Homework 3

Ethan Chang - ehc586

This homework is due on Jan. 31, 2023 at 11:00pm. Please submit as a pdf file on Canvas.

Problem 1: (4 pts) For problem 1, we will work with the `diamonds` dataset. See here for details: https://ggplot2.tidyverse.org/reference/diamonds.html.

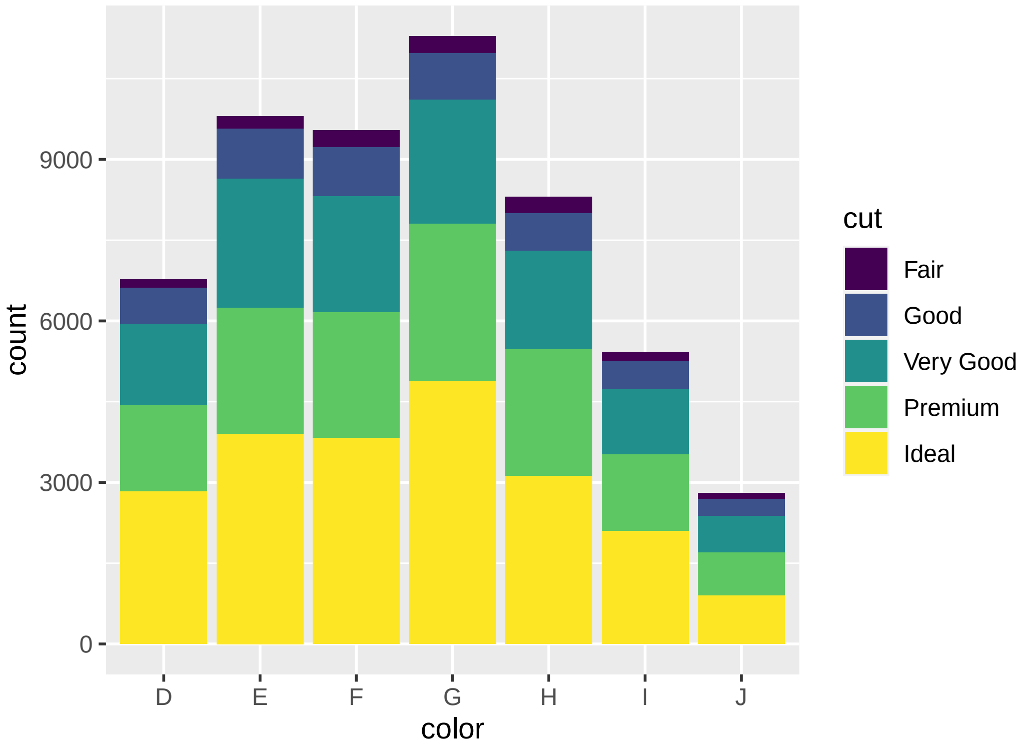
```
diamonds
```

```
## # 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?

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



One would use `geom_bar()` instead of `geom_col()` when graphing bar plots that rely on count with bar heights proportional to the number of cases in each group. The `geom_bar()` function requires only an `x` or `y` variable as it only needs one, while `geom_col()` needs both.

Problem 2: (4 pts) For problem 2 and 3, we will work with the dataset `OH_pop` that contains Ohio 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. `OH_pop` contains two columns: `county` and `poptotal` (the county's total population), and it only contains counties with at least 100,000 inhabitants.

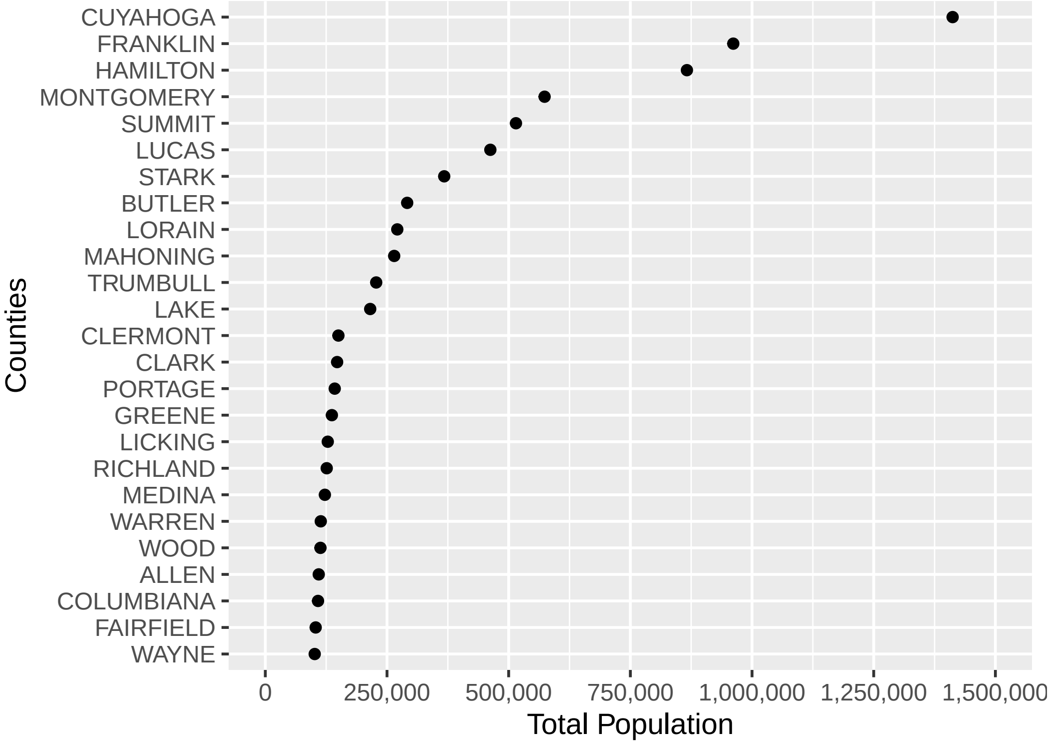
```
OH_pop
```

