Midterm Presentation

Kathryn Doorley

Math 421

Loading and Cleaning Data

- Used haven library to load the SAS dataset
- Filtered for patients discharged in 2018
- Selected only 49 columns
- Removed the following columns with missing values
  - Payfix
  - Preopday
  - Obs_hour
  - Nicu_day

Exploratory Data Analysis

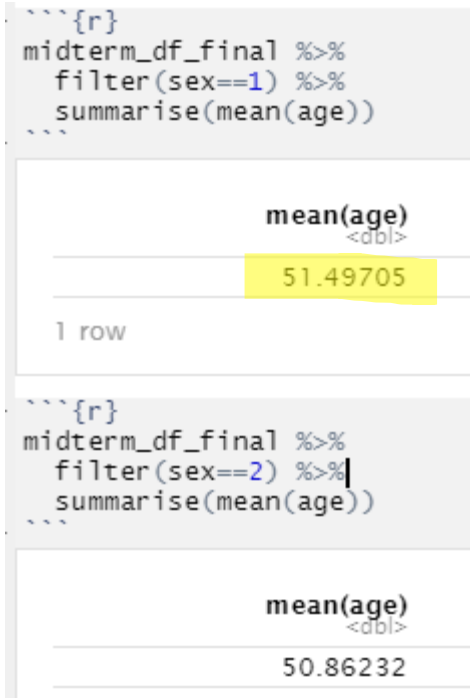
- What month saw the most amount of patients?

```{r}
midterm_df_final %>%
  group_by(moa) %>%
  count()
```

moa <dbl>	n <int>
1	11309
2	10259
3	11073
4	11174
5	11404
6	10690
7	10962
8	11199
9	10666
10	11408
11	10636
12	10698

EDA Continued

- What sex has the higher mean age?



```r
midterm_df_final %>%
  filter(sex==1) %>%
  summarise(mean(age))
```

mean(age) <dbl>
51.49705

1 row

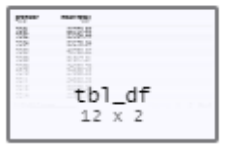
```r
midterm_df_final %>%
  filter(sex==2) %>%
  summarise(mean(age))
```

mean(age) <dbl>
50.86232

EDA Continued

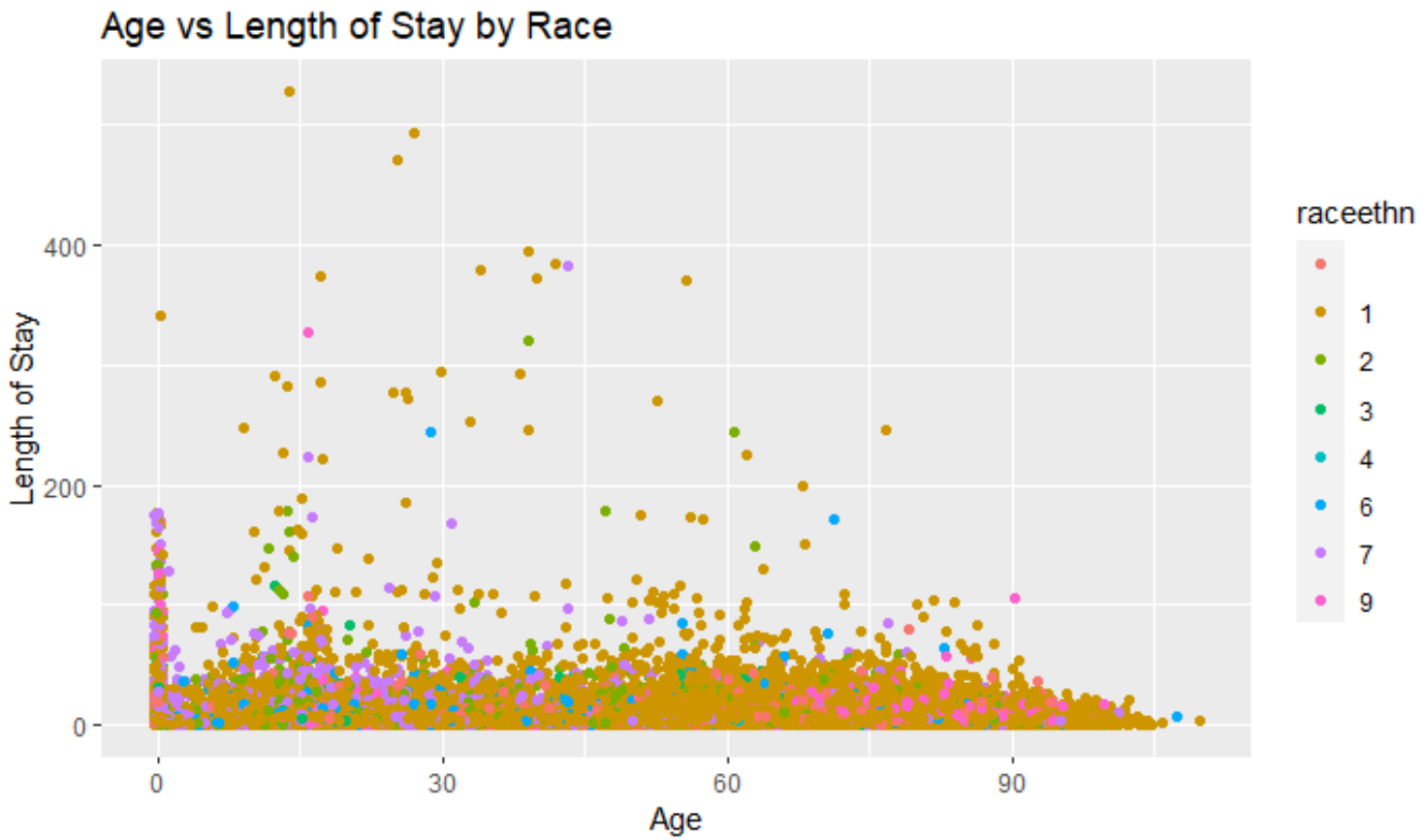
- What provider has the highest cost?

