Megan Moore
1014735
CIDM 5310

Final Project

This project was a long journey which I find myself wishing was far more fruitful in my deliverables. For the material I first came across on Data.gov regarding emerging engineering technologies, I regrettably was unable to find a useful dataset to use for this project within our short time span. As excited as I was to do some research on this topic, most of the data was research type essays and not csv or flat file types which can be used for SQL type queries. I had to recalibrate my approach within our short window. As such, I spent time going through Data.gov specifically searching for topics associated with csv type files. Doing so revealed a much shorter list for use but could still of interest. I decided to choose date on another topic that was not only relevant to myself but that I was curious to dive further into. This topic is Housing Affordability. As my family only became first time homeowners about 4 years ago, I wanted to see how our age and income data compared to the most recent available set of data. The Housing Affordability Data System (HADS) contains data up to 2013 which honestly is very close to when we ourselves bought our home in 2016.

The specific questions I wanted to research and review the data on are as follows:

- Research Question 1: The relationship between an individual's age and whether they rent or own.
- Research Question 2: The relationship between income and if it leads to owning or renting.
- Research Question 3: The frequency of owning compared to the age of a house.

The main purpose of this project is to see overall be able to see the affordability of a house. The ability to afford a home stems from many different variables: How people are able to buy house? Is it because there are many members on the house who has individual income or the houses are transferred from parents' ownership to the next generation? Home ownership has long been one of the priorities of every individual and part of the 'American Dream'. The higher number of family members in a household who work only increases the amount that can contribute towards buying a house, thus speeding up the time it takes to own a house. Larger households maybe reflect a lower age of home ownership. Affordability of house also represents the individual has an income source. This income is a result of being employed. That means the economy of the state or country is good because being employed means one can support a family otherwise, due to family costs running so high, renting an apartment might be more favorable and share the rent cost with the other family members who also may not be employed. So, owning a house somewhat relates to a good economy. It also may relate to the education level of that person. If someone attended college, this would usually be reflected in a higher salary and thus greater discretionary income. Not to say those who don't attend college aren't homeowners, but a trend might be seen where higher incomes correlate to newer or most expensive properties.

I felt the information pulled from studying the queries for the question I selected data sets from will be highly representative of what the average affordability is for individuals and paint a clear picture to the long-discussed rent or buy dilemma most first-time home buyers find themselves in. Also, if the business intelligence infrastructure is setup correctly, this project could be expanded to encompass additional years of information which could be used to study trends over time. With inflation and other variables such as home prices drastically increasing over the years, more studies could be compiled and further research questions asked. Also, further insights could be derived reviewing poverty thresholds

and the trends over time. Further, this project could be expanded to review the data located at OCED to compare housing affordability between nations on a global level.

In terms of relevance and data preparation, the Housing Affordability Data System (HADS) is a set of files derived from the 1985 and later national American Housing Survey (AHS) and the 2002 and later Metro AHS. The HADS files are located on the U.S. Department of Housing and Urban Development (HUD) government website. The HUD provides interested researchers with access to the original data sets generated by PD&R-sponsored data collection efforts, including the American Housing Survey, median family incomes and income limits, as well as microdata from research initiatives on topics such as housing discrimination, the HUD-insured multifamily housing stock, and the public housing population. This work is very important to understanding the overall health of our nation in housing and urban development. Their main concern is the ups and downs of the ownership or rent of any particular house or an apartment. If more people are renting the apartment means people are moving from different locations to this place or if more people are buying house means they want to settle down with their family. The HADS data is more like an observational data collection.

Per the HUD website, the HADS system categorizes housing units by affordability and households by income, with respect to the Adjusted Median Income, Fair Market Rent (FMR), and poverty income. It also includes housing cost burden for owner and renter households. These files have been the basis for the worst case needs tables since 2001. The data files are available for public use, since they were derived from AHS public use files and the published income limits and FMRs. The main data sources are the American Housing Survey (AHS) national sample microdata, for the odd numbered years in 1985-2009 and the AHS metropolitan sample microdata for 2002-2009. Poverty income is based on the Census Bureau's official thresholds with Area median income (AMI) and Fair Market Rent (FMR) data originating from HUD calculations. The HADS datasets contain no proprietary or confidential data. As stated on the website, "The purpose of these datasets is to provide housing analysts with consistent measures of affordability and burdens over a long period. By using these available files, this gives the community of housing analysts the opportunity to use a consistent set of affordability measures."

