Data Project: Data Skills Demonstration Project Part I

Shihan Ding 3033649363, Stephen Hwang, 3034721210, Mark Gutierrez, 3032918371, Leanne Bardoles, 3

Dec 3

```
install.packages("reshape2")
```

```
## Installing package into '/usr/local/lib/R/site-library'
## (as 'lib' is unspecified)
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(readr)
library(reshape2)
```

Part I

1. [2 marks] The first part of our PPDAC framework is to identify the problem you are addressing with these data. State the question you are trying to answer and let us know what type of question this is in terms of the PPDAC framework. A question statement should be as specific as possible. For example: Do students who regularly get 8 hours of slepe have fewer visits to the health center? This question is an example of an etiologic or casual question.

The problem our data is addressing is the distribution of drinking water around the world in areas which the United Nations denotes as Sustainable Development Goal (SDG) regions and its relationship with diarrheal disease. The question we are addressing is: How do percentages regarding what portion of the population has access to drinking water differ across SDG regions? And are the distribution related with diarrheal death? This is a descriptive question.

2. [2 marks] Why is this question interesting or important? You can talk here about how existing data/studies suggest this might be important, how the findings might make an impact, how the findings might be used, or why you are personally interested in this question.

Access to clean water is an important issue that can have severe health consequences that can even lead to death. It's important for drinking water to be adequately distributed around the globe, so that we all have the opportunity to live healthy lives. The findings of this study would highlight the parts of the world that need help gaining access to clean drinking water, specifically between urban and rural areas.And we expect to see a poor water accessibility will influence diarrheal death.

3. [2 marks] What is the target population of your project? Why was this target chosen? (i.e., what was your rationale for wanting to answer this question in this specific population?)

The target population for our project are regions that are considered Sustainable Development Goal (SDG) regions. These regions are geographic groups which are categorized as "more" or "less" developed for statistical purposes. We chose this population, because we are trying to extrapolate whether or not there are significant differences in the access to clean drinking water based on the prevalence of safe management practices. As some regions are considered more developed than others, it is important to evaluate potential inequities in distribution of clean drinking water through subsequent research and analysis of the data.

4. [2 marks] What is the sampling frame used to collect the data you are using? It may be helpful here to read any protocol papers, trial registration records, '.Readme' files or documentation that are associated with your dataset. If you have trouble identifying how the records/individuals were sampled, confirm with your supporting GSI that your dataset will be usable for the purposes of the class. Describe why you think this sampling strategy is appropriate for your question. To what group(s) would you feel comfortable generalizing the findings of your study and why?

The sampling frame used is just about every country where there was obtainable data on drinking water. Since 2000, the program has primarily used household surveys to create their estimates, but they also employ census and administrative data to aid their monitoring. This sampling strategy is appropriate for our question because we want to see differences on a regional level. We feel comfortable generalizing our findings for countries and global regions because that is the given granularity of our data, that is as detailed as our data gets.

5. [2 marks] Write a brief description (1-4 sentences) of the source and contents of your dataset. Provide a URL to the original data source if applicable. If not (e.g., the data came from your internship), provide 1-2 sentences saying where the data came from. If you completed a web form to access the data and selected a subset, describe these steps (including any options you selected) and the date you accessed the data.

The data is from the WHO/UNICEF Joint Monitoring Programme for Water Supply, Sanitation and Hygiene. The program has monitored and collected estimates on the quality of countries' drinking water, sanitation, and hygiene since its creation in 1990. Given the vast amount of data available, we chose to focus on household water in 2020 and have cleaned up the dataset to reflect that. URL to original data: https://washdata.org/data/household#!/

6. [1 mark] Write code below to import your data into R. Assign your dataset to an object. Make sure to include and annotate this code in your submission (you can use a # to comment out regular text within code chunks to annotate).

```
# Uploading the data onto R
sani_data <- read_csv("part2_sanitation_and_diarrheal.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   Country = col_character(),
##   Abbreviation = col_character(),
##   Region = col_character(),
##   Year = col_double(),
##   Population_in_thousands = col_double(),
##   percent_urban = col_double(),
##   diarrheal_death_2017 = col_double(),
##   water_at_least_basic = col_double(),
##   water_limited = col_double(),
##   water_unimproved = col_double(),
##   water_surface_water = col_double()
## )
```

7. [3 marks] Write code in R (included in your submission with annotation) to answer the following questions:

i) What are the dimensions of the dataset?

```
# Running dim() to find the dimensions in the dataset
dim(sani_data)
```

```
## [1] 211  11
```

```
# The dataset has 11 columns (variables) and 234 rows (observations, or countries in this
# dataset).
```

ii) What are the variable names of the variables in your dataset?

```
# Running names() to find all variable names in the dataset
names(sani_data)
```

```
##  [1] "Country"                "Abbreviation"
##  [3] "Region"                 "Year"
##  [5] "Population_in_thousands" "percent_urban"
##  [7] "diarrheal_death_2017"   "water_at_least_basic"
##  [9] "water_limited"          "water_unimproved"
## [11] "water_surface_water"
```

The dataset contains: **"country"**: Names of countries. **"abbreviation"**: Abbreviations for country names. **"region"**: The global region each country belongs to. **"year"**: The year from which sanitation data was collected. **"pop_in_thousands"**: The population of that country in thousands **"percent_urban"**: The percentage of land in that country considered "urban". **"diarrheal_death_2017"**: the percentage death due to diarrheal in 2017 **"water_at_least_basic"**: % of population utilizing water from at least an improved source, with a collection time less than 30mins with a roundtrip. Includes basic and safely managed sources. **"water_limited"**: % of population utilizing water from an improvedsource, but with a collection time more than 30mins with a roundtrip. **"water_unimproved"**: % of population utilizing water from an unprotected dug well or unprotected spring. **"water_surface_water"**: % of population utilizing water directly from a river, dam, lake, pond, stream, canal, or irrigation canal.

iii) Print the first six rows of the dataset.

```
# Running head() with argument 6 to print the first six rows of the dataset.
head(sani_data, 6)
```

```
## # A tibble: 6 x 11
##   Country    Abbreviation Region           Year Population_in_tho~ percent_urban
##   <chr>      <chr>        <chr>           <dbl>              <dbl>         <dbl>
## 1 Afghanist~ AFG          Central and So~  2020              38928            26
## 2 Albania    ALB          Northern Ameri~  2020               2878            62
## 3 Algeria    DZA          Western Asia a~  2020              43851            74
## 4 American ~ ASM          Oceania          2020                 55            87
## 5 Andorra    AND          Northern Ameri~  2020                 77            88
## 6 Angola     AGO          Sub-Saharan Af~  2020              32866            67
## # ... with 5 more variables: diarrheal_death_2017 <dbl>,
## #   water_at_least_basic <dbl>, water_limited <dbl>, water_unimproved <dbl>,
## #   water_surface_water <dbl>
```

```
# Since the countries are sorted alphabetically, the first six rows of the dataset
# contain data for Afghanistan, Albania, Algeria, American Samoa, Andorra, and Angola.
```

8. [2 marks] Use the data to demonstrate a data visualization skill we have covered during Part I of the course. Choose a visualization relevant to your stated problem. Include your code in your submission. For example, you could visualize the distribution of our outcome with a histogram, or use a bar graph to represent the distribution of your exposure variable.

