MIS 6346.502 - Big Data - S19

Big Data Project Report

Analysis of customer complaints using Big Data Analytical techniques

Snehal Thorat
5-6-2019

Contents

1. Introduction and problem description

Introduction:

The Consumer Financial Protection Bureau (CFPB) is a U.S. government agency that makes sure banks, lenders, and other financial companies treat you fairly. Each week CFPB send thousands of consumers' complaints about financial products and services to companies for response. Those complaints are published here after the company responds or after 15 days, whichever comes first. By adding their voice, consumers help improve the financial marketplace. To add some value to my analysis I shall be joining it with IRS data of 2016 and US Census data of 2010.

Problem Description:

a) Finding the top values for each of the following attributes based on the number of complaints: Date Received, Product, Sub-product, Issue, Sub-issue, Company Public Response, Company, State, Tags, Consumer Consent Provided, Submitted Via, Date Sent to Company, Company Response, Timely Response, Consumer Disputed

b) Comparing all the above factors for two companies having the maximum number of complaints for the same type of Product

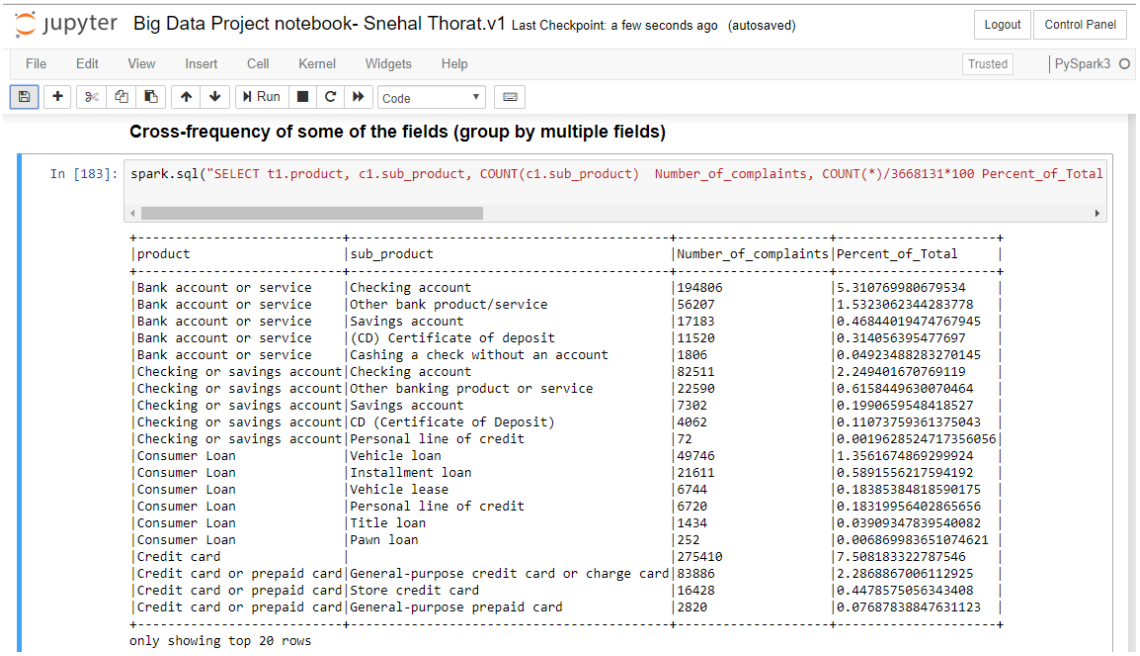
c) Finding the average number of complaints for each state based on the population

d) Checking if there is a correlation between average income in a zipcode and the number of complaints in that zipcode

2. Related work

All the related work done in Spark can be found in an IPython notebook and can be accessed at this link.

a) Finding the top values for each attribute based on the number of complaints:
Used Spark SQL queries displaying the top 20 results. Used SQL aggerate functions and CROSS JOIN.



b) Comparing all the above factors for two companies having the maximum number of complaints for the same type of Product
I ranked the companies based on the number of complaints by narrowing down the result to the product type having keyword credit.

company	Num_of_complaints	Rank_No_of_complaints
EQUIFAX, INC.	167092	1
Experian Information Solutions Inc.	149902	2
TRANSUNION INTERMEDIATE HOLDINGS, INC.	146188	3
CAPITAL ONE FINANCIAL CORPORATION	8364	4
CITIBANK, N.A.	5196	5

For doing a competitive analysis for two companies, I have selected 'EQUIFAX, INC.' and 'Experian Information Solutions Inc.' Following table shows the key distinguishing factors for both the companies:

Category	EQUIFAX, INC.	Experian Information Solutions Inc.
Product	Credit reporting has 154K complaints Credit reporting, credit repair services, or other personal consumer reports has 167K complaints	Credit reporting has 149K complaints Credit reporting, credit repair services, or other personal consumer reports has 153K complaints
Sub-Product	Credit reporting 165K complaints NULL(Missing) 154K complaints	Credit reporting 149K complaints NULL(Missing) 153K complaints
Top Issue	Incorrect information on credit report	Incorrect information on credit report
Top Sub-Issue	Information Belongs to someone else	Information Belongs to someone else
Date received	Maximum complaints were received in April 2018 On April 10, 2018: 564 complaints On April **19**, 2018: 420 complaints	Maximum complaints were received in April 2018 On April 10, 2018: 600 complaints On April **12**, 2018: 504 complaints
Company's public response	For almost 100% complaints: Company's public response is missing	For 65% complaints: Company has responded to the consumer and the CFPB and chooses not to provide a public response For 25%: Company's public response is missing
Tags	Missing 90% times When tag was provided: 55% times it was servicemember	Missing 90% times When tag was provided: 53% times it was servicemember
Consumer consent provided	Missing 43% times 50% times consent was not provided	Missing 43% times 50% times consent was not provided
submitted via	Use postal mail communication channel less number of times (used 10%) than Experian Information Solutions Inc.	Use postal mail communication channel more number of times (used 12%) than Experian Information EQUIFAX, INC.
Company response	66 complaints were closed with monetary relief which are only 0.02% of the total complaints	1650 complaints were closed with monetary relief which is 0.2% of the total complaints
Timely response	98.7% times	100% times
Consumer disputed	38.5% times	44% times
Zip_code	Maximum complaints are in 33122 - Postal code in Miami-Dade County, Florida 5th highest complaints are from 18936 - Postal code in the Montgomery County, Pennsylvania	Maximum complaints are in 33122 - Postal code in Miami-Dade County, Florida 5th highest complaints are from 79105 - Postal code in Moore County, Texas

c) Finding the average number of complaints for each state based on the population

I plotted the total number of complaints aggregating for each state and observed that for both Experian Information Solutions Inc. & EQUIFAX, INC. the maximum complaints came from California and Texas. This result is skewed because of the population in these two states is the maximum. Hence for doing an geographical analysis, I have calculated the average number of complaints based on the population at the granularity of zip code.

For finding the average number of complaints I joined the complaints dataset with the census data on the 5digit zip code. I divided the number of complaints for a zip code by the population in that zip code using a scaling factor of 10000 and rounded it to the nearest integer number using Ceil function.

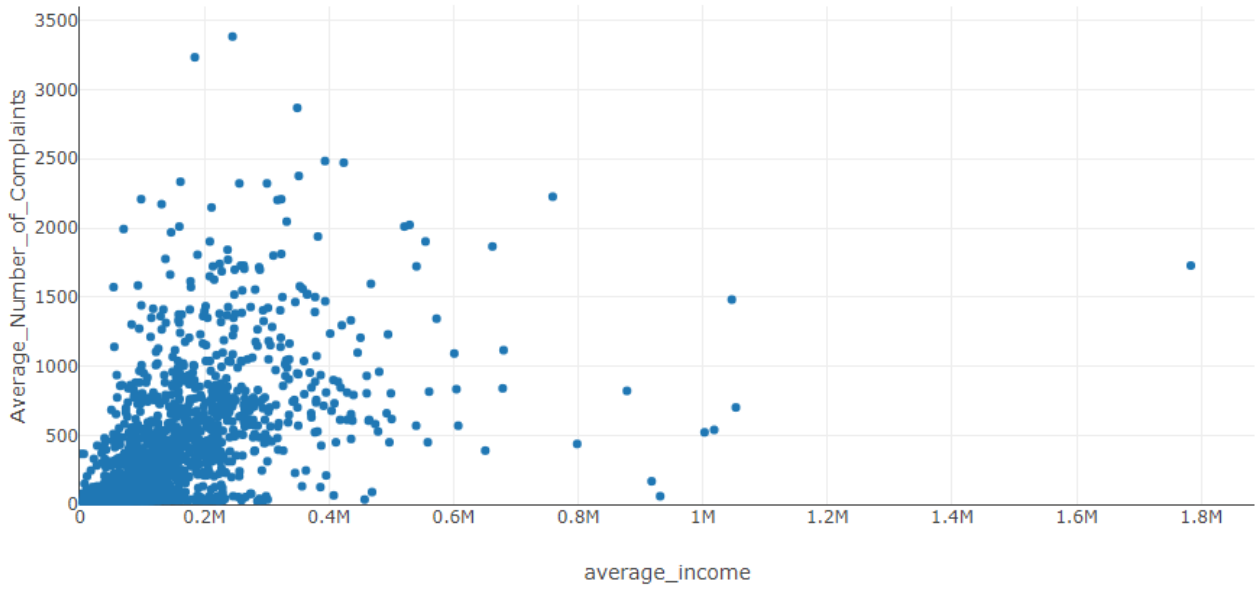
Based on the average number of complaints the top 5 states having maximum average number of complaints are Florida, California, Texas, Georgia, New York for both EQUIFAX, INC. and Experian Information Solutions Inc.

d) Checking if there is a correlation between average income in a zip code and the number of complaints in that zip code

I wanted to check if income affects the number of complaints, hence I joined the complaints dataset with the IRS dataset on the zip code and calculated the average income for a zipcode

I selected a new aggregated DataFrame for average number of complaints and average income.

