CLOUD TECHNOLOGIES AND BIG DATA FRAMEWORK ASSIGNMENT 2

Ziad El Harairi

DATASET FOR USED CAR PRICES

The dataset I chose can be found on kaggle using the following link: https://www.kaggle.com/datasets/ayaz11/used-car-price-prediction

The dataset contains data for used cars in the United States of America from 2000-2023.

The dataset contains data about differents brands, miles, prices, car color (exterior and interior), and condition (number of owners and accidents)

DATAFRAMES

- I divided my data into 3 dataframes
  - Model manufactured in 2022 and 2023 (dataframe 1)
  - Model manufactured in 2020 and 2021 (dataframe 2)
  - Model manufactured in 2018 and 2019 (dataframe 3)
- This way I could focus more on the change through the previous 6 years and the impact of different conditions on the price of a car

QUERIES 1 AND 2: MAXIMUM AND MINIMUM

- I started with finding the maximum and minimum of prices for cars in each year, this was done to find out if the prices for each dataframe would be in order
- My hypothesis was that the maximum would be higher for dataframe 1 than dataframe 2 and dataframe 2 would be higher than dataframe 3. Vice versa for the minimum
- For the minimum my hypothesis was correct however for the maximum in dataframe 3 the maximum belonged for a Rolls Royce car and there were no other Rolls Royce cars in dataframe 1 and 2 therefore my hypothesis for the maximum was rejected.

QUERY 3: MOST CAR BRAND AND MODEL FOR SALE EACH YEAR

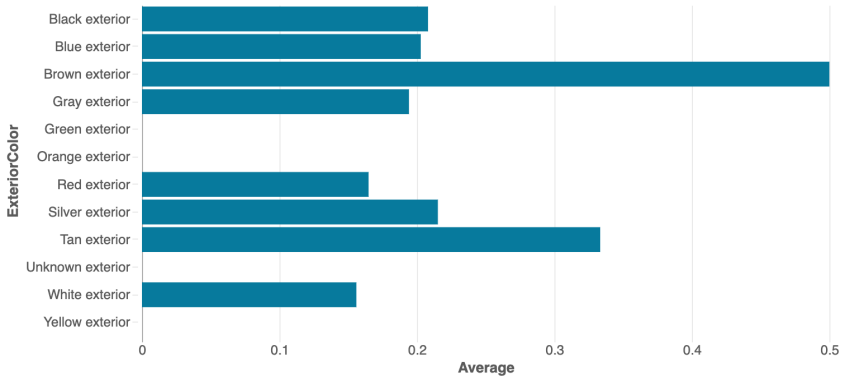
- I then wanted to find out from my data which car brand and model that are for sale that were manufactured from 2018-2023
- I started by grouping by each car brand and model and counting the number occurances for each year
- Furthermore, we retrieve the maximum of each count for each year
- As a result we found these findings:

```
For 2023, the max number of car sales was 19 for Audi A3
For 2022, the max number of car sales was 22 for Chevrolet Malibu
For 2021, the max number of car sales was 54 for Toyota Corolla
For 2020, the max number of car sales was 24 for Honda Civic
For 2019, the max number of car sales was 14 for Ford F-150
For 2018, the max number of car sales was 14 for Jeep Wrangler
```

Query 4: Ratio of Accidents to Car Color

- In some countries it said that red cars result in more accidents. As a result I wanted to test this
- I started by created two functions, first was exteriorColor() for splitting the values in the color column and retreiving the exterior color of the car. The second function was numAccidents(), for retrieving the number of accidents for a specific car
- I then combined all the three dataframes and grouped by the car exterior and the number of cars that have been in accidents
- At the end I divided the number of cars that had an accident over the overall number of cars with that color
- As a result we got this data (next slide) and we can also see in the graph that a higher percentage of grey cars were in accidents compared to the total number of grey cars

