

# NCQA Health Insurance Plan Ratings 2018-2019



Sharad Nirmalkumar Bajpai

DA 5020 | Fall 2019



# Project Overview

For my final project, I chose to prepare a prototype of a database that stores three different types of insurance related data. I extracted the health plan data from the NCQA website which publishes health plan reports every year to observe the quality of health plans in the country. I have scraped the data from the webpage using import.io and downloaded the CSV (Comma Separated Value) file. The next step included cleaning of the data in R using tidyverse library. The cleaned data consisted of three data frames, one for each type of insurance and possessing eight variables each. The data frames were then uploaded on the SQL database and multiple queries were run to check the database functionality. Finally, a few analyses were performed to find answers to the proposed research questions.

My motivation for this project comes from my love for healthcare. As a doctor, I have pledged to work towards treating and saving people's lives. And now, as a Health Informatics student, I pledge to provide safe and secure health related information that not only saves people's lives but also improves their quality of care. Thus, this project can provide me an in depth understanding of how the health plans are rated over the country and how this can help people choose the right health plan based on their requirements and location. This database can help people choose their health plan by providing them the quality of their customer satisfaction, prevention and treatment rates.

## Background

The **National Committee for Quality Assurance** (NCQA) is an independent nonprofit organization in the United States that works to improve health care quality through the administration of evidence-based standards, measures, programs, and accreditation. Health plans seek accreditation and measure performance through the administration and submission of the Healthcare Effectiveness Data and Information Set (HEDIS) and Consumer Assessment of Healthcare Providers and Systems (CAHPS) survey. NCQA Health Plan Accreditation builds upon more than 25 years of experience to provide a current, rigorous and comprehensive

framework for essential quality improvement and measurement. It is the only program in the industry that **bases results on clinical performance and consumer experience** (HEDIS<sup>®</sup> and CAHPS<sup>®</sup>).

## NCQA Health Plan Report Data

The NCQA Health Plan Report Data of **2018-2019** was scraped from the NCQA website:

<http://healthinsuranceratings.ncqa.org/2019/search>. The data used in this project comes from this website solely.

Three data frames were constructed through this website and they are **Private Insurance**, **Medicare** and **Medicaid**. The extraction of the data is limited to 5 states with most health plans in the United States. The states are **California** (CA), **New York** (NY), **Pennsylvania** (PA), **Virginia** (VA) and **Wisconsin** (WI). This decision is purely based on the motive of collecting as many data points as possible.

**Private Insurance** is any health insurance plan that is not run by the federal or state government.

**Medicare** is the federal government program that provides health care coverage to people who are 65 or older, younger people with disabilities and people with End-Stage Renal Disease.

**Medicaid** provides health coverage to millions of Americans, including eligible low-income adults, children, pregnant women, elderly adults and people with disabilities.

**Insurance Type:** Insurance Type Table (Ins) is the fourth data frame that forms the central entity for the other three data frames. This data frame consist of two variables. They are **Insurance Id**(Ins\_Id) AND **Insurance type** (Ins\_type)

Each data frame consists of ten variables. They are:

**HealthPlanId:** Health Plan Id is the primary key for the three report tables. Health Plan Id is unique for each health plan.

**Ins\_Id:** Insurance Id is the primary key for the central table, i.e. Insurance type. Insurance Id describes which health plans are associated with which insurance.

**Overall Rating:** Overall plan rating is the variable that provides combined score of the plan. This variable is numeric in nature. There are NAs, ‘Partial reported Data’ and ‘No reported Data’ available in the values.

**Plan Name:** Plan Name variable describes the name of the health plan. This variable is categorical in nature.

**State:** State variable displays the state in which the health plan is available. This variable is categorical in nature.

**Type:** This variable describes the type of organization associated with the health plan. This variable is categorical in nature.

**NCQA:** This variable demonstrates which health plan is accredited by NCQA. This variable is binary/Boolean. The values are Yes and No.

**Consumer Satisfaction:** Consumer Satisfaction is a variable that displays the satisfaction of the consumer with the health plan’s care, services and physicians. This variable is numeric in nature.

**Prevention:** Prevention variable demonstrates how well plans provide screenings, immunizations and other preventive services. This variable is numeric in nature.

**Treatment:** Treatment variable indicates a plan’s performance in treating chronic and acute conditions. This variable is numeric in nature.

	Private Insurance	Medicare	Medicaid	Insurance type
Total Variables	10	10	10	2
Total observations	100	99	73	3

