DATA 607 Final Project: Hospital Rating and Income

Jian Quan Chen

2023-04-28

Introduction

For this project, I will be analyzing the Overall Hospital Quality Star Rating provided by the Center for Medicaid and Medicare Services (CMS). The CMS dataset contains ratings from over 5,000 hospital across the United States. It summarizes 5 keys measures (mortality rate, safety, readmission, patient experience, and timely & effective care) into a 5 star rating system.

Along with the hospital data, I will also examine a dataset from the Census Bureau that contains the median income of households by zip code. By merging these two datasets based on the zip code field, I hope to gain insights into the relationship between hospital quality ratings and income levels in the areas where the hospitals are located. In addition, I will explore any trends or patterns that may exist among hospitals with high and low ratings.

Being as I work in a hospital, I think it is important to understand why there are discrepancies between hospitals quality in healthcare. Identifying these factors can not only improve hospital ratings but improve patient outcomes as well.

Sources:

- “https://data.cms.gov/provider-data/sites/default/files/resources/092256becd267d9eeccf73bf7d16c46b_1681243512/Hospital_General_Information.csv (https://data.cms.gov/provider-data/sites/default/files/resources/092256becd267d9eeccf73bf7d16c46b_1681243512/Hospital_General_Information.csv)”
- “https://data.census.gov/table?q=median+income&g=010XX00US$8600000&tid=ACSST5Y2021.S1901 (https://data.census.gov/table?q=median+income&g=010XX00US$8600000&tid=ACSST5Y2021.S1901)”

Importing the Libraries

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.2     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.2     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force all conflicts to become errors
```

Extracting the Sources

```
#Extracting hospital data
hospital_df_raw <- read.csv("https://data.cms.gov/provider-data/sites/default/files/resources/092256becd267d9eeccf73bf7d16c46b_1681243512/Hospital_General_Information.csv")

#Extracting income data - The dataset was large and the page crashed when I tried viewing the table on the website so it was easier to download it and upload it to github.
income_df_raw <- read.csv("https://raw.githubusercontent.com/LeJQC/MSDS/main/DATA%20606/Final%20Project/Income%20data.csv")
```

Data Transformation

```
# Tidying the hospital data frame and selecting the columns I need
hospital_df <- hospital_df_raw %>%
  select("Facility.Name","City", "State", "ZIP.Code", "Hospital.Type", "Hospital.Ownership", "Hospital.overall.rating") %>%
  rename(HospitalName = Facility.Name, HospitalRating = Hospital.overall.rating, ZipCode = ZIP.Code)

# Tidying the income data frame. Looking to make a data frame with 2 columns: zip code and median income
income_df <- income_df_raw %>%
  select("NAME","S1901_C01_012E") %>%
  slice(-1) %>%
  mutate(NAME = as.numeric(substr(NAME, nchar(NAME) - 4, nchar(NAME))),
         S1901_C01_012E = as.numeric(S1901_C01_012E)) %>%
  rename(ZipCode = NAME, Income = S1901_C01_012E)
```

```
## Warning: There was 1 warning in `mutate()`.
## ℹ In argument: `S1901_C01_012E = as.numeric(S1901_C01_012E)`.
## Caused by warning:
## ! NAs introduced by coercion
```

```
# Merging both the hospital and income data frames together
merged_df <- merge(income_df, hospital_df, by ="ZipCode")

# Cleaning out the NA values
ratings_df <- merged_df %>%
  filter(HospitalRating != "Not Available", Income >= 0) %>%
  group_by(HospitalRating) %>%
  mutate(HospitalRating = as.numeric(HospitalRating))
```

Summary Statistics