Some background regarding HADS proves to be very interesting. The HADS grew out of a project to provide similar tabulations to the Millennial Housing Commission (MHC) for the years 1985, 1995, and 1999. As described in the HADS documentation file, the Millennial Housing Commission was established by the U.S. Congress in 2000 with the mission "to identify, analyze, and develop recommendations that highlight the importance of housing, improve the housing delivery system, and provide affordable housing for the American people, including recommending possible legislative and regulatory initiatives." The strength and value of the HADS is that it incorporates more than twenty years of housing data using assumptions and computations consistent with the practice of the housing analysts that contributed to the MHC.

Regrettably, I ran into many technical difficulties these last couple weeks aside from not having prior background in this material and dealing with a sick family member. It truly was a perfect storm for such a short semester. The first issue began with being unable to connect to a REST API. I thankfully was able to use SQL and the VPS server to upload my data set file to myPHPadmin using the Filezilla method. Soon after this occured, I ended up needing to reset my root password and this cascaded in a long spiral of needing to reinstall almost everything from the beginning in order for it to function properly. After this lengthy and tedious process, I was able to install Jupyter, then the next issue that confounded me

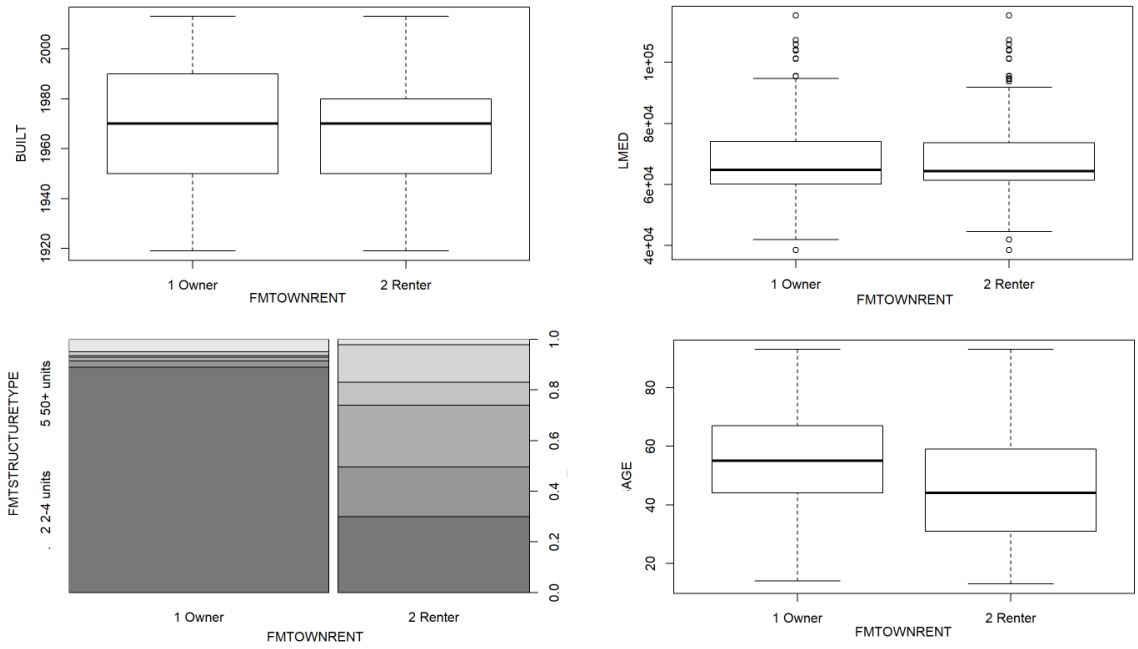
Megan Moore
1014735
CIDM 5310
was the actual notebook would never open in my browser. After researching and trying multiple browser types and online jupyter help forum suggestions, I found myself still in the same spot. Jupyter would show it was running with a local host address but that is where things would stop. I could start and stop the notebook session without issue using Putty but simply could not view it or build the visualizations I strived for. Per the Jupyter help site, this is a windows specific issue but there was no easy 'fix'. If the semester was longer, I feel I could have finally worked this kink out and had a true interactive dashboard with plots and slide bar style widgets that would allow the consumer to adjust the range of data they wished to study based on certain variables.

