Executive Summary: EDA of Medicare Part D

Data Summary

This analysis focuses on Medicare Part D, which provides coverage for prescription drugs in the U.S., particularly for individuals aged 65 and older, as well as for younger individuals with disabilities. The dataset consists of 50,000 records with 26 features, of which 18 are categorical variables (such as gender and procedure type) and 8 are numerical variables (such as charge amounts). The primary variables of interest include Gender of the Provider, Average Submitted Charge Amount, and the Procedures. These variables were used to explore potential disparities between male and female medical providers in terms of the fees they charge for Medicare-covered services.

Methodology

The study began with an Exploratory Data Analysis (EDA) to identify patterns, anomalies, and key relationships between the variables. The steps followed in the analysis included:

- **Data Cleaning:** The first column of the original dataset was removed as it doesn't have a column header and no necessary information at all. No blanks, NULLs, NA values we're cleaned off from the dataset.

- **Summary of Statistics:** Calculating the central tendency (mean, median) and chi-square statistic for key numerical & categorical variables, particularly charges by provider gender on specific procedures.
    a. Medicare Participation – a frequency & proportions table was used to check a supporting data of the objective. It was found that 99.95% of providers in the dataset accept Medicare, confirming that non-participation had a negligible impact on the analysis.
    b. Individual vs. Organization – a proportion analysis was used and showed that 95.62% of the providers in the dataset were individual practitioners, with only a small percentage representing organizations. Furthermore,
    c. Gender Distribution – a proportion analysis was used and showed that 67.69% of the providers in the dataset were male, 27.93% are female, while the remaining 4.37% were organizations whose genders can't be generalized. Organizations were excluded in further gender analysis.
    d. Categorical Frequency Distribution – A one categorical variable analysis where the 2029 unique procedures from 50,000 rows were categorized into 3: High Frequency for > 100k avg. procedures/ day, Low Frequency for < 10k avg. procedures/ day and Mid Frequency for in between. The frequency of procedures were added up and bucketed into each category.
    e. Gender vs. High Frequency Procedure – A two categorical variable analysis where we check both the presence of both genders per procedure for the study and the relationship between. The computed chi-square statistics is 79.818 and visually shows that the two variables are not independent.
    f. Medical Charge vs Gender per Procedure – A one numerical and one categorical variable analysis to compare the median of the average submitted charge amount between the male and female medical practitioners.

- **Data Visualization:** Creating charts to depict the frequency of procedures performed by both male and female providers and to visualize charge amounts across different services.

