DA5020 – Practicum 1

Rebecca Weiss

10/13/2021

```
#load all necessary libraries
library(RCurl)
library(XML)
library(tidyverse)
library(stringr)
```

Clear the workspace:

```
rm(list = ls())
```

1. Load the data, directly from the URL, into your R environment

```
# get XML from URL and parse
url <- getURL("https://data.ny.gov/api/views/ngbt-9rwf/rows.xml")
admits <- xmlParse(url)

## load XML data into dataframe
df <- xmlToDataFrame(admits, nodes=getNodeSet(admits, "//response/row/row"))
```

```
# Make sure that the data was parsed as expected
head(df)
```

```
##   year county_of_program_location program_category
## 1 2007                      Albany           Crisis
## 2 2007                      Albany           Crisis
## 3 2007                      Albany           Crisis
## 4 2007                      Albany           Crisis
## 5 2007                      Albany           Crisis
## 6 2007                      Albany           Crisis
##                      service_type    age_group primary_substance_group
## 1 Medical Managed Detoxification     Under 18                  Heroin
## 2 Medical Managed Detoxification 55 and Older                 Alcohol
## 3 Medical Managed Detoxification 55 and Older              All Others
## 4 Medical Managed Detoxification 55 and Older                  Heroin
## 5 Medical Managed Detoxification 55 and Older           Other Opioids
## 6 Medical Managed Detoxification   45 thru 54                 Alcohol
##   admissions
```

```
## 1          4
## 2        192
## 3          1
## 4         30
## 5          4
## 6        402
```

2. Evaluate the dataset to determine what data preparation steps are needed and perform them. At a minimum, ensure that you discuss the distribution of the data, outliers and prepare any helpful summary statistics to support your analysis

```
# convert to correct datatypes
df$year <- as.numeric(paste(df$year))
df$admissions <- as.numeric(paste(df$admissions))

str(df)
```

```
## 'data.frame':    92907 obs. of  7 variables:
##  $ year                       : num  2007 2007 2007 2007 2007 ...
##  $ county_of_program_location: Factor w/ 61 levels "Albany","Allegany",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ program_category           : Factor w/ 6 levels "Crisis","Inpatient",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ service_type               : Factor w/ 28 levels "Community Residential",..: 9 9 9 9 9 9 9 9 9 9 .
##  $ age_group                  : Factor w/ 6 levels "18 thru 24","25 thru 34",..: 6 5 5 5 5 4 4 4 4 3
##  $ primary_substance_group    : Factor w/ 7 levels "Alcohol","All Others",..: 4 1 2 4 7 1 2 4 7 1 ...
##  $ admissions                 : num  4 192 1 30 4 402 6 152 14 376 ...
```

```
summary(df)
```

```
##       year      county_of_program_location                 program_category
##  Min.   :2007   New York: 4741             Crisis                   :17272
##  1st Qu.:2010   Suffolk : 3645             Inpatient                :12762
##  Median :2014   Queens  : 3502             Opioid Treatment Program: 3386
##  Mean   :2014   Dutchess: 3355             Outpatient               :32742
##  3rd Qu.:2017   Erie    : 3344             Residential              :26284
##  Max.   :2020   Onondaga: 3265             Specialized              :  461
##                 (Other) :71055
##                          service_type          age_group
##  Outpatient Clinic              :25841   18 thru 24  :18782
##  Inpatient Rehabilitation       :12752   25 thru 34  :20046
##  Community Residential          :10659   35 thru 44  :18664
##  Outpatient Rehabilitation      : 6773   45 thru 54  :16979
##  Intensive Residential          : 6029   55 and Older:13223
##  Medical Managed Detoxification: 5952   Under 18    : 5213
##  (Other)                        :24901
##            primary_substance_group   admissions
##  Alcohol               :18709      Min.   :   1.00
##  All Others            :11825      1st Qu.:   3.00
##  Cocaine incl Crack    :15180      Median :   8.00
##  Heroin                :18117      Mean   :  42.81
##  Marijuana incl Hashish:13299      3rd Qu.:  29.00
```

```
##  None                  :    1      Max.   :2862.00
##  Other Opioids         :15776
```

From the NYS website (https://data.ny.gov/Human-Services/Chemical-Dependence-Treatment-Program-Admissions-B/ngbt-9rwf), we know that this dataset contains information on admissions to certified chemical dependence treatment programs in NY, that can be aggregated by county of program, age group, program category, and primary substance used. Each entry is the total admissions to treatment by admission year, county of program location, program category, service type, age group, and primary substance group. Because these are many options to aggregate by, I will leave *df* as is, and create other variables to group by/ pivot depending on the analyses needed.

