Final Project - Health Expenditure

Yohan Kim, Ming Qiu, Kevin Lam, Khiem Pham

Introduction

Healthcare policy refers to the laws for managing a nation's healthcare system.[1] The US implements a third-party payer system, in which the health insurance (third parties) reimburse most of the cost of healthcare services for patients to the hospitals that provide the services.[1] The US uses a mix of public and private insurances. The main public programs include Medicaid (for low-income/disabled individuals) and Medicare (people over 65 or people and people with certain disabilities).[1] The rest of Americans use private insurance through their employer.[1]

Healthcare policy in the United States has become complex over the years. Following Obama's signing of the Affordable Care Act (law to make health insurance coverage accessible to more Americans by lowering healthcare spending and costs) in 2010, it's been met with a multitude of legal challenges (https://ballotpedia.org/Obamacare_lawsuits), among them being based on religious values or state policies.[1] Additionally, skeptics assert that the Affordable Care Act will increase the costs but lower the quality of the healthcare being provided.[1]

It's important to investigate the relationship between the health conditions of Americans and the policies that are enacted to determine if they're actually beneficial to the nation. To do this, we should first get acquainted with the healthcare economics in the United states.[2] For this case study, we will analyze the relationship between how much is being spent on healthcare (healthcare expenditure) and healthcare coverage and how it changes in certain years, and how the expenditure varies from different regions.

Load packages

```
library(OCSdata)
library(tidyverse)
library(pdftools)
library(tesseract)
library(magick)
library(stringr)
library(ggrepel)
library('Kendall')
library('tidymodels')
```

Questions

1. Is there a relationship between healthcare coverage and healthcare spending in the United States?
2. How does the spending distribution change across geographic regions in the United States?
3. Does the relationship between healthcare coverage and healthcare spending in the United States change from 2013 to 2014?

The Data

Dataset comes from the Henry J Kaiser Family Foundation, an organization dedicated to providing information regarding national health issues.

- https://www.kff.org/other/state-indicator/health-care-expenditures-by-state-of-residence-in-millions/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D (https://www.kff.org/other/state-indicator/health-care-expenditures-by-state-of-residence-in-millions/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D)
- https://www.kff.org/other/state-indicator/total-population/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D (https://www.kff.org/other/state-indicator/total-population/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D)

We manually downloaded dataset from their website (by each year), renamed to distinguish which year this dataset came from.

The first dataset is each state's health care expenditures. It has 2 variables: Location and Total Health spending. The observations are all the states in America.

The second dataset is the health insurance coverage of the total population for each state in America. It has 8 variables: Location, Employer, Non-group, Medicaid, Medicate, Military, Uninsured, and Total. The observations are all the states in America.

Data Import

```
# read in CSVs
coverage_data <- list.files("data/raw/coverage/", pattern="*.csv", full.names = TRUE) |>
  map(~ read_csv(., skip=2))

# Get Names
coverage_data_names <- list.files("data/raw/coverage/", pattern="*.csv") |>
  str_extract("raw_data_200[8-9]|raw_data_201[0-9]")
# Apply names
names(coverage_data) <- coverage_data_names
```

Above code imports the health coverage dataset from 2008 to 2019

```
# read in CSVs
expenditure_data <- list.files("data/raw/expenditure/", pattern="*.csv", full.names = TRUE) |>
  map(~ read_csv(., skip=2))

# Get Names
expenditure_data_names <- list.files("data/raw/expenditure/", pattern="*.csv") |>
  str_extract("raw_data_200[8-9]|raw_data_201[0-4]")

# Apply names
names(expenditure_data) <- expenditure_data_names
```

Above code imports the health care spending by state from 2008 to 2014.

Data Wrangling

```
remove_invalid <- function(dataset) {
  dataset <- subset(dataset, Location!="Puerto Rico" , select=-Footnotes) |> # Remove unnecessary column
    drop_na("Total")   # Remove notes and references
}

# Apply function
coverage_data <- map(coverage_data, remove_invalid)
# Test to see if worked properly
coverage_data[["raw_data_2019"]]
```

```
## # A tibble: 52 × 8
##    Location     Employer `Non-Group` Medicaid Medicare Military Uninsured  Total
##    <chr>           <dbl>       <dbl>    <dbl>    <dbl>    <dbl>     <dbl>  <dbl>
##  1 United Sta… 158000000    18728800 63145700 45286700  4393600  29349300 3.19e8
##  2 Alabama       2250900      263400   929500   763800    99000    460400 4.77e6
##  3 Alaska         339800       24700   149400    70200    37100     80500 7.02e5
##  4 Arizona       3202000      366500  1489600  1145300   105600    789100 7.10e6
##  5 Arkansas      1226300      157300   767000   464200    41800    265800 2.92e6
##  6 California   18538700     2569600  9790000  4388900   350200   3005400 3.86e7
##  7 Colorado      2997100      389900   942300   718500   127200    436700 5.61e6
##  8 Connecticut   1827200      165900   743900   488000    23900    204500 3.45e6
##  9 Delaware       467700       38900   191400   162500    17300     62500 9.40e5
## 10 District o…    368400       43400   171000    55300     9000     24200 6.71e5
## # … with 42 more rows
```

