City of South Bend

Residential profile analysis

Ashley Klesmit, Joseph Kosteck, Daniel Malee, Patrick McCullough | April 16, 2019

# Executive Summary

The City of South Bend requested an assessment of the city’s current housing situation from both a rental and owner-occupied perspective. The research team analyzed social, demographic, and economic data available from the U.S. Census as well as residential housing data sources (e.g. Zillow). With this data, models were created to find commonalities within the data at a Census Block Group level to group them into clusters or “profiles”. The team then used these “profiles” to perform statistical and inferential analyses to illustrate similarities and differences among granular areas of the City of South Bend.

The results shown in greater detail in this paper found five clusters of similar Census Block Groups. These five clusters have unique and differentiating characteristics, and they span many areas of the greater South Bend area. While some are naturally grouped in close geographic proximity, it is enlightening to see Block Groups from geographically different areas of the city and outlying areas clustered together. This paper will explain the process used to extract the data, the methodology behind the models created to cluster the data, and the statistical analysis done on the clusters to create “profiles”.

# Introduction and Question of Interest

The research team initially investigated creating an affordability calculator to help residents from different areas of the city determine if it was more beneficial to rent or buy in a particular neighborhood. This proved problematic for several reasons, chief among them being that many different types of financial analysis calculators regarding home affordability already exist and re-creating another is not the core competency of this team. Instead, the direction of the team changed to search for commonalities among areas of the city from a housing, economic, and geographic perspective. This would help address several questions of interest including the following:

* Are there areas of the city that have similar characteristics in terms of housing options?
* Are there housing or economic similarities among parts of the city that are not close in terms of geographic proximity?
* Are there areas of the city that are close in terms of geographic proximity, but differentiated by some other characteristics?

These questions are compelling because their answers can potentially help determine *why* there are differences. The results of our analysis will be included at the end of this paper, but we will be looking to extract patterns or anomalies in the profile data that may be useful for the city to further assess. The goal of the analysis is to give the research and tools to the experts who will hopefully find appropriate uses for it.

# Data Description

The primary source of data for this analysis was the American Community Survey (ACS) conducted by the U.S. Census Bureau. The ACS is an ongoing survey that provides vital information on an annual basis about the nation and its people. Information from the survey generates data that helps drive federal and state funding decisions. The data is made available via API from the U.S. Census Bureau website, and for our analysis we used the 2013-2017 ACS 5-year Estimates for St. Joseph County – which includes South Bend.

The U.S. Census Bureau divides geographic areas into two primary categories: Census Tracts and Census Block Groups. These are simple geographic perimeters defined by census officials. Tracts can be made up of one or more Block Groups. For St. Joseph County, there are a total of 76 Census Tracts and 227 Census Block Groups. The ACS provides a vast number of variables from which we selected approximately 138 variables for our analysis. We focused on using the Census Block Groups in our modeling as that provided more granular areas for potential clustering; however, Exploratory Data Analysis (EDA) was conducted on both Block Group and Tract data. Please see the appendices for a full list of variables used.

In addition to data from the U.S. Census Bureau, data was also extracted from the popular housing aggregator Zillow. The data was primarily used for exploratory purposes (next section) and not in the profile modeling due to its relative incompleteness. However, it was valuable to use this Zillow data as a mapping overlay to review how Zillow-defined boundaries interact with the Census Block Groups and Profiles. Similarly, neighborhood mapping information was pulled from the “Neighborhood Resources Connection (NRC)” as an additional geographic overlay for comparison purposes.

# Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is the process of analyzing data sets, often with visual methods. EDA is a way of becoming familiar with the data prior to creating any algorithmic models. Here, we will show the distribution of several attributes and their relationship to each other. The purpose of this exploration is to understand the main characteristics of the data so that we have an intuitive sense of the overall dataset. We may also discover issues with the data, such as missingness or outliers that may have to be taken into consideration when conducting analysis. We will also include several maps to show the spatial nature of the data we will be researching.

#### Data Boundaries

As stated above, both the Census Bureau and Zillow data is available at different geographic sizes. Shown below are the boundaries for these different geometries.

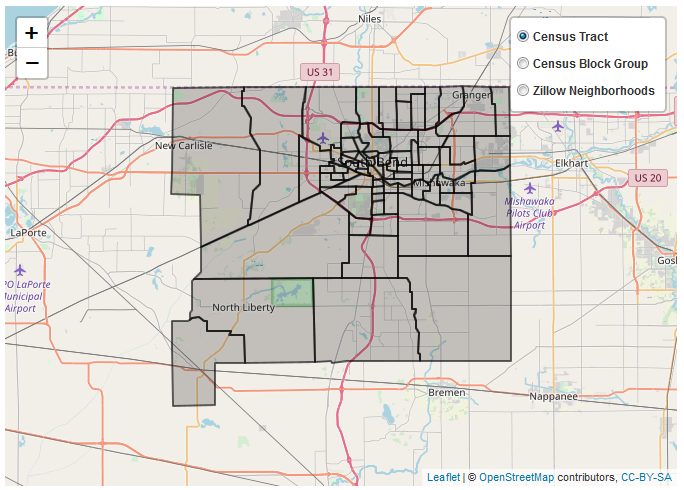


Figure 1: Census Tracts



Figure 2: Census Block Groups

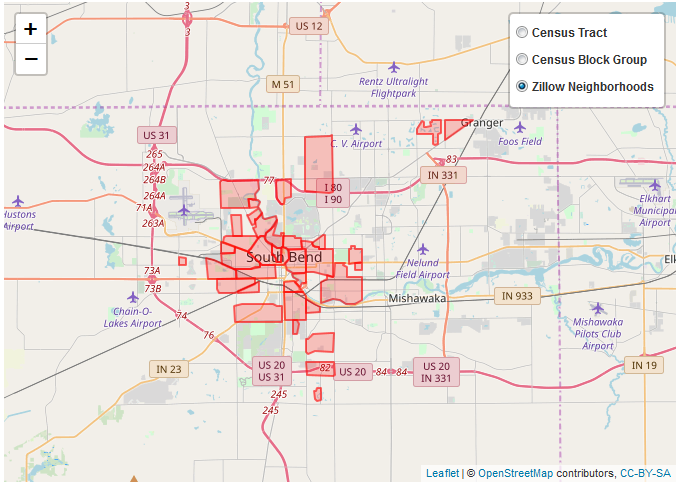


Figure 3: Zillow Neighborhoods

These boundaries illustrate that Tracts are geographically larger than Block Groups; in fact, Tracts are composed of multiple Block Groups. The boundaries also illustrate that neighborhoods (as defined by Zillow) fit within the Census definitions and in some cases may overlap different Census defined boundaries. The spatial data is presented here by itself, but the results of our analysis will show how these boundaries overlap with each other in the different profiles.

