DA5020.P1.Josh.Okon

Hsiao-Yu Peng, Chandani Shrestha, and Josh Okon

2023-10-05

Part 1: Answer the following questions.

1. Create a dataframe with 4 variables. The 4 variables should be doctor_type, doctor_lastname, location, AVG_Rating. The variable doctor_type should contain 4 inputs (PCP, Psychiatrist, Surgeon, Anesthesia). The variable doctor_lastname should contain 4 inputs (Smith, Dame, Jones, Zayas). The variable location should contain 4 inputs (MA, ME, NH, VT). The variable AVG_Rating should contain 4 inputs (7,9,8,9). Print the dataframe and include a screenshot.

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0
## v readr     2.1.4
## -- 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

##    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
```

2. Using the dataframe above... Select row 1 in column 2, what was selected? Select rows 2 through 4, what was selected? Select the last column, what was selected?

Selecting row 1 in column 2 is the doctor last name of "Smith." Selecting rows 2 through 4 provides all the information for the doctors with the last names of "Dame," "Jones," and "Zayas." Selecting the last column provides the average rating for each doctor.

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

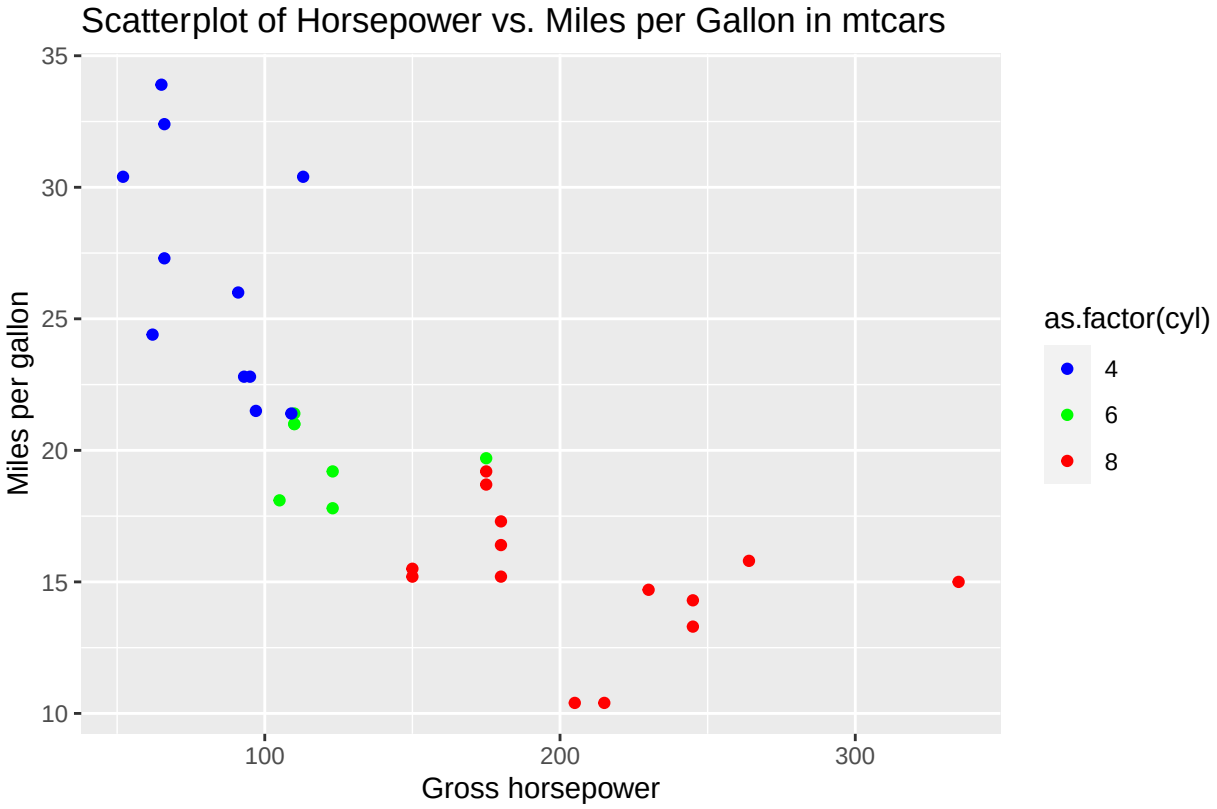
```
##    doctor_type doctor_lastname location AVG_Rating
## 2 Psychiatrist            Dame       ME          9
## 3      Surgeon           Jones       NH          8
## 4   Anesthesia           Zayas       VT          9
```

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

3. Using the dataset Mtcars create a scatterplot showing the relations between any two variables of your choosing. Explain why you picked those variables. Each point should also be based on a color scheme of your choosing.

We chose hp and mpg and colored our scatter plot based on the number of cylinders. We chose colors of our own for the number of cylinders, blue for 4, green for 6, and red for 8. Whenever we think of buying cars, we mainly consider two things - how powerful the car is and what's its mileage. So, we chose these variables to understand how an increase or decrease in horsepower relates to mileage and the number of cylinders.

As per the scatter plot, we can see that there is a negative relationship between horsepower and miles per gallon i.e as the horse power increases, there is a decrease in fuel efficiency (mpg). Likewise, as the number of cylinders increases, there is a decrease in mileage as well.



Analysis of the relationship between horsepower and fuel efficiency

4. Using the dataset MTcars perform a summary statistic on the dataset and calculate the pearson coefficient of the correlation R picking two variables of choice. Explain why you picked those variables and explain the purpose of a pearson coefficient.

We chose mpg and hp because these are important factors in evaluating a car's performance and efficiency. The mpg represents fuel efficiency and hp represents engine power. The pearson coefficient between mpg

and hp is "-0.7761684" which indicates a strong negative correlation. The pearson correlation coefficient is a valuable parameter to understand/measure linear relationships between variables. This coefficient measures strength and direction of the relationship between two continuous variables. The value lies between -1 and 1. -1 means strong negative relation, 1 indicates strong positive relation and 0 represents no relation at all.

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
##        am              gear            carb      
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
##  Median :0.0000   Median :4.000   Median :2.000  
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812  
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

