# CUSTOMER SEGMENTATION IN MORTGAGE INDUSTRY

Capstone Project Milestone Report

Pundir, Saurabh 6/30/18 SB-DataScience Course

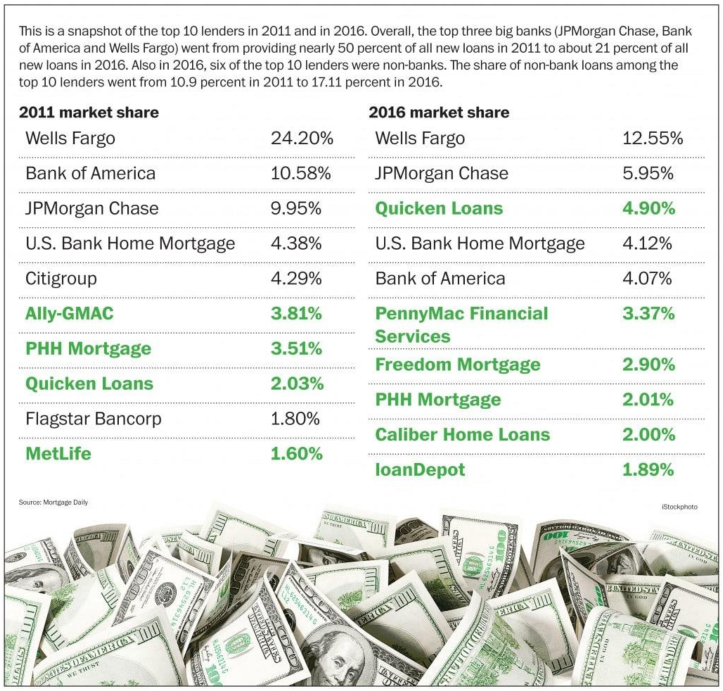
# Table of Contents

# 1. Introduction

## 1.1. About:

The mortgage market is undergoing dramatic transformations. The overall mortgage market today is huge, with **$ 10 trillion-plus** in outstanding mortgages and projections of $2 trillion for this year.

In 2011, 50 percent of all new mortgage money was loaned by the three biggest banks in the United States: **JPMorgan Chase, Bank of America and Wells Fargo**. But by **September 2016**, the share of loans by these three big banks dropped to **21 percent**. At the same time, six of the top 10 largest lenders by volume were non-banks, such as **Quicken Loans,** Loan Depot, and PHH Mortgage, compared with just two of the top 10 in 2011.



This is a snapshot of the top 10 lenders in 2011 and in 2016. Overall, the top three big banks (JPMorgan Chase, Bank of America and Wells Fargo) went from providing nearly 50 percent of all new loans in 2011 to about 21 percent of all new loans in 2016. Also in 2016, six of the top 10 lenders were non-banks. The share of non-bank loans among the top 10 lenders went from 10.9 percent in 2011 to 17.11 percent in 2016.

| **2011 market share** | | **2016 market share** | |
|---|---|---|---|
| Wells Fargo | 24.20% | Wells Fargo | 12.55% |
| Bank of America | 10.58% | JPMorgan Chase | 5.95% |
| JPMorgan Chase | 9.95% | **Quicken Loans** | **4.90%** |
| U.S. Bank Home Mortgage | 4.38% | U.S. Bank Home Mortgage | 4.12% |
| Citigroup | 4.29% | Bank of America | 4.07% |
| **Ally-GMAC** | **3.81%** | **PennyMac Financial Services** | **3.37%** |
| **PHH Mortgage** | **3.51%** | **Freedom Mortgage** | **2.90%** |
| **Quicken Loans** | **2.03%** | **PHH Mortgage** | **2.01%** |
| Flagstar Bancorp | 1.80% | **Caliber Home Loans** | **2.00%** |
| **MetLife** | **1.60%** | **loanDepot** | **1.89%** |

Source: Mortgage Daily

iStockphoto

In such a competitive and dynamic market its critical to obtain a deep understanding of consumer needs and behaviors to quickly find and close high-quality loans

### 1.1..1. Formal Definition - Customer Segmentation:

Customer Segmentation is the subdivision of a market into discrete **customer groups** that share **similar characteristics**. Customer Segmentation can be a powerful means to identify unmet customer needs.