During the course we learned about bar graphs. We will visualize the distribution of different qualities of sanitation water with bar graphs.

```
# First, we remove the countries with no data
sani_data_clean <- sani_data %>%
  filter(water_at_least_basic != "") %>%
  filter(water_limited != "") %>%
  filter(water_unimproved != "") %>%
  filter(water_surface_water != "") %>%
  filter(diarrheal_death_2017 != "")
```

```
# Then, we organize the data by world region and aggregate each water quality category by summing all t
sani_data_regions <- sani_data_clean %>%
  group_by(Region) %>%
  summarize(mean_waterp_basic = mean(water_at_least_basic), mean_waterp_limited = mean(water_limited), m

head(sani_data_regions, 8)
```

```
## # A tibble: 8 x 5
##   Region    mean_waterp_basic mean_waterp_lim~ mean_waterp_uni~ mean_waterp_sur~
##   <chr>                 <dbl>            <dbl>            <dbl>            <dbl>
## 1 Australi~             100              0                0                0
## 2 Central ~              92.6            1.86             2.93             2.36
## 3 Eastern ~              92.4            1.69             3.88             1.88
## 4 Latin Am~              95.5            0.964            2.46             0.857
## 5 Northern~              99.0            0.535            0.419            0
## 6 Oceania                86.8            2.82             5.91             4.18
## 7 Sub-Saha~              67.6           13.1             13.8              5.43
## 8 Western ~              94.5            3.88             0.917            0.542
```

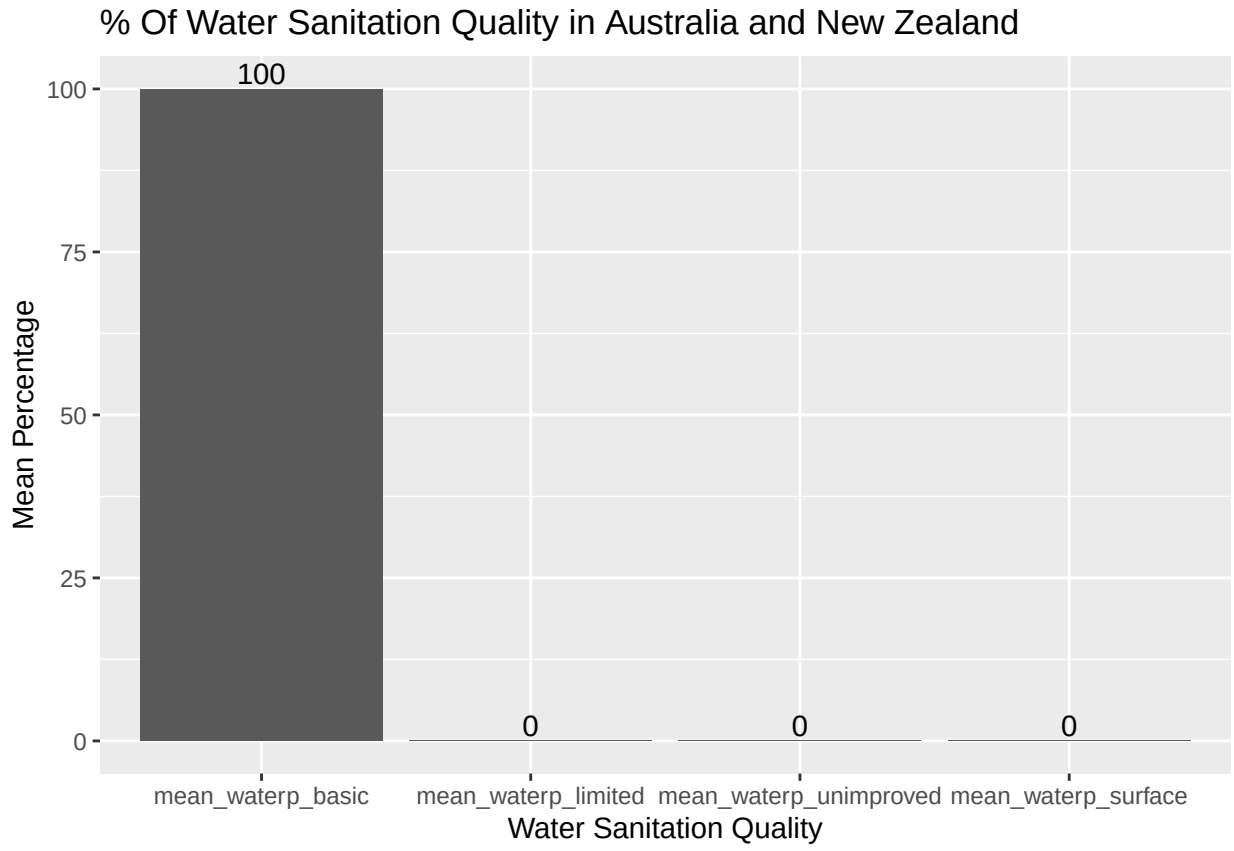
```
# We then use melt from reshape2 to reformat the dataset into a long format.
sani_data_regions_melt <- melt(sani_data_regions, id.vars = "Region")
head(sani_data_regions_melt)
```

```
##                              Region          variable     value
## 1       Australia and New Zealand mean_waterp_basic 100.00000
## 2       Central and Southern Asia mean_waterp_basic  92.57143
## 3  Eastern and South-Eastern Asia mean_waterp_basic  92.37500
## 4 Latin America and the Caribbean mean_waterp_basic  95.50000
## 5     Northern America and Europe mean_waterp_basic  98.97674
## 6                         Oceania mean_waterp_basic  86.81818
```

