**Publicly Available Data Project**

Garret Nunez

Colorado State University Global

MIS 581: Business Intelligence and Data Analytics

Dr. Jamia Mills

09/04/2022

**Publicly Available Data Project**

Abstract:

The data collection process is, at its root, a lengthy and time-consuming process that often requires a number of resources to complete thoroughly; however, data itself is the lifeline of any good research project and is tantamount only to the conclusion on how to best use the knowledge gained from said project. Due to its difficulty in acquiring, but also its necessity in having, the use of publicly available data is often in contention for acquisition when a research project is in its infancy. This fact is no different in the realm of home loans, which itself is a vastly regulated marketing – making public data (and lots of it) readily available to any consumer looking to discover patterns/trends and even predict future ones. The question at hand is whether or not the use of publicly available data is ethical in nature, and how big of an advance is it really when compared to internal data acquisitions. We are all aware that nearly every corporation collects and uses key pieces of user & sales data, and with the technology bubble ever-expanding the race to be “the next big thing” is a constant, and daily, struggle for nearly any company looking to get ahead. The home loans industry is no different, and the following paper will not only show the significance and implications that publicly available data has on the home loans vertical, but will also result in a ruling of contention more akin to willful, general use and admiration for public data.

Within the last two decades the home loans industry has seen a surge in demand for single-family homes, so sourcing and analyzing meaningful data has become a key priority for most financial institutions that offer this type of loan service. This is due to the fact that lenders that provide home loans want to provide said loan to a borrower that, simply put, going to pay back to the loan without defaulting or declaring bankruptcy – this is where the quality of data, and the effectiveness of the analysis process, can make or break the potential financial risk that the lender is foregoing when funding the home loan. One method for analyzing the potential risk of a home loan is loan prediction. Loan prediction processes the likelihood that a loan will be paid back in full (plus interest, etc.) by the borrower by taking in several variables and matching those results against a predetermined index of successful versus unsuccessful borrowers. While loan prediction is not a guarantee of a borrower’s worthiness to pay back the loan it is a metric for judgement based off of learning history that is constantly being collected, analyzed, and updated within the financial institution. For example, a borrower that is living paycheck-to-paycheck but makes all of their mortgage payments on-time will have a far greater loan worthiness score than a borrower with finances to spare but misses payments often due to traveling abroad. The purpose of this research project will be to take publicly available data and clean it accord to industry-level standards, and then perform predictive analysis using the proper tools to see how effectively historic data can be in creating future use-case predictions within the home loans industry. Addressing this problem will help the organization drastically cut costs by obtaining legally sound and free public data, as opposed to needing to increase cots by generating the data internally. In addition, using simple means of analysis can create a “good enough” result when compared to other methods on the basis of a cost-benefit analysis - i.e., if a simple analysis can increase business by 11% and a complex analysis can increase business by 15% but the simple analysis costs half the resources then it may be feasible for a business to employ simple analysis at certain points in the cycle to increase business while keeping costs low.

Comparing the variables tied to each of these hypothetical borrowers would undoubtedly produce a prediction that heavily favored the globe trotter; however, the loan prediction process is meant to assist a lender in matching a borrower’s variable data against those of the average, successful borrower and is used as one of many metrics when approving or denying a loan request. Some of the variables to be included in this research project will include borrower income, current loan-to-value, ethnicity, gender, and even unit affordability (whether the family’s income is likely lower or higher than others in the same area). The data to be surveyed will include over 1 million data points of public use data from Fannie Mae and Freddie Mac single-family mortgage owner-occupied single-unit properties in the year 2020. “The datasets supply mortgage lenders, planners, researchers, policymakers, and housing advocates with information concerning the flow of mortgage credit in America’s neighborhoods.” (Public Use Database – Fannie Mae and Freddie Mac, pg. 2)

