Question 2 [50 points]

We will re-use the same `midwest_modified` data that was used in Question 1, with all the modifications from the other question parts. The description is repeated below for your convenience.

```
str(midwest_modified)
```

```
spec_tbl_df [437 x 11] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ county       : chr [1:437] "ADAMS" "ALEXANDER" "BOND" "BOONE" ...
 $ state        : chr [1:437] "IL" "IL" "IL" "IL" ...
 $ popdensity   : num [1:437] 1271 759 681 1812 324 ...
 $ popwhite     : num [1:437] 63917 7054 14477 29344 5264 ...
 $ popblack     : num [1:437] 1702 3496 429 127 547 ...
 $ popamerindian: num [1:437] 98 19 35 46 14 65 8 30 8 331 ...
 $ popasian     : num [1:437] 249 48 16 150 5 ...
 $ popother     : num [1:437] 124 9 34 1139 6 ...
 $ inmetro      : num [1:437] 0 0 0 1 0 0 0 0 0 1 ...
 $ Metro        : chr [1:437] "NonMetro" "NonMetro" "NonMetro" "Metro" ...
 $ HighDens     : chr [1:437] "NotHigh" "NotHigh" "NotHigh" "High" ...
 - attr(*, "spec")=
  .. cols(
  ..   county = col_character(),
  ..   state = col_character(),
  ..   popdensity = col_double(),
  ..   popwhite = col_double(),
  ..   popblack = col_double(),
  ..   popamerindian = col_double(),
  ..   popasian = col_double(),
  ..   popother = col_double(),
  ..   inmetro = col_double(),
  ..   Metro = col_character(),
  ..   HighDens = col_character()
  .. )
```

```
midwest_modified %>% slice(1:5) %>%
  select(county:popblack)
```

```
# A tibble: 5 x 5
  county    state popdensity popwhite popblack
  <chr>     <chr>      <dbl>    <dbl>    <dbl>
1 ADAMS     IL         1271.    63917     1702
2 ALEXANDER IL          759      7054     3496
3 BOND      IL          681.    14477      429
4 BOONE     IL         1812.    29344      127
5 BROWN     IL          324.     5264      547
```

```
midwest_modified %>% slice(1:5) %>%
  select(county,popamerindian:HighDens)
```

```
# A tibble: 5 x 7
  county    popamerindian popasian popother inmetro Metro    HighDens
  <chr>             <dbl>    <dbl>    <dbl>   <dbl> <chr>    <chr>
1 ADAMS                98      249      124       0 NonMetro NotHigh
2 ALEXANDER            19       48        9       0 NonMetro NotHigh
3 BOND                 35       16       34       0 NonMetro NotHigh
4 BOONE                46      150     1139       1 Metro    High
5 BROWN                14        5        6       0 NonMetro NotHigh
```

The dataset contains population data from midwest counties in five states in the United States from an unspecified year. There are identifying variables for both the `county` (the name) and the `state` (the postal abbreviation).