```
## [1] -0.7761684
```

Part 2: Practicum Tasks

The Office of Addiction Services and Support publishes a dataset on reported admissions of people in certified chemical dependence treatment programs throughout New York State (NYS). This dataset includes the number of admissions to certified treatment programs aggregated by the program category, county of the program location, age group of client at admission, and the primary substance of abuse group. For more information on the dataset, visit the following website.

You are given the task of performing a comprehensive analysis of the admission statistics from 2007 to 2019 and summarize your findings with an accompanying narrative that explains your process-flow.

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

```
##   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 18 through 24              All Others
## 3 Medical Managed Detoxification 18 through 24           Other Opioids
## 4 Medical Managed Detoxification 18 through 24                  Heroin
## 5 Medical Managed Detoxification 18 through 24                 Alcohol
## 6 Medical Managed Detoxification 25 through 34              All Others
##   Admissions
## 1          4
## 2          2
## 3          6
## 4        132
## 5         35
## 6          8
```

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.

We used dim(), glimpse(), head(), str(), and summary() to inspect the data. There are 99367 rows and 7 columns. We looked for NA values as well and there were none. While doing summary, we could see some extreme values in the Admissions column, minimum being 1 and maximum being 2861.

```
## [1] 99367     7
```

```
## Rows: 99,367
## Columns: 7
## $ Year                        <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2~
## $ County.of.Program.Location <chr> "Albany", "Albany", "Albany", "Albany", "Al~
## $ Program.Category           <chr> "Crisis", "Crisis", "Crisis", "Crisis", "Cr~
## $ Service.Type               <chr> "Medical Managed Detoxification", "Medical ~
## $ Age.Group                  <chr> "Under 18", "18 through 24", "18 through 24~
## $ Primary.Substance.Group    <chr> "Heroin", "All Others", "Other Opioids", "H~
## $ Admissions                 <int> 4, 2, 6, 132, 35, 8, 1, 11, 276, 135, 11, 1~
```

```
## 'data.frame':    99367 obs. of  7 variables:
##  $ Year                        : int  2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 ...
##  $ County.of.Program.Location: chr  "Albany" "Albany" "Albany" "Albany" ...
##  $ Program.Category          : chr  "Crisis" "Crisis" "Crisis" "Crisis" ...
##  $ Service.Type              : chr  "Medical Managed Detoxification" "Medical Managed Detoxification"
##  $ Age.Group                 : chr  "Under 18" "18 through 24" "18 through 24" "18 through 24" ...
##  $ Primary.Substance.Group   : chr  "Heroin" "All Others" "Other Opioids" "Heroin" ...
##  $ Admissions                : int  4 2 6 132 35 8 1 11 276 135 ...
```

