Overview of Data Science

Cayla Burch

Bowie State University

Symbolic Computation Using Big Data

CTEC 298

Fall 2021

Jesse Bemley

**Introduction**

Data science is a field that involves other disciplines to focus on one study. Data science can include programming, statistics, analytics, data mining, data visualization, machine learning, and other related disciplines. Data science uses these disciplines to withdraw valuable insights from structured and unstructured data and use those valuable insights across various application domains and also to inform decisions. CTEC 298 is an extension of CTEC 128 in regards to using additional tools and processes to gain more valuable and actionable insights from data. This paper will discuss and show various tools used throughout CTEC 298, as well as, the previous tools and dataset that were used in CTEC 128.

**Summary of CTEC 128 Paper**

Across the country, there are many factors that are taken into consideration in homeownership. To own a home, many factors are taken into account such as income, credit, age, location, ethnicity/race, demographics, accessibility to resources, and interest rates. Owning a home is an important milestone for many people. The purpose of my CTEC 128 project is to examine the different factors that contribute to homeownership and to analyze which factors contribute the most in home buying. The data sources used in my project were datasets from the American Housing Survey (AHS) and the International Monetary Fund (IMF) to evaluate factors related to homeownership in the United States and in other countries around the world. A few main factors that influence homeownership are demographics, location, credit, income and age. The problem that was being addressed in my CTEC 128 project was gaps in homeownership, if any. Due to the fact that there are many factors that influence homeownership and that every individual American comes from different backgrounds and environments, there could be gap, disparity, or similarities in regard to homeownership. The main project objectives were to discover which factors influence homeownership the most; research homeownership gaps among accessibility to resources, loans, location, and income; and to research homeownership rates and factors in other countries and compare that with the United States. The main project questions were what can be done to decrease gaps in homeownership; what specific factors contribute the most in homeownership; does the location, race/ethnicity, and occupation affect the tax rate of a potential home you want to own; and based on the information found on homeownership factors in other countries around the world, what conclusions can be made? The first dataset that was used is the “American Housing Survey 2019 National Public Use File”. The obtained data file was for only the year of 2019, although the Census Bureau website has data files available for a five year time frame. This dataset comes from the American Housing Survey (AHS). AHS is promoted by the Department of Housing and Urban Development which is overseen by the United States Census Bureau. AHS conducts surveys every other year and includes current information about the cost and quality of housing in the United States and major metropolitan areas. Government and private organizations use the AHS data to influence decisions about programs that affect citizens’ income, age, and race. The second dataset that was used derived from the International Monetary Fund (IMF). The IMF is an organization composed of 190 countries. Their goal is to promote global monetary cooperation, establish financial stability, facilitate increasing employment and economic development, and to minimize poverty all around the world. IMF’s main goal is to guarantee the stability of the international monetary system. The IMF conducts their research on a quarterly basis. The specific datasets we are using from IMF are “Price to Income Ratio”, “Credit Growth Annual Percentage”, and “House Prices Around the World”. In conclusion, only one of the objectives was able to be answered which was comparing homeownership factors in other countries to the United States. The dataset that was used only gave information on finances such as: poverty level threshold, market values, household and family income, housing costs, etc. The dataset did not include the ages, races/ethnicities, or specific locations in the United States, therefore I could not definitively answer our analytic question and objectives that involved age, race/ethnicity, or location. However, by analyzing the datasets, I got a better understanding of a realistic financial standpoint when it comes to homeownership. As far as the IMF datasets, I was able to get a better idea of homeownership factors as far as credit, house prices, and price-to-income ratio. From the dataset I understood that as credit increased house prices increased as well, however, house prices have been increasing more that annual incomes. Due to the datasets, there isn’t any much correlation or similarity between homeownership in other countries compared to the United States.

**Description of CTEC 298 Material Submitted**

\*insert screenshot of CTEC 298 assignments submitted\*

For week 1 and 2, the activities were the Dataquest Analyst Tutorials. The tutorials included Python Missions, programming in Python, variables and data types, lists and for loops, conditional statements, dictionaries and frequency tables, and fundamental functions. These tutorials were essentially a refresher for Python. For weeks 2 and 3, the activities were the installation of Jupyter Notebook, GitHub installation, and the Matplotlib tutorial. The Matplotlib tutorial demonstrated how to create data visualizations. The Jupyter Notebook installation was installed so that we can create the visualizations in and the GitHub installation was for all of the CTEC 298 assignment to be stored in. For week 4, the activities were the Panda, Numpy, and Matplotlib assignments. The Panda and Numpy assignments gave insight into functions and code on how to manipulate files for data extraction. The Matplotlib assignment was where we used our data from CTEC 128 to make six visualizations which included: bar graph, histogram, area plot, multiplot, and a pie chart. For weeks 5 and 6, the activities were the six plots, the tableau install, and the tableau plots. The six plots assignment was a continuation of the Matplotlib visualizations where we discussed the graphs we made. The tableau assignments consisted of installing the Tableau application and creating two Tableau visualizations.

