|  |
| --- |
| Photo displaying partial image of two pie charts on a canvas-textured page |
| CUSTOMER SEGMENTATION IN MORTGAGE INDUSTRY  Capstone Project Final Report |
| |  |  |  | | --- | --- | --- | | Pundir, Saurabh | 6/30/18 | SB-DataScience Course | |

Table of Contents

[**1.** **Introduction** 4](#_Toc519709545)

[1.1. About: 4](#_Toc519709546)

[1.1..1. Formal Definition - Customer Segmentation: 4](#_Toc519709547)

[1.2. Problem: 5](#_Toc519709548)

[1.3. Clients and Audiences 5](#_Toc519709549)

[1.4. Scope: 5](#_Toc519709550)

[1.5. Question Client(s) really care about 5](#_Toc519709551)

[**2.** **Data Wrangling** 5](#_Toc519709552)

[2.1. Getting File 6](#_Toc519709553)

[2.2. File Content: 6](#_Toc519709554)

[2.3. Choosing Yearly Files: 6](#_Toc519709555)

[2.4. 3Step Approach 6](#_Toc519709556)

[2.5. Extracting Data: 6](#_Toc519709557)

[2.6. Cleaning Data: 6](#_Toc519709558)

[2.7. Wrangling for data story & inferential statistics: 6](#_Toc519709559)

[2.8. Wrangling for machine learning: 7](#_Toc519709560)

[2.9. Storing Data: 7](#_Toc519709561)

[2.10. Data Dictionary: 8](#_Toc519709562)

[**3.** **Exploratory Data Analysis** 12](#_Toc519709563)

[3.1. Data wrangling: 12](#_Toc519709564)

[3.2. Exploring data relationship 12](#_Toc519709565)

[3.2..1. Applicant Ethnicity 12](#_Toc519709566)

[3.2..2. Applicant Race 15](#_Toc519709567)

[3.2..3. Applicant Gender 17](#_Toc519709568)

[3.2..4. Property Type 19](#_Toc519709569)

[3.2..5. Loan Purpose 21](#_Toc519709570)

[3.2..6. Owner Occupancy 23](#_Toc519709571)

[3.2..7. Loan Amount 24](#_Toc519709572)

[3.2..8. Applicant Income 26](#_Toc519709573)

[3.2..9. Loan distribution in US States 27](#_Toc519709574)

[**4.** **Inferential Statistics** 28](#_Toc519709575)

[4.1. Exploring Application gender 28](#_Toc519709576)

[4.1..1. Exploring loans provided to Male 28](#_Toc519709577)

[4.1..2. Exploring loans provided to Female. 28](#_Toc519709578)

[4.1..3. Exploring loans approved based on gender. 29](#_Toc519709579)

[4.2. Hypothesis Testing 29](#_Toc519709580)

[**5.** **Base Analysis -K Means** 30](#_Toc519709581)

[5.1. Exploring Single State -California 30](#_Toc519709582)

[5.2. Extracting county data 30](#_Toc519709583)

[5.3. Salary and Loan Amount 30](#_Toc519709584)

[5.4. Columns for cluster 30](#_Toc519709585)

[5.5. Choosing number of cluster 30](#_Toc519709586)

[5.5..1. Elbow Method 30](#_Toc519709587)

[5.5..2. Silhouette Method 31](#_Toc519709588)

[5.6. Visualizing clusters for clusters 32](#_Toc519709589)

[**6.** **Cluster Analysis -3 Clusters** 34](#_Toc519709590)

[6.1. Exploring Cluster 1 34](#_Toc519709591)

[6.1..1. Applicant Ethnicity 34](#_Toc519709592)

[6.1..2. Applicant Race 35](#_Toc519709593)

[6.1..3. Applicant Gender 36](#_Toc519709594)

[6.1..4. Property Type 37](#_Toc519709595)

[6.1..5. Loan Purpose 38](#_Toc519709596)

[6.1..6. Owner Occupancy 39](#_Toc519709597)

[6.1..7. Loan Amount 40](#_Toc519709598)

[6.1..8. Applicant Income 41](#_Toc519709599)

[6.2. Exploring Cluster 2 42](#_Toc519709600)

[6.2..1. Applicant Ethnicity 43](#_Toc519709601)

[6.2..2. Applicant Race 44](#_Toc519709602)

[6.2..3. Applicant Gender 45](#_Toc519709603)

[6.2..4. Property Type 46](#_Toc519709604)

[6.2..5. Loan Purpose 47](#_Toc519709605)

[6.2..6. Owner Occupancy 48](#_Toc519709606)

[6.2..7. Loan Amount 49](#_Toc519709607)

[6.2..8. Applicant Income 50](#_Toc519709608)

[6.3. Exploring Cluster 3 50](#_Toc519709609)

[6.3..1. Applicant Ethnicity 51](#_Toc519709610)

[6.3..2. Applicant Race 52](#_Toc519709611)

[6.3..3. Applicant Gender 53](#_Toc519709612)

[6.3..4. Property Type 54](#_Toc519709613)

[6.3..5. Loan Purpose 55](#_Toc519709614)

[6.3..6. Owner Occupancy 56](#_Toc519709615)

[6.3..7. Loan Amount 57](#_Toc519709616)

[6.3..8. Applicant Income 58](#_Toc519709617)

[6.4. Comparing Clusters 59](#_Toc519709618)

[6.4..1. Applicant Ethnicity 60](#_Toc519709619)

[6.4..2. Applicant Race 62](#_Toc519709620)

[6.4..3. Applicant Gender 63](#_Toc519709621)

[6.4..4. Property Type 64](#_Toc519709622)

[6.4..5. Loan Purpose 65](#_Toc519709623)

[6.4..6. Owner Occupancy 66](#_Toc519709624)

[6.4..7. Loan Amount 67](#_Toc519709625)

[6.4..8. Applicant Income 68](#_Toc519709626)

# **Introduction**

## 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.



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

## 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.

## 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 the 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 the last three years?”***

## 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.

## Scope:

The customer segmentation is a 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.

## 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?”**

# **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.

## 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.

## 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.

## 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 file for further processing

## 3Step 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.

## 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 are removed. This is decided because these are important features 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](https://github.com/saurabhspundir/MortgageCustomerSegmentation/blob/master/CapstoneII_DataWrangling_I_ReadFiles.ipynb) file.

## 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](https://github.com/saurabhspundir/MortgageCustomerSegmentation/blob/master/CapstoneII_DataWrangling_II_DataUpdate.ipynb) file.

## 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

## 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. State Code
2. Applicant Ethnicity
3. Applicant Race
4. Applicant Sex
5. Occupancy
6. Property Type
7. LoanPurpose

The salary and loan amount columns are converted into dummy based on amount range. The following ranges in (1000 USD) are used 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](https://github.com/saurabhspundir/MortgageCustomerSegmentation/blob/master/Capstone2_DataWrangling_III_ML.ipynb) file.

## Storing Data:

**Challenge:**

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

After cleaning the data and creating the 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.

## 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) |

# **Exploratory Data Analysis**

The section involves exploring the 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.

## Data wrangling:

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

The information extracted in this step

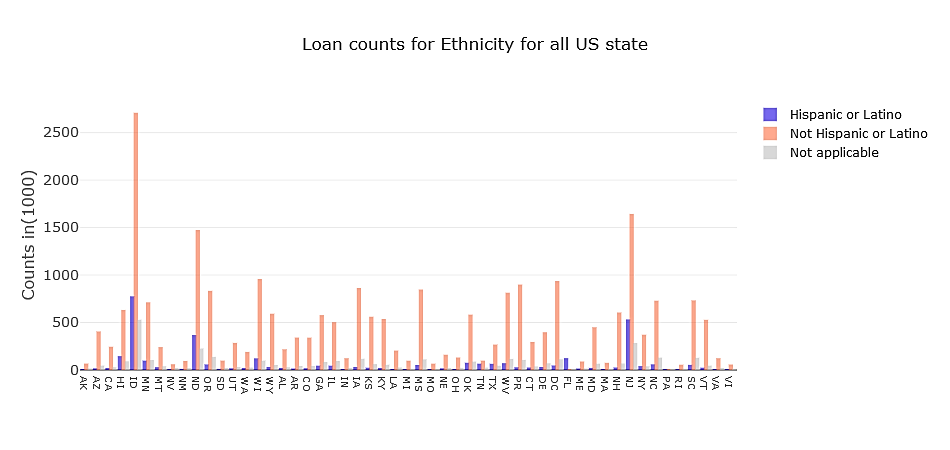
* + - The data for the salary is extracted using ranges for building a bar graph
    - The data for the loan amount is extracted using ranges for building a bar graph
    - The common methods are created to create the similar graphs based on columns
    - The common methods are created to use Plotly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

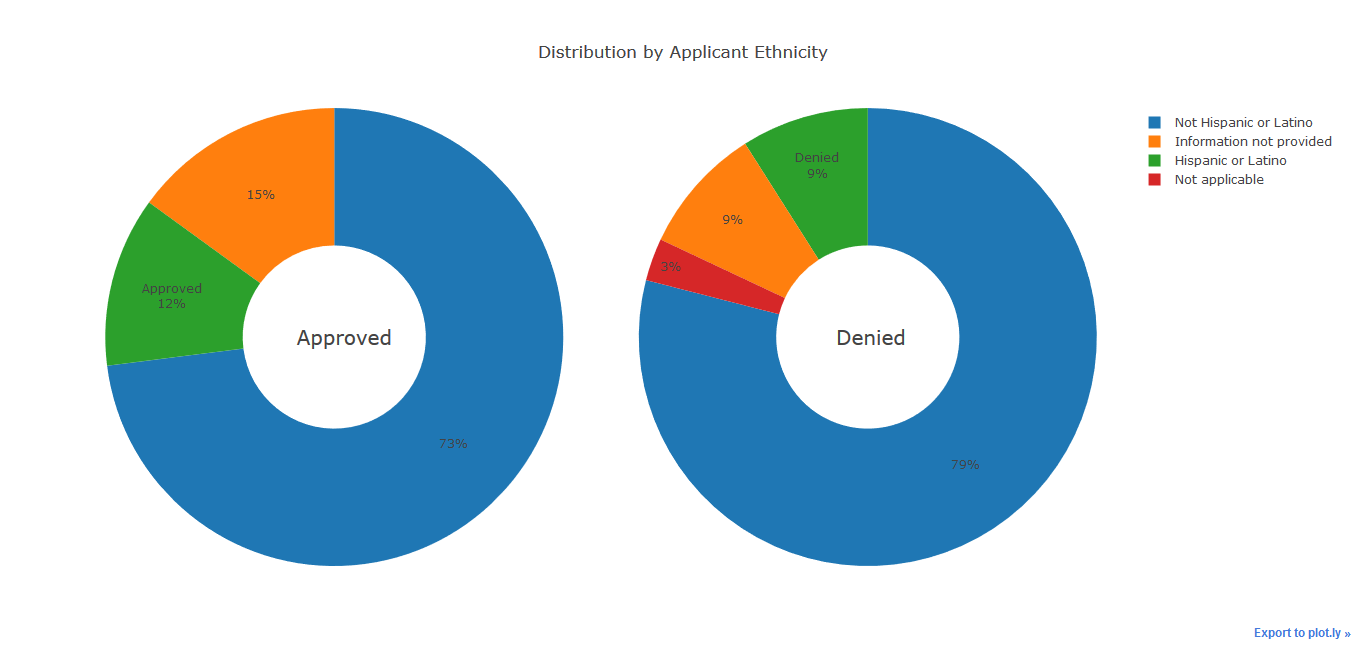
The code related to this part is in [CapstoneII\_DataStoryFile.ipynb](https://github.com/saurabhspundir/MortgageCustomerSegmentation/blob/master/CapstoneII_DataStoryFile.ipynb) file.

## Exploring data relationship

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

## Applicant Ethnicity

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

 *Figure 2: Loan status for Ethnicity*

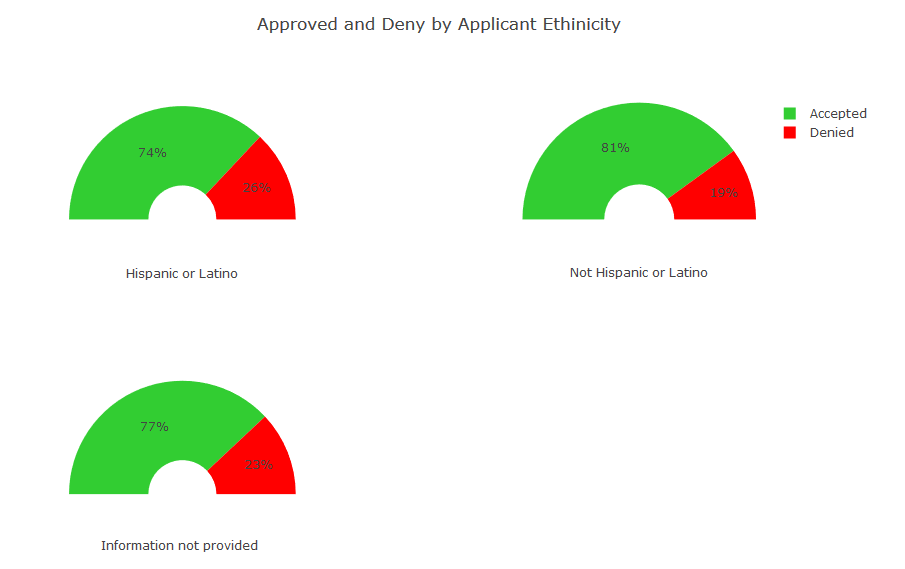


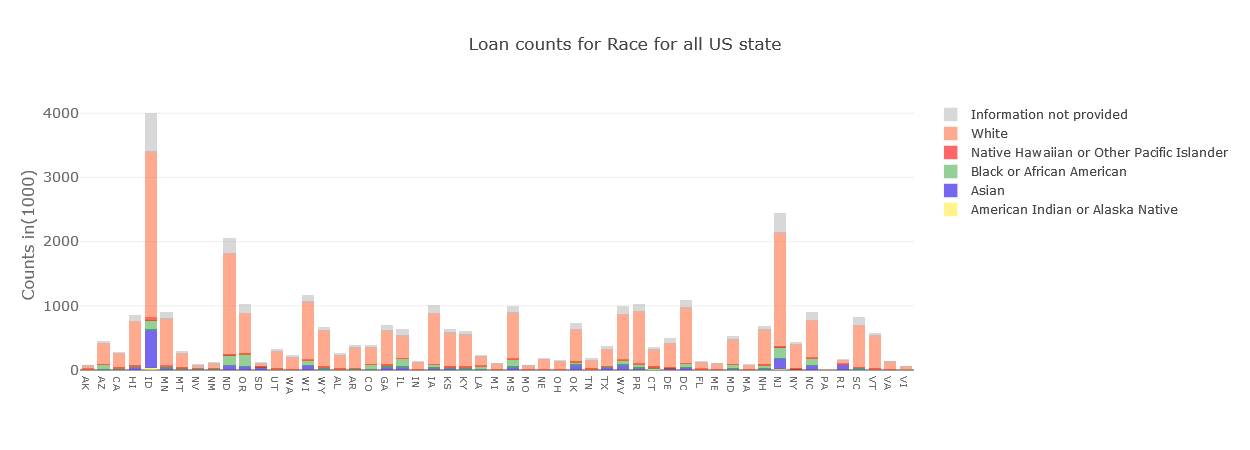
Figure 3: Ethnicity type by loan status

The above diagram on applicant ethnicity shows the following information:

1. Most states show non-Hispanic applicant more than Hispanic, but this can be due to fact that non-Hispanic covers a lot of other ethnicities.
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 ratio of approval for Hispanic is 7% less than not Hispanic people. The ratio of approval for Hispanic and not providing information is almost similar.

