**Assignment 2**

**Advance in Data Sciences and Architecture**

**INFO 7390 - SPRING 2017**

PROFESSOR:

**SRIKANTH KRISHNAMURTHY**

TEAM MEMBERS:

**HARSHAD PARDESHI (**[**www.linkedin.com/in/harshad-pardeshi/**](file:///C:\Data\Important\Study\ADS\Presentations\Kubernete\kubernets_datascience-master\kubernets_datascience-master\www.linkedin.com\in\aashritandon\)**)**

**PRITHVI KAMATH (**[**www.linkedin.com/in/**](http://www.linkedin.com/in/)**prithvipkamath/)**

**TEJESH RATHI (www.linkedin.com/in/tejesh-rathi/)**

Table of Contents

[Problem Statement 3](#_Toc479378863)

[Part 1: Data wrangling and exploratory data analysis 4](#_Toc479378864)

[Data downloading and storing 4](#_Toc479378865)

[Data cleaning 6](#_Toc479378866)

[Exploratory data analysis in Power BI 16](#_Toc479378867)

[Storing clean data on Amazon S3 bucket 20](#_Toc479378868)

# **Problem Statement**

You are working at a bank and you are considering investing in Lending club. Since there are no standard models, you are expected to build prediction models that will help you predict the interest rates based on various parameters users would input.

Your first challenge is to programmatically download the data from https://www.lendingclub.com/info/download-data.action

Your goal is to download the data programmatically from the website and create one dataset for the entire database and perform the following tasks:

* Data download: How will you download all loan data and create one dataset
* Missing data analysis: How will you handle missing data?
* Feature engineering: What variables do you need to predict interest rates? Ensure users would be able to give you that information to help you predict rates
* Pipeline: Using Luigi/Pinball/Airflow automate the above 3 steps.
* You need to create one more pipeline to do this for the “Declined Loan data”. Repeat above steps

You should dockerize the whole project (excluding the Power BI dashboard) and write clear instructions on how to run the docker image. Research how you would schedule this on Amazon/Azure/BlueMix so that this image would run and execute the Luigi/Pinball/Airflow pipeline. After running this pipeline, the clean pre-processed data should be stored on S3/Object Storage/Blob storage(AWS/BlueMix/Azure)

* Write a Jupyter notebook using R/Python to graphically represent different summaries of data. Summarize your findings in this notebook.
* Summarize your key insights about different user profiles, states, loan amounts etc.
* Create a Data scientist view of Power BI dashboards to illustrate your key insights

# **Part 1: Data wrangling and exploratory data analysis**

In this section, we will perform the data the following operations:

* Data downloading and storing
* Data cleaning
* Exploratory data analysis in Python
* Exploratory data analysis in Power BI
* Storing clean data on Amazon S3 bucket

## **Data downloading and storing**

Our first step is to get the Lending club dataset from the Lending Club web site. To do this we have two options.

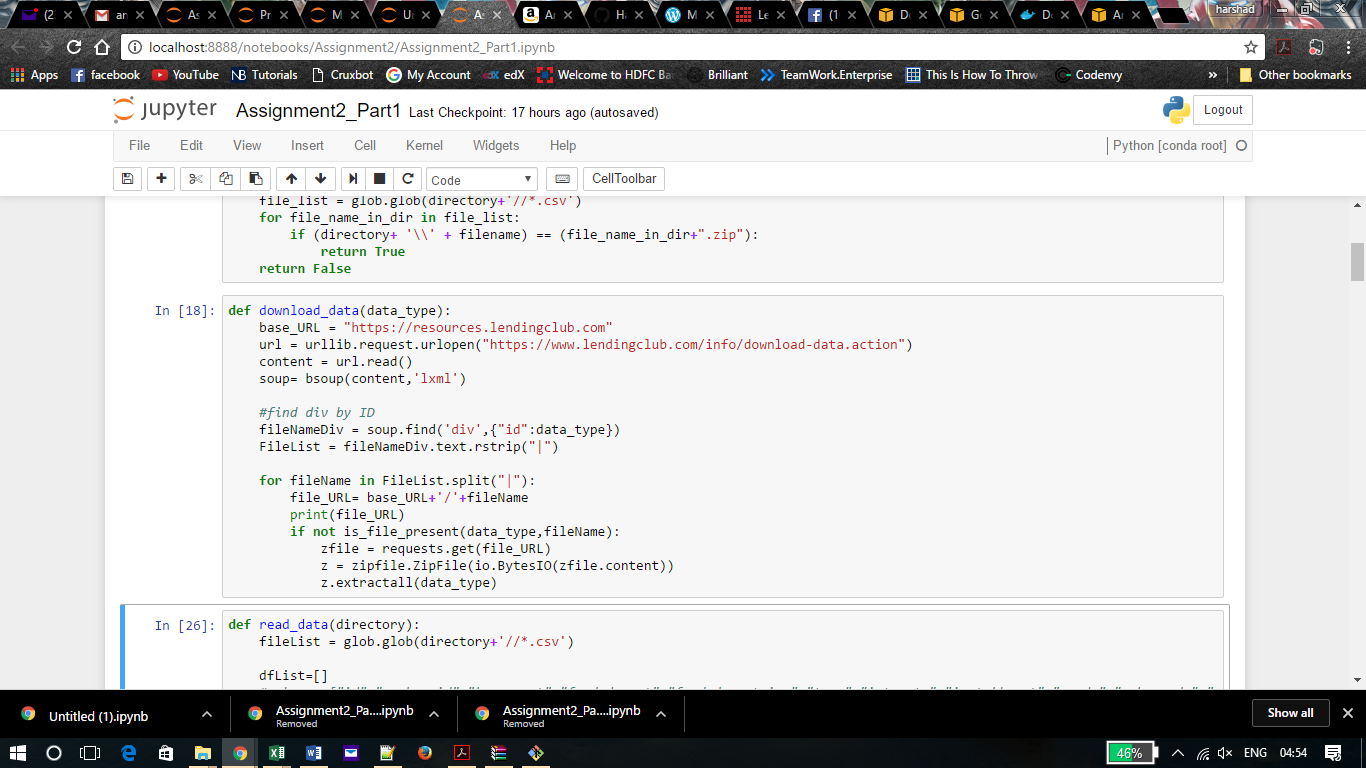
1. To download the data that is available publicly from [lending club site](https://www.lendingclub.com/info/download-data.action).
2. Login to the lending club website by providing all personal details in apply for loan section and then download the complete data set.

We are considering the first option to download both approved loans data as well as rejected loan data. We have gone through the second step to download which requires all your personal details (including your Social Security Number). The terms and conditions of the site nowhere mentions that they are obliged to not share any personal information to any other third party and does not clearly specify how they will use our data.

The data is divided into multiple files based on the year in which loan was issued/rejected. To download multiple files in python we first need to get the URLs of all the files. For this we use *urllib3* library. File name of all the files is fetched from the a hidden div variable. Hidden div for Accepted loan data is ‘*loanStatsFileNamesJS’* and for rejected loan data is *‘rejectedLoanStatsFileNamesJS’*. This variable contains all the file names separated by ‘|’. From this names we form the URL and download the file using *requests.get(url)* method. Then we extract the downloaded zip files using the *extractall()* function *zipfile* library.

