Home Mortgage Disclosure Act (HMDA) 2017 Alaska Mortgage Analysis - Written Summary

Overview

This project implements a complete, reproducible data workflow to analyze the Home Mortgage Disclosure Act (HMDA) 2017 mortgage dataset for Alaska, focusing on first lien, owner-occupied, 1-4 family homes. The analysis examines relationships between applicant income, loan amounts, and various demographic and loan characteristics using Python data science libraries including Pandas, NumPy, Matplotlib, and Seaborn. The dataset was obtained from the Consumer Financial Protection Bureau (CFPB) Historic HMDA Data Portal (https://www.consumerfinance.gov/data-research/hmda/historic-data/). The workflow demonstrates fundamental data science practices including data ingestion, cleaning, exploratory analysis, and visualization.

Dataset Description

The HMDA dataset contains mortgage application and origination records collected under the Home Mortgage Disclosure Act, which requires financial institutions to maintain and report data about mortgages to help identify possible discriminatory lending patterns (Consumer Financial Protection Bureau, 2017). The Alaska 2017 dataset initially contained 12,579 mortgage records across 78 variables including borrower demographics (income, ethnicity, race, sex), loan characteristics (amount, type, purpose), property information, and geographic data at the county and census tract level.

The dataset uses standardized column naming conventions with descriptive field names and coded categorical values. For example, columns containing numeric values have the `_000s` suffix (e.g., `applicant_income_000s`, `loan_amount_000s`) to indicate thousands of dollars, while categorical fields use both numeric codes and plain language labels. Numeric codes represent specific categories: loan purpose (1=Home purchase, 2=Home improvement, 3=Refinancing), loan type (1=Conventional, 2=FHA-insured, 3=VA-guaranteed, 4=FSA/RHS-guaranteed), and action taken (1=Loan originated, 2=Application approved but not accepted, 3=Application denied). The HMDA historic data dictionaries provide comprehensive documentation of all field definitions and code values (Consumer Financial Protection Bureau, 2017).

Key variables analyzed include applicant income (in thousands), loan amount (in thousands), loan purpose (home purchase, refinancing, home improvement), loan type (Conventional, FHA-insured, VA-guaranteed, FSA/RHS-guaranteed), and geographic identifiers (county name, MSA/MD designation). The dataset provides comprehensive information about the mortgage lending landscape in Alaska for 2017, enabling analysis of lending patterns across different borrower profiles and geographic regions.

Workflow Description

Ingestion

The data ingestion process began by downloading the HMDA 2017 Alaska dataset in CSV format with plain language labels and HMDA codes from the CFPB Historic Data Portal. The dataset was loaded into a Pandas DataFrame using `pd.read_csv()` with the `low_memory=False` parameter to handle mixed data types efficiently (McKinney, 2017). Initial inspection confirmed successful loading with 12,579 rows and 78 columns, and the first few rows were displayed to verify data structure and content using the `head()` method.

Cleaning

The data cleaning process followed a systematic approach to ensure data quality and prepare the dataset for analysis. First, columns with excessive missing values (95% missing) were identified and removed, reducing the dataset from 78 to 51 columns. This included completely empty columns like `application_date_indicator`, `sequence_number`, and all `denial_reason` fields. Second, duplicate rows were identified using `DataFrame.duplicated()`, revealing 5 exact duplicates that were subsequently removed. Third, a focused subset of 11 analytically relevant columns was selected based on the research question about loan amounts and applicant income. Finally, rows with missing values in critical fields (`applicant_income_000s` and `county_name`) were removed, resulting in a final cleaned dataset of 12,009 records. Categorical variables were converted to the category data type for memory efficiency and proper handling in analysis (VanderPlas, 2016).

Exploratory Analysis

Exploratory data analysis employed both univariate and multivariate techniques to understand the data distribution and relationships. Univariate analysis included computing summary statistics using `describe()` for both numeric and categorical features, revealing median applicant income of $86,000 and median loan amount of $250,000. Correlation analysis using `df.corr()` identified a moderate positive correlation (r=0.53) between applicant income and loan amount. Multivariate analysis used group-by operations to compute mean income and loan amounts across categorical variables including loan purpose, loan type, county, and borrower demographics. Custom functions were developed to summarize extreme values (maximum and minimum) for each grouping variable to identify notable patterns.

