A05 - Suthi de Silva - CSC 285 - 15th Jan 2024

Loading the data

```
icd <- read.table('/srv/R/CSC285_May21/Data/IdahoCitiesData.txt', header = TRUE, sep = '\t')
```

Idaho Cities and Axis Transformations

Transforming our data and axes has a large impact on visualizations. In the /srv/R/CSC_May21/Data folder, we have fairly recent data on the population of cities in Idaho. The data contains information from all Idaho cities with a population of at least 100. The variables are described below:

1) **name** – Name of the city
2) **pop2021** – population in 2021
3) **pop2010** – population in 2010

Create an clear, nicely commented R Markdown based on the instructions below. The population of Caldwell in 2021 has an extra 0 at the end! Use R to correct this typo.

```
icd$pop2021[icd$name == 'Caldwell'] <- icd$pop2021[icd$name == 'Caldwell'] / 10
icd[icd$name == 'Caldwell',]
```

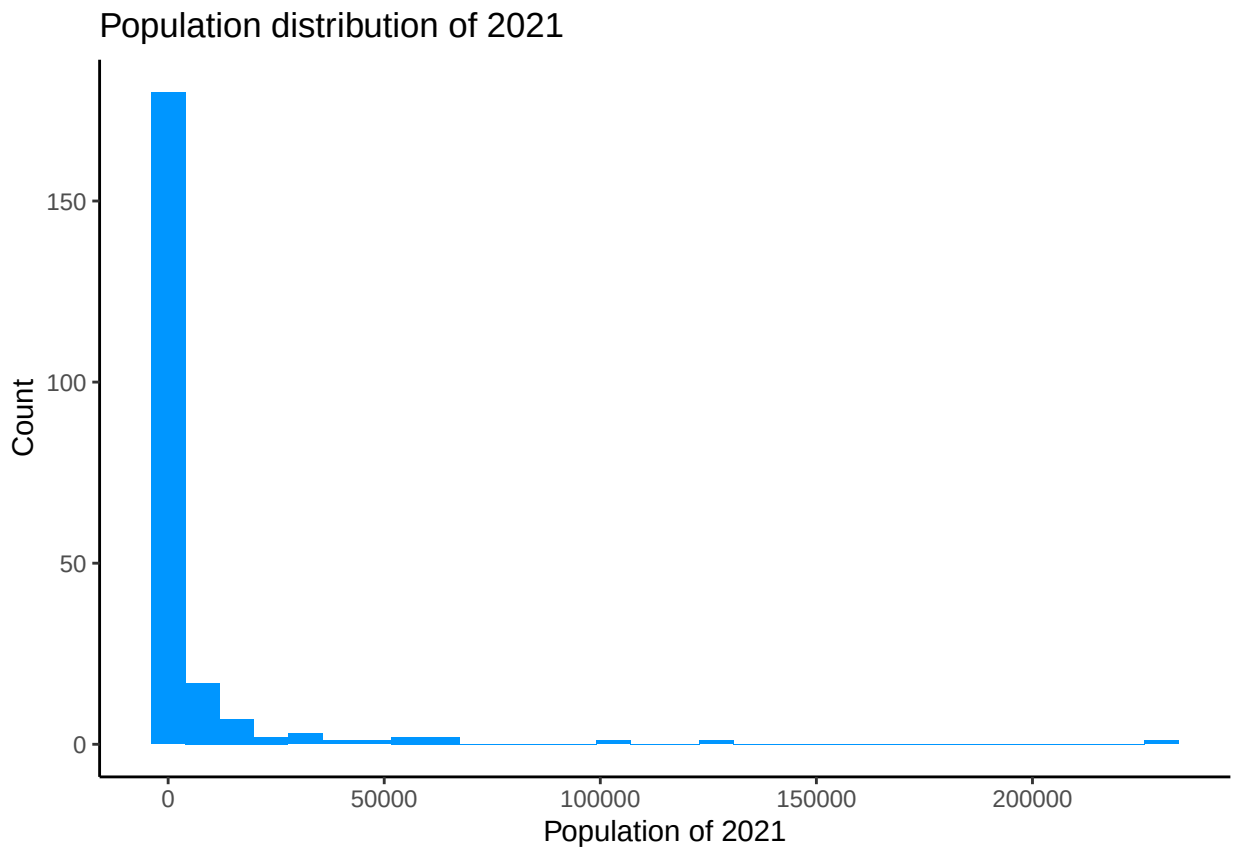
1. Print out just the row with Caldwell to confirm that you edited the data correctly

```
##       name pop2021 pop2010
## 5 Caldwell   62487   46535
```

2. Visualize the 2021 data alone (One quantitative variable)

Visualize the distribution of the pop2021 data.

```
ggplot(data = icd, aes(x=pop2021)) +
  geom_histogram( fill = "#0096FF") +
  labs(title="Population distribution of 2021",
       x ="Population of 2021", y = "Count") + theme_classic()
```




```
icd$log_pop2021 <- log(icd$pop2021)
head(icd)
```

3. Transform the pop2021 variable

```
##          name pop2021 pop2010 log_pop2021
## 1       Boise  229993  209576    12.34580
## 2    Meridian  129555   77428    11.77186
## 3       Nampa  105405   81998    11.56557
## 4 Idaho Falls   64618   57995    11.07625
## 5    Caldwell   62487   46535    11.04271
## 6   Pocatello   57947   54335    10.96728
```

a. Describe your transformation - Why did you choose this transformation.

