Healthcare Costs Analysis

Maya Reese Farmer

October 22, 2021

Introduction I'm going to be analyzing healthcare data and determining the key factors influencing healthcare costs using the Healthcare Cost dataset from kaggle.com

The sales dataset contains 500 observations and has 6 columns with data on:

- AGE
- FEMALE: Binary variable that indicates if the patient is female
- LOS: length of stay in days
- RACE
- TOTCHG: hospital discharge costs
- APRDRG: All Patient Refined Diagnosis Related Groups

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
## importing dataset into r
hospitalcosts <- read_csv("HospitalCosts.csv")
```

```
## Rows: 500 Columns: 6

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (6): AGE, FEMALE, LOS, RACE, TOTCHG, APRDRG
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
## view data
head(hospitalcosts,10)
```

```
## # A tibble: 10 x 6
##      AGE FEMALE   LOS  RACE TOTCHG APRDRG
##    <dbl>  <dbl> <dbl> <dbl>  <dbl>  <dbl>
##  1    17      1     2     1   2660    560
##  2    17      0     2     1   1689    753
##  3    17      1     7     1  20060    930
##  4    17      1     1     1    736    758
##  5    17      1     1     1   1194    754
##  6    17      0     0     1   3305    347
##  7    17      1     4     1   2205    754
##  8    16      1     2     1   1167    754
##  9    16      1     1     1    532    753
## 10    17      1     2     1   1363    758
```

```r
library(dplyr)

## change column names
hospitalcosts = hospitalcosts %>%
    rename(age = AGE, gender = FEMALE, staylength = LOS, race = RACE,
        cost = TOTCHG, diagnosis = APRDRG)

## change 0 and 1 to Male and Female
hospitalcosts$gender[hospitalcosts$gender == 0] <- "Male"
hospitalcosts$gender[hospitalcosts$gender == 1] <- "Female"

## create age groups
hospitalcosts = hospitalcosts %>%
    mutate(age.group = case_when(hospitalcosts$age <= 3 ~ "0-3",
        hospitalcosts$age >= 4 & hospitalcosts$age <= 7 ~ "4-7",
        hospitalcosts$age >= 8 & hospitalcosts$age <= 11 ~ "8-11",
        hospitalcosts$age >= 12 & hospitalcosts$age <= 17 ~ "12-17"))
```

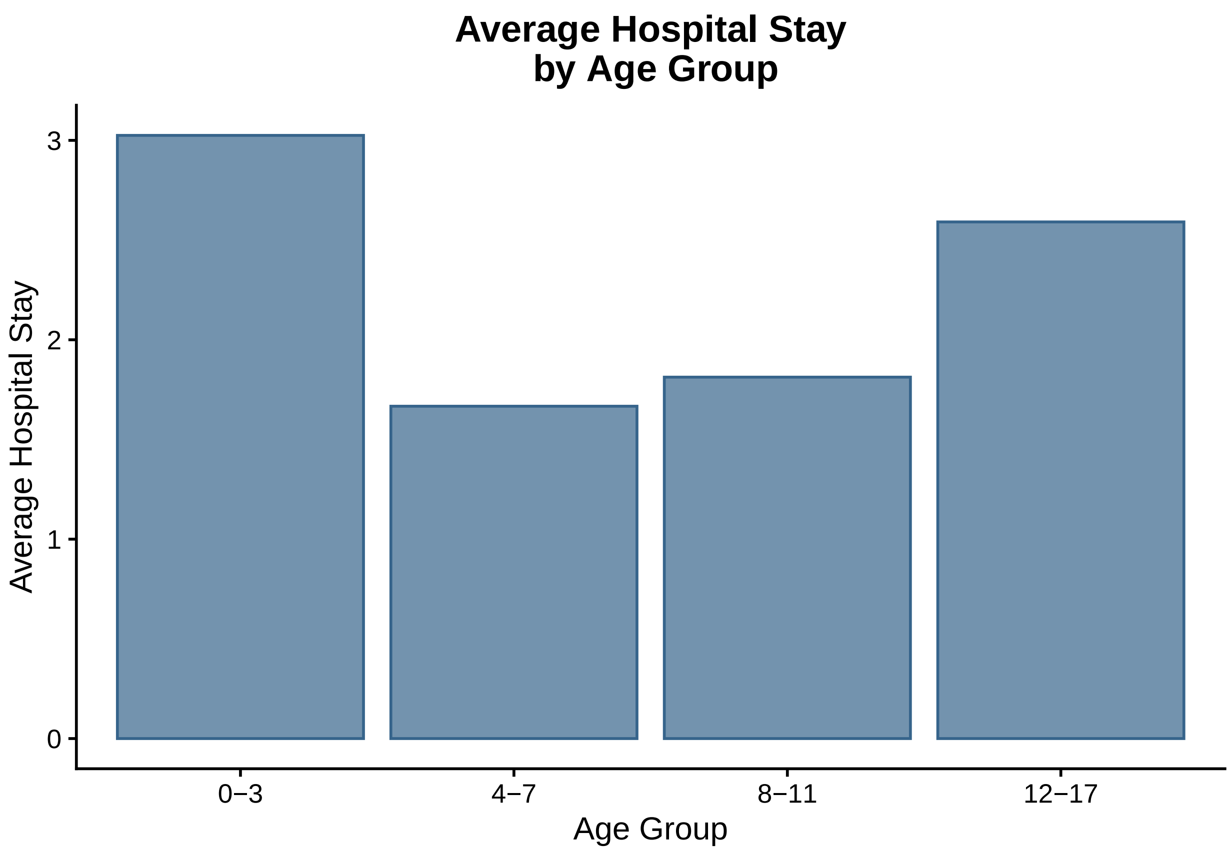
Data Prep

Age, Hospital Costs, and Length of Stay First I want to determine how age group influence patient costs and their length of stay.

