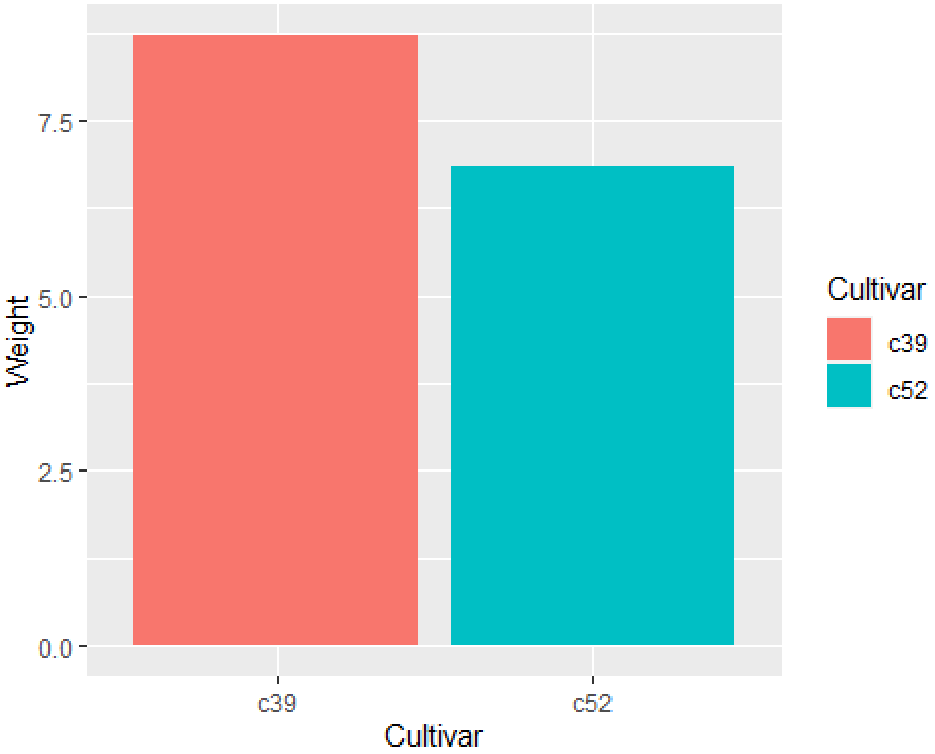
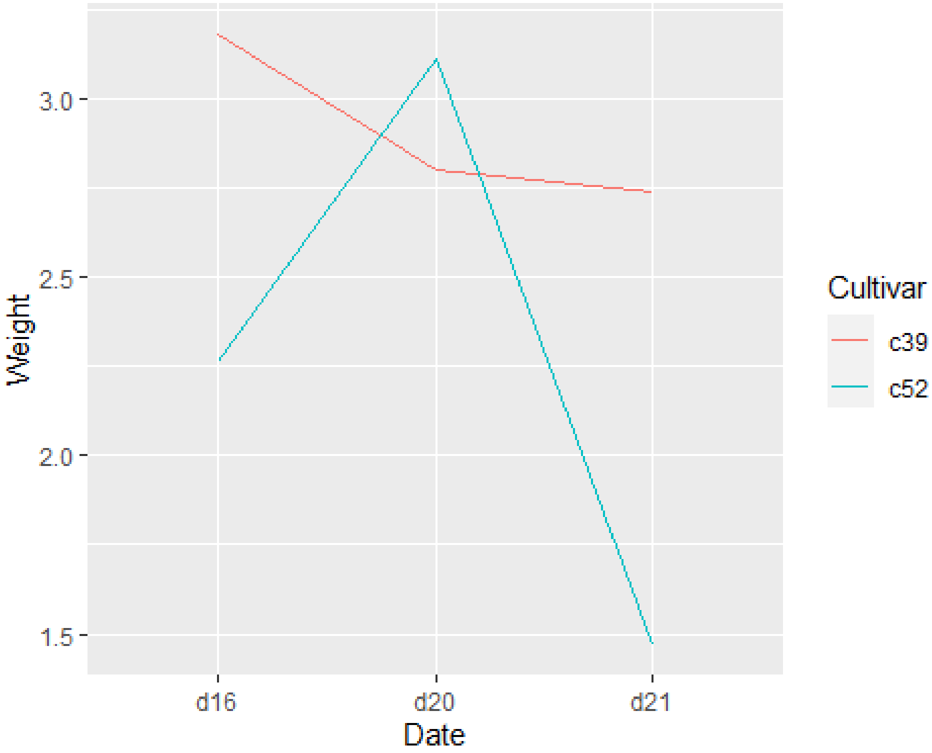
2021 DSO 545 Spring Midterm Exam

