**Analyzing Local Market Trends to Predict Maryland Property Prices**



Group A: Emilee Sheaffer, Justin Mejia, Matthew Rutigliano

## **ORIGINAL WORK STATEMENT**

We the undersigned certify that the actual composition of this proposal was done by us and is original work.

|  | **Typed Name** | **Signature** |
| --- | --- | --- |
| Contact Author | Emilee Sheaffer |  |
|  | Justin Mejia |  |
|  | Matthew Rutigliano | Matthew Rutigliano |

1. **Executive Summary**

Our study investigates what housing market factors impact the average sales price of housing in Maryland, utilizing data measured from January 2022 to May 2024 and collected by the U.S. Bureau of Economic Analysis (BEA) and distributed by the Maryland Board of Realtors. Although the number of observations is relatively small, we covered 2.5 years of housing data. Those 2.5 years are a significant portion of the post-COVID market and capture how recent housing sales behavior may inform future trends in housing prices.

The ideal goal of the study is to effectively determine which market characteristics have an impact on the average sales price of housing. We predicted that region, personal income per capita, and season would have the most predictive impact on housing sales prices. In terms of the relationships between our primary characteristics, we’re estimating that new listings and median days on market will decrease, while income will increase, as housing prices increase. The aim is to build a model that successfully predicts the sale prices of new properties that enter the market based on local market characteristics and housing market trends. Additionally, we also aim to predict whether the housing market in a county is “Hot” or “Cold” based on home sales volumes, average sales prices, and median days on the housing market. By combining these two viewpoints, we hope to provide a unique perspective to studying what drives housing prices.

The study’s findings confirmed that county, income, and season were the independent variables most likely to have a predictive influence on the average sale prices for homes in Maryland. Additionally, through classification models, we uncovered that within the three year span of data collected, a majority of counties in Maryland were classified as a “Hot” housing market, with prices falling above average and median days on market below average. These findings have potential implications for real estate professionals, investors, as well as current and future homeowners, with the opportunity to more accurately predict and classify trends in housing markets across the United States.

1. **Data Description**

Data was collected from three primary sources: The Maryland Board of Realtors, The U.S. Bureau of Economic Analysis (BEA), and The Maryland Office of Tourism. The Maryland Board of Realtors releases a housing statistics update every month (data sourced from the BEA), which details home sales data per county, including a review of units sold, average sales prices, as well as information surrounding active inventory and median days on market.

Source Description

* Sources consist of statistics collected from January 2022 - May 2024, and are segmented by county. In addition to home sales, the housing data also includes sales of condos and co-ops.
* Data of yearly income per county collected by the BEA was included to take a look at the effects of average personal income per capita on home sales. The BEA data included personal income for 2021 and 2022. The average annual growth rate per county from 2021 and 2022 was calculated to generate estimates of what the average personal income per capita would be for 2023 and 2024.
* The Maryland Office of Tourism grouped together counties of Maryland into geographical regions, which allowed the study of regional impact on the average sales prices of homes.

Data Variable Information:

**Housing Statistics (Maryland Board of Realtors, sourced from the BEA):**

* COUNTY - Counties of Maryland, Categorical - factored.
* MONTH - Months of the year, Categorical - factored.
* YEAR - Years (2022 - 2024), Numerical.
* UNITS\_SOLD - Number of homes sold per county by month, Numerical.
* AVG\_SALE\_PRICE - Average sales price of homes sold per county by month, Numerical.
* MEDIAN\_SALE\_PRICE - Median sales price of homes sold per county by month, Numerical. Not included in this study.
* UNITS\_PENDING - Number of homes under contract but not yet sold per county by month, Numerical.
* ACTIVE\_INVENTORY - Number of homes on the market per county by month, Numerical.
* MONTHS\_INVENTORY - Measures the rate at which homes are sold per county by month. Notes the relationship between the number of homes sold in a month by the total number of homes for sale at the end of the month, Numerical.
* MEDIAN\_DAYS\_MARKET - Measures the median days that a home is listed on the market per county by month, Numerical.
* NEW\_LISTINGS - Measures the number of new listings on the market per county by month, Numerical.

**Personal Income per Capita** (**BEA**):

* PERSONAL\_INCOME\_PER\_CAPITA - Measures the personal income for the average person per county by year, Numerical. Shortened to INCOME in the script.

