**Population Density Shifts and Unemployment**

**By**

**Sneha Godbole, Thomas Hauger, and Brett Steele**

1. *Purpose and Description of Project*

Our original goal with this project was to explore the relationship between population density shifts and unemployment rates within the geography of the US Census Zip Code Tabulation Areas (ZCTA’s). We chose to use ZCTA’s because this is one of the geographies that changes the least through time. We wanted to see if there was any correlation between the two variables. We did not approach this issue having a particular expected outcome, nor did we approach our research with any additional research questions in mind. We decided to let the data dictate to us which direction our additional research would go, however, we did want to answer the fundamental question at hand concerning the relationship.

1. *Problem Statement- Obtaining Data*

Starting off we researched from where we can find the two most recent datasets that included both population density and unemployment rates by ZCTA. After researching, it turned out population density was not directly available, but working age population numbers and unemployment rates were. Therefore, these would be the main variables we would focus on. Our original goal was to do a span of 5 years, however, the two most recently published datasets that contained both unemployment rates and working age population for ZCTA geography were from the 5 Year American Community Survey (ACS) from 2007-2011 and the 2000 Decennial Census Summary File 3. [1] Therefore, these were to two datasets we chose to download and compare.

After choosing our data we still had an issue; we did not have population density. We only had working age population. However, we did have the working age population tied to a ZCTA geographic area. Therefore, in order to develop population density all we needed was the land area for each ZCTA. This could also be obtained, (and was obtained) from the US Census Bureau via shapefiles available for download. [2]

After downloading the shapefile geography for the ZCTA’s we needed to make sure the area was given in square miles. Alas, it was not. Therefore we had to transform our data. We did this via ESRI’s ArcGIS software program. [3] The original file came with the Geographic Coordinate System North American 1983 format which measures distance based on decimal degrees. Since we wanted the area in square miles we had to utilize the Project tool in ArcGIS program. With this we were able to transform our dataset into the World Geodetic System 1984 World Mercator projection that provides measurement availability in miles. Thus the step needed to find the square mileage of a ZCTA after this was to select the Calculate Geometry tool within the ArcGIS program. Now we had our ZCTA’s in square miles. All that we needed to do was export this data to a .csv file to match our Census data.

1. *Objective- Code Development and Execution*

Now that we had the refined geographic file we were able to move on to developing our code to process the data. We developed our original code on a local machine because we found the data was too large to add to GitHub, (the data was later added directly to India at Future Systems). We chose to write our code in the Python language and we did so utilizing the Enthought Canopy software IDE. [4]

Originally our code started off with working with three csv files. These were the two Census files and the ZCTA geography file. In developing our code we first imported the csv and pandas modules within Python. Then, we transformed our three files eliminating columns we did not need, and giving the columns heading names that we desired. From here we fused the data together into one dataset by joining the data based on the ZCTA name which we labeled “AFFGEOID10.” After this, we eliminated geographies that had missing data. We chose to do this so we did not have skewed results, and because of the potential of not being able to move forward due to columns not calculating correctly. Moving on, we then converted our fields to the desired field type for processing. This would ensure we could perform the necessary math on the number fields which we did in the next step. This step was to calculate our population densities for each dataset and to calculate the density changes. After this was performed, we finally added export options for visualization. We choose to have the dataset export as both a .csv file and a .json file. Therefore, this was our original code.

Going forward we later realized we needed more data in our dataset for desired visualization and statistical development. We would first utilize ArcGIS again on our geography file to obtain XY centroid coordinates for our ZCTA’s instead of recreating the geography file. We decided just to add this in as an additional .csv file. Therefore, we now had four .csv files. The same steps detailed previous were used to add this file in. Also, we decided since more data such as age group information was available for the Census dataset that it would benefit us to add in some analysis as for age groups. It would also provide a better visualization. Therefore, the addition of these variables was added to the original Python code.

From there, we leveraged Ansible to orchestrate and deploy our software configuration on a remote VM (‘projdata’). Relevant files of note include: inventory, enable-root-access.yml, site.yml as well as the projectdata playbook directory. Within this directory, we have tasks and defaults (variables) defined in the respective subdirectories.

