Inpatient Charge Data 2016

Antonio Avila

April 6, 2019

Begin by loading in the data

```
med_data = read_csv("medicare_data.csv", guess_max = 112000)
```

```
## Parsed with column specification:
## cols(
##   `DRG Definition` = col_character(),
##   `Provider Id` = col_double(),
##   `Provider Name` = col_character(),
##   `Provider Street Address` = col_character(),
##   `Provider City` = col_character(),
##   `Provider State` = col_character(),
##   `Provider Zip Code` = col_double(),
##   `Hospital Referral Region (HRR) Description` = col_character(),
##   `Total Discharges` = col_number(),
##   `Average Covered Charges` = col_character(),
##   `Average Total Payments` = col_character(),
##   `Average Medicare Payments` = col_character()
## )
```

```
real_names = names(med_data)
names(med_data) <- c("DRG", "ID", "Provider", "Address", "City", "State", "Zip", "HRR", "Discharges", "
```

There seems to be a problem parsing the data. The variable "Total Discharges" Doesnt read in a few of the observations correctly because they're value is above 1,000. The commas seem to be affecting the parsing of those particular observations. In addition, The charges and payments are being parsed in as character types instead of numeric (or doubles) because of the dollar sign.

```
parse2num <- med_data %>%
    select("AvgCharge":"AvgMedPmts") %>%
    map(parse_number) %>%
    as_tibble()

med_data2 <- med_data %>%
  select(-("AvgCharge":"AvgMedPmts")) %>%
  bind_cols(parse2num)
```

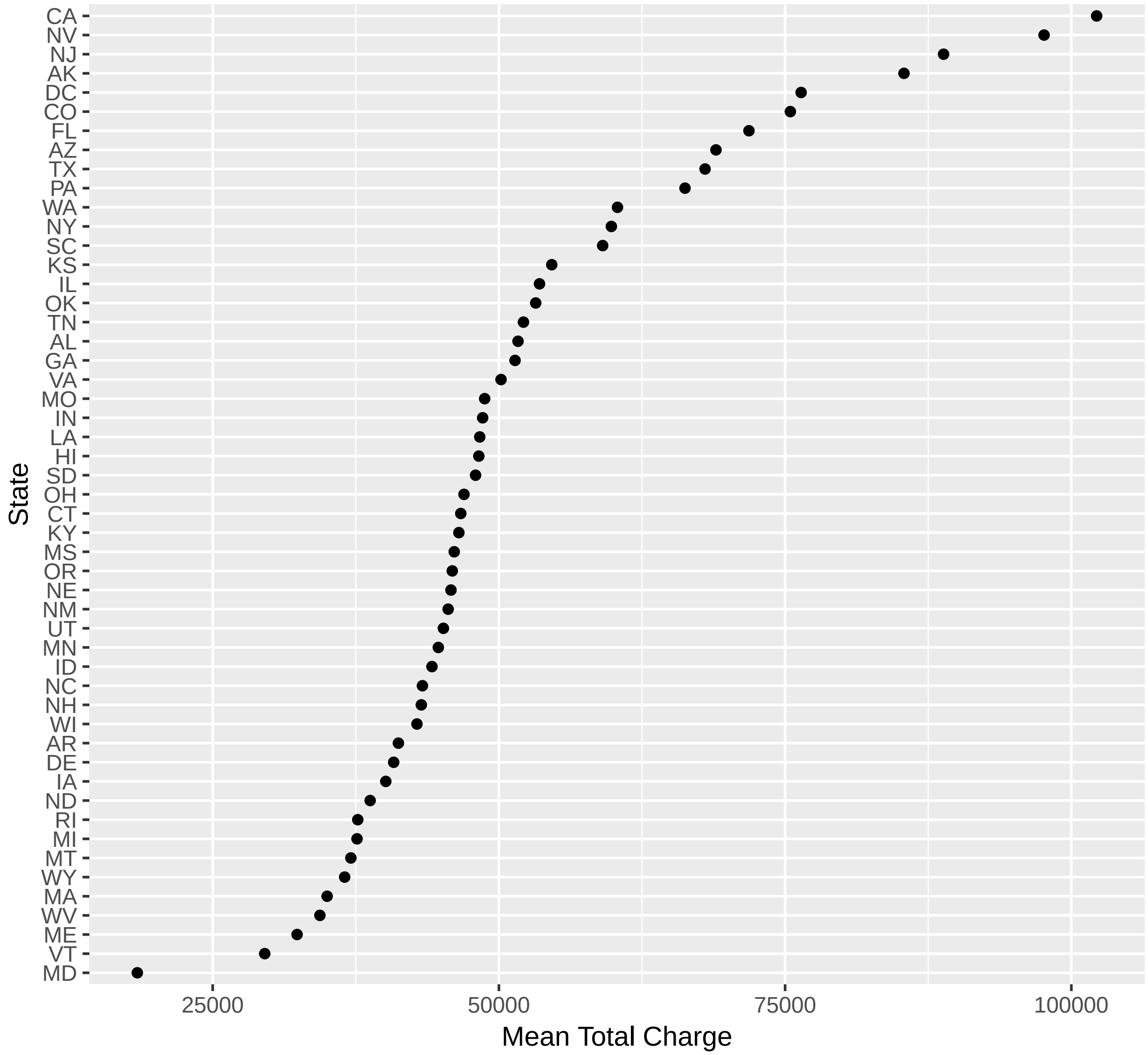
Fixed the parsing issue for the Total Discharges column by extending the number of rows the read_csv() function reads in to determine the type of column it is to 120,000 and by converting the Average dollar pament columns into numeric columns, dropping the dollar symbol. Extending the number of rows to read in worked since the first occurence of a discharge being over 1,000 was at about the 117,000th row, thus removing the comma as a grouping mark and changing it into a normal double.