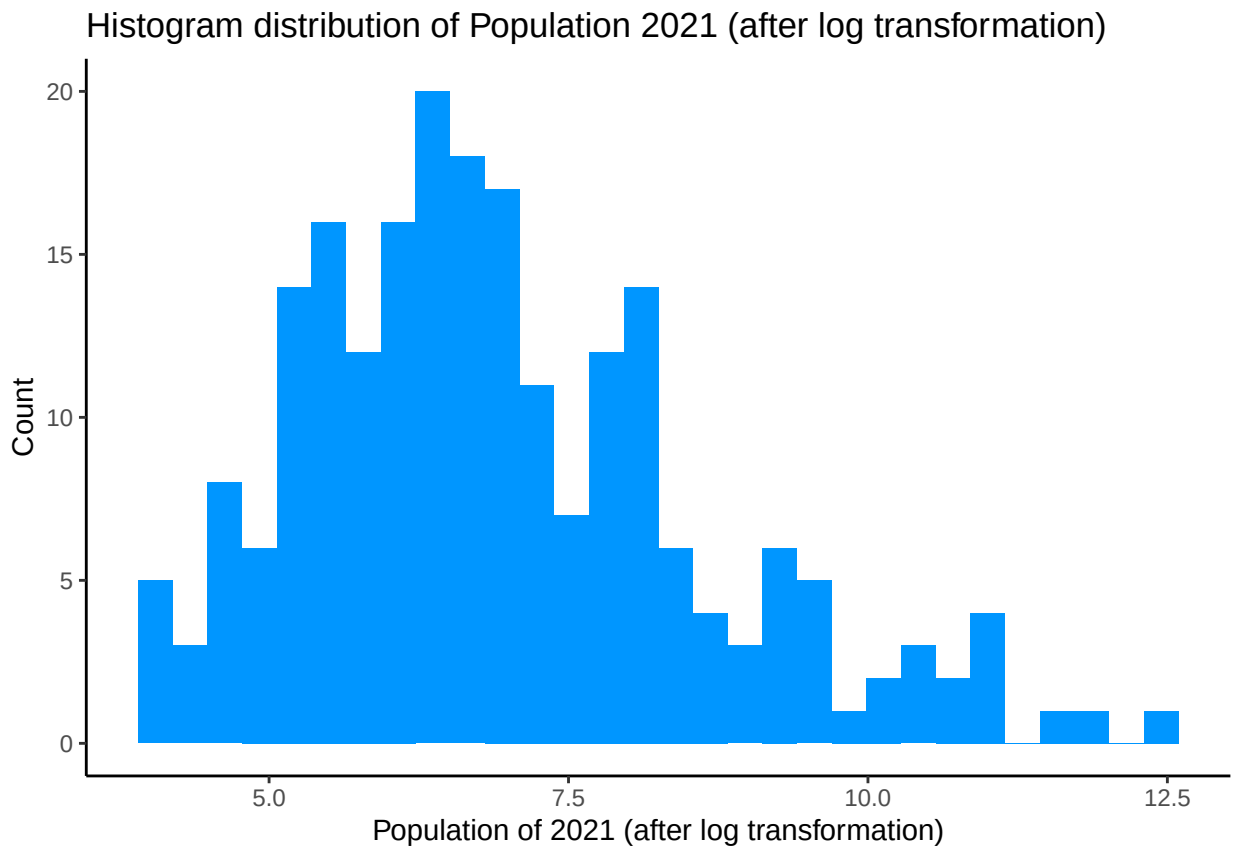
Log transformation is used here to normalize data that exhibits a skewed distribution (as we can see in the previous plot with "pop2021"). Log-transformed data can be more interpretable in certain situations. Log transformation can mitigate the impact of outliers. Extreme values in the original data in "pop2021" can disproportionately affect statistical analyses, and log transformation can help make the effects of outliers more manageable.

b. Visualize the distribution of the transformed pop2021 data.

c. Make sure you have informative and correct axis labels!

ii. Do you think the graph is more or less clear than the original?

