Predicting Mortgage Rates From Government Data

A report designed by Ekaterina Marinova, November 2019

Executive Summary

This document presents and analysis of data with the goal to predict the rate spread of mortgage applications according to the given dataset, which is adapted from the Federal Financial Institutions Examination Council's (FFIEC). The analysis is based on 200 000 observations, which has been initially manipulated through data preparation best practices as part of the data cleansing and exploration phase.

The approach that I have followed was the CRISP-DM methodology in which I went through all the six phases in order to do the complete prediction model for the mortgage rate. During the data and business understanding phases I have went through the description of the variables provided and through the data exploration analysis, which I have conducted in R studio. Following, I have gone through the Data Cleansing part where I have dealt with the missing values and the outliers in the dataset. In this part I have calculated the means, standard deviations and summary characteristics of the variables. Through this phase I have taken care of the factors in both the test and the train datasets.

Immediately after the data exploration phase I have conducted a Regression analysis to test my hypothesis for significance of the inspected variables and it turned out that the variables that were not marked as statistically important were:

- Race
- Ethnicity
- County code
- Population
- Gender
- Area

As initially considered only factors that are directly correlated with the lander were important, as race, ethnicity and gender are characteristics of the person, they are not considered to be of a significant importance and are not correlated. Additionally, the area of living is also not a factor. The financial stability and security in family aspect were the factors of bigger importance.

Finally, I predicted the rate_spread variable for each row of the test data set by training a Neural Network Regression model using the inputs in the train_values dataset, and predicting them for the test_values.

Data Exploration Analysis

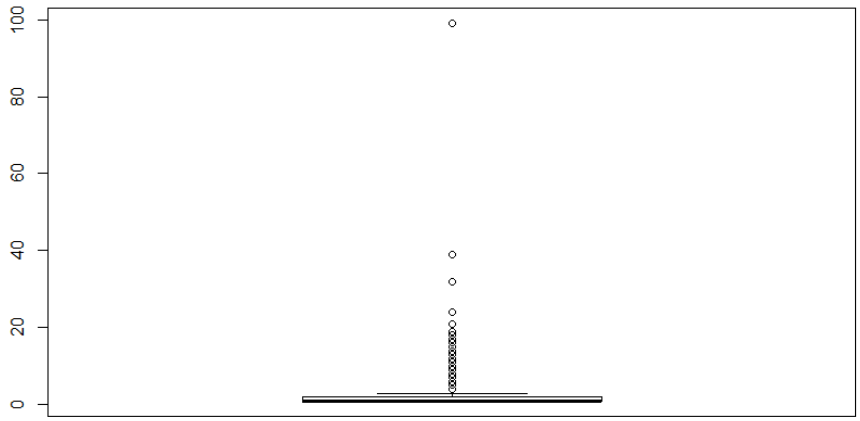
The predictor

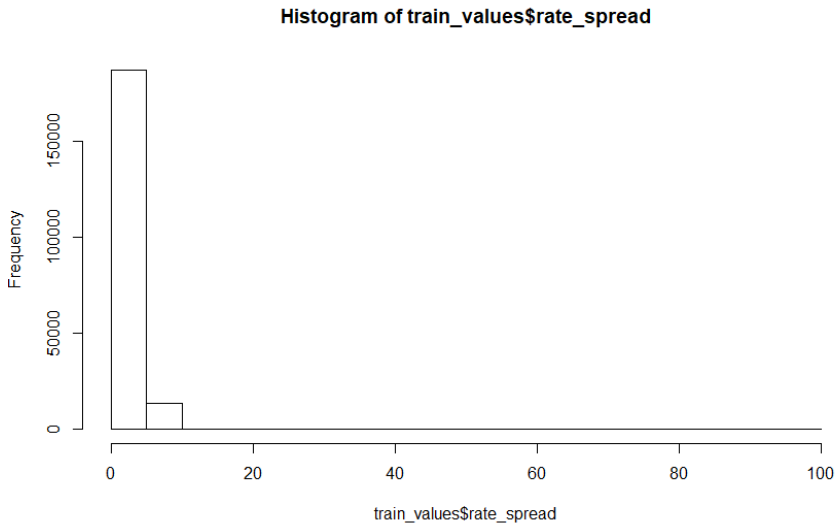
Index and target variables are used as the output variables that need to be submitted. The rate_spread one is to be predicted:

- `row_id` - A unique identifier
- `rate_spread` - Indicates the difference between the offered mortgage rate for the applicant and the standard rate for a comparative mortgage

```
> summary(train_values$rate_spread)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   1.979   2.000  99.000
```

In the below graphs we can see the distribution in boxplot and histogram of the rate_spread variables:





The variables

There are 21 variables in this dataset. Each row in the dataset represents a HMDA-reported loan application, and the dataset we are working with covers one particular year. The unique identifier is the lender variables, which serves for each individual loan-making institution.

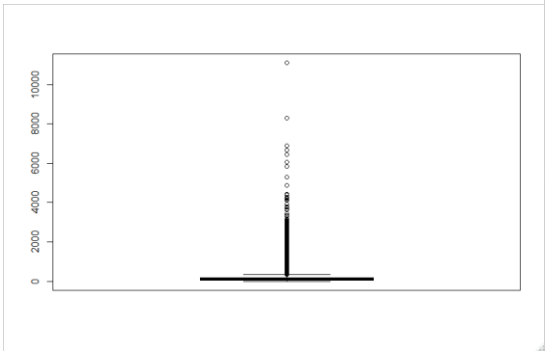
Location variables:

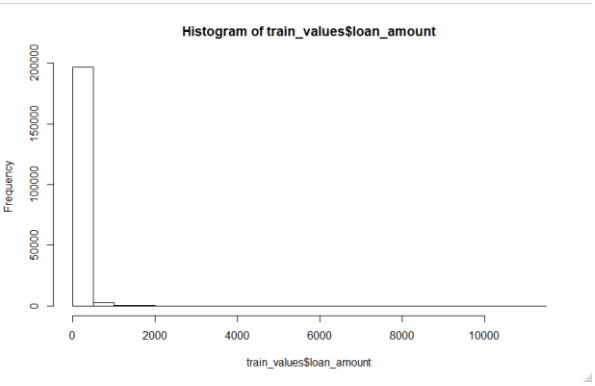
- `msa_md` (categorical) - A categorical with no ordering indicating Metropolitan Statistical Area/Metropolitan Division where a value of `-1` indicates a missing value
- `state_code` (categorical) - A categorical with no ordering indicating the U.S. state where a value of `-1` indicates a missing value
- `county_code` (categorical) - A categorical with no ordering indicating the county where a value of `-1` indicates a missing value

Loan information:

- `lender` (categorical) - A categorical with no ordering indicating which of the lenders was the authority in approving or denying this loan
- `loan_amount` (int) - Size of the requested loan in thousands of dollars

```
> summary(train_values$loan_amount)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    1.0    67.0   116.0   142.6   179.0 11104.0
```



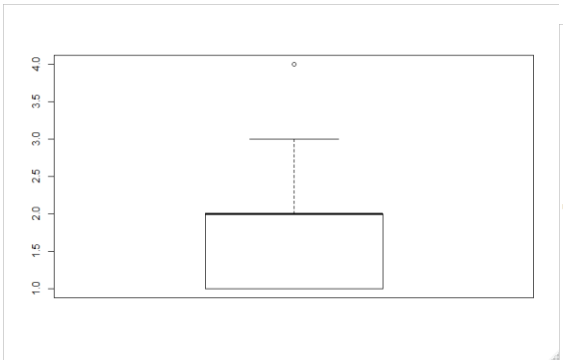


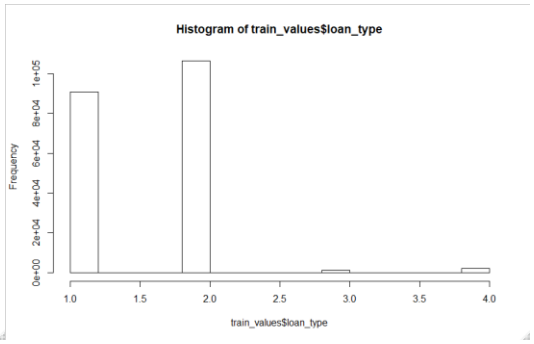
- `loan_type` (categorical) - Indicates whether the loan granted, applied for, or purchased was conventional, government-guaranteed, or government-insured; available values are:

- `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 (Farm Service Agency or Rural Housing Service)

```
> summary(train_values$loan_type)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   2.000   1.571   2.000   4.000
```

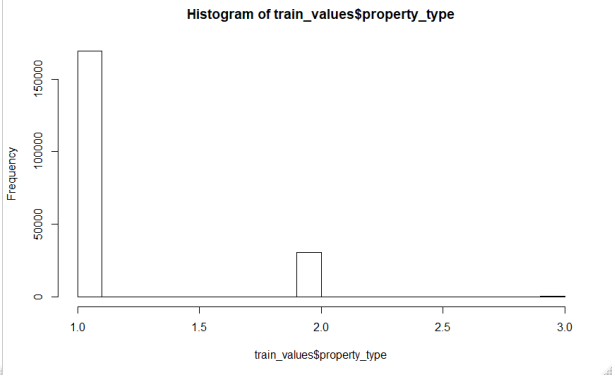




- property_type (categorical) - Indicates whether the loan or application was for a one-to-four-family dwelling (other than manufactured housing), manufactured housing, or multifamily dwelling; available values are:

- 1 -- One to four-family (other than manufactured housing)
- 2 -- Manufactured housing
- 3 -- Multifamily

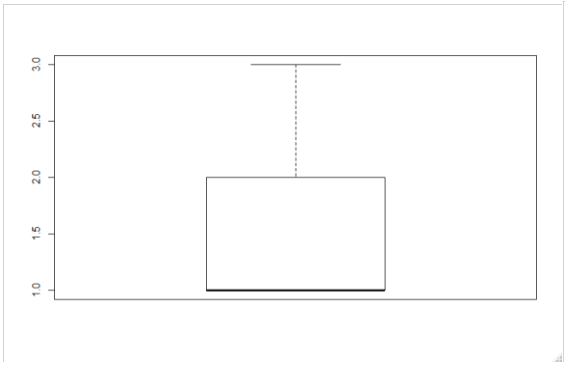
```
> summary(train_values$property_type)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   1.155   1.000   3.000
```

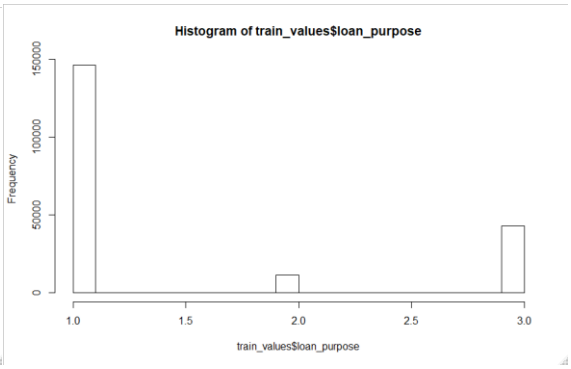


- loan_purpose (categorical) - Indicates whether the purpose of the loan or application was for home purchase, home improvement, or refinancing; available values are:

- 1 -- Home purchase
- 2 -- Home improvement
- 3 -- Refinancing

```
> summary(train_values$loan_purpose)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   1.483   2.000   3.000
```

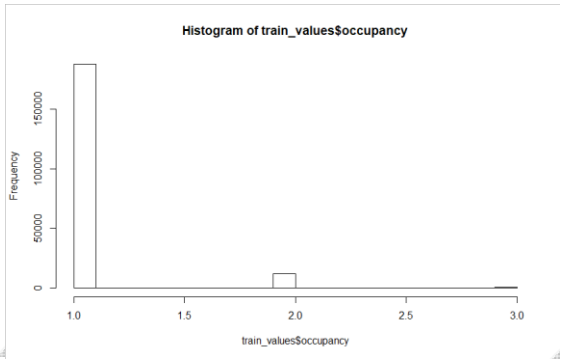




- occupancy (categorical) - Indicates whether the property to which the loan application relates will be the owner's principal dwelling; available values are:

- 1 -- Owner-occupied as a principal dwelling
- 2 -- Not owner-occupied
- 3 -- Not applicable

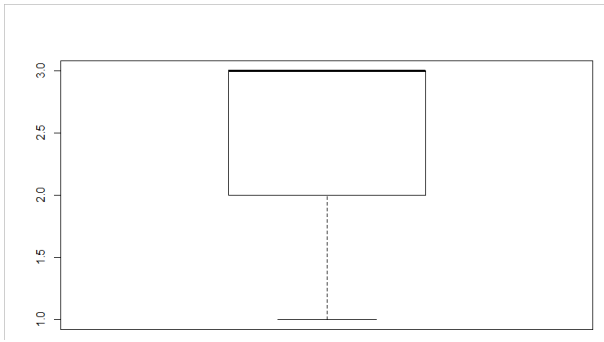
```
> summary(train_values$occupancy)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   1.061   1.000   3.000
```

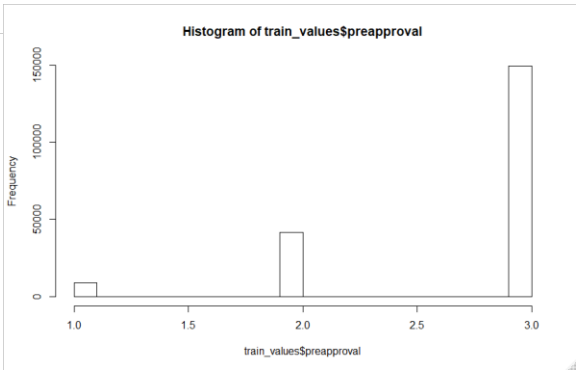


- `preapproval` (categorical) - Indicate whether the application or loan involved a request for a pre-approval of a home purchase loan; available values are:

- `1 -- Preapproval was requested`
- `2 -- Preapproval was not requested`
- `3 -- Not applicable`

```
> summary(train_values$preapproval)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   3.000   2.703   3.000   3.000
```

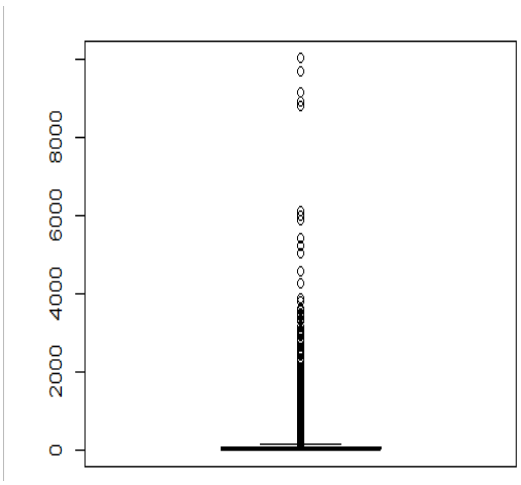


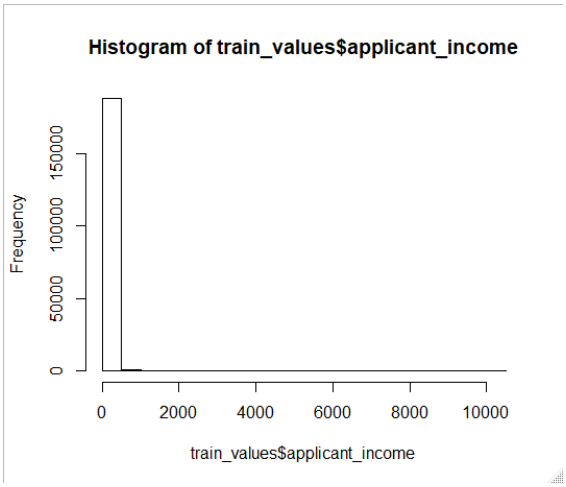


Applicant Information:

- `applicant_income` (int) - In thousands of dollars

```
> summary(train_values$applicant_income)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
    1.00    39.00    56.00    73.62    83.00 10042.00    10708
```