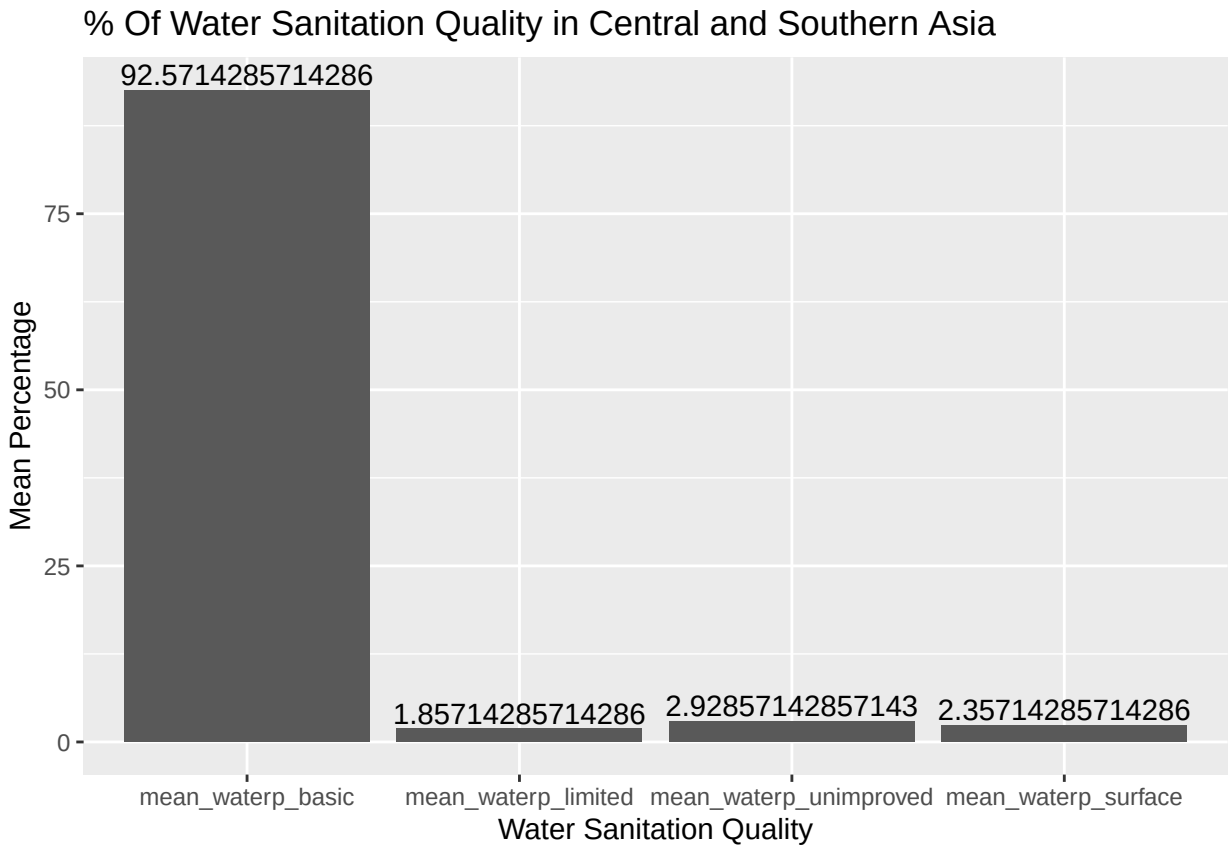
```
# We also separate so each region has their own dataset to draw data from

data_1 = sani_data_regions_melt %>% filter(Region == "Australia and New Zealand")
data_2 = sani_data_regions_melt %>% filter(Region == "Central and Southern Asia")
data_3 = sani_data_regions_melt %>% filter(Region == "Eastern and South-Eastern Asia")
data_4 = sani_data_regions_melt %>% filter(Region == "Latin America and the Caribbean")
```

```
data_5 = sani_data_regions_melt %>% filter(Region == "Northern America and Europe")
data_6 = sani_data_regions_melt %>% filter(Region == "Oceania")
data_7 = sani_data_regions_melt %>% filter(Region == "Sub-Saharan Africa")
data_8 = sani_data_regions_melt %>% filter(Region == "Western Asia and Northern Africa")
```

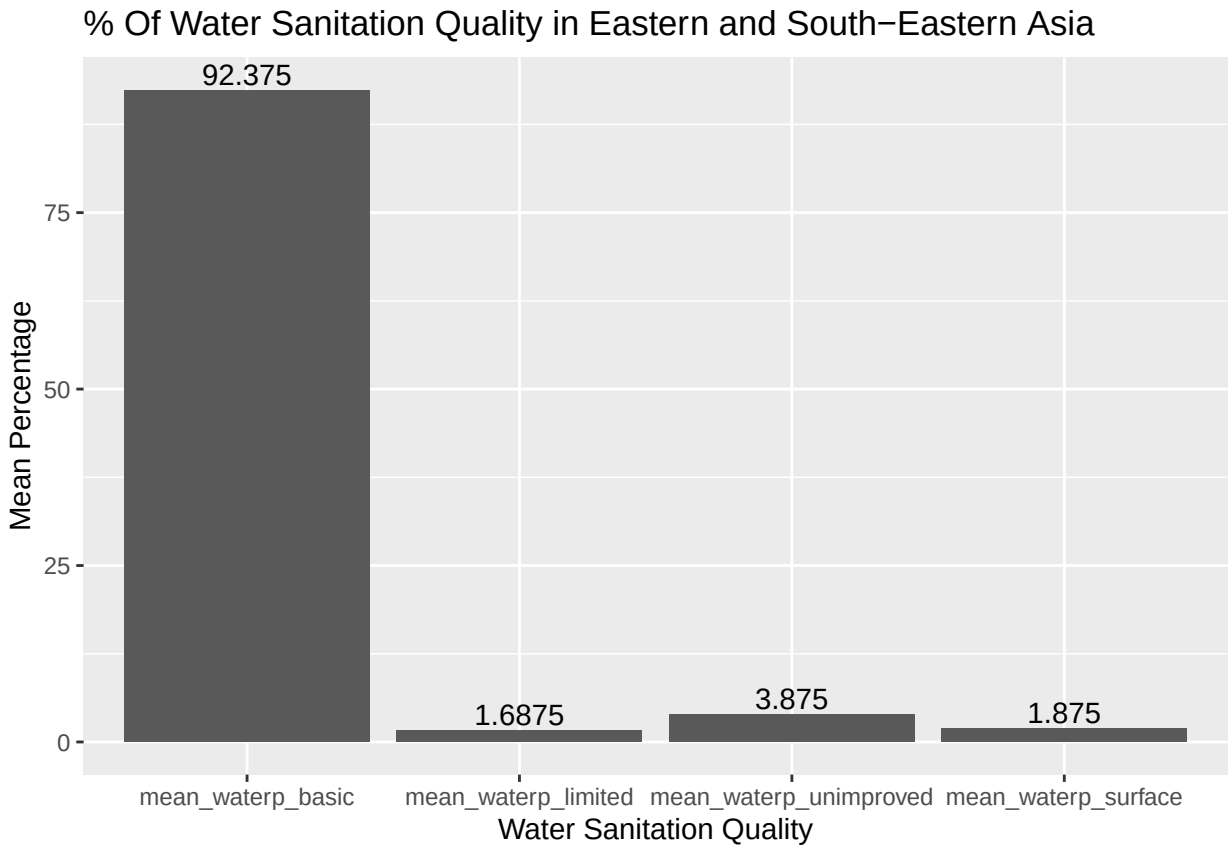
```
# Finally, we use ggplot to plot a bar graph for every region.
library(ggplot2)
aus_and_nz <- ggplot(data_1, aes(x = variable, y = value,)) +
  ggtitle("% Of Water Sanitation Quality in Australia and New Zealand") +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label=value), position=position_dodge(width=0.9), vjust=-0.25) +
  xlab("Water Sanitation Quality") +
  ylab("Mean Percentage")
aus_and_nz
```