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

## Applicant Race

 *Figure 4: Count of Race for all US state*

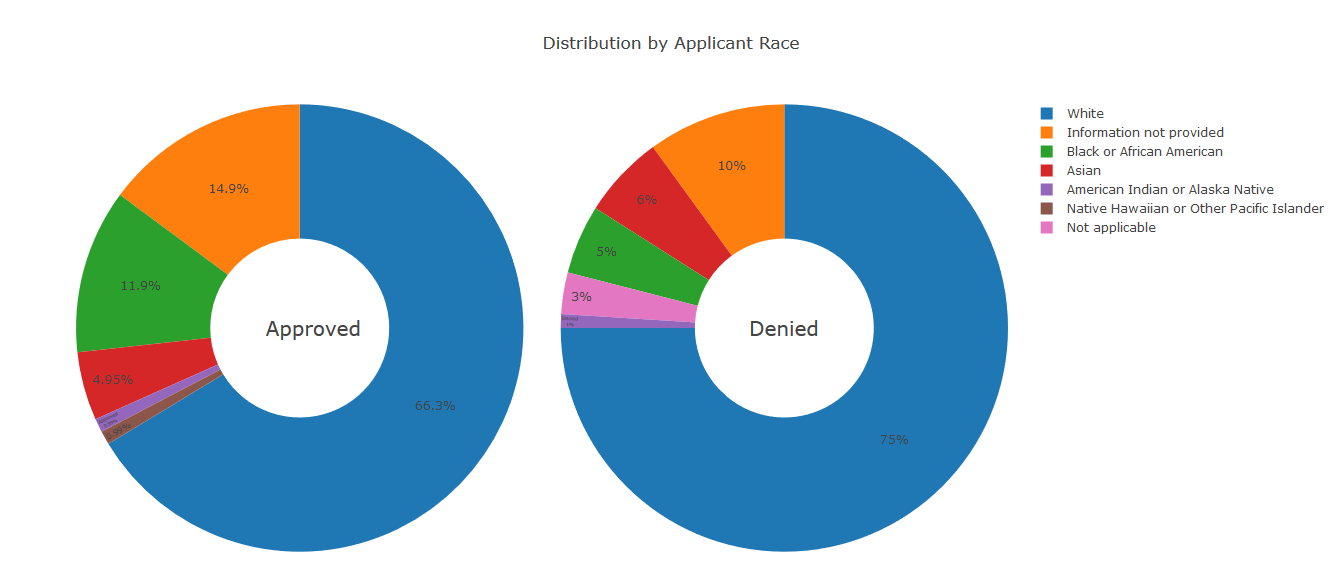


Figure 5: Loan status for Race



Figure 6: Race type by loan status

The above diagram on applicant race shows the following information:

1. In most states, the number of white applicants is higher as compared to other races. This might be also due to fact that there are more white people in the US.
2. A number of applications for loan denials are higher for Black or African American race followed by Asian race.
3. Number of approvals for loans are highest -82% and denials are at 18% as per the data study

## Applicant Gender

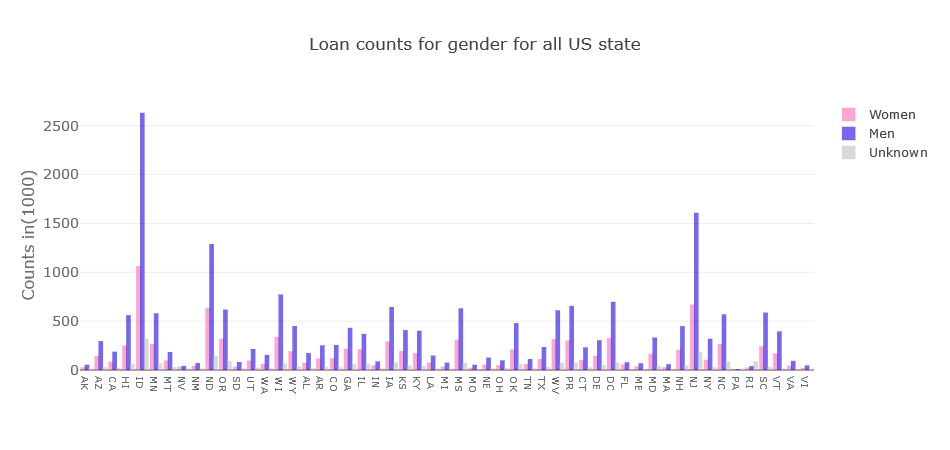


Figure 7: Count of Gender for all US state

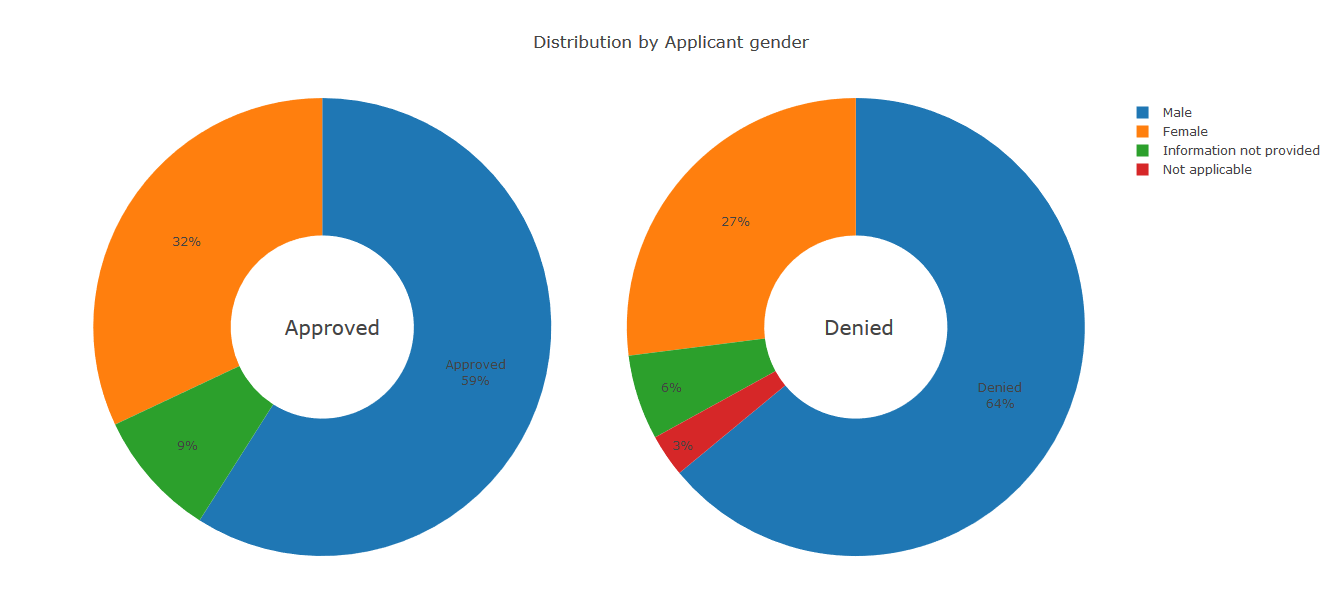


Figure 8: Loan status for Gender

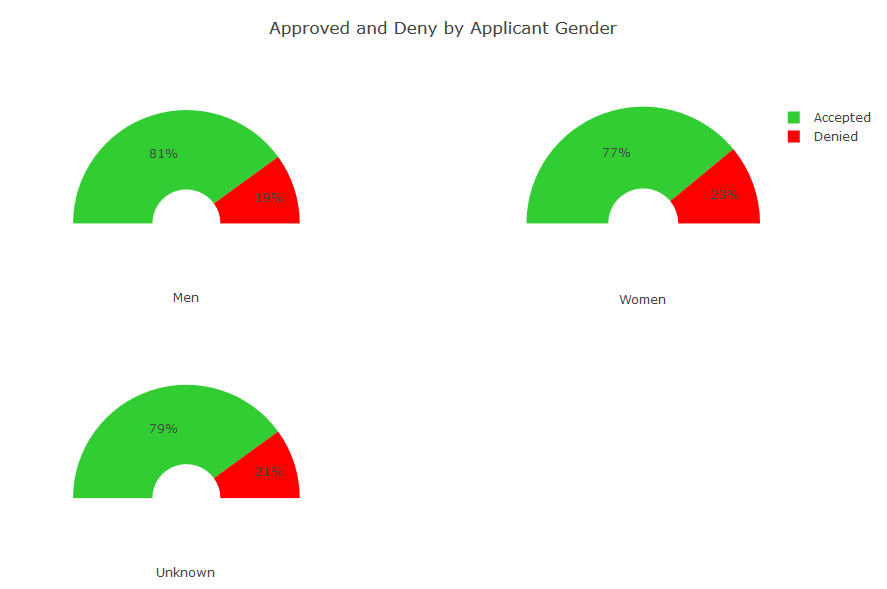
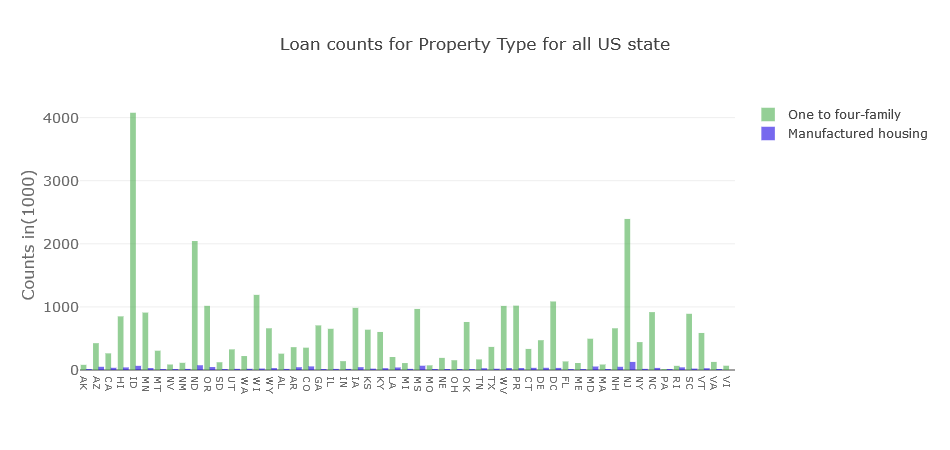


Figure 9: Gender type by loan status

The above diagram on applicant gender shows the following information:

1. The number of male applicants is higher as compared to female applicants for all of US state. This could be due to the reason that a lower number of female applicants have applied for loans.
2. A number of Approved male applicants are higher and the rate of denial for male applicants is also on the higher side.
3. 81% of male applications were accepted and 19% denied visa-via 77% of female applications were accepted and 23% denied

## Property Type

 *Figure 10: Count of Property Type for all US state*

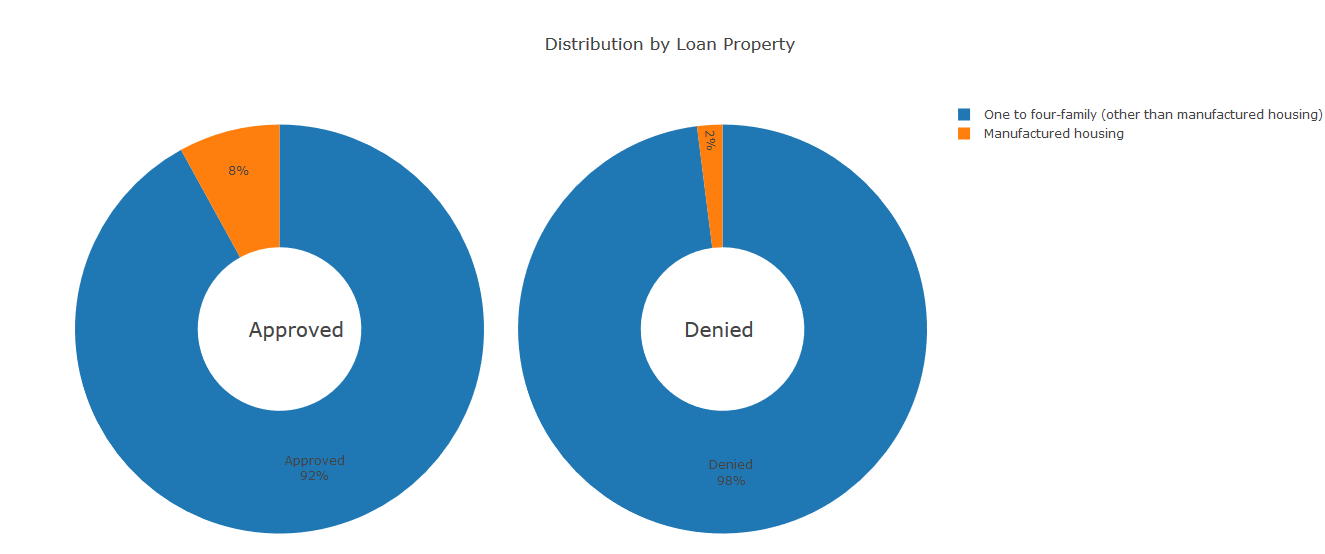


Figure 11: Loan Distribution by Property

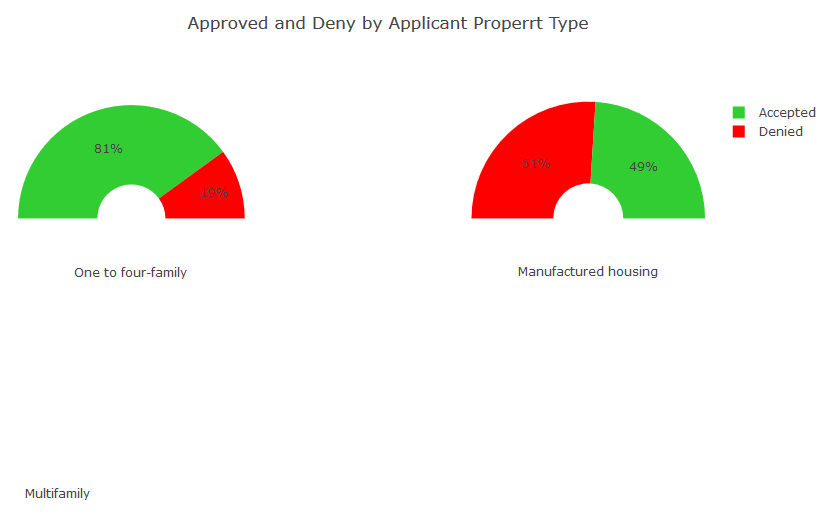
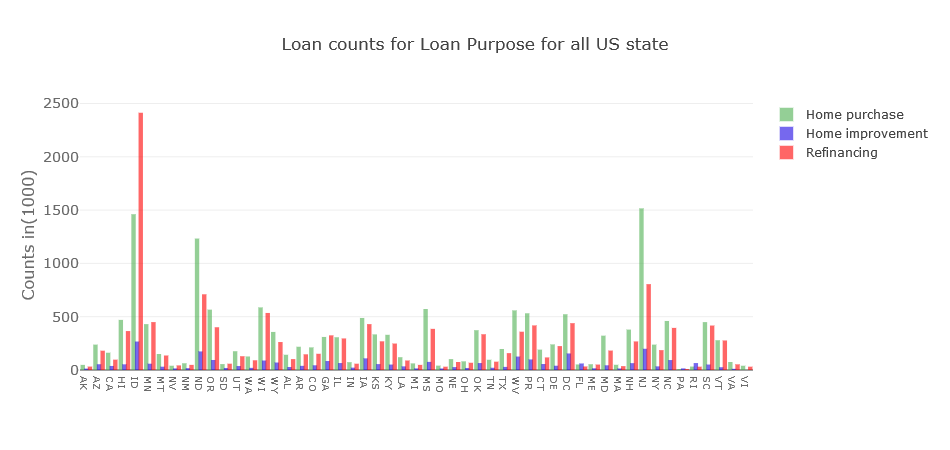
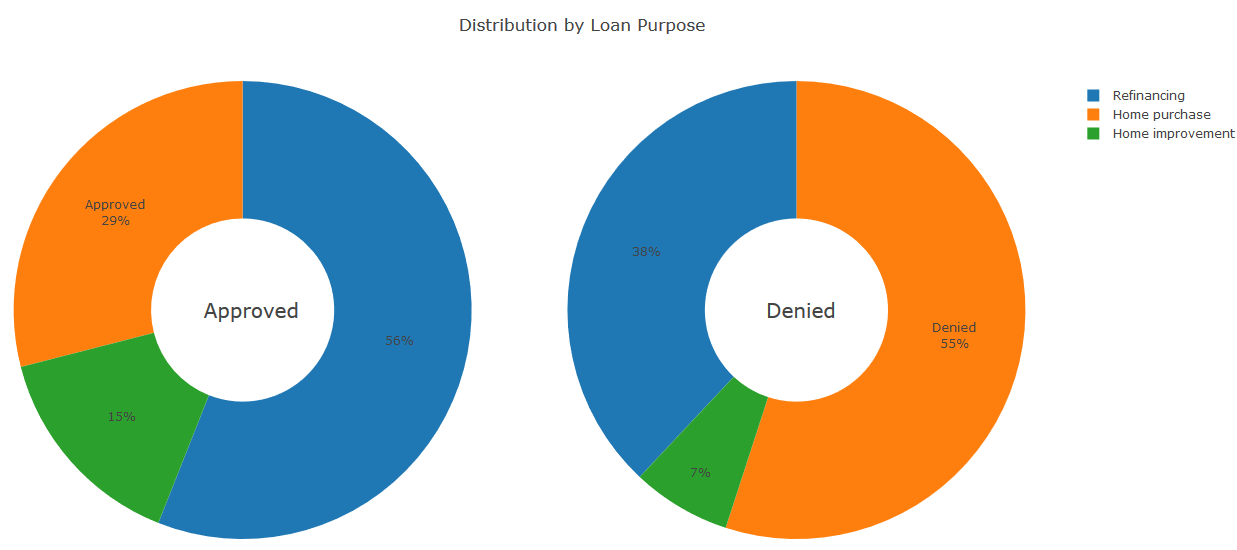