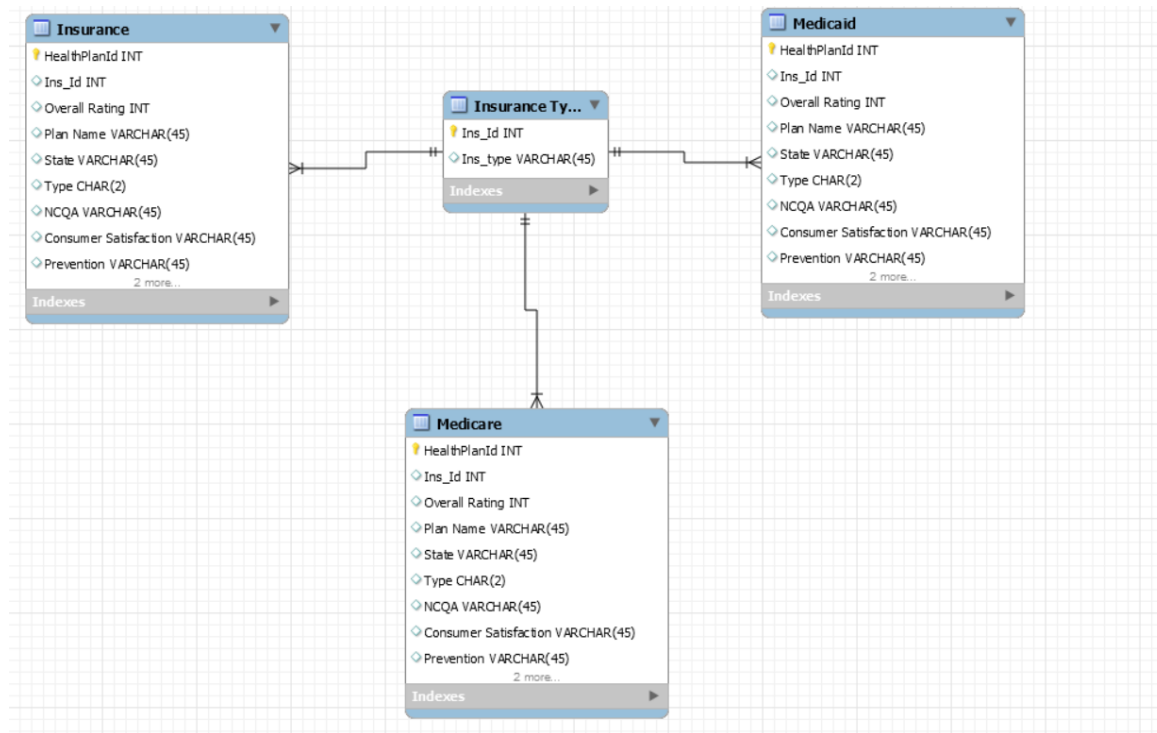
## Method Selection

Data collected for this project is solely through web scraping. I used import.io for scraping data from the NCQA website. I also used its interface for changing column names and added columns that weren't targeted by the css. I planned on scraping data from the five states that has the maximum number of health plans. The states were California, New York, Pennsylvania, Virginia and Wisconsin. I extracted data from each state thrice, one for private insurance, one for Medicare and one for Medicaid health plans (Note: Private, Medicare and Medicaid, they all have different health plans). Thus, in total, I extracted 15 tables such that (Private) Insurance had five tables, Medicare had five and so did Medicaid.

Once the raw data frames were downloaded and imported in R in the CSV format, I began tidying the data using a tidyverse library (mostly dplyr package). I imported data sets based on their insurance type. The cleaning process included removal of multiple values and substitution with the required value, removal of special characters, the addition of new columns, changes in the data type, normalizing the data and combining the data frames to one single table for Private Insurance, Medicare and Medicaid, respectively. The fourth table was also constructed to contain Insurance Id and Insurance Type.

I used MySQLWorkbench to create an entity-relationship model. Insurance Type table became the central entity of the database. (Private) Insurance, Medicare, and Medicaid tables are connected to the central entity. The relationship between the central entity and the other tables is one to many relationships. For example, each

Insurance Id may be in one or many health plans whereas each health plan may have only one Insurance Id.



In addition to the entity-relationship model, RSQLite library was used to create a database and add the data frames as tables. Overall four tables were added to the database and custom queries were run to test the database functionality.

```

Connecting to database:
```{r}
library(RSQLite)
database <- dbConnect(SQLite(), dbname="US_Insurance")
dbwriteTable(conn = database, name = 'Insurance', value = Insurance, row.names=F, header=T, overwrite=T)
dbwriteTable(conn = database, name = 'Medicare', value = Medicare, row.names=F, header=T, overwrite=T)
dbwriteTable(conn = database, name = "Medicaid", value = Medicaid, row.names=F, header=T, overwrite=T)
dbwriteTable(conn = database, name = "Ins", value = Ins, row.names=F, header=T, overwrite=T)
dbListTables(database)
```

[1] "Ins"      "Insurance" "Medicaid" "Medicare"

Checking the database by running custom queries:
```{r}
dbGetQuery(database, "select Max([Overall Rating]), [Plan Name], State
                      from Insurance
                      where (State='CA')")
```

  
```

| Max([Overall Rating]) | Plan Name  | State |
|-----------------------|--|-------|
| 4.5                   | Kaiser Foundation Health Plan Inc. - Southern California | CA    |