Above code creates a function called **remove_invalid()** to remove NA row values, as well as remove row that has Location - Puerto Rico, which is outside U.S. We then apply this function to all Health Care Coverage list. After that, we then group all coverage data into one single vector to use.

```
remove_invalid_expenditure <- function(dataset) {
  dataset <- dataset |>
    rename(Spending=`Total Health Spending`) |>
    drop_na(Spending)
  dataset$Spending <- substring(dataset$Spending, 2)
  dataset <- dataset |>
    mutate(Spending=as.numeric(Spending)*1000000) # As title of this dataset said, it is in millions of dollars
}

# Apply function
expenditure_data <- map(expenditure_data, remove_invalid_expenditure)
# Test to see if worked properly
expenditure_data[["raw_data_2014"]]
```

```
## # A tibble: 52 × 2
##    Location                 Spending
##    <chr>                       <dbl>
##  1 United States        2562824000000
##  2 Alabama                35263000000
##  3 Alaska                  8151000000
##  4 Arizona                43356000000
##  5 Arkansas               21980000000
##  6 California            291989000000
##  7 Colorado               36398000000
##  8 Connecticut            35413000000
##  9 Delaware                9587000000
## 10 District of Columbia    7871000000
## # … with 42 more rows
```

Above code creates a function called **remove_invalid_expenditure()** to remove NA row values, as well as change the column type to numeric, to be able to use it when graphing, etc. Notice that we multiply spending by 1 million, since dataset website said it is in unit of million dollar. We then apply this function to all Health Care Coverage list. We then group all expenditure data into one vector to use.

```
coverage_data <- coverage_data |>
  map_df(bind_rows, .id="Year") |>
  mutate(Year=as.numeric(str_remove(Year, "raw_data_")))

expenditure_data <- expenditure_data |>
  map_df(bind_rows, .id="Year") |>
  mutate(Year=as.numeric(str_remove(Year, "raw_data_")))

health_care <- inner_join(coverage_data, expenditure_data, by=c("Year", "Location")) # to automatically avoid 201
5-2019 data that does not have expenditure spending
health_care
```

```
## # A tibble: 364 × 10
##     Year Location      Employer `Non-Group` Medicaid Medicare Military Uninsured
##    <dbl> <chr>            <dbl>       <dbl>    <dbl>    <dbl>    <dbl>     <dbl>
##  1  2008 United States   1.58e8    16956300 39474800 32115800  4111400  44759100
##  2  2008 Alabama         2.35e6      245000   646300   578200    78700    632500
##  3  2008 Alaska          3.39e5       23600    74600    38500    43000    136000
##  4  2008 Arizona         2.96e6      390000   974300   752300    99300   1191900
##  5  2008 Arkansas        1.21e6      150800   495800   372900    44200    495500
##  6  2008 California      1.79e7     2630300  5468100  3147800   356300   6394600
##  7  2008 Colorado        2.64e6      394500   418400   433500   107300    819700
##  8  2008 Connecticut     2.13e6      163000   394600   379800    19000    301400
##  9  2008 Delaware        4.83e5       34800   116900   103500    14800     90800
## 10  2008 District of …   3.04e5       34500   130000    41000     4500     45000
## # … with 354 more rows, and 2 more variables: Total <dbl>, Spending <dbl>
```

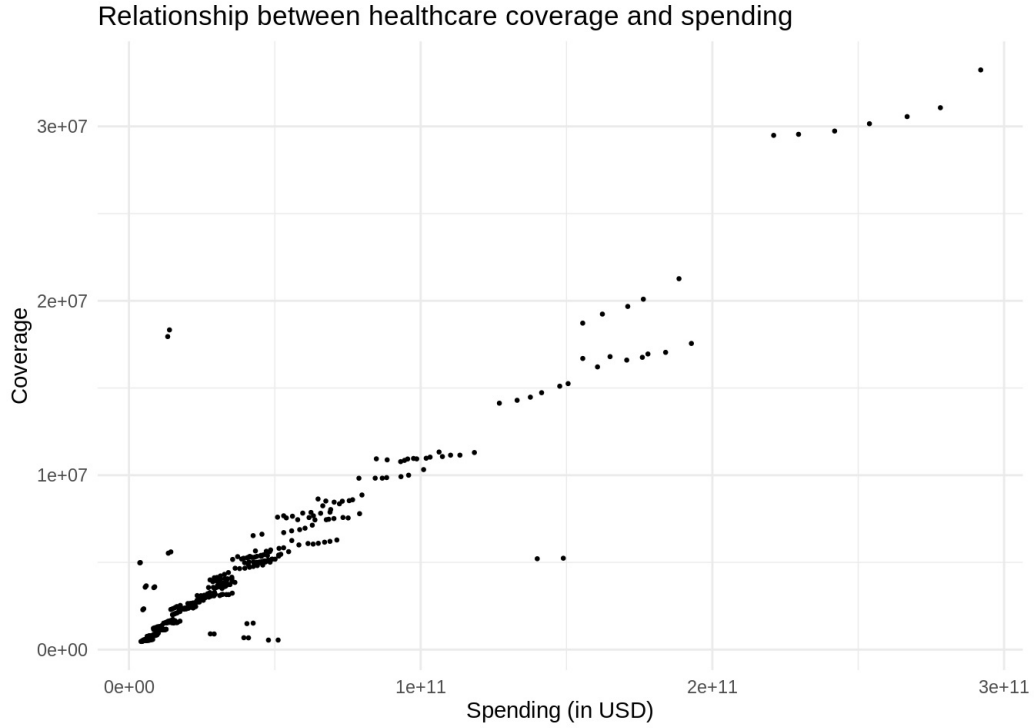
Above code first combine each expenditure / coverage dataset into one whole dataset, with creating Year column as to give distinction between each row. We then combine expenditure / coverage dataset into one whole dataset, and we use join (similar to merge() function in basic R) to conditionally combine if their Year and Location column matches together.

