



## Project Proposal- Complaint analysis



Manisha Gupta

MXG176230 UTD ID - 2021399372

# Manisha Gupta

## [Contents](#)

Introduction.....	2
Project Description.....	2
Descriptive analysis .....	2
Additional Datasets used .....	3
Techniques used .....	4
Performed Work.....	4
Python code to change the delimiter of the file.....	4
Preprocessing techniques .....	4
Preliminary study of the Datasets.....	5
Univariate Analysis.....	5
Products .....	5
Subproduct.....	6
Companies.....	6
State.....	7
Timely response .....	8
Tags .....	8
Issue .....	8
Sub Issue .....	8
Company's public response .....	9
Consumer disputed .....	9
Submitted Via.....	10
Company Response .....	10
Time taken to send the complaint to company.....	10
Average time taken to send the complaints per state .....	12
ZipCode .....	12
Bivariate Analysis.....	13
Products against Timely response.....	13
Products with submission medium.....	13
Top 3 customer concerns related to Products per state .....	13
Top 3 customer concerns per state .....	14
Products Vs Year.....	14
Submitted vs Year .....	14
Timely Response for each Year.....	14
Consumer Dispute - Year .....	15
Issue with Timely response: Generally issues with high complaints are getting timely response.....	15
Consumer Complaints with IRS Datasets.....	15
Sneak Peak in to IRS Dataset.....	15
Consumer Complaints with Population Dataset.....	16
Sneak Peak in to Population Data.....	16
Total complaints and population per area .....	17
Using Hive .....	18
Reference .....	20

## Introduction

The Consumer Complaint Database is a collection of complaints on a range of consumer financial products and services, sent to companies for response. It was designed to connect consumers with financial companies to understand issues with their mortgages, fix errors on their credit reports, stop harassment from debt collectors, and get direct responses about problems with their credit cards, checking and savings accounts, student loans, and more.

The objective of this database is, by analyzing the data, to identify trends and problems in the marketplace and to help to do a better job in supervising companies, enforcing federal consumer financial laws and writing rules and regulations.

- Each complaint is forwarded to the appropriate company for a response.
- The complaint data is shared with state and federal agencies.
- The complaint data is then analyzed to help with our work to supervise companies, enforce federal consumer financial laws, and write better rules and regulations.

Financial products and services sector have significant impact on our lives as they are directly related to our finances. It becomes need of hour to have proper understanding about how we can make correct decisions to buy a product or a service. Here in this project we analyze Customer complaints about various financial products to answer these questions.

## Project Description

The objective of this project Complaint analysis is to track, categorize and handle customer complaints. When a customer registers a complaint, he or she is voicing a concern in relation to a product or service. However, not all complaints are to be treated equally and there are several questions to ask before taking action such as

- Has this happened before?
  - Have the complaints been recorded?
  - How often does the same complaint arise?
  - Is there a pattern to this complaint in how it was received?
  - Has the same customer reported this previously?
1. Analyzing the major customer concerns with banks in each state. We aim to gain an insight into the problems coming from retail banking in each state in the US. This can be derived from the available dataset and the goal is to use classifying the customer concerns based on the state, will give the major concerns in each state and we aim to find the top 3 concerns of each customer.
  2. Derive business impact of customer concerns to banking institutions Classifying the available data based on whether customers' concerns were addressed or not will provide valuable insight into the effect of them on institutions' business. For example, a large number of unresolved complaints can be taken to mean that the specific customers have taken their business elsewhere.
  3. Develop a performance metric based on time taken to resolve concern Depending on the type of concern and the time taken to resolve it, we can come up with a standardized performance metric, taking into account the various factors of the case and apply the metric across all future solutions. This will help pinpoint those problems that take longer to resolve and therefore can be given a higher priority.
  4. Perform the analysis to study the relationship of income with number of complaints.
  5. Study the impact of age, gender or household on complaints datasets in detail.

## Descriptive analysis

Data sets is available on <https://www.consumerfinance.gov/data-research/consumer-complaints/> and its field description is <https://www.consumerfinance.gov/complaint/data-use/>.

## Manisha Gupta

Old Field Name	FIELD NAME	DESCRIPTION
Date received	datereceived	The date the CFPB received the complaint. For example, "05/25/2013."
Product	product	The type of product the consumer identified in the complaint. For example, "Checking or savings account" or "Student loan."
Sub-product	subproduct	The type of sub-product the consumer identified in the complaint. For example, "Checking account" or "Private student loan."
Issue	issue	The issue the consumer identified in the complaint. For example, "Managing an account" or "Struggling to repay your loan."
Sub-issue	subissue	The sub-issue the consumer identified in the complaint. For example, "Deposits and withdrawals" or "Problem lowering your monthly payments."
Consumer complaint narrative	narrative	Consumer complaint narrative is the consumer-submitted description of "what happened" from the complaint. Consumers must opt-in to share their narrative. We will not publish the narrative unless the consumer consents, and consumers can opt-out at any time. The CFPB takes reasonable steps to scrub personal information from each complaint that could be used to identify the consumer.
Company public response	publicresponse	The company's optional, public-facing response to a consumer's complaint. Companies can choose to select a response from a pre-set list of options that will be posted on the public database. For example, "Company believes complaint is the result of an isolated error."
Company	company	The complaint is about this company. For example, "ABC Bank."
State	state	The state of the mailing address provided by the consumer.
ZIP code	zipcode	The mailing ZIP code provided by the consumer. This field may: i) include the first five digits of a ZIP code; ii) include the first three digits of a ZIP code (if the consumer consented to publication of their complaint narrative); or iii) be blank (if ZIP codes have been submitted with non-numeric values, if there are less than 20,000 people in a given ZIP code, or if the complaint has an address outside of the United States).
Tags	tags	Data that supports easier searching and sorting of complaints submitted by or on behalf of consumers.
		For example, complaints where the submitter reports the age of the consumer as 62 years or older are tagged "Older American." Complaints submitted by or on behalf of a servicemember or the spouse or dependent of a servicemember are tagged "Servicemember." Servicemember includes anyone who is active duty, National Guard, or Reservist, as well as anyone who previously served and is a veteran or retiree.
Consumer consent provided?	consent	Identifies whether the consumer opted in to publish their complaint narrative. We do not publish the narrative unless the consumer consents, and consumers can opt-out at any time.
Submitted via	submitted	How the complaint was submitted to the CFPB. For example, "Web" or "Phone."
Date sent to company	datesent	The date the CFPB sent the complaint to the company.
Company response to consumer	responsetoconsumer	This is how the company responded. For example, "Closed with explanation."
Timely response?	timelyresponse	Whether the company gave a timely response. For example, "Yes" or "No."
Consumer disputed?	disputed	Whether the consumer disputed the company's response.
Complaint ID	complaintid	The unique identification number for a complaint.

All the fields except complaints ID, consumer complaint narrative are important as we can analyze this dataset by all such as Year, Product, Company, state, Complaints over time-

Most important fields are state, zipcode, product, issue, timely response and how the complaints are submitting to the database to get the more valuable insights. we will be performing analysis by following-

- Total number of complaints
- Types of complaints
- Total complaints by year

## Additional Datasets used

1. **IRS data by zip code:** A detailed description of the columns contained within the dataset can be found at: <https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-2016-zip-code-data-soi>
2. **2010-census-populations-by-zip-code-** Data can be found on <https://www.kaggle.com/census/us-population-by-zip-code/version/1> and its field names are self-explanatory.