```r
## average hospital stay by age group
hospitalcosts %>%
    group_by(age.group) %>%
    summarise(stay.mean = mean(staylength))
```

```
## # A tibble: 4 x 2
##   age.group stay.mean
##   <chr>         <dbl>
## 1 0-3            3.02
## 2 12-17          2.59
## 3 4-7            1.67
## 4 8-11           1.81
```

```
library(ggplot2)
library(cowplot)
library(forcats)

hospitalcosts %>%
    group_by(age.group) %>%
    summarise(stay.mean = mean(staylength)) %>%
    mutate(age.group = fct_relevel(age.group, "0-3", "4-7", "8-11",
        "12-17")) %>%
    ggplot(aes(x = age.group, y = stay.mean)) + geom_bar(stat = "identity",
    fill = alpha("steelblue4", 0.7), color = "steelblue4") +
    theme_cowplot(12) + labs(x = "Age Group", y = "Average Hospital Stay",
    title = "Average Hospital Stay\n by Age Group") + theme(plot.title = element_text(hjust = 0.5))
```




```
## anova testing differences in hospital stay by age group
summary(aov(hospitalcosts$staylength ~ hospitalcosts$age.group))
```

```
##                             Df Sum Sq Mean Sq F value Pr(>F)
```

```
## hospitalcosts$age.group   3     50   16.58    1.47  0.222
## Residuals               496   5595   11.28
```

We can see that young individuals (those in the 0-3 age group) tend to have the longest hospital stays on average. However, analysis of variance (ANOVA) shows that these differences are not signifcant.

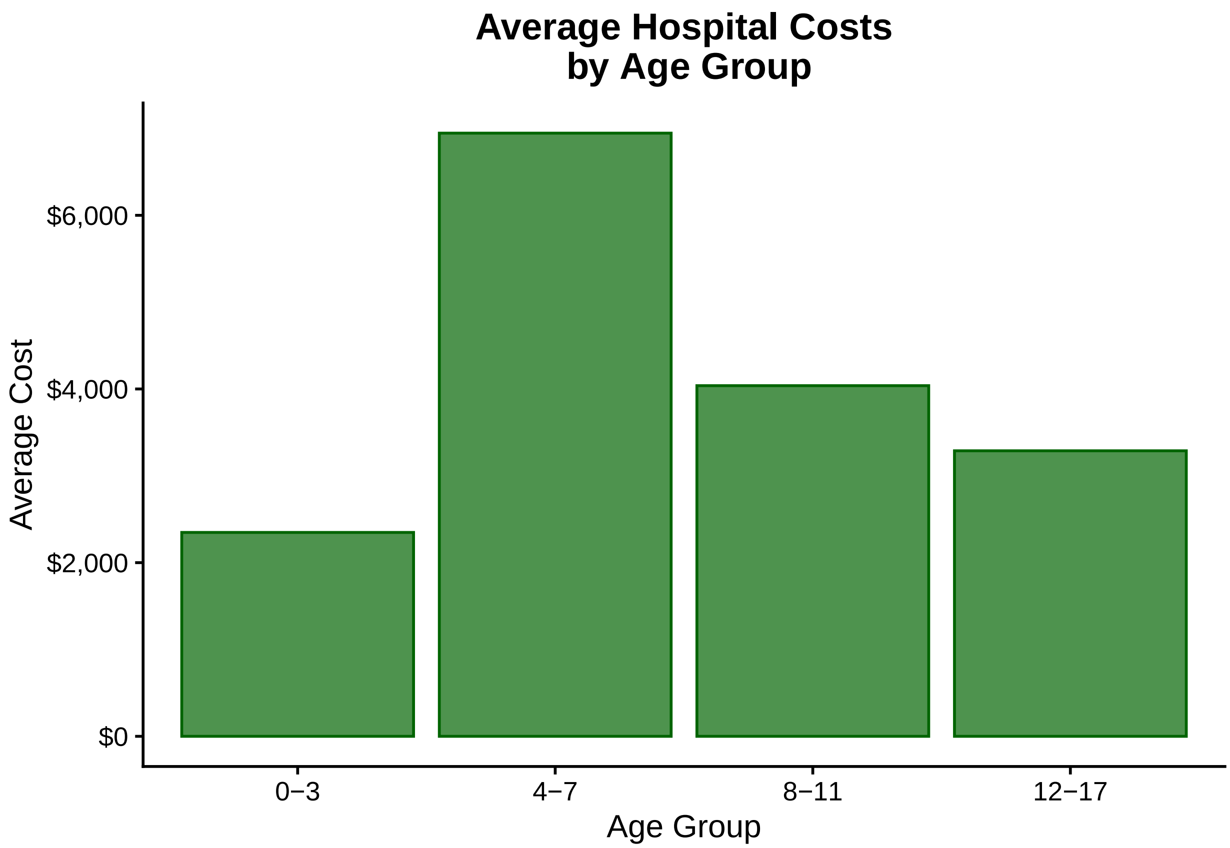
```r
library(scales)
```

```
##
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':
##
##     discard

## The following object is masked from 'package:readr':
##
##     col_factor
```

```r
## cost based on age group
hospitalcosts %>%
    group_by(age.group) %>%
    summarise(age.cost = mean(cost))
```

```
## # A tibble: 4 x 2
##   age.group age.cost
##   <chr>        <dbl>
## 1 0-3          2348.
## 2 12-17        3288.
## 3 4-7          6946
## 4 8-11         4038.
```

```r
hospitalcosts %>%
    group_by(age.group) %>%
    summarise(age.cost = mean(cost)) %>%
    mutate(age.group = fct_relevel(age.group, "0-3", "4-7", "8-11",
        "12-17")) %>%
    ggplot(aes(x = age.group, y = age.cost)) + geom_bar(stat = "identity",
    fill = alpha("darkgreen", 0.7), color = "darkgreen") + theme_cowplot(12) +
    labs(x = "Age Group", y = "Average Cost", title = "Average Hospital Costs\n by Age Group") +
    theme(plot.title = element_text(hjust = 0.5)) + scale_y_continuous(labels = dollar_format())
```




```
## anova testing differences in hospital costs by age group
summary(aov(hospitalcosts$cost ~ hospitalcosts$age.group))
```

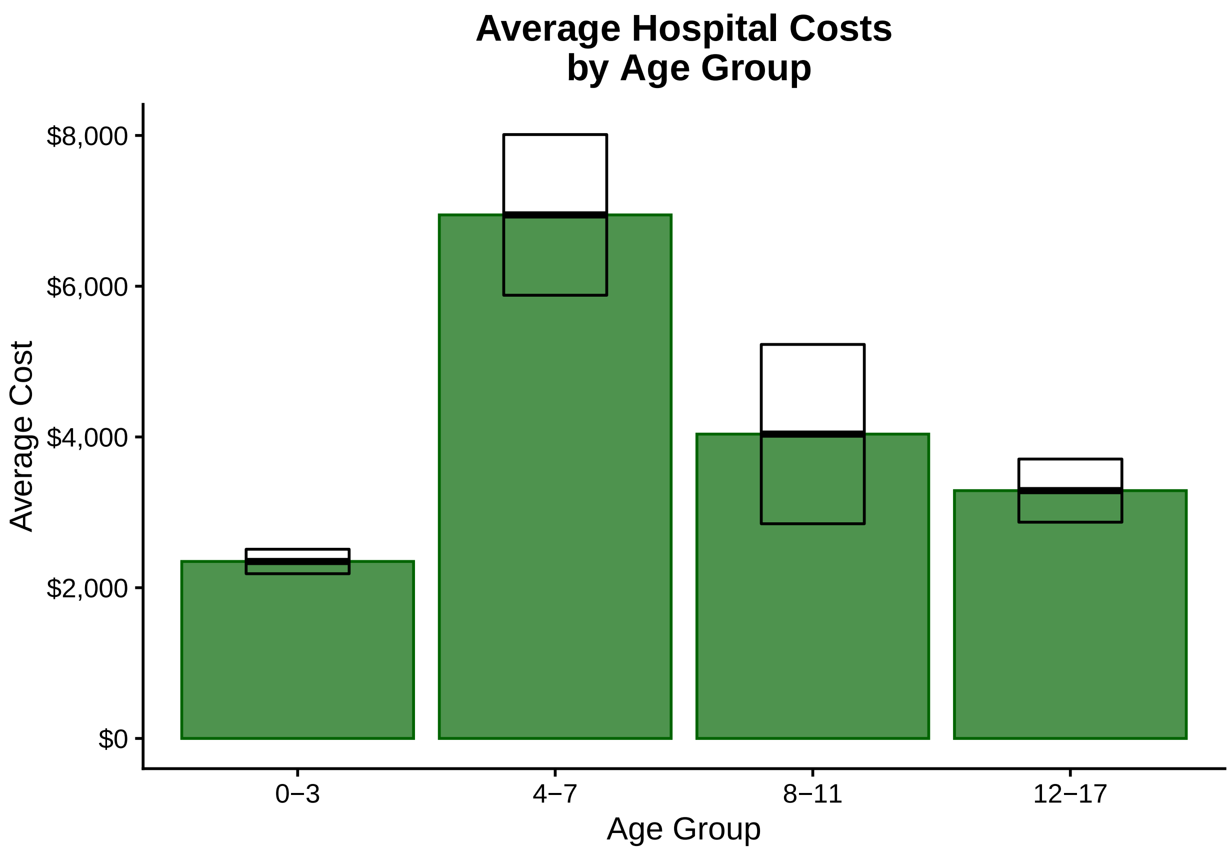
```
##                          Df    Sum Sq  Mean Sq F value   Pr(>F)    
## hospitalcosts$age.group   3 2.812e+08 93720125     6.4 0.000293 ***
## Residuals               496 7.264e+09 14644306                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Although the 0-3 age group has the longest hospital stays, it looks like the 4-7 age group accrues the highest hospital costs on average. Here, the ANOVA test shows that there are signicant differences in costs between age groups. This shows that patients in the 4-7 age group are being charged significantly more than the other age groups (*we can perform post-hoc analyses and comparisons using a Tukey test). This age group should be assessed further.

We can visualize the amount of variation or standard error there is in cost for each age group to make our visualization more complete.

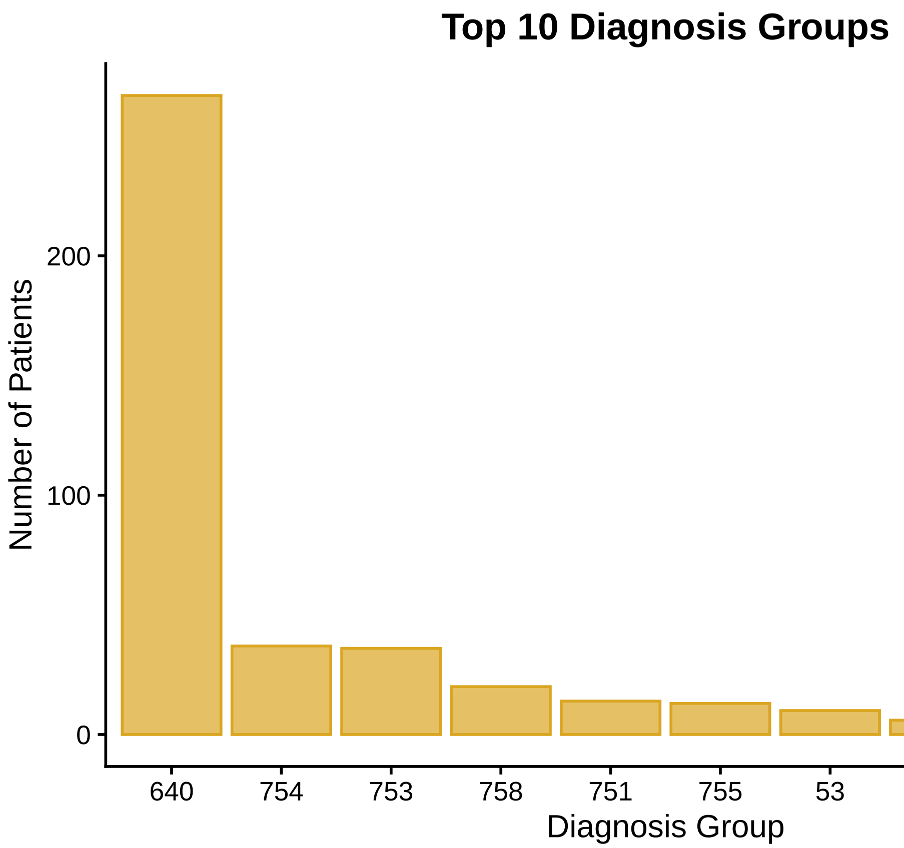
```
hospitalcosts %>%
    group_by(age.group) %>%
    summarise(age.cost = mean(cost), sd.cost = sd(cost), n = n(),
        se.cost = sd.cost/sqrt(n)) %>%
    mutate(age.group = fct_relevel(age.group, "0-3", "4-7", "8-11",
        "12-17")) %>%
    ggplot(aes(x = age.group, y = age.cost)) + geom_bar(stat = "identity",
    fill = alpha("darkgreen", 0.7), color = "darkgreen") + geom_crossbar(aes(x = age.group,
```

```
    y = age.cost, ymin = age.cost - se.cost, ymax = age.cost +
        se.cost), width = 0.4, colour = "black", alpha = 0.9,
    size = 0.5) + theme_cowplot(12) + labs(x = "Age Group", y = "Average Cost",
    title = "Average Hospital Costs\n by Age Group") + theme(plot.title = element_text(hjust = 0.5)) +
    scale_y_continuous(labels = dollar_format())
```



Here, we've added the crossbar plot to our original gglpot. The thick center line represents the mean and the top and bottom lines represent +/- the standard error. Typically, non-overlapping standard errors represent means that are significantly different from one another, which is what we see here. Our previous ANOVA confirms these significant differences.

```
## top 10 hospital diagnoses
hospitalcosts %>%
    count(diagnosis) %>%
    arrange(desc(n)) %>%
    head(10) %>%
    mutate(diagnosis = as.factor(diagnosis), diagnosis = fct_reorder(diagnosis,
        desc(n))) %>%
    ggplot(aes(x = diagnosis, y = n)) + geom_bar(stat = "identity",
    fill = alpha("goldenrod", 0.7), color = "goldenrod") + theme_cowplot(12) +
    labs(x = "Diagnosis Group", y = "Number of Patients", title = "Top 10 Diagnosis Groups") +
    theme(plot.title = element_text(hjust = 0.5))
```



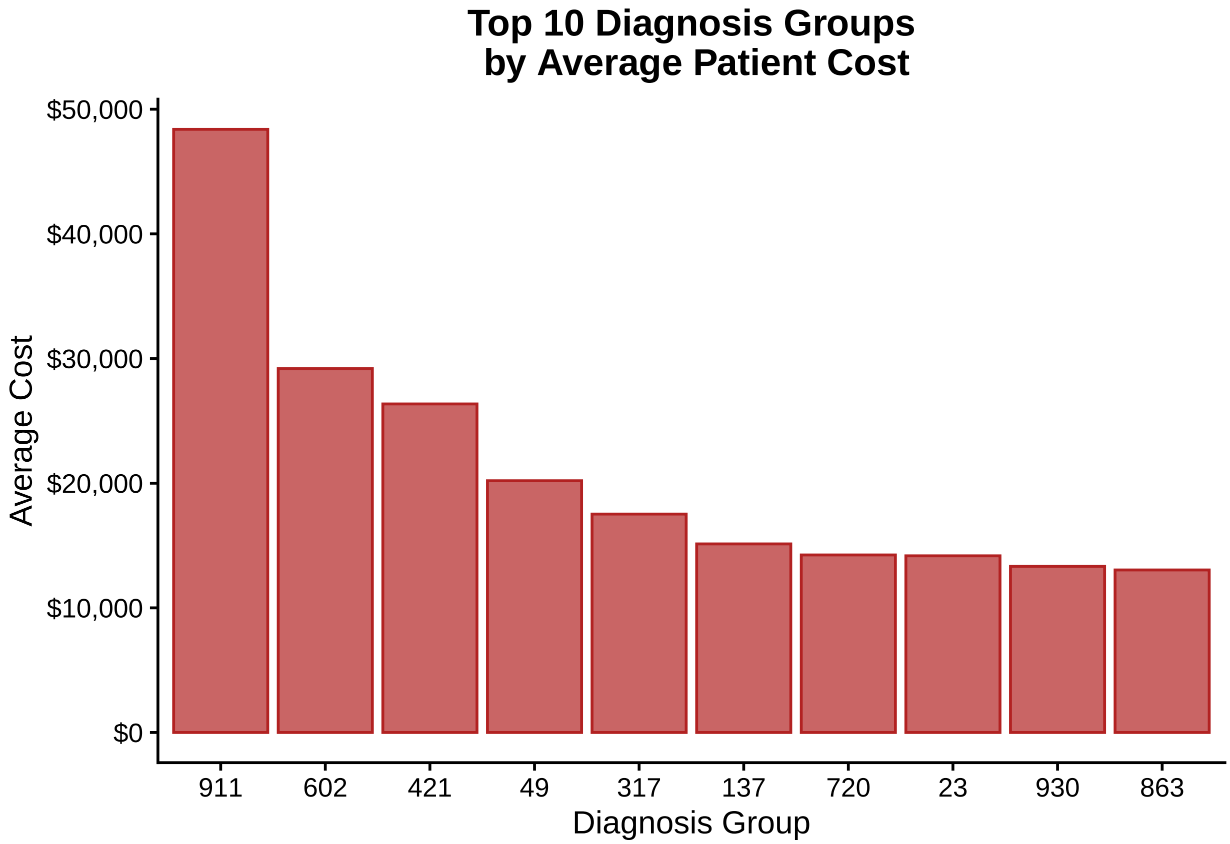
Diagnosis Group and Associated Costs

```
## most expensive diagnoses on average
hospitalcosts %>%
    group_by(diagnosis) %>%
    summarise(avg.cost = mean(cost)) %>%
    arrange(desc(avg.cost))
```

```
## # A tibble: 63 x 2
##    diagnosis avg.cost
##        <dbl>    <dbl>
##  1       911    48388
##  2       602    29188
##  3       421    26356
##  4        49    20195
##  5       317    17524
##  6       137    15129
##  7       720    14243
##  8        23    14174
##  9       930    13327
## 10       863    13040
## # ... with 53 more rows
```

```
hospitalcosts %>%
    group_by(diagnosis) %>%
    summarise(avg.cost = mean(cost)) %>%
    arrange(desc(avg.cost)) %>%
    head(10) %>%
```

```
    mutate(diagnosis = as.factor(diagnosis), diagnosis = fct_reorder(diagnosis,
        desc(avg.cost))) %>%
    ggplot(aes(x = diagnosis, y = avg.cost)) + geom_bar(stat = "identity",
    fill = alpha("firebrick", 0.7), color = "firebrick") + theme_cowplot(12) +
    labs(x = "Diagnosis Group", y = "Average Cost", title = "Top 10 Diagnosis Groups\n by Average Patien
    theme(plot.title = element_text(hjust = 0.5)) + scale_y_continuous(labels = dollar_format())
```



Most individuals are in the 640 diagnosis group, however, that group doesn't fall within the top 10 diagnoses based on cost. None of the other top 10 diagnosis groups by number of patients overlap with the most costly diagnoses either. Individuals in the 911 group are being charged, on average, about $50,000. However, there are likely very few patients in this group.

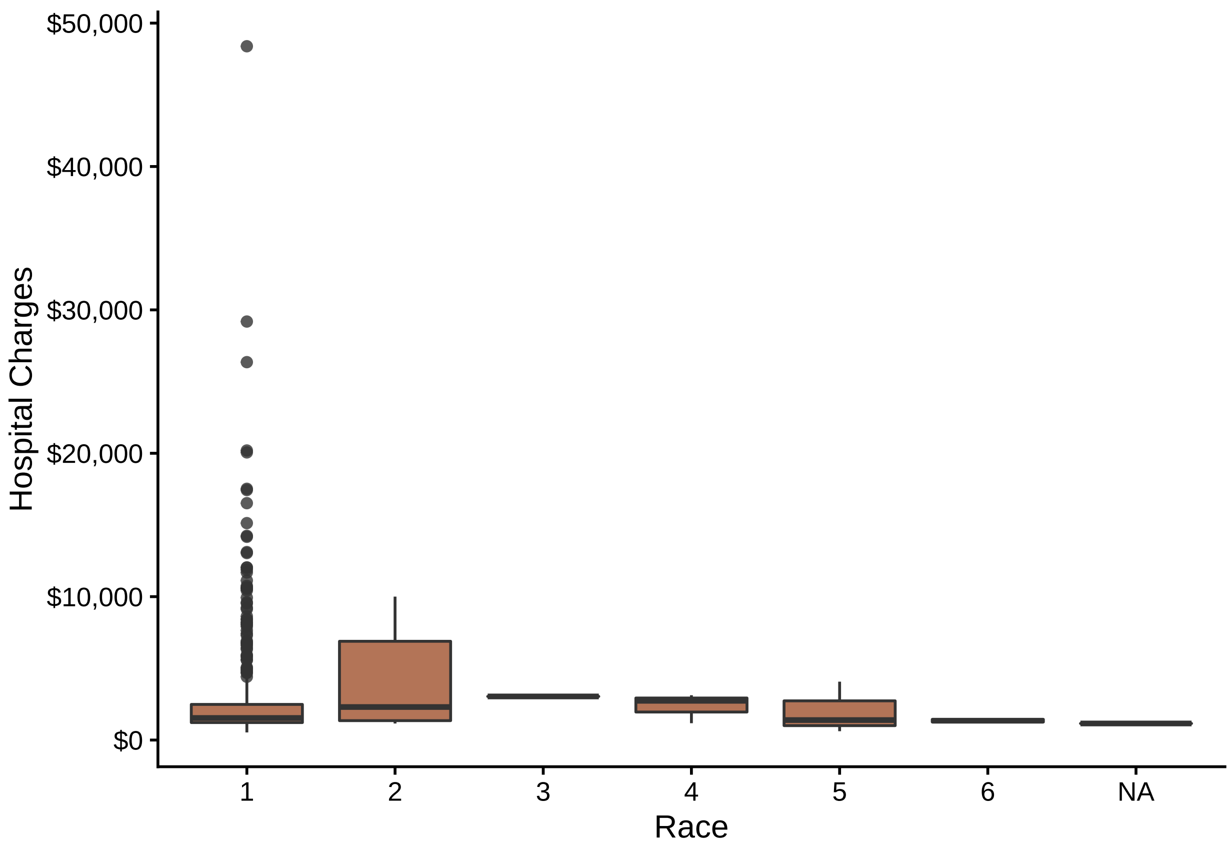
```
## compare difference in hosptial charges by race
hospitalcosts %>%
    group_by(race) %>%
    summarise(avg.cost = mean(cost))
```

Race, Gender, and Hospital Costs

```
## # A tibble: 7 x 2
##    race avg.cost
```

```
##   <dbl>    <dbl>
## 1     1    2773.
## 2     2    4202.
## 3     3    3041
## 4     4    2345.
## 5     5    2027.
## 6     6    1349
## 7    NA    1156
```

```
ggplot(hospitalcosts, aes(x = as.factor(race), y = cost)) + geom_boxplot(fill = "sienna",
    alpha = 0.8) + theme_cowplot(12) + labs(x = "Race", y = "Hospital Charges") +
    scale_y_continuous(labels = dollar_format())
```




```
## test whether different races are charged differently
summary(aov(hospitalcosts$cost ~ hospitalcosts$race))
```

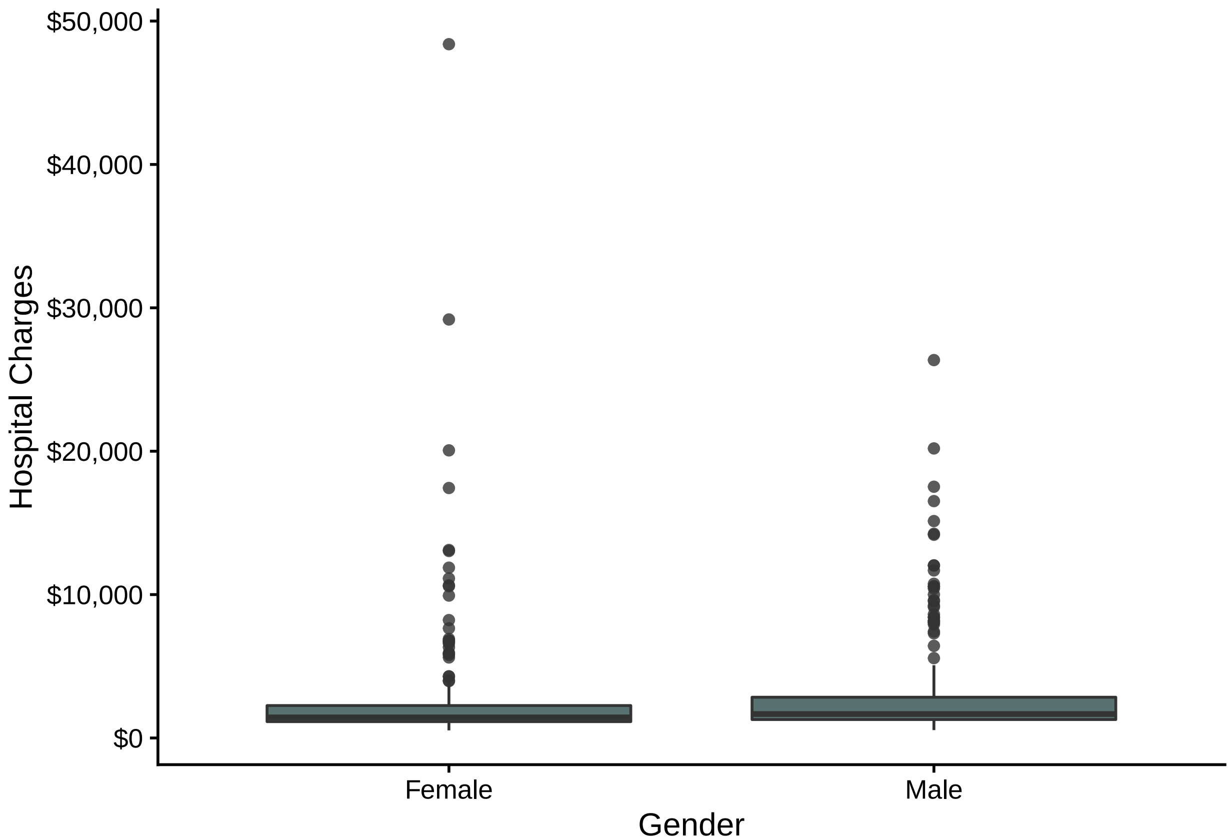
```
##                      Df    Sum Sq  Mean Sq F value Pr(>F)
## hospitalcosts$race    1 2.488e+06  2488459   0.164  0.686
## Residuals           497 7.540e+09 15170268
## 1 observation deleted due to missingness
```

Although there appears to be some variation in costs between different races, our ANOVA (analysis of variance) shows that there is not a significant relationship between race and cost. This means that there aren't significant differences in hospital charges between different races. However, there are a lot of outliers in race group 1 that may require further analysis (i.e. are these individuals more prone to certain illnesses,etc),

```r
## hospital costs based on gender
hospitalcosts %>%
    group_by(gender) %>%
    summarise(gender.charge = mean(cost)) %>%
    arrange(desc(gender.charge))
```

```
## # A tibble: 2 x 2
##   gender gender.charge
##   <chr>          <dbl>
## 1 Male           3014.
## 2 Female         2546.
```

```r
ggplot(hospitalcosts, aes(x = gender, y = cost)) + geom_boxplot(fill = "darkslategrey",
    alpha = 0.8) + theme_cowplot(12) + labs(x = "Gender", y = "Hospital Charges") +
    scale_y_continuous(labels = dollar_format())
```




```r
summary(aov(hospitalcosts$cost ~ hospitalcosts$gender))
```

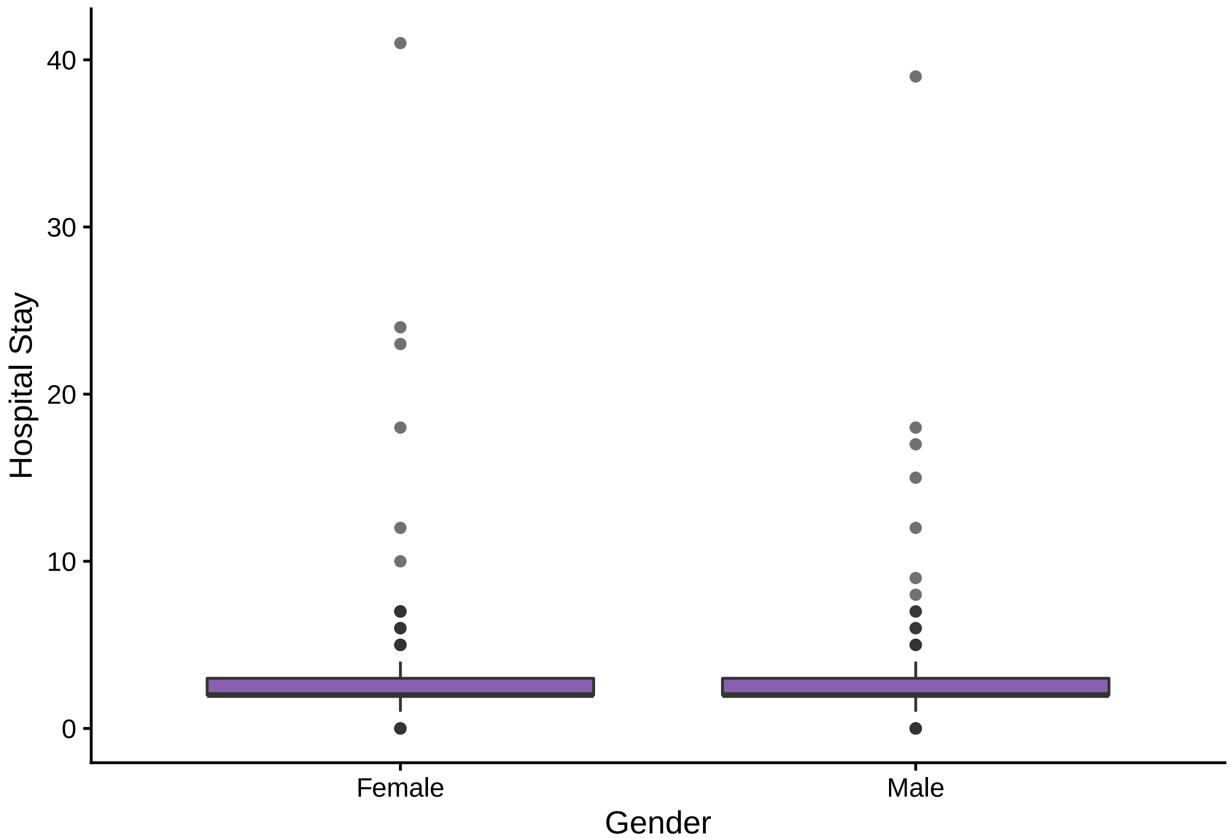
```
##                       Df    Sum Sq  Mean Sq F value Pr(>F)
## hospitalcosts$gender   1 2.734e+07 27337922   1.811  0.179
## Residuals            498 7.517e+09 15095177
```

When you compare costs across different genders, there don't appear to be any significant differences between male and female patients in terms of costs.

```r
hospitalcosts %>%
    group_by(gender) %>%
    summarise(avg.stay = mean(staylength))
```

```
## # A tibble: 2 x 2
##   gender avg.stay
##   <chr>     <dbl>
## 1 Female     2.95
## 2 Male       2.70
```

```r
ggplot(hospitalcosts, aes(x = gender, y = staylength)) + geom_boxplot(fill = "purple4",
    alpha = 0.7) + theme_cowplot(12) + labs(x = "Gender", y = "Hospital Stay")
```