Analysis

Exploratory Data Analysis

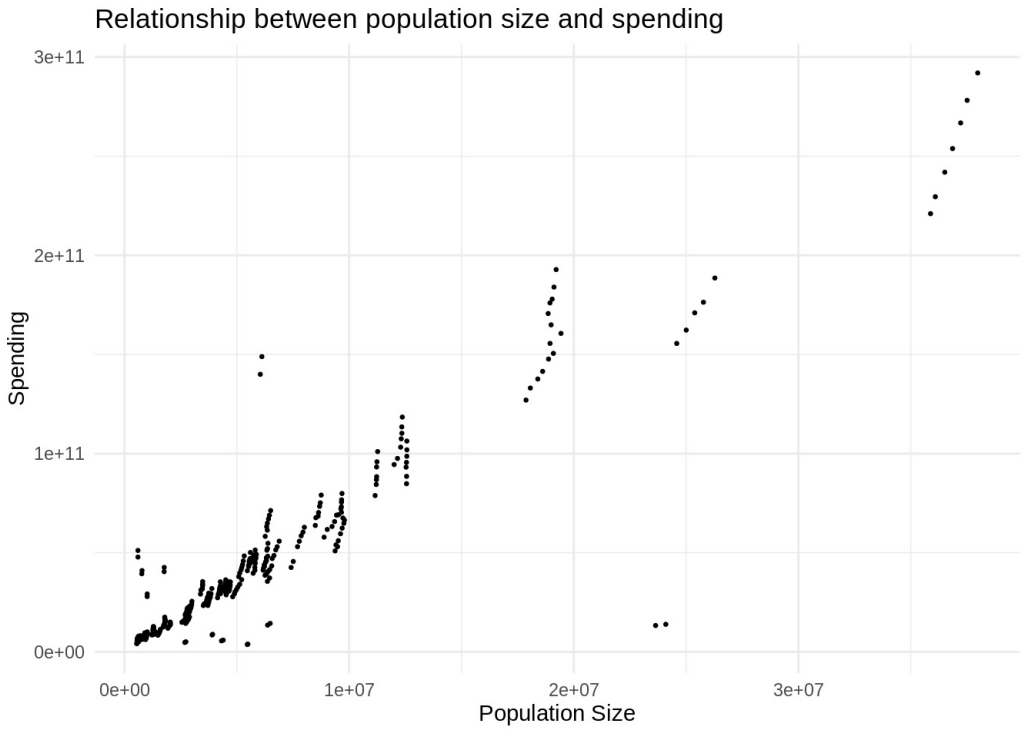
Q1: Is there a relationship between healthcare coverage and healthcare spending in the United States?

```
health_care |>
  filter(Location!="United States") |>
  ggplot(aes(x=Spending, y=Total-Uninsured)) + geom_point(size=0.5) + labs(title="Relationship between healthcare
coverage and spending", x="Spending (in USD)", y="Coverage") + theme_minimal()
```



