Milestone 2: Factors Associated with High Hospital Utilization Among Children with Sickle Cell Disease

Kristina Lai (kristina.lai@choa.org)

March 30, 2020

```
#Import SCD Registry Dataset



Cleaned_dataset <- read_excel("~/NSRG 741 Final
Project/N741_ProjectWorkspace-kristinawlai/hiddenfiles/Cleaned dataset_2020
03 30.xlsx")
View(Cleaned_dataset)
```

GitHub repository: https://github.com/Emory-NRSG-741-Spring-2020/N741_ProjectWorkspace-kristinawlai/tree/master/Milestone%202

Project Background

As the sickle cell disease (SCD) epidemiologist at Children's Healthcare of Atlanta (CHOA), there has been a small but significant increase in the number of patients who are admitted more than 5 times in a given calendar year. In fact, 2% of our patient population accounts for 40% of all hospital admissions, but only in the past 4-5 years. This phenomenon began in 2014 and remains elevated. We have anecdotal evidence of similar patterns from SCD programs in California, however little investigation has been done on factors that may be influencing this rise. I originally presented this data on a poster at the American Society of Hematology conference in 2018. This project aims to expand on this previous work by identifying and hopefully preventing continued escalation in higher hospital utilizers.

My objectives did not change from my original proposal in Milestone 1. However, I will be subsetting the data to only look at 2 years (2018-2019) for the exploratory analysis and then will be expanding to the full 10 years of data in the final analysis.

Total Sickle Cell Patients by Year (2010-2019)

Year	Total Patients
2010	1596
2011	1678
2012	1707
2013	1783
2014	1773
2015	1802
2016	1877

2017	1914
2018	1956
2019	2046

Data acquisition and cleaning

Data was acquired from the CHOA SCD Clinical Database for years 2010-2019. It was then subset to years 2018-2019 for exploratory analysis. The data was previously cleaned for other projects, however I cleaned several missing values for dichotomous variables (bmt_yn, deceased_yn). Specifically, I changed missing values to '0' when I was sure that they were not actually missing.

Because the categorical variables were stored in numeric codes, I applied formats in SAS before exporting the final dataset. The original coding was as follows:

Value	Genotype
1	SS
2	S BETA ZERO THAL
3	SS OR S BETA ZERO THAL
4	S BETA PLUS THAL
5	SC
6	SD
7	SE
8	S O-ARAB
9	SC HARLEM
10	S HPFH
11	FS
12	SV - OTHER
99	NON SCD

Inclusion/Exclusion

The original dataset had 3,778 unique patients in our sickle cell registry. However, I needed to exclude several patients based on various criteria. First, I removed any patients who were non-sickle cell. These patients may have mistakenly been added to the database and should be excluded. Secondly, our database has utilization data from 2010 to present, however, we have patients in our registry who were only active before that time. Because of that, only patients with at least one encounter between 2010-2019 were included (n=3,619 patients). Additionally, I excluded all encounters occurring after a curative bone marrow transplant (BMT) as well as encounters occurring 21 days prior to the transplant date in order to avoid bias from extended hospital admissions for BMT. The final cohort included 3,595 patients with a total of 117,239 encounters. When I limited the dataset to only years 2018-2019, there were 2,306 unique patients with 25,692 encounters.

Exploring the Data

To explore the initial dataset, I restricted the full 10-year cohort to only patients and encounters in 2018-2019 (as described above).

```
# SUbset dataset to only years 2018 and 2019 for this exploratory analysis.
subset<- filter(Cleaned_dataset, dsch_year > 2017)
```

