Capstone Project

**Analysis of wages in Colorado**

**James Flanagin**

**August 2019**

**Table of Contents**

**Introduction 4**

**Databases 5**

**Client 6**

**Data Wrangling 7**

**Data Exploration 8**

Wage and Metropolitan Statistical Area

Wage and Experience

Wage and Employee Count

Wage and percentile

Wage and Population

Job keywords and MSA

**Data Modeling (supervised Learning) 21**

Wages and Population, Experience, Employee

**Data Modeling (Unsupervised Learning)26**

Logistic Regression

Kmeans

Principal Component Analysis

Word Frequency

**Assumptions and Limitations 35**

**Recommendations and Future Work 36**

**Conclusions37**

**Introduction**

Colorado is one of the most desired places to live in the US. According to the US Census Beureu Colorado had the 5th highest total net migration in the US despite being the 22nd most populated state in the US. The nature and booming tech/ Cannabis industry have driven thousands of people to the state. The time frame we are going to explore is the growth from 2009-2015 which saw a positive migration of 484,379 new Colordan’s a 10% increase from 4,972,195 to 5,456,574 residents in 2015. The goal of this study is to see what the effect on wages looks like during an increase in population.’

**Databases**

<https://data.colorado.gov/Labor-Employment/Employment-Wages-in-Colorado/busm-qa5b>

Employment wages by occupation, year, and area, from Colorado Department of Labor and Employment (CDLE), 2009-2015

<https://data.colorado.gov/Labor-Employment/Population-Estimates-by-Year-for-Counties-in-Color/bu8h-8sux>

Annual population estimates for each year by county for the state of Colorado, from US Census Bureau, from 1900 to 2015 provided by the Colorado Department of Labor and Employment (CDLE).

The main Metropolitan Statistical Area (MSA) which are defined by the CDLE and featured in the provided dataset:

1. Pueblo MSA (Pink)
2. Greeley MSA (Brown)

3. Grand Junction MSA (Purple)

4 Fort Collins-Loveland MSA (Red)

5. Denver - Aurora MSA (Green)

6. Colorado Springs MSA (Orange)

7. Boulder-Longmont MSA (Blue)

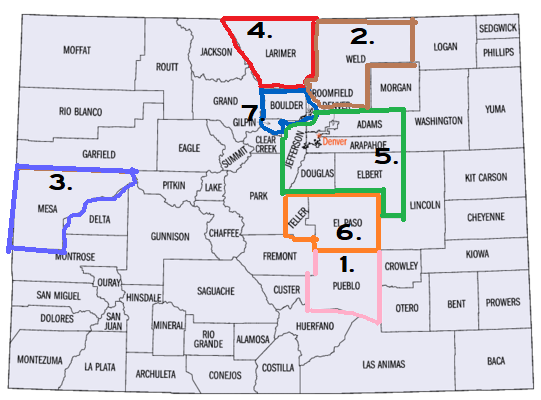


Fig1.

**Client**

There are several clients who would be interested in the outcome of this project

* A college Sophomore who is attending one of Colorado’s great universities who intends to stay in the state once they are finished. They could get a good feel for what occupation and location would be a great fit for them. Once they graduate they could have a basis for salary negation as well as how they should expect it to increase over the time period that they are working in a given job/ metropolitan area
* An Entrepreneur who is writing a business plan that is planning on setting up a business. Knowing what the mean salary given their projected business size as well as their office location will be key in understanding overhead. This can also be used for forecasting financial projections.
* Wages are the basis for standard of living in a given MSA. The city government can use this in instances of rent control and understanding tax revenue.
* Developers and residential real estate companies assessing rent and appraising property can learn what to expect from the average citizen of different MSA’s.

**Data Wrangling**

**Employment Wages:**

* A good deal of redundancies and useless columns were dropped i.e panelcode,
* any of the user defined columns were dropped as they presented conflicting information
* Any duplicate and summarized information. In order to avoid assessing a sum vs it’s parts I dropped any row that contained summarized information. i.e State Totals, hourly wage vs annual salary.
* After the Merge any descriptive data with one or few unique values ‘areatyname’, or any data that didn’t provide useful information eg. ‘indcodty’ was dropped

**Population Data:**

* Any summary data was dropped. Colorado and US totals were removed and County based data was all that remained
* Counties were grouped into their respective MSA
* Dates were adjusted to line up with Wage data (2009-2015)
* Columns that were useless were dropped. Population, periodyear and areaname were kept for merge purposes
* All county based data was summarized to reflect the total population of each MSA
* Only a portion of Colorado data is included in the major MSA’s. Counties not included were scrubbed.
* 2015 was missing data. Data was extrapolated based on the % change from prior year
* Data was merged based on Areaname

The final shape of the data= 19598 rows × 13 columns

**Data Exploration**

The final shape of the data= 19598 rows × 13 columns**.** There are 7 MSA’s as show above (Fig.1). Each row represents a unique year 2009-2015, and a unique occupation 836. This means that some occupations are missing a few years, or an MSA.



In the first part of analysis we look at the total Colorado change in average mean revenue.

