Healthcare fraud

Economic Impact

Everyone Shares the Burden of Healthcare Fraud

In 2018, **$3.6 trillion** was spent on health care in the United States, representing billions in health insurance claims. It is an undisputed reality that some of these claims are fraudulent. Although they constitute only a small fraction, those fraudulent claims carry a very high price tag, both financially and in how they impact our perception of the integrity and value of our healthcare system.

The National Health Care Anti-Fraud Association (NHCAA) estimates that the financial losses due to healthcare fraud are in the tens of billions of dollars each year. A conservative estimate is **3% of total healthcare expenditures**, while some government and law enforcement agencies place the loss as high as 10% of our annual health outlay, which could mean **more than $300 billion**.

NOTE: I would like to clarify that my intention in pursuing this project is **primarily exploratory in nature**. My goal is to delve into the unknown, navigating the complexities of the data and, in the process, enhancing my understanding of data science. It's worth mentioning that I am embarking on this journey with no prior experience in data science.

Types:

Billing: Billing for services that were never rendered—by using **genuine patient information**, sometimes obtained through **identity theft**, to fabricate entire claims or by padding otherwise legitimate claims with charges for procedures or **services that did not take place**.

Upcoding: Billing for more expensive services or procedures than were provided or performed, commonly known as “upcoding”—i.e., **falsely billing for a higher-priced treatment** than was provided (which often requires the accompanying “inflation” of the patient’s diagnosis code to a more serious condition consistent with the false procedure code).

Unnecessary: Performing medically unnecessary services solely for the purpose of generating insurance payments—this is seen very often in diagnostic-testing schemes such as nerve conduction and genetic testing.

Misrepresenting: Misrepresenting non-covered treatments as medically necessary covered treatments for purposes of obtaining insurance payments—this is widely seen in cosmetic-surgery schemes, in which non-covered cosmetic procedures such as “nose jobs” are billed to patients’ insurers as deviated-septum repairs.

Falsifying: Falsifying a patient’s diagnosis and medical record to justify tests, surgeries, or other procedures that aren’t medically necessary.

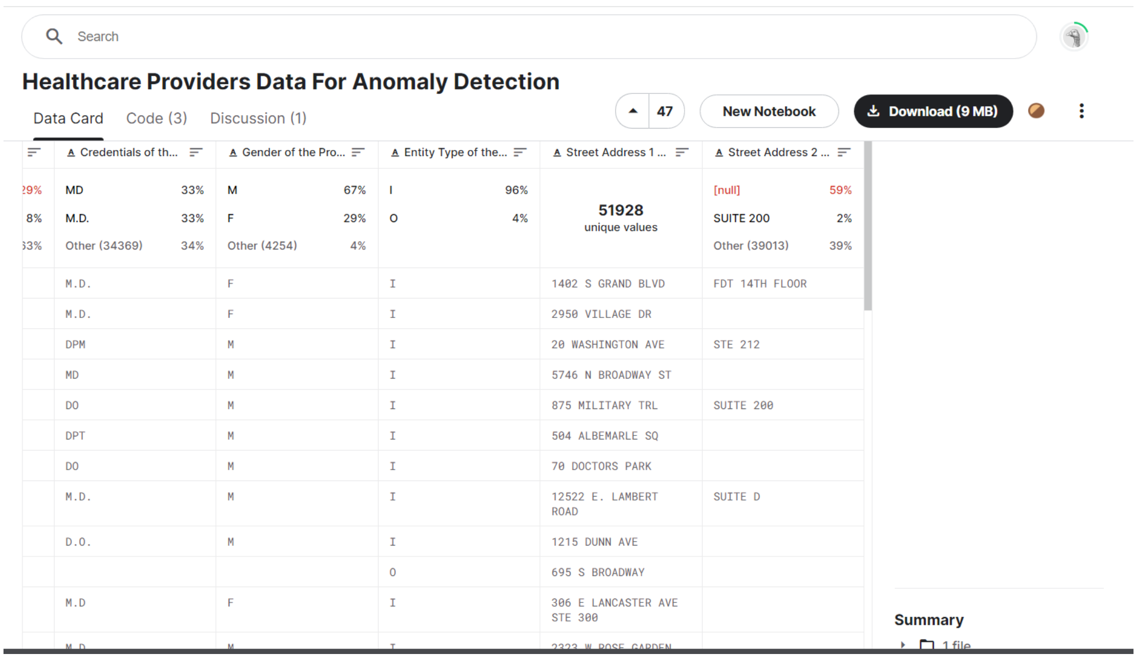
Unbundling: Unbundling—billing for each step of a procedure as if they are separate procedures.

Over-billing: Billing a patient more than the required co-pay amount for services that were prepaid or paid in full by the benefit plan under the terms of a managed care contract.