I also wanted to look at all of the objects and determine what class theur were (i.e. numeric, charachter, etc).

```
str(subset)

## Classes 'tbl_df', 'tbl' and 'data.frame':    914 obs. of  17 variables:
##  $ unique_id     : num  1001 1001 1004 1004 1005 ...
##  $ genotype_char : chr  "SS" "SS" "SC" "SC" ...
##  $ genotype_other: logi  NA NA NA NA NA NA ...
##  $ sex_char      : chr  "F" "F" "F" "F" ...
##  $ bmt_yn        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ bmt_date      : POSIXct, format: NA NA ...
##  $ deceased_yn   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ deceased_date : logi  NA NA NA NA NA NA ...
##  $ dsch_year     : num  2018 2019 2018 2019 2018 ...
##  $ OPvisit       : num  9 35 3 0 0 2 4 2 5 5 ...
##  $ EDonly        : num  5 6 0 0 1 0 1 1 0 2 ...
##  $ EDany         : num  5 10 0 1 1 0 1 3 0 2 ...
##  $ IPvisit       : num  0 5 0 1 0 0 0 2 0 0 ...
##  $ IP_ACS        : num  0 0 0 1 0 0 0 0 0 0 ...
##  $ IP_pain       : num  0 4 0 0 0 0 0 1 0 0 ...
##  $ IP_elective   : num  0 1 0 0 0 0 0 0 0 0 ...
##  $ los           : num  0 8 0 1 0 0 0 2 0 0 ...
```

Next, I wanted to look at each variable to get a feel for types and means, followed by looking at the frequencies of categorical variables.

```
#Look at all variables
summary(subset)

##    unique_id    genotype_char      genotype_other   sex_char        
##  Min.   :1001   Length:914         Mode:logical   Length:914        
##  1st Qu.:1169   Class :character   NA's:914       Class :character  
##  Median :1326   Mode  :character                  Mode  :character  
##  Mean   :1332                                                       
##  3rd Qu.:1494                                                       
##  Max.   :1659                                                       
##                                                                     
##      bmt_yn     bmt_date     deceased_yn       deceased_date    dsch_year   
##  Min.   :0   Min.   :NA    Min.   :0.000000   Mode:logical   Min.   :2018
##  1st Qu.:0   1st Qu.:NA    1st Qu.:0.000000   NA's:914       1st Qu.:2018
##  Median :0   Median :NA    Median :0.000000                  Median :2018
##  Mean   :0   Mean   :NA    Mean   :0.003282                  Mean   :2018
##  3rd Qu.:0   3rd Qu.:NA    3rd Qu.:0.000000                  3rd Qu.:2019
##  Max.   :0   Max.   :NA    Max.   :1.000000                  Max.   :2019
##              NA's   :914
##     OPvisit           EDonly          EDany           IPvisit
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.00   Min.   : 0.0000
##  1st Qu.: 2.000   1st Qu.: 0.000   1st Qu.: 0.00   1st Qu.: 0.0000
##  Median : 3.000   Median : 1.000   Median : 1.00   Median : 0.0000
##  Mean   : 4.217   Mean   : 1.125   Mean   : 1.86   Mean   : 0.9628
##  3rd Qu.: 5.000   3rd Qu.: 2.000   3rd Qu.: 3.00   3rd Qu.: 1.0000
##  Max.   :35.000   Max.   :13.000   Max.   :25.00   Max.   :20.0000
##
##      IP_ACS          IP_pain         IP_elective          los
##  Min.   :0.0000   Min.   : 0.0000   Min.   : 0.0000   Min.   :  0.000
##  1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.:  0.000
##  Median :0.0000   Median : 0.0000   Median : 0.0000   Median :  0.000
##  Mean   :0.1783   Mean   : 0.3468   Mean   : 0.1083   Mean   :  3.172
##  3rd Qu.:0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.:  3.000
##  Max.   :5.0000   Max.   :11.0000   Max.   :11.0000   Max.   :178.000
##
```

```
#Look at the frequencies of categorical variables
#Genotype
freq(subset, var=genotype_char,  report.nas = FALSE,
     cumul = FALSE)
```

```
## Frequencies
## subset$genotype_char
## Type: Character
##
##                                  Freq          %
## ----------------------------- ------ --------
##                            FS    154     17.26
##              S BETA PLUS THAL     58      6.50
##              S BETA ZERO THAL      2      0.22
##                        S HPFH      3      0.34
##                      S O-ARAB      1      0.11
##                            SC    245     27.47
##                            SE      3      0.34
##                            SS    406     45.52
##        SS OR S BETA ZERO THAL     20      2.24
##                         Total    892    100.00
```

```
#Sex
freq(subset, var=sex_char,  report.nas = FALSE,
     cumul = FALSE)
```