**Regions of Maryland** (**The Maryland Office of Tourism**):

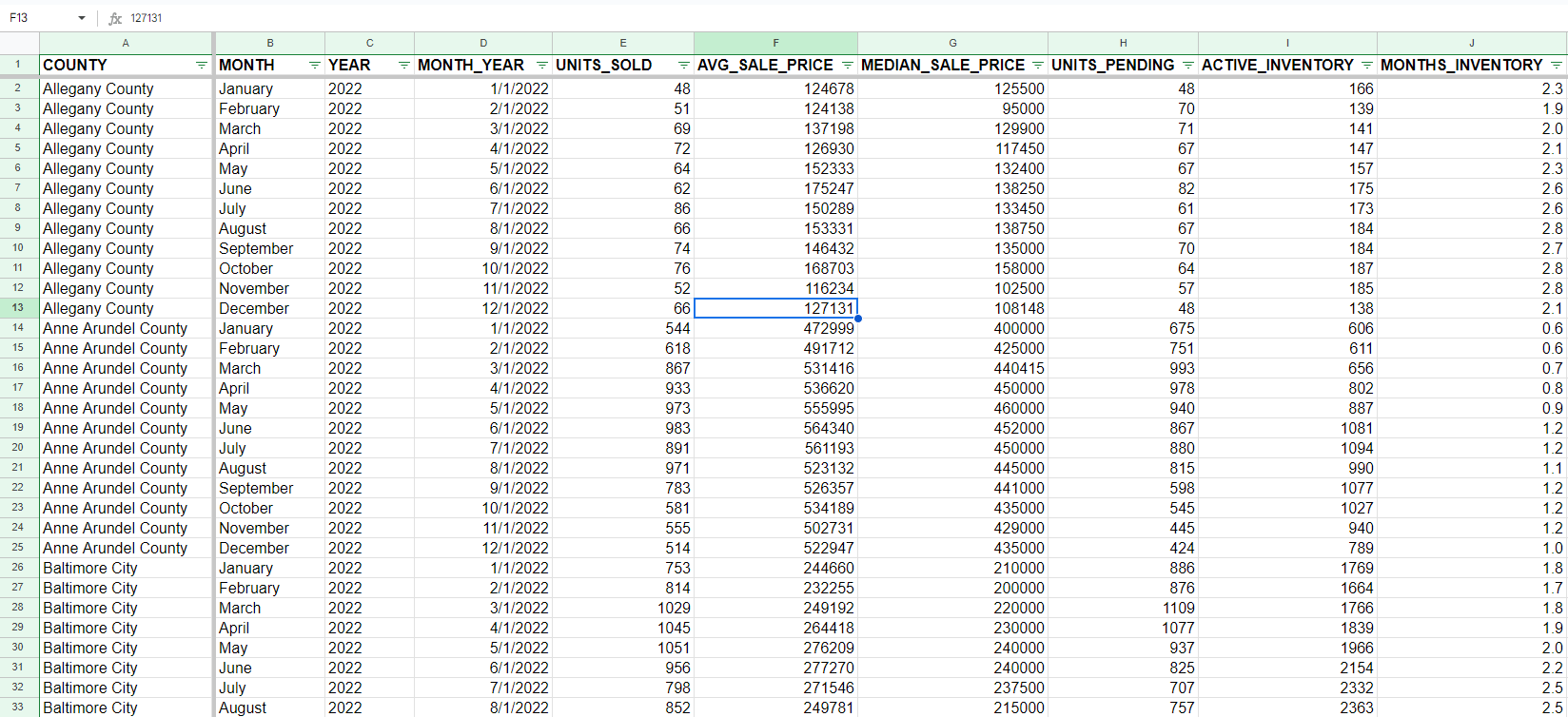
* REGION - Geographic regions of Maryland, Categorical - factored.

**Organizational Fields** (**derived from data**):

* MONTH\_YEAR - calculated off of MONTH variable, Numerical. Not included in this study.
* QUARTER - Calendar quarters of the year, Categorical - factored - calculated based on MONTH variable.
* SEASON - Calendar season of the year, Categorical - factored - calculated based on MONTH variable.

Sample Information:

* Our sample size (n) is 696 total months of housing data broken out by each county of Maryland. Our number of variables (k) is 16, with three categorical variables and 13 numerical variables.



Interest:

* We are interested in understanding what predictor variables have the most significant effect on our outcome variable of average sale prices. We are also interested in generating predictions of what the average sale price would be of a housing unit given a set of predictor variables like local personal income per capita, active inventory, and geographical region. Our goal with this research is to develop a statistical model using machine learning to predict the average sales price of new housing for future years based on a variety of predictor variables.

Data References:

* Housing Statistics (Includes: COUNTY, MONTH, YEAR, UNITS\_SOLD, AVG\_SALE\_PRICE, MEDIAN\_SALE\_PRICE, UNITS\_PENDING, ACTIVE\_INVENTORY, MONTHS\_INVENTORY, MEDIAN\_DAYS\_MARKET, NEW\_LISTINGS)
  + <https://www.mdrealtor.org/News-and-Events/Housing-Statistics>
* Personal Income Per Capita (Includes: PERSONAL\_INCOME\_PER\_CAPITA)
  + <https://www.bea.gov/data/income-saving/personal-income-county-metro-and-other-areas>
* Regions of Maryland data (Includes: REGION)
  + <https://www.visitmaryland.org/article/maryland-regions>

**III. Research Questions**

1. What are Maryland’s average housing sales price ranges by region and county?

Our first question examines the distribution of property prices by their region and county, which offers insights on the market from a geographical perspective. We plan to visualize the data through a boxplot and scatterplot to provide digestible results and information. The boxplot will provide key statistics that can easily show central tendency of prices, variability in the spread of price, and if any regions have significant outliers that would impact the mean or data trends. The scatterplot will show the relationship between price, region, and counties.

By plotting these variables, we can see how prices are distributed across different locations and identify any clustering and locational patterns. This will further provide insight on pricing trends and determine if there are any concentrations of high or low-priced properties based on geographical area.

1. What indicators in the Maryland housing market best predict average sales prices?

The purpose of this question is to explore the impact of different independent variables on average housing sales price, such as active inventory, units sold, median days on the market, month, and personal income. The goal of this investigation is to identify which variables are the most impactful for home owners and real estate professionals to consider when assessing housing market sales price trends. Assuming that our dataset met certain linearity assumptions, we will use a linear regression model to observe what predictors had the greatest influence.

Once executed, we will evaluate the model’s predictive capability by splitting our data set into a training and test data set, and tabulating the error of each regression model. We hope that our linear regression model allows buyers and sellers to make more knowledgeable decisions and adjustments.

1. What will the average sale price be with specified criteria?

As alluded to in question two, our third question focuses on predicting the average price based on certain parameters. Building a model that allows for predicting prices provides flexibility for stakeholders in the housing market. These calculations would allow for more accurate predictions in price and would give buyers and sellers an idea in what areas they are able to purchase, when they should begin trying to purchase a home, and what price range they may be able to afford.

1. Which counties in Maryland are considered “hot” and “cold” markets by sales volume?

This final question proves that the core data provided can be manipulated to add context to sales prices and their corresponding variables. The derived metrics of “Hot” and “Cold” are based on the variable “Market Status”. When labeled “Hot”, the average sales price is above the state’s median and its median days on the market are below the state’s median. This classification helps those in the real estate market to derive information regarding high demand and quick turnover.

**IV. Methodology**

**Data Exploration and Initial Prediction Model Development**

We began our data mining process by partitioning the dataset into a training set (70%) and a test set (30%). This initial step ensured that our models could be evaluated on unseen data to prevent overfitting. Using data visualization techniques, we explored the distribution and relationships within our dataset. A **histogram (See figure AA in the appendix)** of average sales price allowed us to observe the overall distribution, confirming that the price distribution is positively skewed. A **boxplot (See figure AB in the appendix)** of average sales price by region identified the central tendency and variability of prices across different regions, noting that most regions had no significant outliers. A **scatter plot** **(See Figure 1 on pg. 8)** of average sales price by county helped us quantify regional trends and understand how prices vary across different counties.