## Manisha Gupta

### Techniques used

- Hive aggregation queries
- HDFS
- Pyspark
- Graphs using aggregated data(optional)
- Segmentation analysis using ML(optional)

### Performed Work

#### Python code to change the delimiter of the file

```
import os as os

import pandas as pd

import datetime as dt

import numpy as np

import sys

import glob

import csv

import re

if __name__ == '__main__':

    filename = r"C:\Users\abhin\Documents\Manisha\utd\bigdata\complaints.csv"

    dfctest = pd.read_csv(filename, header = 0, quoting=csv.QUOTE_ALL )

    dfcolumns = dfctest.columns

    cnt = 0

    for tp in dfctest.dtypes:

        colname = dfcolumns[cnt]

        if 'object' in str(tp):

            if "date" not in colname.lower():

                dfctest[colname] = dfctest[colname].astype(str).replace('~', ' ').replace("'", "").map(lambda x: re.sub(r'\W+', ' ', x))

                dfctest[colname] = dfctest[colname].apply(lambda x: x if x!='nan' else '')

            cnt = cnt + 1

    dfctest.to_csv(path_or_buf = "complaintsdata.csv", sep='~', index = False)
```

#### Preprocessing techniques

Removed null values from all the columns and uses of split, cross tab, pivot, string, and date function used to get the relevant information of each column using Spark API, Spark SQL. Pyspark stats libraries has also been used for analysis.

**Important Columns: -**

Manisha Gupta

- 1. Consumer Complaints data sets:** - Most important fields are state, zipcode, product, issue, timlyresponse, company, date received, tags, issue and how the complaints are submitting to the database to get the more valuable insights.
- 2. IRS Data sets:** -
  - column "N1" – total number of return filed
  - NUMDEP - total number of dependent
  - A02650 – Total income amount

These three columns have been used to study the relationship between income data with number of complaints

### 3. Zipcode population Data:-

- Population – Total population per zip code
- Gender

These two columns have been used to study the relationship between population data with number of complaints

## Preliminary study of the Datasets

Consumer complaints data sets has 18 columns and 1257521 records.

```
In [3]: df_complaints.cache()
df_complaints.count()

Out[3]: 1257521
```

```
In [27]: df_complaints.columns

Out[27]: ['Datereceived',
'Product',
'Subproduct',
'Issue',
'Subissue',
'narrative',
'Comppublicresponse',
'Company',
'State',
'ZIPcode',
'Tags',
'consent',
'Submittedvia',
'Datesentcomp',
'Compresponseconsumer',
'Timelyresponse',
'Consumer disputed',
'ComplaintID']
```

## Univariate Analysis

### Products

What are the different products?

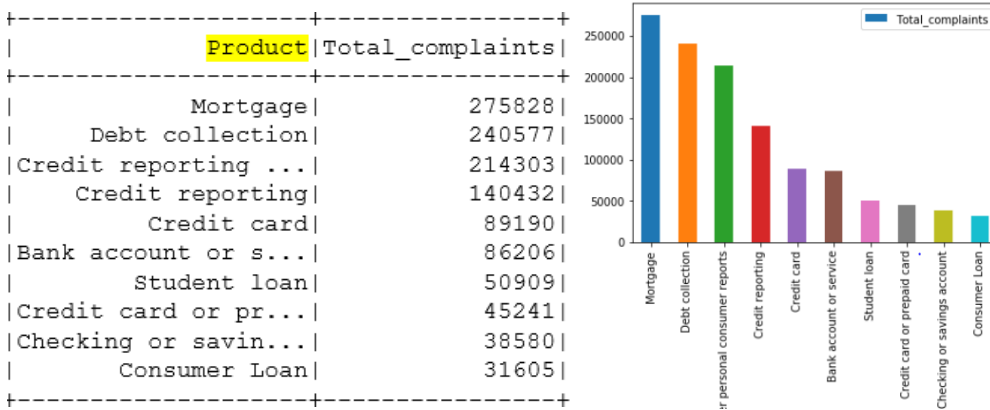
```
+-----+
|count(DISTINCT Product)|
+-----+
|                        |18|
+-----+
```

How many complaints raised against each product?

```
+-----+-----+
|Product|Total_complaints|
+-----+-----+
|Virtual currency|18|
|Other financial s...|1059|
|Prepaid card|3819|
|Money transfers|5354|
|Payday loan|5544|
|Payday loan title...|8286|
|Money transfer vi...|9699|
|Vehicle loan or l...|10871|
|Consumer Loan|31605|
|Checking or savin...|38580|
|Credit card or pr...|45241|
|Student loan|50909|
|Bank account or s...|86206|
|Credit card|89190|
|Credit reporting|140432|
|Credit reporting ...|214303|
|Debt collection|240577|
|Mortgage|275828|
+-----+-----+
```

Top products with highest number of complaints

Manisha Gupta

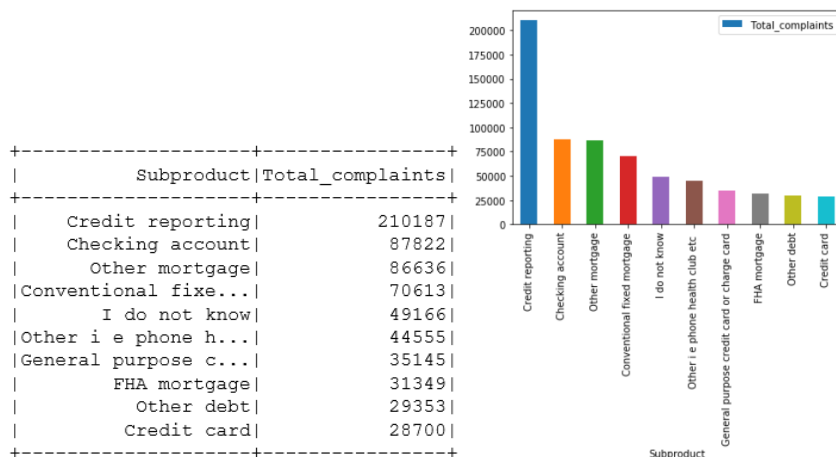


### Top 5 Products with Lowest number of complaints received

Product	Total_complaints
Virtual currency	18
Other financial s...	1059
Prepaid card	3819
Money transfers	5354
Payday loan	5544

## Subproduct

There are 76 sub products. Credit card and checking account have a greater number of complaints.