Regarding the hypothesis for this academic project, we will seek to explore the outcomes derived from the statistical tests as to create the most effective outcome from the dataset. The null hypothesis (H0), as it stands, is that the usage of inexpensive, public data has no effect on the overall effectiveness of the data set when it comes to creating predictive results. The alternative hypothesis (H1), as it stands is that the usage of inexpensive, public data has an effect on the overall effectiveness of the data set when it comes to creating predictive results. When looking at the end results from predictive testing, it is important to note that setting an accepted level of significance for each of the results is crucial as it will determine whether or not the tests actually provided key insight into the data. On average, analysts will judge this “p-value” using a value of 0.05 (which equates to about a 1 in 20 chance). “A key aspect of [this theory] is that only the null-hypothesis is tested, and therefore p-values are meant to be used in a graded manner to decide whether the evidence is worth additional investigation and/or replication.” (Pernet, 2016)

When it comes to data forecasting, there is no escaping potential issues that may arise even within a lucrative industry like home loans. Forecasting is not an exact science. Completely predicting the future is not simply something that can be done with some data and basic probability and statistics methods, and may not even be something attainable in our lifetime. Even with solid data and a great forecasting method it is conceivable that some markets simply have a high level of volatility as well as an endless number of variables that influence change. (Galt, 2020) Forecasting can also be a resource-intensive and time-consuming as it requires an exorbitant amount of data and organization to not only bring fresh data in but also clean and analyze it according to the business/project goal(s). Additionally, the data collection and analysis process are typically done manually as automation will not inherently produce desirable results without further investment into algorithms specifically designed to create results. Lastly, the forecasting process is a costly venture from hiring a knowledgeable team to acquiring quality tools and good quality data the upfront cost to even launch a forecasting project can be staggering, especially considering that acquiring good teammates, tools, and data still does not guarantee the desired result from the venture as a whole. (Galt, 2020)

With numerous advancements in the field of data analytics, and an ever-present fear of “falling behind the competition”, the home loans industry has seen its fair share of triumphs via the use of publicly available data and data forecasting. As clearly stated within every publication on the topic, data forecasting is (and never will be) an exact science – regardless as to the types of growth experienced in our lifetime, predicting future trends will not see an equal breaking point. Due to this fact, statistical methods need to be clearly observed and adhered to when it comes to parsing through data sets in order to identify as much potential insight as possible while not letting noise cloud the end results. While forecasting relies on solid probability and statistics to guide the user to their desired result, one other key ingredient is sometimes overlooked and can certainly ruin any well-intentioned project – good data. The problem at-hand, and the topic of discussion, is whether or not publicly-available data can be used to produce beneficial results for a business (in this case, within the home loans industry).

Another problem to face users that try to use public data for forecasting purposes is the flurry of financial elements that play a key role in determining the success or failure of the venture in itself. Organizational mishaps, forecasting inefficiencies, lack of data credibility, and even operational issues are a few of some of the tedious problems that come along with forecasting on a large scale that prevent companies from reaping the benefits of good forecasting. The study conducted by Supriya et al. stated it best when they wrote, “with the enhancements in the banking sector lots of people are applying for bank loans but the bank has its limited assets which it has to grant to limited people only, so finding out to whom the loan can be granted... [requires] to reduce this risk factor behind selecting the safe person so as to save lots of bank efforts and assets.” (Supriya et al. 2019)

The dataset itself was taken from the Federal Housing Finance Agency (FHFA) and contains simplified single-family mortgage-level owner-occupied 1-unit property data collected directly from Fannie Mae and Freddie Mac. As most people know, both Fannie Mae and Freddie Mac are federally back home mortgage companies that were created by Congress in order “to provide liquidity, stability, and affordability to the mortgage market... Fannie Mae and Freddie Mac buy mortgages from lenders and either hold these mortgages in their portfolios or package the loans into mortgage-backed securities that may be sold.” (About Fannie Mae and Freddie Mac, pg.3) Within the dataset are numerous variables and bits that will be helpful for the course of the overall project. Most notably are the following: tract income ratio, borrower income ratio, LTV ratio, federal guarantee, and unit affordability. Tract income ratio refers to the 2010 census tract median income to determine whether the area qualifies as low-income. Borrower income ratio is the ratio of the borrower’s annual income in comparison to the area’s median family income and is used to determine if the borrower qualifies for an income-based mortgage. LTV ratio is simply the loan-to-value ratio that is determined at origination to online how much of the cost of the home needs to be financed in order for the borrower to acquire the home. Federal guarantee is simply an identifier to determine whether the loan that was taken to purchase the home was an FHA/VA loan, or another type of federally guaranteed loan in order, or a conventional loan (i.e., a loan with no federal guarantee). Lastly, unit affordability covers whether the borrower is a low-income family or very low-income family in a low-income area, a very low-income family not in a low-income area, or otherwise.

