**BIG DATA PROJECT ON KIVA LOANS ANALYSIS**

Submitted for the partial fulfillment of Degree

of

Bachelor of Technology

(Computer Science and Engineering)



**Submitted By:- Submitted to:-**

Abhinav Gupta Prof. Kapil Sharma

Uni Roll No. 1606629 CSE Department

D3CSE (E1)

Department of Computer Science and Engineering

Guru Nanak Dev Engineering College

Ludhiana 141006

**Company Certificate**



# ABSTRACT

This project is based on big data analytics , and the analytics is done on the kiva loans. This project is to help kiva to estimate the poverty levels of the residents in the regions where kiva has active loans. As kiva has an online crowding platform to extend financial services to poor and financially excluded people around the world so by analyzing the collected data will help the organization in many ways ,like, where too lend loans , for which sector more loan is taken ,which gender takes and returns loan properly , etc.

**Purpose & Scope:**

The purpose of this project is to analyze the data set and find out the different ways to get profit for the organization and to find the reasons for the loss . Also we can examine the data for giving loan ( whom to give and to whom not give ) , we can even study the problems for returning the loan, or for what reason loan is taken more etc. these all analysis will increase the demand of the organization

# ACKNOWLEDGEMENT

I am highly grateful to Dr. Amrit pal Singh, Director, Guru Nanak Dev Engineering College, Ludhiana, for providing opportunity to carry out six weeks industrial training at Center for Development of Advanced Computing, Mohali from May-July 2016.

Mr. Sunny (Hadoop Trainer), Sanjay Madan (R Trainer) have provided great help in carrying out the my work and is acknowledged with reverential thanks. Without the wise counsel and able guidance, it would have been impossible to complete the thesis in this manner.

I would like to express thanks profusely to Dr. Parminder Singh, Professor and Head (Computer Science & Engineering) for stimulating me time to time. I would also like to thank Err. Kapil Sharma, Training and Placement coordinator (Computer Science and Engineering) and to entire faculty, staff of computer science and engineering. I also thanks my friends who devoted their valuable time and helped me in all possible ways towards successful completion of this work.

(Abhinav Gupta)

Table of Figures

Figure 3.1(Preferred Tools For big data analytics Project) 6

Figure 3.2(Apache Hadoop Ecosystem) 9

Figure 3.3(High Level Architecture of Hadoop) 10

Figure 3.4(Hadoop Hive Connectivity) 11

Figure 3.5(Apache Pig Hadoop Connectivity) 13

Figure 3.6(Apache Pig vs Hive) 13

Figure 3.7(Tools Used) 14

Figure 4.1(Code Snapshot 1) 15

Figure 4.2(Code Snapshot 2) 15

Figure 4.3(Bar graph of Regions Kiva Funded) 16

Figure 4.4(Bar graph of Countries Kiva Funded) 16

Figure 4.5(Number Of Loans Per Country) 17

Figure 4.6(Multidimension poverty graph region wise) 17

Figure 4.7(Loans By Activity) 18

Figure 4.8(Loans By Sector) 18

Figure 4.9(Pie chart of Agriculture Sector) 19

Figure 4.10(Pie chart of Wholesale Sector) 19

Figure 4.11(Loans by Gender) 20

Figure 4.12(Loans repayment by Gender) 20

Figure 4.13(Pie chart of Men V/S Women Sector) 20

Figure 4.14(Bar graph of Loan repayment by Females) 21

Figure 4.15(Bar graph of Loan repayment by Females) 21

Contents

[ABSTRACT 3](#_Toc533110143)

[ACKNOWLEDGEMENT 3](#_Toc533110144)

[Chapter 1 3](#_Toc533110145)

[Chapter 2 4](#_Toc533110146)

[2.1 Problem Formulation : 4](#_Toc533110147)

[2.2 Objectives : 5](#_Toc533110148)

[Chapter 3 6](#_Toc533110149)

[3.1. Big data analytics : 6](#_Toc533110150)

[3.2.1. Big data analytics technologies and tools : 7](#_Toc533110151)

[Chapter 4 15](#_Toc533110152)

[References 22](#_Toc533110153)

# 

# Chapter 1

Introduction to Organization

I am having my Six Weeks Training at Centre For Development Of Advanced Computing (C-DAC), Mohali. Center for Electronics Design & Technology of India (CEDTI), Mohali was setup in May 1989. Primarily with the mission to train manpower in electronic design & technology by offering a variety of training programmes in diverse aspects of electronics design, product development, production technology, maintenance engineering, information technology and quality control, etc. In December 2002, CEDTI Mohali merged with C-DAC with a primary mandate to promote high end R&D along with education and training.

The centre is engaged in design and deployment of world class IT and electronics solutions in the following domains:

* Health Informatics
* Multilingual Technologies
* Professional Electronics
* Software Technologies
* Cyber Forensics and Security
* Multimedia Technologies

Centre continues to play a leading role in human resource development and training in Information Technology (IT) sector in the northern region. Center offers high-end courses like M.Tech in VLSI as well as ME in Electronic Product Design Technology (EPDT). Short term value added courses and diploma are designed for knowledge based skill development. It also offers courses for foreign participants, sponsored by MEA under ITEC/SCAAP programs.

Centre operates from its own impressive building located in the ELTOP (Electronics Town of Punjab) Complex amidst a large number of industries, manufacturing electronic products relating to computers, peripherals, communication equipment and components, offering a great professional challenge to the faculty and staff of the Centre.

# Chapter 2

Introduction to Project

Kiva Micro funds (commonly known by its domain name, Kiva.org) is a non-profit organization that allows people to lend money via the Internet to low-income entrepreneurs and students in over 80 countries. Kiva's mission is "to connect people through lending to alleviate poverty.

Kiva relies on a network of field partners to administer the loans on the ground. These field partners can be microfinance institutions, social businesses, schools or non-profit organizations. Kiva includes personal stories of each person who needs a loan because they want their lenders to connect with their entrepreneurs on a human level.

## Problem Formulation :

2.1.1 PEOPLE:

Who is the borrower and what do they do?

* Type of Business/Industry
* Organizational Structure
* Payment History with Your Institution
* Credit Report
* Information from Other Lenders/Parties
  + 1. PURPOSE OF LOAN:

What will the money be used for?

* Purchase of Specific Assets
* General Working Capital
* Asset-Based Lending
* Refinance (loan) Other Debt
  + 1. PAYMENT**:**

What are the terms of the debt and the borrowers ability/willingness to pay?

