**Group 3 Project:**

**Austin Affordable Housing & Demographics**

By Mike Owona Ongbwa, Celina Turner, & Mychael Solis-Wheeler

**Initial Analysis of the Data**

The purpose of this project was to examine the relationship between demographics and affordable housing for the Austin area based on zipcodes. The Affordable Housing Directory dataset was based upon public data from 2019 by the City of Austin, which is a listing of income-restricted affordable housing funded and/or incentivized by the City of Austin and/or the Austin Housing Finance Corporation (AHFC). This dataset lists only completed projects or those in which a building permit was issued by the City of Austin [1]. It was assumed that this dataset was primarily based on affordable housing units for rent and no warranty was made regarding the specific accuracy or completeness of this dataset by the City of Austin, which our team was able to download from the data.gov public data website.

However, this dataset contained valuable items such as zipcodes, zipcode per total affordable units, and types of household incomes per total affordable units, companies that manage those units, and location. It did not contain data based upon the demographics in each zipcode zone or in data on what demographic communities have affordable housing units available to them the most within each zipcode zone. Additionally, this dataset also lacked other publically funded buildings and/or projects available in each zipcode zone.

Interesting questions that arose were many from this data. For instance, were there trends of low-affordable housing rates to low demographics of racial minorities in zipcode zones or vice versa? Was there a shortage of low-affordable housing units based from population density zipcode zones or vice versa? What were the publicly available resource allocations of buildings available within each zipcode zone and what demographics are they made available to the most in those zipcode zones? Therefore, we were curious to see the demographics makeup for the different zipcodes across the Austin region, which initiated our team to seek out four other datasets. The Neighborhood Reporting Areas based on the 2010 US Census data was downloaded directly from the City of Austin’s public website [2]. In contrast to the Affordable Housing Directory dataset, this data on housing units was assumed to be both publicly and privately owned, inclusive of rent housing. The Austin High School Graduation Rate from 2016 [3] , Austin Public Health Locations from 2019 [4] , and the Austin Public Internet Locations (Austin Google FiberGB Speed Internet Locations) sites signed up for internet speed through Google from 2019 [5] were the three other datasets publicly available from the City of Austin and were also downloaded from the data.gov public website. The valuable items from Neighborhood Reporting Areas dataset were the demographic percentages within each neighborhood that could be categorized by zipcodes. Other valuable items also included public building information based on address or zipcode and types of buildings present per zipcode zones from the three other datasets.

By examining the relationship between demographics and affordable housing available by zipcode zones of the Austin area, insights could be gained and applied to many business applications for the City of Austin, private businesses, and to the general public seeking housing information. For instance, direct business applications could be for the City of Austin to determine and track if their housing policies and current planning align with their Austin city goals and outcomes. Additionally, Austin city planners or contractors can determine what public resource allocations are needed the most in each zipcode zones based on affordable housing income and/or by the total affordable housing available. Furthermore, these insights could also directly inform potential shoppers or renters of housing units seeking the most affordable housing options in the Austin area. The indirect business applications could also impact and benefit real estate businesses, buyers and sellers of private property near these affordable housing units or near public buildings, city contractors, and private businesses seeking to economically impact, support additional affordable housing, and/or to public resources to their employees all based the insights found for each zipcode zones.

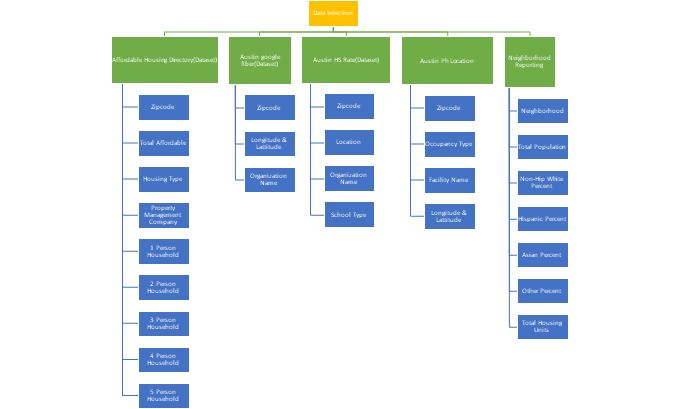
The potential uses from the different variables within these datasets include combining their data to create new variables that could better encompass or concise these variables into one. For instance, having a dataset with a new variable named SiteType to put data from School Type, Housing Type, Occupancy Type, etc., found across these datasets could designate what type of public buildings are most distributed within each zipcode zone. Additionally, using Total Affordable Units from the Affordable Housing Directory dataset to match with Total Housing Units by zipcode (by combining all neighborhoods assigned assigned to their respective zipcode) could lead to calculating an affordable housing rate for each zipcode within the Austin area. These are just a few examples in how using the different variables from the dataset can create additional value and insights not initially present or considered with just one of these datasets on their own.