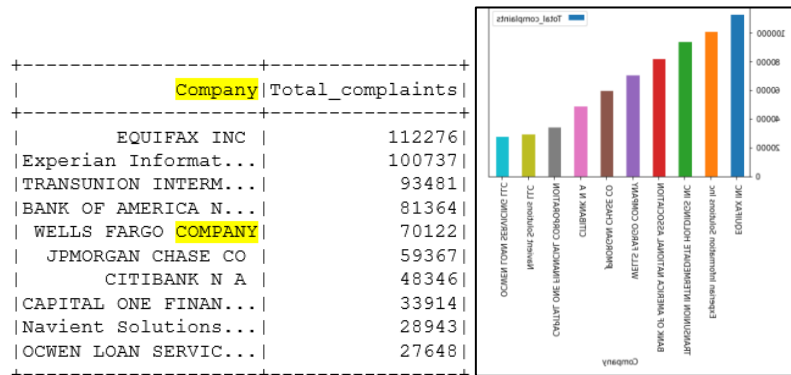
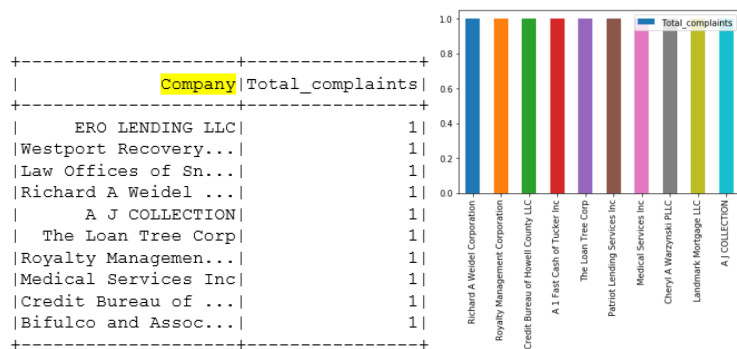
## Companies

- How many different companies?

```
count (DISTINCT Company) |
5254 |
```

- What are the top 10 and bottom 10 companies based on the number of complaints?

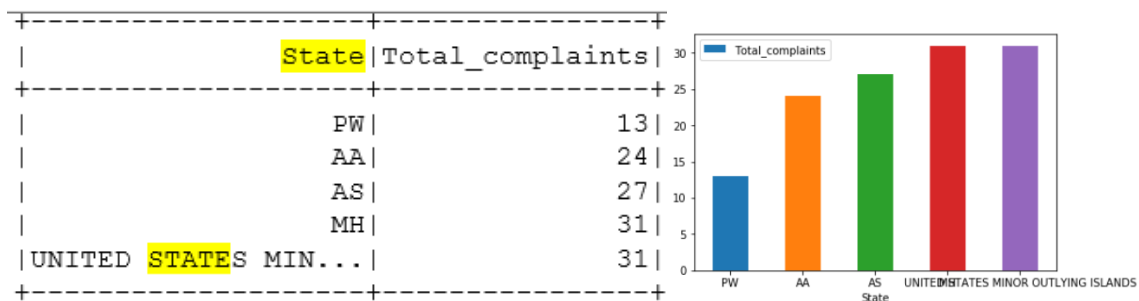
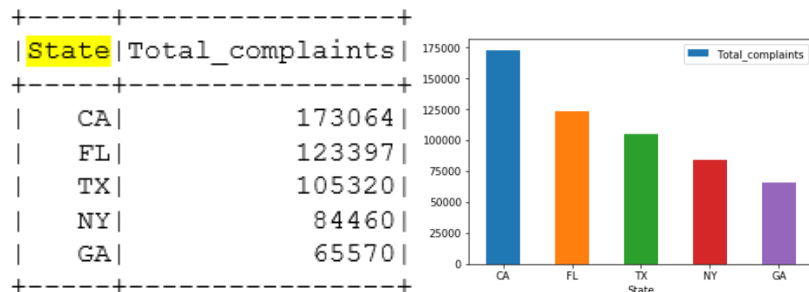
## Manisha Gupta



Biggest banks have thousands of complaints while the ones at the bottom are smaller ones which is obvious.

## State

- Top 5 states with highest and lowest number of complains



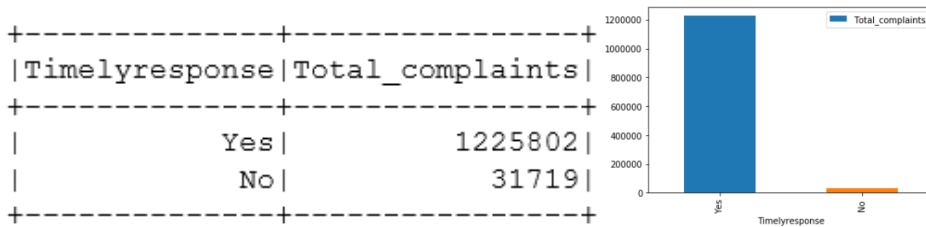
There are more than 50 states in the State column. These include even the federal as well as inhabited territories. Also, the states with most complaints look like the ones which are highly populated and have strong economy in USA



## Manisha Gupta

### Timely response

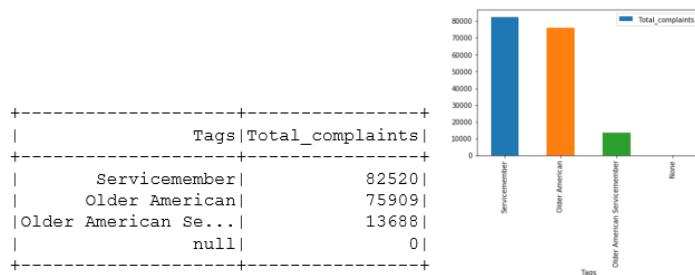
- How many complaints have received timely response?



Almost 97.5% of the complaints have been addressed timely. Given the competition in the financial sector and also the challenges posed by the FINTECH start-ups established players have to do lot more to resolve the complaints.

### Tags

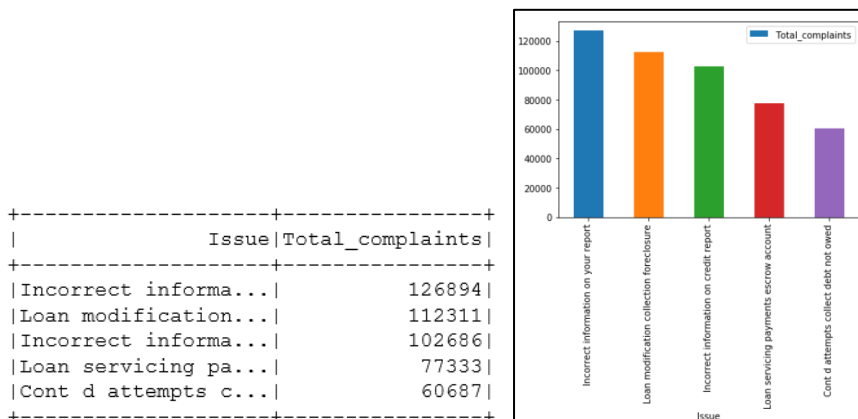
Servicemember use to file more complaints as compared to others.



### Issue

There are total of 165 type of issues. Incorrect informaton and loan modification are the main concerned for the consumers.

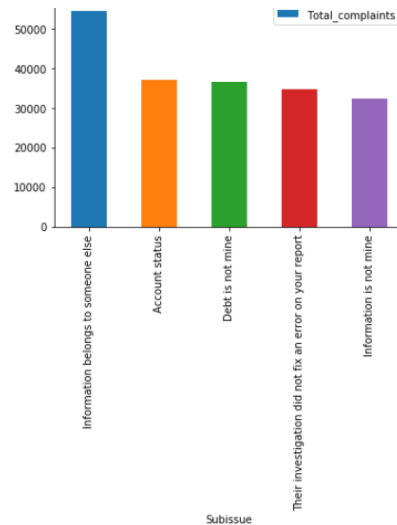
count(DISTINCT Issue)
165



### Sub Issue

There are total 218 type of sub issue have been registered from 2011 to March 2019. Sub issue" Information belongs to someone else" got maximum number of complaints during 2011 to till March 2019

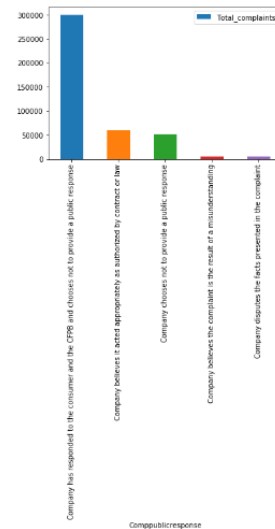
Subissue	Total_complaints
Information belongs to someone else	54677
Account status	37057
Debt is not mine	36731
Their investigation did not fix an error on your report	34765
Information is not mine	32384



## Company's public response

23% of total complaints - companies has responded to consumer and CFPB but did not provide public response approx and approx. 4% of total complaints, they have not provided public response at all.

