STAT 410 Lab Mod 2

Kristin Duncan

2/3/2020

Some exercises from "Introduction to Data Science" by Irizarry

1

Load the dplyr package and the dslabs package. Access the murders dataset from dslabs. You'll see it contains variables state, abb, region, populations, and total.

- Use the function mutate to add a murders column named rate with the per 100,000 murder rate. Use assignment so that this new column remains accessible in the murders data frame.
- rank(x) gives you the ranks of x from lowest to highest, and rank(-x) gives you the ranks from highest to lowest. Use the function mutate to add a column rank containing the rank from highest to lowest murder rate. Use assignment so that this new column remains accessible in the murders data frame.

```
library(dplyr)
```

```
##
## 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
```

```
library(dslabs)
library(ggrepel)
```

```
## Loading required package: ggplot2
```

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:dplyr':
##
##     nasa
```

```
murders
```

```
##                   state abb        region population total
## 1               Alabama  AL         South    4779736   135
```

```
## 2                Alaska  AK          West     710231    19
## 3               Arizona  AZ          West    6392017   232
## 4              Arkansas  AR         South    2915918    93
## 5            California  CA          West   37253956  1257
## 6              Colorado  CO          West    5029196    65
## 7           Connecticut  CT     Northeast    3574097    97
## 8              Delaware  DE         South     897934    38
## 9  District of Columbia  DC         South     601723    99
## 10              Florida  FL         South   19687653   669
## 11              Georgia  GA         South    9920000   376
## 12               Hawaii  HI          West    1360301     7
## 13                Idaho  ID          West    1567582    12
## 14             Illinois  IL North Central   12830632   364
## 15              Indiana  IN North Central    6483802   142
## 16                 Iowa  IA North Central    3046355    21
## 17               Kansas  KS North Central    2853118    63
## 18             Kentucky  KY         South    4339367   116
## 19            Louisiana  LA         South    4533372   351
## 20                Maine  ME     Northeast    1328361    11
## 21             Maryland  MD         South    5773552   293
## 22        Massachusetts  MA     Northeast    6547629   118
## 23             Michigan  MI North Central    9883640   413
## 24            Minnesota  MN North Central    5303925    53
## 25          Mississippi  MS         South    2967297   120
## 26             Missouri  MO North Central    5988927   321
## 27              Montana  MT          West     989415    12
## 28             Nebraska  NE North Central    1826341    32
## 29               Nevada  NV          West    2700551    84
## 30        New Hampshire  NH     Northeast    1316470     5
## 31           New Jersey  NJ     Northeast    8791894   246
## 32           New Mexico  NM          West    2059179    67
## 33             New York  NY     Northeast   19378102   517
## 34       North Carolina  NC         South    9535483   286
## 35         North Dakota  ND North Central     672591     4
## 36                 Ohio  OH North Central   11536504   310
## 37             Oklahoma  OK         South    3751351   111
## 38               Oregon  OR          West    3831074    36
## 39         Pennsylvania  PA     Northeast   12702379   457
## 40         Rhode Island  RI     Northeast    1052567    16
## 41       South Carolina  SC         South    4625364   207
## 42         South Dakota  SD North Central     814180     8
## 43            Tennessee  TN         South    6346105   219
## 44                Texas  TX         South   25145561   805
## 45                 Utah  UT          West    2763885    22
## 46              Vermont  VT     Northeast     625741     2
## 47             Virginia  VA         South    8001024   250
## 48           Washington  WA          West    6724540    93
## 49        West Virginia  WV         South    1852994    27
## 50            Wisconsin  WI North Central    5686986    97
## 51              Wyoming  WY          West     563626     5
```