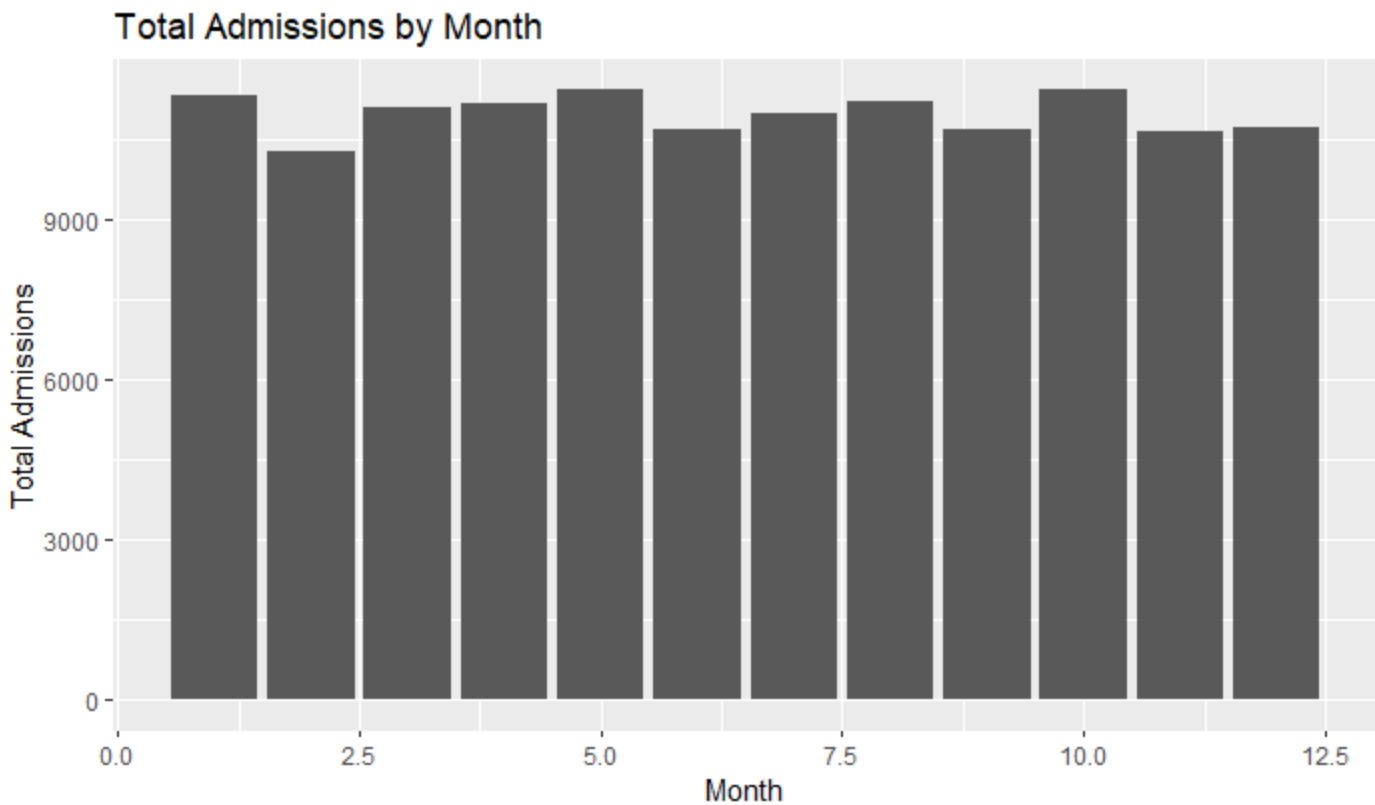
```r
midterm_df_final$total <- as.numeric(midterm_df_final$total)
midterm_df_final %>%
  group_by(provider) %>%
  summarise(mean(total))
```