Comppublicresponse	Total_complaints
Company has responded to the consumer and the CFPB and chooses not to provide a public response	300863
Company believes it acted appropriately as authorized by contract or law	59925
Company chooses not to provide a public response	52473
Company believes the complaint is the result of a misunderstanding	5574
Company disputes the facts presented in the complaint	5171

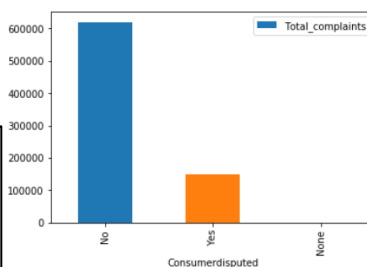


## Consumer disputed

- How many consumers disputed?

Over 20% of the users disputed.

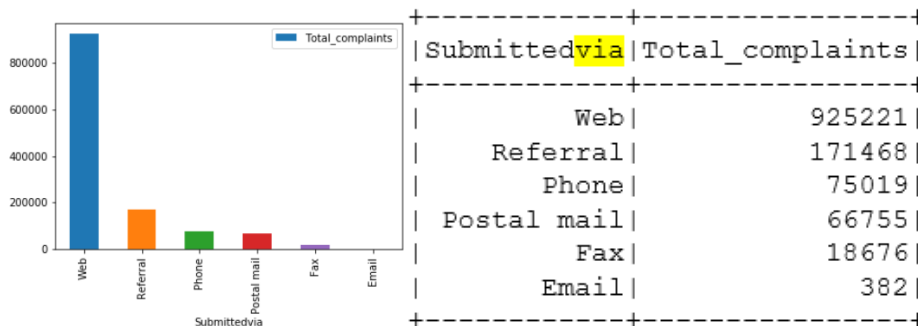
Consumerdisputed	Total_complaints
No	620127
Yes	148378



## Manisha Gupta Submitted Via

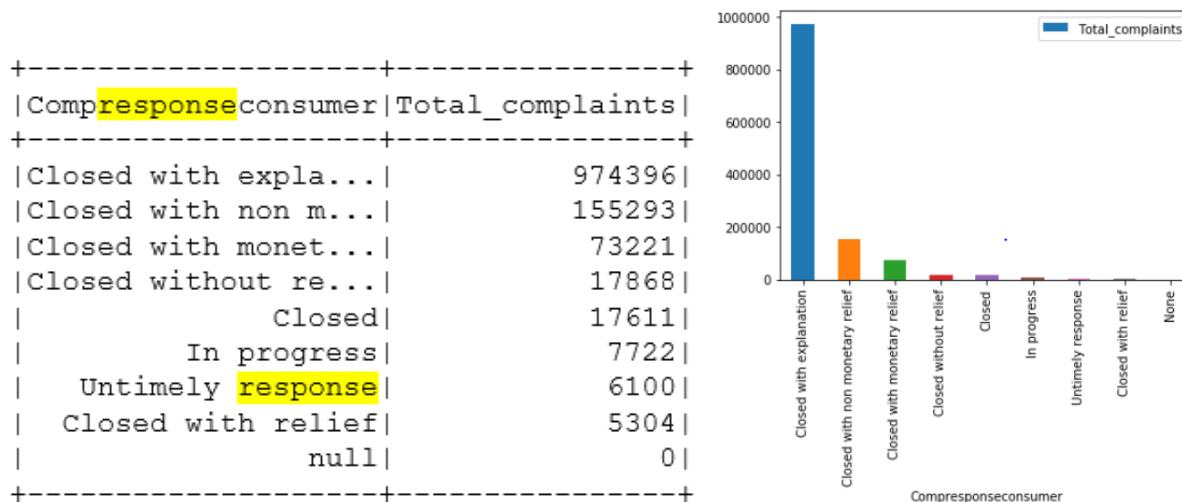
- How are these complaints submitted?

Majority of the complaints are filed through the web and least method used is email.



## Company Response

- What are the different responses users get from the companies?



Majority are closed with explanation and around 12% are closed with non-monetary relief.

## Time taken to send the complaint to company

- Average time taken to send the complaint to the respective company is approx. 2 business day. One complaint took 1847 days to be sent to company.

summary	dif
count	1257521
mean	2.4643166992837497
stddev	30.28485509611261
min	-336
max	1847

## Percentile distribution of Time taken to send the complaint to company

## Manisha Gupta

```
+-----+
|percentile(dif, array(0.05, 0.10, 0.25, 0.50, 0.75, 0.80, 0.95, 0.99), 1)|
+-----+
|[-10.0, 0.0, 0.0, 0.0, 1.0, 2.0, 7.0, 20.0]|
+-----+
```

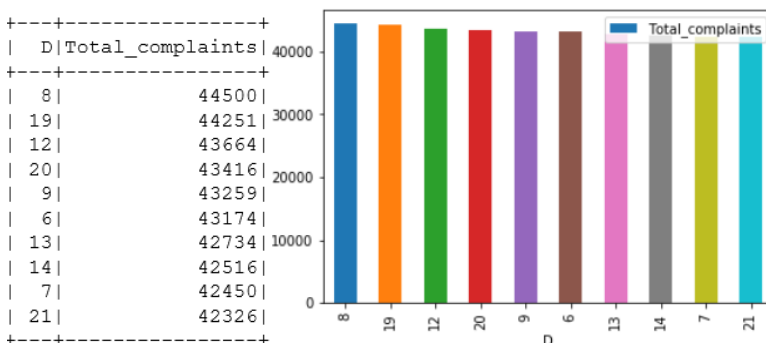
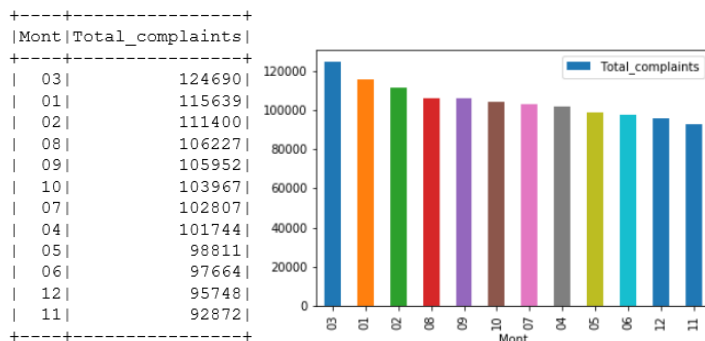
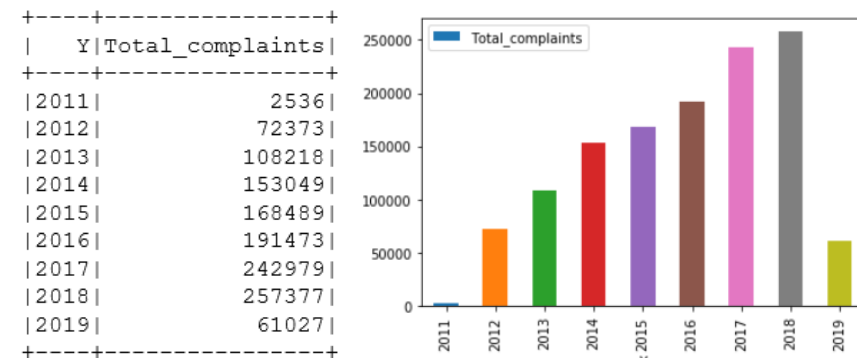
- There are also some data entry issues where the days took is in negative. Date received is after the date sent which is a data entry issue.

```
: df2.where(col("Datereceived") == col("Datesentcomp")).count()
: 776286
```