One of the biggest contributions for our team in how to proceed with data cleaning, data merging, and data visualization were the creations of logic flow diagrams to guide our project’s progress. Although we initially had a plan to begin the project with Python coding, we did not start a logic flowchart until after merging our data initially. It was after the viewing of the results of that merge was when the realization that logic flow diagrams were needed to fully visualize and plan what our team wanted to achieve and visualize through our output file results. Even when our first logic flowchart was developed, it was continually updated and refined to best capture how we wanted our post-merged results to be in our output files before visualizing the results. Below are the finalized logic flow diagrams (Figures 1-3) that we developed to help our team navigate what data we wanted to use and what wanted to achieve for our final data visualizations.

**Project Flow Diagrams**

*Original Datasets* *Flow Diagram* (shown below)

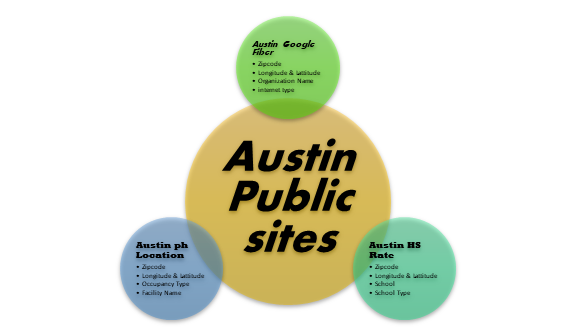
* 5 original raw files
* Data cleaning involved manual cleaning, python cleaning, creating and dropping columns, and renaming columns needed for our analysis.
* Common element (as a column) between all files was Zipcode.
* 2 file outputs were created as products from th 5 original raw files.



**Figure 1-Flow Diagram of Original Datasets from Affordable Housing Directory RAW 2019, Neighborhood Reporting Areas Table I RAW 2010, Austin HS Grad Rates RAW 2016, Austin Google FiberGB Speed Internet Locations RAW 2019, Austin\_Public\_Health\_Locations\_RAW\_2019 csv files**

*File 1 Output Flow Diagram: Austin Public Sites* (shown below)

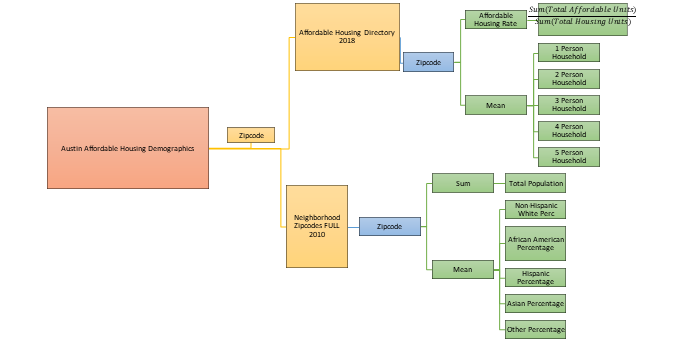
* File 1 output (Austin Public Sites) was produced mainly from Python coding language by combining data from revised files of Affordable Housing Directory 2019, Austin High School Graduation Rate 2016, Austin Public Health Locations 2019, and Austin Google FiberGB Speed Internet Locations 2019.
* Data output primarily consisted of public building names, building types, longitude, and latitude listed by the common element of Zipcode.
* File 1 data composed of both numerical and of characters with 5 variables (columns): Zipcode, SiteName, SiteType, Longitude, and Latitude.
* No calculations were made in this file.



**Figure 2-Flow Diagram of File 2 (Austin Public Sites) from** **Affordable Housing Directory 2019, Austin High School Graduation Rate 2016, Austin Public Health Locations 2019, and Austin Google FiberGB Speed Internet Locations 2019 csv files**

*File 2 Output Flow Diagram: Austin Affordable Housing Demographics* (shown below)

* File 2 output (Austin Affordable Housing Demographics) was produced mainly from Python coding language by combining data from revised files of Affordable Housing Directory 2019 and Neighborhood Zipcodes FULL 2010.
* Most file 2 data was numerical with 15 variables (columns): 1PersonHousehold, 2PersonHousehold, 3PersonHousehold, 4PersonHousehold, 5PersonHousehold, Zipcode, TotalAffordableUnits, NonHispanicWhitePercentage, AfricanAmericanPercentage, HispanicPercentage, AsianPercentage, OtherPercentage, Total Population, TotalHousingUnits, and AffordableHousingRate.
* Means were calculated for each household incomes for racial demographic percentages per Zipcode.
* Sums were calculated for total population, total affordable housing units, and for total housing units per Zipcode.
* AffordableHousingRate was a new variable to calculate the rate of affordable housing present per Zipcode by dividing the sum of total housing units from the sum affordable housing units previously calculated.