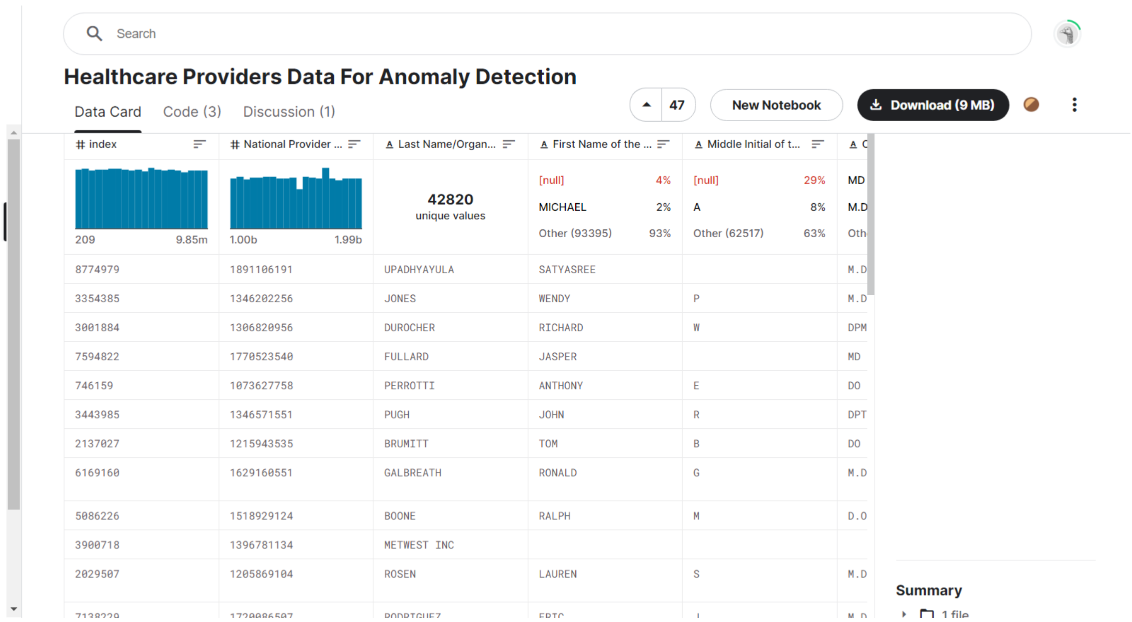
Kickbacks: Accepting kickbacks for patient referrals.

Reference: National Health Care Anti-Fraud Association (NHCAA). (2018). The Challenge of Health Care Fraud. Retrieved from [https://www.nhcaa.org/tools-insights/about-health-care-fraud/the-challenge-of-health-care-fraud/].

The Dataset:



Credentials of th...	Gender of the Pro...	Entity Type of the...	Street Address 1 ...	Street Address 2 ...
MD 33%	M 67%	I 96%	51928 unique values	[null] 59%
M.D. 33%	F 29%	O 4%		SUITE 200 2%
Other (34369) 34%	Other (4254) 4%			Other (39013) 39%
M.D.	F	I	1402 S GRAND BLVD	FDT 14TH FLOOR
M.D.	F	I	2950 VILLAGE DR	
DPM	M	I	20 WASHINGTON AVE	STE 212
MD	M	I	5746 N BROADWAY ST	
DO	M	I	875 MILITARY TRL	SUITE 200
DPT	M	I	504 ALBEMARLE SQ	
DO	M	I	70 DOCTORS PARK	
M.D.	M	I	12522 E. LAMBERT ROAD	SUITE D
D.O.	M	I	1215 DUNN AVE	
		O	695 S BROADWAY	
M.D	F	I	306 E LANCASTER AVE STE 300	



Source: https://www.kaggle.com/datasets/tamilsel/healthcare-providers-data

This is a big data set with 100,000 rows and 27 variables in total, the variables range from identifiers to demographic details of the provider, their participation in Medicare along with financial information about claims and billing.

Variables are as follows:

Npi

nppes_provider_last_org_name

nppes_provider_first_name

nppes_provider_mi

nppes_credentials

nppes_provider_gender

nppes_entity_code

nppes_provider_street1

nppes_provider_street

nppes_provider_city

nppes_provider_zip

nppes_provider_state

nppes_provider_country

provider_type

medicare_participation_indicator

medicare_participation_indicator

place_of_service

hcpcs_code

hcpcs_description

hcpcs_drug_indicator

line_srvc_cnt

bene_unique_cnt

bene_day_srvc_cnt

average_Medicare_allowed_amt.

stdev_Medicare_allowed_amt.

average_submitted_chrg_amt.