```
murders1 = mutate(murders, rate = total/(population/(100000)))
murders1
```

```
##                  state abb        region population total       rate
```

```
## 1               Alabama  AL         South    4779736   135  2.8244238
## 2                Alaska  AK          West     710231    19  2.6751860
## 3               Arizona  AZ          West    6392017   232  3.6295273
## 4              Arkansas  AR         South    2915918    93  3.1893901
## 5            California  CA          West   37253956  1257  3.3741383
## 6              Colorado  CO          West    5029196    65  1.2924531
## 7           Connecticut  CT     Northeast    3574097    97  2.7139722
## 8              Delaware  DE         South     897934    38  4.2319369
## 9  District of Columbia  DC         South     601723    99 16.4527532
## 10              Florida  FL         South   19687653   669  3.3980688
## 11              Georgia  GA         South    9920000   376  3.7903226
## 12               Hawaii  HI          West    1360301     7  0.5145920
## 13                Idaho  ID          West    1567582    12  0.7655102
## 14             Illinois  IL North Central   12830632   364  2.8369608
## 15              Indiana  IN North Central    6483802   142  2.1900730
## 16                 Iowa  IA North Central    3046355    21  0.6893484
## 17               Kansas  KS North Central    2853118    63  2.2081106
## 18             Kentucky  KY         South    4339367   116  2.6732010
## 19            Louisiana  LA         South    4533372   351  7.7425810
## 20                Maine  ME     Northeast    1328361    11  0.8280881
## 21             Maryland  MD         South    5773552   293  5.0748655
## 22        Massachusetts  MA     Northeast    6547629   118  1.8021791
## 23             Michigan  MI North Central    9883640   413  4.1786225
## 24            Minnesota  MN North Central    5303925    53  0.9992600
## 25          Mississippi  MS         South    2967297   120  4.0440846
## 26             Missouri  MO North Central    5988927   321  5.3598917
## 27              Montana  MT          West     989415    12  1.2128379
## 28             Nebraska  NE North Central    1826341    32  1.7521372
## 29               Nevada  NV          West    2700551    84  3.1104763
## 30        New Hampshire  NH     Northeast    1316470     5  0.3798036
## 31           New Jersey  NJ     Northeast    8791894   246  2.7980319
## 32           New Mexico  NM          West    2059179    67  3.2537239
## 33             New York  NY     Northeast   19378102   517  2.6679599
## 34       North Carolina  NC         South    9535483   286  2.9993237
## 35         North Dakota  ND North Central     672591     4  0.5947151
## 36                 Ohio  OH North Central   11536504   310  2.6871225
## 37             Oklahoma  OK         South    3751351   111  2.9589340
## 38               Oregon  OR          West    3831074    36  0.9396843
## 39         Pennsylvania  PA     Northeast   12702379   457  3.5977513
## 40         Rhode Island  RI     Northeast    1052567    16  1.5200933
## 41       South Carolina  SC         South    4625364   207  4.4753235
## 42         South Dakota  SD North Central     814180     8  0.9825837
## 43            Tennessee  TN         South    6346105   219  3.4509357
## 44                Texas  TX         South   25145561   805  3.2013603
## 45                 Utah  UT          West    2763885    22  0.7959810
## 46              Vermont  VT     Northeast     625741     2  0.3196211
## 47             Virginia  VA         South    8001024   250  3.1246001
## 48           Washington  WA          West    6724540    93  1.3829942
## 49        West Virginia  WV         South    1852994    27  1.4571013
## 50            Wisconsin  WI North Central    5686986    97  1.7056487
## 51              Wyoming  WY          West     563626     5  0.8871131
```