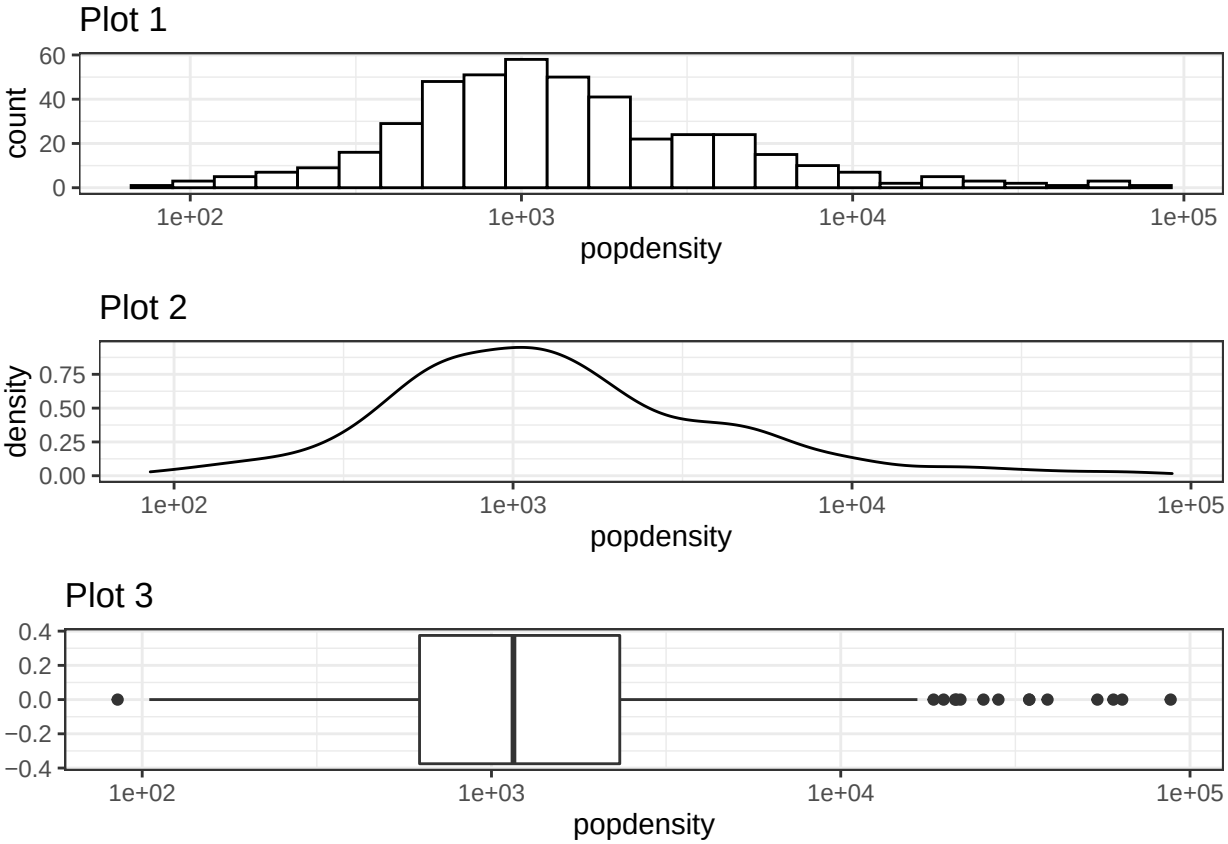
The variable `popdensity` is a measure of density (population per unspecified area units). The variable `inmetro` is equal to 1 if the county is classified as a metropolitan area and 0 otherwise. The other variables contain counts of population size within self-identified racial classifications.

CONTINUED ON NEXT PAGE

(a) **[6 pts]** Below are partially obscured code and three plots of the values of the *log (base 10)* of the population density for all counties:

```
p1<-ggplot(midwest_modified,aes(x=popdensity)) + geom_XXXXX(nbins=30,fill="white",col="black") +
   ggtitle("Plot 1") + theme_bw() + scale_x_log10()
p2<-ggplot(midwest_modified,aes(x=popdensity)) + geom_YYYYY() +
   ggtitle("Plot 2") + theme_bw()+ scale_x_log10()
p3<-ggplot(midwest_modified,aes(x=popdensity)) + geom_ZZZZZ() +
   ggtitle("Plot 3") + theme_bw()+ scale_x_log10()
grid.arrange(grobs=list(p1,p2,p3),nrow=3,ncol=1)
```