Closely observing the dataset will provide a number of opportunities to perform industry-level analytics for the purpose of determining whether or not publicly available data has any value within the private sector for predicting future trends. For example, what percentage of conventional loans were completed by very low-income families in a low-income tract? Questions such as this can be answered by cleaning and analyzing the data accordingly. Speaking of, the analysis and comprehension will come in two parts – data analysis and data visualization. In order to accomplish this, SAS will be used for statistical and predictive analysis and R will be used to create data visualizations/representations with the full force of customization packages such as ggplot and seaborn.

As previously mentioned, SAS will be used to complete the predictive analytics portions of the research project by running the dataset through a number of functions to both clean and parse the data according to key specifications. SAS itself offers quite a bit of support when it comes to research data analysis tools and a large advantage is that it provides GUI’s and multi-application access points so that the analysis performed is as thorough as possible. Because of this and more, SAS will be used to break down the data using the following predictive models: glmselect, pls, and transreg. GLMSELECT procedure will be used in order to effectively select from within the generalized linear model – within this procedure a training model is created from the data set to act as a base and a test model is then created in order to determine the predictive performance qualities that the dataset may contain. PLS procedure is used to perform simple components of regression such as enabling the user to choose how many extracted factors from within the data set are then cross validated during analysis. Lastly, TRANSREG procedure fits linear models with the most optimal variables in order to create nonlinear transformations. This procedure is effectively a tester function that will be used to experiment with different designs prior to their actual analytical use.

Another key portion of this research project will be the data visualizations that will be created in order to showcase very descriptive and telling points made by the dataset itself. In order to complete this task, R will be used to handle, store, and analyze the data before adapting it into a graphical model. R’s integrated development environment allows nearly every part of the graphic to be accessible for customization and even includes a package that will help with the proper code execution for a number of customizations. To most effectively convey the message that exemplifies the course of this project, the following visualizations will be used: scatter plot, box plot, and histogram. The scatter plot will be ideal for showcasing how the divide is between the census tract income ration and the borrower income ratio. The box plot will be used to detail out the number of buckets that exact across one million data points in the loan-to-value ratio variable column. A histogram conveying the breakdown of purpose of loan versus federal guarantee should illuminate any reader accordingly.

While the number of variables (columns) within the dataset it greater than what is showcased below, the following data dictionary represents a detail of the types of values and descriptions of the variables that will be effectively used over the course of the research project:

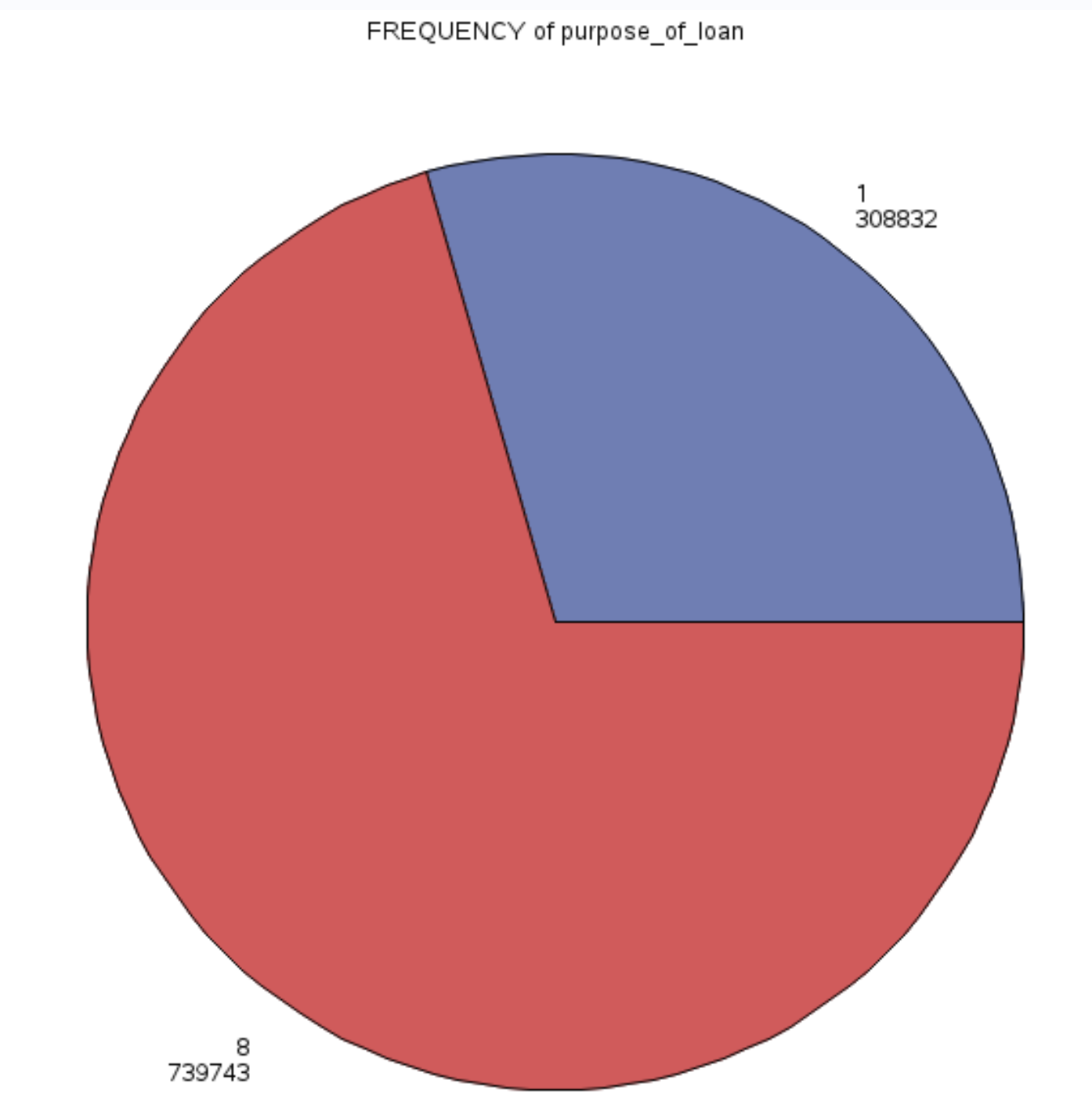
|  |  |  |
| --- | --- | --- |
| Variable Name | Value Range | Description |
| Tract Income Ratio | * 1 = <=80% * 2 = 80% - 120% * 3 = >120% | * Ratio from the 2010 census tract median income to the local area median income |
| Borrower Income Ratio | * 1 = <=50% * 2 = 50% - 80% * 3 = >80% | * Ratio of the borrower’s annual income to the area median family |
| Loan-to-Value Ratio | * 1 = <=60% * 2 = 61% - 80% * 3 = 81% - 90% * 4 = 91% - 95% * 5 = >95% | * Ratio of borrower’s loan amount in conjunction with the amount of income they obtain |
| Purpose of Loan | * 1 = Purchase * 8 = Other * 9 = Not applicable | * Reason for obtaining loan in conjunction with single-family property |
| Federal Guarantee | * 1 = FHA/VA * 2 = RHS * 3 = HECM * 4 = Conventional * 5 = Title 1 | * Type of federally guaranteed loan acquire for acquisition of single-family property |
| Unit Affordability | * 1 = Low-income family in a low-income area * 2 = Very low-income family in a low-income area * 3 = Very low-income family not it a low-income area * 4 = Other | |

As mentioned previously, the organization will benefit from this project by finding an alternative way to engage in data forecasting without having to indulge in the large upfront costs that are usually associated with this type of venture. Freeing up these resources will also benefit the organization by allowing it to divert the funds and personnel to other projects or departments that may also need attention. Allocating resources effectively throughout an organization is necessary, especially when it comes to larger operations, as there may be more than one active project that needs attending to. This also saves money by ensuring that all of the available resources within the organization are being properly utilized when and where they are needed. Allocation also boosts productivity and also improves time management and strategic planning by ensuring that the right personnel and funds are available immediately when deemed necessary, as opposed to having to share these resources thusly stalling the process along the way.