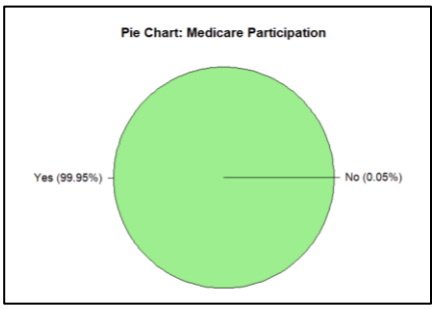


Fig. A: Pie Chart: Medicare Participation

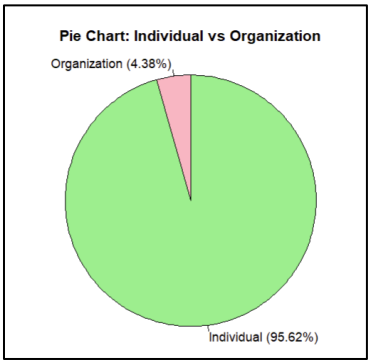


Fig. B: Pie Chart: Individual vs Organization

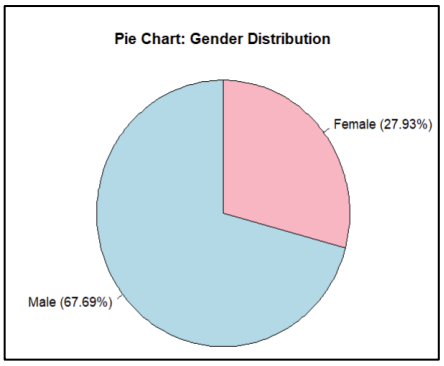


Fig. C: Pie Chart: Gender Distribution

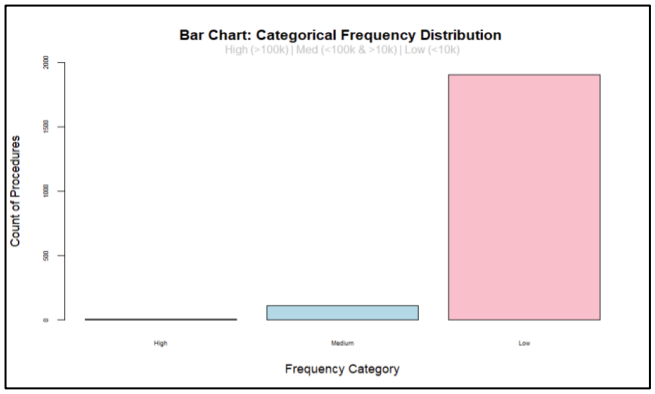


Fig. D: Bar Chart: Categorical Frequency Distribution

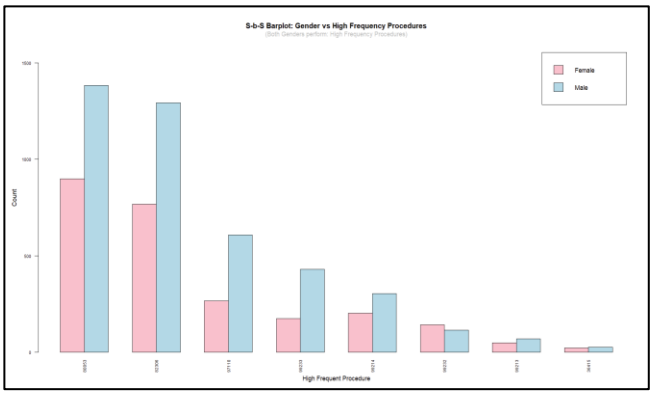


Fig. E: Side-by-Side Bar plot: Gender vs High Freq. Procedures

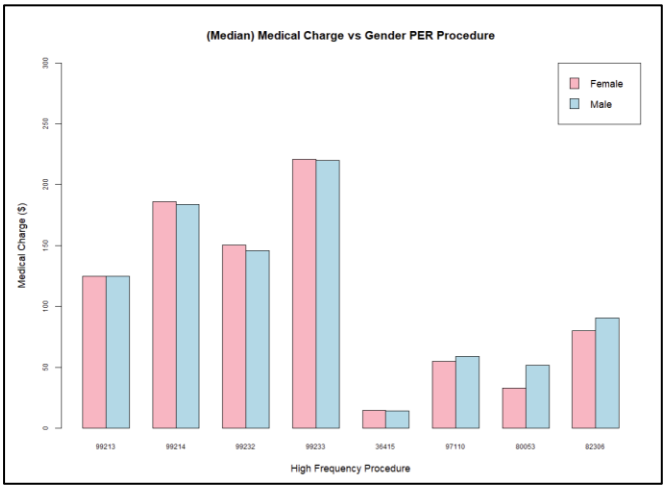


Fig. F: Side-by-Side Bar plot: Medical Charge vs Gender per Procedure

Output & Results

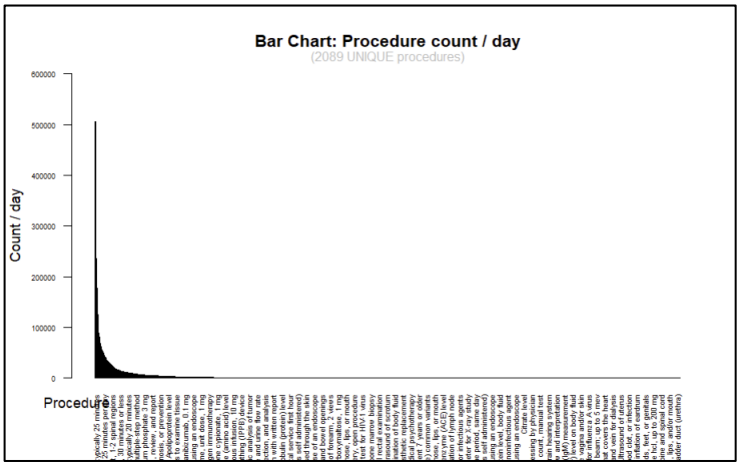
Gender Disparity in Charges

Female Providers Charge More in some Procedures: For 4 of the 8 high-frequency procedures, female providers charged slightly more than their male counterparts. However, the difference was marginal and not significant.
Male Providers Charge Significantly More in Other Procedures: In 3 of the 8 procedures, male providers charged significantly more than female providers. The disparities were notable in certain outpatient and inpatient services, indicating a gender gap in those specific procedures as summarized on Figures E and F.