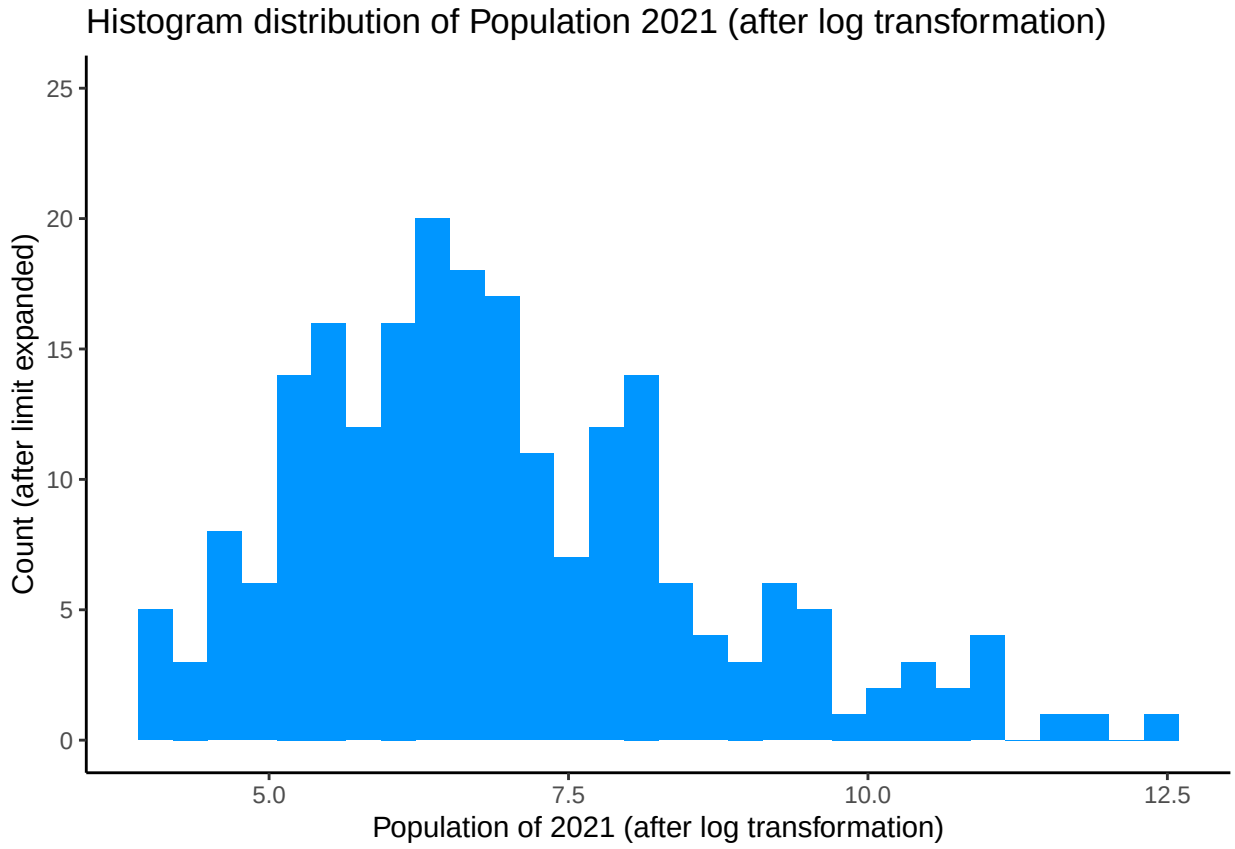
```
ggplot(data = icd, aes(x=log_pop2021)) +
  geom_histogram( fill = "#0096FF") +
  labs(title="Histogram distribution of Population 2021 (after log transformation)",
       x ="Population of 2021 (after log transformation)",
       y = "Count") + theme_classic()
```



It is more clearer as we can see a shape close to a normal distribution.

```
ggplot(data = icd, aes(x=log_pop2021)) +
  geom_histogram( fill = "#0096FF") + ylim(0, 25) +
  labs(title="Histogram distribution of Population 2021 (after log transformation)",
       x ="Population of 2021 (after log transformation)",
       y = "Count (after limit expanded)") + theme_classic()
```

4. Change the range of the y axis in some way (it doesn't necessarily need to be helpful!)



a. Does your new plot accurately portray the data? **Yes, it looks more accurate in the sense of interpretation, as the shape of the curve getting close to a normal distribution.**

b. Does the new plot have the same message as the first plot? **Yes, it is just that the numbers are transformed according to a log base.**

