Measles Data Cleaning

11/3/2017

1. Read the data into R. What class are most of the columns of the data? What issues can you find with the data?

a. Most of the columns of the data are factors. This is an issue since the values in the columns should be values or numeric/integer values since they likely indicate the number of cases of measeles within a given state at a specific point in time.

```
library(dplyr)
measles = read.csv('MEASLES_Cases_1928-2003.csv')
```

2. Change the column names from upper case to lower case.

```
head(colnames(measles))
```

[1] “YEAR” “WEEK” “ALABAMA” “ALASKA” “ARIZONA” “ARKANSAS”

```
colnames(measles) <- tolower(colnames(measles))

head(colnames(measles))
```

[1] “year” “week” “alabama” “alaska” “arizona” “arkansas”

3. Transform the data into long format from wide format.

```
library(tidyr)
measles2 = gather(measles, key = 'state', value = 'cases', alabama:wyoming)

head(measles2)
```

```
##   year week   state cases
## 1 1928    1 alabama    97
## 2 1928    2 alabama   165
## 3 1928    3 alabama   210
## 4 1928    4 alabama   332
## 5 1928    5 alabama   212
## 6 1928    6 alabama   192
```

4. Re-name the missing measles values as NA.

```
library(stringr)
head(which(measles2$cases=="-"))
```

```
## [1]  91  92 104 193 197 201
```

```
measles2$cases = str_replace(measles2$cases, '-','NA')
measles2$cases[91]
```

```
## [1] "NA"
```

5. Change the values for the measles cases from characters to numbers.

```
measles2 = tbl_df(measles2)
measles2$cases = as.numeric(measles2$cases)
```

6. Find the largest number of measles cases for a single week in any state. Which state was it?

a. Kentucky was the state with the largest number of measles cases with 10402 cases in 1954

```
max(measles2$cases, na.rm=T)
```

```
## [1] 10402
```

```
which.max(measles2$cases)
```

```
## [1] 68544
```

```
measles2[68544,]
```

```
## # A tibble: 1 x 4
##    year  week    state cases
##   <int> <int>    <chr> <dbl>
## 1  1954     8 kentucky 10402
```

7. In the step (8) we will want to plot the time series of measles cases. In order to do that, we need the weeks to indicate the time between years. Thus they must be scaled to between 0 and 1, with the first week of the year equaling 0 and the last week of the year coming just before 1.

```
max(measles2$week)
```

```
## [1] 52
```

```
measles2 = mutate(measles2, scale_week = (week-1)/max(week), new_year = year + scale_week)

max(measles2$scale_week)
```

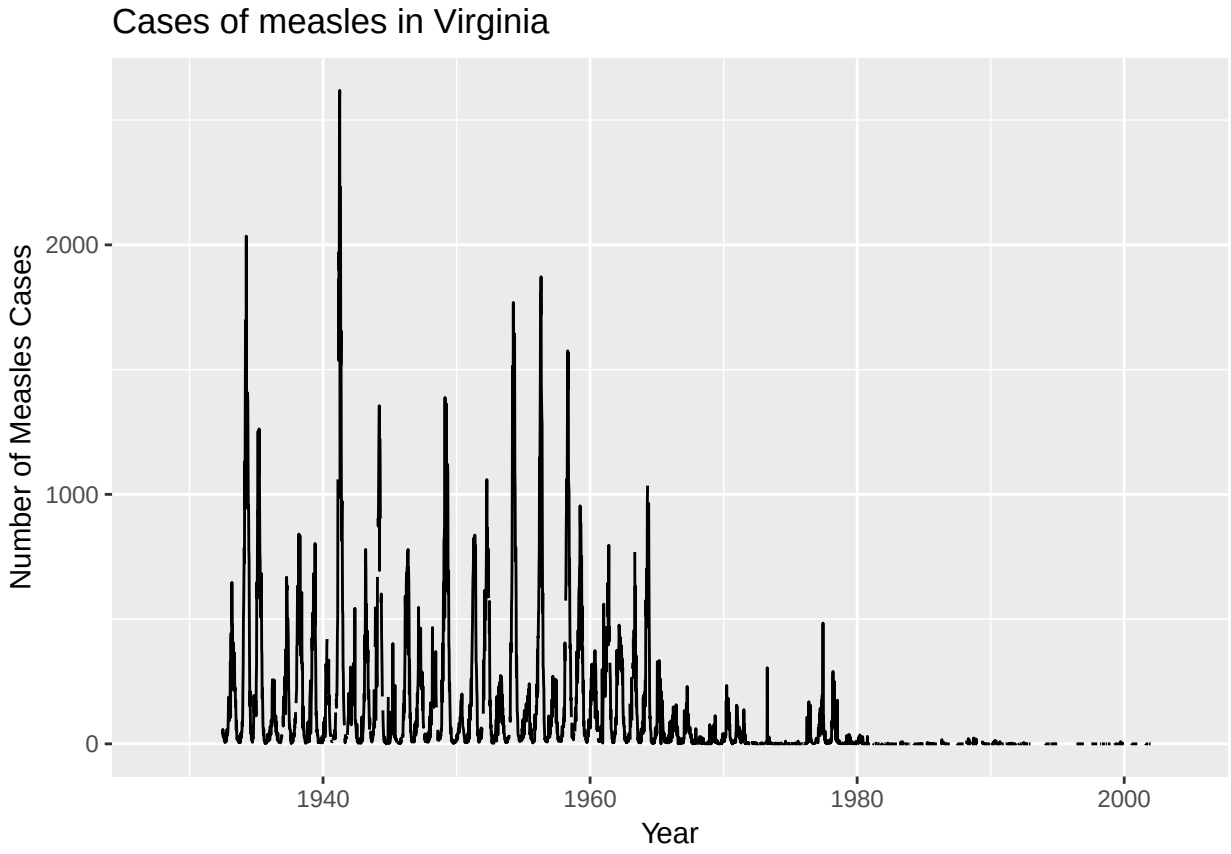
```
## [1] 0.9807692
```

8. Choose one state and make a line plot with the state's measles cases on the y-axis versus year and time-in-year on the x-axis.

```
measles2_va = filter(measles2, state == 'virginia')

library(ggplot2)

ggplot(measles2_va, aes(x = new_year, y = cases)) +
    geom_line() +
    labs(title = 'Cases of measles in Virginia', x = 'Year',y= 'Number of Measles Cases')
```



Bonus 9. Make a table or data frame that shows the number of NA values for each state.

```
measles2 %>%
    group_by(state) %>%
    summarise(missing = sum(is.na(cases)))
```

```
## # A tibble: 51 x 2
##                    state missing
##                    <chr>   <int>
##  1               alabama    1225
##  2                alaska    2593
##  3               arizona     878
##  4              arkansas    1423
##  5            california     288
##  6              colorado     798
##  7           connecticut     717
##  8              delaware    1338
##  9 district.of.columbia    1506
## 10               florida     551
## # ... with 41 more rows
```