**Figure 3-Flow Diagram of File 2 (Austin Affordable Housing Demographics) from** **Affordable Housing Directory 2019 and Neighborhood Zipcodes FULL 2010 csv files**

**Phase I Data Cleaning: Manual Cleaning**

Generally, all of the datasets used for this project were of decent quality, comprehensive, and had most columns filled with data. However, manual cleaning methods were still necessary before our team used any code to further clean and process the datasets used.

When downloading and opening each file to preview the data, there were a couple of issues noticed. For instance, in the most the headers in each file, there were spaces between words or even after words. Therefore, we removed all spaces found in each header, between and after character words.

Additionally, some datasets had combined data in one column that required separating that data and placing into new columns along with columns repeating redundant data. For instance, in the Austin Public Internet Locations dataset, the Location column had the Street Address, the Longitude, and the Latitude all together for each row. The Austin Public Health Locations dataset also had a similar column but instead under a street address column. Therefore, we created new columns such as Longitude and Latitude to separate the Location data from the Street Address data within those respective datasets. Additionally, the Austin High School Graduation Rate dataset also had no Street Address, Longitude, or Latitude columns, which had to be created (for future merging) by finding the address for each high school and then searching their longitude and latitude data by their newly listed respective address through an online site that converts addresses to GPS locations [6].

Furthermore, the Neighborhood Reporting Areas dataset had commas within its reported numbers under the column of Total Housing Units that were manually removed by each row for our coding to fully read, calculate, report the correct results within output files as the commas were acting as delimiters otherwise. Finally, this neighborhood dataset also had to be converted from a xlsx to a csv file. While some missing values were noticed directly from manually opening the datasets, our team decided to use Python coding to clean and fill those missing values instead of manually filling them.

Since other necessary columns for merging were missing in a few of the datasets, additional columns were created. For instance, a Zipcode column was added to the Neighborhood Reporting Areas and the Austin High School Graduation Rate datasets respectively to have make the Zipcode column a common element across all five datasets for future merging. The other three datasets already included a Zipcode column. Out of the all of the new columns created, the Zipcode column for the two mentioned datasets were the most time consuming to create as each neighborhood and school row had to be actively searched online to assign a zipcode for each of their respective rows. Our team used GoogleMaps to look up zipcodes for each high school and used the Travis County Zipmap online to look up zipcodes by neighborhoods [7].

**Phase II Data Cleaning: Python Cleaning**

Through Python coding, our team was able to import pandas, file access each dataset, read each dataset as a dataframe, and then check data organization to detect if any delimiters still present. No delimiters were found at this step. However, when checking each dataset for missing values or nulls, most of our datasets had nulls present for either missing values or descriptions (the exception being Neighborhood Reporting Areas dataset). For instance, for the Affordable Housing Directory dataset, three columns had nulls present, such as for PropertyManagementCompany that were filled with ‘Other’ to replace their null descriptor names, and for Phone and Website columns that were filled with zeros to replace their their null values. This technique was also applied to the other datasets missing descriptors characters or missing values as this was the most efficient way in removing nulls (inclusive of missing characters and values) while maintaining the integrity of the data without furthering unnecessary manipulations for our project. The Austin Public Internet Locations dataset, had six columns with nulls present in ConnectionStatus, ConnectionColor, LocationName, FileName, CouncilDistrict, and AppplicationDocument. The Austin Public Health Locations dataset had eleven columns with nulls present in Hours, Website, PhoneNumber, OtherPhone, BuildingID, OwnershipStatus, Owner, OccupyingDivision, OccupancyType, SqFt, and YearBuilt. The Austin High School Grad Rate dataset had thirteen columns with nulls present in District, 2015ClassSize, 2015Graduated, 2015ClassSize, 2014ClassSize, 2014Rate, 2013ClassSize, 2013Graduated, 2013Rate, 2012ClassSize, 2012Graduated, and 2012Rate.

Since our objective was to create two output files from selected columns for data processing and visualization, any columns across the datasets not needed for the goal in examining the relationship between demographics and affordable housing had to be removed. Therefore, any columns from the datasets not supporting that goal were dropped.