#### Data Patterns

Plotting multiple variables together can reveal patterns or trends in the data that otherwise might not be noticed. Figure 4 below shows a plot of Median Home Value versus Median Rent, with population density coloring the dots. The higher the population of the Census Tract, the darker the dot on the plot. This shows an unsurprising trend of increasing rent prices along with increasing home values; however, there do appear to be some outliers, such as the highest rent of $2,000 in an area with home values near $175,000.

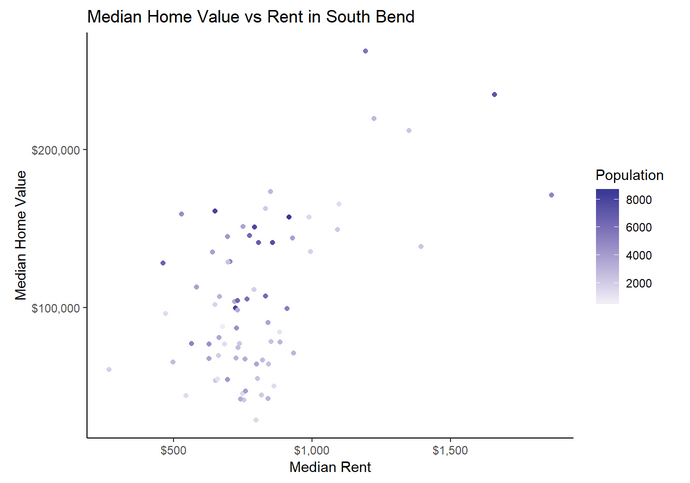


Figure 4: Median Home Value vs. Median Rent

Another interesting plot, Figure 5, shows the rent by number of bedrooms. The median value is the dark horizontal line in the middle of the white rectangles representing each number of bedrooms. The rectangle stretches vertically to cover the upper and lower quartiles, while the lines and dots cover relative outliers. This plot shows an expected increase in rental prices proportionally to the number of bedrooms; however, there are outlier and “NA” values to consider in our modeling. “NA” values are missing data that must be handled prior to any modeling on the data.

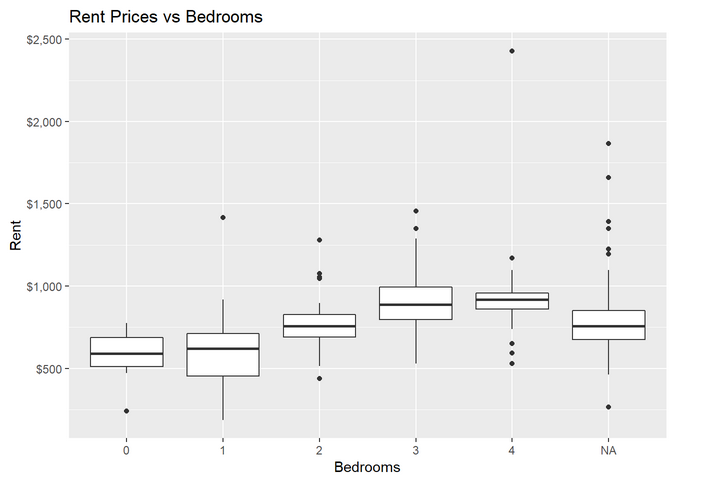


Figure 5: Rent Prices by Number of Bedrooms

While the Census data can show detailed information for a given point in time, other data sources such as Zillow can show how prices change over time. Figure 6 below shows how rental prices have changed over time for many neighborhoods of South Bend.

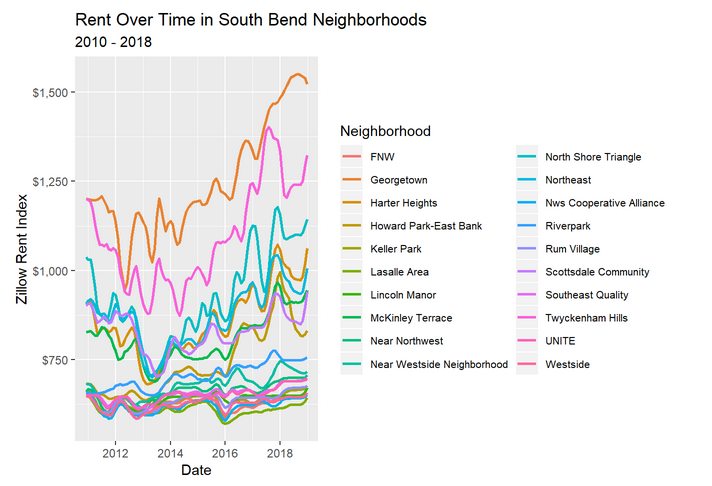


Figure 6: Rental Prices Over Time in South Bend

Since Figure 6 is quite busy and difficult to read, Figure 7 below displays similar information in a more aesthetically pleasing manner. Each dot represents a data point in the time series. The relative width of a given line shows how much that neighborhood’s rent has changed since 2010.

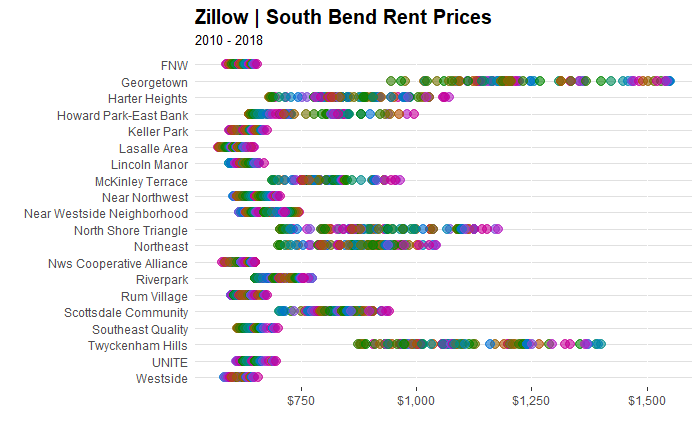


Figure 7: Rental Prices by Neighborhood

These plots show that there has been a wide range in prices in some neighborhoods (e.g. Georgetown), while very little change in others (e.g. Lincoln Manor).