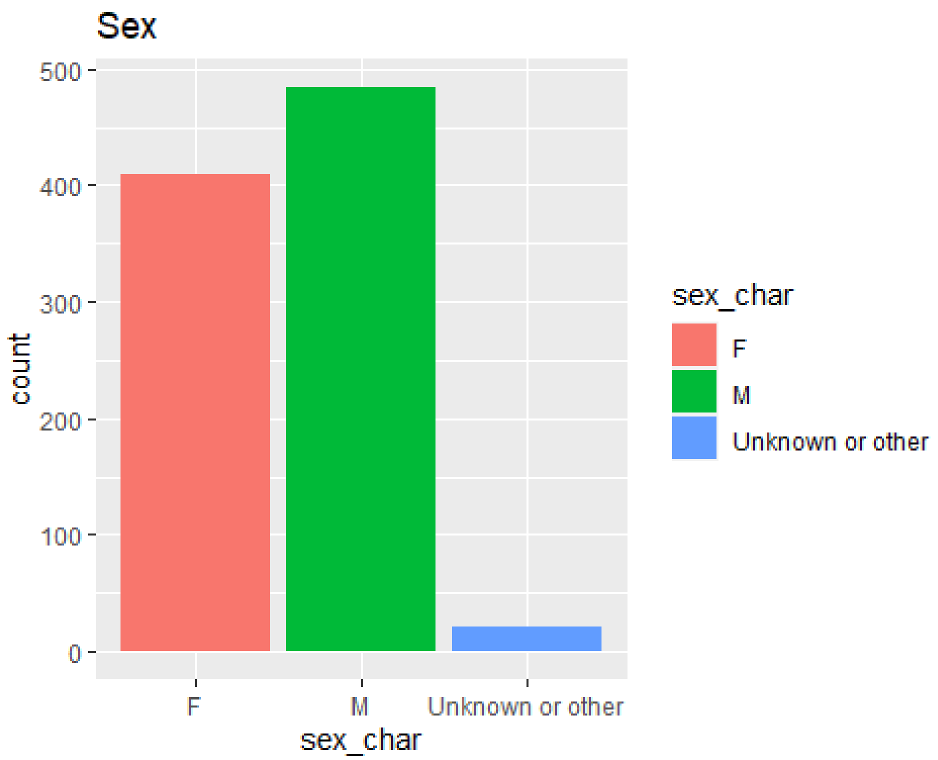
```
## Frequencies
## subset$sex_char
## Type: Character
##
##                          Freq         %
## ---------------------- ------ --------
##                      F    409    44.75
##                      M    484    52.95
##       Unknown or other     21     2.30
##                  Total    914   100.00
```

```
#Deceased
freq(subset, var=deceased_yn,  report.nas = FALSE,
     cumul = FALSE)
```

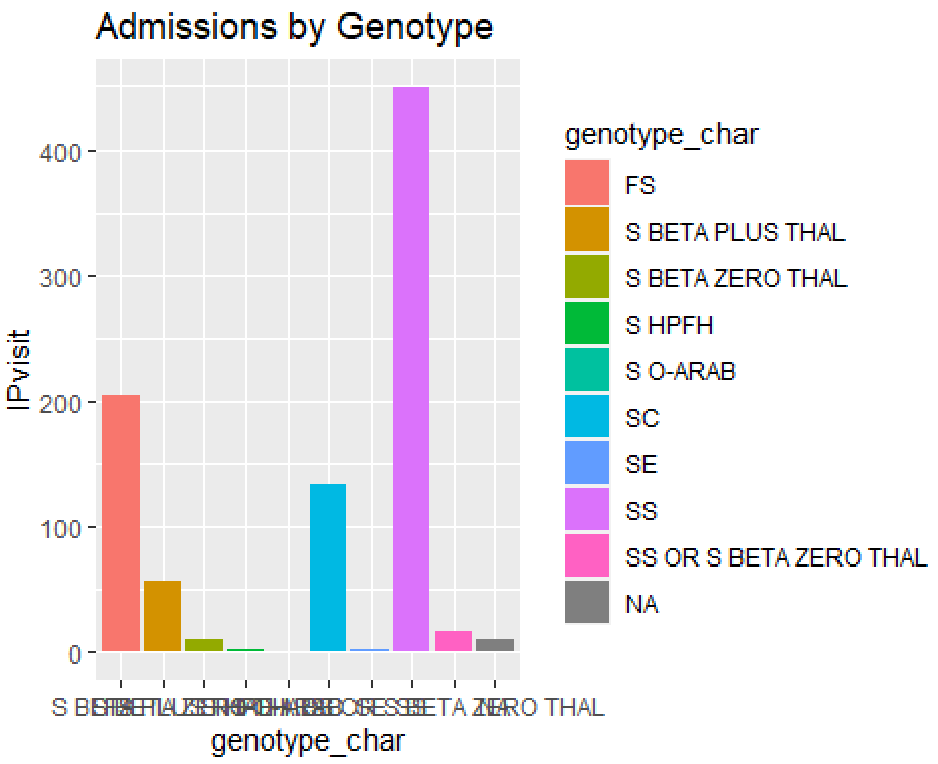
```
## Frequencies
## subset$deceased_yn
## Type: Numeric
##
##               Freq         %
## ----------- ------ --------
##           0    911    99.67
##           1      3     0.33
##       Total    914   100.00
```