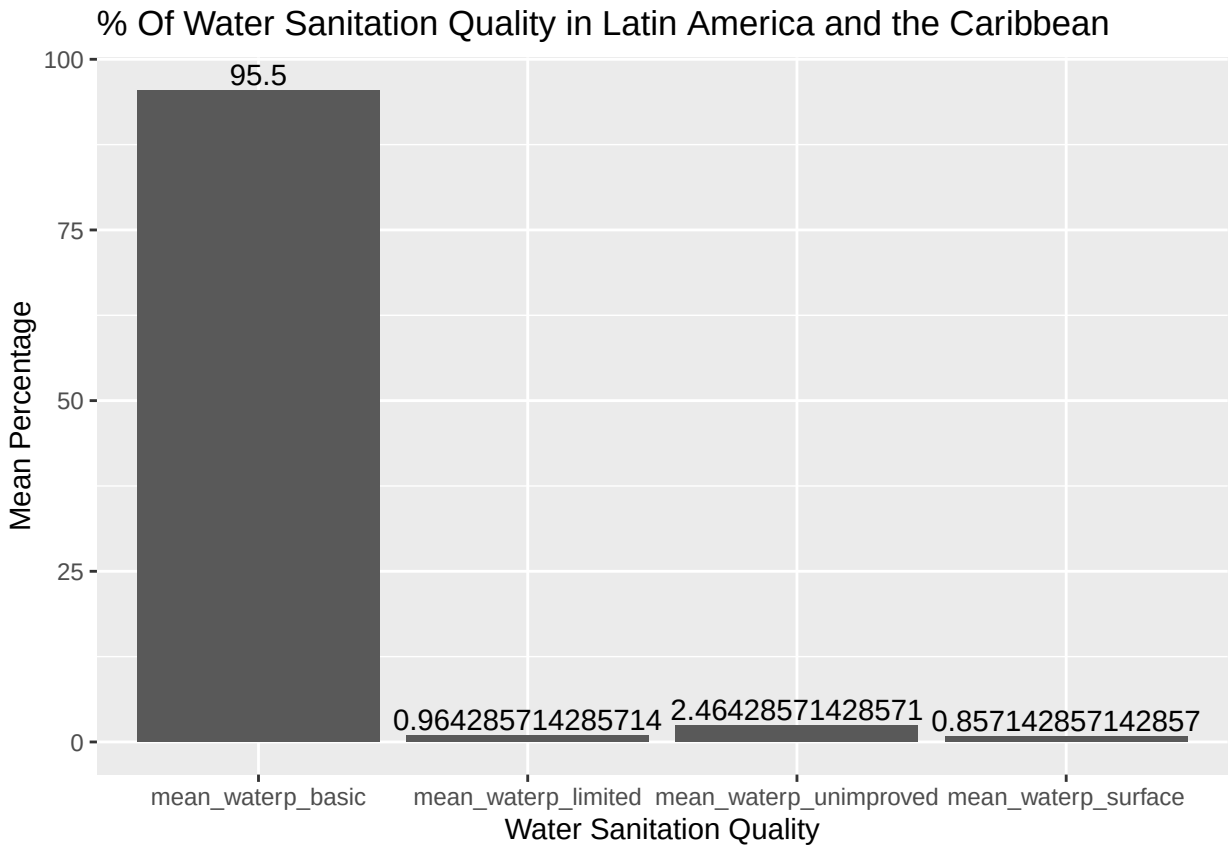

```
cen_and_south_asia <- ggplot(data_2, aes(x = variable, y = value,)) +
  ggtitle("% Of Water Sanitation Quality in Central and Southern Asia") +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label=value), position=position_dodge(width=0.9), vjust=-0.25) +
  xlab("Water Sanitation Quality") +
  ylab("Mean Percentage")
cen_and_south_asia
```