* Terms of Debt
* Source of Repayment
* Income to Debt Service Requirements
* Living/Other Expenses
* Other Sources of Income
  + 1. PROSPECTS:

What is the likelihood of the borrower

continuing to pay?

* Company Trends/Future Prospects
* Industry Trends/Future Prospects
* Economic Forecasts
* Succession/Continuation Plans
* New Developments
* New Laws and Legislation

## Objectives :

* Countries Where Kiva Need Promotion.
* Which sector is growing..
* MALE V/S FEMALE

# Chapter 3

Training Work Undertaken

Figure 3.1 :- Preferred Tools For big data analytics Project

## 3.1. Big data analytics :

Big data analytics is the process of examining large and varied data sets -- i.e., big data -- to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful information that can help organizations make more-informed business decisions.

Big data analytics applications enable [data scientists](https://searchbusinessanalytics.techtarget.com/definition/Data-scientist), predictive modelers, statisticians and other analytics professionals to analyse growing volumes of structured transaction data, plus other forms of data that are often left untapped by conventional business intelligence [(BI)](https://searchdatamanagement.techtarget.com/definition/business-intelligence) and analytics programs.

That encompasses a mix of [semi-structured](https://whatis.techtarget.com/definition/semi-structured-data) and [unstructured data](https://searchbusinessanalytics.techtarget.com/definition/unstructured-data) -- for example, internet [clickstream](http://searchsoa.techtarget.com/definition/click-stream) data, web server logs, social media content, text from customer emails and survey responses, mobile-phone call-detail records and machine data captured by sensors connected to the [internet of things](https://whatis.techtarget.com/definition/Internet-of-Things).

On a broad scale, [data analytics](https://searchdatamanagement.techtarget.com/definition/data-analytics) technologies and techniques provide a means of analyzing data sets and drawing conclusions about them to help organizations make informed business decisions. BI queries answer basic questions about business operations and performance. Big data analytics is a form of [advanced analytics](https://searchbusinessanalytics.techtarget.com/definition/advanced-analytics), which involves complex applications with elements such as [predictive models](https://searchdatamanagement.techtarget.com/definition/predictive-modeling), statistical algorithms and what-if analyses powered by high-performance analytics systems.

## 3.2.1. Big data analytics technologies and tools :

Many organizations that collect, process and analyze big data turn to [NoSQL](https://searchdatamanagement.techtarget.com/definition/NoSQL-Not-Only-SQL) databases as well as Hadoop and its companion tools, including:

* [**MapReduce**](https://searchcloudcomputing.techtarget.com/definition/MapReduce)**:** a software framework that allows developers to write programs that process massive amounts of unstructured data in parallel across a distributed cluster of processors or stand-alone computers.
* [**Spark**](https://searchbusinessanalytics.techtarget.com/definition/Apache-Spark)**:**an open-source parallel processing framework that enables users to run large-scale data analytics applications across clustered systems.
* [**HBase**](https://searchdatamanagement.techtarget.com/definition/Apache-HBase)**:** a column-oriented key/value data store built to run on top of the Hadoop Distributed File System (HDFS).
* [**Hive**](https://searchdatamanagement.techtarget.com/definition/Apache-Hive)**:** an open-source data warehouse system for querying and analyzing large datasets stored in Hadoop files.
* [**Pig**](https://searchdatamanagement.techtarget.com/definition/Apache-Pig)**:** an open-source technology that offers a high-level mechanism for the parallel programming of MapReduce jobs to be executed on Hadoop clusters.

**R LANGUAGE :**

R is a language and environment for statistical computing and graphics. It is a [GNU project](http://www.gnu.org/) which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.

One of R’s strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control.

R is available as Free Software under the terms of the [Free Software Foundation](http://www.gnu.org/)’s [GNU General Public License](https://www.r-project.org/COPYING) in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS.

**PYHTON:**

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed.

Often, programmers fall in love with Python because of the increased productivity it provides. Since there is no compilation step, the edit-test-debug cycle is incredibly fast. Debugging Python programs is easy: a bug or bad input will never cause a segmentation fault. Instead, when the interpreter discovers an error, it raises an exception. When the program doesn't catch the exception, the interpreter prints a stack trace. A source level debugger allows inspection of local and global variables, evaluation of arbitrary expressions, setting breakpoints, stepping through the code a line at a time, and so on. The debugger is written in Python itself, testifying to Python's introspective power. On the other hand, often the quickest way to debug a program is to add a few print statements to the source: the fast edit-test-debug cycle makes this simple approach very effective.

**Apache Hadoop :**

Apache Hadoop is an open source software framework for storage and large scale processing of data-sets on clusters of commodity hardware. Hadoop is an Apache top-level project being built and used by a global community of contributors and users. It is licensed under the Apache License 2.0.

The Apache Hadoop framework is composed of the following modules:

* Hadoop Common: contains libraries and utilities needed by other Hadoop modules
* Hadoop Distributed File System (HDFS): a distributed file-system that stores data on the commodity machines, providing very high aggregate bandwidth across the cluster
* Hadoop YARN: a resource-management platform responsible for managing compute resources in clusters and using them for scheduling of users' applications
* Hadoop MapReduce: a programming model for large scale data processing

All the modules in Hadoop are designed with a fundamental assumption that hardware failures (of individual machines, or racks of machines) are common and thus should be automatically handled in software by the framework.

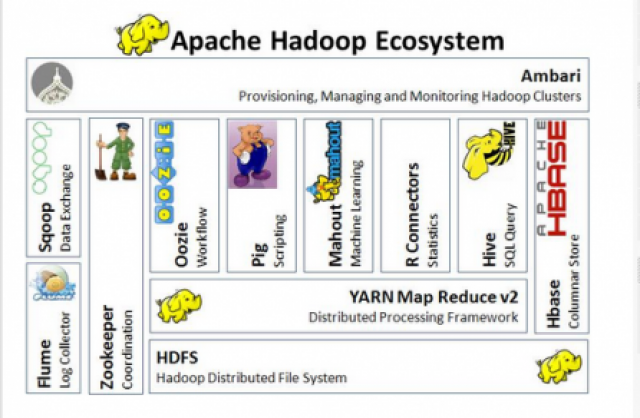


Figure 3.2:- Apache Hadoop Ecosystem

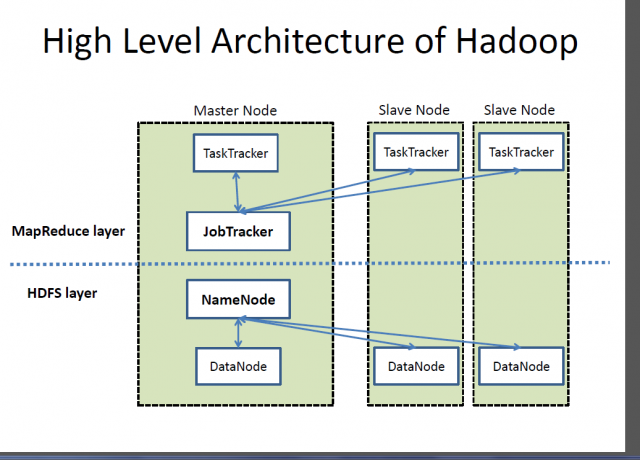
****

Figure 3.3:- High Level Architecture of Hadoop

**Hive :**

Hive is a data warehousing infrastructure built on top of apache Hadoop.