Install the following packages: ggplot2, gcookbook, dplyr, dslabs, gridExtra, lme4

1. (5 points) Within Rstudio there is a dataset called **cabbage_exp,** save it as an object called **cabbagedf.** Then create the visualization below with the Cultivar and Weight variables and fill by Cultivar. Generate and replicate the graph below.

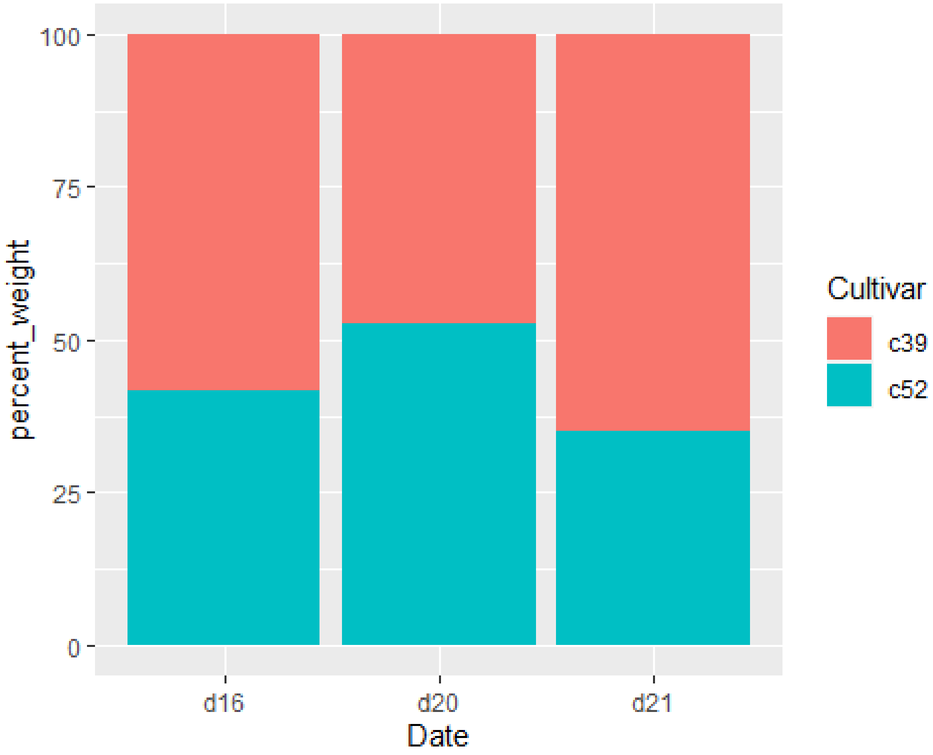


2. (5 points) Now generate the multiple line graph as shown below with the cabbagedf. Generate and replicate the graph below.