```
murders2 = mutate(murders1, rank = rank(-rate))
murders2
```

```
##                   state abb        region population total       rate rank
## 1               Alabama  AL         South    4779736   135  2.8244238   23
## 2                Alaska  AK          West     710231    19  2.6751860   27
## 3               Arizona  AZ          West    6392017   232  3.6295273   10
## 4              Arkansas  AR         South    2915918    93  3.1893901   17
## 5            California  CA          West   37253956  1257  3.3741383   14
## 6              Colorado  CO          West    5029196    65  1.2924531   38
## 7           Connecticut  CT     Northeast    3574097    97  2.7139722   25
## 8              Delaware  DE         South     897934    38  4.2319369    6
## 9  District of Columbia  DC         South     601723    99 16.4527532    1
## 10              Florida  FL         South   19687653   669  3.3980688   13
## 11              Georgia  GA         South    9920000   376  3.7903226    9
## 12               Hawaii  HI          West    1360301     7  0.5145920   49
## 13                Idaho  ID          West    1567582    12  0.7655102   46
## 14             Illinois  IL North Central   12830632   364  2.8369608   22
## 15              Indiana  IN North Central    6483802   142  2.1900730   31
## 16                 Iowa  IA North Central    3046355    21  0.6893484   47
## 17               Kansas  KS North Central    2853118    63  2.2081106   30
## 18             Kentucky  KY         South    4339367   116  2.6732010   28
## 19            Louisiana  LA         South    4533372   351  7.7425810    2
## 20                Maine  ME     Northeast    1328361    11  0.8280881   44
## 21             Maryland  MD         South    5773552   293  5.0748655    4
## 22        Massachusetts  MA     Northeast    6547629   118  1.8021791   32
## 23             Michigan  MI North Central    9883640   413  4.1786225    7
## 24            Minnesota  MN North Central    5303925    53  0.9992600   40
## 25          Mississippi  MS         South    2967297   120  4.0440846    8
## 26             Missouri  MO North Central    5988927   321  5.3598917    3
## 27              Montana  MT          West     989415    12  1.2128379   39
## 28             Nebraska  NE North Central    1826341    32  1.7521372   33
## 29               Nevada  NV          West    2700551    84  3.1104763   19
## 30        New Hampshire  NH     Northeast    1316470     5  0.3798036   50
## 31           New Jersey  NJ     Northeast    8791894   246  2.7980319   24
## 32           New Mexico  NM          West    2059179    67  3.2537239   15
## 33             New York  NY     Northeast   19378102   517  2.6679599   29
## 34       North Carolina  NC         South    9535483   286  2.9993237   20
## 35         North Dakota  ND North Central     672591     4  0.5947151   48
## 36                 Ohio  OH North Central   11536504   310  2.6871225   26
## 37             Oklahoma  OK         South    3751351   111  2.9589340   21
## 38               Oregon  OR          West    3831074    36  0.9396843   42
## 39         Pennsylvania  PA     Northeast   12702379   457  3.5977513   11
## 40         Rhode Island  RI     Northeast    1052567    16  1.5200933   35
## 41       South Carolina  SC         South    4625364   207  4.4753235    5
## 42         South Dakota  SD North Central     814180     8  0.9825837   41
## 43            Tennessee  TN         South    6346105   219  3.4509357   12
## 44                Texas  TX         South   25145561   805  3.2013603   16
## 45                 Utah  UT          West    2763885    22  0.7959810   45
## 46              Vermont  VT     Northeast     625741     2  0.3196211   51
## 47             Virginia  VA         South    8001024   250  3.1246001   18
## 48           Washington  WA          West    6724540    93  1.3829942   37
## 49        West Virginia  WV         South    1852994    27  1.4571013   36
## 50            Wisconsin  WI North Central    5686986    97  1.7056487   34
## 51              Wyoming  WY          West     563626     5  0.8871131   43
```

2

Use filter to show the top 5 states with the highest murder rates. Use select so that only state, rate, and rank are printed.

```
murderstop5 = murders2 %>%
  filter(rank <= 5) %>%
  select(state, rate, rank)
murderstop5
```

```
##                  state      rate rank
## 1 District of Columbia 16.452753    1
## 2            Louisiana  7.742581    2
## 3             Maryland  5.074866    4
## 4             Missouri  5.359892    3
## 5       South Carolina  4.475323    5
```

3

We can use %in% to filter with dplyr.

- Create a new data frame called murders_nw with only the states from the Northeast and the West.

- How many state are in this category? (use code to answer)
- Use filter and select to create a data frame called my_states of states in these regions with murder rates less than 1. Include columns state, rate, and rank.
- Use arrange to print mystates in descending order of rank.