```
# Summary statistics by rating(`group1`)
stats <- psych::describeBy(ratings_df$Income, group = ratings_df$HospitalRating, mat = TRUE)
rownames(stats) <- NULL
stats <- stats %>%
  select(c(group1, n, mean, sd, median, min, max, range, se)) %>%
  rename("Rating" = "group1")

knitr::kable(stats)
```

Rating	n	mean	sd	median	min	max	range	se
1	180	59232.96	27205.07	53540.0	16764	160890	144126	2027.7458
2	663	61283.45	24826.76	54567.0	20239	174419	154180	964.1914
3	851	62800.59	23562.27	57324.0	19083	194462	175379	807.7041
4	851	68834.50	25469.45	62054.0	15833	232400	216567	873.0817
5	408	75001.75	30628.49	66046.5	11404	216286	204882	1516.3362

Surprisingly, the zip code with the lowest income has a hospital with 5 stars.

Data Visualization

To get a general sense of income and how hospitals are rated by state, let's plot this using `ggmap`.

Plotting Average Rating by State

```
library(ggmap)
```

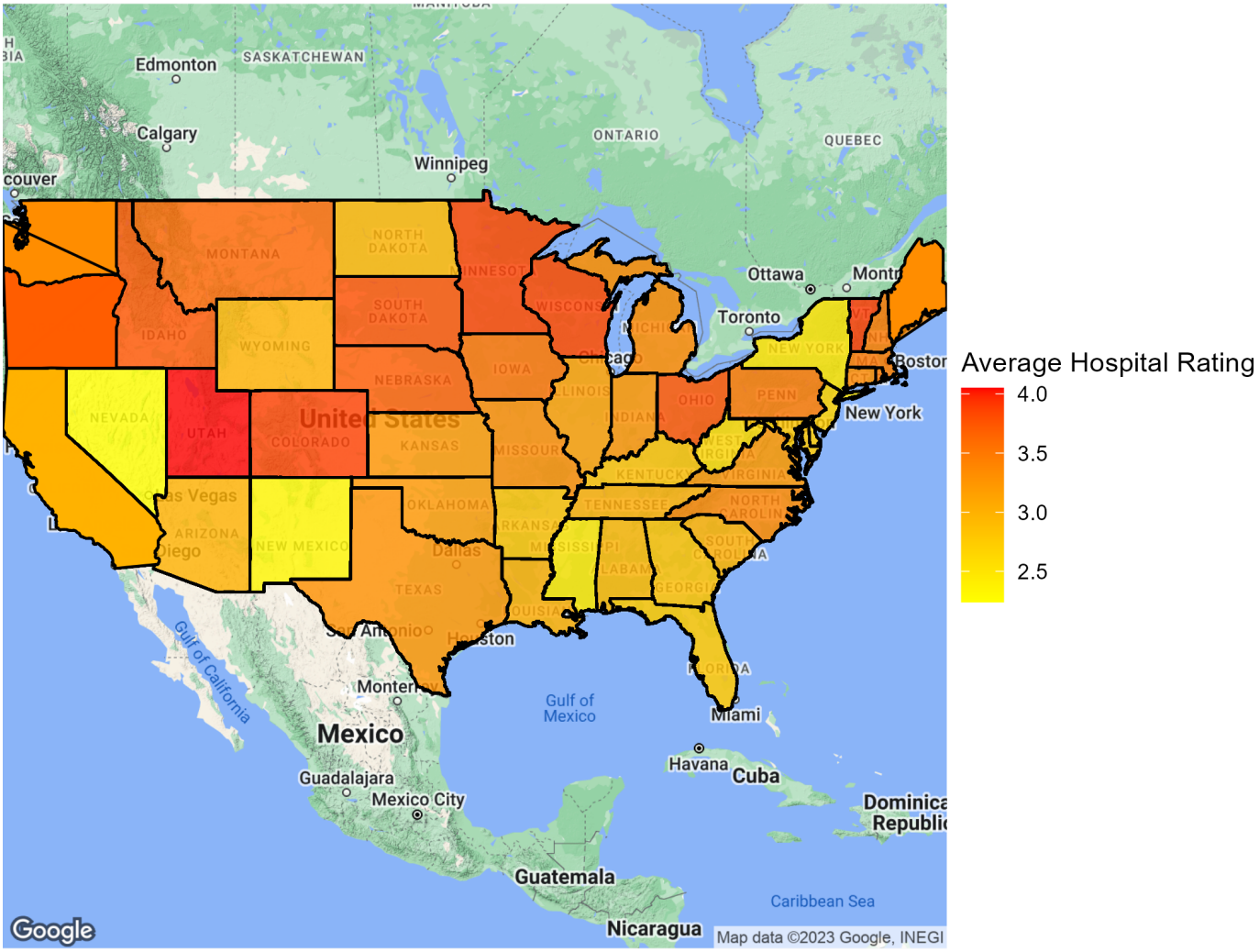
```
## ℹ Google's Terms of Service: ]8;;https://mapsplatform.google.com<https://mapsplatform.google.com>]8;;
## ℹ Please cite ggmap if you use it! Use `citation("ggmap")` for details.
```