Companies that identify **segments** can then outperform the competition by **developing uniquely appealing products** and services. This prioritization can help companies develop marketing campaigns and pricing strategies to **extract maximum value** from both high- and low-profit customers. A company can use Customer Segmentation as the principal basis for **allocating resources** to product development, marketing, service and delivery programs.

### 1.2. Problem:

The top three lenders have a total share of $150 million, with the leader having a share of $80 million. Rest of the pie is distributed amongst 400 small and medium-sized lenders.

According to the problem defined by top players, the mortgage companies have not been able to understand their target customers and provide suitable loan products.

As per a recent study conducted by J.D. Powers, **63%** of customers would **leave their mortgage servicer** for better customer service. The same study shows that **27% of first-time buyers and 21% of all borrowers regret their choice of lender.**

Mortgage companies want to increase their market share and for doing so they need to understand their customers better. This project aims to use data from Consumer Financial Protection Bureau and build an unsupervised machine learning model to segment their customer base.

The problem in this project to be addressed is that ***"Is it possible to identify the segment of mortgage customer based on the loans provided to the customer in last three years?"***

### 1.3. Clients and Audiences

Mortgage companies like **Quicken Loans, Wells Fargo, Chase** and credit unions, etc. can use such a model to understand their customer base. They can create products catering to segments, decide marketing allocation, and formulate the strategy for reaching out to the segmented customers in a unique way.

The data also has features like the type of homes, size of family and features related to clients etc. This segmentation can also provide insights for **Real Estate companies** as every Mortgage customer is also a Real Estate customer

### 1.4. Scope:

The customer segmentation is the very wide topic and can involve various data points but this project is scoped to use only loan data available from HMDA. Also, this is scoped to only take data from 2014-2016.

The technology is currently scoped to use local computer machine for all processing. The processing can be enhanced by performing the wrangling and running machine learning on a cloud with more processing capabilities.

### 1.5. Question Client(s) really care about

All the above clients and audience have a concerning question...

**"What are the segments of mortgage customer identified based on previous loan customer to identify the common pattern for better target marketing and servicing?"**

## 2. Data Wrangling

The data wrangling section involved getting data and extracting useful columns or variable from the files for further processing. The individual section explains the approach applied for the cleaning and wrangling of data to extract data for all further analysis.

## 2.1. Getting File

Data is available as CSV file on Consumer Financial Protection Bureau website. The data is available based on a yearly basis. Each month file size is around 2.5 GB.

The mapping file from county FIPS code to county and state standard census code.

## 2.2. File Content:

Each yearly file contains information of 1-2 million loans. There are 36 distinct fields for each project. The field details are mentioned in the data dictionary section.

## 2.3. Choosing Yearly Files:

Due to the large volume of data and processing restriction data for only 2015-2017 is considered. Each file is read and appended in the data frame to create one final for further processing

## 2.4. 3 Step Approach

Data wrangling was performed and included three stage process.

1. Extracting data for each year.
2. Data Cleaning and consolidation into a single file.
3. Storing data for further processing.

## 2.5. Extracting Data:

There are 3 files consolidated into a single file. Reading files inside the folder is performed using "*os.path*" module.

The data with a null value for significant columns like loan amount Applicant gender and income etc are removed. This is decided because these are an important feature in understanding data and these cannot be filled with any other average or guess value.

The code related to this part is in CapstoneII_DataWrangling_I_ReadFile.ipynb file.

## 2.6. Cleaning Data:

After reading data into a data frame further processing is performed to get the state and county name from FIPS code. The data from the census is merged with loan data and extracted based on first county FIPS code and further for county value null directly on state FIPS code.

The columns are converted from object type to integer, Boolean and category for faster processing and reducing file size while saving.

Another round of null value cleaning is performed after this on columns like state, gender, ethnicity, and result. All the columns which will be used as a feature in further processing are considered in clean up null values.

The Code related to this part is in the CapstoneII_DataWrangling_II_DataUpdate.ipynb file.

## 2.7. Wrangling for data story & inferential statistics:

The new column result is created to categories data into Loan Approved (1), Loan Denied (0) and NA (null).