Hadoop provides massive scale-out and fault-tolerance capabilities for data storage and processing (using the MapReduce programming paradigm) on commodity hardware.

Hive enables easy data summarization, ad-hoc querying and analysis of large volumes of data.

It is best used for batch jobs over large sets of immutable data (like web logs).

It provides a simple query language called Hive QL, which is based on SQL and which enables users familiar with SQL to easily perform ad-hoc querying, summarization and data analysis.

At the same time, Hive QL also allows traditional MapReduce programmers to be able to plug in their custom mappers and reducers to do more sophisticated analysis that May not be supported by the built-in capabilities of the language.

Hive query language provides the basic SQL like operations. These operations work on tables or partitions.

* Ability to create and manage tables and partitions (create, drop and alter).
* Ability to support various Relational, Arithmetic and Logical Operators.
* Ability to do various joins between two tables.
* Ability to evaluate functions like aggregations on multiple “group by” columns in a table.
* Ability to store the results of a query into another table.
* Ability to download the contents of a table to a local directory.
* Ability to create an external table that points to a specified location within HDFS
* Ability to store the results of a query in an HDFS directory.
* Ability to plug in custom scripts using the language of choice for custom map/reduce jobs.

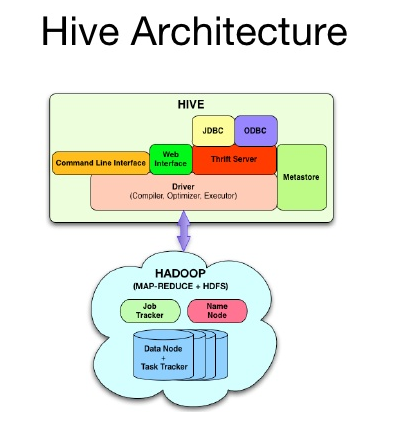


Figure 3.4:- Hadoop Hive Connectivity

Pig:

Apache Pig is an open-source technology that offers a high-level mechanism for the parallel programming of [MapReduce](https://searchcloudcomputing.techtarget.com/definition/MapReduce) jobs to be executed on [Hadoop clusters](https://searchbusinessanalytics.techtarget.com/definition/Hadoop-cluster). Pig enables developers to create query execution routines for analyzing large, distributed data sets without having to do low-level work in MapReduce, much like the way the [Apache Hive](https://searchdatamanagement.techtarget.com/definition/Apache-Hive)data warehouse software provides a SQL-like interface for [Hadoop](https://searchcloudcomputing.techtarget.com/definition/Hadoop) that doesn't require direct MapReduce programming,

The key parts of Pig are a compiler and a scripting language known as Pig Latin. Pig Latin is a data-flow language geared toward parallel processing. Managers of the [Apache Software Foundation](https://whatis.techtarget.com/definition/Apache-Software-Foundation)'s Pig project position the language as being part way between declarative [SQL](https://searchsqlserver.techtarget.com/definition/SQL) and the procedural [Java](https://www.theserverside.com/definition/Java) approach used in MapReduce applications. Proponents say, for example, that data joins are easier to create with Pig Latin than with Java. However, through the use of user-defined functions (UDFs), Pig Latin applications can be extended to include custom processing tasks written in Java as well as languages such as [JavaScript](https://searchmicroservices.techtarget.com/definition/JavaScript) and [Python](https://whatis.techtarget.com/definition/Python).

Apache Pig grew out of work at Yahoo Research and was first formally described in a paper published in 2008. Pig is intended to handle all kinds of data, including structured and unstructured information and relational and nested data. That omnivorous view of data likely had a hand in the decision to name the environment for the common barnyard animal. It also extends to Pig's take on

application frameworks; while the technology is primarily associated with Hadoop, it is said to be capable of being used with other [frameworks](https://whatis.techtarget.com/definition/framework) as well.

The underlying [Hadoop](https://searchcloudcomputing.techtarget.com/definition/Hadoop) framework grew out of large-scale Web applications whose architects chose non-SQL methods to economically collect and analyze massive amounts of data. It has lots of add-on help for handling [big data](https://searchdatamanagement.techtarget.com/definition/big-data) applications because Apache Pig is just part of a long list of Hadoop ecosystem technologies that also includes Hive, [HBase](https://searchdatamanagement.techtarget.com/definition/Apache-HBase), [ZooKeeper](https://searchcloudcomputing.techtarget.com/definition/Apache-ZooKeeper) and other utilities intended to fill in functionality gaps in the framework

