Wrangle and Tidy Data

Stat 198 Project

Maya Ghanem and Isabelle Xiong

11/1/2021

Set Up

```
library(readr)
library(tidyverse)
library(readxl)
library(dplyr)
```

Load Data

```
insurance <- read_excel("~/ Stats 198 Project/data/insurance_500_cities.xlsx")
```

```
visits_to_doctor <- read_excel("~/ Stats 198 Project/data/visits_to_doctor.xlsx")
```

```
medicine_high_bp <- read_excel("~/ Stats 198 Project/data/medicine_high_bp.xlsx")
```

```
smoking <- read_excel("~/ Stats 198 Project/data/smoking.xlsx")
```

```
binge_drinking <- read_excel("~/ Stats 198 Project/data/binge_drinking.xlsx")
```

```
physical_activity <- read_excel("~/ Stats 198 Project/data/physical_activity.xlsx")
```

```
heart_disease <- read_excel("~/ Stats 198 Project/data/heart_disease.xlsx")
```

```
diabetes <- read_excel("~/ Stats 198 Project/data/diabetes.xlsx")
```

```
kidney_disease <- read_excel("~/ Stats 198 Project/data/kidney_disease.xlsx")
```

Select, Filter, Summarize, and Rename Data

```
edit_insurance <- insurance %>%
  select(StateAbbr, StateDesc, CityName, Data_Value_Type, Data_Value, GeoLocation) %>%
  filter(Data_Value_Type %in% c("Age-adjusted prevalence")) %>%
  group_by(CityName) %>%
  mutate(mean_value = mean(Data_Value)) %>%
  select(StateAbbr, StateDesc, CityName, mean_value, GeoLocation) %>%
  distinct(CityName, .keep_all = TRUE) %>%
  rename("insurance" = mean_value)
```

```r
edit_visits_to_doctor <- visits_to_doctor %>%
  select(StateAbbr, StateDesc, CityName, Data_Value_Type, Data_Value, GeoLocation) %>%
  filter(Data_Value_Type %in% c("Age-adjusted prevalence")) %>%
  group_by(CityName) %>%
  mutate(mean_value = mean(Data_Value)) %>%
  select(StateAbbr, StateDesc, CityName, mean_value, GeoLocation) %>%
  distinct(CityName, .keep_all = TRUE) %>%
  rename("visits_to_doctor" = mean_value)
```

```r
edit_medicine_high_bp <- medicine_high_bp %>%
  select(StateAbbr, StateDesc, CityName, Data_Value_Type, Data_Value, GeoLocation) %>%
  filter(Data_Value_Type %in% c("Age-adjusted prevalence")) %>%
  group_by(CityName) %>%
  mutate(mean_value = mean(Data_Value)) %>%
  select(StateAbbr, StateDesc, CityName, mean_value, GeoLocation) %>%
  distinct(CityName, .keep_all = TRUE) %>%
  rename("medicine_high_bp" = mean_value)
```

```r
edit_smoking <- smoking %>%
  select(StateAbbr, StateDesc, CityName, Data_Value_Type, Data_Value, GeoLocation) %>%
  filter(Data_Value_Type %in% c("Age-adjusted prevalence")) %>%
  group_by(CityName) %>%
  mutate(mean_value = mean(Data_Value)) %>%
  select(StateAbbr, StateDesc, CityName, mean_value, GeoLocation) %>%
  distinct(CityName, .keep_all = TRUE) %>%
  rename("smoking" = mean_value)
```

```r
edit_binge_drinking <- binge_drinking %>%
  select(StateAbbr, StateDesc, CityName, Data_Value_Type, Data_Value, GeoLocation) %>%
  filter(Data_Value_Type %in% c("Age-adjusted prevalence")) %>%
  group_by(CityName) %>%
  mutate(mean_value = mean(Data_Value)) %>%
  select(StateAbbr, StateDesc, CityName, mean_value, GeoLocation) %>%
  distinct(CityName, .keep_all = TRUE) %>%
  rename("binge_drinking" = mean_value)
```

```r
edit_physical_activity <- physical_activity %>%
  select(StateAbbr, StateDesc, CityName, Data_Value_Type, Data_Value, GeoLocation) %>%
  filter(Data_Value_Type %in% c("Age-adjusted prevalence")) %>%
  group_by(CityName) %>%
  mutate(mean_value = mean(Data_Value)) %>%
  select(StateAbbr, StateDesc, CityName, mean_value, GeoLocation) %>%
  distinct(CityName, .keep_all = TRUE) %>%
  rename("physical_activity" = mean_value)
```

```r
edit_heart_disease <- heart_disease %>%
  select(StateAbbr, StateDesc, CityName, Data_Value_Type, Data_Value, GeoLocation) %>%
  filter(Data_Value_Type %in% c("Age-adjusted prevalence")) %>%
  group_by(CityName) %>%
  mutate(mean_value = mean(Data_Value)) %>%
  select(StateAbbr, StateDesc, CityName, mean_value, GeoLocation) %>%
  distinct(CityName, .keep_all = TRUE) %>%
  rename("heart_disease" = mean_value)
```

```
edit_diabetes <- diabetes %>%
  select(StateAbbr, StateDesc, CityName, Data_Value_Type, Data_Value, GeoLocation) %>%
  filter(Data_Value_Type %in% c("Age-adjusted prevalence")) %>%
  group_by(CityName) %>%
  mutate(mean_value = mean(Data_Value)) %>%
  select(StateAbbr, StateDesc, CityName, mean_value, GeoLocation) %>%
  distinct(CityName, .keep_all = TRUE) %>%
  rename("diabetes" = mean_value)
```

```
edit_kidney_disease <- kidney_disease %>%
  select(StateAbbr, StateDesc, CityName, Data_Value_Type, Data_Value, GeoLocation) %>%
  filter(Data_Value_Type %in% c("Age-adjusted prevalence")) %>%
  group_by(CityName) %>%
  mutate(mean_value = mean(Data_Value)) %>%
  select(StateAbbr, StateDesc, CityName, mean_value, GeoLocation) %>%
  distinct(CityName, .keep_all = TRUE) %>%
  rename("kidney_disease" = mean_value)
```

