K-Means Clustering Analysis of Vehicle Fuel Type Correlations with Asthma Related Deaths in the Bay Area

By: Mihir Thakar

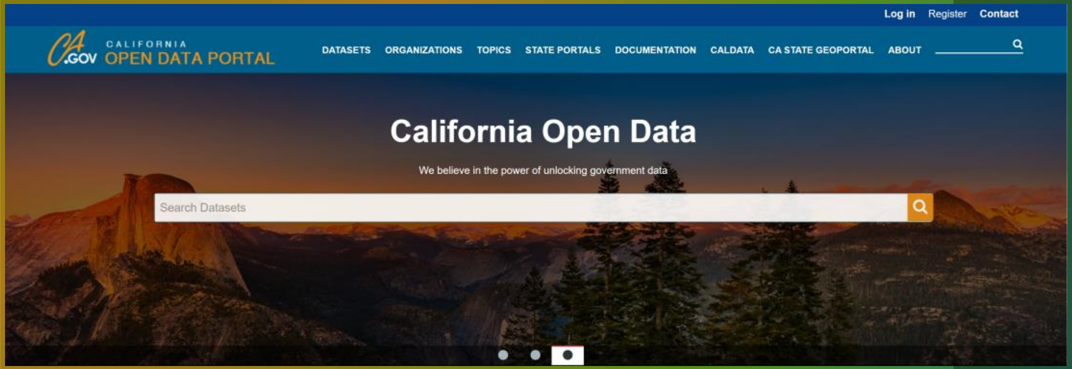
Work Pipeline

DATA ACQUISITION, SELECTION, AND CLEANING

DATA EXPLORATION AND ANALYSIS

DATA VISUALIZATION

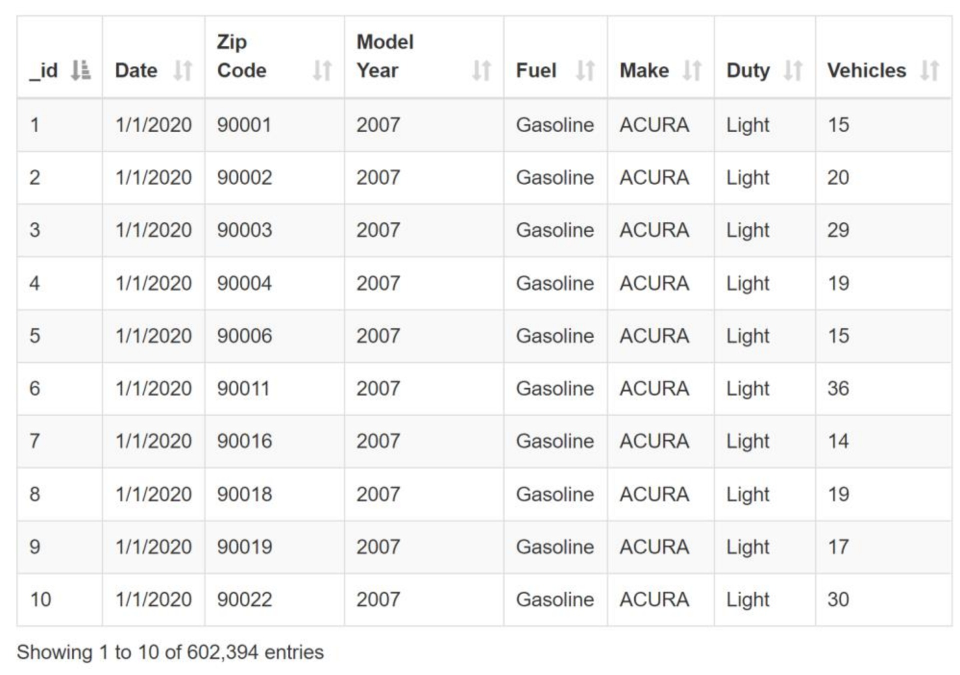
Datasets



California Open Data

We believe in the power of unlocking government data

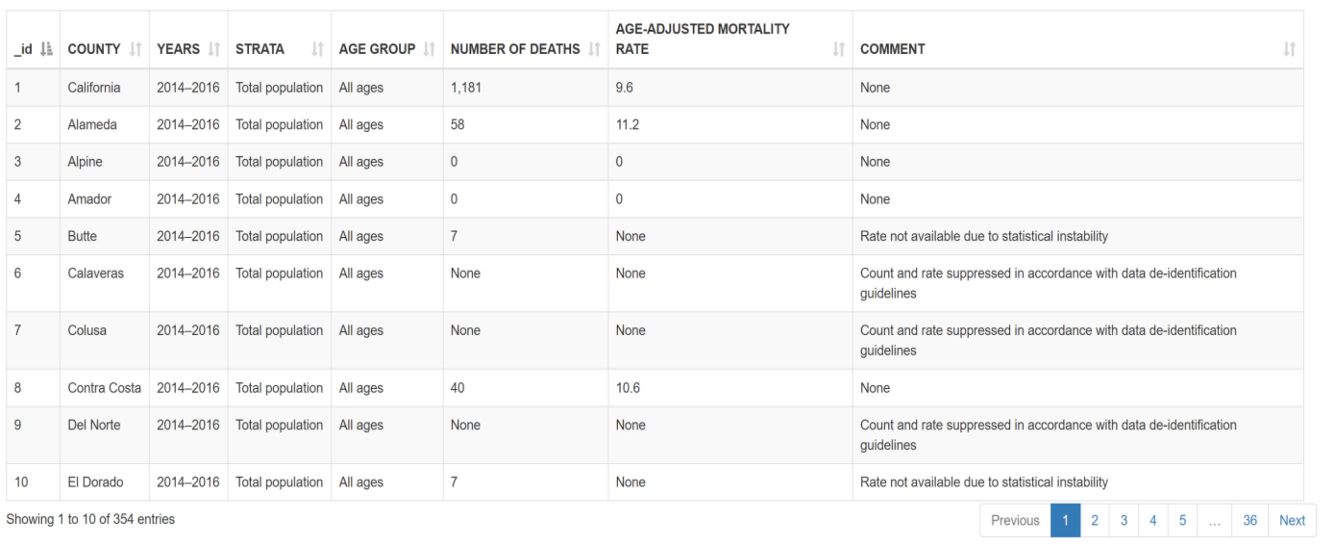
Vehicle Fuel Type Count by Zip Code



_id	Date	Zip Code	Model Year	Fuel	Make	Duty	Vehicles
1	1/1/2020	90001	2007	Gasoline	ACURA	Light	15
2	1/1/2020	90002	2007	Gasoline	ACURA	Light	20
3	1/1/2020	90003	2007	Gasoline	ACURA	Light	29
4	1/1/2020	90004	2007	Gasoline	ACURA	Light	19
5	1/1/2020	90006	2007	Gasoline	ACURA	Light	15
6	1/1/2020	90011	2007	Gasoline	ACURA	Light	36
7	1/1/2020	90016	2007	Gasoline	ACURA	Light	14
8	1/1/2020	90018	2007	Gasoline	ACURA	Light	19
9	1/1/2020	90019	2007	Gasoline	ACURA	Light	17
10	1/1/2020	90022	2007	Gasoline	ACURA	Light	30

