Mortgage Loan Application Data Exploratory Analysis

Finding meaning and importance within Mortgage Loan Application Dataset

Introductory Story

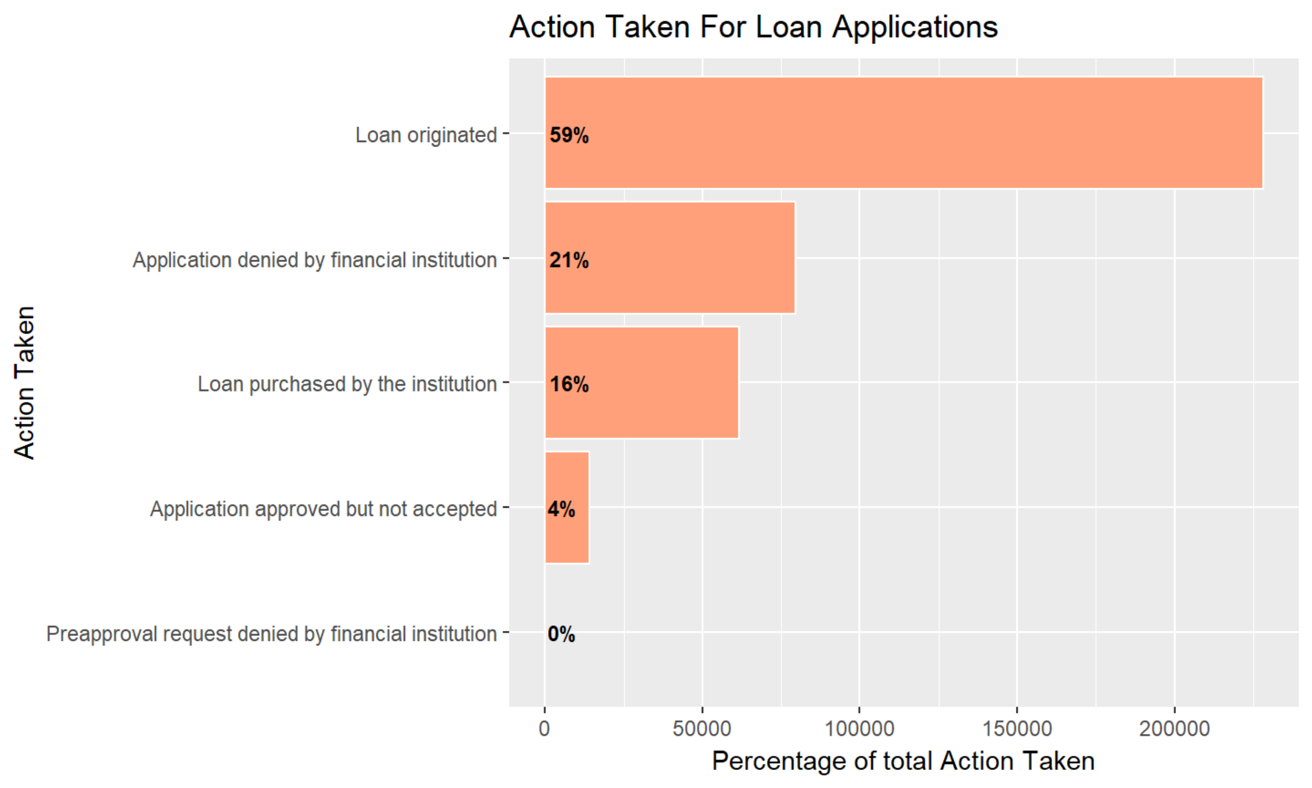
- A client wishes to buy a home, but they do not have full cash-in-hand. So the client must apply for a mortgage loan. They proceed to tell the bank all their financial matters and the home that they wish to purchase. Once the bank makes its decision to approve the loan, the client will sign all necessary papers to close the deal. This is called **Loan Origination**.

Understanding data by Top Categories

- The data can be grouped into 5 major categories:
  - **Loan**
  - **Applicant Information**
  - **Property Type**
  - **Location**
  - **Lender Information**

