DA5020 Practicum 1

Katie McCreedy, Melissa Miller, Maia Woodard

Contents

Part 1

Question 1

```
# Part I: Question 1 - create new dataframe & variables

doctor_type <- c("PCP", "Psychiatrist", "Surgeon", "Anesthesia")
doctor_lastname <- c("Smith", "Dame", "Jones", "Zayas")
location <- c("MA", "ME", "NH", "VT")
AVG_Rating <- c(7,9,8,9)

doctors_df <- data.frame(doctor_type, doctor_lastname, location, AVG_Rating)
print(doctors_df)
```

```
##    doctor_type doctor_lastname location AVG_Rating
## 1          PCP           Smith       MA          7
## 2 Psychiatrist            Dame       ME          9
## 3      Surgeon           Jones       NH          8
## 4   Anesthesia           Zayas       VT          9
```

Question 2

```
# Part I: Question 2 - select rows/columns

# select row 1, column 2
doctors_df[1, 2]
```

```
## [1] "Smith"
```

Smith was selected.

```
# select rows 2-4
doctors_df[2:4, ]
```

```
##    doctor_type doctor_lastname location AVG_Rating
## 2 Psychiatrist            Dame       ME          9
## 3      Surgeon           Jones       NH          8
## 4   Anesthesia           Zayas       VT          9
```

Rows 2-4 were selected: the psychiatrist, surgeon, and anesthesiologist.

```
# select last column in AVG_Rating
doctors_df[, 4]
```

```
## [1] 7 9 8 9
```

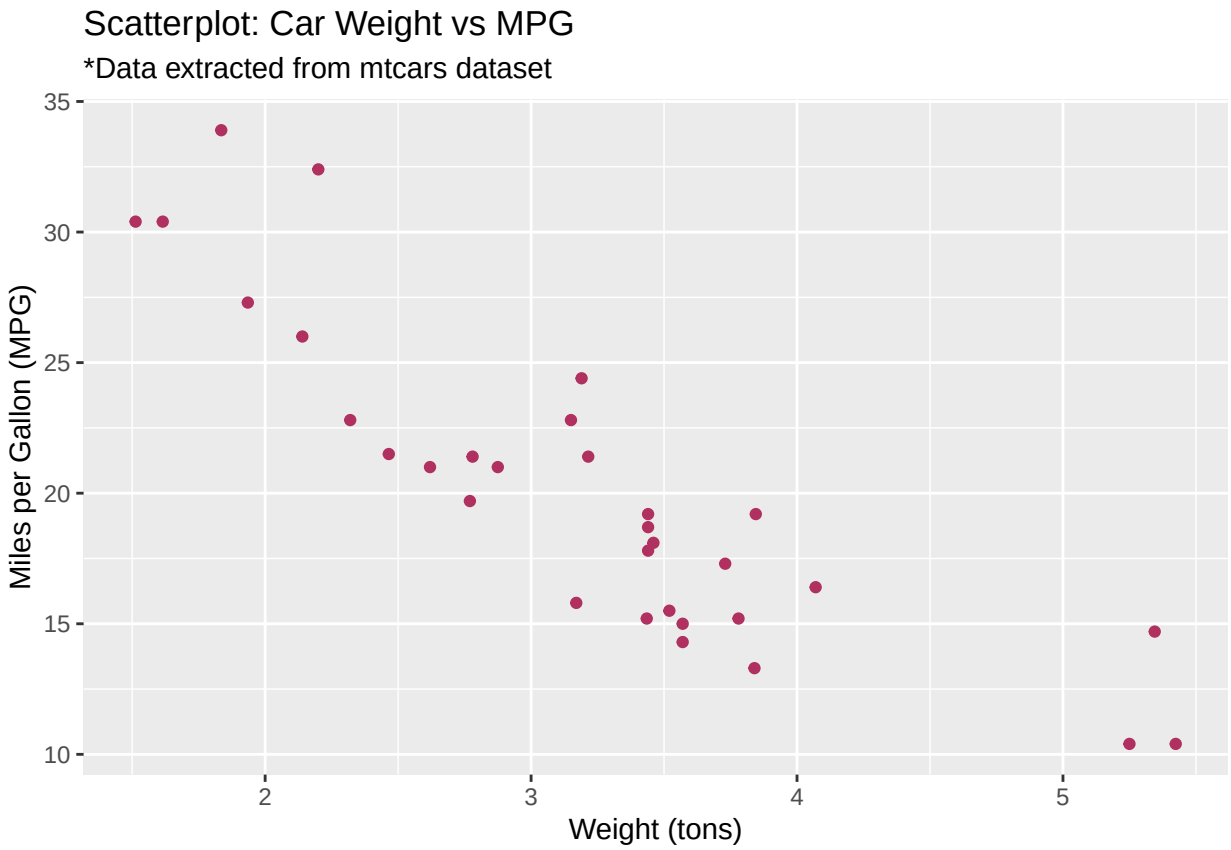
The last column, with values 7, 9, 8, 9 was selected.