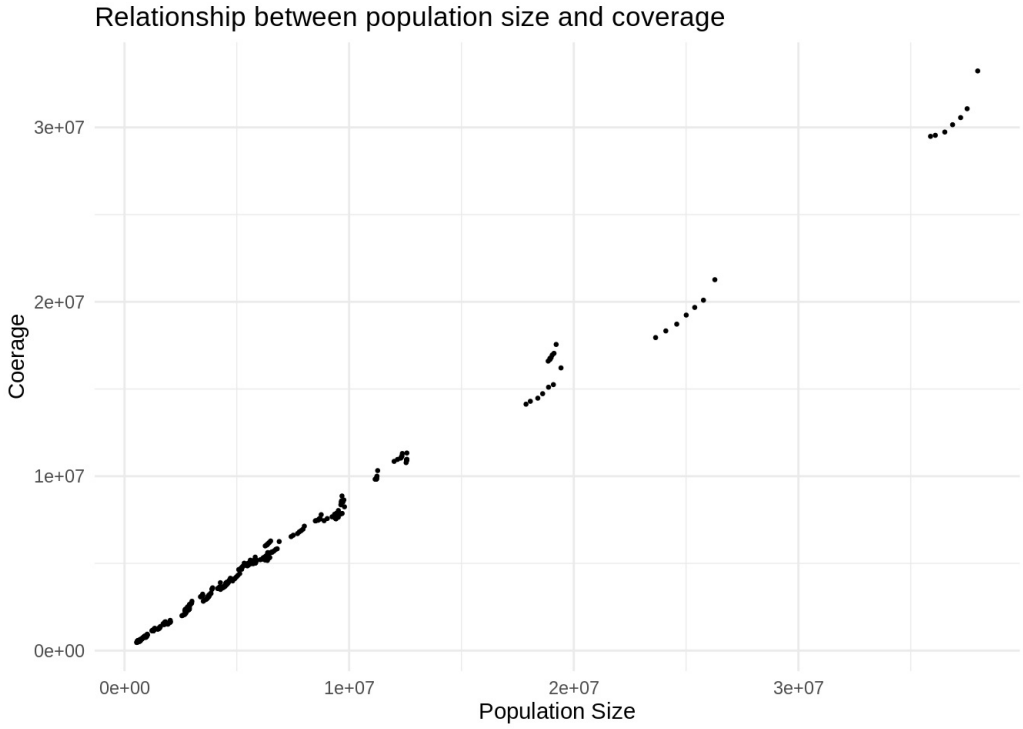
The above graph shows the relationship of spending and coverage. To get the number of people covered, we subtract Uninsured from Total. We can see that lower spending and coverage has more observations, and that there's a strong direct relationship, in that as spending increases, coverage increases. However, population size could have an effect on this relationship. Let's see the relationship between coverage and spending on population size.

```
health_care |>
  filter(Location!="United States") |>
  ggplot(aes(x=Total, y=Spending)) + geom_point(size=0.5) + labs(title="Relationship between population size and
spending", x="Population Size", y="Spending") + theme_minimal()
```



The above graph shows the relationship between population size and spending for each state. Similar to the relationship between coverage and spending, most observations are on the bottom left and it has a strong direct relationship.

```
health_care |>
  filter(Location!="United States") |>
  ggplot(aes(x=Total, y=Total-Uninsured)) + geom_point(size=0.5) + labs(title="Relationship between population si
ze and coverage", x="Population Size", y="Coerage") + theme_minimal()
```



The above graph shows the relationship between population size and coverage for each state. Similar to the relationship between coverage and spending and population size and coverage, it has a strong direct relationship.

Since coverage and spending are both strongly directly related to population size, we need to account for the population size when we compare healthcare coverage and spending. To do this, we first create another variable that represents the proportion of people covered and total population.

```
health_care_prop_coverage <- health_care |>
  mutate(prop_coverage = (Total-Uninsured)/Total)

health_care_prop_coverage
```

```
## # A tibble: 364 × 11
##     Year Location      Employer `Non-Group` Medicaid Medicare Military Uninsured
##    <dbl> <chr>            <dbl>       <dbl>    <dbl>    <dbl>    <dbl>     <dbl>
##  1  2008 United States   1.58e8    16956300 39474800 32115800  4111400  44759100
##  2  2008 Alabama         2.35e6      245000   646300   578200    78700    632500
##  3  2008 Alaska          3.39e5       23600    74600    38500    43000    136000
##  4  2008 Arizona         2.96e6      390000   974300   752300    99300   1191900
##  5  2008 Arkansas        1.21e6      150800   495800   372900    44200    495500
##  6  2008 California      1.79e7     2630300  5468100  3147800   356300   6394600
##  7  2008 Colorado        2.64e6      394500   418400   433500   107300    819700
##  8  2008 Connecticut     2.13e6      163000   394600   379800    19000    301400
##  9  2008 Delaware        4.83e5       34800   116900   103500    14800     90800
## 10  2008 District of …   3.04e5       34500   130000    41000     4500     45000
## # … with 354 more rows, and 3 more variables: Total <dbl>, Spending <dbl>,
## #   prop_coverage <dbl>
```

Next, we create another variable that represents the proportion of total spend and total population.