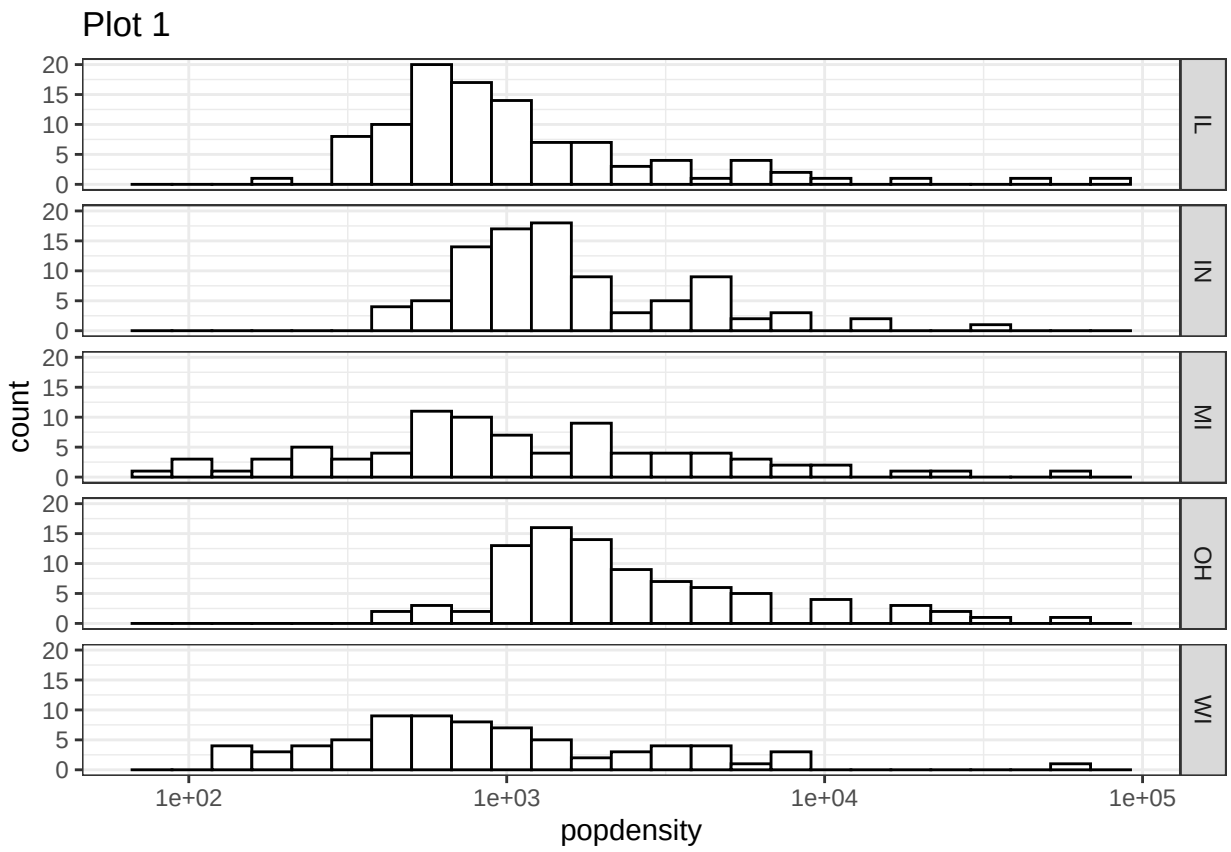
CONTINUED ON NEXT PAGE



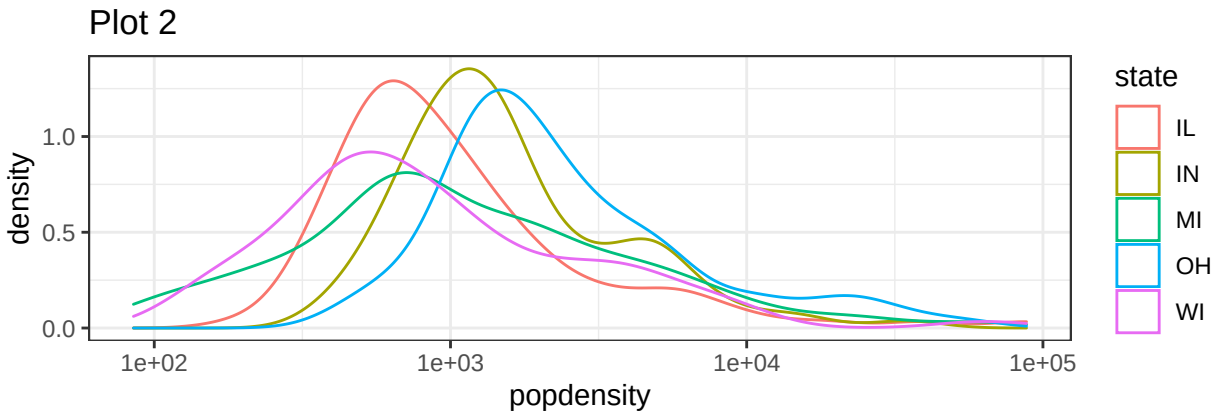
Identify these three plots by name:

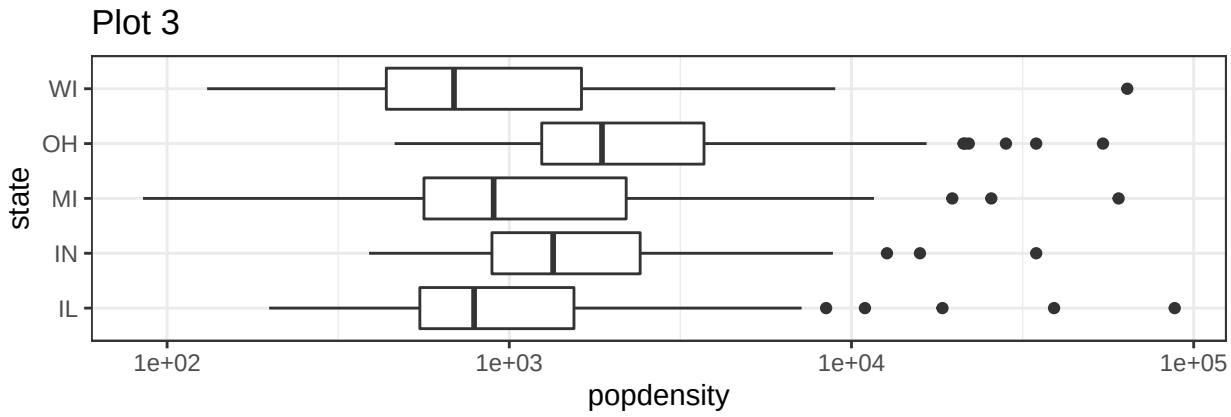
Plot 1 Plot 2 Plot 3

(b) **[10 pts]** Now we make the same plots, but for each state. Do you believe there is evidence of an association between state and population density? In particular, do we see differences in the distributions of population density by state?




```
grid.arrange(grobs=list(p2,p3),nrow=2,ncol=1)
```





CONTINUED ON NEXT PAGE

(c) **[4 pts]** Which plot(s) do you think best shows the association between state and population density? Which plot(s) do you think does not shows the association between state and population density as clearly? Explain your answer and reasoning in a few sentences.

(d) **[5 pts]** Which of the following plots could also be used to assess the association between the `popwhite` and `popblack` variables? List all that apply (or say None if none would be appropriate).

A. 2-d density plot B. Barplot C. Boxplot D. 2-d histogram

We now would like to make plots to take a different look at the population variables. Unfortunately, the format of the `midwest_modified` data needs to be further changed so that we can use it in a `ggplot`.

