Neighborhood Recommendation using Visualizations in D3.js

1. Basic Information

- Team Member 1
  - Name : Tanveen Singh Bharaj
  - U-Id : U1126027
  - U-Mail : tanveen.singh@utah.edu
  - Course : MSIS – Fall 2017
- Team Member 2
  - Name : Vishal Pandey
  - U-Id : U1126029
  - U-Mail : vishal.pandey@utah.edu
  - Course : MSIS – Fall 2017

- GitHub Repository : https://github.com/vishalpandey2192/DV_Project.git

2. Background and Motivation

This projects aims at visualizing the trends of employment, average salary, mortality rate, population and price per square area for the year 2016 in all states of USA. As per the study these factors are most important which directly impacts the habitability in a particular state. We aim at improving decision making and recommend best states to live through visualization of trends and weighted mean algorithm.

The motivation for this project was the struggle that people especially international students/employees have in evaluating that which state is right for them for studies or job. We our self being international students know how much research we went through, researching different websites and articles, reading reviews to know that how life could be in the city we choose to live. Therefore to easy this risky process of selecting the most favorable city to live we leverage the power of data visualizing to analyze the tenability of each US state.

3. Project Objectives

This project is divided into 3 sections:

1. Visualization of magnitude of different parameters in different states of US.
   In this section we analyze different states of the United States on the basis of the top 5 parameters while making a decision of relocating to any other state.
   These parameters are:
   i. Population Density
   ii. Job Opportunities
   iii. Mortality Rates
   iv. Average Pay
   v. Cost of Living / Per sq feet price
   In this section we create visualization of US Map using geoJSON where we show the impact of each factor on each state for year 2016 using color coding channel/color scale that makes it easier to visualize that how big or small is the impact of each factor on each state of US.

   Further for each parameter, we will use area channel encoding to represent the trend of this parameter in the year 2016 across all states using area chart, where the

bars will be arranged in descending order going from state having maximum value to state having minimum value. The aim is to make information grasping easier for users. The area chart will also have marker which would tell the national average of each parameter so that we could compare the score of each states in each factor with the national average and draw better inferences.

2. In the second part of visualization, we aim to create a visual recommendation system where we recommend the top 3 states to users based on the user preference.
   We achieve this by using the concept of weighted mean algorithms.
   In this approach we provide list of major parameters to users which he should consider while selecting a state, and the user specifies the degree of importance of each factor to him by assigning a number from 1-5 to them , 1 being the least important and 5 being the most important.
   Using the weight assigned by the user to each factor we compute a variable called the suitability factor which is calculated by the weighted mean algorithm.
   The suitability factor is calculated by multiplying the weight assigned by the user to each factor with the value of factor and then dividing number obtained for each state by the minimum value obtained.

S.No	Factors	Assumed Weights (1-5: 1 lowest, 5 highest)
1.	Prices per sq. area	2
2.	Mortality	1
3.	Population	3
4.	Jobs	5
5.	Salary Data	4

Sample Data for Year 2016

S.No	State	Prices per sq. area ($)	Mortality Rate	Population	Jobs	Average Salary Data($)
1.	Utah	172	4.2	193744	70000	70000
2.	New York	400	6.2	1059852	200000	100000
3.	Nevada	250	4.8	452351	30000	60000
4.	Kansas	150	7	337896	40000	65000
5.	Arizona	300	5	478952	150000	80000

On Applying weighted mean:

S.No	State	Summation of values*weight	Calculated Sum	Final value= sum/min(sum)
1.	Utah	1*4.2+2*172+3*193744+4*70000+5*70000	1211580.2	1
2.	New York	1*6.2+2*400+3*1059852+4*100000+5*200000	4580362.2	3.78
3.	Nevada	1*4.8+2*250+3*452351+4*60000+5*30000	1747557.8	1.44
4.	Kansas	1*7+2*150+3*337896+4*65000+5*40000	1473995	1.22

5.	Arizona	1*5+2*300+3*478952+4*80000+5*150000	2507461	2.07

Top 3 states:

1. New York
2. Arizona
3. Nevada

Now, the top 3 states with the maximum value of suitability factor are suggested to the user.