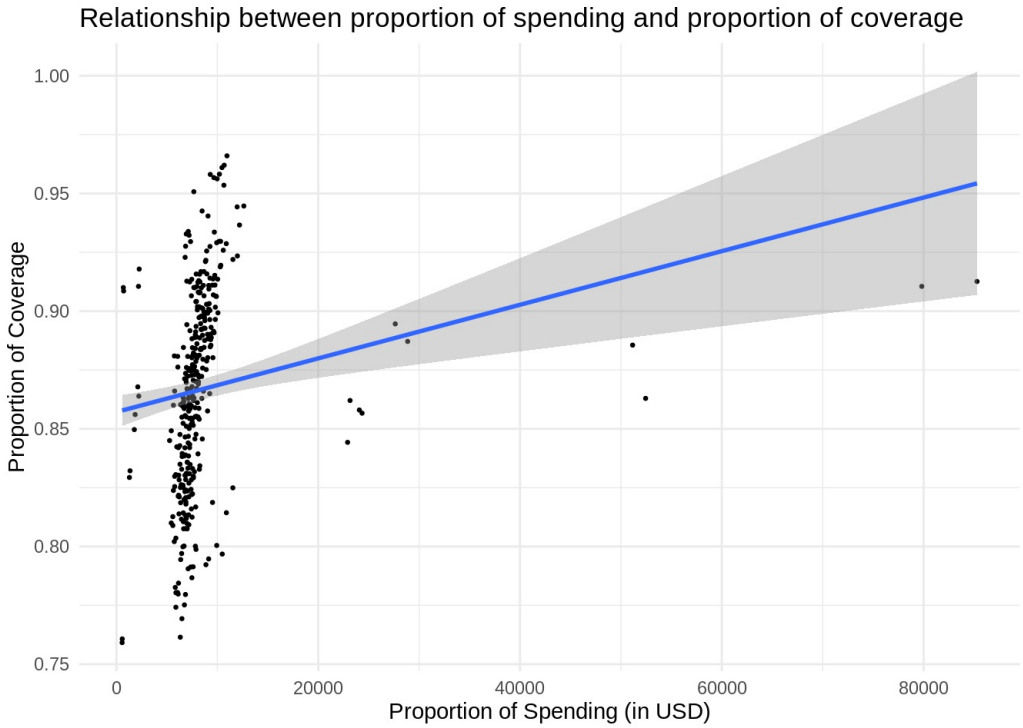
```
health_care_prop_coverage_prop_spending <- health_care_prop_coverage |>
  filter(Location!="United States") |>
  mutate(prop_spending = Spending/Total)

health_care_prop_coverage_prop_spending
```

```
## # A tibble: 357 × 12
##     Year Location      Employer `Non-Group` Medicaid Medicare Military Uninsured
##    <dbl> <chr>            <dbl>       <dbl>    <dbl>    <dbl>    <dbl>     <dbl>
##  1  2008 Alabama        2346200      245000   646300   578200    78700    632500
##  2  2008 Alaska          339000       23600    74600    38500    43000    136000
##  3  2008 Arizona        2955200      390000   974300   752300    99300   1191900
##  4  2008 Arkansas       1210500      150800   495800   372900    44200    495500
##  5  2008 California    17884900     2630300  5468100  3147800   356300   6394600
##  6  2008 Colorado       2643300      394500   418400   433500   107300    819700
##  7  2008 Connecticut    2126000      163000   394600   379800    19000    301400
##  8  2008 Delaware        483400       34800   116900   103500    14800     90800
##  9  2008 District of …   303800       34500   130000    41000     4500     45000
## 10  2008 Florida        8039600     1147900  1960100  2666200   316600   3743900
## # … with 347 more rows, and 4 more variables: Total <dbl>, Spending <dbl>,
## #   prop_coverage <dbl>, prop_spending <dbl>
```

After accounting for population size, our plot should now be more accurate. We view the relationship between prop_spending and prop_coverage.

```
health_care_prop_coverage_prop_spending |>
  ggplot(aes(x=prop_spending, y=prop_coverage)) + geom_point(size=0.5) + labs(title="Relationship between proport
ion of spending and proportion of coverage", x="Proportion of Spending (in USD)", y="Proportion of Coverage") + t
heme_minimal() + geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Compared to the other graphs, the relationship is a little more mixed in the above graph. From 0-20000 for proportion of spending, there's a dramatic increase in coverage. However, a little past 20000, the coverage goes down. However, from 20000 onwards, it follows a gradually increasing trend. We can conclude that there's a direct relationship between health care spending and coverage in the United States.

Q2: How does the spending distribution change across geographic regions in the United States?

To have an exploration on the difference of average healthcare spending from 2008 to 2014 between each regions in the United States, we plan to make a state heatmap first.

```
# calculate the sum of spending from 2008 to 2014
H <- new.env(hash = TRUE)
for(i in 1:nrow(health_care)) {
  state <- as.character(health_care[i,"Location"]) # get names for key
  H[[state]] <-0 # initialize value for hash map
}
for(i in 1:nrow(health_care)) {
  state <- as.character(health_care[i,"Location"])
  value <- as.numeric(health_care[i,"Spending"])
  H[[state]] <- H[[state]] + value
}
```

In order to get the average spending of each state/region, we first need to get the total spending. Above code uses hash map to find the cumulative spending value for each state/region.

```
# get state names
state_list <- unique(health_care$Location)
state_list <- as.vector(state_list)
state_list <- state_list[-1] # remove "United States"
```

```
# get values from Hash map
ave_value <- c()
for (i in state_list){
    ave_value <- c(ave_value, H[[i]]/7) # sum/7 = average spending
}
```

```
# combine state_list and ave_value
ave_spend_df <- data.frame(state_list, ave_value)
ave_spend_df <- ave_spend_df |>
    rename('region' = state_list) # rename state_list to prepare later merging
```

