Assignment 1 Report Robert Hall

Dataset 1: Housing Data for Australian Select Market

Description

The original data-set is from Kaggle and consists of a time-series for the housing market in Australia. The data does not fully encompass the Australian housing market, rather it is a subset for specific area codes (as defined in the data). The first and foremost piece is the time-series which starts as early as February 2nd 2007 and goes all the way to July 26th 2019. The data can be accessed via Kaggle from the website:

https://www.kaggle.com/htagholdings/property-sales

Dataset Format and Type

The data from Kaggle is a flat table that consists of the 13 years of housing market data throughout certain districts in Australia which includes the date, area code, price of the home, type of property purchased, and the number of bedrooms the property has. Each record in the data-set is a single instance of a property purchase which includes the respective information about the home. The “Price” column is a numerical number for the purchase price, the “propertyType” column is a binary string of either ‘house’ or ‘unit’ , the “bedrooms” column is a numerical number for the number of bedrooms, and the “datasold” column is the Time-Series instance of when it was sold. The “postcode” column is a numerical number for the area code. We are not including the postcode in this analysis. Further analysis using map visualization will use this postcode.

Attribute Types and Semantics

The most important pieces from the data that I decided to visualize is of course the time series data but also the overall trend of the housing market. The semantics of the “Price” column is as it implies, the price of the home purchased. There is also the “propertyType” column which has to do with whether the purchased property was either a house or a unit (another word for apartment). There are only two values for this column and they are : House , and Unit. The last column used in the analysis and visualization is the “Bedrooms” column. The bedroom column is a numerical number

representing each data-piece as the number of bedrooms the property has. The values range from 1-bedroom all the way up to 5-bedrooms. The postcode is the area code for which that property was purchased, but since we are not visualizing the area we are not including this in the analysis. From viewing the data it does appear that each of these data-points are single instances and further analysis should use each data-point together to paint the whole picture of the data.

Preprocessing

Since the data consists of multiple values (“Bedrooms” column can have up to 5 , “propertyType” column can have either ‘unit’ or ‘house’) it was decided to split the data up into their respective groups to make the visualization more viewable. The meaning behind this is that the data was sectioned off based on the number of bedrooms and the housing type. I.e Time-Series analysis for 1-Bedroom Houses , or Time-Series analysis for 2-Bedroom Units . It is important to note the “postcode” column in the data-set was not touched due to us not using it for this analysis. The data was processed using pandas to subset the data into their own respective data-frames for analysis. Using pandas subsetting technique like shown below:

```
one_bedroom_data = df.loc[(df["bedrooms"] == 1) &
(df["propertyType"] == "house")]
```

Subsets the data where we can change the bedroom number and the property type to slice the data how we need it. This was done in order to have an easier data-set to work with and cut down on the processing time due to not looping over the data each time we want to subset and visualize.