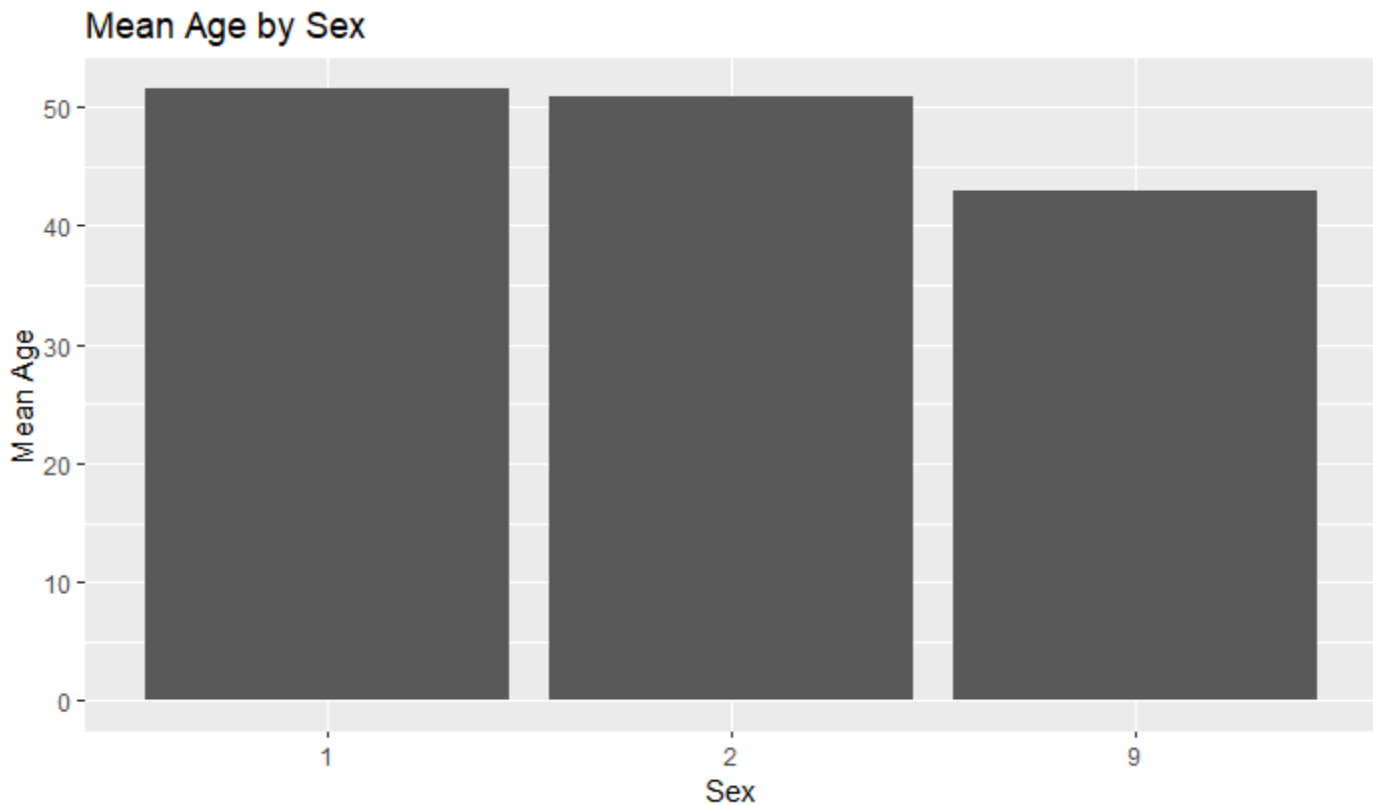
provider <chr>	mean(total) <dbl>
7201	22775.33
7202	35504.44
7204	35276.94
7205	48738.82
7206	31017.81
7209	24538.79
7210	27690.88
7211	24088.58
7213	38200.23
7214	22362.14
7215	69945.55
7216	17781.83

Plots

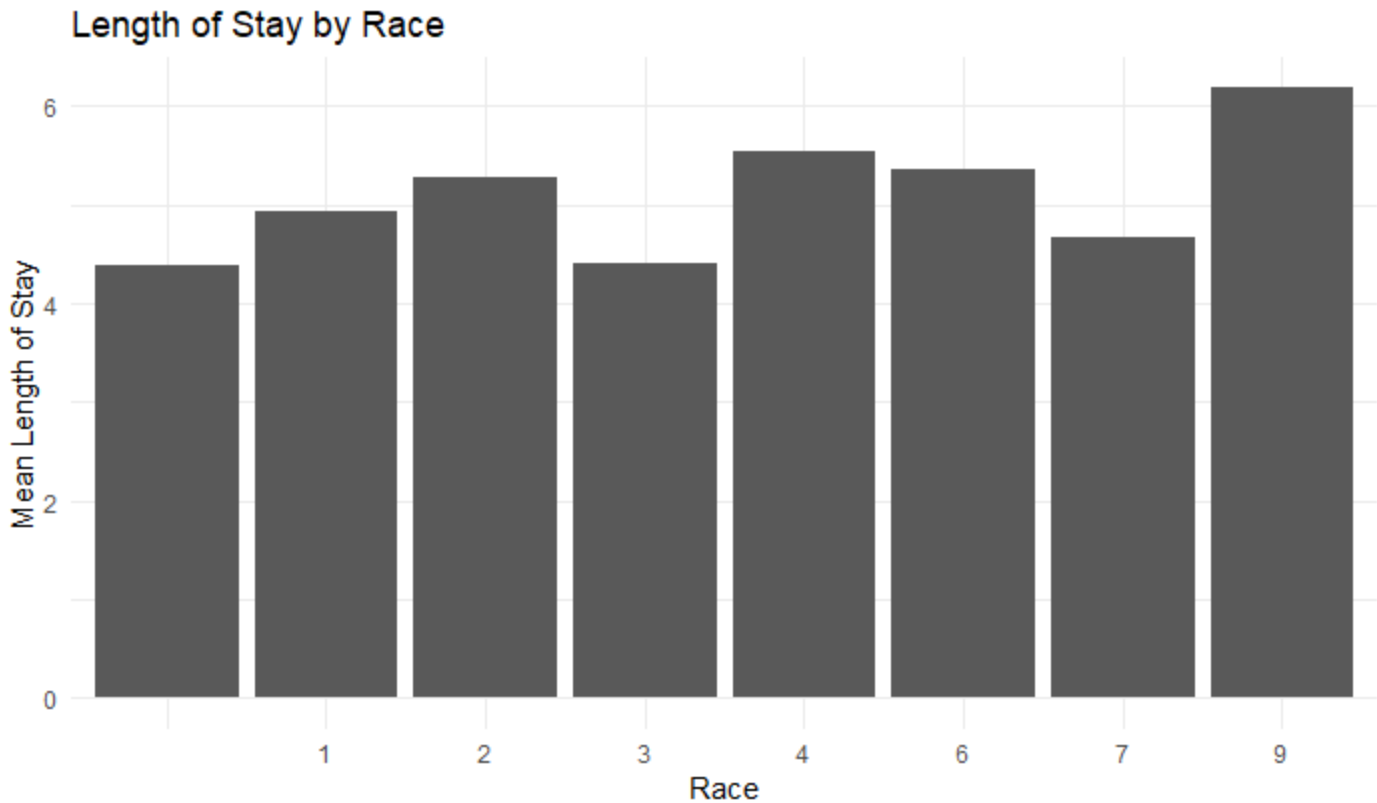

Age vs Length of Stay by Race
Length of Stay
400
200
0
0
30
60
90
Age
raceethn
1
2
3
4
6
7
9