```
library(maps)
```

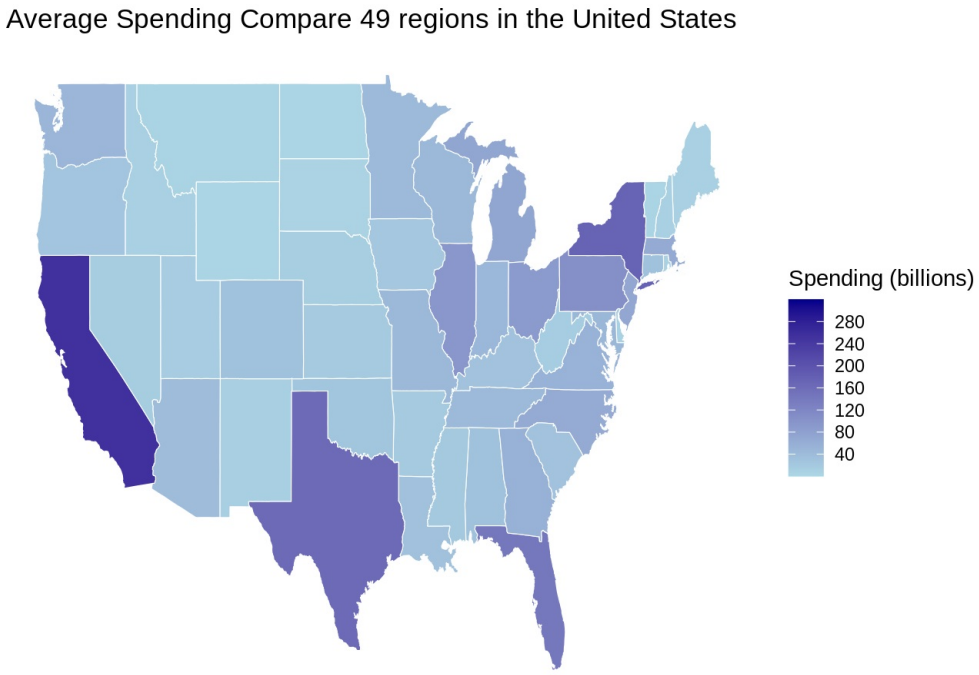
```
##
## Attaching package: 'maps'
```

```
## The following object is masked from 'package:purrr':
##
##     map
```

```
# merge us_state and ave_spend_df
us_states <- map_data("state") # get state map data
us_states$region <- str_to_title(us_states$region, locale = "en") # Capitalize the State names
state_map_aveSpend_df <- inner_join(us_states, ave_spend_df, by = "region", copy = TRUE)
```

Above 2 chunks create a dataframe that match states/regions' names with their corresponding average spending values and combine this dataframe with the data that forms the base for state map.

```
# plot
p2 <- ggplot() +
  geom_polygon( data=state_map_aveSpend_df,
          aes(x=long, y=lat, group=group, fill = ave_value/1000000000), # show in billions
          color="white", size = 0.2) +
  scale_fill_continuous(name="Spending (billions)",
            low = "lightblue", high = "darkblue",limits = c(0,320),
            breaks=c(40,80,120,160,200,240,280), na.value = "grey50") +
    labs( x = "", y = "",
    title=" Average Spending Compare 49 regions in the United States")+
  scale_x_continuous(breaks = NULL) +
  scale_y_continuous(breaks = NULL) +
  theme(panel.background = element_blank())
p2
```

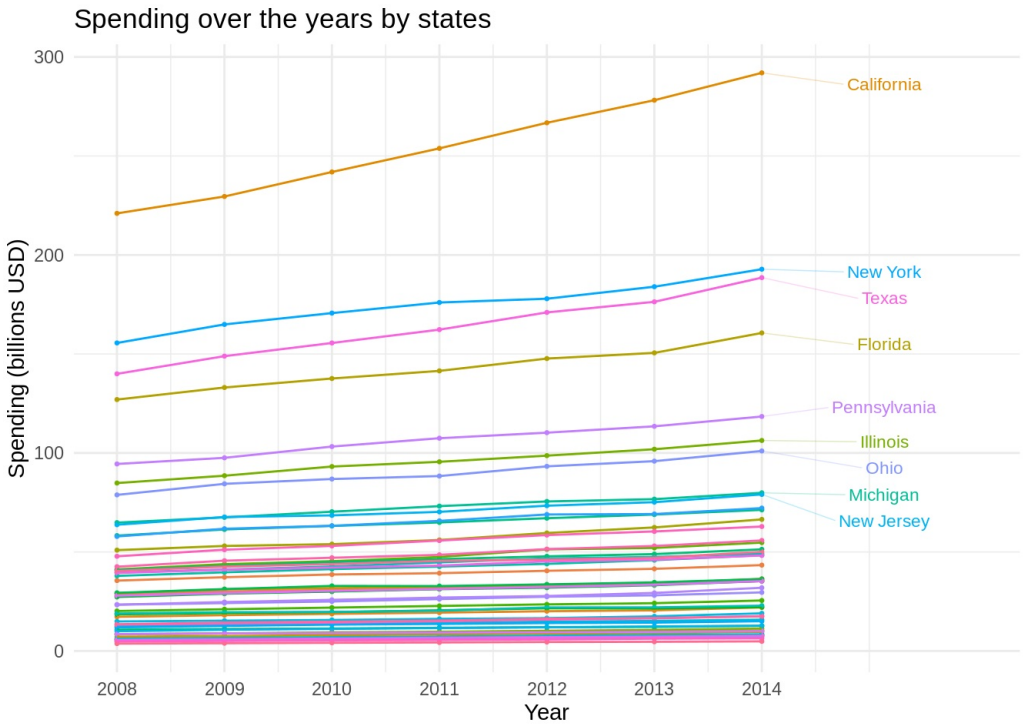


Above graph reports the Average Spending of each region in the mainland United States from 2008 to 2014. We use the shades of color to represent the change of average spending. The darker the color, the greater amount of the average spending is for a state. Since the maps package only include data to show the mainland, we fail to include the state of Alaska and Hawaii in above map.

As we can see, majority of regions has average spending less than 120 billion dollars. The state of California has the darkest color which shows its largest average spending among different state and region. The state of Taxes, New York and Florida also have fairly large average spending (> 160 billions).