Visualization

Figure 1. One-Bedroom House Time Series

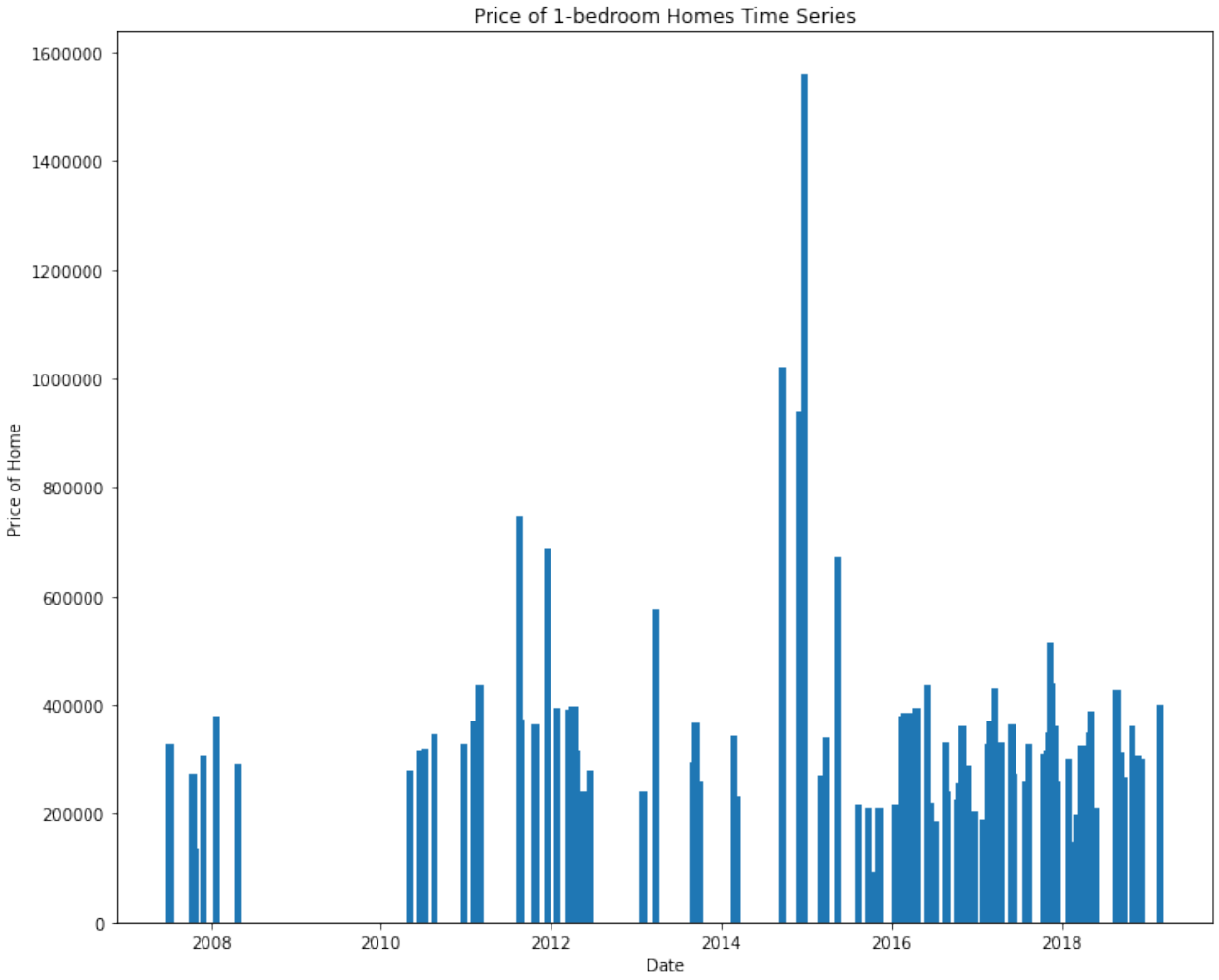


Figure 2. Two-Bedroom House Time Series

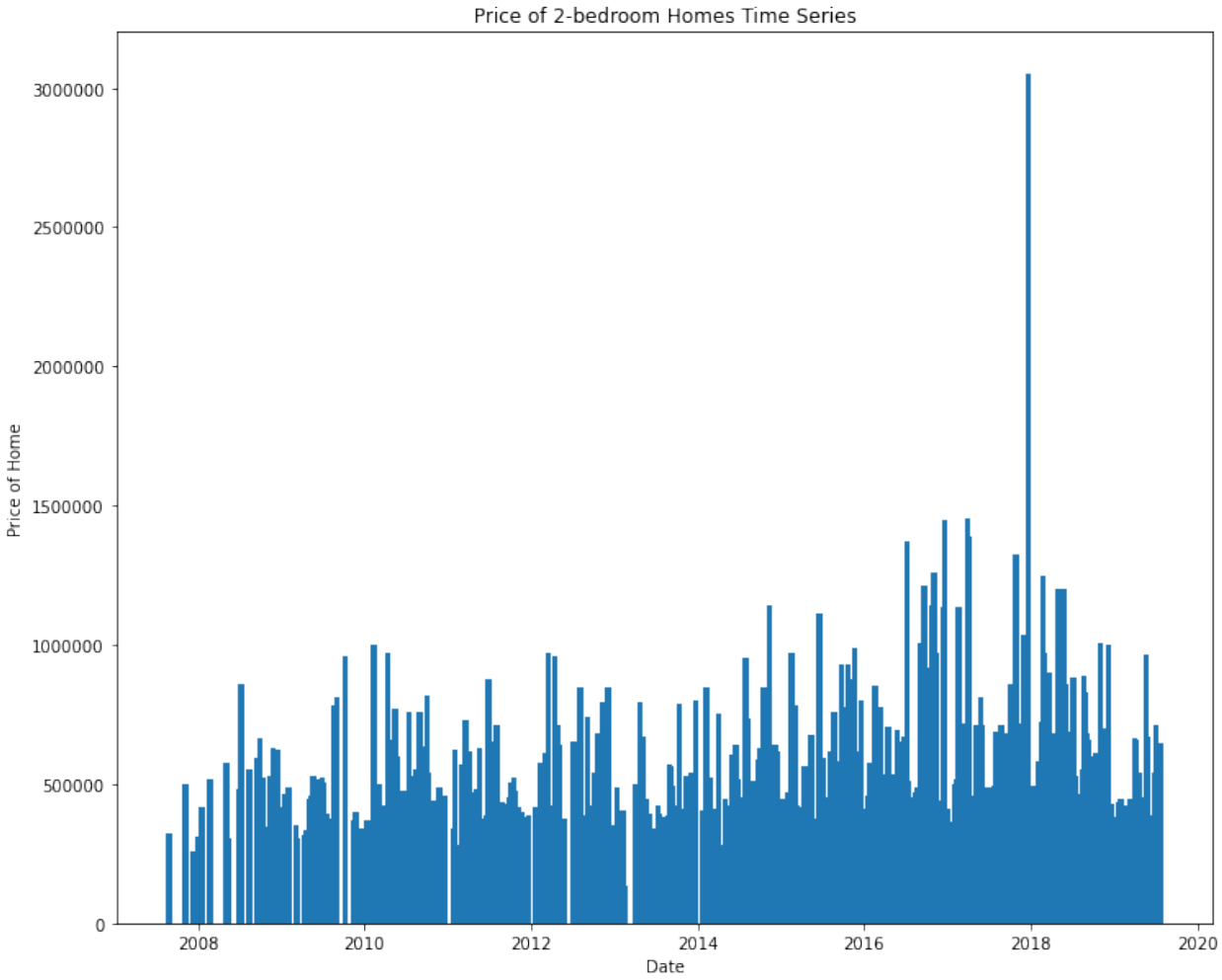


Figure 3. Three-Bedroom Home Time Series

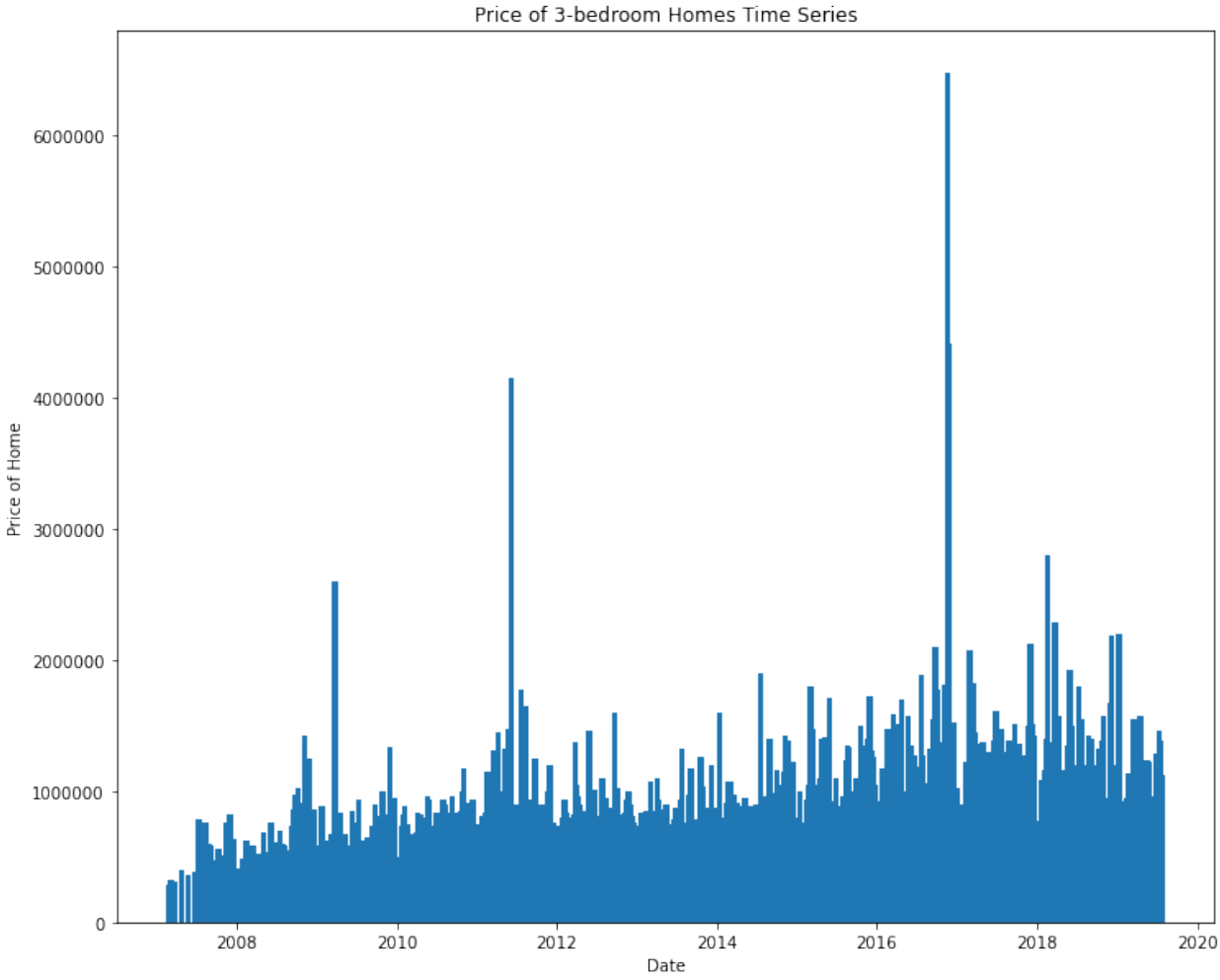


