**SAS Enterprise Miner Research Project:**

**Analyzing and Predicting Housing Listings in California**

Group 2: Thomas Greve, Erik Lapszynski, Ariel Stanalonis, & Zachary Vozzo

Dr. Sedat Cevikparmak

MI-353: Statistical Analysis and Predictive Modeling

May 5, 2022

**EXECUTIVE SUMMARY**

Our group explored a dataset of California housing listings with the intention of determining the variables that contribute to the current housing crisis. First, we had to scrub and clean the data of outliers and missing values. Then by utilizing StatExplore and GraphExplore, we determined which variables most significantly contribute to rising housing prices in California. We used regression analysis to determine which variables demonstrated the highest importance in the predictive model. We then applied this predictive model to see if the variables contributing to higher levels of inflation could be identified and correctly predicted through training and testing sets. By utilizing cluster analysis, we identified housing characteristics where inflation is most prevalent. Lastly, we used a decision tree to determine appropriate price ranges for houses based on particular attributes in the housing listings. This will all help diagnose the root causes of the housing crisis. Through these analyses, we found a multitude of acceptable ranges of housing prices based on particular listing characteristics. This will allows consumers to stay informed about the state of the California housing market while helping them make accurate predictions of housing prices to prevent them from getting exploited by sellers.

**PROJECT MOTIVATION AND BACKGROUND**

The problem that we want to explore through data mining is: How can we discover the root causes of the California housing crisis and prevent it from continuing? We utilized a mix of qualitative and quantitative information to help predict the prices of houses on the market in California. Ultimately, we hope to explore if there are any solutions to help prevent homeowners and renters in California from getting exploited by their real estate agents and landlords. California is the most populated state in the United States of America with over 39 million residents. The next closest state is Texas with 29 million residents. In addition, California has 3 of the 10 largest cities in the United States (Los Angeles, San Diego, San Jose). They also have 5 more of the 50 largest cities in the US (San Francisco, Fresno, Sacramento, Long Beach, Oakland). Despite these statistics, population growth has practically come to a standstill in California. With all these large urban areas sprawling with business and life, the California real estate industry is the largest and most chaotic in the country. The only problem is that there are still not nearly enough new properties built to satisfy demand. This shortage constantly drives up costs for renters and homeowners in heavily populated areas. In addition to housing inflation, California has the highest rate of functional poverty and the second-lowest rate of homeownership in the nation. This means indicated that the majority of residents in California are renting their houses on short-term leases because they cannot afford them in the first place. Research and analysis of relevant data that represent why these issues persist may help diagnose potential solutions to the housing crisis.

**DATA DESCRIPTION**

To explore the solution to our problem, we retrieved the following [table](https://www.kaggle.com/yellowj4acket/real-estate-california) from Kaggle to be downloaded as a .csv file. This dataset (“Real Estate California”) shows all real estate listings for the state of California in the first six months of 2021. There are around 35,390 rows or listings and 39 columns, which contain both qualitative and quantitative data for our analysis. Some dimensions within the data include but are not limited to: housing prices, location, area, number of floors, housing type, and number of rooms. A major cause for concern within this dataset was the number of variables that are derivative of each other (FIGURE 1). For example, city, county, longitude, and latitude all represent very similar attributes. In addition, some variables were entirely irrelevant for analysis, such as addresses and other identification codes. To combat this, we rejected certain variables to avoid repeating them in our analysis. We will use these different variables to help us obtain more concise and relevant information for our project. Attached below is a picture of all the dimensions within the dataset, including their type and whether they were utilized. We hope to find relationships between our variables after cleaning and exploring our data, allowing us to accurately predict the prices of houses on the market.

**DATA PREPARATION ACTIVITIES**

The very first data preparation activity that we had to conduct was to assign roles and levels to our variables. Since the data was imported from a .csv file, SAS Enterprise Miner automatically assigned roles and levels to the variables, some of which were entirely inaccurate. We started by selecting the target variable: price. This was the easiest dimension to set because we want to explore which variables will help predict prices amidst the California housing market crisis. Selecting the variables to take the input role was much more difficult. The first issue we had to resolve was the derivative variables. We decided to minimize the number of unique input variables to exclude irrelevant information and avoid double counting. For instance, the country, latitude, longitude, state, streetAddress, and zipcode dimensions were rejected in favor of the city and county variables, which were set as nominal level variables. The next issue we had to resolve was the use of identification variables. We rejected certain identifiers, such as cityID, countyID, stateID, lotAreaUnits, VAR1, and time because they did not add anything meaningful to the analysis of our dataset. We kept the ID and the datePostedString available as the lone identifiers for potential identification purposes. The final variable assignment issue we had to resolve was the application of binary variables. We went through and changed the variable level from nominal to binary for the following variables: hasGarage, hasPetsAllowed, isNewConstruction, is\_bankOwned, is\_forAuction, parking, pool, and spa. The binary level variable hasBadGeocode was rejected because it did not offer anything of value. In total, we ended up with 17 different input variables that we would use for our analysis (FIGURE 2).

Some other problems we ran into were missing or inaccurate values within the dataset. Our next task was to scrub the data using Microsoft Excel to fix some of the underlying issues in the source data. We decided to remove any duplicate values based on the ID column to ensure that listings were not included in the dataset multiple times. We found 4,151 duplicate values and removed them from the dataset so we could focus on unique listings (FIGURE 3). After this, the most prominent issue within the dataset was the number of missing values in the various input columns. Some listings included “0” in place of a missing value, which initially threw off our analysis. So, we went through and searched the Excel file for instances of a 0 for all dimensions (except for the binary inputs) and removed them. After filtering through the various dimensions, we removed over 30,000 rows with missing values in one or more of the input dimensions. With the newly scrubbed dataset, there are only 3,152 unique housing listings. We imported the updated .csv file into SAS Enterprise Miner to continue our analysis.

**ENTERPRISE MINER MODELS**

We compiled several modeling and classification nodes within our SAS Enterprise Miner workflow to analyze our data (FIGURE 4). We first examined the summary data and then constantly compared the accuracy of different models to create the best version of each classification method. We first utilized StatExplore to examine the results of our dataset. The output for the StatExplore node showed the correlation of our inputs to the target variable of price. The results of the StatExplore output show that livingArea (.66), pricePerSquareFoot (.63), and bathrooms (.60) had the highest correlation with the target variable (FIGURE 5). Again, examining the variable worth of our input variables towards the target variable of price, it was found that the top three inputs in terms of variable worth were living area, price per square foot, and bathrooms. These valuations perfectly match up with the correlations. For the target price, living area, price per square foot, and bathrooms are some of the most important variables in indicating the level at which a home is listed for in the California market. One surprising finding from the StatExplore output is that the variable worth and the correlation coefficient of the yearBuilt input was one of the worst predictors of price. In fact, there was virtually no correlation between the two variables (.03). We assumed that new constructions would tend to have a higher asking price, but clearly this is not always the case. Another interesting aspect of the StatExplore output was the summary statistics of the interval inputs. Here, we can view the mean, standard deviation, minimum, median, maximum, skew, and kurtosis of each input and the target variable. By evaluating the means and the standard deviations of our inputs, we may determine which housing listings fall within the acceptable range and which are outliers.

Next, we utilized the GraphExplore node to visually display housing prices compared to their frequencies represented in percentages (FIGURE 6). An important thing to note from the GraphExplore output is the very high standard deviation of the housing prices: the standard deviation is $1.86 million while the mean is only $1.24 million. This information, paired with the skewness and kurtosis coefficients of 9.29 and 138.17, respectively, tells us that the distribution of houses is heavily skewed to the right with a high kurtosis representing a large concentration of prices among few data points. In particular, the $357,666.70 and $681,733.30 price points alone account for over 50% of all housing prices, which leads to a steep decline of values, despite the fact there is a considerable amount of additional price points represented in the graph. Together, these results represented through the GraphExplore node output aided our research by illustrating the most common price ranges of houses in California and some potential outliers.

The introductory analysis of these dimensions within the California housing listings laid the groundwork for further analysis with SAS Enterprise Miner. We then used regression analysis to determine which variables demonstrated the highest importance in the predictive model. We then applied this predictive model to see if the variables contributing to higher levels of inflation could be identified and correctly predicted through training and testing sets. By utilizing cluster analysis, we identified areas where significant levels of inflation are most prevalent. Lastly, we used a decision tree to determine appropriate price ranges for houses based on particular attributes in the housing listings. We will evaluate these results in the next section.

**RESULTS AND FINDINGS**

In the regression analysis (FIGURE 7.2,) our significant variables (with P-value less than α = 0.05) turned out to include bathrooms, city, home type, living area, parking, pool, price per square foot, and spa availability. We can compile a rough estimate of housing prices from this regression output by utilizing the regression equation that is produced. The F-values associated with each statistically significant variable would be multiplied by the listings of each significant characteristic to generate a rough estimate of a price for houses on the market in California. The R-square value of this regression equation is .8139, indicating that 81% of the variation of housing prices in California can be explained by these statistically significant variables. These results were not very surprising, because when choosing a place to live, these significant variables will need to be considered. It also appears that the more expensive homes consisted of high levels of these significant variables from the regression analysis. These statistically significant variables are a basis as to whether or not a home is going to be more expensive or not. However, this does not justify home price equality. So, just because a home has these features, does not mean that its price listing is still justified. This is where the analysis digs down deeper.