Question 3

```
# load libraries
library(ggplot2)

# load data
data("mtcars")

# Scatter Plot Mtcars
ggplot(mtcars, aes(x=wt, y=mpg)) +
  geom_point(color = "Maroon") +
  labs(x = "Weight (tons)", y = "Miles per Gallon (MPG)",
       title = "Scatterplot: Car Weight vs MPG",
       subtitle = "*Data extracted from mtcars dataset")
```



This scatterplot depicts the relationship between weight & MPG in the mtcars dataset. I chose these variables because they have the clearest logical connection without subject matter knowledge about car functionality — i.e. our hypothesis was that as the weight of a car increases, the miles per gallon achieved should genreally decrease in a negatively linear fashion. Indeed, this scatterplot depicts such a relationship where heavier cars have lower MPG. There are no signficant outliers.

Question 4

```
summary(mtcars$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   15.43   19.20   20.09   22.80   33.90
```

```
# Median = 19.20 mpg
summary(mtcars$wt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.513   2.581   3.325   3.217   3.610   5.424
```

```
# median = 3.325 half-tons
cor(mtcars$mpg, mtcars$wt, method = "pearson")
```

```
## [1] -0.8676594
```

The pearson coefficient of the correlation R = -0.8676. (R) measures the linear correlation between weight and mpg in the mtcars dataset. It can only be a number between –1 and 1 — such that it measures the direction/strength of the linear relationship between weight and mpg. Since the R = -0.8676, weight and mpg are strongly, negatively, linearly correlated. We picked these variables again to test if the R score was consistent with the scatterplot data and it was as they both show negative, linear correlation.

Part 2

Question 1

```
# load libraries
library(readr)
library(dplyr)
library(ggplot2)
library(psych)
```

```
# Importing data from .csv due to broken link
SUD_data <- read_csv("Substance_Use_Disorder_Treatment_Program_Admissions__Beginning_2007 (2).csv")
```

