**Data 670 Data Analytics**

**David Hannon**

**Professor: Dr. Hany Saleeb**

**Housing Affordability Across America**

**June 14th, 2021**

**Executive Summary**

It has become the standard within the apartment application process that applicants are able to prove that their monthly income is greater than or equal to three times the monthly rent cost. While this is meant to lower eviction rates and ensure the renters can afford the space, this rule provides a barrier to many who might be moving to secure a new job or are looking to start a new career in a specific city or county. Due to this standard, it should be expected that the areas with notoriously high rent prices also have higher average incomes to make the area affordable to its citizens.

This purpose of this project is to create a model that predicts what the average rent for an apartment (from a studio to four-bedroom) should be given an array of economic factors about the area in which the apartment is located, in hopes of defining an affordable rental market for all. These factors include median household income, unemployment, education levels, poverty levels, and population. The created model should lead to more informed pricing by property managers to ensure their properties are appropriately priced, more informed apartment hunting by prospective renters to know they are not overpaying, and more informed policy making by local governments in order to make their areas more livable.

Table of Contents

[Project Scope 5](#_Toc513933193)

[Problem Description 5](#_Toc513933194)

[Business Understanding 6](#_Toc513933195)

[Organization 6](#_Toc513933196)

[Stakeholders 7](#_Toc513933197)

[Define Business Area 8](#_Toc513933198)

[Business Objectives 8](#_Toc513933199)

[Business Success Criteria 9](#_Toc513933200)

[Background 9](#_Toc513933201)

[Research 10](#_Toc513933202)

[Gaps in this Problem Resolution 10](#_Toc513933203)

[Proposed Project 10](#_Toc513933204)

[Key Performance Indicators 11](#_Toc513933205)

[Project Insights of your Data Analysis 12](#_Toc513933206)

[Project Milestones 12](#_Toc513933207)

[Completion History 13](#_Toc513933208)

[Lessons Learned 13](#_Toc513933209)

[Data Set Description 15](#_Toc513933210)

[Data Set Description 15](#_Toc513933211)

[High-Level Data Diagram 16](#_Toc513933212)

[Data Definition/Data Profile 17](#_Toc513933213)

[Data Preparation/Cleansing/Transformation 22](#_Toc513933214)

[Data Preparation 22](#_Toc513933215)

[Data Cleansing 23](#_Toc513933216)

[Data Transformation 23](#_Toc513933217)

[Data Analysis 25](#_Toc513933218)

[Data Visualization 27](#_Toc513933219)

[Data Visualization 1 27](#_Toc513933220)

[Data Visualization 2 29](#_Toc513933221)

[Data Visualization 3 33](#_Toc513933222)

[Proposed Visualizations 35](#_Toc513933223)

[Predictive Models 37](#_Toc513933224)

[Predictive Model 1 37](#_Toc513933225)

[Predictive Model 2 41](#_Toc513933226)

[Predictive Model 3 44](#_Toc513933227)

[Predictive Model 4 46](#_Toc513933227)

[Predictive Model Review 48](#_Toc513933228)

[Final Results 50](#_Toc513933229)

[Analysis Justification 50](#_Toc513933230)

[Findings 51](#_Toc513933231)

[Review of Success 53](#_Toc513933232)

[Recommendations for Future Analysis 54](#_Toc513933233)

[References 56](#_Toc513933234)

# Project Scope

# Problem Description

Based on RentCafe estimates, the average monthly rent for a person living in the Navy Yard neighborhood in Washington, D.C. is $2,078. The average monthly rent just north of D.C. in Montgomery County, Maryland is $1,793. The average monthly rent just east of D.C. in Prince George’s County, Maryland is $1,415. A difference of over $600 in rent is equivalent to a $7,200 yearly difference in housing costs which can be a significant percentage of someone’s income. Such a disparity in renting prices within a 10 mile radius shows that there are other important factors at play in determining rent prices other than geography.

The proposed factors most driving the differences in rent pricing are median household income, urban influence, unemployment rates, labor force, population, education levels, and poverty levels. The provided datasets from the Department of Housing and Urban Development breakdown America on a county-by-county level. The rent estimates cover the years from 2017-2020 while the driving factor datasets provided by the USDA vary from covering 1970-2019 (education) to only covering the year of 2019 (poverty estimates). Due to these fluctuations, several models will be created to account for the differences in time periods. These models will also fluctuate in their target variable as well since the data provides average rent prices for studio, 1-bedroom, 2-bedroom, 3-bedroom, and 4-bedroom apartments individually. While the focus will be linear regression models, some will also model a created metric of affordability (average monthly household income – 3 times the average monthly rent price) which should either create a positive or negative affordability value which will be used for a logistic regression model.

In attempting to better describe the relationships between all of these economic factors with rent prices, there are three proposed benefactors. The model should be able to intake the various economic factors of the county or area in which an apartment is located and output what the appropriate price would be for the area. This should benefit property manager in being able to set a base rent price for their units, although it will obviously fluctuate based on differences in amenities and features of each location. It should also benefit governments and policymakers who are concerned with housing costs within their cities by informing them of which economic factors to focus on most in order to indirectly adjust rent prices. Most importantly, it will inform prospective renters as to what they should expect to pay within an area given its socio-economic conditions so they can make informed decisions on where to live.

# Business Understanding

In the New York Times article *How Does Your Landlord Calculate Your Rent?* (2018), Kim Velsey describes the way in which many local New York City landlords determine rent prices compared to how larger property management companies would use an algorithm. Some local landlords seem to have set their price for whenever their first resident moved in and refused to raise the rates despite the market around them almost doubling. Some landlords state that they will only increase rent on a second renewal by 1 to 3% because the thing they value most is maintaining good tenants and not having to worry about the repair costs of switching tenants. Within the Rent Stabilization Association, 70% of the members are considered to be small-property owners, meaning a majority of landlords, at least in New York City, tend to be these more personal type landlords that determine rent in the ways previously described.

The other 30% of landlords account for a greater number of total properties as they are the large-scale property managers with at a minimum of 49 units to rent. A case study of San Francisco shows how changing median household incomes led to increased rent: in 2011, the average two-bedroom apartment in the Mission District was $1,900 per month. Due to the tech boom, companies like Google, Facebook, and Apple all established a presence in San Francisco, increasing housing demand and also increasing the average income. By 2012, average rent prices were $3,500 and has continued to fluctuate, peaking at almost $5000. These increasing prices have led to many people not being able to afford the rent, as well as a doubling of the eviction rate, and there has been a spike in homelessness as a result. Similar, but less extreme trends have been seen in many major cities since the tech boom (Lien, 2018). Population also seems to be a major factor in landlords determining of large-scale rent. Due to so many jobs becoming remote during the coronavirus pandemic, San Francisco’s average rents have decreased by 35% as an estimated 20% of tech workers have moved out of the city (Bloomberg, 2020).

### Organization

From the government’s perspective, the relationship between increasing rents coupled with stagnant economic factors has led to increased discussion of rent control. Rent control consists of “caps on price increases for the duration of a tenancy, and sometimes beyond the duration of a tenancy, as well as restrictions on eviction” (Diamond, 2018). This type of legislation has been widely used to protect renters during the coronavirus pandemic but was being heavily debated prior to the shutdown as well. While helping tenants’ financial stability, this type of legislation tends to hurt landlords and the real estate market by depreciating the value of owning rental properties and causing less people to buy a home of their own. It can also hurt the tenants though by causing them to maintain their same apartment despite a change in living needs or by allowing more affluent renters to rent multiple properties, thus decreasing the supply of housing for others.

With the ongoing debate about rent control legislation, rent prices have risen to a point where “the poorest 20 percent of American families cannot afford rent on standard apartments without some subsidy from the government” (Schuetz, 2019). Despite the “income greater than 3x the rent” rule, almost half of all renters spent greater than 30 percent of their income on rent, meaning most renters are barely meeting or breaking that rule. There are some areas in the U.S. where average rental prices are at least six times greater than average incomes. However, “in the Midwest and many smaller metros in the Heartland, prices are less than twice median incomes” (Schuetz, 2019). Since prices tend to vary so much by location, and in some locations are unaffordable to its citizens, it seems that local government should be tasked with finding a way to make their areas more affordable. Schuetz suggests that recent government policies on zoning have only increased rental prices by not allowing the size of the market to increase with the demand and so they mainly rely on providing subsidies to help people afford housing.

### Stakeholders

The greatest stakeholders in this project are the renting population or tenants. The results of this project should lead to more informed spending, in terms of finding the most affordable location for living as well as being knowledgeable as to what a rent should be in a given area. Someone moving to a new city or state for a new job should be able to compare their salary to that of the median household income of the area and have a better idea of how affordable the area is to them. In contrast, someone deciding where they would like to work may use the results of this to compare their expected income to what they should expect to pay within a city or county. Economists could further this to predict future rents based on their predictions for wage increases and population increases within an area.

The other stakeholders in this are the landlords and local governments. The results will allow landlords to see how what is being charged in their area compares to the most comparable areas across the company and inform them if they are either overcharging or undercharging. Governments can see just how affordable their local housing is to its citizens and enact policy to assist. Most research suggests this would lead to subsidized housing, however research like this could be used within the argument for increased minimum wages to drive up median household income or even within immigration debates to show how changes in population size change the overall economy of an area. Also, there is research to suggest that rezoning land would change the market drastically by allowing the development of more housing units to keep prices lower with supply and demand while not forcing others out of their homes.

# Define Business Area

# The business area for this project is within property management and the relationship between property managers, renters, and the government in setting reasonable housing prices for specific areas of the U.S. The government defines anyone paying greater than 30% of their income towards housing costs as “burdened” and anyone paying greater than 50% as “severely burdened”. To avoid burdening anyone with housing costs but still allowing landlords to maintain profit, reasonable rents must be set for the area in which the property is located. This should be driven by what the average person in that area earns each month as well as population (supply and demand). This should also see a cyclical relationship with homelessness as more appropriate rents should lead to less evictions/homelessness and less homelessness should lead to a more economically stable area.