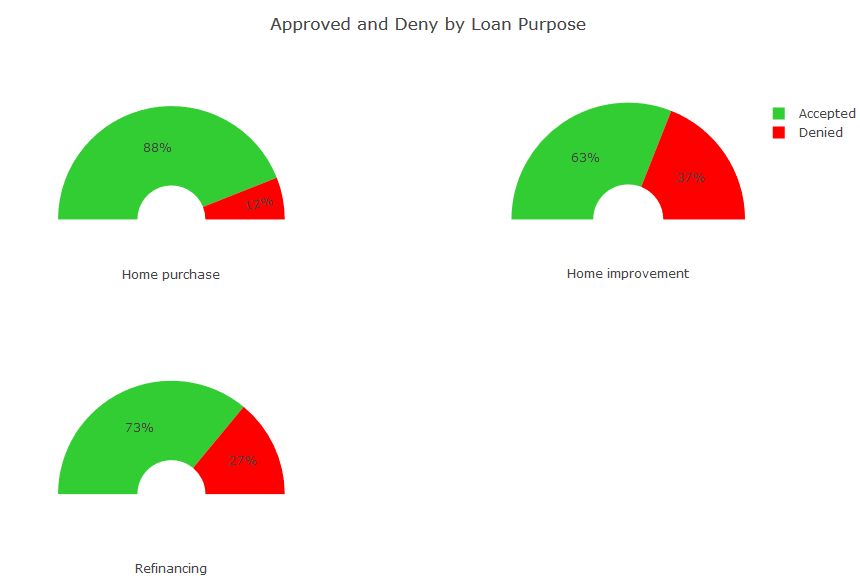
Figure 12: Loan approval and denial by Applicant Property Type

The above diagram on applicant property type shows the following information:

## Loan Purpose

 Figure 13: Loan counts for Loan Purpose for all US States

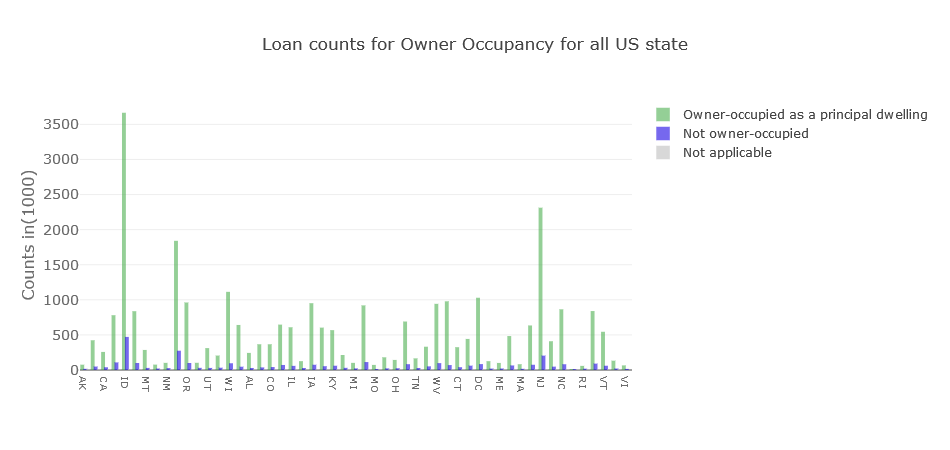
 Figure 13: Loan Distribution for Loan Purpose for all US States

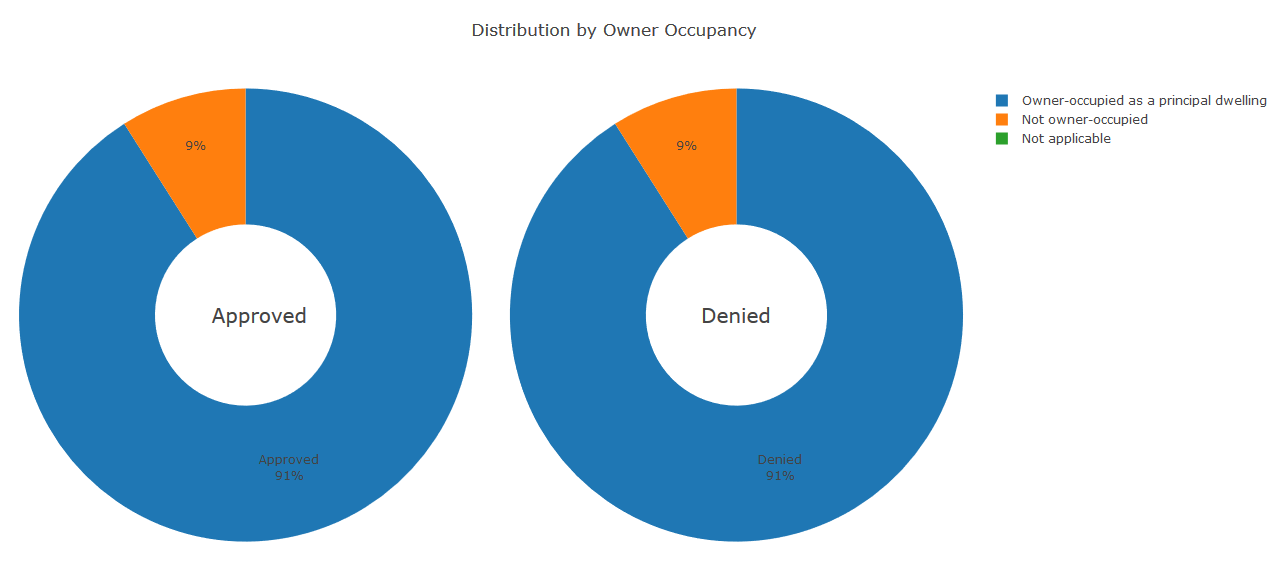
 *Figure 13: Approvals and denials by Loan Purpose*

The above diagram on Loan shows the following information:

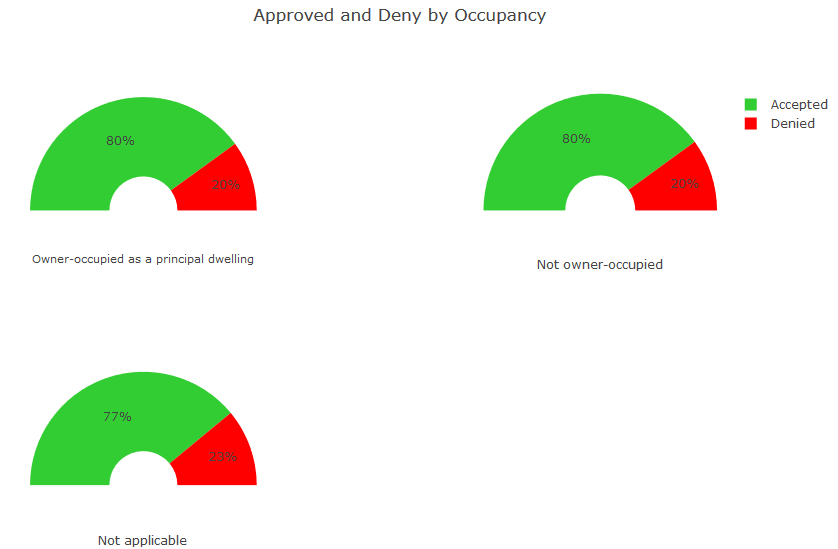
1. The number of male applicants is higher as compared to female applicants for all of US state. This could be due to the reason that a lower number of female applicants have applied for loans.

## Owner Occupancy

 *Figure 14: Loan counts for Owner Occupancy*

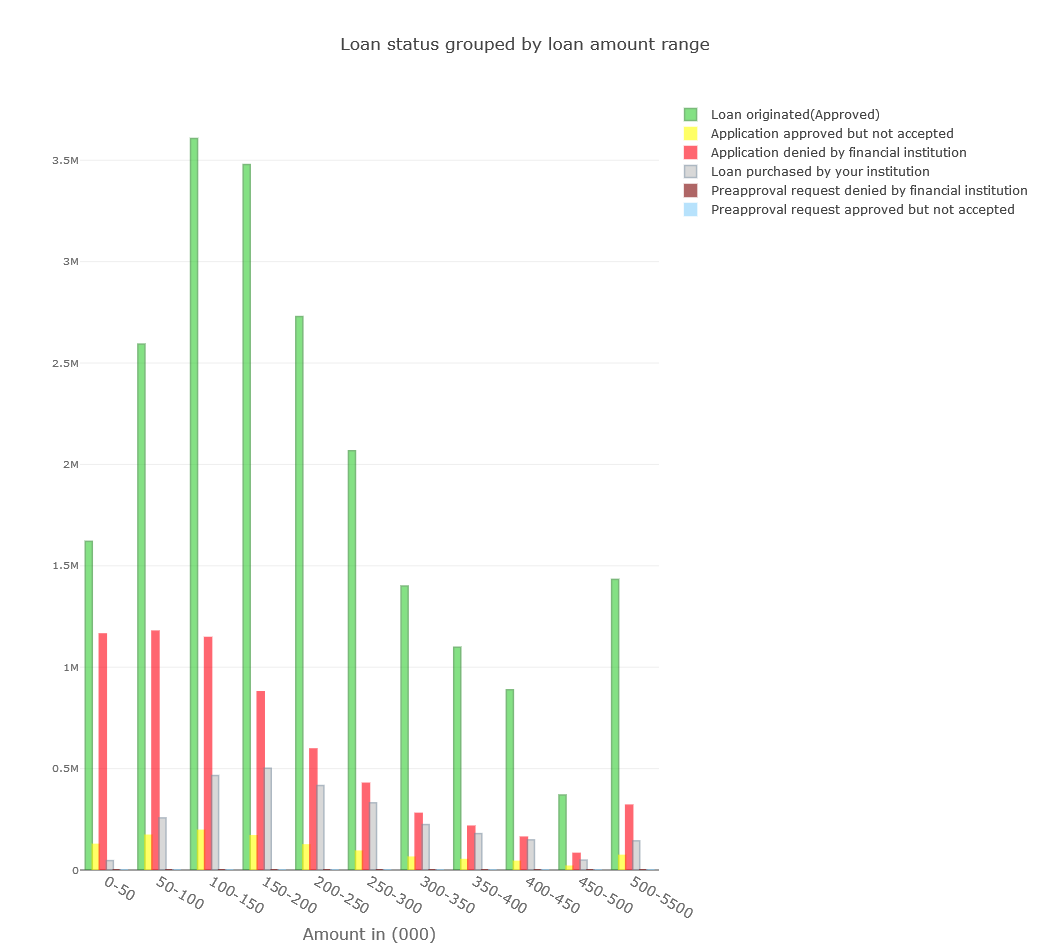


*Figure 15: Distribution by Owner Occupancy*



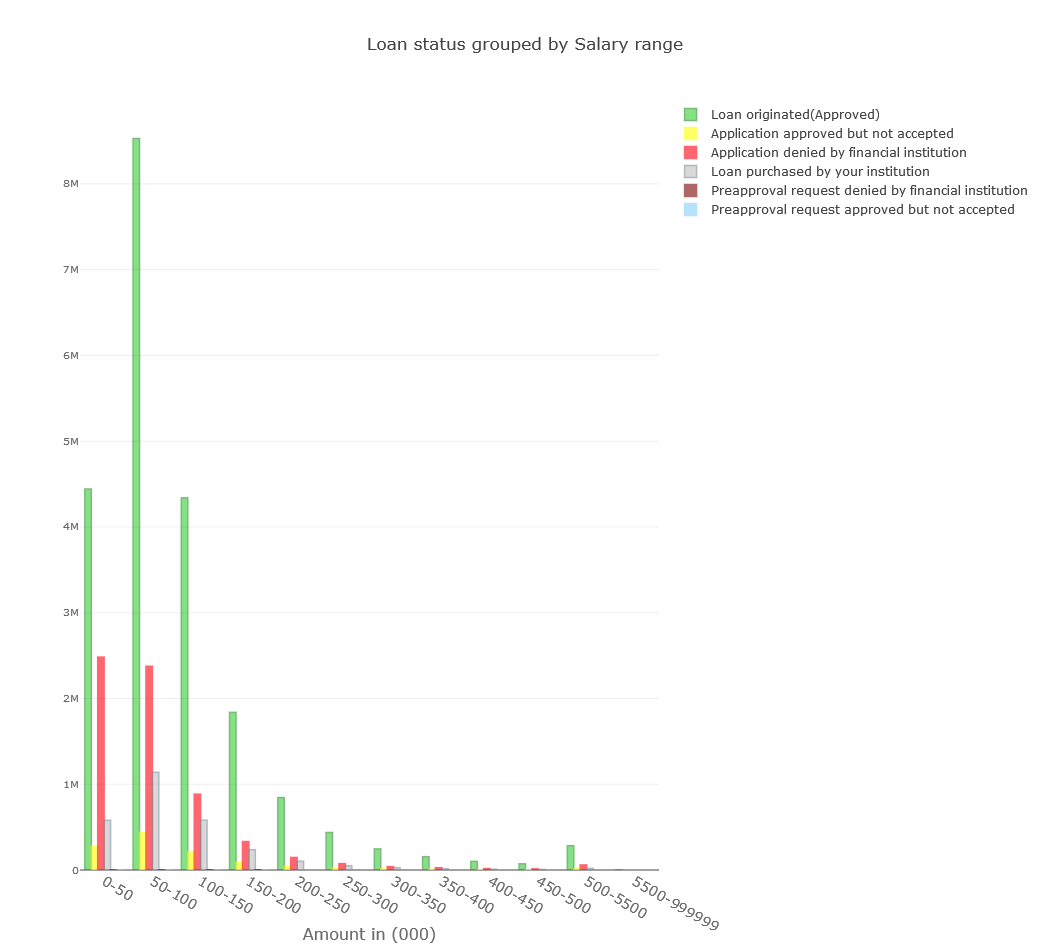
*Figure 16: Approved and Denials by Owner Occupancy*

## Loan Amount

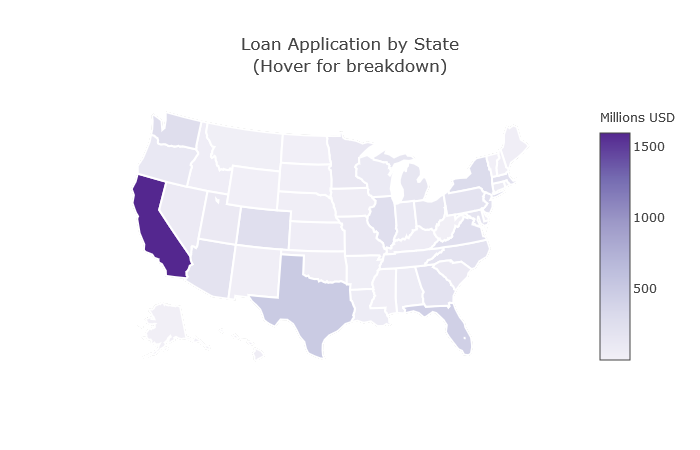


*Figure 17: Loan status grouped by Loan Amount*

## Applicant Income

 *Figure 18: Loan status grouped by Applicant Income*

## Loan distribution in US States

 *Figure 19: Loan distribution in the USA*

# **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.

## 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 |

## 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 |

## 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 |

## 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.

## 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 a 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](https://github.com/saurabhspundir/MortgageCustomerSegmentation/blob/master/CapstoneII_Inferential_Statistic.ipynb) file.