Based on the correlation matrix I got the Pearson correlation as 0.6 and Spearman correlation as 0.8. This indicates that there is a strong correlation between income and the average number of complaints for most of the zip codes. The similar result can be seen in the scatter plot below:



3. Pre-processing techniques

Pre-processing steps can be found in IPython Notebook by following this link of my GitHub repository.

While creating the EMR cluster, I entered following property. This helped in Spark memory allocation, specially while writing the table in parquet format and storing in local HDFS or S3 bucket:

```
[
 {
   "Classification": "spark",
   "Properties": {
    "maximizeResourceAllocation": "true"
   }
 }
]
```

Preprocessing:

For CFPB dataset:

- Downloaded the CFPB data on HDFS directory
- Cleaned the data by running a Python file from HDFS command line
- Stored processed file in my S3 bucket: 'snehalrawinput'.
- Created an external Hive table pointing to this S3 bucket and loaded the data in Hive table.
- Stored the data from Hive table to Spark Dataframe
- Saved the DataFrame as a Parquet file in my S3 bucket: 'snehalpreprocessedfiles'

For Census Data:

- Downloaded the census csv file for 2010 from Kaggle
- Uploaded this file in my S3 bucket: 'snehalrawinput'.
- Created an external Hive table pointing to this S3 bucket and loaded the data in Hive table.
- Stored the data from Hive table to Spark Dataframe

For IRS data:

- Downloaded the IRS data for 2016 from the IRS website
- Uploaded this file to my S3 bucket: 'snehalrawinput'
- Loaded the csv file in Jupyter notebook's Spark DataFrame using spark.read.format

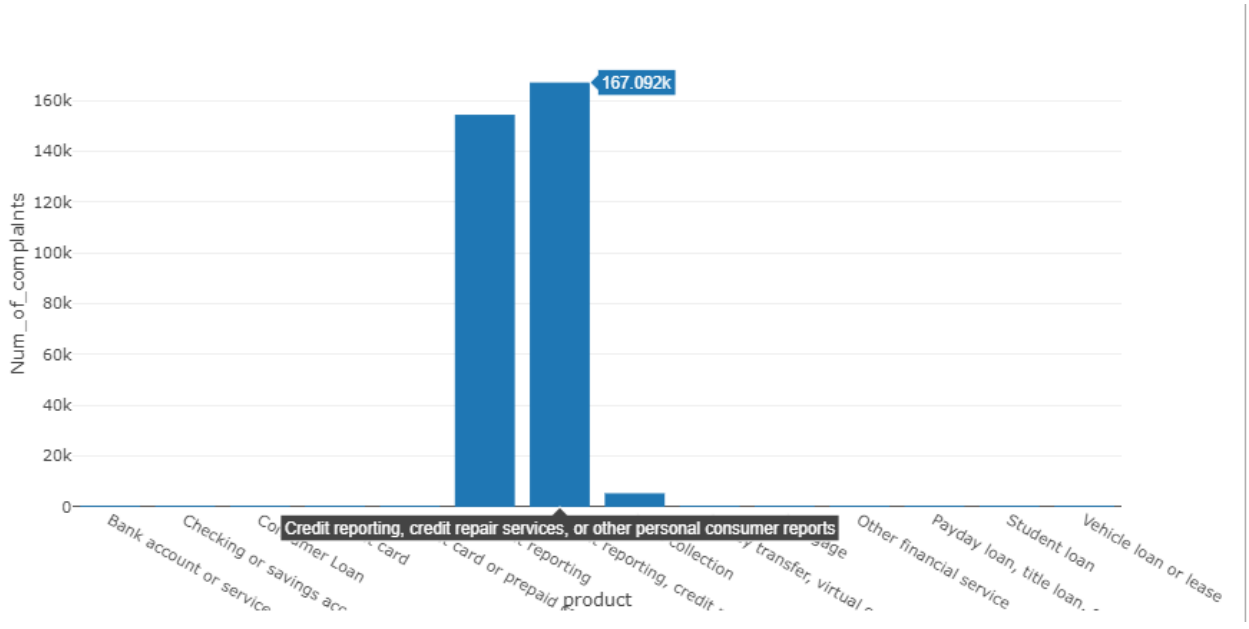
Transformation:

- Combined the CFPB dataset with population and Income dataset using right join using Spark DF APIs, to retain the information of the primary CFPB dataset
- On the subset dataset of 'EQUIFAX, INC.' and 'Experian Information Solutions Inc.', used regexp_replace to replace '/' with '-' in the date received column. Used F.split to extract the year and month part of the date received
- Renamed the columns of the Spark DataFrame
- Created calculated columns using Spark SQL queries for average_income, Number_of_Complaints, and Average_Complaints
- Saved the transformed DataFrame as a Parquet file in my S3 bucket: 'snehalpreprocessedfiles' for each step and version

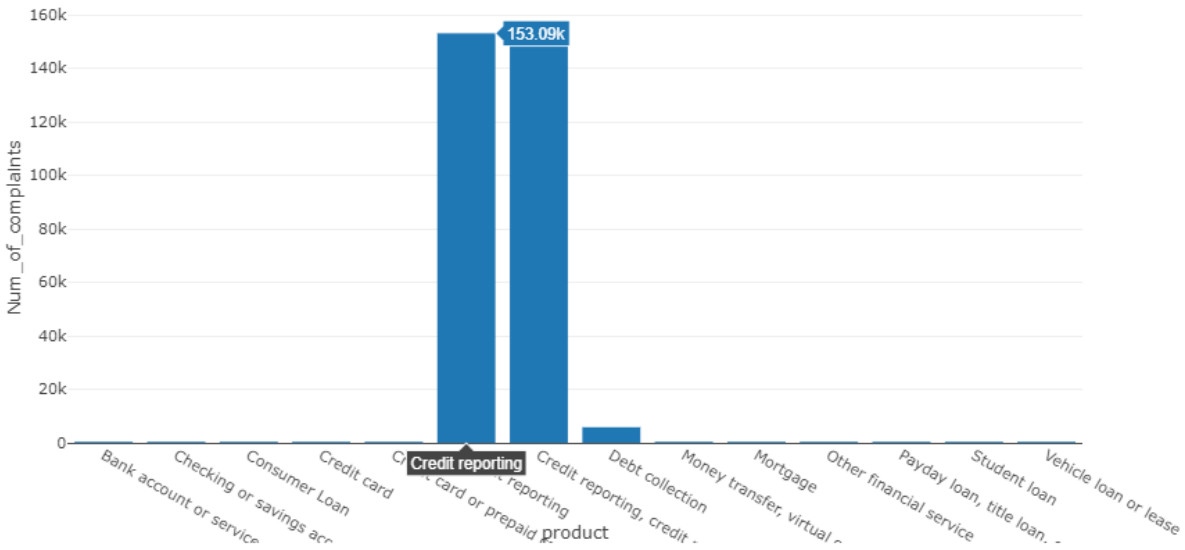
4. Visualizations

Products

EQUIFAX, INC.

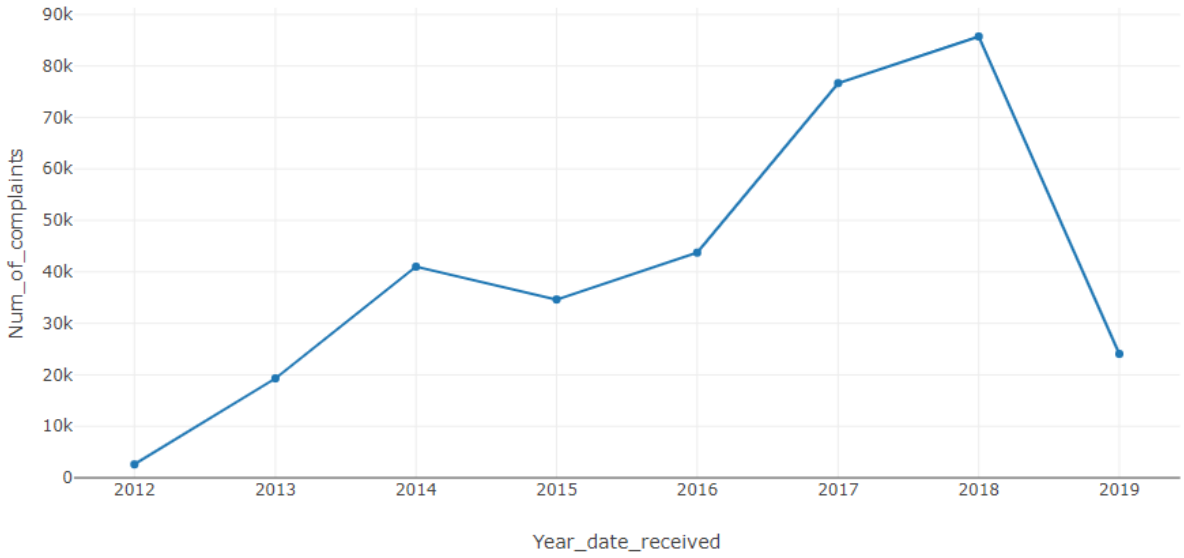


Experian Information Solutions Inc.