I also wanted to look at plots of the vairables to understand their general distributions.

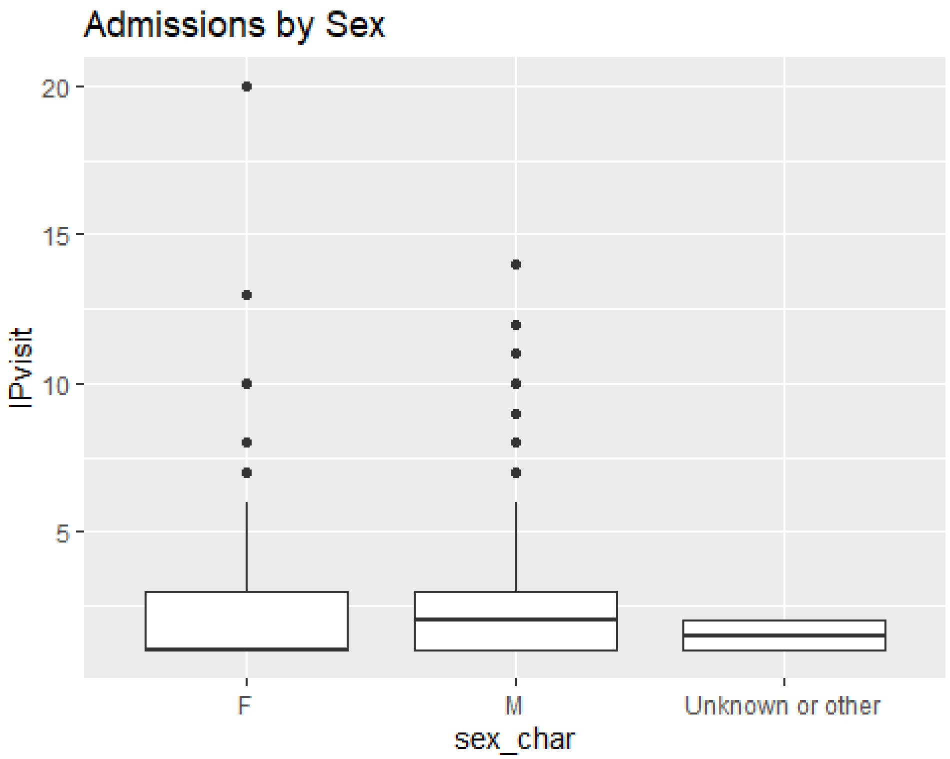
```
attach(subset)

ggplot(subset) + geom_bar(aes(x=sex_char, fill=sex_char)) + ggtitle("Sex")
```