```
# Creating an average hospital rating by state
avg_state <- ratings_df %>%
  group_by(State) %>%
  mutate(avg_rating = round(mean(HospitalRating),2)) %>%
  mutate(avg_income = round(mean(Income),2))

# Using the Google Maps API
register_google(key = "AIzaSyBCuGJujBhbM3SiJMW07WCEM0YFAKvG320")

# Getting the US map from Google Maps API
us_map <- get_map(location = "united states", zoom = 4, maptype = "terrain")
```

```
## ℹ <]8;;https://maps.googleapis.com/maps/api/staticmap?center=united%20states&zoom=4&size=640x640&scale=2&maptype=terrain&
language=en-EN&key=xxxhttps://maps.googleapis.com/maps/api/staticmap?center=united%20states&zoom=4&size=640x640&scale=2&mapt
ype=terrain&language=en-EN&key=xxx]8;;>
## ℹ <]8;;https://maps.googleapis.com/maps/api/geocode/json?address=united+states&key=xxxhttps://maps.googleapis.com/maps/ap
i/geocode/json?address=united+states&key=xxx]8;;>
```

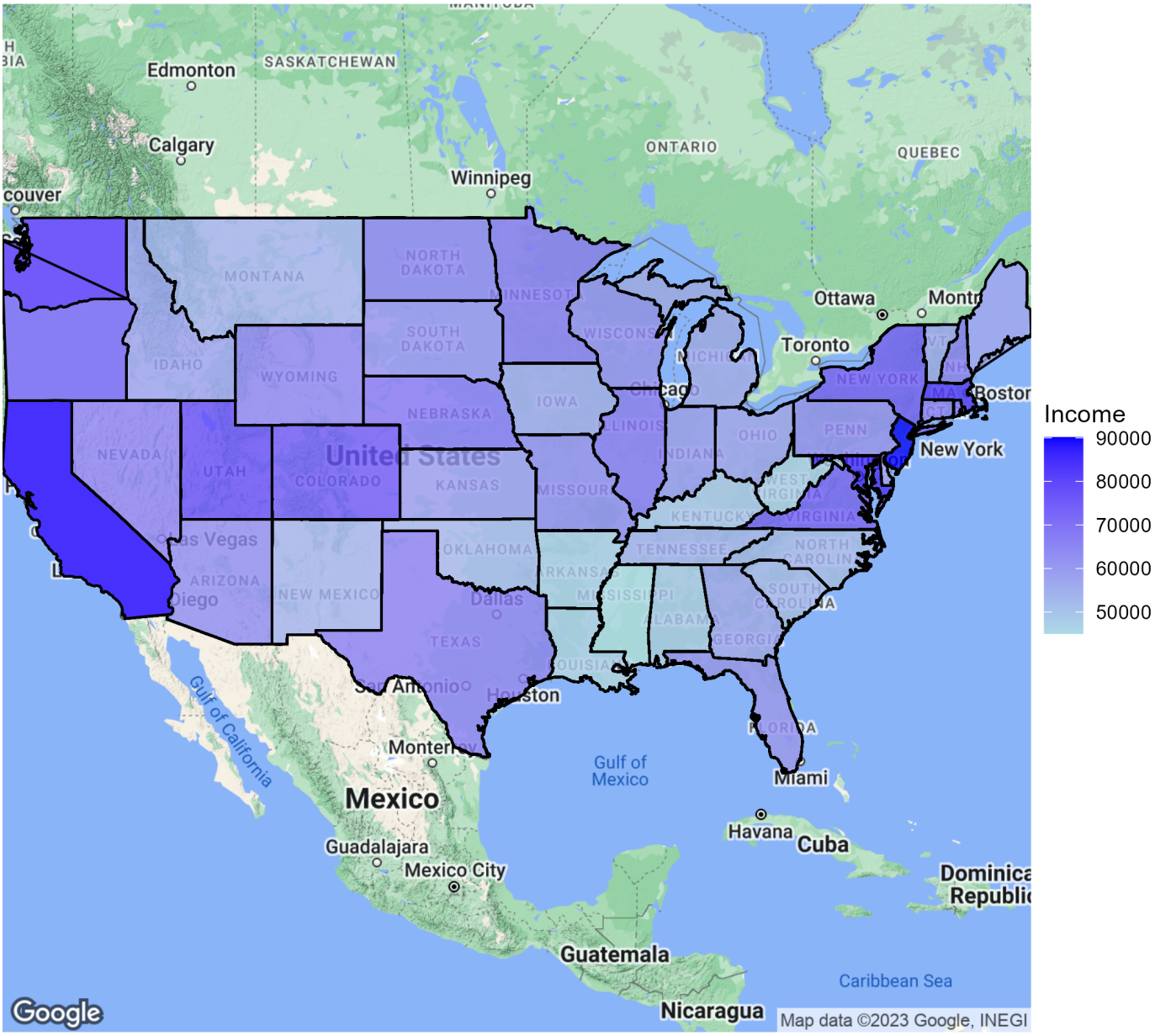
```
# This has the lat and long of all the states
mapdata <- map_data("state")