The variables I determined would be as follows. The dependent variable in this study is whether one owns or rents the house or apartment where they live. In this project, my main concern was to see if the houses or apartments are owned or rented and then, if possible, determine if there are certain factors which contribute to these outcomes. The independent variables I reviewed were age, number of people in the household, income and the age of the house. There are many independent variables compiled in the HADS dataset but due to file size limits and time constraints in myPhpAdmin, I trimmed the file to reflect only the first 20,000 records of the full 60K records for the year 2013 and then ran SQL code through my VPS connection to calculate answers for the questions I posed at the beginning of this project.

Reviewing the results of the SQL queries, a few insights were able to be determined. There was no difference to be found in the minimum or maximum ages of people who owned compared to rented. The was however a distinct different in the average age regarding the same own vs rent. The average age of an owner showed to be 53.8 years old while the average renter was 38 years old. This makes sense based on older individuals have had more time to find better paying employment based on their years of experience as well have larger savings accounts and perhaps a greater need for a home to support a family. The inference can be made that definitely people aging 50+ own houses in higher number than younger people because of money. But income is not only the main source to own the house. There could very well be other reasons that people own houses such as a shared family house, or a relative's house which was passed to nearest alive relative. Next, the average income of the entire dataset was compared to the average income of only those who own. Again, a difference was noticeable. Those who owned saw a significant income increase. Lastly, reviewing the age of the property based on the year the structure was built and seeing if any difference can be seen between own and rent proved surprisingly fruitless. The age of either a home or apartment which was owned or rented was within 4 years. I was estimate that if more datasets from other years was pulled for this same query, a trend could develop which describes the current housing market crisis. The structures are becoming older and newer properties are less affordable. An enhanced BI Dashboard could be constructed to help convey these relationships and plot out the trends based on age, income and property age. This dashboard could be easily updated to include other years data and be kept current so users see a more real time display of the environment around them. The business intelligence which could be provided and insights derived from larger data sets would likely be key to painting stories for the decision makers needing answers to their critical questions.

As an example, I built a few charts from excel and arranged them in the next image which would be representative of interactive dashboard I would have built in Jupyter had I not been plagued with technical access dilemmas.

Megan Moore
1014735
CIDM 5310



To conclude this project, it is well understood that every economic crisis causes not only a contraction of the main indicators of wealth but also changes the distribution, accentuating or diminishing the social equity. The key problem in pinpointing the underlying cause is the sheer number of decision makers in the development value chain from planning agencies, organizations of infrastructure and builders, and then the levels of complexity of the relationships among them all. A future BI DSS system could be able to cross-cut masses of data, pointing to areas best capable of supporting large-scale development. In an ideal world, this system could even possibly point out infrastructural obstacles on a detailed level as failing sewage systems or inadequate transportation solutions which inhibit development at several sites at once.

Deliverables: https://m3gan.xyz/project.html

Login information for VPS server lex jp4eva	Login for http://m3gan.xyz hammond Password123#@!	Login information for https://m3gan.xyz/phpmyadmin/index.php tim Password123#@!

A backup of MySQL database file and Sherman BI Roadmap template included with this project submission.

Works Cited (APA format)

American Housing Survey: Housing Affordability Data System. American Housing Survey: Housing Affordability Data System | HUD USER. (n.d.). http://www.huduser.gov/portal/datasets/hads/hads.html.

Vanderbroucke, D. A. (2011, January 28). *Housing Affordability Data System Documentation File*. https://www.huduser.gov/portal/datasets/hads/HADS_doc.pdf.