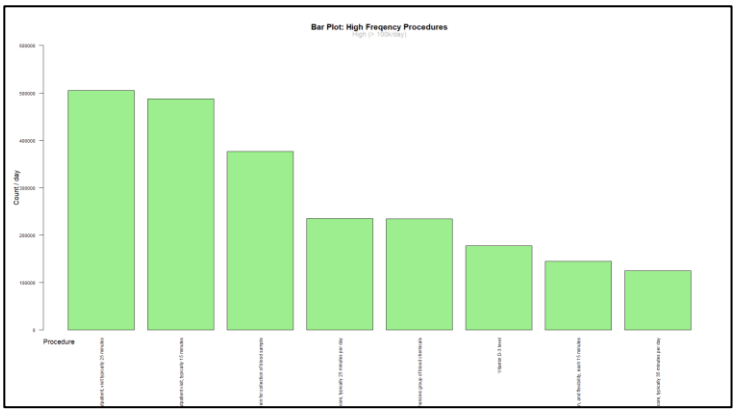
Appendix:

Other supporting Data Analysis Methodology & Visualizations:

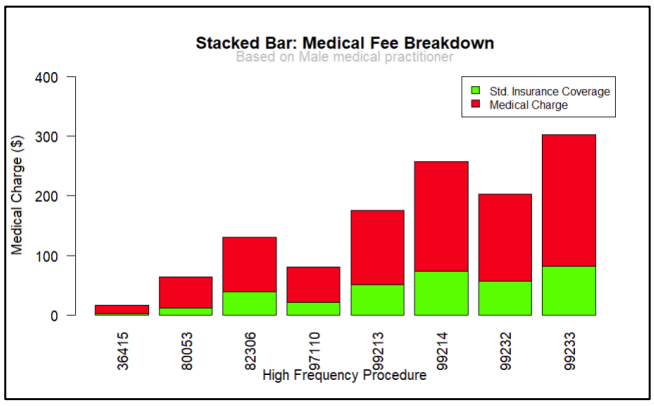
a. Procedure Count per day – A one categorical and one numerical variable analysis that summarized the 50,000 procedures into 2029 unique HCPCS Codes where the average count/day we're summed giving the sum of average service frequency.

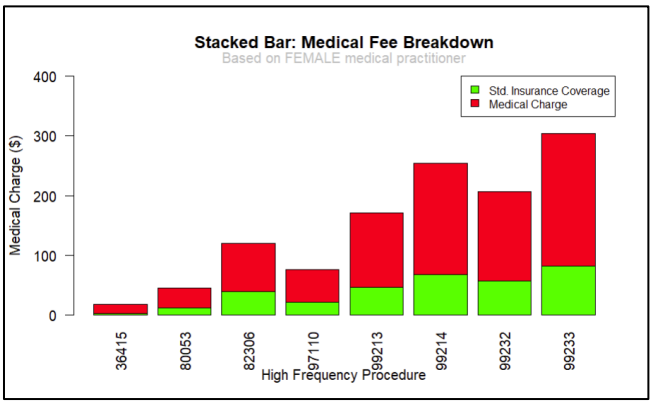


b. High Frequency Procedures Distribution – A one categorical variable and one numerical variable analysis where we took a closer look at the high frequency procedures and their avg. procedure count per day.



c. Medical Fee Breakdown – A one numerical and one categorical variable analysis to highlight the proportion of costs covered by Medicare versus the out-of-pocket expenses of the beneficiaries.





Dataset Variables and interpretations:

Variables	Remarks
National.Provider.Identifier	Basically, just an identifier Duplicates are present but differs on the HCPCS Code – this means they do multiple procedures
Last.Name.Organization.Name.of.the.Provider	Main Name
First.Name.of.the.Provider	When Blank – Last name is 100% an organization, company (LLC), foundation/ inc.
Middle.Initial.of.the.Provider	When Blank – Last name is 100% an organization, company (LLC), foundation/ inc.
Credentials.of.the.Provider	When Provider is an org – Credentials are blank, could be indicator of multi disciplined company When Provider is an individual – Credentials may vary from holding 1 or more discipline, even those who input blank
Gender.of.the.Provider	When Provider is an org – Gender is blank When Provider is an individual – Gender is not blank
Entity.Type.of.the.Provider	If I – Individual, O – Organization
Street.Address.1.of.the.Provider	Address 1, some did not provide specific address, instead provided hospital or clinic name
Street.Address.2.of.the.Provider	Address 2
City.of.the.Provider	City
Zip.Code.of.the.Provider	ZIP- random checked 10
State.Code.of.the.Provider	State Code
Country.Code.of.the.Provider	ALL US
Provider.Type	Specific type of service they provide
Medicare.Participation.Indicator	Indicator if provider accepts insurance claims specifically from Medicare
Place.of.Service	O – non-facility F – Facility
HCPCS.Code	Healthcare Common Procedure Coding System – can't detect a pattern
HCPCS.Description	HCPCS direction description of procedure administered. While Medicare Part D primarily covers prescription drugs, it may also cover certain medical supplies and services related to drug administration or management
HCPCS.Drug.Indicator	Indicator if the procedure involves a specific drug
Number.of.Services	Count of number of times the service was administered
Number.of.Medicare.Beneficiaries	Number of distinct Medicare beneficiaries receiving the service.