A single function is written which can download both loan (accepted/rejected) data and store separately. We call this function once for loan accepted data with *loanStatsFileNamesJS* parameter and once for rejected loan dataset with parameter *rejectedLoanStatsFileNamesJS*.

This will download and extract all the data into separate folders.



Further we read these files in individual Dataframes and add all the dataframes in a list. After reading all the files in list we simply concatenate all the dataframes.

A single function is written which can read the data from both loan (accepted/rejected) data and store separately. We call this function once for loan accepted data with *loanStatsFileNamesJS* parameter and once for rejected loan dataset with parameter *rejectedLoanStatsFileNamesJS*.

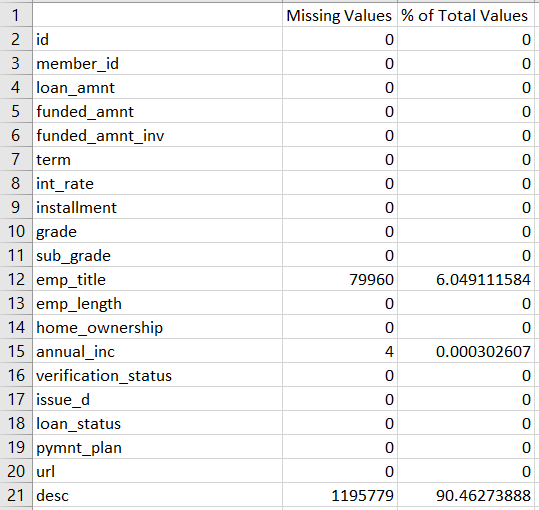
The combined dataframe is then written into a csv file.

This entire process is considered as a single Luigi task. This process will give output as the two folders that will contain all the downloaded csv files.

## **Data cleaning**

This section comprises of handling some special and important missing data. We get the concatenated csv file from the previous part, on which we perform various alteration and changes on the blank data without changing the meaning of the data available and has some logical significance.

We first analyzed all the columns having empty values as shown below:



This was created by a special function we created, which gives us a csv having the count of null values and percentage of empty rows in file.

We started off with feature selection and based on that we found columns which does not have any impact on the prediction of interest rates. If some of these columns were null, we directly removed the column as they were of no use in determining the interest rates.

**Columns deleted:**

* emp\_title
* desc
* last\_pymnt\_d
* next\_pymnt\_d
* last\_credit\_pull\_d
* url
* desc
* title
* emp\_title
* chargeoff\_within\_12\_mths
* mths\_since\_last\_record
* mths\_since\_last\_major\_derog
* open\_il\_12m

As part of feature selection, we are selecting all features except the one mentioned above.

Another document named “FeatureSelection.docx” has been added for better explanation.



**Columns Updated and cleaned:**

* **emp\_length** – As per the Data Dictionary, emp\_length should have only values ranging from 0 to 10, whereas there were few records with value of “10+ years” and “<1 year”, which we replaced with 10 and 0 respectively.
* **Title** – Title is the reason for a person to borrow money. There are various categories based on different reasons provided by the users; since blank means they haven’t specified the reason for borrowing the money, we assumed that it was something they didn’t wanted anybody to know and had some personal reasons, so we assigned them the “personal” category to such rows.
* **mths\_since\_last\_delinq** – This field shows the number of months passed by since a person was marked delinquent. Here, we made use of another column which is correlated to this column i.e. delinq\_2yrs. If the value in delinq\_2yrs = 0, that means there has been no delinquencies recorded for a person in the last 2 years. So, making use of this analysis, we replaced all the empty data of mths\_since\_last\_delinq which has delinq\_2yrs as 0, as 24 (resembling 2 years).
* **mths\_since\_last\_record** – This field indicates the number of months passed by since a person had any public record. This value is provided by the Lender’s based on the data they have. All public record is noted and present in their database. Since, the field is left blank means, that a person doesn’t have any public record yet. So, we marked it as -1 to show that a person doesn’t have any public record.
* **annual\_inc** – Annual income is empty just for 4 records in the data. On analyzing we found that all these 4 records don’t have any employer, which means that they don’t earn so we replaced all blanks by 0.
* **delinq\_2yrs** – delinq\_2yrs column has maximum of 39 and minimum of 0. On reading the data, we found that all the empty columns are for people who are not valid and does not meet the credit policy, so we replaced them by the mean.
* **revol\_util –** These are the percent values, so we first stripped all the “%” and then replaced it by the mean percentage and processed it in a new column “derived\_revol\_util”.
* **delinq\_amnt** – 99.98% of the data is zero. Only negligible records have any other values, so we replaced it by zero in a new column “derived\_delinq\_amnt”.
* **pub\_rec\_bankruptcies** – All the empty rows are the ones which do not meet the credit policy and all are non-verified. With only negligible blanks, we replaced it by mean.
* **tax\_liens** – These are the same 29 empty records as pub\_rec\_bankruptcies and so we processed a new column as “derived\_tax\_liens”.
* **Interest\_rate** – Stripped the “%”, so that analysis can be done on top of that and put the values in a new column “derived\_interest\_rate”

**Data pre-processing**

We are excluding the below rows that summarize the dataset.

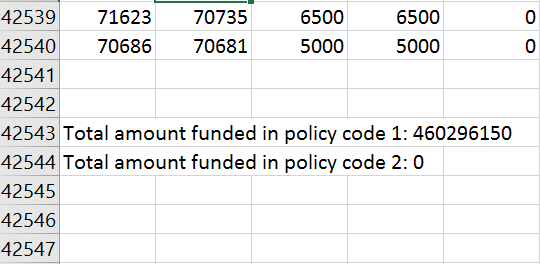
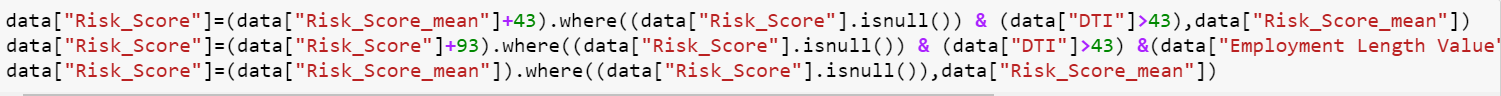


Figure : Summary rows that are excluded in the dataframe.

Converting ‘issue\_d’ to datetime format using pd.to\_datetime

We are creating new columns in the dataframe to store the below values:

* Interest rate value as Interest\_Rate
* Issue month as issue\_month
* Issue year as issue\_year
* For the rejected file, we observe that comparable number of observations have Null in their Risk Score.
* We are filling the null values by considering the Debt-To-Income-Ratio, Employment Length and the mean Risk Score for the individual’s state.
* We have learnt that Debt-To-Income-Ratio above 43% leads to higher risk score and hence leads to loan rejection.
* Hence for employers, who don’t have Risk Score we have replaced them by the addition of mean Risk Score for the state and increased it by 43 if Debt-To-Income-Ratio is above 43%.
* We have also added 50 points to the risk score if the employment length of an individual is less than the average employment rate for that state.
* Hence the lower limit for the value would be the mean of the state-wise risk score and higher limit would be mean of state-wise risk score + 93



We are deriving two variables:

* Lending\_Club\_Interest: This is the amount Lending Club makes on every transaction. It is calculated by subtracting the amount paid back to the investor from the total amount paid by the loan borrower.
* credit\_age: This variable shows the duration since when the user is using credit. It is calculated by subtracting the first date when credit was checked from the date when credit/loan was issued.

