Lab 2: Basic Data Transformation

Anthony Hu

May 28, 2021

Introduction

The task is to explore the US census population estimates by county for 2015 from the package `usmap`. The data frame (`countypop`) has `3142` rows and `4` variables:

- `fips` is the 5-digit FIPS code corresponding to the county;
- `abbr` is the 2-letter state abbreviation; `county` is the full county name;
- `pop_2015` is the 2015 population estimate (in number of people) for the corresponding county.

Each row of the data frame represents a different county or a county equivalent. For the sake of simplicity, when we say a county, that also includes a county equivalent and when we say a state, that also includes the District of Columbia. Answer the following questions.

You will need to modify the code chunks so that the code works within each of chunk (usually this means modifying anything in ALL CAPS). You will also need to modify the code outside the code chunk. When you get the desired result for each step, change `Eval=F` to `Eval=T` and knit the document to HTML to make sure it works. After you complete the lab, you should submit your HTML or PDF file of what you have completed to Sakai before the deadline.

Excercises

```
?countypop
countypop
```

```
## # A tibble: 3,142 x 4
##    fips  abbr  county          pop_2015
##    <chr> <chr> <chr>              <dbl>
##  1 01001 AL    Autauga County     55347
##  2 01003 AL    Baldwin County    203709
##  3 01005 AL    Barbour County     26489
##  4 01007 AL    Bibb County        22583
##  5 01009 AL    Blount County      57673
##  6 01011 AL    Bullock County     10696
##  7 01013 AL    Butler County      20154
##  8 01015 AL    Calhoun County    115620
##  9 01017 AL    Chambers County    34123
## 10 01019 AL    Cherokee County    25859
## # ... with 3,132 more rows
```

Part 1: Length and Unique

a. How many unique 2-letter state abbreviations are there (2 point)? Use `length` and `unique` functions.

```
length(unique(countypop$abbr))
```

```
## [1] 51
```

51 states(including DC)

b. What is the total number of counties in the US (2 point)? Use `length` and `unique` functions.

```
length(unique(countypop$fips))
```

```
## [1] 3142
```

3142 tota number of counties in US

c. How many unique county names are there (2 point)? Use `length` and `unique` functions.

```
length(unique(countypop$county))
```

```
## [1] 1877
```

1877 unique county names

Part 2: Count and Arrange

d. What are the top 10 most common county names (2 points)? `count` number of different county names, `arrange` in descending order and show the first 10 observations.

```
countypop %>%
  count(county) %>%
  arrange(desc(n)) %>%
  head(10)
```

```
## # A tibble: 10 x 2
##    county                n
##    <chr>             <int>
##  1 Washington County    30
##  2 Jefferson County     25
##  3 Franklin County      24
##  4 Jackson County       23
##  5 Lincoln County       23
##  6 Madison County       19
##  7 Clay County          18
##  8 Montgomery County    18
##  9 Marion County        17
## 10 Monroe County        17
```

e. Which state has the smallest number of counties (2 points)? `count` number of observations in each state, `arrange` the data in ascending order and show the first observation.

```
countypop %>%
  count(abbr) %>%
  arrange(n) %>%
  head(1)
```

```
## # A tibble: 1 x 2
##   abbr      n
##   <chr> <int>
## 1 DC        1
```

DC is the state with the smallest number of counties, with 1 county.

Part 3 Group_by and Summarize

f. How many people live in each of the states (2 points)? Group the observation by the variable that serves as state identifier then summarize the data to get total number of people in each state.

```
countypop %>%
  group_by(abbr) %>%
  summarise(total_pop=sum(pop_2015))
```

```
## # A tibble: 51 x 2
##    abbr  total_pop
##    <chr>     <dbl>
##  1 AK       738432
##  2 AL      4858979
##  3 AR      2978204
##  4 AZ      6828065
##  5 CA     39144818
##  6 CO      5456574
##  7 CT      3590886
##  8 DC       672228
##  9 DE       945934
## 10 FL     20271272
## # ... with 41 more rows
```

g. What is the average population of a county in North Carolina (2 points)? `filter` the data to keep observations from 'NC', `summarise` the data to get average population.

```
countypop %>%
  filter(abbr=="NC") %>%
  summarise(mean(pop_2015))
```

```
## # A tibble: 1 x 1
##   `mean(pop_2015)`
##              <dbl>
## 1          100428.
```

Average population of a county in North Carolina is 100428 people.