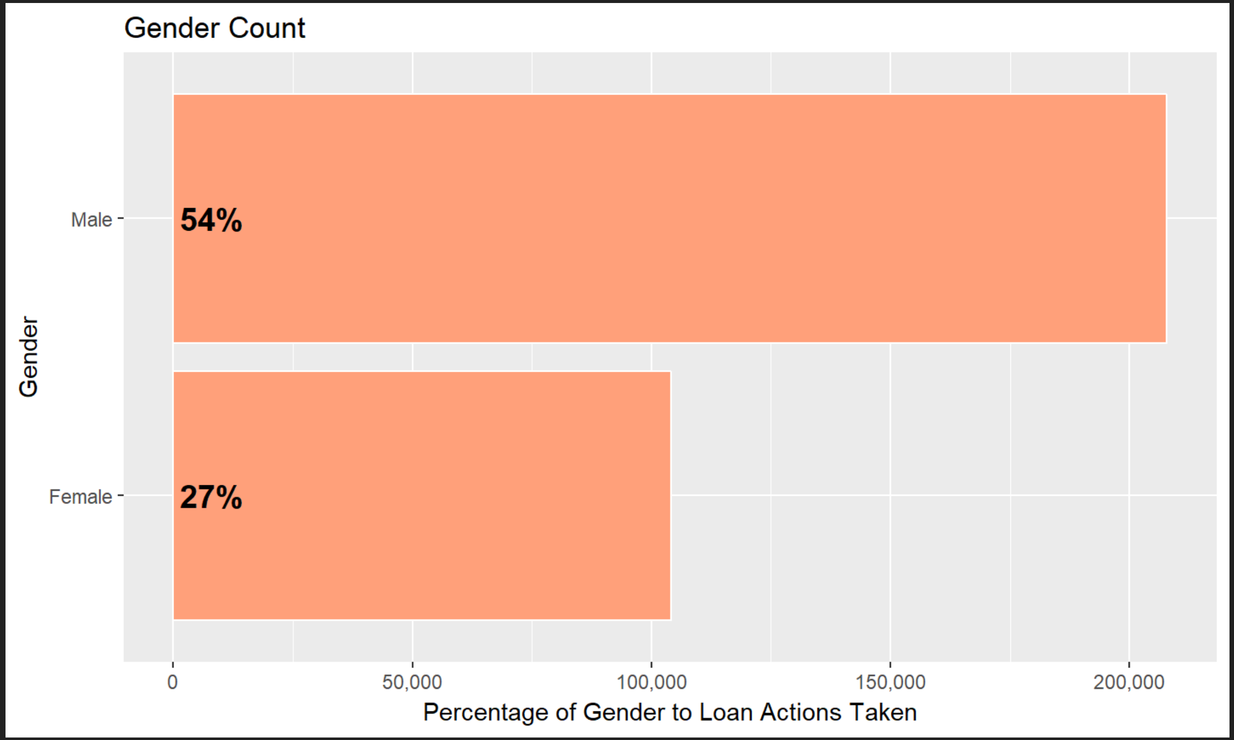
1. Loan Category

- Responses to loan applications.
- 59% of loans have "originated".