Overall, the EDA shows that we have relatively complete data with somewhat expected patterns in the data. There are some NA values that will have to be handled prior to conducting any further analysis. In addition, there appears to be enough variation in the data to make cluster modeling an approach worth pursuing. These concepts and processes will be discussed in the Methods and Models section.

# Methods and Models

Clustering is a method of grouping observations together in a way that aggregates observations according to their similarities; observations in a given cluster have more in common with each other than observations in other clusters. We used k-means clustering to create distinct groups of the block groups. K-means clustering partitions the observations by iteratively establishing a center for each cluster, referred to as a centroid, and assigns each observation to a cluster by choosing the cluster whose centroid is closest—in terms of distance—to that observation. There are k clusters and the centroid of each cluster is the mean of all the values in the cluster, hence the name k-means. K-means clustering is a method of unsupervised learning and is a common technique for statistical data analysis used in many fields. Given our need to build profiles of geographic areas of the city, this was an ideal fit to quantitatively aggregate block groups with similar properties.

The goal of our data modeling exercise was to have to a relatively low number of clusters/profiles of South Bend census data. Because the Census Block Group data was more granular, we decided to use that dataset as opposed to the Census Tract data. The variables are the same, but there are more observations to work with in the block groups. Since our dataset was relatively wide, many variables relative to observations, a dimensionality reduction technique was required to eliminate correlation that would be problematic for a clustering algorithm. Correlated variables can introduce computational complexity into a clustering algorithm and potentially exert undue influence on the distance measurements used by the algorithm to group observations into clusters.

The approach we took was two-fold. First, we chose Principal Component Analysis (PCA) to reduce the dimensions of our census data. Second, we executed k-means clustering to group the data by common elements, using the results from the PCA analysis.

Before any modeling work could begin, the data needed to be made “complete”. Making a dataset complete means that all values with “NA” as the entry must be accounted for and handled in some way. There are many techniques used to handle NA data, from simply omitting any observations with NA entries to using complex imputation algorithms to populate the NA entries with sensible values. Dropping data is not acceptable in a dataset with a limited number of observations and where each one is singularly important. Instead, we chose to impute our missing data with the MICE statistical package.

MICE is an acronym for Multivariate Imputation by Chained Equations. It is a statistical package that creates multiple replacement values for multivariate missing data which helps account for the statistical uncertainty that could result from single imputations. Rather than taking a simple approach like applying the mean value for a particular variable to the missing data, the MICE package uses contextual information from other observations in making its imputed values. Essentially this balances statistical randomness with contextual observations to complete missing values with values that make sense relative to the rest of the data.

After imputing all data using the MICE package, the next step was to perform dimensionality reduction to eliminate correlation among the variables. As noted, we had 138 variables in the Census Block Group data, some of which were correlated. We applied a technique called Principal Component Analysis (PCA) which takes a wide dataset like ours and narrows it to a set of uncorrelated components. PCA takes real variables (e.g. income and race) and reduces them into components (e.g. PC1 and PC2) that are independent of each other. Essentially these are “new” variables that are independent yet are not obviously interpretable. In our case, PCA took our 138 variables and combined them into 35 components. The number of components (35) is determined by the PCA algorithm and is driven by the goal of accounting for as much variability within the data as possible. With the confidence that these 35 components were all independent of each other, we could now move on to a clustering model.

The first step was to determine the optimal number of clusters for a dataset. This is done by using the components created during PCA and splitting it into k sets, where k ranges from one to the total number of observations. The optimal number of clusters is when the creation of more sets does not significantly improve the variance within each cluster. In this dataset, the optimal number of clusters was five. The number of requested clusters (5) is then used in the k-means algorithm, which partitioned our 227 block group observations into five clusters, in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This is a complex way of saying it attempts to find its most similar neighbors and groups them together. Recall that we are using the combined “components” from PCA as opposed to readable variables.

Once complete, the observations are all assigned to a cluster: one through five. We added this value back into the original dataset as a new variable. The end result is our 227 block groups, now with 139 variables, with the final variable being the cluster (or “profile”) to which they most closely align.

# Model Results and Conclusions

The goal of the analysis and model building was to produce “profiles” of the different areas of South Bend. A “profile” in this context is simply a grouping of Census Block Groups that shared enough statistical similarities that the cluster modeling algorithm determined they should be grouped together. Below is a view of all five profiles plotted together on a map, along with an outline of the South Bend city limits:

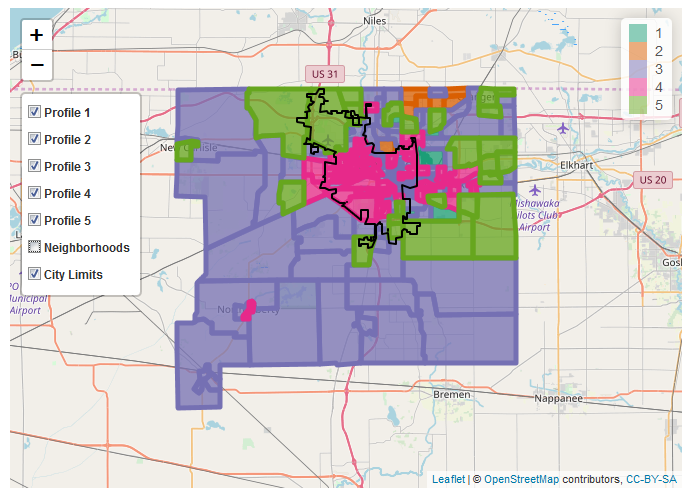
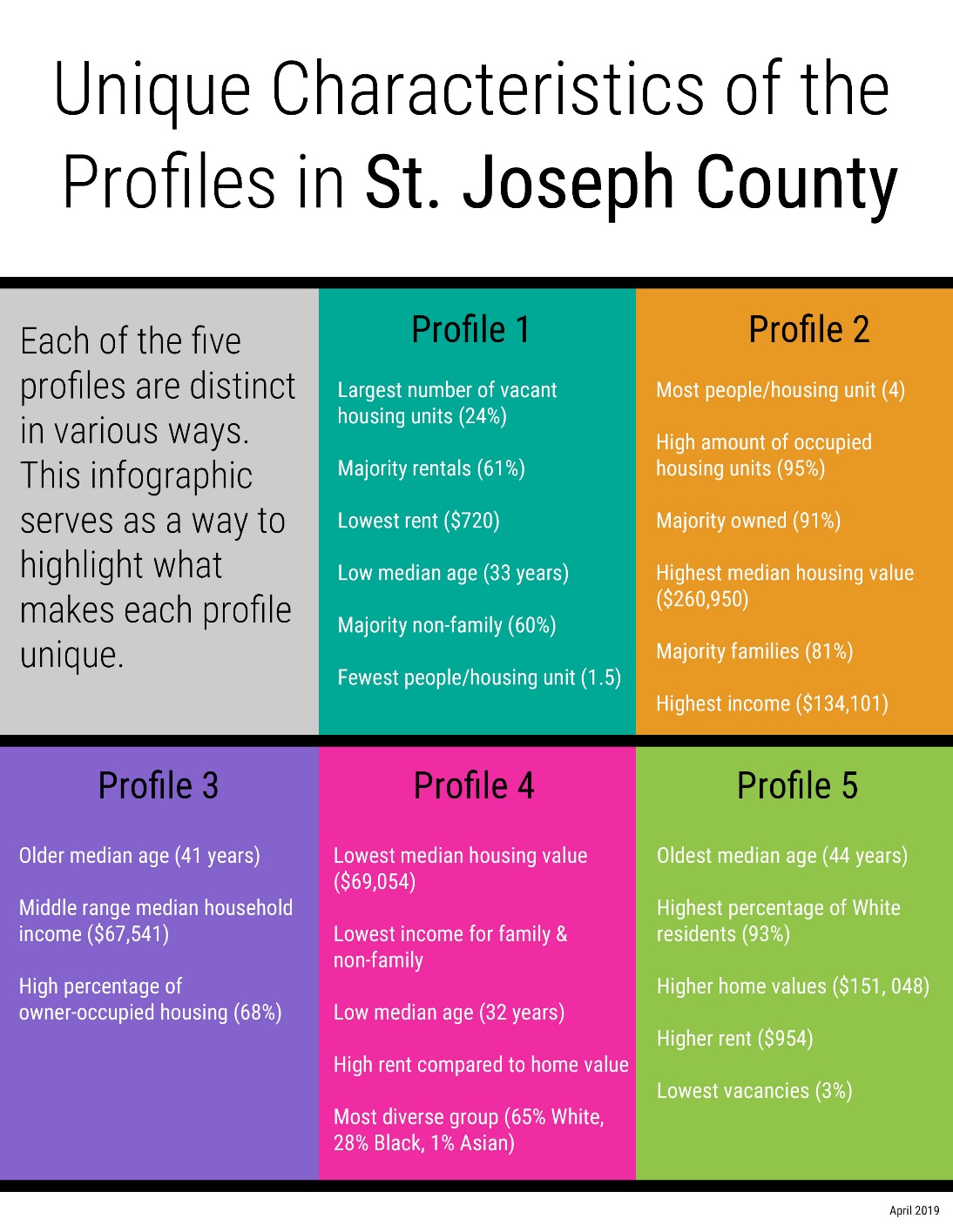
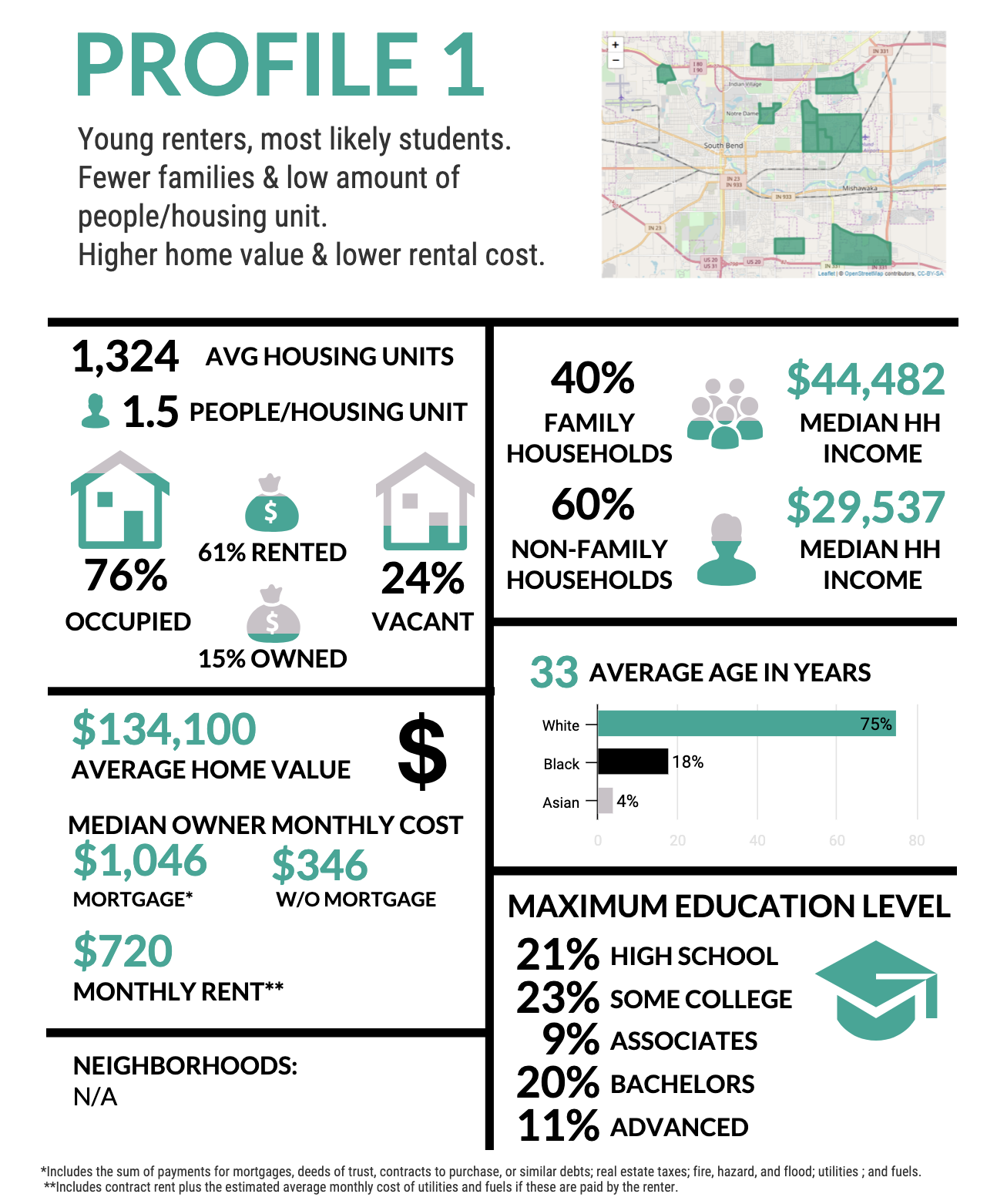