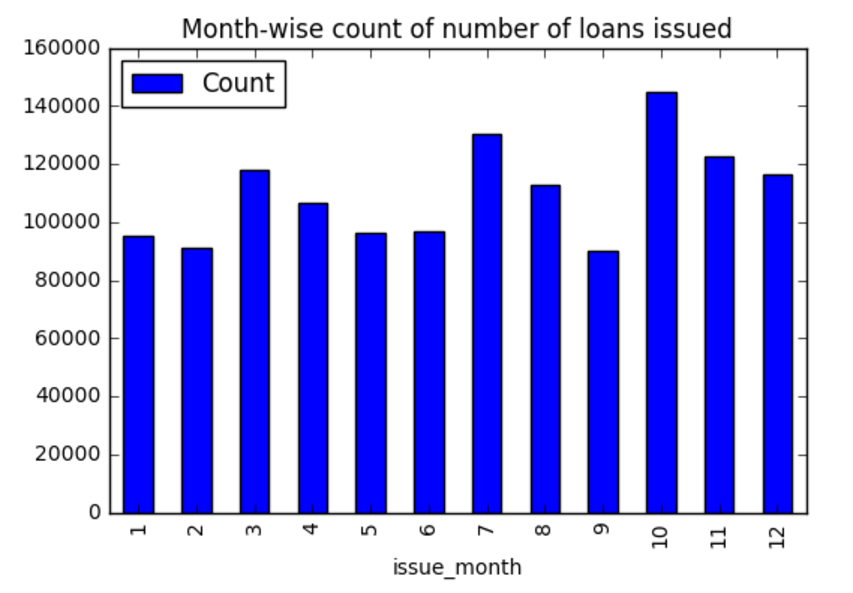
**Exploratory data analysis in Python**

**Analysis 1:** Find the month-wise count of number of loans issued

**Approach:**

Group the above dataframe for every month and count the number of loans issued in each month.

The result is saved in Analysis\_1.csv and a graph is drawn which represent this flow.



**Conclusion:** This graph shows us the month wise distribution on the number of loan issued.

**Analysis 2:** State-wise count of number of loans issued

**Approach:**

Group the dataframe for every state and count the number of loans issued in each state.

The result is saved in Analysis\_2.csv and a graph is drawn which represent this flow.

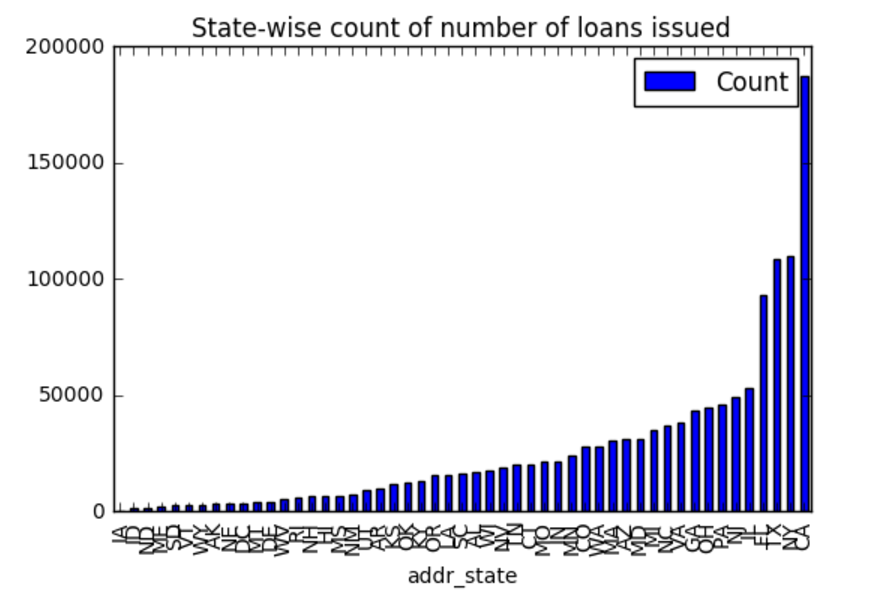
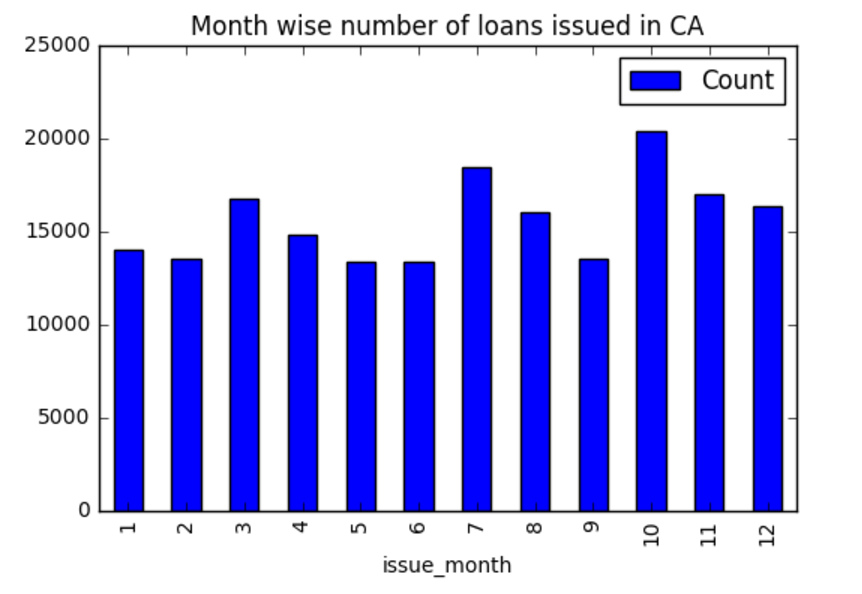
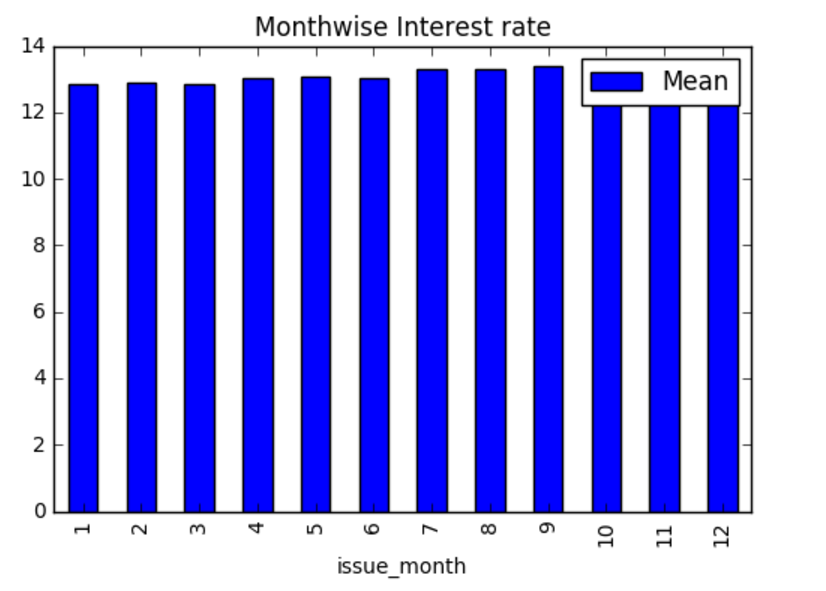
****

Figure : State wise loan issued distribution.