# **Base Analysis -K Means**

The baseline analysis is performed by creating clusters using K means machine learning model. The model obtained is considered as a baseline for further analysis.

## Exploring Single State -California

The clustering is done on the state because of the amount of data and in the real world, the identification of all the customer based on states can help in defining more precise analysis. The California state is considered due to a vast number of loans coming for that single state and also the state has a good spread of data across every feature.

To choose 100000 records custom method is created and with use of test and train from sklearn to pick random.

## Extracting county data

All the county data for the state is transformed using a dummy column. The data without county is excluded.

## Salary and Loan Amount

The salary and loan amount is used in categorical data as the customer with salary 200K and 210 K are not very different and can be considered under one bin. This also helps in finding cluster based on a range of salary. The similar approach is utilized for the loan amount.

## Columns for cluster

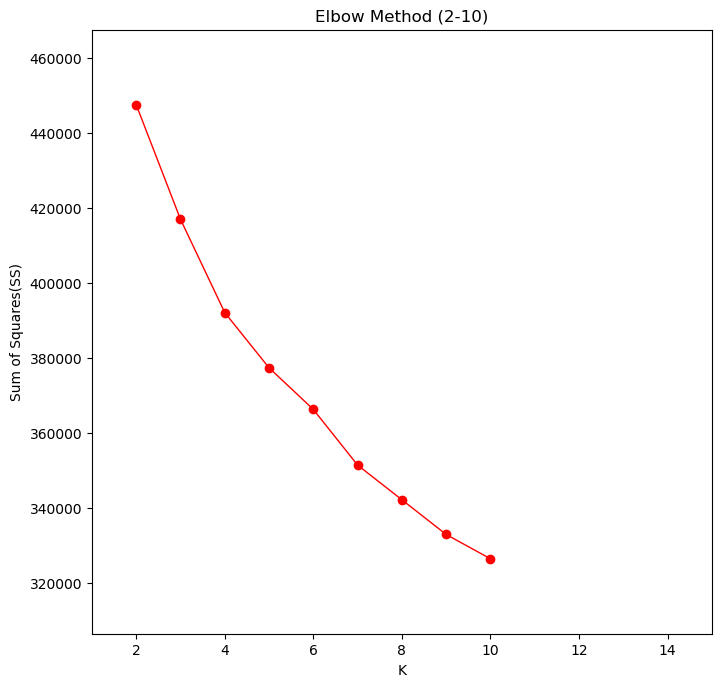
There are 55 columns consider in the cluster. These are columns converted into dummies for Applicant Ethnicity, Applicant Race, Applicant Sex, Occupancy , Property Type, Loan Purpose, County, Salary and Loan Amount.

## Choosing the number of clusters

For K means finding the best value number of clusters. There is a various method available based on the label or non-label data. The data available is not labeled and hence two methods are utilized to choose.

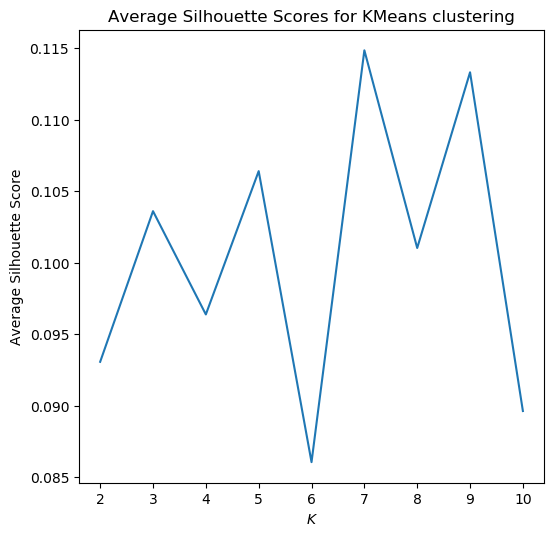
## Elbow Method

The elbow method uses value inertia\_ to determine the elbow visually. The idea is to visualize the graph for all possible value of K and observe the sudden change in the sum of the square of values.



## Silhouette Method

The silhouette method is based on the silhouette score for each cluster in K mean. The strong silhouette score helps in choosing the desired cluster value.

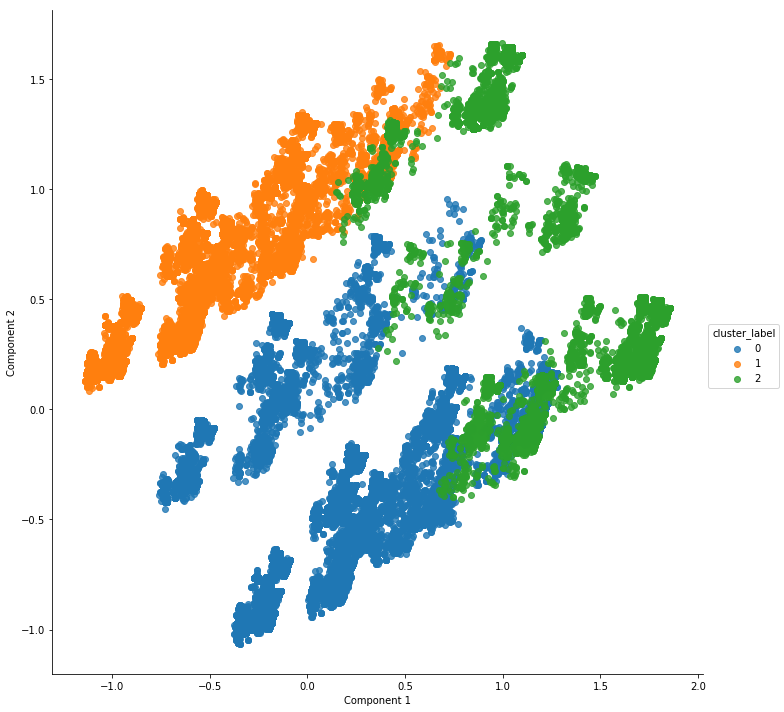
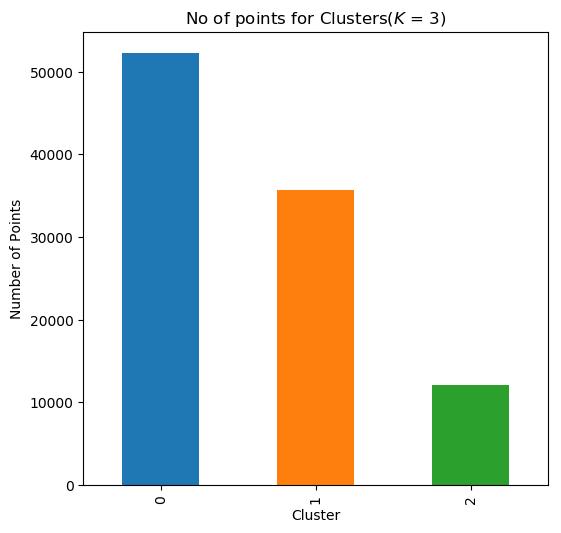


## Visualizing clusters for clusters

Apart from the above method visualizing all the cluster for a number of items in each cluster and the 2D display using PCA components.

All the cluster from 2 -11 are visualized and can be seen in workbook [Capstone2\_ML\_Clustering\_KMeans.ipynb](https://github.com/saurabhspundir/MortgageCustomerSegmentation/blob/master/Capstone2_ML_Clustering_KMeans.ipynb).

The below diagram display the values for the three clusters.



# **Cluster Analysis -3 Clusters**

The K means with 3 clusters is considered for visually exploring the k means for all the features for further analysis. All the features for each cluster are explored to understand more about the clusters.

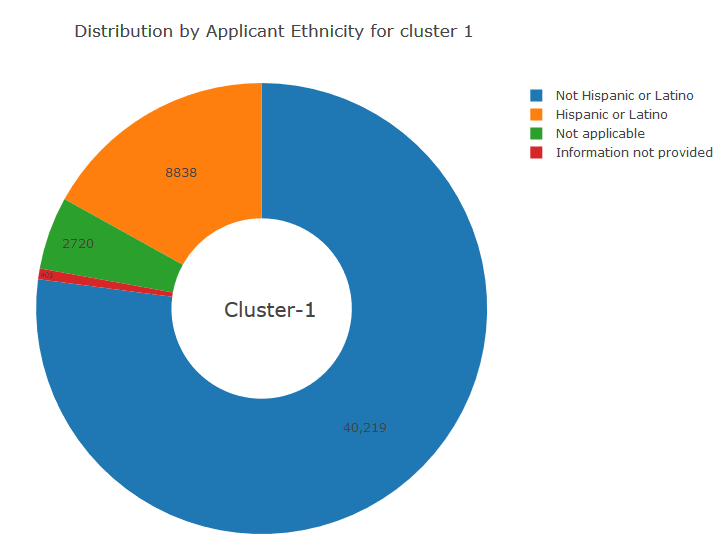
The total count in each cluster

|  |  |
| --- | --- |
| **Cluster Number** | **No of Datapoints** |
| Cluster 1 | 52180 |
| Cluster 2 | 35683 |
| Cluster 3 | 12099 |

## Exploring Cluster 1

The cluster 1 is explored for all the features.

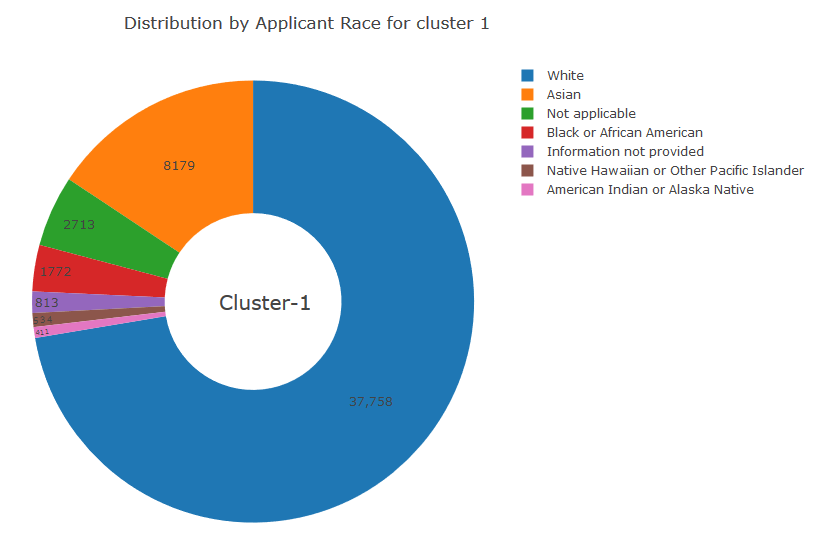
## Applicant Ethnicity

 *Figure 20: Distribution by Applicant Ethnicity*

The above diagram on Applicant Ethnicity shows the following information:

1. The not Hispanic ethnicity is dominating the cluster. The not Hispanic is overall all dominating ethnicity in the whole data set.

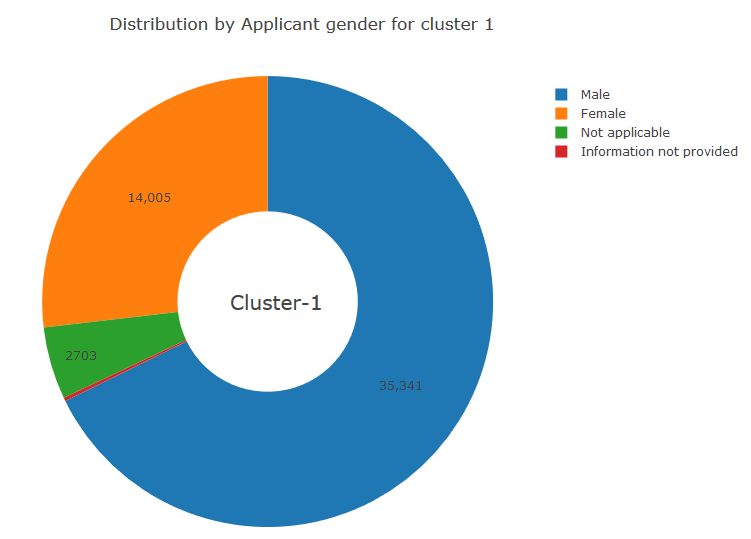
## Applicant Race

 *Figure 21: Distribution by Applicant Race*

The above diagram on Applicant Race shows the following information:

1. The white race is dominating the cluster. The white is overall all dominating race in the whole data set.

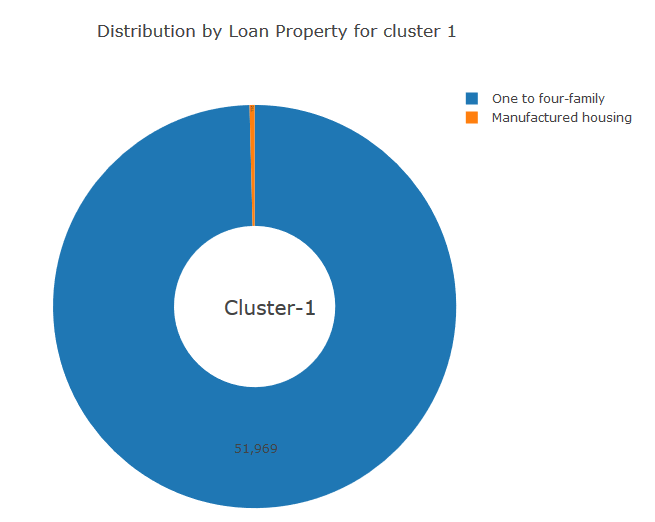
## Applicant Gender

 *Figure 22: Distribution by Applicant Gender*

The above diagram on Applicant Gender shows the following information:

1. The male gender is dominating the cluster. The male is overall all dominating gender in the whole data set.

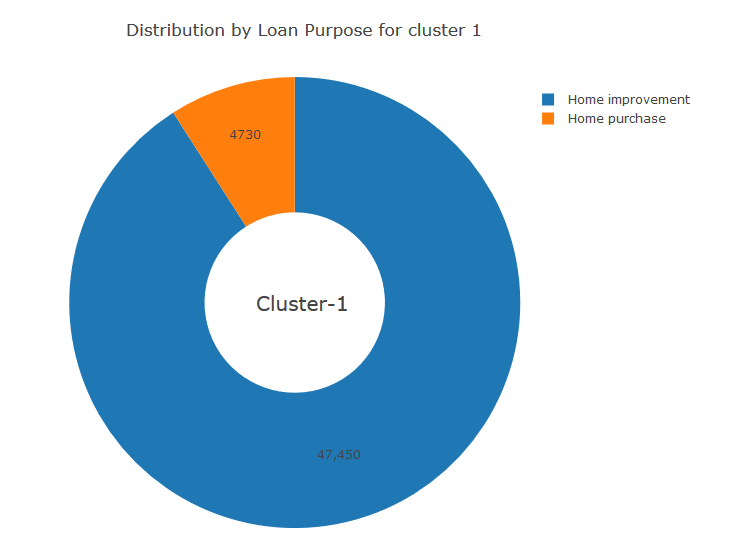
## Property Type

 *Figure 23: Distribution by Loan Property*

The above diagram on distribution by Loan Property shows the following information:

1. The one to four families is dominating the cluster. The one to four families is overall all dominating property type in the whole data set.