Showing 1 to 10 of 602,394 entries

Asthma Deaths by County

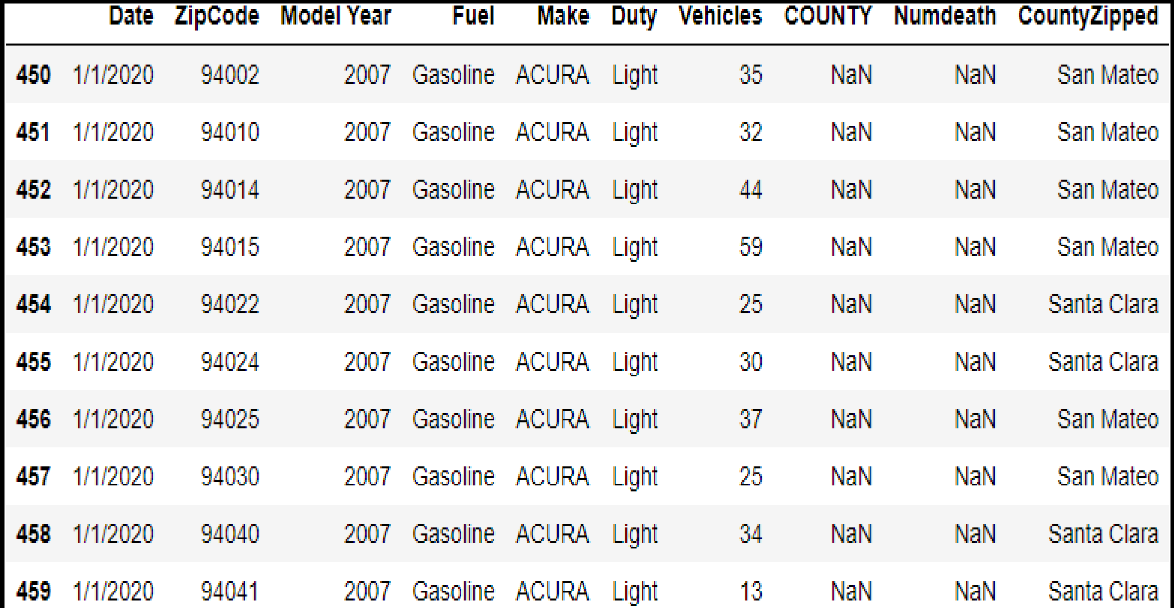


_id	COUNTY	YEARS	STRATA	AGE GROUP	NUMBER OF DEATHS	AGE-ADJUSTED MORTALITY RATE	COMMENT
1	California	2014–2016	Total population	All ages	1,181	9.6	None
2	Alameda	2014–2016	Total population	All ages	58	11.2	None
3	Alpine	2014–2016	Total population	All ages	0	0	None
4	Amador	2014–2016	Total population	All ages	0	0	None
5	Butte	2014–2016	Total population	All ages	7	None	Rate not available due to statistical instability
6	Calaveras	2014–2016	Total population	All ages	None	None	Count and rate suppressed in accordance with data de-identification guidelines
7	Colusa	2014–2016	Total population	All ages	None	None	Count and rate suppressed in accordance with data de-identification guidelines
8	Contra Costa	2014–2016	Total population	All ages	40	10.6	None
9	Del Norte	2014–2016	Total population	All ages	None	None	Count and rate suppressed in accordance with data de-identification guidelines
10	El Dorado	2014–2016	Total population	All ages	7	None	Rate not available due to statistical instability

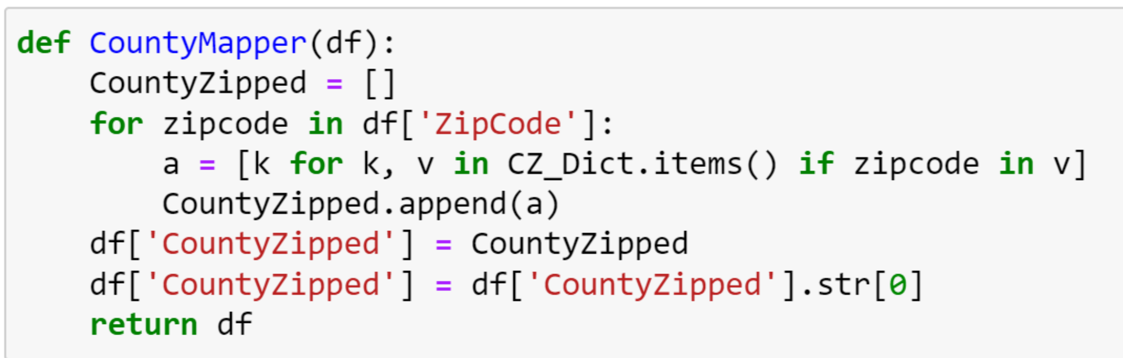
Showing 1 to 10 of 354 entries

Data Cleaning Before

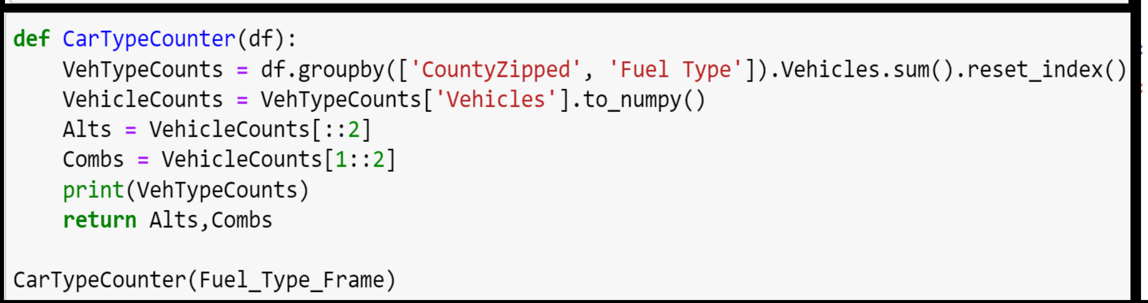
- Functions:
  - CountyMapper: Maps each ZipCode in Fuel Dataset to it's corresponding County
  - CarCounter: Counts the cars in each County
  - CarTypeCounter:
    - Combustion = ['Gasoline', 'Diesel and Diesel Hybrid', 'Hybrid Gasoline', 'Flex-Fuel', 'Natural Gas', 'Other']
    - Alternative = ['Battery Electric', 'Plug-in Hybrid', 'Hydrogen Fuel Cell']