### Business Objectives

The main business objective of this project is to provide a fair and stable market for rent prices across the U.S. The created model should output a calculated, fair price for a given county in the United States to provide a baseline for price. Obviously, there are some unaccounted-for factors such as amenities, parking, etc. that may cause for fluctuation. The model should set the base price given the socio-economic conditions of the area, and then these differences in property could be factored in using different modelling techniques. The objective of this is to calculate a housing market which renters can afford while landlords can still profit.

From the government perspective, the objective is to identify which socio-economic conditions most indirectly affects rent prices. In identifying this, policy makers will be more effective in making change regarding rent-prices since they can create policies that target that specific area rather than attack rent head on with rent control or subsidies.

The final objective is to create an interactive visualization to help prospective renters in budgeting their housing costs. Ideally the visualization could be used on rental websites to give a geographic visual of each county’s “affordability”, a metric that factors in both what the average citizen should expect to earn while living there along with what they should expect to pay in rent. This could hopefully be improved upon given other useful data not currently available, such as factoring in certain amenities or applying certain filters for things like transportation needs, etc.

### Business Success Criteria

The main criteria for business success is being able to produce an accurate model. Creating a model with low squared residuals would mean that it can be used moving forward to predict how much rent should increase or decrease based on the ebb and flow of the population and the economy. Landlords should be able to make informed decisions about how population increase should increase their prices by a certain percentage or how something like increased unemployment should perhaps lower their rates by a certain percentage.

The second main success criteria would be to use correlation coefficients to identify three to five socio-economic factors as being the most related to rent prices and affordability. This would allow local policy makers to reconsider the way the approach creating affordable housing by having them focus on some potential causes rather than on the housing cost itself. Lastly, creating a visual/model that informs prospective renters as to what they should be paying in particular areas would also be considered a success.

# Background

Rents have been following an increasing trend across the United States for the past several decades. While the coronavirus pandemic may have temporarily created an increase of available housing leading to lower rents, in general rents have been increasing at a rate much faster than that of the average salary. This makes finding affordable housing very difficult for people in certain areas. While people may be inclined to move to larger cities to become part of a larger job market, housing costs often create a barrier in doing so. In order to help adjust the market to be in-line with average income as well as other socio-economic factors in a given area, insights must be gained as to which factors most “set the market” for rental prices in an area.

### Research

The census calculates a statistic known as the Income-Rent Gap and has reported on its increase overtime meaning rent prices are increasing faster than renters’ incomes (Mazzara, 2019). The proposed model should help to determine the link between these two variables as well as show changes overtime. A similar model was created by The Urban Institute trying to relate unemployment, income, and mortgages entering forbearance (Goodman & Choi, 2020). This is similar in that it is relating larger economic factors with individual’s affordability of housing, but takes a more specified approach as it is looking at how the COVID pandemic has affected the mentioned data factors. My project fills a gap once again by adding emphasis to location and focusing on renters rather than homeowners.

### Gaps in this Problem Resolution

The gap in the previous research is a lack of determining what factors of an area are causing the rent to be the price that it is. Most articles found state simply that the rent is “driven by the market” but what is driving the market? The case studies previously mentioned show that as more affluent people moved to the area, rent went up. It also showed a positive correlation between homelessness and higher rent as a result. Also, this would imply that greater population led to increased rent as well. The gap is that none of the previous research quantified the effect that these different factors had on the changing rent prices across counties.

# Proposed Project

The proposed project is to create a linear regression model to predict rent price given a variety of socio-economic factors for a specified county in the U.S. The ability of the model to accurately predict what a given rent price should be will both be telling of the power of the model as well as the consistency of rent prices across similar counties across the United States. An analysis of the correlation coefficients will reveal which socio-economic factors are the most correlated with rent prices. The model could be used in the present to assist property managers in determining a ballpark for their market in terms of rent as well as for informing tenants what they should expect to pay in a given area. Also, the model will be used by economists for predictive purposes to determine future rents based on population trends and changes in average salaries.

For the second part of the project, the rent variable will be adjusted to a binary “affordability” variable that links together rent and median household income to determine a positive of negative affordability for a given area. A logistic regression model will then be used to predict positive or negative affordability. The results of this will be analyzed to determine which factors make an area affordable as some of the factors like education may more so effect the income levels while things like population may more so effect the rental prices. This can then be used by governments to test and predict what changes they would need to make to make their city or county more affordable.

### Key Performance Indicators

The most important performance indicator for this project will be an analysis of the correlation coefficients for each variable within the linear regression. Determining each variables relevance in determining rent will be the most useful piece to come out of this analysis for any government concerned about increasing rent, considering rent control, or just wanting to make their city more affordable to decrease subsidies. If the correlation coefficients yield extremely positive between population and rent, then perhaps governments will try to find ways to limit population, etc. Also, for the logistic regression, if the correlation between education and affordability is positive, this could lead to governments providing more educational assistance.

To determine overall model accuracy, a residual plot paired with the average squared error will be a useful KPI. This will show how far off the predictions are from the actual rents charged. In this case though, one could think of the model as trying to set the baseline for what rents should be and thus a large average squared error may be more telling of just how off the rental market is in the United States. If the plot shows a lot of high outliers above the average line, these will likely represent cities like New York or San Francisco with notoriously high rents. Any points below the line may represent smaller towns with comparable socio-economic feature that are “undercharging” their rents.

Various other performance indicators will be used as well. A Normal Q-Q plot of the standardized residuals will help show the normality of the data. One may expect this data to be a bit skewed since most data points will represent non-metropolitan counties that may charge lower rent, with the fewer in quantity, metropolitan counties raising the expected rent calculation. Similarly, leverage may be calculated on some of the outliers as it is expected that the metropolitan cities will have much higher rent than should be expected.

### Project Insights of your Data Analysis

I expect the results of this analysis will showcase the importance of median household income, population size, urban influence, unemployment, and other socio-economic factors in determining average rent prices within each county. It is unclear as to which factors will show a positive or negative relationship since, for example, people coming off unemployment may be settling for low wages and driving down the median household income or the increased employment could be adding to the overall economy and indirectly increasing median household income. However, I believe the results will be useful in informing prospective renters with a baseline of what they should be paying for a comparable area.

I also believe the visualizations combining rent prices with household incomes will show that metropolitan/urban areas are less affordable despite higher incomes, and thus show that the factor variable of urban influence has high correlation with increased rent price. In showing this relationship, as well as how other socio-economic factors seem to drive up rent prices in cities, I hope that city governments could use the correlations of the various factor variables to determine new methods to keep rent more affordable rather than with subsidies and rent control. Perhaps their most effective strategy in enacting changed can be found by trying control a different socio-economic variable with the community.

# Project Milestones

1. Create several linear regression models to predict average two-bedroom rent prices dependent upon unemployment trends, urban influence, etc.
2. Create a logistic regression model for the metric of “affordability”
3. Create a naïve bayes model to compare with the logistic model
4. Create a visualization showing affordability of each county (some metric of the difference between median income and average yearly rent)
5. Identify socio-economic factors that show the strongest relationships with rent prices and affordability

# Completion History

|  |  |
| --- | --- |
| **Week 1** | Find possible datasets. |
| **Week 2** | Define Scope of project. |
| **Week 3** | Background research on problem domain. |
| **Week 4** | Define problem in terms of business. |
| **Week 5** | Prep/Cleanse data. Combine into one Master dataset. |
| **Week 6** | Begin visualizations and transforming the data. |
| **Week 7** | Visualizations created using Tableau. |
| **Week 8** | Visualizations Analyzed. Began model creation. |
| **Week 9** | Modeling- LinReg, LogReg, Naïve Bayes |
| **Week 10** | Conclusions and proposed further research |
| **Week 11** | Editing and revisions |

# Lessons Learned

|  |  |
| --- | --- |
| **Week 1** | Many possible options for project and scope but severely limited by availability of large, public datasets with comparable factors. |
| **Week 2** | Difficult to narrow down a broad idea for a project into one specific research question. |
| **Week 3** | Various ways the variable of rent is calculated and how it effects different groups of people. |
| **Week 4** | Reformatting a social-economic issue into a business lens. |
| **Week 5** | Struggled with getting data input into SQL, was able to use R to complete same tasks. |
| **Week 6** | Refreshed Tableau skills. |
| **Week 7** | Not all variables have correlation to major cities as expected. |
| **Week 8** | Research on necessary packages for model creation. |
| **Week 9** | Experimentation with different modelling types. Research on how to interpret and compare outputs. |
| **Week 10** | Major takeaways from modelling- seeing how the significance levels of each variable with target variable. |
| **Week 11** | Seeing how the scope of the project has changed over the course of the semester. |
| **Week 12** | Learned how to take a broad problem and narrow it down to a specific scope for analysis. Improved skills in outside research. Learned how to weave together multiple modelling techniques to mesh into final conclusions. Improved R and Tableau skills. |

# Data Set Description

The main dataset proposed is provided by The Office of Policy Development and Research under The Department of Housing and Urban Development. It contains the median rent price for every county in the United States and its territories from the years 2017-2021, totaling 23,841 observations. It covers rent prices of studio, 1-bedroom, 2-bedroom, 3-bedroom, and 4-bedroom apartments in each area along with information about the county, area, state, zip code, and population (2010). This data is crucial to this project as it informs on the market value of apartments in all counties across the specified time frame. The data alone for a studio apartment ranges from $319 to $2447 and from $565 to $5225 for 4-bedroom apartments.