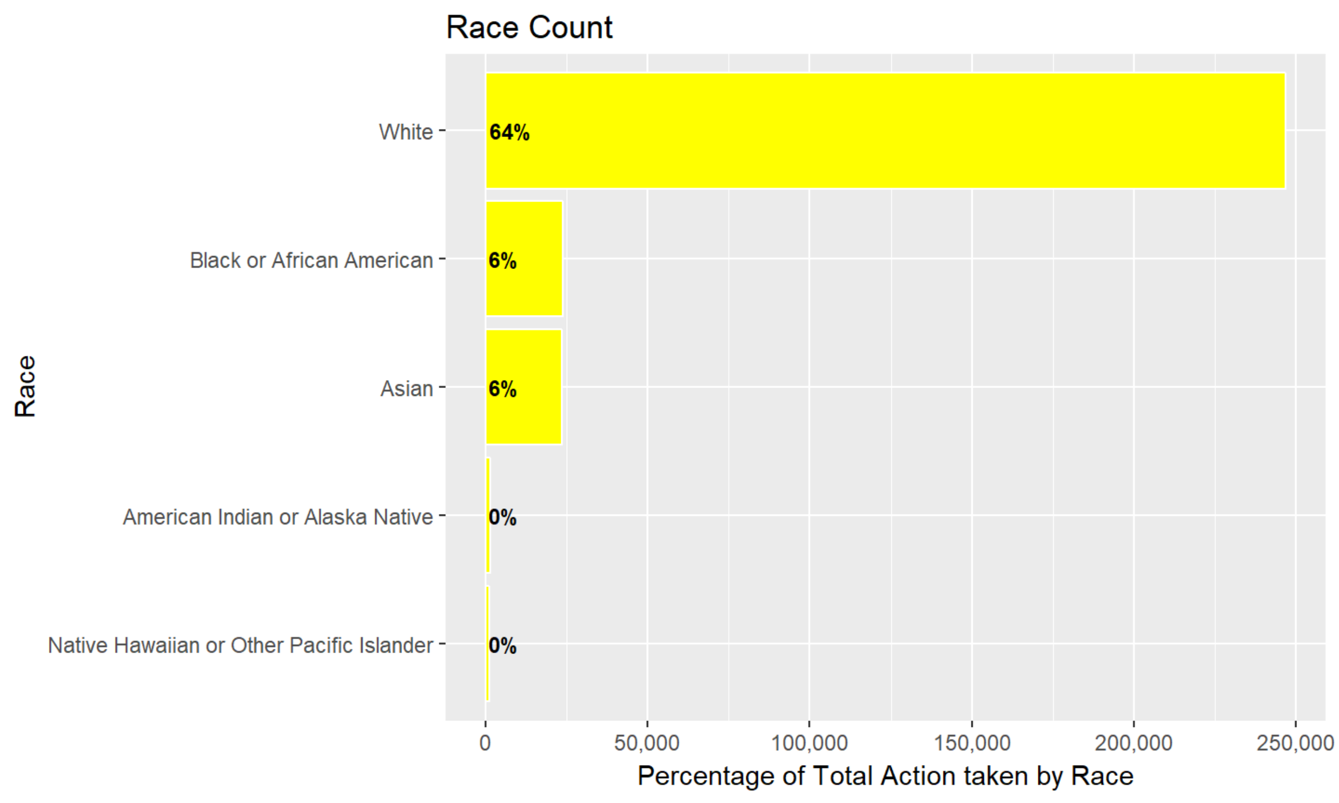
2. Applicants

- Male close to double rate of applicants than their counterpart



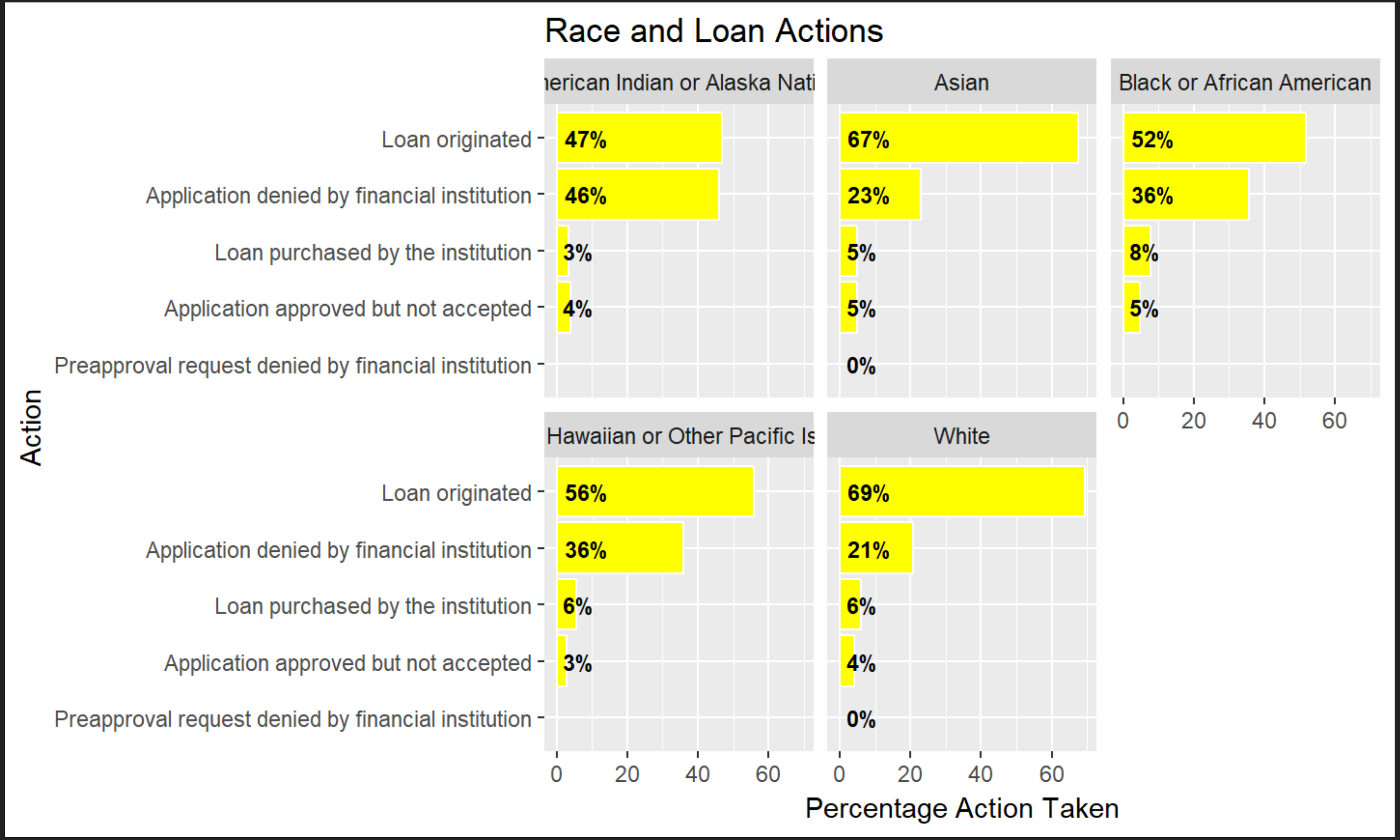
3. Race

- Majority of the loan applicants are "White" in NY.