The specific tasks carried out with our Ansible playbook include:

* Enabling root access on the VM
* Installing and configuring pip and python (including numpy and pandas python packages)
* Executing the ProjectScriptPandas.py script that was previously described to facilitate data ingestion and processing data disparate data sources resulting in a final, formatted and aggregate .csv file which was later used to visualize via D3.js.

In totality, the following tools and software were used to construct a solution for analyzing and understanding the impact of unemployment rates on population density shifts in Zip Code Tabulation Areas:

* ArcGIS
* Python
* Ansible
* D3.js
* Openstack Compute Resources (VM- ‘projdata’)

Next, we took the processed .csv file and graphically represented this data via D3.js.

Using D3.js we have created multiple dashboards. For first two dashboards (Unemployment rate and population density change) we have used color coding of US States map. As data was big so formatting data was most important part. So first we have formatted data by assigning key, value pair to data. Here key is ‘state’ and value is average of unemployment rate/population density change.so now for each state we have average value. We chose to show values at the state level for this project due to visibility in a presentation. However, the ZCTA’s were still the basis for this information, and future work could include focusing on ZCTA’s. To show this on map we have used color coding according to average values. Unemployment rate/Population density changes from color ‘yellow’ to ‘red’.

In third dashboard we have sorted the data according to age groups. We have used grouped bar chart to show the data. Here we have calculated average of each age group per state and then showed into grouped bar chart. We have also added mouse events (mouse over and mouse out) on each dashboard. When you mouse over on particular state on first dashboard it will show you unemployment rate of that state. We have attached snapshot of dashboards in result section below.