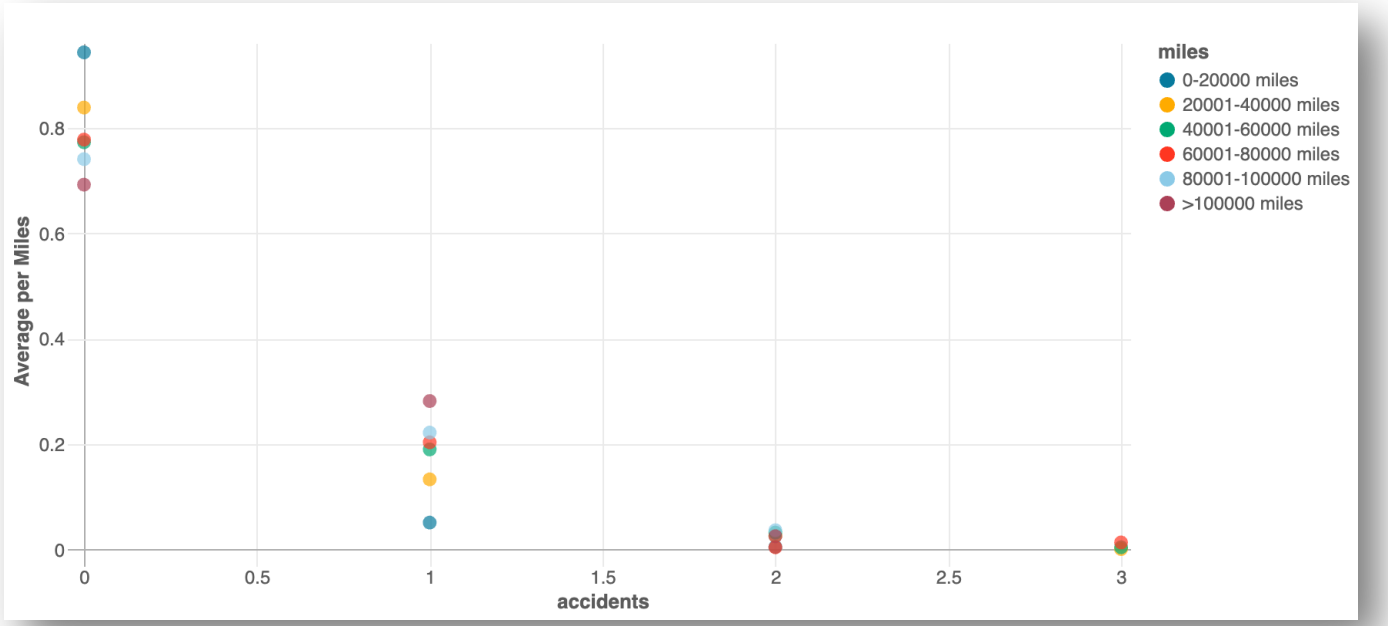
	ExteriorColor	Average
1	White exterior	0.1559633027522936
2	Red exterior	0.16483516483516483
3	Green exterior	0
4	Silver exterior	0.21524663677130046
5	Unknown exterior	0
6	Brown exterior	0.5
7	Black exterior	0.20809248554913296
8	Blue exterior	0.20279720279720279
9	Gray exterior	0.1942257217847769
10	Yellow exterior	0
11	Tan exterior	0.3333333333333333
12	Orange exterior	0



- As a result, we got this data (next slide) and we can also see in the graph that a higher percentage of brown cars were in accidents compared to the total number of brown cars

QUERY 5: RATIO OF ACCIDENTS TO CAR MILES

- For this query I wanted to test how the number of miles would affect the number of accidents a car has
- I first created a function miles() to which each car would fall in a specific car range. The ranges were 0-20000,20001-40000,40001-60000, 60001-80000,80001-100000 and greater than 100000
- Then I created a function to change the accidents to a number
- I then created 3 columns, one for the mile range and another for the accidents and the last for the average of accidents which is computed by dividing the grouped by count over the total number of cars in that range
- We can see from the results in the next slide that in the range 0-20000 miles 0 was the most with 95%. Cars having 1 accident were cars that exceeded 100000 miles. Cars having 2 accidents were cars in the 800001-100000 mile range. Cars having 2 accidents were cars in the 600001-80000 mile range.