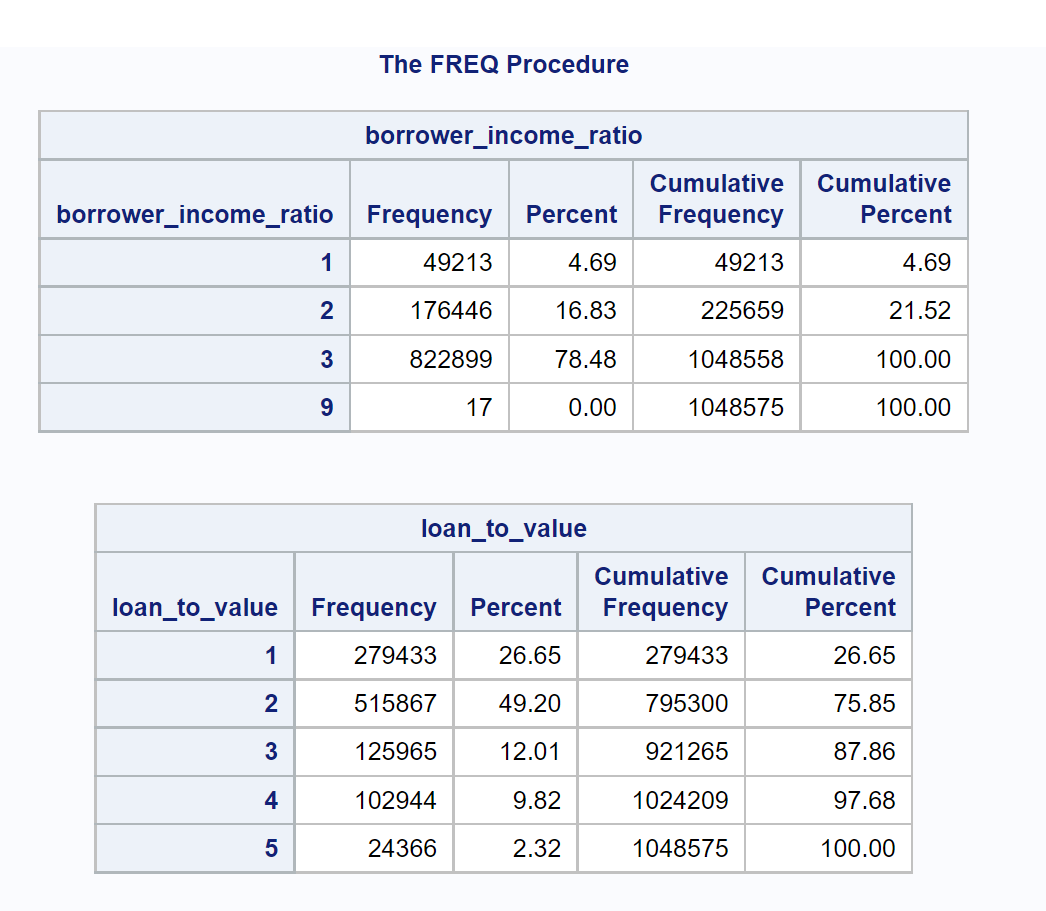
Over the course of the project, we will also take into consideration, and be highly sensitive of, the concerns regarding the efficacy of using publicly available data. As a portion of the analytic community agrees, “the more one knows about people and their condition(s), actions, needs, preferences, beliefs and the like, the better one can provide services to them... however, there are problems or ethical issues with such data, and with capturing, the processes for capturing, and using the data.” (Cooper 2020). The ethical dilemma starts with understanding how and where the data was ascertained, and ensuring that this process was completed in an upstanding manner. The next concern is for what purposes the data is intended to be used for. Lastly, the user should (with no small emphasis) understand how the publication of this data could affect those involved. Ethical concerns surrounding publicly available data will be held in high regard over the course of this project using a mixture of the following approaches. Virtue ethics places an emphasis onto assessing both actions and non-actions accordingly and regardless to preference. Consequentialism is often boiled down to the colloquial phrase of the value of ethical transgressions being justified by the end results – in which a positive net gain, or good results, making the assortation of said results ethically OK. Lastly, deontological ethics emphasizes boiling ethical situations down to a check list of how to respond in/to any given situation. Laws are often considered to be a form of deontological ethics because they clearly set out to list and label the types of actions that can and cannot be taken in most life situations as is presently known. (Cooper 2020)