The secondary datasets are county-level datasets provided by the USDA which cover each county’s unemployment, median household income, educational attainment, poverty levels, urban-influence, and population estimates from 2010-2019. This information is across 4 datasets with a total of 13,042 observations (around 3,200 each). The combination of these data sets is important to this project as they provide an array of socio-economic factors that could be correlated with rent prices. These, along with other factors not readily available in a dataset may have undiscovered correlation with rent price that may help in trying to keep them under control within given areas.

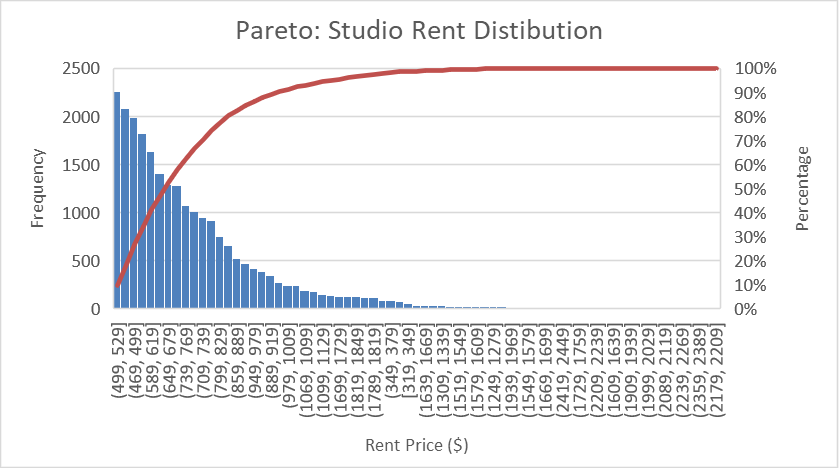
To combine these datasets, each contains a FIPS code assigned by the government that identifies each specific county. Using these, I should be able to use join functions to reformat the data of interest into one dataset containing all of the socio-economic factors, as well as the rent prices for a particular sized apartment. The HUD data has multiple repetitions of each county since this covers different years and thus there may be some issues with repeating the data multiple times and ensuring that the correct data for the correct years align. Once all data is combined with in one, master dataset, it will be divided using a 70/30 split into training and test data with which to create a model.

### High-Level Data Diagram

### 

Figure

The above figure is a scatterplot showing the distribution of average rent (2 bedroom) and population. It shows that the vast majority of rents are below $2500 with a few outliers above the $3000 mark. Looking at the counties with population as an outlier, it appears they are towards the middle of the spread but really they are towards the higher end for rent prices. While it is not a very strong linear correlation, there is a positive trend between these two variables.



Figure

The above diagram is a Pareto showing the general distribution of studio apartment rents across all data. While a vast majority of these appear to be “affordable”, the distribution has a strong rightward skew for these many outliers with extremely high rent.

### Data Definition/Data Profile

|  |  |  |  |
| --- | --- | --- | --- |
| Rent Data |  |  |  |
| Variable | Definition | Data Type | Quality Issues |
| Fips2010 | Government given county code | Integer | Most end in 99999 |
| Year | Year of data collection | Integer | NA |
| Rent50\_0 | 50th pct studio price | Integer | NA |
| Rent50\_1 | 50th pct 1 bed price | Integer | NA |
| Rent50\_2 | 50th pct 2 bed price | Integer | NA |
| Rent50\_3 | 50th pct 3 bed price | Integer | NA |
| Rent50\_4 | 50th pct 4 bed price | Integer | NA |
| State | State/Area # | Integer | NA |
| Areaname | Name of Area | Character | Some duplicates |
| County | County # | Integer | Inconsistent format |
| Countyname | Name of county | Character | Some duplicates |
| Pop2010 | 2010 Population est. | Integer | Some listed as 0 |
| State\_alpha | Abbr of state | Character | NA |

The pertinent data variables in the HUD Rent Data are listed above. The target variable(s) for this project are any of the Rent variables, either for a studio, 1-bedroom, 2-bedroom, 3-bedroom, or bedroom. For consistency moving forward, the 2-bedroom rent data will be selected as the primary target variable. Several other of these variables are the location variables- state, areaname, county, and state alpha. Most importantly, this data has each county’s FIPS number which is consistent with the county FIPS number in the rest of the datasets. The only issue is that the FIPS numbers in this dataset are correct for the first 5 digits but then all end in 99999 which will need to be truncated to make it consistent with the other datasets.

The other datasets described below are the extra USDA datasets that will be used through a bit of cut and pasting to bring together the socioeconomic variables of interest within each county. The first brings in the factor variables of median household income, unemployment, urban influence, and civilian labor force. The second set bring in poverty levels. The third bring in education level measurables such as percentage of population with a high school diploma and percentage of population with a bachelor’s degree. All of these are expected to have some correlation with the target variable of rent.

|  |  |  |  |
| --- | --- | --- | --- |
| **MHI/Unemployment Data** |  |  |  |
| **Variable** | **Description** | **Data Type** | **Issues** |
| FIPS\_Code | State-county FIPS code | Integer |  |
| State | State abbreviation | Char |  |
| Area\_name | State or county name | Char | Includes diff. scale for entire state |
| Rural\_urban\_continuum\_code\_2013 | Rural-urban Continuum Code, 2013 | Integer |  |
| Urban\_influence\_code\_2013 | Urban Influence Code, 2013 | Integer |  |
| Metro\_2013 | Metro nonmetro dummy 0=Nonmetro 1=Metro (Based on 2013 OMB Metropolitan Area delineation) | Integer |  |
| Civilian\_labor\_force\_2017 | Civilian labor force annual average, 2017 | Integer |  |
| Employed\_2017 | Number employed annual average, 2017 | Integer |  |
| Unemployed\_2017 | Number unemployed annual average, 2017 | Integer |  |
| Unemployment\_rate\_2017 | Unemployment rate, 2017 | Integer |  |
| Civilian\_labor\_force\_2018 | Civilian labor force annual average, 2018 | Integer |  |
| Employed\_2018 | Number employed annual average, 2018 | Integer |  |
| Unemployed\_2018 | Number unemployed annual average, 2018 | Integer |  |
| Unemployment\_rate\_2018 | Unemployment rate, 2018 | Integer |  |
| Civilian\_labor\_force\_2019 | Civilian labor force annual average, 2019 | Integer |  |
| Employed\_2019 | Number employed annual average, 2019 | Integer |  |
| Unemployed\_2019 | Number unemployed annual average, 2019 | Integer |  |
| Unemployment\_rate\_2019 | Unemployment rate, 2019 | Integer |  |
| Civilian\_labor\_force\_2020 | Civilian labor force annual average, 2020 | Integer |  |
| Employed\_2020 | Number employed annual average, 2020 | Integer |  |
| Unemployed\_2020 | Number unemployed annual average, 2020 | Integer |  |
| Unemployment\_rate\_2020 | Unemployment rate, 2020 | Integer |  |
| Median\_Household\_Income\_2019 | Estimate of median household Income, 2019 | Integer |  |
| Med\_HH\_Income\_Percent\_of\_State\_Total\_2019 | County household median income as a percent of the State total median household income, 2019 | Integer |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Poverty Data** |  |  |  |
| **Variable** | **Description** | **Type** | **Issues** |
| FIPStxt | State-county FIPS code | Int |  |
| State | State abbreviation | Char | Some lines for whole state and not just county |
| Area\_name | Area name | Char |  |
| Rural-urban\_Continuum\_Code\_2003 | Rural-urban Continuum Code, 2003 | Int |  |
| Urban\_Influence\_Code\_2003 | Urban Influence Code, 2003 | Int |  |
| Rural-urban\_Continuum\_Code\_2013 | Rural-urban Continuum Code, 2013 | Int |  |
| Urban\_Influence\_Code\_2013 | Urban Influence Code, 2013 | Int |  |
| POVALL\_2019 | Estimate of people of all ages in poverty 2019 | Int |  |
| PCTPOVALL\_2019 | Estimated percent of people of all ages in poverty 2019 |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Education Data** |  |  |  |
| Variable | Description | Type | Issues |
| FIPS Code | State-county FIPS code | Integer |  |
| State | State abbreviation | Char |  |
| Area Name | Area name | Char | Some listed areas are entire states |
| 2013 Urban | Urban Influence Code, 2013 | Integer |  |
| High school diploma only, 2015-19 |  | Integer | Some missing data |
| Some college or associate's degree, 2015-19 |  | Integer | Some missing data |
| Bachelor's degree or higher, 2015-19 |  | Integer | Some missing data |
| Percent of adults with less than a high school diploma, 2015-19 |  | Integer | Some missing data |
| Percent of adults with a high school diploma only, 2015-19 |  | Integer | Some missing data |
| Percent of adults completing some college or associate's degree, 2015-19 |  | Integer | Some missing data |
| Percent of adults with a bachelor's degree or higher, 2015-19 |  | Integer | Some missing data |

# Data Preparation/Cleansing/Transformation

### Data Preparation

The two major steps in preparing the data will be to identify the key variables that are desired from each dataset, including dealing with redundancies, and then unifying the variable for joining the datasets all together. Looking across all 4 datasets, the most important factor to maintain amongst all 4 is the FIPS code as this is how they will all be joined together. For the rent data, all variables will be maintained as listed in the Rent Data Chart above. For the unemployment data, all years prior to 2017 will be removed using Excel as the aligned time period for the rent data only reaches back to 2017. The initial set had data all the way through the year 2000. Also, some redundant variables can be removed since the rent data already contains variables of state name and county name which are steadily aligned with FIPS code.

For the poverty data, the desired variables to keep are those listed in the Poverty Data Chart above. Since the unemployment data already contains urban influence code and rural urban continuum, these variables will be removed along with the state and area name variables that are already listed within the rent data. Like the unemployment data, the education data also had variable measures dating way before the period of interest. There were 8 different measurements of education for each decade from 1970 to 2019, and so the data for the 70s, 80s, 90s, and 00s were all omitted to keep consistent with the time period. While this did remove 32 variables from this dataset, it still leaves the data with 10 education-specific measures for each area.