```
##   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 18 through 24              All Others
## 3 Medical Managed Detoxification 18 through 24           Other Opioids
```

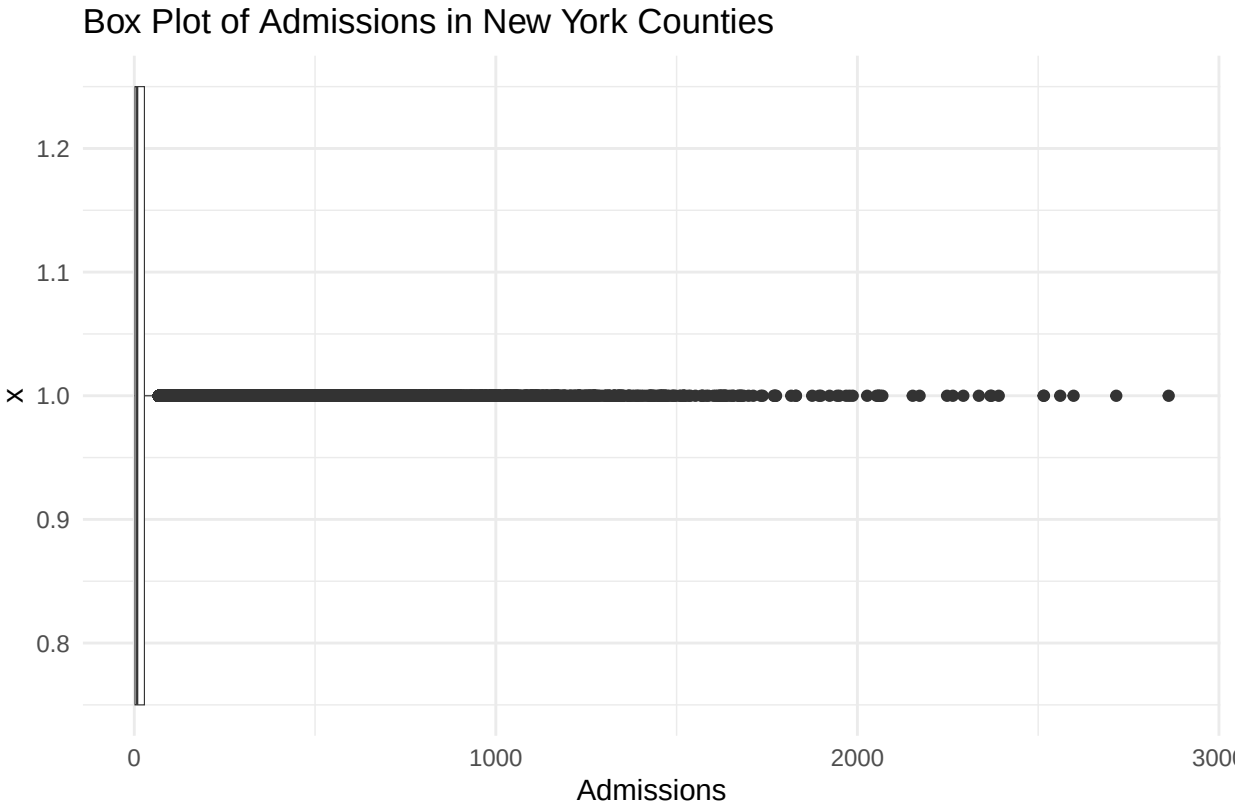
```
## 4 Medical Managed Detoxification 18 through 24                  Heroin
## 5 Medical Managed Detoxification 18 through 24                 Alcohol
## 6 Medical Managed Detoxification 25 through 34              All Others
##   Admissions
## 1          4
## 2          2
## 3          6
## 4        132
## 5         35
## 6          8
```

```
##       Year      County.of.Program.Location Program.Category
##  Min.   :2007   Length:99367               Length:99367
##  1st Qu.:2010   Class :character           Class :character
##  Median :2014   Mode  :character           Mode  :character
##  Mean   :2014
##  3rd Qu.:2018
##  Max.   :2021
##  Service.Type        Age.Group         Primary.Substance.Group
##  Length:99367       Length:99367       Length:99367
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##    Admissions
##  Min.   :   1.00
##  1st Qu.:   2.00
##  Median :   8.00
##  Mean   :  41.91
##  3rd Qu.:  28.00
##  Max.   :2861.00
```

```
## [1] 0
```

We did some data preparation by renaming the column names to a more standard name format. We looked for outliers by using boxplot and we could see many points outside the whiskers of the boxplot. Hence, we use z-score to find and remove those. We considered a z-score > 3 as an outlier and removed the data points. We first created a function called z_score to do all the calculations and used that in the mutate function. After removing outliers, the 99367 observations decreased to 97450 observations i.e 1917 observations were outliers.

The type of the "Year" and "Admissions" were correctly specified as "int" and the rest were of type "char". So, we looked for the variables of type "char" if there were any which can be considered factors. We used the unique() function to get unique observations of each variable. Based on that, we considered County_of_Program_Location, Age_Group, Program_Category and Primary_Substance_Group variables to be converted to factors.



Visualizing Addiction Services Admissions Distribution to look for outliers

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

```
## $Year
##  [1] 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021
##
## $County_of_Program_Location
##  [1] "Albany"         "Allegany"       "Bronx"          "Broome"
##  [5] "Cattaraugus"    "Cayuga"         "Chautauqua"     "Chemung"
##  [9] "Chenango"       "Clinton"        "Columbia"       "Cortland"
## [13] "Delaware"       "Dutchess"       "Erie"           "Essex"
## [17] "Franklin"       "Fulton"         "Genesee"        "Greene"
## [21] "Herkimer"       "Jefferson"      "Kings"          "Lewis"
## [25] "Livingston"     "Madison"        "Monroe"         "Montgomery"
## [29] "Nassau"         "New York"       "Niagara"        "Oneida"
## [33] "Onondaga"       "Ontario"        "Orange"         "Orleans"
## [37] "Oswego"         "Otsego"         "Putnam"         "Queens"
## [41] "Rensselaer"     "Richmond"       "Rockland"       "Saint Lawrence"
## [45] "Saratoga"       "Schenectady"    "Schoharie"      "Schuyler"
## [49] "Seneca"         "Steuben"        "Suffolk"        "Sullivan"
## [53] "Tioga"          "Tompkins"       "Ulster"         "Warren"
## [57] "Washington"     "Wayne"          "Westchester"    "Wyoming"
## [61] "Yates"
##
## $Program_Category
## [1] "Crisis"                  "Inpatient"
```