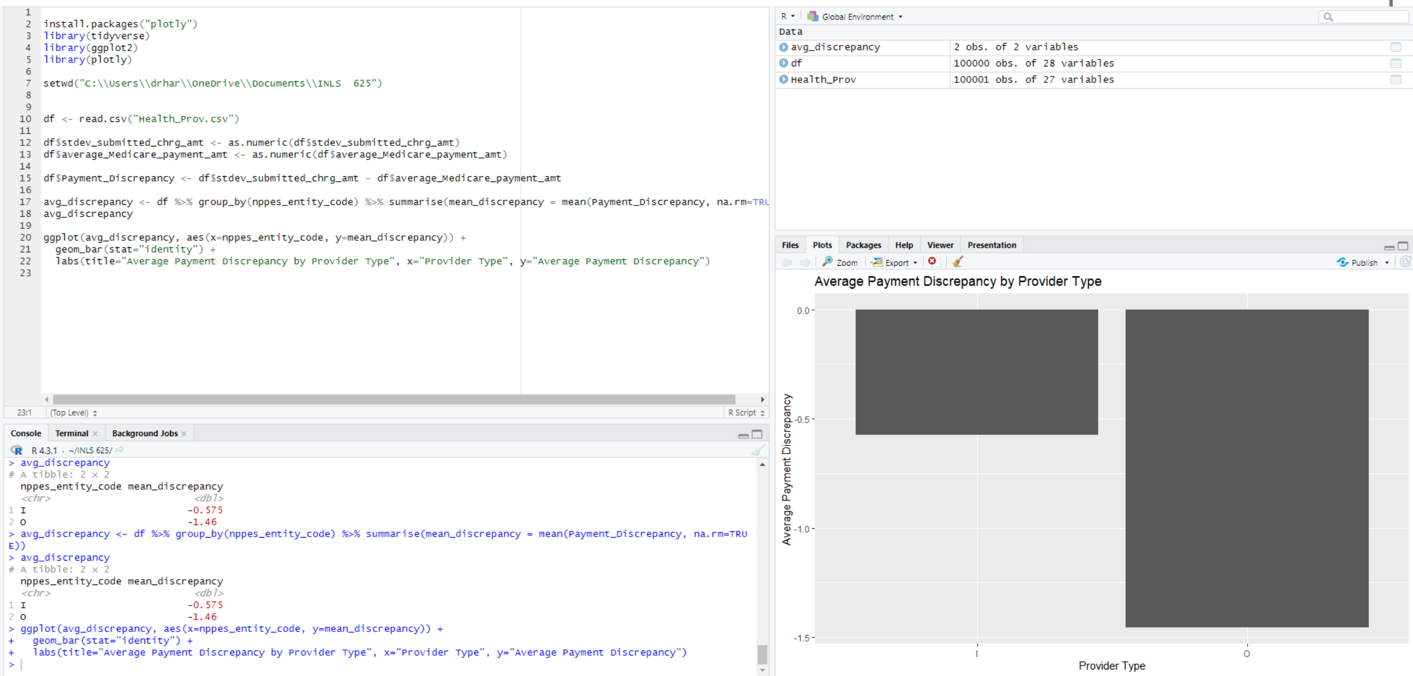
stdev_submitted_chrg_amt.

average_Medicare_payment_amt

The R studio Code:

```
# Loading necessary libraries

install.packages("plotly")

library(tidyverse)

library(ggplot2)

library(plotly)

# Setting the working directory

setwd("C:\\Users\\drhar\\OneDrive\\Documents\\INLS  625")

# Read the data

df <- read.csv("Health_Prov.csv")

# Convert columns to numeric

# Analysis 1: determining if individual providers have more
Payment_Discrepancy compared to organizational providers

'''The focus of my 1st analysis is on 4 variables: Provider
type(Individual or Organisational):stdev_submitted_chrg_amt, Actual
payment for a service recieved
```

```
by a provider and Average amount paid by Medicare for their purticular
service:average_Medicare_payment_amt .

'''

df$stdev_submitted_chrg_amt <- as.numeric(df$stdev_submitted_chrg_amt)

df$average_Medicare_payment_amt <-
as.numeric(df$average_Medicare_payment_amt)

# Payment_Discrepancy is the difference between the two variables, It
helps us understand the discrepancy in the payment with

# regards to various factors(ex: Individual provider or a provider under
an organization )


df$Payment_Discrepancy <- df$stdev_submitted_chrg_amt -
df$average_Medicare_payment_amt


# Grouping by nppes_entity_code (Code for Individual provider or a
provider under an organization) and calculate the mean
Payment_Discrepancy

avg_discrepancy <- df %>% group_by(nppes_entity_code) %>%
summarise(mean_discrepancy = mean(Payment_Discrepancy, na.rm=TRUE))

avg_discrepancy

# Plot

ggplot(avg_discrepancy, aes(x=nppes_entity_code, y=mean_discrepancy)) +

  geom_bar(stat="identity") +

  labs(title="Average Payment Discrepancy by Provider Type", x="Provider
Type", y="Average Payment Discrepancy")
```



'''

The results provided show the average Payment_Discrepancy for individual providers (I) and organizational providers (O).

interpreting the results:

Individual Providers (I):

The average Payment_Discrepancy is -0.575. This means that, on average, individual providers receive $0.575 less than the amount they submitted as charges to Medicare.

Organizational Providers (O):