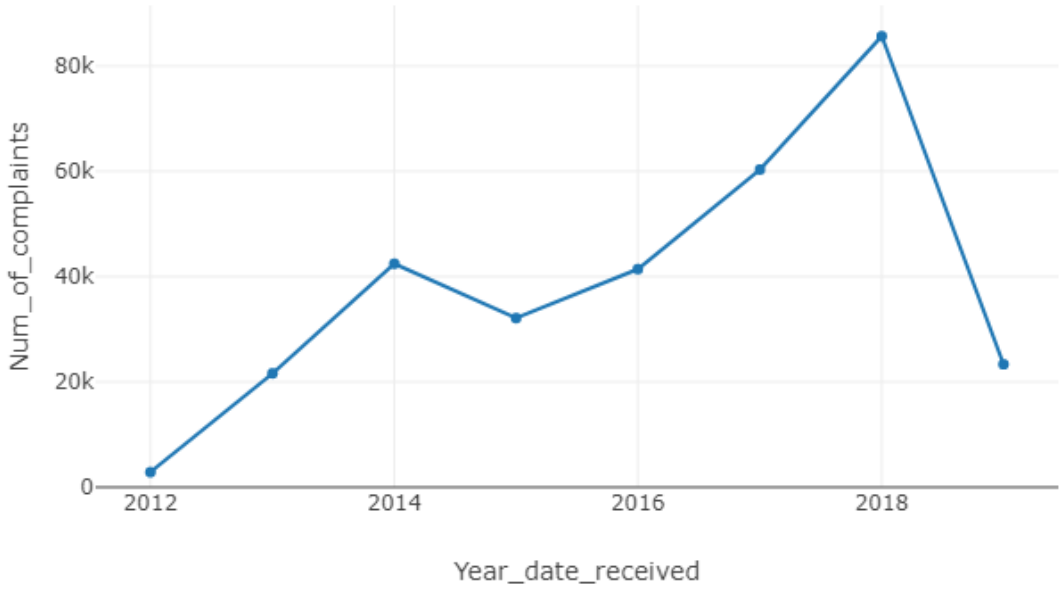


Year

EQUIFAX, INC.

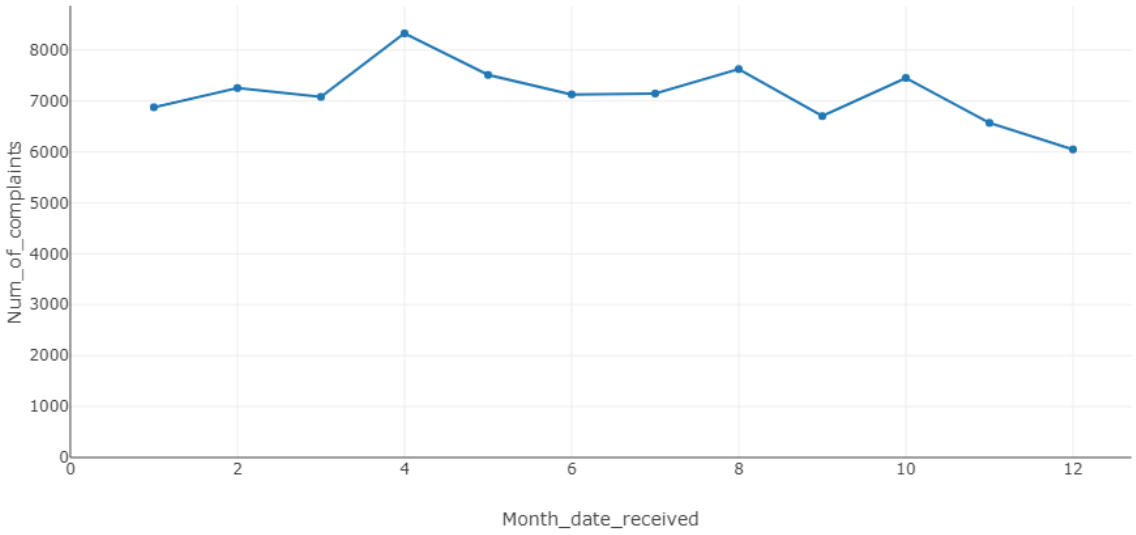


Experian Information Solutions Inc.