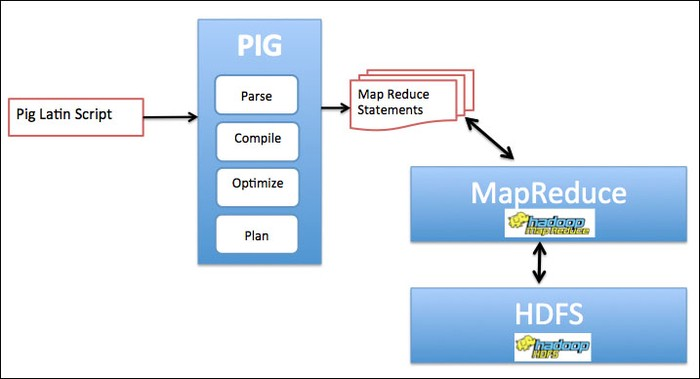


Figure 3.5;- Apache Pig Hadoop Connectivity

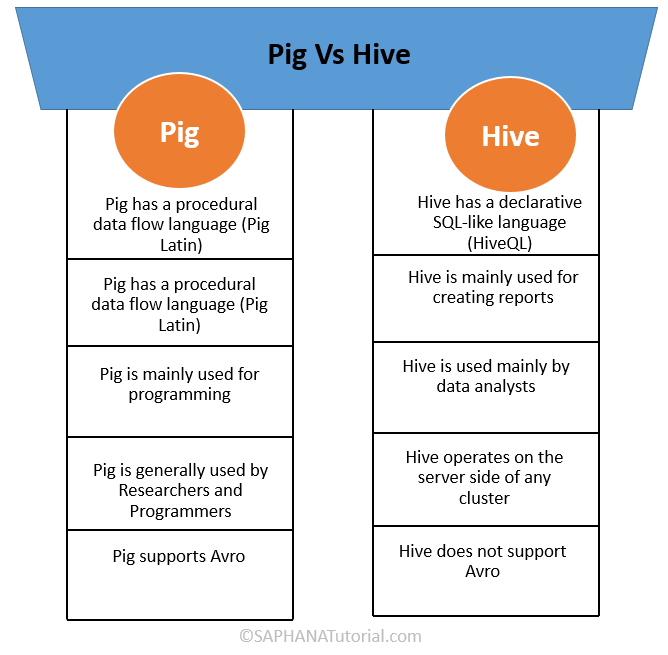
Figure 3.6: -Apache Pig vs Hive

Figure 3.7:- Tools Used

# Chapter 4

Results And Discussion

****

Figure 4.1:- Code Snapshot 1

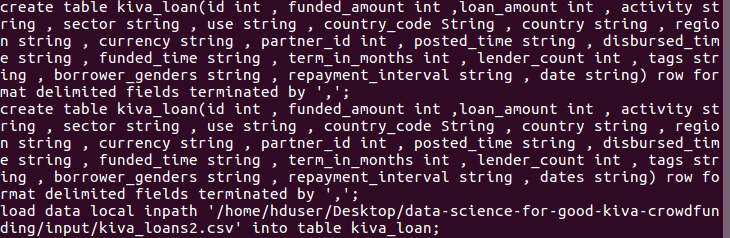
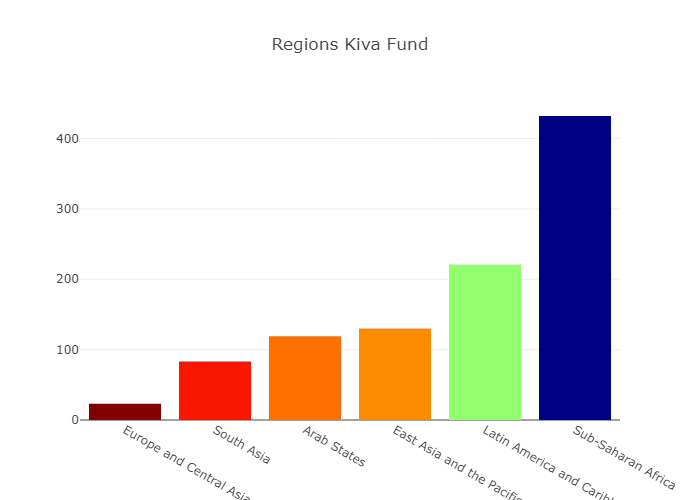
****