As for the cluster analysis, we utilized the centroid clustering method to produce 7 unique clusters of housing characteristics (FIGURE 8). This method of clustering truly produced 4 clusters with reasonable frequencies. We found that within these clusters, the most expensive home characteristics were listings with 4 bedrooms, 3 bathrooms, garage accessibility, and built after 1987. Ultimately, the newer the house with the greater amenities, the more expensive. Cluster 1 is classified by these newer houses with over 5 bedrooms and bathrooms. Cluster 4 is classified by the oldest and smallest living spaces with the fewest number of rooms. Cluster 5 consists of moderately sized houses that are newer. Lastly, Cluster 6 consists of newer and smaller living spaces. We found these combinations to be interesting, because traditionallythe most expensive combinations would consist of spas, pools, additional square feet, but this was proven not to be the case. In California, there are large homes with great amenities, but these homes do not engulf the majority of the housing population. Most of the homes in California are connected and are very similar, so there is a lack of diversity, which leads homebuyers to question the price and relative value of these homes.

We can see with the decision tree results (FIGURE 9), the binary attributes of the dataset are broken between if they have a particular amenity or not. Within each node, there is a count of the number of homes with or without the binary variable paired with the average price of the count of houses meeting the criteria. One thing that stands out and makes sense for the data is the isNewContruction binary variable has far less listings than older properties. This can be seen based on the count of the tree splits. The ones that are older properties have a count of 2,338 while there are just 54 new constructions. This highlights the problem regarding the development of new homes, which may be due to a low supply of or too expensive materials. This logically points to the rising housing prices in California. Taking a look at other values, they indicate that most homes that have amenities like a garage or pool have an average value that is greater than the ones that do not. It is difficult to make assumptions based on the splitting rules of the decision tree, since it does not explain the context of the results. However, if we follow the bold line for the decision outcomes, we see that the majority of houses within the dataset do not have a spa, do not have a pool, have parking, and are older constructions. We see that the average price of these houses is just over $1.1 million. We can assume that houses within three standard deviations of this mean meet the acceptable pricing range, while those that lie outside of the range would be considered outliers. If the listing criteria of these houses are compared with a test set against the results of the training set, then we may reasonably conclude if the listed prices fit the acceptable range, or if there are other underlying issues.

**MANAGERIAL IMPLICATIONS AND CONCLUSIONS**

These models could help prospective buyers determine the acceptable range of housing prices given certain listing properties. The goal of these models is to educate buyers on the results of this analysis and could help future home buyers find cost-saving methods by eliminating additions they do not need, such as an extra bedroom or a pool that can be added if desired. This could save homeowners money in the short-term and long-term if these applications are properly applied. Since California home prices are on the rise with increased demand, it is likely that many “dream homes” are out of the price ranges of buyers. Thus, buyers must sacrifice their wants for necessities to bring their potential costs down. At the same time, if the few homeowners in California pay off their mortgage, it may be wise to think about investing in home additions and upgrades. The data presented in this project can be used to look at areas with lower housing prices to find the optimal areas to make additions to that will increase the value of an existing home. The most valuable sources of those additions come from amenities such as pools, bathrooms, extra rooms, garages, and things of the like. Homeowners can use this data to better prepare for market changes. With this information, they can choose to add if they want to buy a home based on its existing value or buy a house for cheaper and then adding on to it later. If a consumer buys a home relatively under their budget, even with inflated housing prices, the value of that home with additions could rise depending on the area where that home is located. Given this factor, it is imperative to look at key values that may influence housing prices and the demand for certain amenities that could add value to a purchased home in the future. With greater demand for homes and less supply, there is a need for more flippers or property developers to come into the market. Looking at these analyses can help buyers figure out which houses have optimal pricing based upon their desired characteristics. Targeting certain characteristics of homes that increase value may be a key player in driving more supply into a market that desperately needs it. This may not be a short-term fix, but enough supply in the market would likely incite competition and as a result drive prices down in the long-run.

Our analysis has just brushed the surface of the situation surrounding the California housing crisis. There are many opportunities for further research that could be even beneficial. For instance, looking into different pricing based on locational information may be useful for diagnosing areas where the inflation of housing prices is greater. If an area is found to be more expensive than other areas, new home buyers with a stricter budget might want to look in an area that is less expensive. The data can help buyers figure out which areas fit them the best financially. Further applications of this research could be used to track price increases and decreases within certain areas to target specific areas. For example, if an area does not have as many properties, then the price of homes may be higher, so developers might want to build in that area or county. There are certainly limitations to the extent and the validity of this research, but the potential for further research on this topic can reveal more relevant findings to help keep consumers informed about the volatile California housing market.

**APPENDIX**

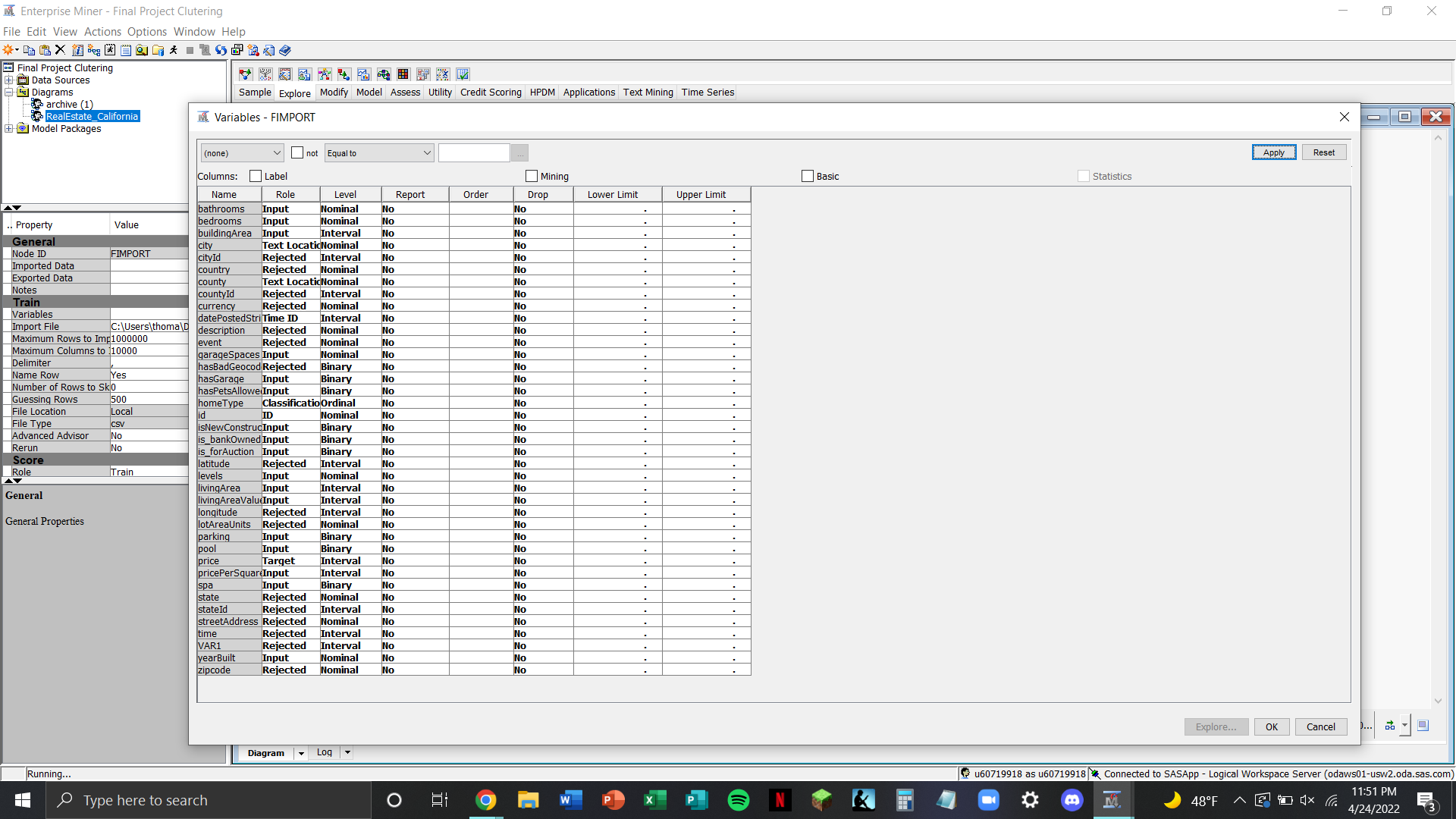
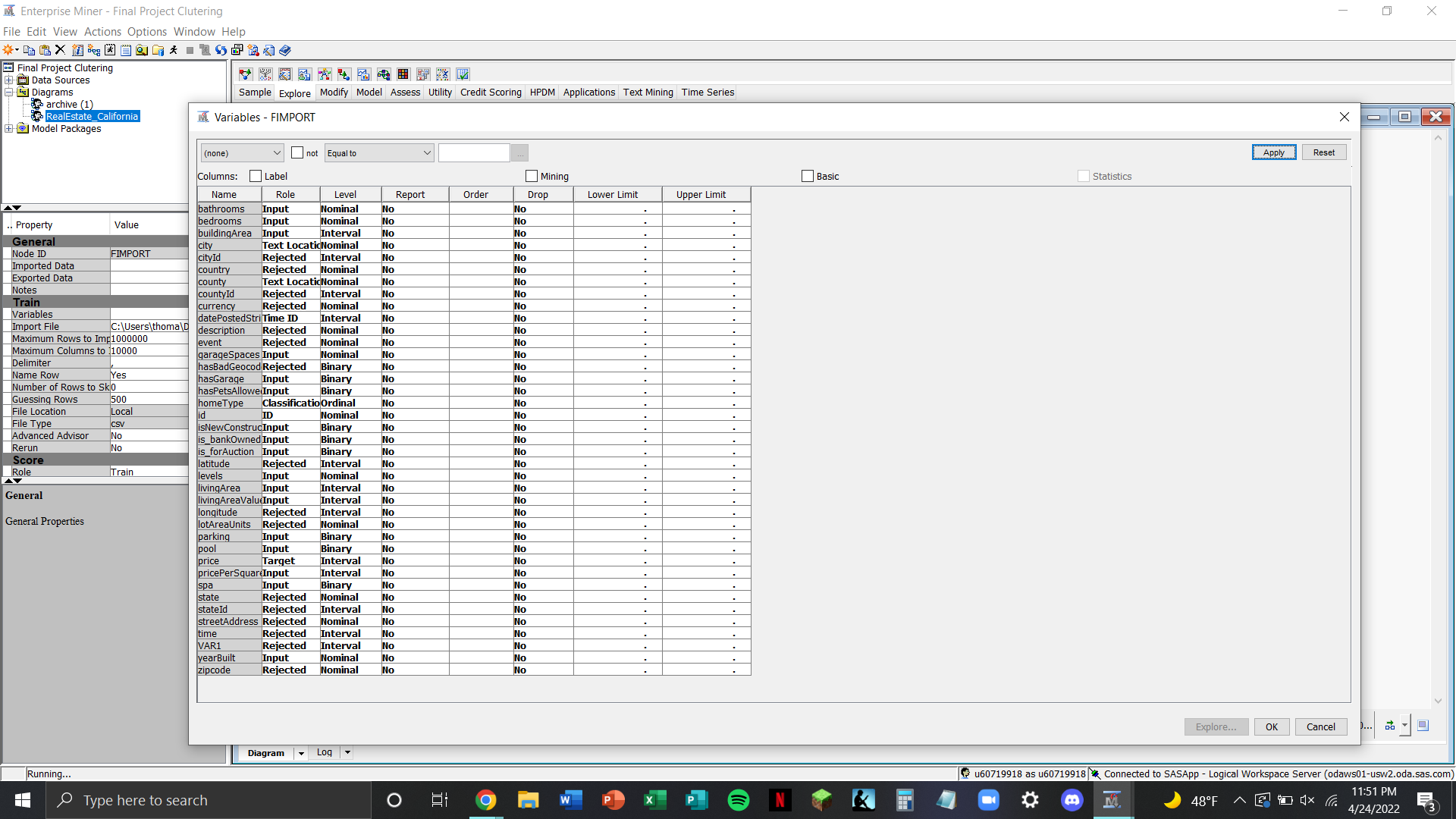