Variables	Remarks
Number.of.Distinct.Medicare.Beneficiary.Per.Day.Services	Number of distinct Medicare beneficiary/per day services. Since a given beneficiary may receive multiple services of the same type (e.g., single vs. multiple cardiac stents) on a single day, this metric removes double counting from the line service count to identify whether a unique service occurred.
Average.Medicare.Allowed.Amount	Average of the Medicare allowed amount for the service; this figure is the sum of the amount Medicare pays, the deductible and coinsurance amounts that the beneficiary is responsible for paying, and any amounts that a third party is responsible for paying. a) Amount Medicare pays: The portion that Medicare directly covers. b) Deductible: The amount the beneficiary must pay before Medicare starts covering costs. c) Coinsurance: The percentage of costs the beneficiary is responsible for after meeting the deductible. d) Third-party payments: Any amounts that another insurance or entity is responsible for paying. Importance: It's used as a benchmark for pricing and can affect how much beneficiaries and healthcare providers can expect to pay or receive for services.
Average.Submitted.Charge.Amount	Average of the charges that the provider submitted for the service.
Average.Medicare.Payment.Amount	Average amount that Medicare paid after deductible and coinsurance amounts have been deducted for the line item service
Average.Medicare.Standardized.Amount	Average amount that Medicare paid after beneficiary deductible and coinsurance amounts have been deducted for the line item service and after standardization of the Medicare payment has been applied. Standardization removes: Geographic differences in payment rates for individual services such as those that account for local wages or input prices and makes Medicare payments across geographic areas comparable, so that differences reflect variation in factors such as physicians' practice patterns and beneficiaries' ability and willingness to obtain care.