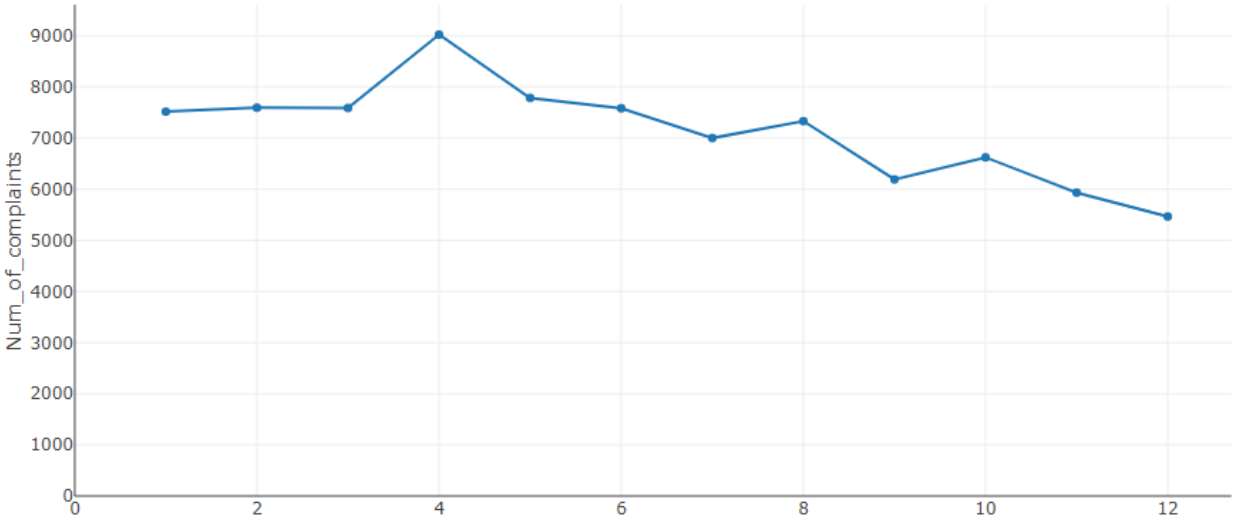


Monthly analysis for April 2018

EQUIFAX, INC.

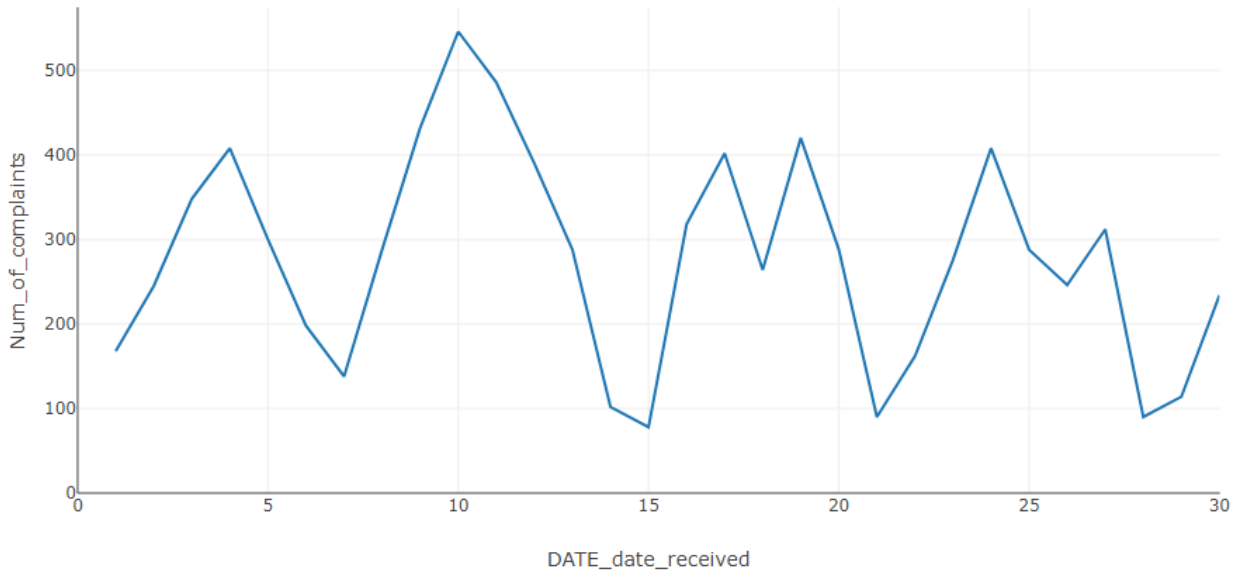


Experian Information Solutions Inc.