Visualizations

Three primary visualization types were created to communicate data patterns effectively. Histograms were generated for the two continuous variables (applicant income and loan amount) using Matplotlib to display their distributions, revealing right-skewed distributions with potential outliers (Hunter, 2007). Pie charts with custom legends were created for categorical variables with fewer categories to show the composition of loan purposes, loan types, and borrower demographics. All visualizations included proper titles, axis labels, and formatting to enhance readability and interpretation.

Summary

The analysis revealed several key patterns in Alaska's 2017 mortgage market. Home purchases dominated the loan purpose category (8,150 of 12,009 records) with the highest average loan amount ($284,000) and largest loan-to-income gap ($182,000). Conventional loans were most common (6,694 records), but VA-guaranteed loans showed the highest average loan amounts ($306,000). Geographic concentration was evident, with Anchorage Municipality accounting for nearly half of all records (5,522). Demographic patterns showed male applicants with higher average incomes ($107,000) and loan amounts ($278,000) compared to female applicants.

Key Decisions and Assumptions

Cleaning Choices

The decision to remove columns with 95% missing data was based on the principle that variables with such extreme missingness provide minimal analytical value and can introduce noise into the analysis. The 95% threshold was selected as a standard cutoff point used in data quality assessment (Tabachnick & Fidell, 2013). For the `msamd_name` variable with ~21% missing values, records were retained because missing values represent non-metropolitan areas, which is meaningful geographic information rather than data quality issues. The removal of only 5 duplicate rows (0.04% of data) had negligible impact on the dataset while ensuring each observation represents a unique mortgage application or origination.

EDA Focus

The exploratory analysis focused on understanding the relationship between applicant income and loan amounts because this relationship is fundamental to understanding lending patterns and affordability. The analysis was structured to examine how this relationship varies across loan purposes and types to identify whether different mortgage products serve different market segments. The inclusion of geographic variables allowed assessment of regional variations in lending patterns, which is relevant for understanding market dynamics and potential geographic disparities in access to credit.

Purpose of Each Visualization

The histograms of applicant income and loan amount were created to visualize the distributions of these key continuous variables, identify the presence of outliers, and understand the typical ranges. The right-skewed distributions indicated that most loans and incomes cluster toward the lower end with some high-value outliers. Pie charts were selected for categorical variables to clearly show the proportional breakdown of categories and identify dominant groups (e.g., Conventional loans, Home purchases). The combination of these visualization types provides both distributional understanding and categorical composition insights necessary for comprehensive data understanding.

Results and Interpretation

The analysis revealed a moderate positive correlation (r=0.53) between applicant income and loan amount, indicating that higher-income applicants generally receive larger loans, though the relationship is not deterministic. This aligns with lending standards that consider income in loan qualification. The average loan-to-income ratio varied significantly by loan purpose: home purchases showed the largest gap ($182,000), while home improvements showed the smallest ($124,000), reflecting the different purposes and property values associated with each loan type.

Loan type analysis revealed that VA-guaranteed loans had the highest average loan amounts ($306,000) despite borrowers having lower average incomes ($98,000) compared to Conventional loan borrowers ($111,000). This pattern likely reflects the zero or low down payment feature of VA loans, which allows qualified veterans to purchase higher-priced properties. Geographic concentration in Anchorage Municipality (46% of records) reflects Alaska's population distribution, as Anchorage is the state's largest city.

Demographic patterns showed gender disparities, with male applicants having higher average incomes ($107,000 vs. $91,000) and loan amounts ($278,000 vs. $247,000) compared to female applicants. These differences could reflect broader economic patterns, joint applications where male income is recorded as primary, or other factors that would require additional analysis to fully understand.

Responsible Practice (Bias & Data Quality)

Potential Sources of Bias

Several potential sources of bias exist in the HMDA data and this analysis. Selection bias may be present because the dataset only includes mortgage applications that reached financial institutions, excluding potential borrowers who did not apply due to perceived ineligibility or other barriers (Bhutta & Ringo, 2014). Missing data on applicant income (526 records, 4.2%) could introduce bias if missingness is related to factors like loan denial or applicant characteristics. The demographic variables show high rates of "Information not provided" categories, which could mask important patterns if refusal to provide information is systematically related to loan outcomes or borrower characteristics.