The data is further processed to filter out columns which are not being explored in data story to get the data frame and data store

## 2.8. Wrangling for machine learning:

For machine learning, all the columns containing feature are converted into dummy columns. The following columns are converted into dummy columns

1. StateCode
2. ApplicantEthnicity
3. ApplicantRace
4. ApplicantSex
5. Occupancy
6. PropertyType
7. LoanPurpose

The salary and loan amount columns are converted into dummy based on amount range. The following ranges in (1000 USD) are uses to convert into dummy columns.

1. 0-50
2. 50-100
3. 100-150
4. 150-200
5. 200-250
6. 250-300
7. 300-350
8. 350-400
9. 400-450
10. 500-5500
11. 5500-999999

Another variation of using salary and loan amount as a feature without converting into ranges is also been explored.

Due to processing restriction in the local computer, the data is further restricted to use **one million records** from **California** state.

After extracting these columns. Only above columns are kept in a data frame for machine learning. The data is stored in the separate file only for the machine learning purposes.

The Code related to this part is in Capstone2_DataWrangling_III_ML.ipynb file.

## 2.9. Storing Data:

**Challenge:**

The challenge in cleaning and wrangling and cleaning this data is the processing time. Processing some columns take 3-4 hours on medium configuration machine (laptop) available.

After cleaning data and creating final data frame for the file. The file is stored on the hard disk to avoid re-running the time-consuming process. The pickle object sterilization is utilized to store data frame.

This file is unpickled and processed in further steps. The process of unpickling and making the file available is under 5 minutes.

This approach provided the benefit of processing the data once and then utilizing it further at a faster pace.

**Further Work:**

Hadoop and map -reduce technology can be utilized to make this process better. But currently, this was not the scope of our project.

## 2.10. Data Dictionary:

The data will be acquired from the Consumer Financial Protection Bureau. In this project, we will mainly focus on the data collected during 2015-2017. There is a possibility to use census data to gather more demographic information.

The data is available in CSV and similar delimiter format. During our study, we will consider the below-mentioned data fields.

| Sr. No | Field Name | Field Type | Valid Values | Descriptions and Examples |
|---|---|---|---|---|
| 1 | Record Identifier - Value is 2 | Numeric | 2 | |
| 2 | Respondent-ID | Alphanumeric | | Please see?RID for 2017 HMDA Filers? table above |
| 3 | Agency Code | Numeric | 1 2 3 5 7 9 | Descriptions: 1. Office of the Comptroller of the Currency (OCC) 2. Federal Reserve System (FRS) 3. Federal Deposit Insurance Corporation (FDIC) 5. National Credit Union Administration (NCUA) 7. United States Department of Housing and Urban Development (HUD) 9. Consumer Financial Protection Bureau (CFPB) |
| 4 | Loan Type | Numeric | 1 2 3 4 | Descriptions:1. Conventional (any loan other than FHA VA FSA or RHS loans) 2. FHA-insured (Federal Housing Administration) 3. VA-guaranteed (Veterans Administration) 4. FSA/RHS-guaranteed (Farm Service Agency or Rural Housing Service) |
| 5 | Property Type | Numeric | 1 2 3 | Descriptions: 1. One to four-family (other than manufactured housing) 2. Manufactured housing 3. Multifamily |
| 6 | Loan Purpose | Numeric | 1 2 3 | Descriptions: 1. Home purchase 2. Home improvement 3. Refinancing |

