**Data Wrangling, Pre-processing and Exploratory Data Analysis**

**By Harshad Pardeshi, Prithvi Kamath and Tejesh Rathi**

**PROBLEM STATEMENT 1: Data Wrangling Edgar data from text files**

Extract tables for 10q and 10k filings using R or Python for a given CIK and Document Accession Number. Parse through the files and extract all the tables and save them in a CSV format. The program should log the activities, zip the tables and upload the log file and zip file to Amazon S3. Also dockerize the entire pipeline to automate the process and take inputs from a config file.

**SOLUTION:**

We have followed the below steps to complete the above problem:

1. A log file is created at the beginning and all the below steps are logged in the same.
2. For a given CIK and Document Accession Key in the config file, a dynamic URL is generated which we parse through to find the 10q or 10k URL.
3. Once the 10q or 10k URL is found, we parse this URL to find all the tables and save them as a CSV in the current directory.
4. Once all the CSVs are downloaded, a ‘zip\_CIK.zip’ file is created.
5. The generated Zip file and the log file are then uploaded on an Amazon S3 bucket whose Access Key and Secret Key are also given by the user in a config file.

The entire project has been dockerized and the same can be executed by executing the below commands:

1. Pull the docker image

$ docker pull prithvik/adsassignment11

Change the parameters in config.json file as per your requirements

The parameters are ordered as CIK, Document Accession Key, Amazon Key ID, Amazon Secret Access Key

1. Run the docker file

$ docker run adsassignment11

**PROBLEM STATEMENT 2: Missing Data Analysis and Visualization**

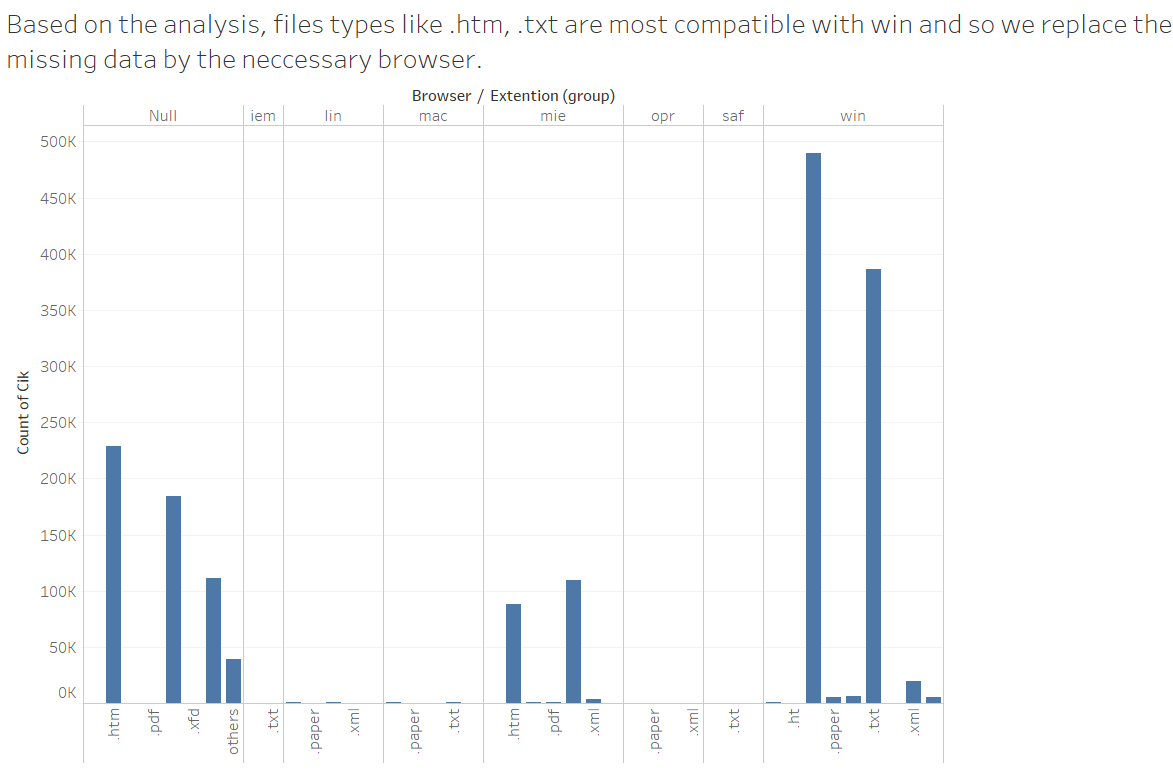
For a given year, your program in R or Python should get data for the first day of the month for every month by handling missing data, computing summary metrics, checking observable anomalies, generating log files, zipping year wise files and uploading the compiled data on Amazon S3 bucket. Build a dashboard on Tableau and analyze the data in Amazon S3 bucket.