From the website and visual inspection, *admissions* and *year* should be numeric, so we begin by making that conversion. Then, to evaluate the data distributions and get a sense of the patterns, `str(df)` and `summary(df)` are run and learn that there are 92,907 observations/rows in the entire dataframe, and 7 variables/columns.

It is noted on the website that "significant others are excluded from this dataset", and while I don't know exactly what variable(s) that refers to, it appears from the output of `summary(df)` that this is the case. In other words, it appears that all the data is aggregated into a category and accounted for, and that there is no missing data. One thing that does stick out to me is that there is only one patient that has "None" in the *primary_substance_group* category. To look at this individual, we filter to see what other data we can gather about them, since it seems odd that they do not have any primary chemical used for admission into a chemical dependence treatment:

```
filter(df, primary_substance_group == "None")
```

```
##   year county_of_program_location program_category          service_type
## 1 2007                   Sullivan      Residential Intensive Residential
##   age_group primary_substance_group admissions
## 1  Under 18                   None          1
```

We learn that this person was admitted to an intensive residential service program in Sullivan in 2007, and they were a minor. Since that information does not appear to drastically alter the dataset, I will leave it in for now, but it is certainly something I would ask the owner of the data about: was this coded as none in error, or is it related to them being a minor and their parents not willingly signing off on collecting that information? Other than this one case, however, it appears all data is as expected and that we will move forward with the analyses.

3. Structure the data relationally, at a minimum, you should have four tibbles or data frames as follows:

• *county* which contains the name of all counties and their respective county code (which is the primary key). When creating the county codes, you can use the data from the NYS County Codes in the Useful Resources section or create your own unique code** for each county. Note: ensure that your data frame does not contain duplicate counties and that it contains all counties in the data.

```
# create county data
# note: data was downloaded as csv, attached here

# load downloaded county codes from NYS open data portal
nys_codes <- read_csv("New_York_State_ZIP_Codes-County_FIPS_Cross-Reference.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   `County Name` = col_character(),
##   `State FIPS` = col_double(),
##   `County Code` = col_character(),
##   `County FIPS` = col_double(),
##   `ZIP Code` = col_double(),
##   `File Date` = col_character()
## )
```

```
# select/name only values we care about, eliminate duplicates
nys_codes <- nys_codes %>%
  select("County Name", "County Code") %>%
  distinct() %>%
  rename("county_code" = "County Code") %>%
  rename("county_name" = "County Name")

# merge with county name in df
county <- left_join(df, nys_codes, by = c("county_of_program_location" = "county_name"))

# name and format as in assignment example, with unique values
county <- as_tibble(county) %>%
  select("county_of_program_location", "county_code") %>%
  unique() %>%
  relocate("county_name" = "county_of_program_location", .after = last_col())

# view to make sure looks correct
head(county)
```

```
## # A tibble: 6 x 2
##   county_code county_name
##   <chr>       <chr>
## 1 001         Albany
## 2 003         Allegany
## 3 005         Bronx
## 4 007         Broome
## 5 009         Cattaraugus
## 6 011         Cayuga
```

• *program_category*: which contains a unique identifier and the name of the program category. Note: ensure that your data frame does not contain duplicates. The program codes can be alphanumeric.

```
# create data frame using list of factor levels + identifier
program_category <- tibble(
  program_code = c("CR", "INPT", "OPT", "OUTPT", "RES", "SPEC"),
  program_category = levels(df$program_category))

# view to make sure looks correct
program_category
```

```
## # A tibble: 6 x 2
##   program_code program_category
```

```
##   <chr>        <chr>
## 1 CR           Crisis
## 2 INPT         Inpatient
## 3 OPT          Opioid Treatment Program
## 4 OUTPT        Outpatient
## 5 RES          Residential
## 6 SPEC         Specialized
```

• *primary_substance_group*: which contains a unique identifier and the name of the substance. Note: ensure that your data frame does not contain duplicates. The substance codes can be alphanumeric.

```
# create data frame using list of factor levels + identifier
primary_substance_group <- tibble(
  substance_code = c("ALC", "AO", "CC", "H", "MH", "NO", "OO"),
  primary_substance_group = levels(df$primary_substance_group))

# view to make sure looks correct
primary_substance_group
```

```
## # A tibble: 7 x 2
##   substance_code primary_substance_group
##   <chr>          <chr>
## 1 ALC            Alcohol
## 2 AO             All Others
## 3 CC             Cocaine incl Crack
## 4 H              Heroin
## 5 MH             Marijuana incl Hashish
## 6 NO             None
## 7 OO             Other Opioids
```