Figure 4.2:- Code Snapshot 2

**COUNTRIES WHERE KIVA NEEDS PROMOTIONS:**

**[](https://plot.ly/~abhi11nav/3.embed)** Figure 4.3:- Regions Kiva Funded

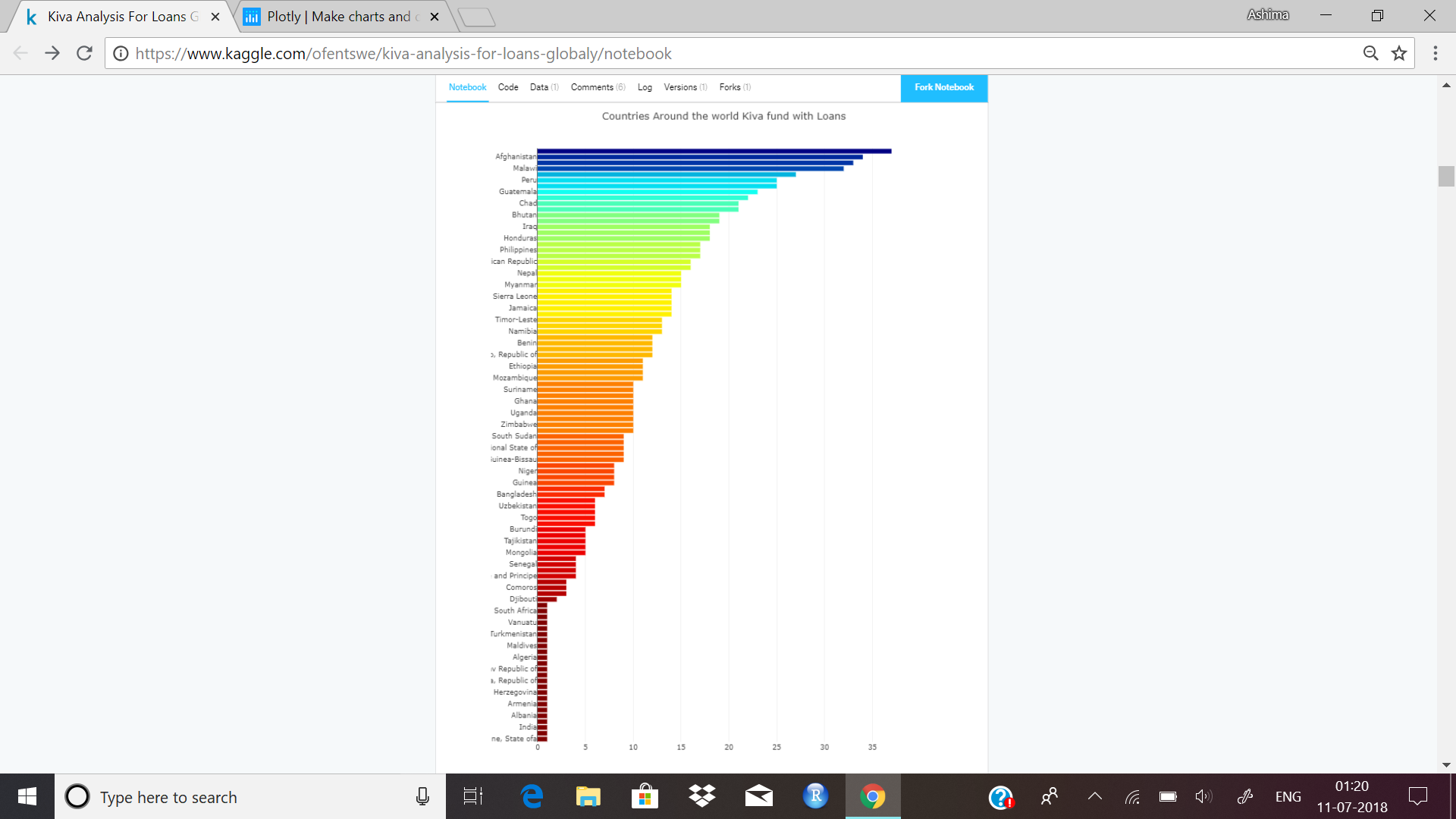
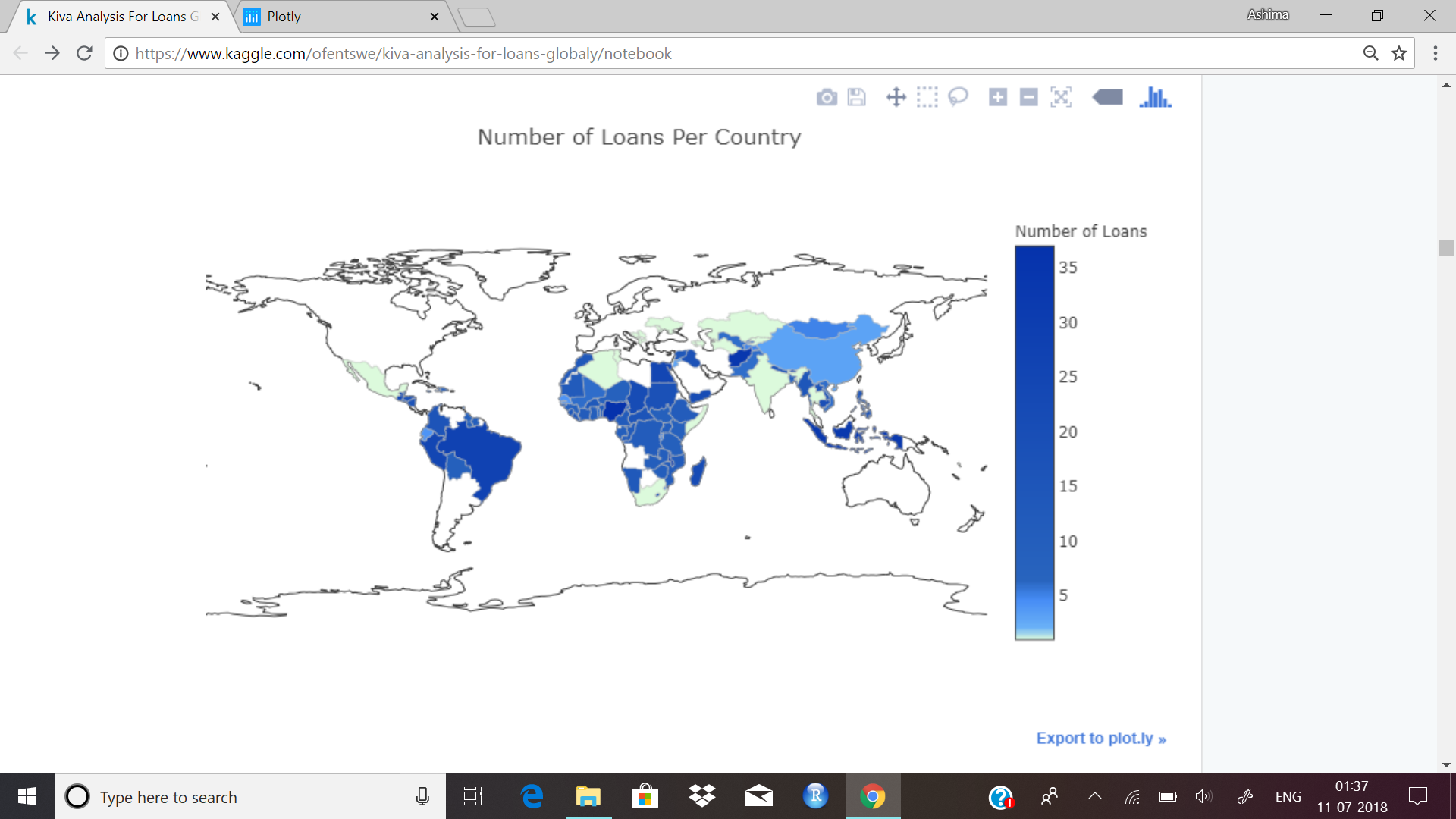
**[](https://plot.ly/~abhi11nav/1.embed)**

Figure 4.4:- Countries Kiva Fund

**[](https://plot.ly/~abhi11nav/5.embed)**Figure 4.5:- Number of Loans Per Country

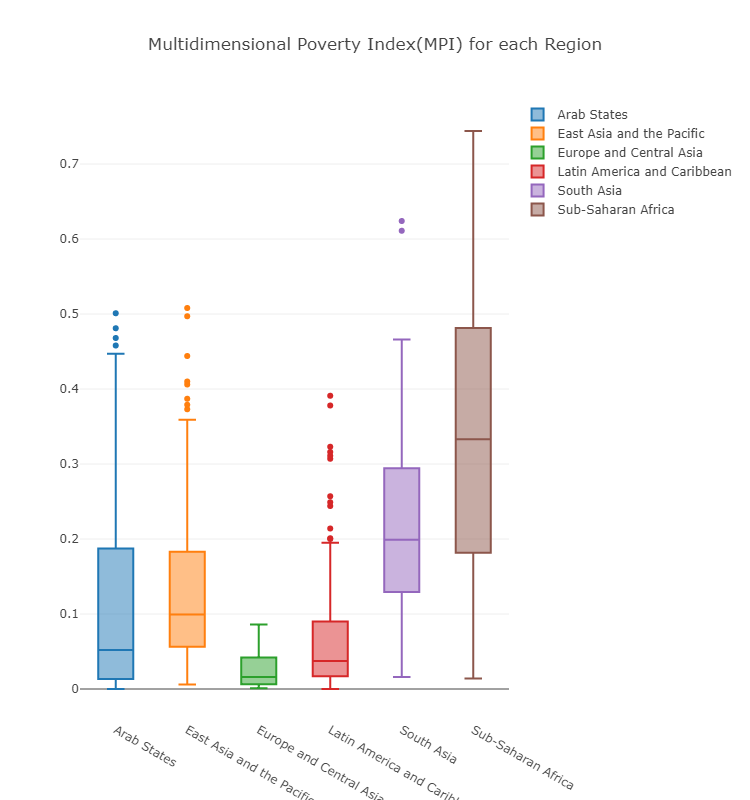
**[](https://plot.ly/~abhi11nav/9.embed)**