Question 2

```
any(is.na(SUD_data))
```

```
## [1] FALSE
```

There are no `NA` values in this dataset, so we don't have to worry about removing them.

```
str(SUD_data)
```

```
## spc_tbl_ [99,367 x 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Year                      : num [1:99367] 2007 2007 2007 2007 2007 ...
##  $ County of Program Location: chr [1:99367] "Albany" "Albany" "Albany" "Albany" ...
##  $ Program Category          : chr [1:99367] "Crisis" "Crisis" "Crisis" "Crisis" ...
##  $ Service Type              : chr [1:99367] "Medical Managed Detoxification" "Medical Managed Detoxi
##  $ Age Group                 : chr [1:99367] "Under 18" "18 through 24" "18 through 24" "18 through 2
##  $ Primary Substance Group   : chr [1:99367] "Heroin" "All Others" "Other Opioids" "Heroin" ...
##  $ Admissions                : num [1:99367] 4 2 6 132 35 8 1 11 276 135 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Year = col_double(),
##   ..   `County of Program Location` = col_character(),
##   ..   `Program Category` = col_character(),
##   ..   `Service Type` = col_character(),
##   ..   `Age Group` = col_character(),
##   ..   `Primary Substance Group` = col_character(),
##   ..   Admissions = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

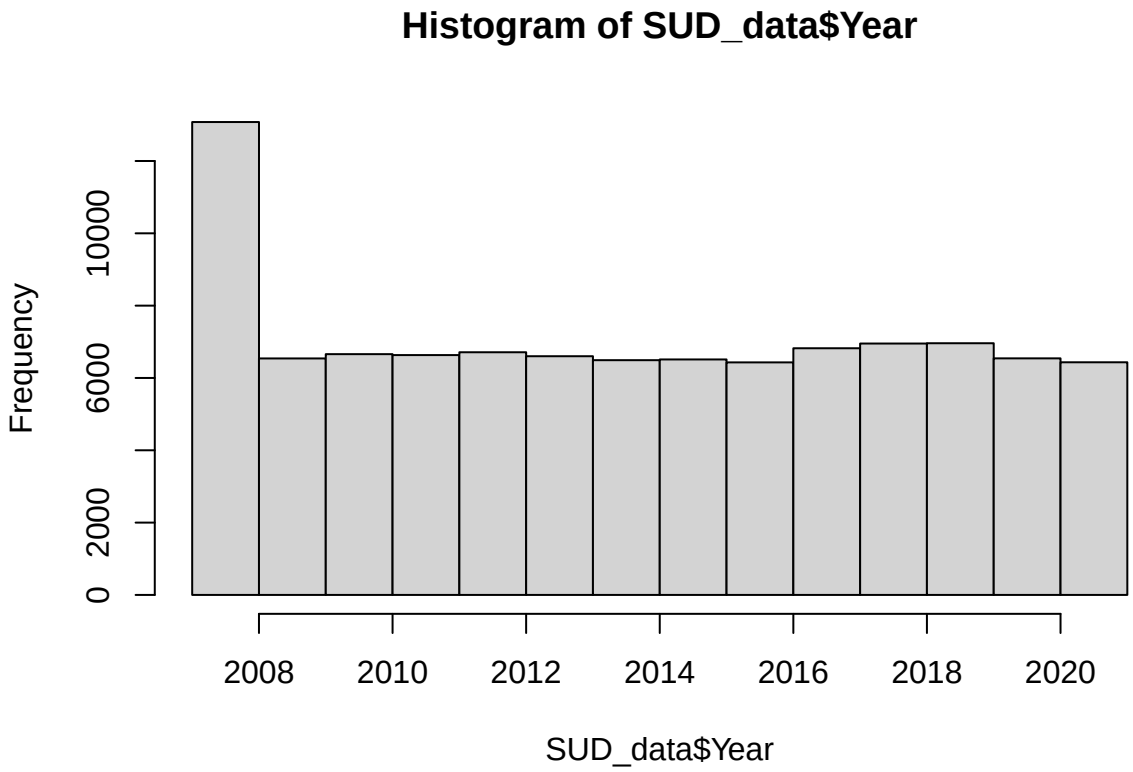
All columns seem to be of an appropriate type, so they don't need to be converted.