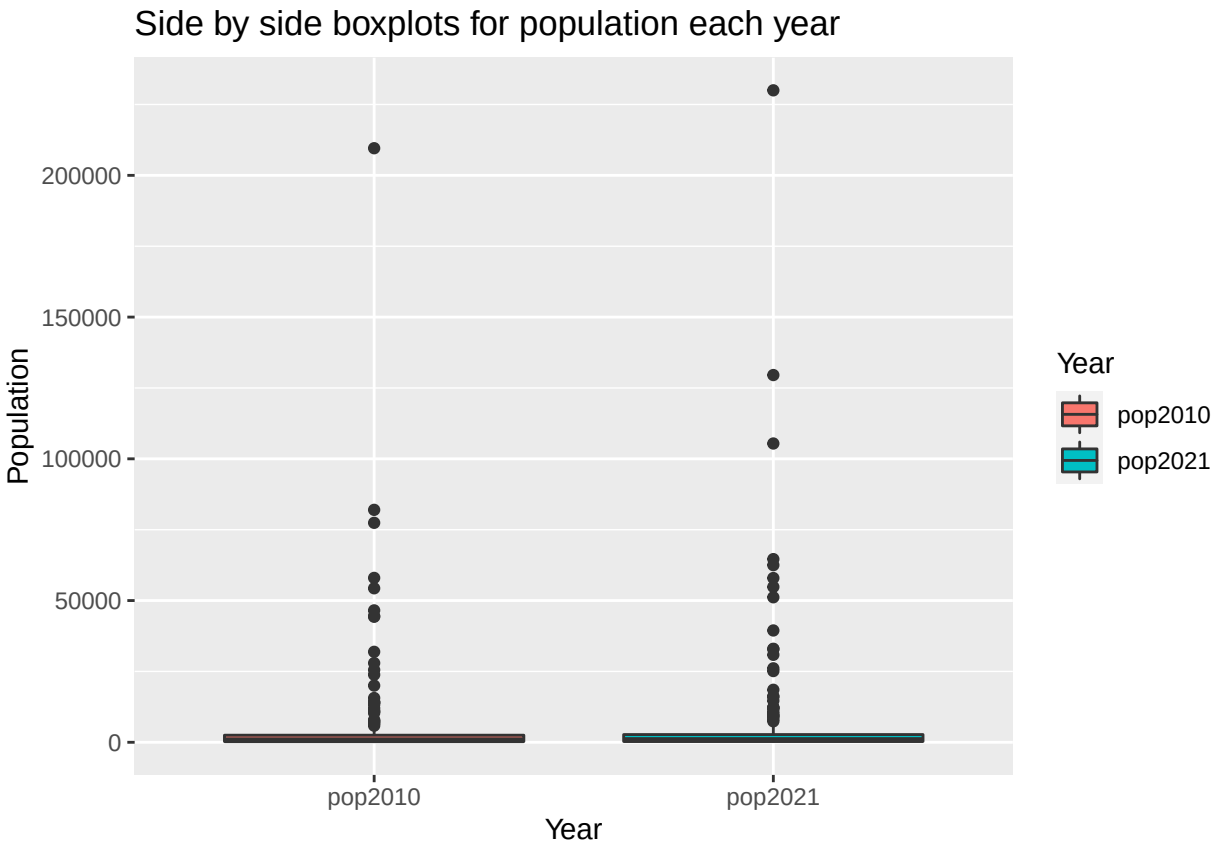
Visualize the 2010 and 2021 data together

5. Create side by side box plots for the populations of Idaho cities in the two different years.

a. First with the original data

```
# making pop2021 and pop2010 to longer format
icd2 <- icd %>% pivot_longer(cols=c('pop2021', 'pop2010'),
                             names_to='Year',
                             values_to='Population')

# plotting side by side box plots
ggplot(icd2, aes(x= Year, y= Population, fill= Year)) +
  geom_boxplot() +
  ggtitle('Side by side boxplots for population each year')
```



b. Then with transformed data

c. Hint: make sure you transform both the 2010 and 2021 data in the same way

ii. Make sure you have informative and correct axis labels!

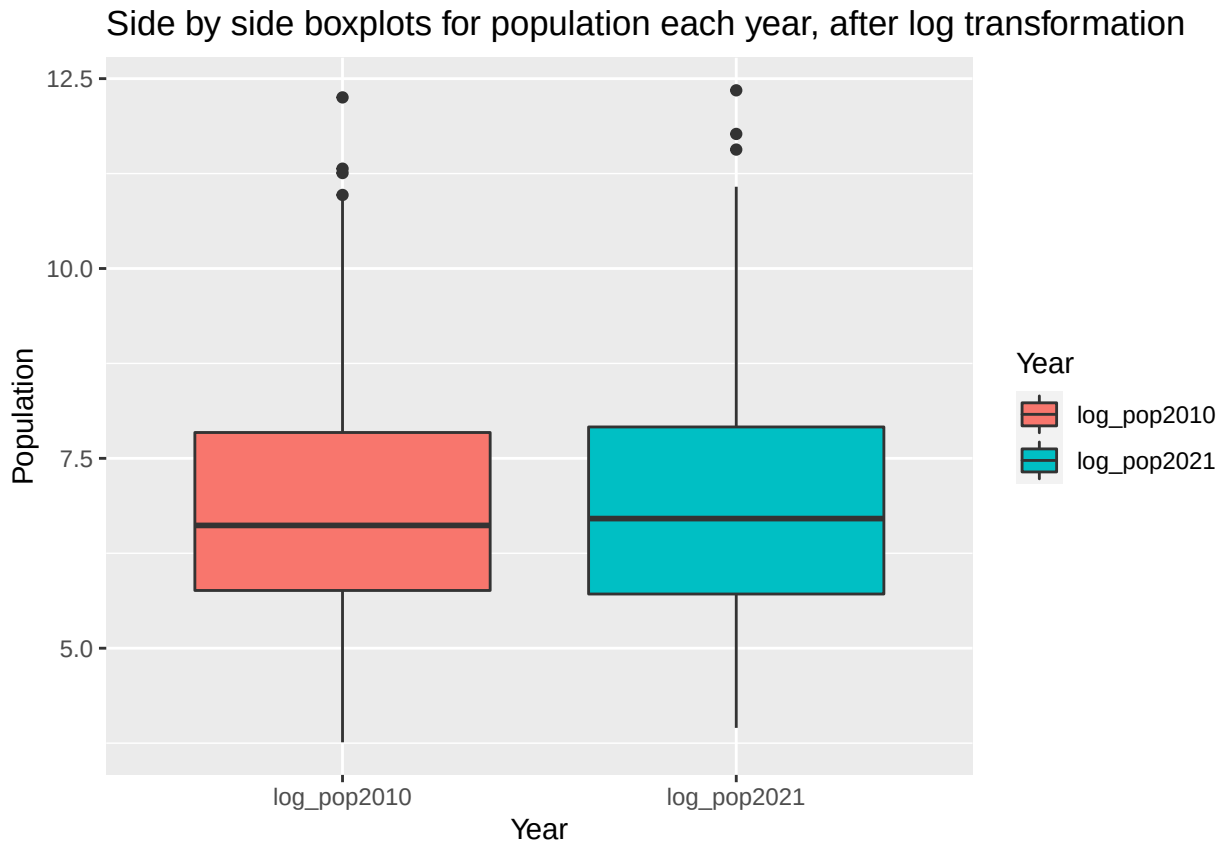
```
# log transforming pop2010
icd$log_pop2010 <- log(icd$pop2010)
head(icd)
```

```
##          name pop2021 pop2010 log_pop2021 log_pop2010
## 1       Boise  229993  209576    12.34580    12.25284
## 2    Meridian  129555   77428    11.77186    11.25710
## 3       Nampa  105405   81998    11.56557    11.31445
## 4 Idaho Falls   64618   57995    11.07625    10.96811
## 5    Caldwell   62487   46535    11.04271    10.74796
## 6   Pocatello   57947   54335    10.96728    10.90292
```

```
# making log_pop2021 and log_pop2010 to longer format
icd3 <- icd %>% pivot_longer(cols=c('log_pop2021', 'log_pop2010'),
                             names_to='Year',
                             values_to='Population')

# plotting side by side box plots
ggplot(icd3, aes(x= Year, y= Population, fill= Year)) +
  geom_boxplot() +
  ggtitle('Side by side boxplots for population each year, after log transformation')
```

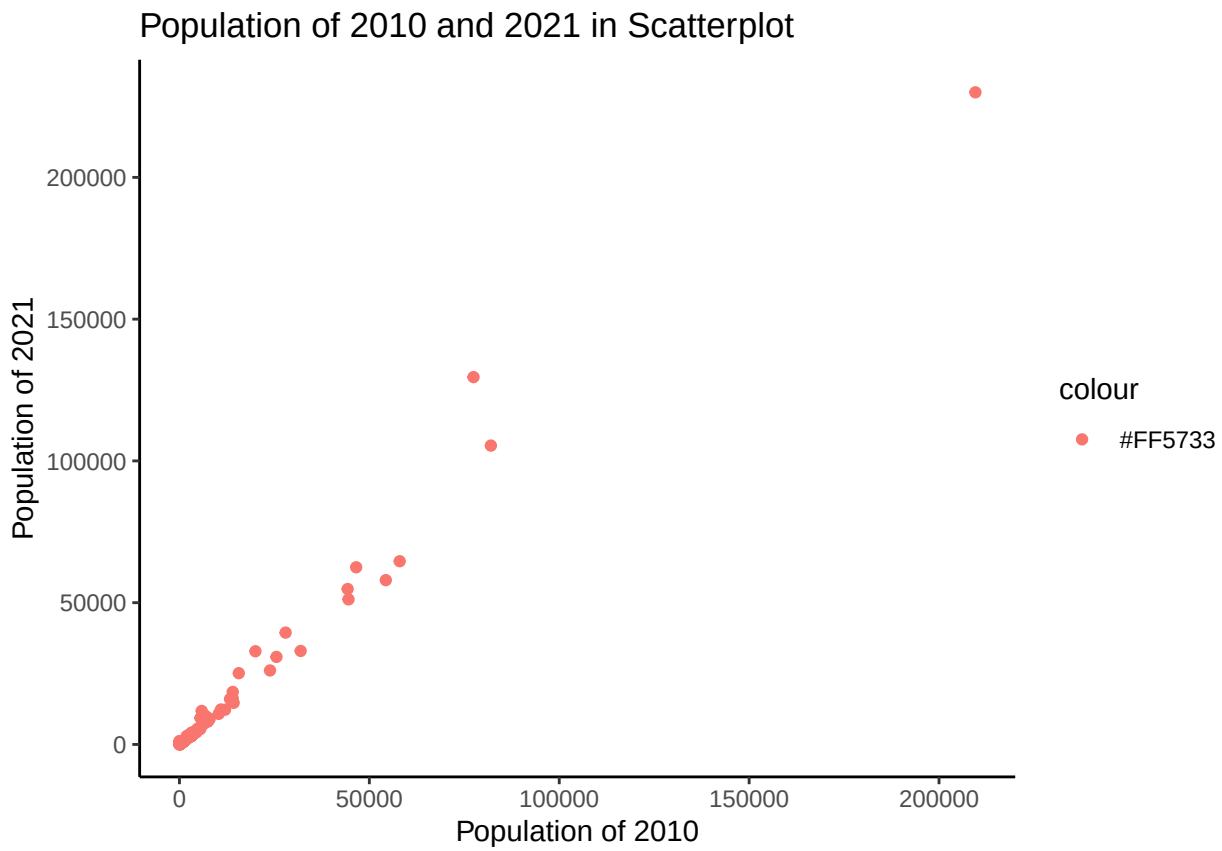
```
## Warning: Removed 6 rows containing non-finite values (stat_boxplot).
```



