Mengze Yin

Professor Wright

MA415

10/24/2016

Midterm Project

Dataset Cleaning Using R

1. Introduction

In this assignment, the dataset *“County-level oil and gas production” (Reference:* [www.ers.usda.gov](http://www.ers.usda.gov)*)* is chosen as the raw dataset, which records each county’s FIPS code, county level, including population level, urban influence, size of county, etc., annual oil and gas production from 2000 to 2011, and the categorical group based upon change in the dollar value of the oil and gas production of the county to explore the relationship between population level and oil and gas production. For this project using R, the objectives are: first, cleaning the dataset and giving a tidy dataset with annual production of oil and gas for each county; second, summarizing the dataset and provide some basic statistics that may be used for future exploration on this dataset; finally, visualizing some relationship between county, county level, the production amount and year. The completion of this project is under collaboration of three members: Yilin Li, Mengze Yin and Tianchi Zhang, who had equally contributed to the coding in R.

1. Tidy Dataset

In this part, I aim to reorganize the raw dataset and turn it into a tidy dataset for further analysis. First, read the csv. Table into R and revise columns’ names for readily. Then, when viewing the dataset, I find it consists of two parts: one with the information of the county level, the other one with the production amount for each year. So, I spilt the dataset into two tables. First, I select the information of County from the raw dataset to form a table with county information using function *“select”*, and get a table called “county”. Second, I select information on annual production for each county from the raw dataset and get a table called “oil\_gas0”. Since there are multiple variables stored in columns such as “oil2000” which contains the type of production and year, I firstly *gather* up the non-variable columns and secondly *separate* column headers to get a tidy dataset called “oil\_gas\_tidy”.

1. Dataset Summary

In this part, I aim to provide some basic statistic summary of this dataset for further analysis. The first summary I do is the average annual oil and gas production for each county, which may help to compare the production difference between each county. The second one I do is the national wide mean production (per county) in each year which aim to trace the change of nationwide oil and gas production for each year. Thirdly, I give a tidy table with average oil and gas production for counties by each population level, which may help to explore the relationship between county level and the oil and gas production.

1. Visualization

Based on those summarized table created in part 2, I use *“ggplot”* to visualize three type of relationship discovered in this *“County-level oil and gas production”* dataset. The first on is a bar graph with oil and gas side by side to illustrate how the national mean production varied year to year and compare the differences between oil and gas production.

1. Questions