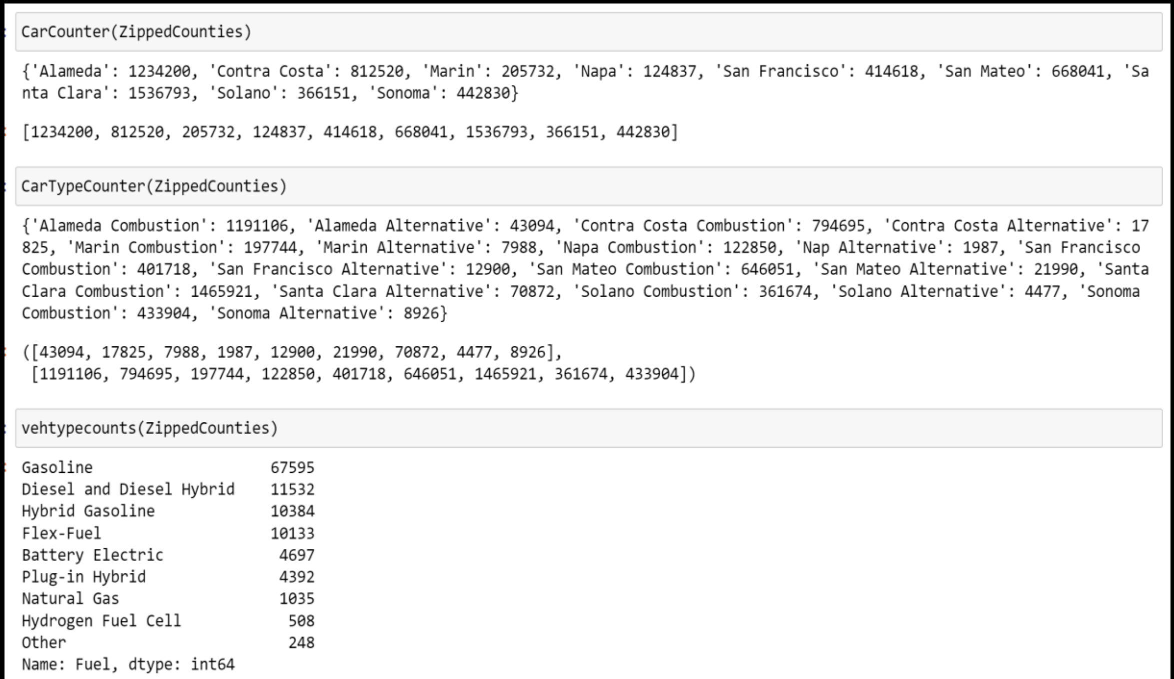
	Date	ZipCode	Model Year	Fuel	Make	Duty	Vehicles	COUNTY	Numdeath	CountyZipped
450	1/1/2020	94002	2007	Gasoline	ACURA	Light	35	NaN	NaN	San Mateo
451	1/1/2020	94010	2007	Gasoline	ACURA	Light	32	NaN	NaN	San Mateo
452	1/1/2020	94014	2007	Gasoline	ACURA	Light	44	NaN	NaN	San Mateo
453	1/1/2020	94015	2007	Gasoline	ACURA	Light	59	NaN	NaN	San Mateo
454	1/1/2020	94022	2007	Gasoline	ACURA	Light	25	NaN	NaN	Santa Clara
455	1/1/2020	94024	2007	Gasoline	ACURA	Light	30	NaN	NaN	Santa Clara
456	1/1/2020	94025	2007	Gasoline	ACURA	Light	37	NaN	NaN	San Mateo
457	1/1/2020	94030	2007	Gasoline	ACURA	Light	25	NaN	NaN	San Mateo
458	1/1/2020	94040	2007	Gasoline	ACURA	Light	34	NaN	NaN	Santa Clara
459	1/1/2020	94041	2007	Gasoline	ACURA	Light	13	NaN	NaN	Santa Clara



```python
def CountyMapper(df):
    CountyZipped = []
    for zipcode in df['ZipCode']:
        a = [k for k, v in CZ_Dict.items() if zipcode in v]
        CountyZipped.append(a)
    df['CountyZipped'] = CountyZipped
    df['CountyZipped'] = df['CountyZipped'].str[0]
    return df
```