Figure 8: “Profile” Geographic Overlay

It is clear that the profiles cover different areas of the city, with some interesting overlaps. We see that the majority of the city is covered by profiles 3, 4, and 5. It is also clear that profiles 3 and 4 have some interesting overlaps. This is more easily viewed using the included clickable HTML file, which will allow for selection of the Profile check boxes and the ability to zoom to a lower level.

Each profile will be detailed with common statistics to allow for simple comparison among profiles. For further analysis of each profile, please refer to the detailed profile infographics and summaries. In addition, it is insightful to compare the profile’s unique characteristics side-by-side in the infographic below. In the summaries, if a specific variable or statistic is not mentioned, it can be assumed that it is otherwise consistent with St. Joseph County as a whole. By using the observation with the minimum distance from each observation to the mean of its profile calculated in the clustering process, we were able to get an idea of which block group best represented its profile. Likewise, by using the observations with larger distances, we specified block groups that stood out relative to the profile for one reason or another.





Relative to St. Joseph County, Profile 1 is characterized by:

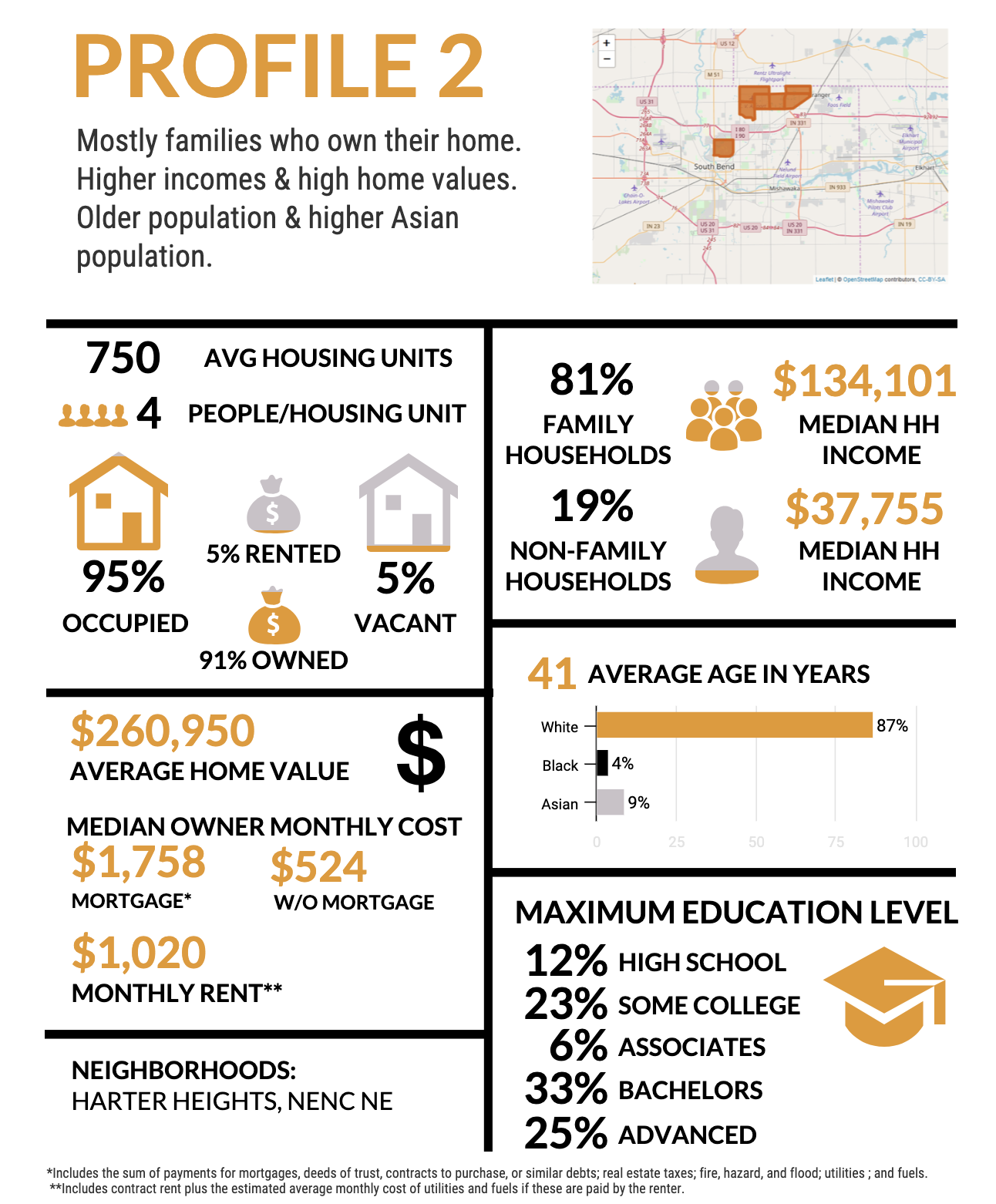
* Higher home values ($134k vs $106k)
* Higher home ownership costs with a mortgage ($1,046 vs $972)
* Lower family household incomes ($44k vs $62k)
* Slightly lower rental costs ($720 vs $773)
* Younger age (33.4 years vs 38.5 years)
* Lower number of people per housing unit (1.5 vs 2.3)
* Lower percentage of occupied housing units (76% vs 87%)
* Much lower percentage of owner-occupied housing units (15% vs 59%)
* Much higher percentage of renter-occupied housing units (61% vs 28%)
* Much lower percentage of family households (40% vs 64%)
* Higher labor force participation rate (55% vs 50%)

Based on these observations, Profile 1 can be described as block groups with a higher proportion of young workers who primarily rent and are less likely to have families.