To understand the relationship between predictors and average sale price, we built three linear regression models. This modeling process involved several key assumptions:

1. **Linearity**: We assumed a linear relationship between the predictors (independent variables) and the average sale price (dependent variable). This assumption was checked by examining residual plots.
2. **Independence**: We assumed that the observations were independent of each other, meaning the sale price of one property does not influence the sale price of another.
3. **Homoscedasticity**: We assumed that the variance of the errors was constant across all levels of the independent variables. This was assessed by inspecting the scatterplots of residuals versus fitted values.
4. **Normality of Errors**: We assumed that the residuals (the differences between observed and predicted values) followed a normal distribution, which we verified by normalizing RMSE and MAE.

Given the overlapping nature of some variables (e.g., Season, Month, Year, and Quarter), it was crucial to identify which variables had the highest correlation with the average sale price while ensuring these assumptions were met. We evaluated the performance of these models using the root mean square error (RMSE) and mean average error (MAE) on both the training and test datasets. This comparison helped us select the most accurate linear regression model for prediction.

**Market Classification, Development and Evaluation**

We aimed to classify Maryland’s counties into "Hot" and "Cold" categories using various machine learning approaches. To start off, we engineered a binary target variable, "Market Status," classifying a county as a hot market if its average sale price was above the state median and its median days on market were lower. Otherwise, the county was classified as cold. The average sale price is a direct indicator of the market's demand and overall strength. A higher average sale price relative to the state median suggests that homes in the county are selling for more. By comparing the county's average sale price to the state median, we normalize the classification across different regions, allowing us to account for statewide economic variations. This approach helps to identify counties that are outperforming others within the state, classifying them as hot. Median days on market is a key measure of how quickly homes are being sold. A lower median days on market implies that properties in the county are selling quickly. This rapid turnover is a clear signal of a hot market.

To ensure model integrity, we excluded the Average Sale Price and Median Days in Market variables. We partitioned the dataset into training (70%), validation (15%), and test (15%) sets. The validation set was necessary to fine-tune the models (such as the K-Nearest Neighbors model) while still having a test set unobserved to determine a reliable error rate.

Techniques utilized to analyze the Hot/Cold classifications:

1. **Logistic Regression:**
   * Predictor variables included Region, Units Sold, Units Pending, Active Inventory, Months Inventory, New Listings, Season, and Income.
   * We evaluated this model using confusion matrices, ROC curves, and performance metrics such as accuracy and sensitivity, demonstrating its strong classification performance.
2. **K-Nearest Neighbors (KNN):**
   * We standardized numerical variables and created dummy variables for categorical predictors.
   * Different values of K were tested to minimize validation error.
   * ROC curves and Lift charts were used to visualize and compare model performance.
3. **Linear Probability Model (LPM):**
   * The binary target variable was converted to a numeric format to fit a linear model.
   * Validation and test error rates were computed to assess model accuracy.
4. **Classification Tree:**
   * We constructed a classification tree to categorize markets, pruning the tree based on cross-validation results to prevent overfitting.
   * This method allowed us to interpret the decision-making process behind market classification.
5. **Random Forest:**
   * A random forest model was also developed, which provided a robust ensemble approach to classification.

**V. Results and Findings**

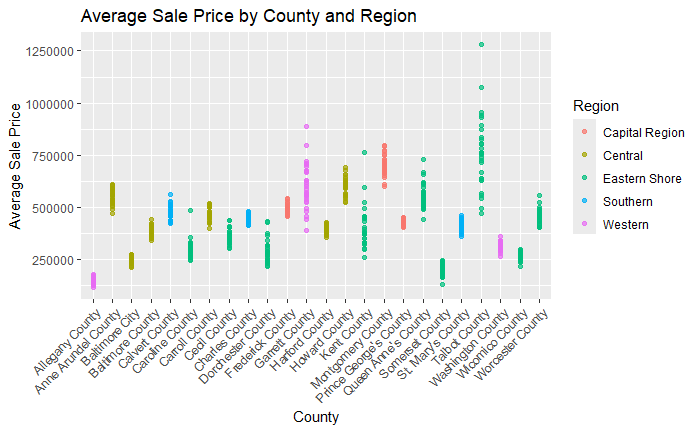
**Descriptive Statistical Findings - Distribution of Prices by County and Region:**

Taking an initial look at our data visualizations, it appeared that individual county had more of an impact on the average sales price of a house when compared to region. As seen in the scatterplot in Figure 1, in addition to the boxplot in Figure AB in the appendix, average sales prices in the Eastern Shore and Western Maryland experienced a large variance. For example, the Eastern Shore saw the state’s largest average sales price overall in Talbot County, but also had the second lowest median average sales price due to lower sales prices in Somerset and Washington Counties. We were surprised to see the large disparity in average sales prices in two of Maryland’s more rural regions, but were not surprised to see that both regions included the counties with the first and second lowest average sales prices in Maryland. The distribution of our data was right skewed (as shown in our histogram in Figure AA), showing the impact of outlier average sales prices in counties like Talbot and Garrett counties.

Generally speaking, a majority of sales prices hovered around the mean of average sales price, which was about $430,000 (also shown in Figure AA). The median average sales price was largest in the Capital Region of Maryland. This trend was particularly noticeable in the average sales price of housing in Montgomery County, the largest for an individual county in the state and home of major economic centers like Chevy Chase. However, average sales prices in Frederick and Prince George’s County in the Capital Region were unexpectedly lower than counties in regions with lower average sales prices.

Knowing the distribution of average sales price by county and region allowed us to know more about the characteristics of our dataset before running linear regression to see what independent variables had the most impact on the average sales price.