FIGURE 1. The list of all dimensions from the California housing database, including their role and level.

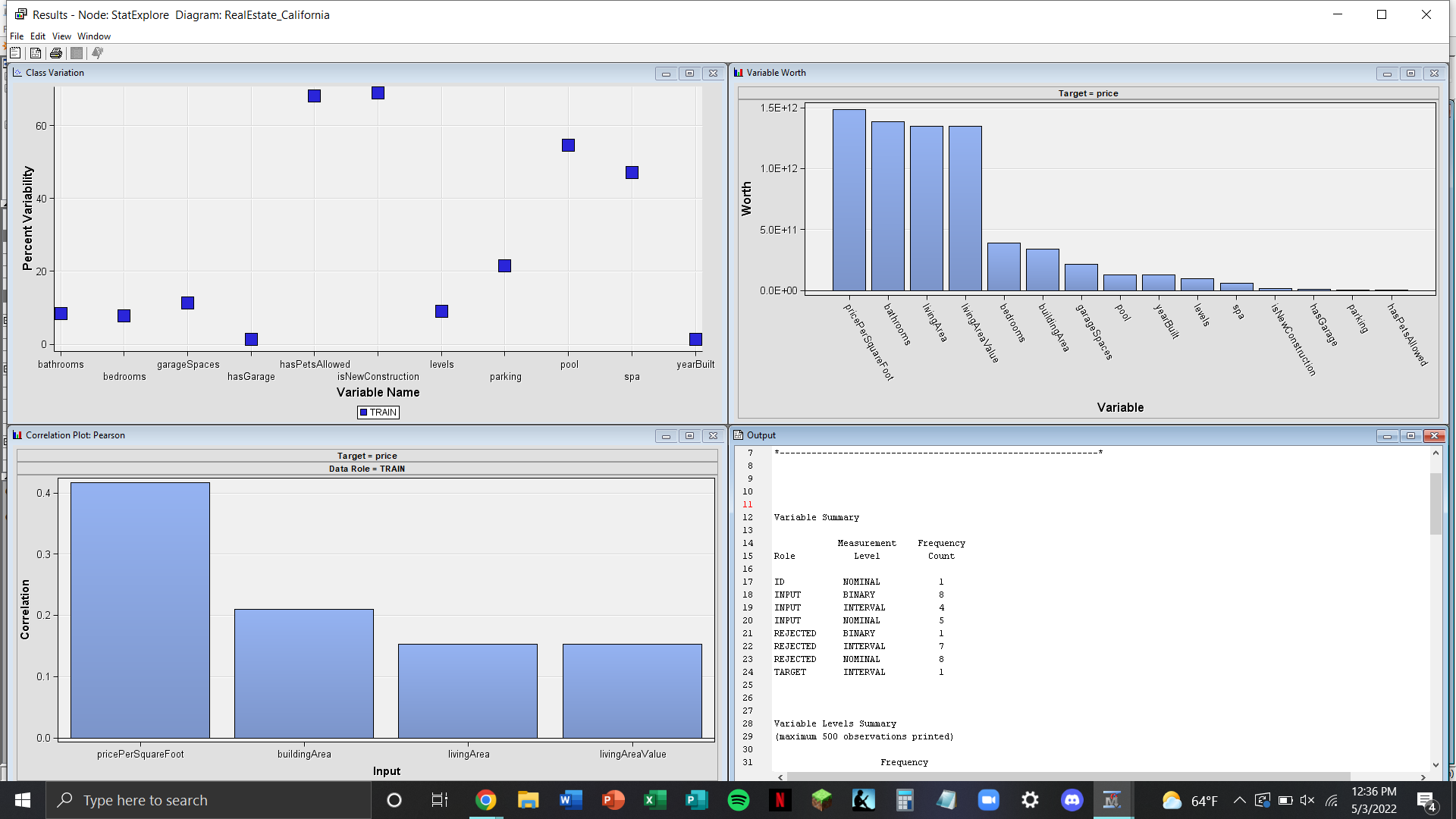
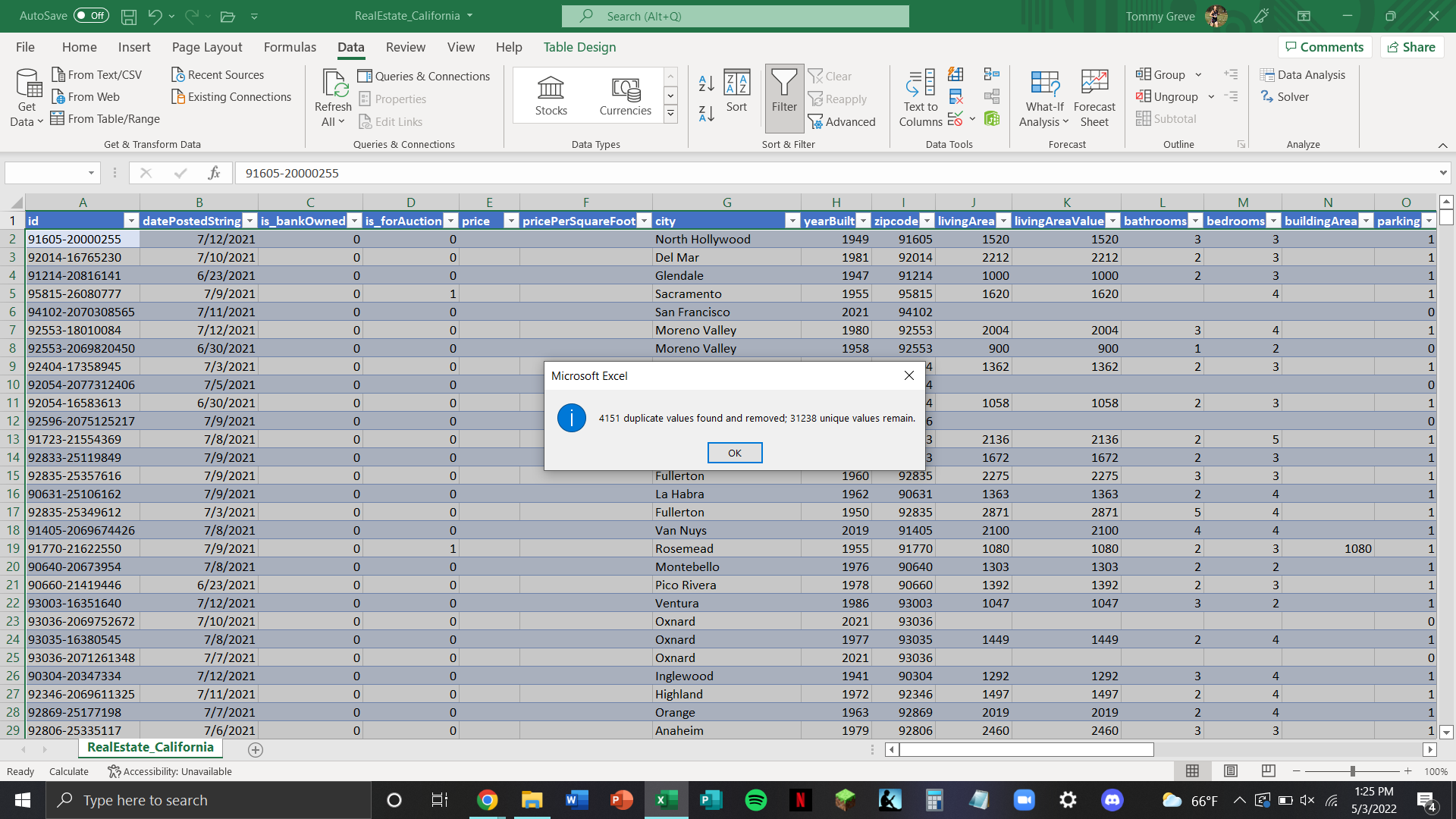


FIGURE 2. A summary of the count of variables based on their role and levels.