```
murders_nw = subset(murders2, region %in% c("Northeast", "West"))
murders_nw
```

```
##            state abb    region population total      rate rank
## 2         Alaska  AK      West     710231    19 2.6751860   27
## 3        Arizona  AZ      West    6392017   232 3.6295273   10
## 5     California  CA      West   37253956  1257 3.3741383   14
## 6       Colorado  CO      West    5029196    65 1.2924531   38
## 7    Connecticut  CT Northeast    3574097    97 2.7139722   25
## 12        Hawaii  HI      West    1360301     7 0.5145920   49
## 13         Idaho  ID      West    1567582    12 0.7655102   46
## 20         Maine  ME Northeast    1328361    11 0.8280881   44
## 22 Massachusetts  MA Northeast    6547629   118 1.8021791   32
## 27       Montana  MT      West     989415    12 1.2128379   39
## 29        Nevada  NV      West    2700551    84 3.1104763   19
## 30 New Hampshire  NH Northeast    1316470     5 0.3798036   50
## 31    New Jersey  NJ Northeast    8791894   246 2.7980319   24
## 32    New Mexico  NM      West    2059179    67 3.2537239   15
## 33      New York  NY Northeast   19378102   517 2.6679599   29
## 38        Oregon  OR      West    3831074    36 0.9396843   42
## 39  Pennsylvania  PA Northeast   12702379   457 3.5977513   11
## 40  Rhode Island  RI Northeast    1052567    16 1.5200933   35
## 45          Utah  UT      West    2763885    22 0.7959810   45
## 46       Vermont  VT Northeast     625741     2 0.3196211   51
## 48    Washington  WA      West    6724540    93 1.3829942   37
## 51       Wyoming  WY      West     563626     5 0.8871131   43
```

```
my_states = murders_nw %>%
  filter(rate <= 1)%>%
```

```
  select(state, rate, rank)
my_states
```

```
##           state      rate rank
## 1        Hawaii 0.5145920   49
## 2         Idaho 0.7655102   46
## 3         Maine 0.8280881   44
## 4 New Hampshire 0.3798036   50
## 5        Oregon 0.9396843   42
## 6          Utah 0.7959810   45
## 7       Vermont 0.3196211   51
## 8       Wyoming 0.8871131   43
```

4

Use group_by to find the average and standard deviation of murder rates by region