Notable Block Groups:

**Block Group 2, Census Tract 115.05** is the best representation of Profile 1 due to a low median age (30.5 years), higher home values ($162,800 median home value), and the percentage of family households (40% of family households).

**Block Group 3, Census Tract 113.01** is the most distinct block group of Profile 1 due to a very low median age (25.7 years), a low number of people per housing unit (.98), a very low percentage of family households (19%), and higher rental costs ($883).



**Profile 2:**

Relative to St. Joseph County, Profile 2 is characterized by:

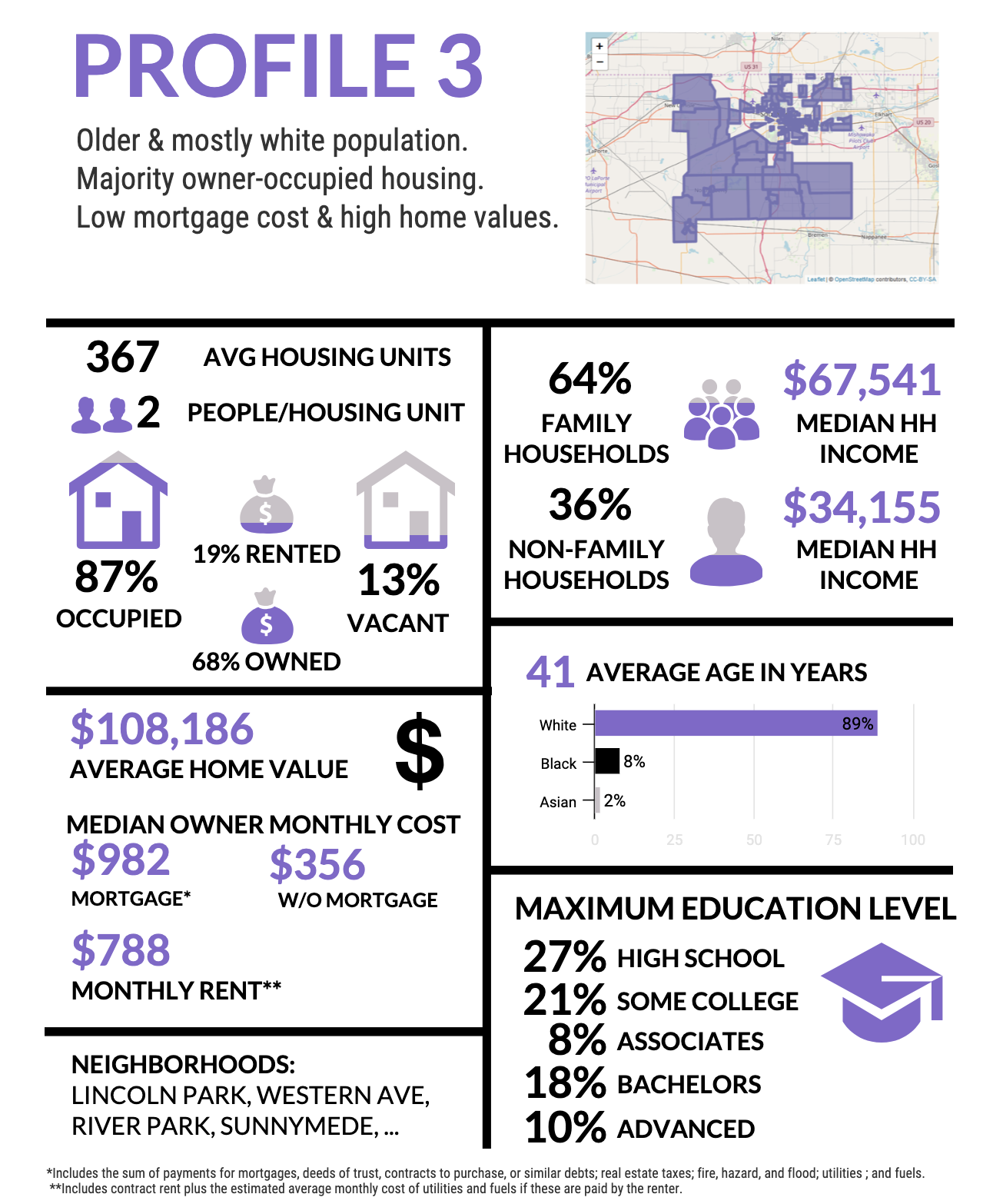
* Much higher home values ($260k vs $106k)
* Much higher home ownership costs with a mortgage ($1,758 vs $972)
* Much higher home ownership costs without a mortgage ($524 vs $349)
* Much higher family incomes ($134k vs $62k)
* Much higher rent ($1,020 vs $773)
* Slightly older age (41.0 years vs 38.5 years)
* Much higher people per housing unit (4.1 vs 2.3)
* Higher percentage of occupied housing units (95% vs 87%)
* Much higher percentage of owner-occupied housing units (91% vs 59%)
* Much lower percentage of renter-occupied housing units (5% vs 28%)
* Lower percentage of Black residents (4% vs 13%)
* Higher percentage of Asian residents (9% vs 2%)
* Much higher percentage of family households (81% vs 64%)
* Higher percentage of people not in the labor force (36% vs 29%)

Based on these observations, Profile 2 can be described as block groups with higher home values, more people per home, and whose inhabitants own their own homes, have higher incomes, and tend to have families.

Notable Block Groups:

**Block Group 1, Census Tract 113.04** is the best representation of Profile 2 due to a higher median family income ($121,616), a higher percentage of owner-occupied housing units (94%), a higher percentage of family households (78.8%), and a higher median age (49.2 years old).

**Block Group 1, Census Tract 113.05** is the most distinct block group of Profile 2 due to a lower median home value ($199,700), lower median family income ($120,179), and a higher percentage of family households (86%).



**Profile 3:**

Relative to St. Joseph County, Profile 3 is characterized by:

* Slightly older age (41.4 years vs 38.5 years)
* Slightly higher percentage of owner-occupied housing units (68% vs 59%)
* Slightly lower percentage of renter-occupied housing units (19% vs 28%)
* Higher percentage of White residents (87% vs 79%)
* Lower percentage of Black residents (7% vs 13%)

Based on these observations, Profile 3 can be described as block groups that are representative of St. Joseph County as a whole, albeit with a slightly older population that is more likely to own their homes and has a higher proportion of White residents.