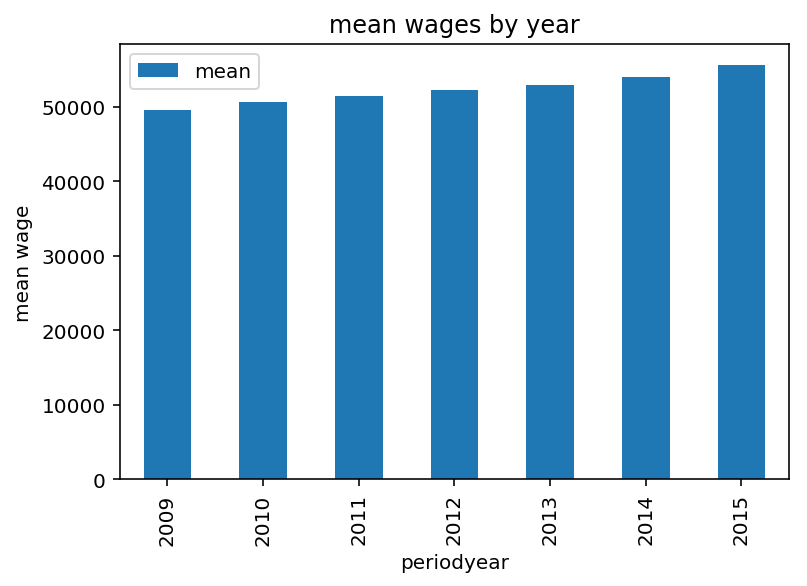


Fig.2 shows the mean wage by year is increasing slightly year over year on a steady trend

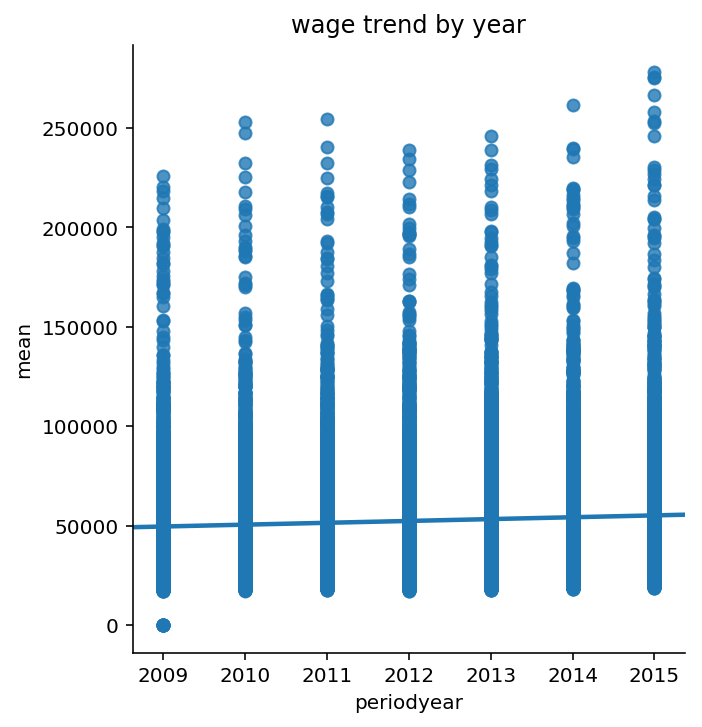


Fig.3 a lmplot shows that there are a few less outliers in 2012 and 2013 but still indicates a positive trendline for the state

When we divide it into MSA we can see a key distinction in mean wage

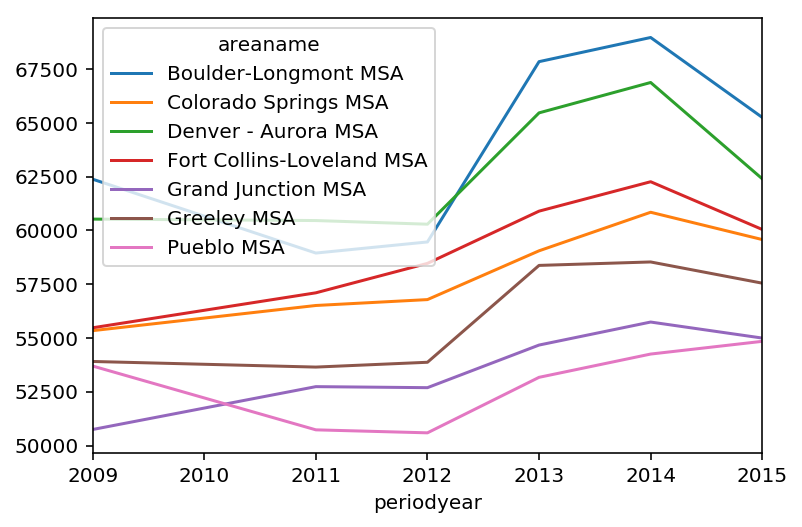


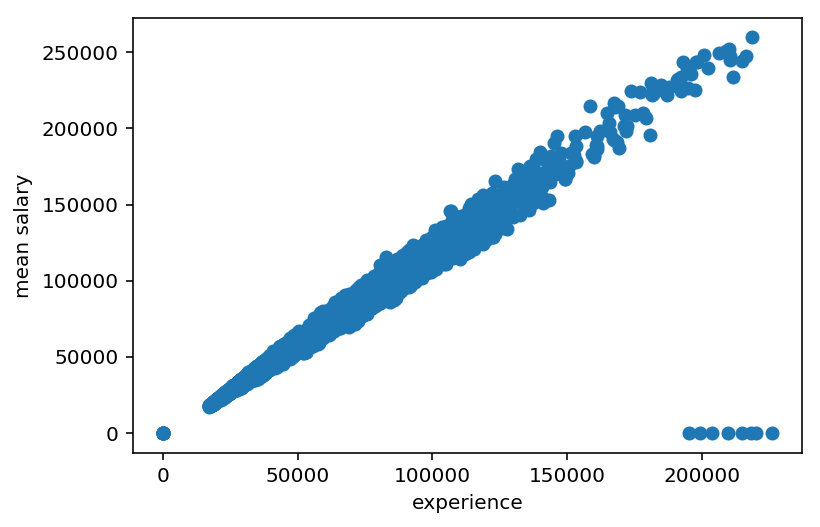
Fig.4 shows the trend in mean revenue across MSA. Boulder and Denver battle for first place as B-L pulls away in recent years. We show a slight dip in major locations except for Colorado springs, Fort Collins and Grand Junction (all agriculture heavy locations) after the recession in 2008 but increasing since 2012 in each location