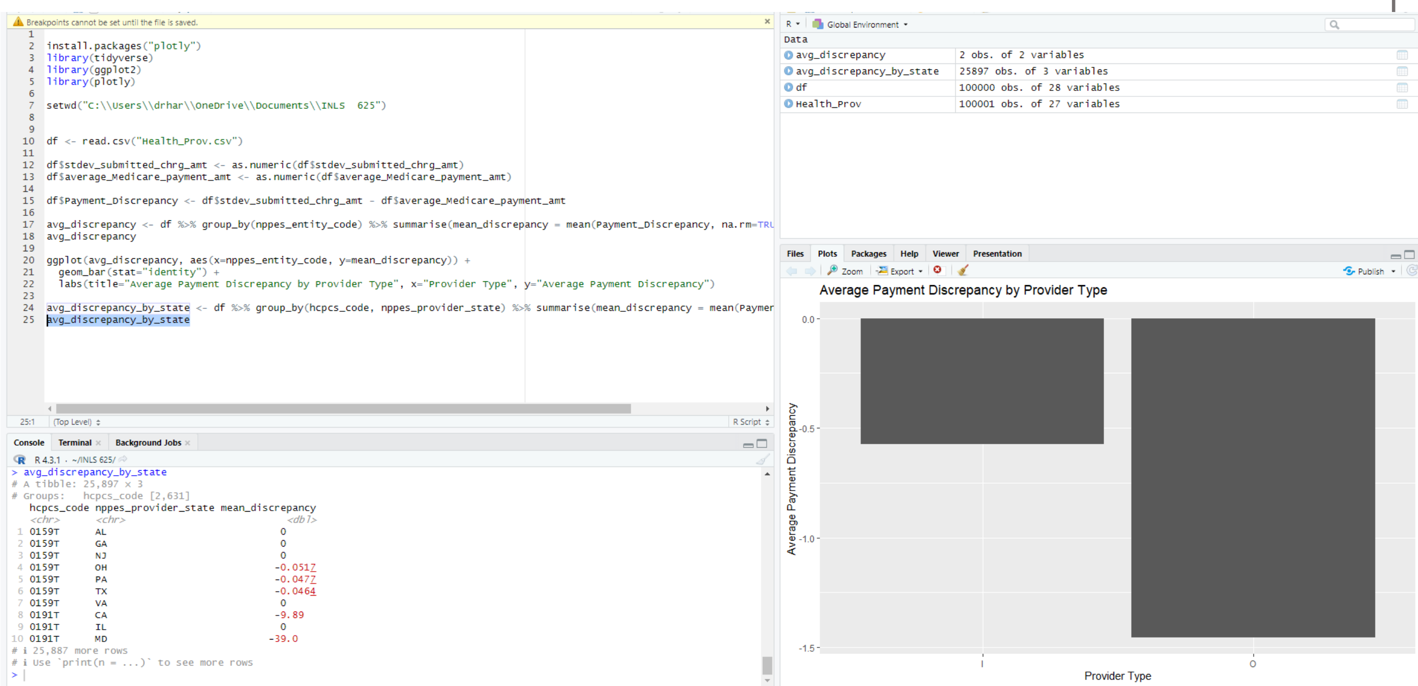
The average Payment_Discrepancy is -1.456773. This indicates that, on average, organizational providers receive approximately $1.457 less than the amount they submitted as charges to Medicare.

Comparison:

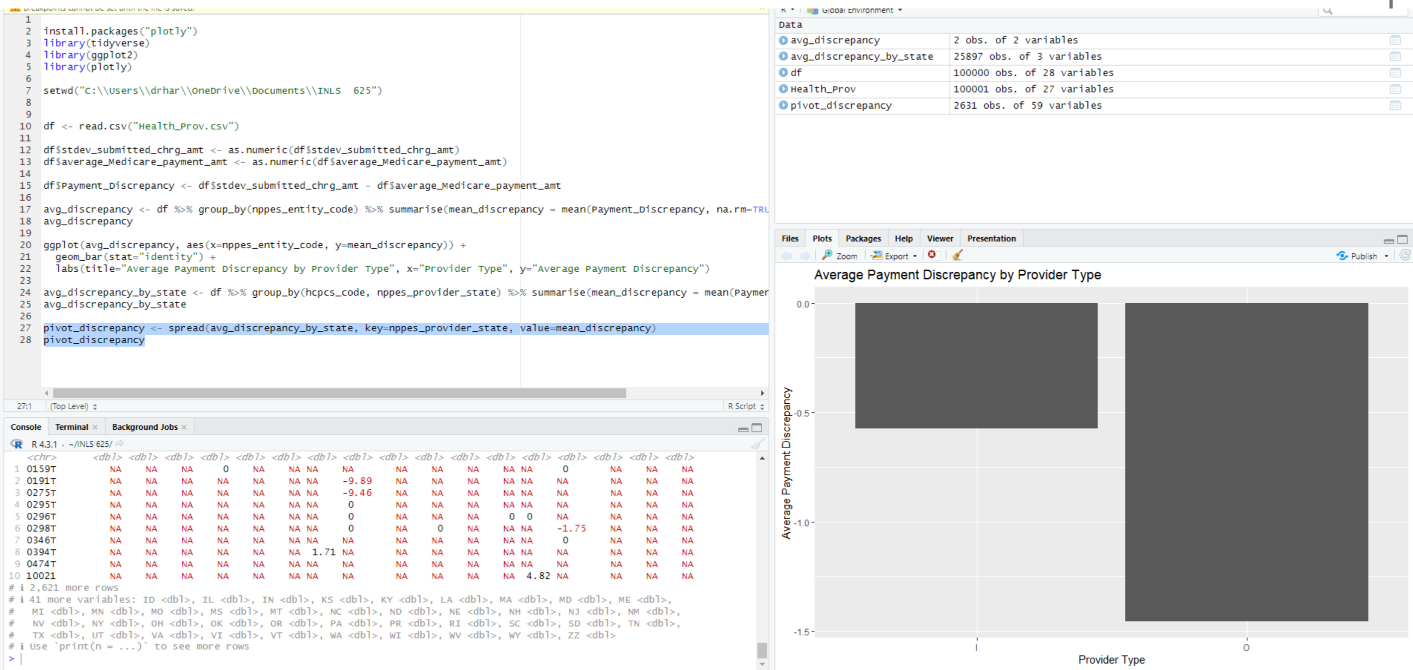
Organizational providers have a larger average discrepancy between the amount they charge and the amount they receive from Medicare compared to individual providers. This suggests that organizational providers might be charging more than individual providers, but they are also receiving a slightly lesser proportion of their charges from Medicare.

In simpler terms, both individual and organizational providers are receiving less than what they charge, but the discrepancy is larger for organizations.