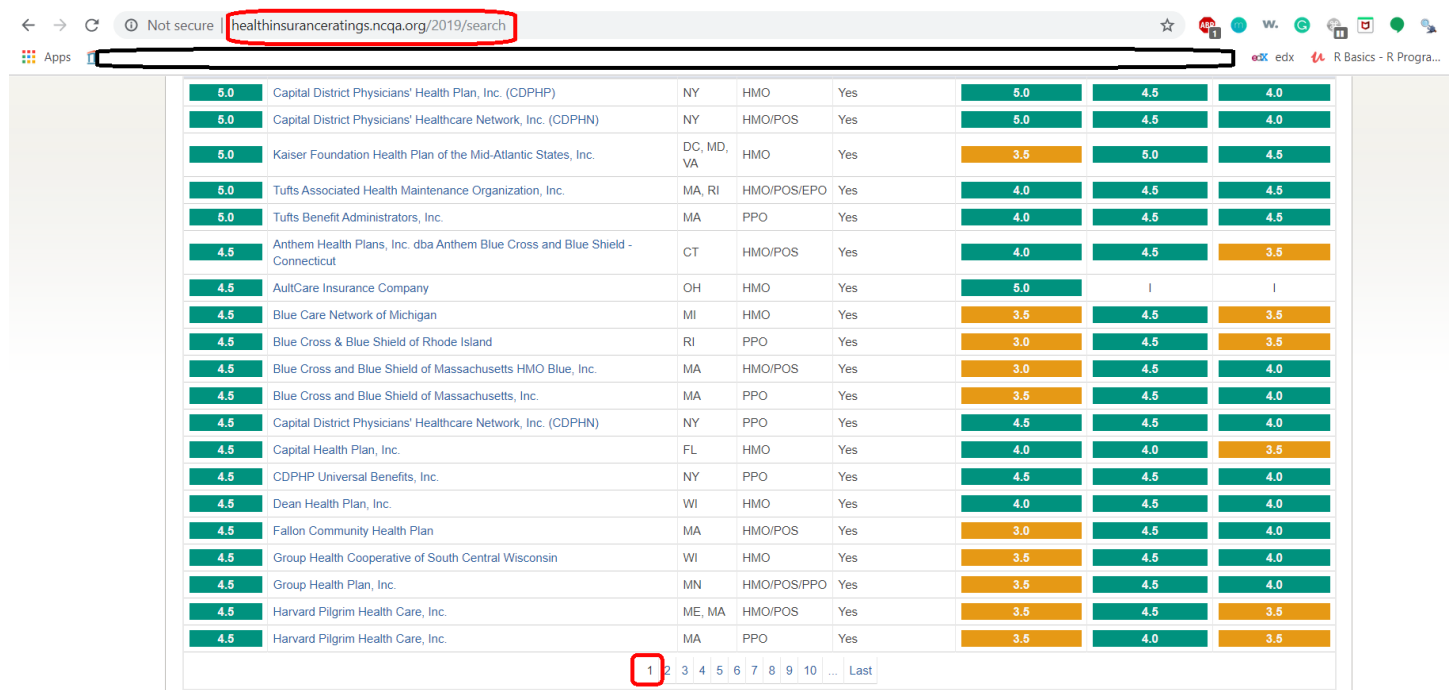
```

1 row
  
```

## Issues Encountered

The issue encountered during the project was related to the NCQA web URL. As my Project Proposal stated, I planned on extracting the data for all the states. On selecting 'All States', when one press to see the 2nd page of the list, there is no change in the URL. This is a common phenomenon as few website developers don't want the whole page to load when requested. There is a change in the JavaScript upon request and thus, only a portion of the webpage is loaded, instead of the whole page. This phenomenon does not allow the web page to reload. Hence, I could not extract data from the 2nd page onwards using the HTML syntax. Therefore, I thought of an alternative without jeopardizing the time spent on the idea. I planned on scraping data by choosing specific State names, which did show a change in the URL. Thus, I was able to proceed with my project.

In the picture below, one can see that the page is on the 1<sup>st</sup> list, and the url is <http://healthinsuranceratings.ncqa.org/2019/search>



|     |   |            |             |     |     |     |     |
|-----|---|------------|-------------|-----|-----|-----|-----|
| 5.0 | Capital District Physicians' Health Plan, Inc. (CDPHP)                        | NY         | HMO         | Yes | 5.0 | 4.5 | 4.0 |
| 5.0 | Capital District Physicians' Healthcare Network, Inc. (CDPHN)                 | NY         | HMO/POS     | Yes | 5.0 | 4.5 | 4.0 |
| 5.0 | Kaiser Foundation Health Plan of the Mid-Atlantic States, Inc.                | DC, MD, VA | HMO         | Yes | 3.5 | 5.0 | 4.5 |
| 5.0 | Tufts Associated Health Maintenance Organization, Inc.                        | MA, RI     | HMO/POS/EPO | Yes | 4.0 | 4.5 | 4.5 |
| 5.0 | Tufts Benefit Administrators, Inc.  | MA         | PPO         | Yes | 4.0 | 4.5 | 4.5 |
| 4.5 | Anthem Health Plans, Inc. dba Anthem Blue Cross and Blue Shield - Connecticut | CT         | HMO/POS     | Yes | 4.0 | 4.5 | 3.5 |
| 4.5 | AultCare Insurance Company  | OH         | HMO         | Yes | 5.0 | I   | I   |
| 4.5 | Blue Care Network of Michigan   | MI         | HMO         | Yes | 3.5 | 4.5 | 3.5 |
| 4.5 | Blue Cross & Blue Shield of Rhode Island                                      | RI         | PPO         | Yes | 3.0 | 4.5 | 3.5 |
| 4.5 | Blue Cross and Blue Shield of Massachusetts HMO Blue, Inc.                    | MA         | HMO/POS     | Yes | 3.0 | 4.5 | 4.0 |
| 4.5 | Blue Cross and Blue Shield of Massachusetts, Inc.                             | MA         | PPO         | Yes | 3.5 | 4.5 | 4.0 |
| 4.5 | Capital District Physicians' Healthcare Network, Inc. (CDPHN)                 | NY         | PPO         | Yes | 4.5 | 4.5 | 4.0 |
| 4.5 | Capital Health Plan, Inc.   | FL         | HMO         | Yes | 4.0 | 4.0 | 3.5 |
| 4.5 | CDPHP Universal Benefits, Inc.  | NY         | PPO         | Yes | 4.5 | 4.5 | 4.0 |
| 4.5 | Dean Health Plan, Inc.  | WI         | HMO         | Yes | 4.0 | 4.5 | 4.0 |
| 4.5 | Fallon Community Health Plan  | MA         | HMO/POS     | Yes | 3.0 | 4.5 | 4.0 |
| 4.5 | Group Health Cooperative of South Central Wisconsin                           | WI         | HMO         | Yes | 3.5 | 4.5 | 4.0 |
| 4.5 | Group Health Plan, Inc.   | MN         | HMO/POS/PPO | Yes | 3.5 | 4.5 | 4.0 |
| 4.5 | Harvard Pilgrim Health Care, Inc.   | ME, MA     | HMO/POS     | Yes | 3.5 | 4.5 | 3.5 |
| 4.5 | Harvard Pilgrim Health Care, Inc.   | MA         | PPO         | Yes | 3.5 | 4.0 | 3.5 |