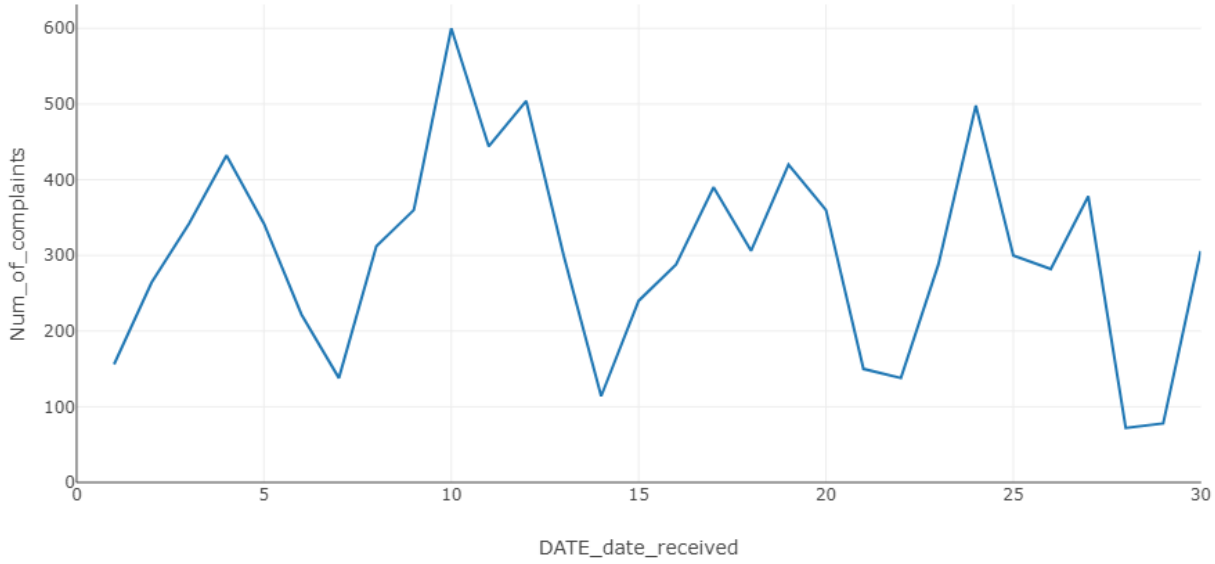


Daily analysis for April 2018

EQUIFAX, INC.



Experian Information Solutions Inc.



company_public_response

EQUIFAX, INC.

company_public_response for both companies

```
%%sql -o query2

select company_public_response, COUNT(*) Num_of_complaints FROM experian_equi_date_year_month \
WHERE company = 'EQUIFAX, INC.' \
GROUP BY company_public_response ORDER BY company_public_response
```

Type: Table Pie Scatter Line Area Bar

company_public_response	Num_of_complaints
	327710
Company has responded to the consumer and the ...	6

Experian Information Solutions Inc.

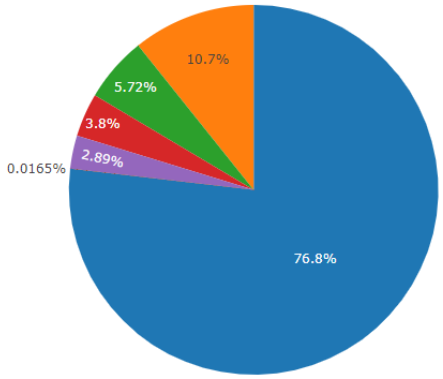
```
%%sql -o query2

select company_public_response, COUNT(*) Num_of_complaints FROM experian_equi_date_year_month \
WHERE company = 'Experian Information Solutions Inc.' \
GROUP BY company_public_response ORDER BY company_public_response
```

Type: Table Pie Scatter Line Area Bar

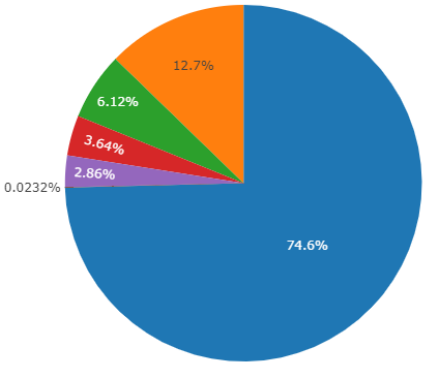
company_public_response	Num_of_complaints
	79428
Company believes complaint caused principally ...	174
Company believes it acted appropriately as aut...	24
Company believes the complaint is the result o...	18
Company can't verify or dispute the facts in t...	6
Company chooses not to provide a public response	26388
Company has responded to the consumer and the ...	203752

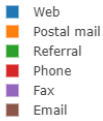
submitted_via

EQUIFAX, INC.