Notable Block Groups:

**Block Group 1, Census Tract 12** is best representative of Profile 3 due to due to a higher median home value ($120,100), a median family income ($64,500) consistent with the rest of Profile 3, percentage of owner-occupied homes (91%), percentage of White residents (86%), and percentage of family households (64%).

**Block Group 3, Census Tract 10** is a distinct block group in Profile 3 due to lower home values (median value of $33,000), higher rental costs ($904), a lower percentage of owner-occupied housing units (11%), a lower percentage of family households (14%), and a lower percentage of White residents (77%).

**Block Group 1, Census Tract 7** is a distinct block group in Profile 3 due to lower home values (median home value of $73,600), a lower median age (34.8 years), a higher percentage of Black residents, and a higher percentage of vacant units (21%).



**Profile 4:**

Relative to St. Joseph County, Profile 4 is characterized by:

* Much lower home values ($69k vs $106k)
* Lower home ownership costs with a mortgage ($817 vs $972)
* Lower rental costs ($728 vs $773)
* Much lower family incomes ($40k vs $61k)
* Much younger age (32.5 years vs 38.5 years)
* Lower percentage of owner-occupied housing units (45% vs 59%)
* Higher percentage of renter-occupied housing units (38% vs 28%)
* Lower percentage of White residents (62% vs 79%)
* Higher percentage of Black residents (27% vs 13%)
* Higher unemployment rate (10% vs 6%)

Based on these observations, Profile 4 can be described as block groups with lower home values, lower home ownership costs with a mortgage, slightly lower rental costs, and much lower family incomes. The population of Profile 4 tends to be younger, contains a higher percentage of Black residents, a lower percentage of White residents, and is more likely to be unemployed.

Notable Block Groups:

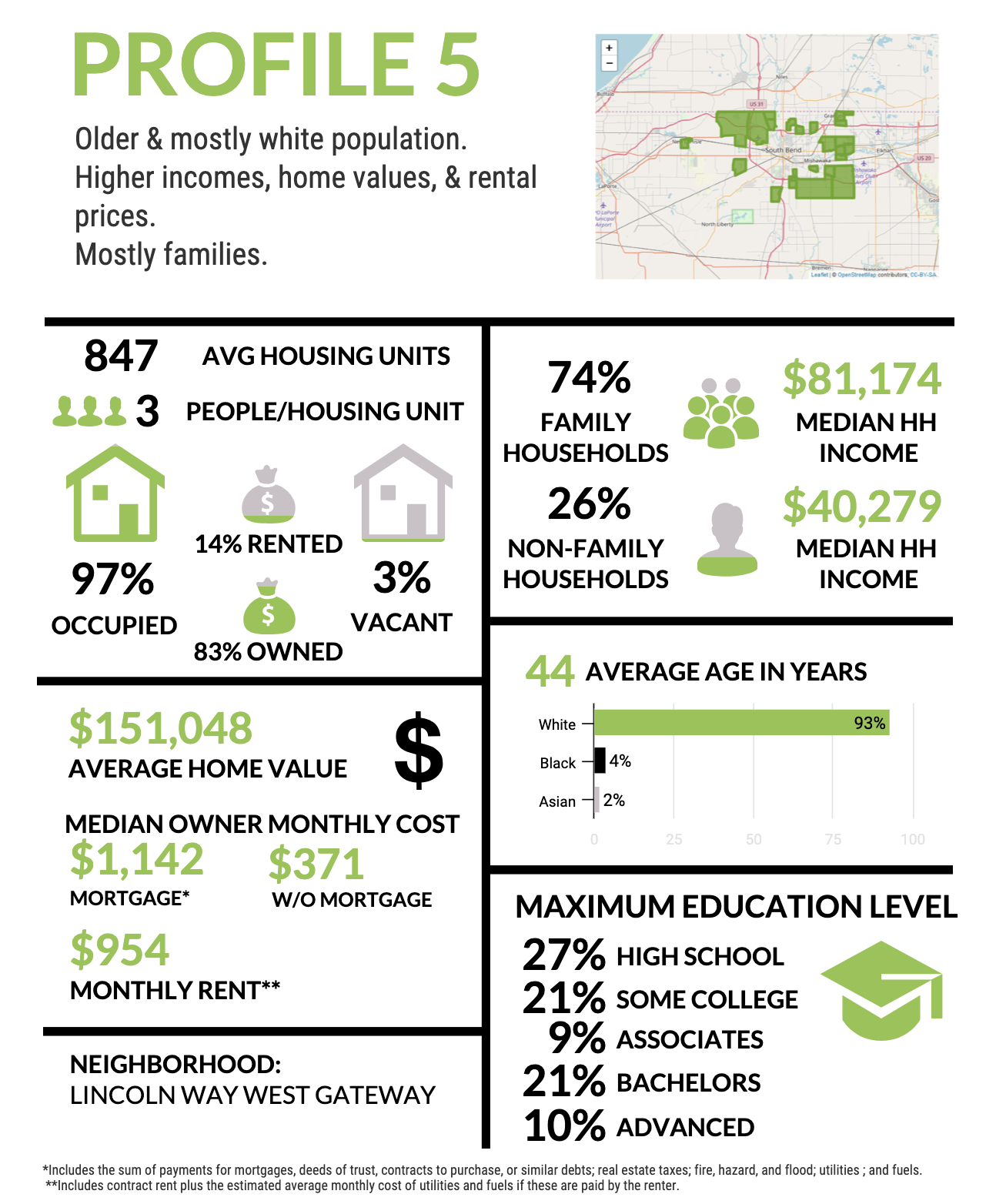
**Block Group 1, Census Tract 29** is best representative in Profile 4 due to lower home values (median home value of $54,600), lower rental costs ($658), a lower median age (29.1), a lower percentage of occupied housing units (77.9%), and a high unemployment rate (89%). However, this block group is not representative of Profile 4 due to its higher percentage of families (82%) and a very high percentage of Black residents (52%).