# Create lookup table for state abbreviations
state_lookup <- tibble(State = state.abb, region = state.name)
state_lookup$region <- tolower(state_lookup$region)

# Merge lookup table with  state ratings data frame
state_ratings <- avg_state %>%
  left_join(state_lookup, by = "State") %>%
  full_join(mapdata, by = "region") %>%
  filter(avg_rating != 0, long != 0, Income != 0)

# Plot of US and average rating
map1_rating <- ggmap(us_map) +
  geom_polygon(data = state_ratings, aes(x = long, y = lat, group = group, fill = avg_rating), color = "black", linewidth =
0.5, alpha = 0.8) +
  theme_void()

map2_rating <- map1_rating +
  scale_fill_gradient(name = "Average Hospital Rating", low = "yellow", high = "red", na.value = "grey50")

#map2_rating

# Saving the map because it took a long time to load
#ggsave("US Hospital rating.png", map2_rating)

# Plot of US and Income
map1_income <- ggmap(us_map) +
  geom_polygon(data = state_ratings, aes(x = long, y = lat, group = group, fill = avg_income), color = "black", linewidth =
0.5, alpha = 0.8)+
  theme_void()

map2_income <- map1_income +
  scale_fill_gradient(name = "Income", low = "lightblue", high = "blue", na.value = "grey50")