Figure 4. Four-Bedroom House Time Series

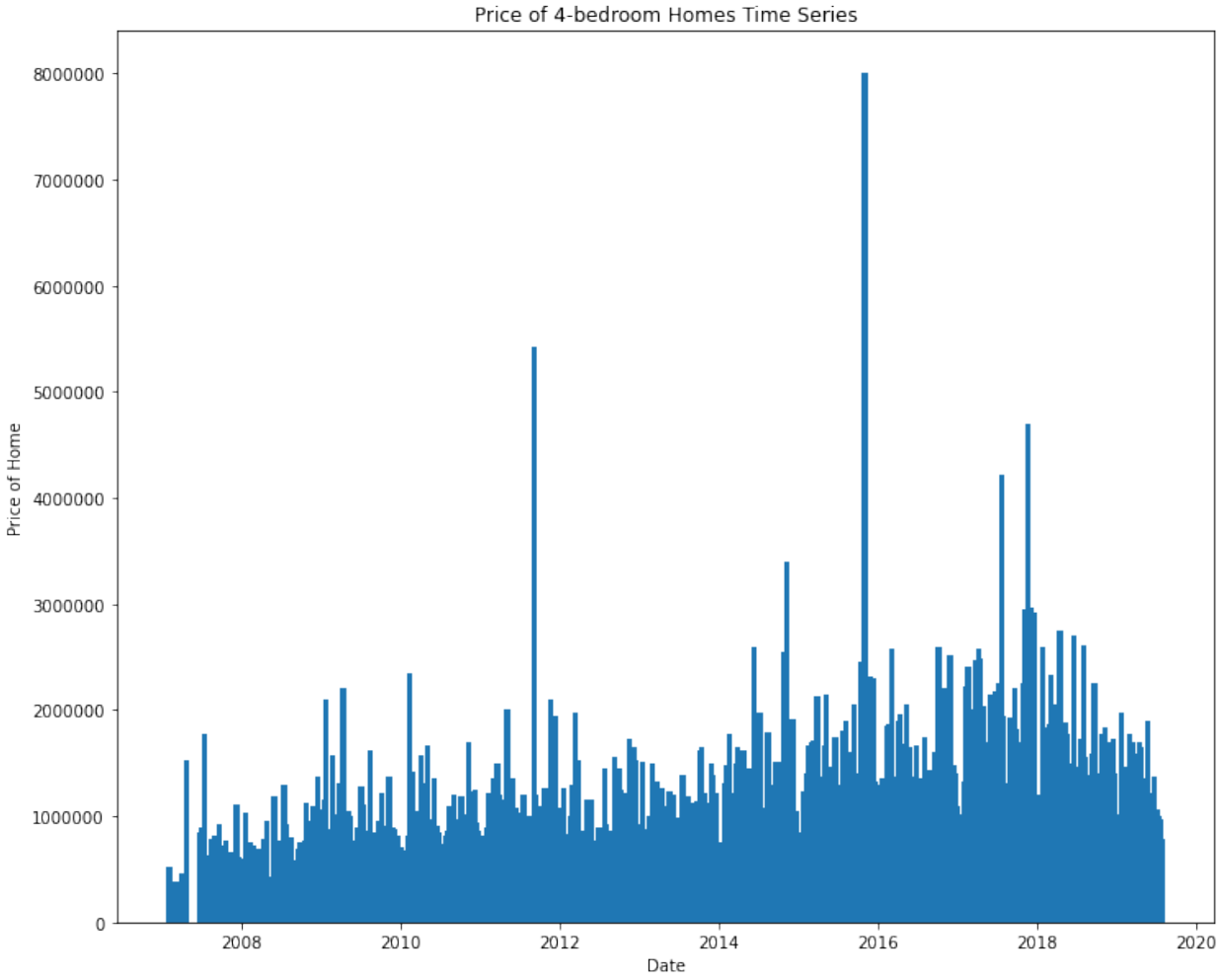


Figure 5. Five-bedroom Home Time Series

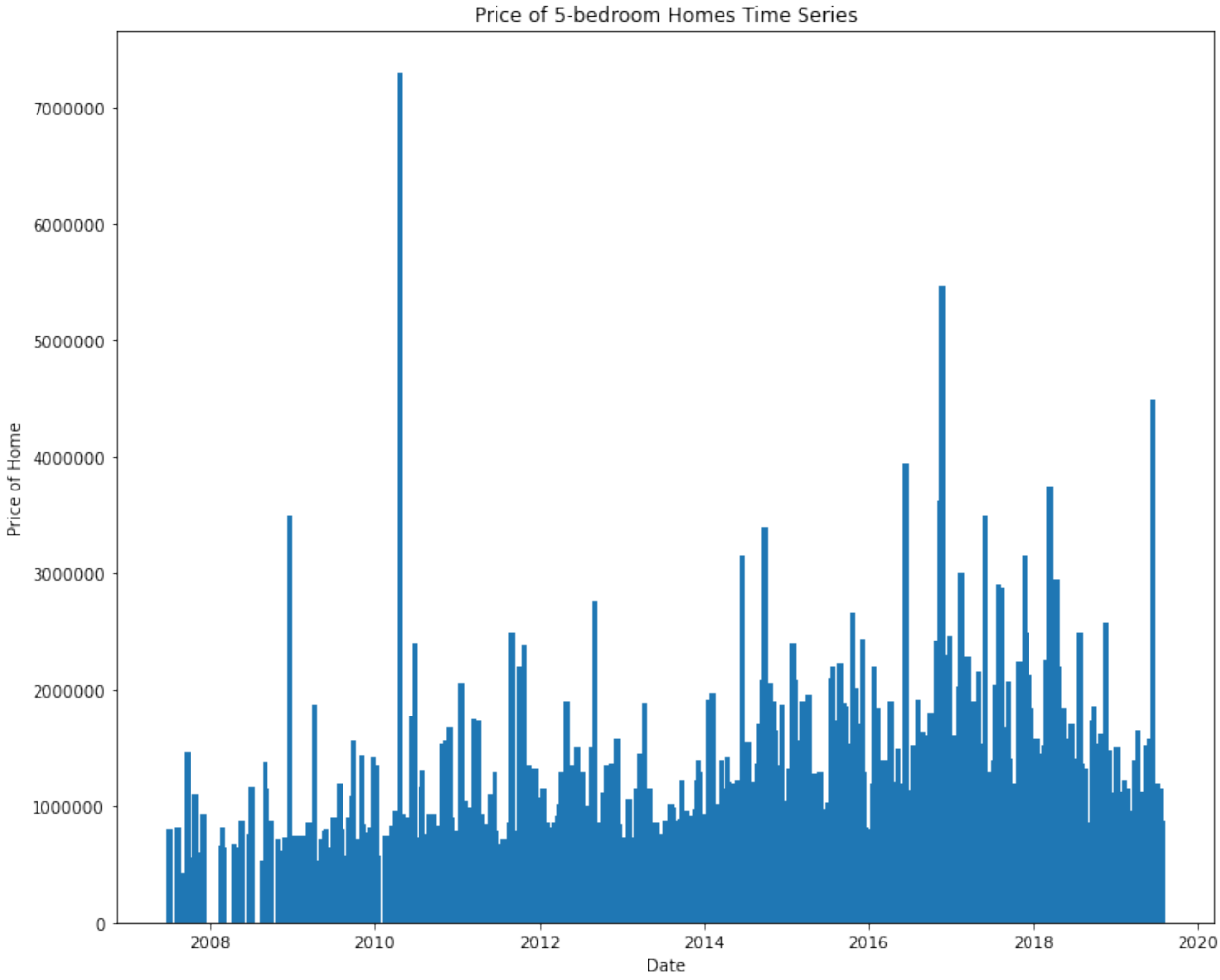


Figure 6. One-Bedroom Unit Time Series

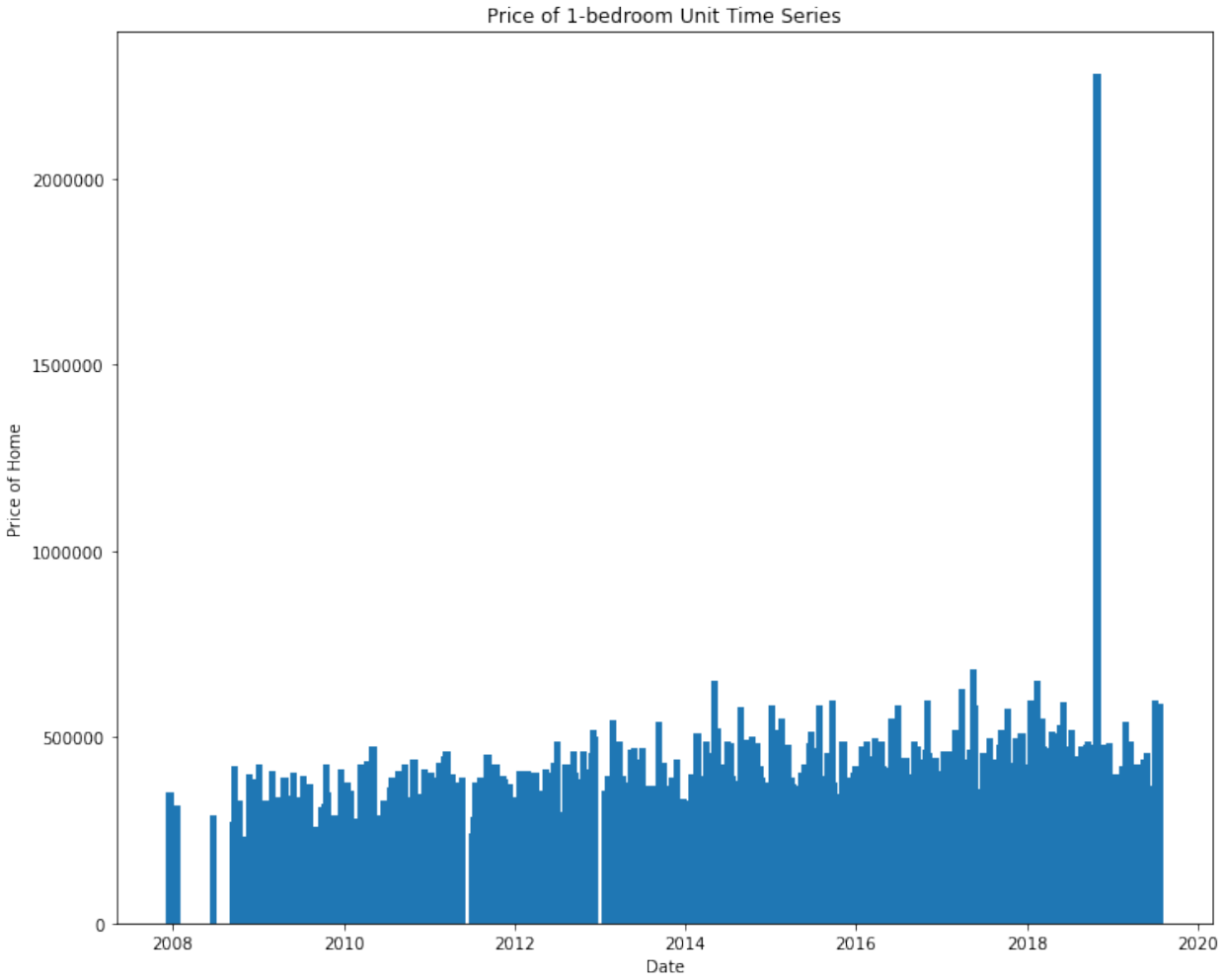


Figure 7. Two Bedroom Unit Time Series