```
summary(SUD_data$Year)
```

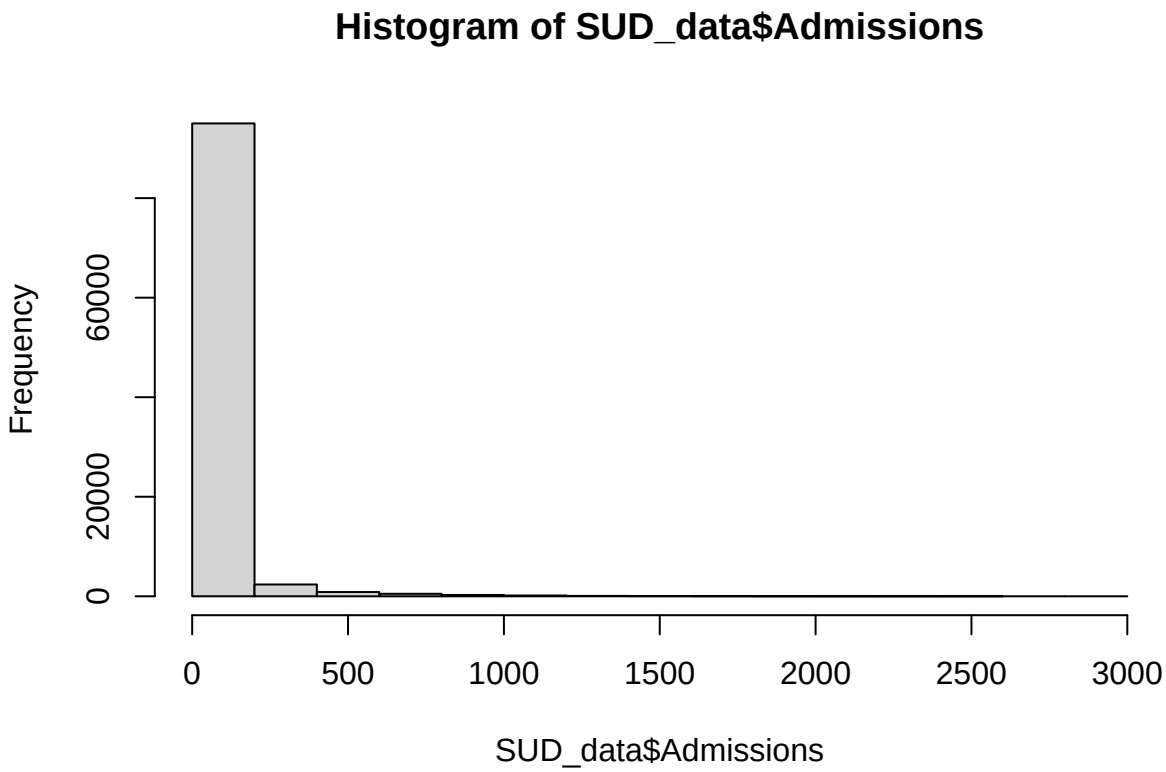
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2007    2010    2014    2014    2018    2021
```

We can see that the data is from the years 2007 to 2021.

```
# Check the distribution of the numerical variables
hist(SUD_data$Year)
```




```
hist(SUD_data$Admissions)
```



Neither admissions nor year appear to be normally distributed based on their histograms. As such, outliers will be removed to adjust this distribution. The interpretation of the results will also take into account that the 2008 data is twice as common as all of the other years.

```r
# summary statistics
summary(SUD_data$Admissions)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    2.00    8.00   41.91   28.00 2861.00
```

```r
sd(SUD_data$Admissions)
```

```
## [1] 122.8758
```

Finally, based on these summary statistics we can see that there is a strong right skew. The "max" value of 2861 is far above the median value of 8. The mean's deviation from the median also indicates a skew. The high standard deviation also indicates that the data is widely distributed, likely owing to the presence of major outliers.

```r
# The minimum is not entirely relevant as it returns a negative number
outlier_min <- mean(SUD_data$Admissions) - 3*sd(SUD_data$Admissions)
outlier_max <- mean(SUD_data$Admissions) + 3*sd(SUD_data$Admissions)
```

```r
length( SUD_data$Admissions[which(SUD_data$Admissions < outlier_min |
                                  SUD_data$Admissions > outlier_max)] )
```

```
## [1] 1917
```

There are 1917 outliers. If we were to do follow-up analyses that might be affected by the presence of outliers, they might have to be identified and removed if necessary.