Geographic bias is evident in the concentration of lending in Anchorage, which could reflect both population distribution and potential differences in lending practices or access to credit between urban and rural areas. The dataset's focus on approved and originated loans (rather than all applications) means denied applications are underrepresented, limiting analysis of potential discriminatory patterns in loan approval.

Mitigation Strategies

This analysis mitigated bias through transparent documentation of data quality issues and limitations. Records with missing critical variables were removed rather than imputed, preventing artificial pattern creation while acknowledging the reduction in sample size. The analysis examined multiple demographic and geographic dimensions to identify potential disparities that could warrant further investigation. All data cleaning decisions and their rationale were documented to enable transparency and reproducibility.

Future work could better address bias by including denied applications to examine approval patterns, analyzing change over time to identify trends, and using statistical methods like regression analysis to control for confounding variables when examining demographic disparities. Additionally, qualitative research with lenders and borrowers could provide context for quantitative patterns, and comparison with other states could identify Alaska-specific patterns versus national trends.

Reproducibility

This analysis was designed with reproducibility as a core principle. The complete workflow is documented in a Jupyter Notebook (`data_workflow.ipynb`) that can be executed sequentially to reproduce all results. All dependencies are specified in `requirements.txt` generated using `pip freeze`, allowing exact recreation of the Python environment using `pip install -r requirements.txt` (Python Packaging Authority, 2023).

The project uses Python 3.13 as specified in the README. All data transformations are implemented as reusable functions with docstrings explaining parameters, return values, and behavior. The original dataset is preserved in its downloaded form, and all transformations are applied programmatically rather than manually, ensuring the analysis can be rerun on the same or updated data.

Version control using Git tracks all project files and changes, with the repository hosted on GitHub for accessibility. The README provides step-by-step instructions for running the analysis, including how to install dependencies, launch Jupyter, and execute the notebook. This multi-layered approach to reproducibility (code, environment, documentation, version control) ensures that other researchers or practitioners can verify and build upon this work (Wilson et al., 2017).

Sources and Citations

Official Documentation

Consumer Financial Protection Bureau. (2017). *HMDA Historic Data.* Retrieved from https://www.consumerfinance.gov/data-research/hmda/historic-data/

Consumer Financial Protection Bureau. (2017). *HMDA LAR Record Codes.* Retrieved from https://files.consumerfinance.gov/hmda-historic-data-dictionaries/lar_record_codes.pdf

Consumer Financial Protection Bureau. (2017). *HMDA LAR Record Format.* Retrieved from https://files.consumerfinance.gov/hmda-historic-data-dictionaries/lar_record_format.pdf

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90-95. https://doi.org/10.1109/MCSE.2007.55

McKinney, W. (2017). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython* (2nd ed.). O'Reilly Media. ISBN: 978-1491957660. Retrieved from https://www.oreilly.com/library/view/ for-data/9781491957653/

Python Packaging Authority. (2023). *pip documentation.* Retrieved from https://pip.pypa.io/

VanderPlas, J. (2016). *Python Data Science Handbook: Essential Tools for Working with Data.* O'Reilly Media. ISBN: 978-1491912058. Retrieved from https://www.oreilly.com/library/view/python-data-science/9781491912126/

Academic and Research Sources

Bhutta, N., & Ringo, D. R. (2014). The 2013 Home Mortgage Disclosure Act data. *Federal Reserve Bulletin*, 100(6), 1-43. Retrieved from https://www.federalreserve.gov/pubs/bulletin/2014/pdf/2013_HMDA.pdf

Tabachnick, B. G., & Fidell, L. S. (2013). *Using Multivariate Statistics* (6th ed.). Pearson. ISBN: 978-0205849574. Retrieved from https://api.pageplace.de/preview/DT0400.9781292034546_A24616694/preview-9781292034546_A24616694.pdf

Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., & Teal, T. K. (2017). Good enough practices in scientific computing. *PLOS Computational Biology*, 13(6), e1005510. https://doi.org/10.1371/journal.pcbi.1005510