Figure 1) Scatterplot displaying average sales price by county and region:



**Linear Regression Findings - Predicting Housing Prices:**

After analyzing our dataset descriptively and learning more about the distribution of our data, we ran a linear regression model (LRMPrice\_0) using average housing sales prices as our dependent variables and all other variables in our dataset as independent variables. As we stated in the methodology section above, we split our data into a training and validation data set, with 70% of data belonging to the training data set.

We discovered that there were multiple cases of multicollinearity present in LRM\_0, particularly in the region, season, and quarters variables. Indeed, regions correlated with our county level data, while season and quarter data correlated with our month level data. County level data saw the most significant relationship with average housing sales prices (at the 99% level of confidence). Year and winter months (January and February) were both significantly related to sales price at the 95% confidence level, though year was positively related and the winter months were negatively. Given these results, we decided to run two more linear regression models (LRMPrice\_1 and LRMPrice\_2), with one excluding region level data, and the other excluding county level data. We included season in these two models to get a less granular view of the impact of time on sales price.

Of the two models, LRMPrice\_1 saw the largest R^2 value, indicating that the model explained more of the variance in average sales prices. Similarly to our first model, LRMPrice\_0, we saw a similar negative linear relationship between average sales price and winter that was significant at the 95% level of confidence. Additionally, in LRMPrice\_1, income was shown to also be significant at the 95% level of confidence and had a positive linear relationship with average sales price.In answering our second research question, we can conclude with high confidence that county level data generally had a greater predictive influence on average sales price than region level data did. Additionally, season and month level data, as well as income, both did a good job at predicting the average sales price.

To further evaluate the predictive performance of our linear regression models on average housing sales prices, we calculated the RMSE and MAE on both our training and test data sets, as seen in Figure 2. Subsequently, for purposes of comparison between all three models, we normalized both error measures. Coinciding with the R^2 values in explaining the variance in average sales price, both LRMPrice\_0 and LRMPrice\_1 had smaller normalized RMSE and MAE values when compared to LRMPrice\_2. Both LRMPrice\_0 and LRMPrice\_1 were more consistent in their predictive performance when being introduced to a new set of unseen data. The predictions from both LRMPrice\_0 and LRMPrice\_1 were off by about 12% of the average housing sales price for the RMSE and 7% for the MAE.

Given this information, we ultimately decided to use LRMPrice\_1 to answer our third research question and predict three specific prices given three different sets of parameters to fit into the model. Feeding our model county level parameters, along with other parameters based on the variables in our model (Units\_Sold, Units\_Pending, Active\_Inventory, Months\_Inventory, Median\_Days\_Market, New\_Listings, Season, and Income) we were successfully able to generate three predicted housing sales prices for Anne Arundel, Baltimore, and Montgomery Counties that all aligned within the range of prices for all three counties ($394,498, $371,397, and $550,485). As stated in the regression output of LRM\_1, County, Season, and Income, are among the variables that have the largest impact on predictive performance.

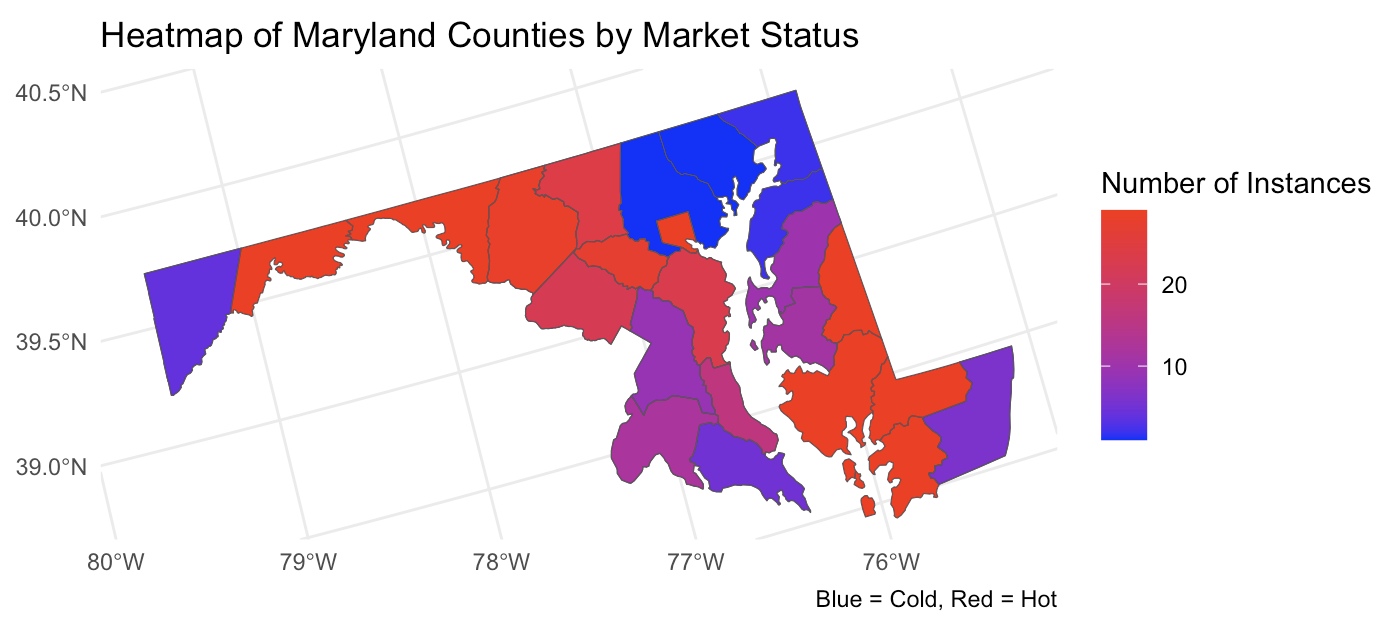
Figure 2) Normalized Residual Error Comparison of Linear Regression Models:

| Model | Train Normalized RMSE (% of avg price) | Test Normalized RMSE (% of avg price) | Train Normalized MAE (% of avg price) | Test Normalized MAE (% of avg price) |
| --- | --- | --- | --- | --- |
| 0 | 12.17 | 11.81 | 6.72 | 7.13 |
| 1 | 12.41 | 12.04 | 7.10 | 7.17 |
| 2 | 20.05 | 22.68 | 13.97 | 15.51 |

**Classification Findings - Identifying Hot and Cold Markets by County:**

In our analysis of hot and cold real estate markets (as referenced in Figure 3 below), the logistic regression model emerged as the most effective at first, demonstrating high accuracy across both the validation and test datasets. The model's ROC curves confirmed its strong discriminatory power, with an AUC score of 0.959 in the validation run, indicating excellent performance in distinguishing between hot and cold markets. Logistic regression was chosen for its interpretability and effectiveness in binary classification tasks, making it a robust choice for our analysis. It proved true with error rates of 10.6% and 7.6% in the validation and test datasets.