| | | | | |
|---|---|---|---|---|
| 7 | Owner Occupancy | Numeric | 1 2 3 | Descriptions: 1. Owner-occupied as a principal dwelling 2. Not owner-occupied 3. Not applicable |
| 8 | Loan Amount | Numeric | | A report in thousands. Round to the nearest thousand without leading zeros and without commas. Example: 111 |
| 9 | Preapprovals | Numeric | 1 2 3 | Descriptions: 1. Preapproval was requested 2. Preapproval was not requested 3. Not applicable |
| 10 | Type of Action Taken | Numeric | 1 2 3 4 5 6 7 8 | Descriptions: 1. Loan originated 2. Application approved but not accepted 3. Application denied by financial institution 4. Application was withdrawn by applicant 5. File closed for incompleteness 6. Loan purchased by your institution 7. Preapproval request denied by financial institution 8. Preapproval request approved but not accepted (optional reporting) |
| 11 | Metropolitan Statistical Area/Metropolitan Division | Alphanumeric | | Metropolitan Statistical Area or Metropolitan Division (if appropriate) code or NA. Example: 40900 |
| 12 | State Code | Alphanumeric | | FIPS code or NA. Example: 06 |
| 13 | County Code | Alphanumeric | | FIPS code or NA. Example: 113 |
| 14 | Census Tract | Alphanumeric | | Include decimal point or NA. Example: 0109.02 |
| 15 | Applicant Ethnicity | Numeric | 1 2 3 4 | Descriptions: 1. Hispanic or Latino 2. Not Hispanic or Latino 3. Information not provided by the applicant in mail Internet or telephone application (see App. A I.D.2.) 4. Not applicable |
| 16 | Co-applicant Ethnicity | Numeric | 1 2 3 4 5 | Descriptions: 1. Hispanic or Latino 2. Not Hispanic or Latino 3. Information not provided by the applicant in mail Internet or telephone application (see App. A I.D.2.) 4. Not applicable 5. No co-applicant |
| 17 | Applicant Race: 1 | Numeric | 1 2 3 4 5 6 7 | Descriptions: 1. American Indian or Alaska Native 2. Asian 3. Black or African American 4. Native Hawaiian or Other Pacific Islander 5. White 6. Information not provided by the applicant in mail Internet or telephone application (see App. A I.D.2.) 7. Not applicable |
| 18 | Applicant Race: 2 | Numeric | 1 2 3 4 5 | Descriptions: 1. American Indian or Alaska Native 2. Asian 3. Black or African American 4. Native Hawaiian or Other Pacific Islander 5. White. If this data field does not contain an entry leave it blank |
| 19 | Applicant Race: 3 | Numeric | 1 2 3 4 5 | Descriptions: 1. American Indian or Alaska Native 2. Asian 3. Black or African American 4. Native Hawaiian or Other Pacific Islander 5. White. If this data field does not contain an entry leave it blank |
| 20 | Applicant Race: 4 | Numeric | 1 2 3 4 5 | Descriptions: 1. American Indian or Alaska Native 2. Asian 3. Black or African American 4. Native Hawaiian or Other Pacific Islander 5. White. If this data field does not contain an entry leave it blank |
| 21 | Applicant Race: 5 | Numeric | 1 2 3 4 5 | Descriptions: 1. American Indian or Alaska Native 2. Asian 3. Black or African American 4. Native Hawaiian or Other |

| | | | | |
|---|---|---|---|---|
| | | | | Pacific Islander 5. White. If this data field does not contain an entry leave it blank |
| 22 | Co-applicant Race: 1 | Numeric | 1 2 3 4 5 6 7 8 | Descriptions: 1. American Indian or Alaska Native 2. Asian 3. Black or African American 4. Native Hawaiian or Other Pacific Islander 5. White 6. Information not provided by the applicant in mail Internet or telephone application (see App. A I.D.2.) 7. Not applicable 8. No co-applicant |
| 23 | Co-applicant Race: 2 | Numeric | 1 2 3 4 5 | Descriptions: 1. American Indian or Alaska Native 2. Asian 3. Black or African American 4. Native Hawaiian or Other Pacific Islander 5. White. If this data field does not contain an entry leave it blank |
| 24 | Co-applicant Race: 3 | Numeric | 1 2 3 4 5 | Descriptions: 1. American Indian or Alaska Native 2. Asian 3. Black or African American 4. Native Hawaiian or Other Pacific Islander 5. White. If this data field does not contain an entry leave it blank |
| 25 | Co-applicant Race: 4 | Numeric | 1 2 3 4 5 | Descriptions: 1. American Indian or Alaska Native 2. Asian 3. Black or African American 4. Native Hawaiian or Other Pacific Islander 5. White. If this data field does not contain an entry leave it blank |
| 26 | Co-applicant Race: 5 | Numeric | 1 2 3 4 5 | Descriptions: 1. American Indian or Alaska Native 2. Asian 3. Black or African American 4. Native Hawaiian or Other Pacific Islander 5. White. If this data field does not contain an entry leave it blank |
| 27 | Applicant Sex | Numeric | 1 2 3 4 | Descriptions: 1. Male 2. Female 3. Information not provided by the applicant in mail Internet or telephone application (see App. A I.D.2.) 4. Not applicable |
| 28 | Co-applicant Sex | Numeric | 1 2 3 4 5 | Descriptions: 1. Male 2. Female 3. Information not provided by applicant in mail Internet or telephone application (see App. A I.D.2.) 4. Not applicable 5. No co-applicant |
| 29 | Applicant Income | Alphanumeric | | A report in thousands round to the nearest thousand and without commas or NA. Example: 36 |
| 30 | Type of Purchaser | Numeric | 0 1 2 3 4 5 6 7 8 9 | Descriptions: 0. The loan was not originated or was not sold in the calendar year 1. Fannie Mae 2. Ginnie Mae 3. Freddie Mac 4. Farmer Mac 5. Private securitization 6. A commercial bank savings bank or savings association 7. Life insurance company credit union mortgage bank or finance company 8. Affiliate institution 9. Other type of purchaser |
| 31 | Denial Reason: 1 | Numeric | 1 2 3 4 5 6 7 8 9 | Descriptions: 1. Debt-to-income ratio 2. Employment history 3. Credit history 4. Collateral 5. Insufficient cash (down payment closing costs) 6. Unverifiable information 7. Credit application incomplete 8. Mortgage insurance denied 9. Other. If this data field does not contain an entry leave it blank |