Join Datasets

```
data_500_cities <- edit_insurance %>%
  left_join(edit_visits_to_doctor) %>%
  left_join(edit_medicine_high_bp) %>%
  left_join(edit_smoking) %>%
  left_join(edit_binge_drinking) %>%
  left_join(edit_physical_activity) %>%
  left_join(edit_heart_disease) %>%
  left_join(edit_diabetes) %>%
  left_join(edit_kidney_disease) %>%
  select(c(StateAbbr, StateDesc, CityName, insurance, visits_to_doctor, medicine_high_bp, smoking, bing
```

Glimpse Final Dataset

```
glimpse(data_500_cities)
```

```
## Rows: 475
## Columns: 13
## Groups: CityName [475]
## $ StateAbbr         <chr> "US", "MO", "ND", "OH", "OK", "TN", "AL", "AL", "AZ"~
## $ StateDesc         <chr> "United States", "Missouri", "North Dakota", "Ohio",~
## $ CityName          <chr> NA, "Springfield", "Fargo", "Columbus", "Broken Arro~
## $ insurance         <dbl> 15.20000, 14.66667, 9.80000, 17.00000, 13.50000, 17.~
## $ visits_to_doctor  <dbl> 69.0, NA, 61.5, NA, 66.6, 73.0, 74.9, 70.5, NA, 66.9~
## $ medicine_high_bp  <dbl> 56.20, NA, 59.20, NA, 59.30, 62.80, 67.60, 62.70, 54~
## $ smoking           <dbl> 16.70, NA, 18.50, NA, 17.70, 19.70, 22.20, 19.00, NA~
## $ binge_drinking    <dbl> 18.1, NA, 23.2, NA, 16.4, 14.5, 11.5, 13.4, NA, 19.0~
## $ physical_activity <dbl> 26.20, NA, 25.80, NA, 28.70, 28.60, 38.10, 30.70, NA~
## $ heart_disease     <dbl> 5.60, NA, 5.40, NA, 5.50, 6.10, 7.30, 6.40, NA, 4.30~
## $ diabetes          <dbl> 9.6, NA, 8.4, NA, 9.0, 11.4, 16.2, 11.4, NA, 7.2, 6.~
## $ kidney_disease    <dbl> 2.90, NA, 2.70, NA, 2.70, 3.10, 4.00, 3.10, 3.25, 2.~
## $ GeoLocation       <chr> NA, "(37.1942661484, -93.2914273656)", "(46.86524578~
```