There are 776286 complaints – (62% of total complaints) which were sent to company on the same day.

### Create new variable month, year and Day

- Extract month, year and day from the date complaint received.
- Distribution of the complaints across different months and years



- First three months of the year have the most complaints.
- Complaints are increasing over the years

Manisha Gupta

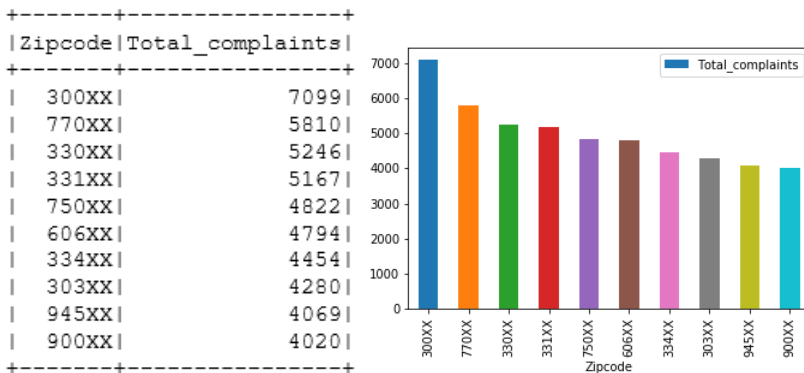
- 2018 being the worst year
- Will be interesting to see how many ends up in 2019

## Average time taken to send the complaints per state

State	avgtime
PR	4.882182052937379
null	4.00838779956427
SD	3.821218074656189
HI	3.680452093501156
NM	3.587267136914451
MS	3.326530612244898
FM	3.3217391304347825
KY	3.1855388813096863
MT	3.134924753502854
OK	3.0352527913687117

The average time taken is high for those states which are expansive, sparsely populated such as Puerto Rico, South Dakota.

## ZipCode



The maximum number of complaints are coming from Zip code starts “300”.

Manisha Gupta

## Bivariate Analysis

### Products against Timely response

Product_Timelyresponse	No	Yes
Mortgage	5516	270312
Debt collection	15067	225510
Credit reporting credit repair services or other personal consumer reports	3023	211280
Credit reporting	297	140135
Credit card	982	88208
Bank account or service	2226	83980
Student loan	765	50144
Credit card or prepaid card	267	44974
Checking or savings account	314	38266
Consumer Loan	1241	30364
Vehicle loan or lease	387	10484
Money transfer virtual currency or money service	167	9532
Payday loan title loan or personal loan	556	7730
Money transfers	146	5208
Payday loan	600	4944
Prepaid card	50	3769
Other financial service	113	946
Virtual currency	2	16

The products which are having most complaints got timely response in general.

### Products with submission medium

Product_Submittedvia	Email	Fax	Phone	Postal mail	Referral	Web
Checking or savin...	2	361	3533	1199	13440	20045
Bank account or s...	67	1183	11758	3725	28656	40817
Money transfers	0	54	1328	134	199	3639
Student loan	11	222	1775	990	2499	45412
Payday loan	0	27	311	73	142	4991
Payday loan title...	0	166	570	261	650	6639
Credit reporting	22	2085	3089	18125	9604	107507
Consumer Loan	22	360	2792	1309	3481	23641
Other financial s...	1	22	163	102	180	591
Prepaid card	0	18	509	102	492	2698
Vehicle loan or l...	1	128	759	393	1040	8550
Credit card or pr...	3	428	3032	2150	5502	34126
Debt collection	13	2409	13844	9599	12857	201855
Mortgage	179	5571	16989	15744	73334	164011
Money transfer vi...	0	59	741	125	468	8306
Credit card	43	901	6675	5341	14722	61508
Credit reporting ...	18	4682	7151	7383	4202	190867
Virtual currency	0	0	0	0	0	18

The above table shows how many complaints are coming from which medium for each product and it is evident that people are using web more and more now days which is obvious due to advancement of digital technology.

### Top 3 customer concerns related to Products per state

State	Product	count	rnk
WY	Debt collection	258	1
WY	Mortgage	245	2
WY	Credit reporting	163	3
WV	Debt collection	576	1
WV	Mortgage	513	2
WV	Credit reporting	430	3
WI	Mortgage	2822	1
WI	Debt collection	2551	2
WI	Credit reporting ...	1918	3
WA	Mortgage	6483	1
WA	Debt collection	4594	2
WA	Credit reporting ...	2747	3
VT	Mortgage	428	1
VT	Debt collection	221	2
VT	Credit card	204	3
VI	Credit reporting	52	1
VI	Mortgage	50	2
VI	Bank account or s...	34	3
VA	Mortgage	8658	1
VA	Debt collection	6525	2

Credit reporting, Debt collection and Mortgage are top 3 products in which customers are having concern in every state.

Manisha Gupta

## Top 3 customer concerns per state

Incorrect information, Loan modification and debt related concern are common in each states

State	Issue	count	rnk
UNITED STATES MIN...	Trouble during pa...	5	1
UNITED STATES MIN...	Incorrect informa...	4	2
UNITED STATES MIN...	Communication tac...	3	3
AZ	Loan modification...	2535	1
AZ	Incorrect informa...	2151	2
AZ	Incorrect informa...	2024	3
SC	Incorrect informa...	2951	1
SC	Incorrect informa...	1627	2
SC	Loan modification...	1161	3
LA	Incorrect informa...	2035	1
LA	Incorrect informa...	1306	2
LA	Problem with a cr...	934	3
MN	Loan modification...	1388	1
MN	Incorrect informa...	971	2
MN	Loan servicing pa...	964	3
AA	Loan servicing pa...	3	1
AA	Loan modification...	3	2
AA	Trouble during pa...	3	3
NJ	Loan modification...	5232	1
NJ	Incorrect informa...	3919	2