```
murdersregion = murders2 %>%
  group_by(region) %>%
  summarize(mean = mean(rate), sd = sd(rate))
murdersregion
```

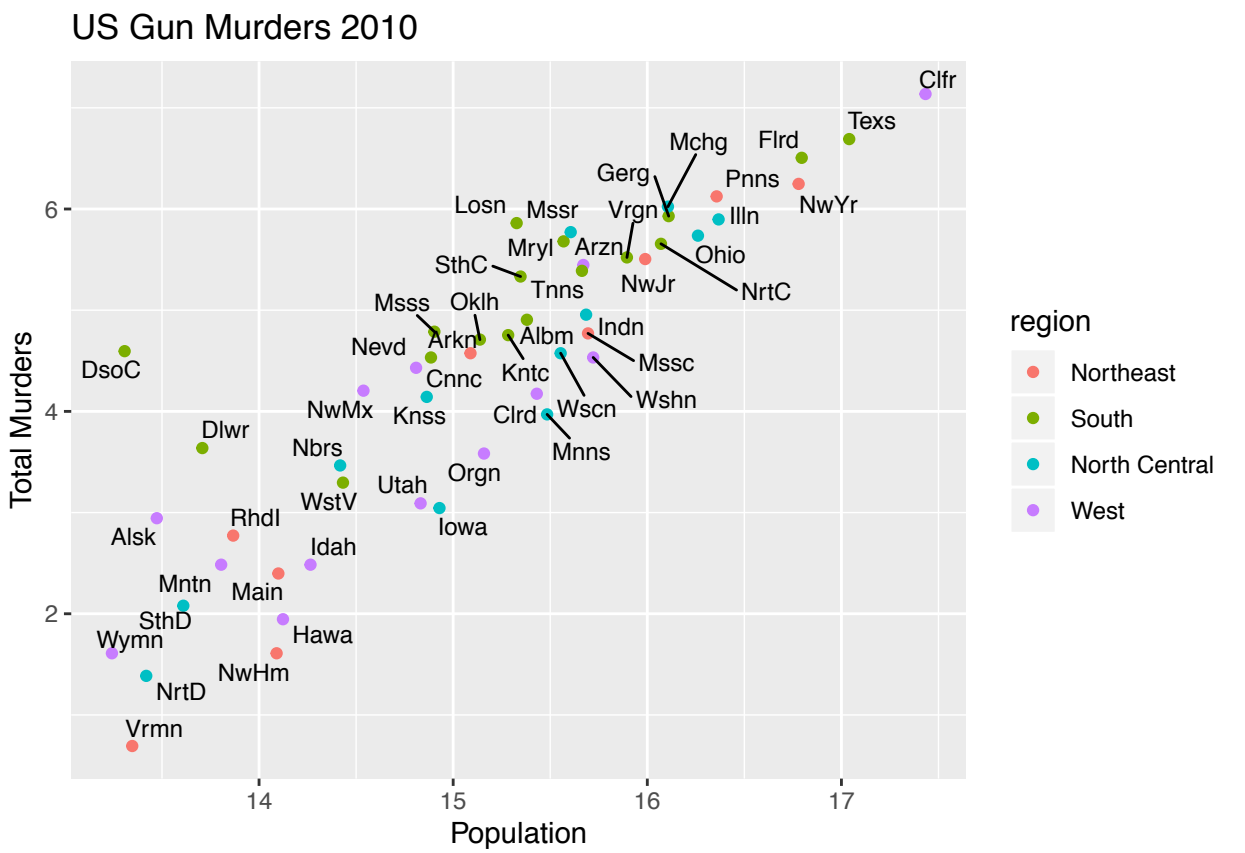
```
## # A tibble: 4 x 3
##   region         mean    sd
##   <fct>         <dbl> <dbl>
## 1 Northeast      1.85  1.17
## 2 South          4.42  3.37
## 3 North Central  2.18  1.44
## 4 West           1.83  1.17
```

5

Load ggplot2 and ggrepel. Use ggplot2 to make a scatterplot of population and total for the murders data set.

- Transform both population and murders total to log scale and replot.
- Add a title (US Gun Murders 2010) and axis labels
- Add the state abbreviations as labels using the repel package
- Color code the points by region