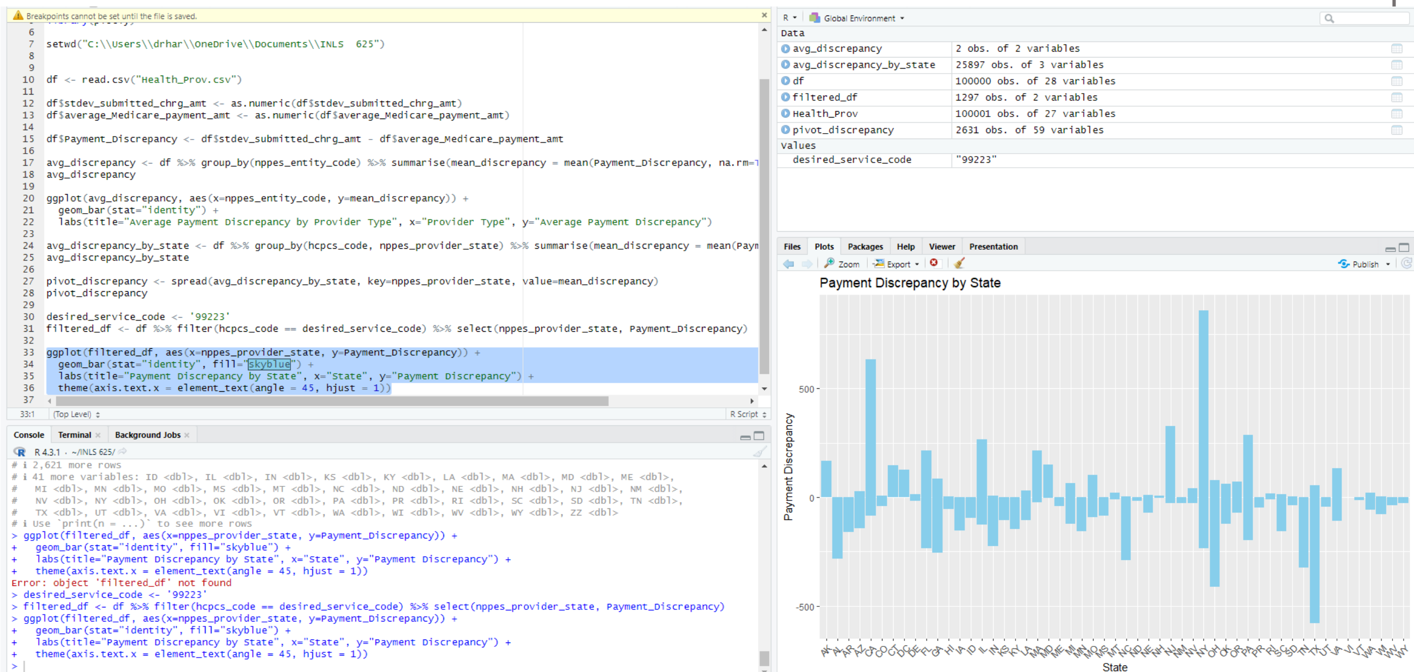
```
'''
 #   Analysis 2: determining the minimum and maximum discrepancies for a particular service code in different states.

# Grouping by service code and state and calculate the average Payment_Discrepancy

avg_discrepancy_by_state <- df %>% group_by(hcpcs_code, nppes_provider_state) %>% summarise(mean_discrepancy = mean(Payment_Discrepancy, na.rm=TRUE))
```