Here we see that CA has the highest count compared to other states and hence we dive into CA state issuers. On getting a month-wise graph, we observe that it is similar to the overall month-wise distribution.



Since there is no anomalies or sudden rise or fall in the graph we extract the interest rate in CA for each month.



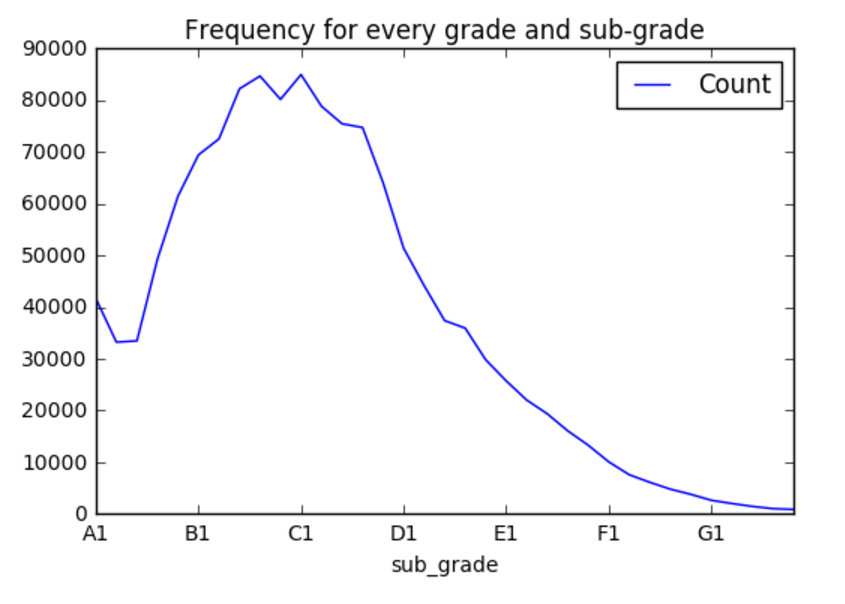
**Conclusion:** We observe that even though the interest rate is increasing over the months, there seems to be a rise in the number of people who are issued loans. This concludes that the Lending Club is targeting the right market in CA.

**Analysis 3:** Find the frequency for every grade and sub-grade

**Approach:**

Group the above dataframe for every grade and sub-grade and count the number of loans issued.

The result is saved in Analysis\_3.csv and a graph is drawn which represent this flow.

****

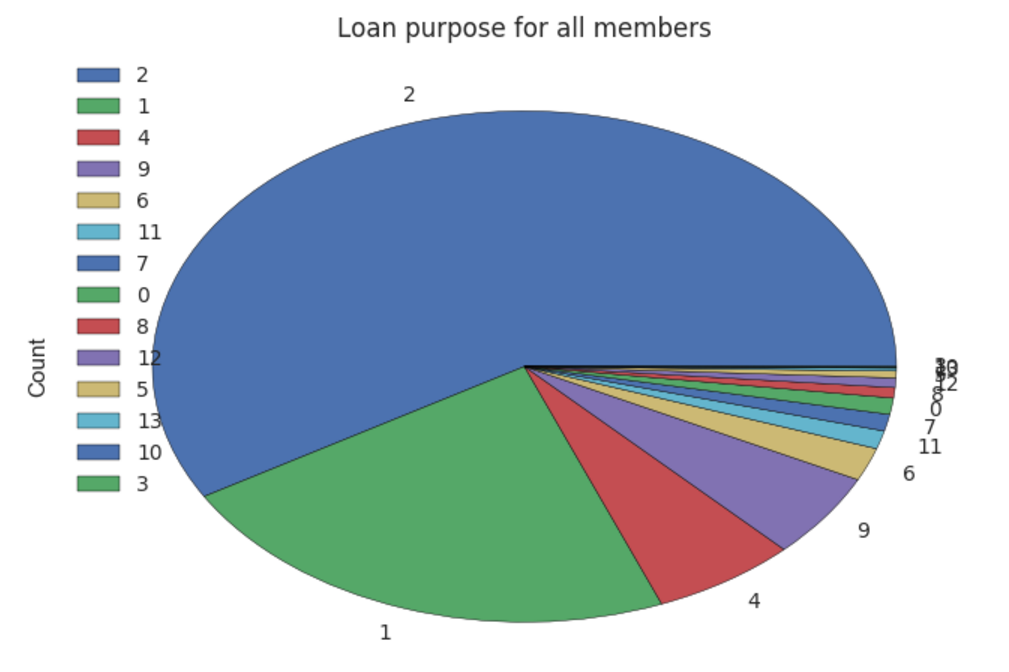
**Conclusion:** We observe that the maximum number of loan issuers lie in grades B3, B4, B5, C1 and C2.

**Analysis 4:** Find the loan purpose for all members

**Approach:**

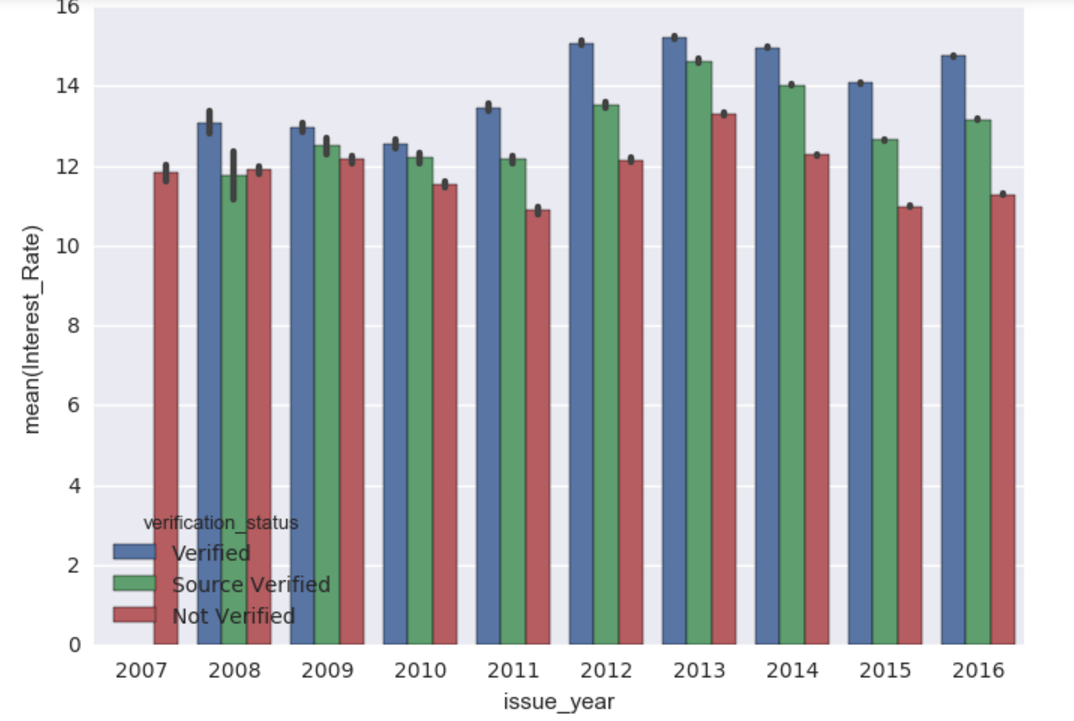
Group the above dataframe for every purpose and count the number of loans issued.

