Midterm Project Summary Report

Wayne Wang
ID: 79612177
Dataset used: County-level Oil and Gas Production in the US
(website: http://www.ers.usda.gov/data-products/county-level-oil-and-gas-production-in-the-us.aspx)

The dataset selected for this project is ñCounty-level Oil and Gas Production in the USò, compiled and published by the Economic Research Service of the United States Department of Agriculture. The dataset is freely available for public use on the web address provided above. The website also provides a good description of the dataset, which is directly quoted below:

> The dataset contains ñ*County-level data from oil and/or natural gas producing States – for onshore production in the lower 48 States only. Most States have production statistics available by county, field, or well, and these data were compiled at the county level to create a database of county-level production, annually for 2000 through 2011. Raw data for natural gas is for gross withdrawals, and oil data almost always include natural gas liquids. Note that State-provided natural gas withdrawals were not available for Illinois or Indiana; those estimates were produced using geocoded wells and State total production reported by the U.S. Department of Energy's Energy Information Agency*ò (source:http://www.ers.usda.gov/data-products/county-level-oil-and-gas-production-in-the-us.aspx).

The dataset has only a handful of variables, most of which are oil (in units of barrels) and natural gas (in units of thousand cubic feet) withdrawals from each county in each year during the period of 2000 ï 2011. A Microsoft Excel file can be found on the website with a detailed description of each variable used in this dataset. The variable description is reproduced below:

Variable Name	**Description and Value Label**
FIPS	Five digit Federal Information Processing Standard (FIPS) code (numeric)
geoid	FIPS code with leading zero (string)
Stabr	State abbreviation (string)
County_Name	County name (string)
Rural_Urban_Continuum_Code_2013	Rural-urban Continuum Code 2013 (see code descriptions)
Urban_Influence_2013	Urban Influence Code, 2013 (see code descriptions)
Metro_Nonmetro_2013	Metro-nonmetro 2013 (0=nonmetro, 1=metro)
Metro_Micro_Noncore_2013	Metro-Micro-Noncore Indicator 2013 (0=nonmetro noncore, 1=nonmetro micropolitan, 2=metropolitan)
oil2000, oil2001, ..., oil2011	Annual gross withdrawals (barrels) of crude oil, for the year specified in the variable name

gas2000, gas2001, ..., gas2011	Annual gross withdrawals (thousand cubic feet) of natural gas, for the year specified in the variable name
oil_change_group	Categorical variable based upon change in the dollar value of oil production, 2000-11. Values are H_Growth (>=$20 million),H_ Decline (<=-$20 million), Status Quo (change between +/- $20 million)
gas_change_group	Categorical variable based upon change in the dollar value of natural gas production, 2000-11. Values are H_Growth (>=$20 million), H_Decline (<=-$20 million), Status Quo (change between +/- $20 million)
oil_gas_change_group	Categorical variable based on the change in the dollar value of the sum of oil and natural gas production, 2000-11. Values are H_Growth (>=$20 million),H_ Decline (<=-$20 million), Status Quo (change between +/- $20 million)

There are a total of 35 variables (35 columns) and 3109 observations (counties) in the raw data. Some of the variables identify the particular geographical areas from which data were obtained, such as *Stabr* and *County_Name*. But since many county names can be found in different States, the variables *FIPS* and *geoid* uniquely identify each county, and thus should be used as the ñidentifierò for each county. This unique identifier will serve as the reference when we merge frames in the data-tidying process. The columns that contain useful information are the columns labeled *oil2000*, *gas2001*, etc., which record the oil and natural gas production in each county. However, this is not tidy because each of these data columns consists of two variables: year and oil(gas). An immediate task is to break up these columns and re-assemble them into just three columns: year, oil, and gas for each county. The strategy to accomplish this is divided into three steps:

1. Break up the entire raw dataset into three smaller datasets:
   i) id ï dataset that houses variables containing geographical information and identifier information, which can be left unchanged for the most part
   ii) oil ï dataset that contains variables related to oil production in each county, these are the variables with names oil2000 ï oil2011
   iii) gas ï dataset that contains variables related to gas production in each county, these are the variables with names gas2000 ï gas2011
2. For the oil and gas data frames, use tidyr tools to separate the year from the oil/gas production data
3. Merge the three datasets into one dataset, which now has the year variable separated from the oil and gas production variables for each observation

The following code is used to divide the raw data into three smaller datasets:

```
varnames <- names(oilgascounty)
id <- oilgascounty[varnames[c(1:8,33:35)]] %>% data.table() %>% setkey(FIPS)
oil <- select(oilgascounty, +FIPS, contains("oil2"))
gas <- select(oilgascounty, +FIPS, contains("gas2"))
```

The dataset ñidò contains columns 1 to 8 and 33 to 35 of the raw data and is made into a data.table, which has the advantage of setting a reference for each observation using the unique identifier FIPS. The data frame ñoilò gathers every column with variable name starting with ñoil2ò in the raw dataset, thereby segregate the oil production information from the rest of the dataset. The unique identifier FIPS is also included as part of the oil dataset. The same was done with the gas columns.