1. *Results*

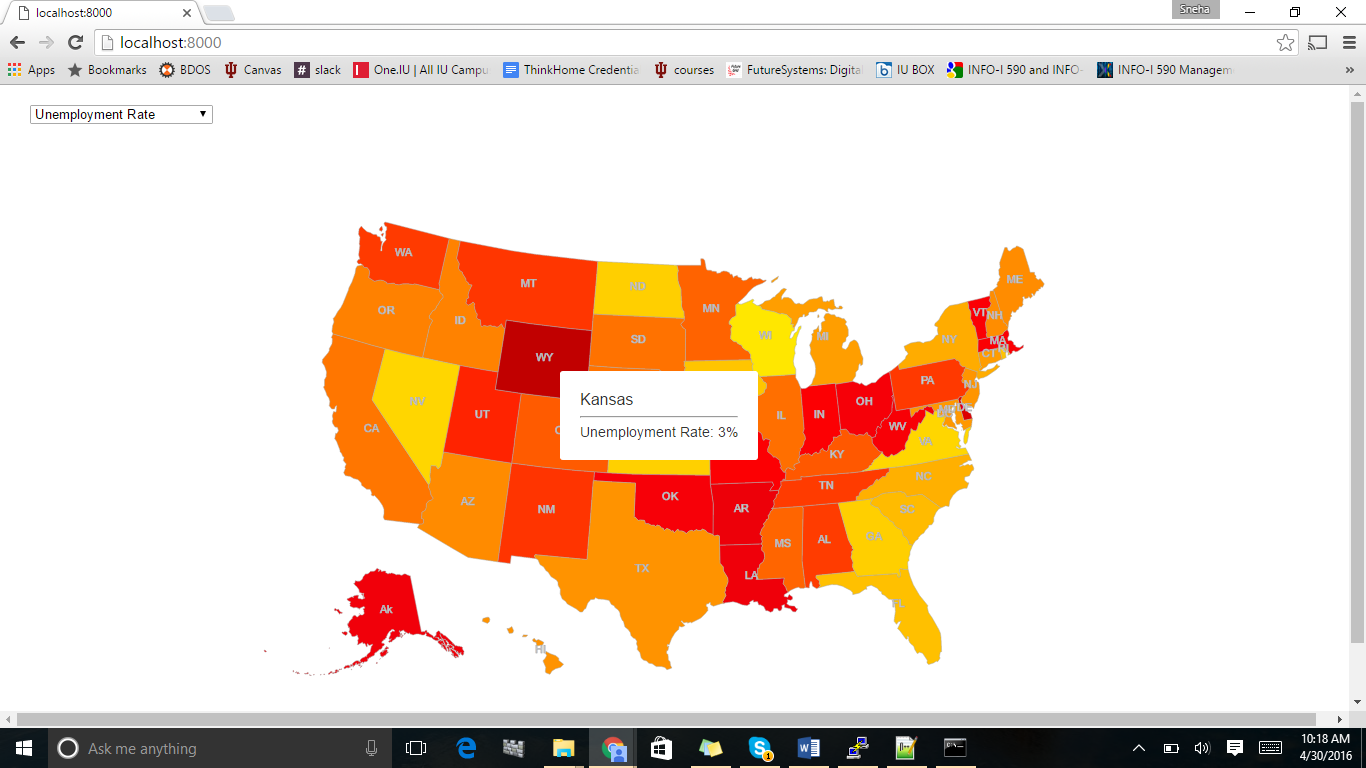


Figure 1. Unemployment Rate Across US.