FIGURE 3. A look at the data scrubbing process to remove duplicate listings.

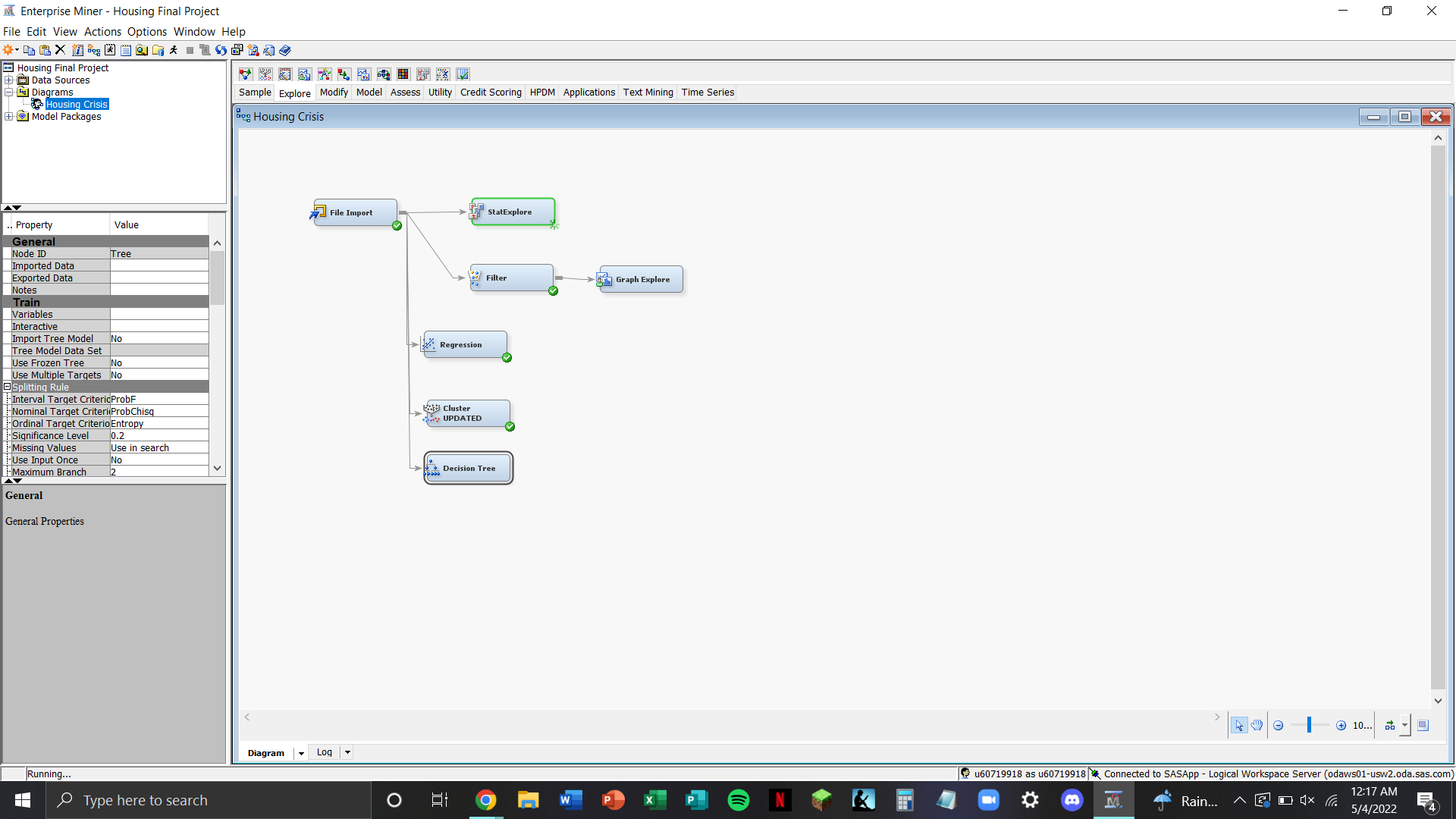


FIGURE 4. A SAS Enterprise Miner workflow displaying our analysis of California housing prices.