Figure 4.6:- Multidimension poverty graph region wise

From the above analysis we get to know that kiva needs to promote itself more in the countries like algeria, somalia and saudi arabia

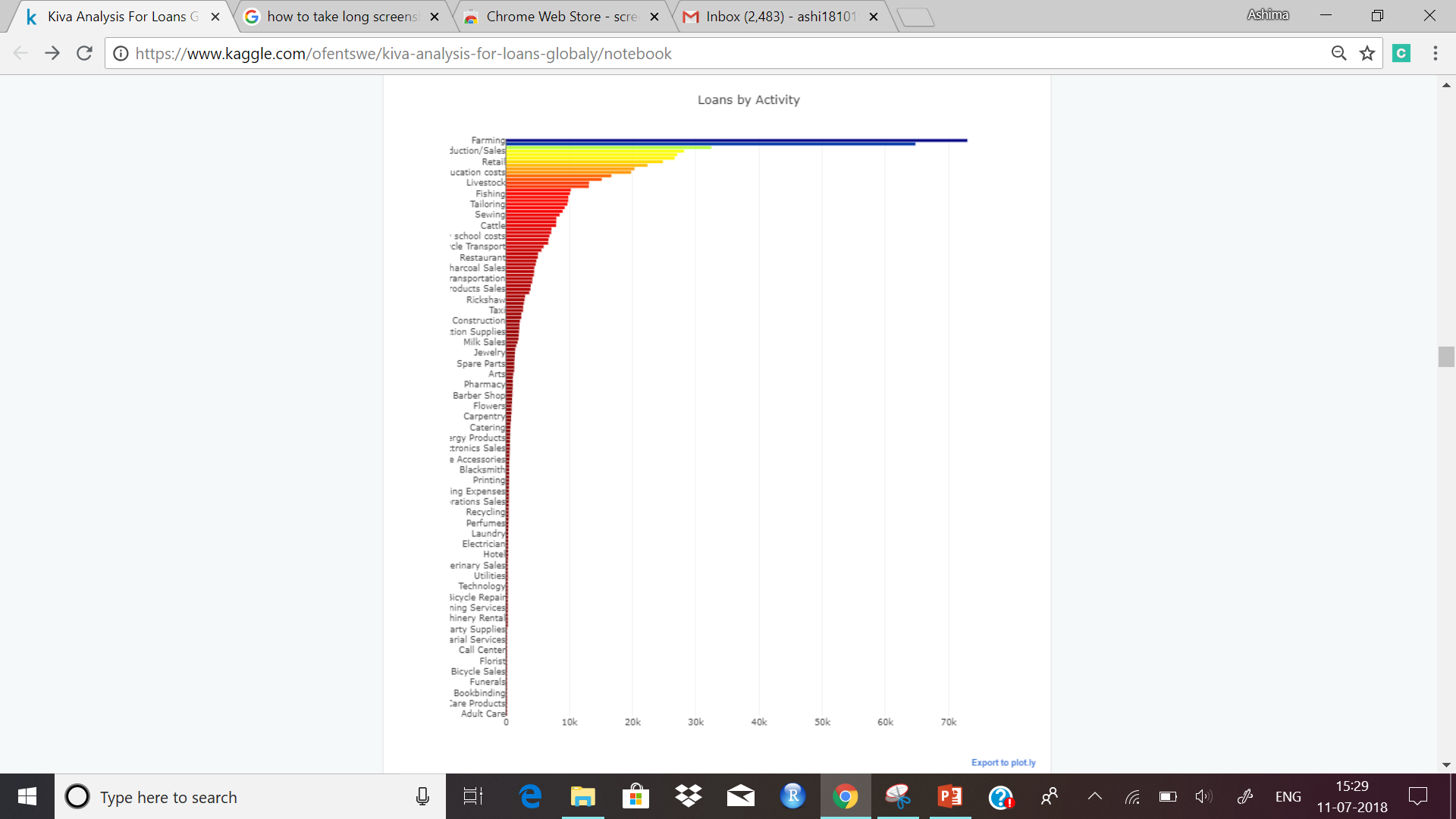
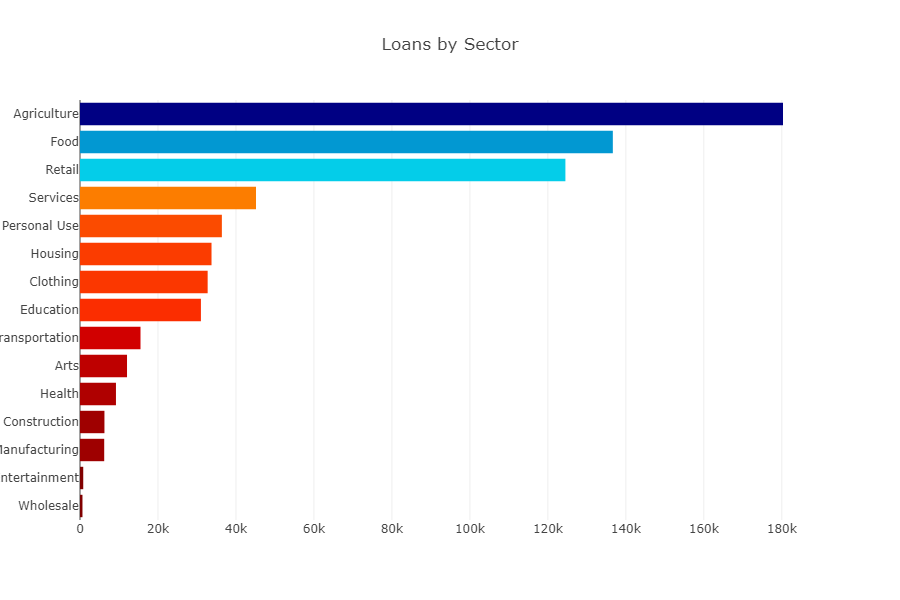
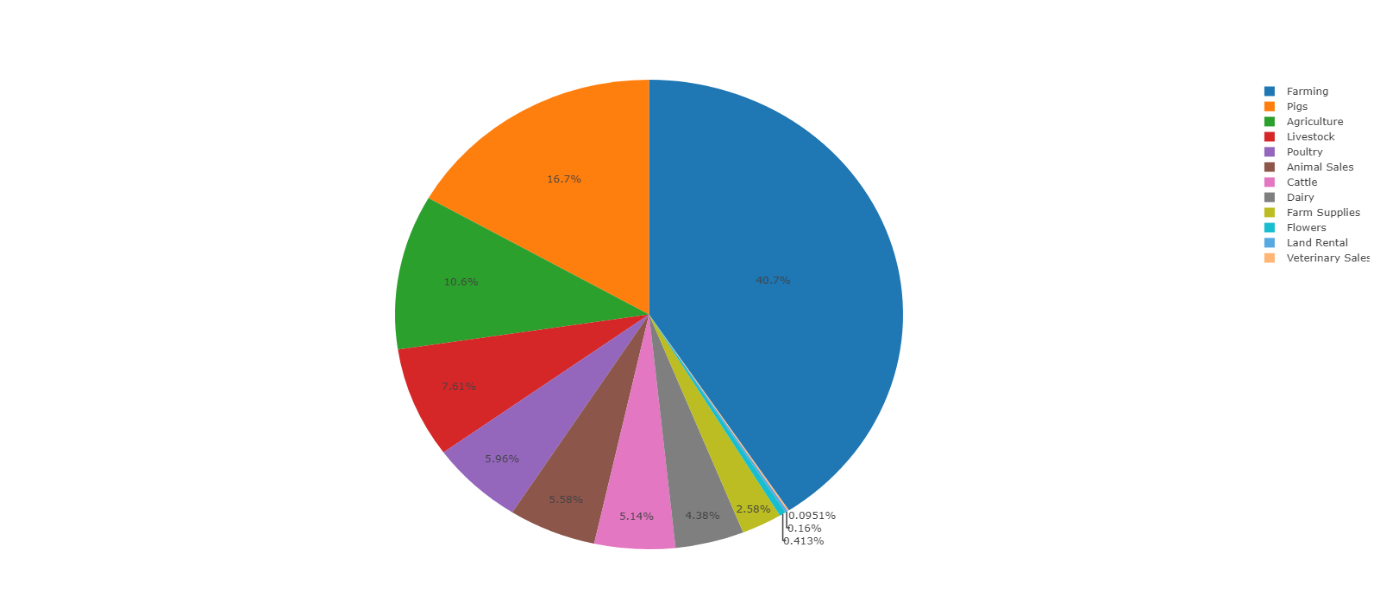
**WHICH SECTOR IS GROWING[](https://plot.ly/~abhi11nav/11.embed)**