The last step in preparing the data was ensuring that the join variable of FIPS code was uniform across all 4 datasets. Looking at the rent data, the FIPS code was listed as either an 8 or 9 digit number, with most codes ending in “99999”. Comparing this to the same variable in the other datasets, FIPS codes were either a 4- or 5-digit number. Using Excel, the value was truncated for the rent data to a 4- or 5- digit number dependent on whether it was initially 8 or 9 digits. Then the only issue to address was that a join operation could potentially join non-matching line with a 4-digit number to a line with a 5-digit number if the first 4 digits match (Bruni et al., 2011). To address this, all FIPS code in all 4 datasets were reformatted into 5 digit numbers, with a leading 0 for the 4 digit numbers.

### Data Cleansing To begin cleansing the data, the first step taken was to identify missing data in any of the 4 datasets prior to joining. Since the data sources were the USDA and HUD, the data was for the most part very clean and complete. Using R to count missing observations per variable showed that the Rent data was only missing 4 observations for population count. Further examination showed that these observations came from 2 counties within Maine for the years of 2017 and 2018. Due to the small number of missing variables for this set, these values were imputed. For the unemployment data, R showed no missing values whatsoever and so this dataset was left alone for the time being.

Looking at the education dataset, 6 lines of data were missing variables across 4 different variables. Further inspection of this data showed that these specific missing values were from Puerto Rico. For the purposes of this analysis, these lines were omitted from further analysis. For the poverty dataset, the variables that measures 2019 poverty estimates for children aged 0 to 4 was missing thousands of values. For the purposes of this analysis, rather than omit those thousands of observations, it was decided to just throw out those specific variables as there must be inconsistencies in its reporting across all counties.

After checking that the 4 original datasets all reported 0 missing values now, each was joined individually onto the rent dataset to create 1 master dataset. Now checking this for missing values showed many variables reporting up to 435 missing values as a result of discrepancies with the join function. Once again, inspection showed that this was a result of some of the data including Puerto Rico and some of them not. For the purposes of continuing analysis, it was decided to eliminate Puerto Rico from the scope of this project by omitting those lines with missing data as they were missing a significant number of variables. The final Master dataset after cleansing includes 23,376 observations of 67 variables with no missing data.

### Data Transformation

The first step in transforming the data was to determine if any variables required normalization due to significant skew. A quick visualization of the target variable, median rent for a two-bedroom apartment, was made to check the skew, as shown in Figure 3. This boxplot shows a large spread and heavy rightward skew. While this may hint to a need for normalization, the decision was made to leave the data as is, since identifying the unique factors leading to these high outliers may be the most significant takeaways from this analysis.

Diagram, box and whisker chart

Description automatically generated

Figure

Combining the variables of two-bedroom rent, median household income, and the concept that household income should be greater than 3 times someone’s rent, the variable of affordability was created. This was done in Excel by dividing the MHI variable by 12 (to represent monthly income) and subtracting 3 times the median rent for a two-bedroom apartment per county. The affordability values range from -$3333 to $7412 which in context would represent the level of burden the average person could expect per county due to their rent. Figure 4 shows a quick glimpse of the spread of this variable, and it appears that by combining the 2 variables, the result is a much more normal distribution.

Chart, histogram

Description automatically generated

Figure

To further this concept, a Binary variable of affordability was created to allow for the possibility of logistic regression. This was done using an if statement in Excel which would output NO if the county’s affordability value was negative and a YES if the county’s affordability value was positive. At initial glance, this value is imbalanced and will require some adjusting as only 160 observations were deemed unaffordable compared to 23,315 deemed affordable and so perhaps the cutoff will be shifted slight above 0. Or perhaps it could be expanded from binary and instead listed as unaffordable, semi-affordable, affordable, and very affordable. Taking a quick glimpse to see what states contain the highest quantity of unaffordable counties, the R table shows 91 of the observations were from Massachusetts, followed by 14 in California, 13 in Virginia, 12 in Alaska, and 11 in New York.

### Data Analysis

For full analysis of the data, both R a Tableau will be used. R will be used to create simplistic visualizations such as 1 variable boxplots or histograms such as the ones above. Tableau will be the key in creating visualizations that tell a fuller story. Tableau’s ability to map geographic data will be a key in an analysis such as this so that a map of the country can be shown with each county’s rent. This visualization can be furthered by recreating it for MHI as well as the affordability statistics. It is expected that these statistics on a map will really highlight the issues of affordability in certain metropolitan areas, but perhaps uncovering affordability or high rent issues in other parts of the country may give further insight into the socio-economic factors that most affect these variables.

Beyond Tableau’s mapping ability, it will also provide easier and more user-friendly visualizations for picking and choosing specific variables to compare with the rent disparities. While visualizations of every single county will be very cluttered due to the high data count, Tableau can be used to group the data by state and then compare each state’s rent, unemployment, education levels, poverty levels, etc. to further discussions of how these socio-economic factors are affecting rents. This will likely be done with some type of stacked or side-by-side bar graphs.

For completing a linear regression to model the rent price based on the other socio-economic factors, R will be used and specifically the arules package within R along with ggplot and dplyr. R is an appropriate tool for this type of modelling as it is easily accessible, free, able to handle data this size, able to join multiple datasets using a key, and able to output very informative statistics regarding the regression and goodness of fit. The first model will be a multiple regressions model so that relationships between dependent variables will also be considered throughout the model. Initially, the model may be made using just a few of the available independent variables to try to zero in on some specific relationships but a model using all available data will also be made to identify any key variables. Also, a similar linear regression model will be run of the calculated variable of affordability as this factors in both rent price and income into the dependent variables.

Lastly, a logistic regression will also be run in R, specifically using the e1071 package. This will be run on the binary variable created based on whether a county yielded a positive or negative affordability score. There is concern over the low quantity of observations with negative affordability and thus the cutoff may be teased up from 0. The resulting model should give insight as to which specific socio-economic factors are most predictive in determining whether a city will be “affordable” based on the three times rent tule. For this model, rent prices, affordability, and MHI can not be factor variables since they are already calculated into the binary affordability. A naïve bayes will also be run on the binary variable as a control.

# Data Visualization

### Data Visualization 1- Heat Map of Target VariableMap Description automatically generated

Figure a

Figure 5a shows the heat map of median 2 bedroom rent per county in the United States. Since the dataset covers 3 years’ worth of data per county, this shows the average median rent price over those 3 years. The values of rent range from $593 per month to $3,424 per month. It is evident that the “red” areas seem to correlate with major cities as San Francisco, New York, D.C., Chicago, Miami, Seattle, etc. are all highlighted in red. Investigating some of more of the “anomalies” of red in the center of the country that might not immediately strike as a city, Teton County in Wyoming contains Jackson, Coconino County in Arizona contains Flagstaff, and Summit County in Utah does not quite contain Salt Lake City but contains major suburbs. The two most interesting anomalies are Dunn County, ND and San Miguel/Ouray County, CO as all show above average rent but are not necessarily linked to any major metropolitan areas.

Map

Description automatically generatedMap

Description automatically generated The highest rent values exhibited are in Marin and San Mateo County, CA. These counties surround San Francisco, and both average a median 2 bedroom rent of $3,424. The counties surrounding this area- Contra Costa, Alameda, Santa Cruz, and Santa Clara- all also have rents above $2,467. These are the 5 most expensive counties in the country. Of the other major cities, Boston and its surrounding counties range in rent from $1900 to $2100, New York and its surrounding counties range in rent prices from $1800 a month to $2,100, Seattle is roughly $2000, D.C. ranges from $1450 to $1850, Miami and Denver are roughly $1500, and Chicago is roughly $1300. The cheapest counties in America in terms of rent are all located within Alabama. There are 9 counties in the state that each averaged a median 2 bedroom rent of $620 over the course of 2017-2020.

Figure 5b.

Figure 5b shows the same heat map expanded to show Alaska and Hawaii. Both states show predominantly above average rent prices, with Hawaii being entirely red. The island of Honolulu County has the highest rent of $2,195 while the rest of the islands range from $1,562 to $1,823. In Alaska, the highest average rent is $1,816 in the Aleutians West Census Area, which is the strip of islands in the SW of the state. The counties containing Juneau and Anchorage have rent of $1,539 and $1,375 respectively.

Insights provided by this graph heavily emphasize the importance of urban influence on rent prices. It is evident from Figure 5a that the highest concentrations of above average rents occur in or around the major U.S. metropolitan areas. Moving forward, the factor variable of rural urban continuum code will be looked at more in depth since this is clearly a major factor in rent prices. This may also be expanded to the variable of population. While most major cities would be expected to have a greater supply of rentable housing, it is quite possible that the concentration of the population in major cities is a reason for the higher rents.

While the visualization’s main insight is focused on urban influence, it also showcased that a few anomalies do exist across the United States that cannot be explained by metropolitan influence. Specifically, Dunn County, ND and most of the counties in Alaska. Since these observations are mixed in with 20,000 others, they may not carry much weight in the overall output of the model but may be worth further exploration to determine what are the driving factors of their rent prices.

Due to the evident relationship between urban influence and rent, some of the scope of this project may be shifted to observing how the other factor variables of unemployment, poverty, education, etc. correlate with rural urban continuum codes to ensure that any underlying relationships between factor variables is also uncovered and discussed in the final analysis. By running a multiple linear regression model, these relationships should be calculated, but rather than just seeing each factor variables relationship with rent, it will be important to see their relationships with other factor variables as well.