3. As, an add-on once we have given the top 3 suggestions to the user, we further use more visualizations to give him a clear picture of his choice.
   We will use line chart to show the trends of the any parameter chosen by user over the past 10 years over the top 3 suggested states. Such visualizations helps us to understand the changes which happened to the state over the past 10 years and also predict that seeing the past what could happened in the future.

4. Data

For each parameter -- population, employment, average salary, mortality rate and price per square area, the data sources are:

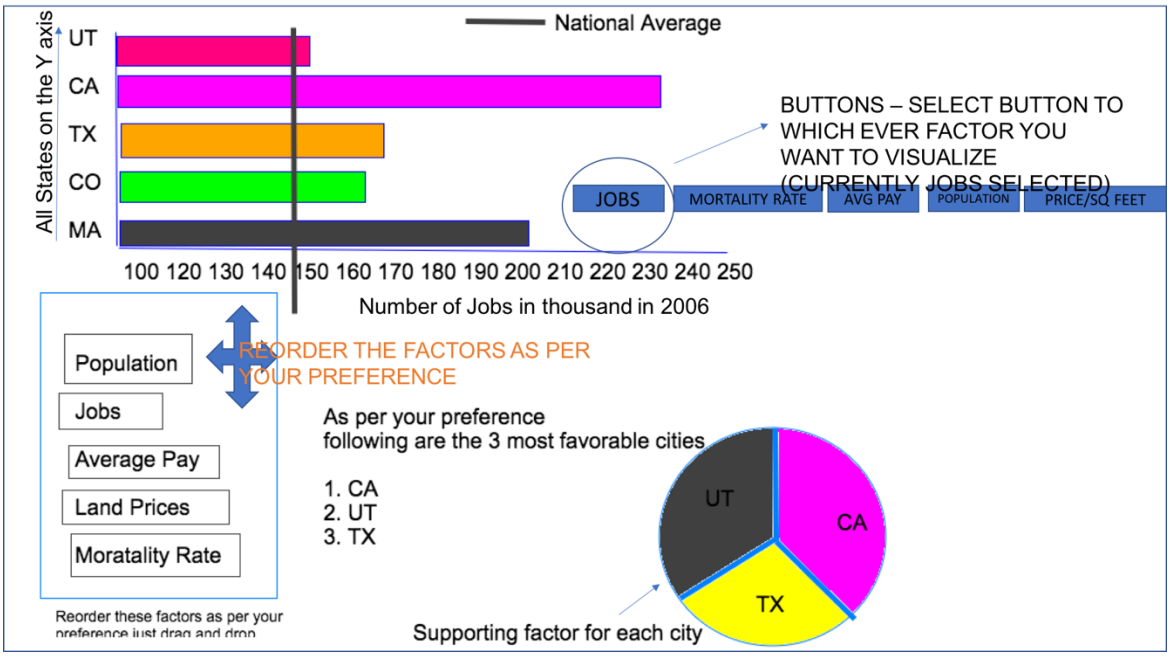
Population - https://www2.census.gov/programs-surveys/popest/datasets/
Price per square area - https://www.zillow.com/research/data/#rental-data
Employment - https://data.bls.gov/map/MapToolServlet?survey=la&map=county&seasonal=u
Mortality Rate – http://ghdx.healthdata.org/us-data
Average Salary – https://www2.census.gov/programs-surveys/acs/summary_file/2015/data/5_year_by_state/

5. Data Processing

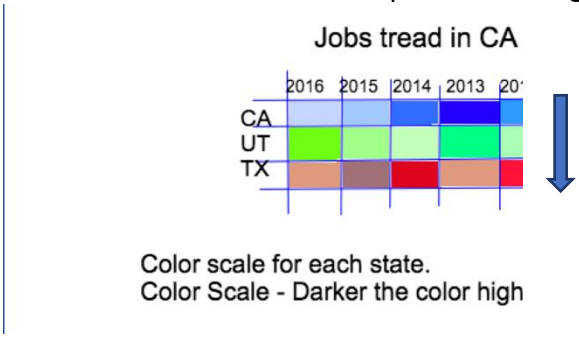
Most of the data sets we received were of county level and our visualizations are state level so we have to convert county level data to state level
In some cases we have data for last 10 years for each parameter in separate files so we have to convert it into single file
In some datasets the value give are not standardized like for the rents data set in some files the rent is in per square feet while at some places meter square so we need to standardize it.