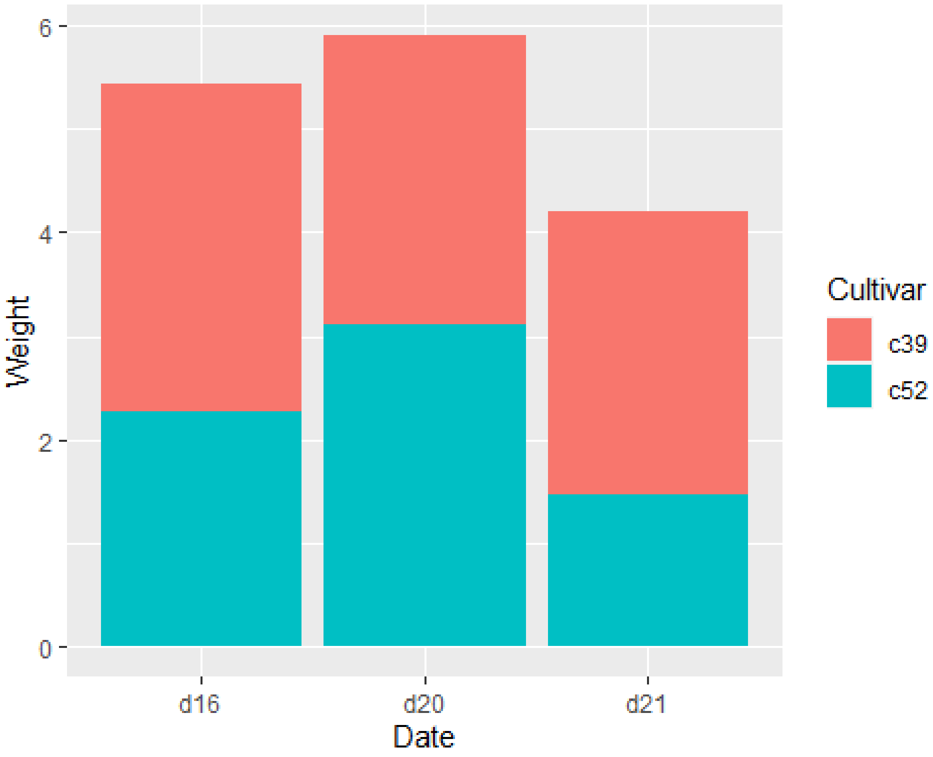


3. (5 points) Now create a new dataframe called **ce** that has cabbbage_exp using dplyr verb and with a new variable called **percent_weight** that entails the calculation of the percent weight grouped by date. Then check to see that the new variable is in the dataframe.

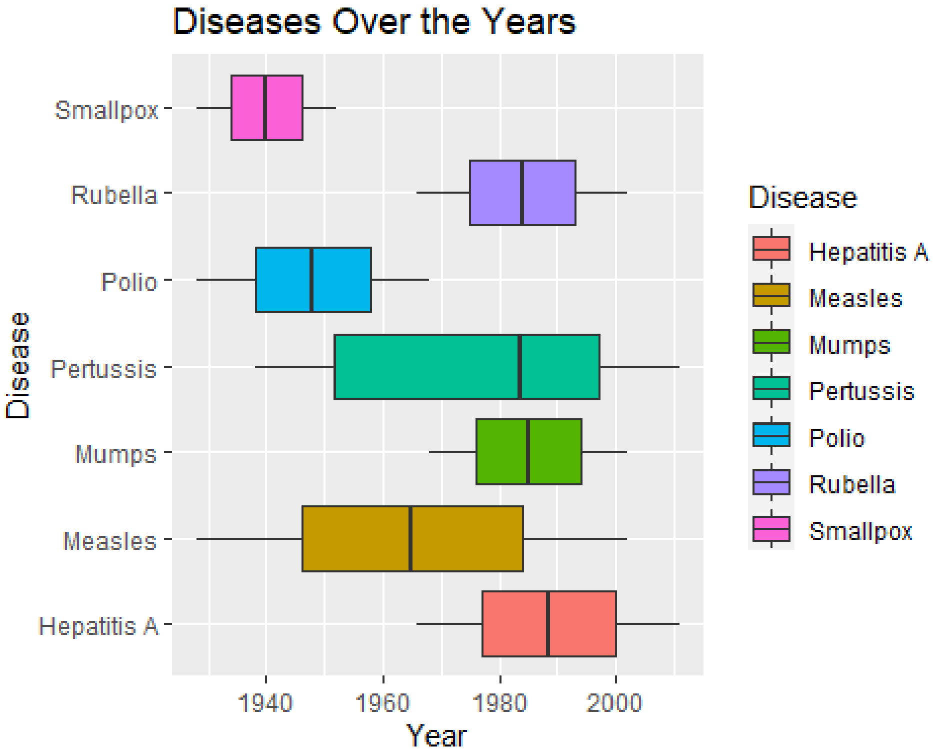
4. (5 points) After computing the new column, make a stacked bar chart with the ce dataframe and the percent_weight variable and fill by Cultivar. Generate and replicate the graph below.