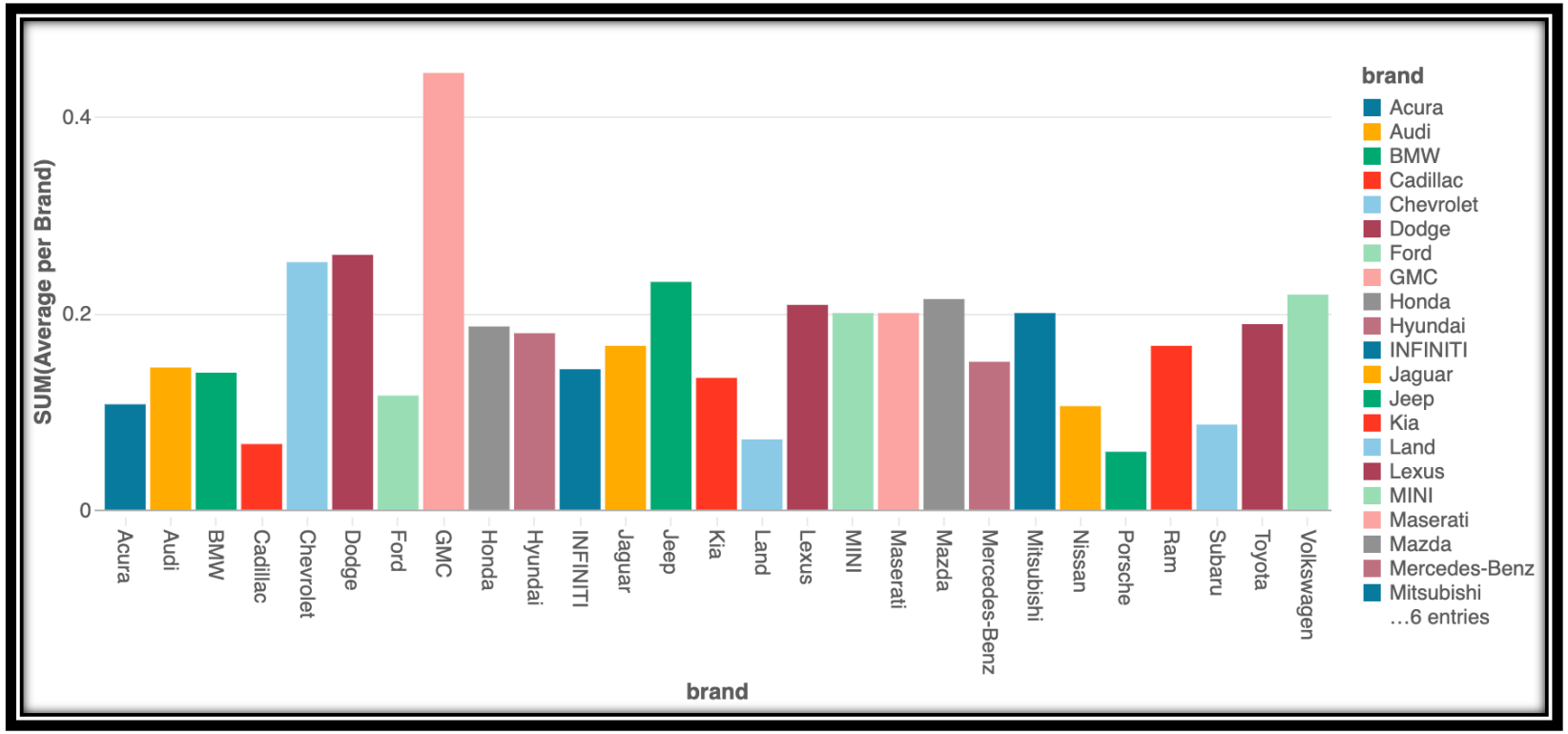


	miles	accidents	Average per Miles
1	0-20000 miles	0	0.9431616341030196
2	0-20000 miles	1	0.05150976909413854
3	0-20000 miles	2	0.0053285968028419185
4	20001-40000 miles	0	0.8384512683578104
5	20001-40000 miles	1	0.13351134846461948
6	20001-40000 miles	2	0.0267022696929239
7	20001-40000 miles	3	0.0013351134846461949
8	40001-60000 miles	0	0.7726218097447796
9	40001-60000 miles	1	0.1902552204176334
10	40001-60000 miles	2	0.03248259860788863
11	40001-60000 miles	3	0.004640371229698376
12	60001-80000 miles	0	0.7777777777777778
13	60001-80000 miles	1	0.2037037037037037
14	60001-80000 miles	2	0.004629629629629629
15	60001-80000 miles	3	0.013888888888888888
16	80001-100000 miles	0	0.7407407407407407
17	80001-100000 miles	1	0.2222222222222222
18	80001-100000 miles	2	0.037037037037037035
19	>100000 miles	0	0.6923076923076923
20	>100000 miles	1	0.28205128205128205
21	>100000 miles	2	0.02564102564102564

RESULTS FOR QUERY 5

QUERY 6: RATIO OF ACCIDENTS TO CAR BRAND

- For this query I wanted to test whether the brand of the car results in more accidents
- I began by creating a function car brand to split the brand of the car from its model in the first column
- I then computed the ratio of the accidents to the car whether 0,1,2,3 by grouping by the accidents and the brand
- As we can see in the next slides, GMC cars had 44% car accidents and surprisingly Porsche had the least car accidents with only 5%



RESULTS FOR QUERY 6

- The table for the results can be found in the file named result6.csv

QUERY 7: AVERAGE OF MILES FOR CARS EACH YEAR

- For my final query I want to know what the average number of miles for each manufacture year is. My assumption was that average would increase as the model year got older
- I started by creating the a function called miles_int(), which takes the miles column string and returns it as integer
- I then grouped by the manfacturing year and aggrgated with avg function in sql and printed out the result for each year

```
Average of miles for cars sold in 2023: 6730.937799043062
Average of miles for cars sold in 2022: 23621.300330033002
Average of miles for cars sold in 2021: 34246.58658346334
Average of miles for cars sold in 2020: 39405.172557172555
Average of miles for cars sold in 2019: 53368.174468085104
Average of miles for cars sold in 2018: 56504.96174863388
```