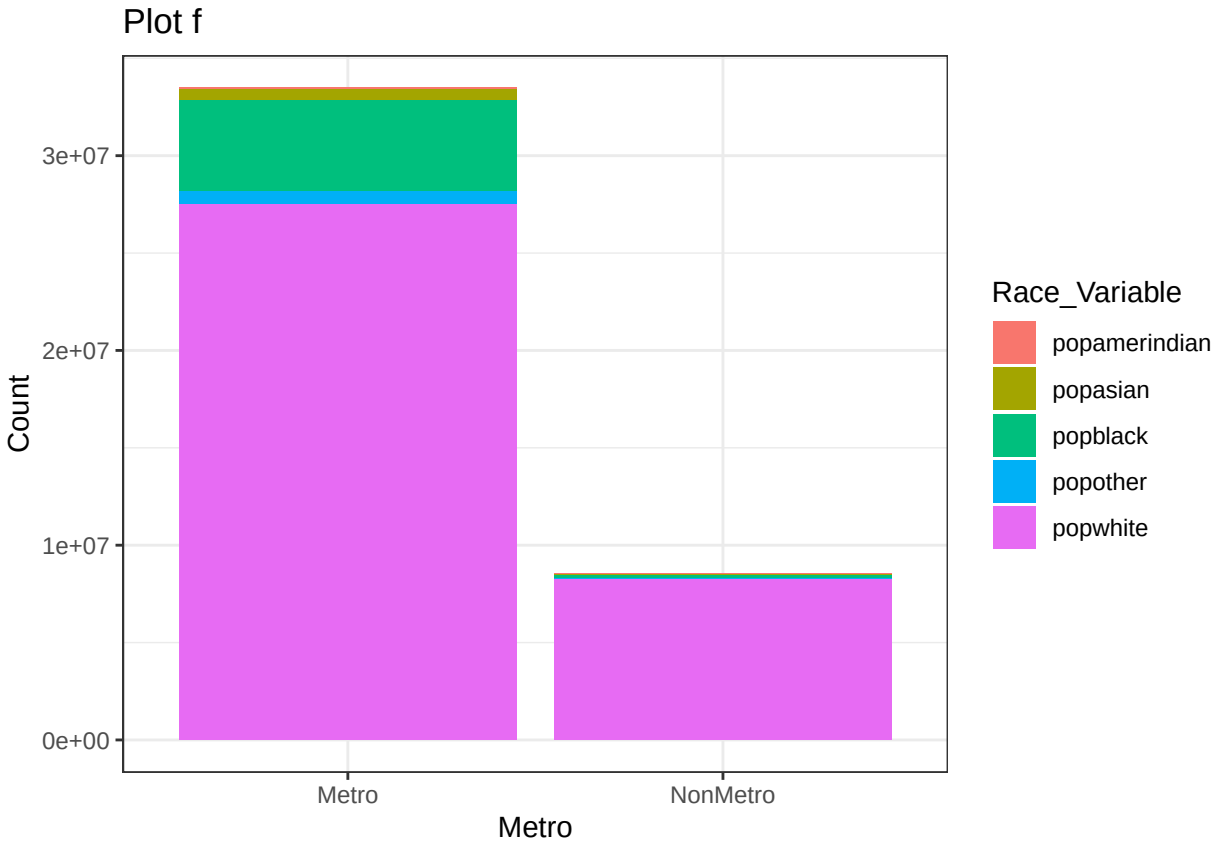
(e) **[5 pts]** Write a line of code that will create a new `tibble` converts the `midwest_modified_new` to "long" format where each row contains a population count for a specific racial group called `Count`, and the variable from where that count originated (e.g. `popwhite`) as well as the `state`, `county`, and `Metro` information for that population group. You should not include the columns for `HighDens`, `inmetro` or `popdensity`. The first 10 rows of the new tibble are below

```
midwest_modified_new %>% slice(1:10)
```

```
# A tibble: 10 x 5
   county    state Metro    Race_Variable Count
   <chr>     <chr> <chr>    <chr>         <dbl>
 1 ADAMS     IL    NonMetro popwhite      63917
 2 ADAMS     IL    NonMetro popblack       1702
 3 ADAMS     IL    NonMetro popamerindian    98
 4 ADAMS     IL    NonMetro popasian        249
 5 ADAMS     IL    NonMetro popother        124
 6 ALEXANDER IL    NonMetro popwhite       7054
 7 ALEXANDER IL    NonMetro popblack       3496
 8 ALEXANDER IL    NonMetro popamerindian    19
 9 ALEXANDER IL    NonMetro popasian         48
10 ALEXANDER IL    NonMetro popother          9
```

CONTINUED ON NEXT PAGE

Below is a figure along with the code (partially obscured) which generated it.




```
ggplot(midwest_modified_new,aes(x=______,fill=______,y=______,
                       )) +
  geom_XXXXXX(stat="identity") +  ggtitle("Plot f")  + theme_bw()
```