In the picture below, one can see that the page is on the 4th list, and still, the web URL is the same as the above screenshot.

← → ↻ Not secure | healthinsuranceratings.ncqa.org/2019/search

Apps MBA + MSc in Heal... . St... edX edX R Be

| Rating | Plan Name  | States | Type    | Accreditation | Satisfaction | Prevention | Treatment |
|--------|--|--------|---------|---------------|--------------|------------|-----------|
| 4.0    | First Priority Life Insurance Company Inc. (FPLIC)                     | PA     | PPO     | Yes           | 3.0          | 3.5        | 3.5       |
| 4.0    | Geisinger Health Plan  | PA     | HMO/POS | Yes           | 2.5          | 3.5        | 3.5       |
| 4.0    | Geisinger Quality Options  | PA     | PPO     | Yes           | 3.5          | 3.5        | 3.5       |
| 4.0    | Group Health Cooperative of Eau Claire                                 | WI     | HMO     | No            | 3.5          | 4.0        | 4.0       |
| 4.0    | Harvard Pilgrim Health Care of New England, Inc.                       | NH     | HMO/POS | Yes           | 3.0          | 3.5        | 3.5       |
| 4.0    | Hawaii Medical Service Association (HMSA)                              | HI     | HMO/POS | Yes           | 4.0          | 4.0        | 3.0       |
| 4.0    | Hawaii Medical Service Association (HMSA)                              | HI     | PPO     | Yes           | 4.5          | 3.5        | 3.5       |
| 4.0    | Health Alliance Medical Plans  | IL     | HMO/POS | Yes           | 3.0          | 4.0        | 3.5       |
| 4.0    | Health Alliance Midwest, Inc.  | IA     | HMO/POS | Yes           | 3.0          | 4.0        | 3.5       |
| 4.0    | Health Alliance Plan of Michigan                                       | MI     | HMO/POS | Yes           | 3.5          | 3.5        | 3.5       |
| 4.0    | Highmark Choice Company  | PA     | HMO     | Yes           | 4.0          | 4.0        | 3.5       |
| 4.0    | Highmark Health Insurance Company (HHIC)                               | PA     | PPO/EPO | Yes           | 3.0          | 3.0        | 3.5       |
| 4.0    | HMO Colorado, Inc.   | CO     | HMO     | Yes           | 2.5          | 4.0        | 3.5       |
| 4.0    | HMO of Northeastern Pennsylvania, Inc. d/b/a First Priority Health     | PA     | HMO     | Yes           | 3.5          | 3.5        | 3.5       |
| 4.0    | Humana Health Plan, Inc. (Illinois)                                    | IL     | HMO/POS | Yes           | 3.5          | 3.5        | 3.5       |
| 4.0    | Humana Health Plan, Inc. (Kansas City)                                 | KS, MO | HMO/POS | Yes           | 4.0          | 3.5        | 3.0       |
| 4.0    | Humana Insurance Company (Illinois)                                    | IL     | PPO     | Yes           | I            | 4.0        | 3.5       |
| 4.0    | Humana Wisconsin Health Organization Insurance Corporation (Wisconsin) | WI     | HMO/POS | Yes           | I            | 3.5        | 3.5       |
| 4.0    | Independence Hospital Indemnity Plan                                   | PA     | PPO     | Yes           | 4.0          | 4.0        | 3.0       |
| 4.0    | Kaiser Foundation Health Plan of Georgia, Inc.                         | GA     | HMO     | Yes           | 2.0          | 4.0        | 3.5       |

1 2 3 4 5 6 7 8 9 10 ... Last

## What I Learned

This project provided a comprehensive understanding of the steps involved in the collection, storage and retrieval of the data. There are numerous things that I have learned during this project and I would like to list them out.

- 1) Web-scraping using import.io.
- 2) Use of dplyr package for cleaning and tidying, the data this includes changing column names, row values, creating new variables, substituting values (gsub), and binding two or more data frames into one.
- 3) Use of regular expressions to describe the patterns of the set of strings.
- 4) Use of RSQLite Package to create a SQL database, and store tables in it.
- 5) Learned how to carry out custom queries to check the database functionality.
- 6) Gained experience by performing complex analyses.