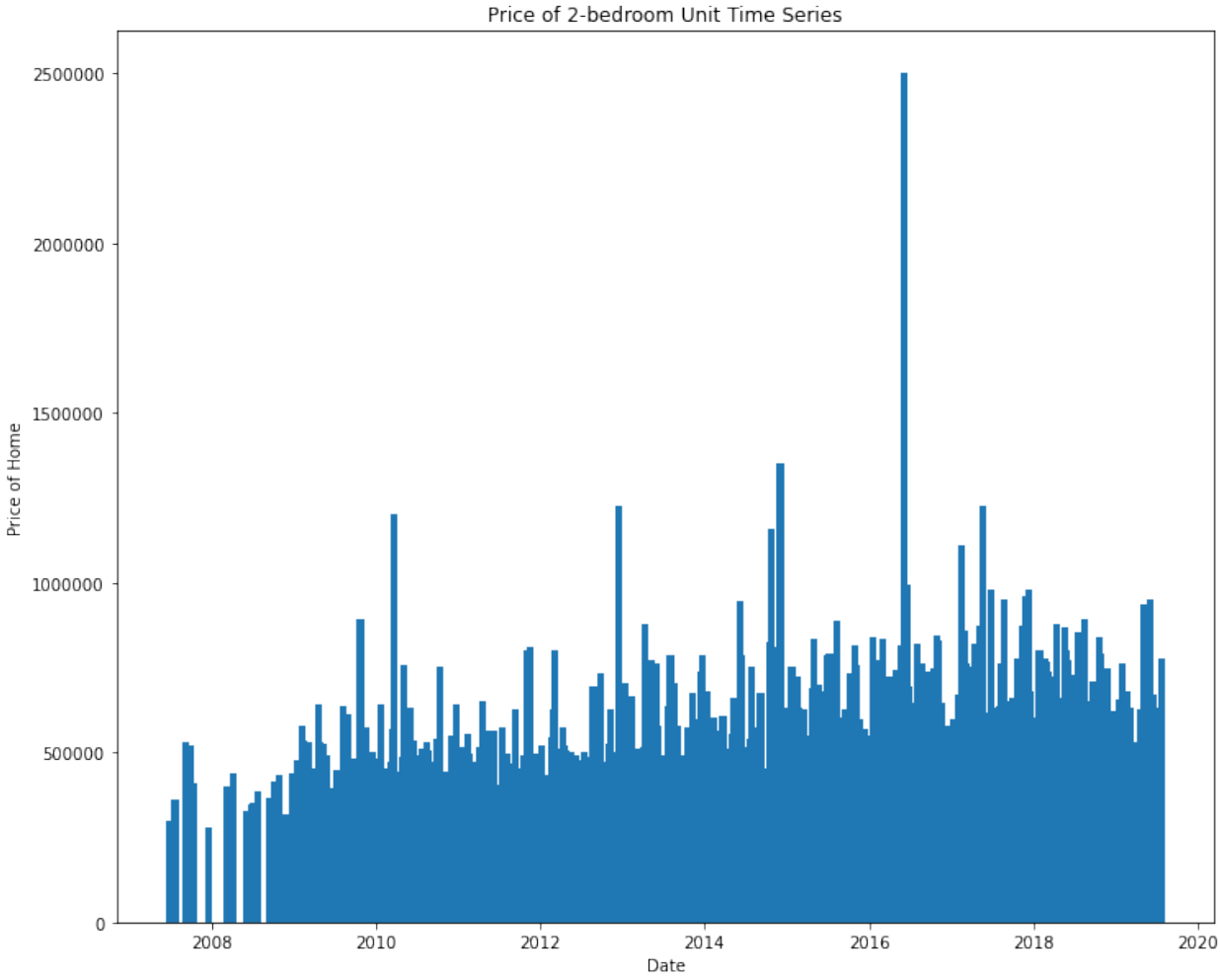


Figure 8. Three-Bedroom Unit Time Series

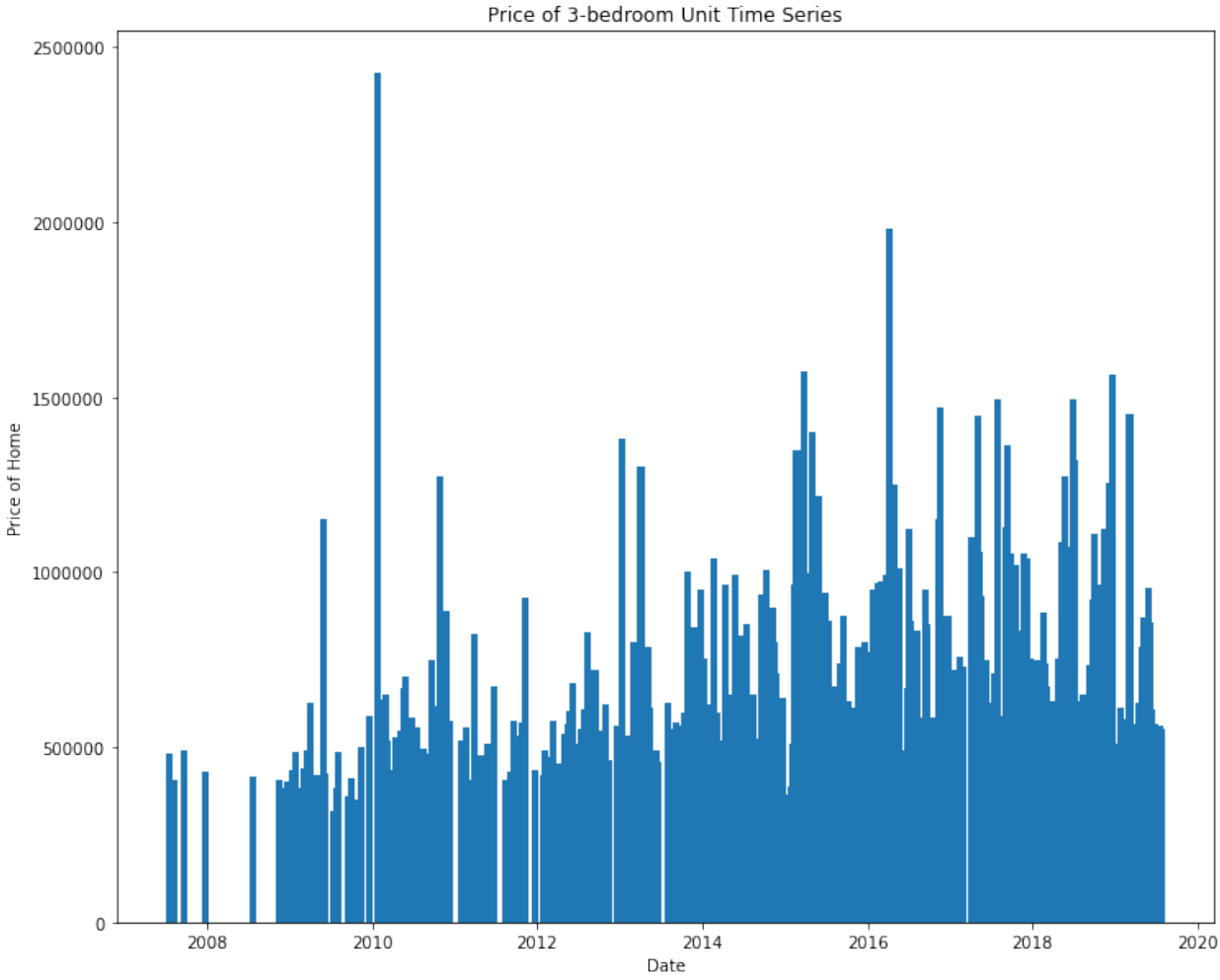


Figure 9. Four-Bedroom Unit Time Series

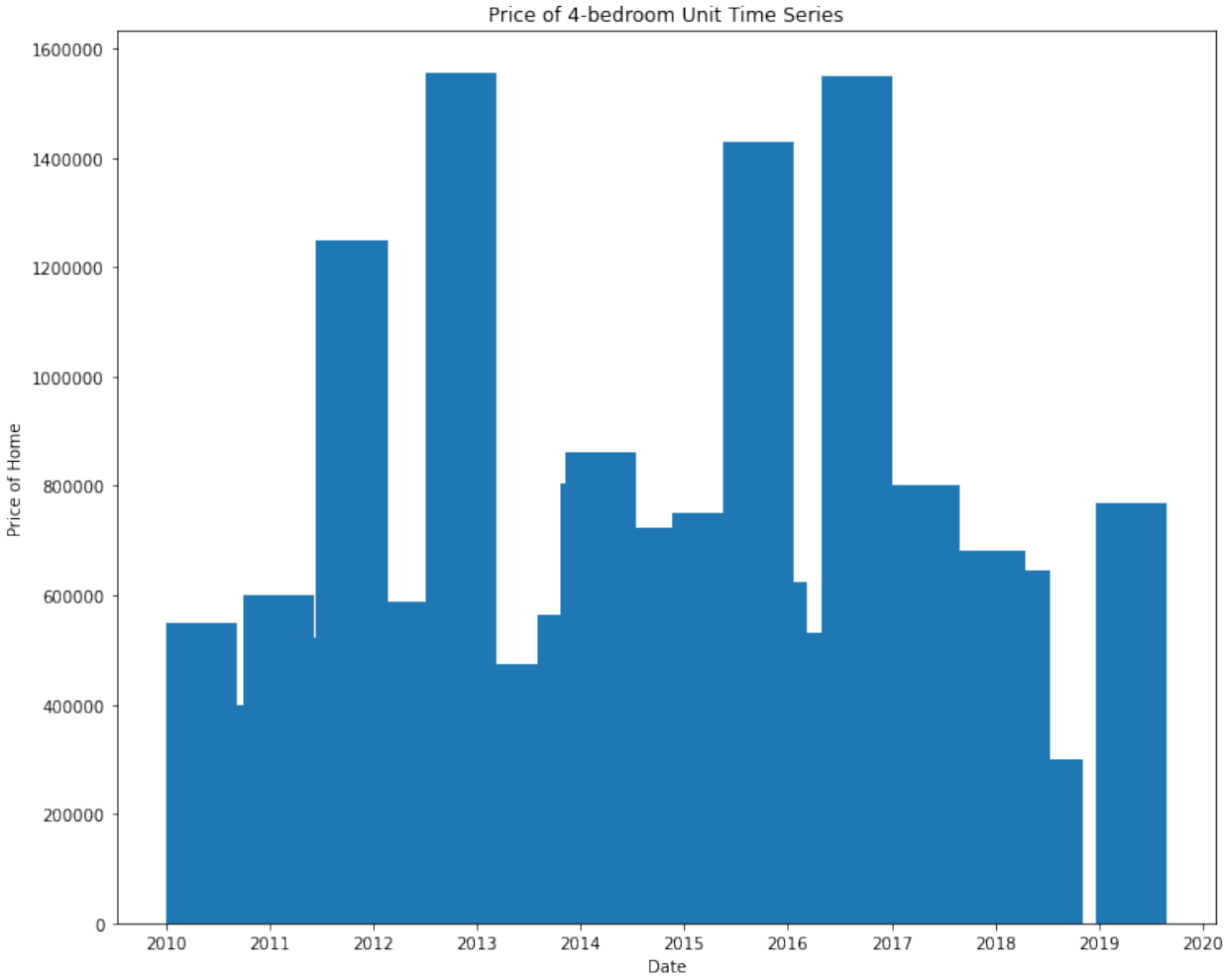
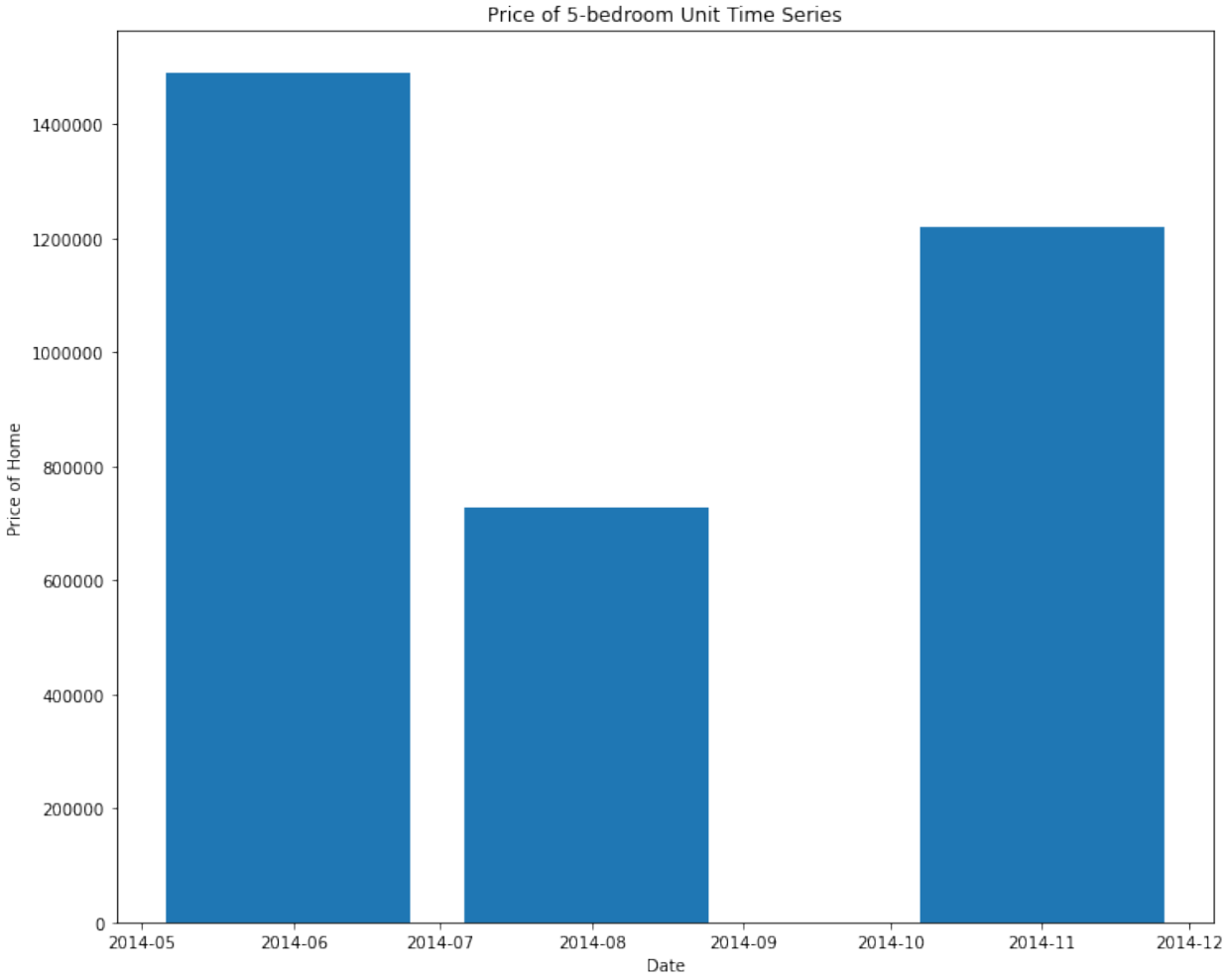