6. Visualization Design

1. **Design 1**
   We first analyzed our design in 2 phases where in the first phase which show the visualizations and data for all the states for the year 2016 so that the user can analyze all the states at once to get a sound understanding.
   Whereas in the second phase we show data corresponding only to the 3 suggested states as per the user preferences.
   Our main aim is to show trends in parameters over past 10 years so the user has better understanding of those states.



Once the user receives the top 3 suggestions , he can select to view the yearly trends in those 3 states for each of the parameter to get more deeper insights of those states



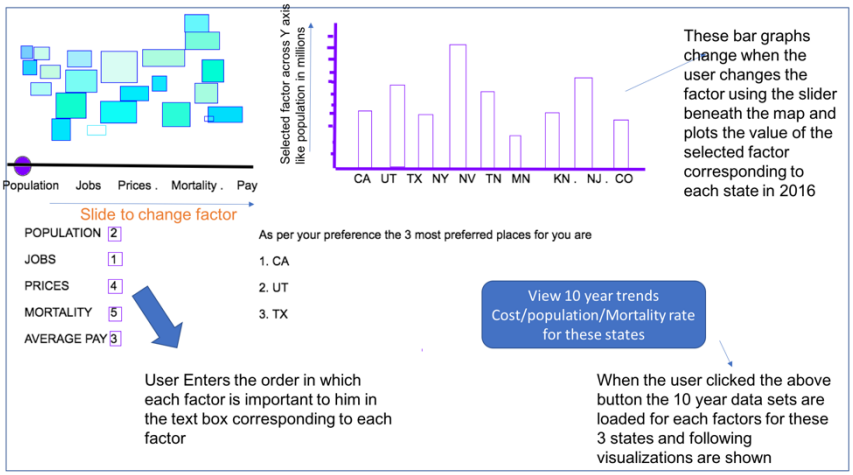
In this frame the user sees color grid map for each parameter he chooses to view and analyze.

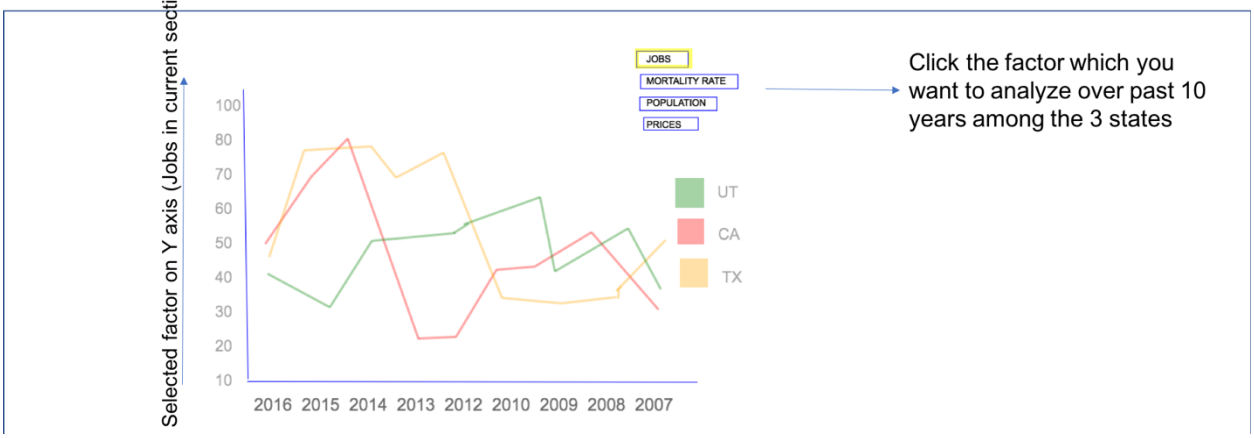
As in the above visualization we plot the employment from past 10 years and their number (denoted by the color channel) in each of the 3 states.