## Future Implementations

As described, the purpose of this project is to obtain more information about the health plans in the five most populous states. This project will answer the following two research questions which can help in further investigation:

1. What's the average state ratings (of all the three insurance types) based on consumer satisfaction, prevention and treatment?
2. Which insurance plan provides the best customer satisfaction, treatment, and prevention?

The information collected can be used to analyze which states have above average quality of health plans. This can also be used as a guide to choose health plans by people. Private insurance companies can use this database to observe the quality of health plans and take relevant measures to improve their health plans. States with poor treatment or prevention scores can come under the radar of private & federal companies, which can ultimately improve the quality of care. Another future implementation of this project could be to create a database that has discrete tables that can allow live flow of data through the system.

## References:

1. <https://www.ncqa.org/>
2. <http://healthinsuranceratings.ncqa.org/2019/search>

# Appendix: Methods and Figures

## Methods

I first explored the NCQA website to observe the report of the health plans

(<http://healthinsuranceratings.ncqa.org/2019/search>). Once, I established a detailed plan, I started scraping data from the website using import.io. I am attaching a few screenshots that can display my work.

1. The picture below is the final product of data scraping for the Medicaid table of California State.

| Plan Name   | State | Type | NCQA Accreditation | Consumer Satisfaction | Prevention | Treatment |
|---|-------|------|--------------------|-----------------------|------------|-----------|
| Alliance for Health   | CA    | HMO  | Yes                | 3.0                   | 4.5        | 3.0       |
| City Health Group   | CA    | HMO  | Yes                | 3.0                   | 4.0        | 3.5       |
| County Health Authority, dba L.A. Care Health Plan                              | CA    | HMO  | Yes                | 2.0                   | 3.5        | 3.5       |
| County Health Authority - dba CalOptima   | CA    | HMO  | Yes                | 2.5                   | 4.0        | 3.5       |
| Cisco Community Health Authority  | CA    | HMO  | Yes                | 3.0                   | 4.0        | 3.5       |
| Costa Health Plan   | CA    | HMO  | Yes                | 2.0                   | 3.5        | 3.0       |
| Empire Health Plan  | CA    | HMO  | Yes                | 3.0                   | 3.5        | 3.0       |
| Healthcare of California Partner Plan Inc.                                      | CA    | HMO  | Yes                | 1.5                   | 3.5        | 3.0       |
| Health of California Partnership Plan   | CA    | HMO  | Yes                | 1.5                   | 3.0        | 2.5       |
| Health of California Promise Health Plan  | CA    | HMO  | Yes                | 2.0                   | 3.5        | 2.0       |
| California Health & Wellness  | CA    | HMO  | Yes                | 2.5                   | 2.0        | 3.0       |
| Health Net of California, Inc.  | CA    | HMO  | Yes                | 1.5                   | 2.5        | 2.5       |
| San Joaquin County Health Commission dba Health Plan of San Joaquin             | CA    | HMO  | Yes                | 1.5                   | 2.5        | 2.5       |
| Santa Clara County Health Authority, dba Santa Clara Family Health Plan (SCFHP) | CA    | HMO  | No                 | 1.5                   | 2.5        | 1.0       |
| Special Project / Area: CA Medicaid / Santa Clara                               |       |      |                    |                       |            |           |

As one can see in the picture below, all the column names are incorrect. While using this application, I had to change the column names after extraction. There were instances where the app didn't select all the rows, and I had to be cautious while using this application.

The screenshot shows the import.io website interface for extracting data from the NCQA Health Insurance Plan Ratings 2019-2020. The main content area displays the NCQA logo and a search interface for health insurance plans by state, plan name, or plan type. A table of results is shown, listing plan names, states, and types. A dropdown menu is open over the table, showing a list of ratings from 3.5 to 5.0.

2. After Scraping, the data frame was downloaded from import.io website in the CSV format and it was imported to R using readr library.

The screenshot shows the import.io website interface for the 'Insurance\_California' extractor. The main content area displays the 'Run History' section, showing a table of runs with columns for Date/Time, Duration, URLs, Success, Failed, and Total Rows. A dropdown menu is open over the table, showing options for downloading the data in Excel, CSV, NDJSON, or Images and Files format.

```

19 Importing datasets for Private Insurance:
20 {r}
21 library(readr)
22 California <- read_csv("C:/Users/bajpa/Downloads/Collecting, Storing, Retrieving Data/Project/Data/Private/California.csv")
23 view(California)
24
25 library(readr)
26 NewYork <- read_csv("C:/Users/bajpa/Downloads/Collecting, Storing, Retrieving Data/Project/Data/Private/NewYork.csv")
27 view(NewYork)
28
29 library(readr)
30 Pennsylvania <- read_csv("C:/Users/bajpa/Downloads/Collecting, Storing, Retrieving Data/Project/Data/Private/Pennsylvania.csv")
31 head(Pennsylvania)
32
33 library(readr)
34 Virginia <- read_csv("C:/Users/bajpa/Downloads/Collecting, Storing, Retrieving Data/Project/Data/Private/Virginia.csv")
35 view(Virginia)
36
37 library(readr)
38 wisconsin <- read_csv("C:/Users/bajpa/Downloads/Collecting, storing, Retrieving Data/Project/Data/Private/wisconsin.csv")
39 view(wisconsin)
40