Base-R Code:

```
setwd("C:/Users/Patrick/Desktop/LIFE/School_PDD DA/Term1_Fall2024/DANA
4800/Projects/Project 1A")
#Medicare Part D Study

#load libraries needed
library(readr)

#load data set
mc_data <- read.csv("Data_D_Medicare_0920update.csv") #excel file already with original
column A removed.

#replace M/F/<Blank> to Male, Female, Org
mc_data$Gender.of.the.Provider[mc_data$Gender.of.the.Provider == "M"] <- "Male"
mc_data$Gender.of.the.Provider[mc_data$Gender.of.the.Provider == "F"] <- "Female"
mc_data$Gender.of.the.Provider[mc_data$Gender.of.the.Provider == ""] <- "Org"

#One Categorical: Entity.Type.of.the.Provider
table_entity <- round(proportions(table(mc_data$Entity.Type.of.the.Provider)),4)
pie(table_entity,
    main = "Pie Chart: Individual vs Organization",
    clockwise = TRUE,
    radius = 1,
    col = c("lightgreen","lightpink"),
    labels = c("Individual (95.62%)","Organization (4.38%)"))

#One Categorical: Medicare Participation
table_participation <-
round(proportions(table(mc_data$Medicare.Participation.Indicator)),4)
pie(table_participation,
    main = "Pie Chart: Medicare Participation",
    clockwise = FALSE,
    radius = 1,
    col = c("lightpink","lightgreen"),
    labels = c("No (0.05%)","Yes (99.95%)"))

#With minimal amount of organization, we can remove and just consider individuals
#One Categorical: Gender.of.the.Provider

table_gender <- round(proportions(table(mc_data$Gender.of.the.Provider)),4) #create
proportions from table function
table_gender <- data.frame(table_gender) #transform as dataframe
table_gender <- table_gender[-3,] #remove org

pie(table_gender$Freq,
    main = "Pie Chart: Gender Distribution",
    radius = 1,
    clockwise = TRUE,
    col = c("lightpink","lightblue"),
    labels = c("Female (27.93%)","Male (67.69%)")) #Pie Chart (Clockwise)


#Exploring the which procedures are the most common
```

```
#HCPCS.Description & Number.of.Distinct.Medicare.Beneficiary.Per.Day.Services
#There are 2029 distinct Procedures

table_HCPCS_desc <-
aggregate(mc_data$Number.of.Distinct.Medicare.Beneficiary.Per.Day.Services, by =
list(mc_data$HCPCS.Description), FUN = sum) #aggregate or pivot by unique HCPCS
description where values of distinct services are added
colnames(table_HCPCS_desc) <- c("Procedure", "Frequency") #name of columns
table_HCPCS_desc <- table_HCPCS_desc[order(table_HCPCS_desc$Frequency, decreasing =
TRUE),] #sort by ASC

par(cex.axis = 0.5) #reduce font size
options(scipen = 999) #expand y-axis tick labels
barplot(table_HCPCS_desc$Frequency,
        main = "Bar Chart: Procedure count / day",
        ylab = "Count / day",
        xlab = "",
        ylim = c(0,600000),
        names.arg = table_HCPCS_desc$Procedure,
        las = 2) #barplot (vertical)
mtext("(2089 UNIQUE procedures)", side = 3, line = 0.6, col = "gray",cex = 0.9) #Sub text
mtext("Procedure", side = 1, line = 1, adj = -.05) #move "fake" X-lab to the left

#Insight/s: There are significant outliers - let's set our thresholds
#>100,000 - most common procedures = High Freq
#In between = Medium Freq
#<10,000 - least common procedures - 1908 = Low Freq

cmn_table_HCPCS_desc <- table_HCPCS_desc[table_HCPCS_desc$Frequency > 100000,
c("Procedure","Frequency")] #filter frequency > 100k to make a new table
rows_to_remove <- which(cmn_table_HCPCS_desc$Procedure %in% c("Ground mileage, per
statute mile", "Travel allowance one way in connection with medically necessary
laboratory specimen collection drawn from home bound or nursing home bound patient;
prorated trip charge.")) #set procedures that are org based
cmn_table_HCPCS_desc <- cmn_table_HCPCS_desc[-rows_to_remove, ] #remove org based
procedures

#let's create a separate table to categorize the frequencies from High, Mid, Low to show
that the

#Insight:dataset have a majority low frequency procedures

high_freq <- table_HCPCS_desc$Frequency > 100000 #set condition: High Freq
low_freq <- table_HCPCS_desc$Frequency < 10000 #set condition: Low Freq

filter_table_HCPCS_desc <- table_HCPCS_desc  #creating a new table
filter_table_HCPCS_desc$Procedure <- factor(ifelse(high_freq, "High",
                                            ifelse(low_freq, "Low", "Medium")),
                                            levels = c("High", "Medium", "Low"))
#categorize each procedure by frequency

table_filtered <- table(filter_table_HCPCS_desc$Procedure) #create a frequency table

#One Categorical Variable: Frequency Category
barplot(table_filtered,
```

```
        main = "Bar Chart: Categorical Frequency Distribution ",
        xlab = "Frequency Category",
        ylab = "Count of Procedures",
        col = c("lightgreen","lightblue","pink"),
        ylim = c(0,2000))
mtext("High (>100k) | Med (<100k & >10k) | Low (<10k)", side = 3, line = 0.5, col =
"gray",cex = 0.9)

#We have a clear visual that most procedures are low frequency.
#Let's shift our focus to the high frequency procedures (8 procedures)

par(cex.axis = 0.7) #adjust text size