## Products Vs Year

Product_Y	2011	2012	2013	2014	2015	2016	2017	2018	2019
Checking or savin...	0	0	0	0	0	0	12764	21212	4604
Bank account or s...	0	12212	13388	14662	17140	21848	6956	0	0
Money transfers	0	0	559	1169	1619	1567	440	0	0
Student loan	0	2840	3005	4283	4501	8087	17174	8781	2238
Payday loan	0	0	194	1706	1585	1566	493	0	0
Payday loan title...	0	0	0	0	0	0	2960	4373	953
Credit reporting	0	1873	14380	29239	34272	44081	16587	0	0
Consumer Loan	0	1986	3117	5457	7886	9602	3557	0	0
Other financial s...	0	0	0	116	312	465	166	0	0
Prepaid card	0	0	0	336	1784	1250	449	0	0
Vehicle loan or l...	0	0	0	0	0	0	3694	5893	1284
Credit card or pr...	0	0	0	0	0	0	15404	24251	5586
Debt collection	0	0	11069	39145	39738	40469	47958	51210	10988
Mortgage	1276	38109	49401	42961	42345	41466	30580	24577	5113
Money transfer vi...	0	0	0	0	0	0	3266	5435	998
Credit card	1260	15353	13105	13974	17300	21065	7133	0	0
Credit reporting ...	0	0	0	0	0	0	73395	111645	29263
Virtual currency	0	0	0	1	7	7	3	0	0

As per the above table, we can see from 2011 to 2016, number of complaints for bank account was increasing but it starts declining and eventually disappear from 2017. For complaints related to student load was worst in year 2017.

## Submitted vs Year

Submittedvia_Y	2011	2012	2013	2014	2015	2016	2017	2018	2019
Fax	10	959	2068	2318	2143	2592	2770	5460	356
Postal mail	47	4572	7992	11256	9487	11872	11291	8908	1330
Email	10	154	151	17	8	8	8	21	5
Referral	795	27930	28178	23625	21542	24005	19715	22260	3418
Phone	218	6324	9489	10268	10270	12317	11242	11730	3161
Web	1456	32434	60340	105565	125039	140679	197953	208998	52757

The above table shows that consumers are using web to file their complaints and this trend has increasing order by each year.

## Timely Response for each Year

Timelyresponse_Y	2011	2012	2013	2014	2015	2016	2017	2018	2019
No	251	2508	1270	3639	4837	6676	7025	4657	856
Yes	2285	69865	106948	149410	163652	184797	235954	252720	60171

Manisha Gupta

Companies are now more cognizant towards complaints and providing timely response to consumer and this trend is increasing by each passing year.

## Consumer Dispute - Year

Consumerdisputed_Y	2011	2012	2013	2014	2015	2016	2017	2018	2019
No	1963	55946	85624	123338	134206	156685	62365	0	0
Yes	573	16427	22594	29711	34283	34788	10002	0	0
null	0	0	0	0	0	0	170612	257377	61027

Consumer were not happy with the company's response and it was worst in 2015 and 2016.

**Issue with Timely response:** Generally, issues with high complaints are getting timely response.

Issue_Timelyresponse	No	Yes
Incorrect informa...	1735	125159
Loan modification...	2155	110156
Incorrect informa...	218	102468
Loan servicing pa...	1511	75822
Cont d attempts c...	4131	56556
Problem with a cr...	660	47870
Attempts to colle...	1482	39428
Account opening c...	1033	36928
Communication tac...	2452	32485
Improper use of y...	485	31781
Disclosure verifi...	2035	28765
Managing an account	192	23846
Deposits and with...	557	22294
Written notificat...	868	22006
Trouble during pa...	299	21610
Managing the loan...	701	18539
False statements ...	1358	18514
Struggling to pay...	195	17809
Dealing with my l...	274	17356
Credit reporting ...	20	16863

## Consumer Complaints with IRS Datasets

### Sneak Peak in to IRS Dataset

```
print("Total column in IRS: ", len(df_irs.columns), "Total records in IRS : ", df_irs.count())
```

Total column in IRS: 147 Total records in IRS : 29974

Combining IRS data with Consumer Complaint dataset to study the relationship of income with consumer complaints. Extracted 3 digits of zip code to combine IRS data with Consumer compliant datasets.

### Study the relationship with IRS

Below table shows records of total return filed and average income amount for each zip code. "i\_zip" column has been made by extracting first 3 digit of zip code columns.

*C\_zip* – Consumer datasets; *i\_zip*- IRS dataset



## Manisha Gupta

c_zip	Total_complaints	i_zip	totalreturnfiled	toal_dep	Avg_income	ratio
300	21508	300	1062740	809380	1286266.92	2.02
606	15600	606	1286600	821000	1436147.29	1.21
770	15354	770	1337490	1030700	903301.53	1.15
331	14144	331	934790	536810	978877.73	1.51
330	14008	330	860550	530610	858324.57	1.63
900	13594	900	1072740	715020	1313384.84	1.27
750	12907	750	1135360	861360	1389927.18	1.14
112	12611	112	1198290	759430	1845402.26	1.05
334	11722	334	726880	381750	1210929.49	1.61
945	11552	945	1155000	786900	1469183.56	1.0
303	10792	303	481230	280160	1088971.11	2.24
207	9752	207	537380	336110	626448.54	1.81
100	9390	100	836790	301060	3390827.81	1.12
891	9206	891	698660	506480	991633.84	1.32
302	9139	302	468080	369250	477004.5	1.95
070	8664	null	null	null	null	null
191	8575	191	694620	425980	774071.92	1.23
282	8375	282	431920	278590	1155779.66	1.94
913	8371	913	639550	445370	1544983.39	1.31
926	8350	926	643830	374970	1887072.74	1.3

Using spark Stats library, we calculated the correlation coefficient and we found that there is positive co-relation between number of complaints with number of return files.

- Total number of complaints are highly positive correlated with number of returns filed which means if number of return increases, then number of complaints will also increase.
- Average income amount has positive correlation with number of complaints filed
- One more interesting fact we found is that number of dependents has high positive correlation with number of complaints that means if number of dependents is increasing of any location, the number of complaints are also increasing of that specific location.

```
In [208]: df_join_irs.stat.corr('Total_complaints','totalreturnfiled')
Out[208]: 0.8334642490188765

In [209]: df_join_irs.stat.corr('Total_complaints','Avg_income')
Out[209]: 0.5102003225461417

In [210]: df_join_irs.stat.corr('Total_complaints','toal_dep')
Out[210]: 0.8327416816076443
```

## Consumer Complaints with Population Dataset

### Sneak Peak in to Population Data

There 6 columns and 1622831 records in population data sets

```
print("Total column in Population: ", len(df_pop.columns), "Total records in Population : ", df_pop.count())
Total column in Population: 6 Total records in Population : 1622831
```

We have combined the population data set with consumer dataset based on first 3 digit of zip code.

