**Option #1: Capstone Project—Business Intelligence Solution for U.S. Organization**

Daniel Folsom

Colorado State University - Global

MIS480: Capstone – Business Analytics and Information Systems

**Option #1: Capstone Project—Business Intelligence Solution for U.S. Organization**

**Background of State of Connecticut – Office of Policy and Management**

The Office of Policy and Management (OPM) was created in 1977 and “plays a central role in [Conneticut’s] state government, providing information and analysis used to formulate public policy for the State and assisting State agencies and municipalities” (OPM Background, n.d.). The office consists of 7 divisions which are: Administration, Budget and Financial Management, Criminal Justice Policy and Planning Division, Finance, Intergovernmental Policy and Planning Division, Labor Relations, and Health and Human Services Policy and Planning Division. The Finance division is likely the division where the dataset of *Real Estate Sales 2001-2018* resides. Likely the information was given to help explain possible policy initiatives for the Governor’s office in the State of Connecticut.

**Business Problem Description**

The Real Estate Sales dataset was likely initiated to help solve policy questions when it comes to housing. The housing market has a significant impact on the citizens of the state and trying to help encourage more homeownership and more affordable housing is a policy initiative of every governor. Along with the desire to achieve those two goals, one of the possible business or policy problems could be phrased as what factors led to an increase in home sale prices. Additionally, is there a model that can accurately represent these factors and predict home sale price?

**Data Set Explanation and Real-Time Applications**

The dataset used to address this problem, *Real Estate Sales 2001-2018,* has several factors that could potentially influence sale price. Possible variables to consider in the dataset were *list year, sale recorded date, town, address, assessed value, sales ratio, property type, residential type, non-use code, assessor remarks, location* and *opm remarks*. Some of these variables were excluded from analysis for various reasons. Sale-recorded date was excluded in favor of just using the list year, sales ratio was excluded because it is a direct ratio of assessed value and sale amount (including the ratio a second time would be redundant). Address was not used because of processing limitations. Town was changed to a variable to assess if the town was a coastal city or not (yes or no). Property type was excluded in favor of filtering for only residential properties and then having categorical variables for apartment, condo, single family home, and multi family home. Non-use code was used as a variable but was changed to a binary option of having a non-use code or not having a code. This was because of the hundreds of possible categories for non-use codes. Assessor remarks were not deemed usable because of the many possible varying remarks. Location was excluded because of the lack of necessary programs to assess coordinate locations.

The real-time applications of this real-estate data are very real. The state most likely receives paperwork on each real estate sale in the state. If the data is useful for policy initiatives it could be tracked in semi-real time as soon as the paperwork is uploaded to state systems. This information could be given to a real-time dashboard, but may also be possible to have the information given to policy makers on a weekly, monthly, or quarterly basis.

**Description of BI Tools Used**

Two business intelligence tools used were Microsoft Excel and SAS software. These tools were used for two distinct purposes: data cleaning and for multiple regression analysis. One author noted the importance of data cleaning by saying, “The first task in [analyzing] any set of data is to check the data for obvious errors and anomalies” (Hunt & Tyrrell, 2002). The author continues to explain how Microsoft Excel is a perfect candidate for this task. I used Excel’s filtering features to filter out data like non-residential real estate listings. I also used it to change the town variable to a coastal town or not variable. This was done by using VLOOKUP and using a table of potential coastal cities and allowing a yes or no response to that question for each sale. The VLOOKUP function was also used to have population of the town be a variable. Instead of the town name, VLOOKUP would put the population of the town for each house listing.

For regression analysis, a program was needed that would have the capability to process and interpret the variables we would test against it. Excel’s functionality can be limited on the number of variables so another program, SAS was chosen. One author explained that “SAS is a powerful statistical software which caters to the demands in high-performance computing environments” (Ganesan et al., 2004). This high-performance is what was needed for our regression analysis.

**Description of Data Visualization Tools Used**

Data visualization was performed by using Tableau to look at the variables in the dataset. As one author explained, “Tableau’s key benefit is that users can transform large amounts of data into effective visualization reports with a quick and easy-to-use interface” (Eaton & Baader, 2018). This was done originally for this dataset to get a good idea what the variables in the dataset looked like before additional analysis was chosen to be performed.

**Code Explanation**

To perform multiple regression analysis on the dataset, SAS software was used to determine significant factors and the viability of the model. Categorical variables were changed to dummy variables so that regression would be possible. The first program run was a regression program with all possible regression variables that would influence the sale amount.

The first regression analysis with all of the variables did not deliver a very promising R-squared. Which did not give us a good feeling about the current model explaining the variation in the model. Our hope was that by looking for the strongest predictors that we could have the model explain more of the variation. It became apparent that many of the variables that I might have expected to be strong predictors were not, such as *population or being in a coastal city or not*. The strongest indicators were actually the list years of 2015 and 2017. So, year was the only strong predictor in this model. But even with the strongest predictors, the model still only accounted for 0.1% of the variation in the data.