par(mar = c(10, 5, 5, 2)) #adjust chart margins
barplot(cmn_table_HCPCS_desc$Frequency,
        main = "Bar Plot: High Freqency Procedures",
        xlab = "",
        ylab = "Count / day",
        names.arg = cmn_table_HCPCS_desc$Procedure,
        las= 2,
        horiz = FALSE,
        ylim = c(0,600000),
        col = "lightgreen")
mtext("High (> 100k/day)", side = 3, line = 1, col = "gray")
mtext("Procedure", side = 1, line = 1, adj = 0) #new x-label


#Now that we've identified the most common procedures - let's check the distribution of
genders to check if both genders even do them.
#Create new table and group by Gender / Procedure / Frequency / Price
#Gender.of.the.Provider
#HCPCS.Description
#Number.of.Distinct.Medicare.Beneficiary.Per.Day.Services
#Average.Medicare.Standardized.Amount

new_mc_data <- mc_data[, c("Gender.of.the.Provider","HCPCS.Code",
"Number.of.Distinct.Medicare.Beneficiary.Per.Day.Services",

"Average.Medicare.Standardized.Amount","Average.Submitted.Charge.Amount")] #create new
table with 4 columns needed

grouped_new_mc_data <- aggregate(new_mc_data$Gender.of.the.Provider, by = list(Procedure
= new_mc_data$HCPCS.Code), FUN = table) #group by or pivot by HCPCS Codes / "Procedures"

cmn_procedures <- c("99214",
                    "99213",
                    "36415",
                    "99232",
                    "80053",
                    "82306",
                    "97110",
                    "99233") #the HCPCS codes from table_filtered

fil_cmnproc_newmcdata <- grouped_new_mc_data[grouped_new_mc_data$Procedure %in%
cmn_procedures, ] #pipe the common HCPCS codes
```

```
female_counts <- sapply(fil_cmnproc_newmcdata$x, function(g) g["Female"]) #extract the
Female counts from the aggregated table
male_counts <- sapply(fil_cmnproc_newmcdata$x, function(g) g["Male"]) #extract the Male
counts from the aggregated table

total_counts <- female_counts + male_counts #add male and female counts

sorted_indices <- order(total_counts, decreasing = TRUE) #sort Descending

sorted_procedures <- cmn_procedures[sorted_indices] #sort procedures
sorted_female_counts <- female_counts[sorted_indices] #sort female counts
sorted_male_counts <- male_counts[sorted_indices] #sort male counts



#Two Cat: Gender vs Top Frequent Procedure
par(cex.axis = 0.7)  # adjust text size
par(mar = c(5, 5, 5, 2))  # adjust chart margins
barplot(rbind(sorted_female_counts, sorted_male_counts), beside = TRUE,
        names.arg = sorted_procedures,
        col = c("pink", "lightblue"),
        legend = c("Female", "Male"),
        las = 2,
        main = "S-b-S Barplot: Gender vs High Frequency Procedures",
        xlab = "High Frequent Procedure",
        ylab = "Count",
        ylim = c(0, 1600)) #create a side by side bar plot of high frequency procedures
vs gender
mtext("(Both Genders perform: High Frequency Procedures)", side = 3, line = 0.8, col =
"grey") #sub main title

#We can observe that the top frequency procedures are being performed by both genders
#Additionally, it seems that the two variables are not independent to each other

#Now let's see how much medical providers typically charge for those top frequency
procedures
#side-by-side bar chart to compare male and female charge code per high frequency
procedure

#Group by mean - there seem to be outliers, mean value may be skewed to the right.
charge_data <- mc_data[, c("HCPCS.Code", "Average.Submitted.Charge.Amount")] #create new
table
grouped_charge_data <- aggregate(charge_data$Average.Submitted.Charge.Amount, by =
list(Procedure = charge_data$HCPCS.Code), FUN = "mean") #group by/ Pivot by HCPCS code
fil_chrg_data <- new_mc_data[new_mc_data$HCPCS.Code %in% cmn_procedures, ] #filter to
only have high freq procedures and create new table
fil_grpchrgdata_chargedata<- grouped_charge_data[grouped_charge_data$Procedure %in%
cmn_procedures, ]

#using median - considered using median dew to right hand outliers
desired_order <- c("99213", "99214", "99232", "99233", "36415", "97110", "80053",
"82306")

median_subcharges <- aggregate(Average.Submitted.Charge.Amount ~ HCPCS.Code +
Gender.of.the.Provider,
```

```
                                data = fil_chrg_data,
                                FUN = median, na.rm = TRUE) #group by median

median_subcharges_wide <- reshape(median_subcharges,
                                  idvar = "HCPCS.Code",
                                  timevar = "Gender.of.the.Provider",
                                  direction = "wide") #transform into 4 columns (gender
based)

# Combine the male and female median into a single vector for sorting
# Assuming the male median is always in the first column and female in the second
median_subcharges_wide$TotalMedian <- rowMeans(median_subcharges_wide[, 2:3], na.rm =
TRUE)

# Sort by TotalMedian in descending order and then by desired order
median_subcharges_wide_sorted <-
median_subcharges_wide[order(median_subcharges_wide$TotalMedian,
match(median_subcharges_wide$HCPCS.Code, desired_order)), ]

