Homework 3

Enter your name and EID here

Connor Hanna cdh3663

This homework is due on Feb. 7, 2022 at 11:00am. Please submit as a pdf file on Canvas.

Problem 1: (2 pts) For problem 1, we will work with the `diamonds` dataset. See here for details: https://ggplot2.tidyverse.org/reference/diamonds.html.

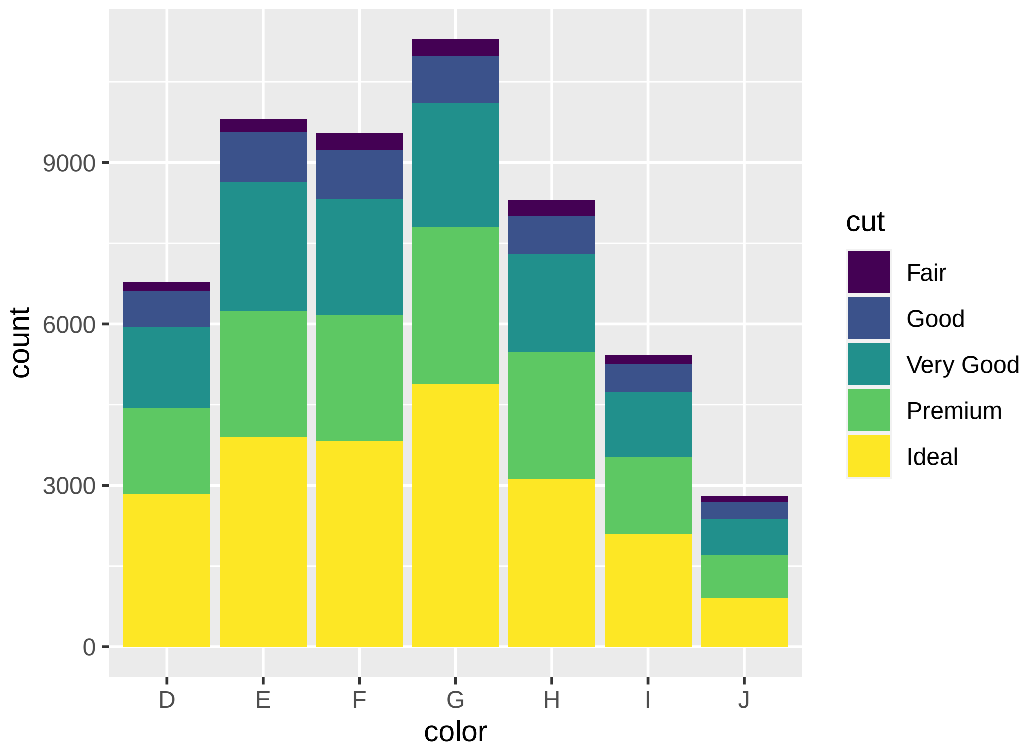
```
diamonds
```

```
## # A tibble: 53,940 x 10
##    carat cut       color clarity depth table price     x     y     z
##    <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
##  1  0.23 Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43
##  2  0.21 Premium   E     SI1      59.8    61   326  3.89  3.84  2.31
##  3  0.23 Good      E     VS1      56.9    65   327  4.05  4.07  2.31
##  4  0.29 Premium   I     VS2      62.4    58   334  4.2   4.23  2.63
##  5  0.31 Good      J     SI2      63.3    58   335  4.34  4.35  2.75
##  6  0.24 Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48
##  7  0.24 Very Good I     VVS1     62.3    57   336  3.95  3.98  2.47
##  8  0.26 Very Good H     SI1      61.9    55   337  4.07  4.11  2.53
##  9  0.22 Fair      E     VS2      65.1    61   337  3.87  3.78  2.49
## 10  0.23 Very Good H     VS1      59.4    61   338  4     4.05  2.39
## # ... with 53,930 more rows
```

(a) Use ggplot to make a bar plot of the total diamond count per `color` and show the proportion of each `cut` within each `color` category.