Figure 10. Five-Bedroom Unit Time Series



Analysis

From viewing the data on the graphs, one can tell that there are plenty of outliers, specifically the property price when it jumps anywhere above the $4,000,000 range (It is also important to note the numbers representing the price of the properties is in Australian currency rather than the United States currency we are use to, where $1 USD = $1.40 AUD) . The graphs we need to address are the 1-Bedroom Houses, 4-Bedroom Units, 5-Bedroom Units. From seeing these graphs there is very little data-points to consistently point towards an overall trend. From what the data reports one can see that there are very few instances of four and five bedroom units so the visualizations reflect that as well. The Table below outlines these low-reported data points in more detail for reference.

Type	Number of Instances
1 - Bedroom House	95
4 - Bedroom Unit	38
5 - Bedroom Unit	3

Due to the low-number of instances it cannot be reported with certainty that the prices of these property types are increasing or decreasing with the data we are given. However, with the other visualizations that have a larger number of instances we can see some more information. More specifically, each visualization that does not have a low number of instances can be seen as increasing over the time-series. What is meant by this is that the data shows that over time the price of each type of property (Both houses and units) are increasing at a steady rate. Even if we bar the outliers from the analysis the trend does appear to be going upwards for the Australian housing market. It is important to note that this is just a simple visualization and further analysis of the overall Australian Housing market will be needing more specific data and more information rather than just the raw numbers. In conclusion, in each visualization over time (excluding the three property types with low number of instances) does show an increasing cost in regards to the price of the home.

Dataset 2: COVID-19 PCR Testing Diagnostic Laboratory Testing Results for the US.

Description

The data used in this analysis is a sample of Covid-19 PCR Testing data from over 1,000 U.S laboratories and testing locations. The data presented here is a representation of diagnostic specimens being tested rather than individual people. From the data-source it mentions that the data may not represent the current count due to the time it takes to report testing information. Data can be accessed from the website via:
https://catalog.data.gov/dataset/covid-19-diagnostic-laboratory-testing-pcr-testing-time-series-80872

Dataset Format and Type

The data-set used in this analysis is a table consisting of multiple columns of Covid-19 PCR testing data which includes the date of test, the state in which the test was run, the overall outcome of the test, the new results reported, and the total results reported over that time period. Each row in this table corresponds to that region's reported Covid data. What is meant by this is that each row is the total count for that specific day for that specific region in the state. This analysis only looks at the overall trend rather than each state.

Attribute Types and Semantics

The semantics of each of the data-points corresponds as follows: The “overall_outcome” column relates to whether the Covid-19 test was positive, negative or inconclusive. Positive meaning that the test indicated Covid-19 was present, Negative meaning no Covid-19 was present, and Inconclusive meaning that the test did not correspond to a positive or negative reaction, possibly from lack of data in the test itself or other outside factors not included in this analysis due to no data present. The “date” column relates to the date in which the data was submitted where it begins at 2020/03/01 and results are current to 2022/02/02. The “state_name” and “state” columns correspond to which state these tests occurred in.

Preprocessing

The data presented here is similar to the previous analysis in this report on

how it is broken down into different categories (I.e the “overall_outcome” column is broken down into the strings : Negative, Positive, or Inconclusive). For this reason the data was split into multiple groups specifically for the different categories for the outcome variable so negative instances are in their own group and so on and so forth. There are other columns in the dataset that were not used for example such as the “fema_region” , “geocoded_state” , and the states themselves since this analysis is looking at the overall trends compared to a state-by-state basis. The data was processed using the same tool as the previous analysis: pandas in order to subset the data into their respective groups. From this subset, the visualizations created used the subsetted data for analysis for:

- Negative New Results Time-Series
- Positive New Results Time-Series
- Inconclusive New Results Time-Series.

Visualization

Figure 1. Overall Total Reported No Separation of Outcome

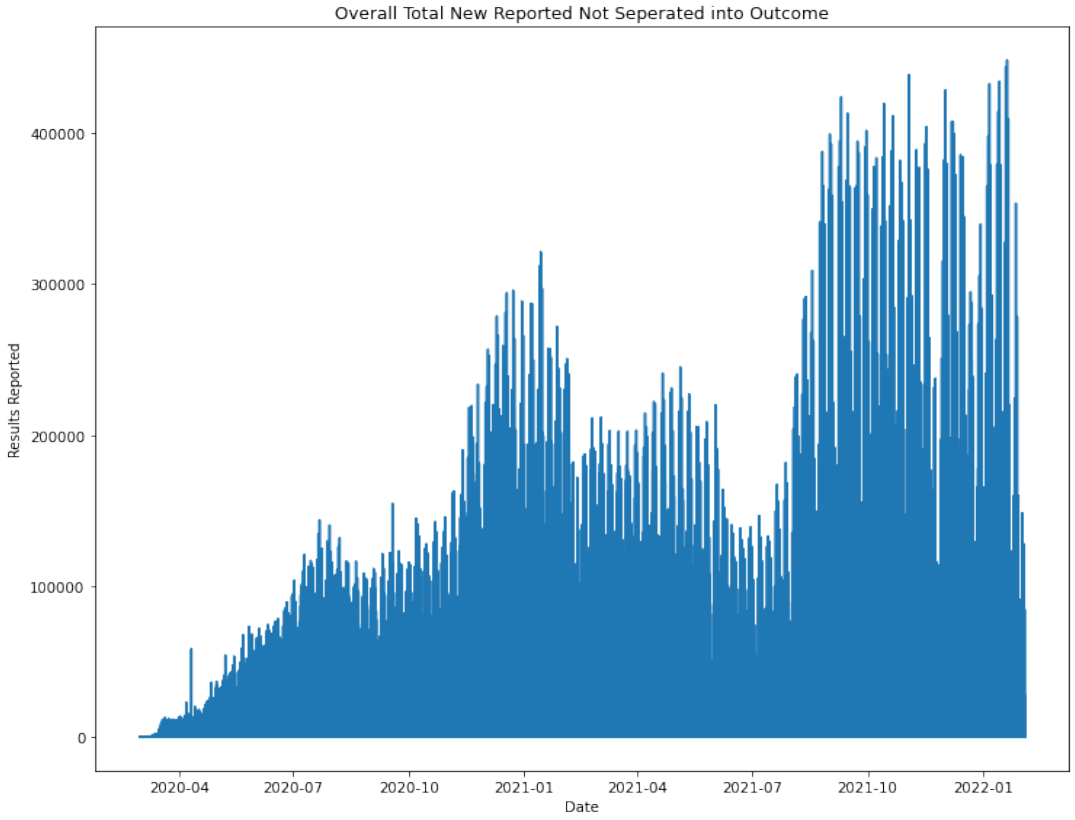


Figure 2.