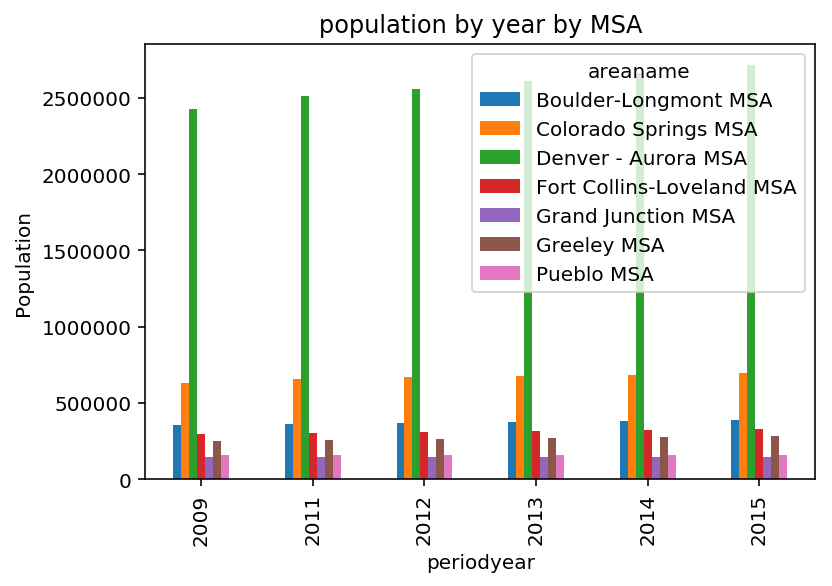
Fig.5 Compares experience and mean salary. We can see a tightly distributed positive correlation with some outliers where the mean salary is missing.

Fig.6 shows the population by year by MSA. This shows it steadily rising, with Denver-Aurora MSA being the clear leader. Colorado Springs is in second.

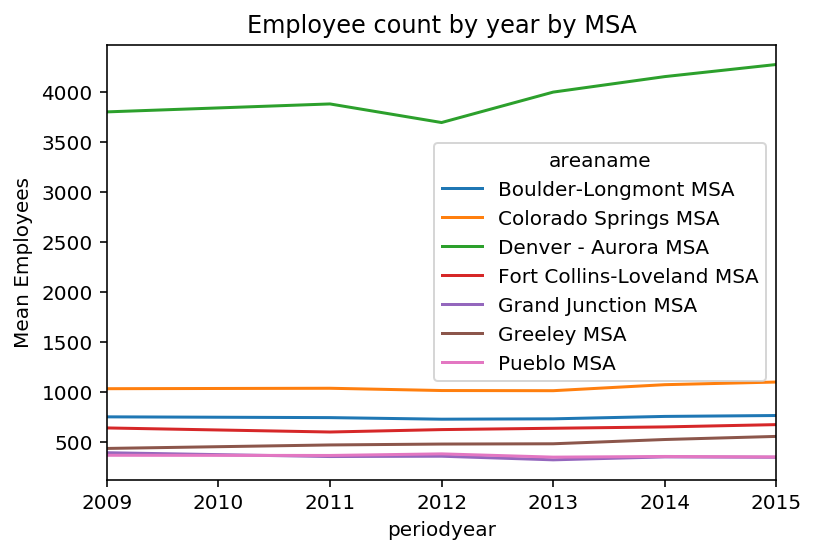


Fig.7 shows average employee count by locale. The more populated regions logically have higher concentration of employees by occupation.

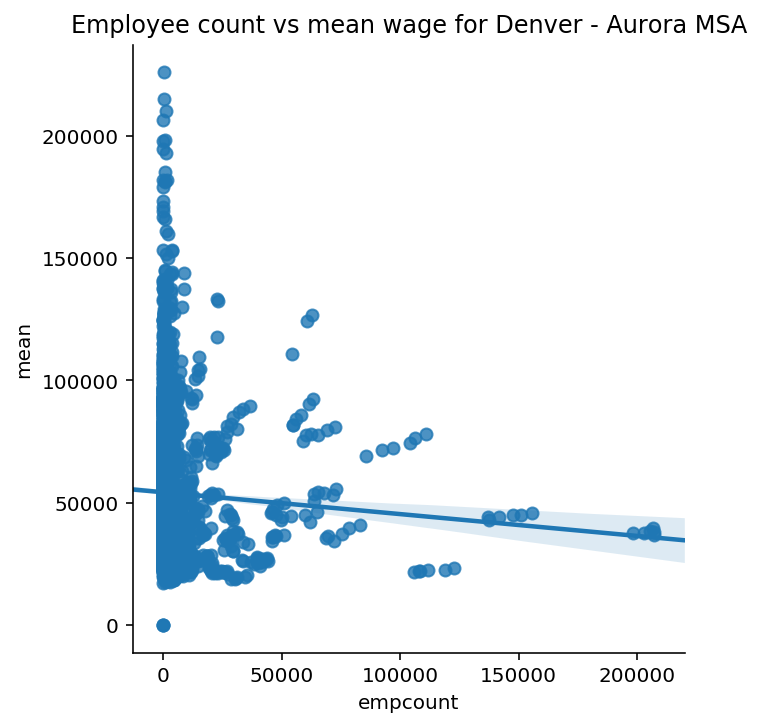


Fig.8 Shows Denver – Aurora MSA mean wage compared against the employee count for each occupation. Shows a slight negative correlation