We will have 4 similar color grid map for each of the other parameter which are population, average salary, mortality rate and price per square area.

2. Design 2

In design 2 we have the first illustration where we use a tile chart of the united state and follow intensity based color coding to mark the value of the selected parameter on each state.
The parameter to view is selected from horizontal line chart below the map by moving the marker towards the parameter.



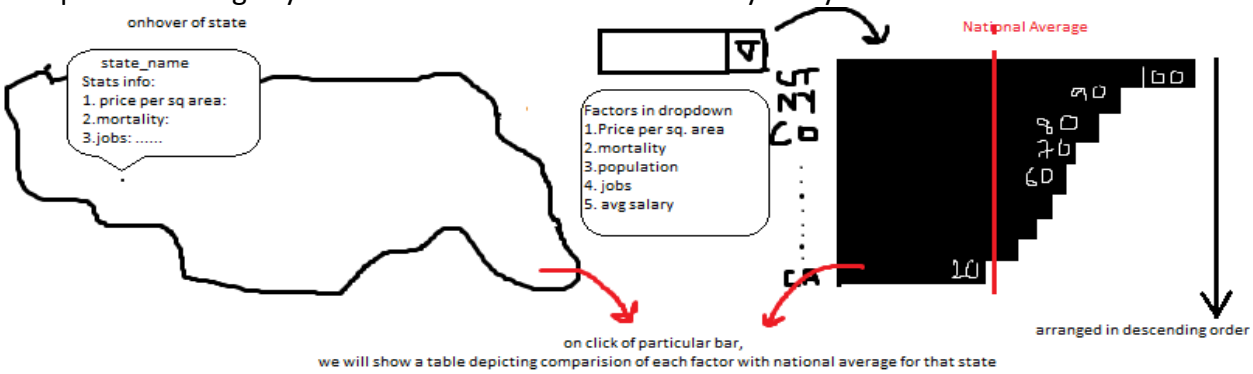


As mentioned earlier that the above visualization is for the 3 suggested states which we got after applying weighted mean algorithm. In this visualization the user can select the parameter he wants to analyze, view and compare among the 3 states over 10 years.