6. Create a scatter plot (Hint: Scatter plots are best to look at two quantitative variables)

a. First with the original data

```
ggplot(icd, aes(x = pop2010 , y = pop2021)) +
  geom_point(aes(color = "#FF5733"))  +
  labs(title="Population of 2010 and 2021 in Scatterplot",
       x="Population of 2010", y = "Population of 2021") + theme_classic()
```

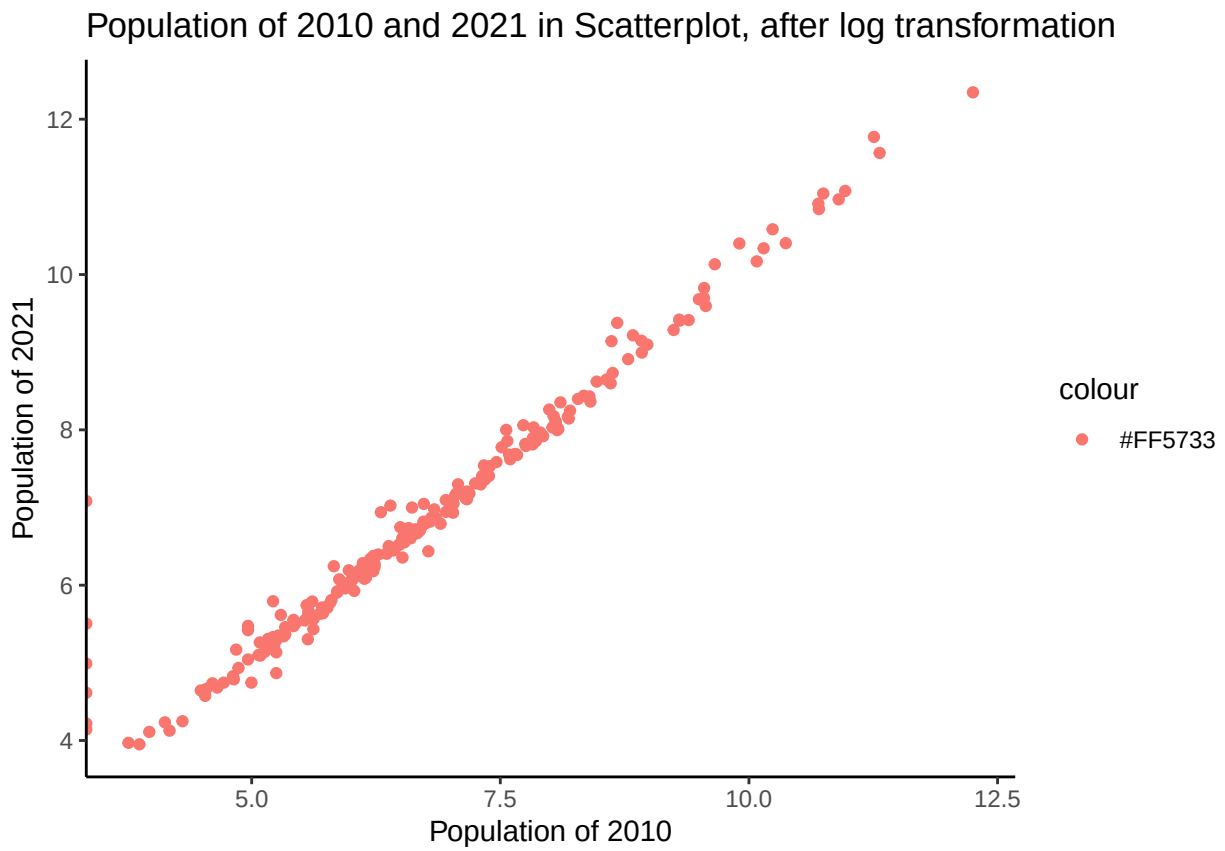


b. Then with transformed data

c. Hint: make sure you transform both the 2010 and 2021 data in the same way

ii. Make sure you have informative and correct axis labels!

```
ggplot(icd, aes(x = log_pop2010 , y = log_pop2021)) +
  geom_point(aes(color = "#FF5733"))  +
  labs(title="Population of 2010 and 2021 in Scatterplot, after log transformation",
       x="Population of 2010", y = "Population of 2021") + theme_classic()
```



7. Create a new variable to represent the change from 2010 to 2021

a. Difference (2021 – 2010)