median_subcharges_wide_sorted$HCPCS.Code <-
factor(median_subcharges_wide_sorted$HCPCS.Code,
                                                    levels = desired_order) #sort by
desired_order

# Clean column names for plotting
names(median_subcharges_wide_sorted) <- gsub("Average.Submitted.Charge.Amount.", "",
names(median_subcharges_wide_sorted))

median_subcharges_wide_sorted <- subset(median_subcharges_wide_sorted, select = -4)
#remove total median column
median_subcharges_wide_sorted <- subset(median_subcharges_wide_sorted, select = -4)
median_subcharges_wide_sorted <-
median_subcharges_wide_sorted[order(match(median_subcharges_wide_sorted$HCPCS.Code,
desired_order)), ] #force sort

barplot(
  t(as.matrix(median_subcharges_wide_sorted[, -c(1)])), # Exclude the HCPCS code
  beside = TRUE,
  col = c("lightpink", "lightblue"),
  main = "(Median) Medical Charge vs Gender PER Procedure",
  xlab = "High Frequency Procedure",
  ylab = "Medical Charge ($)",
  ylim = c(0, 300),
  names.arg = median_subcharges_wide_sorted$HCPCS.Code,
  legend.text = c("Female", "Male"),
  args.legend = list(x = "topright"))

#stacked bar of submitted charge vs standardized medicare payment of male and female
median_charges_payment <- aggregate(list(Average.Medicare.Standardized.Amount =
fil_chrg_data$Average.Medicare.Standardized.Amount,
                                         Average.Submitted.Charge.Amount =
fil_chrg_data$Average.Submitted.Charge.Amount),
                                    by = list(HCPCS.Code = fil_chrg_data$HCPCS.Code,
                                              Gender.of.the.Provider =
fil_chrg_data$Gender.of.the.Provider),
```

```
                                    FUN = median, na.rm = TRUE) #aggregate both medical
charge and std. medicare by Median

plot_data_male <- median_charges_payment[,
c("Gender.of.the.Provider","Average.Medicare.Standardized.Amount",
"Average.Submitted.Charge.Amount")] #create new table for male
plot_data_male <- plot_data_male[plot_data_male == "Male",] #filter to only have male

unique_hcpcs_codes <- unique(median_charges_payment$HCPCS.Code) #get unique HCPCS codes
rownames(plot_data_male) <- unique_hcpcs_codes #replace rownames w/ HCPCS codes

plot_matrix_male <- t(as.matrix(plot_data_male)) #transpose matrix
plot_matrix_male <- plot_matrix_male[-1, ] #remove first row (male)

barplot(plot_matrix_male,
        col = c("green", "red"),
        names.arg = unique_hcpcs_codes,
        xlab = "High Frequency Procedure",
        ylab = "Medical Charge ($)",
        main = "Stacked Bar: Medical Fee Breakdown",
        las = 2,
        ylim = c(0,400)) #barplot (2-way)
legend("topright", legend = c("Std. Insurance Coverage", "Medical Charge"),
       fill = c("green", "red"),
       cex = 0.8) #create legend
mtext("Based on Male medical practitioner", side = 3, line = .7, col = "grey") #sub main
title

plot_data_female <- median_charges_payment[,
c("Gender.of.the.Provider","Average.Medicare.Standardized.Amount",
"Average.Submitted.Charge.Amount")]
plot_data_female <- plot_data_female[plot_data_female == "Female", ]

rownames(plot_data_female) <- unique_hcpcs_codes

plot_matrix_female <- t(as.matrix(plot_data_female))
plot_matrix_female <- plot_matrix_female[-1, ]

barplot(plot_matrix_female,
        col = c("green", "red"),
        names.arg = unique_hcpcs_codes,
        xlab = "High Frequency Procedure",
        ylab = "Medical Charge ($)",
        main = "Stacked Bar: Medical Fee Breakdown",
        las = 2,
        ylim = c(0,400)) #barplot (2-way)
legend("topright", legend = c("Std. Insurance Coverage", "Medical Charge"),
       fill = c("green", "red"),
       cex = 0.8) #create legend
mtext("Based on FEMALE medical practitioner", side = 3, line = .6, col = "grey") #sub
main title


#getting the chi-square statistics for Gender vs their count per procedure
```

```
gdr_count_data <- matrix(c(
  304, 71, 28, 115, 1384, 1293, 609, 431,
  203, 49, 23, 144, 900, 768, 268, 176
), nrow = 2, byrow = TRUE)

print(chisq.test(gdr_count_data))

#END
#Back-up Data Analysis (Patrick Salazar)

# mean average submitted charge amount vs high frequency (Back-up)
# Insert after line 196
# barplot(fil_grpchrgdata_chargedata$x,
#         names.arg = fil_grpchrgdata_chargedata$Procedure,
#         main = "Bar Graph: Medical Charge vs High Frequency Procedure",
#         xlab = "High Frequency Procedures",
#         ylab = "Charge Amount ($)",
#         ylim = c(0, 250),
#         col = "lightgreen")

#Group by median submitted charge vs high frequency (Back-up)
#
# grouped_charge_data_median <- aggregate(charge_data$Average.Submitted.Charge.Amount, by
= list(Procedure = charge_data$HCPCS.Code), FUN = "median") #group by HCPCS code, take
median.
# fil_grpchrgdata_chargedata_median<-
grouped_charge_data_median[grouped_charge_data_median$Procedure %in% cmn_procedures, ]
#filter to only show the high frequency procedures and create a new table
#
# barplot(fil_grpchrgdata_chargedata_median$x,
#         names.arg = fil_grpchrgdata_chargedata_median$Procedure,
#         main = "Bar Graph: Medical Charge vs High Frequency Procedure",
#         xlab = "High Frequency Procedures",
#         ylab = "Medical Charge($)",
#         ylim = c(0, 250),
#         col = "lightgreen")

#side-by-side bar chart to compare male and female and standardized charge amount (back-
up)
# Insert after 196
# mean_charges <- aggregate(Average.Medicare.Standardized.Amount ~ HCPCS.Code +
Gender.of.the.Provider,
#                           data = fil_chrg_data,
#                           FUN = mean, na.rm = TRUE) #pivot by mean
#
# rows_to_remove2 <- which(mean_charges$Gender.of.the.Provider %in% "Org")
# mean_charges <- mean_charges[-rows_to_remove2, ] #remove org
#
# mean_charges_wide <- reshape(mean_charges,
#                              timevar = "Gender.of.the.Provider",
#                              idvar = "HCPCS.Code",
#                              direction = "wide")
# barplot(
#   t(as.matrix(mean_charges_wide[, -1])),
#   beside = TRUE,
```

```
#   col = c("lightblue", "lightpink"),
#   main = "Average Standardized Medicare vs Gender and HCPCS Code",
#   xlab = "HCPCS Code",
#   ylab = "Average Medicare Standardized Amount",
#   names.arg = mean_charges_wide$HCPCS.Code,
#   legend.text = c("Male", "Female"))

#using the high freq procedure - check Submitted Charge Amount vs Gender per Procedure
#using mean (backup)
# mean_subcharges <- aggregate(Average.Submitted.Charge.Amount ~ HCPCS.Code +
Gender.of.the.Provider,
#                            data = fil_chrg_data,
#                            FUN = mean, na.rm = TRUE)
#
# mean_subcharges <- mean_subcharges[-rows_to_remove2, ]
#
# mean_subcharges_wide <- reshape(mean_subcharges,
#                                 timevar = "Gender.of.the.Provider",
#                                 idvar = "HCPCS.Code",
#                                 direction = "wide")
# barplot(
#   t(as.matrix(mean_subcharges_wide[, -1])),
#   beside = TRUE,
#   col = c("lightblue", "lightpink"),
#   main = "Average Submitted Charge vs Gender & Procedure",
#   xlab = "HCPCS Code",
#   ylab = "Average Submitted Charge",
#   ylim= c(0,300),
#   names.arg = mean_subcharges_wide$HCPCS.Code,
#   legend.text = c("Male", "Female"),
#   args.legend = list(x=25, y= 300, horiz=T))

# #Two Categorical: Gender.of.the.Provider & State.Code.of.the.Provider
# insert after line 45
# table_gp_sp <- table(mc_data$Gender.of.the.Provider,mc_data$State.Code.of.the.Provider)
#create frequency table
# table_gp_sp <- table_gp_sp[, order(colSums(table_gp_sp),decreasing = TRUE)] #sort in
ASC
# table_gp_sp <- data.frame(table_gp_sp) #transform as dataframe
# table_gp_sp <- table_gp_sp[table_gp_sp$Var1 != "Org", ] #remove org
#
# reshaped_table_gp_sp <- reshape(table_gp_sp, timevar = "Var1", idvar = "Var2",
direction = "wide") #reshape to wide format
# colnames(reshaped_table_gp_sp) <- c("State", "Male", "Female") #name the columns
# rownames(reshaped_table_gp_sp) <- reshaped_table_gp_sp$State #name the rows per state
# reshaped_table_gp_sp$State <- NULL #remove State column
# reshaped_matrix_gp_sp <- t(as.matrix(reshaped_table_gp_sp)) #transpose the matrix
#
# par(cex.axis = 0.8) #reduce font size
# barplot(reshaped_matrix_gp_sp,
#         main = "Stacked Bar: Gender Distribution per state",
#         beside = FALSE,
#         ylim = c(0,4000),
#         xlab = "US State Code",
#         ylab = "Frequency",
```

```
#         las = 2,
#         col = c("lightpink","lightblue")) #barplot (vertical)


#Back-up (Derica Gooden)
#insert anywhere
#scatterplot to check correlation of medicare allowed amount and submitted charge amount
# mc_data <- read.csv("Data_D_Medicare_0920update.csv")
# mc_data <- na.omit(mc_data)
# plot(mc_data$Average.Medicare.Allowed.Amount, mc_data$Average.Submitted.Charge.Amount,
#      main = "Scatterplot of Medicare allowed amount vs submitted charge amount",
#      xlab = "Average submitted charge amount($)", ylab = "Average medicare allowed
amount ($)",
#      col = "blue", pch = 17)

# covariance to check the relationship between the two mentioned numerical variable.
# cov(mc_data$Average.Medicare.Allowed.Amount, mc_data$Average.Submitted.Charge.Amount)
```

References:

- Medicare (United States). Wikipedia. 2024. https://en.wikipedia.org/wiki/Medicare_(United_States)
- **The Centers for Medicare and Medicaid Services Office of Enterprise Data and Analytics**. Centers for Medicare & Medicaid Services. Medicare Fee-For-Service Provider Utilization & Payment Data Physician and Other Supplier Public Use File: A Methodological Overview. April 7, 2014. Last Updated: September 22, 2020.
- **Christov-Moore L, Simpson E, Coudé G, Grigaityte K, Iacoboni M, Ferrari P.** Empathy: gender effects in brain and behavior. *Neuroscience & Biobehavioral Reviews*. 2014 Oct;46(4):604-627. doi:10.1016/j.neubiorev.2014.09.001. PMID: 25236781; PMCID: PMC5110041.