## Manisha Gupta

```
In [232]: pivot_pop = df_pop.groupby("p_zip").agg((sum("population").alias("population"))
pivot_pop.show()

+-----+-----+
|p_zip|population|
+-----+-----+
| 691| 239697|
| 296| 2871921|
| 467| 997722|
| 675| 433323|
| 829| 211455|
| 451| 1014138|
| 853| 3949887|
| 800| 2674974|
| 125| 1365336|
| 944| 395403|
| 870| 877317|
| 919| 1633344|
| 926| 4154505|
| 666| 490245|
| 124| 530484|
| 447| 630861|
| 591| 384375|
| 574| 196578|
| 475| 523239|
| 307| 1289226|
+-----+-----+
only showing top 20 rows
```

## Total complaints and population per area

Below table shows the total complaints and population per area. As we can see, the more population- high number of complaints are there.

```
+-----+-----+-----+-----+
|c_zip|Total_complaints|p_zip|population|
+-----+-----+-----+-----+
| 300| 21508| 300| 6533130|
| 606| 15600| 606| 8294418|
| 770| 15354| 770| 8823435|
| 331| 14144| 331| 5657535|
| 330| 14008| 330| 5017686|
| 900| 13594| 900| 7213185|
| 750| 12907| 750| 6610437|
| 112| 12611| 112| 7543050|
| 334| 11722| 334| 4282482|
| 945| 11552| 945| 6822306|
| 303| 10792| 303| 2903100|
| 207| 9752| 207| 2972868|
| 100| 9390| 100| 4741269|
| 891| 9206| 891| 4255950|
| 302| 9139| 302| 3003009|
| 070| 8664| null| null|
| 191| 8575| 191| 4745973|
| 282| 8375| 282| 2438532|
| 913| 8371| 913| 4119525|
| 926| 8350| 926| 4154505|
+-----+-----+-----+-----+
```

## Study the relationship with population

We applied pyspark.stats lib to calculate correlation between population and number of complaints and found both are highly positive correlated.

```
In [237]: df_join_pop.stat.corr('Total_complaints','population')

Out[237]: 0.8612501335153144
```

**Correlation Male and Female with complaints-** both have positive co-relation with number of complaints however, female's correlation is higher than males.

Manisha Gupta

c_zip	Total_complaints	p_zip	null	female	male
300	21508	300	2177710	2242444	2112976
606	15600	606	2764806	2845196	2684416
770	15354	770	2941145	2944366	2937924
331	14144	331	1885845	1944412	1827278
330	14008	330	1672562	1725726	1619398
900	13594	900	2404395	2407226	2401564
750	12907	750	2203479	2241390	2165568
112	12611	112	2514350	2656656	2372044
334	11722	334	1427494	1471532	1383456
945	11552	945	2274102	2316290	2231914
303	10792	303	967700	979488	955912
207	9752	207	990956	1023618	958294
100	9390	100	1580423	1671314	1489532
891	9206	891	1418650	1401214	1436086
302	9139	302	1001003	1037936	964070
070	8664	null	null	null	null
191	8575	191	1581991	1670956	1493026
282	8375	282	812844	838782	786906
913	8371	913	1373175	1384258	1362092
926	8350	926	1384835	1414296	1355374

```
df_join_pop.stat.corr('Total_complaints','female')
0.8631007445786633

df_join_pop.stat.corr('Total_complaints','male')
0.8588500047114318
```

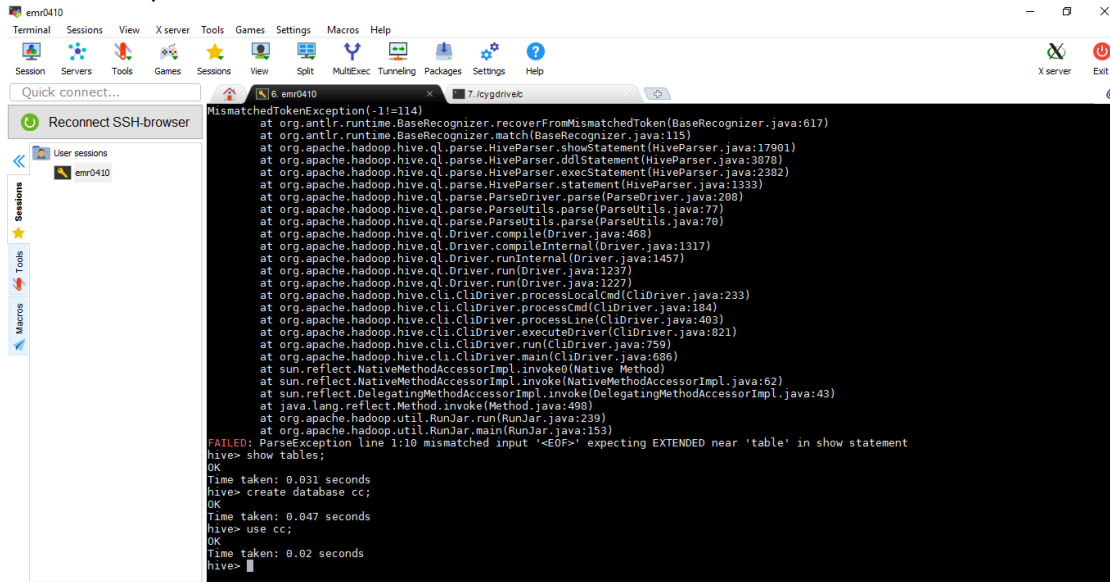
**Relationship with Age:** In the population dataset, minimum and maximum age per zip code was given. Hence, we calculated the average age by (minimum+ maximum)/2 and then try to get average age per area. This calculation does not seem to be correct hence didn't proceed further. Below are the screen shots-

p_zip	population	avg_age
691	239697	40.97727272727273
296	2871921	40.97727272727273
467	997722	40.97727272727273
675	433323	40.97727272727273
829	211455	40.97727272727273
451	1014138	40.97727272727273
853	3949887	40.97727272727273
800	2674974	40.97727272727273
125	1365336	40.97727272727273
944	395403	40.97727272727273
870	877317	40.97727272727273
919	1633344	40.97727272727273
926	4154505	40.97727272727273
666	490245	40.97727272727273
124	530484	40.97727272727273
447	630861	40.97727272727273
591	384375	40.97727272727273
574	196578	40.97727272727273
475	523239	40.97727272727273
307	1289226	40.97727272727273

## Using Hive

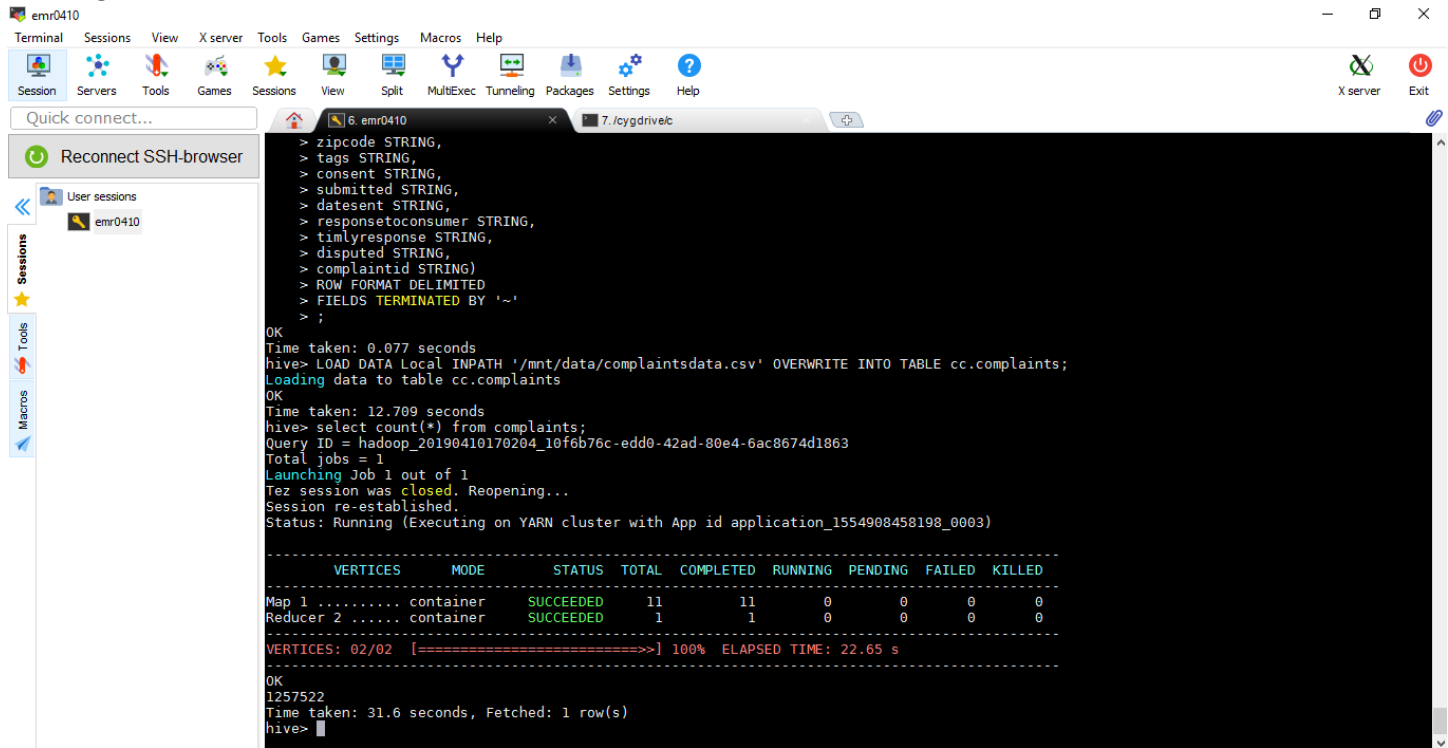
### Creating Database called cc;