- `applicant_ethnicity` (categorical) - Ethnicity of the applicant; available values are:

```
- 1 -- Hispanic or Latino
- 2 -- Not Hispanic or Latino
- 3 -- Information not provided by applicant in mail, Internet, or telephone pplicat
  ion
- 4 -- Not applicable
- 5 -- No co-applicant
```

- `applicant_race` (categorical) - Race of the applicant; available values are:

```
- 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 applicant in mail, Internet, or telephone applica
  tion
- 7 -- Not applicable
- 8 -- No co-applicant
```

- `applicant_sex` (categorical) - Sex of the applicant; available values are:

```
- 1 -- Male
- 2 -- Female
- 3 -- Information not provided by applicant in mail, Internet, or telephone applica
  tion
- 4 or 5 -- Not applicable
```

- `co_applicant` (bool) - Indicates whether there is a co-applicant (often a spouse) or not

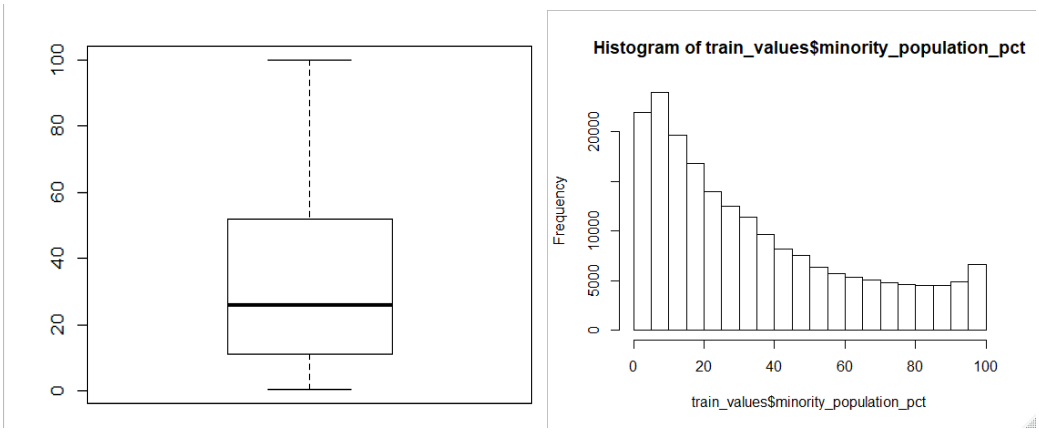
Census information

- `population` - Total population in tract

```
summary(train_values$co_applicant)
   Mode    FALSE     TRUE
logical   123299    76701
```

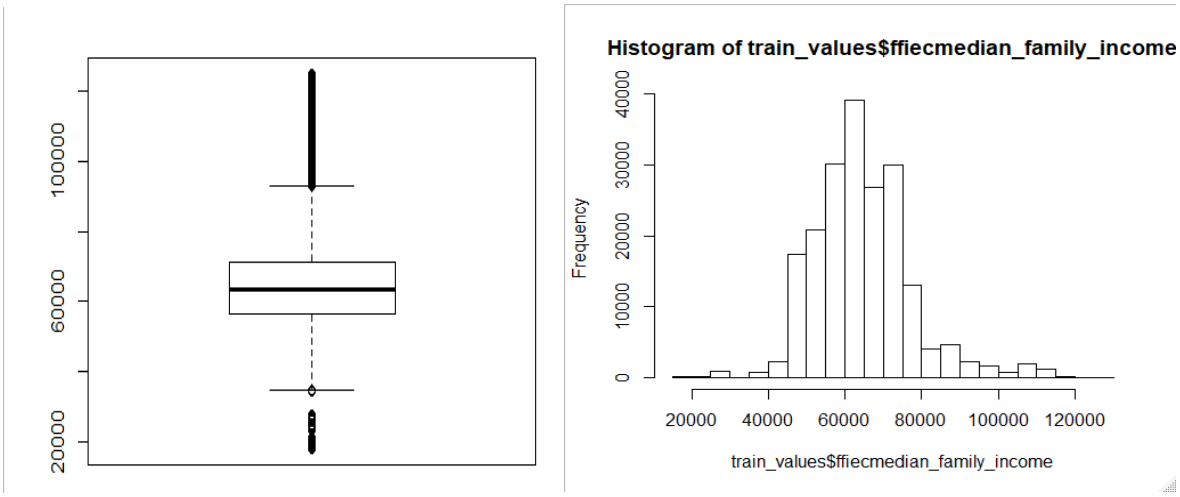
- `minority_population_pct` - Percentage of minority population to total population for tract

```
> summary(train_values$minority_population_pct)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.326  10.928  25.996  34.239  52.000 100.000    1995
```



