How Population Density Affects Living Conditions

Authors: Darwin Lopez, Michael Wong, Kevin Milne

Summary of research questions

1. Looking at data in 2019, how can population density affect the employment rate in the Washington State?

   Answer: It appears that population density *does* affect the employment rate in the Washington State When we plotted our regression line, we found that it was upwards sloping and our model had a low mean-squared error. This low mean-squared error indicates that our model accurately predicts the actual data when we use our training sets to inform our model. For a better understanding of our regression model, examine the graph below.

2. Looking at data in 2019, how is the change in population affecting housing in the Washington State?

   Answer: When we designate our responding variable as housing in Washington state, our model suggests that population density *does* affect the percentage of the population which occupies a house. While our model is fairly heteroskedastic[1], there seems to be an association between higher populations and higher rates of occupancy.

3. Looking at data in 2019, what other factor can affect the employment rate and housing in the Washington State?

   Answer: Keeping the population as a constant variable, out model showed that the medium income directly affects the employment rate and housing. It is seen in the positive upward sloping regression line in the regression plot. From all of our indicators factors the income showed to be important when looking for employment and housing in a county.

Motivation

Our motivation for learning about employment and housing is to gain a better understanding of the factors that can affect our future careers since we are most likely going into the job field after graduation. In order to focus our research, we decided to select one factor that has a lot of data and is likely to affect employment and housing. This factor was population. This will also help others understand whether there will or has been a change in employment, housing, or other factors based on population changes. Our understanding might be that these factors are not being affected but after analyzing the data set they might say otherwise.

Datasets:

1. Population Density 2019:

[1] Heteroskedastic refers to the tendency for our model to be less accurate as X increases. While this doesn't mean our model is wrong or biased, it does affect how accurate it may be for predicting values based on higher populations. Part of this may definitely be related to the outlier in our data, King county.

https://drive.google.com/file/d/1RbSYkJhK7PixlqRPqwLnI5J8AQa3DAKO/view?usp=sharing

This dataset was accessed through DataPlanet. It includes columns of countries in Washington State and population density. The population density by square mile across states in Washington. Through the DataPlanet portal, we can download this file as a CSV file.

2. Selected Economic Characteristics:

https://drive.google.com/file/d/1a6WXKG-QM-rb79gb4UHGEUYAnzKkag3V/view?usp=sharing

This was accessed through the Washington State census database. It has an array of economic indicators for each county in Washington. We used this dataset to see what other indicator affect employment rates and housing statistics. This data is available on the Washington State census database and is available for download as a CSV file.

3. Unit Occupancy Status

https://drive.google.com/file/d/1edvCHWxMQ66Yvb8YHf8qj8daonTB8eQc/view?usp=sharing

This dataset was accessed through DataPlanet. It stores and reports on housing units in the Washington State for specific counties in Washington State. It contains the columns reporting the country, occupied or unoccupied units. The dataset has a housing unit listed as a house, an apartment, a mobile home, a group of rooms, a single room, boats, recreational vehicles (RVs), vans, tents, or railroad cars. This data was collected by the American Community Survey (ACS) conducted by the US Census Bureau and listed under the 2015-2019 Washington State census. It can can be exported as a CSV file through DataPlanet.

Method

Question 1 - How can population density affect the employment rate in the Washington State?:

Developing Code:

a. We imported the Pandas, scikits-learn, and matplotlib library
b. Import CSV of the data file(s) and clean the data.
c. We took the population density in different areas (counties and states), using the 2019 (most recent) dataset, and compare it to the employment rate in the same area and year.
d. Examining Washington as a training dataset, we used population as a feature, and employment rate would be the label.
e. Using the given data, we ran regression analysis over data in other states.
f. Repeat this process with other locations.

Test:

a. Calculate mean squared error for each of our regression analyses.

Question 2 - How is the change in population affecting housing in the Washington State? :

Developing Code:

a. We imported the Pandas, scikits-learn, and matplotlib library

b. Import CVS of the data file(s) and clean the data.
c. We took the population density in different areas (counties and states), using the 2019 (most recent) dataset, and compared it to the housing by occupancy in the same area.
d. Examining Washington as a training dataset, we used population as a feature, and housing by occupancy would be the label.
e. Using the given data, we ran a regression analysis over data in other states.