For the Affordable Housing Directory dataset, the columns dropped were ProjectName, Address, TotalUnits, HousingType, Status, Phone, Website, and Location. Although ProjectName, TotalUnits, and HousingUnits were dropped, they may have some value for future uses. Keeping PropertyManagementCompany rather than the ProjectName was more informative about what companies owned or managed housing units by aggregates in the Austin area, since our project was primarily focused on public resource allocations available in each zipcode zone rather than focued on individual housing units based by name or address. TotalUnits was not necessary to be kept as the Neighborhood Reporting Areas dataset was more comprehensive in for their reported total housing (in their TotalHousingUnits column) and a better choice for our team to calculate the affordable housing rate by. However, future researchers could use those values for measuring how much each housing company allocates affordable housing in percent by housing aggregates. UnitType was more descriptive on the structure of the housing units and complemented Household columns based on household unit/incomes than HousingType did. Furthermore, we wanted just one type from this dataset. Therefore, HousingType was dropped for those reasons. With that said, having HousingType present rather than UnitType could have been more informative in describing the community that inhabited those housing units.

For the Neighborhood Reporting Areas dataset, the columns dropped were Neighborhood, OccupiedHousingUnits, VacantHousingUnits, PercentOwnerOccupiedHousingUnits, PerPerAcre, DensityRanking, and Acres. Although Neighborhood, VacantHousingUnit, and DensityRanking were dropped, they may have some value for future uses such as viewing the overall data by neighborhood rather than zipcode zones, or measuring the rate of vacant housing within zipcode zones to monitor community migrations, gentifications, or comparing density changes by zipcode zones. Additionally, one could also compare this 2010 US Census dataset with the future 2020 US Census dataset based from these dropped variables.

For the Austin Public Internet Locations dataset, the columns dropped were Category, ConnectionStatus, ConnectionColor, LocationName, CouncilDistrict, LocationAddress, LocationCity, LocationState, LocationCountry, FileName, ApplicationDocument, and Location. Although Category, ConnectionStatus, and LocationAddress were dropped, they may have some value for future uses such as further detailing the site type of the building providing public internet service in each zipcode zone. And since more than half of these internet sites were indicated to still be awaiting to become active or updated by Google, one could track which sites are currently active or not by ConnectionStatus. The assumption we placed in considering these internet sites, active or not, was that these sites already had publicly available internet service at their public building site and were simply updated (or still awaiting to be updated) by Google. Additionally, using LocationAddress could provide more resources for the public to find these publicly available internet locations, such as directly from Austin’s local government resources online. One new column that was added to this dataset through Python coding was the InternetType column, which each row was filled with ‘PublicInternet’ as each row represented each site was an public building site for public internet access (based from the previously stated assumption). The main reason for creating this new InternetType column was to have a simplified version of the Category column, which had too much descriptors than was needed, especially for future merging.

For the Austin High School Graduation Rate, the columns dropped were District, StreetAddress, 2016Graduated, 2016ClassSize, 2016Rate, 2015ClassSize, 2015Graduated, 2015Rate, 2014ClassSize, 2014Rate, 2013ClassSize, 2013Rate, 2013Graduated, 2012ClassSize, 2012Rate, 2012Graduated, and Location. Although Category, most of the data was dropped involving the number of students that graduated and the rates of graduation per high school, this data may have some value for future uses, particularly for educational researchers or planners investigating or seeking more public resource allocations for their school districts in Austin.

For the Austin Public Health Locations, the columns dropped were StreetAddress, Hours, Website, PhoneNumber, OtherPhone, BuildingID, OwnershipStatus, Owner, OccupyingDivision, SqFt, YearBuilt, and Location. Although Hours, Owner, and OccupyingDivision were dropped, they may have some value for future uses such as using Hours to indicate operational hours per public health site visually displayed on an online public site map, or using the Owner column to indicate which public entity operates each public health site, or using OccupyingDivision to further indicate what each public health site specializes in services.

**Data Merging**

Although the most common element for all datasets used for data merging was the Zipcode column, some mergings involved not only renaming columns within datasets but also creating new data frames to populate them with the renamed columns to merge, such as for the file 1 output of Austin Public Sites or indexing calculated results by each Zipcode row from datasets, such as for the file 2 output of Austin Affordable Housing Demographics. Therefore, three main methods, dataframe manipulation, indexing, and different joining functions, were used to manipulate the data for merging to acquire our two output files. To prepare for file 1 outputs, PropertyManagementCompany column was renamed as SiteName and UnitType column was renamed to SiteType for the initial housing dataframe, OrganizationName column was renamed to SiteName and the newly created InternetType column was renamed to SiteType from the initial internet dataframe, School column was renamed to SiteName and SchoolType was renamed to SiteType, and FacilityName column was renamed to SiteName and OccupancyType was renamed to SiteType.