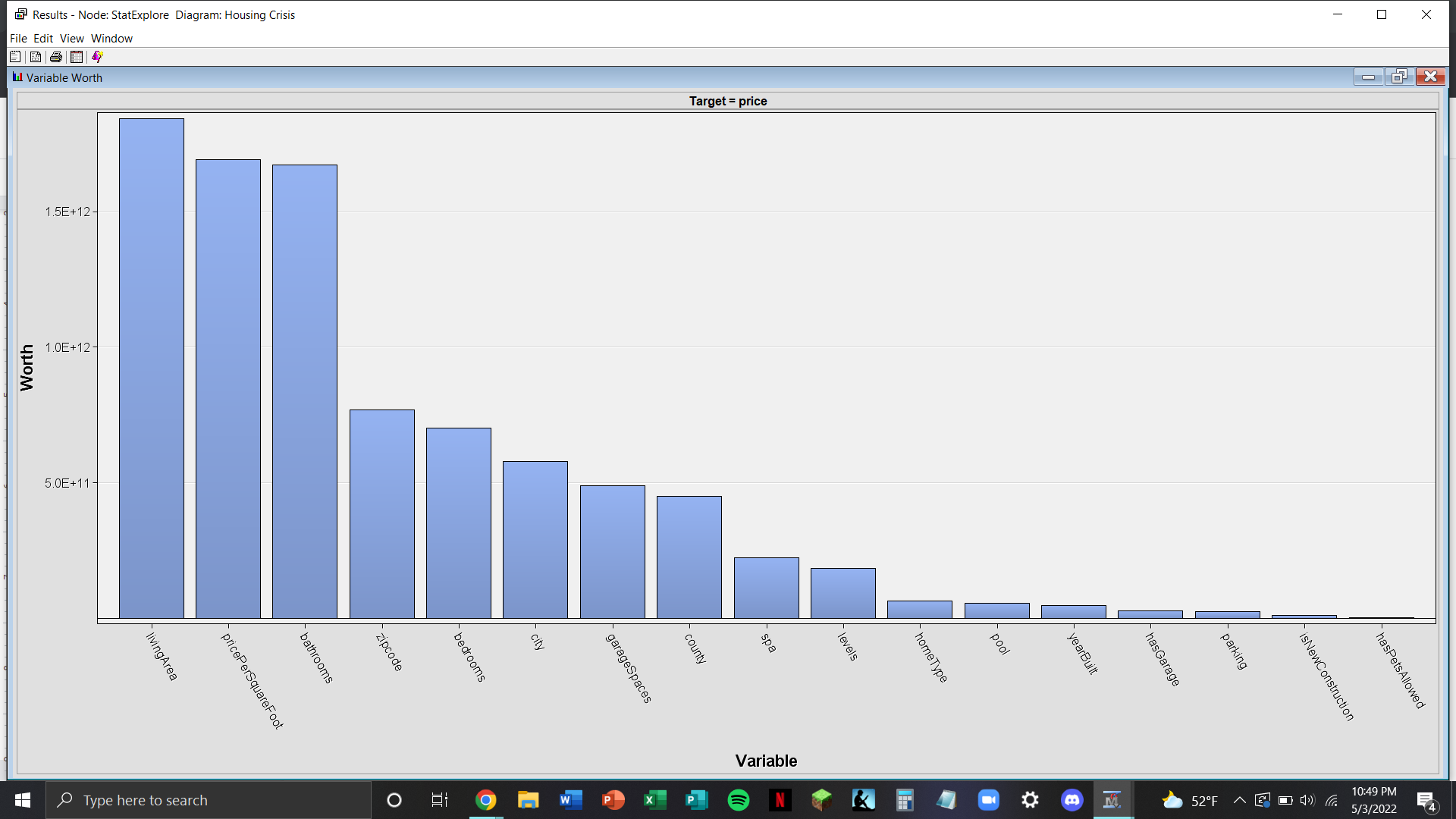
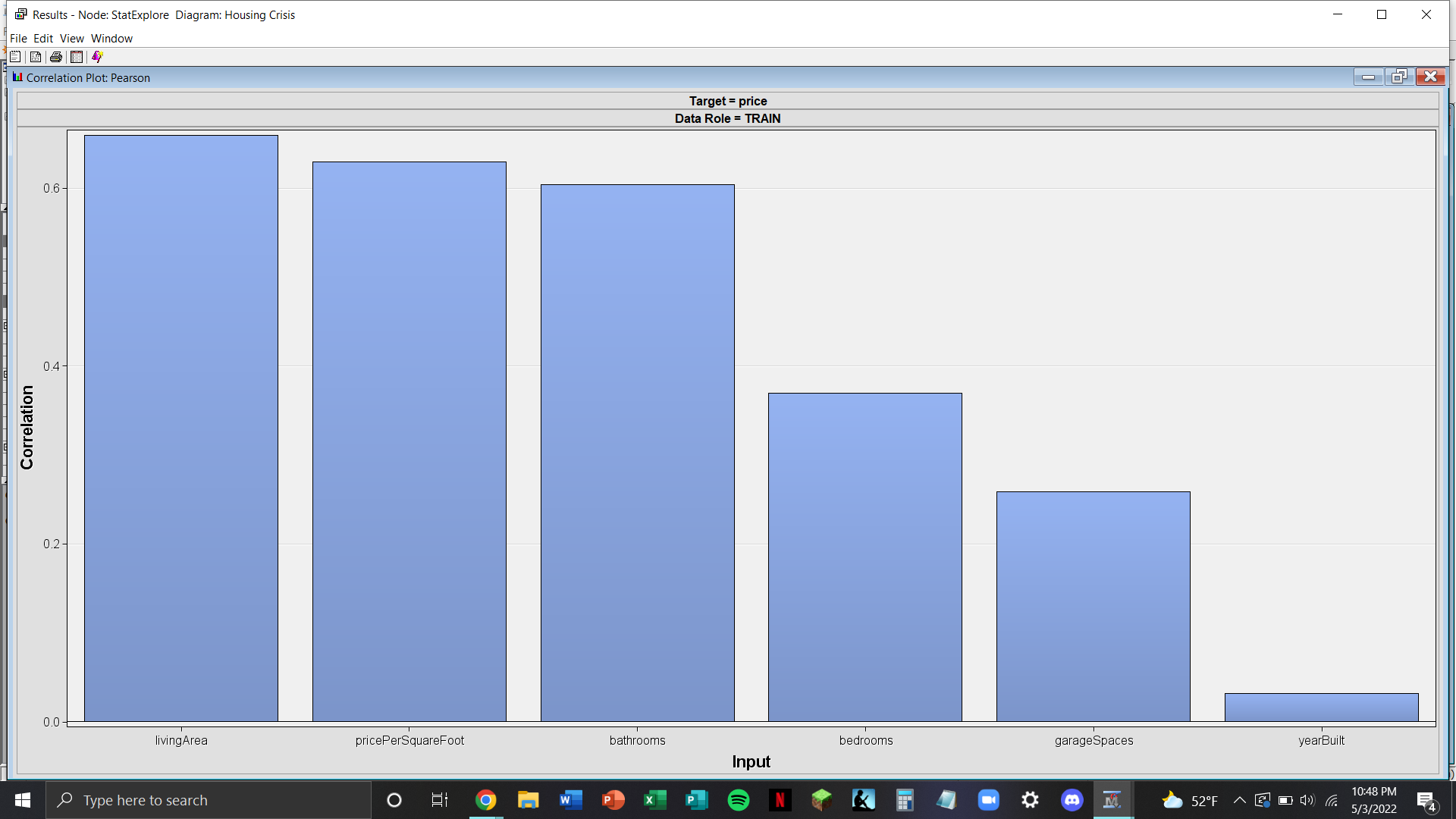
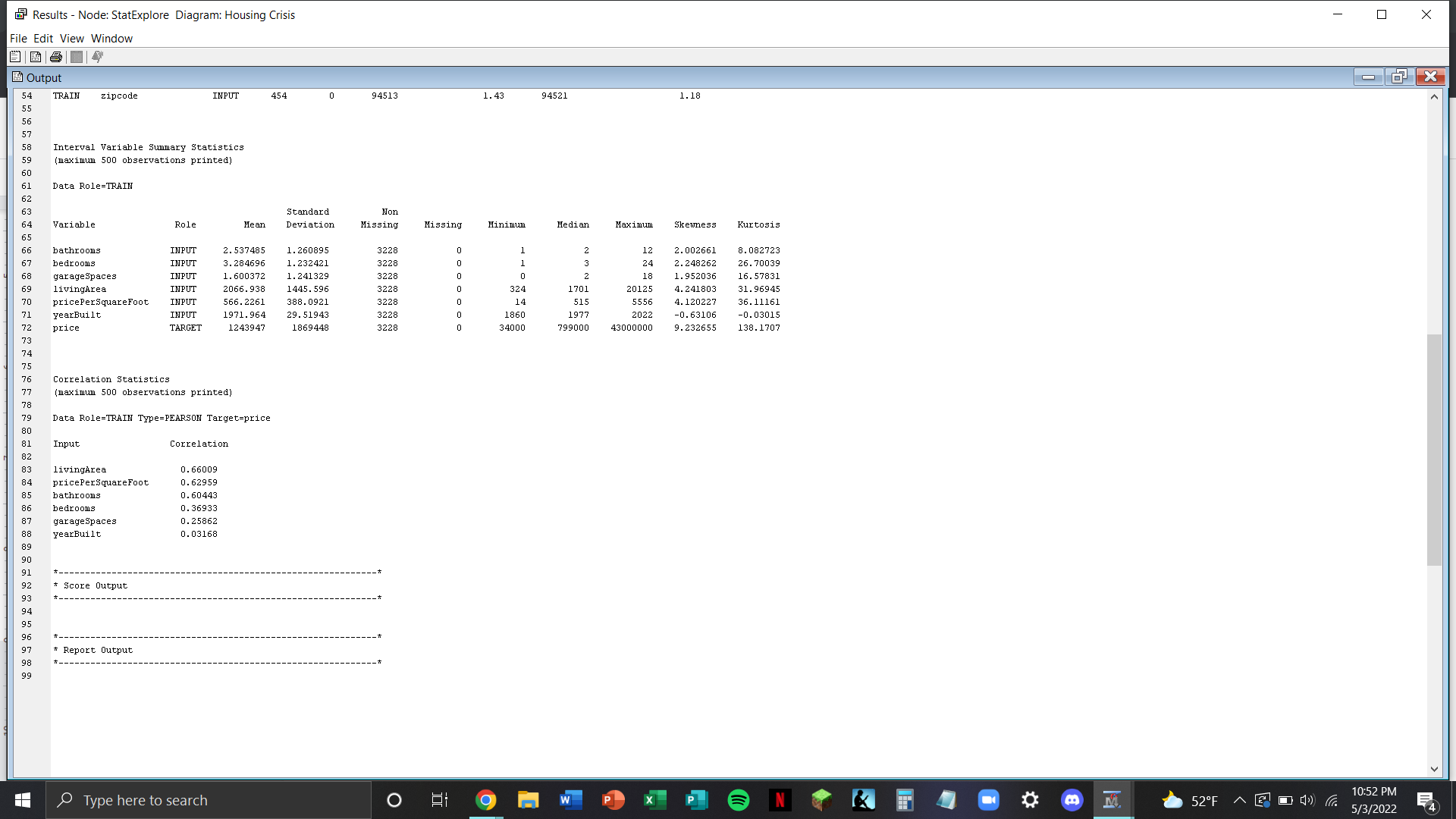
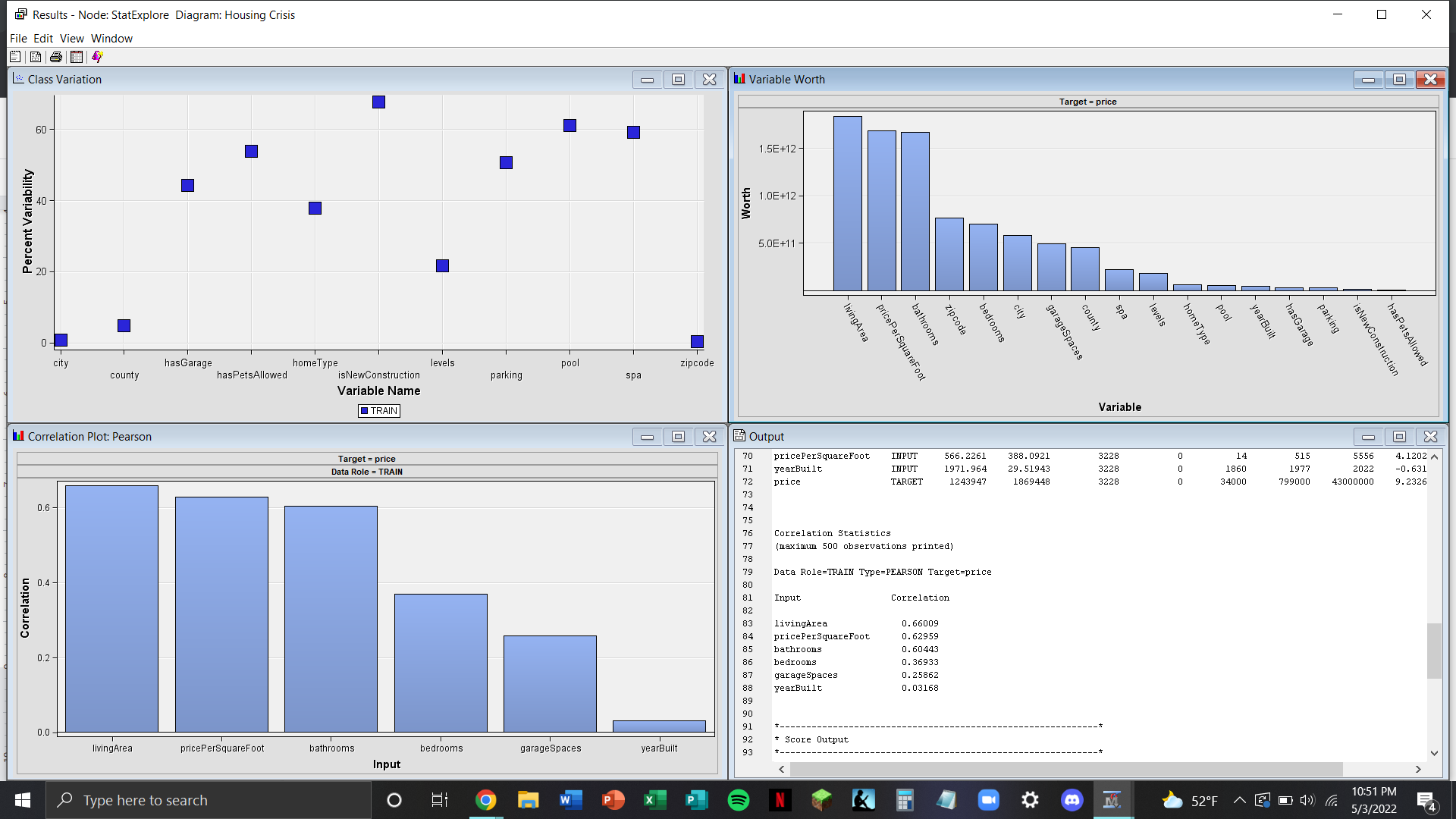
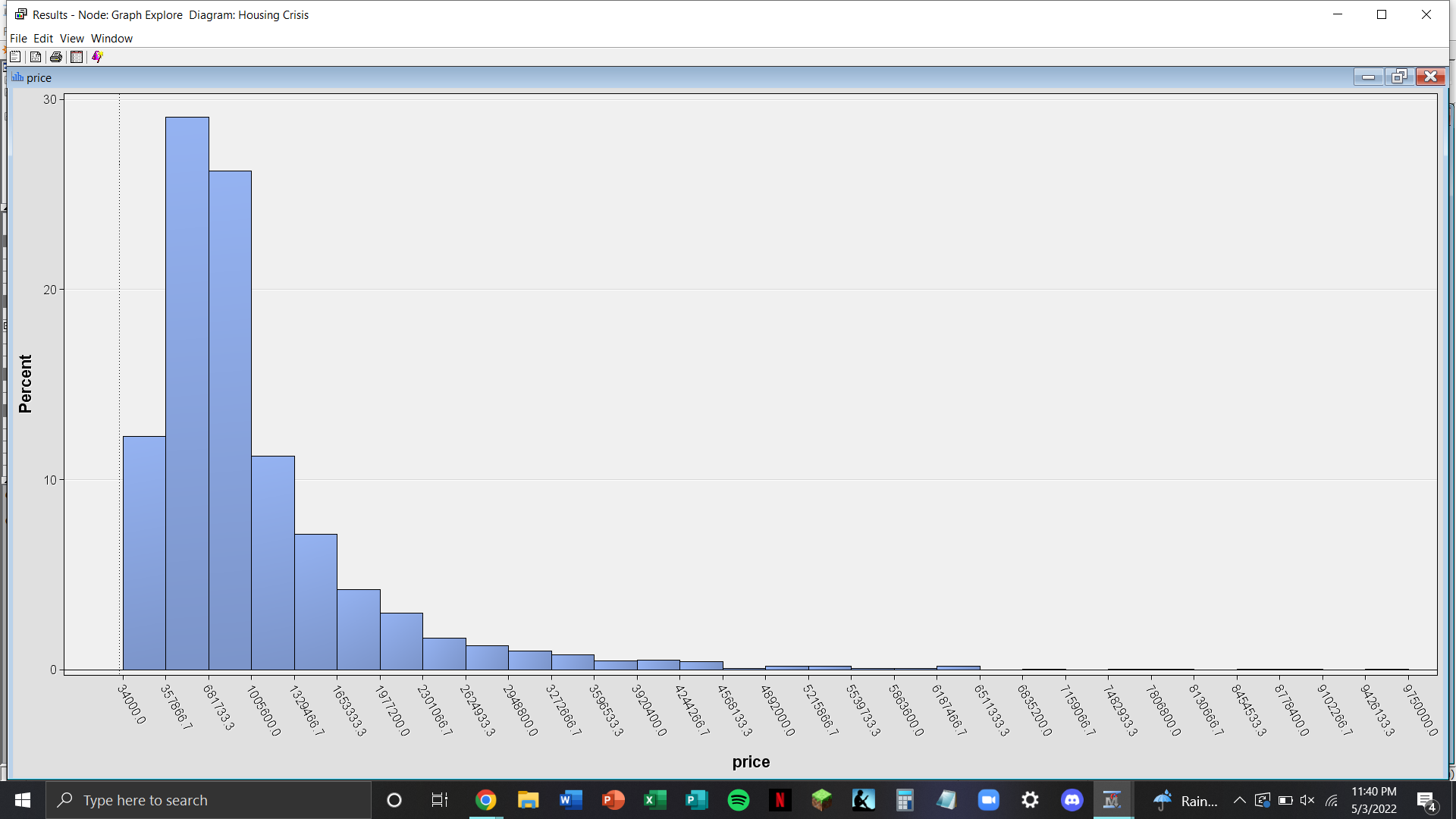