The result is saved in Analysis\_4.csv and a graph is drawn which represent this flow.



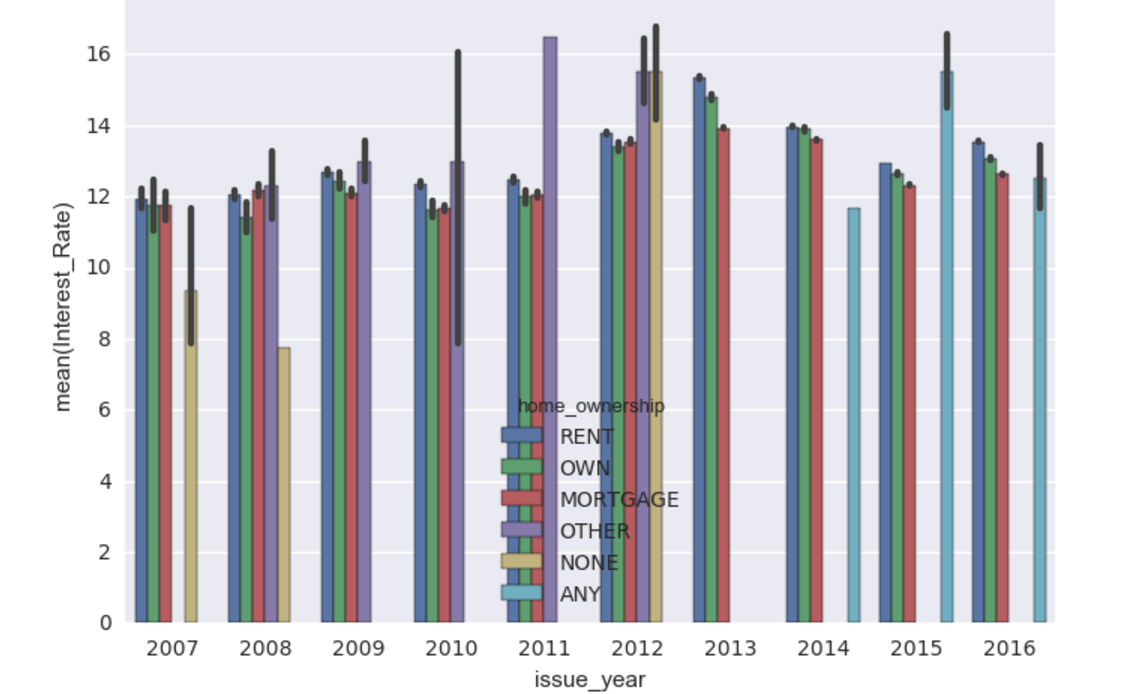
**Conclusion:** We observe that the most mentioned reason for issuing a loan is debt consolidation which is referred by the number 2.

**Analysis 5:** Find the mean Interest Rate for every Verification status over all the years.

****

**Conclusion:** We observe that the difference between various groups of Verification Status grows over the years from 2007 to 2016. Hence we can conclude that the verification team is working efficiently and processing more

**Analysis 6:** House ownership status for all members

****

**Conclusion:** This graph represents the home ownership distribution for each year from 2007 to 2016. As the mean interest rate for people renting, owning and mortgaging a house is almost the same across all the years, we can conclude that this attribute does not make a huge difference to the credit score and interest rate.

## **Exploratory data analysis in Power BI**

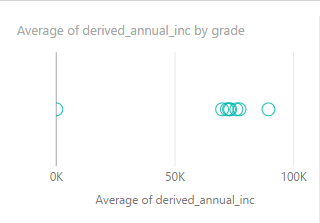
We extracted the data, cleaned and pre-processed it, analyzed the data, completed the feature selection and then the final task of making the data insights visually appealing using PowerBI.

Analysis is done for two reasons –

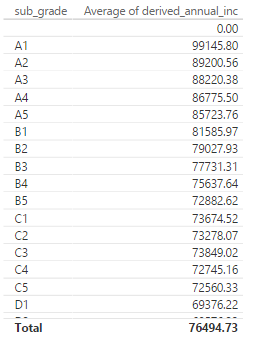
1. to get some favorable insights
2. for statistical purpose

We did a few on both:

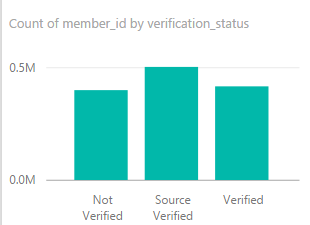
1. Average annual income based on the Grade a person belongs to.
   * This gives the data of the people having an average salary based on the Grades they belong. By looking at the data, we get to know that only Grade A people are class apart and have way better salary than any other Grades. Any Grade other than A, has more or less a similar average salary.



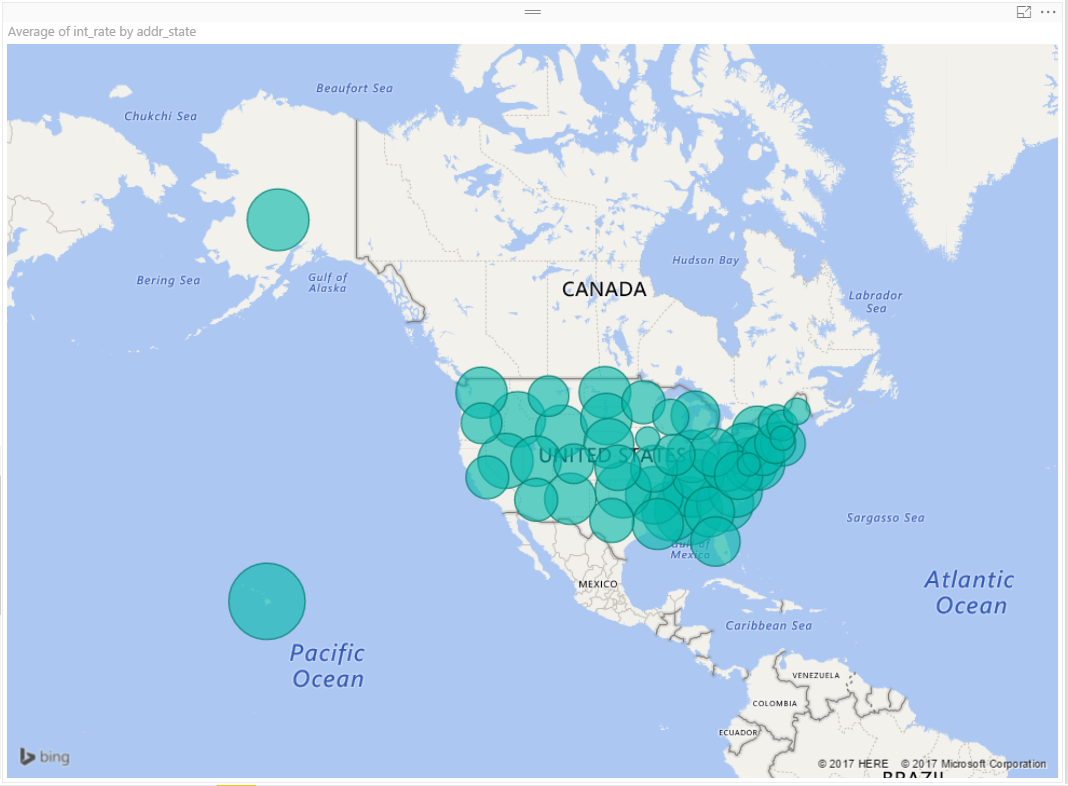
1. Getting even deeper and drilling down the hierarchy, the average salary based on the sub-grades of an individual.