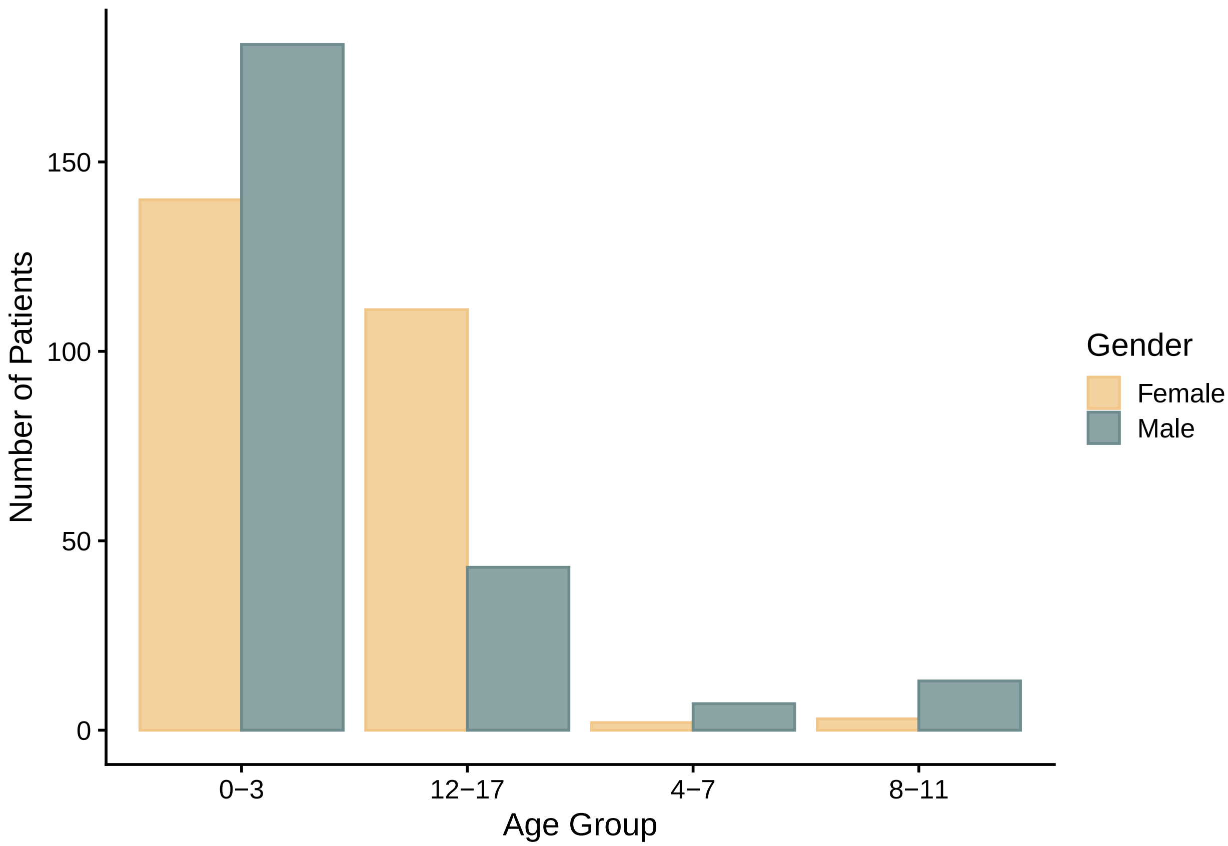
We also see that average hospital stay doesn't differ between genders either.

```r
## breakdown of number of patients by gender by age group
hospitalcosts %>%
    group_by(age.group, gender) %>%
    count(gender)
```

```
## # A tibble: 8 x 3
## # Groups:   age.group, gender [8]
##   age.group gender     n
##   <chr>     <chr>  <int>
## 1 0-3       Female   140
## 2 0-3       Male     181
```

```
## 3 12-17      Female   111
## 4 12-17      Male      43
## 5 4-7        Female     2
## 6 4-7        Male       7
## 7 8-11       Female     3
## 8 8-11       Male      13
```

```
hospitalcosts %>%
    group_by(age.group, gender) %>%
    count(gender) %>%
    ggplot(aes(age.group, n, fill = gender, color = gender)) +
    geom_bar(position = "dodge", stat = "identity") + theme_cowplot(12) +
    labs(x = "Age Group", y = "Number of Patients") + scale_fill_manual(values = alpha(c("#F1C789",
    "#6E8C8E"), 0.8), name = "Gender") + scale_color_manual(values = c("#F1C789",
    "#6E8C8E"), name = "Gender")
```



Based on the figure, each gender is not represented equally in each age group. We can see that there are more males hospitalized in the 0-3 age group (along with 4-7 and 8-11, but less so). And there are many more females hospitalized in the 12-17 age group.

Building a Model Based on the information we've gathered so far, it appears that patient age and diagnosis group are the most significant factors influencing patient costs. First, lets look at how age group and diagnosis influence cost in the top 5 most costly diagnosis groups.

```r
## first create a new vector that only includes the top 5
## diagnosis groups by cost
top5.diagnosiscost = hospitalcosts %>%
    group_by(diagnosis) %>%
    summarise(avg.cost = mean(cost)) %>%
    arrange(desc(avg.cost)) %>%
    top_n(5) %>%
    pull(diagnosis)
```

```
## Selecting by avg.cost
```

```r
top5.diagnosiscost
```

```
## [1] 911 602 421  49 317
```

```r
## filter data to only see patients in the top 5 diagnosis
## groups
hospitalcosts %>%
    filter(diagnosis %in% top5.diagnosiscost)
```

```
## # A tibble: 5 x 7
##     age gender staylength  race  cost diagnosis age.group
##   <dbl> <chr>       <dbl> <dbl> <dbl>     <dbl> <chr>
## 1     0 Female         41     1 29188       602 0-3
## 2     0 Male           39     1 26356       421 0-3
## 3    15 Male            6     1 20195        49 12-17
## 4    10 Male            7     1 17524       317 8-11
## 5    17 Female          7     1 48388       911 12-17
```

Here, we can see that only 5 patients represent the diagnosis groups that have the highest costs (one patient per group). This indicates that most patients aren't paying high hospital costs. This was also indicated when we compared the top 10 diagnosis groups by patient count to the top 10 diagnosis groups by average cost earlier (because there was no overlap in diagnosis group between them).