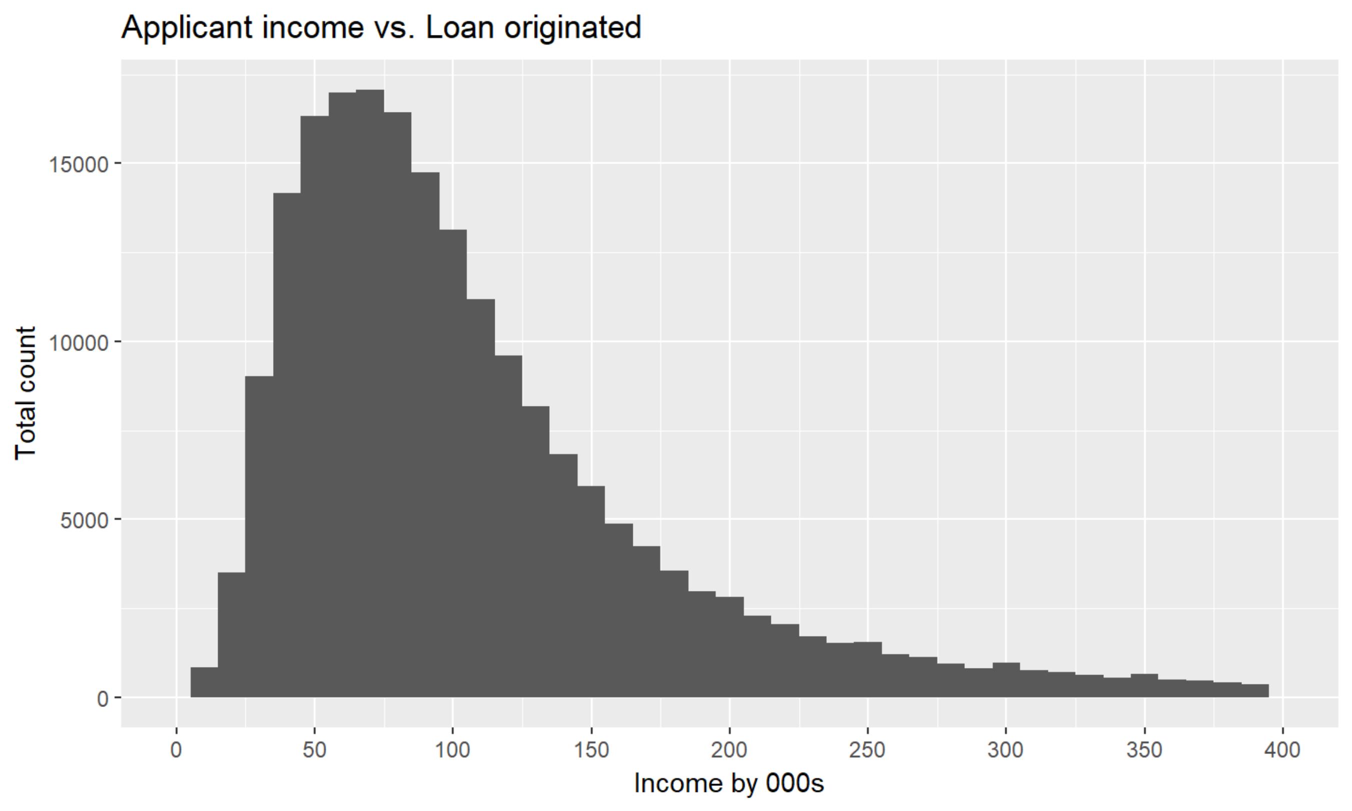
4. Actions taken based on Race

- White/Asians are ~68% loan originated while other race are much lower with higher denial rate.