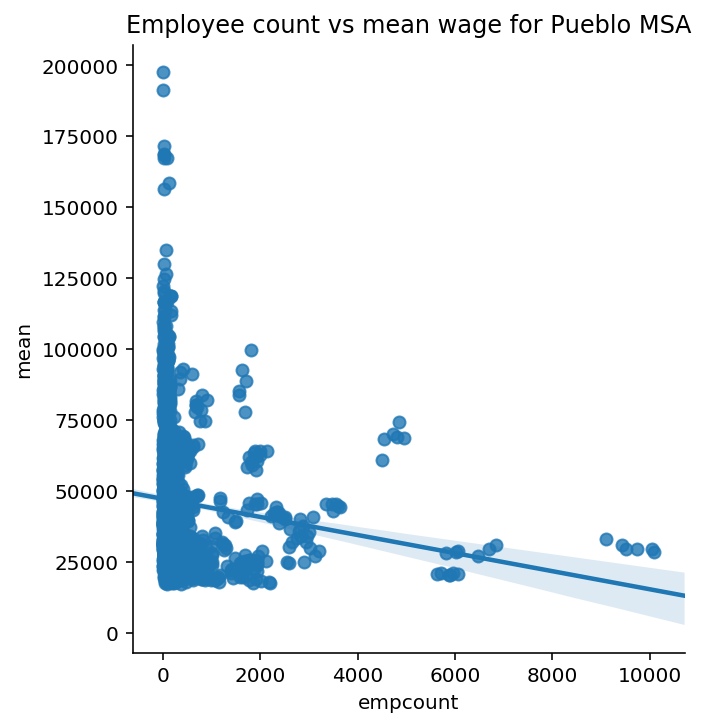


Fig.9 Shows Pueblo MSA mean revenue compared against the employee count for each occupation. Shows a higher negative correlation.

We can see that occupations with more workers tend to have a lower mean wage. We also see that across a high population area(Denver), and a low population area(pueblo) that the majority of jobs have less than 2000 total workers.

Finally let’s look at the ceilings and floors for each job by state

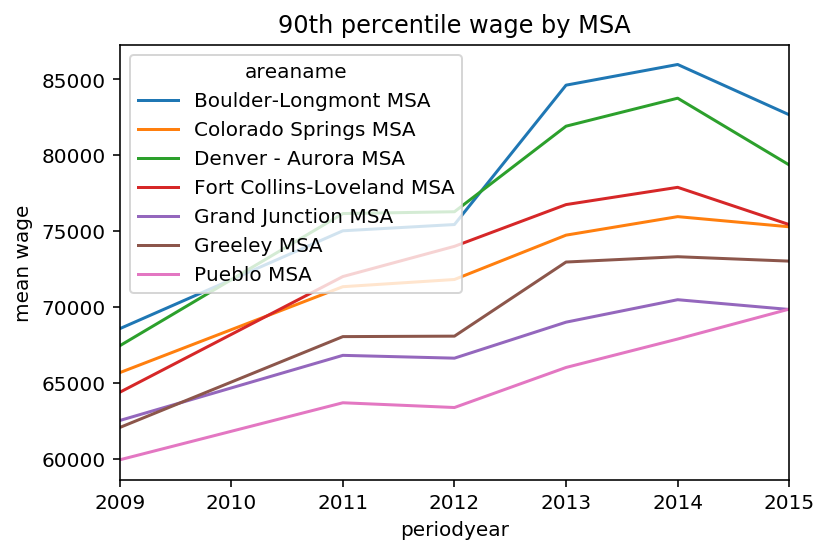


Fig.10 show’s the 90th percentile or the mean top pay by year for each city. Boulder-Longmont leads the pack and is $25k higher in the peak year of 2014. We can see they all follow similar trends, and were relatively stable in a recovering economy

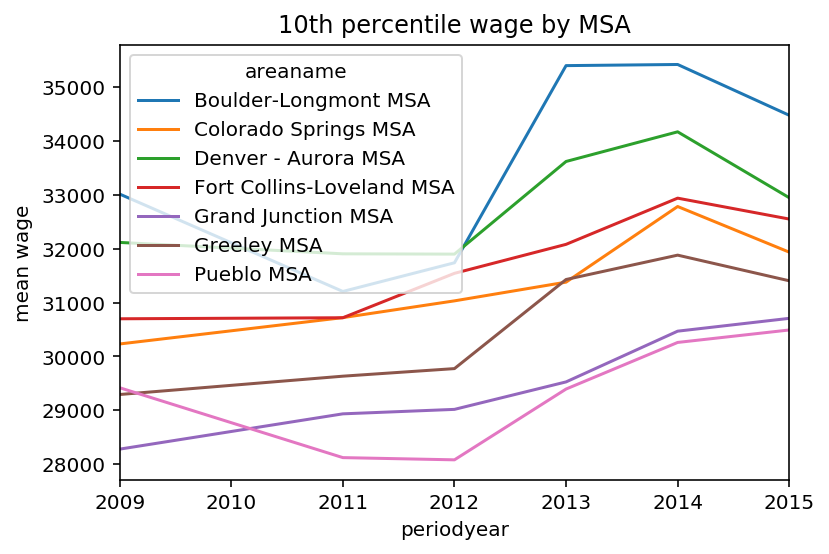


Fig.11 shows the bottom 25th percentile of wage by MSA. We see that the lower paying parts of each occupation are much slower to recover from the recession. They have a consistent negative trend until 2012 where each MSA increases across