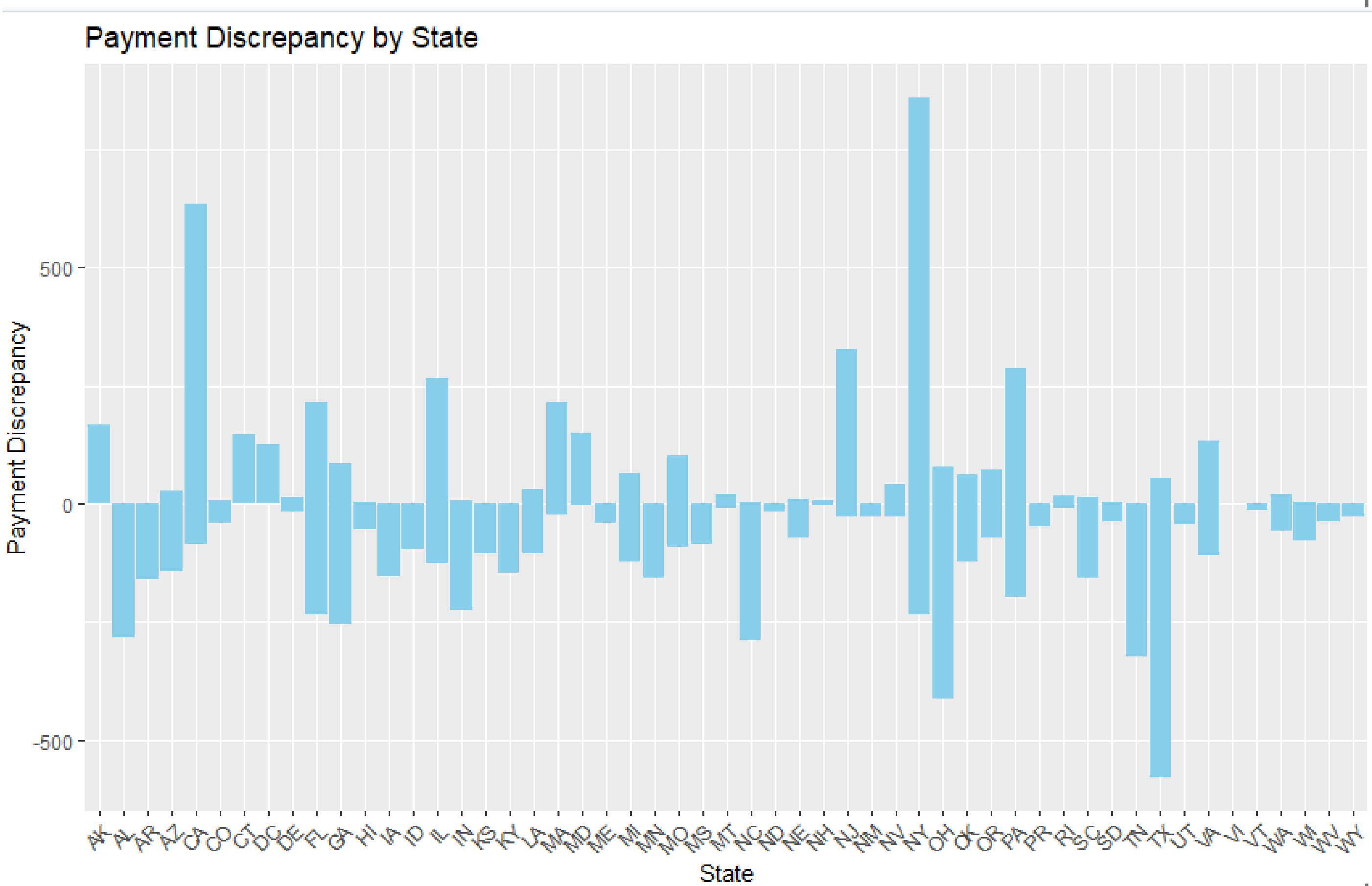

```
avg_discrepancy_by_state

# Pivot the data

pivot_discrepancy <- spread(avg_discrepancy_by_state, key=nppes_provider_state, value=mean_discrepancy)

pivot_discrepancy
```




```
# Filtering the data frame based on the desired service code

desired_service_code <- '99223'

filtered_df <- df %>% filter(hcpcs_code == desired_service_code) %>%
select(nppes_provider_state, Payment_Discrepancy)

# Plot

ggplot(filtered_df, aes(x=nppes_provider_state, y=Payment_Discrepancy))
+

  geom_bar(stat="identity", fill="skyblue") +

  labs(title="Payment Discrepancy by State", x="State", y="Payment
Discrepancy") +

  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

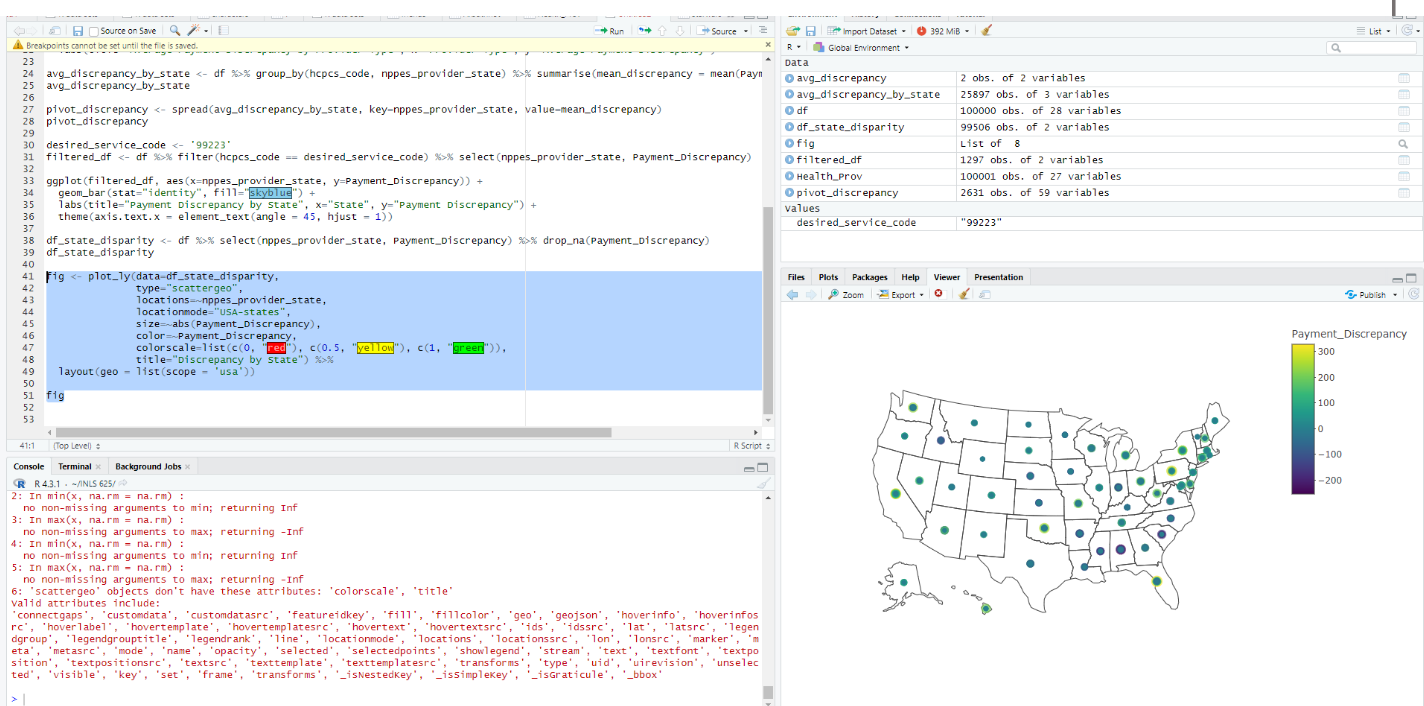


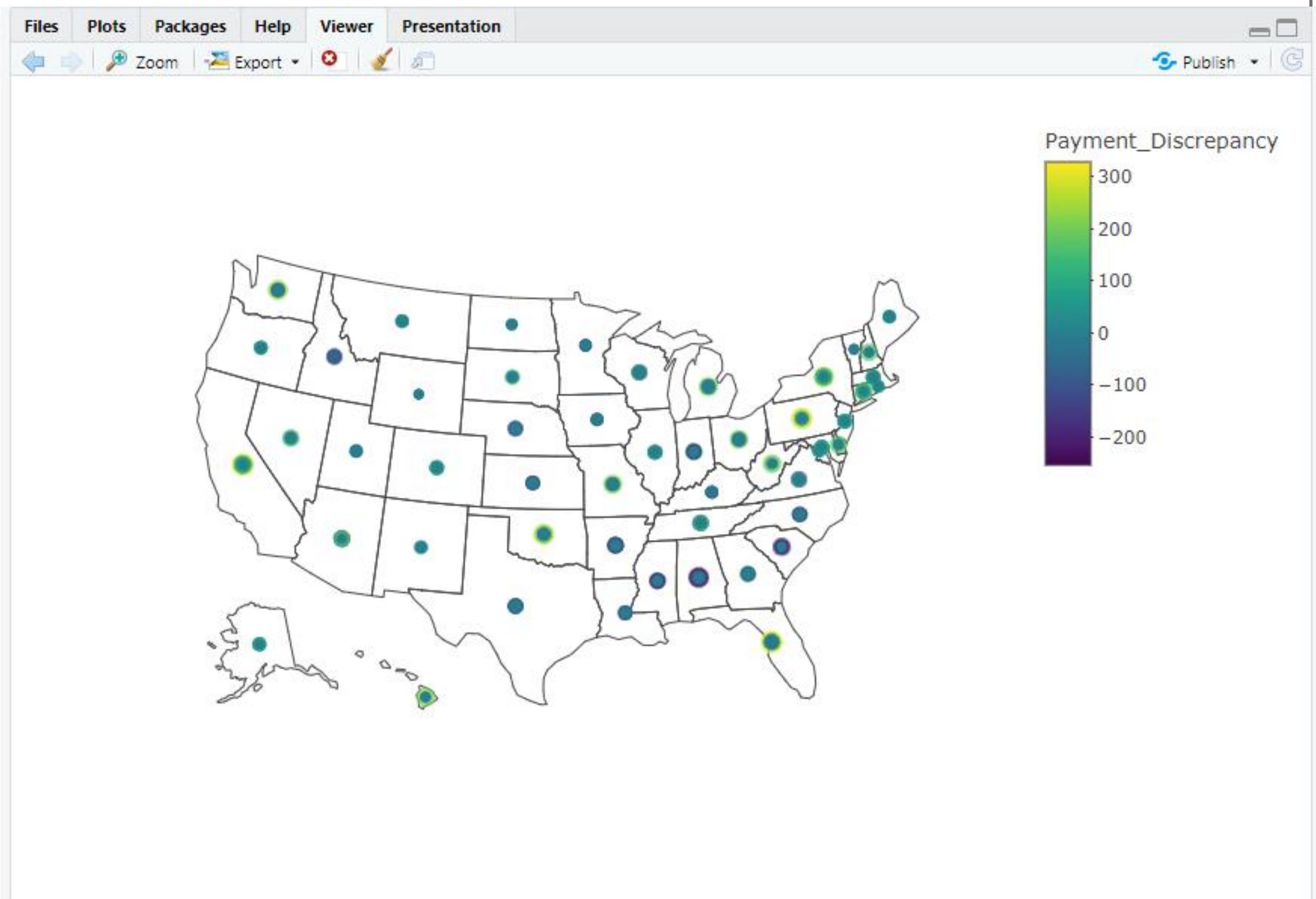
Breakpoints cannot be set until the file is saved.
setwd("C:\\Users\\drhar\\OneDrive\\Documents\\INLS 625")
df <- read.csv("Health_Prov.csv")
df$stdev_submitted_chrg_amt <- as.numeric(df$stdev_submitted_chrg_amt)
df$average_Medicare_payment_amt <- as.numeric(df$average_Medicare_payment_amt)
df$Payment_Discrepancy <- df$stdev_submitted_chrg_amt - df$average_Medicare_payment_amt
avg_discrepancy <- df %>% group_by(nppes_entity_code) %>% summarise(mean_discrepancy = mean(Payment_Discrepancy, na.rm=T
avg_discrepancy
ggplot(avg_discrepancy, aes(x=nppes_entity_code, y=mean_discrepancy)) +
geom_bar(stat="identity") +
labs(title="Average Payment Discrepancy by Provider Type", x="Provider Type", y="Average Payment Discrepancy")
avg_discrepancy_by_state <- df %>% group_by(hcpcs_code, nppes_provider_state) %>% summarise(mean_discrepancy = mean(Paym
avg_discrepancy_by_state
pivot_discrepancy <- spread(avg_discrepancy_by_state, key=nppes_provider_state, value=mean_discrepancy)
pivot_discrepancy
desired_service_code <- '99223'
filtered_df <- df %>% filter(hcpcs_code == desired_service_code) %>% select(nppes_provider_state, Payment_Discrepancy)
ggplot(filtered_df, aes(x=nppes_provider_state, y=Payment_Discrepancy)) +
geom_bar(stat="identity", fill="skyblue") +