3. Design – 3 (Final Design)

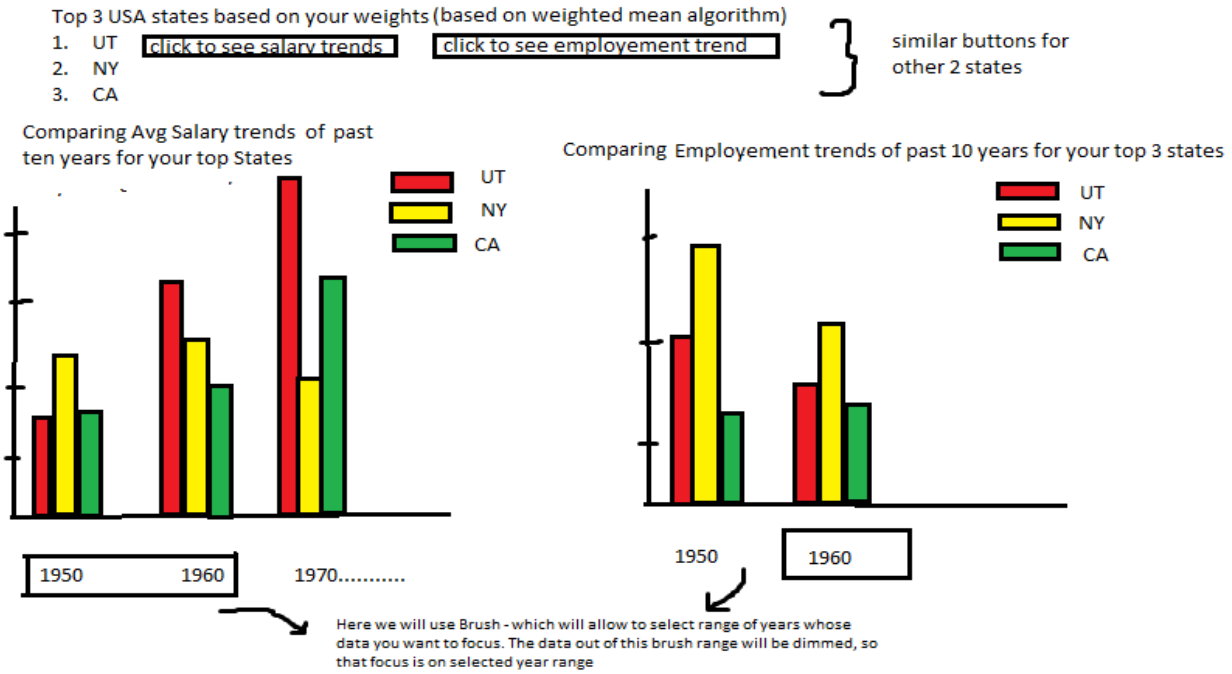
In design 3 , we will create map of USA using geoJSON. This is an interactive map where the visualizations change when you click on the map as well as when you hover over any particular state.

Moreover, as similar to second design, here where we show the trends in the parameters considered over the past 10 years for the suggested states but in addition to previous designs you can use the brush to select the years you want to consider.



	Utah	National Average
Prices per sq. area	172	500
Mortality	4.2	7
Population	193744	120502
Jobs	70000	150000
Salary Data	72000	95000

S.No	Factors	Input Weights (1-5: 1 lowest, 5 highest)
1.	Prices per sq. area	2
2.	Mortality	1
3.	Population	3
4.	Jobs	5
5.	Salary Data	4



We have finalized on the Design 3 due to:

1. The map of USA added will make the visualization very easy to grasp especially when used with color coding for different factors to show the intensity of each factor in each state.

2. Also, the visualizations made using the 10 year datasets for each of the 3 suggested states is very crisp and graspable which helps the user further understand the trends among various factors between these states over number of years helping him to finally select the one state which suits him the most.

7. Must Have Features

The must have visualizations are:

1. The Information box and table representing value of each parameter a particular state clicked or hovered on map.
2. The horizontal bar chart arranged in descending order of values for the parameter selected from dropdown, which shows the value of parameter for each state as well as compares the value with the average value.
3. The rise/fall in values for each parameter for the most preferred states from past 10 years.

8. Optional Features

- After completion of all the must have features, we can have a frame where the user can input states of his choice and compare them for each of the factors like mortality rate, employment and etc.
- We would like to introduce some sort of story telling in the final visualization where we are using the 10 years data sets for each factor. For instance : the land pricing of Utah from 2006 to 2016 , we see a sudden price hike after 2014 which is mainly attributed to the evolving silicon slopes here. We want to trace some more such events in different visualizations to justify the sudden hike / fall in visualizations for different factors.

9. Project Schedule

Week 1 (30 Oct – 5 Nov)	Preparation of Dataset / Data Cleaning
Week 2 (6 Nov – 10 Nov) – 1st deliverable	Creating map of USA and incorporating dropdown. Making maps react to hover and click events as well as using color scale over the map to show intensity of values.
Week 3 (11 Nov – 19th Nov)	Bar chart for each parameter selected from dropdown. Making the bars react to click event.
Week 4 (20 Nov – 26 Nov)	Incorporation of Weighted mean algorithm for state suggestions
Week 5 (27 Nov – 30 Nov) – Final Project	Incorporating line chart to show trends for 10 years for selected factors, for selected states as well as brush to select the year range to concentrate the visualization on.