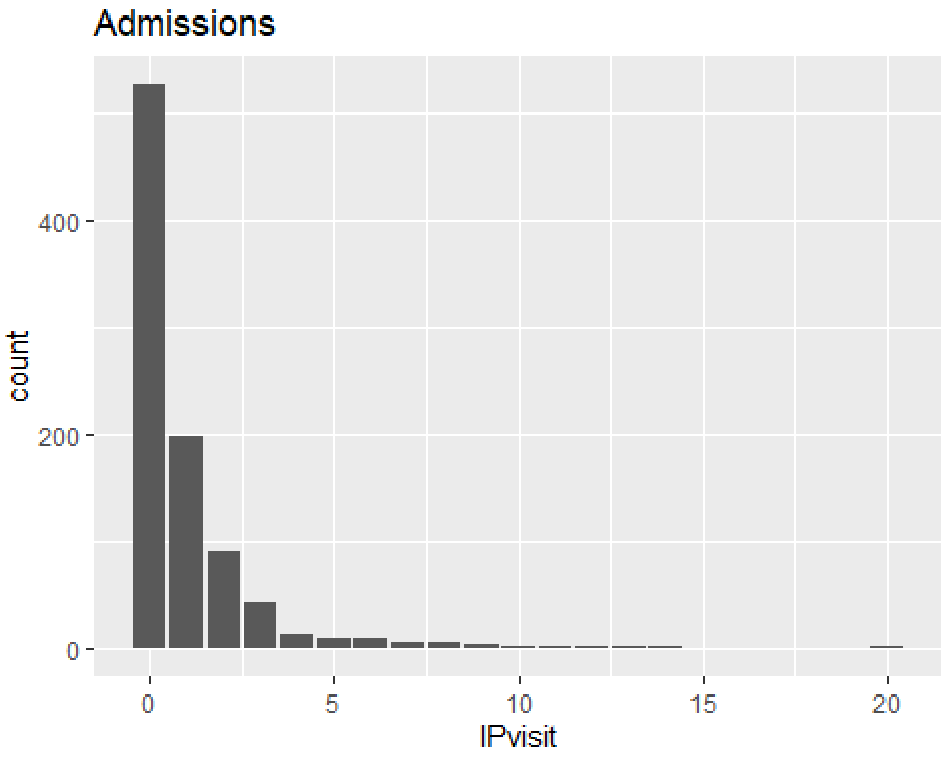
```
ggplot(subset) + geom_col(aes(x=genotype_char, y=IPvisit,
fill=genotype_char)) +ggtitle("Admissions by Genotype")
```



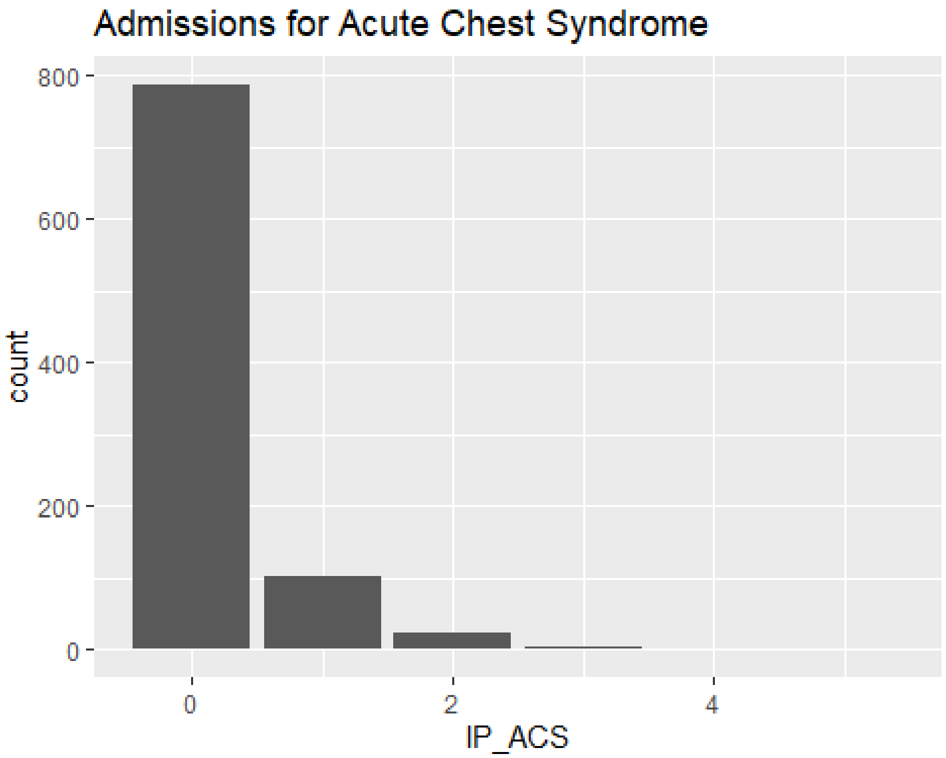
```
test <- subset(subset,IPvisit >=1)
ggplot(test, aes(x=sex_char, y=IPvisit)) + geom_boxplot()
+ggtitle("Admissions by Sex")
```