labs(title="Payment Discrepancy by State", x="State", y="Payment Discrepancy") +
theme(axis.text.x = element_text(angle = 45, hjust = 1))
33:1 (Top Level) R Script
Console Terminal Background Jobs
R 4.3.1 · ~/INLS 625/
i 2,621 more rows
i 41 more variables: ID <dbl>, IL <dbl>, IN <dbl>, KS <dbl>, KY <dbl>, LA <dbl>, MA <dbl>, MD <dbl>, ME <dbl>,
MI <dbl>, MN <dbl>, MO <dbl>, MS <dbl>, MT <dbl>, NC <dbl>, ND <dbl>, NE <dbl>, NH <dbl>, NJ <dbl>, NM <dbl>,
NV <dbl>, NY <dbl>, OH <dbl>, OK <dbl>, OR <dbl>, PA <dbl>, PR <dbl>, RI <dbl>, SC <dbl>, SD <dbl>, TN <dbl>,
TX <dbl>, UT <dbl>, VA <dbl>, VI <dbl>, VT <dbl>, WA <dbl>, WI <dbl>, WV <dbl>, WY <dbl>, ZZ <dbl>
i Use `print(n = ...)` to see more rows
> ggplot(filtered_df, aes(x=nppes_provider_state, y=Payment_Discrepancy)) +
+ geom_bar(stat="identity", fill="skyblue") +
+ labs(title="Payment Discrepancy by State", x="State", y="Payment Discrepancy") +
+ theme(axis.text.x = element_text(angle = 45, hjust = 1))
Error: object 'filtered_df' not found
> desired_service_code <- '99223'
> filtered_df <- df %>% filter(hcpcs_code == desired_service_code) %>% select(nppes_provider_state, Payment_Discrepancy)
> ggplot(filtered_df, aes(x=nppes_provider_state, y=Payment_Discrepancy)) +
+ geom_bar(stat="identity", fill="skyblue") +
+ labs(title="Payment Discrepancy by State", x="State", y="Payment Discrepancy") +
+ theme(axis.text.x = element_text(angle = 45, hjust = 1))
Global Environment
Data
avg_discrepancy 2 obs. of 2 variables
avg_discrepancy_by_state 25897 obs. of 3 variables
df 100000 obs. of 28 variables
filtered_df 1297 obs. of 2 variables
Health_Prov 100001 obs. of 27 variables
pivot_discrepancy 2631 obs. of 59 variables
Values
desired_service_code "99223"
Files Plots Packages Help Viewer Presentation
Zoom Export Publish
Payment Discrepancy by State



Payment Discrepancy by State
Payment Discrepancy
500
0
-500
AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA MA MD ME MI MN MO MS MT NC ND NE NH NJ NM NV NY OH OK OR PA PR RI SC SD TN TX UT VA VI VT WA WI WV WY
State