**Data Modeling (Supervised Learning)**

**Linear Regression Model**

The columns below are the variables we are going to explore. The primary purpose is to understand what effects the mean wage of a given occupation.

|  |  |
| --- | --- |
| **Column** | **Description** |
| mean | Mean wage for the occupation. |
| entrywg | Entry level wage for the occupation, mean of the first third (ALC definition). |
| experience | Experienced level wage for the occupation, mean of upper two thirds (ALC definition). |
| pct10 | Wage at tenth percentile. |
| pct25 | Wage at twenty-fifth percentile. |
| median | Median wage of the occupation; also the wage at fiftieth percentile. |
| pct75 | Wage at seventy-fifth percentile. |
| pct90 | Wage at ninetieth percentile. |
| population | Number representing the population total for the specified geographic area and time period. |

The first step taken was to look at a basic correlation matrix to see how the variables interact. From the exploratory analysis above we can start to assume a few things. One is that experience is going to be highly positively correlated with mean wage, and that the number of employees at a given job will slightly negatively correlate with the mean wage.

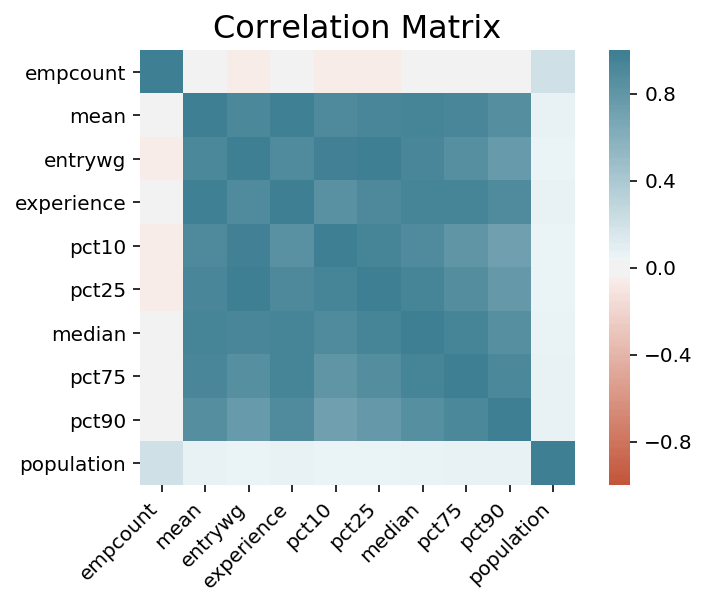


Fig.12 Correlation matrix. We can see that the obvious values that correlate more closely to mean are any of the percentile wages (which makes sense as mean moves in tandem with those variables). Experience is also a highly positively correlated as one would assume, the amount of experience attached to a certain job should inform how the amount the person who works that job is compensated. The employee count seems only slightly negatively correlated, and the population of an MSA in which the job is located shows little correlation.

Using a linear regression model let’s see if we can predict some prices. Out Estimated intercept coefficient is: 963.5609897620525

Our estimated coefficients are listed below:

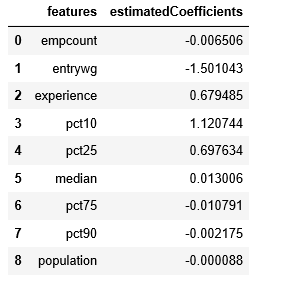


Fig.13 Estimated coefficients

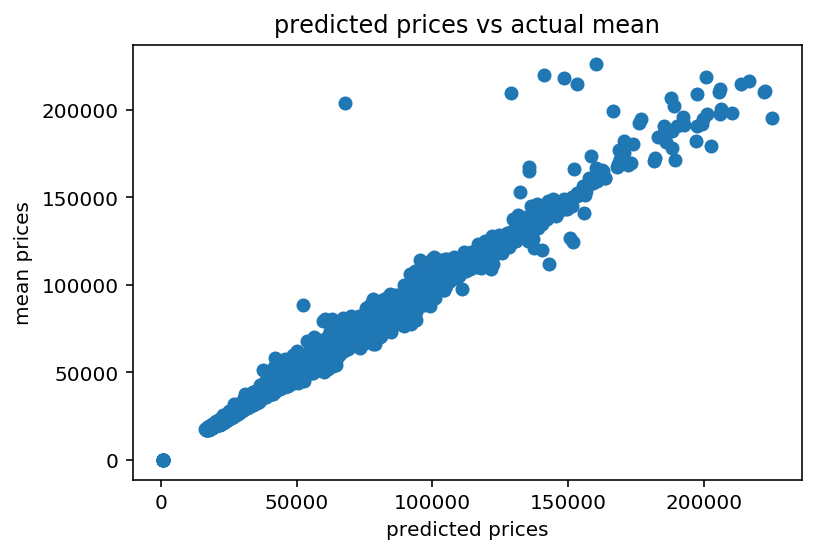


Fig.14 Predicted prices vs actual prices

Using these variables let’s build an Ordinary Least Squares model to estimate the best fit model for simple linear regression.