```
health_care |>
  filter(Location != 'United States') |>
  ggplot(aes(x = Year, y = Spending/1000000000, color = Location)) +
  geom_point(size = 0.5,show.legend = FALSE) +
  geom_line(aes(group = Location),
    size = 0.5,
    show.legend = FALSE) +
  labs(
    title = "Spending over the years by states",
    x = "Year", y = "Spending (billions USD)") +
  theme_minimal() +
  geom_text_repel(data = health_care |>
      filter(Location != 'United States') |>
      filter(Year == last(Year)),
      aes(label = Location,x = Year, y = Spending/1000000000),
      size = 3, alpha = 1, nudge_x = 1, direction = "y",
      hjust = 1, vjust = 1, segment.size = 0.25, segment.alpha = 0.25,
      force = 1, max.iter = 9999,max.overlaps = 3, show.legend = FALSE) +
  scale_x_continuous(
     breaks = seq(2008, 2014, by = 1),
    limits = c(2008, 2016),
    labels = c(seq(2008, 2014, by = 1))
  )
```

```
## Warning: ggrepel: 42 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



Above graph shows the change of Spending in billion dollars from 2008 to 2014 for each region in the United States.

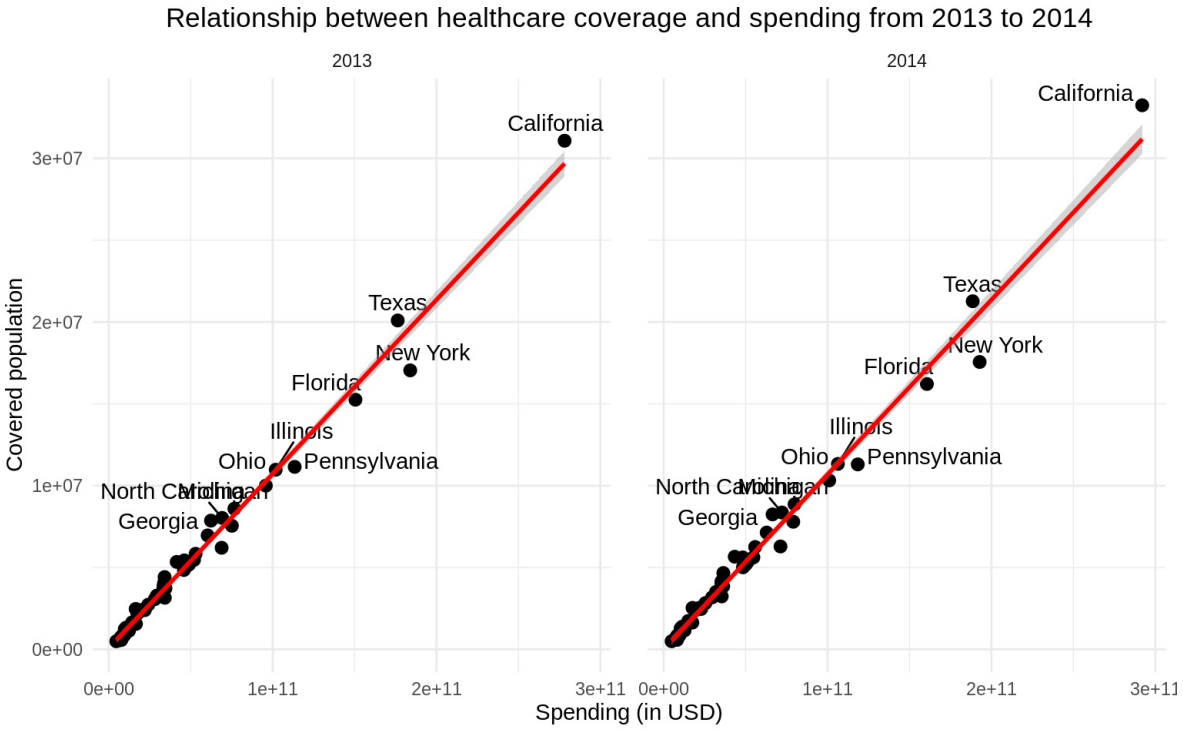
Based on the graph, we learn that all regions seem to have an increase trend of spending over the years. We also noticed that California, New York, Texas, Florida and Pennsylvania are the 5 states with the highest spending, while California has the highest spending overall and most rapid increase among those 5 regions, and the rest of most states have a spending of less than 100 billions over the years.