### Data Visualization 2- Heat Maps of Factor Variables

Map

Description automatically generated

Figure

Figure 6 shows the first of several heat map visualizations of the main factor variables. Figure 6 specifically focuses on Median Household Income (MHI). With the focus of this analysis being the affordability of the various counties, MHI is tied together with rent in that one would hope the areas where rent is more expensive, that MHI is also high to reflect the cost of living. Comparing Figure 6 with Figure 5a, this is not the case in many cities. Once again, many major metropolitan areas- particularly San Francisco, Washington, D.C., and New York. In contrast though, many of the cities like Miami, Chicago, and Portland do not have above average MHI to reflect their above average rents which will reflect poorly in their affordability. In the reverse, some highlighted areas that have high MHI that did not show up as having high rent are counties around Nashville, Indianapolis, and Los Alamos, meaning places like this should in theory be very affordable.

Figure 7 below shows a heat map of unemployment data. This is specifically showing the unemployment rates as a percentage and thus population size should be considered when looking into each of the highlighted counties. The range for unemployment rates goes from 1.7% to 22.5% which is a significant spread. The highest unemployment rate appears in Imperial County, CA, followed by Kusilvak Census Area, AK (not included in Figure 6) with 19.40% and then Jefferson County, MS with 18.40%. The map seems to show a trend of the Map

Description automatically generatedlowest unemployment rates (darkest blue coloring) being concentrated in the middle of the country, around Nebraska and Kansas. The major U.S. metropolitan areas seem to not be associated with either extreme measure for the unemployment data.

Figure

Map

Description automatically generated

Figure

Figure 8 focuses on the poverty distribution across the United States. Once again, there does not seem to be a big influence from metropolitan areas. The range of poverty rates go from 2.7% to 47.7%. Surprisingly, the highest levels of poverty occur in the Dakotas, with Ziebach County reporting 47.7% in poverty and Todd County reporting 43.40% in poverty. Further Research shows that poverty has been a persistent issue in the area and is linked to the American Indian Reservations within these counties (Rapid City Journal, 2012). There is also a concentration of high poverty counties along the southern part of the Mississippi river, amongst Louisiana, Arkansas, and Mississippi all reporting between 30% and 40% impoverished.

Map

Description automatically generated

Figure

### Figure 9 shows the spread of college educated persons in the U.S., showing the percent of the population with a bachelor’s degree or higher as reported from 2015-2019. The percentages range from 0% to 75.3%. Only 3 counties report lower than 5%- Loving County, Texas (0%), Kennedy County, Texas (1%), and Issaquena County, MS (3.2%). The highest reported percentage of population with a bachelor’s degree or higher occurs in Arlington County, VA (75.3%), followed by Los Alamos, NM (67.4%), Howard County, MD (62.6%), and Boulder County, CO (62.1%). Once again, this visualization showcases an immense spread across the United States. It also seems to most align with the heat map for rent and MHI with a lot of the concentrations of red aligning with major cities.

### The combination of these visualizations of various factor variables has provided insight as to which variables might have underlying connections when the results of modelling are reported. It is expected that rent model will report higher expected values in metropolitan areas and these visualizations show that it should also be expected that college-education should also have a high correlation coefficient since it seemingly goes hand in hand with rural urban continuum code. Similar can be said for the variable of Median Household Income. It was initially expected that poverty and unemployment would be heavily linked to urban areas, but these visualizations show that to not be the case.

Adjusting the scope of this project to reflect these discoveries puts more emphasis on how all of the variables tie into population and urbanization. There is also clearly underlying demographic ties since demographics are a main reported reason for high levels of poverty in certain counties. Moving forward, as rent pricing and affordability remain the target variables, extra emphasis will be placed on the factor variables associated with metropolitan areas to ensure the model is not overfit to all of these variables that are already very connected.

### Data Visualization 3

A picture containing chart

Description automatically generated

Figure

# Figure 10 shows multiple bar graphs relating the rural urban continuum codes with several of the factor variables while still showcasing average rent (size of the bar) and population (shading of the bar). Trends seem to arise in respect to all 5 variables paired with these- MHI, Unemployment Rates (2019), Poverty Rates (2019), percent of population with a bachelor’s degree or higher, and the calculated field of affordability. Within each individual bar graph, trends can be seen amongst all variables with the rural urban continuum code. In all graphs, it is emphasized how population is directly related as the shading lightens as the code number increases, showing that the average population for a county with RUUC 1 is 236,746 compared to 5,696 in counties with an RUUC of 9. Also from the widths of the bars, it is clear that rent decreases as RUUC increases.

Focusing on the bar graph for MHI, there is an obvious downward trend with a major drop in MHI between RUUC 1 and 2, decreasing by $13,800. Similarly, between RUUC 2 and 3 there is a drop of about $9000. The shape then steadies out as the MHI for the remaining RUUC’s are all within $50,000-$58,000 with the exception of RUUC 7 which has the lowest MHI at $49,748. Focusing on the graph for unemployment rate, the trend is slightly positive as RUUC increases. However, it should be noted that in general, the average unemployment for all RUUC classifications are between 3.2% and 4%. Considering how the heat map showed some counties experiencing 22.5% unemployment, this suggests that there for the most part, unemployment is independent of RUUC, but is still at its lowest in counties with an RUUC of 1.

The graph for poverty percentages showcases a clear positive trend ranging from 9.4% to 15.5% and showcasing that poverty levels tend to be lower in more urban areas. The trend shown in percentage of college educated adults shows a steep decline between RUUC 1 and 3, then increases slightly for RUUC 4 and 5, then drops roughly 13% for RUUC 6-9. This is a strong enough trend to suggest that major cities and suburbs tend to have a much greater college educated population, however after a certain point (RUUC 6 and beyond) the urban influence loses its correlation with education since these 4 RUUC classifications all seem to have comparable percentages. Lastly, for the affordability graph, once again there is a negative trend between RUUC and affordability. The average “affordability” for RUUC 1 counties is $2,458, meaning that after removing 3 times the average rent from the average monthly income, there is $2,458 leftover. This suggests that even though RUUC 1 counties have the highest rent, their MHI is proportionately large enough to maintain a higher average than higher RUUC counties. A downward trend continues between RUUC’s 1-3 but then average affordability fluctuates between RUUC 4-9, bouncing around between ~$1700 and $2000.

The affordability graph shows that on average, despite having significantly higher average housing costs, renters in larger cities should still come out of each month with more money than those living in counties with higher RUUC codes since the MHI is high enough in major metropolitan areas to cover living costs and then some. However, for the purpose of the logistic regression model, the binary affordability (positive/negative) was imbalanced and so a variable of adjusted affordability was created, which scaled the affordability number down by $1000. In essence this measures MHI -3\*rent-1000, estimating that after paying rent and the other estimated living expenses (3\*rent), who should expect to still have $1000 leftover after each month.

### Proposed Visualizations

The affordability graph shows that on average, despite having significantly higher average housing costs, renters in larger cities should still come out of each month with more money than those living in counties with higher RUUC codes since the MHI is high enough in major metropolitan areas to cover living costs and then some. However, for the purpose of the logistic regression model, the binary affordability (positive/negative) was imbalanced and so a variable of adjusted affordability was created, which scaled the affordability number down by $1000. In essence this measures MHI -3\*rent-1000, estimating that after paying rent and the other estimated living expenses (3\*rent), who should expect to still have $1000 leftover after each month. Using this newly adjusted measure, a proposed visual would be along the lines of Figure 11. It shows a histogram broken down into “rent brackets”, then broken down again by RUUC. It shows roughly what might be considered “monthly gains” for people living in counties by rent price. So, someone living in a county where they are being charges $3,700 a month in an RUUC 1 county for a 2 bedroom apartment could expect to lose $1,847 each month. Someone paying $3,100 a month could expect to gain $626 a month if they are in an RUUC 1 county but expect to lose $3,222 a month if they are in a RUUC 2 county due to the differences in expected MHI. This is proposed as an interesting way to relate what are emerging as 3 of the most significant variables within this analysis. This would be very valuable to prospective renters looking to move as they could find their rent price within the chart and see how to maximize their monthly profit by which level of RUUC to look into since this maximizes their potential earnings. While there are several other factors at play with this, such as that renters career, apartment size, amenities, etc., this seems to be an interesting way to tie together these variables into a budgeting tool.

Chart, bar chart

Description automatically generated

Figure

A final proposed visualization would require the data to be parsed to only include major metropolitan areas and create a stacked bar graph comparing all of their major factor variables. This could likely be done by just parsing the data to include RUUC 1 counties and/or setting a population threshold. The issue is doing this is that many metropolitan areas are spread over multiple counties and thus aggregating these observations into one datapoint for the city would be valuable. The idea behind this visualization is that rent prices tend to be the biggest “problem” within metro areas. Visualizing the metropolitan areas for all their factor variables may start showing linkages between cities. For example, maybe New York City and Chicago are very comparable based on most socio-economic factors and D.C. and San Francisco can be linked as well. Then looking at their rental prices, what underlying factor causes San Francisco to have such higher rent despite higher MHI in the D.C. area, and why is New York City more expensive than Chicago? This idea could be expanded to a dashboard that lets someone compare 2 cities side by side based on all their data.