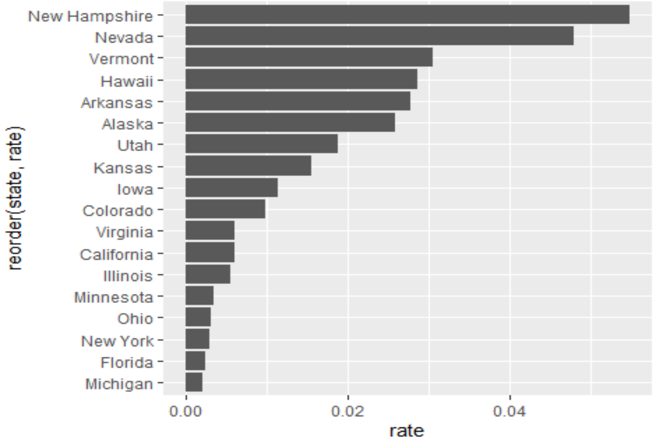
5. (5 points) Make a stacked bar graph that shows proportions of weight per date. Generate and replicate the graph below.



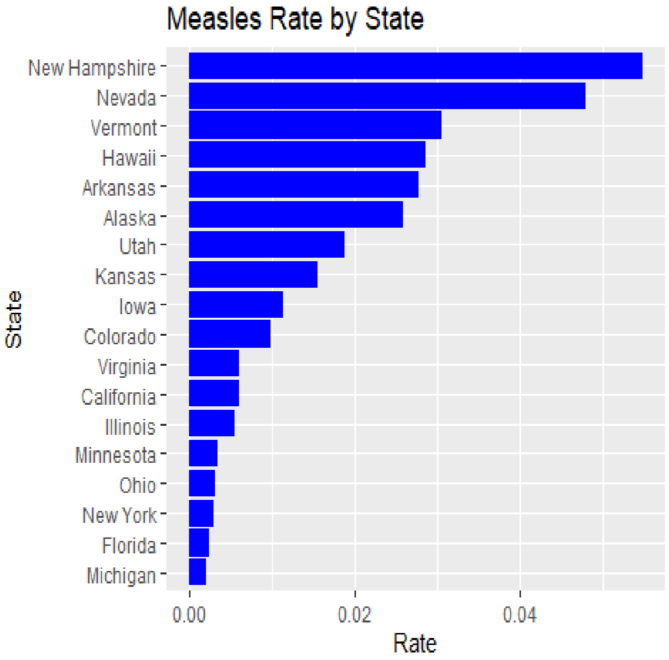
6. (5 points) **New dataset (within R),** now obtain the **us_contagious_diseases** data by typing in us_contagious_diseases and save it in an object called **uscdf**. Generate and replicate the graph shown below that depicts the range of each of contagious disease throughout different periods. In order to make it look like the graph shown, include the following line of code in your ggplot: scale_fill_discrete(name = “Disease”).



7. (5 points) Obtain the information for “Measles” that had a count greater than 0 for the year “2000” and make sure to not include any missing values from the population variable.Then create a variable called rate which will be the result of count/population multiplied by 10000 and multiplied by 52 divided by the weeks_reporting. Create the plot that shows the rate per state and reorder it so that the it shows the state with the highest rates at the top of the graph. Generate and replicate the graph shown below.



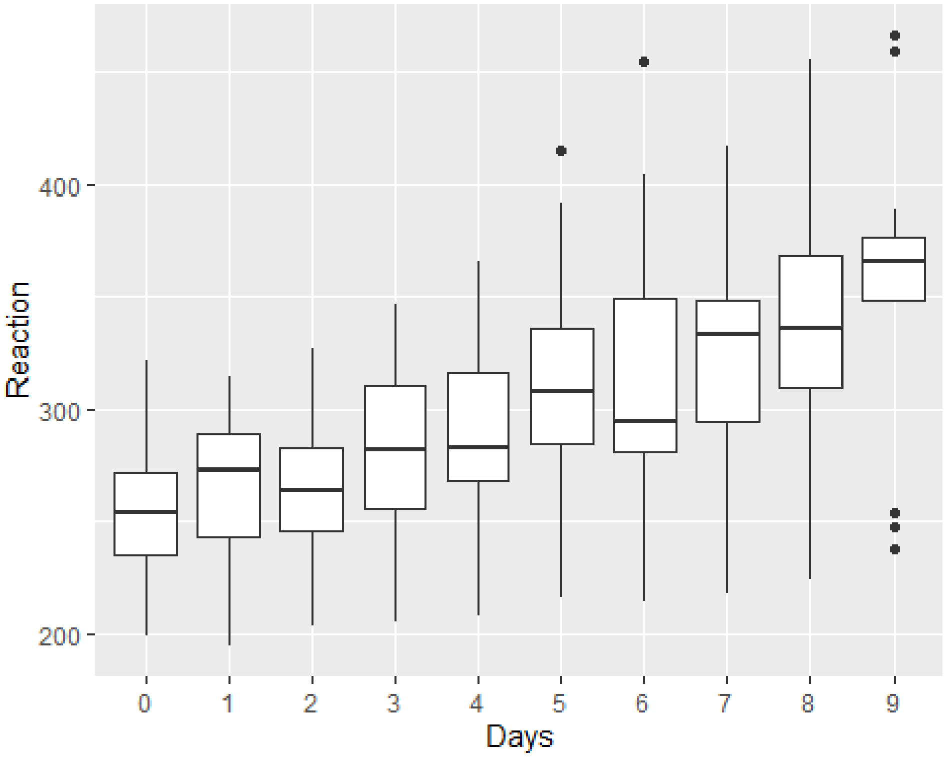
8. (5 points) Copy the code from the previous questions and enhance your plot. Make sure to add the appropriate x axis label and change the y label to “State” and a title “Measles Rate by State” and finally change the color to a color of the graph to a color of your choice. Obtain the information for “Measles” for the year “2000” and make sure to not include any missing values. Generate and replicate the graph shown below.