```
ggplot(murders2, aes(x = log(population), y = log(total))) +
  geom_point(aes(color = region)) +
  geom_text_repel(aes(label = abbreviate(state)), size = 3) +
  labs(title = "US Gun Murders 2010", x = "Population", y = "Total Murders")
```



6

Load the GGally package. Use ggpairs to make a correlation matrix plot for the variables in the swiss data set from the R base package.

- Use mutate to create a new column MajorityCatholic that is 0 when Catholic is less than 50 and 1 when Catholic is more than 50. Store this in a new data frame, swiss1.
- Replot the correlation matrix for the same variables using MajorityCatholic to color the points.

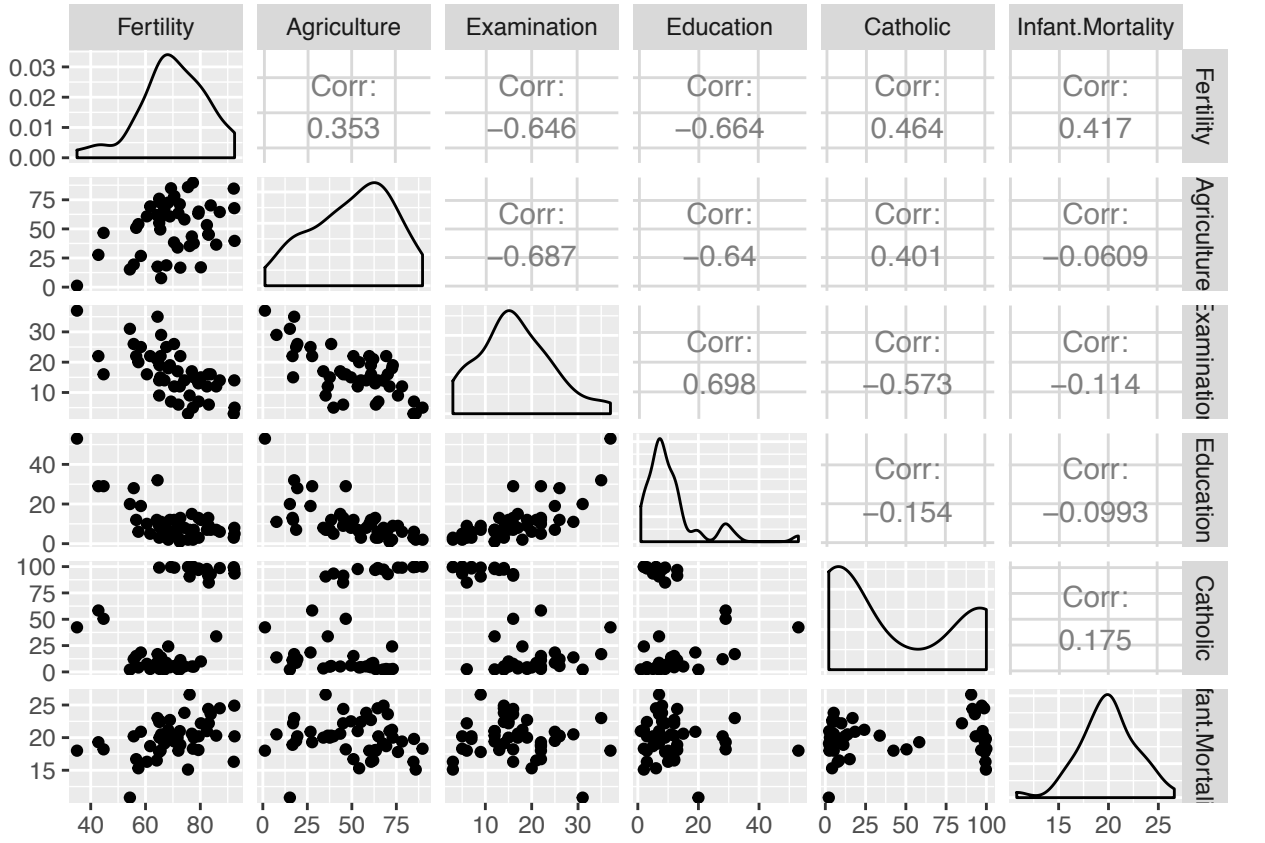
```
swiss
```

```
##              Fertility Agriculture Examination Education Catholic
## Courtelary        80.2        17.0          15        12     9.96
## Delemont          83.1        45.1           6         9    84.84
## Franches-Mnt      92.5        39.7           5         5    93.40
## Moutier           85.8        36.5          12         7    33.77
## Neuveville        76.9        43.5          17        15     5.16
## Porrentruy        76.1        35.3           9         7    90.57
## Broye             83.8        70.2          16         7    92.85
## Glane             92.4        67.8          14         8    97.16
## Gruyere           82.4        53.3          12         7    97.67
## Sarine            82.9        45.2          16        13    91.38
## Veveyse           87.1        64.5          14         6    98.61
## Aigle             64.1        62.0          21        12     8.52
## Aubonne           66.9        67.5          14         7     2.27
## Avenches          68.9        60.7          19        12     4.43
## Cossonay          61.7        69.3          22         5     2.82
## Echallens         68.3        72.6          18         2    24.20
```