| | | | | |
|---|---|---|---|---|
| 32 | Denial Reason: 2 | Numeric | 1 2 3 4 5 6 7 8 9 | Descriptions: 1. Debt-to-income ratio 2. Employment history 3. Credit history 4. Collateral 5. Insufficient cash (down payment closing costs) 6. Unverifiable information 7. Credit application incomplete 8. Mortgage insurance denied 9. Other. If this data field does not contain an entry leave it blank |
| 33 | Denial Reason: 3 | Numeric | 1 2 3 4 5 6 7 8 9 | Descriptions: 1. Debt-to-income ratio 2. Employment history 3. Credit history 4. Collateral 5. Insufficient cash (down payment closing costs) 6. Unverifiable information 7. Credit application incomplete 8. Mortgage insurance denied 9. Other. If this data field does not contain an entry leave it blank |
| 34 | Rate Spread | Alphanumeric | | Enter the rate spread to two decimal places. Include the decimal point and any leading or trailing zeros or NA. Example: 03.29 |
| 35 | HOEPA Status | Numeric | 1 2 | Descriptions: 1. HOEPA loan 2. Not a HOEPA loan |
| 36 | Lien Status | Numeric | 1 2 3 4 | Descriptions: 1. Secured by a first lien 2. Secured by a subordinate lien 3. Not secured by a lien 4. Not applicable (purchased loans) |

# 3. Exploratory Data Analysis

The section involved exploring the various relationship between columns or independent variable in the context of the geographical state of loan origination, the status of the loan and another client related field such as income, loan amount, gender etc.

## 3.1. Data wrangling:

The data needs to be further modified to get some variables in format to provide it to data visualization library.

The information extracted in this step

- The data for the salary is extracted using ranges for building a bar graph
- The data for the loan amount is extracted using ranges for building a bar graph
- The common methods are created to create the similar graphs based on columns
- The common methods are created to use Plotly visualization library

The code related to this part is in CapstoneII_DataStoryFile.ipynb file.

## 3.2. Exploring data relationship

The various columns are visualized to get the data representation of their behaviour with all US states and loan status.

