ISQS 6339

Bingkun, Ma; Nguyen, Nguyen

Texas Tech University | Spring 2019

Quality of life in California

at county level

Contents

[**Introduction** 2](#_Toc7932669)

[**Results** 3](#_Toc7932670)

[Analysis of data 3](#_Toc7932671)

[Data Cleaning 4](#_Toc7932672)

[Data Merging 4](#_Toc7932673)

[Analysis of Visualizations 5](#_Toc7932674)

[Flow diagram of project. 5](#_Toc7932675)

[Instructions for code 5](#_Toc7932676)

# **Introduction**

The aim of this project is to provide a view of quality of life in California at county level with the most recent data. Quality of life is an abstract concept which there is no method to measure. For this purpose, our group focuses on 4 important aspects of life: Housing-Education-Employment-Safety. Through our analysis of 4 aspects, we expect to provide an overall view about current conditions among California’s counties to assess the quality of life and the project’s final report would be also able to assist policy makers, professionals, researchers, or even individuals with information in different areas, e.g.

* Multivariate analysis of different numerical variables: housing price, crime rate, education level, … to evaluate quality of life statistically.
* The causal relationships among different variables, for example: which factors affect housing price.
* Decisions in real estate, business strategy, ...
* The appropriate governmental policies to enhance residents’ satisfaction.

To assess 4 aspects: Housing-Education-Employment-Safety, we have chosen the following datasets:

* **Housing:** County\_MedianValuePerSqft\_AllHomes

(Source: <https://www.zillow.com/research/data/>)

* **Education:** Education2017

(Source: <https://www.ers.usda.gov/data-products/county-level-data-sets/>)

* **Employment:** UnemploymentReport2017Percentage

(Source: <https://www.ers.usda.gov/data-products/county-level-data-sets/>)

* **Safety:** crime\_rates\_by\_county\_2017\_per100k

(Source: <https://www.ppic.org/data-set/crime-rates-in-california/>)

# **Results**

## Analysis of data

In this section you need to discuss issues related to the data you are

gathering. While you are free to address any issues you see fit, I do ask that you address the

following:

◦ What questions can these data potentially answer?

◦ What are the potential valuable data items exists within the data?

◦ How might they be applied for direct business application and indirect business

applications?

◦ What do you suggest as potential usages for different variables within the dataset?

## Data Cleaning

In this section you need to discuss issues related to the quality of your data.

While you are free to address any issues you see fit, I do ask that you address the following:

◦ What is the overall quality of the data?

◦ What variables contained missing data?

◦ What kinds of missing value exists in the dataset and which variables are they related to?

◦ What methods did you use to clean the missing data?

## Data Merging

In this section you need to discuss the methods you used to merge your

datasets. While you are free to address any issues you see fit, I do ask that you address the

following:

◦ What were the common elements between both datasets?

◦ Were there any issues with multilevel measurement in the final dataset?

◦ What variables are more valuable combined than being in separate datasets?

◦ In what ways has the data become more valuable since being merged? i.e. what new

business insights can be generated due to the combined datasets rather than the sets being

separate.

Our topic is about livability index of counties in California. There are many ways to evaluate livability index. But none of them used only one variable or one aspect. In another word, to evaluate livability, we need combined data with variables that from many different aspects.

For example, U.S. News has evaluated some of the most populous metro areas in the U.S. to help people find the best places for them every year. That is a complex work which need five sub-indexes, and each sub-index is computed with many variables with different weight. U.S. News merged data from various sources, such as the U.S. Census and the Federal Bureau of Investigation's Uniform Crime Reports. The analysis of the merged data can help readers make the most informed decision when choosing where to settle down.

We wanted to do the analysis focused on California and to help people decide which part is better to settle down in California. So, we merged dataset about housing price, job market, household income, crime rate and air quality of each county in California. For the datasets that we mentioned above, all of them are based on county level. The county names are the primary key which is used for data merging. And there is no issue with multilevel measurement in the merged dataset. Merged data will help us describe the quality of life from different aspect and make the livability more convincing.

Data merging not only improved the analysis quality, but also saved coding work and improved respondent experience. Referring to different datasets will require extra code and easily get some mistakes. To save the work for referring to different datasets and detect and correct the mistakes that may occur, merging datasets to one is a good option. The data is run in the memory and stored in the disk. Working one table can save the usage of both hardware. The data merging has better quality and can be used more efficient. The merged data can help people have better understanding of the whole picture.

## Analysis of Visualizations

In this section you need to discuss the methods you used to

visualize your datasets. While you are free to address any issues you see fit, I do ask that you

address the following:

◦ How well does your visualization adhere to the principles and characteristics of a good

visualization?

◦ How well does your visualization adhere to the concept of natural processing? Are there

things in your graph that are necessary but do not have a natural processing correlate?

◦ Copies of your visualizations.

Flow diagram of project. (draw.io would work very well for this).

◦ This should illustrate all steps necessary to gather the data to visualizing the data. It should

also illustrate which “files” and “language” are being used at each step.

Instructions for code – This section should detail how to run your code.

◦ Note, this should be explicit step-by-step instructions, including any variables that might

need to be manipulated.

• Report Quality – It is expected that this should be written as a professional document using

correct grammar and layout.