# Predictive Models

### Predictive Model 1

|  |
| --- |
| Call:  lm(formula = myFormula, data = train.Rent)  Residuals:  Min 1Q Median 3Q Max  -888.46 -86.12 -7.99 64.47 1850.90  Coefficients: (3 not defined because of singularities)  Estimate Std. Error t value Pr(>|t|)  (Intercept) 1.315e+04 2.371e+03 5.543 3.01e-08 \*\*\*  pop2010 1.092e-04 5.138e-05 2.124 0.03365 \*  hu2010 -5.823e-04 1.315e-04 -4.427 9.61e-06 \*\*\*  PCTPOVALL\_2019 -2.866e+01 4.145e+00 -6.915 4.86e-12 \*\*\*  CI90LBALLP\_2019 3.923e+01 4.639e+00 8.457 < 2e-16 \*\*\*  PCTPOV017\_2019 2.852e+01 6.016e+00 4.741 2.15e-06 \*\*\*  CI90LB017P\_2019 -4.133e+01 6.178e+00 -6.689 2.32e-11 \*\*\*  PCTPOV517\_2019 8.494e+00 5.216e+00 1.628 0.10344  CI90LB517P\_2019 -1.684e+00 5.356e+00 -0.314 0.75323  Percent.of.adults.with.less.than.a.high.school.diploma..2015.19 -1.305e+02 2.372e+01 -5.504 3.77e-08 \*\*\*  Percent.of.adults.with.a.high.school.diploma.only..2015.19 -1.337e+02 2.373e+01 -5.634 1.79e-08 \*\*\*  Percent.of.adults.completing.some.college.or.associate.s.degree..2015.19 -1.353e+02 2.372e+01 -5.705 1.18e-08 \*\*\*  Percent.of.adults.with.a.bachelor.s.degree.or.higher..2015.19 -1.262e+02 2.373e+01 -5.320 1.05e-07 \*\*\*  Rural\_urban\_continuum\_code\_2013 -1.287e+01 1.548e+00 -8.315 < 2e-16 \*\*\*  Urban\_influence\_code\_2013 9.599e+00 9.428e-01 10.181 < 2e-16 \*\*\*  Metro\_2013 4.640e+01 5.793e+00 8.008 1.24e-15 \*\*\*  Civilian\_labor\_force\_2017 1.959e-03 8.666e-04 2.260 0.02382 \*  Unemployment\_rate\_2017 -3.719e+00 3.878e+00 -0.959 0.33754  Civilian\_labor\_force\_2018 3.054e-03 1.802e-03 1.695 0.09009 .  Unemployment\_rate\_2018 2.780e+01 5.632e+00 4.936 8.05e-07 \*\*\*  Civilian\_labor\_force\_2019 -4.168e-03 1.402e-03 -2.973 0.00295 \*\*  Unemployment\_rate\_2019 -5.837e+01 3.750e+00 -15.564 < 2e-16 \*\*\*  Civilian\_labor\_force\_2020 -3.821e-04 5.020e-04 -0.761 0.44655  Unemployment\_rate\_2020 3.673e+01 1.098e+00 33.444 < 2e-16 \*\*\*  Median\_Household\_Income\_2019 1.203e-02 2.232e-04 53.891 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 170.8 on 16314 degrees of freedom  Multiple R-squared: 0.7536, Adjusted R-squared: 0.7533  F-statistic: 2080 on 24 and 16314 DF, p-value: < 2.2e-16 |

Figure - Mode l1 Output

The results of the first created linear regression model are shown in Figure 10. The model was created using the lm command in R with the median rent price of a two bedroom apartment as the target variable. A few minor adjustments were made to the overall data to create this specific model. It was decided that some variables could be omitted out of redundancy, for example, total unemployment count and unemployment rates within a county may have been redundant and so just the rate was used as it is adjusted for the population size already. The data used for this specific model contained 24 variables and was created on a training data set of 16,339 observations.

The model reports a residual standard error of $170.8, which is not too bad based on the nature of the rent variable- meaning on average, the model should be accurate within $170 on the rent estimate. However, it is shown that the residuals range from -$888.46 to $1850.90, meaning that some predictions are very far off in either direction. The model also reports a multiple R-squared statistic of .7536 which is also moderately good in justifying that 75% of the variations in the rent price are accounted for by the listed independent variables. Looking at the p-values, 23 of the 24 variables reported as statistically significant, with two tailed p-values below .05. Further, 18 of the variables reported as significant at a significance level of .01 and 13 reported significance at a level of essentially 0. This suggests, for most variables, an acceptance of the assumed alternative hypothesis that there does exist some form of linear relationship between the specified independent variable and the dependent variable of rent.

Comparing the coefficient estimates for each variable, the model begins with an estimate y-intercept of $13,150 and then uses a majority of the variables to decrease the rent value from there. The most influential variables in terms of rate of change seem to be the 4 measures of education levels within the county, reporting rates of change between -126.2 and -135.3. This signifies that for each percentage point increase in these measures, the rent ideally would be expected to go down by that much. There is an inherent problem with this mentality since increasing the percentage of the population with a high school diploma would then increase the percentage with less than a high school diploma, but it is worth noting that the greatest rates of change are associated with percentage with a high school diploma and percentage completing some college. Other variables worth noting are unemployment rate for 2019 which reported an estimated coefficient of -58.37, the metropolitan binary variable with an estimated coefficient of positive 46.4, and the Median Household Income with an estimated coefficient of .012 (which is very influential on final estimated rent price due to the nature of the variable.

Chart, scatter chart

Description automatically generated

Figure -Model 1 Residual Plot

Figure 13 above shows the residual plot created when running the model of the test set of data, which contained 7,036 observations. It shows most of the data on the left side clustered around the line, meaning predicted values were fairly close to actual. Then as the eye moves further right, the graph shows a divergence from the model due to the outliers which were actively chosen to be kept. This was one of the purposes of the model- to set a baseline of what rent should be based on socio-economics of an area and then showcase how high or below market value an area is. The data points within the $3000-$3500 range showcase the counties where rent is very overpriced based on the socioeconomics of the area and that singular datapoint towards the top-center of the graph shows a county where rent is actually much cheaper than its socioeconomics would suggest.

Despite having a stronger R-squared than expected, this model changes the scope of the project by showcasing that a majority of the variables showcase a negative relationship with rent prices. It is very interesting that the model suggests a y-intercept of $13,150 and the uses most of the variables to scale back the rent price from there. The variables that tended to have a positive relationship with rent were urban influence code, metropolitan binary, population, and median household income. Interestingly, some years the unemployment statistics showed a positive relationship and some years it reported a negative relationship. Overall, this model did seem to put focus onto the strength of the relationship the education variables have with rent prices.

### Text Description automatically generatedPredictive Model 2

Figure - Model 2 Output

Figure 14 shows the results of Model 2. Model 2 is also a linear regression model with median two-bedroom rent price as its target variable. This model differs from Model 1 in its choice of independent variables. The variables selected for this model were all deemed significant at a significance level of essentially 0) within the first model and showcased some of the highest t-values as well. Furthermore, the variables were chosen to reduce redundancy in some of the variables, i.e., not including all years of poverty data and just trying to choose a representative variable from each original dataset to represent all of the income data or poverty data, etc. In doing so, the model lost some accuracy in terms of its R-squared statistic, reporting at .7342 and also shows a slightly larger range of residuals ranging from -856.02 to 1904.48.

All variables included in this model, except for percent of adults with high school diploma only, reported as statistically significant at an alpha level of essentially 0. It is interesting that in the first model, the high school diploma variable reported significance at this level as well, but in reducing the number of variables, it has lost its significance. The variables with the strongest relationship to rent prices in terms of t-value are median household income 2019 (t=53.493), civilian labor force 2019 (t=34.561), and unemployment rate 2020 (t=31.827). The percentage of adults with a bachelor’s degree or higher as well as population and poverty levels also reported fairly high t-values.

Analyzing the coefficients, this model reports starting with an intercept of -$227.90 for the uniformed rent price, and then gives most variables a positive linear coefficient to essentially increase the rent price as the values of these variables increase. The only variables showcasing a negative linear relationship with rent price are population and the lower bound for the poverty confidence interval. Median household income shows a coefficient of .011 meaning roughly 1% of the median household income for a county gets added into the rent estimate. It is also interesting that the urban influence code showed a positive relationship as this would adjust rent estimates to be higher for less urban counties. The highest coefficient reported is 24.75 associated with the county’s unemployment rate.

Chart, scatter chart

Description automatically generated

Figure - Model 2 Output

Figure 15 shows the residual plot for Model 2. Compared to Model 1’s residual plot, the spread of high outliers, where the model predicts a rent much higher than observed, appears lower and in general seems to predict the rent values to be lower than model 1 as it seems the scatter is shifted a bit lower. This makes sense with the nature of this model, beginning with a negative intercept and then increasing the predictive price per most variables as opposed to the intercept well above the predicted price and then decreasing per variable. Still, most of the data some seem to be clustered around the line towards the left and then begins spreading out as the observed values increase due to more differences amongst the high-rent counties.

This model aligns with the scope of the project more than Model 1 due to the reportability of the coefficients. This model reports that for each percentage increase in unemployment rates, one should expect rent prices to increase by $24.75 and for each percentage increase in poverty percentage, one should expect rent to increase by roughly $10. This is the type of output that could be easily interpreted to policy makers to show that if policies are put in place to attack poverty and unemployment, one should expect rent prices to fall as well. This would be the suggested alternative to the aforementioned rent control policies that can hurt both renters and property managers.

### *Graphical user interface, text, application Description automatically generated*Predictive Model 3

Figure - Model 3 Output

The output for Model 3 is shown in Figure 16. This is a logistic regression model with the target variable of affordability. As previously mentioned, the affordability statistic was calculated as a difference between median monthly income and 3 times the median rent price as a measure essentially of how much money the average person per county should profit at the end of each month. The calculation of the binary target variable was slightly adjusted to result in a roughly 25/75 split of unaffordable/affordable by raising the cutoff from above or below 0, to above or below 1500. With this, the predictive model will be using socioeconomic data for each county to predict if it will be above or below the 25th percentile in affordability. Also, since affordability was calculated from rent and household income, these variables were omitted from this modelling.

The results shown for Model 3 come after taking several iterations to build the minimal adequate model. Because of this method, all variables shown are significant at an alpha level of .001 and most are significant at an alpha of essentially 0. The highest z-value shown corelates with percent of state total median household income (z=23.05). While this is obviously related to standard median household income, this statistic was not used in the calculation and is helpful in representing counties where the median household income is high in states that may not contain a major metropolitan area. Other high z-values are shown for two of the poverty measures, with poverty rates reporting a z value of -18.263 and the lower bound of the poverty confidence interval reporting a z-value of 18.435.