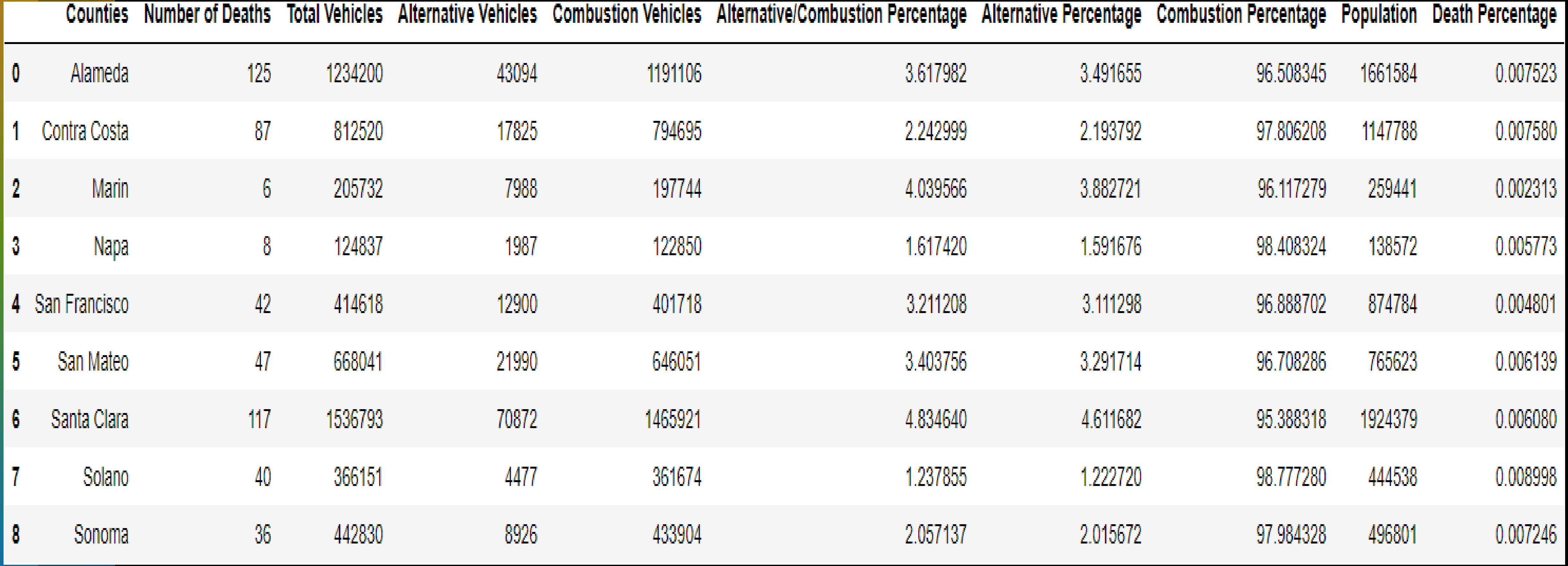


```python
def CarTypeCounter(df):
    VehTypeCounts = df.groupby(['CountyZipped', 'Fuel Type']).Vehicles.sum().reset_index()
    VehicleCounts = VehTypeCounts['Vehicles'].to_numpy()
    Alts = VehicleCounts[::2]
    Combs = VehicleCounts[1::2]
    print(VehTypeCounts)
    return Alts,Combs

CarTypeCounter(Fuel_Type_Frame)
```



```
CarCounter(ZippedCounties)

{'Alameda': 1234200, 'Contra Costa': 812520, 'Marin': 205732, 'Napa': 124837, 'San Francisco': 414618, 'San Mateo': 668041, 'Sa
nta Clara': 1536793, 'Solano': 366151, 'Sonoma': 442830}

[1234200, 812520, 205732, 124837, 414618, 668041, 1536793, 366151, 442830]

CarTypeCounter(ZippedCounties)

{'Alameda Combustion': 1191106, 'Alameda Alternative': 43094, 'Contra Costa Combustion': 794695, 'Contra Costa Alternative': 17
825, 'Marin Combustion': 197744, 'Marin Alternative': 7988, 'Napa Combustion': 122850, 'Nap Alternative': 1987, 'San Francisco
Combustion': 401718, 'San Francisco Alternative': 12900, 'San Mateo Combustion': 646051, 'San Mateo Alternative': 21990, 'Santa
Clara Combustion': 1465921, 'Santa Clara Alternative': 70872, 'Solano Combustion': 361674, 'Solano Alternative': 4477, 'Sonoma
Combustion': 433904, 'Sonoma Alternative': 8926}

([43094, 17825, 7988, 1987, 12900, 21990, 70872, 4477, 8926],
 [1191106, 794695, 197744, 122850, 401718, 646051, 1465921, 361674, 433904])

vehtypecounts(ZippedCounties)

Gasoline                    67595
Diesel and Diesel Hybrid    11532
Hybrid Gasoline             10384
Flex-Fuel                   10133
Battery Electric             4697
Plug-in Hybrid               4392
Natural Gas                  1035
Hydrogen Fuel Cell            508
Other                         248
Name: Fuel, dtype: int64
```

Data Cleaning After



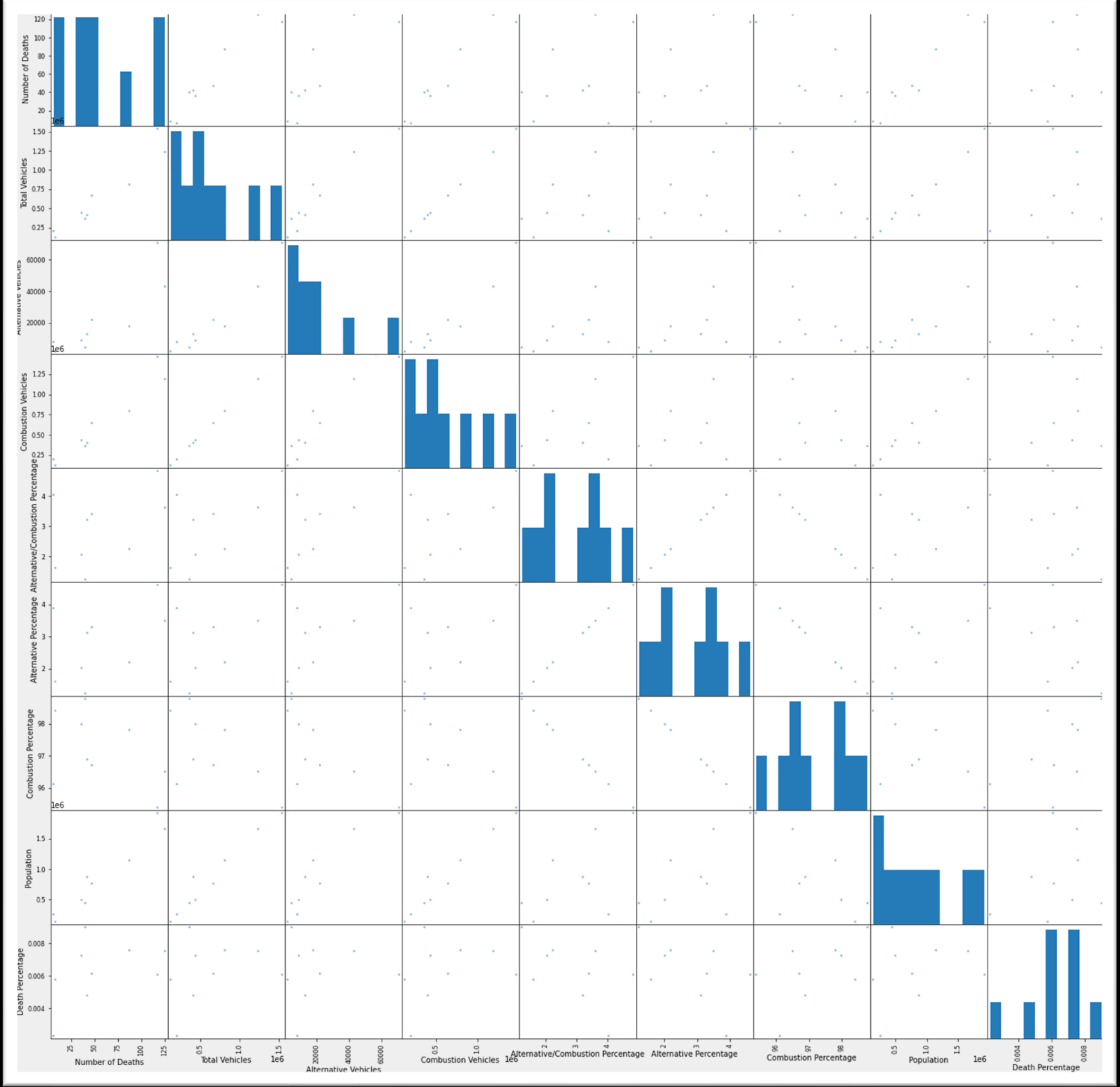
	Counties	Number of Deaths	Total Vehicles	Alternative Vehicles	Combustion Vehicles	Alternative/Combustion Percentage	Alternative Percentage	Combustion Percentage	Population	Death Percentage
0	Alameda	125	1234200	43094	1191106	3.617982	3.491655	96.508345	1661584	0.007523
1	Contra Costa	87	812520	17825	794695	2.242999	2.193792	97.806208	1147788	0.007580
2	Marin	6	205732	7988	197744	4.039566	3.882721	96.117279	259441	0.002313
3	Napa	8	124837	1987	122850	1.617420	1.591676	98.408324	138572	0.005773
4	San Francisco	42	414618	12900	401718	3.211208	3.111298	96.888702	874784	0.004801
5	San Mateo	47	668041	21990	646051	3.403756	3.291714	96.708286	765623	0.006139
6	Santa Clara	117	1536793	70872	1465921	4.834640	4.611682	95.388318	1924379	0.006080
7	Solano	40	366151	4477	361674	1.237855	1.222720	98.777280	444538	0.008998
8	Sonoma	36	442830	8926	433904	2.057137	2.015672	97.984328	496801	0.007246