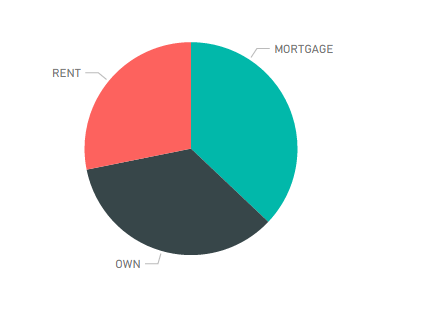
1. Statistical data on count of verified, not verified and Source verified individuals.



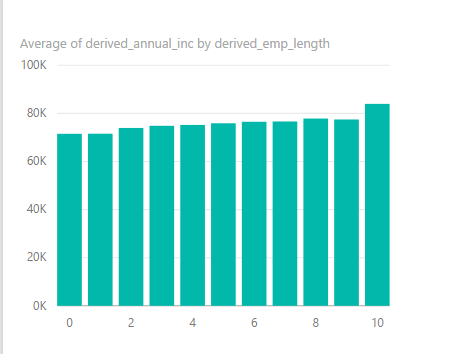
1. The average interest rate based on each state. On visualizing the data from the given link, we concluded that the closer the western part of the States has comparatively less interest rates.



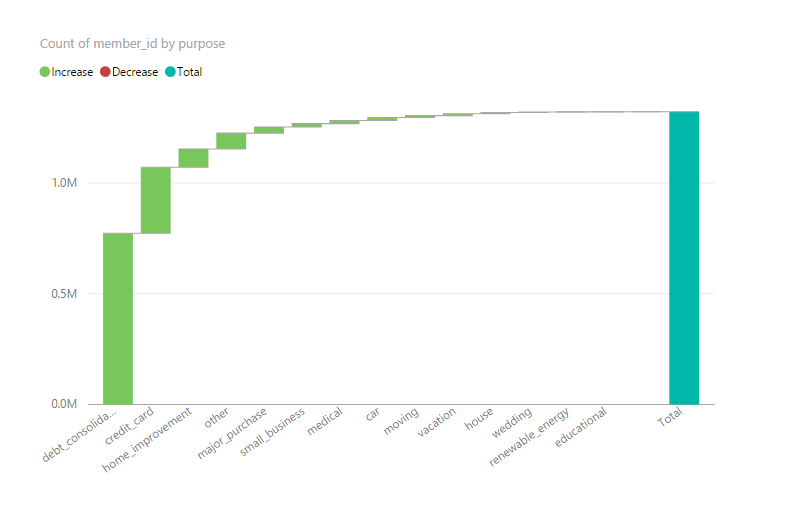
1. The average employee work length to acquire different types of home ownership.



1. Salary to number of years a person has worked. This shows that salary increases with amount of experience.



1. This analysis is probably the most interesting which helps us find the most popular reason for a person to land into Lender’s Club to borrow money.



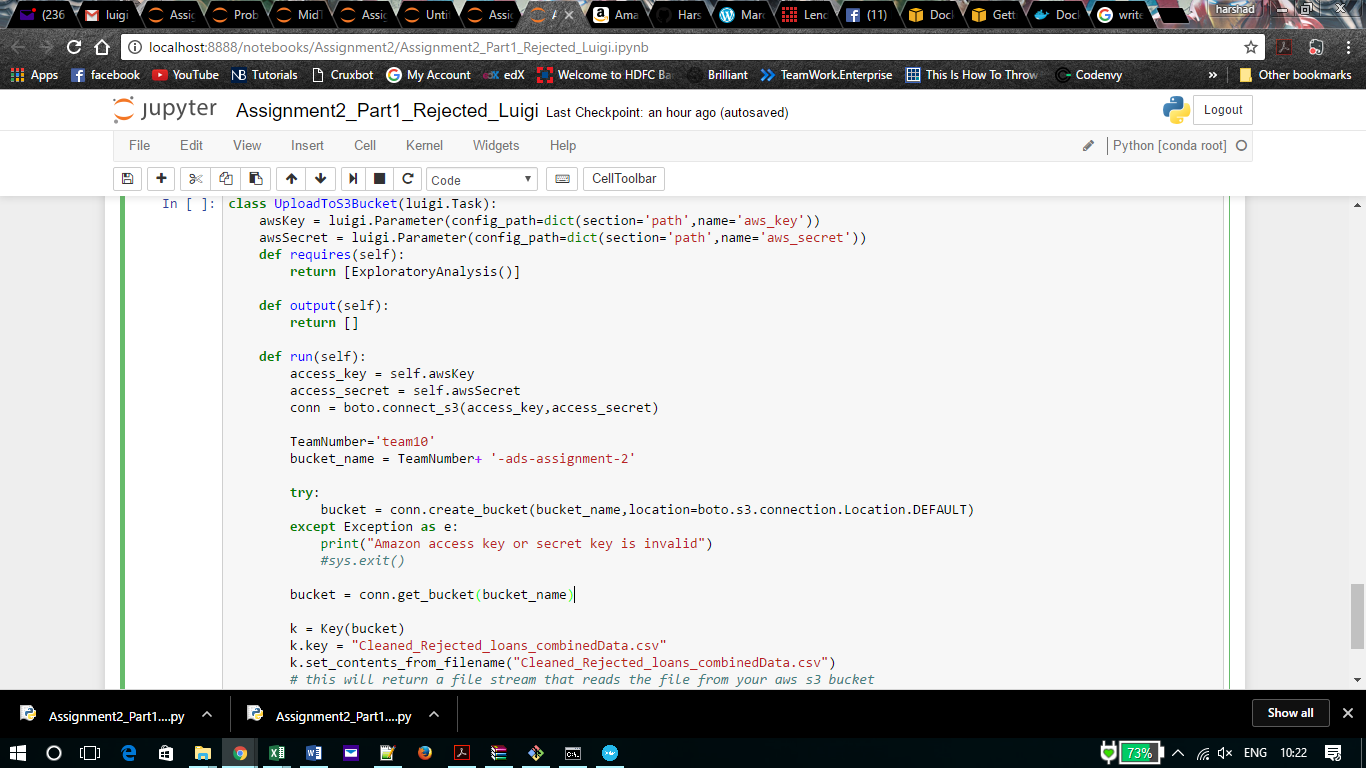
PowerBI dashboards are available in the below link:

<https://app.powerbi.com/groups/me/reports/6a3b7b14-1a9c-44b2-8cf5-9a7db45bb3bf/ReportSection>

## **Storing clean data on Amazon S3 bucket**

Once data cleaning and preprocessing is complete we will push the final clean file to the Amazon S3 bucket. We are using *boto* library to archive this task. We first get the Amazon S3 connection using the AWS\_KEY and AWS\_SECRET\_KEY. These two keys are passed when liuigi scrip is called.