(b) In two sentences, explain when to use `geom_bar()` instead of `geom_col()`. Which of these functions requires only an `x` or `y` variable?

```
# a.
ggplot(diamonds, aes(color, fill = cut)) +
  geom_bar()
```



(b) `geom_col()` can be used to make bar graphs where observations contain values that need to be attributed to both an `x` and `y` axis, especially when each observation requires a separate column as in the `txhouse` data from HW 2. `geom_bar()` is used when counting obervations containing certain values for a specified `x` or `y` variable.

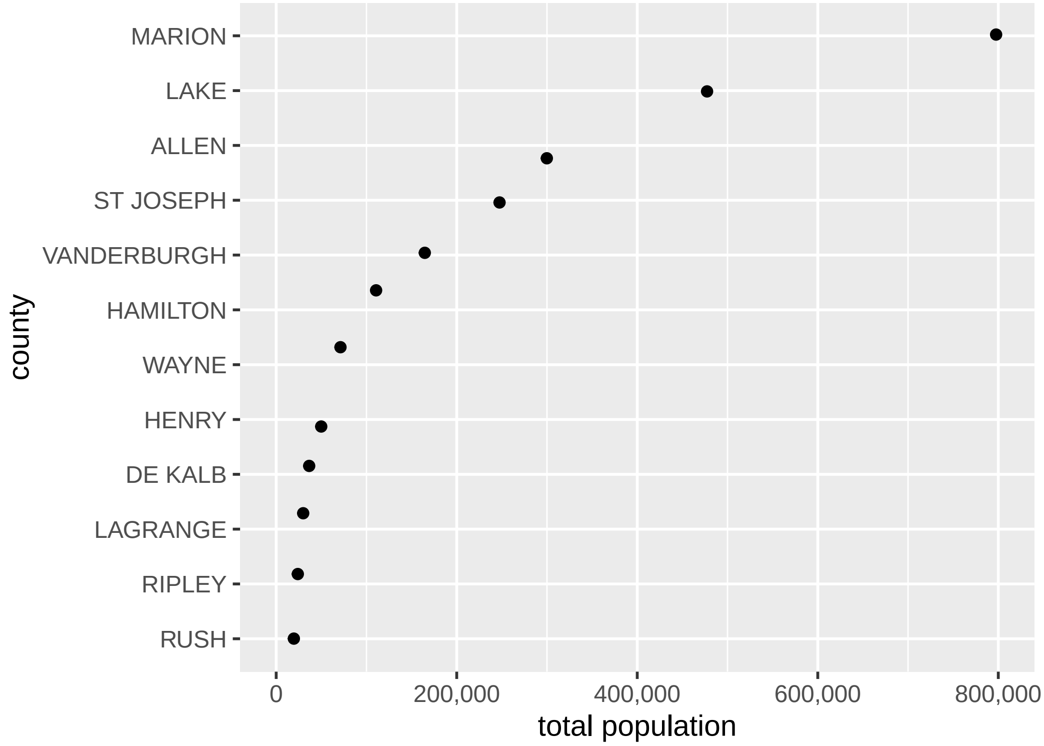
Problem 2: (4 pts) For problems 2 and 3, we will work with the dataset `IL_pop` that contains Illinois state demographics and has been derived from the `midwest` dataset provided by **ggplot2**. See here for details of the original dataset: https://ggplot2.tidyverse.org/reference/midwest.html. `IL_pop` contains two columns: `county` and `poptotal` (the county's total population).

```
IL_pop
```

```
## # A tibble: 12 x 2
##    county      poptotal
##    <chr>          <int>
##  1 MARION        797159
##  2 LAKE          475594
##  3 ALLEN         300836
##  4 ST JOSEPH     247052
##  5 VANDERBURGH   165058
##  6 HAMILTON      108936
##  7 WAYNE          71951
##  8 HENRY          48139
##  9 DE KALB        35324
## 10 LAGRANGE       29477
## 11 RIPLEY         24616
## 12 RUSH           18129
```

(a) Use ggplot to make a scatter plot of `county` vs total population (column `poptotal`) and order the counties by increasing population.
(b) Rename the axes and set appropriate limits, breaks and labels. Note: Do not use `xlab()` or `ylab()` to label the axes.