Test:

a. Calculate mean squared error for each of our regression analyses.

Question 3 - What other factors can affect the employment rate and housing in the Washington State?:

Developing Code:

g. We imported the Pandas, scikits-learn, and matplotlib library
h. Import CSV of the data file(s) and clean the data.
i. Calculate the change in our label relative to the change in our feature (how much does employment rate or housing change by when our features change?)
j. We examined columns other than population density and housing by occupancy and compare the calculations we make in g.
k. Our output will consist of the ratios in descending order in a list.

Test:

l. Create custom datasets that we import into a testing file where we make sure our code is calculating the ratios correctly.

Results:

Our discussion of our results will be sequential; we will start with explaining our findings for our first research question, then move to our second question, and finally we will conclude with a discussion of our last research question.

Question 1- Looking at data in 2019, how can population density affect the employment rate in the Washington State?

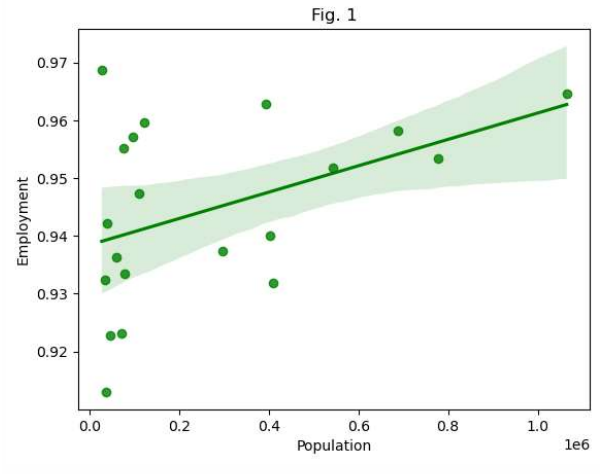

Figure. 1 Comparing Population and Employment

From Figure. 1, which was coded using seaborn and matplotlib, it is shown that population density seems affect the employment rate in the Washington State. The regression line is an upwards slope with our model showing had a -1.4 R-squared value. The model also showed a low mean-squared error of 0.000662. While our R-squared value was pretty low, we still believe, given our mean-squared error, that our model can predict employment based off of probability and that it suggests a positive association. This low mean-squared error indicates that our model accurately predicts the actual data when we use our training sets to inform our model.

Question 2 - Looking at data in 2019, how is the change in population affecting housing in the Washington State?

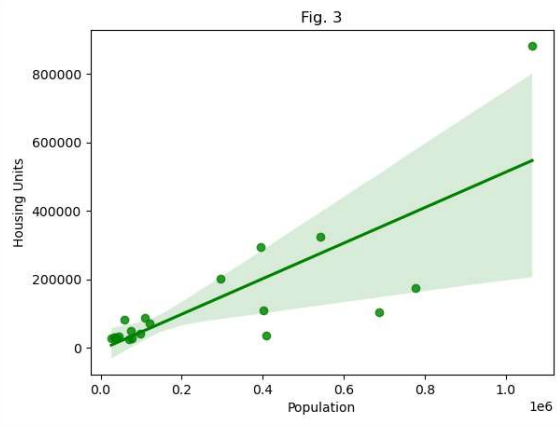

Figure. 3 Comparing Population and Housing

In Figur. 3, which was created using seaborn and matplotlib, a clear upward trend is shown with the regression line slope. From the model created it was determined that the R-squared value is -0.829. For ther mean-squared value it was determined as 0.000396. Therefore these factors suggests that population density *does* affect the residents occupying a housing unit. While our model is fairly heteroskedastic[2], there seems to be an association between higher populations and higher rates of occupancy.

Question 3 - Looking at data in 2019, what other factor can affect the employment rate and housing in the Washington State?

[2] Heteroskedastic refers to the tendency for our model to be less accurate as X increases. While this doesn't mean our model is wrong or biased, it does affect how accurate it may be for predicting values based on higher populations. Part of this may definitely be related to the outlier in our data, King county.