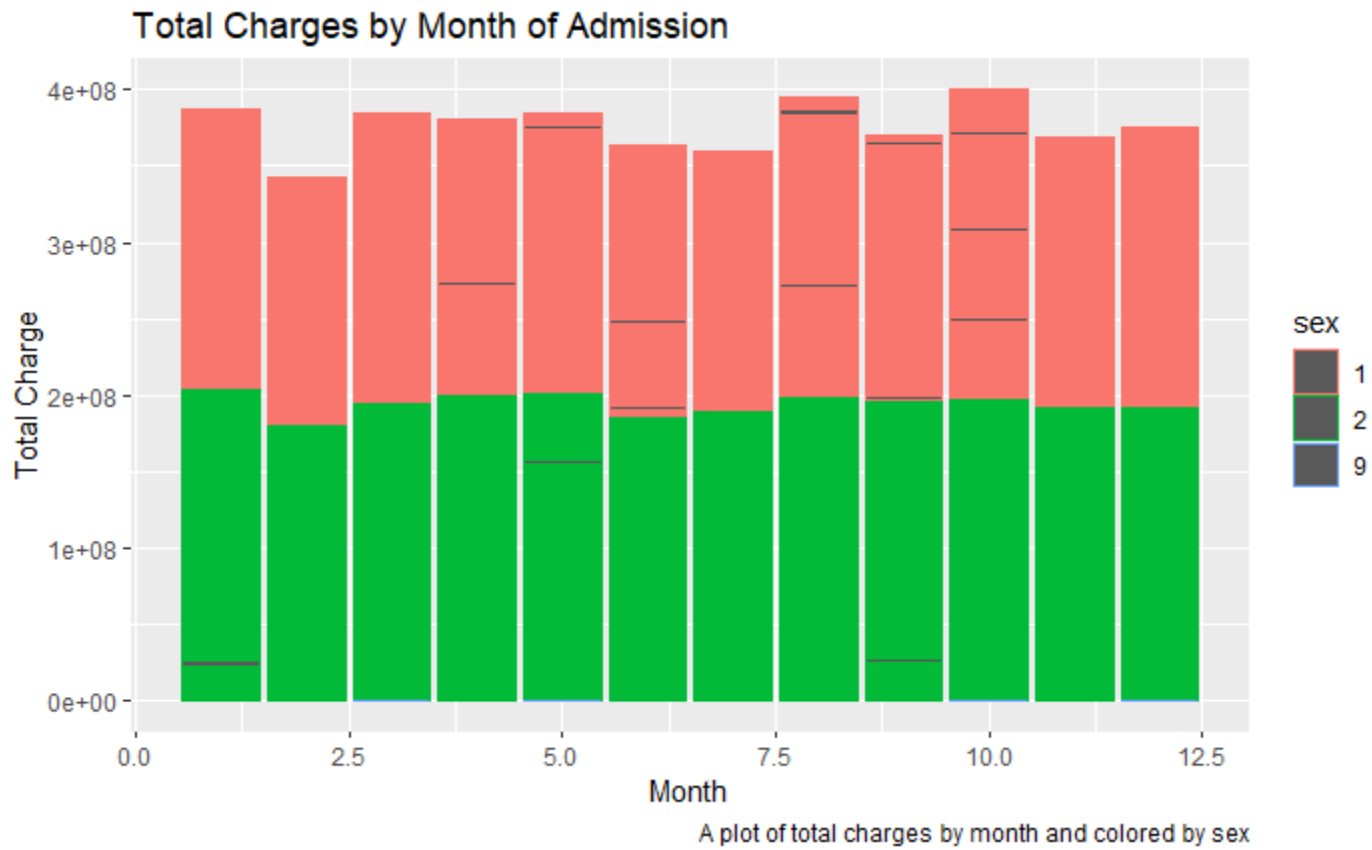
A plot of total admissons by month, January has the highest total where the next month of Feburary has the lowest.