```
## [3] "Opioid Treatment Program" "Outpatient"
## [5] "Residential"              "Specialized"
##
## $Service_Type
##  [1] "Medical Managed Detoxification"    "Medically Monitored Withdrawal"
##  [3] "Inpatient Rehabilitation"          "Opioid Outpatient Treatment"
##  [5] "Outpatient Clinic"                 "Outpatient Rehabilitation"
##  [7] "Community Residential"             "Intensive Residential"
##  [9] "Residential Rehab for Youth"       "Supportive Living"
## [11] "Specialized Outpatient - TBI"      "Med Sup Withdrawal - Inpatient"
## [13] "Limited Outpatient/KEEP"           "MAOT-A-Residential"
## [15] "Outpat Chem Depend for Youth"      "Short Term Res CD/Youth"
## [17] "Long Term Res CD/Youth"            "Med Sup Withdrawal-Outpatient"
## [19] "Non-Med Sup Chem Depend OP"        "OTP Intensive Residential"
## [21] "Specialized Services Outpt Rehab"  "Specialized Outpatient - Mobile"
## [23] "Stabilization Rehab Reintegration" "Residential Stabilization"
## [25] "Rehab and Reintegration"           "Residential Reintegration"
## [27] "Stabilization and Rehab"           "Residential Rehabilitation"
##
## $Age_Group
## [1] "Under 18"      "18 through 24" "25 through 34" "35 through 44"
## [5] "45 through 54" "55 and Older"
##
## $Primary_Substance_Group
## [1] "Heroin"        "All Others"    "Other Opioids" "Alcohol"
## [5] "Cocaine"       "Marijuana"     "None"
##
## $Admissions
##    [1]    4    2    6  132   35    8    1   11  276  135   15  196  377   14
##   [15]  152  401   30  192    9   25   40   22  102   27  592   26   37    5
##   [29]  661  150   33   16   88   10   79    3  129   21  161   67   13  144
##   [43]   12   57  248   49  284  185  137  246   23   52  376  222  115   18
##   [57]  357  118   17   29  296    7   39   19   20   31   38   48   76  106
##   [71]  172   54  133   89   72   47  744  583 1060 1572   93  586 1241  439
##   [85]   69   65   78  189   36   42  116   46   58  141  138  298   91   75
##   [99]  283   81  104  151  112   44  488 1159  890  256   51  354 1115  209
##  [113]  378 1458  331  557  998  575  681  787  656  164  548  627  153  123
##  [127]  258   45   34   59   84   32   55  174  127  110  181   90  146  113
##  [141]   63   24   43   87  128  119   28  160   99  171   74   68   86   64
##  [155]   70  156   50  183  214  163   60   85   53   66   83  107  111   41
##  [169]   77  147  100  227   61   56  101  114  184  231  522  142  402  158
##  [183]  139   94  254  175  240  237  109  187  108  334   80  121  255  318
##  [197]  136  282   73  514  143  805  195  738  371  613  125 1082  695  247
##  [211] 1340  451  122 1130   71  319   62  103  221  405  470  604  262  311
##  [225]  580   92  126  597  394 1028 1173  653 1059  130   98  304  199  347
##  [239]  773  654  194 1164  238 1329  648  918  589  635  963  649  469  887
##  [253]  389  117  140  178  208   82  134  309  219  251  154  588  521  167
##  [267]  482  249  368  431  834  509  236  173  678  700  274 1143  443  893
##  [281]  253   95   96  170  426  658  869  271  324 1036  459 1052  215  888
##  [295]  362  746  854 1341 2861  737 2515  193  960  567  428 2516 1899  516
##  [309]  278  212  197  149  483  211  671 1582 1519 1137  379 1270  265 1128
##  [323] 1490  595  624 1633 1207  616 1524  244  614  198  453  267  168  105
##  [337]  159  273  120  224   97  235  361  349  182  292  513  337  561  179
##  [351]  406  131  148  162  216  145  165  266  232  422  535 1588  327 1285
```