• *admissions_data* which contain the details on the reported number of admissions — excluding the data that resides in the county, program_category and primary_substance_group tibbles/data frames; you should instead include a column with their respective keys

```
# join and mutate other dfs to create admissions_data
admissions_data <- df %>%
  left_join(county, by = c("county_of_program_location" = "county_name")) %>%
  mutate(county_of_program_location = county_code) %>%
  select(-county_code) %>%
  left_join(program_category, by = "program_category") %>%
  mutate(program_category  = program_code) %>%
  select(-program_code) %>%
  left_join(primary_substance_group, by = "primary_substance_group") %>%
  mutate(primary_substance_group = substance_code) %>%
  select(-substance_code)

# view to make sure looks correct
head(admissions_data)
```

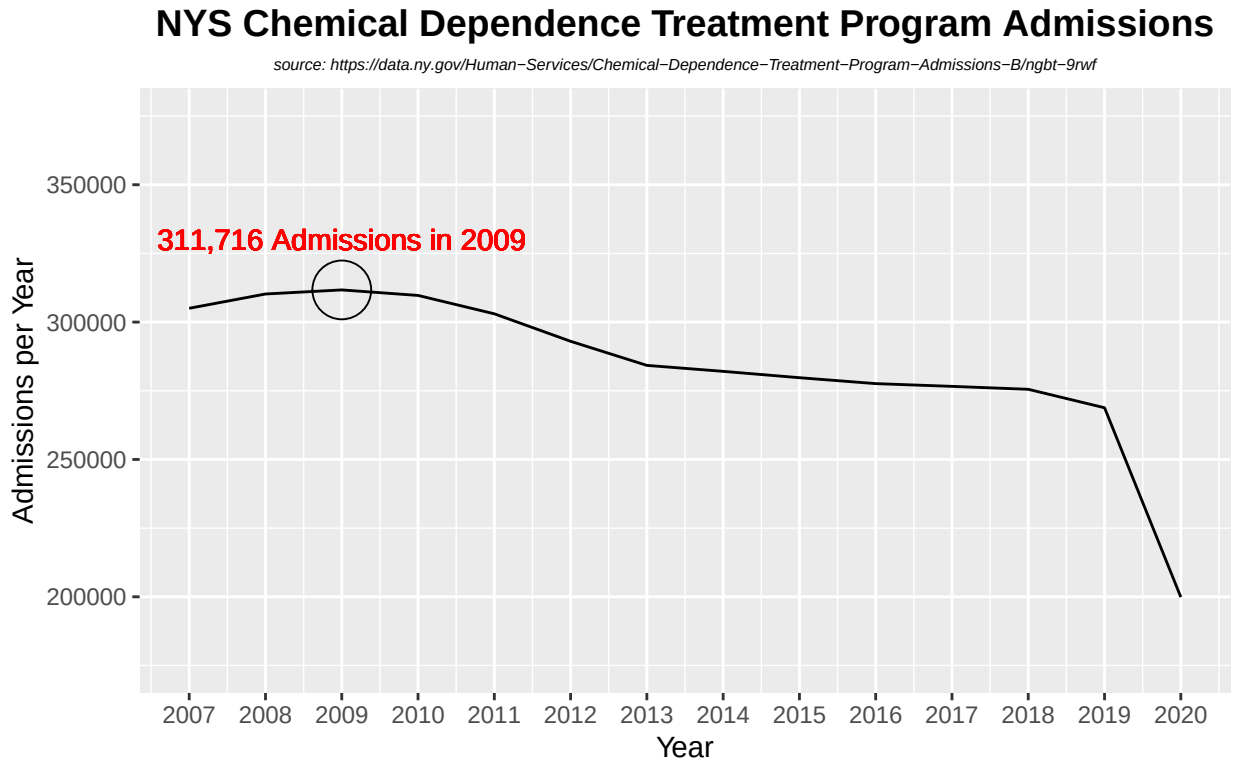
```
##   year county_of_program_location program_category
## 1 2007                         001              CR
## 2 2007                         001              CR
## 3 2007                         001              CR
```

```
## 4 2007                          001               CR
## 5 2007                          001               CR
## 6 2007                          001               CR
##                       service_type    age_group primary_substance_group
## 1 Medical Managed Detoxification     Under 18                         H
## 2 Medical Managed Detoxification 55 and Older                       ALC
## 3 Medical Managed Detoxification 55 and Older                        AO
## 4 Medical Managed Detoxification 55 and Older                         H
## 5 Medical Managed Detoxification 55 and Older                        OO
## 6 Medical Managed Detoxification   45 thru 54                       ALC
##   admissions
## 1          4
## 2        192
## 3          1
## 4         30
## 5          4
## 6        402
```

4. Create a function called annualAdmissions() that derives the total number of reported admissions that transpired each year, for the entire state of NY and displays the results using a line chart. Annotate the chart to show the year with the highest number of admissions. Note: the year should be on the x-axis and the number of admissions on the y-axis. Explain the chart.

```
annualAdmissions <- function(admissions_data)   {
  # create dataframe of total admissions per year
  yearly <- admissions_data %>%
    select(year, admissions) %>%
    group_by(year) %>%
    summarise(admits_per_yr = sum(admissions))
  # return(yearly)

  #create plot
  ggplot(data = yearly, mapping = aes(x = year, y = admits_per_yr)) +
  geom_line() +
  scale_x_continuous(breaks = seq(2005, 2020, by = 1)) +
  ylim(175000, 375000) +
  labs(title = "NYS Chemical Dependence Treatment Program Admissions",
       subtitle = "source: https://data.ny.gov/Human-Services/Chemical-Dependence-Treatment-Program-Admi
       caption = "Annual admissions to chemical dependence treatment programs in New York State
       (NYS), from 2007-2020, showing a steady increase in yearly admissions starting when data
       was collected in 2007, and peaking in 2009 with 311,716 admissions.",
       y = "Admissions per Year",
       x = "Year") +
    theme(plot.subtitle=element_text(size=6, hjust=0.5, face="italic", color="black")) +
    theme(plot.title=element_text(size=14, hjust=0.5, face="bold", color="black")) +
    theme(plot.caption.position = "panel", plot.caption = element_text(size=8)) +
    geom_text(aes(x=2009, y=330000, label="311,716 Admissions in 2009"), size=4, color="red") +
    annotate(geom="point", y=311716, x=2009, size=10, shape=21, fill="transparent")
}
```

```
# make sure it works
annualAdmissions(admissions_data)
```



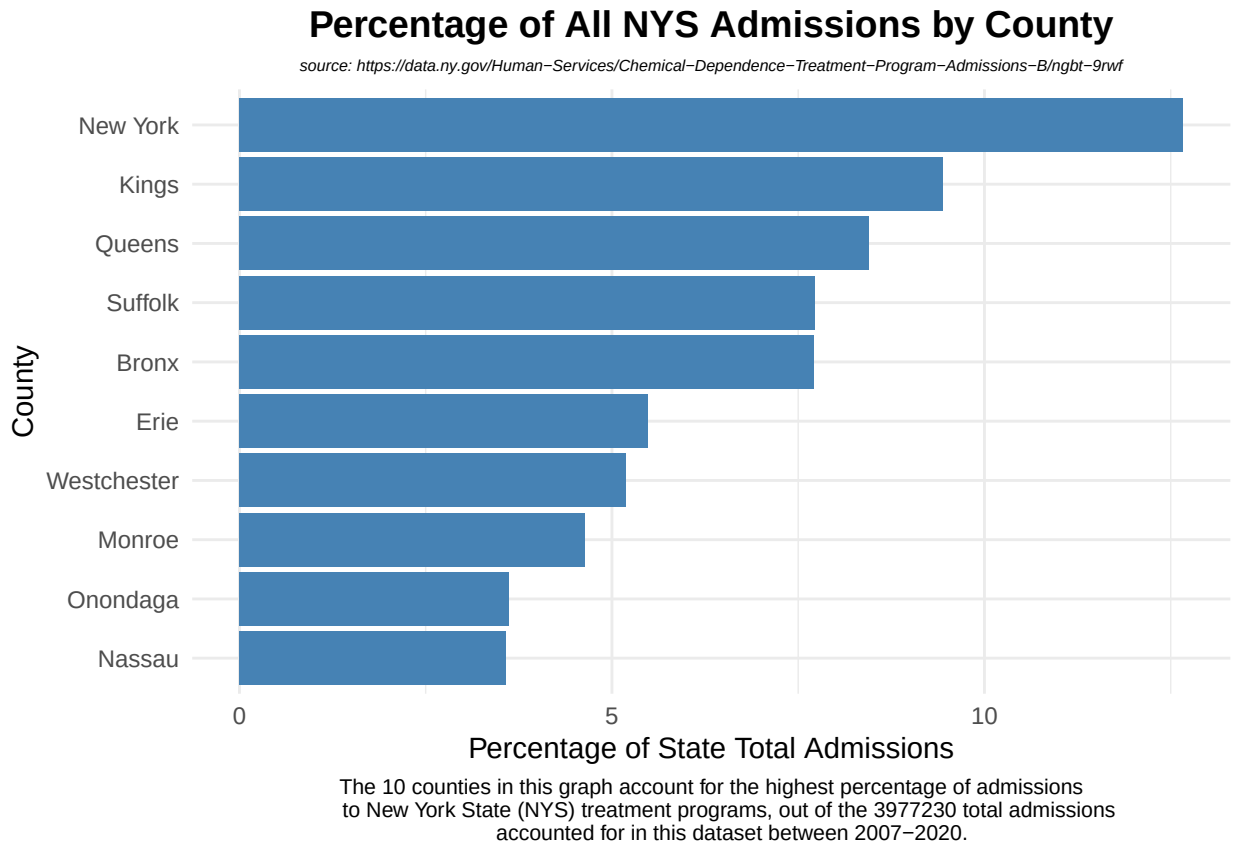
Annual admissions to chemical dependence treatment programs in New York State (NYS), from 2007−2020, showing a steady increase in yearly admissions starting when data was collected in 2007, and peaking in 2009 with 311,716 admissions.