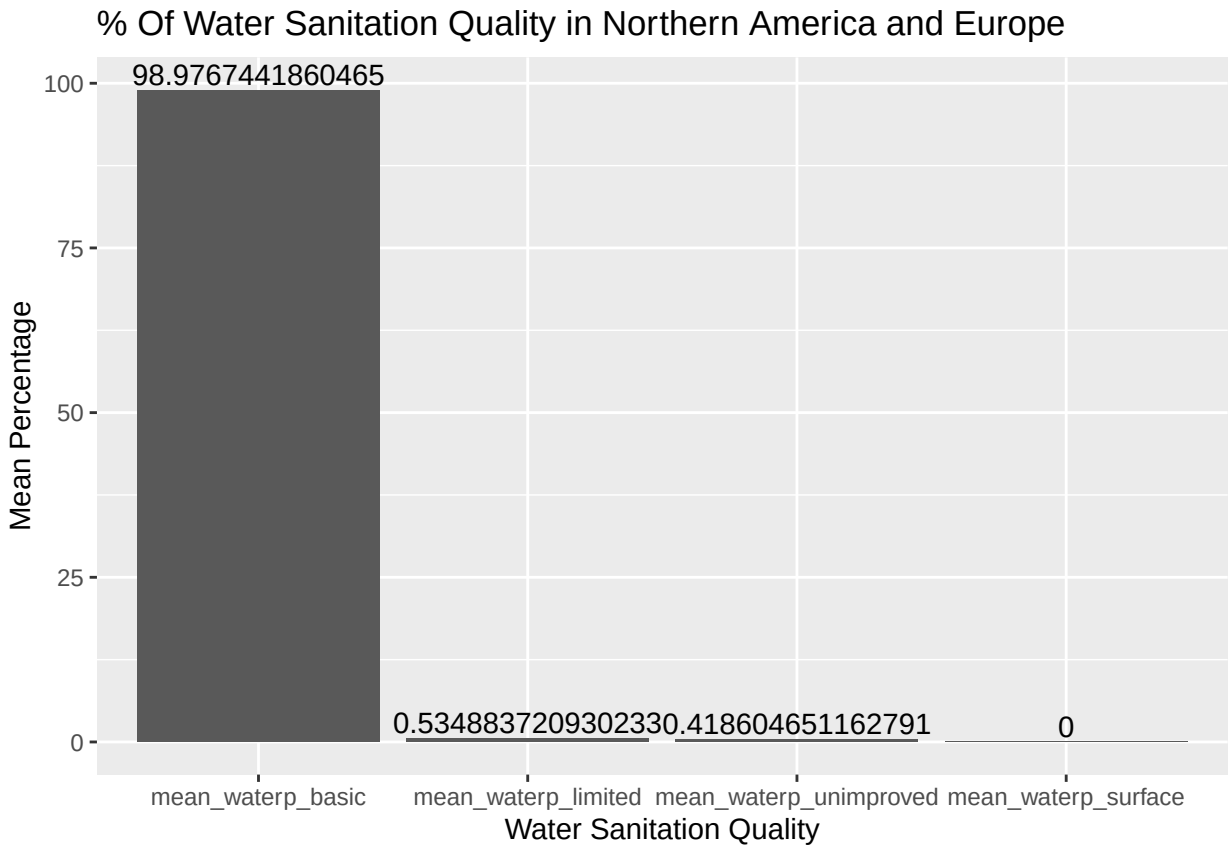

```
east_and_se_asia <- ggplot(data_3, aes(x = variable, y = value,)) +
  ggtitle("% Of Water Sanitation Quality in Eastern and South-Eastern Asia") +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label=value), position=position_dodge(width=0.9), vjust=-0.25) +
  xlab("Water Sanitation Quality") +
  ylab("Mean Percentage")
east_and_se_asia
```



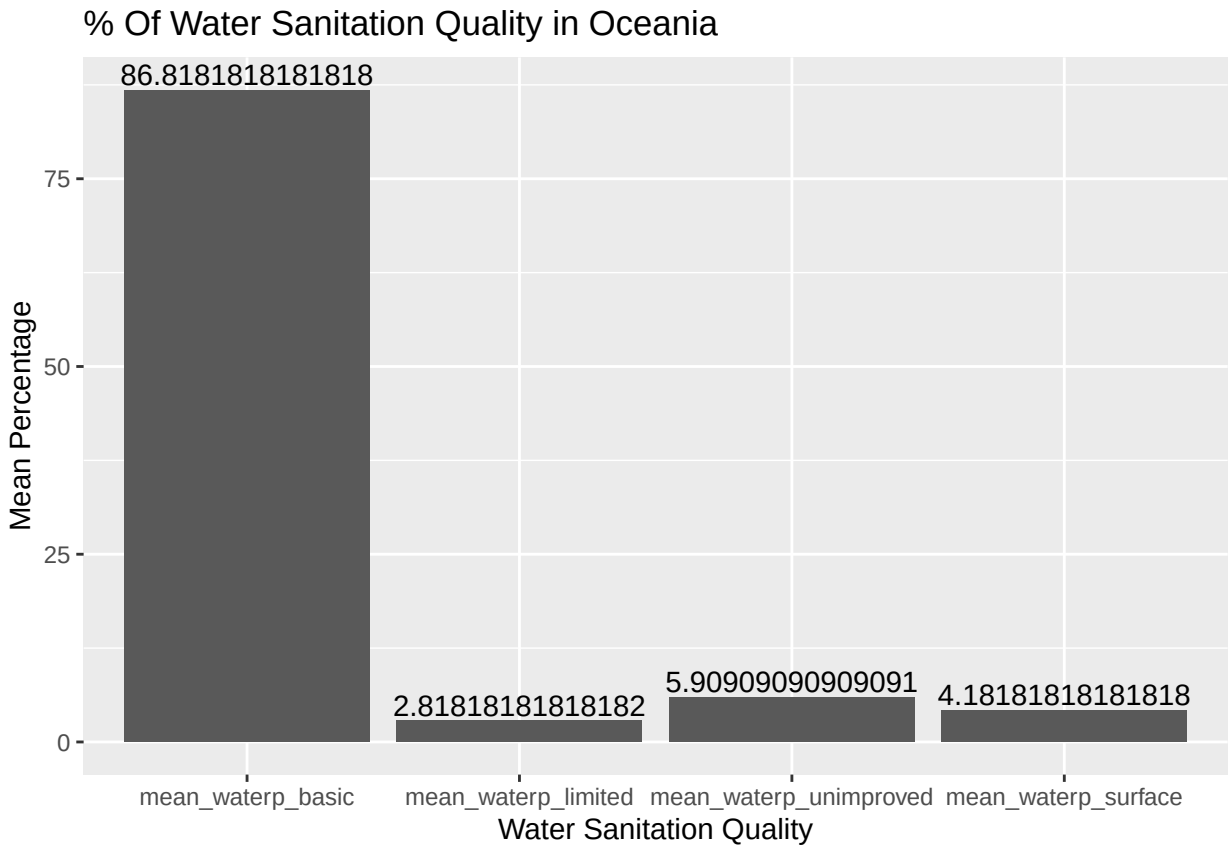

```
latin_am_and_cari <- ggplot(data_4, aes(x = variable, y = value,)) +
  ggtitle("% Of Water Sanitation Quality in Latin America and the Caribbean") +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label=value), position=position_dodge(width=0.9), vjust=-0.25) +
  xlab("Water Sanitation Quality") +
  ylab("Mean Percentage")
latin_am_and_cari
```