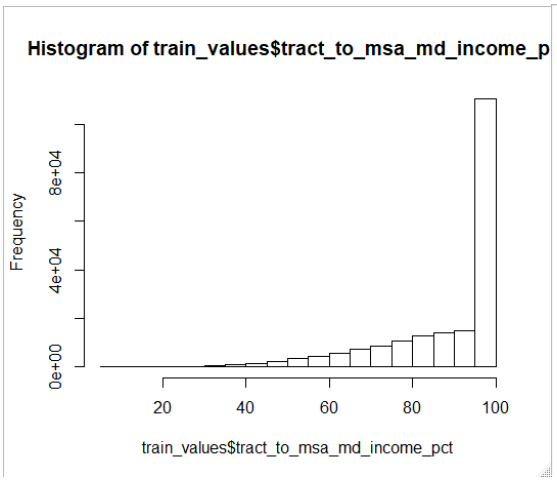
- `ffiecmedian_family_income` - FFIEC Median family income in dollars for the MSA/MD in which the tract is located (adjusted annually by FFIEC)

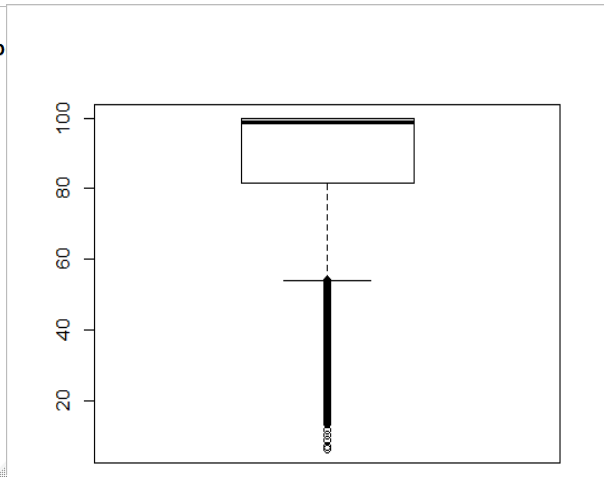
```
> summary(train_values$ffiecmedian_family_income)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  17860   56654   63485   64595   71238  125095    1985
```



- `tract_to_msa_md_income_pct` - % of tract median family income compared to MSA/MD median family income

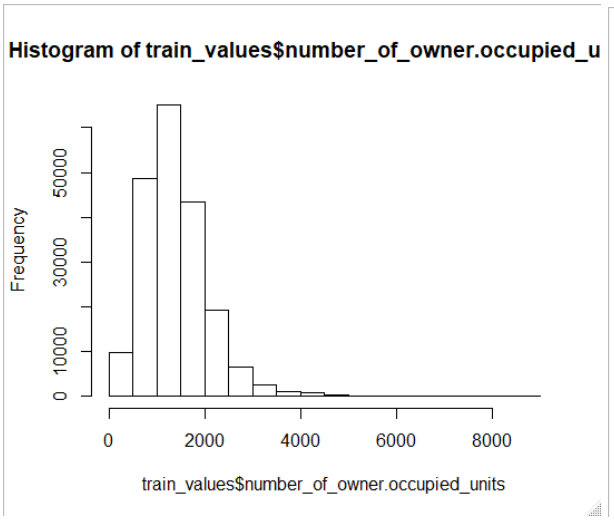
```
> summary(train_values$tract_to_msa_md_income_pct)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  6.193  81.648  98.959  89.283 100.000 100.000    2023
```