Following this, Zipcode, SiteName, SiteType, Longitude, and Latitude columns were only selected to be inputted into second data frame versions from each of their initial dataframes. Then these second dataframes versions (four total) were merged consecutively to each other using the outer join function to merge all data rows under those five common columns that these second data frame versions all shared. In this way, the data rows were, conceptually, stacked on top of each other to ensure all data was collected to their respective columns, which was a main reason for renaming columns to SiteName and SiteType. This approach helped solve a previous issue in merging the data where there were multiple columns unnecessarily present as those columns were redundant to keep. The benefit of having additional common elements created was that the data could be easily merged with the outer function than with other joining functions. The result was an output file that had more value combined with four datasets than separate within their own. The results from this file output already indicated how useful the data from public sites could be used, such as being displayed onto a GPS mapping website, as our team originally intended for, which Austin city government officials may be able to use and present to the public they serve.

In contrast, for file 2 output, data was calculated by averaging person household incomes and racial demographic percentages from values grouped by and sorted by zipcode. Total affordable units, total population, and total housing units were each respectively summed up, grouped by, and also sorted by zipcode. Additionally, a new column was created called the AffordableHousingRate that combined the summed data of TotalAffordableUnits to TotalHousingUnits for calculating an affordable housing rate per zipcode. Then each of these calculations were coded into proper calculated formats, such as for some in percentage formats, rounded, or made into integers for reporting until finally indexed to be merged. Rather than be merged on five columns and with an outer join, file 2 output was merged only on the Zipcode column and with an inner join since mainly two dataframes from Affordable Housing Directory and Neighborhood Reporting Areas datasets were used for creating the output file 2 of Austin Affordable Housing Demographics. The resulting output results also indicated these some interesting insights based on matching racial demographics with affordable housing info, household income, and total population per zipcode as Austin city government officials or city planners may to use this resulting dataset to track what affordable housing projects are most available to certain communities of income, racial demographics, or by total population.

Finally, all two of the finalized merged data frames were checked for duplicated data within the rows, which ten rows were found to be duplicated more than once and then deleted based from the same SiteName, Longitude, and Latitude for the output of Austin Public Site while no duplicates were found based the calculations by Zipcode for the output of Austin Affordable Housing Demographics.