Question 3

county

```r
# From counties of NY DOT, this data is missing Hamilton county, so will omit
# Several counties are listed under NY, so am creating unique keys for those
# Aside from "New York" county which will remain NY
# Bronx - BX; Kings - KI; Queens - QU; Richmond - RI
county_code <- c("AL", "AG", "BX", "BM", "CA", "CY",
                 "CH", "CM", "CN", "CL", "CO", "CR",
                 "DE", "DU", "ER", "ES", "FR", "FU",
                 "GE", "GR", "HE", "JE", "KI", "LE",
                 "LI", "MA", "MO", "MG", "NA", "NY",
                 "NI", "ON", "OD", "OT", "OR", "OL",
                 "OS", "OG", "PU", "QU", "RE", "RI",
                 "RO", "SL", "SA", "SC", "SH", "SY",
                 "SE", "ST", "SU", "SV", "TI", "TO",
                 "UL", "WR", "WS", "WA", "WE", "WY",
                 "YA")
county_name <- unique(SUD_data$`County of Program Location`)
county <- data.frame(county_code, county_name)
```

program_category

```r
program_code <- c("CR", "IN", "OTP",
                  "OUT", "RES", "SP")
program_category <- c("Crisis",
                      "Inpatient",
                      "Opioid Treatment Program",
```

```
                      "Outpatient",