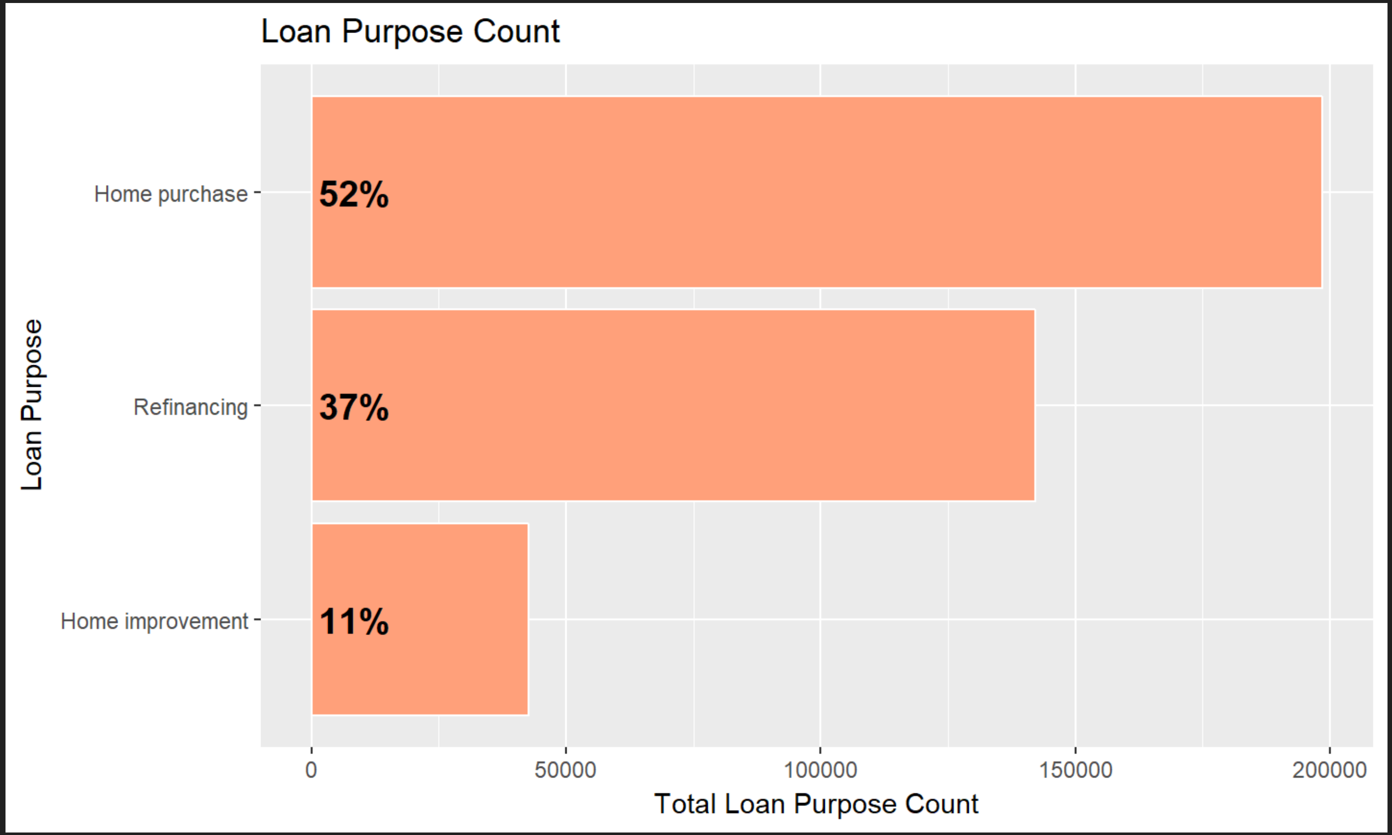
5. Loan Origination based on Income

- Highest Loan Originated is from income of 55K-75K



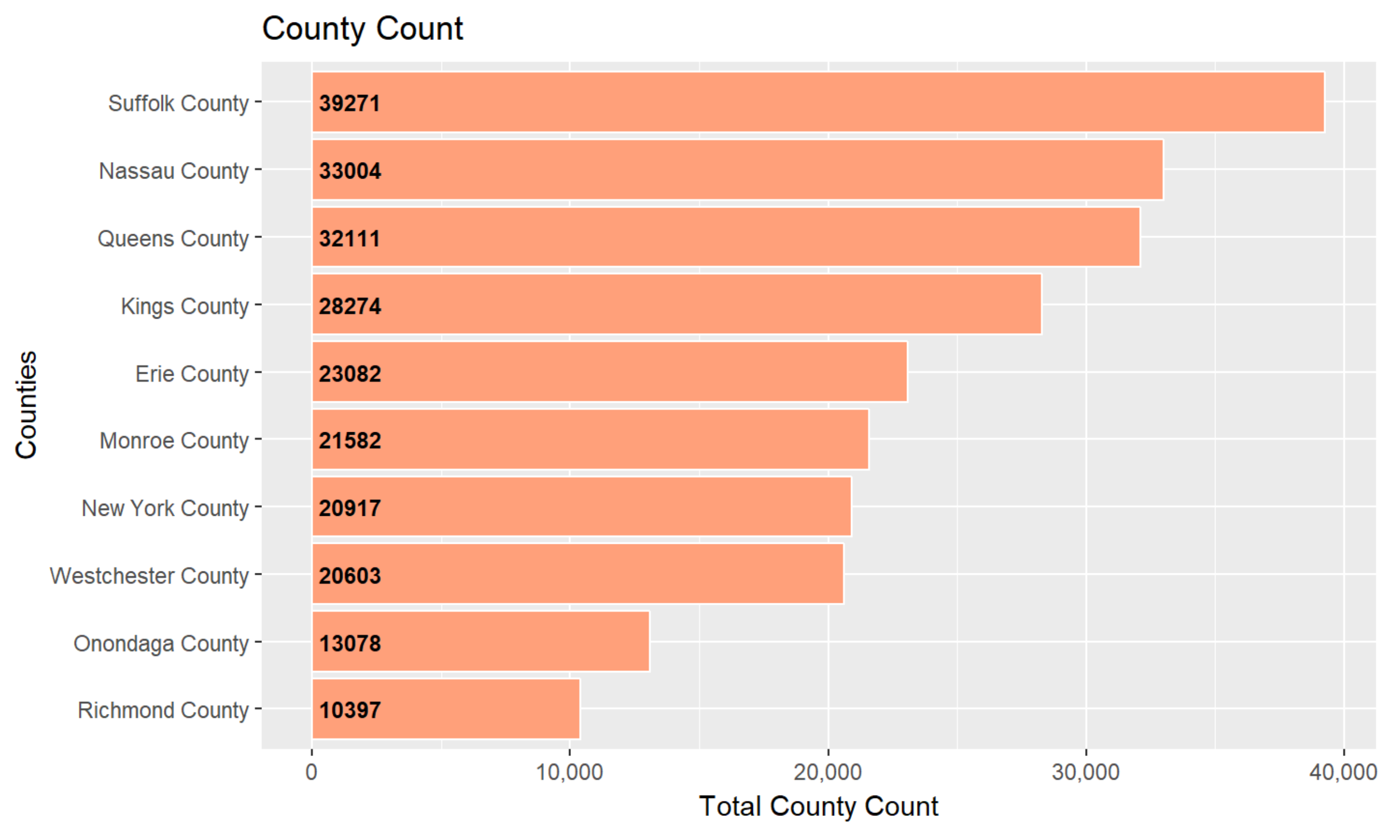
6. Loan Purpose Types