### 3.2..1. Applicant Ethnicity

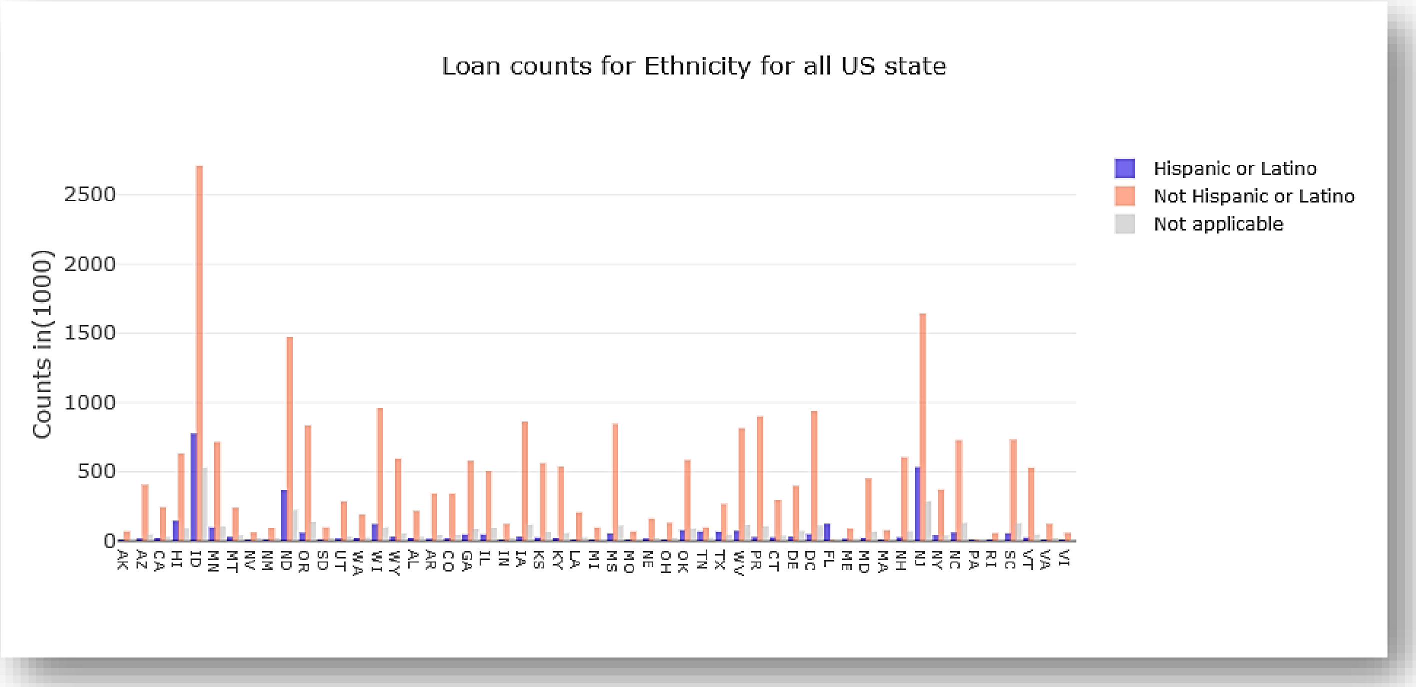


*Figure 1: Count of Ethnicity for all US state*

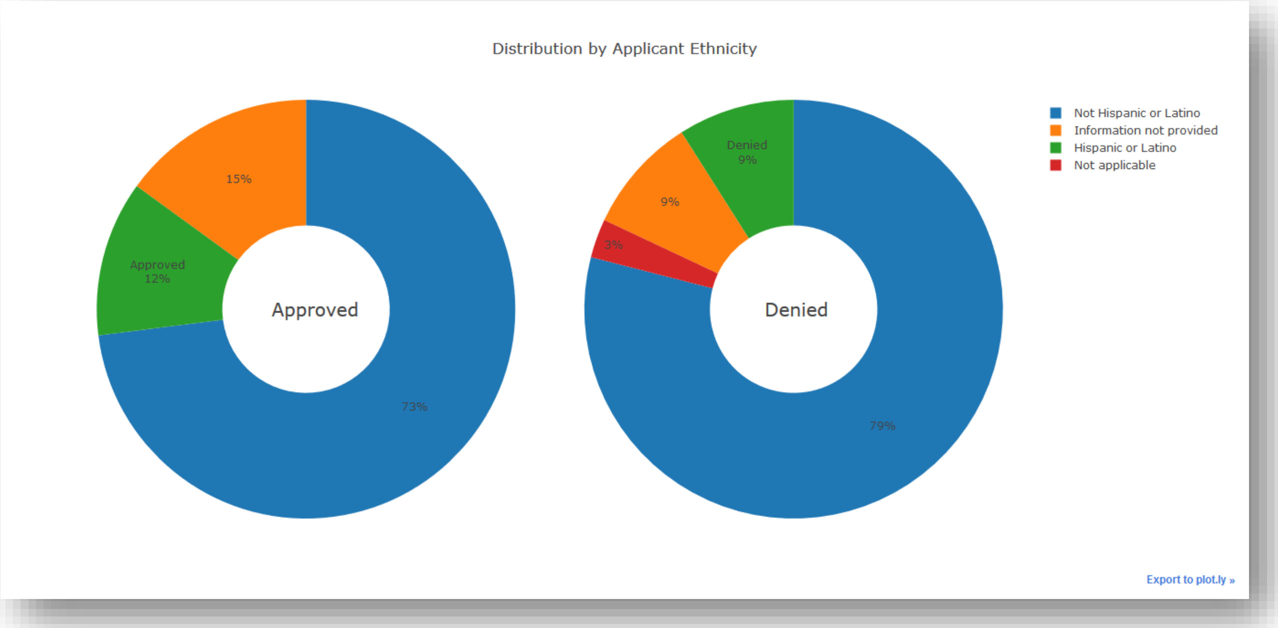


*Figure 2: Loan status for Ethnicity*

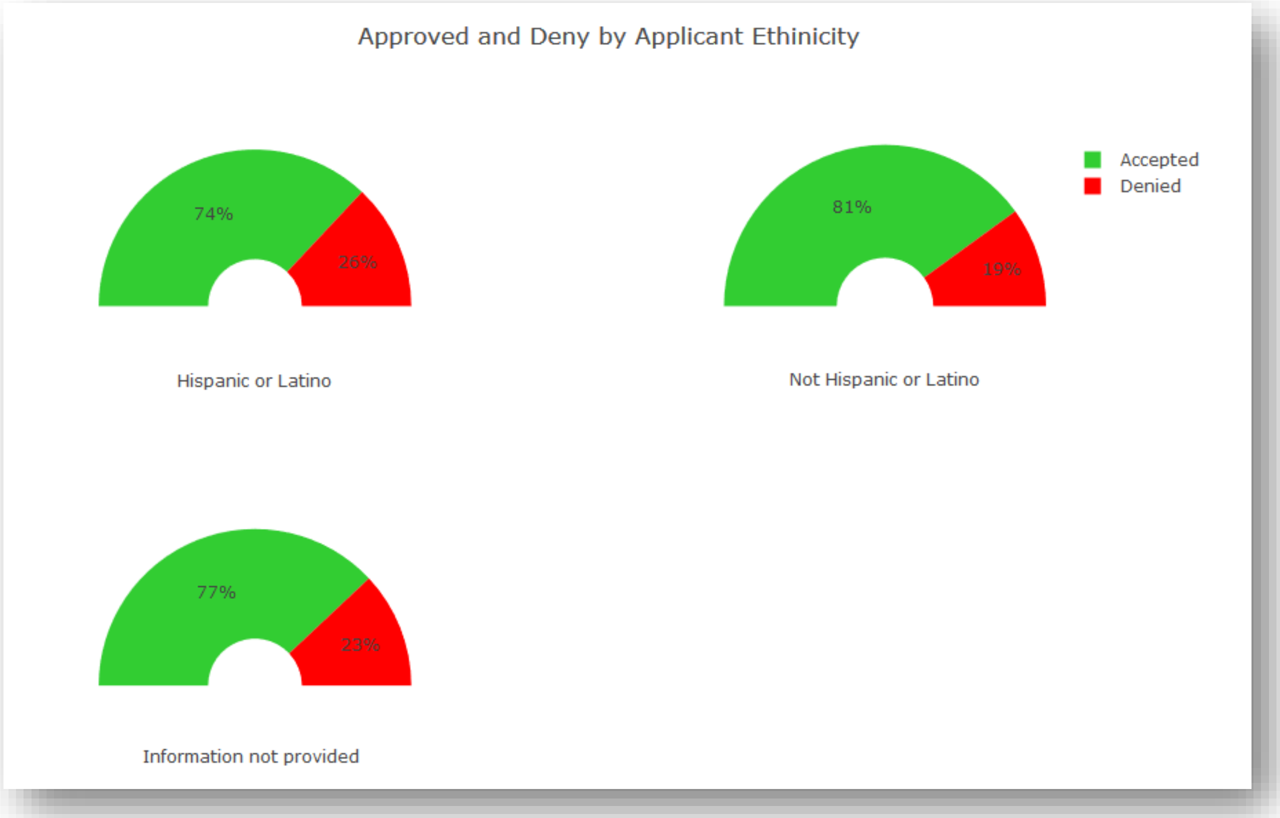


*Figure 3: Ethnicity type by loan status*

The above diagram on applicant ethnicity shows the following information

1. Most state shows not Hispanic applicant more than Hispanic, but this can be due to fact that no Hispanic covers lot of another ethnicity.
2. The 3 % of the application without ethnicity get rejected. But the application without ethnicity do not have a significant effect on approval
3. The ration of approval for Hispanic is 7% less than no Hispanic people. The ratio of approval for Hispanic and not providing information is almost similar.