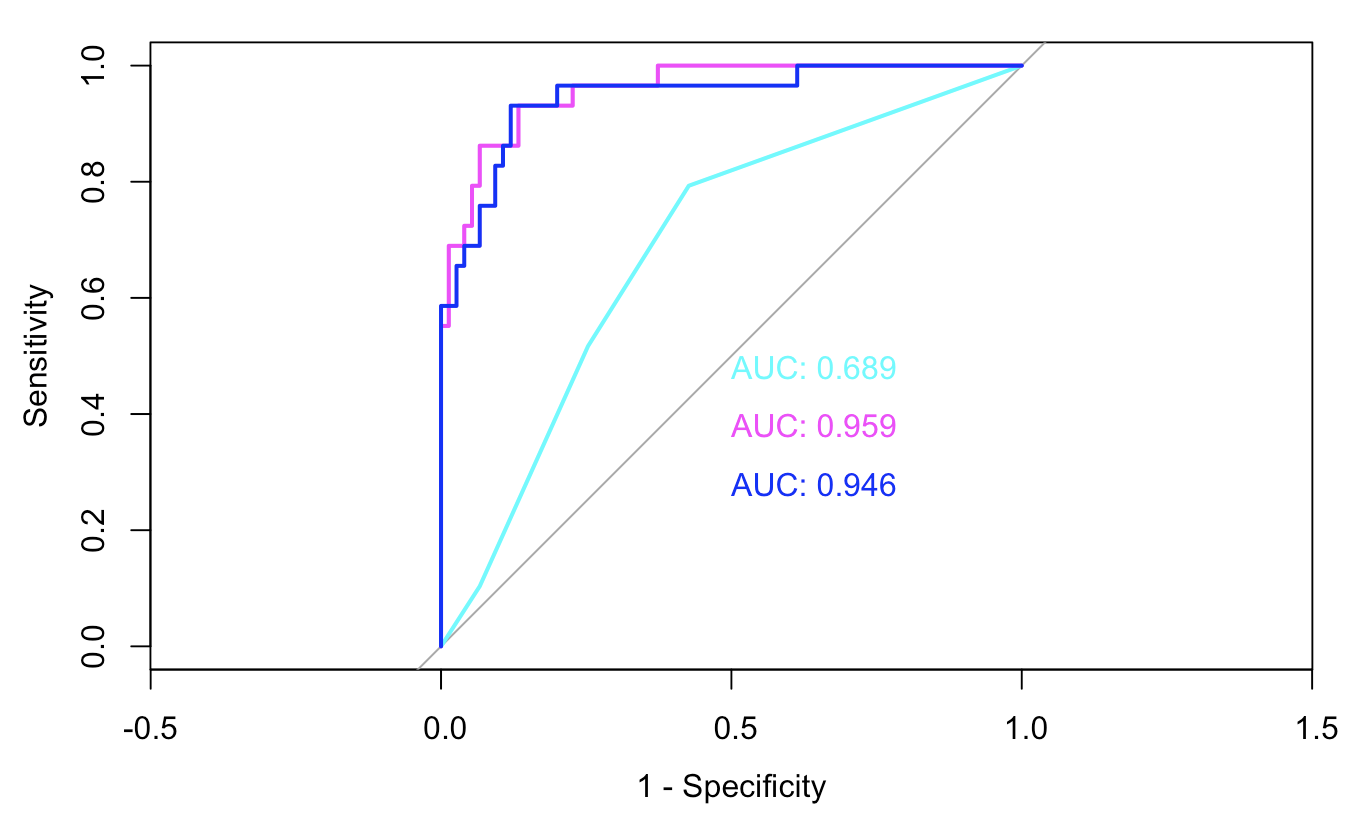
Figure 3) Hot and cold heatmap of market status



The K-Nearest Neighbors (KNN) model's performance was evaluated across different values of K, with the optimal K identified as 6. Despite this optimization, the KNN model's performance was slightly inferior to that of logistic regression, as reflected in its AUC score of 0.689. This suggests that while KNN can be useful in some classification contexts, it may not be as well-suited to this particular dataset and classification task, where the logistic regression model’s linear decision boundaries proved more effective.

The Linear Probability Model (LPM) also showed a marginally higher error rate compared to logistic regression, reinforcing that it might not be as effective for binary classification in this context. Specifically, the validation and test error rates for LPM were 12.5% and 9.5% respectively, and with an AUC of 0.946, indicating a moderately acceptable misclassification rate but still falling short of logistic regression's performance (as displayed in Figure 4).