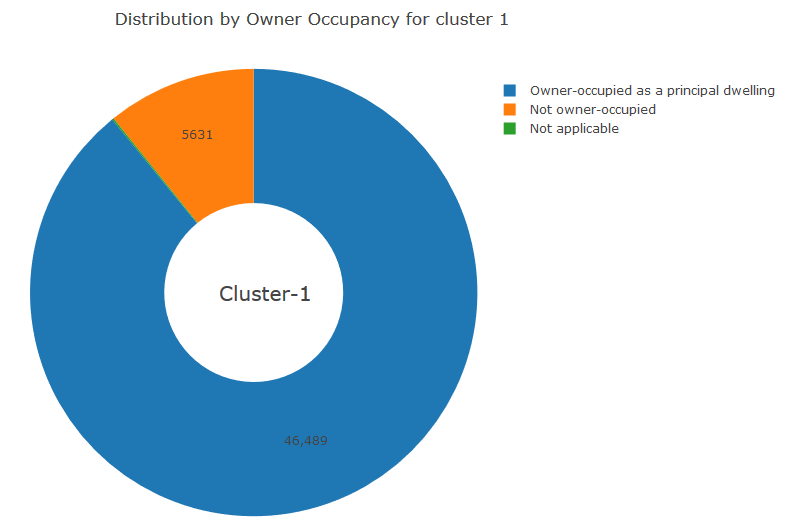
## Loan Purpose

 *Figure 24: Distribution by Loan Purpose*

The above diagram on Loan purpose shows the following information:

1. The home improvement is dominating the cluster.
2. This cluster has no home purchase. The home purchase is overall dominating in the dataset

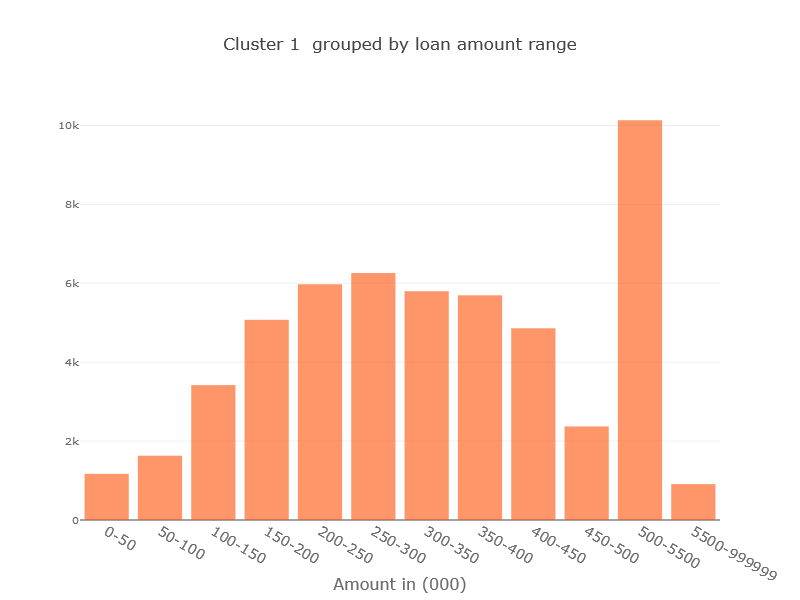
## Owner Occupancy

*Figure 25: Distribution by Owner Occupancy*

The above diagram on Owner Occupancy shows the following information:

1. The owner-occupied is dominating the cluster. The owner-occupied is overall all dominating owner occupancy in the whole data set.

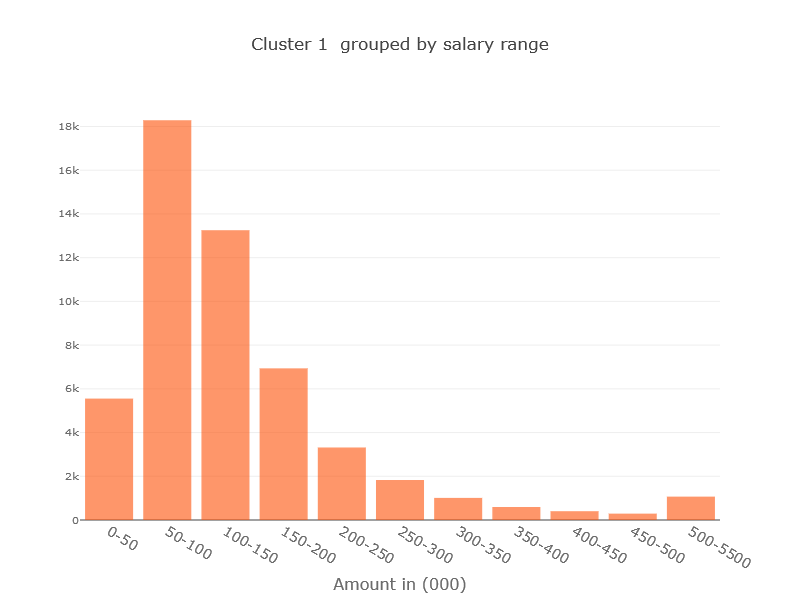
## Loan Amount

 *Figure 26: Cluster 1 grouped by Loan Amount*

The above diagram on Loan Amount shows the following information:

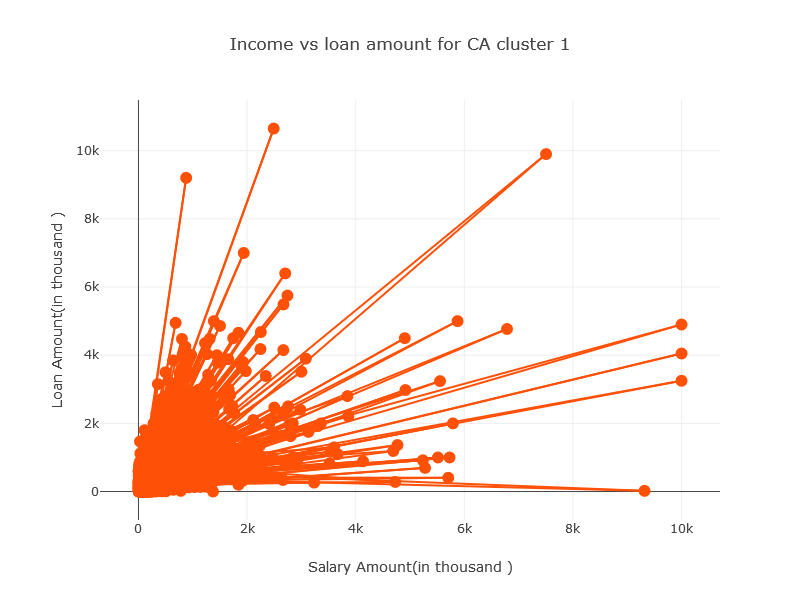
1. This cluster mostly contains all loan less than 5500.
2. The most dominating range is 500-5500.

## Applicant Income

 *Figure 26: Cluster 1 grouped by Income*

The above diagram on Applicant Income shows the following information:

1. This cluster mostly contains all loan less than 5500.
2. The most dominating range is 50-100.



*Figure 27: Income VS Loan Amount for Cluster 1*

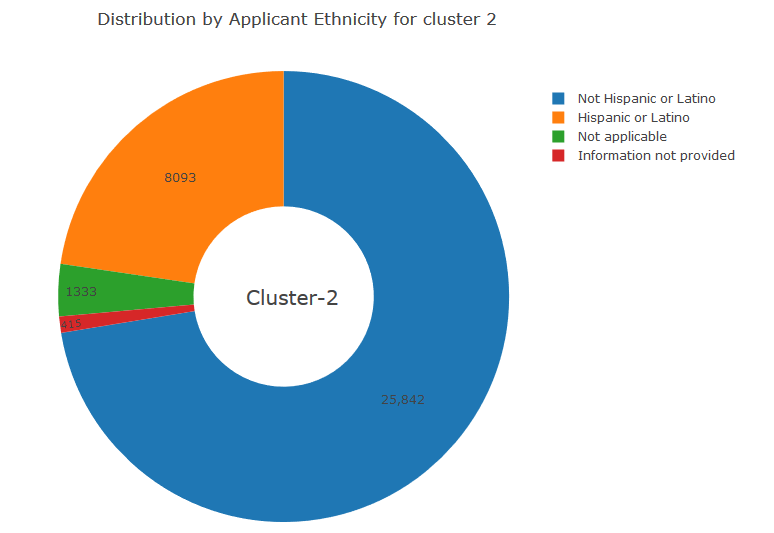
The above diagram on Applicant Income shows the following information:

1. This loan amount increases with salary amount. The most point in 0-2k.

## Exploring Cluster 2

The cluster 2 is explored for all the features.

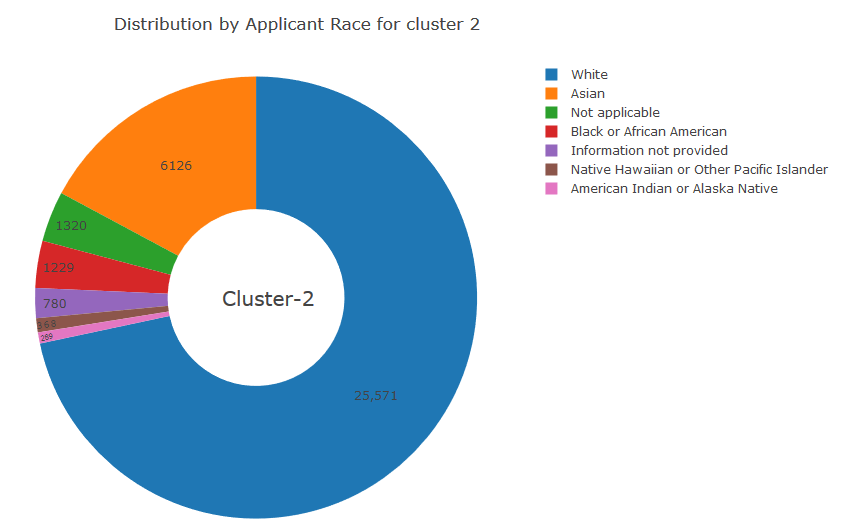
## Applicant Ethnicity

 *Figure 28: Distribution by Applicant Ethnicity for Cluster 2*

The above diagram on Applicant Ethnicity shows the following information:

1. The not Hispanic ethnicity is dominating the cluster. The not Hispanic is overall all dominating ethnicity in the whole data set.

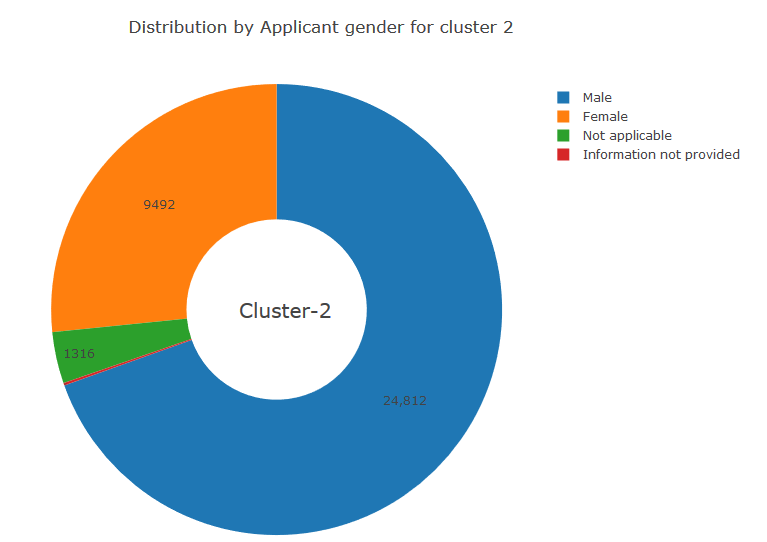
## Applicant Race

 *Figure 29: Distribution by Applicant Race for Cluster 2*

The above diagram on Applicant Race shows the following information:

1. The white race is dominating the cluster. The white is overall all dominating race in the whole data set.

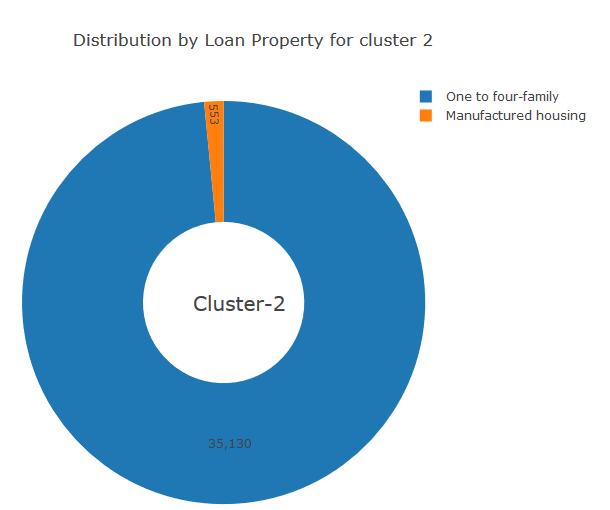
## Applicant Gender

 *Figure 30: Distribution by Applicant Gender for Cluster 2*

The above diagram on Applicant Gender shows the following information:

1. The male gender is dominating the cluster. The male is overall all dominating gender in the whole data set.

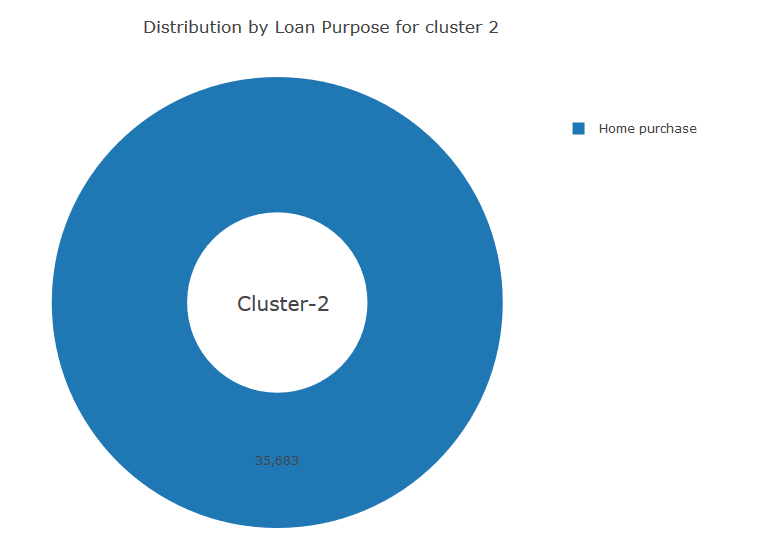
## Property Type

 *Figure 31: Distribution by Loan Property for Cluster 2*

The above diagram on Loan Property for Cluster 2 shows the following information:

1. The one to four family is dominating the cluster. The one to four family is overall all dominating property type in the whole data set.