Figure 4.7:- Loans By Activity

**[](https://plot.ly/~abhi11nav/13.embed)**Figure 4.8:- Loans By Sector

**Agriculture :**

**[](https://plot.ly/~aishrock006/27.embed)** Figure 4.9:- Pie chart of Agriculture Sector

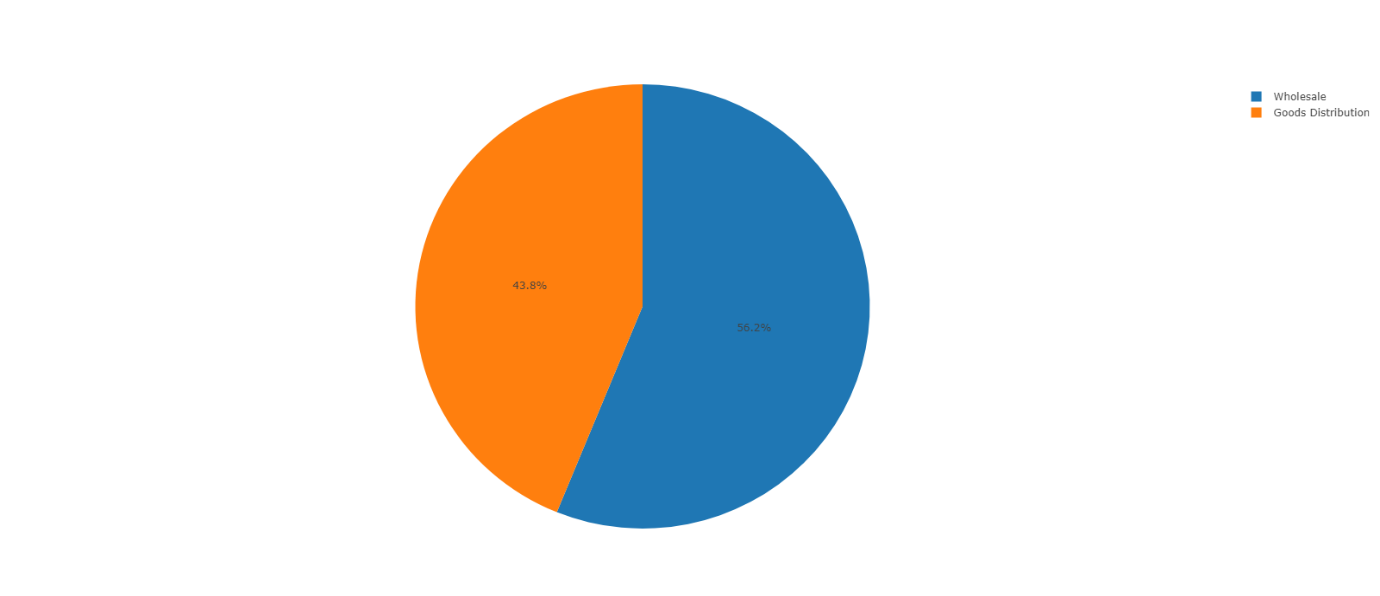
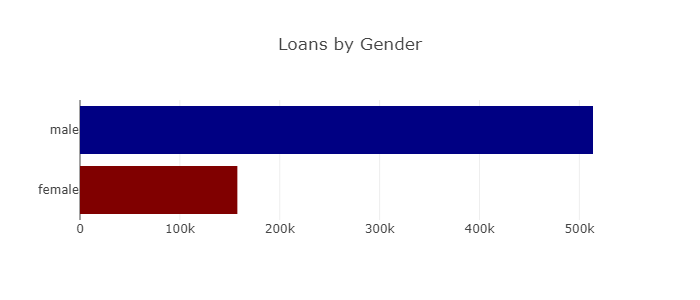
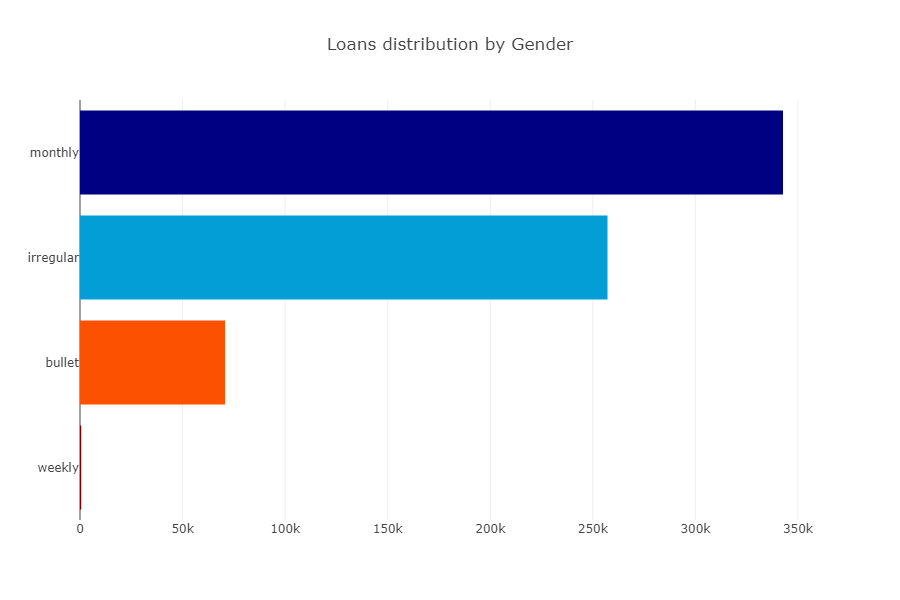
**Wholesale : [](https://plot.ly/~aishrock006/25.embed)**