#map2_income
#ggsave("US Income.png", map2_income)
```

US Hospital Rating



US Hospital Rating

US Income



US Income

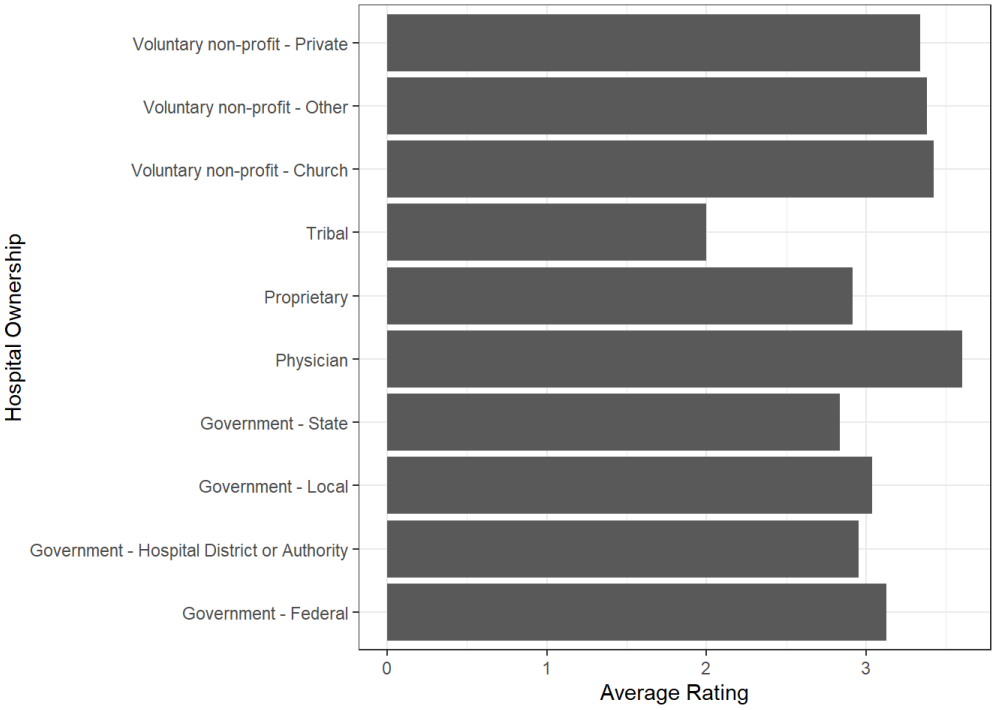
Analyzing Hospital Ownership

```
owner_table <- sort(table(ratings_df$Hospital.Ownership), decreasing = TRUE)

knitr::kable(owner_table)
```

Var1	Freq
Voluntary non-profit - Private	1454
Proprietary	530
Voluntary non-profit - Other	252
Government - Hospital District or Authority	233
Voluntary non-profit - Church	233
Government - Local	197

Var1	Freq
Government - State	24
Physician	20
Government - Federal	8
Tribal	2

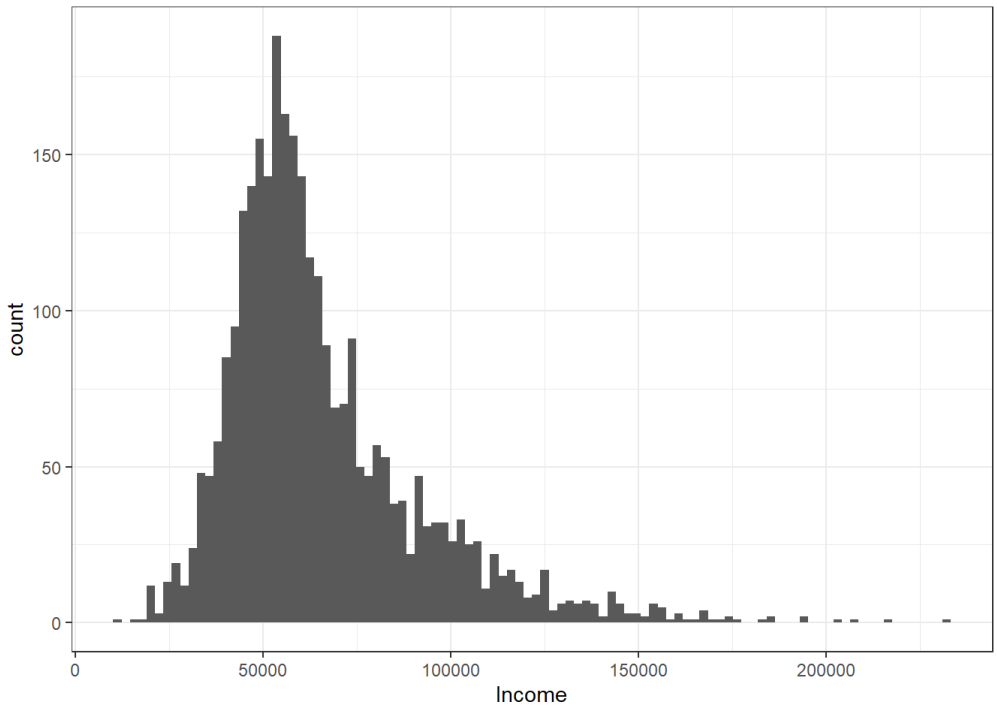
```
# Rating based on hospital ownership
ratings_df %>%
  group_by(Hospital.Ownership) %>%
  ggplot(aes(x = Hospital.Ownership, y = HospitalRating)) +
  coord_flip() +
  geom_bar(stat = "summary", fun = "mean")+
  labs(x = "Hospital Ownership", y = "Average Rating") +
  theme_bw()
```