## Loan Purpose

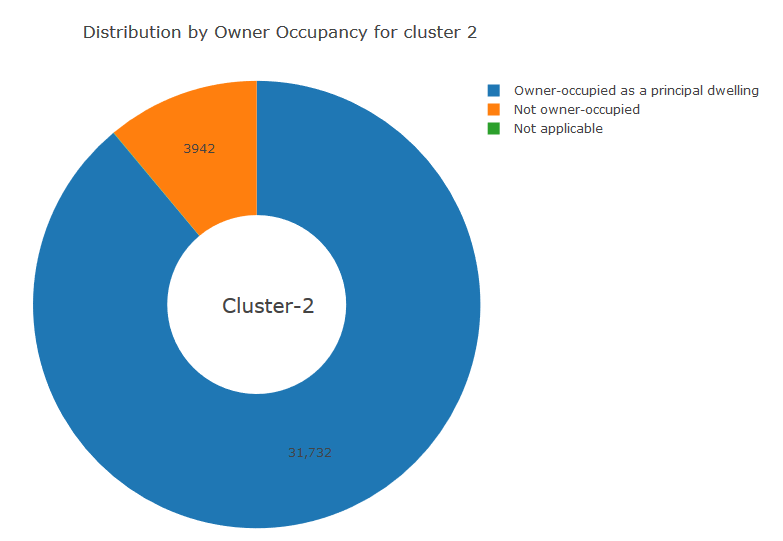


*Figure 32: Distribution by Loan Purpose for Cluster 2*

The above diagram on Distribution by Loan Purpose for Cluster 2 shows the following information:

1. The home purchase is the only purpose in the cluster. The home purchase is overall dominating in the dataset

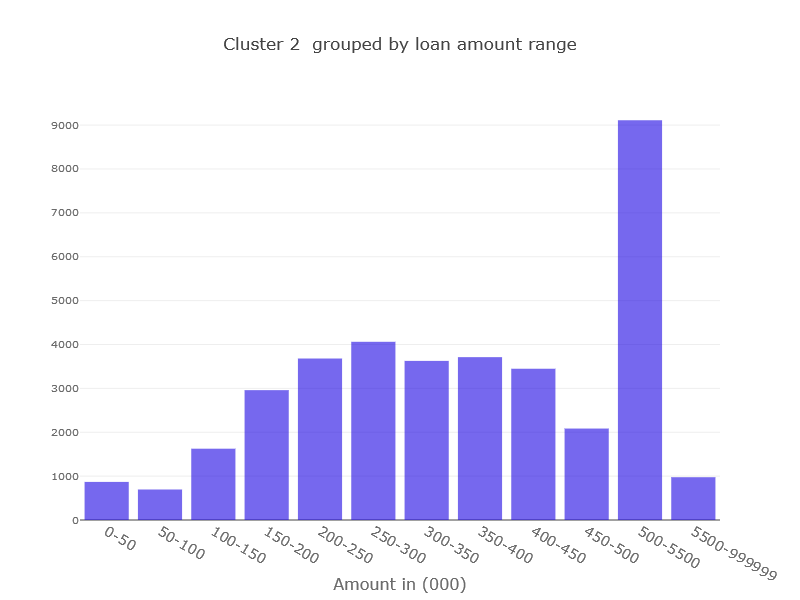
## Owner Occupancy

 *Figure 33: Distribution by Owner Occupancy for Cluster 2*

The above diagram on Distribution by Owner Occupancy shows the following information:

1. The owner-occupied is dominating the cluster. The owner-occupied is overall all dominating owner occupancy in the whole data set.

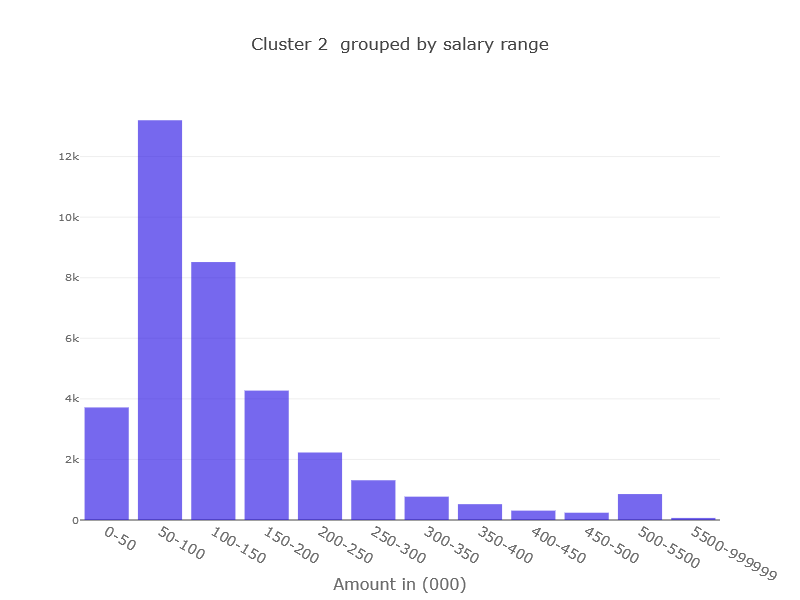
## Loan Amount

 *Figure 34: Cluster 2 grouped by Loan Amount Range*

The above diagram on Distribution by Loan Amount shows the following information:

1. This cluster mostly contains all loan less than 5500.
2. The most dominating range is 500-5500.

## Applicant Income

 *Figure 35: Cluster 2 grouped by Salary Range*

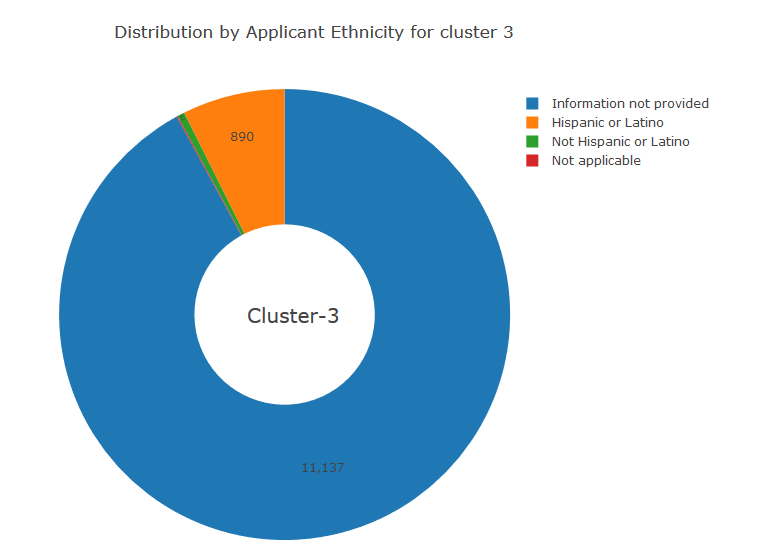
The above diagram on Applicant Income shows the following information:

1. This cluster mostly contains all loan less than 5500.
2. The most dominating range is 50-100.

## Exploring Cluster 3

The cluster 3 is explored for all the features.

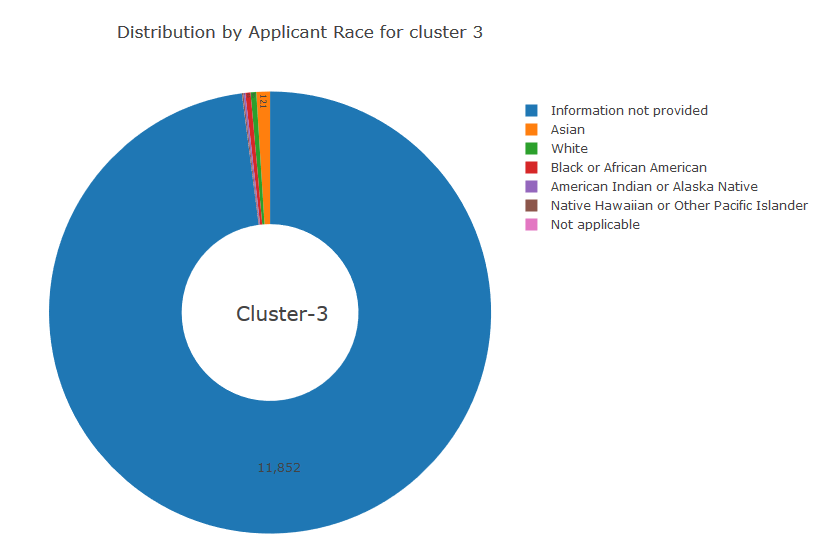
## Applicant Ethnicity

 *Figure 35: Distribution by Applicant Ethnicity for Cluster 3*

The above diagram on Applicant Ethnicity shows the following information:

1. The information not available is dominating in this cluster compare to not Hispanic in other two clusters.
2. .

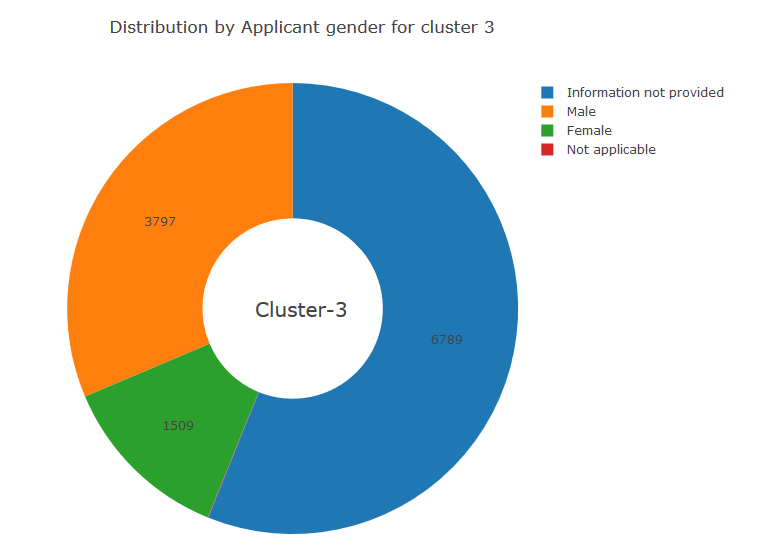
## Applicant Race

 *Figure 36: Distribution by Applicant Race for Cluster 3*

The above diagram on Applicant Race shows the following information:

1. The information not provided is dominating the cluster.
2. All other races have negligible participation in this cluster.

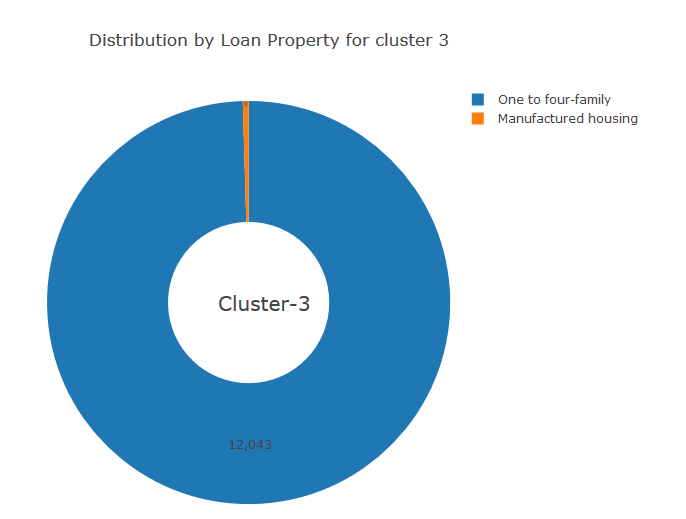
## Applicant Gender

 *Figure 36: Distribution by Applicant Gender for Cluster 3*

The above diagram on Applicant gender shows the following information:

1. The information not available gender is dominating the cluster

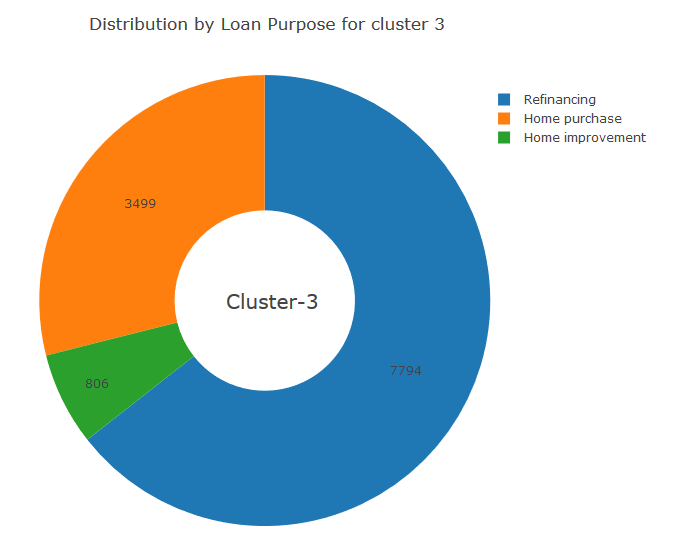
## Property Type

 *Figure 37: Distribution by Loan Property for Cluster 3*

The above diagram on Distribution by Loan Property shows the following information:

1. The one to four family is dominating the cluster. The one to four family is overall all dominating property type in the whole data set.

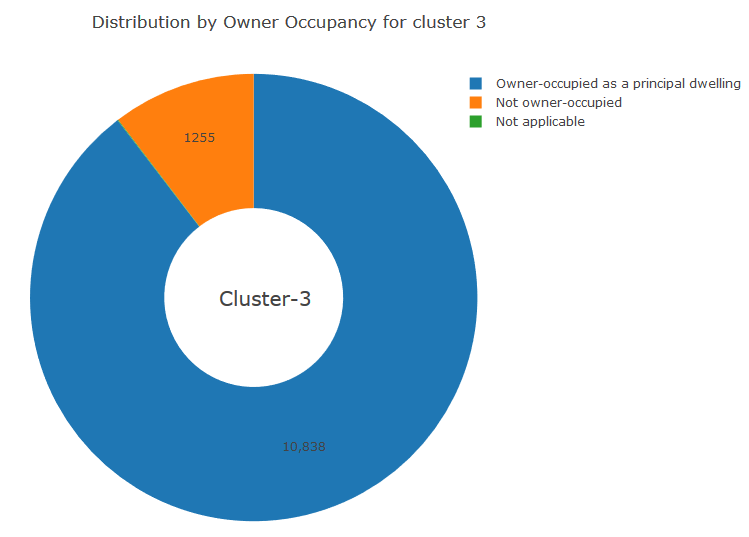
## Loan Purpose

 *Figure 38: Distribution by Loan Purpose for Cluster 3*

The above diagram on Distribution by Loan Purpose shows the following information:

1. The refinance is dominating the cluster.

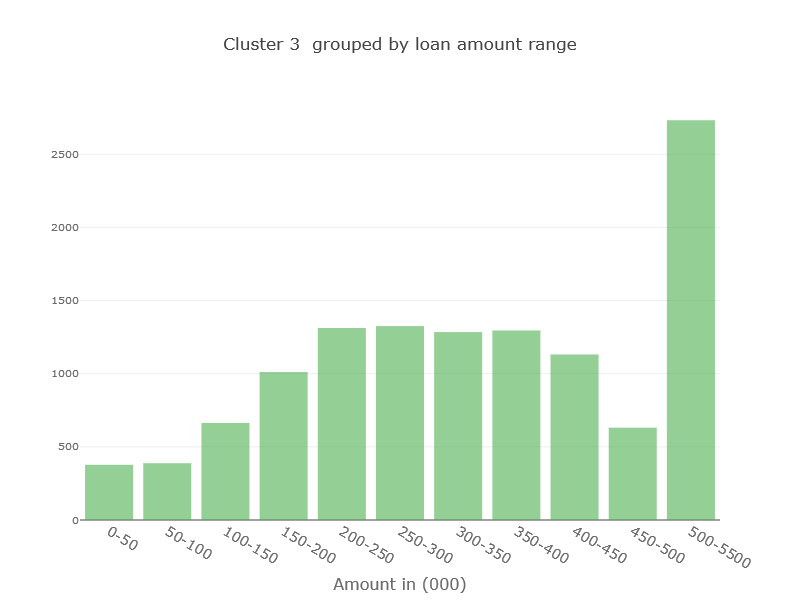
## Owner Occupancy

 *Figure 39: Distribution by Owner Occupancy for Cluster 3*

The above diagram on Distribution by Owner Occupancy shows the following information:

1. The owner-occupied is dominating the cluster. The owner-occupied is overall all dominating owner occupancy in the whole data set.