```

- The picture below shows the before and after of tidying, the Pennsylvania data frame. As one can see in the first half of the picture, under the state column, we have MA & PA (As multiple state values were present). Since this data frame is extracted from Pennsylvania, the desired state value is PA. Thus, in the second half of the picture, gsub() code is run to remove all the values and the next code inserts the PA values. This code helps in providing only one value for the state. Similar approach was taken for all the states.

```
library(readr)
Pennsylvania <- read_csv("C:/Users/bajpa/Downloads/Collecting, Storing, Retrieving Data/Project/Data/

## Parsed with column specification:
## cols(
##   `Overall Rating` = col_double(),
##   `Plan Name` = col_character(),
##   State = col_character(),
##   Type = col_character(),
##   `Consumer Satisfaction` = col_double(),
##   Prevention = col_double(),
##   Treatment = col_double(),
##   NCQA = col_character()
## )

head(Pennsylvania)

## # A tibble: 6 x 8
##   `Overall Rating` `Plan Name` State Type `Consumer Satis- Prevention
##             <dbl> <chr>      <chr> <chr>          <dbl>      <dbl>
## 1             4.5 Martin's P~ MA, ~ HMO             5         3.5
## 2             4.5 UPMC Benef- PA    HMO             3.5        4.5
## 3             4.5 UPMC Healt- PA    HMO             3.5        4.5
## 4             4.5 UPMC Healt- PA    HMO             3.5        4.5
## 5             4    Aetna Heal- PA    HMO/-            3         3.5

Pennsylvania$State<-gsub("[[:upper:]]","",Pennsylvania$State)
Pennsylvania$State<-'PA'
head(Pennsylvania)

## # A tibble: 6 x 8
##   `Overall Rating` `Plan Name` State Type `Consumer Satis- Prevention
##             <dbl> <chr>      <chr> <chr>          <dbl>      <dbl>
## 1             4.5 Martin's P~ PA    HMO             5         3.5
## 2             4.5 UPMC Benef- PA    HMO             3.5        4.5
## 3             4.5 UPMC Healt- PA    HMO             3.5        4.5
## 4             4.5 UPMC Healt- PA    HMO             3.5        4.5
## 5             4    Aetna Heal- PA    HMO/-            3         3.5
## 6             4    Aetna Life- PA    PPO/-           3.5        3.5
## # ... with 2 more variables: Treatment <dbl>, NCQA <chr>
```

- In the picture below, the first line of code is the rbind() code. This code is used to combine two or more data frames. In this scenario, I am combining five data frames that are extracted from the private

insurance database. These data frames are taken from California, New York, Pennsylvania, Virginia and Wisconsin.

Second code is used to create two new variables. The first variable is the Insurance Id which is '1' for Private Insurance and the second variable is the Health Plan Id which is unique for every health plan.

These two variables are combined with the Insurance data frame, thus providing a data frame with ten variables. Similar approach is used to create the master data frame for Medicare and Medicaid.

```
Insurance<-rbind(California,NewYork,Pennsylvania,Virginia, Wisconsin)
head(Insurance)
```

```
## # A tibble: 6 x 8
##   `Overall Rating` `Plan Name` State Type  NCQA  `Consumer Satis-
##           <dbl> <chr>         <chr> <chr> <chr> <chr>
## 1           4.5 Kaiser Fou~ CA    HMO    Yes  2.5
## 2           4.5 Kaiser Fou~ CA    HMO    Yes   3
## 3           4.5 Sharp Heal~ CA    HMO    Yes   4
## 4           3.5 Aetna Heal~ CA   HMO/~  Yes   2
## 5           3.5 Anthem Blu~ CA   PPO/~  Yes  2.5
## 6           3.5 Blue Cross~ CA   HMO/~  Yes  2.5
## # ... with 2 more variables: Prevention <chr>, Treatment <chr>
```

```
Ins_Id <-1
Insurance <- cbind(Ins_Id,Insurance)
HealthPlanId<- c(101:200)
HealthPlanId<-as.data.frame(HealthPlanId)
Insurance <- cbind(HealthPlanId,Insurance)
head(Insurance, 8)
```

```
##   HealthPlanId Ins_Id Overall Rating
## 1          101      1           4.5
## 2          102      1           4.5
## 3          103      1           4.5
## 4          104      1           3.5
## 5          105      1           3.5
## 6          106      1           3.5
## 7          107      1           3.5
## 8          108      1           3.5
##
##                                Plan Name State  Type
## 1 Kaiser Foundation Health Plan Inc. - Southern California  CA    HMO
## 2 Kaiser Foundation Health Plan, Inc. - Northern California  CA    HMO
## 3                                Sharp Health Plan  CA    HMO
## 4                                Aetna Health of California Inc.  CA HMO/POS
## 5          Anthem Blue Cross Life and Health Insurance Company  CA PPO/EPO
## 6                Blue Cross of California dba Anthem Blue Cross  CA HMO/POS
## 7                Blue Cross of California dba Anthem Blue Cross  CA PPO/EPO
## 8                Blue Shield of California  CA HMO/POS
##   NCQA Consumer Satisfaction Prevention Treatment
## 1 Yes                2.5           5           4.5
## 2 Yes                3           5           4.5
```

5. In the picture below, 'gsub()' code is used to substitute the letter 'I' with 'NA'. This code is used in the other master data frames as well.

