again, the generational nature of the automobile sales business is the biggest hindrance to adopting big data. Current automobile dealers learn from their fathers, who learned from their fathers; many have worked in this industry since they were teenagers and only know how to advertise one way. My family has been in the automobile industry for four generations; my hope is by using this information, I can change some minds.

Targeted Advertising for New Vehicle Sales

References

Electric Population Data, Retrieved May 15, 2024

https://catalog.data.gov/dataset/electric-vehicle-population-data

Decibel (November 2, 2021) *Why ZIP code targeting is key for digital campaigns*

https://decibelads.com/zip-code-targeting/

Tesla, Inc. (n.d.) Pricing information

https://www.tesla.com/