**SOLUTION:**

**Handling Missing Data:**

* Missing data in columns ‘CIK’, ‘Size’ and ‘Browser’ have been handled.
* The rows with empty CIK values have been removed, because CIK values could are irreplaceable and also the count of such rows is very small. Hence we have assumed that neglecting such values will not make a major difference.
* Size value has been replaced by the mean size for the same combination IP and CIK. If the mean value is not available, then the size if replaced by zero.

Example: For IP 194:164:39:abc and CIK=100, if the value in Size column is null, then it is replaced by the mean value of Size, calculated for the same IP and CIK combination.

* Similarly, empty values in Browser have been replaced by mode of Browser for the same IP and CIK combination.

**Summary Metrics:**

* We analyzed the data and took the count of all the IP’s based on CIK. We observed the most number of browser compatible with a given company. We also showed this as a part of data in Tableau.

**Log file:**

* We have generated a log file that will record all the activities that we perform. A new file named “edgar-logs.log” is generated at the same location where the program is executed.

**Compile all files into one:**

* We looped across all the files and zipped it into a single folder.

**Amazon S3 Bucket:**

* The output of the program, that is, the zip files which contains combined log files and the log file is uploaded to the new S3 bucket for which the keys were provided.

The entire project has been dockerized and the same can be executed by executing the below commands:

1. Pull the docker image

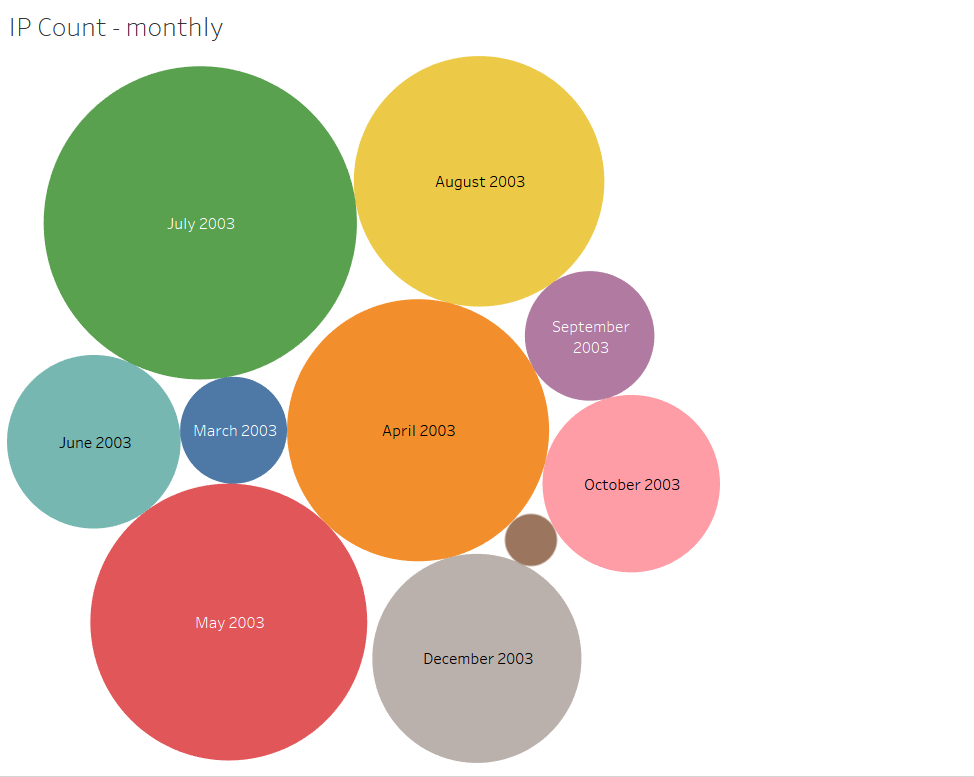
$ docker pull prithvik/adsassignment12

1. Run the docker file

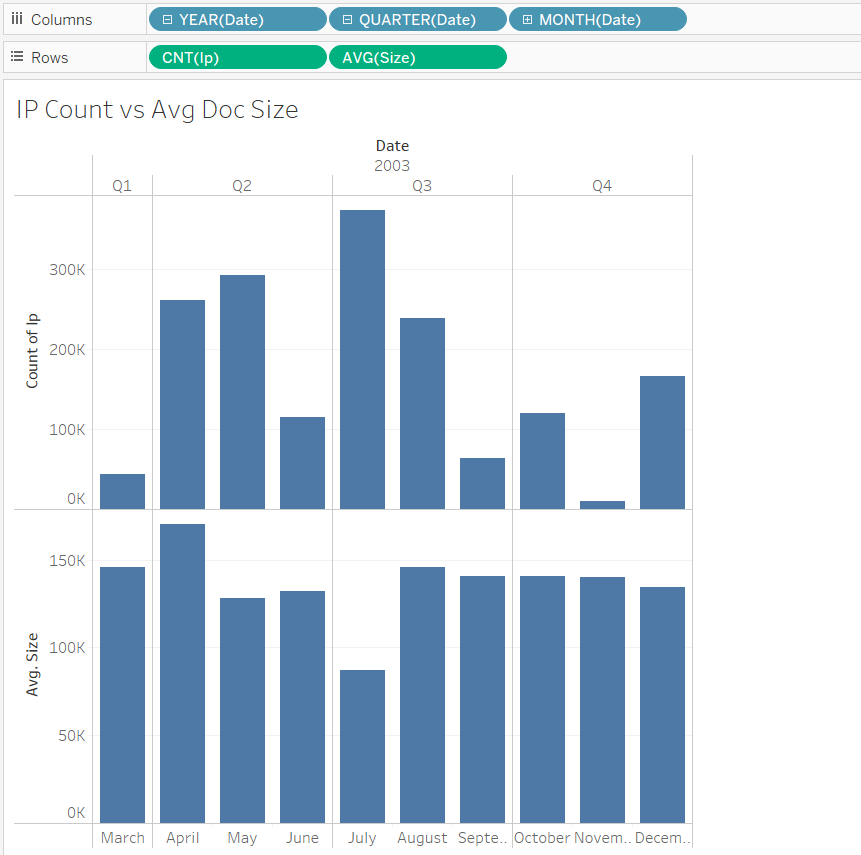
$ docker run adsassignment12

**The Tableau Visualization for the 2003 data is as below:**

1. The below visualization is based on the monthly IP’s accessed over the months. How many IP’s were accessed over in every month of the year 2003.



It gives us a comparison between the number of IP’s accessed in a particular month and the average size of documents uploaded for the same month. This shows that the average size doesn’t really matter with the number of IP’s hitting the website.



The link to the dashboard of our visualization is as below:

https://public.tableau.com/profile/publish/2003\_6/AnalysisDashboard#!/publish-confirm