9. (5 points) **New dataset**, within R use the **iris** dataset. Check to see if the correlations between each combination of variables within the iris dataset excluding the categorical variable Species. Save the **iris** dataset without the Species variable in an object called **numiris** (show your code). Run all the correlations.

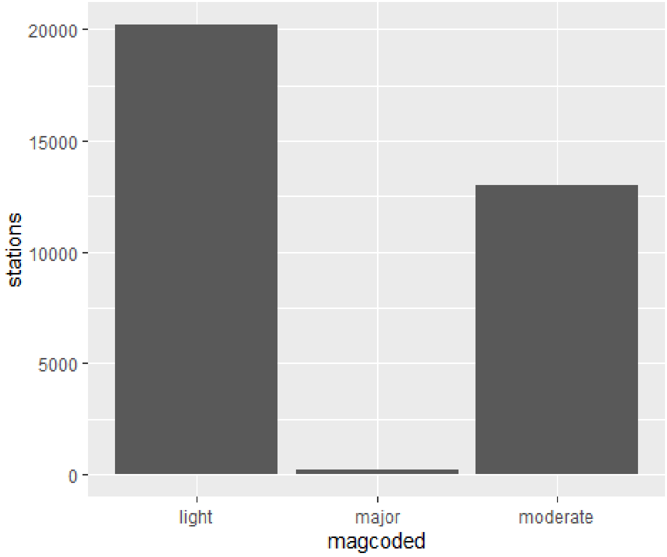
10. (5 points) Which pairing is the has the strongest positive correlation? Then use ggplot to visualize it with the appropriate **geom** type of graph to display the strongest positive correlation.

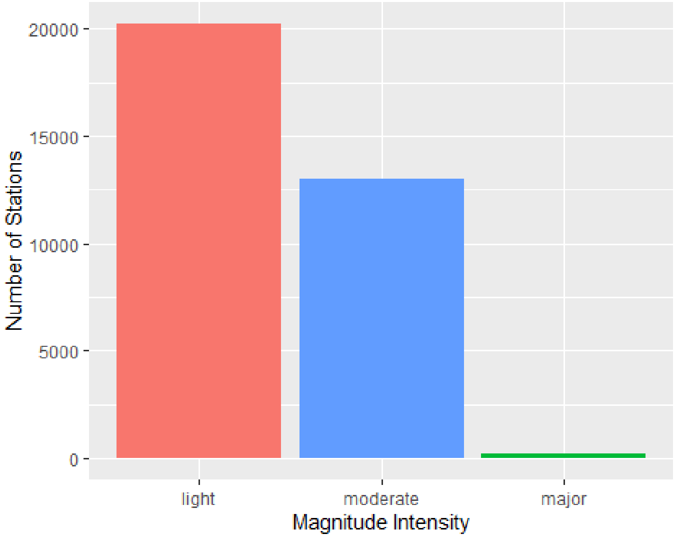
11. (5 points) Within R there is a dataset called **sleepstudy**, go ahead and save it as an object in your global environment as **sleepstudydf.** Then generate and replicate the visualization below.



