MAT259a project1 – MySQL & Knowledge Discovery

Russell Liu
#8314361

Introduction

The gasoline vehicle industry faces serious challenges as the EV (electric vehicle) industry grows rapidly over the last decade. As a person who loves the modification of vehicles and gasoline engines, I would love to have a look at the relationship between checkout times and people's enthusiasm for performance vehicles. It arouses my interest whether people borrow more or fewer books, which related to vehicles fields, from the library. After comparing the checkout times and sales in different brands and related fields, I also pay attention to the stock price market and try to find out if there is a connection between the checkout time and stock price.

Query 1

Here's my first query in Fig 1.1. I would like to see different brand sales in different years and put them together in a form. So, I choose to select year and group by year and track the total number of different brands that appeared in checkout books' titles. For the searching keyword, I separate them into two fields: one is about vehicle brands (includes Challenger, Mustang, Honda, Volkswagen), and the other one is about repair and maintenance (includes car repair, maintenance, and mechanic). I choose several representative brands from different nations: challenger and mustang from the U.S, Honda from Japan, and Volkswagen from Europe. It's quite easy to sum all the cases from 'outraw' list because it already provides me labels/ID for each checkout book.

```
1  SELECT YEAR(cout) as 'Year',
2  sum(CASE WHEN  instr(title, 'car' )>0 THEN   1 ELSE 0    END )'car',
3  sum(CASE WHEN  instr(title, 'vehicle' )>0 THEN   1 ELSE 0    END )'vehicle',
4  sum(CASE WHEN  instr(title, 'automotive' )>0 THEN   1 ELSE 0    END )'automotive',
5  sum(CASE WHEN  instr(title, 'challenger' )>0 THEN   1 ELSE 0    END )'challenger' ,
6  sum(CASE WHEN instr(title, 'mustang' )>0 THEN   1 ELSE 0    END )'mustang',
7  sum(CASE WHEN instr(title, 'honda' )>0 THEN   1 ELSE 0    END )'honda',
8  sum(CASE WHEN  instr(title, 'volkswagon' )>0 THEN   1 ELSE 0    END )'volkswagon',
9  sum(CASE WHEN  instr(title, 'car repair' )>0 THEN   1 ELSE 0    END )'car repair',
10 sum(CASE WHEN  instr(title, 'maintenance' )>0 THEN   1 ELSE 0    END )'maintenance',
11 sum(CASE WHEN  instr(title, 'mechanic' )>0 THEN   1 ELSE 0    END )'mechanic'
12  from  outraw
13  GROUP BY YEAR(cout)
```

Fig 1.1

Here is the result (Fig1.2):

Result Grid Filter Rows: Search Export:

Year	car	vehicle	automotive	challenger	mustang	honda	volkswagon	car repair	maintenance	mechanic
2006	90768	558	321	808	250	63	0	15	1129	1815
2007	89433	520	319	1179	188	75	0	12	1028	1905
2008	117968	666	413	1672	235	130	0	45	1276	2188
2009	124725	693	502	1313	225	94	0	52	1552	1965
2010	110395	587	461	983	179	117	0	19	1522	1574
2011	104095	538	385	1029	144	90	0	15	1323	1360
2012	100159	596	314	759	124	87	0	13	1106	1076
2013	108926	663	312	804	147	66	0	9	1336	1245
2014	100238	787	259	596	165	81	0	8	1078	1302
2015	93288	1065	211	490	153	56	0	4	968	1251
2016	90368	1362	211	628	455	40	0	10	837	1294
2017	80306	1255	216	655	283	49	0	24	625	2537
2018	55189	850	159	421	138	50	0	9	394	1099
2019	76721	1594	190	487	645	91	0	14	538	1290
2020	19950	458	53	180	181	25	0	1	174	315
2021	38443	948	75	342	207	35	0	5	313	611
2022	1764	45	3	16	11	1	0	0	15	31

Fig 1.2

Data visualization by Anaconda (Python):
Analysis - Part 1:
To analyze more detail, I choose to visualize the data and separate it into Brand checkouts and Repair. Firstly, let's look at Fig 1.3, which is the summary of Brand Checkouts. We could get such information for example: Challenger is no doubt American's favorite vehicle, and it makes sense because we are using the data from a U.S library. Take Ford Mustang as an example, from 2006 to 2021, the checkout times of Mustang is steady except in the Year 2014 to 2016. There is a dramatic increase and decrease in this period. Therefore, I fetch the data of Mustang sales from 2008 to 2021 from the "GoodCar Automotive Data&Statistic" website and visualize it in order to find out if such fluctuation is related to sales.