                      "Residential",
                      "Specialized")
program_category <- data.frame(program_code,
                               program_category)
```

primary_substance_group

```
substance_code <- c("HE", "AO", "OPI", "ALC", "CO", "MJ", "NO")
primary_substance_group <- c("Heroin", "All Others", "Other Opioids",
                             "Alcohol", "Cocaine", "Marijuana", "None")
primary_substance_group <- data.frame(substance_code,
                                      primary_substance_group)
```

admissions_data

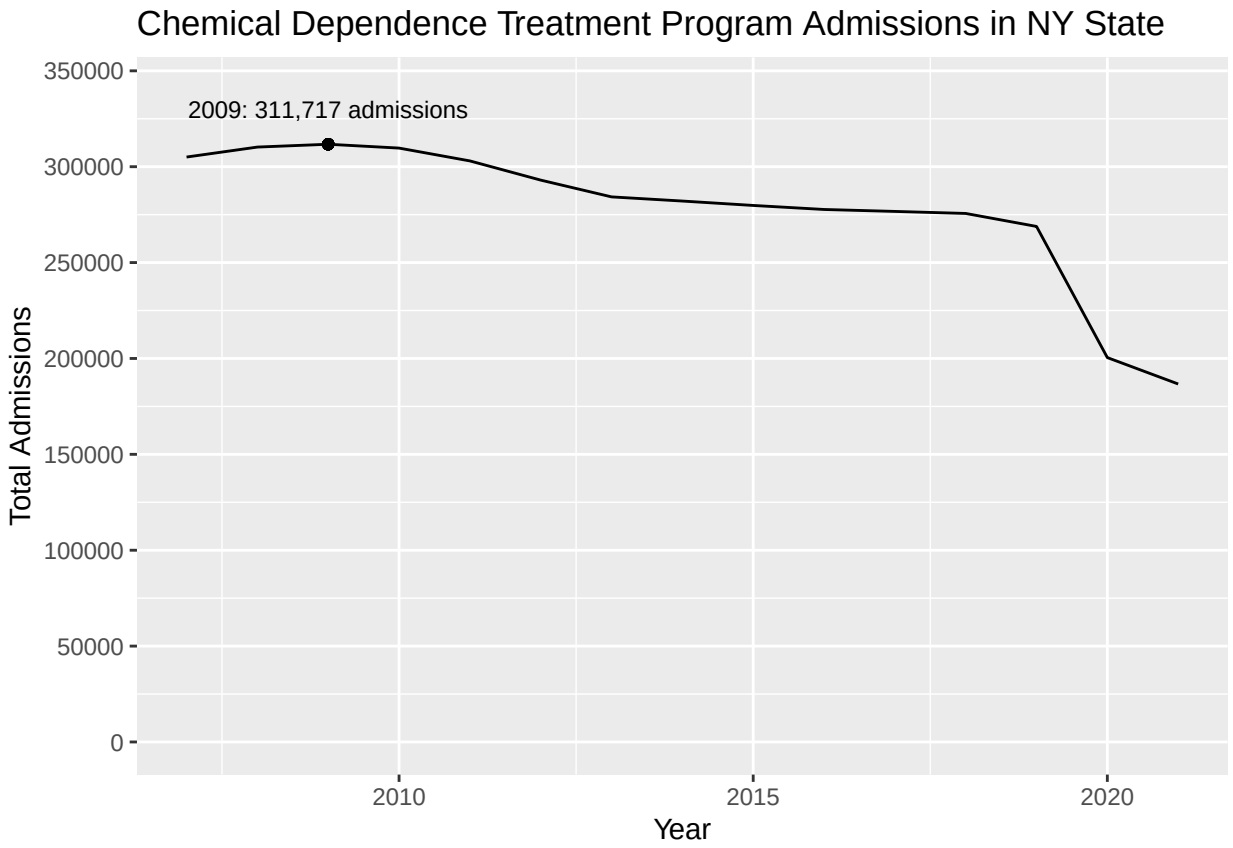
```
# Changing column names of input data to work with easier
colnames(SUD_data) <- c("Year", "county_name", "program_category",
                        "service_type", "age_group",
                        "primary_substance_group", "Admissions")
```

```
admissions_data <- SUD_data %>%
  inner_join(county) %>%
  inner_join(program_category) %>%
  inner_join(primary_substance_group) %>%
  select(Year,
         county_of_program_location = county_code,
         program_category = program_code,
         service_type, age_group,
         primary_substance_group = substance_code,
         Admissions)
```