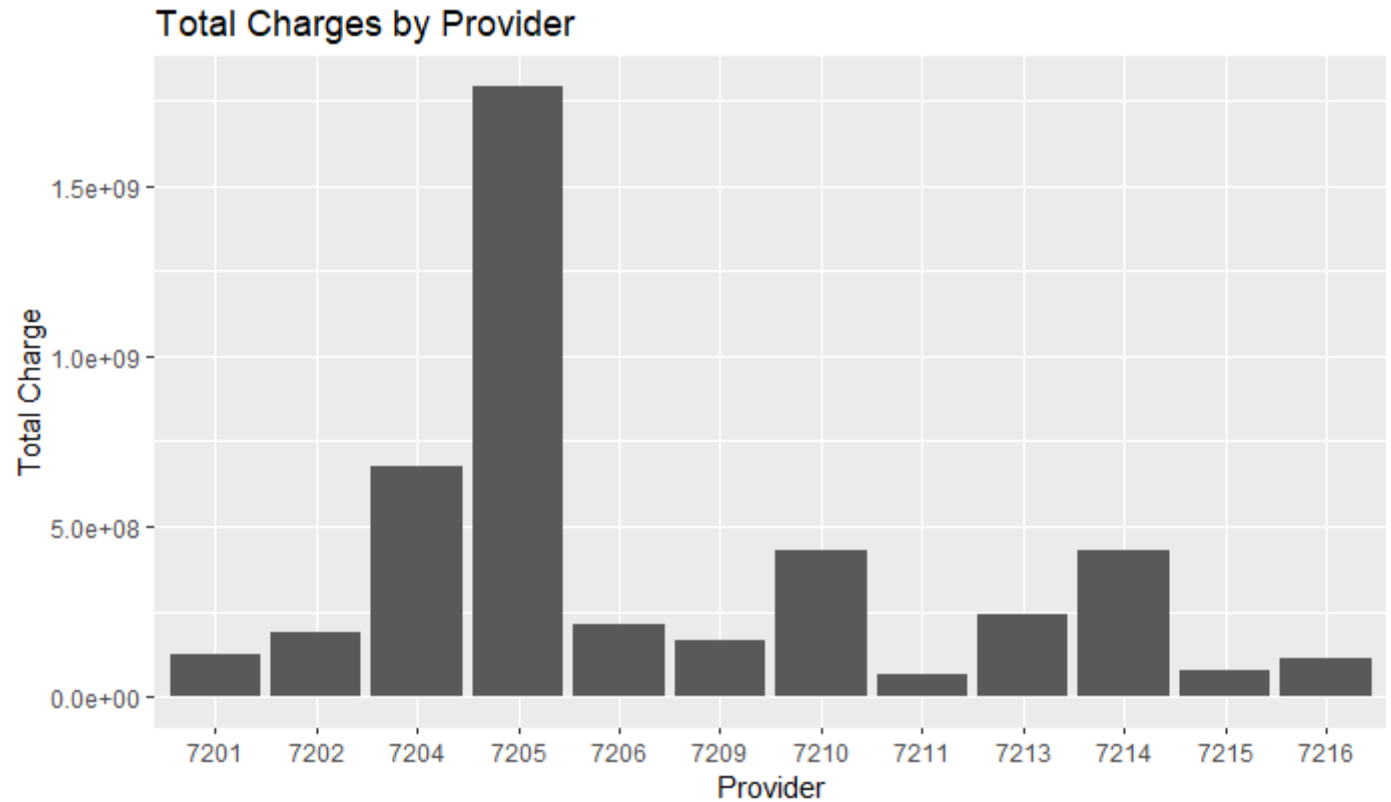
A plot of average age by sex, the female admitted patients have a lower age than the male patients



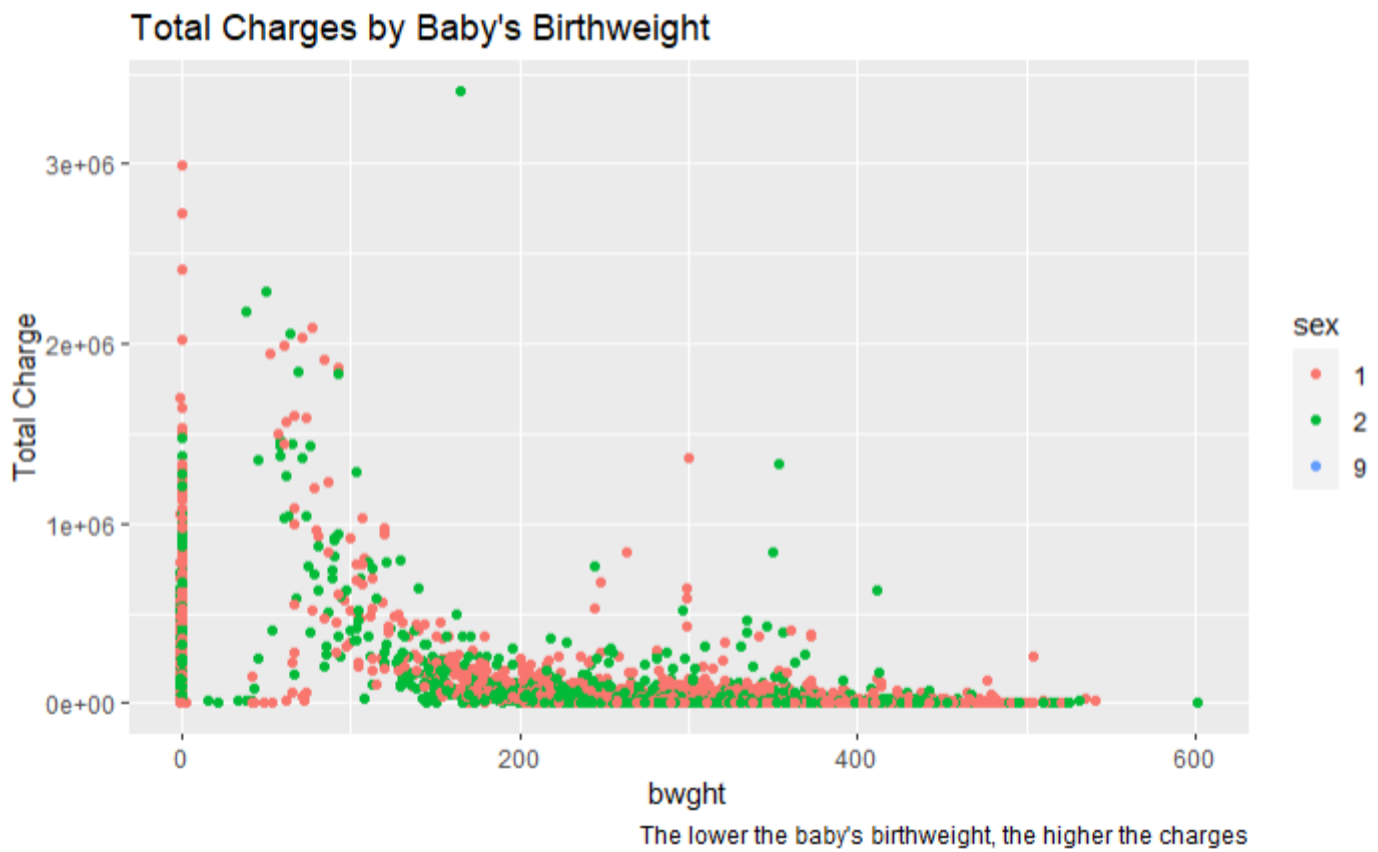
A plot of average length of stay by race.From the plot you can see Unknown race (9) has the highest average length of stay.