```
##  [365]  313  252  506  495  702 1313  312  180  380  287  229  346  502  864
##  [379]  294  719  716  536  858  338  701  643  290  260  186  306  397  507
##  [393] 1468  302 1278  403  280  204  206  223  245  233  176  269  200  321
##  [407]  124  226  203  188  220  314  307  718  706  293  743  749  425  433
##  [421]  155 1273  713 1467  333 1262  177  326  395 1195  275  970  356  308
##  [435]  479  484  351  537  166  442  387  201  646  270  329  396  359  677
##  [449]  877 1712  539 1479  570  391  340  492 1140 1049 1455  404 1620  545
##  [463]  913  712  808  765  692  774  191  423  449  335  332  917  239  753
##  [477]  218  606 1368  419 1264  420  491  289  416  809  400 1047 1076  742
##  [491] 1187  410  241  343  386  735  729 1317  268 1625  741  897 1092  781
##  [505]  261  621 1011  291  468  797  473  764 1095  432  875  234  169  413
##  [519]  909  429  973  901  344  205  544  807 2371  581 2391  882  478 2053
##  [533] 2058  541  418  709 1462 1305 1224  434  538 1251 1343  630  675 1738
##  [547] 1229  721 1630  277  651  217  279  242  213  288  350  543  303  472
##  [561]  612  445 1424  560  766 1132 1109  341  843  708  752  832  690  610
##  [575]  228  441  210  264  358  384 1191  202  272  421  373  257  481  659
##  [589]  310  770  874 1474  618 1461 1274  374 1110 1158  450  519  529  533
##  [603]  369  365  411  190  300  323  353  957 1830 1950  756  364  549 1032
##  [617] 1020 1389  243 1634  355  552  728  665  688  724  259  838  263  157
##  [631]  301  286  984  772 1165 1312 1330  352  512  975  305 1217  375  525
##  [645]  693 1057  556 1460  461  348  471  768 1392  336 1058  320  465  848
##  [659]  822  866 1100  315  928  547  771 1006 2172  751 2598  945  562 1894
##  [673]  317 2336  714 1355 1402  579 1346 1433  299 1067  824  652 1509 1230
##  [687]  370  686 1655  641  250  297  457  367  345  316  383 1511  527 1733
##  [701]  360  230  639  894  462  994  800  733  414  851  759  593 1408  207
##  [715]  225  372  836  857  438 1351 1336  466  381 1000 1171  437  467  385
##  [729]  474  691  696 1685  531 2063  763 1293 1817  622  668  755  806  510
##  [743] 1042  760 1181 1163  903 1326  409  620 1139  524 1670  330  322  487
##  [757]  424 1875  940  845 1106 1209  480 1112  850  415  951  873  784 1053
##  [771]  916  889  720 1923 2561 1026  530 1328 2248  629  626 1086  605 1425
##  [785] 1672 1170  971  657 1533 1260 1774  412  540  497  325  684 1365  642
##  [799] 1648  454  964  669  679  554 1304  550  342  699  447  794  363  633
##  [813]  878  565  476 1446 1268 1350 1010 1348  456  578  440  339  295  392
##  [827]  715 2069  885 1064 1623  573  534  685  655  835  444  281  475  863
##  [841]  408  637 1019 1002 1084  399  867  571 1469  448  566 1657 1677  921
##  [855]  758  993  723 1236 1102  847  802  842  997  736  956 1602 2716 1088
##  [869] 2153  631 1146 1427 1121  705 1377 1089  623 1680  390 1258 1944  489
##  [883]  591  460  856  844  813  732  559  666  518  828  582 1248 1206  931
##  [897]  366 1699  600  436 1231  452  748 1029 1608  564  634  839  417  427
##  [911]  568  922  996  938  558  464  430  503  880 1613  511  846  949  574
##  [925] 1118  819  837  703  961  667  640 1431  625 2293 1114 1024 1383  982
##  [939] 1219  673  988 1253  967 1498  704 1978  455  707  555  985  490  398
##  [953]  285  520  803  532  868 1099 1215  791  435  382 1416  486  676 1040
##  [967] 1520  553  609  870  388  825  939  841  801  731 1770  576  981  328
##  [981] 1370  900  599 1151  546 1113  821  804 1098  761 2027  458 1079  710
##  [995]  972 1153 1033  498 1188  884 1454  463 1987  628  879  777  775 1068
## [1009]  569 1396  779  517  829  860 1439  590  694  647  670  776 1775  508
## [1023]  477  680 1265  898  906 1101  955  826 1168  855 2264 1277 1001 1284
## [1037]  711  861 1177  816  587 1538  762  899  528  876 2059  636  496  526
## [1051]  932  603  795  585 1232 1233  500  523  407  823  607  674  494  683
## [1065]  672  747 1107 1769  786  790 1030  992  942  485  812 1568 1016  923
## [1079] 1169 1048  722 1301  504  615  987 1969  697  393 1069 1152 1136  895
## [1093]  572  611  551  853  783  871 1450  792  914  933  833  689 1306  726
## [1107] 1202  645 1223 1123  660  757  664  632  904 1311 1244  937  584 1345
```