The intercept for this model is reported as -293.3 and the coeffiecient for all of the variables are pretty evenly split between positive and negative relationships. The education statistics for percentage of the population for each education bracket report the four highest coefficients, but it should be remembered that these coefficients multiply with a number bound between 0 and 100. The largest variables by nature, population and civilian labor force, both are adjusted accordingly as their coefficients are and respectively, almost showing a relationship where these two intertwined variables cancel each other out.

|  |  |
| --- | --- |
| Training Data | Test Data |
| NO YES  0 2569 829  1 1837 11104  Accuracy: .8368 | NO YES  0 1023 352  1 824 4837  Accuracy: .8328 |

Figure 7- Model 3 Confusion Matrices

Figure 17 shows the confusion matrices produced by the logistic regression model. The training data reported an accuracy of 83.68% and the test data reported an accuracy of 83.28%. The similarity between these two suggests that the model was not overfit to the training data. It is evident in both confusion matrices that the model is much more accurate when dealing with “affordable counties” as the number of false positives is much greater than the number of false negatives. In this specific case, there is not much consequence in making an error one way or the other so the sensitivity level for this model is fine.

This model enhances the scope of the project as it has created a model, without using the variables of rent or median household income, that can intake the socioeconomics of a given county and output fairly accurately whether the county will be affordable or not. Since this affordability statistic is a measure of the relationship between income and rent, this model is really showing a relationship between all variables with these two very important variables in the scope of this project. Seeing that variables such as poverty and urban influence have such a strong association with affordability could give policy makers ideas as to what their policies’ should focus on to help the community as a whole.

**Predictive Model 4**

Text

Description automatically generated

Figure - Model 4 Partial Output

A fourth model was created for comparison purposes with the logistic regression model, this time using a naïve bayes modelling approach. This technique assumes each independent variable to be independent of each other, which in this case may be add a lot of insight as so many of the variables in this dataset have overlap, this will help separate out each one’s relationship with rent. This method also required that all variables be factor variables and thus all variables were discretized into 3 bins. The output shows a matrix for each variable individually with the probability of being affordable or unaffordable dependent upon the bin that the county falls into. The highest probabilities shown are all associated with poverty data, with 70.1% of “unaffordable” counties falling within the bin of 14.3-47.7% impoverished or 70.3% of “unaffordable” counties having an upper bound for the poverty C.I. between 17.1% and 57%. Similarly, there is some association with the education data as well as 54.49% of the “unaffordable” counties are where 12.2-73.6% of the population has less than a high school diploma. Conversely, 48.8% of the “unaffordable” counties are counties where less than 18.8% of the population has a bachelor’s degree or higher. Also, 55% of “unaffordable counties” are labelled with an urban influence code between 6 and 12 which are the ratings given to the counties with the lease urban influence- implying affordability is a problem in places other than metropolitan areas as well.

|  |  |
| --- | --- |
| Training Data | Test Data |
| NO YES  NO 3123 2528  YES 1283 9405  Accuracy: .7667 | NO YES  NO 1291 1126  YES 556 4063  Accuracy: .7609 |

Figure 9- Model 4 Confusion Matrices

Figure 19 shows the confusion matrices for the Naïve Bayes model created. The training data shows an accuracy of 76.67% and the test data shows an accuracy of 76.09% so there are no signs of overfitting. While there is a drop off in accuracy from the previous model, this model is stronger in correctly identifying unaffordable counties. However, as a tradeoff, it is less accurate in identifying affordable counties along with the drop in overall accuracy.

This model furthers the scope of this project mainly by providing the breakdown shown in Figure 18. This gives an in-depth view at each individual independent variable’s relationship with affordability which provided insight on how overall education and poverty can relate to affordability, which in turn shows how they relate to rent and median household income. The shown probabilities could certainly be analyzed further by looking more in depth at other variables relationships with affordability and also by changing the bin size for discretization. Reporting this to local governments could help them realize ideal levels for the various socioeconomic factors to help make their communities affordable. For example, while 0% unemployment is obviously the goal, this shows that getting it between 3 and 4% can essentially half the chances of being an unaffordable community- from 58.17% to 16.53%. Policy makers could look through these probabilities for each variable to set attainable goals (decrease unemployment and poverty by x% and increase high school graduation rates by y%) that should in turn create more affordable communities and grow wealth within a county.

### Predictive Model Review

Linear-

Based on R-squared values, Model 1 appears to be the superior model. It reports that 75.36% of the variation in median two-bedroom rent prices can be attributed to variations in the included variables while Model 2 reports slightly less at 73.42%. However, for only a difference of roughly 2% in R-squared, the simplicity that the reduction of variables between models 1 and 2 provides does help provide some more clarity in the results. Some of the variables in Model 1 could be considered redundant since they are essentially measuring the same types of socioeconomics. The residual plots are very similar as well.

There is an interesting comparison to be made based on the shape of the regression equations. Model 1 starts its prediction with an intercept of 13,150 and then uses most variable to decrease the predicted value from there. Model 2 starts its prediction with an intercept of -227.90 and then uses most of the variables to increase the prediction value from there.

Chart, scatter chart

Description automatically generated**Chart, scatter chart

Description automatically generated**

Figure .2- Model 2

Figure 0.1- Model 1

# 

# Figure 20 above shows the Residual vs. Fitted charts for both models. Both help to emphasize the existence of a linear relationship due to having essentially flat 0 residual line with the points being randomly distributed about the line. The big takeaway from these is on the x-axis though which emphasizes that Model 2 tends to predict lower rent prices from the data than Model 1.

Due to the scope of the project being to predict accurate rent prices, be able to pick out specific socioeconomics and explain how they affect rent prices, Model 2 is the recommended champion model. Despite having the slightly lower R-squared value, being able to describe most of the variables with a positive coefficient is important in describing the relationships to policy makers. It is more intuitive to describe that as poverty increases, so too does rent and thus policy makers would want to decrease both, rather than say that poverty increasing causes rent to decrease and possibly suggesting a benefit to increasing poverty. Model 2 provides more intuitive relationships for many of the variables, while also just focusing on the main quantifiers of socioeconomics, making the results much more conveyable. For these reasons, Model 2 will be the recommended linear model.

Logistic-

Comparing the models created for the binary variables- the logistic and naïve bayes- the stronger model in terms of classification accuracy is the logistic regression model, Model 3. It attained an accuracy level of 83.28% on the testing data compared to only 76.09% on the test data for Model 4. Neither model showed signs of overfitting or any other significant issues. Model 4 seemed stronger at correctly identifying unaffordable counties. Out of a total of 6,253 total unaffordable counties across the data, Model 4 identified 4,414 correctly across the training and test data while Model 3 only identified 3,592. This increased sensitivity did lead to a loss of accuracy, however, form the perspective of this model, it may be more important to correctly identify unaffordable counties rather than correctly label the counties in general. Knowing that a certain county has the socioeconomics associated with being unaffordable should be a red flag to policy makers while being labelled an affordable county would not inspire any action.

Model 4 also gave the very interesting insights into each variable’s independent relationships with the target variable of affordability with the output from the naïve bayes model. Having one large output that can show which percentage of the unaffordable counties fall within the different categories of the variable provides a lot of insight about how each socioeconomic factor is related to rent prices. Most notably, Model 4 revealed the strength of the relationships between education and rent as well as poverty and rent. While these relationships may not be truly independent of the other variables, this does suggest attainable policies and programs to be put into place that could indirectly decrease the rent.

Despite the lower accuracy, the champion model for the binary modelling is Model 4. Because it is far more important to identify the unaffordable counties rather than affordable ones, the sensitivity shown by Model 4 is critical in being able to enact change and make communities more affordable. Also, the output provided by the naïve bayes method gives more insight into individual relationships with affordability, as previously mentioned. The drop of a few percentages in terms of accuracy from Model 3 is outweighed by the sensitivity and insight provided by Model 4.

# Final Results

### Analysis Justification

Regarding the analysis performed towards the target variable of two-bedroom rent prices, the analysis was appropriate as the modelling technique handled well the nature of the variable and the outputs provided valuable insights into the described goals. With rent prices being a continuous variable, linear regression was the appropriate technique to be used. To further this, the exploratory visualizations showed linear relationships between most of the prominent factor variables and price, whether it was a positive or negative relationship. This justifies the usage of standard linear regression rather than other types such as exponential regression or quadratic regression.

The results further the justification as nearly all factor variables showed significance evidence to reject the null hypothesis that a linear relationship did not exist between the variable and two-bedroom rent. Not only this, but a majority showed significance at an alpha level of essentially 0. This output suggests a strong relationship exists between these variables and shows the appropriateness of the model.

The champion model of the linear regression technique reveals key insights that achieve several goals of this project. Firstly, it yielded a fairly accurate model (in terms of residual squared error) with which to set a baseline of what rent prices should be in a particular county based off of similar counties across the United States. Secondly, analyzing the significance levels and coefficients of each factor variable uncovers which of the socio-economic variables hold the most weight in determining rent prices.

Regarding the analysis performed towards the binary target variable of affordability, the techniques used were appropriate as they created models that could accurately predict whether a county would be affordable or unaffordable and were also able to “red flag” which factor variables tended to be the most influential in determining the outcome. For the logistic regression, the predictive capabilities exceeded expectations and the significance levels for each variable provided almost a ranking of which variables were most influential. For the naïve bayes regression, the accuracy also exceeded expectations and the output gave keen insights of how each factor variables broke down amongst the subgroups of affordable and unaffordable housing. This allowed a visual of how the unaffordable counties were spread amongst each variable to key in on which levels per factor variable were most in common amongst the counties.

### Findings

The linear regression models first and foremost produced a model that could be used to output a fair rent price for any county given its socioeconomic factors. Based on the residual squared error for the champion model, one should expect the output to be accurate within $177 of the actual two-bedroom rent price. This was the important outcome for the renters and property managers. Furthermore, regarding the government lens, the champion linear regression model yielded keen insight on how each factor variable related to the rent price.