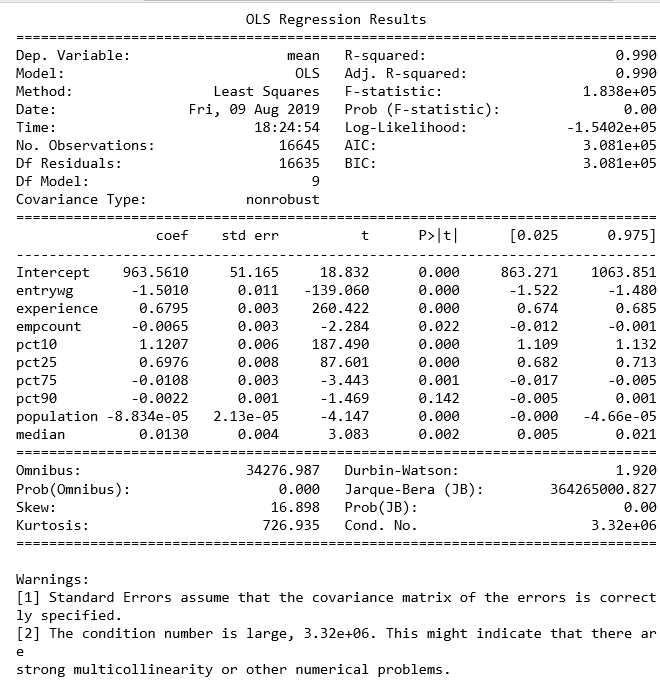


Fig.15 OLS using all variables

The warning below indicates that we may be over fitting as a result of multicollinearity.

The percentile variables have been removed so that only employee count, experience, and population remain.

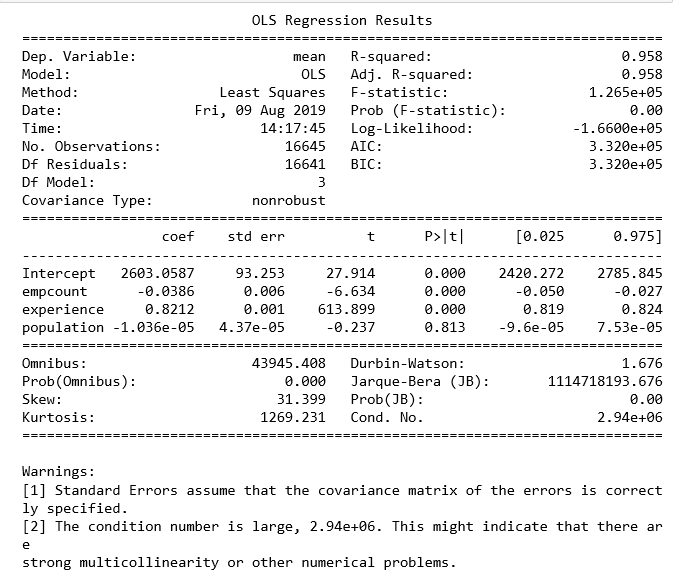


Fig.16 OLS using only empcount, experience and population

We still see a high level of multicollinearity, and high scores for skewness and kurtosis, implying our model is not normally distributed. We also can assume we have some level of overfitting as a result of the high r^2.

**DataModeling (Unsupervised Learning)**

Next we attempt to perform logistic regression with a training and a testing dataset to see if we can predict when mean revenue. We place our data into 3 quantiles and attempt to predict if the data will be placed in the top quantile.

To reduce over fitting we test the model with both experience(the variable with the highest t score indicating it is the most statistically significant.

With experience included we see an r^2 of .48. With experience removed the r^2 increases to .66 indicating that the generalizability of the model with experience added is lower. The overfitting was slightly improved, but we are able to conclude that more variables should be considered when attempting to predict mean wage. Further exploration of factors like population density, housing prices, crime rates, average employee age and other standard of living metrics should be performed to improve this model.

Let’s look at how data can be clustered using kmeans

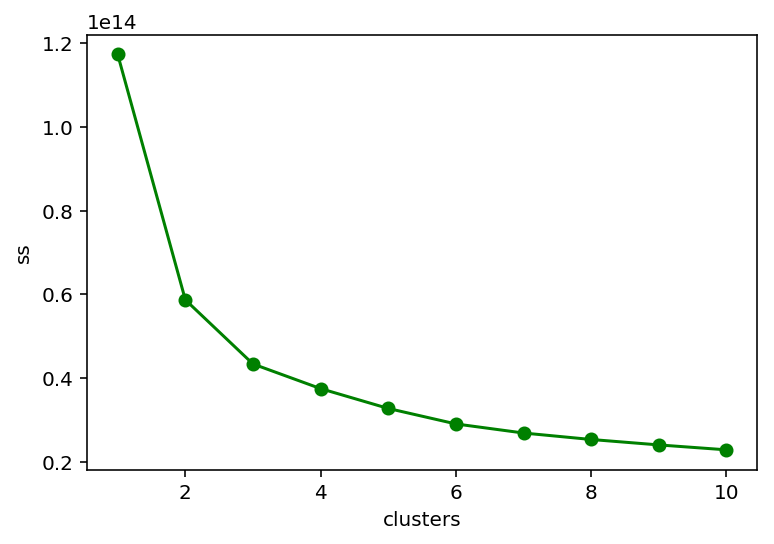


Fig.17 indicates that there is an elbow at 3 showing that the data can be clustered into three groups. Let’s see how they group by size