After getting the connection se simply create a new bucket with team name and assignment2 (team-10-assignment2). After this we simply push the clea file to S3 bucket



**Scheduling and pipelining**

For pipelining the process, we have used Luigi that will run the series of tasks. We have created 4 tasks for each pipeline:

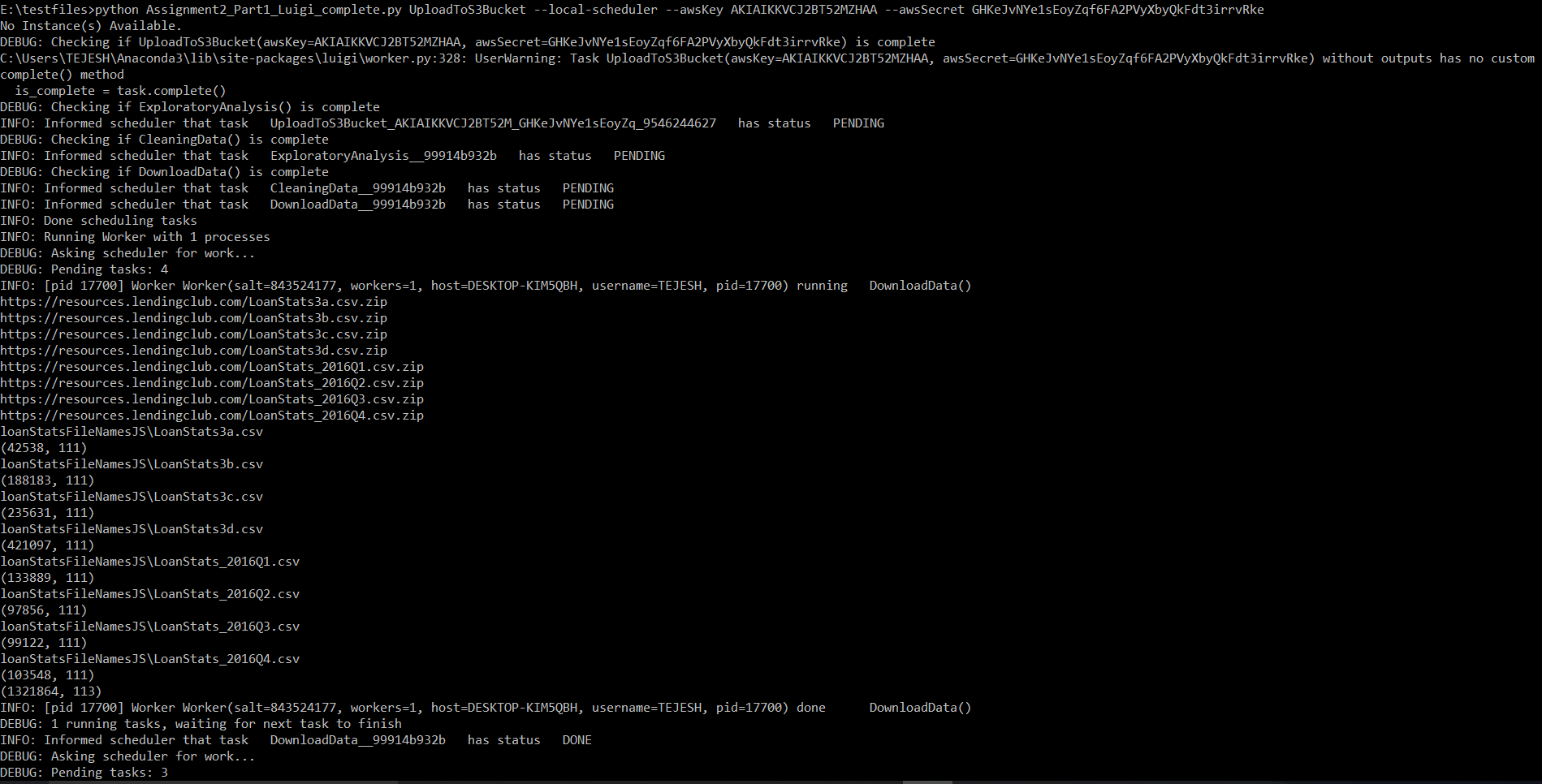
1. DownloadData
2. CleaningData
3. ExploratoryAnalysis
4. UploadToS3Bucket

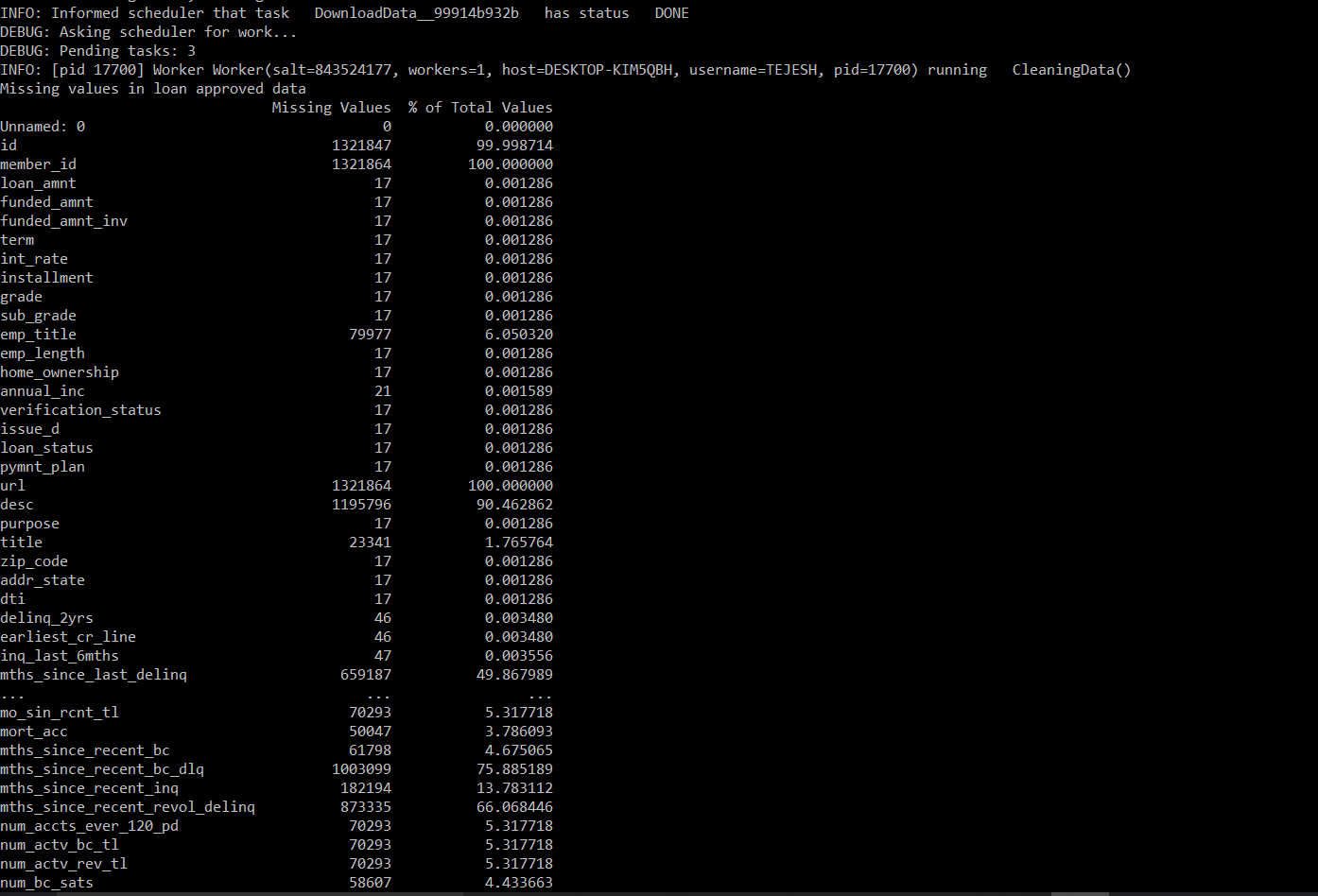
Output of each task is an intermediate processed file which is the output the next task.