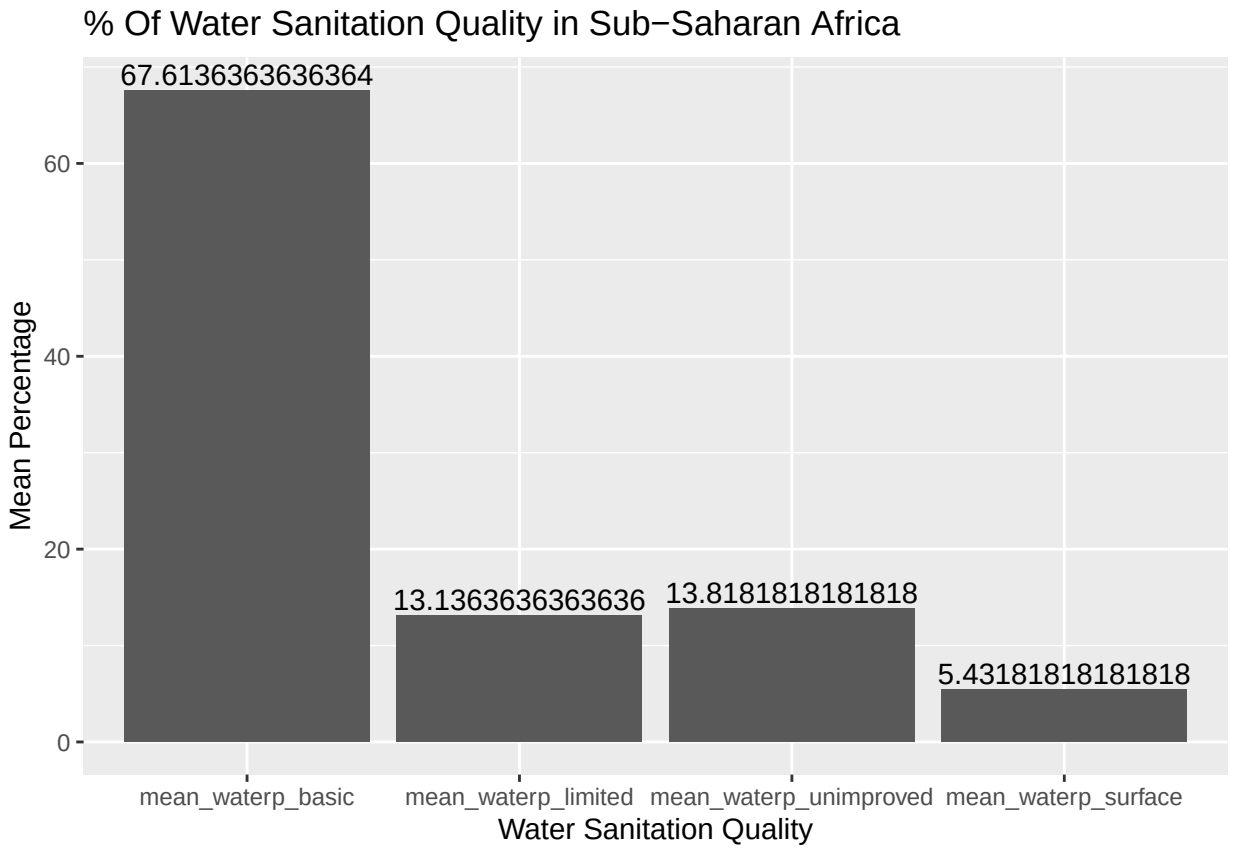

```
na_and_eu <- ggplot(data_5, aes(x = variable, y = value,)) +
  ggtitle("% Of Water Sanitation Quality in Northern America and Europe") +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label=value), position=position_dodge(width=0.9), vjust=-0.25) +
  xlab("Water Sanitation Quality") +
  ylab("Mean Percentage")
na_and_eu
```



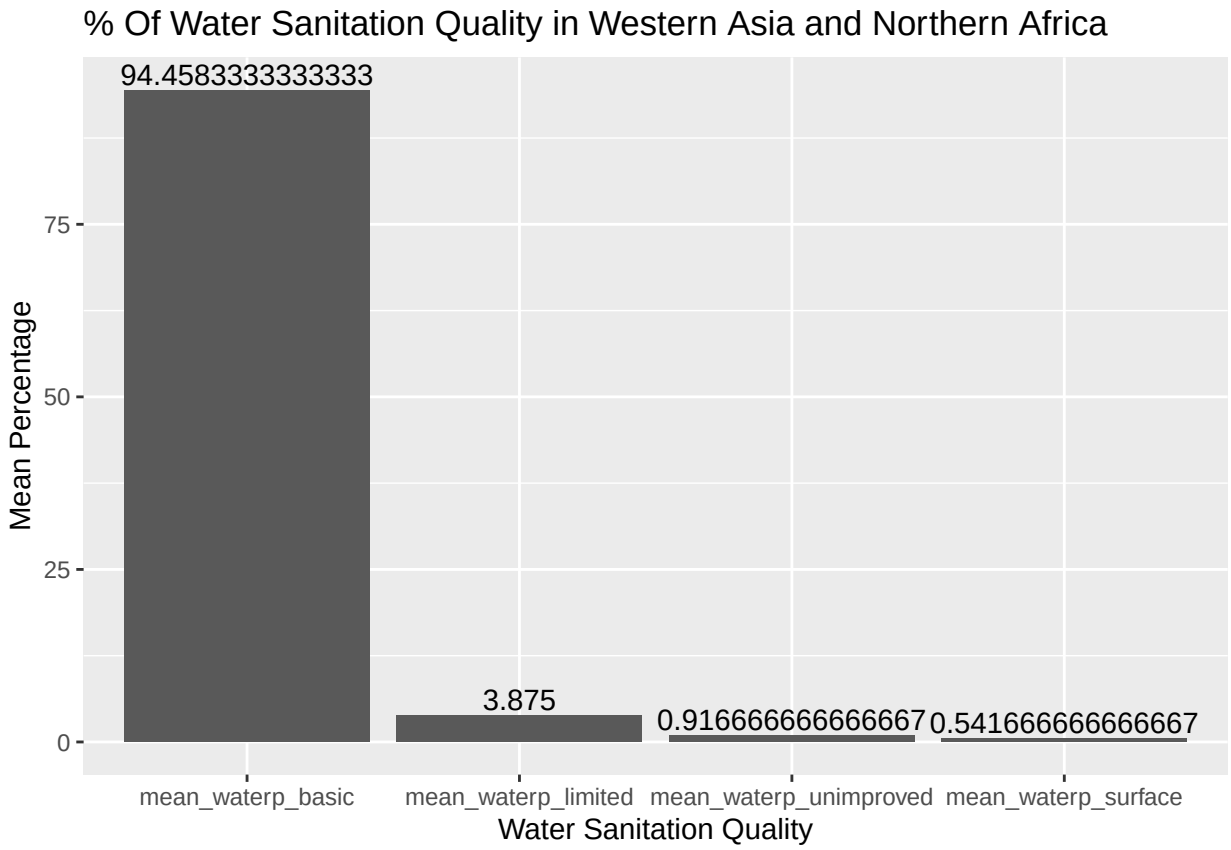

```
oceania <- ggplot(data_6, aes(x = variable, y = value,)) +
  ggtitle("% Of Water Sanitation Quality in Oceania") +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label=value), position=position_dodge(width=0.9), vjust=-0.25) +
  xlab("Water Sanitation Quality") +
  ylab("Mean Percentage")
oceania
```