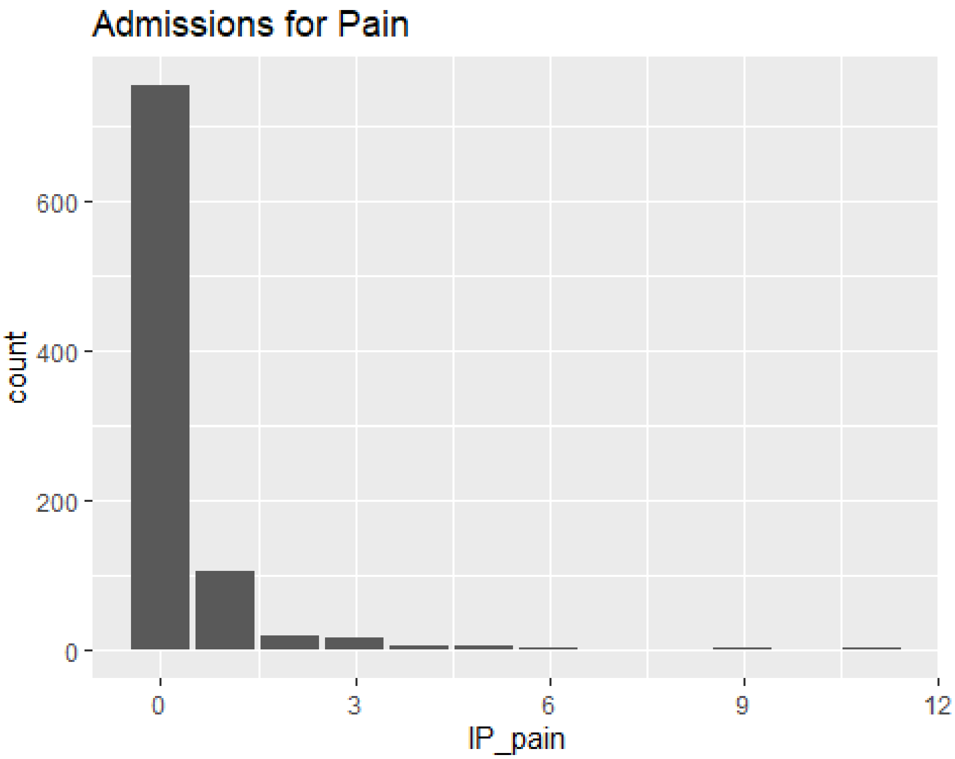
```
ggplot(subset) + geom_bar(aes(x=IPvisit))+ ggtitle("Admissions")
```