Figure 4) ROC Curves for hot and cold classification



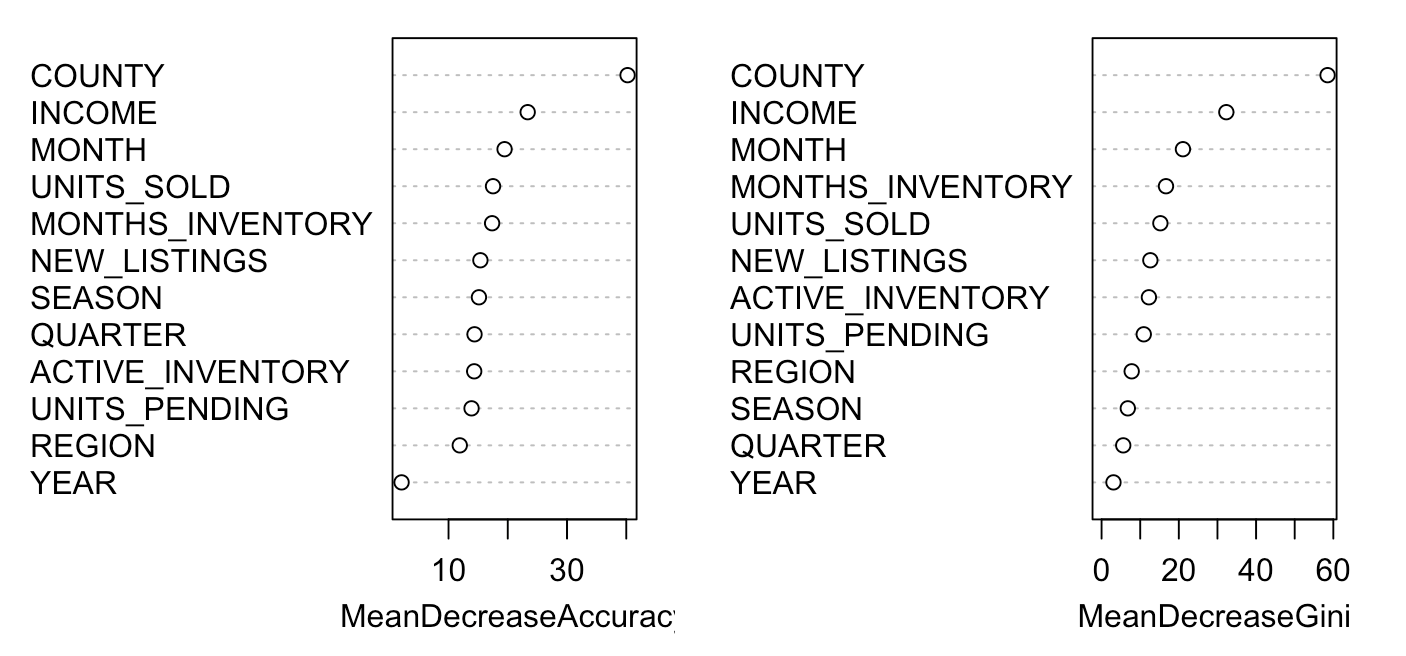
LR in pink, LPM in blue, and KNN in cyan

Finally, we approached ensemble methods to identify effective models for market classification. The pruned classification tree, designed for interpretability, was limited by a higher misclassification rate compared to logistic regression. With an optimal tree size of 4, the model's test error was approximately 3% higher than that of the logistic regression model, emphasizing the decision tree’s limitations in capturing the complexities of the data.

At this point in the analysis, the logistic regression model was the strongest at classifying hot and cold markets. However, the random forest model soon proved to be the best at this classification, as it demonstrated exceptional performance with an error rate of 6.7% on the test set. The model has 100% specificity, meaning it is excellent at correctly identifying cold markets with no false positives. This is crucial in contexts where you want to avoid mistakenly identifying a market as hot when it is actually cold. The sensitivity (approximately 75.86%) is lower, indicating that the model misses about 24% of hot markets by classifying them as cold. The overall accuracy of 93.33% and the low test error of 6.67% suggest that the random forest model is highly effective in classifying market status correctly for the majority of cases.

As shown in Figure 5, the final results (from random forest where m=4) indicate that county and income are the most influential variables in classifying a market as hot or cold. These variables have the highest impact on both the accuracy of the model and the purity of decision trees. Understanding the importance of these features helps in refining the model and guiding stakeholders on the key factors that drive market status. Given the perfect specificity, the model is reliable in not falsely flagging cold markets as hot, which could be particularly valuable if the cost of such a mistake is high.

Figure 5) Random forest model - variable importance plot



**VI. Conclusion**

To summarize, using Maryland housing sales data measured from January 2022 to May 2024, we were able to observe that county level data, in addition to season and income, both played a significant role in predicting average housing sales prices. We observed that average housing sales prices appear to be more determinant from housing market characteristics at the county level rather than at the regional level. Our LRMPrice\_1 linear regression model, which included county and season level data, was the best at predicting average housing sales price given a set of parameters with no multicollinearity.

In attempting to classify each county of Maryland as a hot or cold market based on sales volume, price, and median days on the market, we found that the random forest model had the best performance. By using these two variables to define market status, we effectively captured the dual aspects of market performance—price and speed—offering a robust way to identify and differentiate housing markets across Maryland. This classification helps in understanding regional market differences and provides valuable insights for stakeholders looking to make informed decisions in the real estate market.

To improve the accuracy and depth of future analyses, we recommend several enhancements. First, additional features should be considered, including the incorporation of more granular economic indicators and demographic data. Integrating geospatial data, such as distance to coastlines and mountains, may also be beneficial, as it could better capture location-specific market dynamics and provide a more precise visualization of hot and cold market regions. Moreover, if correctly identifying hot markets is a priority, it might be valuable to adjust the model or use techniques like resampling, adjusting class weights, or tuning hyperparameters to improve sensitivity.

Furthermore, we believe that future models should also incorporate the variety of housing available in these counties, as primary/secondary home status, housing type, and square footage would also help to further illustrate the individual dynamics that drive the unique housing markets of each county in Maryland. Finally, expanding the dataset to include a longer historical range and potentially incorporating data from neighboring states could significantly strengthen all of the models, providing a more comprehensive view of regional market behavior.