**Original Dataset (Head and Tail)**

Table

Description automatically generated

This is the homeownership dataset from AHS before it was wrangled.

**Wrangled Dataset (Head and Tail)**

Table

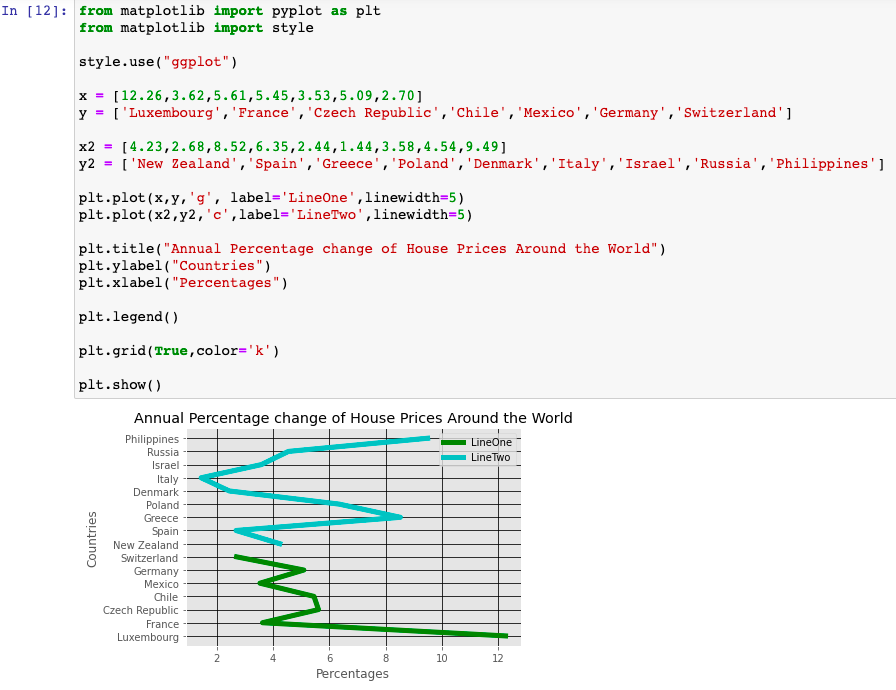
Description automatically generated

This is the homeownership dataset after wrangling.