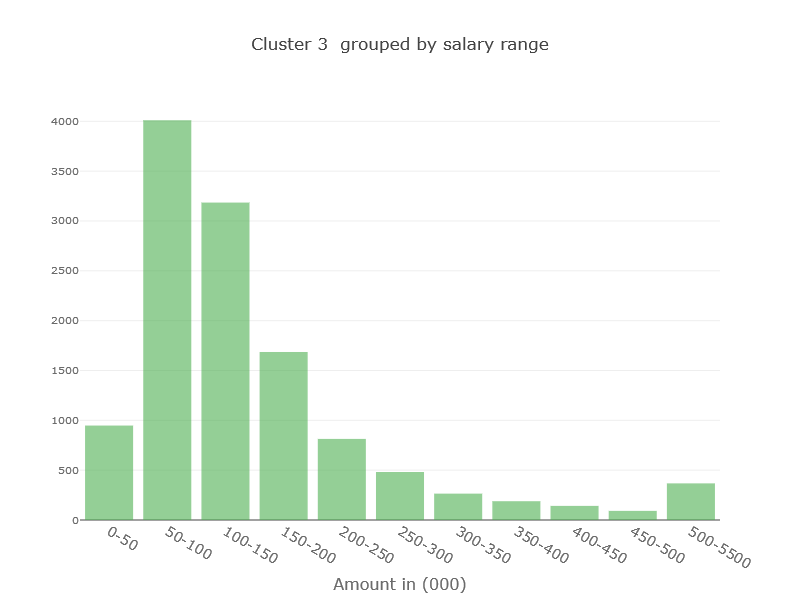
## Loan Amount

 *Figure 40: Distribution by Loan Amount for Cluster 3*

The above diagram on Distribution by Loan Amount shows the following information:

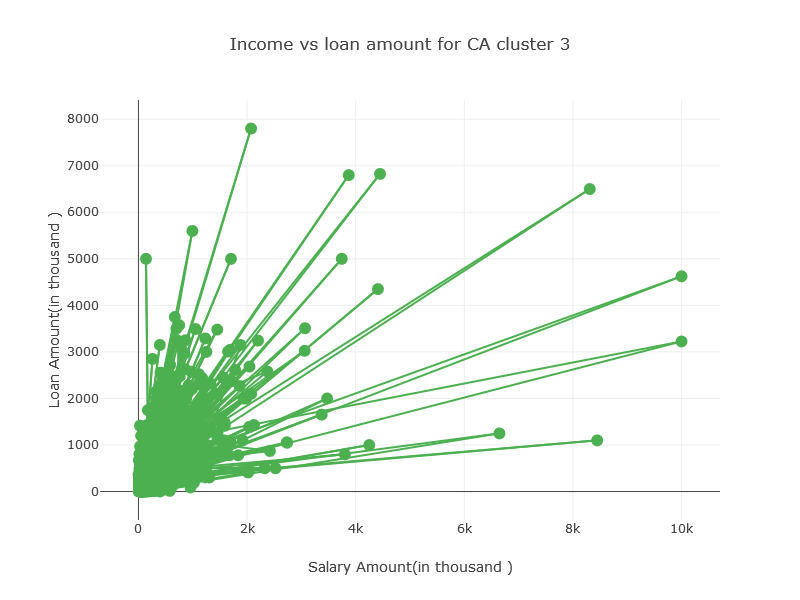
1. This cluster mostly contains all loan less than 5500.
2. The most dominating range is 500-5500.

## Applicant Income

 *Figure 41: Cluster 3 grouped by Salary Range*

The above diagram on Applicant Income shows the following information:

1. This cluster mostly contains all loan less than 5500.
2. The most dominating range is 50-100.

*Figure 42: Income VS Loan Amount for CA Cluster 3*

The above diagram shows the following information:

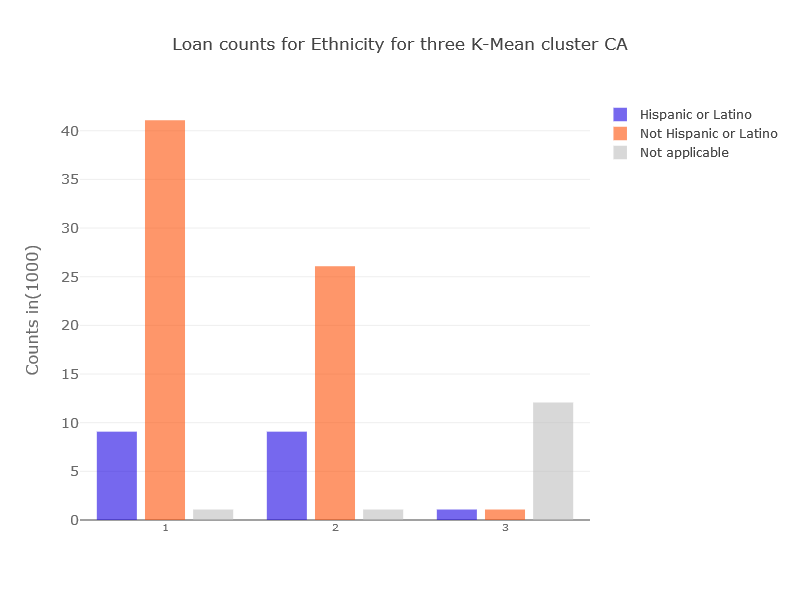
1. This loan amount increases with salary amount. The most point between 0-2k.

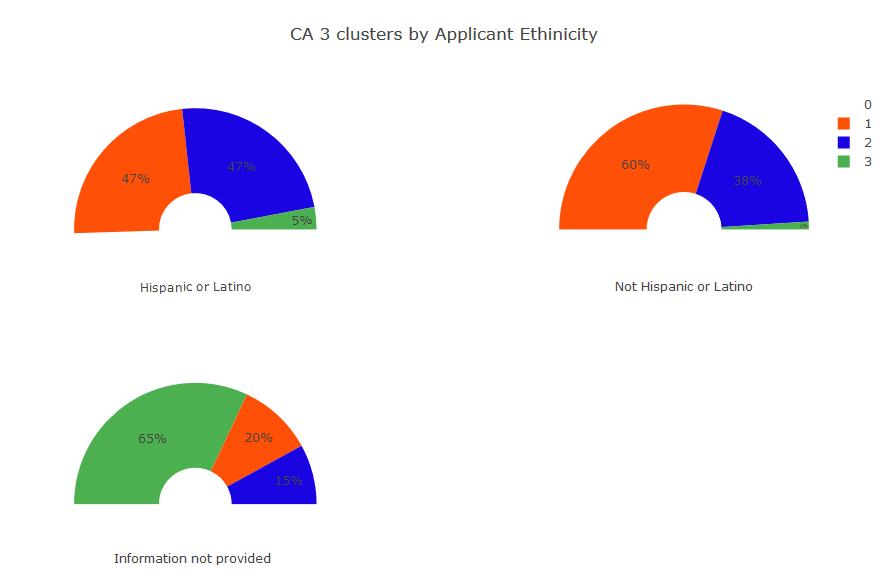
## Comparing Clusters

All the clusters are compared for all the features.

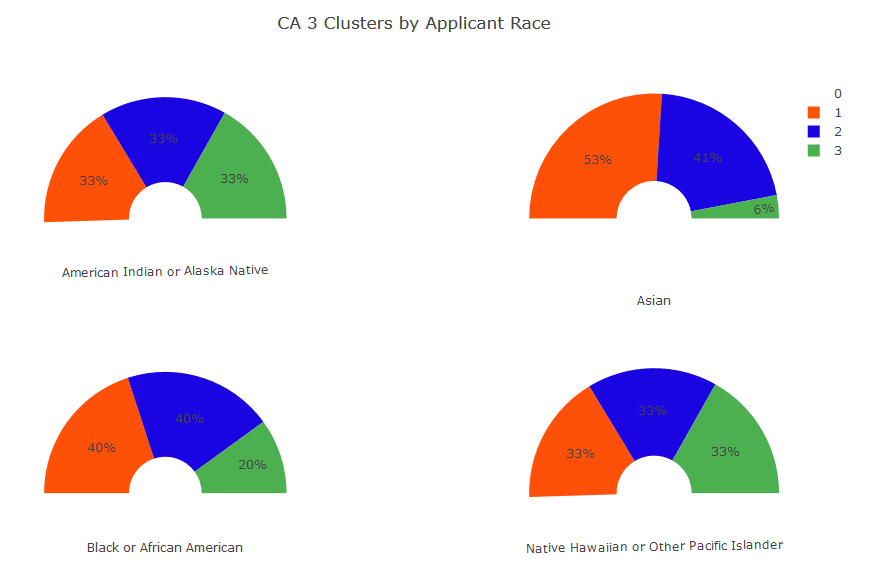
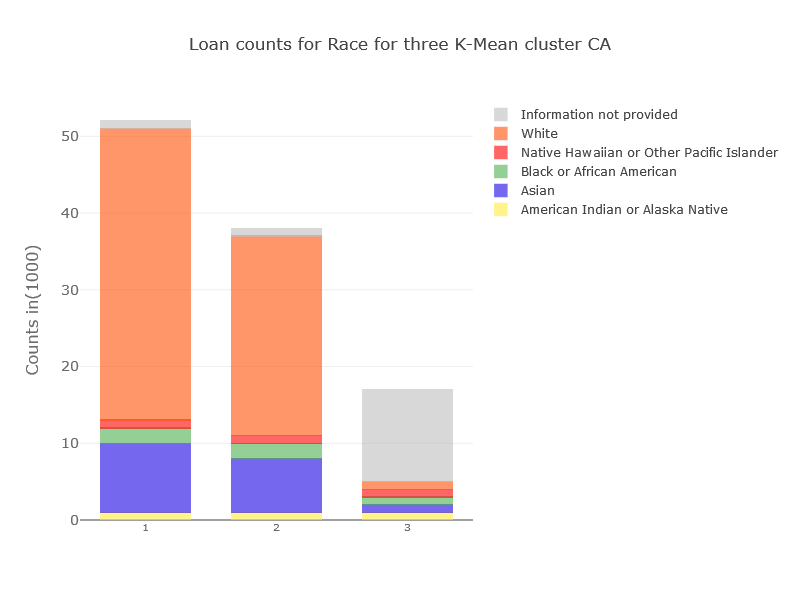
|  |  |  |  |
| --- | --- | --- | --- |
|  | **Cluster 1** | **Cluster 2** | **Cluster 3** |
| **Size** | 52180 | 35683 | 12099 |
| **Applicant Ethnicity** | Not Hispanic | Not Hispanic | Not Applicable |
| **Applicant Race** | White | White | Not Available |
| **Applicant Gender** | Men | Men | Unknown |
| **Property Type** | 1 to 4 Family | 1 to 4 Family | 1 to 4 Family |
| **Loan Purpose** | Home Improvement | Home Purchase | Refinance |
| **Owner Occupancy** | Owner Occupied | Owner Occupied | Owner Occupied |
| **Loan Amount** | Most records in every range | More records in every range | Least records in every range |
| **Applicant Income** | More records in every range | More records in every range | More records in every range |