**Description of Data Analytics Outcomes and Benefits**

This model cannot be used as a predictive model to help explain the sale amount in this real estate dataset. If this question is needed to help policy makers, then it would be necessary to look at changing some of the data collected. Perhaps, other variables like square footage, number of rooms, garage or no, swimming pool or no, yard square footage, and other factors could possibly influence sale price and help to explain the model in a better way. The only benefit from running this model is by realizing that these variables do not bring us closer to arriving at a solution for the business problem. However, it does bring us closer to trying to collect a better dataset that may be able to answer our questions and if new collection requirements are approved by the Office of Policy Management, then we may be able to have a greater outcome and potential benefits.

**ScreenShots of Code and Outputs**

Screenshots of all SAS Code and the SAS outputs can be found in Annex A, but in this part of the report we will explain the process and the outcomes of that process. We first did regression analysis on all possible regression variables. Then realized we had to get rid of one of our categorical variables in list year and for housing type. This did not significantly improve the model, but was an important process because too many variables that are related, it is necessary to take out a category.

**ANNEX A – SAS CODE**

Program 1

proc reg data=WORK.IMPORT1;

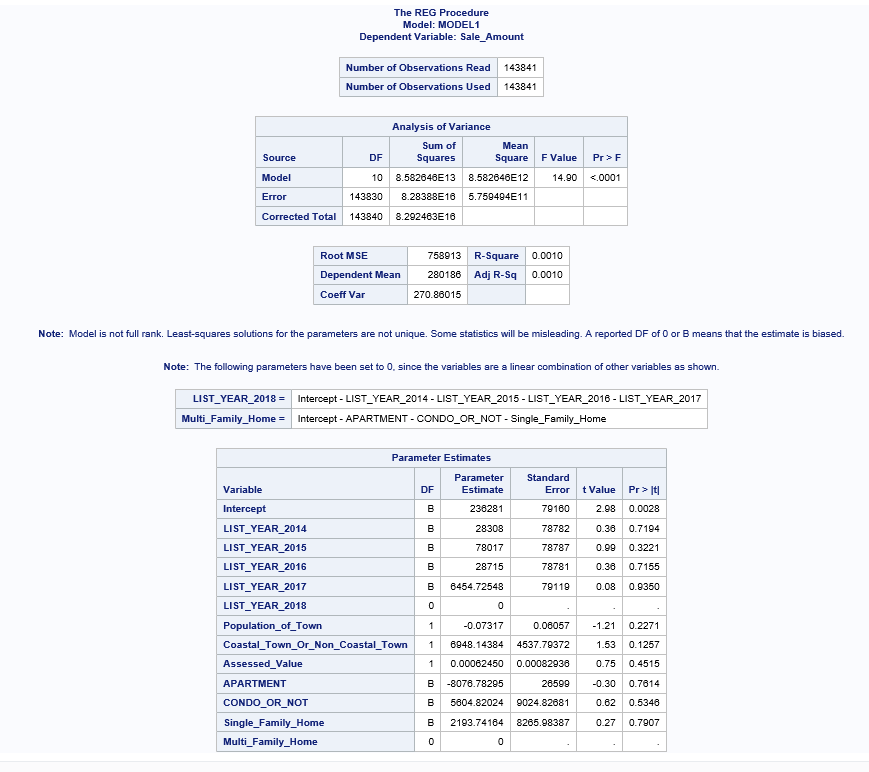
model Sale\_Amount=LIST\_YEAR\_2015 LIST\_YEAR\_2016

LIST\_YEAR\_2017 LIST\_YEAR\_2018 Population\_of\_Town

Coastal\_Town\_Or\_Non\_Coastal\_Town Assessed\_Value CONDO\_OR\_NOT Single\_Family\_Home Multi\_Family\_Home

non\_use\_code;

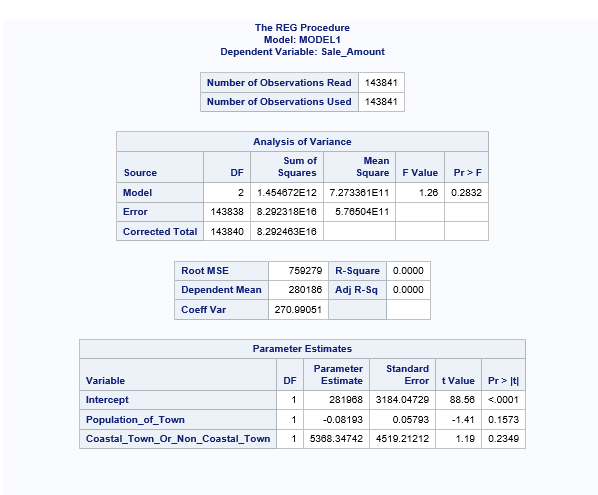
run;



Program 2

roc reg data=WORK.IMPORT;

model Sale\_Amount=Population\_of\_Town Coastal\_Town\_Or\_Non\_Coastal\_Town;



Program 3

proc reg data=WORK.IMPORT1;

model Sale\_Amount=LIST\_YEAR\_2015 LIST\_YEAR\_2016

LIST\_YEAR\_2017 LIST\_YEAR\_2018 Population\_of\_Town

Coastal\_Town\_Or\_Non\_Coastal\_Town Assessed\_Value CONDO\_OR\_NOT Single\_Family\_Home Multi\_Family\_Home

non\_use\_code;

run;



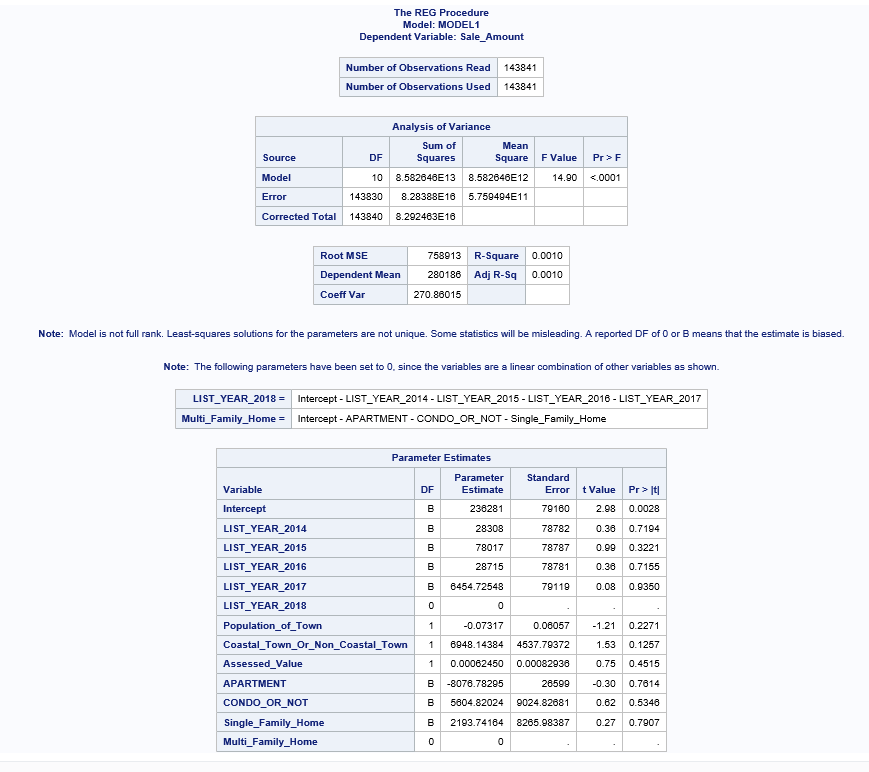
Program 4

proc reg data=WORK.IMPORT;

model Sale\_Amount=LIST\_YEAR\_2014 LIST\_YEAR\_2015 LIST\_YEAR\_2016

LIST\_YEAR\_2017 LIST\_YEAR\_2018 Population\_of\_Town Coastal\_Town\_Or\_Non\_Coastal\_Town Assessed\_Value APARTMENT CONDO\_OR\_NOT Single\_Family\_Home Multi\_Family\_Home;

run;



Program 5

proc reg data=WORK.IMPORT1;

model Sale\_Amount=LIST\_YEAR\_2016 LIST\_YEAR\_2018 Population\_of\_Town

Coastal\_Town\_Or\_Non\_Coastal\_Town Assessed\_Value CONDO\_OR\_NOT Single\_Family\_Home Multi\_Family\_Home;

run;

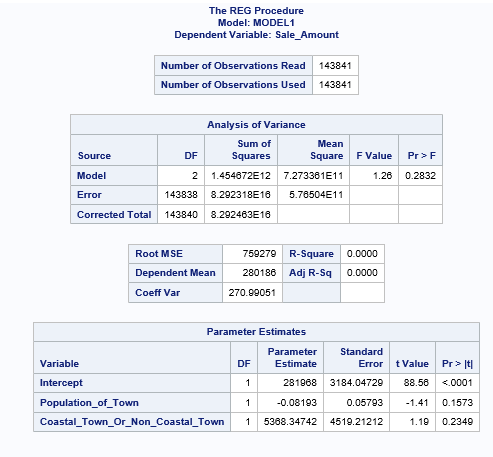


PROGRAM#6

proc reg data=WORK.IMPORT1;

model Sale\_Amount=Population\_of\_Town Coastal\_Town\_Or\_Non\_Coastal\_Town;

run;

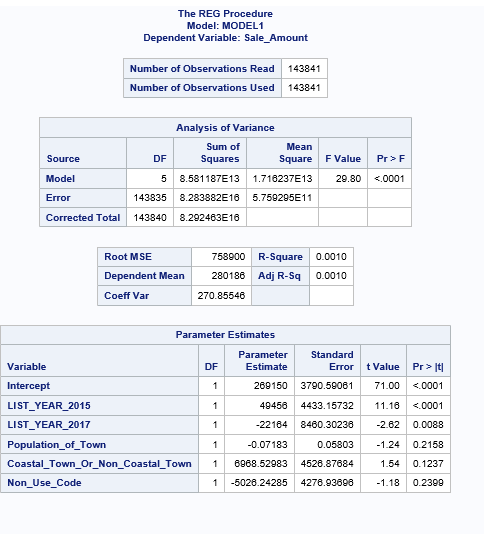


Program 7

proc reg data=WORK.IMPORT1;

model Sale\_Amount=LIST\_YEAR\_2015 LIST\_YEAR\_2017 Population\_of\_Town Coastal\_Town\_Or\_Non\_Coastal\_Town non\_use\_code;

run;

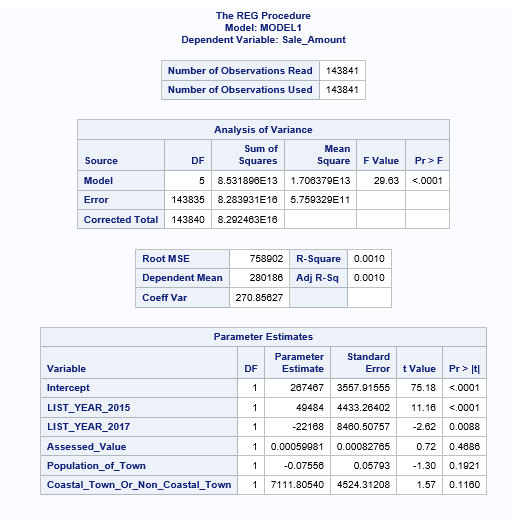


Program 8

proc reg data=WORK.IMPORT1;

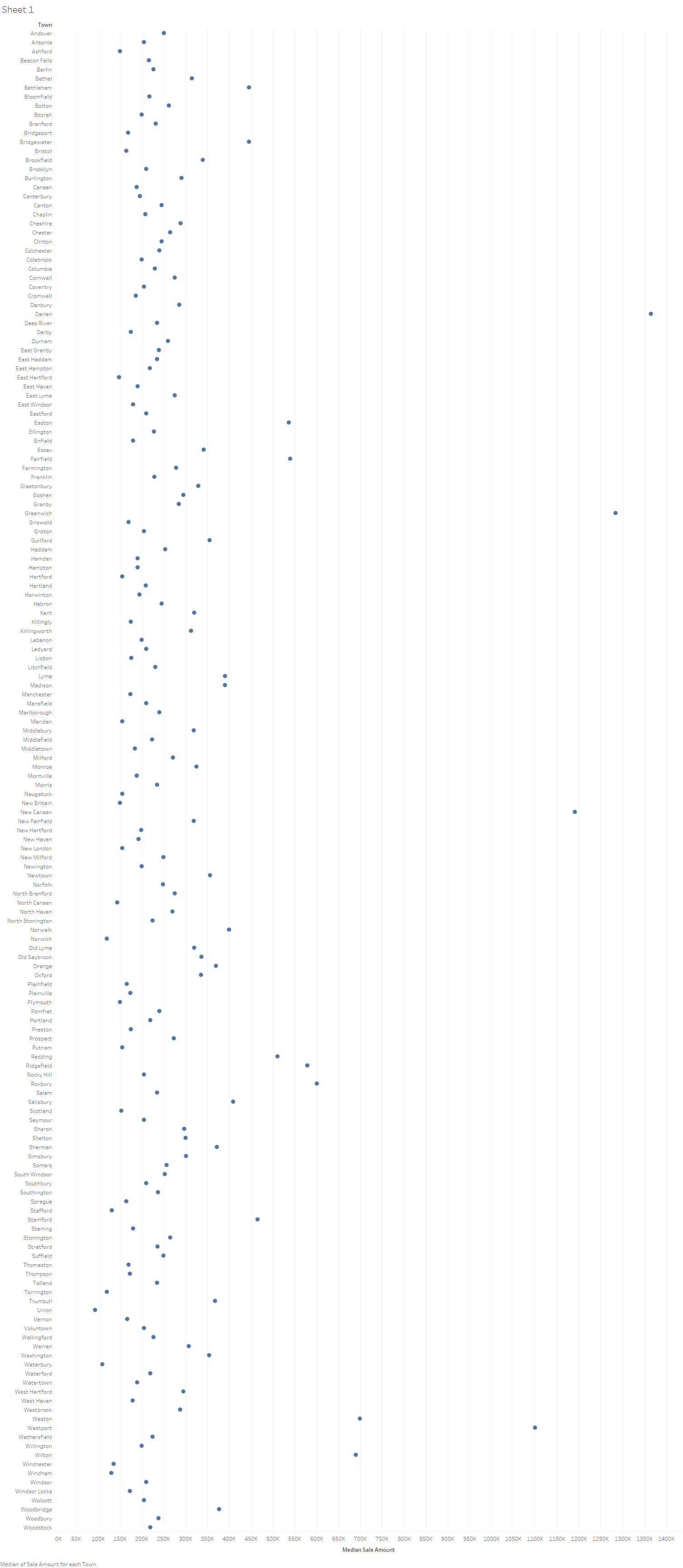
model Sale\_Amount=LIST\_YEAR\_2015 LIST\_YEAR\_2017 assessed\_value Population\_of\_Town Coastal\_Town\_Or\_Non-Coastal\_Town;

run;



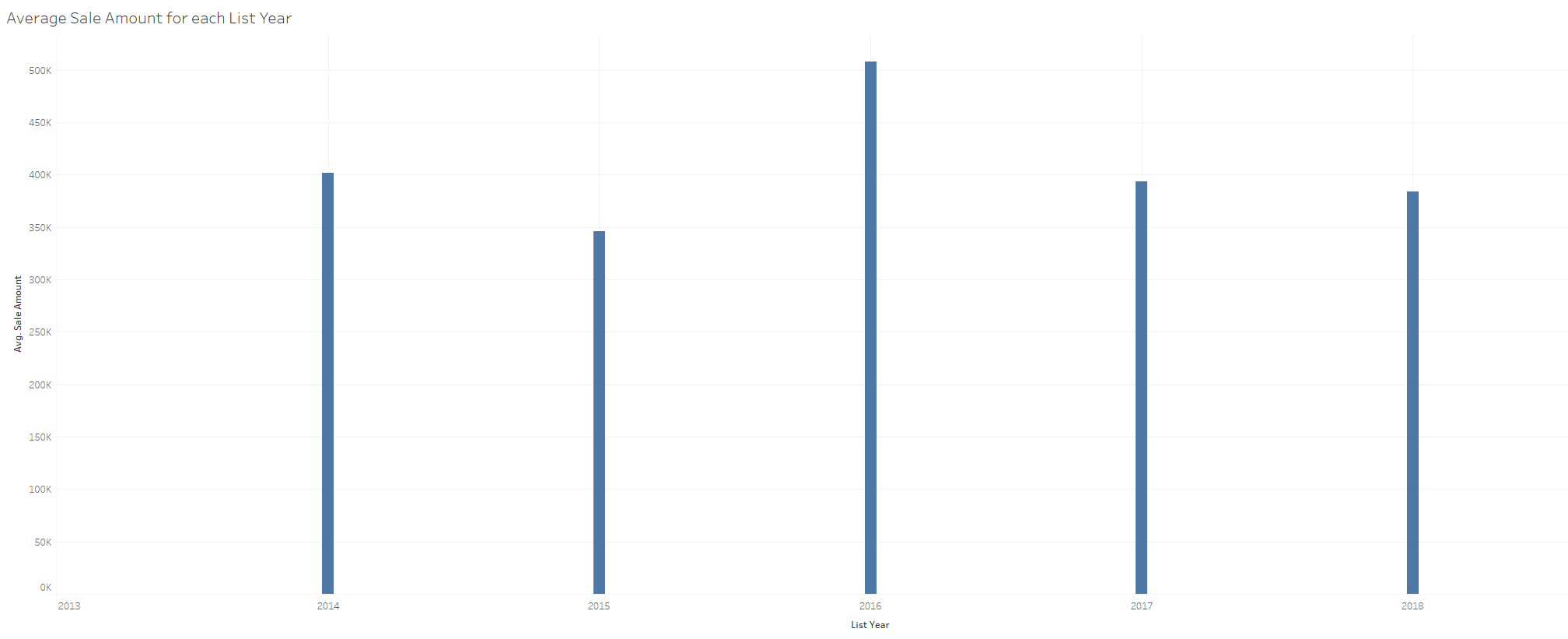
**ANNEX B – TABLEAU VISUALIZATIONS**

Sale Price VS Town

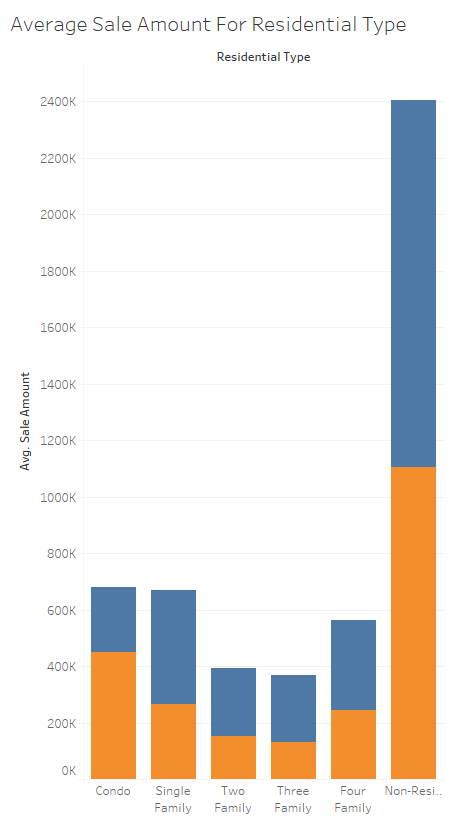


In this visualization, we are able to determine that most of the towns stay in a reasonably tight range of 200K to 500K. This range seems to take into account what we would expect in many situations would be the price of a home. We do see some notable exceptions like Darien which has an average sale amount of 1.6 Million.

Avg. Sale Price VS List Year

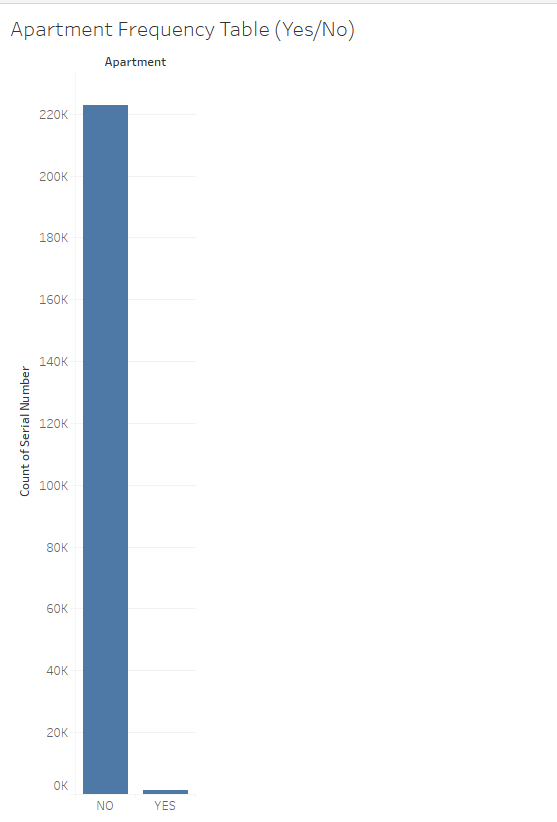
+

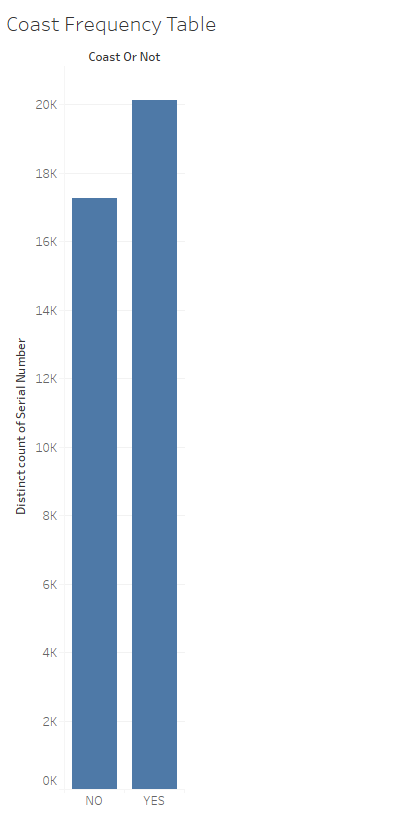
In this bar chart we see the sale prices of the list years of 2014,2015,2016, 2017, and 2018. We can see that 2016 reached the peak average sale’s price and was the most successful list year. Avg. Sale Price VS Residential Type

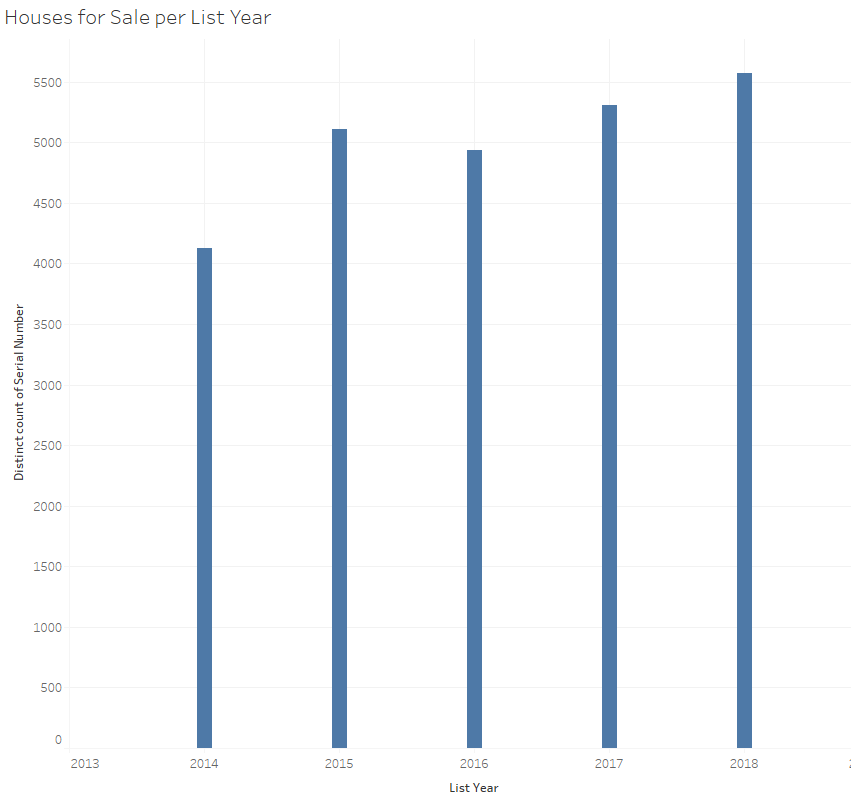


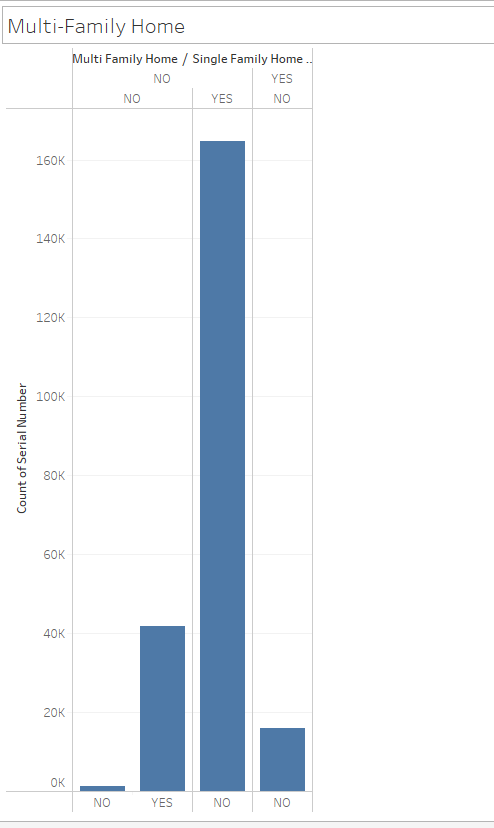
This chart shows us that condos and single family homes may be more expensive than would be expected, but perhaps this has more to do with being in a more populated are. The biggest difference in residential type comes when a real estate asset is categorized as Non-Residential.

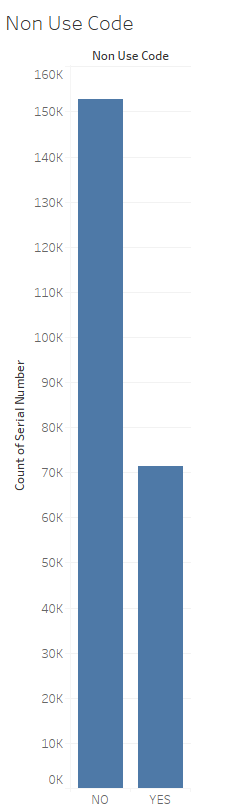
Other Variable Visualizations

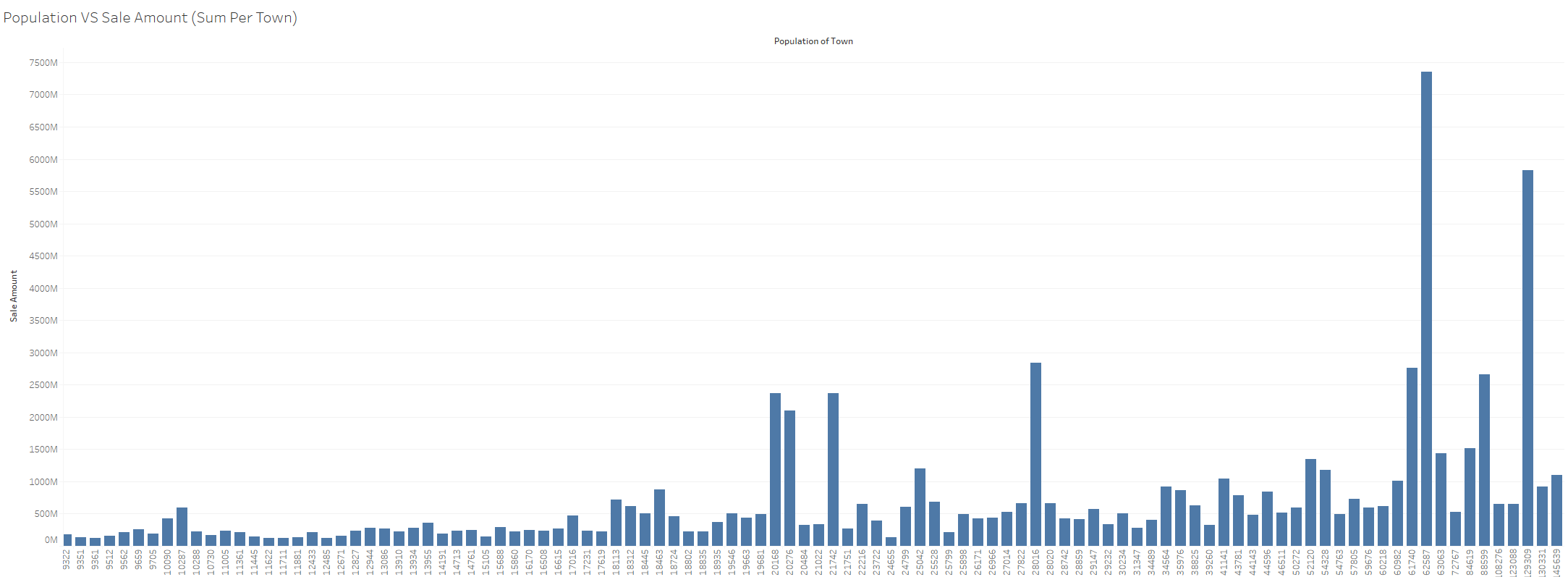


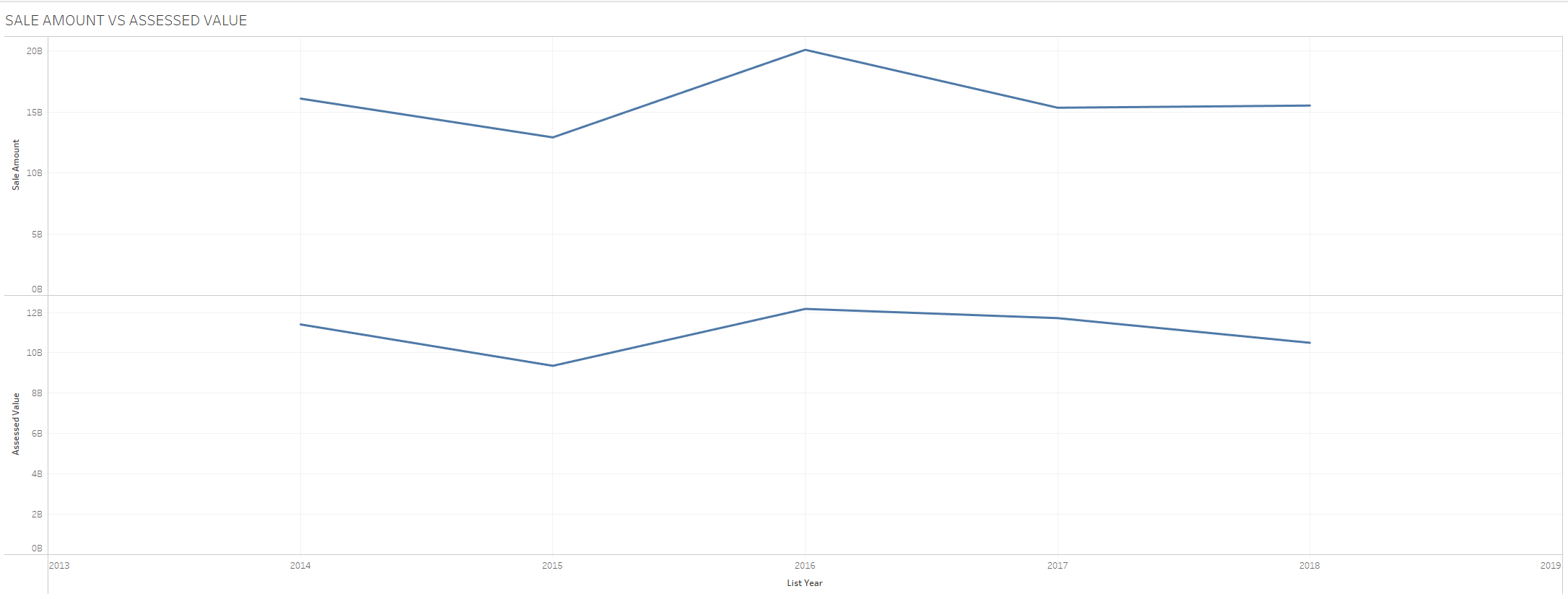












**References**

*Connecticut Cities by Population*. Connecticut Outline. (n.d.). https://www.connecticut-demographics.com/cities\_by\_population.

Dennis, A., Wixom, B., & Roth, R. (2019). Systems analysis & design (7th ed.). Wiley.

Eaton, T. V., & Baader, M. (2018). Data Visualization Software. CPA Journal, 88(6), 50–53.

Ganesan, R., Sampath, V., Okogbaa, O. G., Polavarapu, I., & Weatherspoon, M. (2004). Regression and ANOVA: An Integrated Approach Using SAS Software (Book). IIE Transactions, 36(12), 1211–1216. https://doi.org/10.1080/07408170490507855

Hunt, N., & Tyrrell, S. (2002). Cleaning Dirty Data in Excel. Teaching Statistics, 24(3), 90–92. https://doi.org/10.1111/1467-9639.00096

*OPM Background*. CT.gov. (n.d.). https://portal.ct.gov/OPM/Root/About-OPM/OPM-Background.

*Real Estate Sales 2001-2018 GL: Connecticut Data*. State of Connecticut - Open Data. (2021, February 23). https://data.ct.gov/Housing-and-Development/Real-Estate-Sales-2001-2018-GL/5mzw-sjtu.

Wikimedia Foundation. (2020, April 25). *Coastal Connecticut*. Wikipedia. https://en.wikipedia.org/wiki/Coastal\_Connecticut.