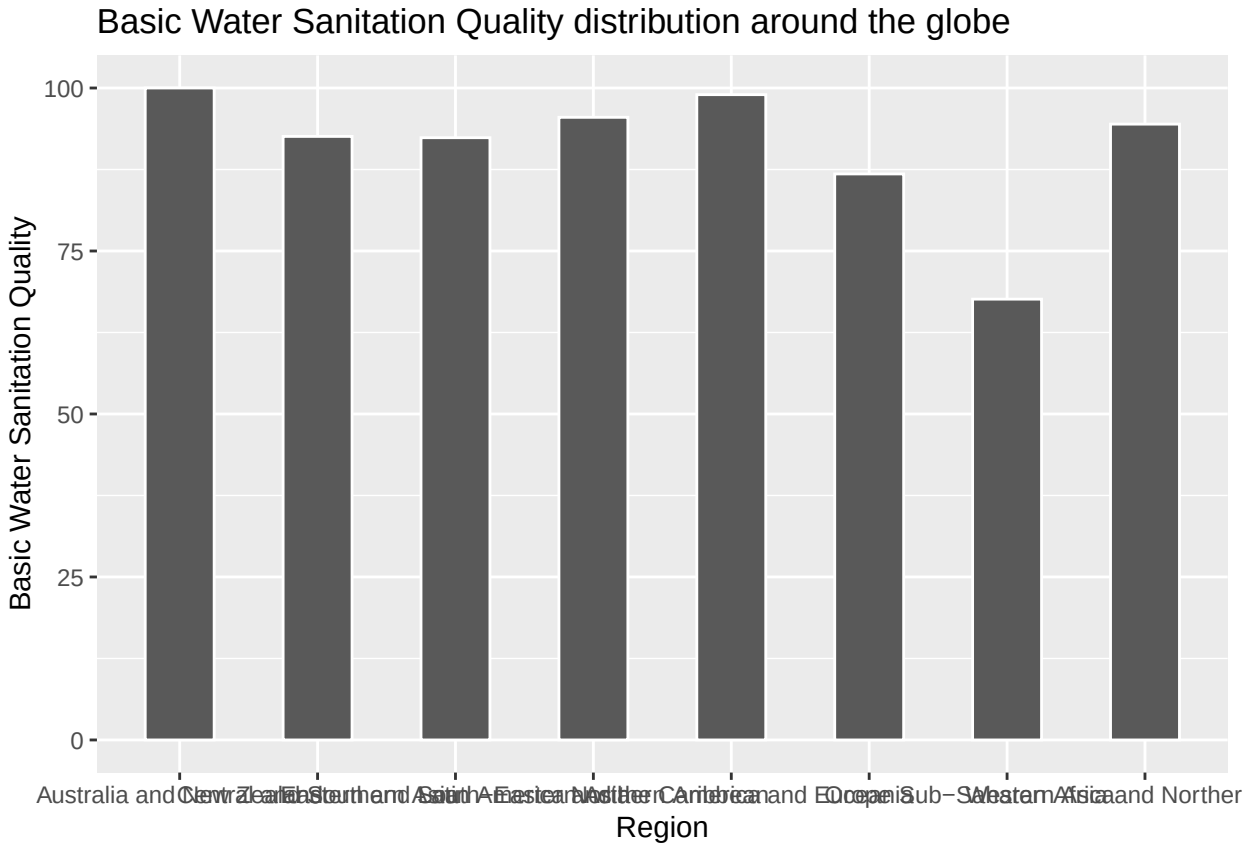

```
subsaharan <- ggplot(data_7, aes(x = variable, y = value,)) +
  ggtitle("% Of Water Sanitation Quality in Sub-Saharan Africa") +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label=value), position=position_dodge(width=0.9), vjust=-0.25) +
  xlab("Water Sanitation Quality") +
  ylab("Mean Percentage")
subsaharan
```




```
oceania <- ggplot(data_8, aes(x = variable, y = value,)) +
  ggtitle("% Of Water Sanitation Quality in Western Asia and Northern Africa") +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label=value), position=position_dodge(width=0.9), vjust=-0.25) +
  xlab("Water Sanitation Quality") +
  ylab("Mean Percentage")
oceania
```




```
water_globe <- ggplot(sani_data_regions, aes(x = Region, y = mean_waterp_basic))+
               geom_bar(stat = "identity", width = 0.5, col = "white")+
               ggtitle("Basic Water Sanitation Quality distribution around the globe")+
               xlab("Region")+
               ylab("Basic Water Sanitation Quality")
water_globe
```


Basic Water Sanitation Quality distribution around the globe
Basic Water Sanitation Quality
100
75
50
25
0
Region

9. [2 marks] Describe the skill that you are demonstrating and interpret your findings. For example, if you have created a histogram, describe the central tendency, shape of the distribution, etc.

We are using bar graphs to compare the differences in water quality in each region and globally. In each region, we can see that the trend is very similar, with most water reaches the quality of basic supply, and small percentage left limited, unimproved, and simply surface water. It means generally, we have a good supply of water to population. But comparing the basic water quality across the globe, the water in access of 30min, we can clearly see that particular region like Sub-Saharan Africa are behind other regions, meaning it does not have a very functional water supply system, and there is till great difficulties for people in these regions to access for the water.

Part II

10. [1 mark] Include your work for Part I.

Part I is attached.

11. [2 marks] Calculate a marginal probability based on your outcome variable. Provide an equation (using probability notation) that describes this probability. For example, if my outcome variable is height in inches, I might calculate the probability that an individual in the dataset has a height of greater than 60 inches. This would be a marginal probability. You may need to first add a new variable to your dataset to calculate your probability of interest, such as a binary variable indicating whether height is greater than 60 inches. There is a resource video about how to code such variables that could be helpful!