Question 4

```
annualAdmissions <- function(){
  grouped_data <- SUD_data %>%
    group_by(Year) %>%
    summarize(total_admits = sum(Admissions))

  ggplot(grouped_data, aes(x = Year, y = total_admits)) +
    geom_line() +
    labs(x = "Year", y = "Total Admissions",
         title = "Chemical Dependence Treatment Program Admissions in NY State") +
    scale_y_continuous(n.breaks=10, limits=c(0, 340000)) +
    geom_text(aes(label = "")) +
    annotate("text", x = 2009, y = 330000, label = "2009: 311,717 admissions",
             size = 3) +
    geom_point(inherit.aes=FALSE, aes(x=2009, y=311717))
}
annualAdmissions()
```

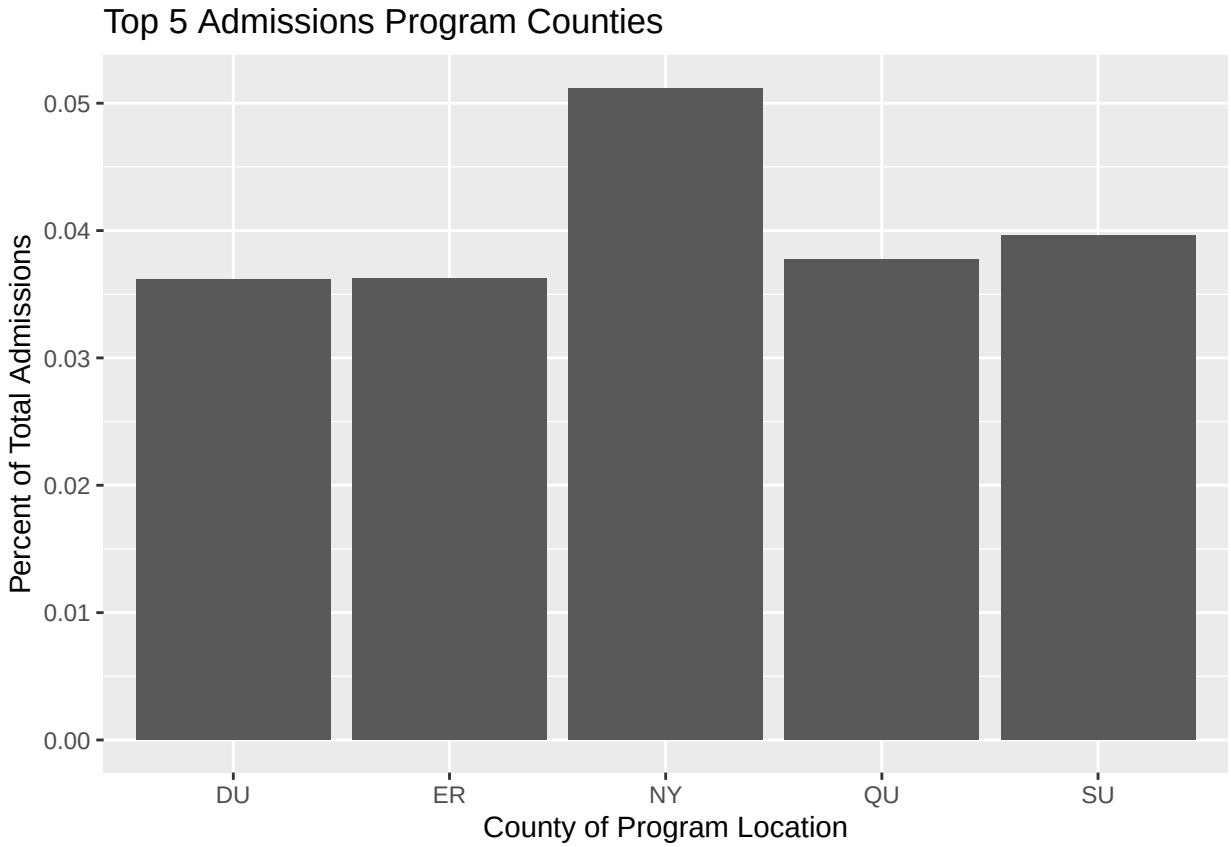


The function takes no arguments to create the chart. This chart depicts the total number of admissions to chemical dependence treatment programs per year for every county in the state of New York from 2007 to 2021. The number of admissions peaked in 2009 at 311,717, and was on a slight but steady decline through 2019. Between 2019 and 2020 the number of admissions dropped sharply, by about 68,000 cases; the number of admissions has continued to drop between 2020 and 2021. It would be interesting to explore if there were any policy or administrative changes between 2019 and 2020 accounting for this sharp drop, or if COVID-19 was at play due to a lack of clinical space or workers at treatment facilities.

Question 5

```
percentage_analysis <- admissions_data %>%
  group_by(county_of_program_location) %>%
  summarise(total=n()) %>%
  distinct() %>%
  mutate("percentage"=total/sum(total))

percentage_analysis %>%
  arrange(desc(percentage)) %>%
  head(5) %>%
  ggplot(aes(x=county_of_program_location,y=percentage)) +
  geom_col() +
  labs(x = "County of Program Location", y = "Percent of Total Admissions",
         title = "Top 5 Admissions Program Counties")
```




```
median(percentage_analysis$percentage)
```

```
## [1] 0.01116065
```

```
mean(percentage_analysis$percentage)
```