```
## Grandson          71.7        34.0          17         8     3.30
## Lausanne          55.7        19.4          26        28    12.11
## La Vallee         54.3        15.2          31        20     2.15
## Lavaux            65.1        73.0          19         9     2.84
## Morges            65.5        59.8          22        10     5.23
## Moudon            65.0        55.1          14         3     4.52
## Nyone             56.6        50.9          22        12    15.14
## Orbe              57.4        54.1          20         6     4.20
## Oron              72.5        71.2          12         1     2.40
## Payerne           74.2        58.1          14         8     5.23
## Paysd'enhaut      72.0        63.5           6         3     2.56
## Rolle             60.5        60.8          16        10     7.72
## Vevey             58.3        26.8          25        19    18.46
## Yverdon           65.4        49.5          15         8     6.10
## Conthey           75.5        85.9           3         2    99.71
## Entremont         69.3        84.9           7         6    99.68
## Herens            77.3        89.7           5         2   100.00
## Martigwy          70.5        78.2          12         6    98.96
## Monthey           79.4        64.9           7         3    98.22
## St Maurice        65.0        75.9           9         9    99.06
## Sierre            92.2        84.6           3         3    99.46
## Sion              79.3        63.1          13        13    96.83
## Boudry            70.4        38.4          26        12     5.62
## La Chauxdfnd      65.7         7.7          29        11    13.79
## Le Locle          72.7        16.7          22        13    11.22
## Neuchatel         64.4        17.6          35        32    16.92
## Val de Ruz        77.6        37.6          15         7     4.97
## ValdeTravers      67.6        18.7          25         7     8.65
## V. De Geneve      35.0         1.2          37        53    42.34
## Rive Droite       44.7        46.6          16        29    50.43
## Rive Gauche       42.8        27.7          22        29    58.33
##              Infant.Mortality
## Courtelary               22.2
## Delemont                 22.2
## Franches-Mnt             20.2
## Moutier                  20.3
## Neuveville               20.6
## Porrentruy               26.6
## Broye                    23.6
## Glane                    24.9
## Gruyere                  21.0
## Sarine                   24.4
## Veveyse                  24.5
## Aigle                    16.5
## Aubonne                  19.1
## Avenches                 22.7
## Cossonay                 18.7
## Echallens                21.2
## Grandson                 20.0
## Lausanne                 20.2
## La Vallee                10.8
## Lavaux                   20.0
## Morges                   18.0
## Moudon                   22.4
```

```
## Nyone                16.7
## Orbe                 15.3
## Oron                 21.0
## Payerne              23.8
## Paysd'enhaut         18.0
## Rolle                16.3
## Vevey                20.9
## Yverdon              22.5
## Conthey              15.1
## Entremont            19.8
## Herens               18.3
## Martigwy             19.4
## Monthey              20.2
## St Maurice           17.8
## Sierre               16.3
## Sion                 18.1
## Boudry               20.3
## La Chauxdfnd         20.5
## Le Locle             18.9
## Neuchatel            23.0
## Val de Ruz           20.0
## ValdeTravers         19.5
## V. De Geneve         18.0
## Rive Droite          18.2
## Rive Gauche          19.3
```

```
ggpairs(swiss)
```



```
swiss1 = mutate(swiss, MajorityCatholic = ifelse(Catholic > 50, 1, 0))
swiss1
```