- 3 types of loan purpose(desc.)
  - Home Purchase
  - Home Refinancing
  - Home Improvement



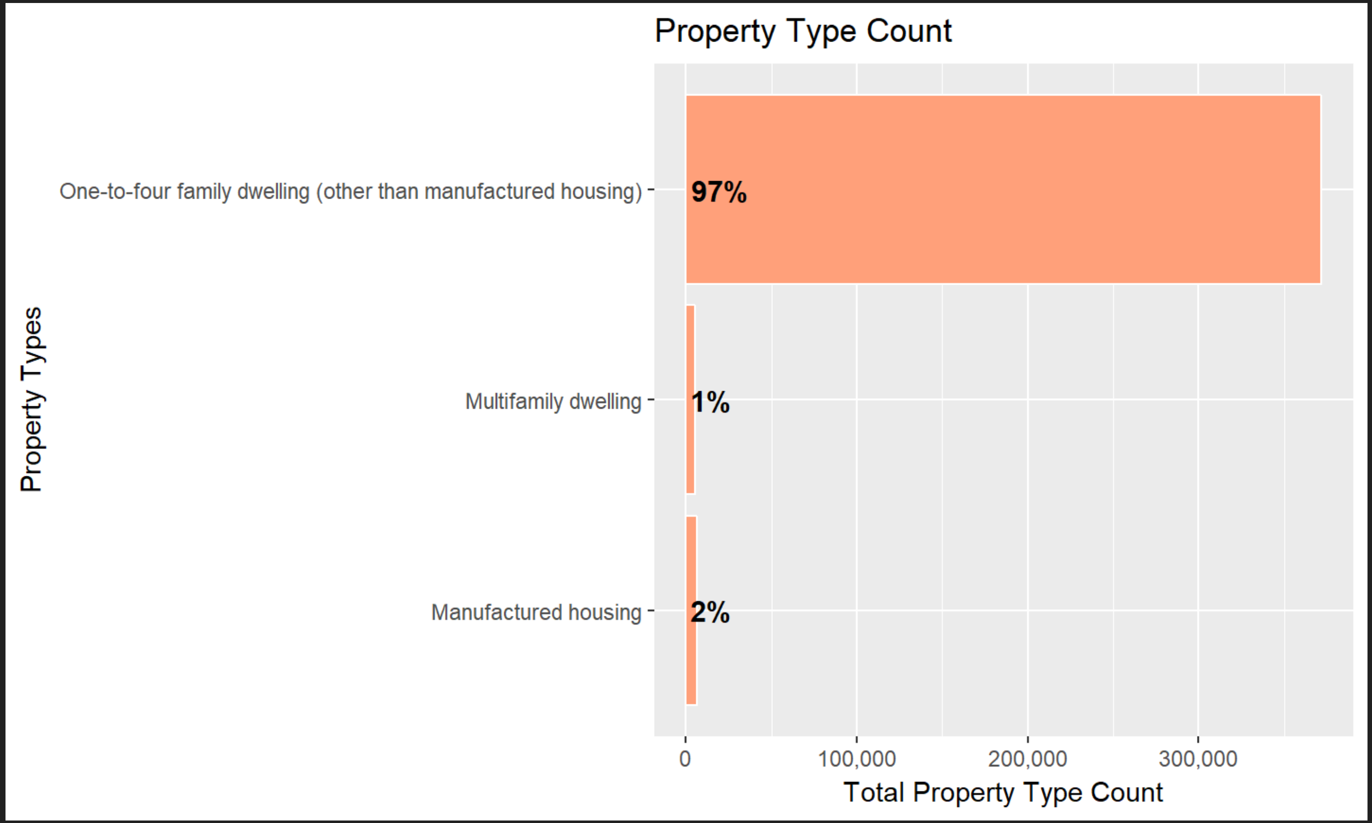
7. Top 10 counties with Loan Origination