Removing Special Characters from the columns:

```
Insurance$`Consumer Satisfaction`<-gsub("I","NA",Insurance$`Consumer Satisfaction`)
Insurance$Prevention<-gsub("I","NA", Insurance$Prevention)
Insurance$Treatment<-gsub("I","NA", Insurance$Treatment)
head(Insurance)
```

6. In the picture below, a new data frame is created. This table consists of two variables, which are Insurance Id and Insurance type. The final data frames that are stored in the database are Insurance, Medicare, Medicaid and Ins.

Creating the Central Entity that connects all the other tables.

```
Ins_Id<-c(1,2,3)
Ins_type<-c("Private_Insurance", "Medicare", "Medicaid")
Ins <- cbind(Ins_Id,Ins_type)
Ins<-as.data.frame(Ins)
head(Ins)
```

```
##   Ins_Id      Ins_type
## 1      1 Private_Insurance
## 2      2      Medicare
## 3      3      Medicaid
```

The four dataframes are Insurance (Private Insurance), Medicare and Medicaid and Ins

```
view(Insurance)
view(Medicare)
view(Medicaid)
view(Ins)
```

7. In the picture below, RSQLite library is used to connect to a database and store the data frames in the form of tables. dbListTables() code results in the list of tables that are present in the database. As one can see, there are four tables present in this database.
- Once the table has been established in the database, there was a need to run a few queries to check the database functionality. Below is one of the custom queries which results in the row that has a maximum overall rating in the California state.

Connecting to database:

```
library(RSQLite)
database <- dbConnect(SQLite(), dbname="US_Insurance")
dbWriteTable(conn = database, name = 'Insurance', value = Insurance, row.names=F, header=T, overwrite=T)
dbWriteTable(conn = database, name = 'Medicare', value = Medicare, row.names=F, header=T, overwrite=T)
dbWriteTable(conn = database, name = "Medicaid", value = Medicaid, row.names=F, header=T, overwrite=T)
dbWriteTable(conn = database, name = "Ins", value = Ins, row.names=F, header=T, overwrite=T)
dbListTables(database)
```

```
## [1] "Ins"      "Insurance" "Medicaid" "Medicare"
```

Checking the database by running custom queries:

```
dbGetQuery(database, "select Max([Overall Rating]), [Plan Name], State
                      from Insurance
                      where (State='CA')")
```

```
## Max([Overall Rating])
## 1 4.5
##                               Plan Name State
## 1 Kaiser Foundation Health Plan Inc. - Southern California CA
```

8. Here are few more queries that test the data base model.

The image displays three sequential screenshots of an RStudio interface, showing SQL queries and their corresponding data results.

**Query 1:** Selects data from the Medicare table based on Overall Rating, Consumer Satisfaction, and Treatment.

```
dbGetQuery(database, "select [Overall Rating], [Plan Name], State, Type, NCQA, [Consumer Satisfaction], [Prevention], Treatment
                        from Medicare
                        where ([Overall Rating]>4 AND [Consumer Satisfaction]>4 AND [Consumer Satisfaction]!='NA') AND [Overall Rating]!='NA' ")
```

| State | Type | NCQA | Consumer Satisfaction | Prevention | Treatment |
|-------|------|------|-----------------------|------------|-----------|
| PA    | HMO  | Yes  | 4.5                   | 3.5        | 3.5       |
| WI    | HMO  | No   | 5                     | 4          | 5         |
| WI    | HMO  | Yes  | 4.5                   | 3          | 4         |
| WI    | HMO  | Yes  | 5                     | 5          | 3.5       |

4 rows | 3-8 of 8 columns

**Query 2:** Selects data from the Medicare table based on Overall Rating, Prevention, and Treatment.

```
dbGetQuery(database, "select [Overall Rating], [Plan Name], State, Type, NCQA, [Consumer Satisfaction], [Prevention], Treatment
                        from Medicare
                        where ([Treatment]>4 AND [Prevention]>4 AND [Treatment]!='NA')")
```

| State | Type | NCQA | Consumer Satisfaction | Prevention | Treatment |
|-------|------|------|-----------------------|------------|-----------|
| CA    | HMO  | Yes  | 3.5                   | 5          | 4.5       |
| CA    | HMO  | Yes  | 3.5                   | 5          | 4.5       |
| VA    | HMO  | Yes  | 3.5                   | 5          | 4.5       |

3 rows | 3-8 of 8 columns

**Query 3:** Selects data from the Medicaid table based on Overall Rating and Prevention.

```
dbGetQuery(database, "select [Overall Rating], [Plan Name], State, Type, NCQA, [Consumer Satisfaction], [Prevention], Treatment
                        from Medicaid
                        where ([Overall Rating]>4 AND [Prevention]<4 AND [Prevention]!='NA' AND [Overall Rating]!='NA')")
```

| Overall Rating | Plan Name  | State | Type | NCQA | Consumer Satisfaction |
|----------------|--|-------|------|------|-----------------------|
| 4.5            | Vista Health Plan DBA AmeriHealth Caritas Pennsylvania | HMO   | Yes  | 4.5  |                       |

1 row | 1-6 of 8 columns

9. In the picture below, Insurance data frame is used to carry out analysis to check the average ratings of the health plans in the states.