*** The other graphs will be added in final report*****

#### 3.2..2. Applicant Race

#### 3.2..3. Applicant Gender

#### 3.2..4. Property Type

#### 3.2..5. Loan Purpose

#### 3.2..6. Owner Occupancy

#### 3.2..7. Loan Amount

#### 3.2..8. Applicant Income

#### 3.2..9. Loan distribution in US States

#### 3.2..10.

# 4. Inferential Statistics

The inferential statistics section involved exploring columns or variables from the data to perform statistical analysis. This section applies some inferential statistical concept to the data.

## 4.1. Exploring Application gender

The column or variable Applicant gender is a very interesting column and can be explored with inferential statistics tools for all loan application and can be further explored.

| | ApplicantGender | Result |
|---|---|---|
| **count** | 28794250.00 | 28794250.00 |
| **mean** | 1.31 | 0.80 |
| **std** | 0.46 | 0.40 |
| **min** | 1.00 | 0.00 |
| **25%** | 1.00 | 1.00 |
| **50%** | 1.00 | 1.00 |
| **75%** | 2.00 | 1.00 |
| **max** | 2.00 | 1.00 |

### 4.1..1. Exploring loans provided to Male

The column or variable Applicant gender is Male (1).

| | |
|---|---|
| **count** | 19915170.00 |
| **mean** | 0.81 |
| **std** | 0.39 |
| **min** | 0.00 |
| **25%** | 1.00 |
| **50%** | 1.00 |
| **75%** | 1.00 |
| **max** | 1.00 |

### 4.1..2. Exploring loans provided to Female.

The column or variable Applicant gender is Female (0).

| | |
|---|---|
| **count** | 8879076.00 |
| **mean** | 0.77 |
| **std** | 0.42 |
| **min** | 0.00 |
| **25%** | 1.00 |
| **50%** | 1.00 |
| **75%** | 1.00 |
| **max** | 1.00 |

### 4.1..3. Exploring loans approved based on gender.

The **mean** for an **approved** loan for a **male** candidate is **0.8074539544570386**
The **standard deviation** for an **approved** loan for a **male** candidate is **0.3942994720520549**
The **variance** for an **approved** loan for a **male** candidate is **0.1554720736605292**
The **mean** for an **approved** loan for a **female** candidate is **0.7707920283597077**
The **standard deviation** for an **approved** loan for a **female** candidate is **0.42032332468866307**
The **variance** for an **approved** loan for a **female** candidate is **0.17667169727733129**

The **difference** of **mean** for **male** & **female** candidate is **0.03666192609733088**
The **total female** with result **approved** in population **6843921.** The **total female** in population **8879076.** The **female** result / total female **0.7707920283597077**
The **total male** with the result in population **16080586.** The total **male** in population **19915174.** The male result / total male **0.8074539544570386**
The **female** variance **0.000000019898. The** male people variance **0.000000007807.** The total population variance **0.000000027704**

The above data shows that there is some difference between the mean of the approved loan based on gender.

## 4.2. Hypothesis Testing

The column or variable Applicant gender is a very interesting column and can be explored with inferential statistic tools.

There is a difference between the **mean of the approved loan** between male and female. We can further analyze this hypothesis. The hypothesis is as follows

**Null Hypothesis:** There is no difference in result in male or female approved loan. Which means for the result for men - means for the result for female equals Zero.

**Alternate Hypothesis:** There is a significant difference in result in male or female approved loan. Which means for the result for men - means for the result for female not equal Zero.

*Calculate Z score and p score for the null hypothesis*
Calculate Z stat using ztest method in a weightstat module with significance level 0.005
The calculated values are as follow for two-sided & larger
t-statistic:  225.7191608052462
p-value:  0.0

*Calculate T score and p value to test the same hypothesis*
Calculate T score using the ttest_ind method in stats module with significance level 0.005
The calculated values are as follow
t-statistic:  220.26328358859774
p-value:  0.0

The above p value is less than our significance value and hence there is enough evidence to **reject** Null hypotheses.

The code related to this part is in the CapstoneII_Inferential_Statistic.ipynb file.