## Manisha Gupta



```
emr0410
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help
Quick connect...
Reconnect SSH-browser
User sessions
emr0410
MismatchedTokenException(-1!=114)
at org.apache.hadoop.hive.ql.parse.HiveParser.showStatement(HiveParser.java:17901)
at org.apache.hadoop.hive.ql.parse.HiveParser.ddStatement(HiveParser.java:3378)
at org.apache.hadoop.hive.ql.parse.HiveParser.execStatement(HiveParser.java:2382)
at org.apache.hadoop.hive.ql.parse.HiveParser.statement(HiveParser.java:1333)
at org.apache.hadoop.hive.ql.parse.ParseDriver.parse(ParseDriver.java:208)
at org.apache.hadoop.hive.ql.parse.ParseUtils.parse(ParseUtils.java:77)
at org.apache.hadoop.hive.ql.Driver.compile(Driver.java:468)
at org.apache.hadoop.hive.ql.Driver.compileInternal(Driver.java:1317)
at org.apache.hadoop.hive.ql.Driver.runInternal(Driver.java:1457)
at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1227)
at org.apache.hadoop.hive.cli.CliDriver.processLocalCmd(CliDriver.java:233)
at org.apache.hadoop.hive.cli.CliDriver.processCmd(CliDriver.java:184)
at org.apache.hadoop.hive.cli.CliDriver.executeDriver(CliDriver.java:821)
at org.apache.hadoop.hive.cli.CliDriver.run(CliDriver.java:759)
at org.apache.hadoop.hive.cli.CliDriver.main(CliDriver.java:686)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:498)
at org.apache.hadoop.util.RunJar.run(RunJar.java:239)
at org.apache.hadoop.util.RunJar.main(RunJar.java:153)
FAILED: ParseException line 1:10 mismatched input '<EOF>' expecting EXTENDED near 'table' in show statement
hive> show tables;
OK
Time taken: 0.031 seconds
hive> create database cc;
OK
Time taken: 0.047 seconds
hive> use cc;
OK
Time taken: 0.02 seconds
hive>
```

## Creating Hive table & load the data and count the number rows



```
emr0410
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help
Quick connect...
Reconnect SSH-browser
User sessions
emr0410
> zipcode STRING,
> tags STRING,
> consent STRING,
> submitted STRING,
> datesent STRING,
> responsestoconsumer STRING,
> timlyresponse STRING,
> disputed STRING,
> complaintid STRING)
> ROW FORMAT DELIMITED
> FIELDS TERMINATED BY '~'
> ;
OK
Time taken: 0.077 seconds
hive> LOAD DATA LOCAL INPATH '/mnt/data/complaintsdata.csv' OVERWRITE INTO TABLE cc.complaints;
Loading data to table cc.complaints
OK
Time taken: 12.709 seconds
hive> select count(*) from complaints;
Query ID = hadoop_20190410170204_10f6b76c-edd0-42ad-80e4-6ac8674d1863
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1554908458198_0003)

-----
VERTICES      MODE      STATUS    TOTAL    COMPLETED    RUNNING    PENDING    FAILED    KILLED
-----
Map 1 ..... container    SUCCEEDED    11         11         0         0         0         0
Reducer 2 ..... container    SUCCEEDED     1         1         0         0         0         0
-----
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 22.65 s
-----
OK
1257522
Time taken: 31.6 seconds, Fetched: 1 row(s)
hive>
```

UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: <https://mobaxterm.mobatek.net>

## Displaying 10 rows

## Manisha Gupta

The screenshot shows a MobaXterm terminal window with the title 'emr0410'. The terminal displays the output of a Hive query: `select * from complaints limit 10;`. The output is a table with columns: Date received, Product, Sub-product, Issue, Sub-issue, Consumer complaint narrative, Company public response, Company State, and a final column with a 'Z'. The data includes entries from CITIBANK N A, CAPITAL ONE FINANCIAL CORPORATION, Navient Solutions LLC, and OCWEN LOAN SERVICING LLC. The terminal also shows the execution time: 34.267 seconds, Fetched: 10 row(s).

## Counting total complaints filed per company

The screenshot shows a MobaXterm terminal window with the title 'emr0410'. The terminal displays the output of a Hive query: `select company, count(issue) as isu from complaints group by company order by isu desc limit 10;`. The query ID is `hadoop_20190410172024_87b97d41-8f76-4bf8-ba4c-218df9a7d61e`. The terminal shows the execution progress of the query, including the number of vertices, mode, status, total, completed, running, pending, failed, and killed. The status is 'SUCCEEDED' and the elapsed time is 25.82 s. The terminal also shows the execution time: 34.267 seconds, Fetched: 10 row(s).

UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: <https://mobaxterm.mobatek.net>

## Reference

<https://medium.com/@JackRemondi/what-cfpb-consumer-data-prescribes-for-student-loans-ab46b8e62179>  
[https://www2.census.gov/programs-surveys/rhfs/cbp/technical%20documentation/2015\\_record\\_layouts/zip\\_totals\\_layout\\_2015.txt?#](https://www2.census.gov/programs-surveys/rhfs/cbp/technical%20documentation/2015_record_layouts/zip_totals_layout_2015.txt?#)  
<https://www.kaggle.com/umeshnarayanappa/exploring-the-data>  
<http://hadooptutorial.info/hive-aggregate-functions/>  
<https://catalog.data.gov/dataset/2010-census-populations-by-zip-code>

**Manisha Gupta**

<https://simplemaps.com/data/us-zips>

<https://github.com/open-austin/project-ideas/issues/107>

<https://www.kaggle.com/umeshnarayanappa/exploring-the-data>