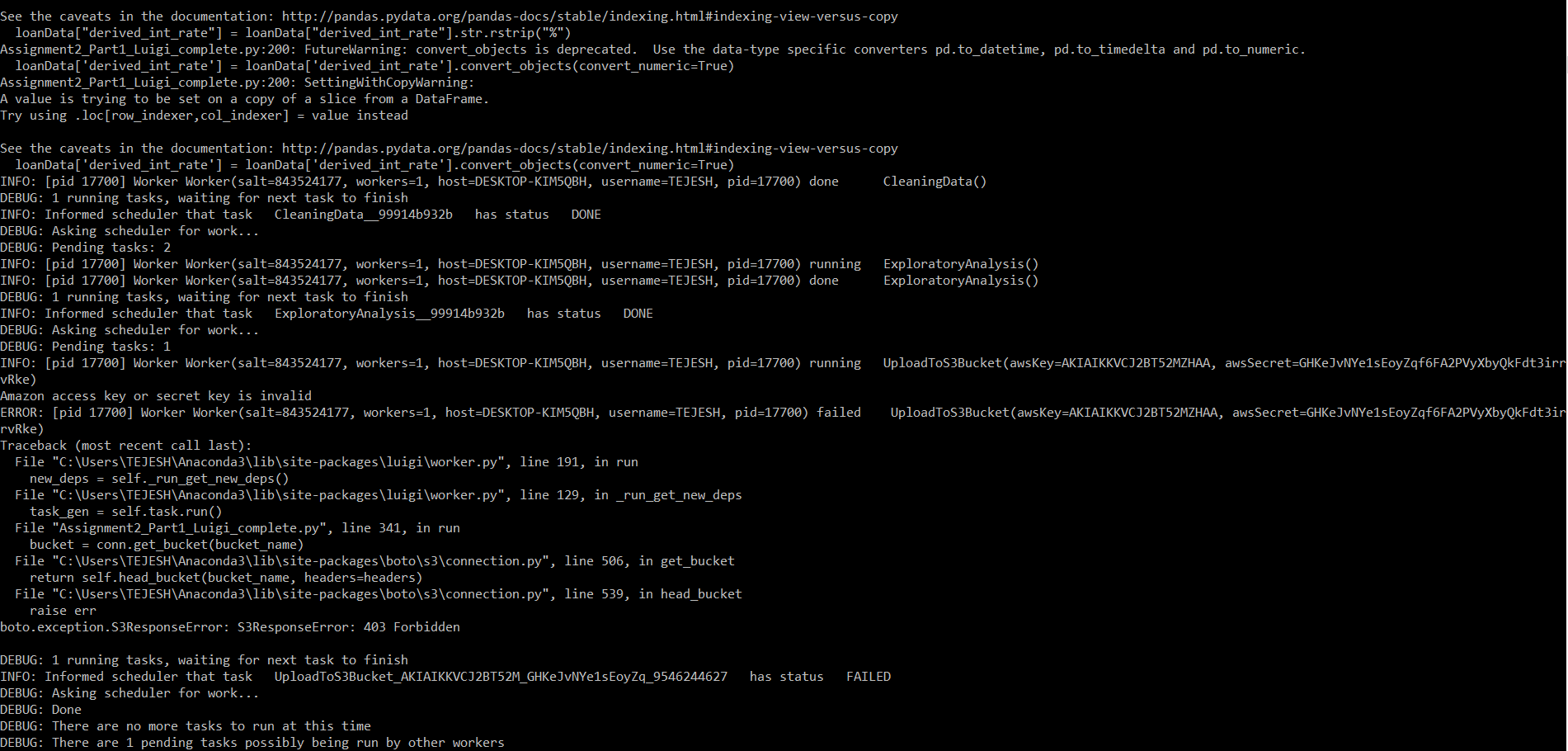
The final task does not have any output as it does not have any other task dependent on it. It directly uploads the data to S3 bucket.

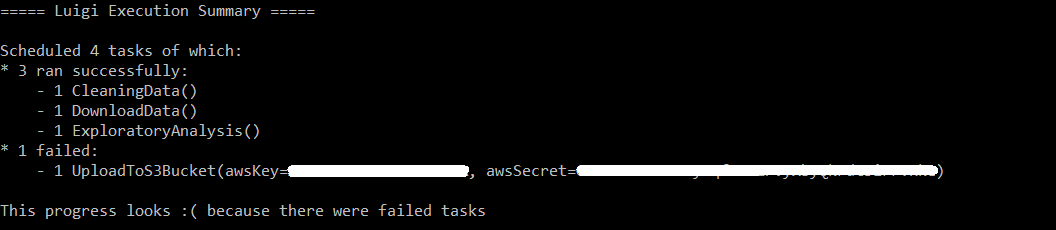
**Luigi Execution:**

We ran the entire process using Luigi workflow. Please find below the screenshots of the same showing the full execution of Luigi.









Steps followed to perform scheduling on Amazon AWS:

* Created an EC2 instance.
* Installed Docker on EC2 instance.
* Created repositories that can hold docker images on AWS.
* Created Task Definitions to run our pipeline.
* Created scheduling tasks to run the above task definitions.

Steps to be followed to replicate the above assignment:

docker pull harshadpardeshi/assignment2

docker run -ti harshadpardeshi/assignment2