The following code is used to split the year from the oil withdrawal information in the "oil" data frame created in the first step:

```
oil_tidy <- gather(oil, "label", "oil", 2:ncol(oil)) %>%
        separate(label, into = c("delete","year"), sep="l", remove=TRUE) %>%
        select(-delete) %>%
        data.table() %>%
        setkey(FIPS, year)
```

All the columns with the exception of the 1st column, which is FIPS, are stacked on top of one another, which creates a new column that is named "label" containing the name of the original column such as "oil2000". The stacked column is named "oil", since it contains the oil production information. The new "label" variable, which contains the year information, is further divided into the "delete" column and the "year" column by splitting every value of the "label" variable after the letter "l", taking advantage of the fact that every value is uniformly presented in the form oil20XX, where 20XX being the year. Such splitting gives a new "delete" variable that contains "oil" for every observation, and the "year" variable that needs to be extracted. The "delete" column is subsequently deleted from the data frame, and the new oil_tidy dataset is converted into a data.table, which makes setting references easier by the setkey function. The same operations are performed on the gas dataset as well, giving rise to the corresponding gas_tidy dataset.

The following code is used to merge all three datasets: id, oil_tidy, and gas_tidy, which are all data.table, into one dataset called oilgascounty_tidy:

```
oilgascounty_tidy <- id[gas_tidy[oil_tidy] %>% setkey(FIPS)]
```

Since oil_tidy and gas_tidy have previously been sorted according to "FIPS" and "year" variables, it is convenient to use the seemingly subsetting operation: `gas_tidy[oil_tidy]` to merge these two datasets, a unique advantage offered by the data.table package. The merged dataset is assigned a new key using FIPS, and be combined with the id dataset according to FIPS. The final tidy dataset is named oilgascounty_tidy.

To avoid potential problems during the analysis stage, every value in the "oil" and "gas" variables are converted into numeric by replacing every comma in the value with "no space", using the following code:

```
oilgascounty_tidy$oil <- gsub(",", "", oilgascounty_tidy$oil) %>% as.numeric()
oilgascounty_tidy$gas <-  gsub(",", "", oilgascounty_tidy$gas) %>% as.numeric()
```

After obtaining a tidy dataset, a variety of descriptive, statistical explorations can be conducted. Since the variable "year" has been separated from the variable name, a national average trend for oil and gas can be plotted by year (see the attached R markdown for the plots with the ggplot2 package). Note that most counties in the dataset have the value 0 in many of the oil and gas withdrawal record. These values are not omitted in the averaging process, because these values could really be 0, and not NA. An interesting observation is that the oil usage shows a dip between 2005 and 2009, a trend that is not seen in the gas usage. That period also coincides with the recession years due to financial crisis.

With the national trends in mind, the second analysis is to look at the oil and gas usage per year categorized by metro status (the variable "Metro_Nonmetro_2013"). This is an indicator

variable with the value ñ0ò being nonMetro and ñ1ò being Metro (see the attached R markdown for the plots with ggplot2). Some interesting trends observed are:

1. While the oil usage in Metro areas remains relatively constant from 2005 onward, oil usage in nonMetro areas increases exponentially from 2004 onward. This trend can be explained by the heavier reliance on public transportation by people working/living in the Metro areas as well as more and more people moving into the suburban areas after having a family.
2. The gas usage trends between Metro and nonMetro areas, however, are both monotonically increasing. Although the rate of increase in gas usage in nonMetro areas is more or less linear during the study period, the rate of increase in Metro areas are more of an exponential growth starting from 2004.

The last type of descriptive exploration is to investigate the percentage of counties in the US that post a high growth, high decline, or status quo in oil and gas usage during the period of 2000 to 2011. It is not necessary to use the tidy data in this exploration so the original raw dataset is used. The number of counties that post high growth, hi decline, or status quo are individually added up and divided by total number of counties in this dataset. A small contingency table is used to list the percentages of each category (see attached R markdown for the table).