## Applicant Ethnicity

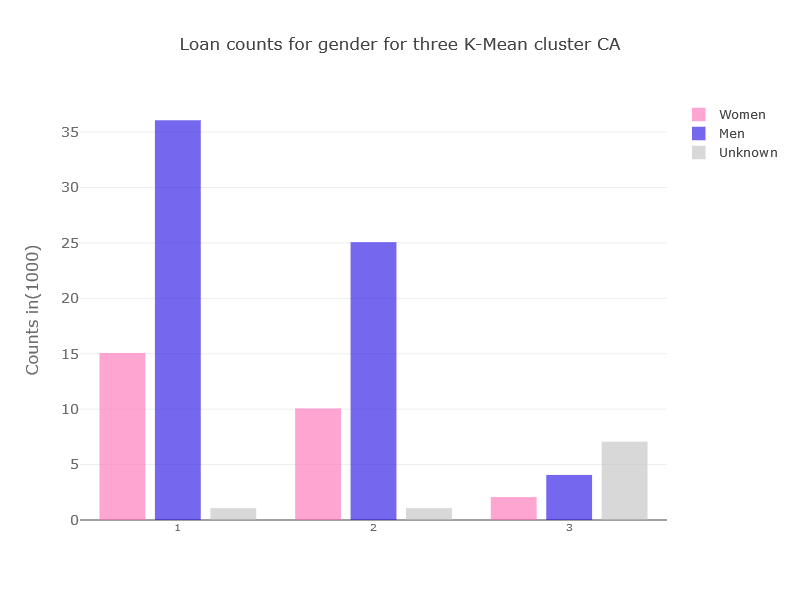


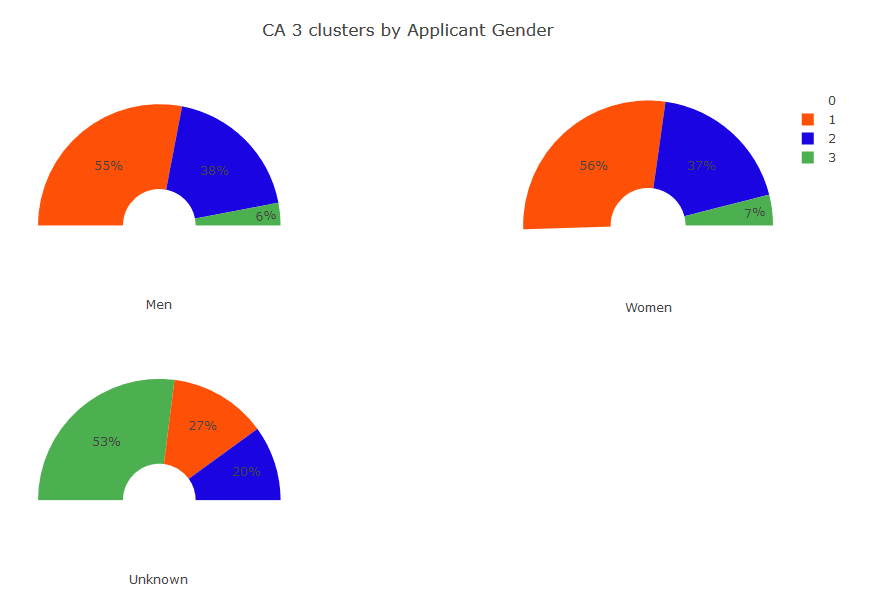


## Applicant Race



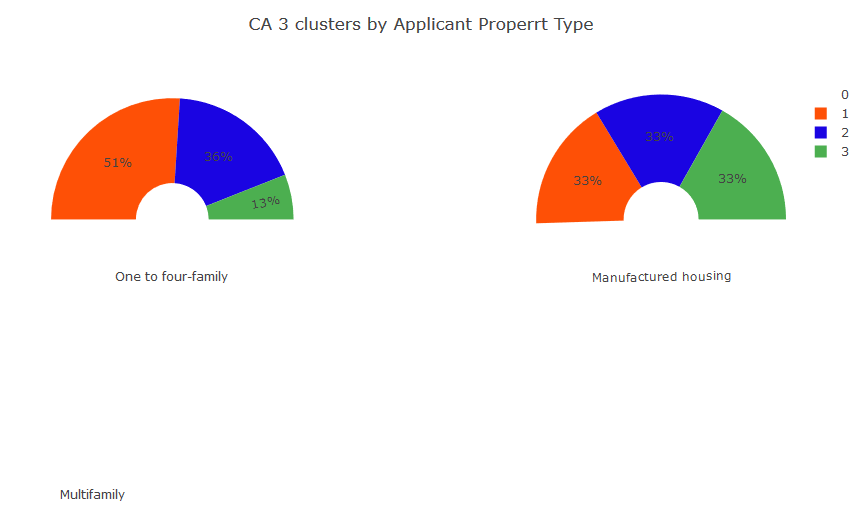
## Applicant Gender



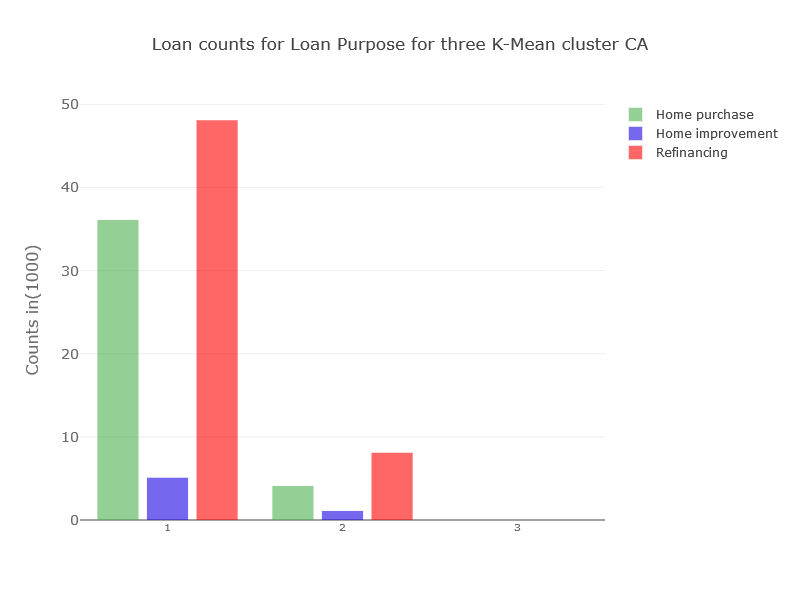


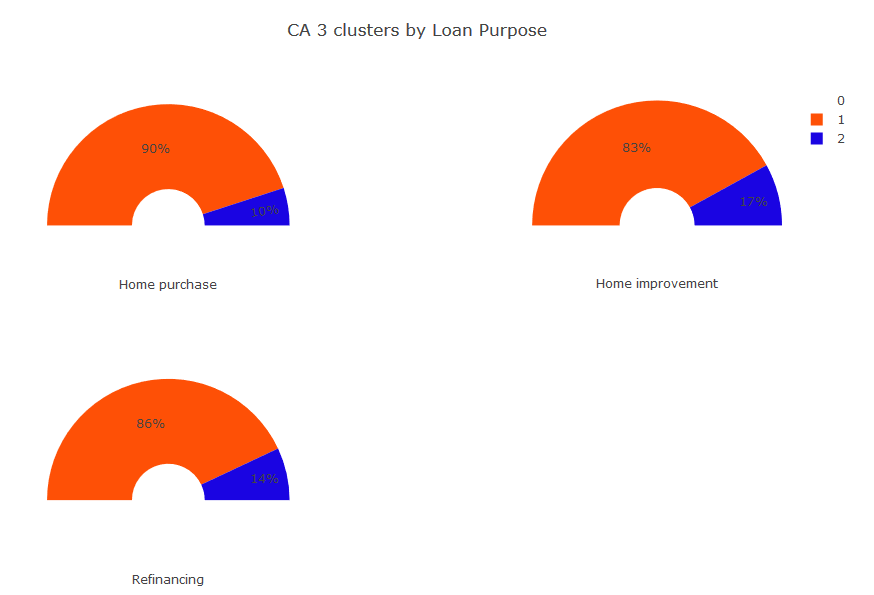
## Property Type



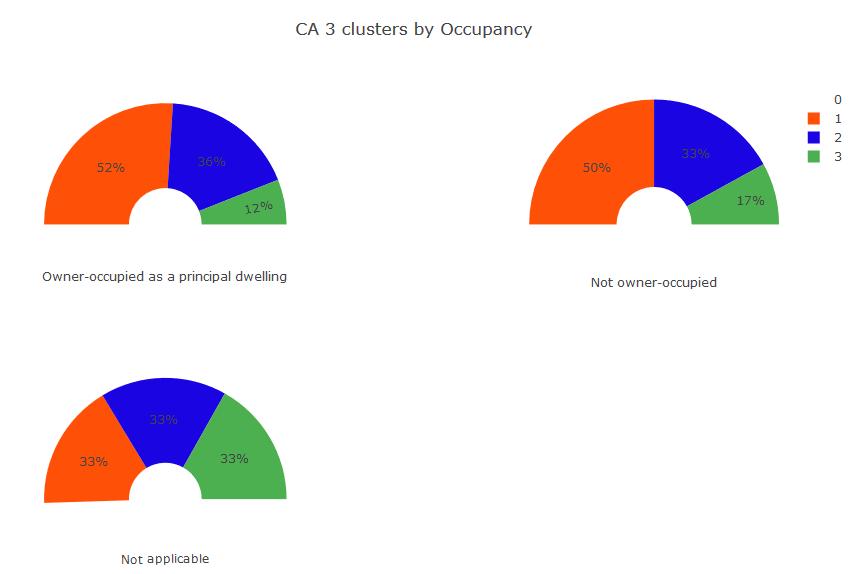


## Loan Purpose

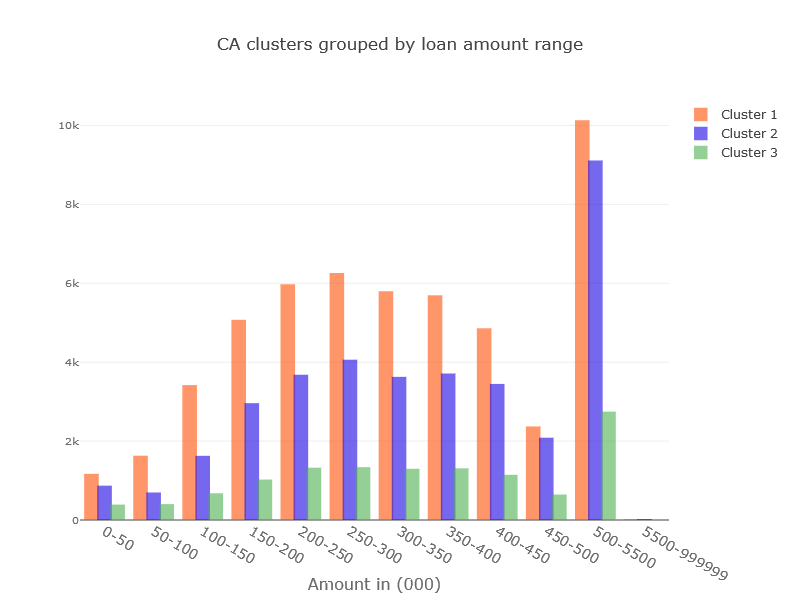


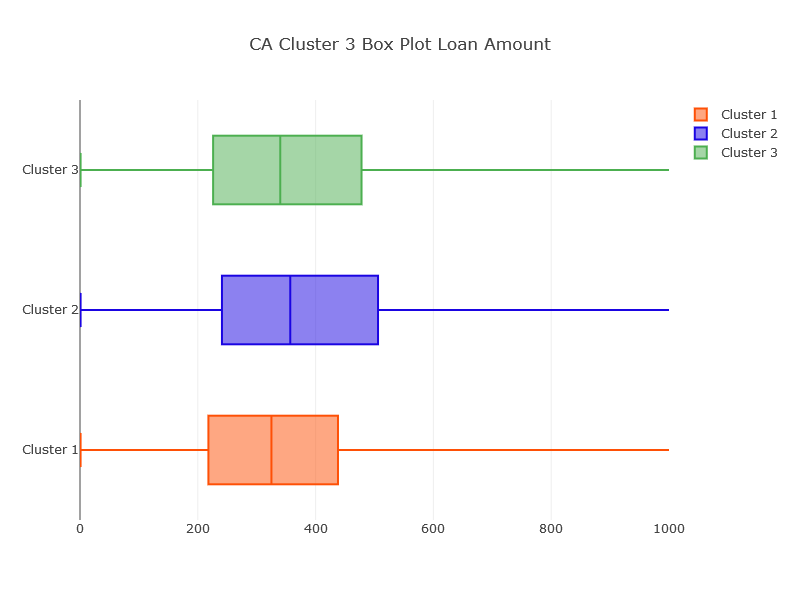


## Owner Occupancy

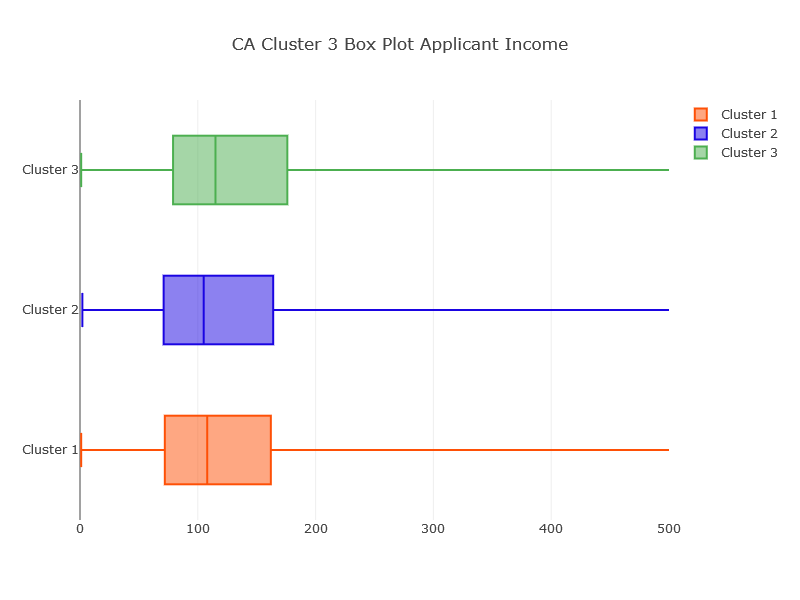
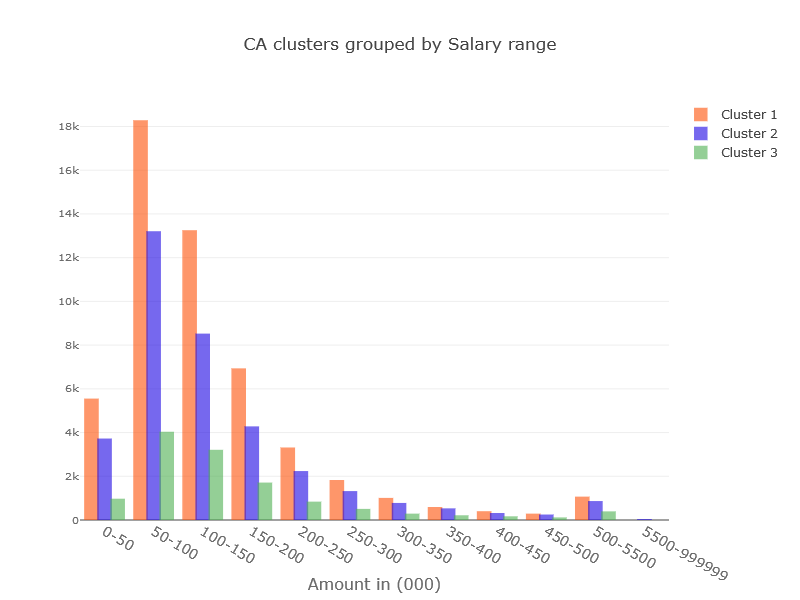


## Loan Amount





## Applicant Income



## Insight from Clusters

After analyzing the all the features within a cluster there is insight extracted from clusters. This can be used as a recommendation to the client.

## Specific products for home improvement loan

The home improvement is in high demand in all ethnicity, race, and genders. The number of the home improvement loan compares to purchase. The approval rates for home improvement is lower compared to other loans, but the products specific to improvement can help.

## Refinance is a new favorite

The refinance request is very close to the new home loan request. There are some areas where refinances are higher than a new home loan.

## Hispanic Vs Non-Hispanic

The non-Hispanic ethnicity people are the majority of customer bases but assuming a non-Hispanic included vast number of races, the number of Hispanic customers are insignificant numbers

## White, Asian & African American

The race should not be considered for discrimination but there can be things like the language of advertisement material which help target customers. The majority of the client base belongs race white. But Asian and African American are respectively two races with the most customer based.

## Male is 70% of the applicant

The most primary applicant for the loan is males. This can help is using male-centric magazine and website and other places as better brand placement.

## Home for a single family in demand

Most people taking a loan for a single family. The 98% of the market is for single-family houses as compare to multi-family or manufactured home.

## People buying a home for themselves

97% of loans are purchased are occupied by the applicants as principal. This might indicate that people are not buying a house for investment purposes. This can help builders and agent to understand the market.

## People financing good amount on home improvement

For all ranges the loan amount number of loans for home purchase and home improvement are close. Defiantly more on the home purchase but the amount on home improvements are also close.

## Middle-class spending more on home loans

Most numbers of people in all cluster are in the salary range of 50-100 thousand. The most number of new home improvement and purchase are dominated by these salary ranges.

## Conclusion and recommendations

We explored data and found out the customer segmentation based on three clusters of K means. After performing machine learning and visual analysis of data the following are conclusion and recommendation

1. The mortgage customer in the state of CA is very good sample space to understand overall. behavior as the visualization of CA and all state data are very similar.
2. This analysis is done using 100,000 sample out of millions of loans in CA.
3. Most of the feature available are significant in contributing towards the clustering.
4. The ML used is K means for the clustering.
5. The data should be at least explored state wise to create better understanding instead of all state as one big data set.
6. The similar exercise should be performed for all other states to create better understanding as per state.
7. The code is usable and can be very well utilized in running for any state, county or area based on the requirement.
8. The visualization created can be presented to the business team to generate more insights from their perspective.

# **Future Work**

The project has tried to cover many aspects of data science. But there is still scope to improve the

methodology or explore more facets of the project. As for endnotes, this will be marked as future

work in reference to this project.

## Exploring more states

The clustering is done on the only state of CA. The work can be extended to include all 50 states.

## Bringing more features in customer data

The clustering is done on data available to the public via govt agency. There better insights can be created based on more feature related to clients available within the company.

## Other clustering algorithms

Due to computation work restriction, the other algorithms like Agglomerative clustering and DBScan cannot be performed. The single computer laptop is unable to process the request. The python scripts are created (available with the code on Github) to run without Juypter notebook but unable to run on server level configuration machine. The more algorithm can be explored to compare clusters.

## Technology

Currently, the files are processed using code running in a Jupiter notebook on the local computer.

The process like data wrangling, data cleaning, and modeling took hours to finish on a local

machine. Cloud technologies like AWS, Azure, and Cloudera etc. can be used to speed up the

processing. The clustering is done on the only state of CA for only 100, 000 records. The work can be extended to include all 50 states with more data on these technologies.