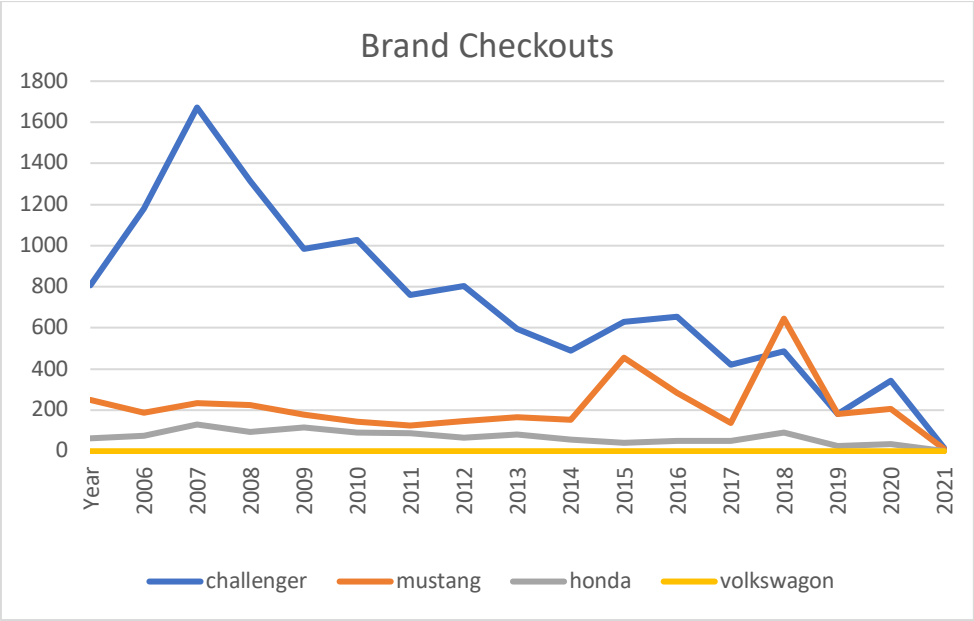

Fig1.3

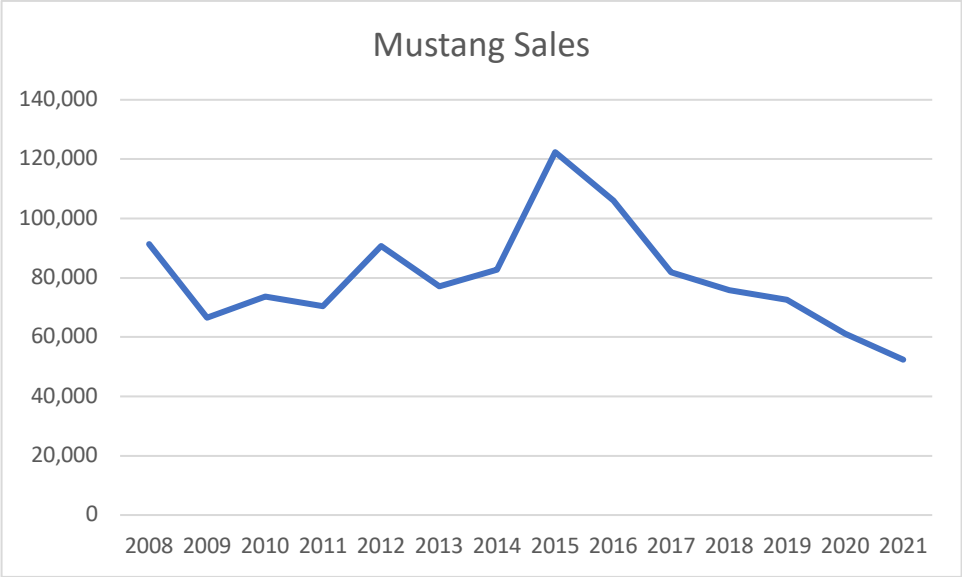

Fig 1.4

As we can see, in the period of 2014 to 2016, the fluctuation of sales looks very similar to the fluctuation of checkout times. And the reason behind this, I think it is all about the new release version of the Sixth generation of Mustang, including the brand new 2.3 Ecoboost I4 and 5.0 V8.

As a student majoring in Financial Mathematics, I wonder does it have some connection to the stock price? Therefore, I, again, fetch Ford's stock price data from Yahoo Finance. And here is the visualized data (Fig 1.5):

```
Out[83]: <AxesSubplot:xlabel='Date'>
```

Fig 1.5

We can clearly see that the stock price is continuously decreasing from 2014 to 2017. Theoretically, it shows that the product of Ford (in this case, Mustang) is not welcome in the market and the sudden increase of sales from 2014 to 2015 is contributed to the new release of the model. After months, it backed normally.

Part 2

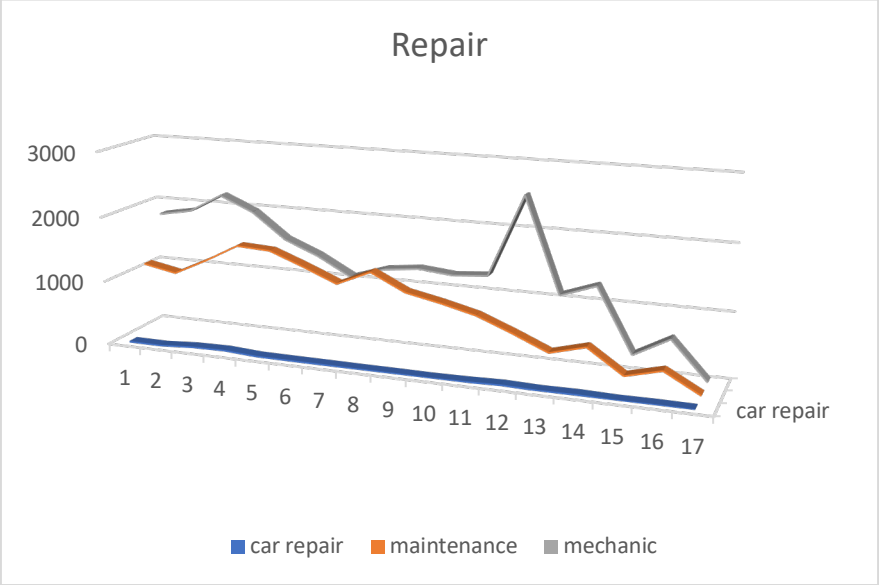

Fig 2.1

Now, here is the visualized data of field 2, which is about repair and maintenance. Similar to the situation of Brand checkout, Repair checkout times drop significantly.

Query 2

To look more directly at how people's general enthusiasm for vehicles correlates to checkout times, I wrote this query code (Fig 2.2) to calculate the sum of all fields that are related to vehicle and brands. To be more specific, it is a simplified version of the code in part 1. From the column of spl_2016.outraw, I try to find out all the checkout books whose titles have the following keywords: car, automotive, Honda, Mustang, Challenger, Volkswagen, and Subaru. Then, I sum the number of counts and groups by year from 2006 to 2022.

```sql
select (YEAR(cout)) AS year, COUNT(*) AS count
from spl_2016.outraw
where (
LOWER(title) like '% car %'
or LOWER(title) like '% automotive %'
or LOWER(title) like '% honda %'
or LOWER(title) like '% mustang %'
or LOWER(title) like '% challenger %'
or LOWER(title) like '% volkswagon %'
or LOWER(title) like '% subaru %'
)
GROUP BY YEAR(cout)
ORDER BY YEAR(cout)
```