In this analysis, Pennsylvania has the highest overall, consumer satisfaction and prevention rating whereas, Wisconsin has the highest treatment rating.

```

94 `}`
95 B<- Insurance
96 B$`Consumer Satisfaction`<-gsub("NA","0",B$`Consumer Satisfaction`)
97 B$`Prevention`<-gsub("NA","0",B$`Prevention`)
98 B$`Treatment`<-gsub("NA","0",B$`Treatment`)
99 B<-na.omit(B)
00 B$`Consumer Satisfaction`<-as.numeric(B$`Consumer Satisfaction`)
01 B$`Prevention`<-as.numeric(B$`Prevention`)
02 B$`Treatment`<-as.numeric(B$`Treatment`)
03 view(B)
04
05 b <- B %>%
06   na.omit() %>%
07   group_by(State) %>%
08   summarise_at(vars(-NCQA, -'Plan Name',-Type,-HealthPlanId, -Ins_Id),fun(mean(., na.rm=TRUE)))
09 view(b)
10
11 Ans1_a<- b[order(-b$`Overall Rating`),]
12 head(Ans1_a)
13

```

| State<br><chr> | Overall Rating<br><dbl> | Consumer Satisfaction<br><dbl> | Prevention<br><dbl> | Treatment<br><dbl> |
|----------------|-------------------------|--------------------------------|---------------------|--------------------|
| PA             | 4.100                   | 3.575                          | 3.775               | 3.450              |
| WI             | 4.000                   | 2.850                          | 3.750               | 3.550              |
| NY             | 3.825                   | 3.300                          | 3.225               | 3.325              |
| CA             | 3.525                   | 2.175                          | 3.350               | 2.950              |
| VA             | 3.450                   | 2.800                          | 3.150               | 3.025              |

5 rows

In the Medicare analysis, New York has the highest overall and prevention rating whereas, Wisconsin has the highest treatment rating and Pennsylvania has highest consumer satisfaction rating.

```

Medicare Analysis:
`}`
C<- Medicare
C$`Consumer Satisfaction`<-gsub("NA","0",C$`Consumer Satisfaction`)
C$`Prevention`<-gsub("NA","0",C$`Prevention`)
C$`Treatment`<-gsub("NA","0",C$`Treatment`)
C$`Overall Rating`<-gsub("NA","0",C$`Overall Rating`)
C<-na.omit(C)
C$`Overall Rating`<-as.numeric(C$`Overall Rating`)
C$`Consumer Satisfaction`<-as.numeric(C$`Consumer Satisfaction`)
C$`Prevention`<-as.numeric(C$`Prevention`)
C$`Treatment`<-as.numeric(C$`Treatment`)
view(C)

c <- C %>%
  na.omit() %>%
  group_by(State) %>%
  summarise_at(vars(-NCQA, -'Plan Name',-Type, -HealthPlanId, -Ins_Id),fun(mean(., na.rm=TRUE)))
view(c)

Ans1_b<- c[order(-c$`Overall Rating`),]
head(Ans1_b)
`}`

```

| State<br><chr> | Overall Rating<br><dbl> | Consumer Satisfaction<br><dbl> | Prevention<br><dbl> | Treatment<br><dbl> |
|----------------|-------------------------|--------------------------------|---------------------|--------------------|
| NY             | 3.850000                | 3.675000                       | 3.950000            | 3.450000           |
| PA             | 3.725000                | 3.700000                       | 3.625000            | 3.250000           |
| WI             | 3.656250                | 3.593750                       | 3.656250            | 3.593750           |
| CA             | 3.500000                | 2.525000                       | 3.400000            | 3.300000           |
| VA             | 2.294118                | 2.117647                       | 2.735294            | 2.852941           |

5 rows

In the Medicaid Analysis, Pennsylvania has the highest Overall Rating and Treatment rating, Virginia has the highest Consumer satisfaction ratings whereas New York has the highest prevention rating.



Medicaid Analysis:

```
```{r}
D<- Medicaid
D$`Consumer Satisfaction`<-gsub("NA","0",D$`Consumer Satisfaction`)
D$`Prevention`<-gsub("NA","0",D$`Prevention`)
D$`Treatment`<-gsub("NA","0",D$`Treatment`)
D$`Overall Rating`<-gsub("NA","0",D$`Overall Rating`)
D<-na.omit(D)
D$`Overall Rating`<-as.numeric(D$`Overall Rating`)
D$`Consumer Satisfaction`<-as.numeric(D$`Consumer Satisfaction`)
D$`Prevention`<-as.numeric(D$`Prevention`)
D$`Treatment`<-as.numeric(D$`Treatment`)
view(D)

d <- D %>%
  na.omit() %>%
  group_by(State) %>%
  summarise_at(vars(-NCQA, -`Plan Name`, -Type, -HealthPlanId, -Ins_Id), funs(mean(., na.rm=TRUE)))
view(d)

Ans1_c<- d[order(-d$`Overall Rating`),]
head(Ans1_c)
```
```

| State<br><chr> | Overall Rating<br><dbl> | Consumer Satisfaction<br><dbl> | Prevention<br><dbl> | Treatment<br><dbl> |
|----------------|-------------------------|--------------------------------|---------------------|--------------------|
| PA             | 4.000000                | 3.111111                       | 3.333333            | 3.555556           |
| VA             | 3.300000                | 3.600000                       | 2.200000            | 2.700000           |
| NY             | 2.423077                | 1.807692                       | 3.423077            | 3.307692           |
| CA             | 2.325000                | 1.525000                       | 2.700000            | 2.300000           |
| WI             | 1.607143                | 1.392857                       | 2.857143            | 3.214286           |

5 rows

Thank You 😊