**VII. Appendix:**

Figure AA)

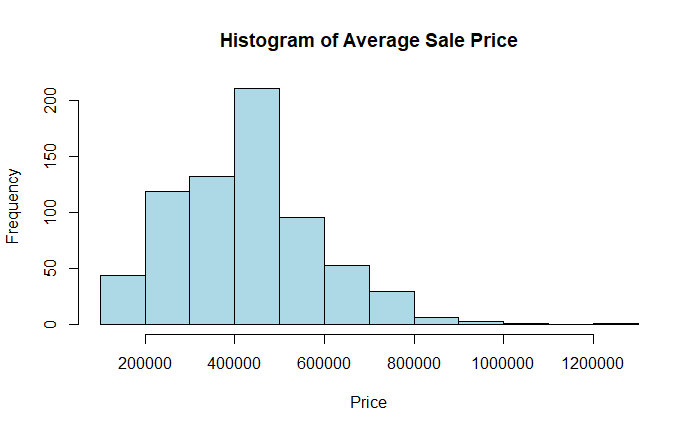


Figure AB)

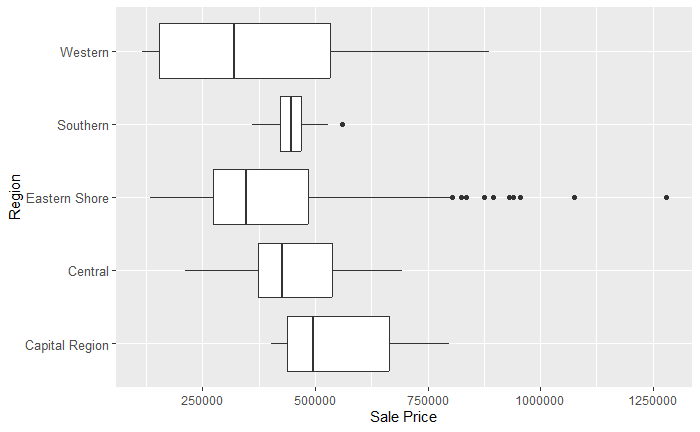


Figure AC) LRM 1 (preferred model) summary

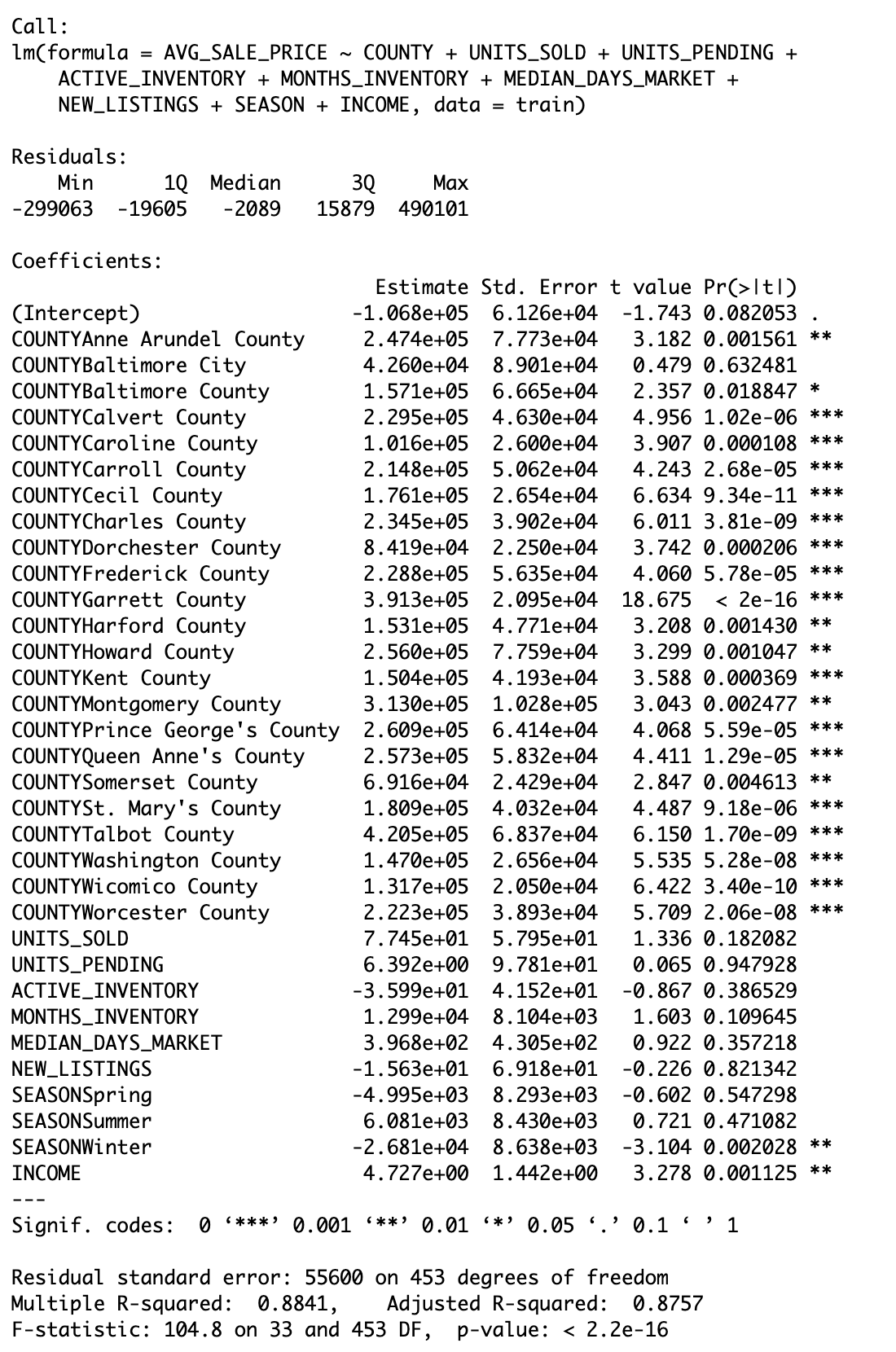


Figure AD) ROC curves for best KNN model. Red reflects our results on the test set using our best k value of 6.