FIGURE 5. A summary of the StatExplore input correlations, variable worth, and a statistical summary.

FIGURE 6. A distribution of the shares of the target variable: price.

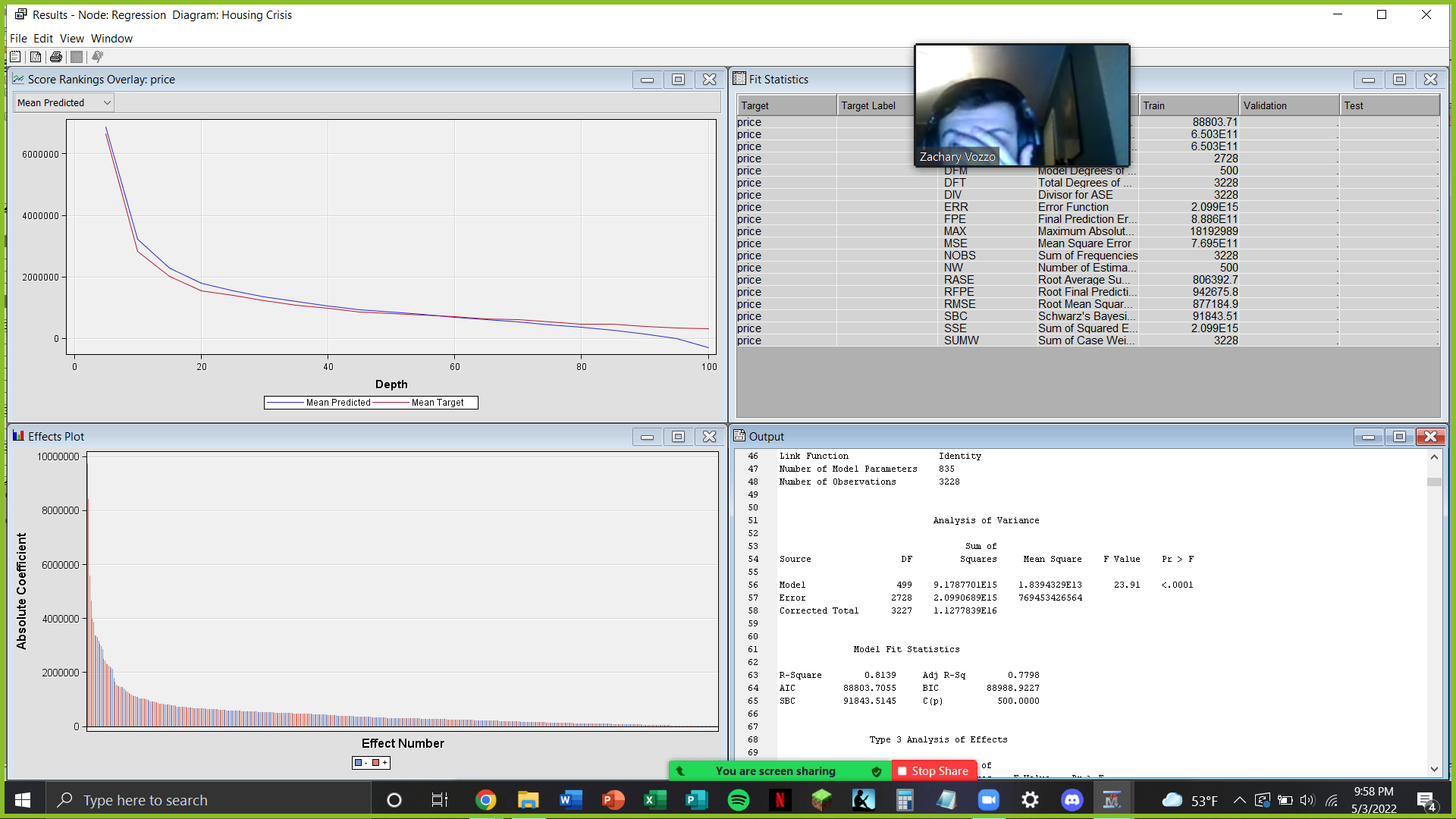


FIGURE 7.1. The output from the logistic regression model.