Having fixed the parsing issues, I can finally begin cleaning the data a little. I will begin by separating the code and decriptions from the DRG column to shorten it.

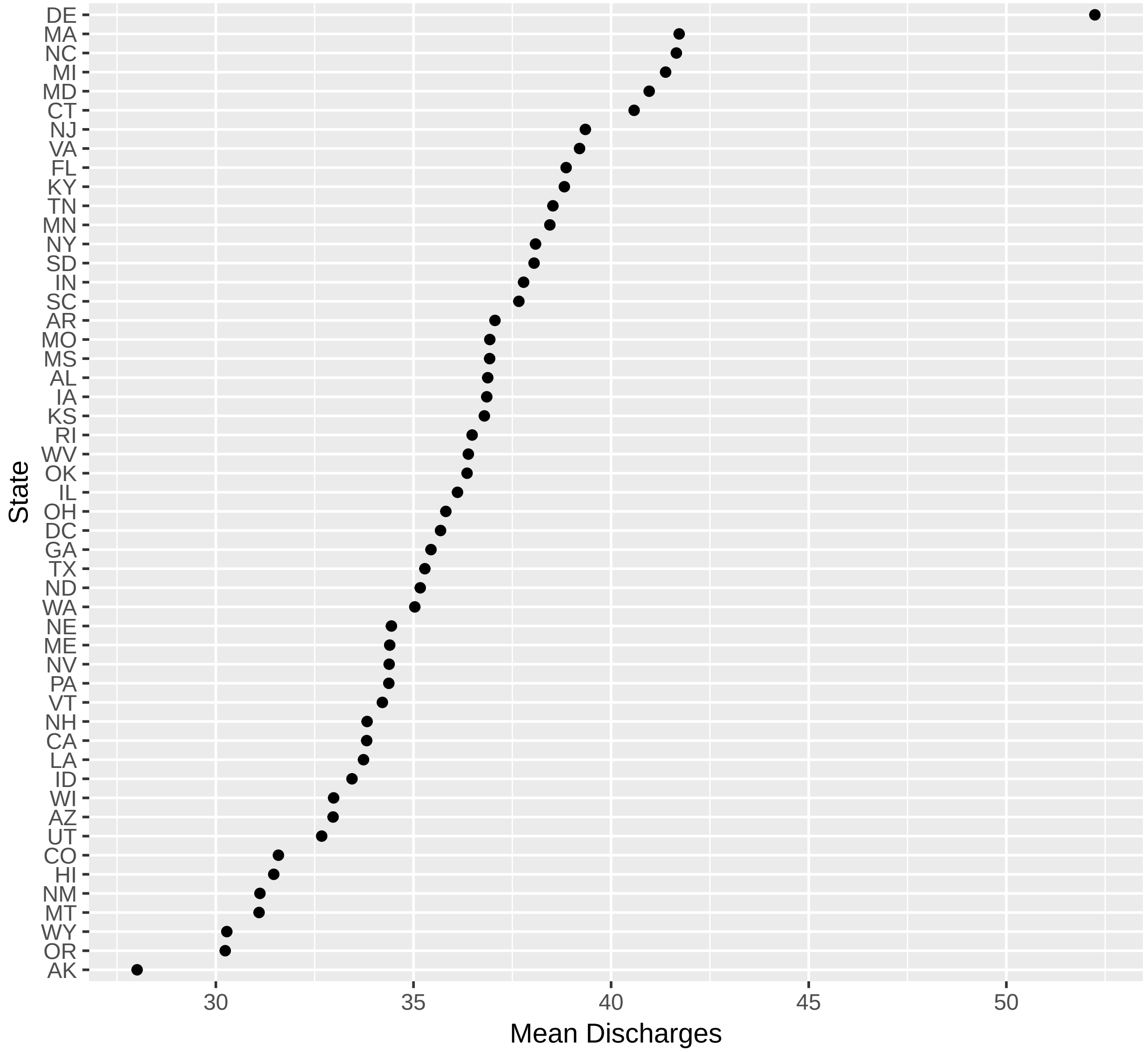
```
med_data3 <- med_data2 %>%
  separate(DRG, c("DRG_Code", "DRG_Descr"), sep = 3)

med_data3$DRG_Descr = str_sub(med_data3$DRG_Descr, 4)
```

```r
med_data3 %>%
  group_by(State) %>%
  summarise(mean_charge = mean(AvgCharge)) %>%
  ggplot(aes(mean_charge, reorder(State, mean_charge))) +
    geom_point() +
    labs(x = "Mean Total Charge", y = "State")
```



```r
med_data3 %>%
  group_by(State) %>%
  summarise(mean_disch = mean(Discharges)) %>%
  ggplot(aes(mean_disch, reorder(State, mean_disch))) +
    geom_point() +
    labs(x = "Mean Discharges", y = "State")
```



```
# med_data3 %>%
#   ggplot(aes(Discharges, AvgCharge)) +
#     geom_point(aes(color = State))
#
```

Intention in the future is to build some kind of heat map to see how location plays into a role. In addition, I also plan to ook into the politics of the states and how their medicare funding has been affected by the support or rejection of ACA policies.