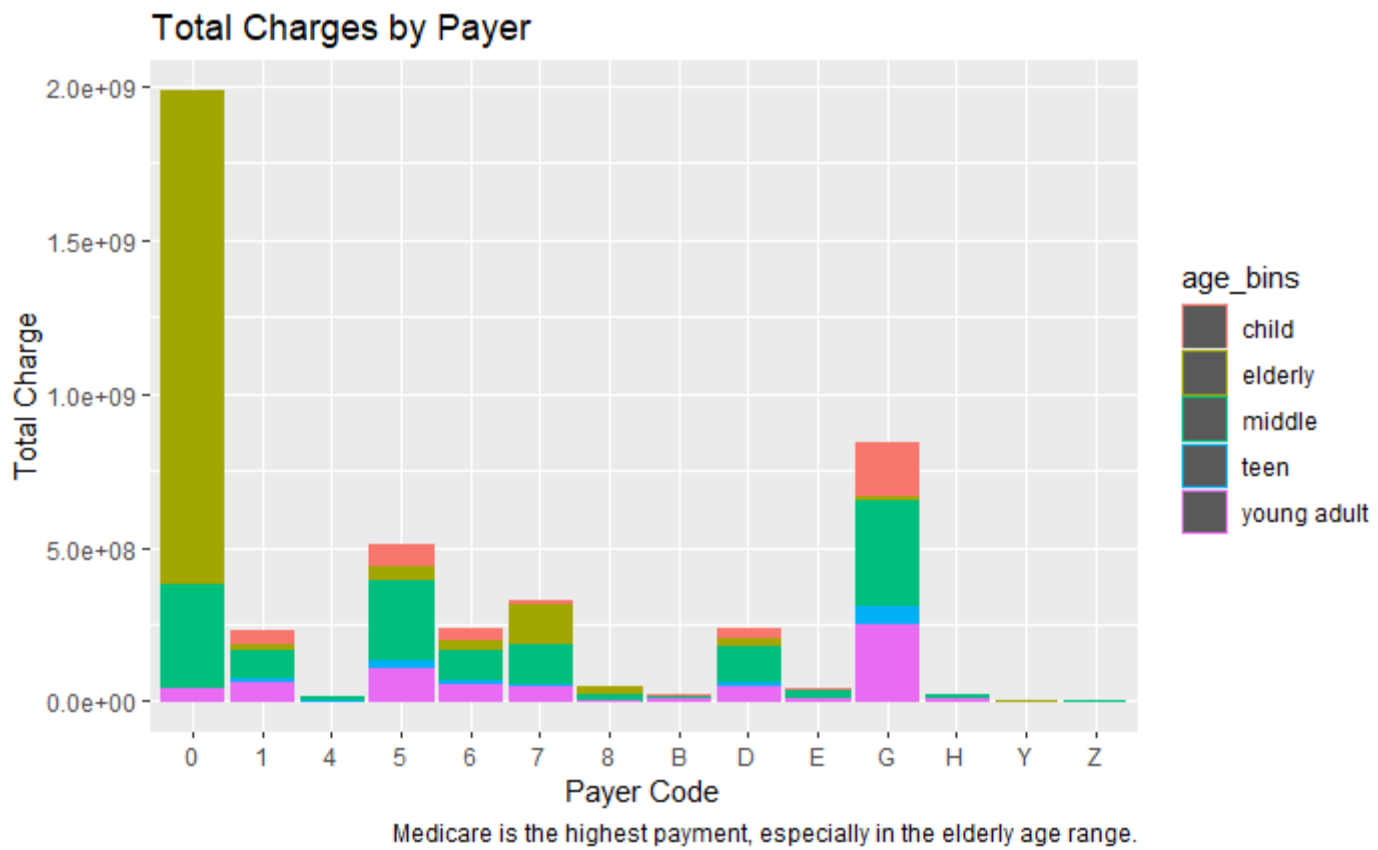

Total Charges by Month of Admission
4e+08
3e+08
2e+08
1e+08
0e+00
Total Charge
0.0
2.5
5.0
7.5
10.0
12.5
Month
sex
1
2
9
A plot of total charges by month and colored by sex

