WASHINGTON STATE HOME LOANS 2016

1 Problem Statement

Factors affecting loan approval decisions?
Is there any area, demographic or gender bias?
Current trends?

2 Data Wrangling Report:

2.1 Dataset source

Washington state Home Loans dataset is obtained from Kaggle in CSV format.

https://www.kaggle.com/miker400/washington-state-home-mortgage-hdma2016

2.2 Dataset details –

Dataset contains following 47 columns and 466,566 records

tract_to_msamd_income
rate_spread
population
minority_population
number_of_owner_occupied_units
number_of_1_to_4_family_units
loan_amount_000s
hud_median_family_income
applicant_income_000s
state_name
state_abbr
sequence_number
respondent_id
purchaser_type_name
property_type_name
preapproval_name
owner_occupancy_name
msamd_name
loan_type_name
loan_purpose_name
lien_status_name
hoepa_status_name
edit_status_name
denial_reason_name_3
denial_reason_name_2
denial_reason_name_1
county_name
co_applicant_sex_name
co_applicant_race_name_5
co_applicant_race_name_4
co_applicant_race_name_3
co_applicant_race_name_2
co_applicant_race_name_1

co_applicant_ethnicity_name
census_tract_number
as_of_year
application_date_indicator
applicant_sex_name
applicant_race_name_5
applicant_race_name_4
applicant_race_name_3
applicant_race_name_2
applicant_race_name_1
applicant_ethnicity_name
agency_name
agency_abbr
action_taken_name

2.3 Data Cleaning Methods

All the columns in the dataset are imported as object type. After inspecting data following observations are found

- Some columns contain very few non-null values
  - denial_reason_name_3 1246 non-null object
  - denial_reason_name_2 6746 non-null object
  - denial_reason_name_1 34499 non-null object
  - co_applicant_race_name_5 14 non-null object
  - co_applicant_race_name_4 21 non-null object
  - co_applicant_race_name_3 105 non-null object
  - co_applicant_race_name_2 1862 non-null object
  - applicant_race_name_5 46 non-null object
  - applicant_race_name_4 68 non-null object
  - applicant_race_name_3 297 non-null object
  - applicant_race_name_2 4478 non-null object
- Some columns missing (NaN) value
- Duplicate columns
  - “applicant_race_name_1” duplicate of “applicant_ethnicity_name”
  - “co_applicant_race_name_1” duplicate of “co_applicant_ethnicity_name”
  - “agency_abbr” duplicate of “agency_name”
  - “state_abbr” duplicate of “state_name”
- Columns with no significant information for statistical analysis
  - sequence_number
  - respondent_id

2.4 Drop Columns

Columns with few data elements are dropped. Following Data columns are dropped.

denial_reason_name_3
denial_reason_name_2
denial_reason_name_1
co_applicant_race_name_5
co_applicant_race_name_4

co_applicant_race_name_3
co_applicant_race_name_2
co_applicant_race_name_1
applicant_race_name_5
applicant_race_name_4
applicant_race_name_3
applicant_race_name_2
applicant_race_name_1
agency_abbr
state_abbr
sequence_number
respondent_id

2.5 Column Data Types

Data type of all the columns with datatype "object" is changed to "category".

2.6 Missing Data

2.6.1 Drop missing rows

Rows in column "**tract_to_msamd_income**" with NaN are dropped, since the missing values are less than 20% of the data. Dropping of NaN values from "**tract_to_msamd_income**", also eliminated missing values from "**population**", "**minority_population**", "**number_of_1_to_4_family_units**", "**loan_amount_000s**" and "**hud_median_family_income**".

2.6.2 Impute missing values

Impute "number_of_owner_occupied_units" and "applicant_income_000s" with median value.

Substitute missing values in msamd_name with value based on value in census_tract_number

2.7 Outliers

Outliers are visually inspected with box plots. Outliers from the following columns are filtered

Column Name
tract_to_msamd_income
population
minority_population
number_of_owner_occupied_units
number_of_1_to_4_family_units
loan_amount_000s
applicant_income_000s

2.8 Duplicate Rows

Remove duplicate rows if any.

2.9 Save data

Save cleaned data to csv for further analysis.

3 Exploratory Data Analysis

3.1 Summary of Findings

3.1.1 Trend Analysis

Seattle-Bellevue-Everett Metropolitan area trends-

1. Has largest population and highest income
2. Most number of applications
3. Among the lowest 1 to 4 family homes.
4. Most of the property are owner occupied

General trends –

1. Most of the loans applications are either for Home purchase or Refinancing
2. Most of the loan applications are conventional loans
3. Most of the loan applicants are Male

Action taken trends –

1. Most of the loan applications are accepted, there seems to be a good acceptance rate.
2. Loan applicants in each action taken category is “Not Hispanic or Latino”
3. Loan applicants in each action taken category is “Male”
4. Seattle-Bellevue-Everett Metropolitan area has the highest loan approval this could because of hi ghest number of loan applications.
5. Most of the loans applications accepted are either for Home purchase or Refinancing
6. Conventional loans types are most accepted.

Distribution of loan amounts are between 150K to 300K across all the Metropolitan areas, loan purpose and action taken.

Please see a detailed report at this link.

https://github.com/khmsha/Springboard/blob/master/Capstone_Project1/Data%20Story/WashingtonHomeLoans%20-Data%20Story.ipynb

3.1.2 Gender Bias

A hypothesis test was conducted to determine Gender bias.

H0: $\boldsymbol{\rho} = 0$ There is no gender bias.
H1: $\boldsymbol{\rho} \neq 0$ There is a gender bias.

A t-statistic and P-value was calculated as per of hypothesis test and determined that there is a gender bias in loan approval.

3.1.3 Racial Bias

Same procedure followed for determining gender bias was followed for determining racial bias.
A Hypothesis test was defined to evaluate if there is racial bias between ‘Not Hispanic or Latino’ and ‘Hispanic or Latnio’.

H0: $\boldsymbol{\rho} = 0$ There is no gender bias.
H1: $\boldsymbol{\rho} \neq 0$ There is a gender bias.

After calculating t-statistic and p-value it was determined that there is no racial bias in loan approval.

3.1.4 Correlation Between Loan Amount and Applicant Income

Pearson correlation coefficient was calculated using np.corrcoef() and a hypothesis test was calculated to test the correlation coefficient.

H0: $\rho = 0$ There is no correlation loan amount and applicant income.
H1: $\rho \neq 0$ There is a correlation loan amount and applicant income.

A t-statistic, confidence interval and P-value was calculated and determined that there was a low chance of getting the observed correlation coefficient.

From the statistical calculation there is no significance between loan amount and applicant income however scatter plot does show some positive correlation between loan amount and applicant income.

Please see a detailed statistical report at this link.

https://github.com/khmsha/Springboard/blob/master/Capstone_Project1/EDA/WashingtonHomeLoans%20-%20EDA.ipynb