```
# Plotting the highest and lowest rated hospital
ratings_df %>%
  filter(HospitalRating == c(1,5)) %>%
  ggplot(aes(x = Hospital.Ownership, fill = Hospital.Type)) +
  facet_wrap(~ HospitalRating)+
  geom_bar()+
  coord_flip()
```

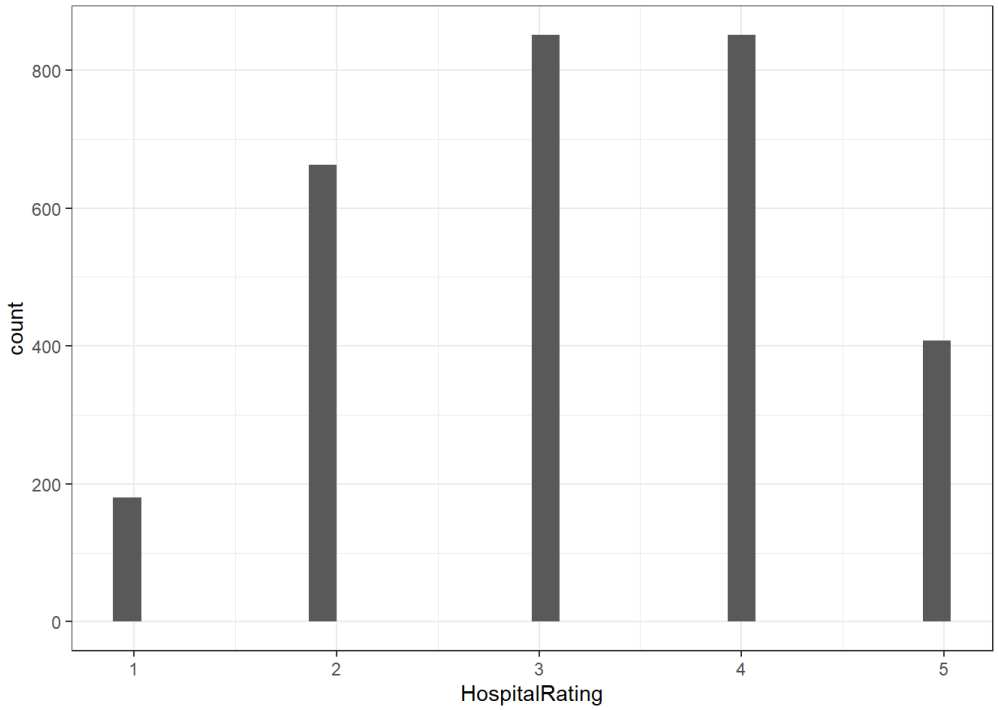
Histogram of Income and ratings