From the output we see that there is a steady increase in yearly admissions starting when data was collected in 2007, peaking in 2009 with 311,716 admissions. After 2009, there are over time more substantial declines in annual admissions, dropping most drastically from 2019-2020. From the graph, we can see that the lowest year of admissions was in 2020 with 199,892 annual admissions.

5. Analyze the percentage of admissions for each county and visualize the results for the top 10 counties using a bar chart. Explain the results. Note: ensure that you join any related dataframes/ tibbles.

```
# establish total admissions:
total_admits = sum(df$admissions)

# create per_county data
per_county <- df %>%
  group_by(county_of_program_location) %>%
  summarise(percent_admits = sum(admissions)/total_admits * 100)

# establish top 10
top10 <- per_county %>%
  arrange(desc(percent_admits)) %>%
  slice_head(n = 10)
```

```r
# create bar chart
ggplot(data = top10, aes(x=reorder(county_of_program_location, percent_admits), y=percent_admits)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  coord_flip() +
  theme_minimal() +
  labs(title = "Percentage of All NYS Admissions by County",
       subtitle = "source: https://data.ny.gov/Human-Services/Chemical-Dependence-Treatment-Program-Admi
       caption = "The 10 counties in this graph account for the highest percentage of admissions
       to New York State (NYS) treatment programs, out of the 3977230 total admissions
       accounted for in this dataset between 2007-2020.",
       y = "Percentage of State Total Admissions",
       x = "County") +
  theme(plot.subtitle=element_text(size=6, hjust=0.5, face="italic", color="black")) +
  theme(plot.caption.position = "panel", plot.caption=element_text(size=8, hjust=0.5, color="black")) +
  theme(plot.title=element_text(size=14, hjust=0.5, face="bold", color="black"))
```



From this graph we see that New York has the most admissions per county in the entire state, making up more than 10% of the total admissions. We can see that once we get past the top 5 highest counties – New York, Kings, Queens, Suffolk, and Bronx – the remaining counties account for ~5% or fewer of the statewide admissions. This could be due to a lot of reasons, from facilities available in a given county, number of referrals made, or just raw population totals. These are all important variables to consider when interpreting this graph, and could be an interesting future direction to include in analyses.

6. Filter the data, using a regular expression, and extract all admissions to the various "Rehab" facilities; i.e. your regex should match all facilities that include the word rehab, rehabilitation, etc. Using the filtered data, identify which substance is the most prominent among each age group. Visualize and explain the results.

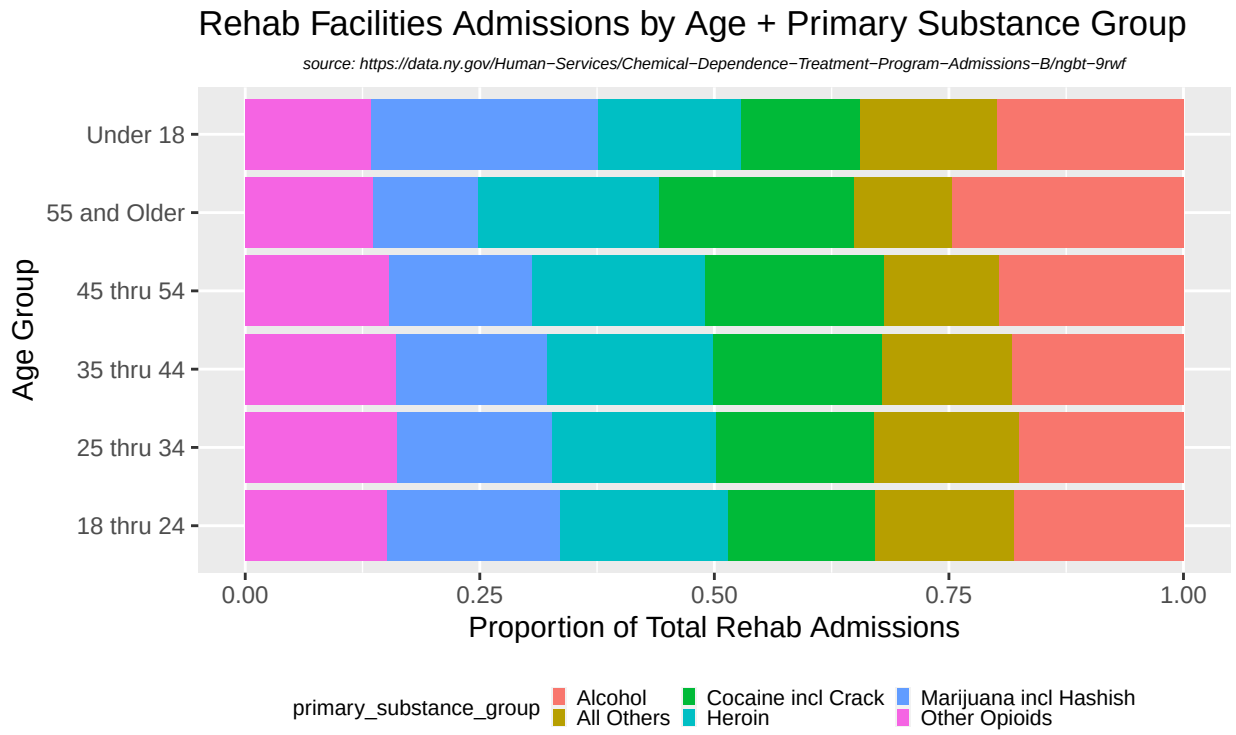
```
# filter for rehab string using regex
rehab <- df %>%
  filter(str_detect(df$service_type, regex("Rehab", ignore_case = TRUE)))

# group_by age group, determine which substance is most prominent
rehab_by_age <- rehab %>%
  group_by(age_group, primary_substance_group) %>%
  summarise(sum=n())
```

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

```
# make sure data looks correct
head(rehab_by_age)
```

```
## # A tibble: 6 x 3
## # Groups:   age_group [1]
##   age_group  primary_substance_group   sum
##   <fct>      <fct>                   <int>
## 1 18 thru 24 Alcohol                   917
## 2 18 thru 24 All Others                750
## 3 18 thru 24 Cocaine incl Crack        798
## 4 18 thru 24 Heroin                    909
## 5 18 thru 24 Marijuana incl Hashish    933
## 6 18 thru 24 Other Opioids             766
```

```
# visualize data
ggplot(rehab_by_age, aes(fill=primary_substance_group, y=sum, x=age_group)) +
    geom_bar(position="fill", stat="identity") +
  labs(title = "Rehab Facilities Admissions by Age + Primary Substance Group",
       subtitle = "source: https://data.ny.gov/Human-Services/Chemical-Dependence-Treatment-Program-Admi
       caption = "The purpose of this graph is to demonstrate the proportion of the primary substance gr
       to visualize what substances account for the highest percentage of Rehab admissions. From this gr
       variations in primary substance groups can be visualized based on the size
       of the stacked bar that represents a given substance.",
       y = "Proportion of Total Rehab Admissions",
       x = "Age Group") +
  theme(legend.position = "bottom") +
  theme(legend.key.height= unit(2, 'mm'),
        legend.key.width= unit(2, 'mm')) +
  theme(plot.subtitle=element_text(size=6, hjust=0.5, face="italic", color="black")) +
  theme(plot.caption.position = "panel", plot.caption=element_text(size=8, hjust=0.5, color="black")) +
  theme(legend.text = element_text(size=8), legend.title = element_text(size=8)) +
  coord_flip()
```



The purpose of this graph is to demonstrate the proportion of the primary substance group within one of the 5 age gro to visualize what substances account for the highest percentage of Rehab admissions. From this graph, the variations in primary substance groups can be visualized based on the size of the stacked bar that represents a given substance.

From running the other analyses above and creating relational data tables, we know that the only variable that is left to contain the term "rehab" would be the *service_type*. Thus, we filter all cases where the string is detected (TRUE), and store in the *rehab* variable. Then we group by age and get a sum of the primary substances in each age group and store it in *rehab_by_age*.

From that variable, we look at the proportion of primary substance group by age group, graphed above. From here, what sticks out most are that Marijuana + Hashish are most prominent in Under 18, and least so in 55 and Older. Also, it appears that Alcohol is the most prominent in the 55 and Older group accounting for ~25% of the rehab admissions for that age group, which is higher proportion than Alcohol rehab admissions in the other age groups.

7. Using the "rehab" data from question 6 above, perform a detailed analysis to identify any patterns or trends with respect to the admission to rehab facilities in certain counties and substance groups. Explain your observations. Note: ensure that you join any related dataframes/tibbles.

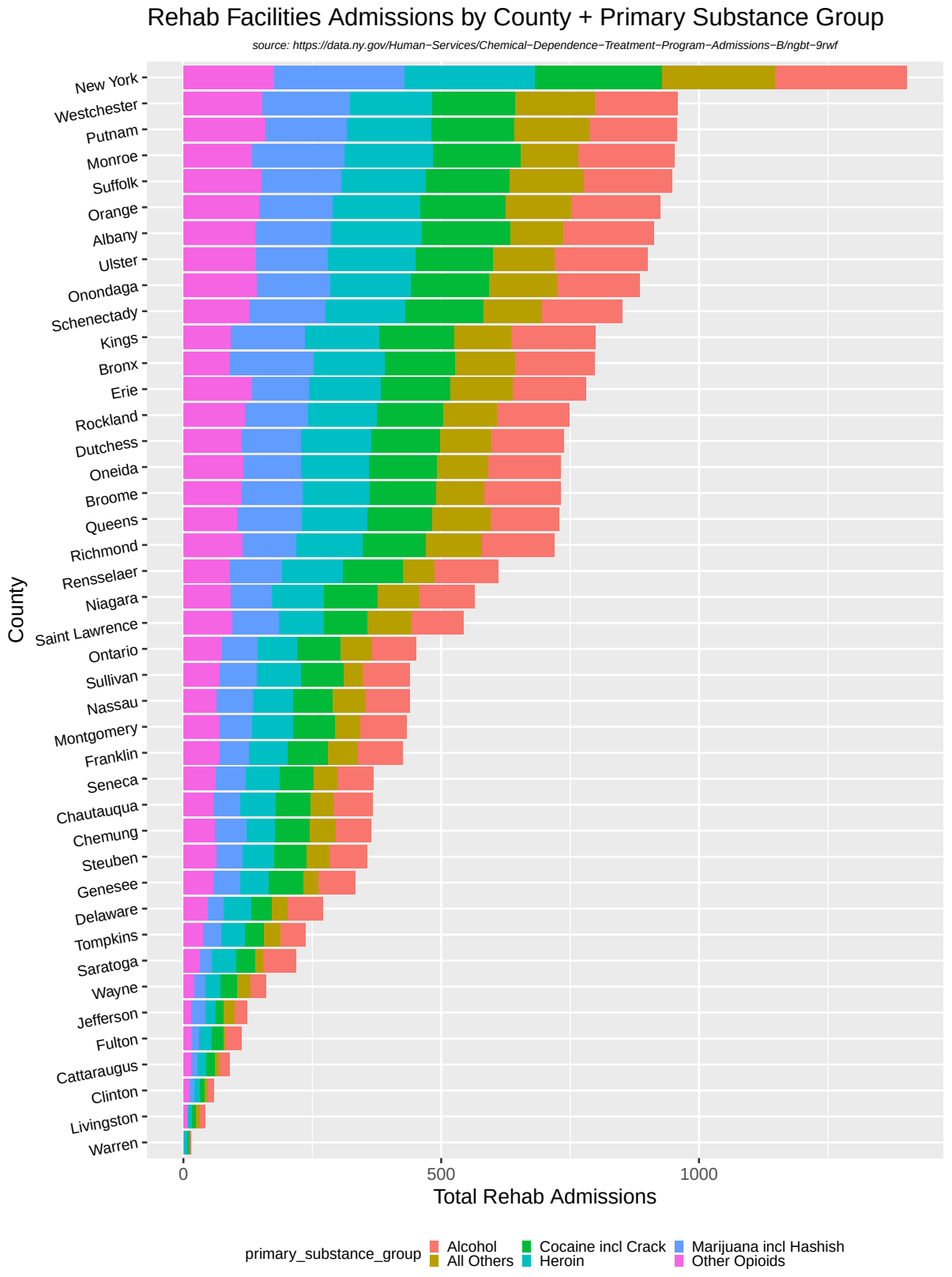
```
rehab_by_county <- rehab %>%
  group_by(county_of_program_location, primary_substance_group) %>%
  summarise(sum=n())
```