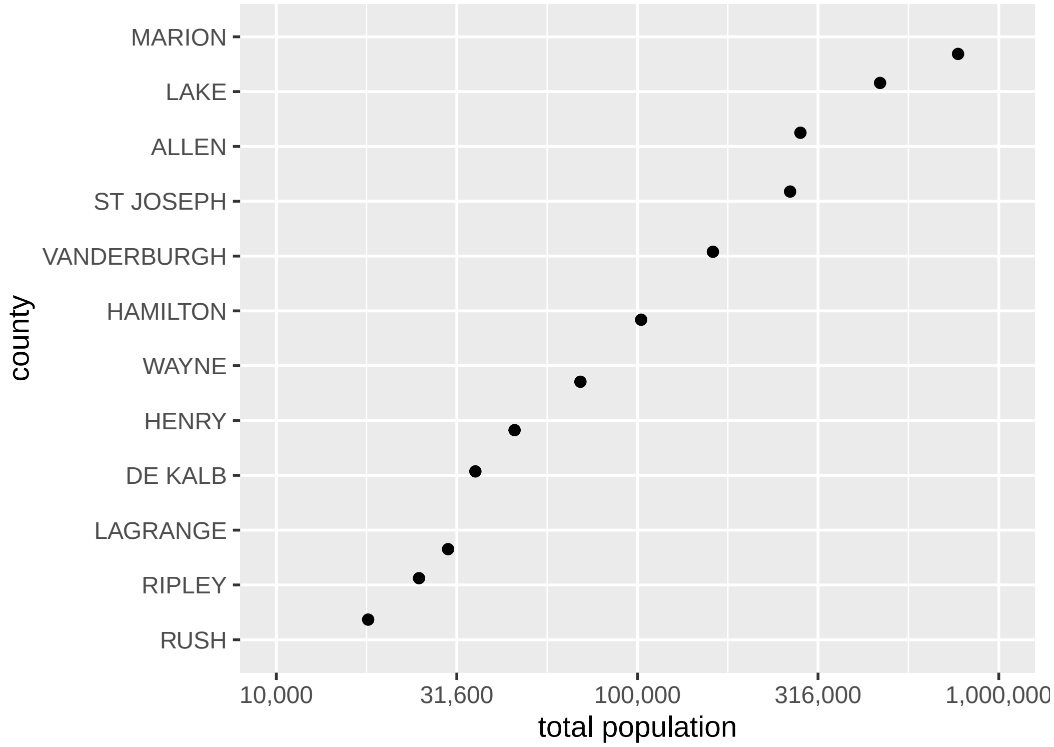
```r
ggplot(IL_pop, aes(poptotal, fct_reorder(county, poptotal))) +
  geom_jitter() +
  scale_x_continuous(
    name = "total population",
    limits = c(0, 800000),
    breaks = c(0, 200000, 400000, 600000, 800000),
    labels = c("0", "200,000", "400,000", "600,000", "800,000")
  ) +
  scale_y_discrete(
    name = "county"
  )
```



Problem 3: (4 pts)

(a) Modify the plot from Problem 2 by changing the scale for `poptotal` to logarithmic.
(b) Adjust the limits, breaks and labels for the logarithmic scale.

```r
ggplot(IL_pop, aes(poptotal, fct_reorder(county, poptotal))) +
  geom_jitter() +
  scale_x_log10(
    name = "total population",
    limits = c(1e4, 1e6),
    breaks = c(1e4, 3.16e4, 1e5, 3.16e5, 1e6),
    labels = c("10,000", "31,600", "100,000", "316,000", "1,000,000")
  ) +
  scale_y_discrete(
    name = "county"
  )
```


county
MARION
LAKE
ALLEN
ST JOSEPH
VANDERBURGH
HAMILTON
WAYNE
HENRY
DE KALB
LAGRANGE
RIPLEY
RUSH
total population