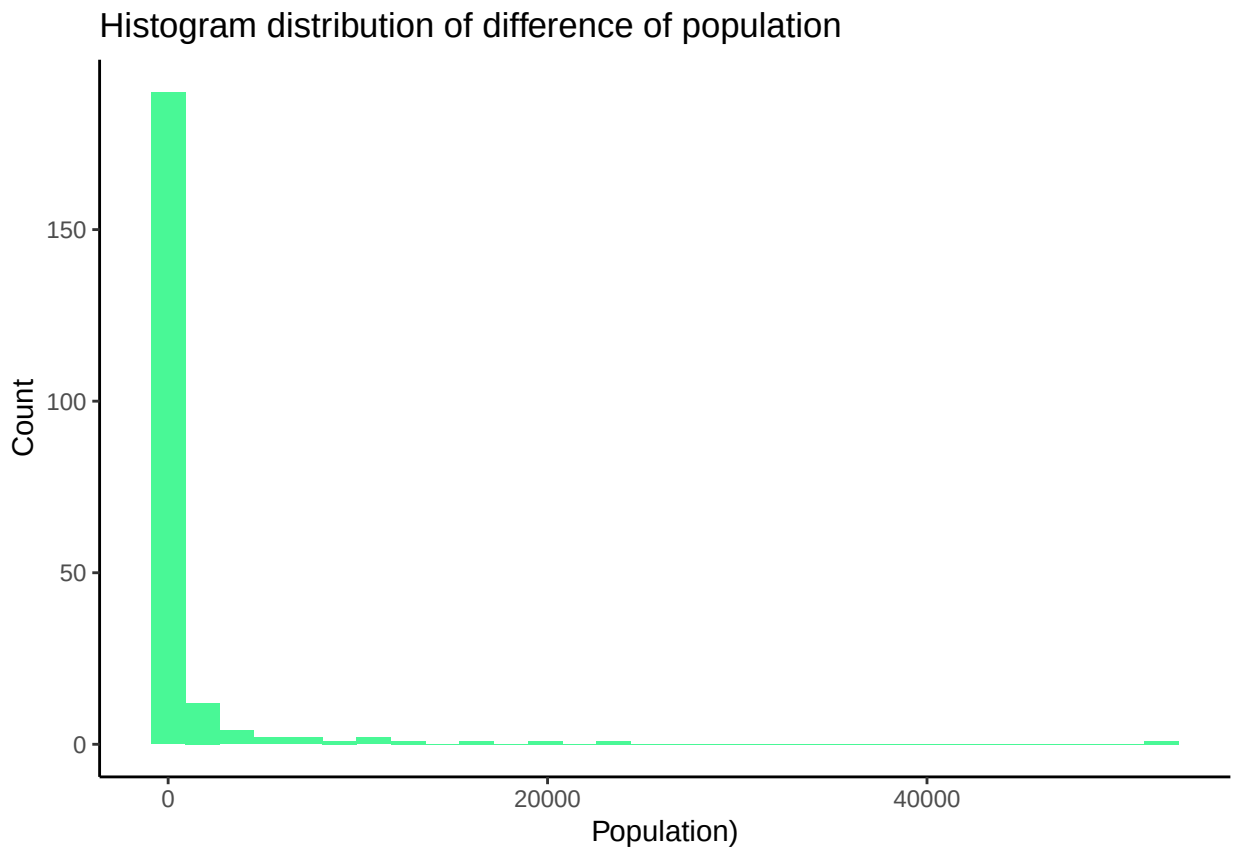
```
icd$Difference <- icd$pop2021 - icd$pop2010
head(icd)
```

```
##          name pop2021 pop2010 log_pop2021 log_pop2010 Difference
## 1       Boise  229993  209576    12.34580    12.25284      20417
## 2    Meridian  129555   77428    11.77186    11.25710      52127
## 3       Nampa  105405   81998    11.56557    11.31445      23407
## 4 Idaho Falls   64618   57995    11.07625    10.96811       6623
## 5    Caldwell   62487   46535    11.04271    10.74796      15952
## 6   Pocatello   57947   54335    10.96728    10.90292       3612
```

i. Visualize the difference with a box plot or histogram

1. Make sure you have informative and correct axis labels!

```
ggplot(data = icd, aes(x=Difference)) +
  geom_histogram( fill = "#49F896") +
  labs(title="Histogram distribution of difference of population",
       x ="Population)",
       y = "Count") + theme_classic()
```



b. Relative difference (2021 – 2010)/2010

```
icd$Relative_Diff <- (icd$pop2021 - icd$pop2010) / icd$pop2010
head(icd)
```

```
##          name pop2021 pop2010 log_pop2021 log_pop2010 Difference Relative_Diff
## 1       Boise  229993  209576    12.34580    12.25284      20417    0.09742051
## 2    Meridian  129555   77428    11.77186    11.25710      52127    0.67323191
## 3       Nampa  105405   81998    11.56557    11.31445      23407    0.28545818
## 4 Idaho Falls   64618   57995    11.07625    10.96811       6623    0.11419950
## 5    Caldwell   62487   46535    11.04271    10.74796      15952    0.34279575
## 6   Pocatello   57947   54335    10.96728    10.90292       3612    0.06647649
```

i. Something odd happened when the relative difference was calculated in R and produced a warning. What does the warning mean?