After thoroughly reviewing the dataset, and passing it through numerous descriptive and analytical functions, we were able to conclude that the null hypothesis was indeed correct and the usage of inexpensive, public data has no effect on the overall effectiveness of the data set when it comes to creating predictive results. The figures below cumulatively show that insight into a problem or situation can be obtained from almost any source as long as the data is thoroughly cleaned and analyzed, and so long as the right questions are being asked. If the expectation of (in this instance) the free data is to provide an enormous amount of insight then the project will fall short; however, if the goal is to better understand the subject, then free data can work just as well as any other ethically-sourced dataset.

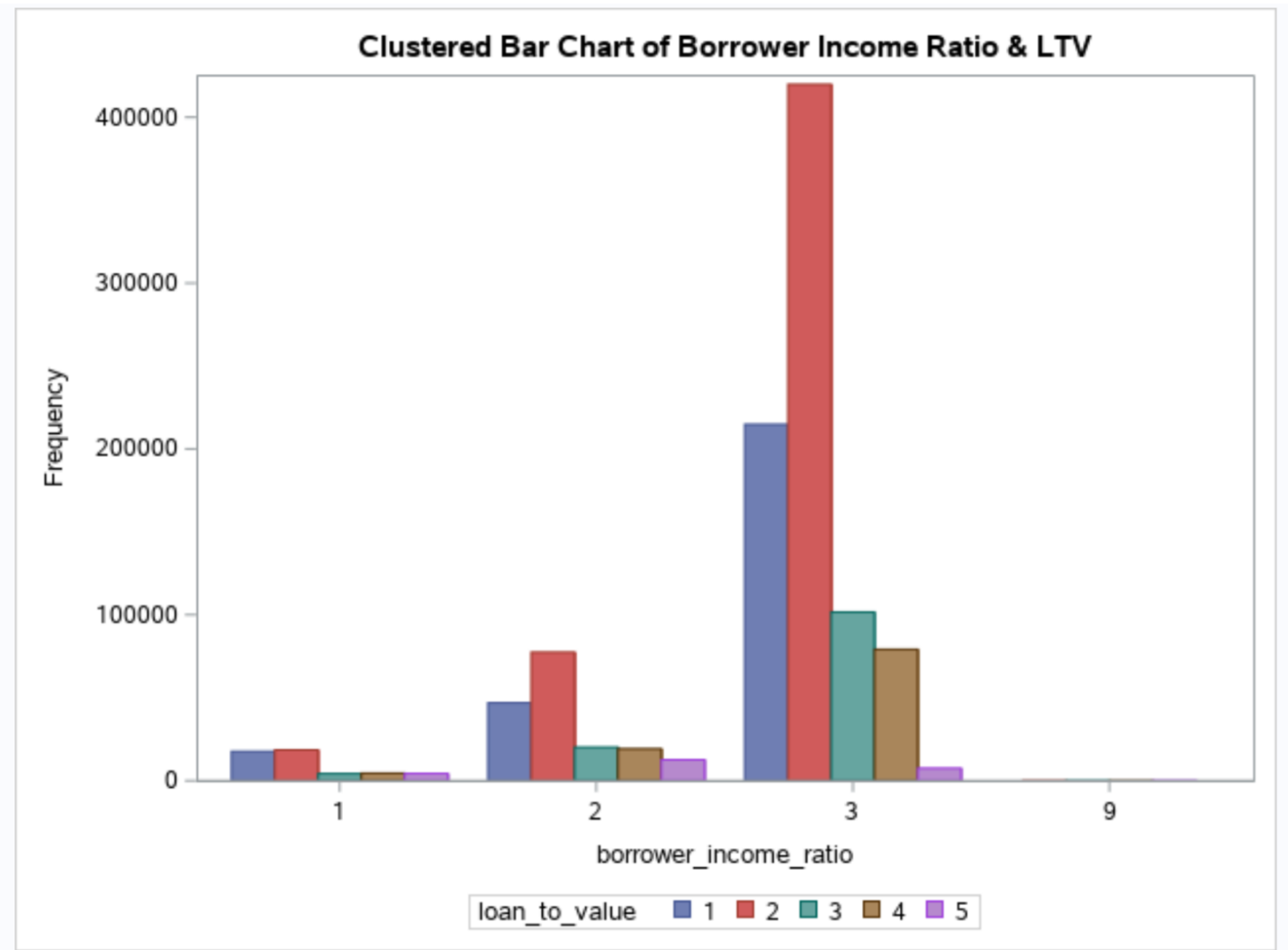
**Figure 1:**

 Figure 1 shows a pie chart that shows the split in data within the purpose\_of\_loan variable. This variable was created to detail out the numerous reasons in which an applicant from this list first sought out a loan in conjunction with a desired single-family property. The category of 1 was a placeholder for borrowers seeking a Purchase loan, and the category of 8 was a placeholder for borrowers seeking a type of loan other than a Purchase loan.

**Figure 2:**

 Figure 2 details out a frequency function called onto the variables borrower\_income\_ratio and loan\_to\_value. The results exemplify a split in the types of income brackets that loan borrowers in this dataset fall into, as well as the buckets of LTV types that are shown. In borrower income ratio, we can see an overwhelming majority of borrowers fall into the 3 bucket which states that over the borrower’s income level is in-line with over 80% of the surrounding areas average income. The loan-to-value frequency table shows a more diverse spread amongst borrowers ranging from less than 60% LTV all the way up to greater than 95% LTV.