Background pattern

Description automatically generated with medium confidence

Figure - Model 2 Influential Variables

Figure 22 shows a ranking of the most influential factor variables from Model 2, the champion linear regression model. The greater the absolute value of the T-value, the greater amount of evidence there is of a relationship with the target variable. This model had enough evidence to suggest a relationship with 9 out of the 10 variables, and this table shows the strongest 5. Median household income appeared as the most influential factor variable in both linear regression models and 2020 unemployment rates appeared as the second most influential in both. These top 2 influential rankings were by a large margin in both models. For Model 2, the size of the civilian labor force appeared as the third most influential while Model 1 had shown urban influence code to be the third most influential. Lastly, the percentage of the population with a bachelor’s degree and poverty rates appeared in the top 5 for both models as well. These findings relay which socioeconomic factors could be taken on to combat rent prices as an alternative to rent control.

The coefficients for each variable suggest the type of relationship with rent. Median household income has a positive relationship with rent, so both should be expected to raise together. Unemployment and poverty rates also both showed a positive relationship with rent prices. This would then suggest that lower poverty and unemployment rates would be associated with lower rent prices. These are two socio-economic factors that local policy makers could attack head on to improve their community while also potentially lowering rent prices and making their community more affordable at the same time.

The logistic regression and naïve bayes analysis both provided valuable insight as to the commonalities of the top 75%, most affordable counties in the United States and commonalities amongst the bottom 25%, least affordable counties in the United States. The output of the logistic regression was similar to that of the linear regression in showing the various t-values that could be used to rank the significance of each variable, as shown in Figure 23. Table

Description automatically generated with medium confidence

Figure - Model 3 Influential Variables

The naïve bayes model provided even deeper insight by breaking down the spread of each factor variable into 3 levels across either the affordable or unaffordable counties, as seen in Figure 18. For example, it showed that amongst unaffordable counties, 69.7% have high poverty levels (between 14.3 and and 47.7), 53.9% have high school dropout rates between 12.2 and 73.6%, and 58.3% have high unemployment (between 4.7 and 19.7%). All of these findings could be presented to policymakers concerned about unaffordability to suggest policies that reduce poverty/unemployment or promote the completion of high school education.

### Review of Success

The main goal stated for this project was to create a model that input any county’s socioeconomic factors and output a fair and accurate rent price for the county. After yielding 2 models linear models with residual standard errors between $170 and $177, this has been deemed a success. This can be interpreted as expecting the output of the model to be within $170 from the true, fair rent price for the county based on socioeconomics across the United States. This would be even more accurate had the outliers from several major cities been removed- however, then the estimates for these specific cities would have been further off.

Secondly, the business success criteria was contingent upon providing a clear picture of which factor variables hold the most weight in determining rent prices. With the outputs from the various models, it is clear that median household income, unemployment rates, poverty, and education levels tend to hold the most weight in determining the fair rent price. This information paired with the specific coefficients for each lays forth a possible plan of action for each in how it could be altered to alleviate rent-related problems within a community. If local policy makers wanted to shift spending away from subsidies, they could invest in things like unemployment (by creating jobs) or incentivizing high school completion to indirectly attack unaffordability issues.

The key performance indicators for all models were adequate and support the usage of each modelling technique for the given data. For the linear regression, the residual standard error was reasonable as previously mentioned, and also the r-squared values for both models were strong enough to support the linearity. The champion model reported an r-squared of .7342. Furthermore, the results showing that 9 out of 10 variables used in Model 2 showed significance, as well as 23 out of 24 variables for Model 1, shows further success than expected as now all variables can be further analyzed for their relationship with rent.

For the logistic regression and naïve bayes model, the accuracy in correctly classifying 83% and 76% of the test data respectively far exceeds expectations. As previously discussed, Model 4’s stronger sensitivity in identifying unaffordable counties (71% vs 57%) was a major factor in championing the naïve bayes model over the logistic, as this also exceeded expectations. The insights gained from the granular look at each factor variable’s relationship with affordability is also a major success from this project, providing ideas for possible policies to improve multiple facets of a community.

### Recommendations for Future Analysis

In considering future research, the main proposal that comes to mind is to take the granular look at each factor variable’s relationship with affordability and apply a similar naivety within the linear model. This could either be done by making many simplistic 2 variable linear regressions with each variable of interest to see which have the strongest linear relationships individually or by running a multiple linear regressions model to also uncover ties between the factor variables. This would enhance the findings of this project since it is very possible that the reason a positive or negative slope for a factor variable was reported may have been based on the slope of another factor variable. This inconsistency was seen for several variables that reported a negative slope in Model 1 but a positive slope in Model 2. Analyzing individual slopes would show a more pure relationship with rent prices and getting an r-squared value for each variable would help show how strongly tied any individual factor variable is to determining rent prices.

If the desire should arise to make this rent calculator available to the public as many home buying or property sites have built in, it would be worthwhile to create a model that factors out the high outliers associated with San Francisco and New York in the interest of creating a model with a smaller residual standard error. Taking this further, knowing the influence metropolitan areas have on rent prices, it may be worth splitting the data by urban influence code and creating two separate models- one for metropolitan areas (urban influence codes 1 and 2) and one for non-metropolitan areas (urban influence codes 3-12). This would help to isolate out the metropolitan variable that seems to spike rents in the city, making everywhere else’s rent seem low.

Another possibility with the available data would be to explore affordability by family type since the data is available for all apartment sizes of studio thru four-bedroom. It is quite possible due to crowding in the cities, zoning, and space limitations that there may be a major difference in affordability of studio, one-bedroom, and two-bedroom apartments versus the availability of larger three- or four-bedroom homes. This analysis could be crucial for informer prospective renters based on their family size and other needs.

Other possible analyses to further understanding of this data could be done through the lens of COVID-19 research. Since the data has the rent prices for each year from 2019-2021 and the socioeconomics covering up through 20202, one could compare how each different type of county was affected by the pandemic. Trends could be shown in how rent prices changed in urban vs rural areas, how unemployment rates rose in different types of counties, how populations shifted in different counties, etc. Research such as this would provide valuable insight into which communities need the most help to recover from the pandemic and also give insight as to which types of communities would need the most help should anything similar occur in the future.

# References

*50th Percentile Rent Estimates: HUD USER. 50th Percentile Rent Estimates | HUD USER. (n.d.). https://www.huduser.gov/portal/datasets/50per.html#null.*

*AHAR Reports: HUD USER*. AHAR Reports | HUD USER. (n.d.). https://www.huduser.gov/portal/datasets/ahar.html.

Alexander, S. (2020, December 8). Bloomberg.com. https://www.bloomberg.com/news/articles/2020-12-08/san-francisco-apartment-rents-drop-35-as-tech-embraces-remote-work-during-covid.

Bruni, P., Bortoletto, F., Kalyanasudaram, R., & McGeoch, G. (2011). *Db2 10 for z/os performance topics*. Vervante.

Centers for Disease Control and Prevention. (2021, June 8). *Rural-Urban Continuum 2013 (Grouped) Variable*. Centers for Disease Control and Prevention. https://www.cdc.gov/cancer/uscs/public-use/dictionary/rural-urban-continuum.htm.

Diamond, R. (2018, October 18). *What does economic evidence tell us about the effects of rent control?* Brookings. https://www.brookings.edu/research/what-does-economic-evidence-tell-us-about-the-effects-of-rent-control/.

Dillon, M. (2020, November 17). *How Do Landlords Determine Rent?* Home Guides | SF Gate. https://homeguides.sfgate.com/landlords-determine-rent-78037.html.

*Download Data*. USDA ERS - Download Data. (n.d.). https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/.

Goodman, L., Neal, M., & Choi, J. H. (2020). Can Unemployment Numbers Predict the Number of Mortgages That Will Go into Forbearance? *Urban Institute*.

Lengerich, R. (2012, January 22). *Nation's top three poorest counties in western South Dakota*. Rapid City Journal Media Group. https://rapidcityjournal.com/news/nation-s-top-three-poorest-counties-in-western-south-dakota/article\_2d5bb0bc-44bf-11e1-bbc9-0019bb2963f4.html.

Lien, T. (n.d.). *Inside San Francisco's housing crisis*. Vox.com. https://www.vox.com/a/homeless-san-francisco-tech-boom.

Local Attractions - Dunn County, ND . (n.d.). https://www.dunncountynd.org/index.asp?SEC=416D1674-A752-4EDA-A7B5-4F7B652B23E1&Type=B\_LOC.

Loving County, Texas Population 2021. (n.d.). https://worldpopulationreview.com/us-counties/tx/loving-county-population.

Mazzara, A. (2019, September 27). *Census: Income-Rent Gap Grew in 2018*. Center on Budget and Policy Priorities. https://www.cbpp.org/blog/census-income-rent-gap-grew-in-2018.

*Montgomery County Maryland Residential Rent and Rental Statistics*. Department of Numbers. (n.d.). https://www.deptofnumbers.com/rent/maryland/montgomery-county/.

Prince George's County, Maryland Cost of Living. (n.d.). https://www.bestplaces.net/cost\_of\_living/county/maryland/prince\_george's.

Schuetz, J. (2020, October 1). *How can government make housing more affordable?* Brookings. https://www.brookings.edu/policy2020/votervital/how-can-government-make-housing-more-affordable/.

Velsey, K. (2018, January 5). *How Does Your Landlord Calculate Your Rent?* The New York Times. https://www.nytimes.com/2018/01/05/realestate/small-landlord-calculate-rent.html.

*Washington, DC Rental Market Trends*. Apartments for rent – RENTCafé. (n.d.). https://www.rentcafe.com/average-rent-market-trends/us/dc/washington/.