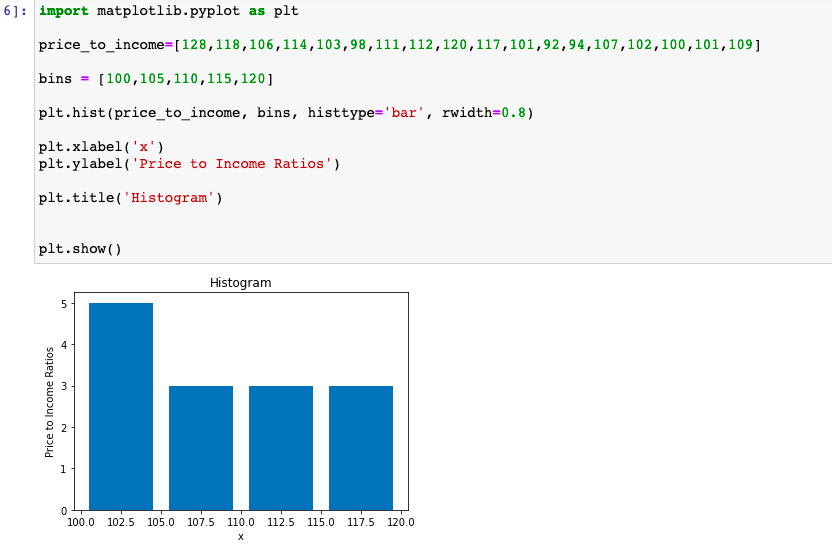
**Plot Deliverables**



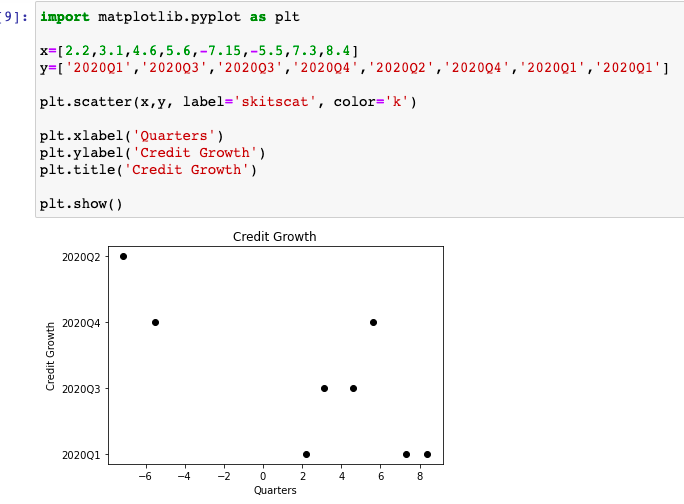
The first plot shown here is the bar graph. These are best used for categorical values. In this data visualization, I chose to represent the averages of first time homebuyer and non first time homebuyers. On the y axis, I put the total averages for monthly housing costs, household income, total remaining debt, and market value. On the x axis the averages are displayed in thousands and the blue represents the first time home buyer and the non first homebuyer is represented by orange. I imported the Pandas and Matplotlib libraries to create the visualization and I used a dataframe for the columns and index.



The second plot that is represented is the multiplot. In this visualization. These are best used for showing change over time. In this data visualization, I chose to represent the annual percentage change of house prices around the world. On the y axis is the countries and the x axis is the annual percentage change. I chose to put the countries on the y axis and to make the visualization horizontal because it looks more visually appealing that way and it is easier to read the countries. I used the Matplotlib library with the functions of pyplot and style to create the visualization.



The third plot shown here is a histogram. Histograms are best used for demonstrating continuous data or determining the distribution of the data or how spread out the data is. I chose to represent the price to income ratio for different countries. On the x axis is the ratio of price of houses to income in different countries. I used only the Matplotlib library to create this visualization. This graph was not well suited for any of my data, so I did the best I could.



The fourth plot shown here is the scatter plot. Scatter plots are best used for when comparing two variables. In this visualization, I chose to represent credit growth in annual percentage change of different countries in different quarters of 2020. On the y axis, are the quarters of 2020. On the x axis the annual percentage change for credit growth. I used on the Matplotlib library to create this visualization.

Chart, pie chart

Description automatically generated

This fifth plot is a pie chart. Pie charts are best used when comparing parts of data to as a whole. In this visualization, I chose to represent price to income ratios as percentages for the second quarter in 2020 in different countries. Most of the countries in this dataset had the same or similar percentages, so I tried to spread the data out by choosing the most countries with differing values. I only used the Matplotlib library to create this visualization. This visualization also was not well suited for my data.

Text

Description automatically generated

The last plot is the stack plot. Stack plots are best used for conveying total amounts over time in subcategories and showing the rise and fall of data over time. I chose to represent amounts in homeownership such as total tax amount, insurance amount, and maintenance amounts. The y axis is the actual total values in thousands. The x axis is just numbers to show the rise and fall of the amounts in relation to each other. I only used the Matplotlib library for this visualization.

**Tableau Plots**

Graphical user interface, application

Description automatically generated

This first Tableau visualization is a bar graph. I chose to represent the mortgages for first time homebuyers which represented in the graph by the number “1” and non first time homebuyers which represented by the number “2” on the x axis. The y axis is displaying monthly mortgages in thousands. I also chose to have the mortgages to be represented by count instead of sum.

Chart, surface chart

Description automatically generated

The second Tableau visualization is another bar graph, but is displayed horizontally. I chose to represent the average credit growth in annual percentage change. On the y axis, the corresponding countries are labeled. On the x axis is the annual percentage change in negative and positive values.

**Conclusion**

Data Science is a powerful field that encompasses other disciplines and also focuses on meaningful data. The goal of each assignment for CTEC 298 was to become more knowledgeable about different Python based programs such as: Anaconda, Pandas, Numpy, and Tableau to assist us in extracting meaningful and valuable insights from data. Furthermore, this class demonstrated how we can create visualizations from data. While becoming more knowledgeable about the various tools, we used our CTEC 128 data to cultivate the visualizations and used the python code to extract data.

**References**