**Block Group 1, Census Tract 27** is an extreme representation of Profile 4 due to a lower median age (30.2 years), lower home values (median home value of $41,200), lower family incomes ($32,0310), a lower percentage of occupied housing units (43%), and a high unemployment rate (18%). This block group also has a higher percentage of family households (72%) and higher rental costs ($754).

**Block Group 3, Census Tract 34** is an extreme representation of Profile 4 due to a lower median home value ($51,700), lower rental costs ($711), a lower percentage of occupied housing units (65%), and a higher percentage of Black residents (32%). This block group also skews older (39.2 years) and has a higher percentage of families (84%).

**Block Group 4, Census Tract 22** is a distinct block group in Profile 4 due to a lower median home value ($42,000), a much lower median age (19.8 years), much higher rental costs ($977), a lower percentage of White residents (46%), and a higher percentage of family households (84%). This block group also has an occupied-housing rate of 88% and a median family income of $42,083.

**Block Group 1, Census Tract 26** is notable for having home values that are lower than the Profile 4 average (median home value of $60,200) yet has higher median rents ($834), has a very low median age (19.7 years), has a lower percentage of White residents (46%), and contains more families (73%) than Profile 4, on average.

****

**Profile 5:**

Relative to St. Joseph County, Profile 5 is characterized by:

* Higher home values ($151k vs $106k)
* Higher home ownership costs with a mortgage ($1,142 vs $972)
* Higher rental costs ($954 vs $773)
* Older age (43.6 years vs 38.5 years)
* Slightly higher number of people per housing unit (2.6 vs 2.3)
* Higher percentage of occupied housing units (97% vs 87%)
* Much higher percentage of owner-occupied housing units (83% vs 59%)
* Lower percentage of renter-occupied housing units (14% vs 28%)
* Much higher percentage of White residents (91% vs 79%)
* Higher percentage of family households (74% vs 64%)

Based on these observations, Profile 4 can be described as block groups with higher home values and correspondingly higher home ownership costs with a mortgage, higher rental costs, and slightly more occupants per home whose inhabitants are more likely to own their homes. The population of Profile 5 tends to be older, with a higher proportion of White residents, and a higher percentage of family households.

Notable Block Groups:

**Block Group 2, Census Tract 113.05** is representative of Profile 5 due to a higher median home value ($146,300), higher rental costs ($1,304), an older median age (44 years), more people per housing unit (2.66), a higher percentage of White residents (85%), and a higher percentage of family households (75%).

**Block Group 4, Census Tract 109** is an extreme representation of Profile 5 due to a higher median home value ($177,700), higher median age (46.3 years), and a higher percentage of owner-occupied housing units (91%).

# Ethical Considerations

As we were working with demographic and economic data, ethical considerations were at the forefront of our mind the entire project. We were very sensitive to any kind of negative portrayal of any group or profile. This included considering statistical presentation and things like color schemes for the maps we created.

That said, it is important to remember that all analysis was conducted using data collected from the U.S. Census Bureau ACS survey, which is voluntary and unverified. That is not to say that the ACS is not a professionally conducted survey, it is merely to point out that there may be gaps in their data collection process. There may also be internal biases baked into the questions asked on the survey. In addition, the ACS provides hundreds of variables for possible analysis. Through thoughtful consideration, we chose 138 of them for analysis, but it must be acknowledged that our unconscious bias may have had an impact on the variables we chose.

Much of the data that we used from the American Community Survey are estimates that the U.S. Census Bureau calculates based on surveys that they conduct. The ACS itself does not send surveys to every household, so the values that are published are calculated estimates rather than the actual survey responses of the entire population. The fact that our data is ACS estimates means that there is some uncertainty, but we do not think that this should be too worrisome as it is simply the best and most representative data that is available, at least publicly.

Another ethical consideration is the data imputation process we chose. Recalling that imputation is the process by which missing data can be handled, we chose a statistical algorithm called MICE to accomplish this task. The package uses statistical techniques to attempt to most appropriately fill in the missing data, but these are not real answers given by human beings, and therefore must be acknowledged. Likewise, when grouping the block groups into profiles, we did not include variables that contained more than 10% missing values because we felt that there was too much missing data for imputations to accurately account for. This should not have had a strong effect on our results as 113 of the 138 variables were still used and many of the omitted 25 variables were rather specific, such as the amount of people who commute by walking or the number of vehicles that a worker had available.

The determination of the appropriate number of PCA components and number of clusters is inherently a subjective task. There is no hard and fast rule to specify exactly how many components or clusters should be used. Rather, it is up to the individual to determine the appropriate number based on when including more clusters would qualify as the point of diminishing returns and the type of analysis to be conducted. We are confident that our methods were appropriate as the PCA components captured approximately 84% of the variation in the original data and our clusters provided meaningful distinctions between Profiles without being too specific.

In conducting the final analysis with summary statistics of each profile, many of the variables that we worked were median values, such as home value, income, and rental costs. Therefore, the profile summaries of these variables are the averages of the medians for each block group. Our final average for each profile may not be approximate of the actual average, but we still feel that this is the best way to convey the statistics of each profile.

Throughout the team’s analysis, we have been very careful to not inadvertently inject bias into our analysis. We ensured that our team did not have a conclusion in mind when conducting analysis so that our biases and opinions would not affect any conclusions. We aimed to determine various aspects of housing and rental statistics in South Bend and to keep any of our individual knowledge about the city out of our work.

# References

Link to U.S. Census Bureau ACS Survey:

<https://www.census.gov/programs-surveys/acs/>

Link to U.S. Census Bureau Variable Glossary:

<https://www.census.gov/glossary/>

Zillow research data:

<https://www.zillow.com/research/data/>

South Bend Neighborhood Resources Connection:

<http://www.nrc4neighborhoods.org/neighborhoods/>

Principal Component Analysis (PCA) reference:

<https://www.dezyre.com/data-science-in-python-tutorial/principal-component-analysis-tutorial>

K-means clustering reference:

<https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68>

<http://www.learnbymarketing.com/methods/k-means-clustering/>

<https://www.dezyre.com/data-science-in-r-programming-tutorial/k-means-clustering-techniques-tutorial>

# Appendices

Census Tract Data:



Census Block Group Data:



Census Variables Pulled for Analysis:



Project Log:



Profile summaries with block group data:



Clickable HTML Document:



Code Github Repository: <https://github.com/jkosteck/DS-Now-Final-Project>

Code Zip File:

