SDS 390 Final Project Topic Selection and Exploratory Analysis

My My Tran, Faith Ndanu

2024-04-16

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(broom)
library(readr)
cms_hospital <- read_csv("cms_hospital.csv")
```

```
## Rows: 121577 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (4): state, race, admission, discharge
## dbl (9): hospID, hospVol, year, female, age, deyo, LOS, T1, T2
##
## 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.
```

The name(s) of the individual(s) in the group.

My My Tran, Faith Ndanu

A brief description of your data and data source.

The Centers for Medicare & Medicaid Services (CMS) is an agency within the United States Department of Health and Human Services (HHS). The CMS provides public datasets regarding the U.S. healthcare system to provide transparency, study health patterns/trends, implement policies, and overall improve outcomes, equity and quality of care. (https://www.cms.gov/about-cms#:~:text=CMS%20is%20the%20federal%20agency,in%20the%20health%20care%20system). There are over 4,000 Medicare-certified and registered hospitals across the United States (https://data.cms.gov/provider-data/search?page-size=50&theme=Hospitals). Our chosen dataset, cms_hospitals.csv provides

information on the hospitals a patient was admitted to (by hospital ID), the year of the patient's admittance, the state of the hospital, demographic information on the patient (i.e. gender, age, race, type of admission, the Charlson-Deyo comorbidity index on admission, and their length of stay). The length of hospital stay is the amount of time between a patient's admission to the hospital and their discharge. The patients are greater than or equal to 65 years of age, were diagnosed with pancreatic cancer between 2000 and 2009, and were diagnosed at a hospital with at least 50 admissions. Looking at our data, there are 121,577 admissions that were reported during those 9 years, as well as 13 columns, including the predictor variables mentioned above.

An exploratory data analysis (EDA) of the data you plan to use.

```
# race count
race.count <- cms_hospital|>
  group_by( race)|>
  summarize(n = n())
race.count
```

```
## # A tibble: 3 x 2
##   race       n
##   <chr>  <int>
## 1 Black  12504
## 2 Other   5100
## 3 White 103973
```

```
# discharge counts

discharges.summary <-cms_hospital|>
  group_by(discharge)|>
  summarize(mean = mean(LOS), variance = var( LOS), n = n())
discharges.summary
```

```
## # A tibble: 9 x 4
##   discharge    mean variance     n
##   <chr>       <dbl>    <dbl> <int>
## 1 1.Home       7.40     30.9 61165
## 2 2.HomeCare  11.5      63.4 22945
## 3 3.SNF/ICF   12.8      98.9 17296
## 4 4.Hospice    9.47     52.2 15714
## 5 5.Rehab     16.8     210.   1279
## 6 6.Inpatient 13.1     123.   1855
## 7 7.LTC       21.2     222.    671
## 8 8.Swing bed 14.6      95.8   120
## 9 9.Other      7.31     35.6   532
```

```
# numerical exposures

mean(cms_hospital$age)
```

```
## [1] 76.47929
```

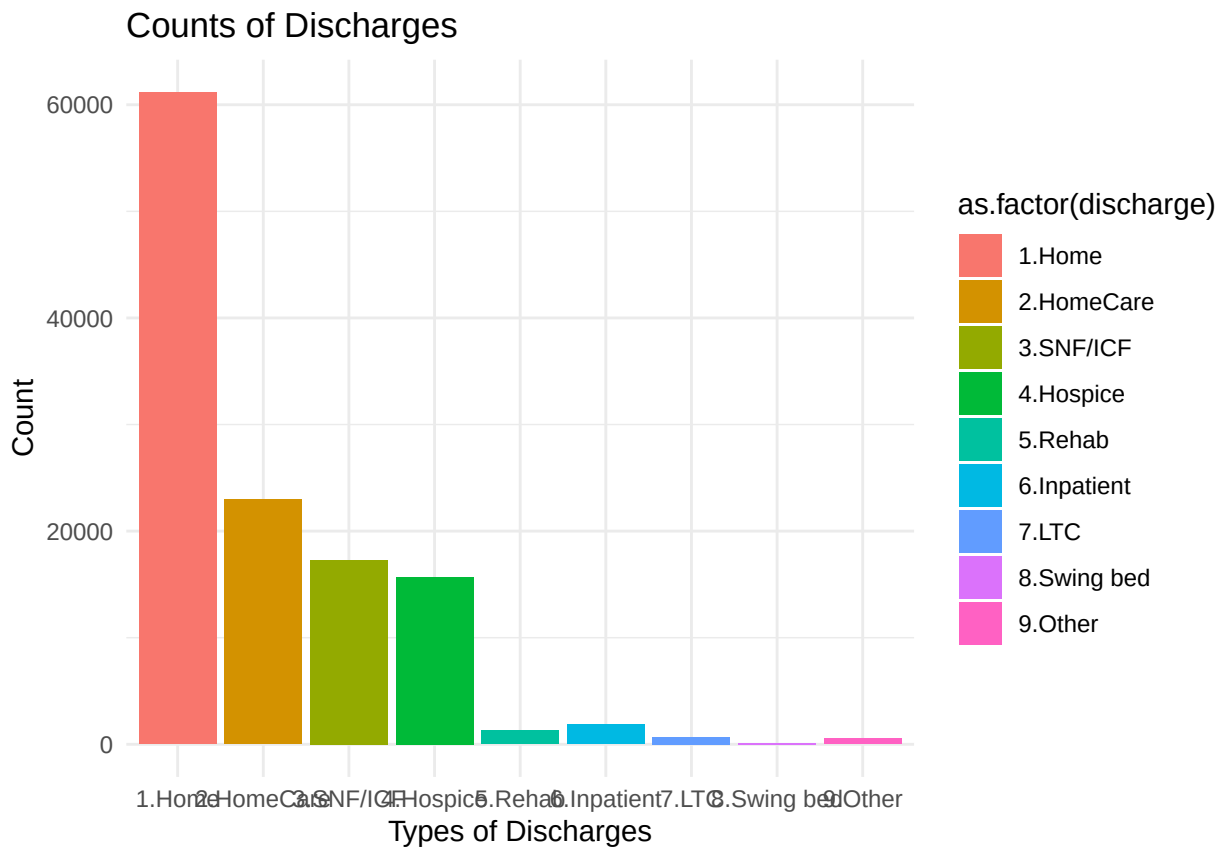
```r
mean(cms_hospital$LOS)
```

```
## [1] 9.484508
```

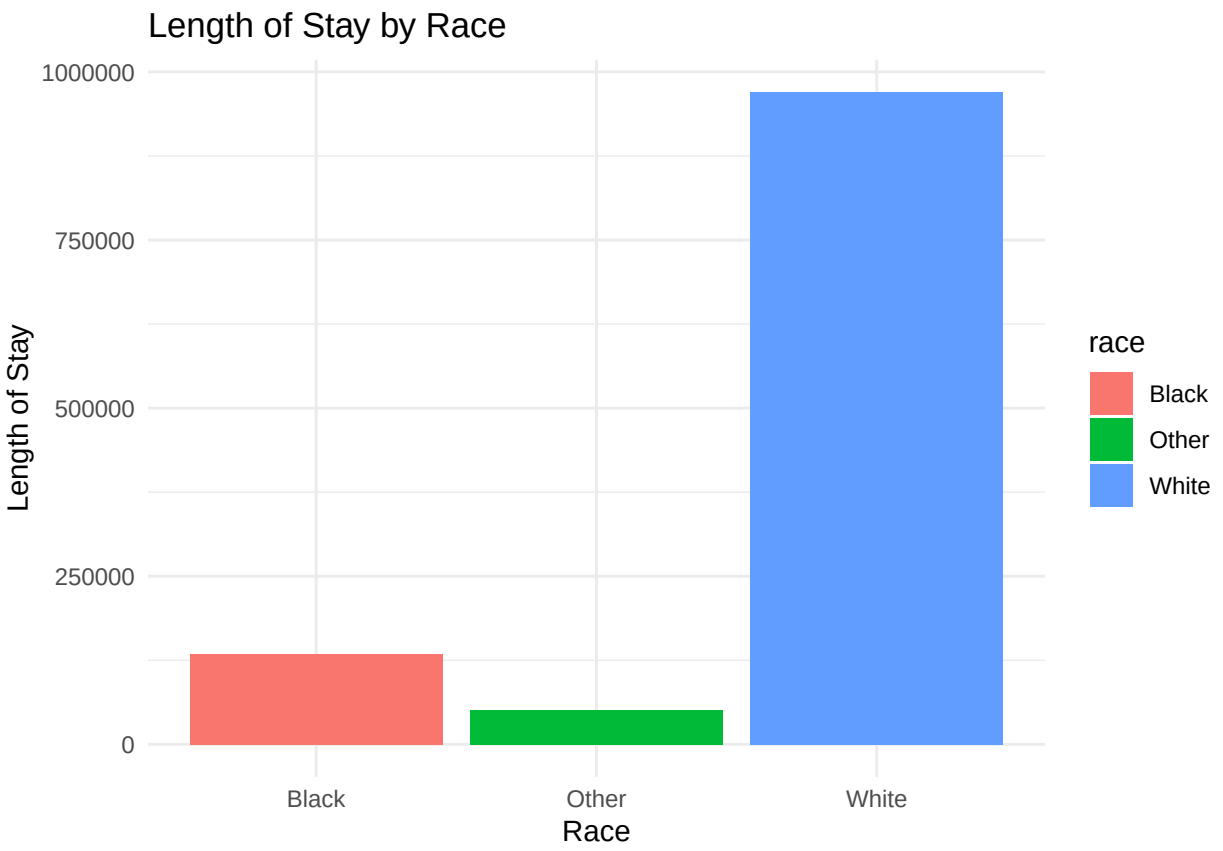
```r
# deyo index
deyo.summary <- cms_hospital|>
  group_by(deyo)|>
  summarize( mean = mean (LOS), variance = var(LOS), n= n())
deyo.summary
```

```
## # A tibble: 8 x 4
##    deyo  mean variance      n
##   <dbl> <dbl>    <dbl>  <int>
## 1     0 10.0      88.2   8694
## 2     1  9.43     57.8 104117
## 3     2  9.47     57.5   6552
## 4     3  9.59     53.2   1762
## 5     4 10.6      78.4    388
## 6     5 10.6      67.4     52
## 7     6  9.18     29.0     11
## 8     7 12        NA        1
```

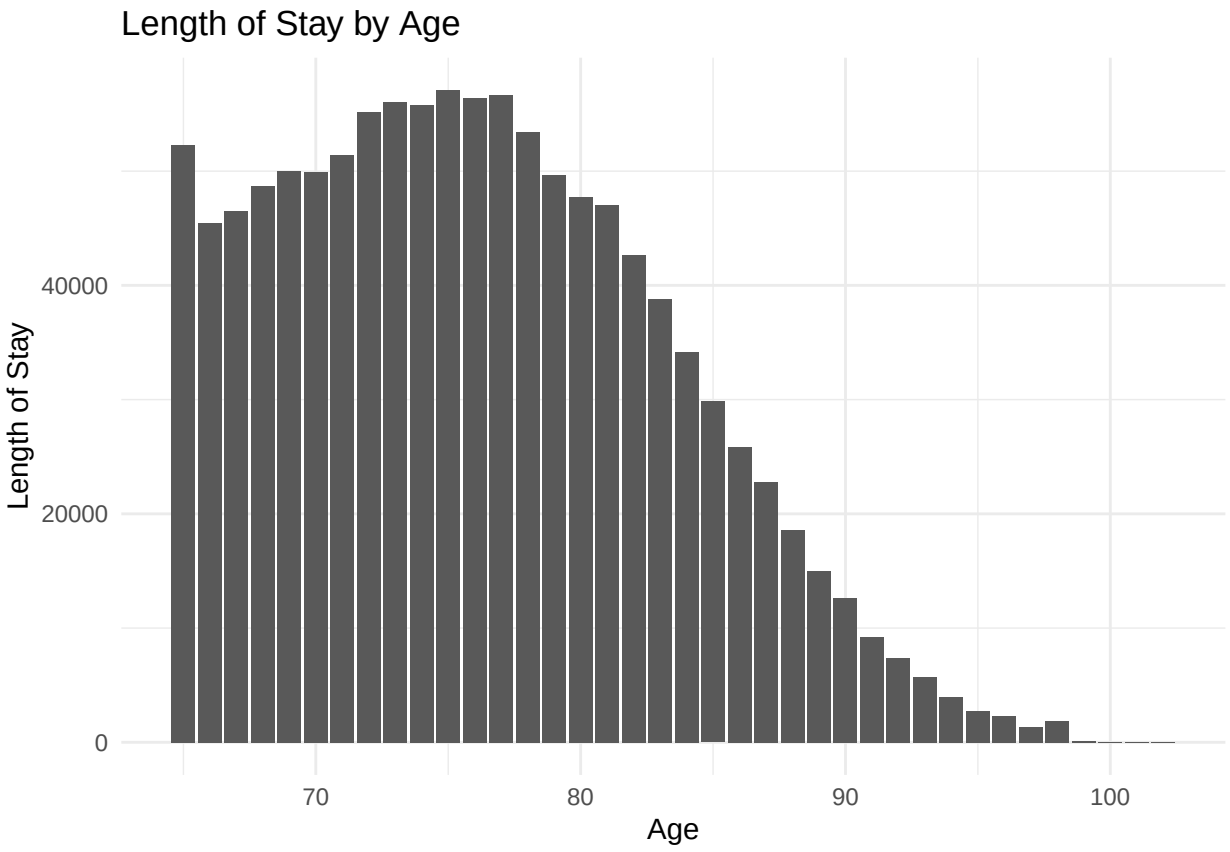
```r
ggplot(cms_hospital, aes(x = as.factor(discharge), fill = as.factor(discharge))) +
  geom_bar() +
  labs(x = "Types of Discharges", y = "Count", title = "Counts of Discharges") +
  theme_minimal()
```



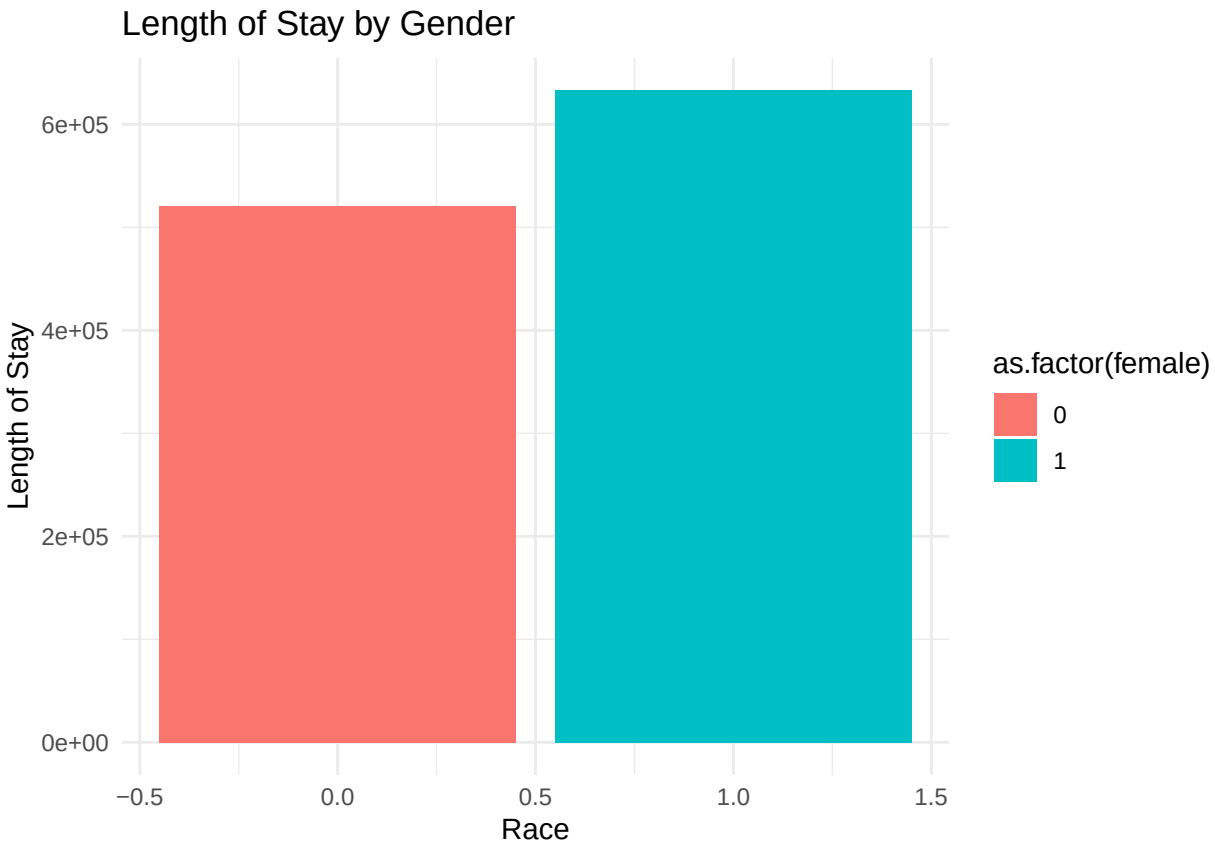

```
ggplot(cms_hospital, aes(x = race, y = LOS, fill = race)) +
  geom_bar(stat = "identity") +
  labs(x = "Race", y = "Length of Stay", title = "Length of Stay by Race") +
  theme_minimal()
```




```
ggplot(cms_hospital, aes(x = age, y = LOS)) +
  geom_bar(stat = "identity") +
  labs(x = "Age", y = "Length of Stay", title = "Length of Stay by Age") +
  theme_minimal()
```




```
ggplot(cms_hospital, aes(x = female, y = LOS, fill = as.factor(female))) +
  geom_bar(stat = "identity") +
  labs(x = "Race", y = "Length of Stay", title = "Length of Stay by Gender") +
  theme_minimal()
```



The primary research question(s) that you aim to explore with your analysis. In particular, what health outcome(s) and exposure(s) do you plan to focus on?

From the provided document, we recognize that length of stay (LOS) is an important indicator of hospital performance, as "unnecessary days spent in the hospital can lead to increased hospital-acquired patient complications (like infections or falls)." Depending on the various models and analysis we are conducting, the health outcomes that we may choose to explore are the types of discharges (a categorical variable), length of stay, and Charlson-Deyo comorbidity index on admission, and focusing on exposures such as race, age, and length of stay (LOS doubling as a predictor variable and outcome).

The following are our research questions:

(Primary) Question 1: Can a patient's length of stay (LOS), race, and age (exposures) determine their type of discharge (health outcome)?

i.e., for individuals who have a longer length of stay (more severity of pancreatic cancer?) indicate a longer type of discharge such as long-term care, or hospice care (anticipated life expectancy of < 6 months), etc. We can also explore the factors of race (if there are differences between 'Black,' 'White,' and 'Other' as well as age (similarly, if someone is older, are they more likely to be discharged to a hospice?)

Question 2: Is there a relationship between the Charson-Deyo comorbidities (7 levels outcome variable-weighted index to predict risk of death within 1 year of hospitalization for patients with specific comorbid conditions) and race, age and the length of stay of an individual in the hospital (exposure variable)? Higher scores indicate a more severe condition and consequently, a worse prognosis. We intend to explore the relationship between the index and the age of an individual, i.e, older individuals are more likely to get the higher levels.

Question 3: Is there an association between the outcome of length of stay (a count) and race, age and gender?

An outline of the extension you plan to pursue.

You should list what resources you will use to learn about your topic (e.g., textbooks, review articles), as well as the R package(s) or function(s) you feel might be relevant. You will also provide a short (two to four sentence) description of what your extension is and how it builds on ideas from class and/or relates to the data you intend to analyze.

Instead of a logistic regression model with a binary outcome of 0/1 we are exploring an outcome variable (discharge) of 9 different levels to which we can predict using a combination of numerical variables (age, length of stay) as well as categorical variables (race, which can be factored into 3 levels, and gender with two levels). Discharges has multiple levels (1. Home, 2. HomeCare, 3. SNF/ICF, 4. Hospice, 5. Rehab, 6. Inpatient, 7. LTC, 8. Swing Bed, 9. Other), thus we can explore this outcome by using a multinomial logistic regression model, which is an extension of logistic regression models to nominal categorical outcomes with more than 2 levels (https://bookdown.org/sarahwerth2024/CategoricalBook/multinomial-logit-regression-r.html) since in our case, we have 9 levels in our outcome variable. R packages include nnet, broom, ggeffects, marginaleffects, from this online textbook.

For our analysis using a method we have already learned from from SDS 390, we will use a Poisson regression model for a count outcome (LOS) (corresponding to question 3). This will help deepen our understanding and analysis when fitting the multinomial logistic regression model.

My My is also taking SDS 293 this semester, and is interested in using LASSO for understanding which predictor variables are more significant in a model (in place of the Likelihood Ratio Tests), but this is just an idea!