Question 2

a.

```
p1<-ggplot(midwest_modified,aes(x=popdensity)) + geom_histogram(bins=30,fill="white",col="black") +
  ggtitle("Plot 1") + theme_bw() + scale_x_log10()

p1

p2<-ggplot(midwest_modified,aes(x=popdensity)) + geom_density() +
  ggtitle("Plot 2") + theme_bw()+ scale_x_log10()

p2

p3<-ggplot(midwest_modified,aes(x=popdensity)) + geom_boxplot() +
  ggtitle("Plot 3") + theme_bw()+ scale_x_log10()

p3
```

P1 -> Histogram
P2 -> Density Plot
P3 -> Boxplot

b. I believe there is some association between state and population density. We can also spot some differences between the states when it comes to their population density and where people like to live in those states.

To elaborate, WI has an overall lower count in terms of population density (likely arising from the lower number of counties in this state) – while IL has an overall higher count due to the higher number of counties.

Furthermore, WI does not have a lot of counties with a population density higher than 1e+04, while OH has more than others – which hints that some counties of OH could be heavily overpopulated.

The median population density for OH and IN is higher than the overall average for all 5 states, while it is lower for the 3 other states.

While WI and MI has a more equal balance between people living in each of its counties, the balance is not as equal for the other 3 states – which means people prefer living in some specific counties in these 3 states a lot more than others. Finally, MI also has a very high variance in data – which means the population density in Michigan is more equally distributed than all other states. It is the least in IN, which confirms the idea that most people live in very few counties in IN.

c. I think all 3 graphs are helpful in showing us the association between state and population density. They present different facts about the data which helps us reconfirm our readings from the other two.
On its own, the density plot possibly has the best representation since it is plotted on the same axes and graph and thus helps us get an overall clearer picture about the association between

states and population density. On the other hand, the histogram, while in this case shows us an association, might not be the best choice since it would be very difficult to compare them if it was not placed in a grid pattern and as separate graphs instead.

d. I believe all 4 of 2-d density plot, Barplot, Boxplot and 2-d histogram would be useful in assessing the association between popwhite and popblack variables.
However, 2D histogram and density plot would be able to show us a direct association between popwhite and popblack variables. On the other hand, we would have to plot popwhite and popblack separately for each state using a stacked/grouped barplot or a boxplot side by side in the same graph and then compare the data to indirectly assess the association between popwhite and popblack.
Thus, 2D density plots and 2D histograms would be better, but the other 2 can also serve the purpose indirectly.

e.

```
midwest_modified_new <- midwest_modified %>% pivot_longer(cols = popwhite:popother, names_to = "Race_Variable", values_to = "Count") %>%
                        select(county:state, Metro, Race_Variable:Count) %>% mutate(Count = as.double(Count))

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

midwest_modified_new <- midwest_modified %>% pivot_longer(cols = popwhite:popother, names_to = "Race_Variable", values_to = "Count") %>%
select(county:state, Metro, Race_Variable:Count) %>% mutate(Count = as.double(Count))

midwest_modified_new %>% slice(1:10)

f.

```
ggplot(midwest_modified_new, aes(x= Metro,fill= Race_Variable,y= Count)) + geom_bar(stat="identity") + ggtitle("Plot f") + theme_bw()
```

ggplot(midwest_modified_new, aes(x= Metro,fill= Race_Variable,y= Count)) + geom_bar(stat="identity") + ggtitle("Plot f") + theme_bw()

This means: x -> Metro, fill -> Race_Variable, y-> Count, and geom_xxxxx -> geom_bar

g.

```
metro_race_summaries <- midwest_modified_new %>% group_by(Metro, Race_Variable) %>% summarise(Race_Count = sum(Count)) %>%
  mutate(Race_Count = as.double(Race_Count)) %>% mutate(Proportion = Race_Count/sum(Race_Count))
```

metro_race_summaries <- midwest_modified_new %>% group_by(Metro, Race_Variable) %>% summarise(Race_Count = sum(Count)) %>%
mutate(Race_Count = as.double(Race_Count)) %>% mutate(Proportion = Race_Count/sum(Race_Count))

h.
```
ggplot(metro_race_summaries, aes(x = Metro, fill = Race_Variable,y = Proportion)) + geom_bar(stat = "identity") + ggtitle("Plot g") + theme_bw()
```

ggplot(metro_race_summaries, aes(x = Metro, fill = Race_Variable,y = Proportion)) +
geom_bar(stat = "identity") + ggtitle("Plot g") + theme_bw()

i. Yes. As we can clearly see from the barplot, the proportion of popwhite is comparatively higher in NonMetro areas compared to Metro areas. On the other hand, the proportion of popblack is comparatively higher in Metro areas than NonMetro areas.
However, the absolute proportion of popwhite is overall higher than all other populations for both Metro and NonMetro areas. On the other hand, populations other than popwhite and popblack comprise very little portions of the dataset – and are close to negligible for both metro and nonmetro areas.