Experian Information Solutions Inc

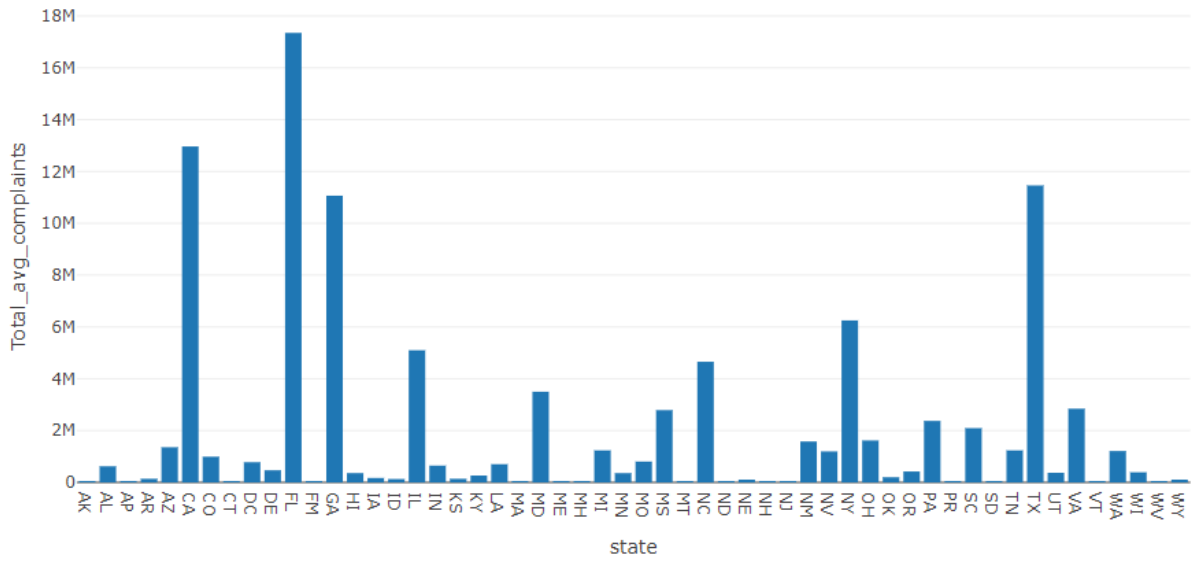




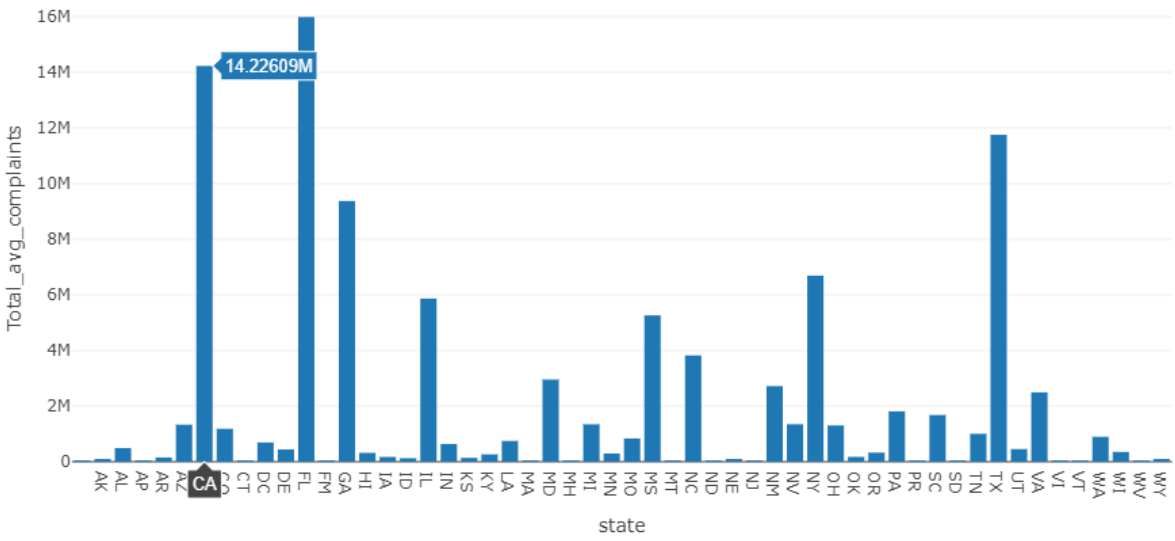
STATE-WISE ANALYSIS

Average complaints

EQUIFAX, INC.



Experian Information Solutions Inc.



5. Conclusion

- EQUIFAX, INC. had more complaints for Credit reporting than Experian Information Solutions Inc.
- On April 10, 2018 Experian Information Solutions Inc. had more complaints than EQUIFAX, INC.
- EQUIFAX, INC. had never provided a public response
- Experian Information Solutions Inc. closed 1650 complaints with monetary relief.
- Experian Information Solutions Inc. always provided timely response
- Average complaints after normalizing based on population provides better insights than using just the count of number of complaints
- Average income and average number of complaints are strongly correlated

6. References

Dataset sources:

- Primary dataset: CFPB data – https://www.consumerfinance.gov/data-research/consumer-complaints/#download-the-data

- Secondary datasets –
    1. IRS data (2016): https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-2016-zip-code-data-soi
    2. Census Data (2010): https://www.kaggle.com/census/us-population-by-zip-code/version/1#

References:

- Big Data class Jupyter notebooks, notes, and assignments
- https://spark.apache.org/docs/2.2.0/ml-statistics.html#correlation
- https://stackoverflow.com/questions/52214404/how-to-get-the-correlation-matrix-of-a-pyspark-data-frame
- Overview - Spark 2.4.2 Documentation - Apache Spark - Spark SQL, DataFrames and Datasets Guide: https://spark.apache.org/docs/latest/sql-programming-guide.html#spark-sql-dataframes-and-datasets-guide