- 7 of 10 counties are within New York City area



8. Property Types for Loan Applied

- High rate of application for single family dwelling



9. Predictive Modeling Categories

- Selected Data for predictive modeling
  - Action Taken
  - Applicant Ethnicity
  - Applicant Income in 000s
  - Applicant Race
  - Co-applicant ethnicity/sex
  - County
  - Hoepa Status
  - Lien Status
  - Loan Purpose
- Loan Types
- MSAMD
- Owner Occupancy
- Preapproval
- Property Type
- Purchaser Type
- Loan amount

10. Logistic Regression Model

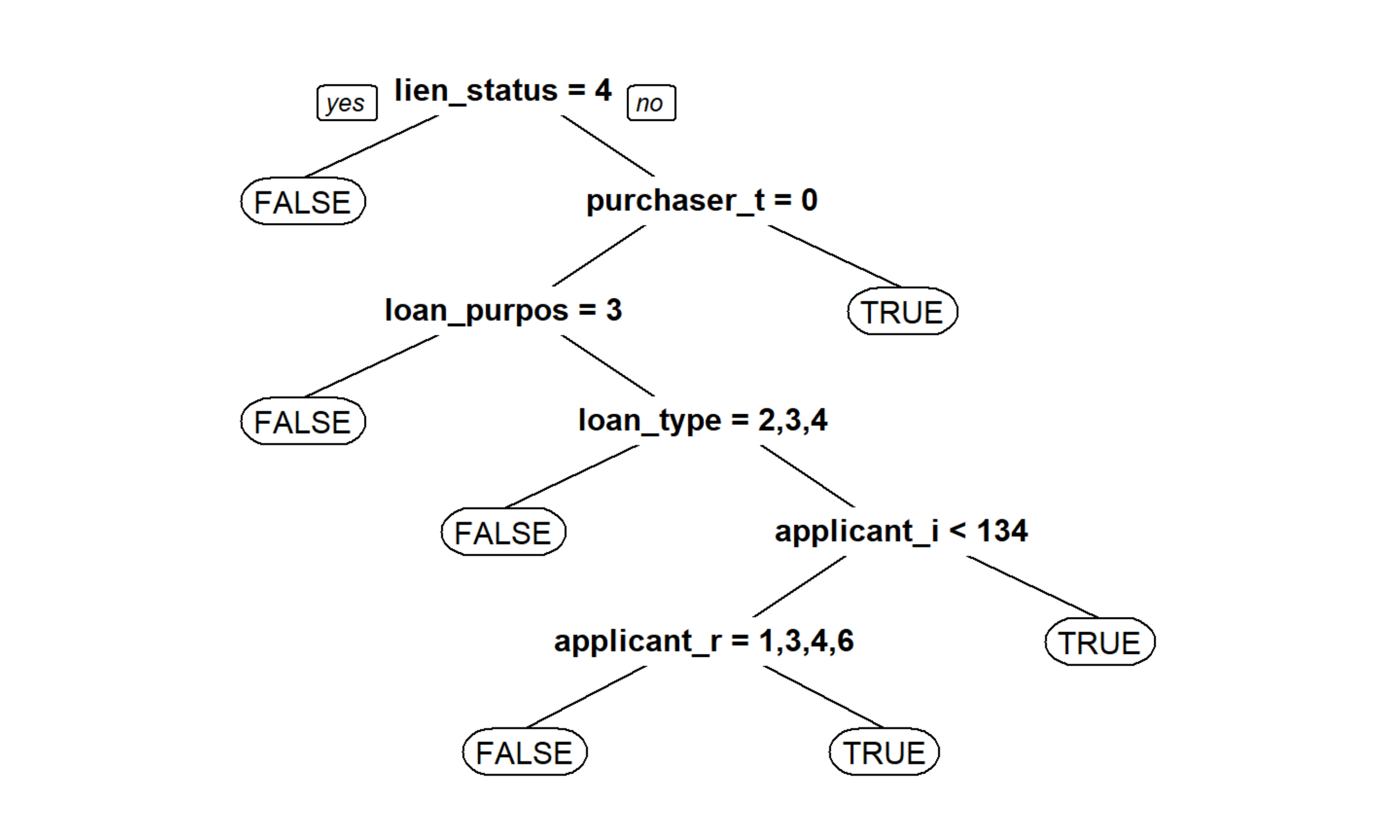
- Based on the logit model we summarise

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.0000  0.3781  0.6787  0.6435  1.0000  1.0000
```

- With an accuracy of:
  - 83.77%

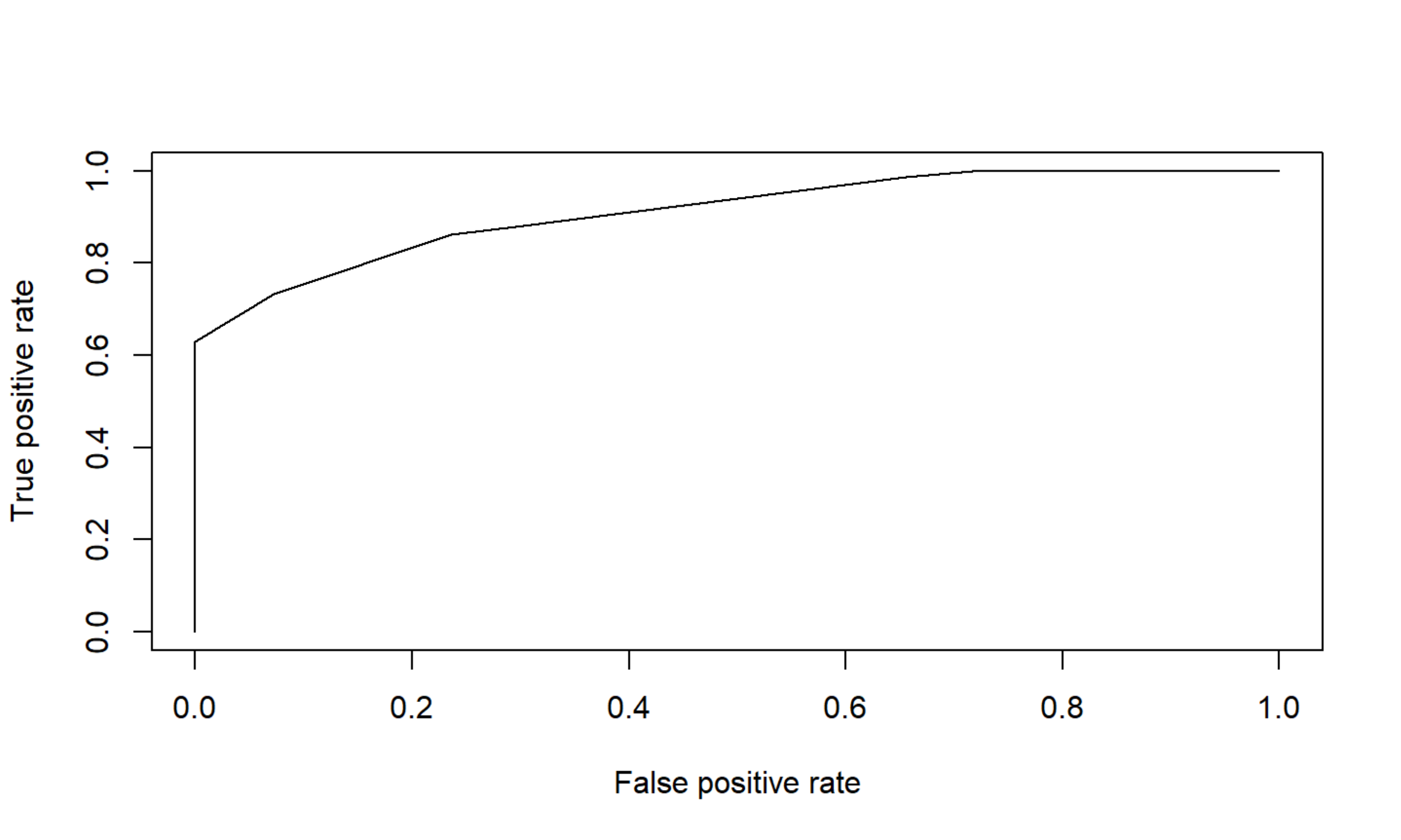
11. CART Predictive Model

- The priority of importance based on decision tree:
  - Lien Status
  - Purchaser Type
  - Loan Purpose
  - Loan Type
  - Applicant Income
  - Applicant Race



12. ROC Curve and AUC Calculation

- The AUC calculation based on the ROC Curve:
  - 91.17%



Logistic Regression VS Decision Tree

- Logistic Regression feels much more closer to a linear dimension, where you pit dependent variable against the listed independent variables.
- Decision Tree seems to calculate as a “big picture”, where you take the prediction into a multiple dimension, able to create a much clearer priority list, then pit the list against the next priority variable.