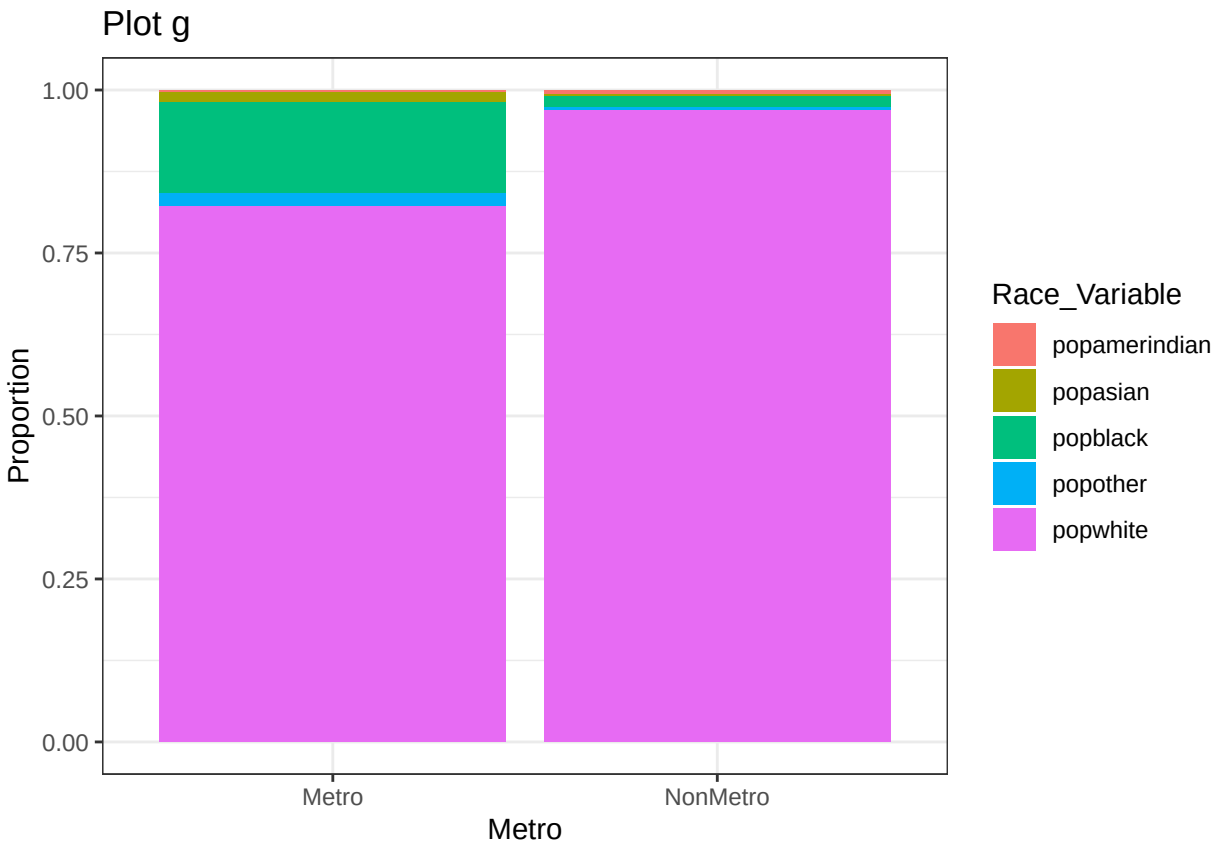
(f) **[5 pts]** What are the missing geometry and aesthetics that generated the figure on the previous page (that is, what are the words that are missing in the code above for Plot f)?

(g) **[5 pts]** Note that the plot in part (f) is a bit difficult to use because it contains the counts, rather than the relative proportions. Write a line of code (or lines of code) to create a new tibble called `metro_race_summaries` which contains each racial population count and proportion relative to the level of the Metro variable as below:

```
metro_race_summaries
```

```
# A tibble: 10 x 4
# Groups:   Metro [2]
   Metro    Race_Variable Race_Count Proportion
   <chr>    <chr>              <dbl>      <dbl>
 1 Metro    popamerindian      99145    0.00296
 2 Metro    popasian          538463    0.0161
 3 Metro    popblack         4672825    0.140
 4 Metro    popother          668449    0.0200
 5 Metro    popwhite        27496337    0.821
 6 NonMetro popamerindian      50794    0.00595
 7 NonMetro popasian           34210    0.00401
 8 NonMetro popblack          144611    0.0169
 9 NonMetro popother           36402    0.00427
10 NonMetro popwhite         8267706    0.969
```

(h) **[5 pts]** Using the tibble from (g), write a line of code that created the barplot below.



(i) **[5 pts]** Based on the plot in part (h), would you conclude that there the population distribution of race varies between Metro and NonMetro areas? Explain your answer in a few sentences.