**Analysis of Visualizations**

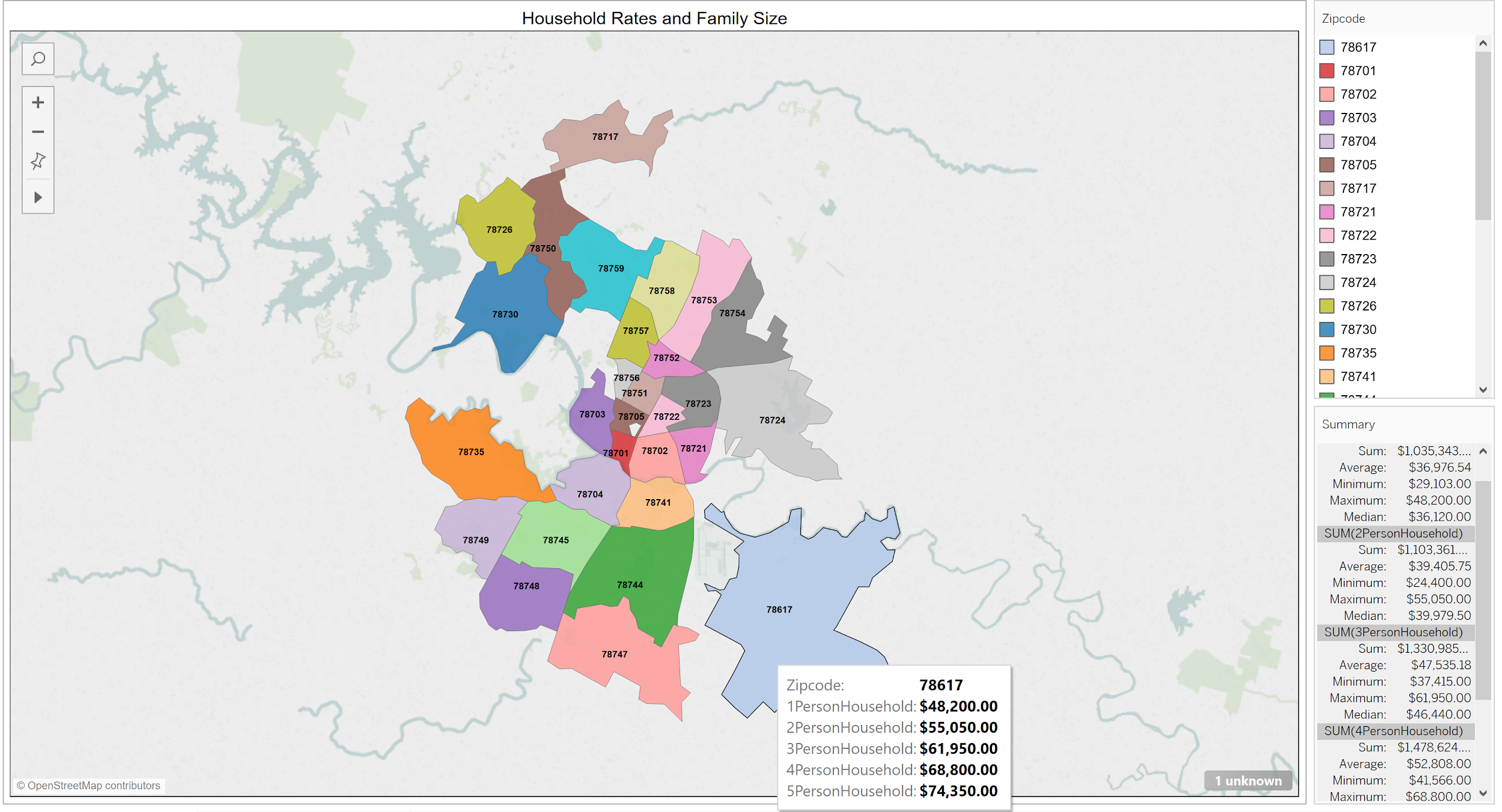
Visualization is one of the best methods of learning, in order to make a great visualization, one must add an appealing design to draw in attention to viewers. It should give concise information while being meaningful with concise yet noticeable design easy to understand when viewing. Visualization enables us to look at data different way and to view data concisely. In this way, we strived to present our data visualizations following that approach.

We focused on Affordable Housing Rates and Demographics and Family Size and income by zipcodes for visualization. By grouping the demographic population rates together as descriptive items for each zipcode we followed the visualization characteristic of association. We chose to color and separate each zipcode area a different color to help illustrate the differentiation of each area easily(Fig. 1 & 2). We show the housing rate and demographics for each zipcode area to show how the rates are correlated and can be compared visually, thus demonstrating quantitative perception. By following all these characteristics, we are able to effectively show our data in a comprehensive way.