```r
# Drop NA values

df_state_disparity <- df %>% select(nppes_provider_state,
Payment_Discrepancy) %>% drop_na(Payment_Discrepancy)

df_state_disparity


# Plot using plotly

fig <- plot_ly(data=df_state_disparity,

               type="scattergeo",

               locations=~nppes_provider_state,

               locationmode="USA-states",

               size=~abs(Payment_Discrepancy),

               color=~Payment_Discrepancy,

               colorscale=list(c(0, "red"), c(0.5, "yellow"), c(1,
"green")),

               title="Discrepancy by State") %>%

  layout(geo = list(scope = 'usa'))


fig
```





List of other project sites that you searched and explored before choosing your project (at least 5):

1. Raw data for Association Between Physician Burnout and Patient Safety, Professionalism, and Patient Satisfaction A Systematic Review and Meta-analysis.
   https://data.mendeley.com/datasets/nk4y768v9c/1
2. Change of Primary Care Physician (PCP)- Identifying why members request a change in PCP
   https://www.kaggle.com/datasets/harshams07/change-of-primary-care-physician
3. Medical Practitioner and Prescription Behavior-National Survey of Physician Characteristics and Prescription Practices in USA
   https://www.kaggle.com/datasets/tubmak/dataset
4. Prescription drug information from the Controlled Substance Monitoring Database (CSMD) for the five most recent years (2018-2022) of available data. Source: **Opioid Prescription data by TDH-office of Informatics & Analytics- TN Department of Health.**
   https://www.tn.gov/content/dam/tn/health/program-areas/reports_and_publications/2022-CSMD-Annual-Report.pdf
5. Health Detection Using Smart Phone Data-mobile devices to track the health of patients.
   https://www.kaggle.com/datasets/vshantam/health-detection-using-smart-phone-data