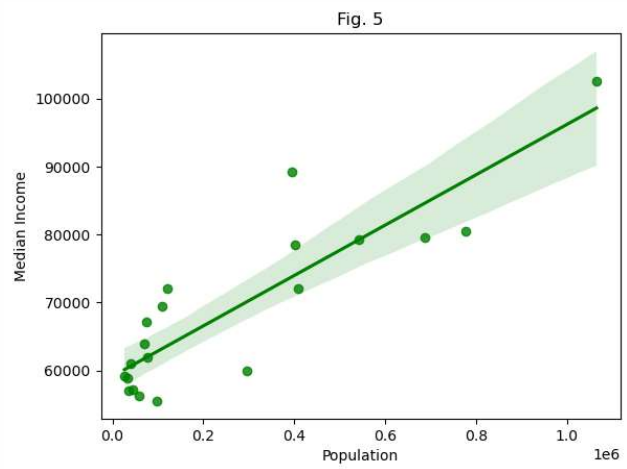

Figure. 5 Comparing Population and Income

In Figure 5 it is hown that there is a clear upward regression line slope. This was created using seaborn and matplolib. In our model the population was kept a contant and the column were examined for a contribution factor and it was determined that the medium income was another factor that was influencing the employment and housing units in Washington state from out regression model. The model showed to have a R-squared value of 0.141 indicating a positive correlation. Although this is a low value it is still a good indicator because the data ranges alot in population and income in Washington state with Seattle being the outlier. The mean-squared value was determined to be 0.000268. Thhis low value indicates taht our model predicts our data as the testing data did aswell.

Impact and Limitations:

This research helps residents living in Washington State understand how population and income are affecting employment and housing. Due to there being a variation in county to county the results should not be tied to a specific county. This shows the results of the state as a whole. Our data seems to suggest positive associations between population and percent of people in housing. This runs contrary to the intuitive notion that there is more homelessness in bigger cities. If higher populations also are tied to

Our data had a significant outlier. Seattle is a tech city and is a hot destination for people to move to with a very high market value for homes. Thus, if while Seattle is associated with more housing and with higher employment, there may be other confounding factors that explain this. In order to account for these hidden variables, further research should focus on multivariate regressions and putting each variable in context with each other. Maybe employment sector has an effect on employment and housing rates, or maybe median income is more important for creating a model than population density is.

Challenges:

- Multiple Datasets

  Our challenge was definitely going through each data set, cleaning it and making sure it had the necessary information to implement in the code. Therefore we merged three

dataset which also included many errors and going back to clean the original data files. We merged these datasets by creating a for loop and passing in a set of datasets which our program merged by matching counties.

- Machine Learning

  Machine Learning was useful since we are trying to compare different sets of data. Machine learning was used to determine the correlations, which in turn helps answer our research question.

Work Plan Evaluation:

Proposed:

a. Access, clean, and read all datasets being used (2 hours)
b. Sort and display desired information (4 hours)
c. Use Machine Learning to make models and computations help answer research questions (3 hours)
d. Reflection (1 hour)
e. Prepare a report (2.5 hours)

Reflection:

Our original work plan was not met and far from expected due to schedule changes and finals. Although we had these set back we managed to still meet for 2-3 hour periods to work. We met at least once a week and kept communication. We did not expect he machine learnin and computation took more time than expected (about 6 hours). The Testing was not included which extended the work plan by 1 hour.

Testing Code:

A smaller data file was used to test the merging of the data. There was no means of testing the ML and main since they both directly depended from the merging of the data. Insead, we wrote a series of testing files which we ran our merge data function in which merged multiple datasets into one. To accomplish this, we transformed the testing dataframe and then dropped all na values within the testing function. We then transformed our merged dataframe into a list then checked if it was equal to our testing list. That is not to mention that we split our data into a training and testing set when performing our ML. Due to our low MSE when our code is run, our model is statistically accurate when run on our data. According to these mathematical formulae, you should trust our findings.

Collaboration:

Only team members and course staff worked on this project.