Q3: Does the relationship between healthcare coverage and healthcare spending in the United States change from 2013 to 2014?

```
health_care |>
  # Want to show only 2013 and 2014 relationship
  filter(Year == 2013 | Year == 2014, Location != "United States") |>
  ggplot(aes(x = Spending, y = Total - Uninsured)) +
  geom_point(size = 2.5) +
  # Makes the labels for the points easier to read
  geom_text_repel(aes(label = Location), nudge_y = 560000) +
  labs(
    title = "Relationship between healthcare coverage and spending from 2013 to 2014",
    x = "Spending (in USD)",
    y = "Covered population") +
  # Add line to connect points
  geom_smooth(method = "lm", col = "red") +
  facet_wrap(~ Year) +
  theme_minimal() +
  theme(panel.spacing = unit(1.25, "lines")) +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: ggrepel: 41 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps

## Warning: ggrepel: 41 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



The graphs above show that from 2013 to 2014, there is an increase in the amount of USD spent for healthcare, and the healthcare coverage in the United States also increased. From there we can deduce a direct relationship between spending and coverage in both 2013 and 2014. If the relationship between the two years changed, we would have seen an indirect relationship where more spending would lead to less coverage.

Data Analysis

Regarding to the question 1 (Is there a relationship between healthcare coverage and healthcare spending in the United States?), we want to see how each different health insurance affect the spending.

```
health_care_exclude_us <- health_care|>
  filter(Location!="United States")

linear_reg() |>
  set_engine("lm") |>
  fit(Spending ~ Medicaid + Military + Employer + Medicare + `Non-Group`, data=health_care_exclude_us)
```

```
## parsnip model object
##
## Fit time:  4ms
##
## Call:
## stats::lm(formula = Spending ~ Medicaid + Military + Employer +
##     Medicare + `Non-Group`, data = data)
##
## Coefficients:
## (Intercept)     Medicaid     Military     Employer     Medicare  `Non-Group`
##  2531486261        16814       -62778         2946        34340        -6086
```

Above code is using a linear regression model to see the trends for each healthcare population relative to the spending. Notice that we have excluded the rows that are United states because rows with the location “United States” are just a summation of all states (with Spending, Healthcare population, etc)

Slope: - for every increase count of people with Medicaid, the healthcare spending increases by 16814 dollars on average. - for every increase count of people with Military, the healthcare spending decreases by 62778 dollars on average. - for every increase count of people with Employer Healthcare, the healthcare spending increases by 2946 dollars on average. - for every increase count of people with Medicare coverage, the healthcare spending increases by 34340 dollars on average. - for every increase count of people with individual health coverage, the healthcare spending decreases by -6086 dollars on average.

Intercept: The healthcare spending for groups that do not have any health coverage are expected, on average, to be 2531486261 dollars.

Equation: Total_Expenditure = 2531486261 + (16814 * # of Medicaid Users) + (-62778 * # of Military Healthcare Users) + (2946 * # of Employer Healthcare Users) + (34340 * # of Medicare Healthcare Users) + (-6086 * # of Non-Group Healthcare Users)

Above is the linear regression model for each health coverage, compared to the total spending. Some remarks we can see is that as , the number of people who have Military and Non-Group healthcare population increase, healthcare spending (in dollars) decreases, whereas healthcare spending (in dollars) increases when Medicaid, Medicare, and Employer healthcare population increase.

Results and discussion of Results

Here are our questions and answers for this project:

1. Is there a relationship between healthcare coverage and healthcare spending in the United States?

- There is a direct relationship between healthcare coverage and healthcare spending. Based on our data analysis, we found that some coverages have more negative relationship to the health spending, such as Military and Non-Group, while the relationship between spending and coverages including Employer Healthcare, Medicaid, Medicare is more positive.

2. How does the spending distribution change across geographic regions in the United States?

- Spending in all states seems to have a increase trend over the years. In particular, we found that California has the fastest increase in Spending and greatest spending overall. From the Mainland United States heatmap, we can see that states like California, Taxes, New York and Florida have the greatest average Spending from 2008 to 2014, while most but those 4 states seem to have less than 120 billion dollars of average spending. This suggested that healthcare spending varies from region to region in the United States.

3. Does the relationship between healthcare coverage and healthcare spending in the United States change from 2013 to 2014?

- We see the same direct relationship of healthcare coverage and healthcare spending from both 2013 and 2014, with just an increase in the values of spending and coverage in 2014 compared to 2013.

Conclusion

From this project, we are able to utilize Healthcare Coverage relative to Healthcare Expenditure throughout the years. We were able to answer several questions like finding the relationship between Healthcare Expenditure and Coverage, as well as how each healthcare expenditure distribution differ across regions, and more. Throughtout the project, we faced some limitations. For example, our dataset lacked details of some columns. For instance, we did not have any specific amount expenditure goes towards each healthcare providers like “Medicaid” or “Medicare” to see how much funding effectively goes to each person. Not only that, we could not find a better heatmap that includes all states with Hawaii and Alaska. Not only that, outside of this dataset at all, it would be nicer to have how each healthcare’s plans are and how much does it cost to specifically view the differences between each healthcare. In all, though we lacked on some details in the dataset, we have seen a positive relationship between healthcare expenditure and coverage that stayed consistent throughout the years.

References

1. https://ballotpedia.org/Healthcare_policy_in_the_United_States (https://ballotpedia.org/Healthcare_policy_in_the_United_States) - Providing background information for introduction
2. https://www.opencasestudies.org/ocs-healthexpenditure/ (https://www.opencasestudies.org/ocs-healthexpenditure/) - Guidance for project