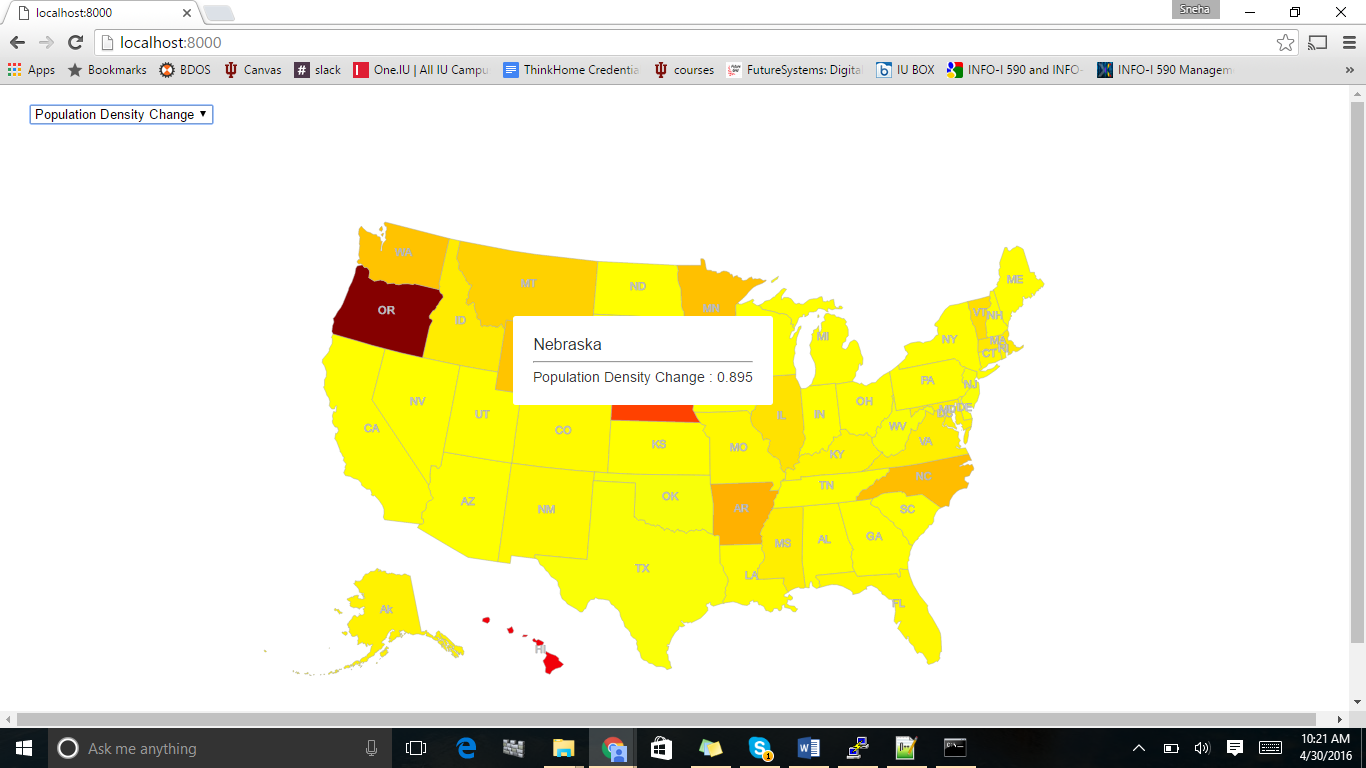


Figure 2. Population Density change across US

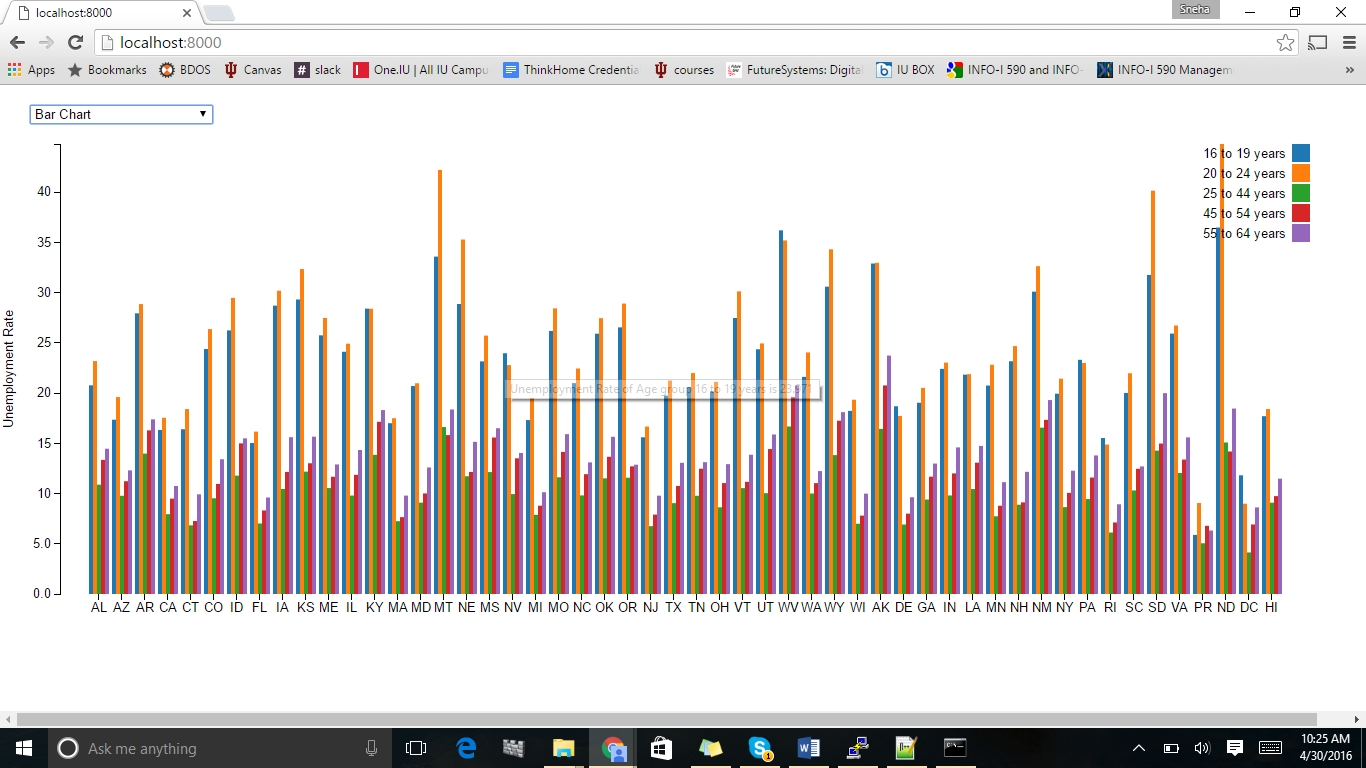


Figure 3. Grouped bar chart

Here are some high level takeaways based on the visualizations analysis –

1. The unemployment rate chart shows most of the states have unemployment rates change under 10%, however some of them have very high unemployment rates ranging from 13-20%.
2. Population density change in most of the states is less than 1%, however few of them have as high as 3%.
3. Unemployment rate for age group of 20-24 is relatively higher than other age groups provided.
4. *Findings*