**Figure 3:**

 Figure 3 lays out a clustered bar chart that compares the borrower income ratio to the loan-to-value ratio of the average borrower within the publicly available dataset. The data shows that borrowers that fall into the 1st bucket of the income ratio (those whose median income is similar to less than half of the area’s population) tend to have much fewer loans taken out, and have much lower LTV amounts. Where the data becomes interesting is within the 3rd bucket that shows borrower whose median income matches over 80% of the area’s population, and what is shown here is that an overwhelming majority of these borrowers held loans with LTV amounts around 60% but predominantly 61%-80%.

**References**

About Fannie Mae & Freddie Mac. *Federal Housing Finance Agency*, <https://www.fhfa.gov/about-fannie-mae-freddie-mac>

Bhanu, Udaya. Customer Loan Prediction Using Supervised Learning Technique. *International Journal of Scientific and Research Publications*, June 2021, <https://www.ijsrp.org/research-paper-0621/ijsrp-p11453.pdf>

Cooper, Antony K., and Serena Coetzee. On the Ethics of Using Publicly-Available Data. *SpringerLink*, Springer International Publishing, 1 Apr. 2020, <https://link.springer.com/chapter/10.1007/978-3-030-45002-1_14>

Fuster, Andreas, et al. The Role of Technology in Mortgage Lending - Federal Reserve Bank *New York Fed*, Feb. 2018, <https://www.newyorkfed.org/medialibrary/media/research/staff_reports/sr836.pdf->

Galt, John. 3 Advantages and 3 Disadvantages of Forecasting. *JohnGalt*, 2020, <https://johngalt.com/learn/blog/3-advantages-disadvantages-of-forecasting>

Pernet, Cyril. Null Hypothesis Significance Testing: A Short Tutorial. *National Library of Medicine*, 10 Oct. 2016, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5635437/>

Public Use Database - Fannie Mae and Freddie Mac. *Public Use Databases | Federal Housing Finance Agency*, <https://www.fhfa.gov/DataTools/Downloads/Pages/Public-Use-Databases.aspx>

Salvi, Rohan, et al. Home Loan Data Analysis and Visualization. *ResearchGate*, Jan. 2021, <https://www.researchgate.net/publication/354967983_HOME_LOAN_DATA_ANALYSIS_AND_VISUALIZATION>

Sugiura, Lisa, et al. Ethical Challenges in Online Research: Public/Private Perceptions. *Research Ethics*, vol. 13, no. 3-4, 2016, pp. 184–199., <https://doi.org/10.1177/1747016116650720>

Supriya, P., et al. Loan Prediction by Using Machine Learning Models: Semantic Scholar. *Semantic Scholar*, Jan. 2019, <https://www.semanticscholar.org/paper/Loan-Prediction-by-using-Machine-Learning-Models-Supriya-Pavani/54646ad5279f94bb717dd57263da4cf360a9a8c8>