```
## [1121]  948  725 1055  499 1012 1627  515  865  493  814  780  817  446  793
## [1135] 1157  687  952 1096  881 1041 1552  976 1022 1436 1044  872 1009 1003
## [1149] 2368  908  778  896  727  818  969 1081  924  977 1430 1256 1039 1493
## [1163] 1339  602  739  840  950  638  577  983  608 1488  769 1116  959 1639
## [1177] 1618  619  730  911 1369  927  617  682  501  601  505  989  594  943
## [1191] 1025 1314
```

```
##   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 18 through 24              All Others
## 3 Medical Managed Detoxification 18 through 24           Other Opioids
## 4 Medical Managed Detoxification 18 through 24                  Heroin
## 5 Medical Managed Detoxification 18 through 24                 Alcohol
## 6 Medical Managed Detoxification 25 through 34              All Others
##   Admissions
## 1          4
## 2          2
## 3          6
## 4        132
## 5         35
## 6          8
```

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). Note: ensure that your data frame does not contain duplicate counties and ensure that your dataframe contains all counties in the data. 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 codes can be numeric (e.g. auto incremented). 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 codes can be numeric (e.g. auto incremented). 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 foreign keys. The names should be substituted with their respective foreign keys.

```
## # A tibble: 5 x 2
##   county_code county_name
##   <chr>       <chr>
## 1 NY          Queens
## 2 NY          Bronx
## 3 NY          New York
## 4 NY          Richmond
## 5 NY          Kings
```

```
## [1] 0
```

```
## # A tibble: 62 x 2
##    county_code county_name
##    <chr>       <chr>
##  1 AL          Albany
##  2 CA          Cattaraugus
##  3 CN          Chenango
##  4 DE          Delaware
##  5 FR          Franklin
##  6 HA          Hamilton
##  7 LE          Lewis
##  8 MG          Montgomery
##  9 ON          Oneida
## 10 OL          Orleans
## # i 52 more rows
```

```
## # A tibble: 6 x 2
##   program_code program_category_name
##   <chr>        <chr>
## 1 CR           Crisis
## 2 IN           Inpatient
## 3 OTP          Opiod Treatment Program
## 4 RES          Residential
## 5 OUT          Outpatient
## 6 SP           Specialized
```

```
## [1] "Heroin"        "All Others"    "Other Opioids" "Alcohol"
## [5] "Cocaine"       "Marijuana"     "None"
```

```
## # A tibble: 7 x 2
##   substance_code primary_substance_group_name
##   <chr>          <chr>
## 1 H              Heroin
## 2 AO             All Others
## 3 OO             Other Opioids
## 4 A              Alcohol
## 5 C              Cocaine
## 6 M              Marijuana
## 7 N              None
```

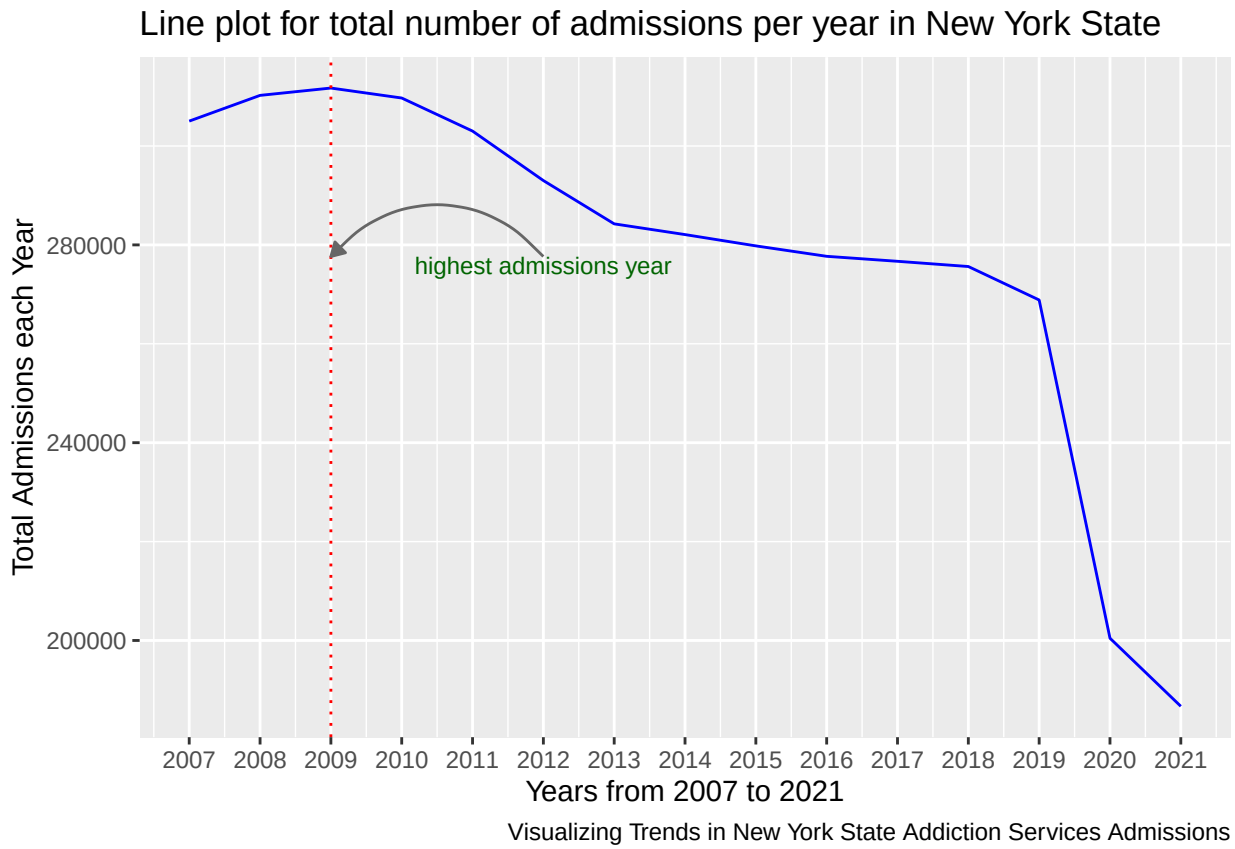
```
##   Year County_of_Program_Location Program_Category
## 1 2007                         AL               CR
## 2 2007                         AL               CR
## 3 2007                         AL               CR
## 4 2007                         AL               CR
## 5 2007                         AL               CR
## 6 2007                         AL               CR
##                      Service_Type     Age_Group Primary_Substance_Group
## 1 Medical Managed Detoxification      Under 18                        H
## 2 Medical Managed Detoxification 18 through 24                       AO
## 3 Medical Managed Detoxification 18 through 24                       OO
## 4 Medical Managed Detoxification 18 through 24                        H
## 5 Medical Managed Detoxification 18 through 24                        A
## 6 Medical Managed Detoxification 25 through 34                       AO
```

```
##   Admissions
## 1          4
## 2          2
## 3          6
## 4        132
## 5         35
## 6          8
```

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.

We created a function called annualAdmissions() to return a plot of total admissions per year. First, we grouped our dataframe by 'Year' and calculated total number of admissions using the summary() function which we put inside 'total_admissions'. We only kept unique rows by using the distinct() function. We selected only 'Year' and 'total_admissions' and arranged 'total_admissions' in descending order. We put this into a new data frame called NY_admissions and then we extracted the first 'Year' to use it for annotation in the line plot.

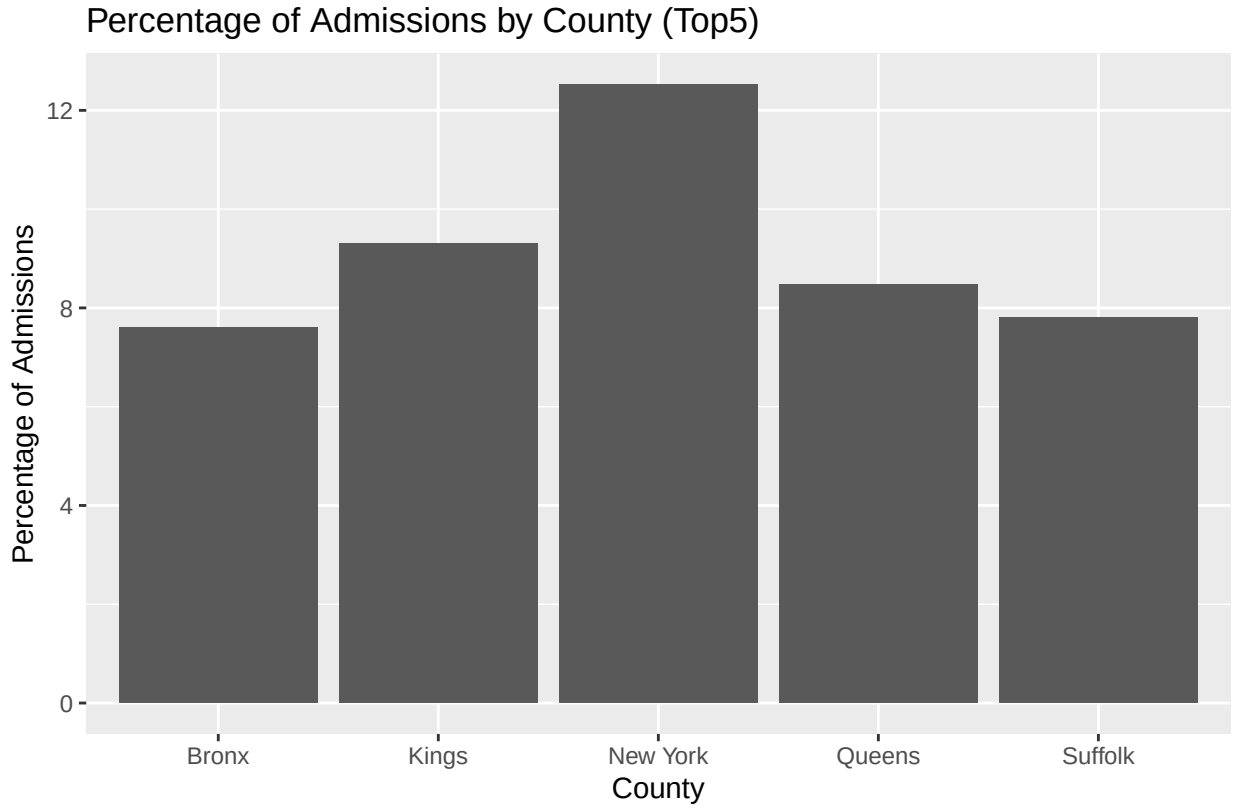
We put year in x-asis and number of admissions on y-axis. We use geom_vline() to place the reference line for the year with the highest number of admissions. We used annotate() to include both text and an arrow to show our year with the maximum admissions on the line plot.