Now, instead of looking at the top 5 diagnoses by cost, let's look at how age group and diagnosis influence cost in the top 20 diagnosis groups with the most patients.

```r
top10.diagnosiscount = hospitalcosts %>%
    count(diagnosis) %>%
    arrange(desc(n)) %>%
    top_n(10) %>%
    pull(diagnosis)
```

```
## Selecting by n
```

```r
top10.diagnosiscount
```

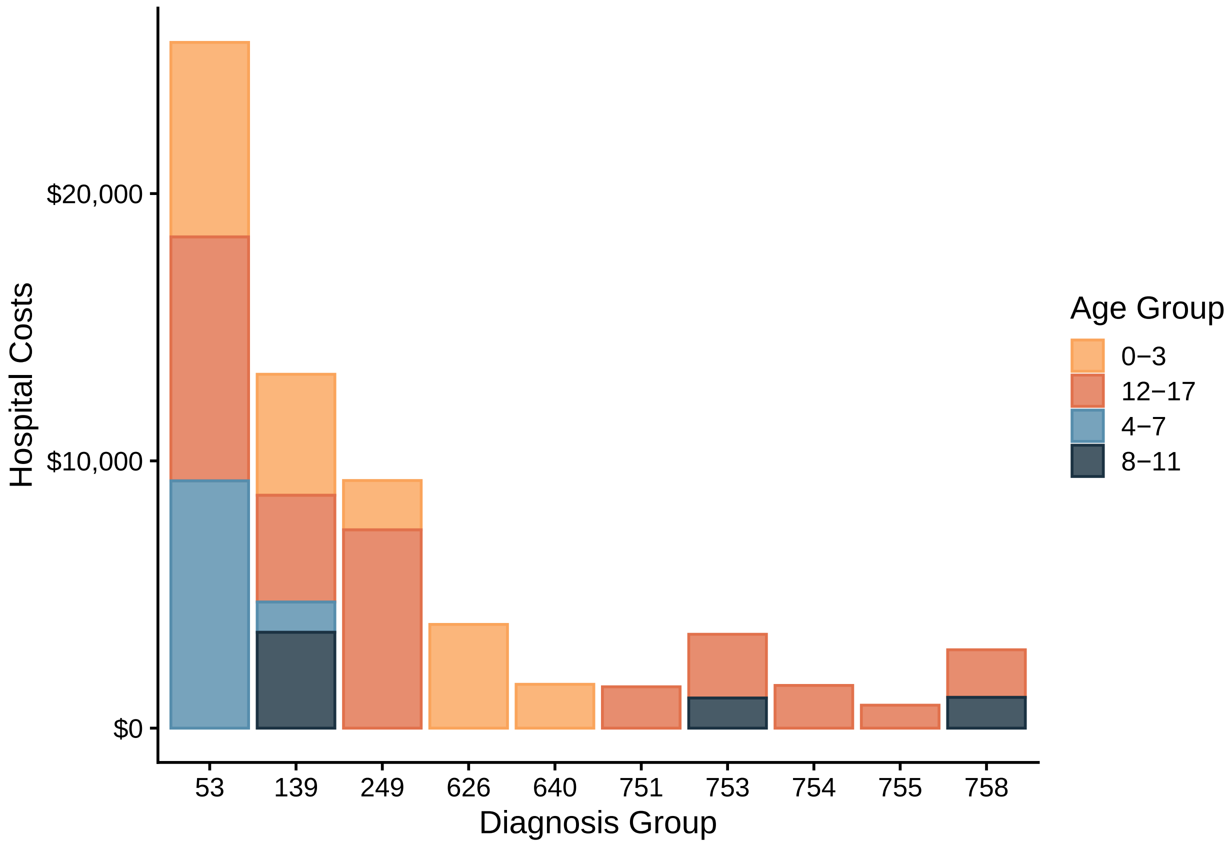
```
##  [1] 640 754 753 758 751 755  53 249 626 139
```

```
## filter data to only see patients in the top 5 diagnosis
## groups
hospitalcosts %>%
    filter(diagnosis %in% top10.diagnosiscount)
```

```
## # A tibble: 414 x 7
##      age gender staylength  race  cost diagnosis age.group
##    <dbl> <chr>       <dbl> <dbl> <dbl>     <dbl> <chr>
##  1    17 Male            2     1  1689       753 12-17
##  2    17 Female          1     1   736       758 12-17
##  3    17 Female          1     1  1194       754 12-17
##  4    17 Female          4     1  2205       754 12-17
##  5    16 Female          2     1  1167       754 12-17
##  6    16 Female          1     1   532       753 12-17
##  7    17 Female          2     1  1363       758 12-17
##  8    17 Female          2     1  1245       758 12-17
##  9    15 Male            2     1  1656       753 12-17
## 10    15 Female          2     1  1379       751 12-17
## # ... with 404 more rows
```

```
hospitalcosts %>%
    filter(diagnosis %in% top10.diagnosiscount) %>%
    group_by(diagnosis, age.group) %>%
    summarise(avg.cost = mean(cost)) %>%
    ggplot(aes(x = as.factor(diagnosis), y = avg.cost, fill = age.group,
        color = age.group)) + geom_bar(position = "stack", stat = "identity") +
    theme_cowplot(12) + labs(x = "Diagnosis Group", y = "Hospital Costs") +
    scale_fill_manual(values = alpha(c("#faa45b", "#e1714c",
        "#568dac", "#1b3242"), 0.8), name = "Age Group") + scale_color_manual(values = c("#faa45b",
    "#e1714c", "#568dac", "#1b3242"), name = "Age Group") + scale_y_continuous(labels = dollar_format()
```

```
## `summarise()` has grouped output by 'diagnosis'. You can override using the `.groups` argument.
```



The first thing we can conclude from the figure is that all age groups are not represented for each diagnosis, For example, the 640 group (which contains the highest number of patients) only has patients in the 0-3 age group. This means that different diagnoses are exclusive to patients of certain ages.

However, among diagnosis groups where there are multiple age groups (e.g. 53, 139, 753, etc), it there seem to be large differences in the hosptical costs between age groups. This indicates that there may be an interactive effect between age and diagnosis group on hospital costs. We can perform a two-way ANOVA to formally test this.

```
twoway.model = aov(hospitalcosts$cost ~ hospitalcosts$age.group *
    as.factor(hospitalcosts$diagnosis))
summary(twoway.model)
```

```
##                                                               Df    Sum Sq
## hospitalcosts$age.group                                        3 2.812e+08
## as.factor(hospitalcosts$diagnosis)                            62 6.419e+09
## hospitalcosts$age.group:as.factor(hospitalcosts$diagnosis)   14 1.010e+08
## Residuals                                                    420 7.435e+08
##                                                               Mean Sq F value
## hospitalcosts$age.group                                      93720125  52.939
## as.factor(hospitalcosts$diagnosis)                          103532641  58.482
## hospitalcosts$age.group:as.factor(hospitalcosts$diagnosis)    7215143   4.076
## Residuals                                                     1770334
##                                                               Pr(>F)
## hospitalcosts$age.group                                      < 2e-16 ***
## as.factor(hospitalcosts$diagnosis)                           < 2e-16 ***
## hospitalcosts$age.group:as.factor(hospitalcosts$diagnosis) 1.16e-06 ***
## Residuals
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on our model, we do indeed find that age group and diagnosis both independently affect hospital costs, and there is also an interactive effect (p value is < 0.001 for each effect). This means that within different diagnosis groups, patients in different age groups are being charged significantly different costs.