```
ratings_df %>%
  ggplot(aes(x = Income)) +
  geom_histogram(bins = 100) +
  theme_bw()
```




```
ratings_df %>%
  ggplot(aes(x = HospitalRating)) +
  geom_histogram() +
  theme_bw()
```

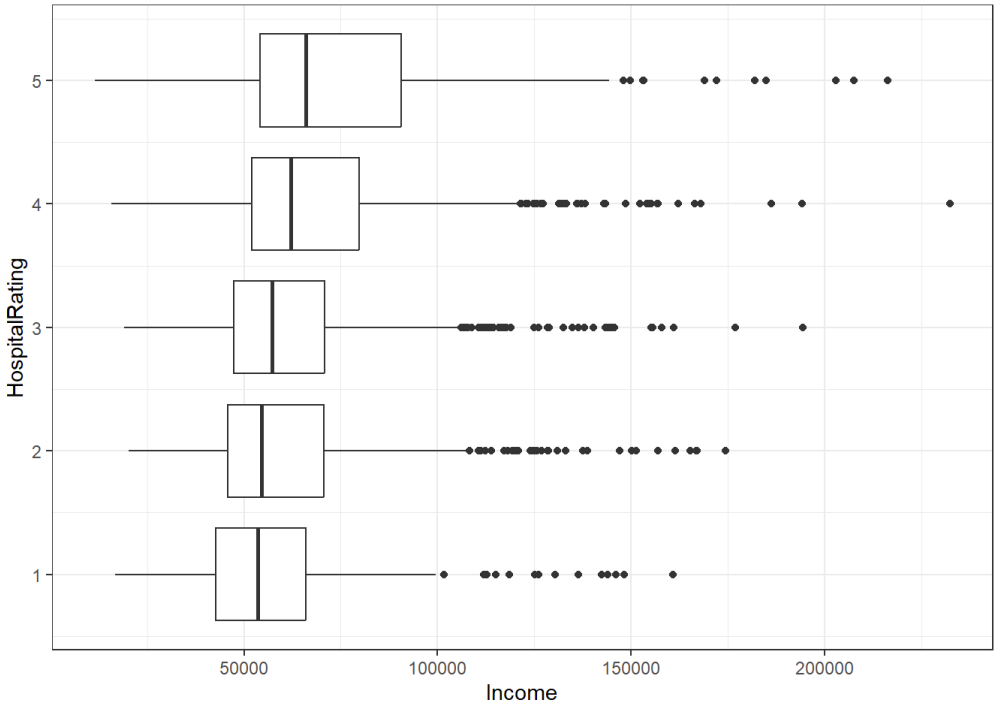
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Let's look at how these two relate!

Boxplot