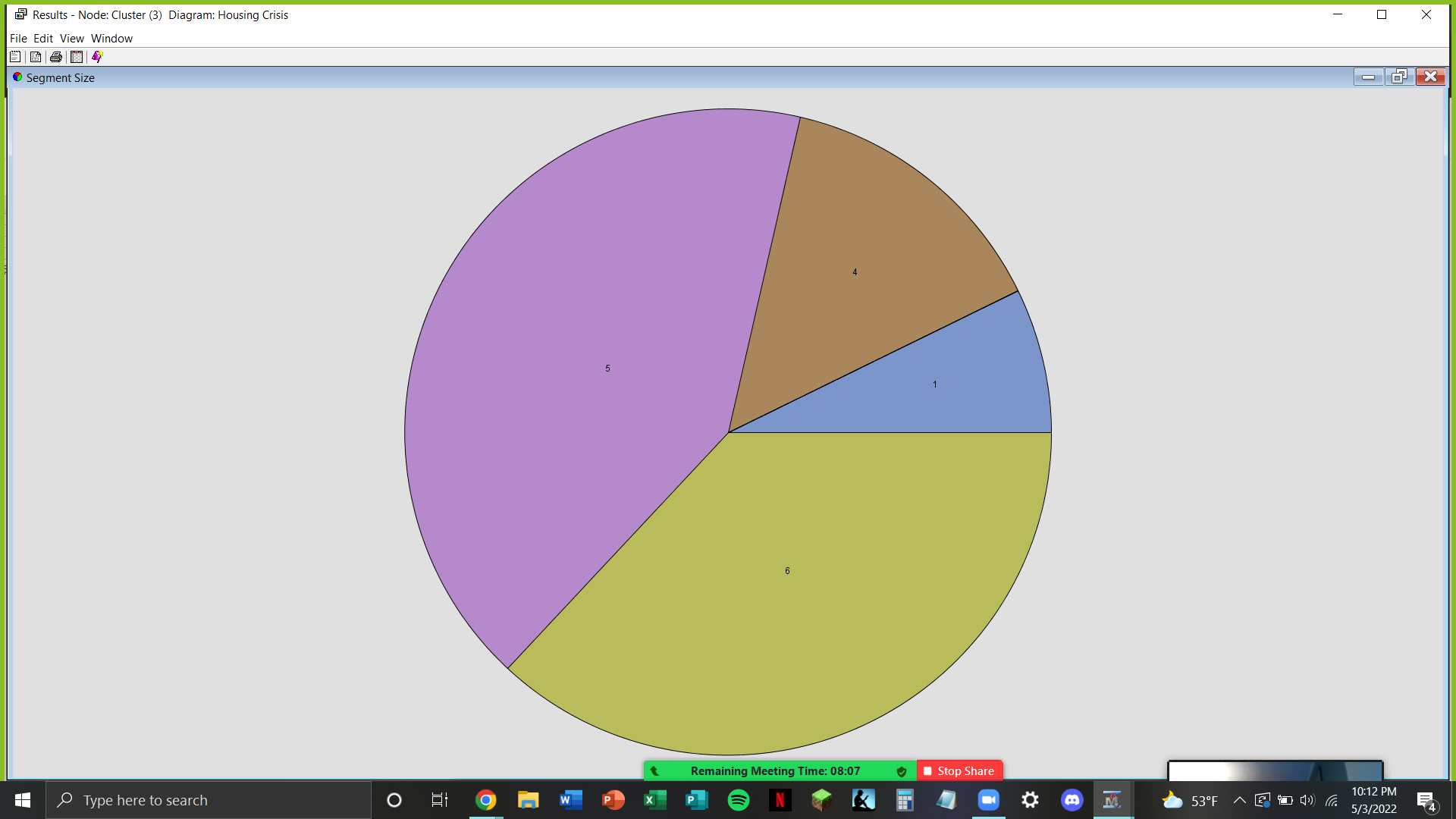
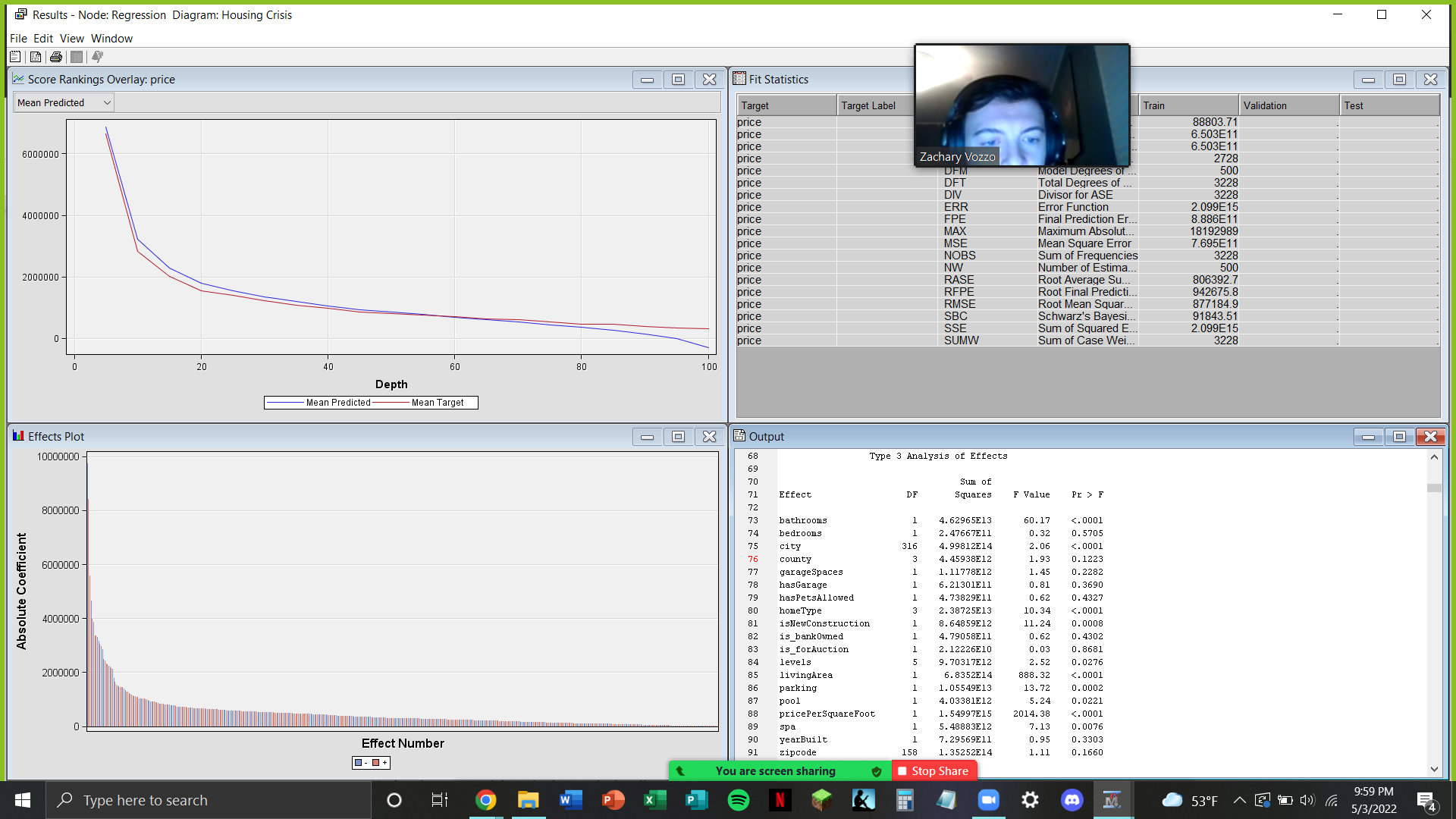
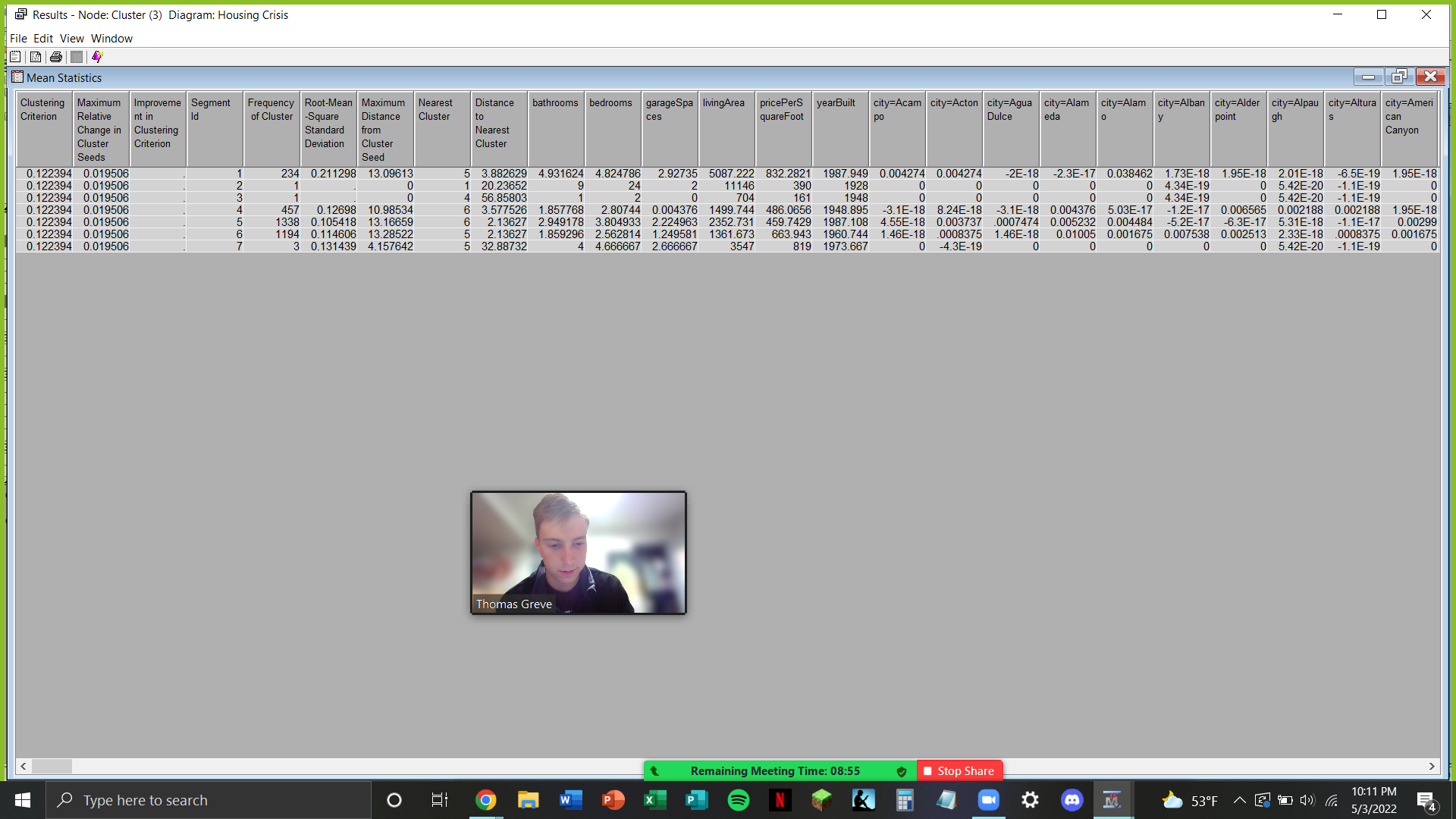
FIGURE 7.2. The output from the logistic regression model.