In the relative difference column, there are some values that goes to infinity (they are represented as "Inf"), which might trouble accurate visualization, as the variable is not 'numeric' anymore, but 'char'.

ii. Visualize the relative difference

1. Make sure you have informative and correct axis labels!

```
# Removing columns with "Inf" values.
icd <- icd[-c(which(icd$Relative_Diff == 'Inf' )), ]

# Now we will check whether "Inf' values exist or not.
icd[c(which(icd$Relative_Diff == "Inf" )), ]
```

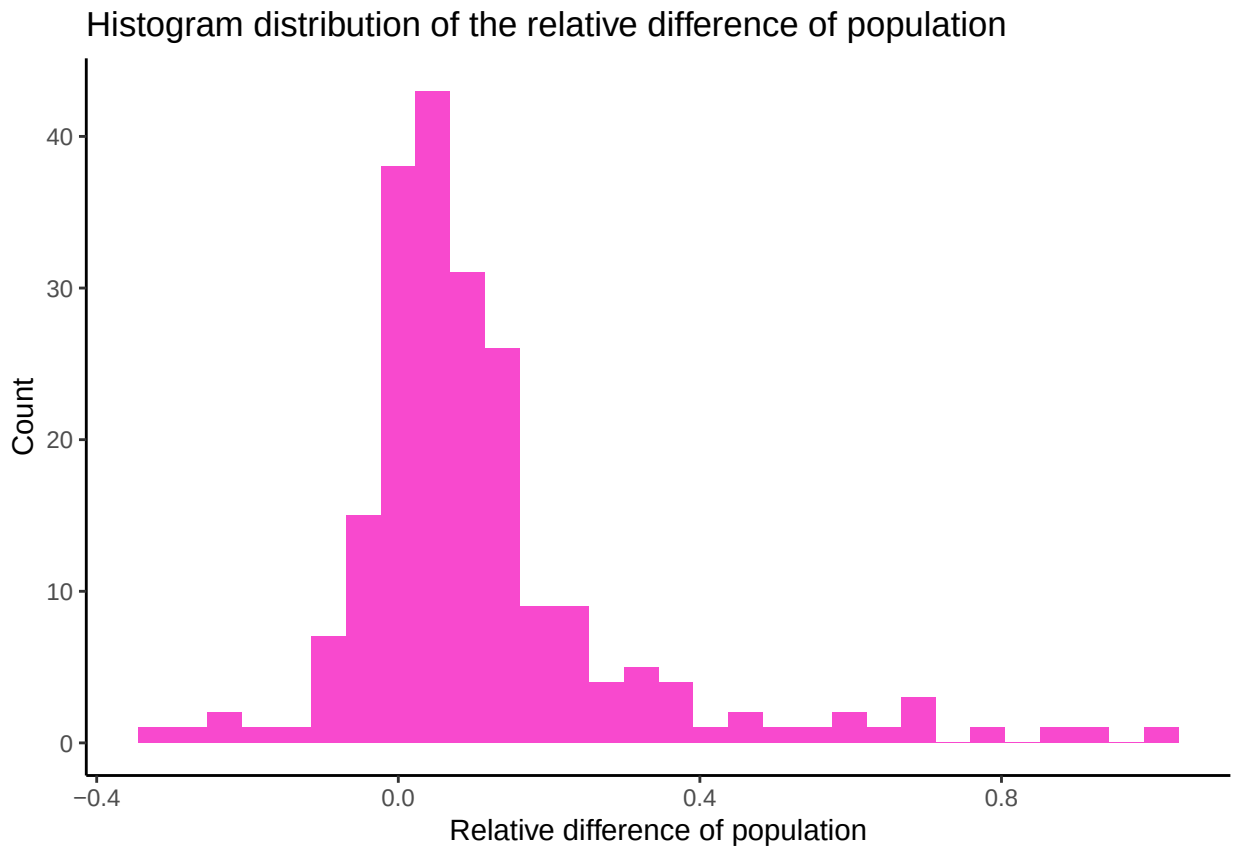
```
## [1] name          pop2021       pop2010       log_pop2021   log_pop2010
## [6] Difference    Relative_Diff
## <0 rows> (or 0-length row.names)
```

```
# Changing the data type of "Relative_Diff" to "numeric".
icd$Relative_Diff <- as.numeric(icd$Relative_Diff)

# Now we will check the data type of "Relative_Diff".
str(icd$Relative_Diff)
```

```
##  num [1:212] 0.0974 0.6732 0.2855 0.1142 0.3428 ...
```

```
# Plotting relative difference histogram
ggplot(data = icd, aes(x = Relative_Diff)) +
  geom_histogram( fill = "#F849CE") +
  labs(title="Histogram distribution of the relative difference of population",
       x ="Relative difference of population",
       y = "Count") + theme_classic()
```



8. Which of the graphs that visualize the change do you believe best represent the data? Relative difference plot

a. Explain why? **Because it is less skewed to right and visualize data meaningfully being bothered almost no outliers, and the shape of the curve getting close to a normal distribution.**