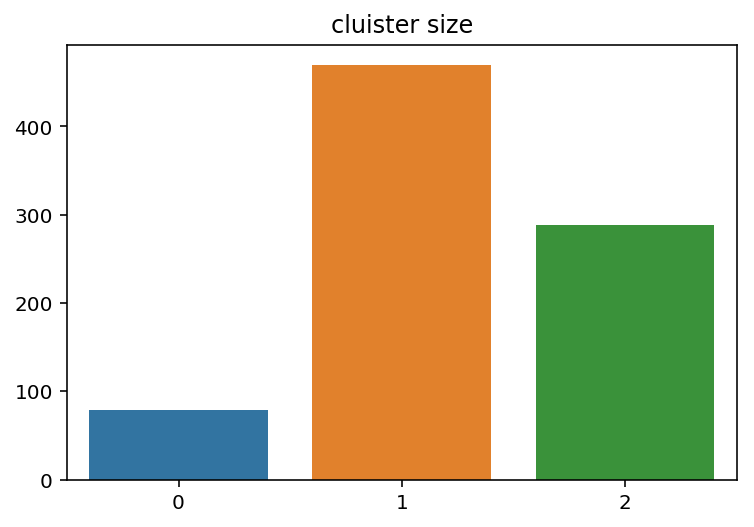


Fig.18 shows the size of each cluster

Next using principle component analysis we can plot the data to see how evenly distributed the clusters are.

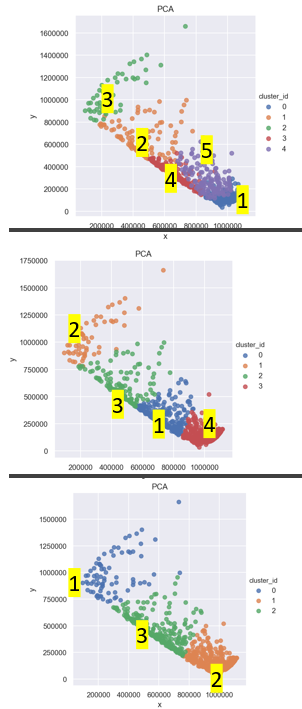


Fig.19 shows the dispersion of each cluster using 3-5 clusters

Let’s see if we can find parse out information that’s involved in each cluster. Is there a common word associated with each cluster? Is one area more popular in a cluster than the other

First we should explore if there are job titles that are outliers. This could unfairly weigh the data against certain words

When arranged by the count of codetitle we see

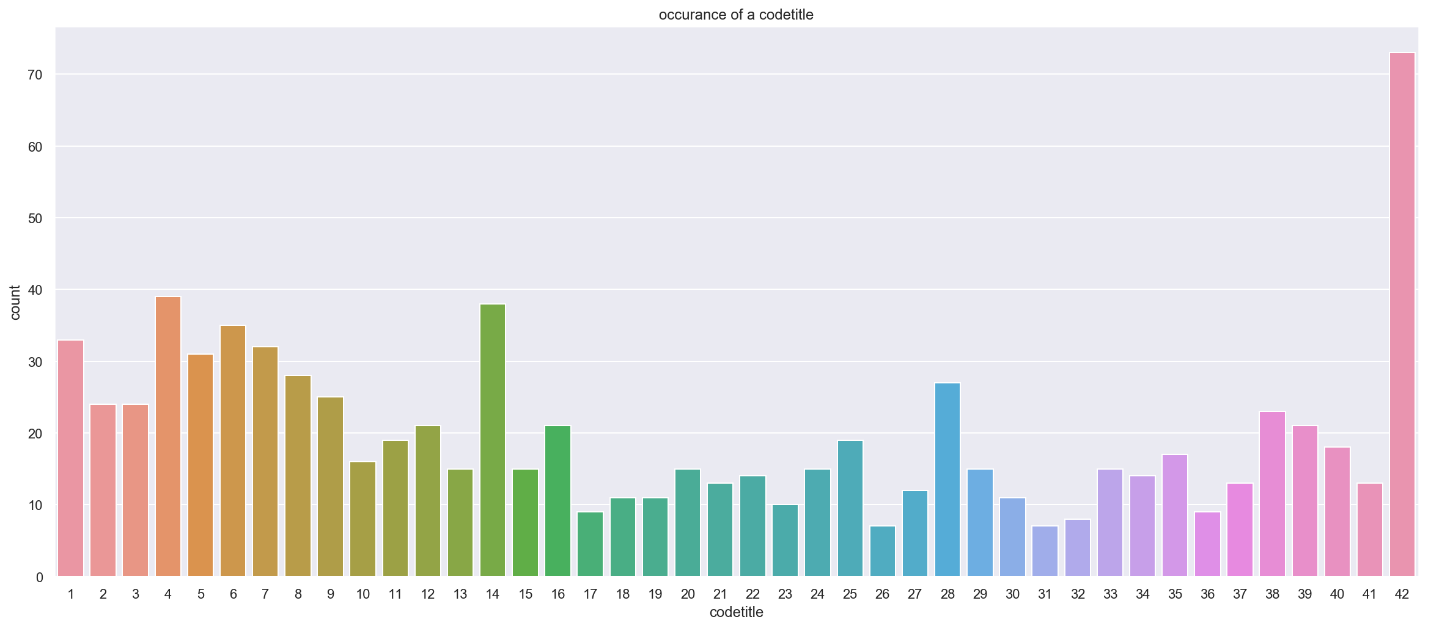
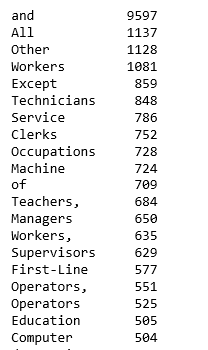


Fig.20 shows that the most an occupation will show up in the data is 42 times. There are about 73 jobs that show up this many times. Let’s see if there is a word that is most common from those jobs



Fig.21 shows the most popular words involved in the top occurring job titles. This will be important as we move into showing what’s in each cluster

We also had to address what the most common words may be in the entire dataframe for codetitle



It’s interesting to note the differences in the most common words above and the list from the entire dataframe. Workers is not included as a common occupation title. Any word above that doesn’t help us understand the clustering wordcloud will be removed. Ie. And,all,other, except.