Data Exploration

Pair Wise Plots

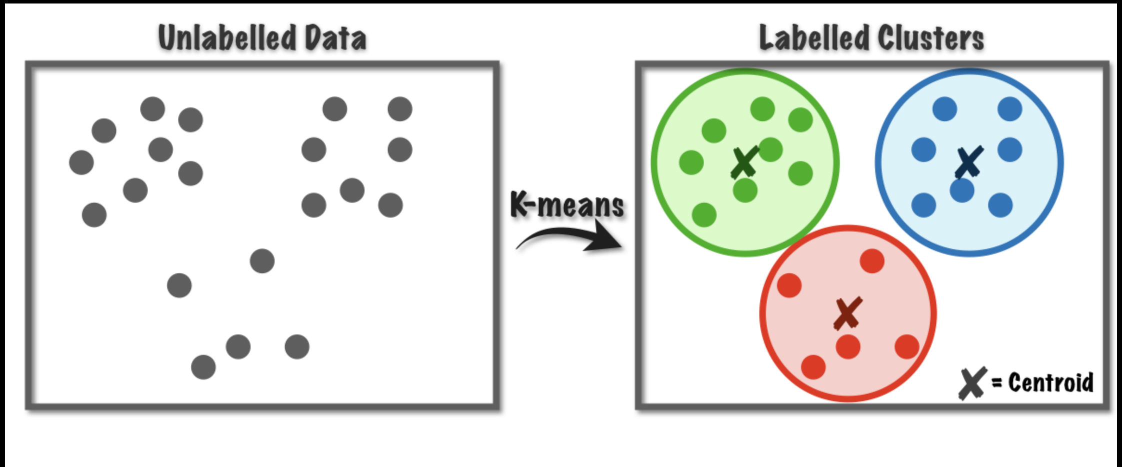
- Matrix of scatterplots meant to display correlations between attributes

- See relationships between 2 variables



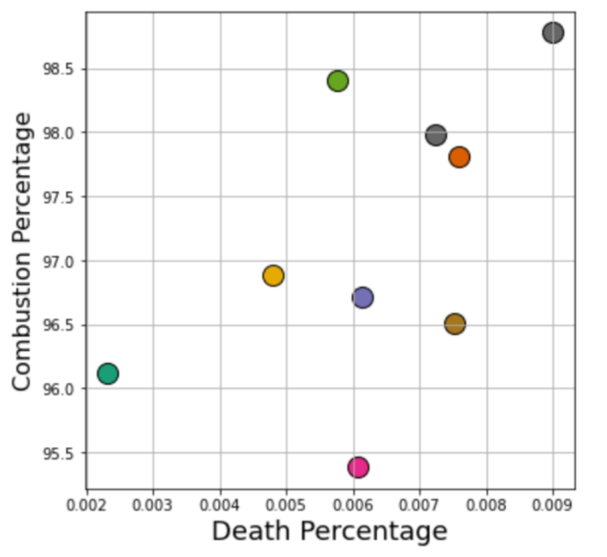
K-Means Clustering

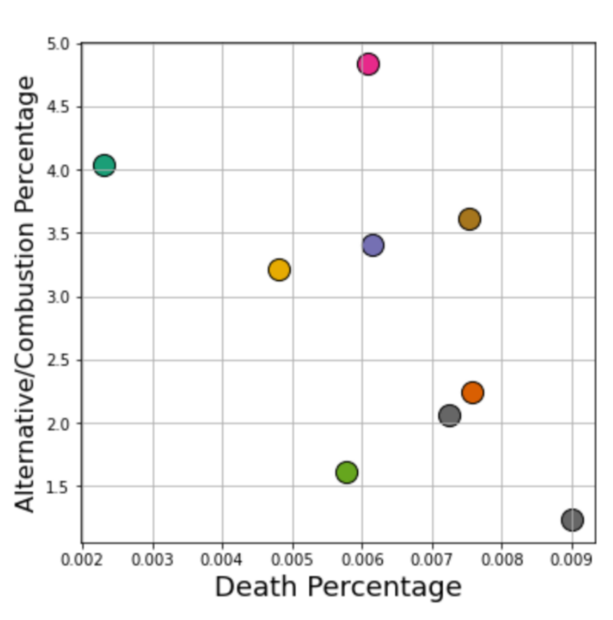
- Unsupervised Machine Learning algorithm: Designated for unlabeled data
- Objective: Group similar data points together to discover patterns
  - Cluster: Collection of data points aggregated because of certain similarities
  - First ran program with k=9 clusters for the 9 different counties in the Bay Area
- Inertia Score determined 3 clusters was the optimal cluster coefficient => 3 triplets of counties were aggregated together based on their similarity

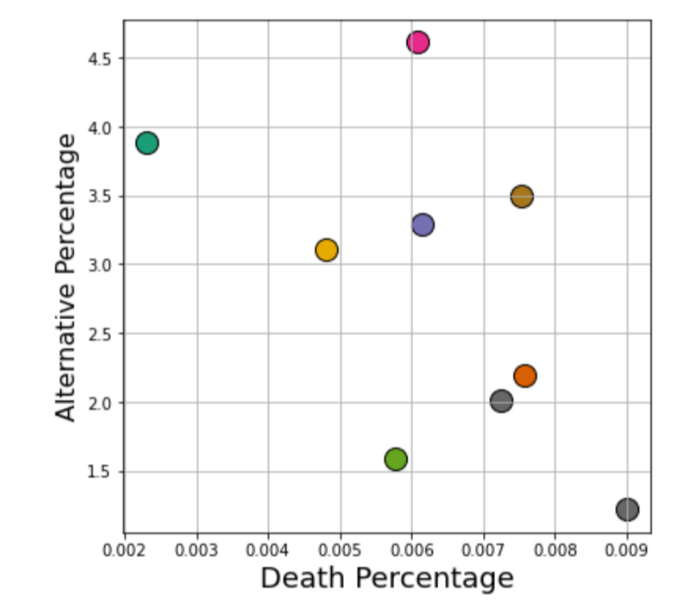


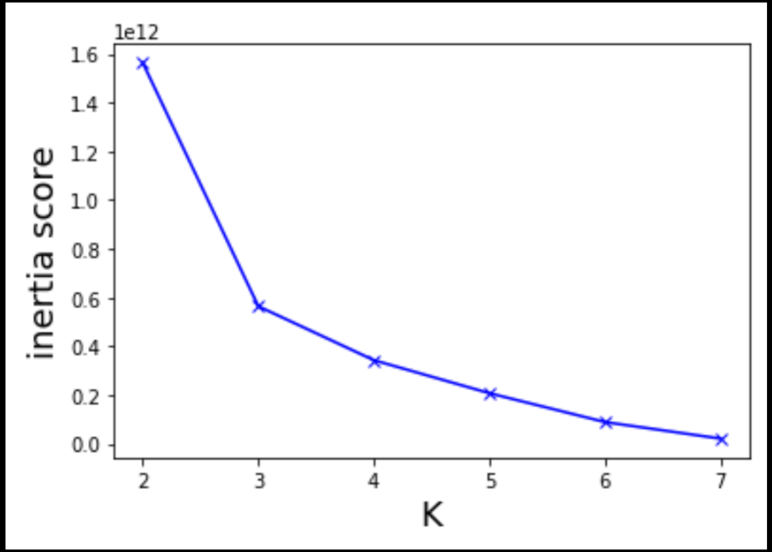
Source: https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

Run 1: K = 9

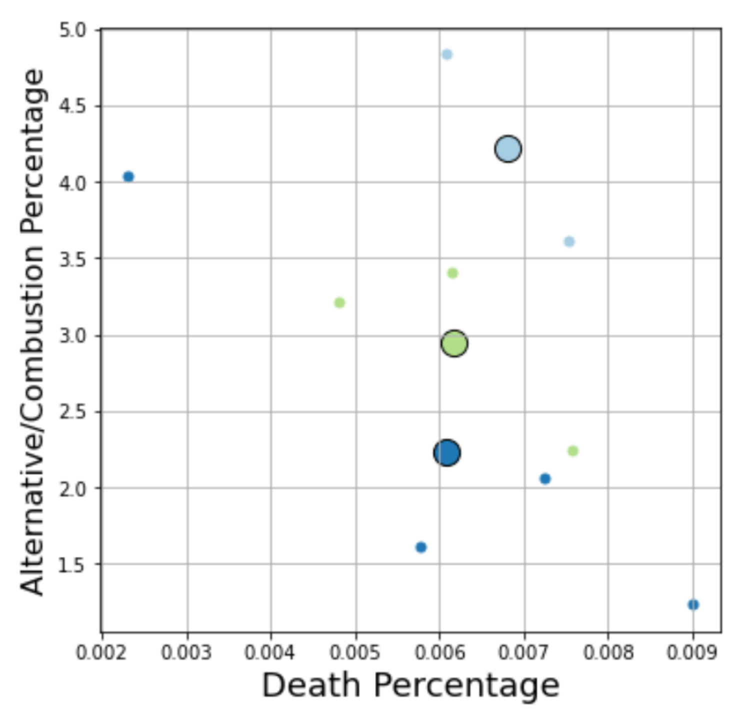

Combustion Percentage
98.5
98.0
97.5
97.0
96.5
96.0
95.5
0.002
0.003
0.004
0.005
0.006
0.007
0.008
0.009
Death Percentage

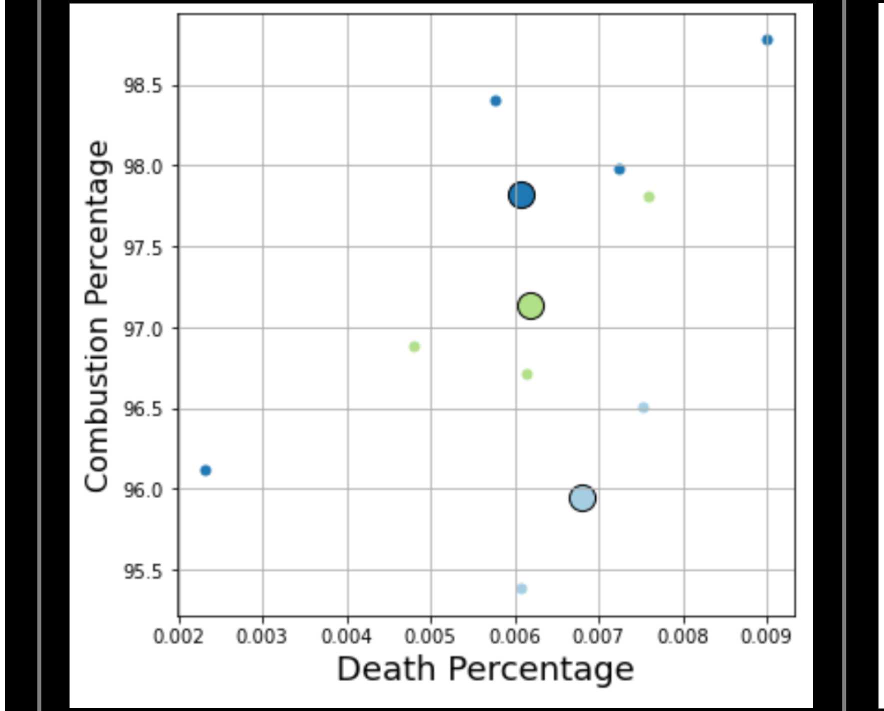

Alternative/Combustion Percentage
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
0.002
0.003
0.004
0.005
0.006
0.007
0.008
0.009
Death Percentage

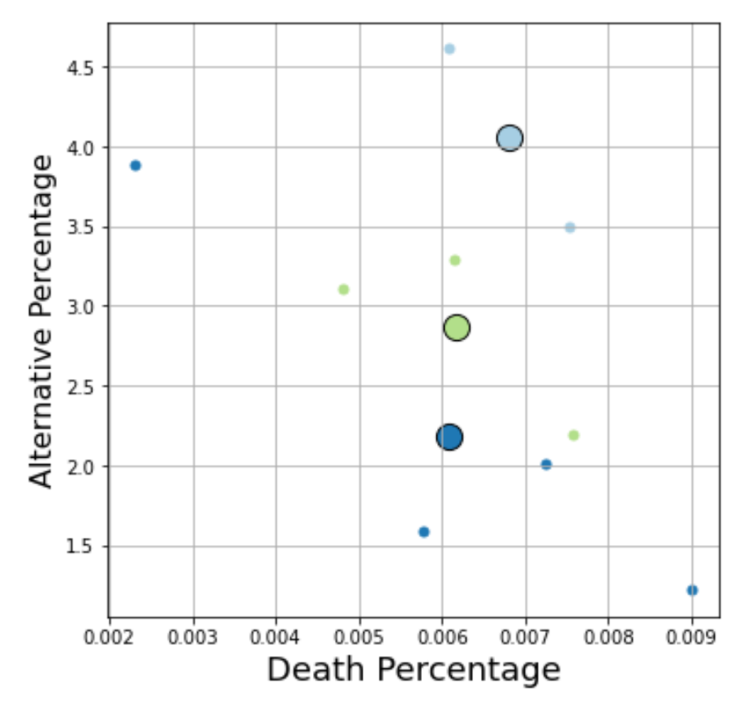

Alternative Percentage
4.5
4.0
3.5
3.0
2.5
2.0
1.5
0.002
0.003
0.004
0.005
0.006
0.007
0.008
0.009
Death Percentage


1e12
inertia score
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
2
3
4
5
6
7
K

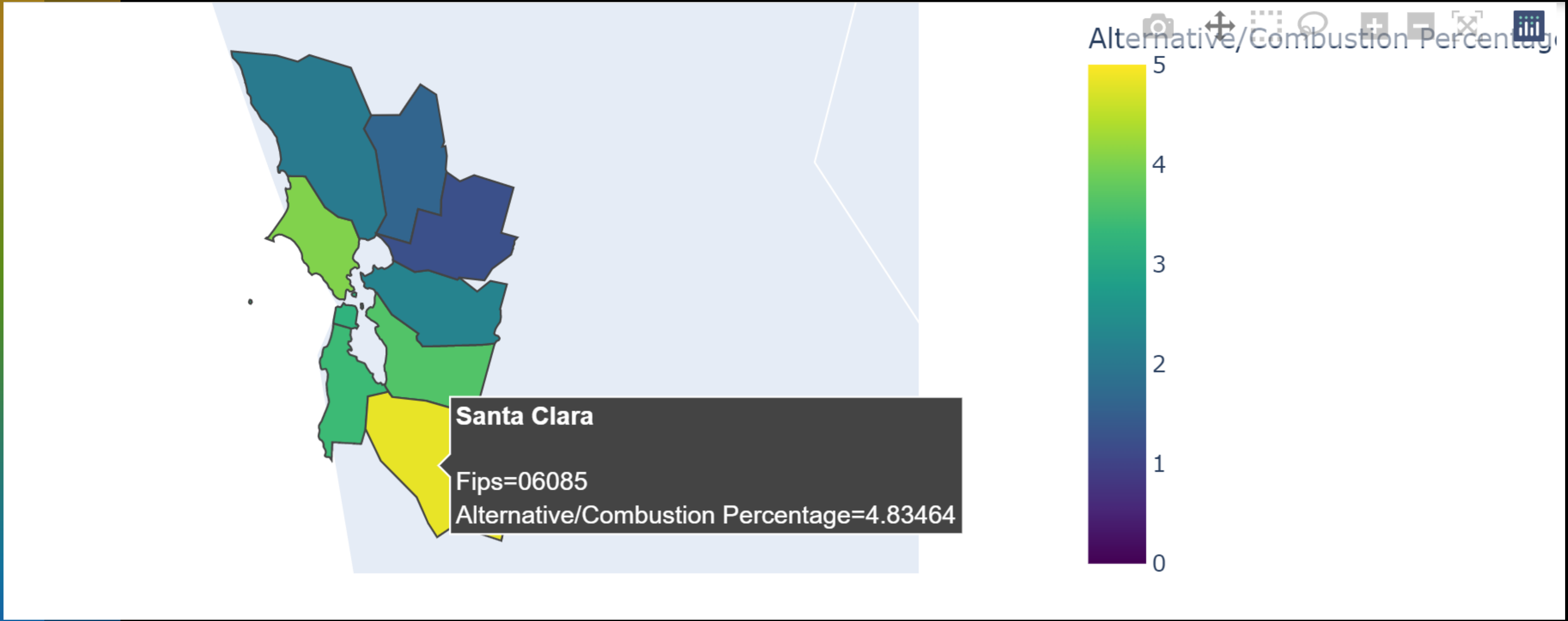
Run 2: K=3


Alternative/Combustion Percentage
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
0.002
0.003
0.004
0.005
0.006
0.007
0.008
0.009
Death Percentage


Combustion Percentage
98.5
98.0
97.5
97.0
96.5
96.0
95.5
0.002
0.003
0.004
0.005
0.006
0.007
0.008
0.009
Death Percentage


Alternative Percentage
4.5
4.0
3.5
3.0
2.5
2.0
1.5
0.002
0.003
0.004
0.005
0.006
0.007
0.008
0.009
Death Percentage