FIGURE 8. A custom centroid method cluster analysis, creating 7 distinct clusters.

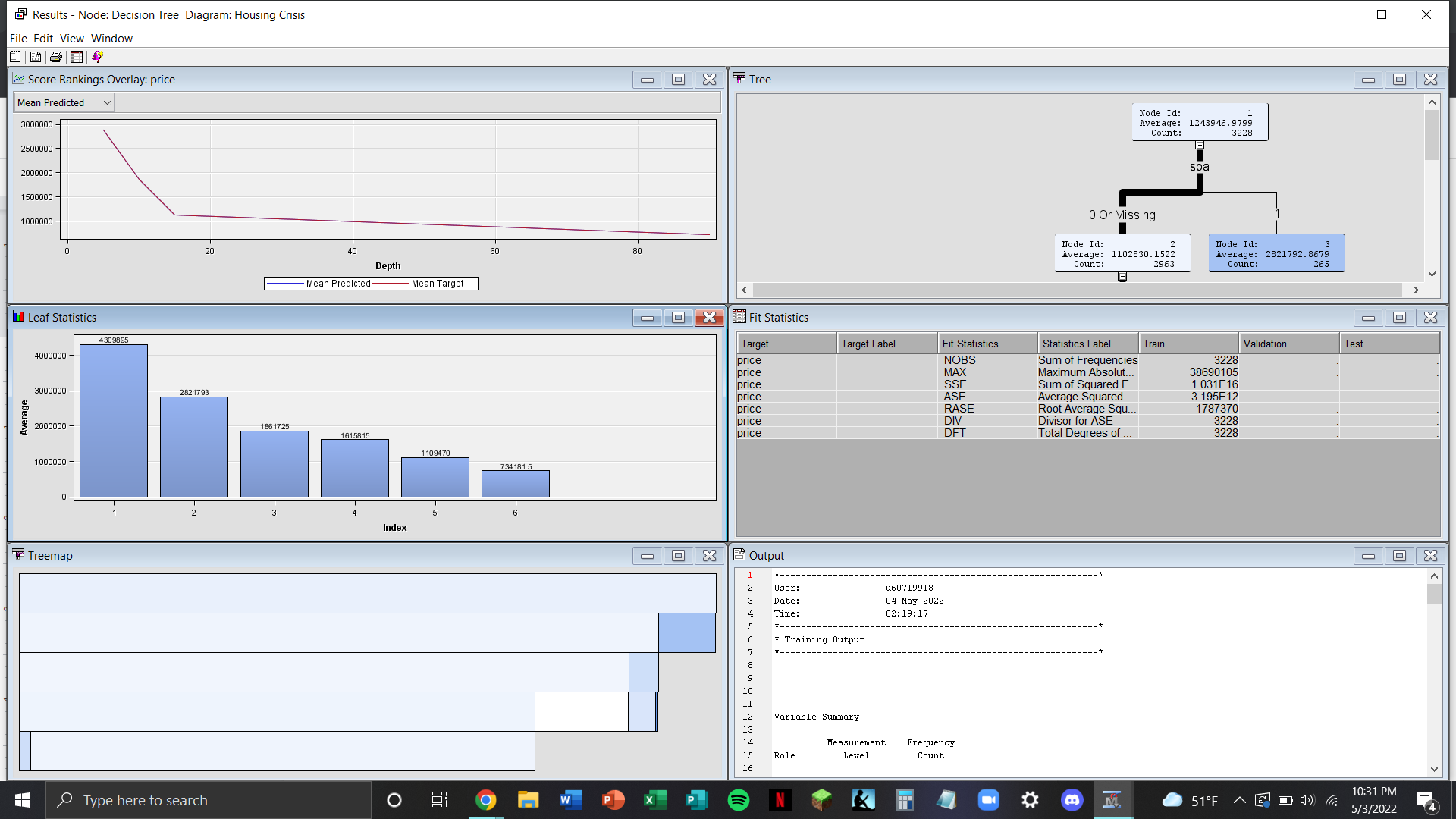
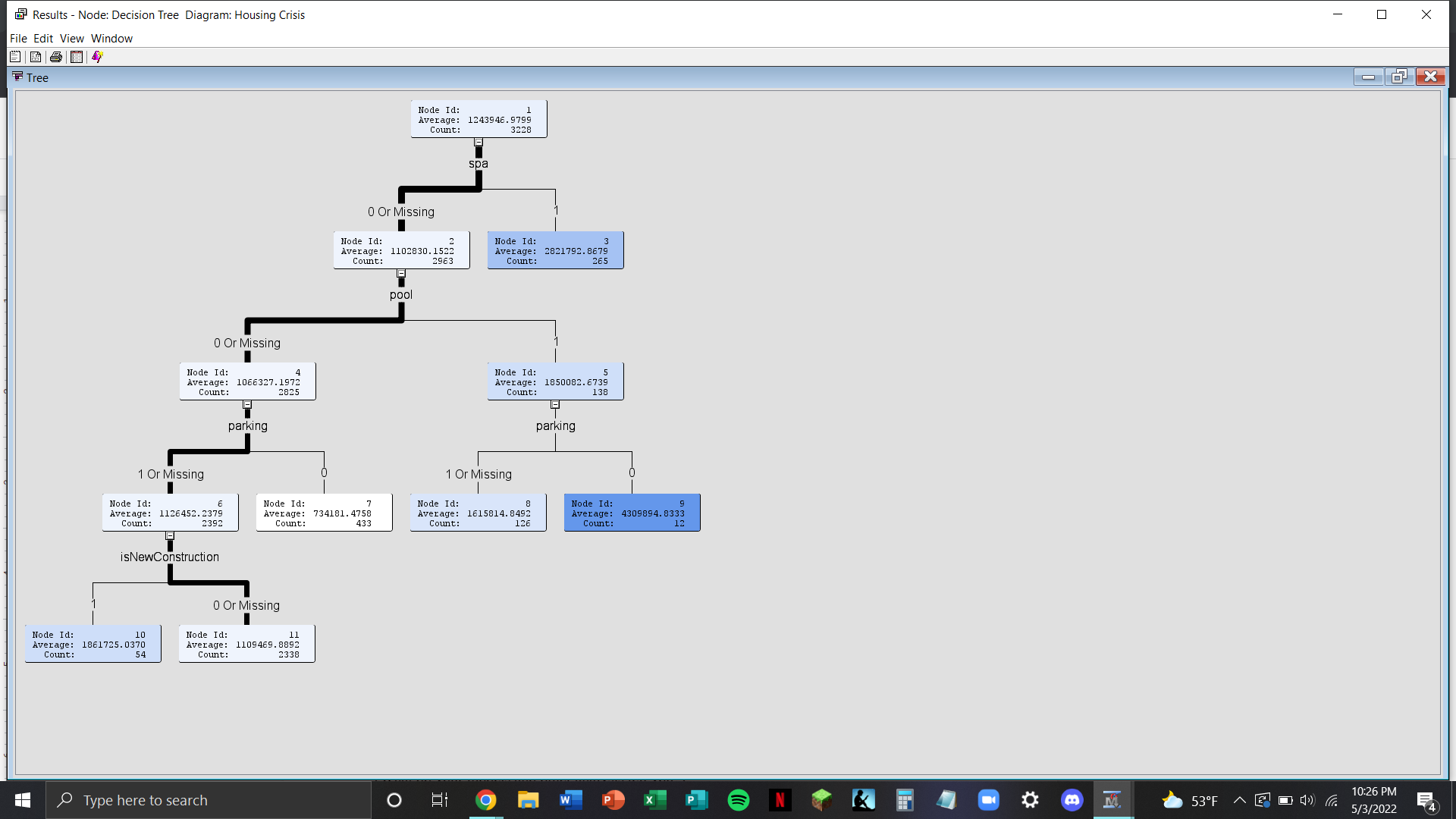


FIGURE 9. Our decision tree output based on the binary inputs in the dataset.