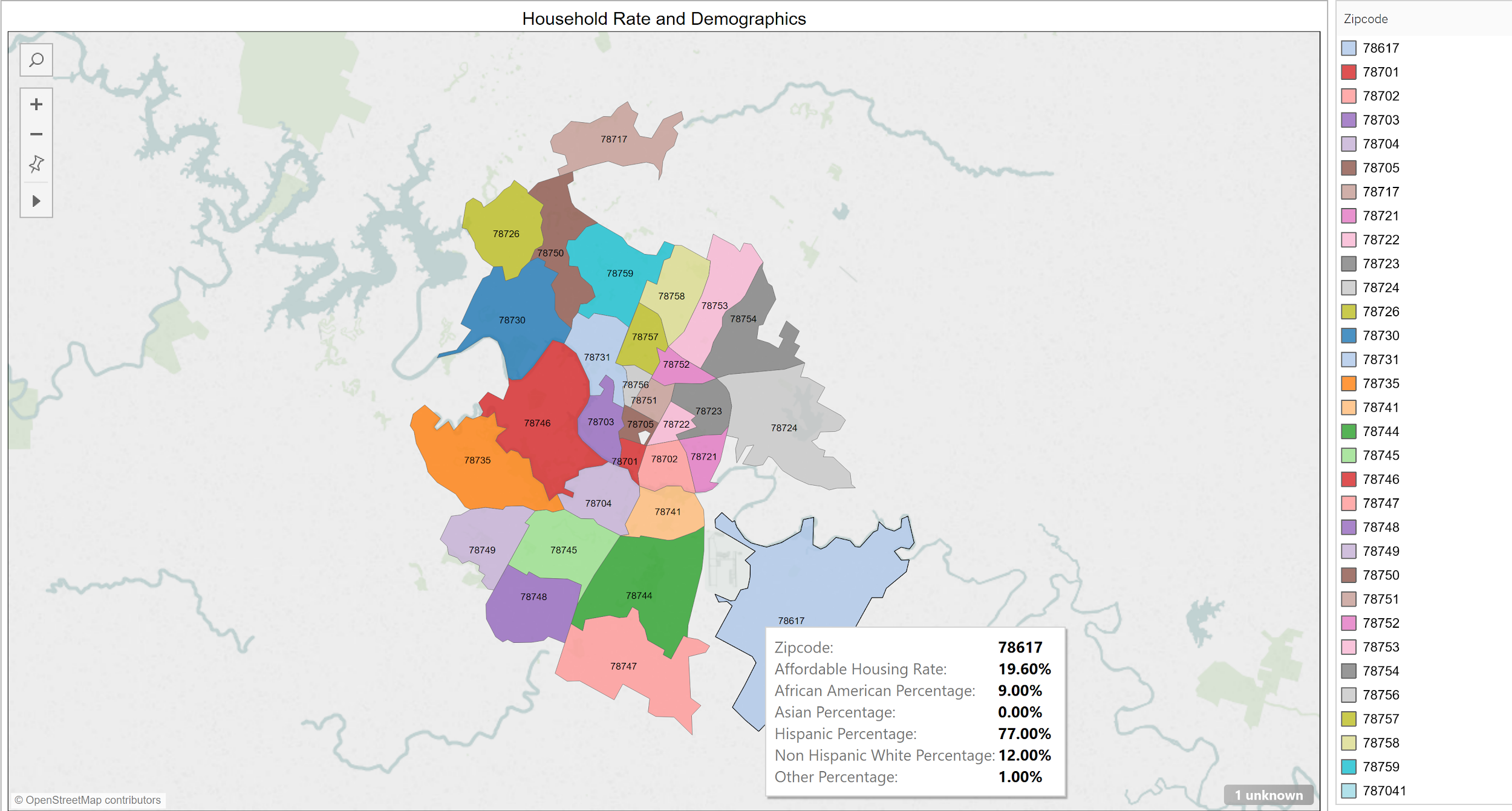
For example, in our building’s visualization(Fig.3), we were able to differentiate every zipcode by giving them various colors to make it appealing to eye and make it easier to differentiate. Adding the buildings in the form of pie charts, this shows all the government buildings within a zipcode, while hovering over the pie chart will show the different buildings. We also added the labels over the colored areas to show the buildings within those zipcodes. The only issue was that our visualization lacked preciseness because we based our analysis on zipcode; making it impossible to pinpoint where each building was located on our map, but grouping the different buildings by zipcode.

We kept our visualizations fairly simple for ease of processing for the audience. The information that we used is combined from several sources, but we focused our research on very specific areas. We narrowed our data to visualize very specific items for comparisons. This lead to very easy to understand concepts with the use of color and associated groups to describe our data. A user can hover over each zipcode to see the itemized data listed for each area. For example, in the Household Rates and Family Size (Fig.2), each area has the average income for different family sizes. This can easily be compared to other areas.

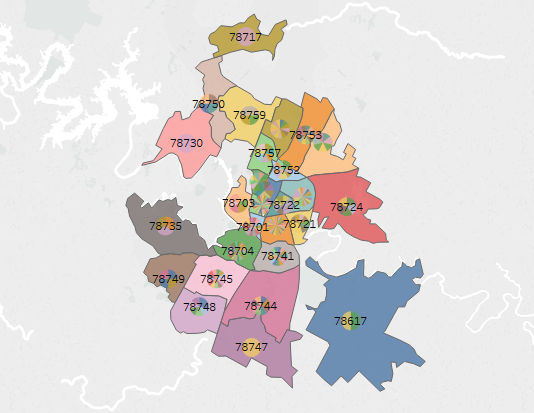
Finally, we also had a representation for the different types of buildings within a zipcode in the Austin area. The areas are delimited with different colors and labelled with the zipcode. There is a circle in the middle of each zipcode area, some have one shade, and others have multiple shades (Pie Chart)(Figure 3). Hovering over the pie charts will show different types of buildings such as libraries, schools, apartment complex owners and other Government buildings (Shelters, Community Care, …). This representation can help someone by providing information online to families about the different high schools within a zipcode to see if it is good enough for their liking. We made this visualization by linking two different maps within one by using the two different flies we created (Austin Affordable Housing Demographic & Austin Public sites).