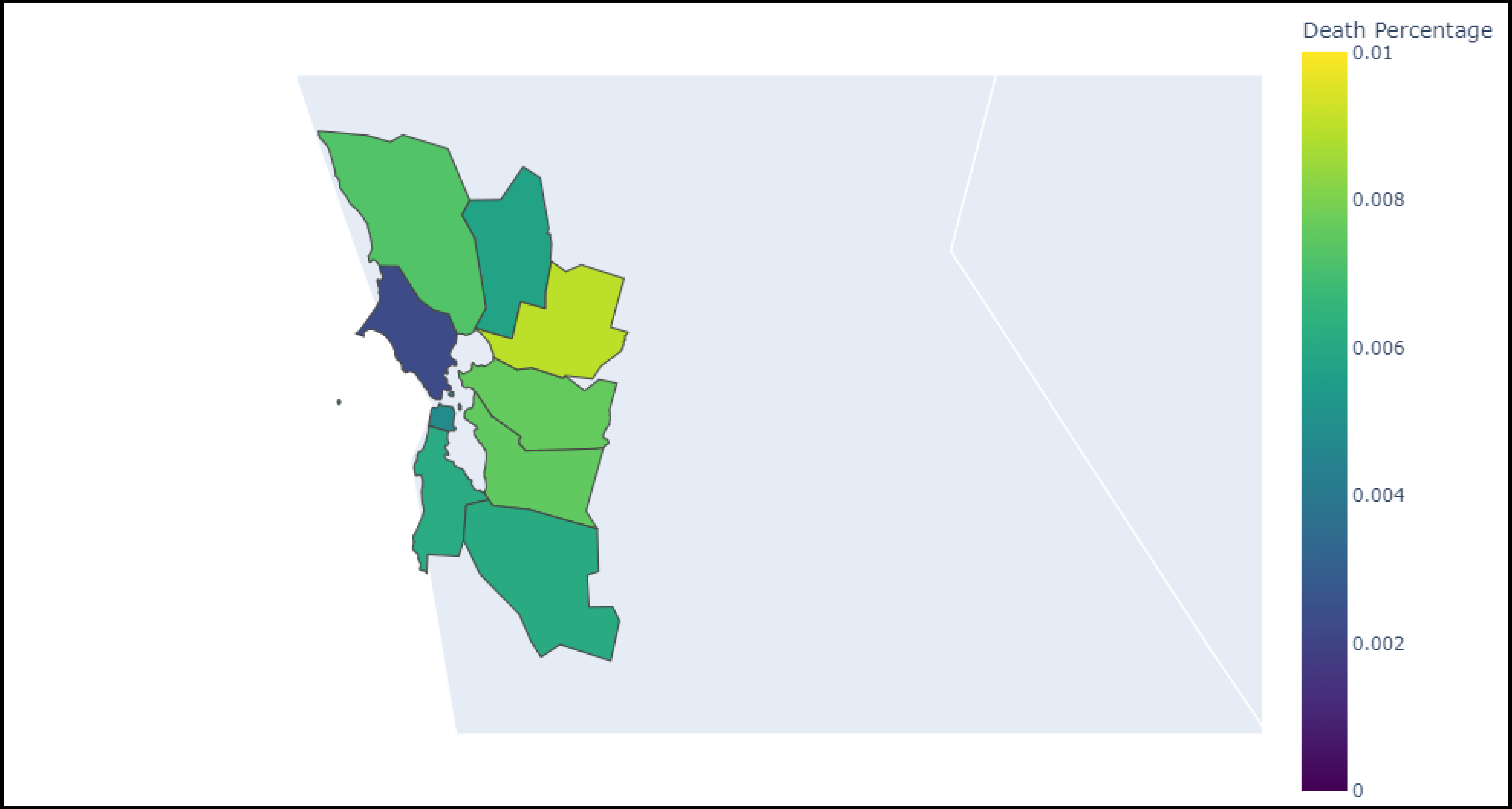
Choropleth Maps: Alternative/Combustion Ratio



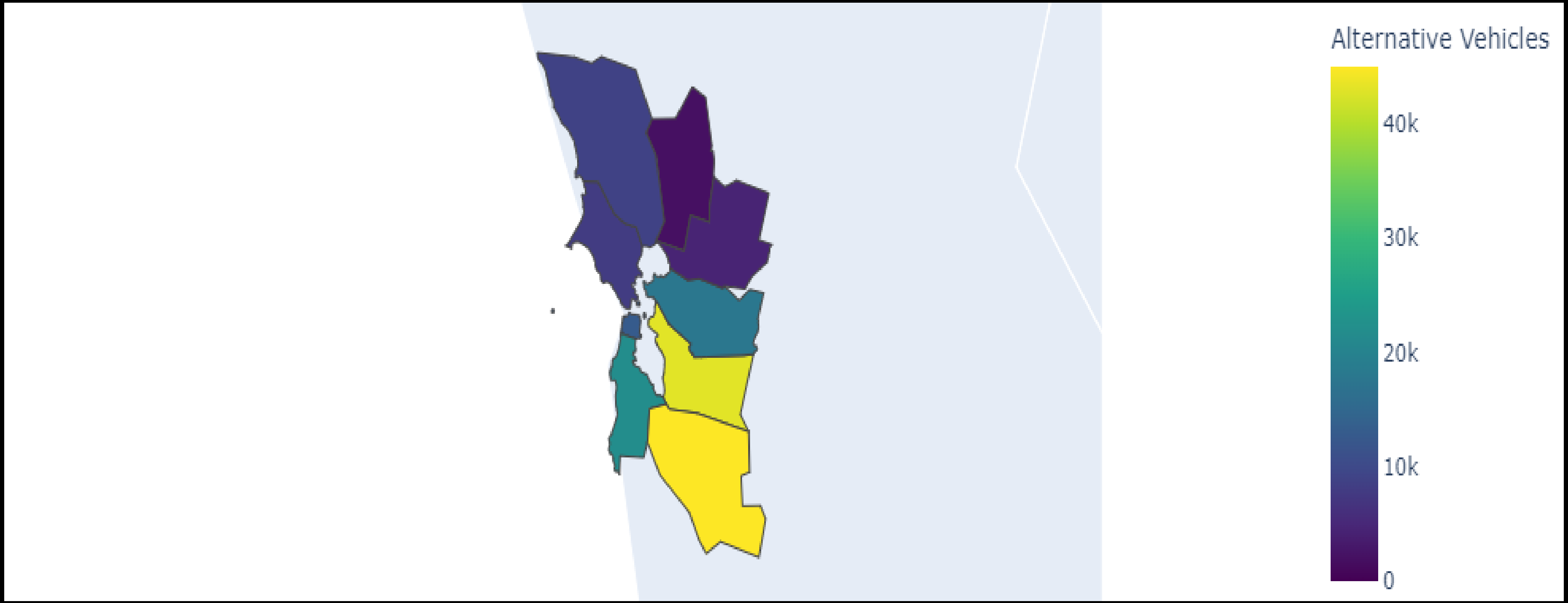
Santa Clara

Fips=06085
Alternative/Combustion Percentage=4.83464

Choropleth Maps: Death % by County



Choropleth Maps: Alternative Vehicles



Thank you for listening!