WASHINGTON STATE HOME LOANS 2016

Data Wrangling Report:

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

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

1. 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. 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

2.1 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

2.2 Column Data Types

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

2.3 Missing Data

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**”.

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**

2.4 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

2.5 Duplicate Rows

Remove duplicate rows if any.

2.6 Save data

Save cleaned data to csv for further analysis.