**Figure 4-Household Rates and Family Size (of household incomes) of BI Visualization Final.twb file based from Austin Affordable Housing Demographics.csv file**



**Figure 5-Household Rates and Demographics of BI Visualization Final.twb file based from Austin Affordable Housing Demographics.csv file**



**Figure 6-Public Building Sites by Zipcodes of Buildings in Zipcode.twb file based from Austin Public Sites.csv file**

**Instructions for Code**

–Steps For Python Coding: Following these steps in order was adopted for all file outputs:

1. Open Python programming & import pandas
2. Define your file variables to be read as dataframes: file\_1 to file\_4, assigned to files, Affordable Housing Directory, Neighborhood Zipcodes, Austin Google Fiber GB Speed Internet Locations, Austin Public Health Locations, and Austin HS Grad Rates
3. Use pandas to read data from csv.s
4. Check how the data is organized by viewing the first 10 rows of data
5. Check for delimiters and null values by viewing the count of rows for each column
6. Remove columns that are not needed for analyzation
7. Check remaining columns are displayed as intended
8. Rename columns to remove spaces and format for consistency
9. Create a new data frame to merge the files (dfhousing2, dfinternet2, dfschools2)
10. Create new calculated fields (average rates by zipcode) for each necessary column
11. Create index to assist in merging new calculated fields and their tables.(G1b, G2b)
12. Merge new tables with adjusted calculated fields(Gmerge1 to Gmerge3 and then gmergedemo and gmergebldgs)
13. Double check for duplicates and remove any found
14. Create a new csv output with cleaned data (Austin\_Public\_Sites and Austin\_Affordable\_Housing\_Demographics)

–Steps For Tableau Visualization: Following these steps in order was adopted for all visualizations, but we are using Buildings\_in\_Zipcode as an example below:

1. Download every excel dataset within folder.
2. Download any Tableau files (Buildings\_in\_Zipcode).
3. Drag Austin Affordable Housing Demographic & Austin Public sites from the files list in Tableau and drop it into the space on top left.
4. Join both files by an inner join.
5. The data will show in Tableau, check between the Tableau data and Excel files just to make sure the data matches.
6. There are two different types of list in the column right (Measures and Dimensions).
7. To create our visualization, we start by dragging the zipcode in Affordable Housing Demographic and dropping into the color button in the mark column. This process will give our zipcode different colors and make them unique.
8. It will create a latitude in the row area and longitude in column area.
9. To label the colored area, drag the zipcode from dimension and drop it into the label button in mark.
10. Left click and hold the latitude in the row and drag it right next to latitude; it will create a duplicate map with the same zip code on top of each with a different tab under our first tab where we dragged our colors before.
11. Take the Site Type and drag into colors.
12. There is a drop down menu over the button selection of our data (Color, size, label, …). Choose pie chart from the menu.
13. Pie chart helps us show every public structure within a zipcode.
14. Go back to rows on top and right click the second latitude and select in the menu dual axis, this will merge both files.
15. The end product will result as a representation just as Figure 3.

# **References**

[1] City of Austin. "Affordable Housing Directory." *data.austintexas.gov*, 29 Mar. 2019, <https://catalog.data.gov/dataset/affordable-housing-listing>

[2] City of Austin. "Neighborhood Reporting Areas Table I." *austintexas.gov*, 29 Mar. 2019, <http://austintexas.gov/page/demographic-data>

[3] Texas Education Agency. "Austin High School Graduation Rates." *data.austintexas.gov*, 29 Mar. 2019, <https://catalog.data.gov/dataset/austin-high-school-graduation-rates>

[4] Austin Public Health. “Austin Public Health Locations.” *data.austintexas.gov*, 29 Mar. 2019, <https://catalog.data.gov/dataset/health-and-human-services-locations>

[5] City of Austin-Office of Telecom & Regulatory Affairs (TARA) and Google Fiber. “Community Connections Program: 100 Public Facilities Signed up for Google Fiber Gigabit Speed Internet” *data.austintexas.gov*, 29 Mar. 2019, <https://catalog.data.gov/dataset/community-connections-program-100-public-facilities-signed-up-for-google-fiber-gigabit-spe>

[6] GPS Coordinates. “Latitude, longitude and address of any GPS location on Google Maps.” *gps-coordinates.net*, 29 Mar. 2019, <https://www.gps-coordinates.net/>

[7] Travis County Zipmap. “Find Zipcode.” *zipmap.net*, 29 Mar. 2019, <https://www.zipmap.net/Texas/Travis_County/Austin.htm>