```
ratings_df %>%
  ggplot(aes(x = Income, y = HospitalRating, group = HospitalRating)) +
  geom_boxplot() +
  theme_bw()
```

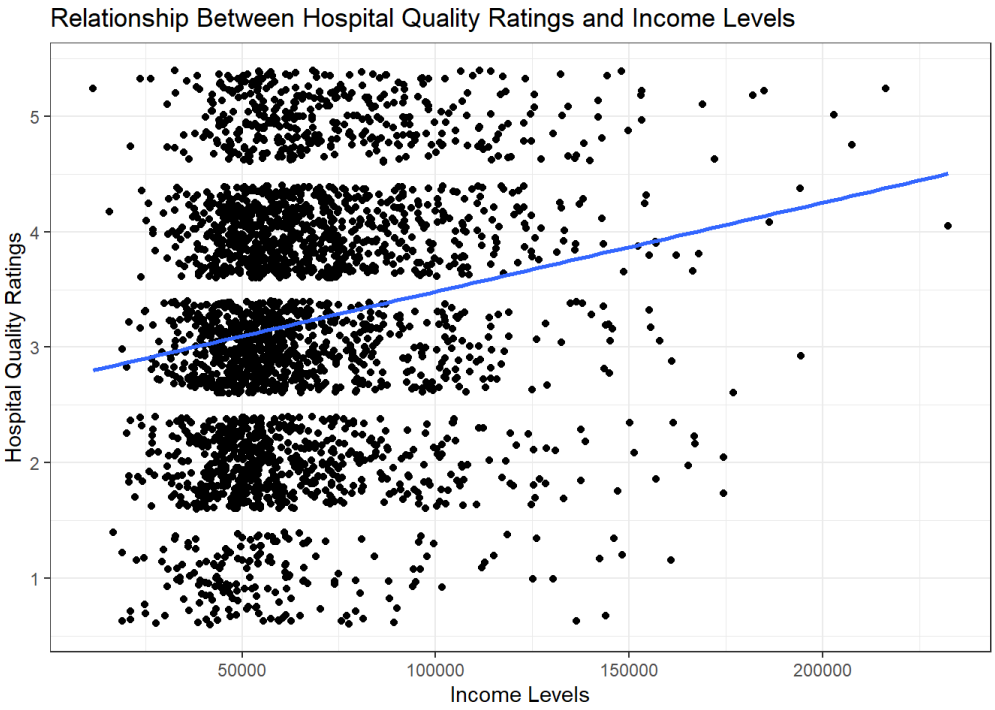


The median income seems to be increasing as the rating increases. Let's run a linear regression analysis!

Scatterplot

```
scatter_plot <- ggplot(ratings_df, aes(x = Income, y = HospitalRating, group = 1)) +
  geom_jitter() +
  labs(title = "Relationship Between Hospital Quality Ratings and Income Levels",
       x = "Income Levels", y = "Hospital Quality Ratings") +
  geom_smooth(method = lm, se = FALSE) +
  theme_bw()

scatter_plot
```

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



Linear Regression Analysis

```
# Fit a linear regression model to the data
lm_model <- lm(HospitalRating ~ Income, data = ratings_df)

# Print the summary of the linear regression model
summary(lm_model)
```

```
##
## Call:
## lm(formula = HospitalRating ~ Income, data = ratings_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.95154 -1.02962 -0.08089  0.84345  2.19988
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.712e+00  5.501e-02  49.304   <2e-16 ***
## Income      7.702e-06  7.784e-07   9.896   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.105 on 2951 degrees of freedom
## Multiple R-squared:  0.03212,    Adjusted R-squared:  0.03179
## F-statistic: 97.93 on 1 and 2951 DF,  p-value: < 2.2e-16
```

The coefficient estimate for the Income is 7.702e-06 so for every dollar increase in Income, the hospital rating increases by 7.702e-06. As for significance, the p-value is very small indicating that there is a statistically significant relationship between income and hospital rating. However, the R-squared value is very small(0.03212), indicating that only 3% of the variation in hospital ratings is accounted for by the Income variable.

Conclusion

Based on the linear regression model, there is statistically significant positive, linear correlation between hospital rating and income. Although, there is a statistically significant relationship between income and hospital rating, the small coefficient estimate indicates income only contributes a small proportion(3%) of the variation in hospital rating. This may suggest that a linear regression model is not a reliable statistical test for this dataset.

There are some important limitations to note with this analysis. The hospital dataset is not inclusive of all the hospitals in the US. Hospitals that are not associated with CMS or do not have a rating were not included in the analysis. Also, the association between income and hospitals is based on zip code and on the assumption that everyone residing in the same zip code as the hospital has the same income level.

Overall, the results from this analysis highlight the need for further research into hospital quality beyond just income. In New York, NYU and Bellevue are located in the same zip code but their hospital ratings are complete opposites. NYU's overall rating is a 5 while Bellevue's is a 1. This illustrates that income is just a small piece of the puzzle. Factors such as demographics, health insurance, and healthcare policies can all impact hospital ratings and need to be researched as well. By identifying factors that influence the variations in hospital ratings, we can lower the discrepancies in healthcare and ensure equal, high quality care to every one.