



Capstone 2 Presentation



Context

- This dataset was county level data for demographics, job, and education.
- The data was put together from two separate datasets that were joined together.
- The first dataset came from census data and the second one was a dataset found on Kaggle.

Criteria of Success

In this analyses I wanted to know which features had the greatest effect on income. I wanted the results of this analyses to be able to form an understanding of what features in a county led to greater prosperity.



Constraints

Environmental and
real estate data
were not present
for this analyses

I did not pull the
education data by
age for this project

I don't have
county tax or
budget data

Data Sources



<https://data.census.gov/table/ACSST1Y2021.S1501> – This is the census data I got the education data from



https://www.kaggle.com/datasets/muonneutrino/us-census-demographic-data?select=acs2015_county_data.csv – Any non-education data came from this data set

Data Wrangling process

There were 3220 counties in the county demographics data set

There 830 counties in the county education data set


From the education dataset I pulled the education data for percent of high school graduates for males and female and then the percent of college graduates for males and females for each county

I also pulled the total high school and college graduates percentage


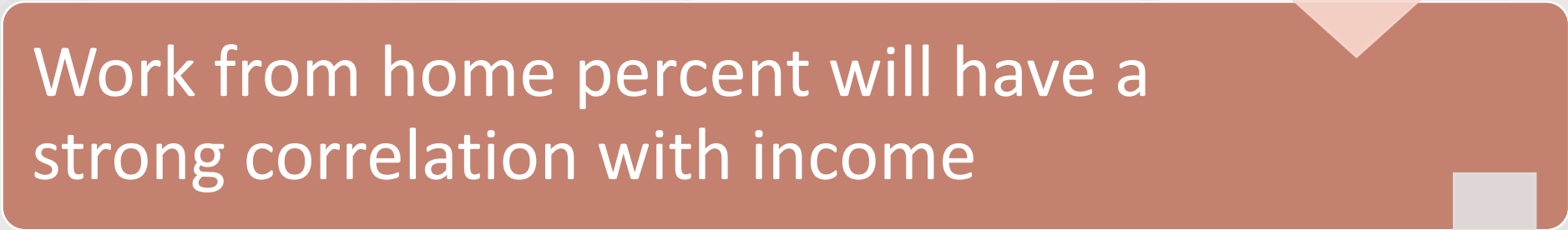
After joining the data there were 802 counties left in the new dataset

Predictions About the Data

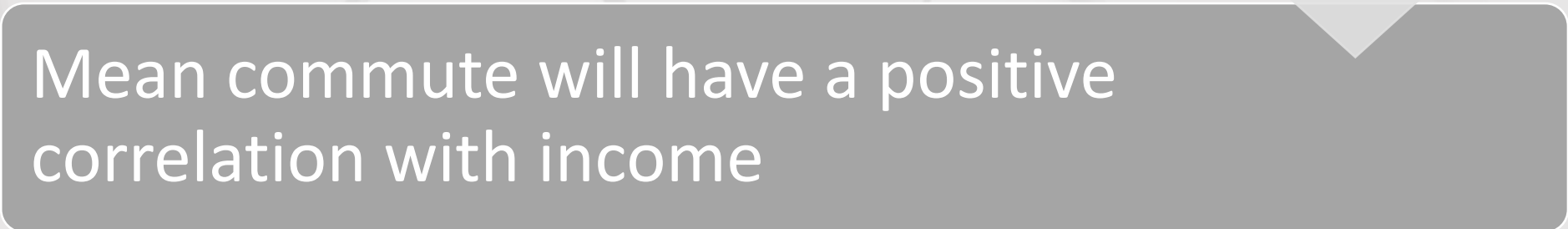
Higher rates of education would lead to higher income



Work from home percent will have a strong correlation with income



Mean commute will have a positive correlation with income



EDA - Correlations



AS EXPECTED, EDUCATION DOES HAVE A STRONG POSITIVE CORRELATION WITH INCOME AND A STRONGER CORRELATION WITH INCOME PER CAPITA



WORK FROM HOME PERCENT SHOWED TO HAVE A MODERATE CORRELATION WITH INCOME

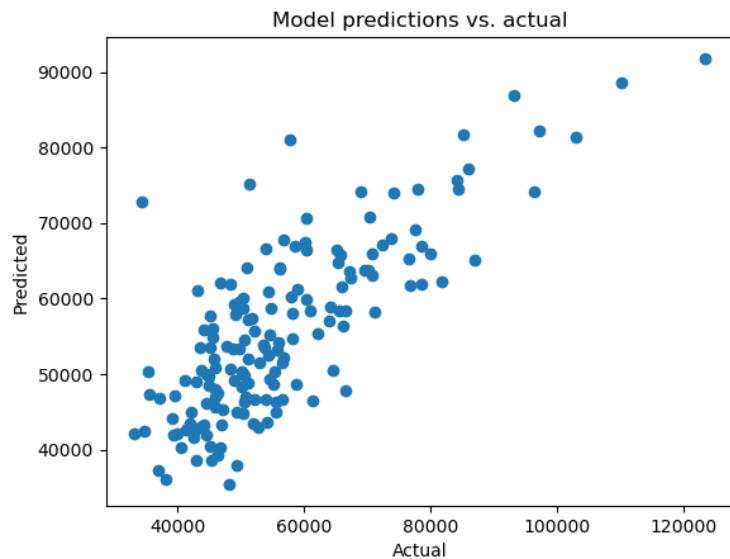


MEAN COMMUTE HAD A MODERATE TO ALMOST STRONG CORRELATION WITH INCOME

Models used

- Regression
- K-Means Classification
- DBSCAN

Regression



- Model used – OLS from the statsmodel.api library
- The regression model showed an R^2 value of .656 using only Total Percent bachelor's degree or higher and MeanCommute features

K-Means Classification

- The highest silhouette score was at 2 clusters followed by a drop off
- The next highest peak was at 6 clusters
- We got the best results using the same features as in our regression model [Total Percent bachelor's degree or higher and MeanCommute]



DBSCAN

- DBSCAN was run using a grid search
- DBSCAN was not able to identify any distinct clusters
- I tried running DBSCAN using multiple variations on used features and there was no satisfactory clustering discovered



Conclusions

Education seems to be the greatest determinant of income in this dataset

This dataset does fit neatly into clusters as demonstrated by the clustering algorithms run

More data is needed to be able to differentiate the counties into groups that can indicate the factors of prosperity

It is also possible that counties are too large, and that the data would be more telling if it were more granular, down to the city level.