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

New York County has the highest admission percentage. According to the admission percentages, from high to low, the next four counties are Kings, Queens, Suffolk, and the Bronx.

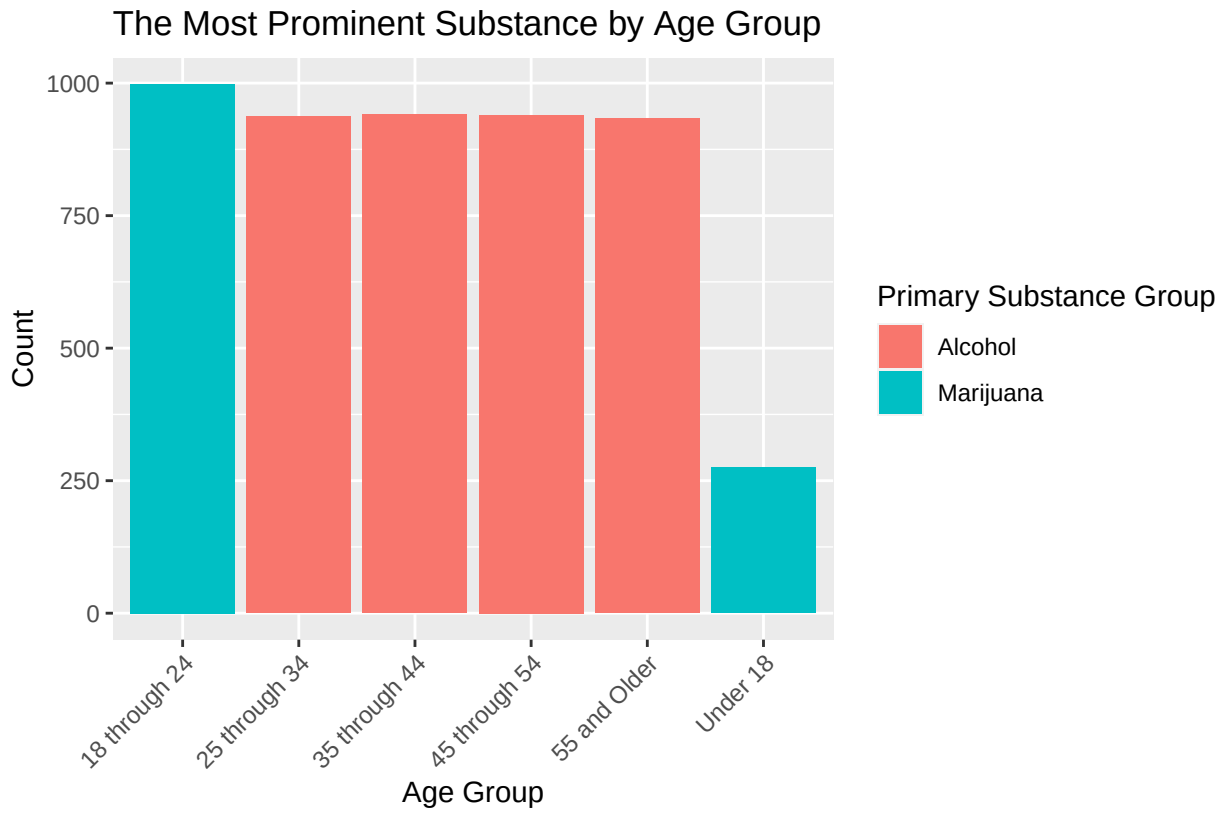
```
## # A tibble: 5 x 2
##   County_of_Program_Location Percentage
##   <chr>                           <dbl>
## 1 New York                        12.5
## 2 Kings                            9.31
## 3 Queens                           8.48
## 4 Suffolk                          7.81
## 5 Bronx                            7.60
```



The Top 5 Counties in New York State with the Highest Addiction Services Admission Rates

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.

Marijuana is the most prominent substance in the "18 through 24" and "Under 18" age groups. In the other four groups, "25 through 34," "35 through 44," "45 through 54," and "55 and Older," alcohol is the most prominent substance in these groups.



most prominent substance among each age group requiring addiction services