Finally lets compare the 3 clusters for codetitle and areaname

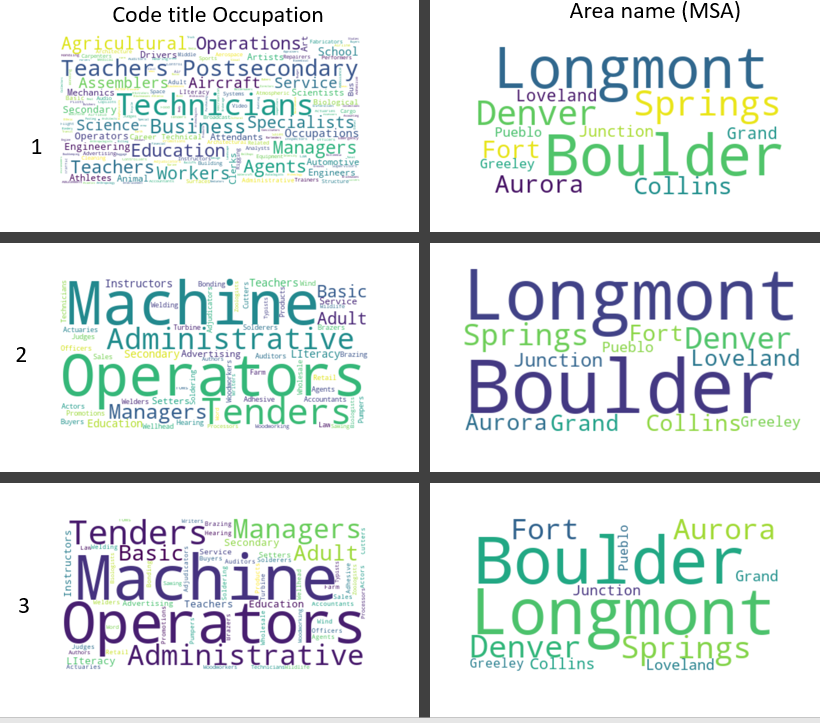


Fig22. Shows the 6 wordclouds and their specific cluster/ areaname.

From this we can loosely definte a few things about each cluster

Cluster1: looks to be more technical. Words like athlete actor, agents aircraft and engineering are prominent and unique. We can also see that they favor urban areas as Boulder and Denver and the springs have the biggest showing here

Cluster 2: looks to be a bit more professional managerial class with some more technical and trade individuals listed. These people heavily favor boulder but are more evenly spread across areaname than the first.

Cluster 3: seems to be much more rural. With machine and operators dominating the word cloud. Tenders also indicates jobs in the Cannabis industry, so this would likely be people who migrated to the state and are in their first year of work or general working class. Boulder is huge, but the other cities seem evenly distributed as well.

**Assumptions and Limitations**

* The Colorado Datasets had some missing information by year so it will be weighted heavily against jobs that have 2009-2015 filled out completely.
* The data had very few variables to analyze in a meaningful manner for regression. They were either too un-correlated or too strongly correlated to the point where it caused overfitting.
* Boulder being the highest PHD per capita location in the US could have a significant skewing factor on the data for wages. It is likely that other low population cities such as that would be weighted towards a high wage. This may be an outlier across the US
* Certain areas of Colorado have significant oil subsidies. Compared to a lot of the rural US there is a decent job market in less populated areas

**Recommendations and Future work**

* Some analysis on standard of living would be a much more reasonable way to analyze the data, as that may have features that can be generalized across the US
* Closer analysis on neighborhoods in Denver could tell a more interesting story. Aurora has lowest educated rates in Colorado, and is ranked quite low in the US, but was grouped in with Denver. An analysis and comparison of Aurora on it’s own, being an urban and lower income location would be interesting in contrast to Boulder
* There are a bunch of counties that were not included in the overall employment data. Some of the most rural parts of Colorado are left out of analysis. Additionally, Pitkin County includes Aspen which is one of the wealthiest areas in the country, which was also not included in the Colorado websites data
* Usable industry codes would have also grouped the data more naturally. Fewer industries would have been more easy to manage than 850 unique occupations

**Conclusions**

* Given the data and variables that I was provided, it is insufficient to build a model of expected wages, simply on population, employee count, and experience alone. Moreresearch needs to be done on education levels of an area and other standard of living metrics
* Despite the limitations we can conclude that Boulder, and Denver perform quite well and increased significantly in wage growth over the last few years as an average
* Bottom 10th percentile see a significant lag in wage growth in comparison to the top 90th and even the mean. They seemed to struggle the worst from the recession even in a part of the country that has strong industries and available jobs.
* Experience is an extremely strong indicator of wage, more information needs to be addressed to see how true that is across all industries. In most cases being very experienced at one occupation can lead to a higher wage.