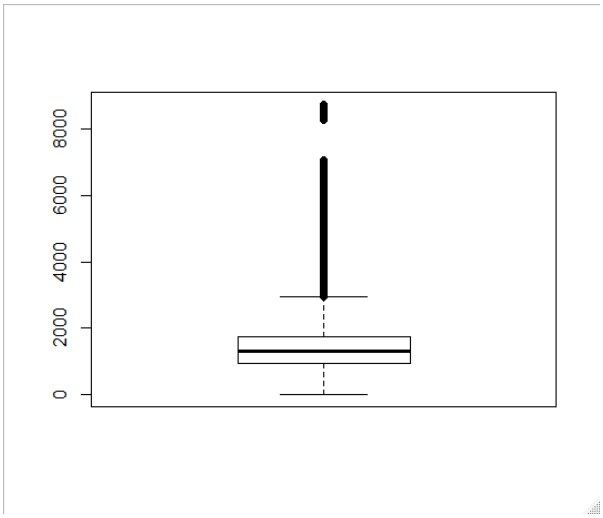




- `number_of_owner-occupied_units` - Number of dwellings, including individual condominiums, that are lived in by the owner

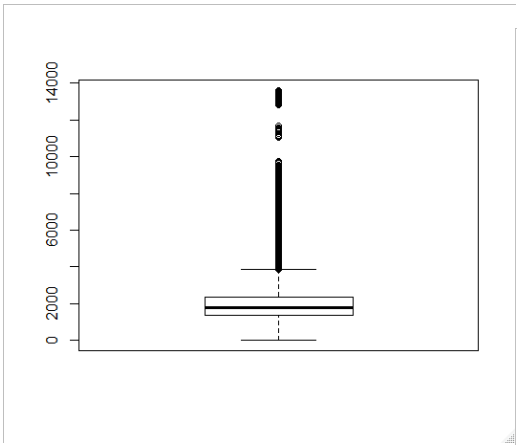
```
> summary(train_values$number_of_owner.occupied_units)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
      3     932    1304    1403    1742    8747    2012
```

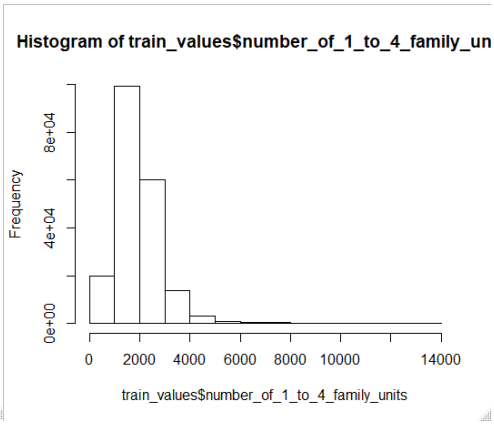




- `number_of_1_to_4_family_units` - Dwellings that are built to house fewer than 5 families

```
> summary(train_values$number_of_1_to_4_family_units)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
      6    1344    1799    1927    2353   13615    2016
```





Then I have continued to the Data Cleansing part in which I have cleaned the outliers that have been identified, and cleared the NAs, and conducted Correlation Analysis.

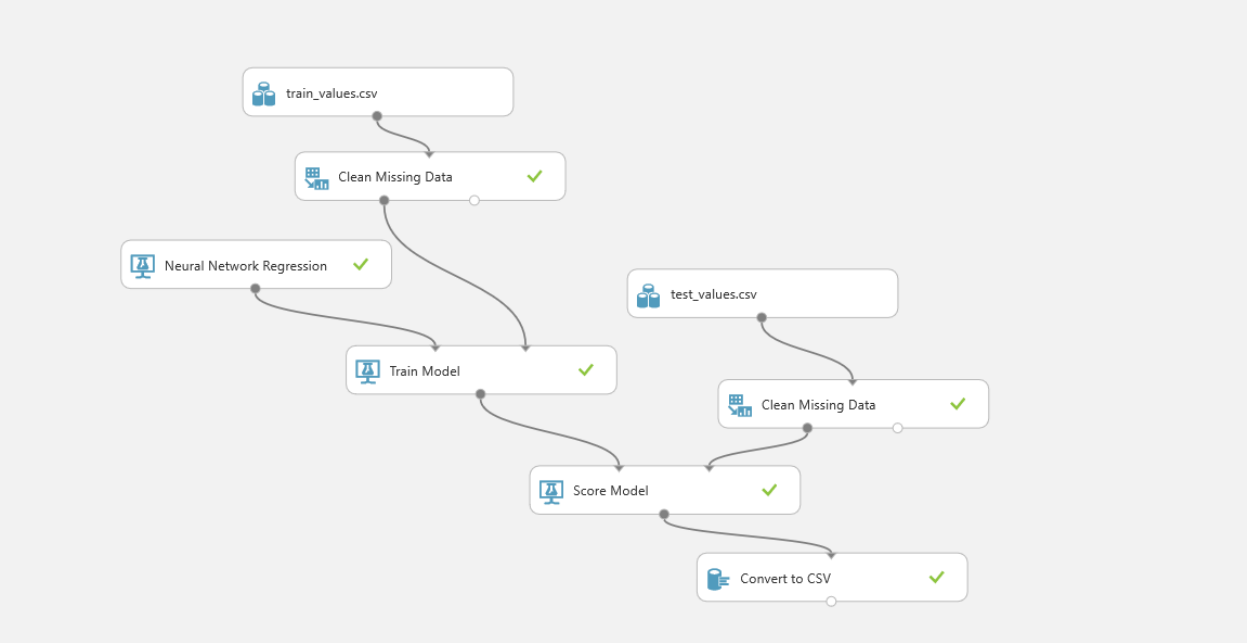
Furthermore, I have conducted a linear regression model to see the importance of the factors:

```
                                          t value Pr(>|t|)
(Intercept)                                 7.679 1.61e-14 ***
train_values$loan_type                   -115.448  < 2e-16 ***
train_values$property_type                196.709  < 2e-16 ***
train_values$loan_purpose                  -9.795  < 2e-16 ***
train_values$occupancy                    -33.852  < 2e-16 ***
train_values$loan_amount                  -77.021  < 2e-16 ***
train_values$preapproval                   38.816  < 2e-16 ***
train_values$msa_md                        -4.213 2.52e-05 ***
train_values$state_code                    10.238  < 2e-16 ***
train_values$county_code                    0.578 0.563254
train_values$applicant_ethnicity           22.637  < 2e-16 ***
train_values$applicant_race                -9.961  < 2e-16 ***
train_values$applicant_sex                 -0.054 0.957207
train_values$applicant_income              29.033  < 2e-16 ***
train_values$population                     3.350 0.000810 ***
train_values$minority_population_pct       16.750  < 2e-16 ***
train_values$ffiecmedian_family_income     13.682  < 2e-16 ***
train_values$tract_to_msa_md_income_pct     9.514  < 2e-16 ***
train_values$number_of_1_to_4_family_units -7.627 2.41e-14 ***
train_values$number_of_owner.occupied_units 3.392 0.000694 ***
train_values$co_applicantTRUE               8.884  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Modelling, Evaluation and Deployment

We're predicting a float value, so this is a regression problem. I have used a Neural Network Regression in order to do the prediction accurately.

Neural networks are reducible to regression models—a neural network can "pretend" to be any type of regression model. It takes the dependent variables = input parameters, multiplies them by their coefficients = weights, and runs them through a sigmoid activation function and a unit step function, which closely resembles the logistic regression function with its error term.



The algorithm runs in Azure ML Studio, where I have preloaded the two cleansed datasets, and have applied once again the double-cleaning procedures. Additionally, each of them has been assigned to the dataset division – Train Model to train the algorithm against e Neural Network Regression and the Test Value as the second value of the output of the scoring of the model.

The result is contained in the scoring of the model, where different weights have been assigned to the different rows, and the result has been produced with pretty high accuracy and R-squared of 0.5296.

Finally, the scored model has been extracted as CSV and loaded.

Setting	Value
Is Initialized From String	False
Is Classification	False
Initial Weights Diameter	0.1
Learning Rate	0.005
Loss Function	CrossEntropy
Momentum	0
Neural Network Definition	
Data Normalizer Type	MinMax
Number Of Input Features	22
Number Of Hidden Nodes	System.Collections.Generic.List`1[System.Int32]
Number Of Iterations	100
Number Of Output Classes	1
Shuffle	True
Allow Unknown Levels	True
Random Number Seed	