Overall Total Reported Separated by Negative Outcome

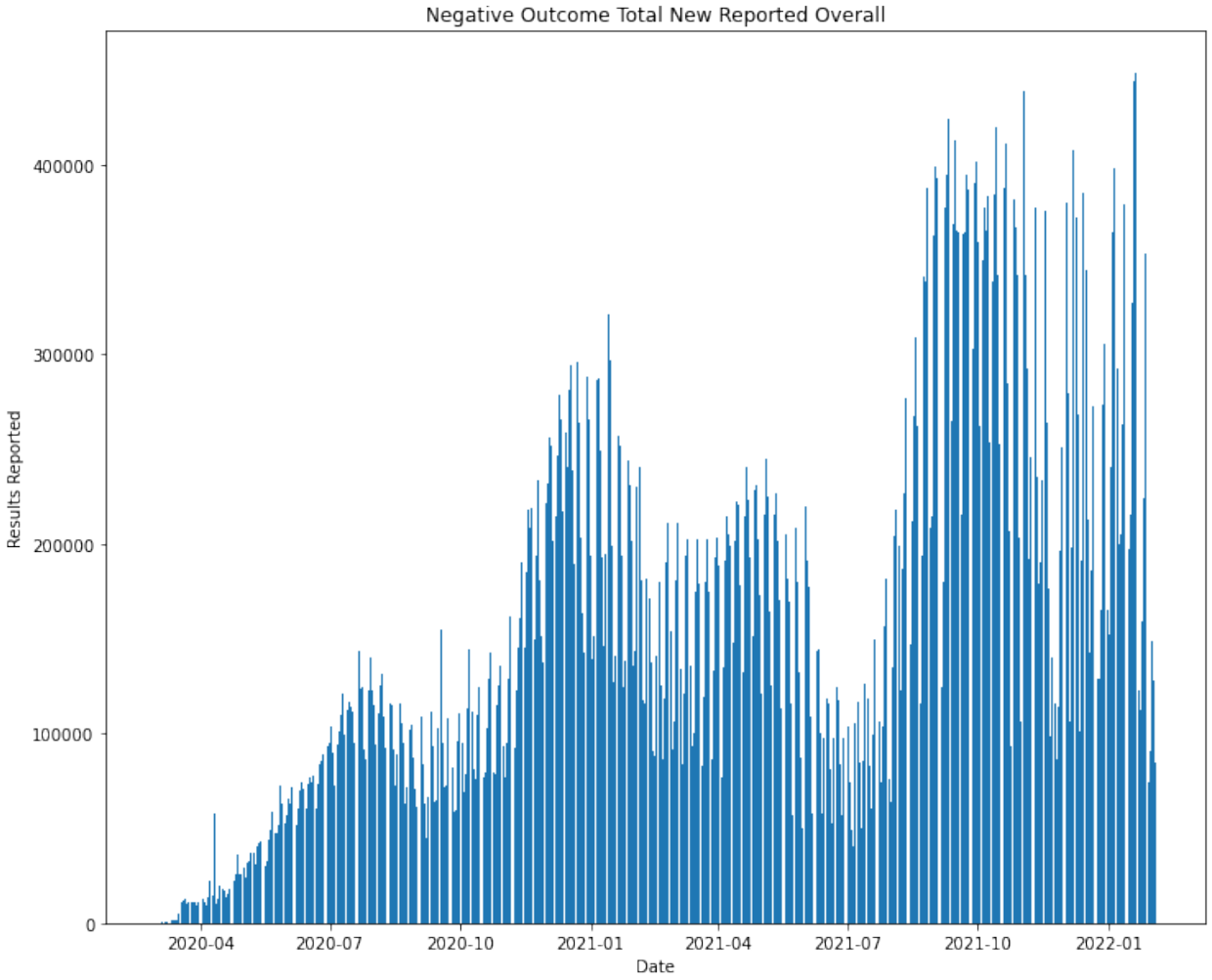


Figure 3.

Overall Total Reported Separated by Positive Outcome

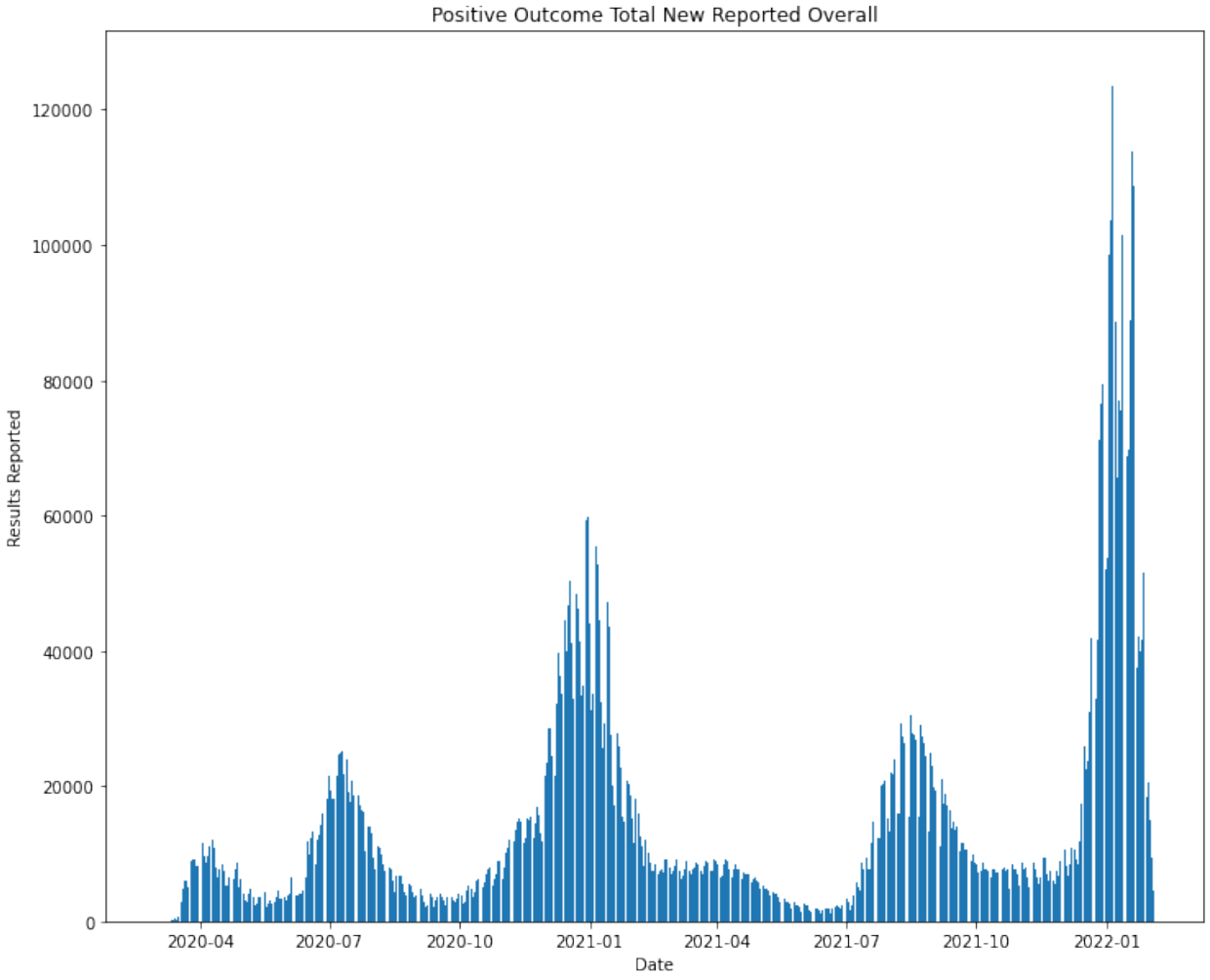
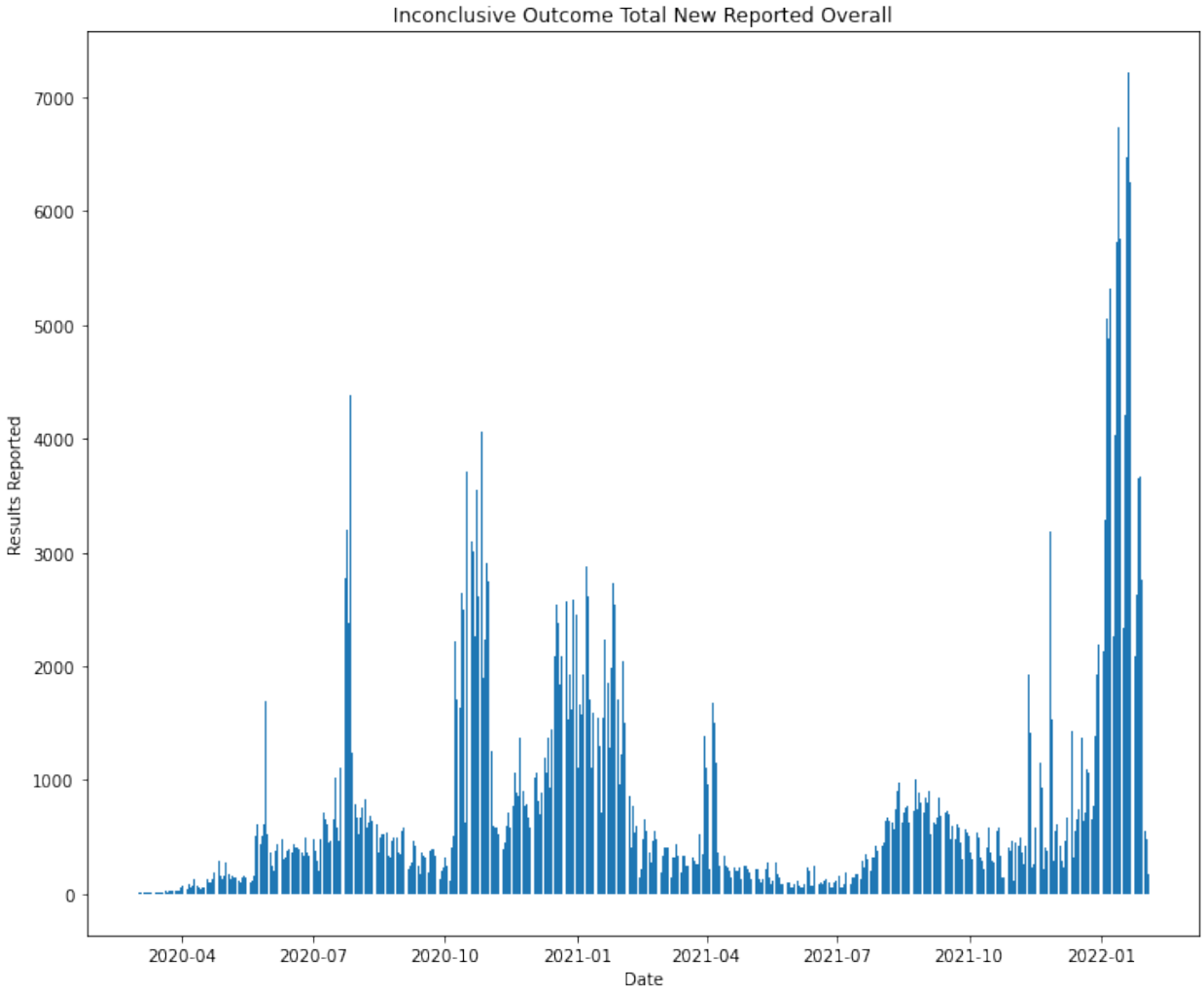