```
## # A tibble: 25 x 2
##    county     poptotal
##    <chr>         <int>
##  1 CUYAHOGA    1412140
##  2 FRANKLIN     961437
##  3 HAMILTON     866228
##  4 MONTGOMERY   573809
##  5 SUMMIT       514990
##  6 LUCAS        462361
##  7 STARK        367585
##  8 BUTLER       291479
##  9 LORAIN       271126
## 10 MAHONING     264806
## # ... with 15 more rows
```

(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.

```
ggplot(OH_pop, aes(poptotal, fct_reorder(county,poptotal))) +
  geom_point() +
  scale_x_continuous(name = "Total Population",
                     limits = c(0,1500000),
                     breaks = seq(0,1500000,250000),
                     labels = c("0","250,000","500,000","750,000","1,000,000","1,250,000",
                                "1,500,000")) +
  scale_y_discrete(name = "Counties")
```

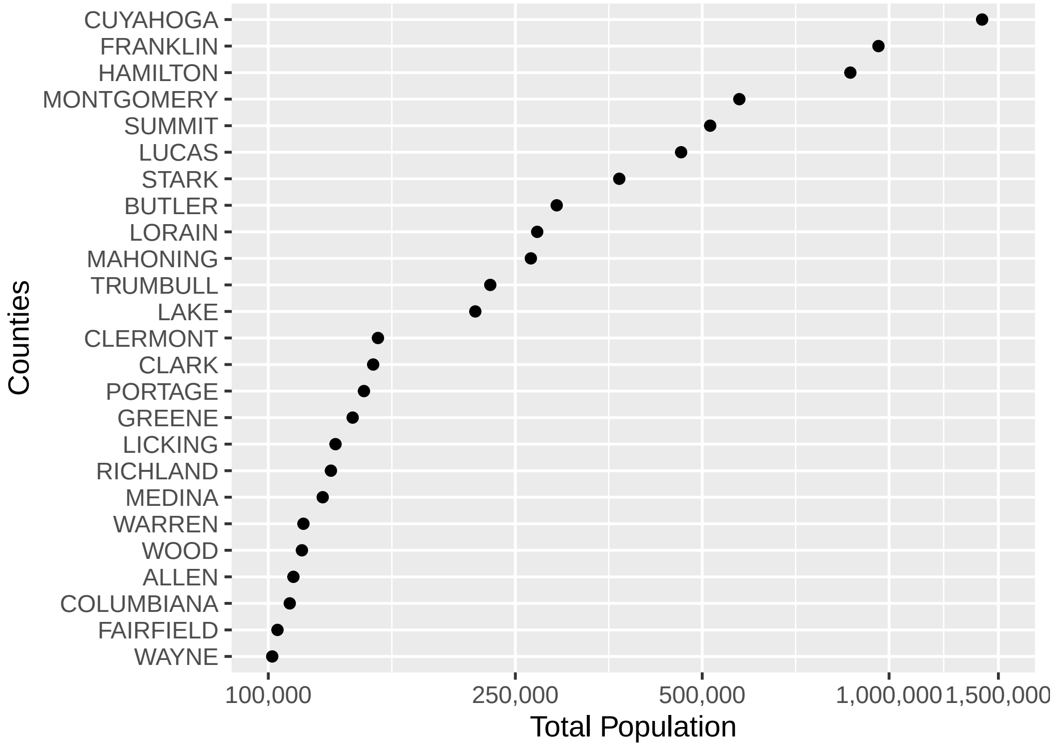


Problem 3: (2 pts)

(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.

```
ggplot(OH_pop, aes(poptotal, fct_reorder(county,poptotal))) +
  geom_point() +
  scale_x_log10(name = "Total Population",
                limits = c(100000,1500000),
                breaks = c(100000, 250000, 500000, 1000000, 1500000),
                labels = c("100,000", "250,000", "500,000", "1,000,000", "1,500,000")) +
  scale_y_discrete(name = "Counties")
```


CUYAHOGA
FRANKLIN
HAMILTON
MONTGOMERY
SUMMIT
LUCAS
STARK
BUTLER
LORAIN
MAHONING
TRUMBULL
LAKE
CLERMONT
CLARK
PORTAGE
GREENE
LICKING
RICHLAND
MEDINA
WARREN
WOOD
ALLEN
COLUMBIANA
FAIRFIELD
WAYNE
Counties
100,000
250,000
500,000
1,000,000
1,500,000
Total Population