Fig 2.2

CSV file Result (Fig 2.3):

year	count
2006	2764
2007	3172
2008	3791
2009	3557
2010	3130
2011	2873
2012	2257
2013	2540
2014	2278
2015	1891
2016	1886
2017	1936
2018	1028
2019	1489
2020	472
2021	949
2022	42

Fig 2.3

Data Visualization by Anaconda (Fig 2.4):

Fig 2.4

From Fig 2.4, we can see that the total number of checkout books related to (performance) vehicles is downtrend especially in year 2020 to 2022. One of the reasonable possibilities is that people are not borrowing books when there was a pandemic that happened in 2020. People stop buying vehicles, stop modifying vehicles, and thus stop borrowing related books from library.

Other than that, as a huge fan of Volkswagen GTI and owner of GTI MK7, it is upset to see that no one is borrowing any book that is related to Volkswagen!

Finally, I attach my code in Anaconda here.

```
In [81]: import pandas as pd
         df2 = pd.read_csv('ford.csv')
         print(df2)

                Date   Open   High    Low  Close  Adj Close    Volume
         0    2015-01-02  15.59  15.65  15.18  15.36  11.183619  24777900
         1    2015-01-05  15.12  15.13  14.69  14.76  10.746758  44079700
         2    2015-01-06  14.88  14.90  14.38  14.62  10.644824  32981600
         3    2015-01-07  14.78  15.09  14.77  15.04  10.950627  26065300
         4    2015-01-08  15.40  15.48  15.23  15.42  11.227304  33943400
         ..          ...    ...    ...    ...    ...        ...       ...
         750  2017-12-22  12.66  12.66  12.56  12.58  10.767381  17876200
         751  2017-12-26  12.57  12.65  12.55  12.60  10.784499  11664600
         752  2017-12-27  12.57  12.58  12.45  12.50  10.698908  17005600
         753  2017-12-28  12.48  12.58  12.47  12.58  10.767381  14793500
         754  2017-12-29  12.58  12.61  12.49  12.49  10.690348  18362500

         [755 rows x 7 columns]

In [82]: price = []
         for i in range(len(df2['Close'])):
             price.append(df2['Close'][i])

In [90]: result2 = pd.DataFrame(price).mean(axis=1)
         result2.index = df2["Date"]
         #result2.plot()

In [91]: import pandas as pd
         df = pd.read_csv('d5.csv')
         #print(df)

In [88]: sum_list = []
         for i in range(len(df['count'])):
             sum_list.append(df['count'][i])

In [92]: import matplotlib.pyplot as plt
         from datetime import datetime, timedelta
         plt.style.use('seaborn')

         dates = [
             datetime(2006, 12, 31),
             datetime(2007, 12, 31),
             datetime(2008, 12, 31),
             datetime(2009, 12, 31),
             datetime(2010, 12, 31),
             datetime(2011, 12, 31),
             datetime(2012, 12, 31),
             datetime(2013, 12, 31),
             datetime(2014, 12, 31),
             datetime(2015, 12, 31),
             datetime(2016, 12, 31),
             datetime(2017, 12, 31),
             datetime(2018, 12, 31),
             datetime(2019, 12, 31),
             datetime(2020, 12, 31),
             datetime(2021, 12, 31),
             datetime(2022, 12, 31),
         ]


         #plt.plot_date(dates, sum_list)
         #plt.tight_layout()
         #plt.plot_date(dates, sum_list, linestyle ='solid')
```

Conclusion

Based on what we analyzed above, it can be inferred that people's enthusiasm for performance cars and modifications is decreasing. It's highly possible that no one is interested in gasoline vehicles anymore in the future decades. In recent years, almost all the leading companies in the industry start working on new electric vehicles but not gasoline ones. However, I believe there is always someone, like me, who are interested in modification and gasoline engines. And I don't believe gasoline and performance vehicles will become history in the future. Finally, here I attach my MK7 with my friend's MK7. (Picture taken in freeway 154)