12. (5 points) Within R there is the **quakes** dataset, save it as an object called **earthquakedf**. Convert the magnitude into the following categories: anything below a 5 = light, anything between 5 and 5.9 = moderate, and anything 6 or above = strong. Save categorized mag as **magcoded**. Check the class and make sure it is a factor variable.

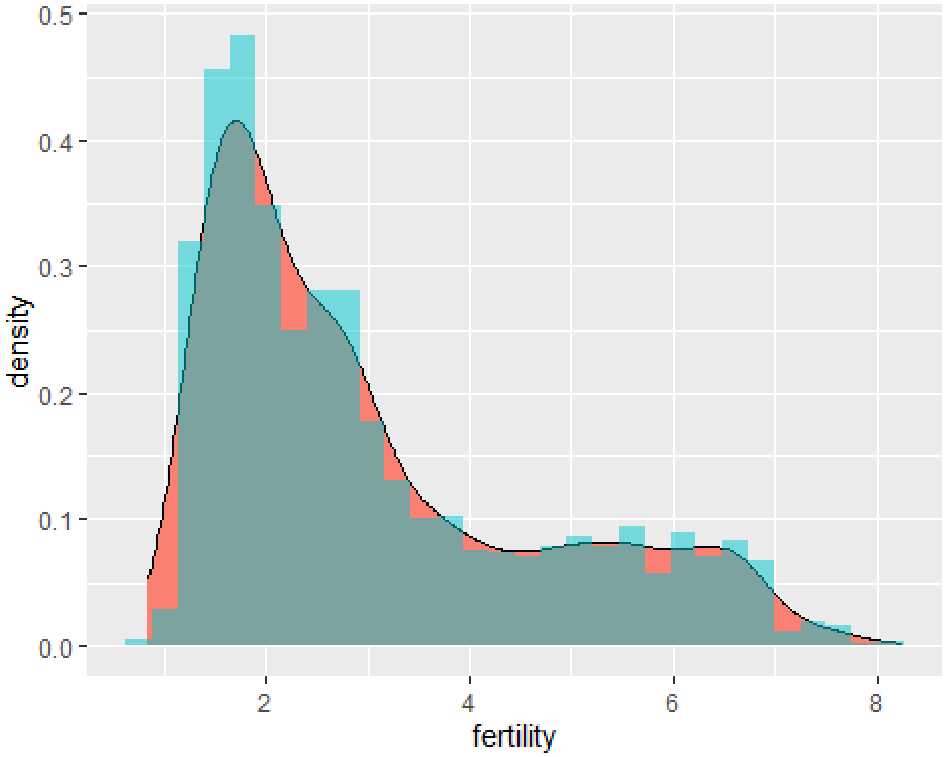
13. (5 points) Then visualize the newly categorized variable of **magcoded** with stations variable, choose the most appropriate graph. Then reorder the visualization so as to generate and replicate the graph below. (to replicate the graph below without the legend, incorporate this line of code at the end of your ggplot: theme(legend.position = "none"))



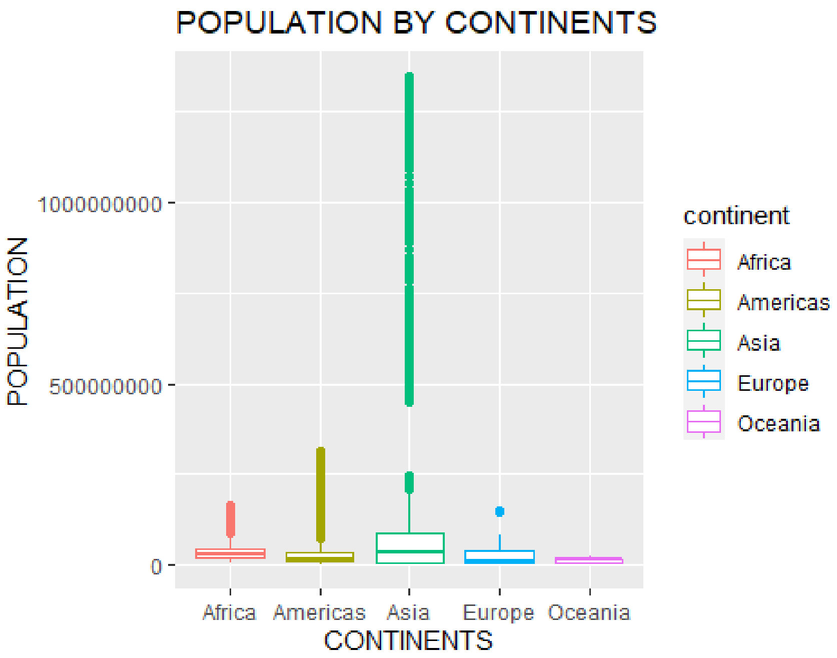


14. (5 points) Now save the **gapminder** dataset as an object called **gadp**. Then create a subset that includes only the following variables: country, year, fertility, population, gdp, continent, region that have a gdp greater than or equal to 10 billion (10000000000).