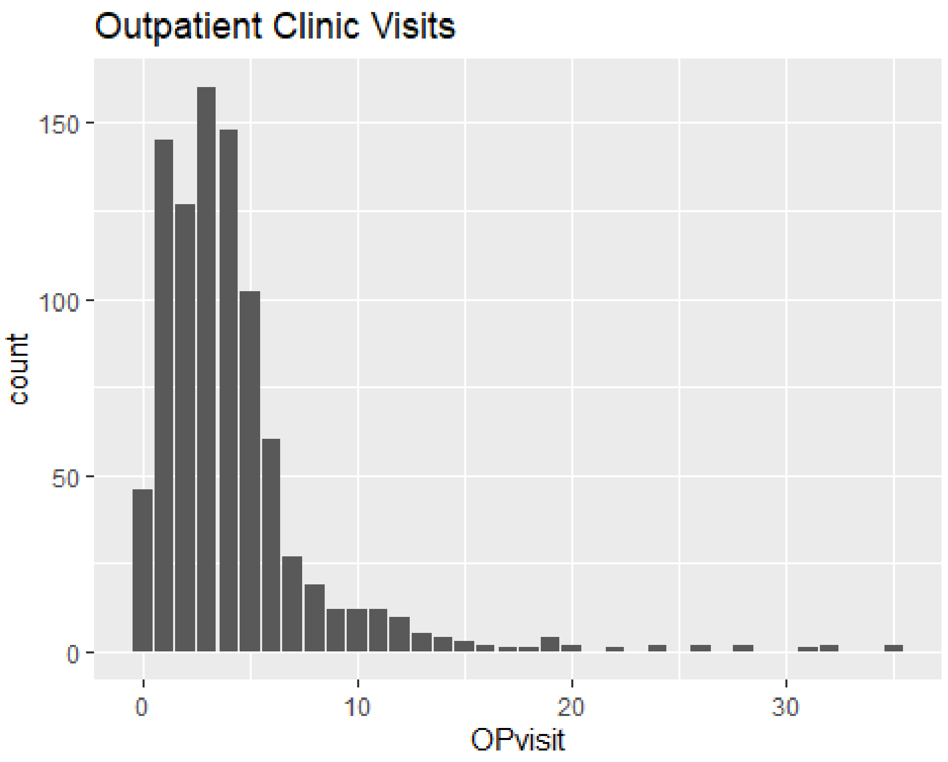
```
ggplot(subset) + geom_bar(aes(x=IP_ACS)) + ggtitle("Admissions for Acute
Chest Syndrome")
```



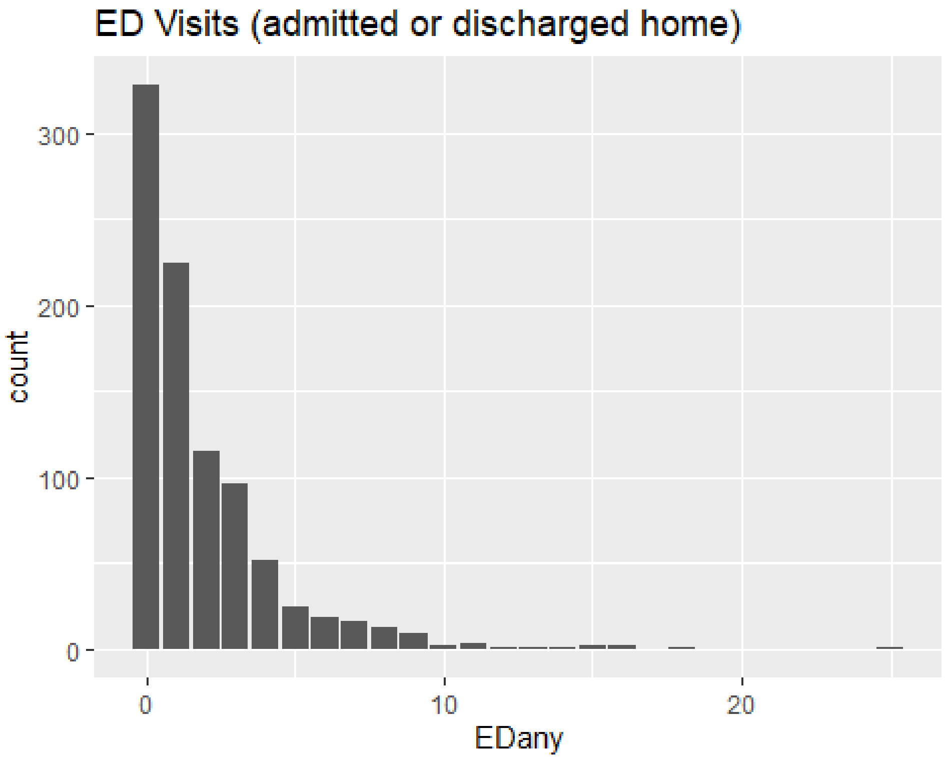
```
ggplot(subset) + geom_bar(aes(x=IP_pain)) +ggtitle("Admissions for Pain")
```