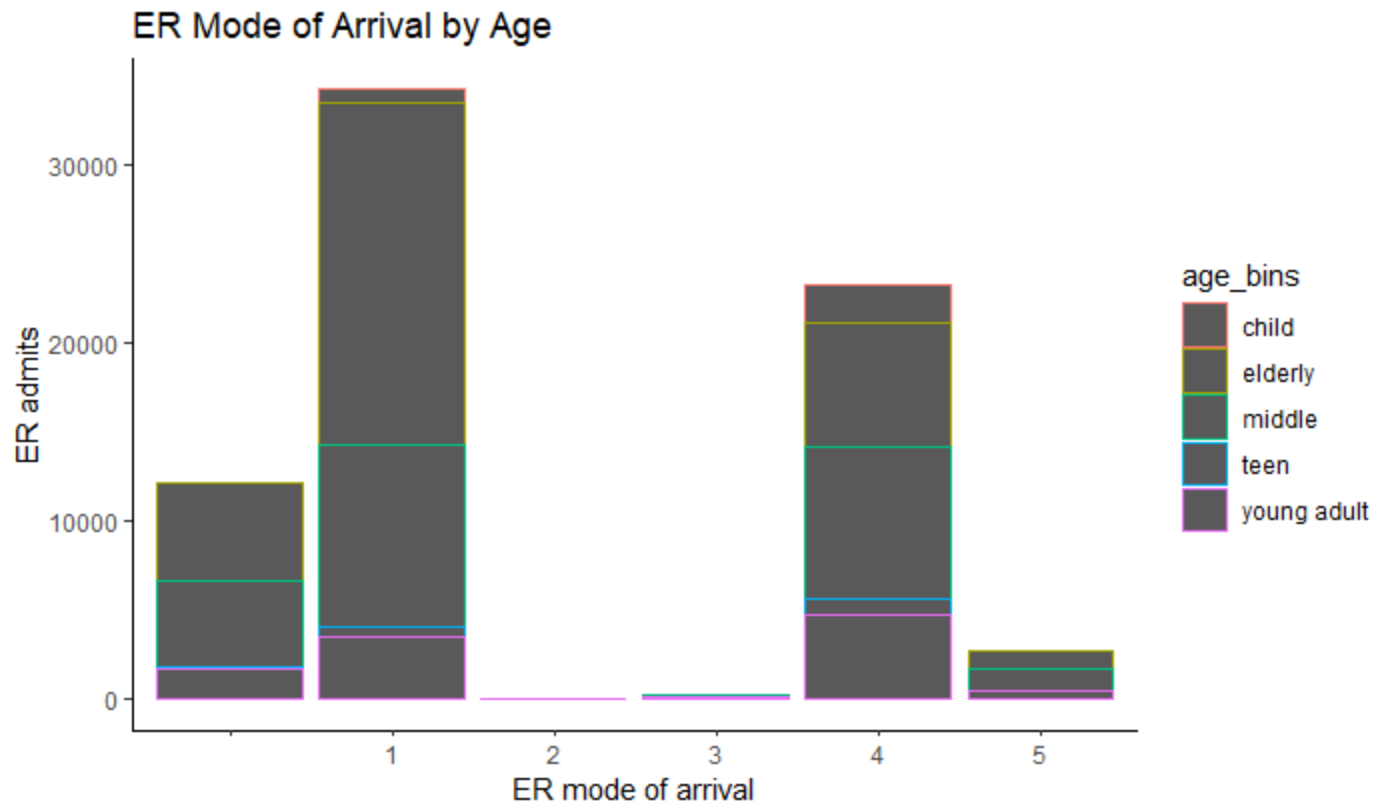
Density plot of the age variable. There are three distinct peaks, one for babies, 30-35 year olds, and elderly people



ıd Hospital, this makes sense because they are the largest hospital in Rhode Island and they do more complicated procedures

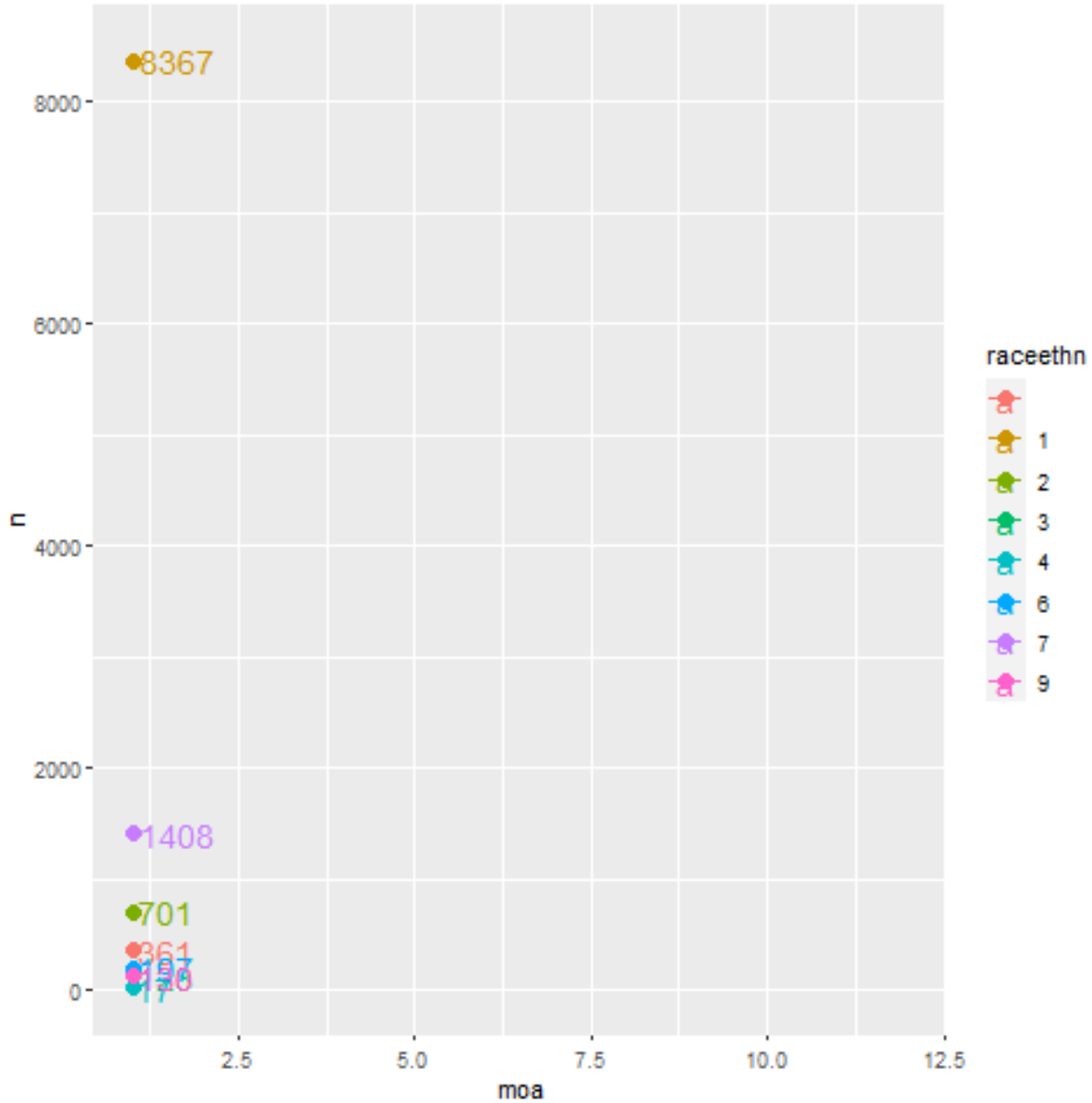

Total Charges by Baby's Birthweight
Total Charge
3e+06
2e+06
1e+06
0e+00
0
200
400
600
bwght
sex
1
2
9
The lower the baby's birthweight, the higher the charges