Figure 4.

Overall Total Reported Separated by Inconclusive Outcome



Analysis

From viewing the visualizations one can see that there are many high - frequency periods of testing and positive results and low-frequency periods as well. From just viewing this and relating back to when certain variants of the Covid-19 virus appeared, it mostly aligns with both the appearance of the new variants but also around winter-time seems as if testing is more frequent. One can see that around the winter holiday periods the total number of tests starts to reach 200,000 and above before the warmer months appear and it settles back below that. We can see this from both instances of around the new year once you look at the dates 2021-01 and 2021-10 the months leading into winter and leaving has the greatest amount of testing done during that period. It is also important to note that overall

testing seems to be increasing as time goes on. From looking at Figure 1. Starting the months between July 2021 and October 2021 the total number of tests appears to be going up, whether it be from people testing to visit family for holidays or more people becoming sick with symptoms of the Omicron variant or more seasonal sickness such as the Flu and the common cold. This would make sense since Covid-19 shares some similarities between the flu and the cold. The more concerning thing from this analysis is from Figure 3 when we start to each into 2022. The overall new positive outcomes appears to be on a rise much greater then we have seen in previous periods or high-frequency such as around 2021-01, 2020-07, and between July 2021 and October 2021. The negative reports around the same periods also have a higher frequency than before but it seems as if the higher-peaks and reported cases are due to more rapid testing measures.

Dataset 3 : Chattanooga City Biking Data

Description

The dataset used for this analysis comes from the city of Chattanooga TN biking data for the bike stations placed along the city. Each time someone rents a bike it takes note of the time they started, the time they ended, where in relation they are, and the actual time they were using the bicycle. The dataset consists of over 500,000 instances of bike rides. This analysis looks at the trips taken over the period of time to see if there are any places of highs and lows. It is estimated that Covid-19 will play a part in some of the time-series due to lockdowns and restrictions. The last time the dataset was updated was from July 2nd, 2021 and is updated every 6 months. The data-set can be accessed from:

https://www.chattadata.org/Recreation/Bike-Chattanooga-Trip-Data/tdrg-39c4

Dataset Format and Type

The format the data is presented to us is in a tabular format more specifically in a .csv file. Each of the rows in the table corresponds to a single instance of a bike ride, this meaning that each row has a start time for the bike ride and an end-time for the bike ride. The data also gives the information about which station it started at, which one it ended and also the start and end locations encoded in longitude and latitude variables. However since we are plotting time-series we are not using the start and end locations (Both addresses and longitude and latitude). This statement also applies to the ‘Start Station ID’ and ‘End Station ID’ since they are numerical representations of the stations and would not benefit from the visualization. The data used in visualization follows the format:

- Member Type : String (Customer or Subscriber)
- Start Time: Datetime Variable
- TripDurationMin: Float variable that can take on numerical values.

Attribute Types and Semantics

Since we are using the data-set there are a few columns we are going to use for the analysis, the first one being the “Start Date” column. There is also the “End Date” column but was not used in processing and visualization for

reasons that will be built upon in the Preprocessing section. The “Start Date” column is as it implies the start time the user rented the bike whereas the “End Date” column is the time the user stopped renting the bicycle. We are also using the “Member Type” column which consists of values of either : ‘Customer’ or ‘Member’. Customers are people who do not have the plan and pay as they go, and members are individuals who have the biking plan. It is also important to note that used in the visualization is the “TripDurationMin” column which is the time in minutes that the trip took place. There is care taken with this column that is mentioned in the Preprocessing section. The overall semantics of the dataset is the total biking data recorded for the city of Chattanooga using the station machines.

Preprocessing

Whenever the initial data was plotted from the time-series it became apparent that there are many erroneous values in the “TripDurationMin” column that skewed the visualizations quite a bit to where any valuable information is lost due to the differences in time. For this reason the data was spliced so that biking instances that were below 2000 minutes were included in the analysis. With this split of the data based on the under 2K minute threshold it was also subsetted into the respective groups of either ‘Customer’ or ‘Subscriber’ and the visualizations included below reflect these groups. Originally whenever the initial data was being plot, the ‘End Time’ column variables where used as the X-axis for the visualizations but once it shown that there is erroneous sections (I.e there are multiple instances in the ‘End Time’ column that have dates going back to 1970 - possible overflow bug or problems with data recording is to be suspected), it was decided to switch over to the ‘Start Time’ column which does not include any inappropriate values such as going back in time. It is also important to note that since we are using barcharts for this analysis the computation of the 500,000 + datapieces took approximately 10 minutes to compute and create graphs for. The data was split using pandas dataframe utility and used the plotting package from MatplotLib for the visualizations.

Visualization

Figure 1. Initial Visualization Showing Erroneous Data

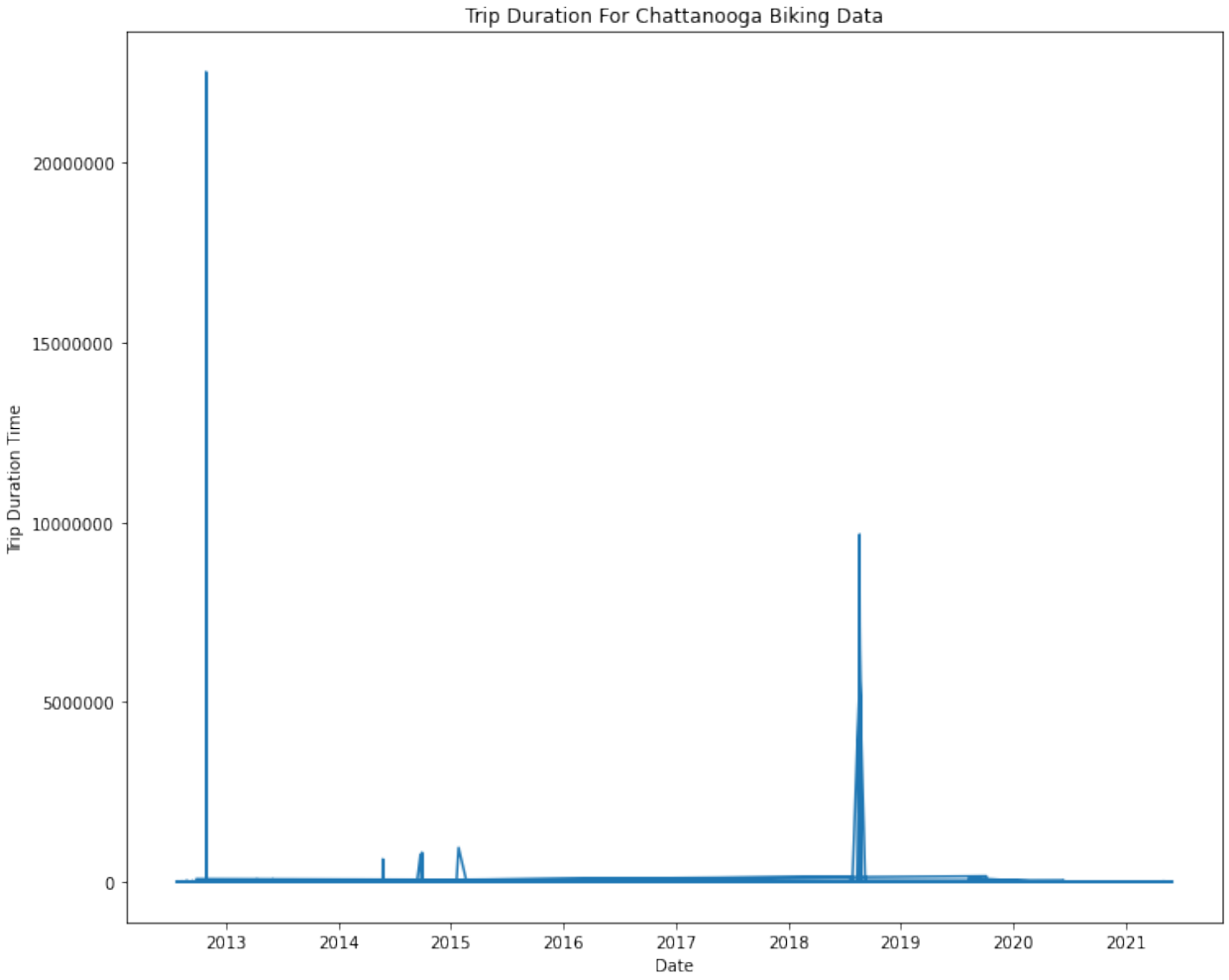


Figure 2. Trip Duration for Customers:

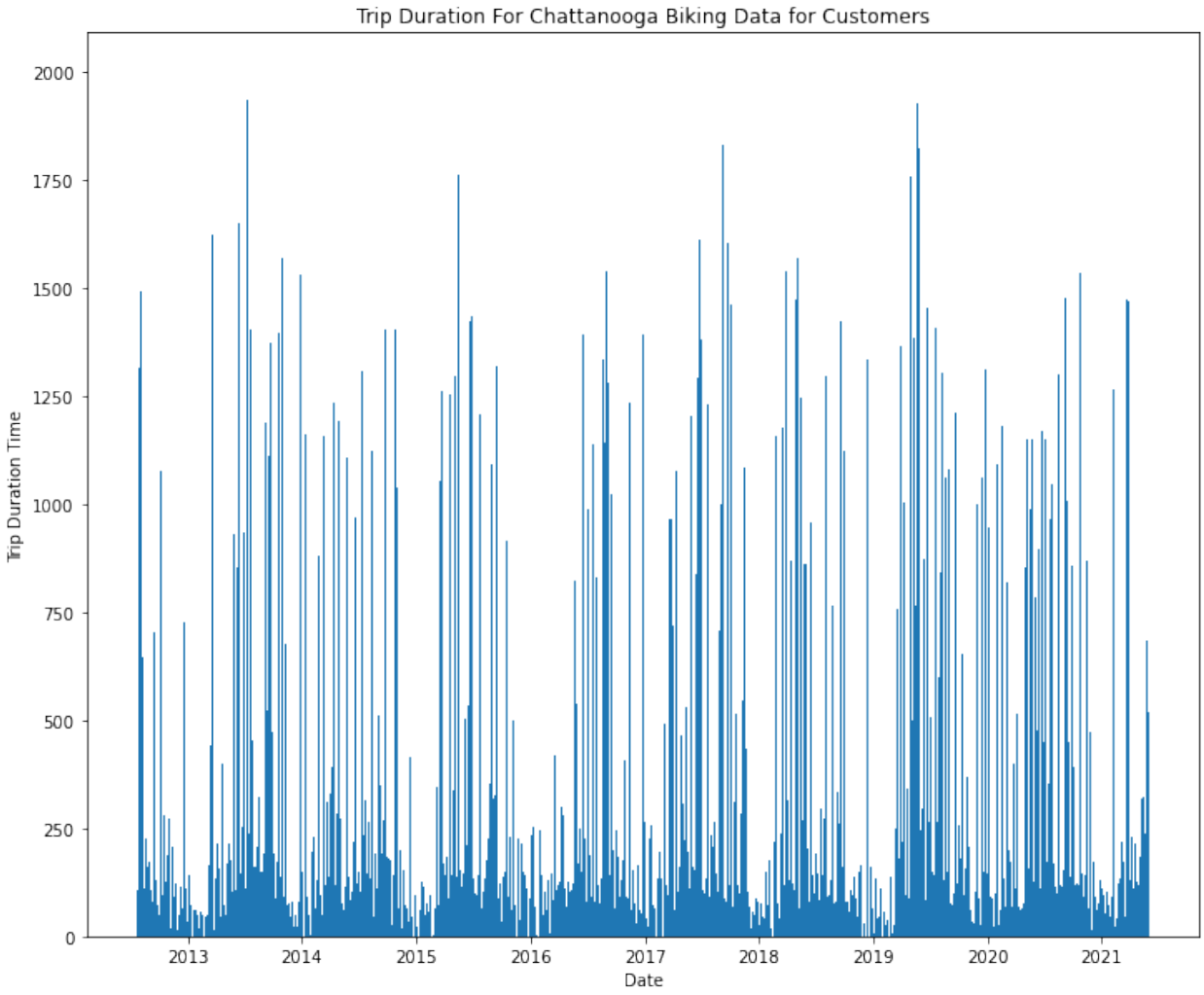
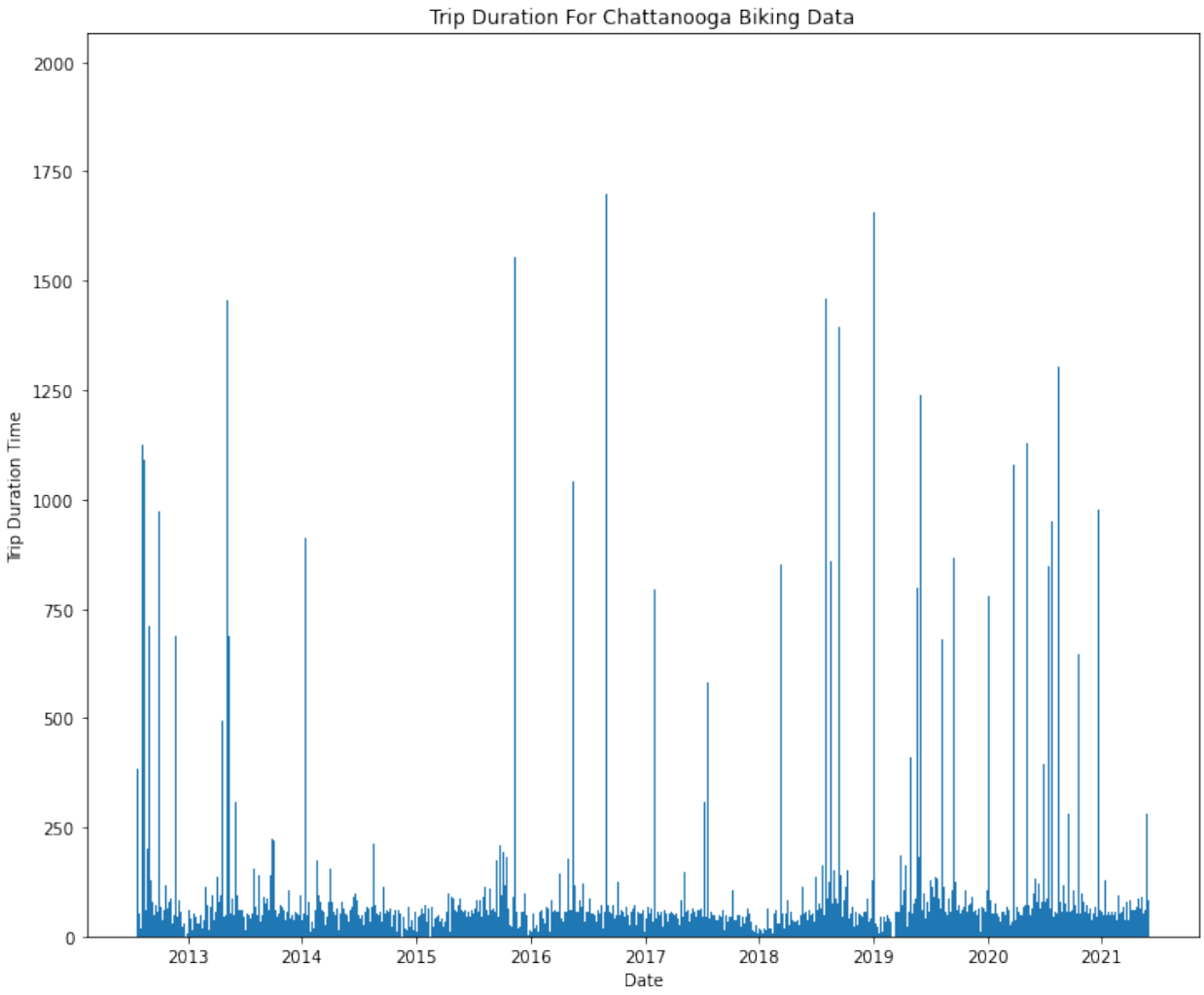


Figure 3. Trip Duration for Members



Analysis

From viewing Figure 1 one can see that there does exist erroneous values in the data set; however, due to this it was decided to create the other visualizations under the 2,000 travel time threshold. From viewing figure 2 one can see that there is more data points corresponding to customers who pay each time to use the bicycles compared to figure 3 where it shows less data points for customers who are subscribers. It is also important to note that both figures 2 and 3 of the data also follows a seasonal cyclicar pattern meaning that during periods leading into the winter months total number of both data points and overall trip duration starts to

lower. The vise-versa of this is true as well where it shows that during the warmer months there is not only more data points but also longer trip times during those months as well. Another observation that one can see is that for the customers that are not subscribers there are more instances of longer trips often reaching above 500 minutes per trip compared to subscribers. It would be estimated that subscribers use the bikes more for commuting / recreation compared to how non-subscribers use the bikes for longer trips. It is also to note that whenever the pandemic started in early 2020 there appears to be no discernable drop in either trip time or amount of trips during this period due to the points in the bar chart being close together and still aparent. The last important piece to see from this data is that at the beginning of 2019 for both customers and subscribers there is a noticable drop in both trip length and total data points leading to small gaps in the graph. While the beginning of 2019 and the surrounding points take place in the winter months it is important to note too that this dip is much larger than the other corresponding time periods.