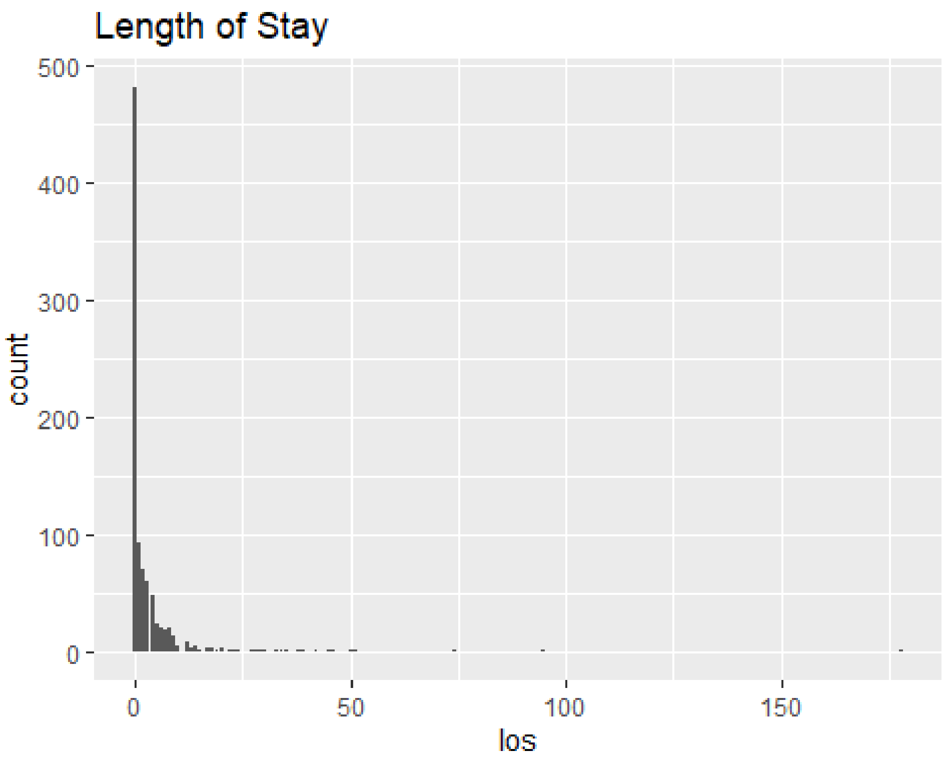
```
ggplot(subset) + geom_bar(aes(x=OPvisit)) + ggtitle("Outpatient Clinic
Visits")
```


Outpatient Clinic Visits

```
ggplot(subset) + geom_bar(aes(x=EDany)) + ggtitle("ED Visits (admitted or
discharged home)")
```


ED Visits (admitted or discharged home)

```
ggplot(subset) + geom_bar(aes(x=los)) +ggtitle("Length of Stay")
```



Final Analysis Plan

Based on my exploratory analysis, I think I will be able to continue my original analysis plan. I was having trouble with assigning age groups and utilization groups in R, so I will likely go back and continue that in SAS if I can't figure it out. I believe that I will still be able to answer my original questions, or at least get close to it. My plan is to continue the analysis once we learn more about regressions and the more advanced techniques. Because I am familair with this type of health services data in SAS, the biggest challenge is to do the same data cleaning and manipulation tasks in R. I was also having trouble with knitting to PDF for this assignment, even though I was able to do it for previous homework assignments.

Final Project Objectives

The following objectives remain relevant and will help us to understand why certain patients develop high utilization and more severe disease than others, and point us toward factors that can help identify patients at risk of becoming high utilizers.

- Objective 1: What are the patterns of high hospital utilization in pediatric SCD patients in the 10-year period from 2010-2019? This will help us understand the trends in our patient population.

- Objective 2: Which factors may be associated with and/or predict high hospital utilization? Does this differ in patients who are consistent high utilizers vs those who are not?

Additional questions that I may explore if time permits are whether there should be more variables included in this analysis to strengthen it and improve the final model(s).