



```
1 !python -V
```

 Python 3.11.11

```
1 import numpy as np
2 import pandas as pd
3 import geopandas as gpd
4 import plotly.graph_objects as go
5 import plotly.express as px
6 import matplotlib.ticker as ticker
7 from datetime import datetime
8
9
10 %matplotlib inline

11 # data recieved from https://catalog.data.gov/dataset/crash-reporting-drivers-cc
12 filename = "/content/maryland_crash_data.csv"
13 df = pd.read_csv(
14     filename,
15     dtype={
16         "Report Number": str,
17         "Local Case Number": str,
18         # "Crash Date/Time": datetime,
19         "Latitude": str,
20         "Longitude": float,
21         "Vehicle Year": int,
22     },
23     parse_dates=[
24         "Crash Date/Time",
25     ],
26 )
```

 <ipython-input-3-1ee1d5b2bc70>:3: UserWarning: Could not infer format, so each
df = pd.read_csv(

✓ Project

Data source

As discussed in class, as the professor advised, I am changing my data source to be aligned with the govt data sources. My data source for this mining operation can be found at [Data.gov site](#)

The data has 192183 rows and 39 columns.

```

1 # Sanitize to remove absurd make year values
2 df = df[df["Vehicle Year"] < 2025]
3 df = df[df["Vehicle Year"] > 1970]

1 df["Vehicle Year"] = pd.to_datetime(df["Vehicle Year"], format='%Y')
2 df["difference_crash_make_year"] = (df["Crash Date/Time"] - df["Vehicle Year"]

1 df = df[df["difference_crash_make_year"] >= 0 ]

1 df.head()

```



	Report Number	Local Case Number	Agency Name	ACRS Report Type	Crash Date/ Time	Route Type	Road Na
0	DM8479000T	210020119	Takoma Park Police Depart	Property Damage Crash	2021-05-27 19:40:00	NaN	Na
1	MCP2970000R	15045937	MONTGOMERY	Property Damage Crash	2015-09-11 13:29:00	NaN	Na
2	MCP20160036	180040948	Montgomery County Police	Property Damage Crash	2018-08-17 14:25:00	NaN	Na
3	EJ7879003C	230048975	Gaithersburg Police Depar	Injury Crash	2023-08-11 18:00:00	NaN	Na
4	MCP2967004Y	230070277	Montgomery County Police	Property Damage Crash	2023-12-06 18:42:00	Maryland (State)	CONNECTIC A

5 rows × 40 columns

✓ Visualization

Here we can plot all the accidents in our dataset and see how clustered the accidents are and we can also see what year cars are more accident prone and if it has anything to do with how old the cars are.

There might be other interesting ideas that might pop up after checking out the visualizations!

```

1 fig_vehicle_year = px.scatter_geo(df, lat="Latitude", lon="Longitude", color="Ve
2 fig_vehicle_year.show()

```

```
1 fig_vehicle_difference_year = px.scatter_geo(df, lat="Latitude", lon="Longitude"  
2 fig_vehicle_difference_year.show()
```

```
1 print(df.groupby(["difference_crash_make_year"])["difference_crash_make_year"].c
```

difference_crash_make_year	difference_crash_make_year
0	8600
1	12440
2	12470
3	12755
4	13047
5	12517
6	11781
7	11071
8	10720
9	10226
10	9520
11	8890
12	8158
13	7453
14	6731
15	6041
16	5234
17	4326
18	3628
19	2796
20	2122
21	1635
22	1220
23	878
24	635
25	471
26	340
27	208
28	180
29	138
30	91
31	61
32	63
33	48
34	41
35	33
36	16

	37		11	
	38		11	
	39		7	
	40		8	
	41		10	
	42		3	
	43		4	
	44		2	
	46		3	
	47		3	
	48		1	
	49		1	

✓ Corr, Mean, Median

I'm still understanding what useful knowledge can be extracted to use to make safer cars.

I will use Corr, Mean, Median to help determine if cars have gotten safer over time.

```
1 df[[
2     "Vehicle Year",
3     "difference_crash_make_year"
4 ]].corr(min_periods=3)
```

```
1 df[[
2     # "Vehicle Year",
3     "difference_crash_make_year"
4 ]].median(
5     axis=0,
6 )
```

```
1 df[[
2     # "Vehicle Year",
3     "difference_crash_make_year"
4 ]].mean(
5     axis=0
```

```
5     axis=0,
6 )
```

```
1 df[[
2     "Vehicle Make",
3 ]].value_counts().idxmax()
```

```
( 'TOYOTA' , )
```

```
1 pd.Timestamp.fromordinal(int(df[["Vehicle Year"]].apply(lambda x: x[1].toordinal(
Timestamp('2012-01-01 00:00:00'))
```

```
1 pd.Timestamp.fromordinal(int(df[["Vehicle Year"]].apply(lambda x: x[1].toordinal(
Timestamp('2012-01-01 00:00:00'))
```

Findings

As we can see, on average, we can expect accidents to start occurring after the car is around 7~8 years of age. This dataset also highlights that the 2012 models are more likely to be the ones in accident. This could also just be a red herring that is throwing off our data since a lot of crashes can go unreported.

Algo selection

I would like to explore a few different ideas. But one of the algorithms we were taught in class was decision tree. I would like to explore more columns with decision tree to determine rates of accidents given the factors, such as:

- Make of the car
- Year of the car
- Age of the car

A major reason to use the decision tree is that it is easy to see the outcome. This will produce a very clean algebraic equation that can help customers buy the right car for them!

This learning can produce a highly detailed shopping guide for our use! I was also reading the syllabus, and in the future I saw that Apriori was an algorithm we might learn. I would also like to come back and re evaluate my decision once I have gone through that algorithm too since it seems to be a shopping related algorithm, however, I still do not fully understand that one just yet.