15. (5 points) Now create a plot that examines the distribution of **fertility** by using a histogram overlayed with a density curve. Generate and replicate the graph below using two colors of your choice to represent the histogram and the density curve.



16. (5 points) Create a boxplot of the population by continent. Then color by continent to *distinguish* between continents. Type options(scipen = 999) prior to running your ggplot code so the Population values don't show up in scientific notation.

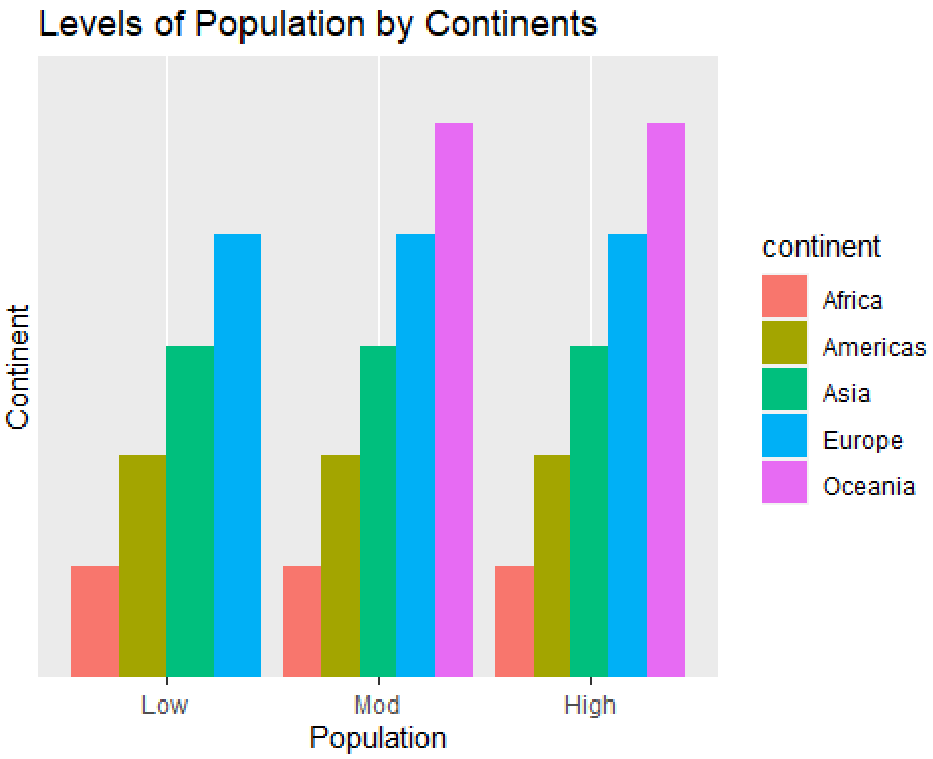


17. (5 points) Using the **piping operator** and appropriate **verb** choose the following variables: country and population. Then specify the **country** to obtain the **mean** of "Thailand" and the **mean** of "Germany" and and save it an object called, **Average_pop** that should have two means.

18. (5 points) Obtain the summary of the gadp population variable and create labels of 'lowpop', 'modpop', and 'highpop' based on anything less than or equal to the 1st quartile as lowpop, anything from lowpop up and including the 3rd quartile as 'modpop' and anything above as 'highpop' (use the summary function to find the quartiles) and save it in an object called **popfactor**. In order to see the value and remove the scientific notation use the function options(scipen = 999).

19. (5 points) Convert the popfactor object to an actual factor variable and have it be an extra variable in the gadp dataset, and finally graph it by country.

20. (5 points) Generate and replicate the graph below.


Levels of Population by Continents
Continent
Low
Mod
High
Population
continent
Africa
Americas
Asia
Europe
Oceania