The data analysis and visualizations done are based on the dataset provided, which contains information / data points for a timespan of a decade. In case we could get more historical data, say around 4-5 decades or so, we would like to design and implement some predictive analytical algorithm / model which will learn from historical data and then can predict the un/employment rate OR population density change over the period of time. Also, we would to put multiple metrics together to figure out whether there are any correlations or causalities among them.

1. *Implementation Instructions*

Below are the following steps for running this project. (Brackets below are used to dictate user-defined inputs.)

**1. Source appropriate files for nova command and select appropriate project number. Ensure openstack is loaded:**

*- source ~/.cloudmesh/clouds/india/kilo/openrc.sh*

*- source ~/.cloudmesh/clouds/india/kilo/fg491*

*- module load openstack*

**2. Start an ssh agent and add your own credentials:**

*- eval $(ssh-agent -s)*

*- ssh-add ~/.ssh/id\_rsa*

**3. Launch a Virtual Machine via nova:**

*- nova boot --flavor m1.medium --image Ubuntu-15.10-64 --key\_name [insert ssh key] [insert virtual machine name] --nic net-id=e5228c15-38af-4f91-a6de-1590d399427e*

**4. Generate a floating ip-address and associate it with your virtual machine:**

*- nova floating-ip-create ext-net*

*- nova floating-ip-associate [insert virtual machine name] [IP address]*

**5. Ensure you can ssh into your machine:**

*- ssh ubuntu@[IP address]*

**6. Clone project repository from github\*:**

*- git clone git@github.iu.edu:tehauger/groupproject\_bdossp.git*

*- cd [insert user github directory]*

*- git branch master*

*- git checkout master*

*- git pull https://github.iu.edu/tehauger/groupproject\_bdossp.git master*

*- git push -u origin master*

**7. Update inventory file with the [IP address] of your virtual machine:**

*- nano inventory*

*- [replace IP address with [VM IP address] of your virtual machine]*

**8. Create a separate virtualenv and install ansible:**

*- virtualenv $HOME/[insert directory of your choice]*

*- source $HOME/[insert directory of your choice that you just created]/bin/activate*

*- pip install --trusted-host pypi.python.org ansible*

**9. Navigate to groupproject\_bdossp directory and run the following code:**

*- cd [file/path/to/groupproject\_bdossp]*

*- export ANSIBLE\_TRANSPORT="ssh"*

*- export ANSIBLE\_SSH\_ARGS="-o ForwardAgent=yes"*

*- ansible-playbook -i inventory -c ssh enable-root-access.yml*

*- ansible-playbook -i inventory -c ssh site.yml*

**10. Now open a browser\*\* and type in the ip address of your virtual machine into the address bar:**

*- http://[VM IP address]*

**11. Enjoy exploring the U.S. unemployment rate data and population density change.**

\*Please check folder directory and ensure relevant files such as inventory, enable-root-access.yml, site.yml, and projectdata directory are displayed. The github code repository can be found at: <https://github.iu.edu/tehauger/groupproject_bdossp>.

\*\*Try Firefox if Chrome does not render the visualization properly.

1. *References*

|  |  |
| --- | --- |
| [1] | US Census Bureau, "American Fact Finder," US Census Bureau, 25 March 2016. [Online]. Available: http://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml. [Accessed 25 March 2016]. |
| [2] | US Census Burueau, "ZIP Code Tabulation Areas," US Census Bureau, 25 March 2016. [Online]. Available: http://www.census.gov/geo/reference/zctas.html. [Accessed 25 March 2016]. |
| [3] | ESRI, *ArcGIS Version 10.3,* Redlands, CA: ESRI, 2016. |
| [4] | Enthought, *Canopy,* Austin, TX: Enthought, 2016. |