Figure 4.10:- Pie chart of Wholesale Sector

Agriculture Sector is growing Fastest.

**[](https://plot.ly/~abhi11nav/17.embed)**Figure 4.11:- Loans by Gender

****Figure 4.12:- Loans repayment by Gender

Men vs women sector :

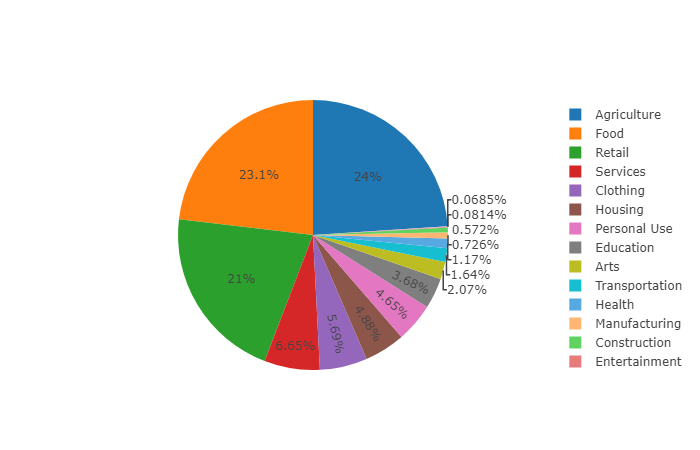
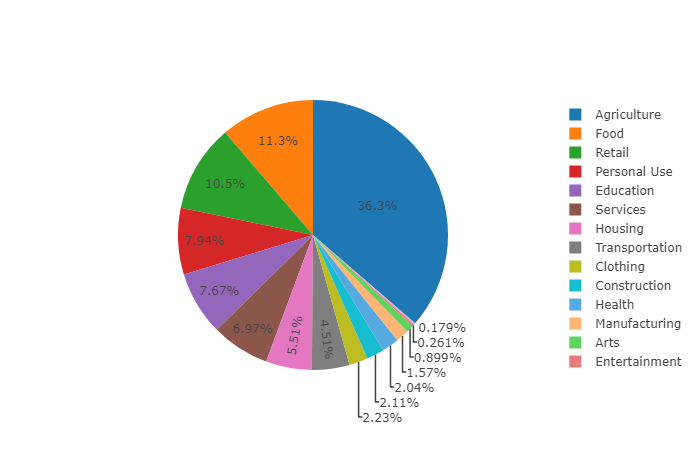
[](https://plot.ly/~abhi11nav/24.embed)**[](https://plot.ly/~abhi11nav/25.embed)**

Figure 4.13:- Pie chart of Men V/S Women Sector

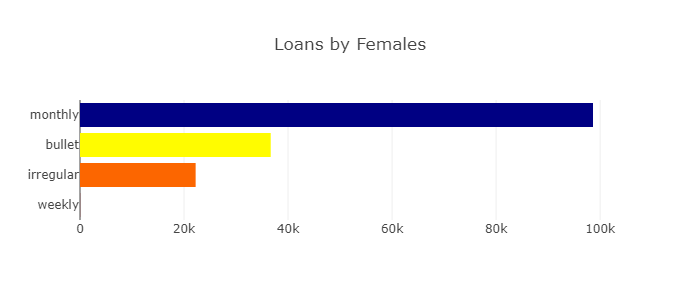
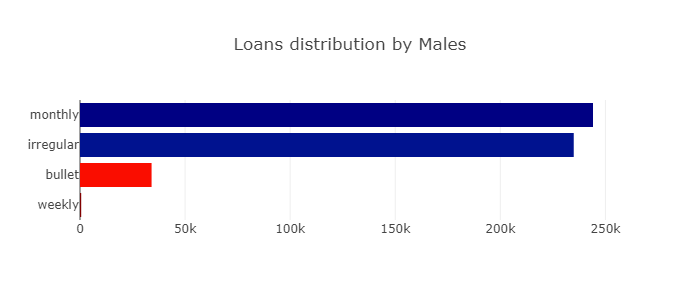
**[](https://plot.ly/~abhi11nav/21.embed)[](https://plot.ly/~abhi11nav/19.embed)**

Figure 4.14:- Bar graph of Loan repayment by Females Figure 4.15:- Bar graph of Loan repayment by Male

# References

1. <https://www.tutorialspoint.com/r/r_overview.htm> (15 June, 2018)
2. <https://www.tutorialspoint.com/data_mining/dm_tasks.htm> (19 June, 2018)
3. <https://www.tutorialspoint.com/hadoop/> (24 June, 2018)
4. <https://www.tutorialspoint.com/hive/> (23 June, 2018)
5. <https://www.tutorialspoint.com/apache_pig/> (25 June, 2018)
6. <https://kaggle.com/datasets> (Downloaded dataset on 1 July, 2018)