Total Charges by Payer
2.0e+09
1.5e+09
1.0e+09
5.0e+08
0.0e+00
Total Charge
0
1
4
5
6
7
8
B
D
E
G
H
Y
Z
Payer Code
age_bins
child
elderly
middle
teen
young adult
Medicare is the highest payment, especially in the elderly age range.



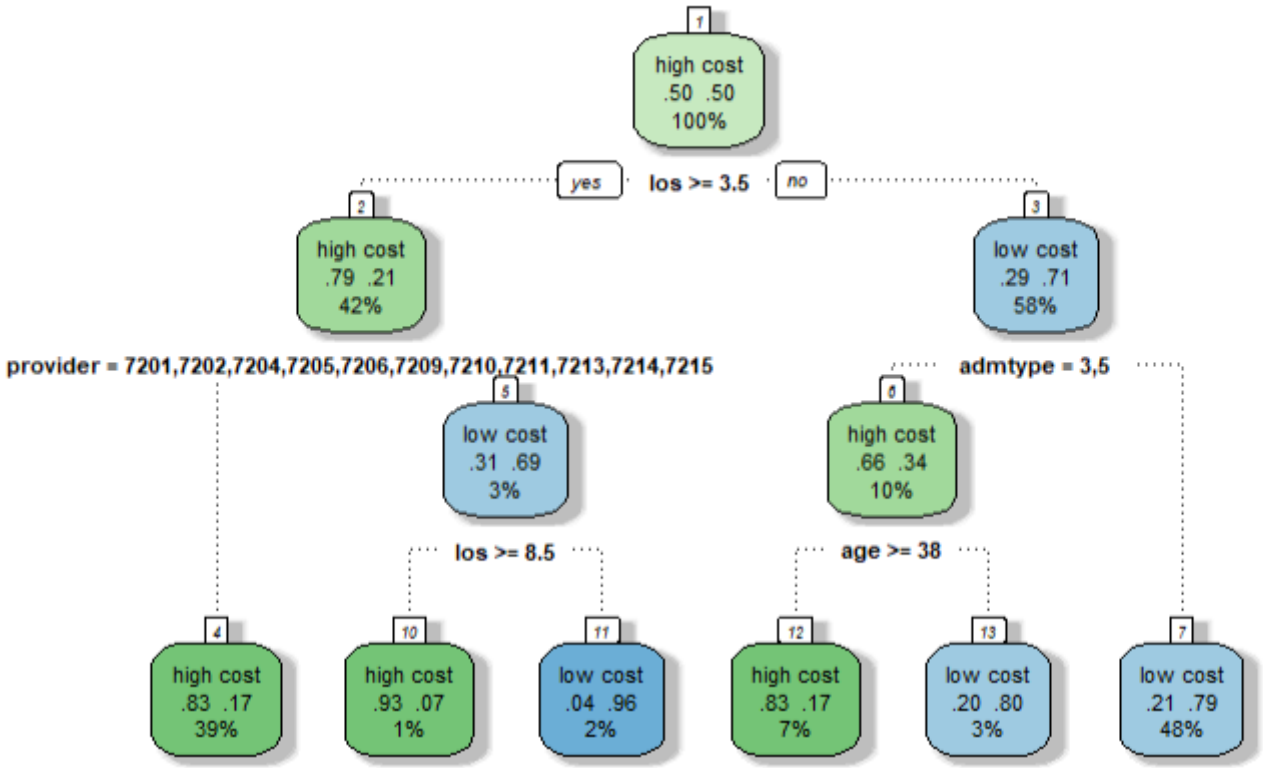
Filtered out for NA and Unknown, the highest was through an ambulance or personal transportation

Hospital Admissions by Month Colored by Race



Predictive Modeling

Tree Model


1
high cost
.50 .50
100%
yes
los >= 3.5
no
2
high cost
.79 .21
42%
3
low cost
.29 .71
58%
provider = 7201,7202,7204,7205,7206,7209,7210,7211,7213,7214,7215
admtype = 3,5
5
low cost
.31 .69
3%
6
high cost
.66 .34
10%
los >= 8.5
age >= 38
4
high cost
.83 .17
39%
10
high cost
.93 .07
1%
11
low cost
.04 .96
2%
12
high cost
.83 .17
7%
13
low cost
.20 .80
3%
7
low cost
.21 .79
48%

Model Choice 1- High/Low Cost

- Random Forest method had highest accuracy:

5. What is your final selection for the model? Test the accuracy of your final model on the test data.

```{r}
pred <- predict(forest_cv, df_test)
cm <- confusionMatrix(data = pred, reference = df_test$target, positive = "high cost")
cm$overall[1]
```

```
  Accuracy
0.8275247
```

Model Choice 2- Old/Young

- Parallel Random Forest had highest accuracy

```r
pred <- predict(parRF_cv2, df_test2)
cm <- confusionMatrix(data = pred, reference = df_test2$target, positive = "old")
cm$overall[1]
```

```
Accuracy
0.867074
```