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

```
# check to make sure it looks alright
head(rehab_by_county)
```

```
## # A tibble: 6 x 3
## # Groups:   county_of_program_location [1]
##   county_of_program_location primary_substance_group   sum
##   <fct>                      <fct>                   <int>
## 1 Albany                     Alcohol                   176
## 2 Albany                     All Others                103
## 3 Albany                     Cocaine incl Crack        172
## 4 Albany                     Heroin                    176
## 5 Albany                     Marijuana incl Hashish    147
## 6 Albany                     Other Opioids             139
```

```
# visualize data
ggplot(rehab_by_county, aes(fill=primary_substance_group,
                            y=sum, x=reorder(county_of_program_location, sum))) +
    geom_bar(position="stack", stat="identity") +
  labs(title = "Rehab Facilities Admissions by County + Primary Substance Group",
       subtitle = "source: https://data.ny.gov/Human-Services/Chemical-Dependence-Treatment-Program-Admi
       caption = "This graph demonstrates the total admissions to Rehab facilities by county, with the
       stacked bars representing the number of admissions for a primary substance group. By looking at
       the data in this way, the number of total Rehab admissions in a given county can be determined, a
       can distinguish which primary substance groups account for the most of the total Rehab admissions
       y = "Total Rehab Admissions",
       x = "County") +
  theme(legend.position = "bottom") +
  theme(legend.key.height= unit(2, 'mm'),
        legend.key.width= unit(2, 'mm')) +
  theme(legend.text = element_text(size=8), legend.title = element_text(size=8)) +
  theme(plot.caption.position = "panel", plot.caption=element_text(size=8, hjust=0.5, color="black")) +
  theme(axis.text.y = element_text(colour = "black", size = 8, angle = 10)) +
  theme(plot.subtitle=element_text(size=6, hjust=0.5, face="italic", color="black")) +
  coord_flip()
```



This graph demonstrates the total admissions to Rehab facilities by county, with the stacked bars representing the number of admissions for a primary substance group. By looking at the data in this way, the number of total Rehab admissions in a given county can be determined, and visually can distinguish which primary substance groups account for the most of the total Rehab admissions in a given county.

To visualize the primary substance group for rehab admissions by county, I created a bar graph that has each county with the height of their total admissions, and the color stacked representing the primary substance group that was to be treated in the admission. The reason that I looked at total numbers by county instead of percentages as in the previous question is because I think this is useful for answering the question of patterns in Rehab admissions by county and primary substance, as opposed to the question above that just asks which is the most prominent substance per age group (thus making the proportion more appropriate).

Here, we see that New York county has the highest number of total rehab admissions, and that the distribution of primary substance group appears to be pretty even throughout. The next 4 highest counties for total number of Rehab admissions – Westchester, Putnam, Monroe, and Suffolk – all have about the same number of admissions, but differ slightly in their distributions. For example, Monroe appears to have more admissions for Marijuana + Hashish and Crack + Cocaine than the other 3, but fewer rehab admissions that fall into the "All other" primary substance group. Additionally, we see that there is a disparity among counties with how many "Other Opioid" rehab admissions there are. For example, New York has the highest total admissions, and yet from the graph we see counties with much smaller total admissions – like Suffolk, Putnam, Erie, Onondaga, etc. – have about the same number of rehab admissions for other opioids. To look at the numbers on this, we confer with the data:

```
# Highest county total with "other opioids"
head(filter(rehab_by_county, primary_substance_group == 'Other Opioids') %>%
  arrange(desc(sum)), n = 10)
```

```
## # A tibble: 10 x 3
## # Groups:   county_of_program_location [10]
##    county_of_program_location primary_substance_group   sum
##    <fct>                      <fct>                   <int>
##  1 New York                   Other Opioids             176
##  2 Putnam                     Other Opioids             158
##  3 Westchester                Other Opioids             153
##  4 Suffolk                    Other Opioids             151
##  5 Orange                     Other Opioids             147
##  6 Onondaga                   Other Opioids             142
##  7 Ulster                     Other Opioids             140
##  8 Albany                     Other Opioids             139
##  9 Erie                       Other Opioids             132
## 10 Monroe                     Other Opioids             132
```

As we can see, there are differences in county patterns of overall Rehab admissions, and what their primary substance was for the given admission. This graph gives an overview of the total Rehab admissions for each county sectioned by primary substance group, and visualizes these differences within the dataset.