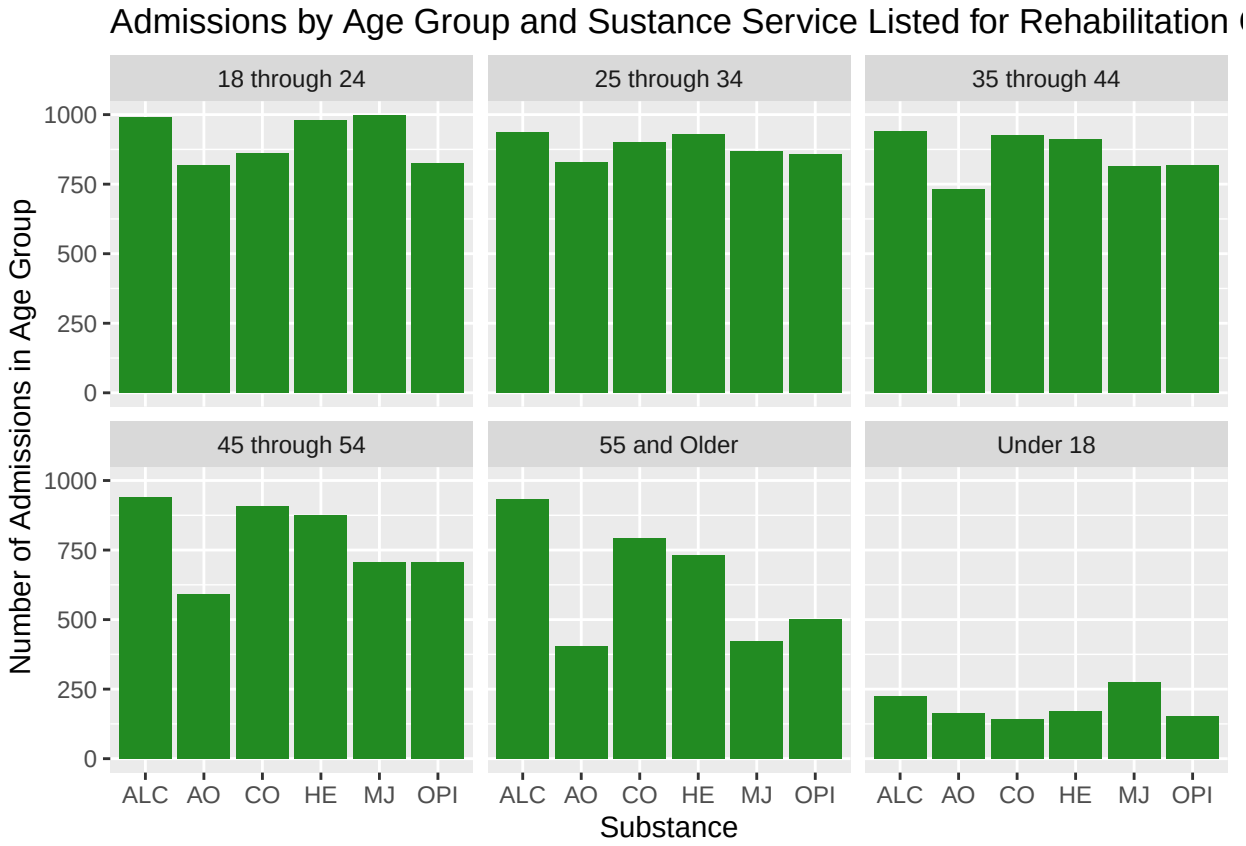
```
## [1] 0.01639344
```

The top counties of admission are Dutchess, Erie, New York, Queens, and Suffolk. Other than New York, all admissions make up less than 5 percent per county. There is over a 1% difference in the overall admissions between New York and the second, Suffolk County. Given that there are 61 counties, and even distribution would result in 1.6% per county. The mean percent of overall admissions per county is 1.1%.

Question 6

```
rehab_centers <- admissions_data %>%
  filter(grepl("ehab",service_type)) %>%
  group_by(age_group,primary_substance_group) %>%
  summarize(count_of_users=n()) %>%
  distinct()

rehab_centers %>%
  ggplot(aes(x=primary_substance_group, y=count_of_users)) +
  geom_bar(stat='identity', fill="forest green")+
  facet_wrap(~age_group) +
  labs(x = "Substance", y = "Number of Admissions in Age Group",
         title = "Admissions by Age Group and Sustance Service Listed for Rehabilitation Centers")
```




```
admissions_by_substance <- admissions_data %>%
  filter(grepl("ehab",service_type)) %>%
  group_by(primary_substance_group) %>%
  summarize(count_of_users=n()) %>%
  summarize(primary_substance_group, count_of_users,
            percent_of_admissions = count_of_users/sum(count_of_users)) %>%
  distinct()
```

For those under 18, marijuana was the most common primary substance. For 18 to 24, both alcohol and marijuana were prominent at nearly 1000 admissions each, marijuana being slightly higher. For the age groups of 25-34, 35-44, 45-54, and 55 plus, alcohol remained as the most common primary substance. Overall, 19.4% of admissions listed alcohol as the primary substance. Both cocaine and heroine were listed individually for over 17% each as the primary substance.