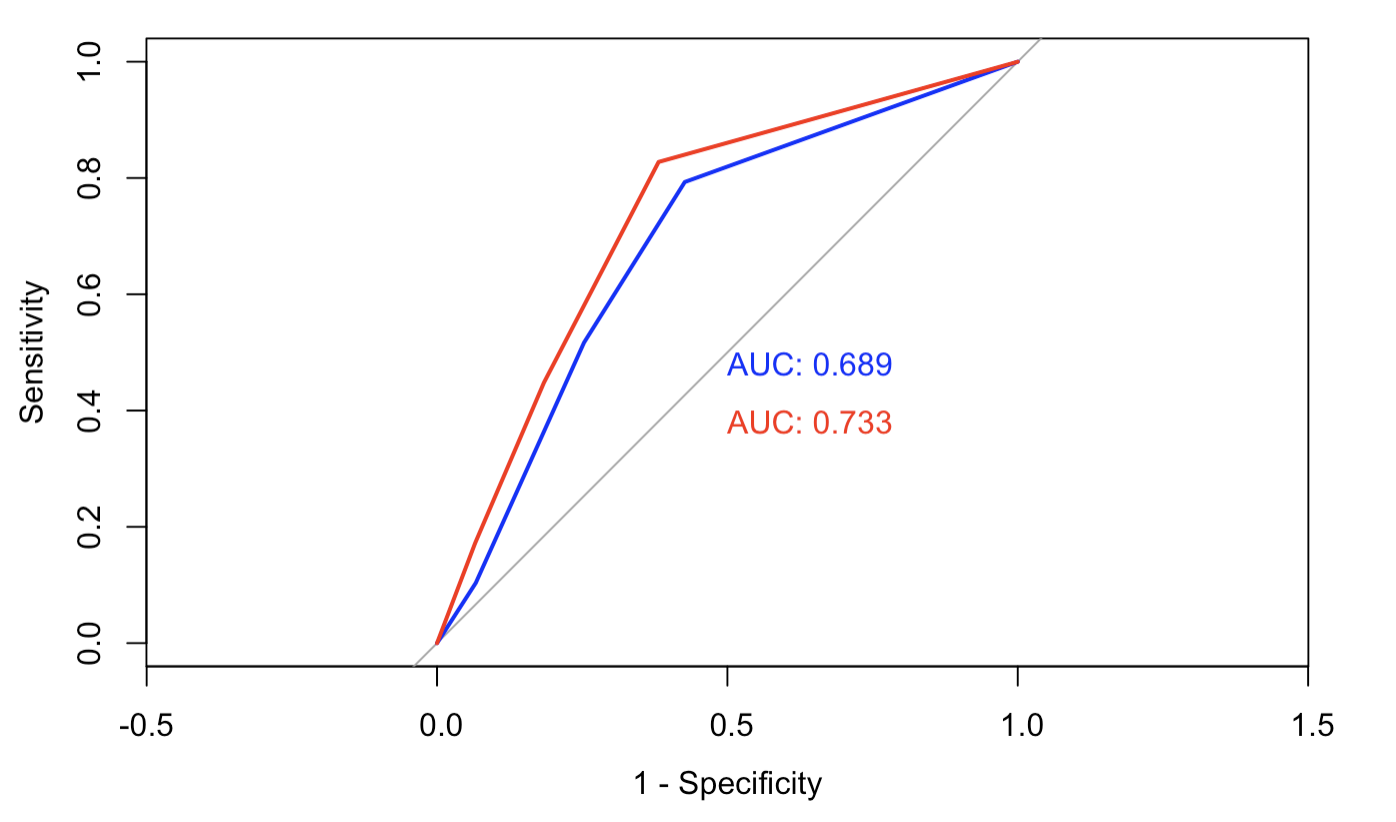
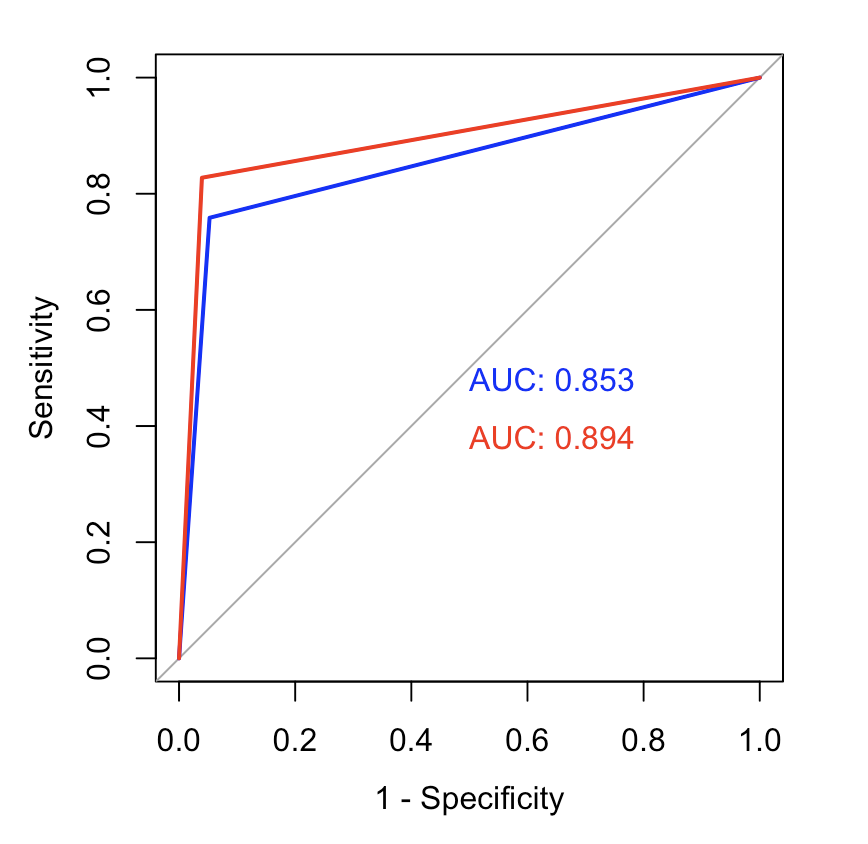


Figure AE) ROC for classification tree and logistic regression model (shows logistic is the better model)

****

**Citations:**

Title Page Image:

"The US Housing Market" Real Data, 9 Oct. 2018, https://real-data.com/statistics/the-us-housing-market/. Accessed 8 Aug. 2024.

* <https://www.mdrealtor.org/News-and-Events/Housing-Statistics>
* <https://www.bea.gov/data/income-saving/personal-income-county-metro-and-other-areas>
* <https://www.visitmaryland.org/article/maryland-regions>