```
##    Fertility Agriculture Examination Education Catholic Infant.Mortality
## 1       80.2        17.0          15        12     9.96             22.2
## 2       83.1        45.1           6         9    84.84             22.2
## 3       92.5        39.7           5         5    93.40             20.2
## 4       85.8        36.5          12         7    33.77             20.3
## 5       76.9        43.5          17        15     5.16             20.6
## 6       76.1        35.3           9         7    90.57             26.6
## 7       83.8        70.2          16         7    92.85             23.6
## 8       92.4        67.8          14         8    97.16             24.9
## 9       82.4        53.3          12         7    97.67             21.0
## 10      82.9        45.2          16        13    91.38             24.4
## 11      87.1        64.5          14         6    98.61             24.5
## 12      64.1        62.0          21        12     8.52             16.5
## 13      66.9        67.5          14         7     2.27             19.1
## 14      68.9        60.7          19        12     4.43             22.7
## 15      61.7        69.3          22         5     2.82             18.7
## 16      68.3        72.6          18         2    24.20             21.2
## 17      71.7        34.0          17         8     3.30             20.0
## 18      55.7        19.4          26        28    12.11             20.2
## 19      54.3        15.2          31        20     2.15             10.8
## 20      65.1        73.0          19         9     2.84             20.0
## 21      65.5        59.8          22        10     5.23             18.0
## 22      65.0        55.1          14         3     4.52             22.4
## 23      56.6        50.9          22        12    15.14             16.7
## 24      57.4        54.1          20         6     4.20             15.3
## 25      72.5        71.2          12         1     2.40             21.0
## 26      74.2        58.1          14         8     5.23             23.8
## 27      72.0        63.5           6         3     2.56             18.0
## 28      60.5        60.8          16        10     7.72             16.3
## 29      58.3        26.8          25        19    18.46             20.9
## 30      65.4        49.5          15         8     6.10             22.5
## 31      75.5        85.9           3         2    99.71             15.1
## 32      69.3        84.9           7         6    99.68             19.8
## 33      77.3        89.7           5         2   100.00             18.3
## 34      70.5        78.2          12         6    98.96             19.4
## 35      79.4        64.9           7         3    98.22             20.2
## 36      65.0        75.9           9         9    99.06             17.8
## 37      92.2        84.6           3         3    99.46             16.3
## 38      79.3        63.1          13        13    96.83             18.1
## 39      70.4        38.4          26        12     5.62             20.3
## 40      65.7         7.7          29        11    13.79             20.5
## 41      72.7        16.7          22        13    11.22             18.9
## 42      64.4        17.6          35        32    16.92             23.0
## 43      77.6        37.6          15         7     4.97             20.0
## 44      67.6        18.7          25         7     8.65             19.5
## 45      35.0         1.2          37        53    42.34             18.0
## 46      44.7        46.6          16        29    50.43             18.2
## 47      42.8        27.7          22        29    58.33             19.3
##    MajorityCatholic
## 1                 0
## 2                 1
```

```
## 3                 1
## 4                 0
## 5                 0
## 6                 1
## 7                 1
## 8                 1
## 9                 1
## 10                1
## 11                1
## 12                0
## 13                0
## 14                0
## 15                0
## 16                0
## 17                0
## 18                0
## 19                0
## 20                0
## 21                0
## 22                0
## 23                0
## 24                0
## 25                0
## 26                0
## 27                0
## 28                0
## 29                0
## 30                0
## 31                1
## 32                1
## 33                1
## 34                1
## 35                1
## 36                1
## 37                1
## 38                1
## 39                0
## 40                0
## 41                0
## 42                0
## 43                0
## 44                0
## 45                0
## 46                1
## 47                1
```

```
ggpairs(swiss1, ggplot2::aes(color = as.factor(MajorityCatholic)))
```

```
## Warning in cor(x, y, method = method, use = use): the standard deviation is zero

## Warning in cor(x, y, method = method, use = use): the standard deviation is zero

## Warning in cor(x, y, method = method, use = use): the standard deviation is zero

## Warning in cor(x, y, method = method, use = use): the standard deviation is zero
```

```
## Warning in cor(x, y, method = method, use = use): the standard deviation is zero

## Warning in cor(x, y, method = method, use = use): the standard deviation is zero

## Warning in cor(x, y, method = method, use = use): the standard deviation is zero

## Warning in cor(x, y, method = method, use = use): the standard deviation is zero

## Warning in cor(x, y, method = method, use = use): the standard deviation is zero

## Warning in cor(x, y, method = method, use = use): the standard deviation is zero

## Warning in cor(x, y, method = method, use = use): the standard deviation is zero

## Warning in cor(x, y, method = method, use = use): the standard deviation is zero
```