Our outcome variable is the diarrheal death rate expressed as the number of diarrheal deaths per year per 100,000 individuals. In our dataset, there were 107 out of 182 countries with available data that had a diarrheal death rate below 5 deaths per year per 100,000 individuals. P(diarrheal death rate < 5) = 107/182 = 0.588 = 58.8%.

$$P(Diarrheal\ Death\ Rate < 5) = \frac{No.\ of\ Countries\ with < 5\ deaths\ per\ 100000}{Total\ no.\ of\ countries\ with\ diarrheal\ data}$$

```
#Using dim() to find the no. of rows in original dataset#
dim(sani_data_clean)
```

```
## [1] 182  11
```

```
#Filtering the dataset to only contain countries with < 5 deaths per 100,000#
lessthanfive <- sani_data_clean  %>%
  filter(diarrheal_death_2017 < 5)
#Using dim() to find the no. of rows in new dataset#
dim(lessthanfive)
```

```
## [1] 107  11
```

```
#Calculating P(Diarrheal Death Rate <5)#
p11 <- 107 / 182
p11
```

```
## [1] 0.5879121
```

12. [2 marks] Using any two variables in your dataset (or derived variables), calculate a conditional probability. Provide an equation (using probability notation) that describes this probability and then use R to calculate it.

By comparing the proportions of countries' access to basic quality water or higher, we were able to create our own binary variable to indicate whether a country had 100% access to at least basic quality water through a new column "full_access_to_basic_water", with an outcome of yes or no. Using this variable in tandem with the diarrheal data from the previous question, we were able to calculate the conditional probability of more than 50% diarrheal death given that a country did NOT have full access to at-least-basic quality water.

P(diarrheal death rate > 5 | not full access to basic water) = P(diarrheal death rate > 5 & not full access to basic water) / P(not full access to basic water)

$P(Diarrheal\,Death\,Rate < 5 || Not\,full\,access\,to\,basic\,water) = \frac{diarrheal\ death\ rate > 5 \cap not\ full\ access\ to\ basic\ water}{No.not\ full\ access\ to\ basic\ water}$

```
# Using mutate() to derive a new column variable #
sanidatap12 <- sani_data_clean %>%
  mutate(full_access_to_basic_water = water_at_least_basic %/% 100)

# Counting how many countries do NOT have full access using dim()#
notfullaccess <- sanidatap12 %>%
  filter(full_access_to_basic_water == 0)
dim(notfullaccess)
```

```
## [1] 121  12
```

```
# Counting how many countries among those that have above 5 diarrheal deaths per 100000
# using dim()
above5andnotfull <- notfullaccess  %>%
  filter(diarrheal_death_2017 > 5)
dim(above5andnotfull)
```

```
## [1] 73 12
```

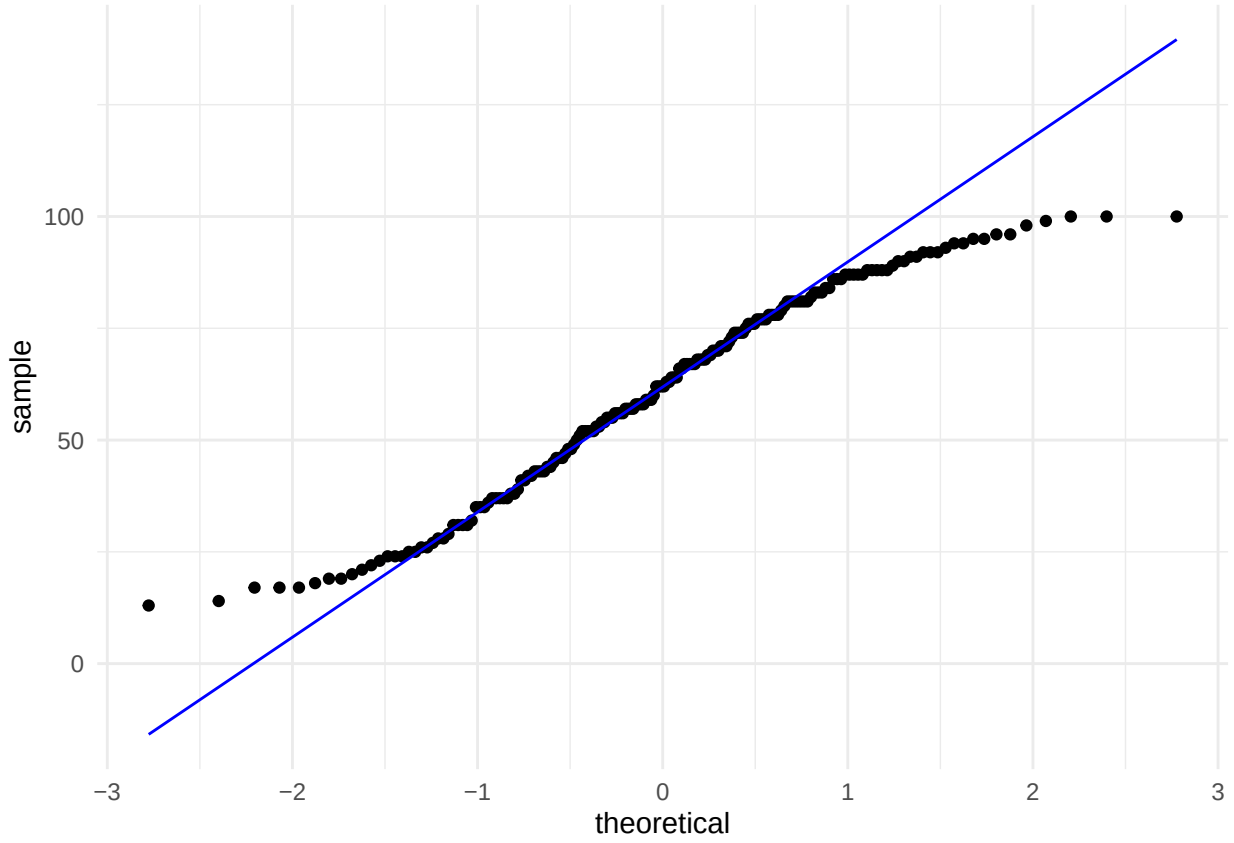
```
# Calculating P(diarrheal death rate > 5 | not full access to basic water) #
p12 <- 73 / 121
p12
```

```
## [1] 0.6033058
```

13. [2 marks] Does your dataset contain a continuous variable? If it does, does the distribution of that variable appear to be normal? Justify your answer using a plot. If your data does not contain a continous variable, give an example related to your dataset of a hypothetical variable that is continuous. That is, imagine what a continuous variable could be in relation to your dataset and topic of interest. For this hypothetical variable, describe what you imagine its shape might be, and how you would check whether or not it is normally distributed.

Yes, the dataset contains continuous variables. One such variable is percent urban, which is the percent of the population in each region that lives in urban areas.To check for normality a QQ plot can be used.

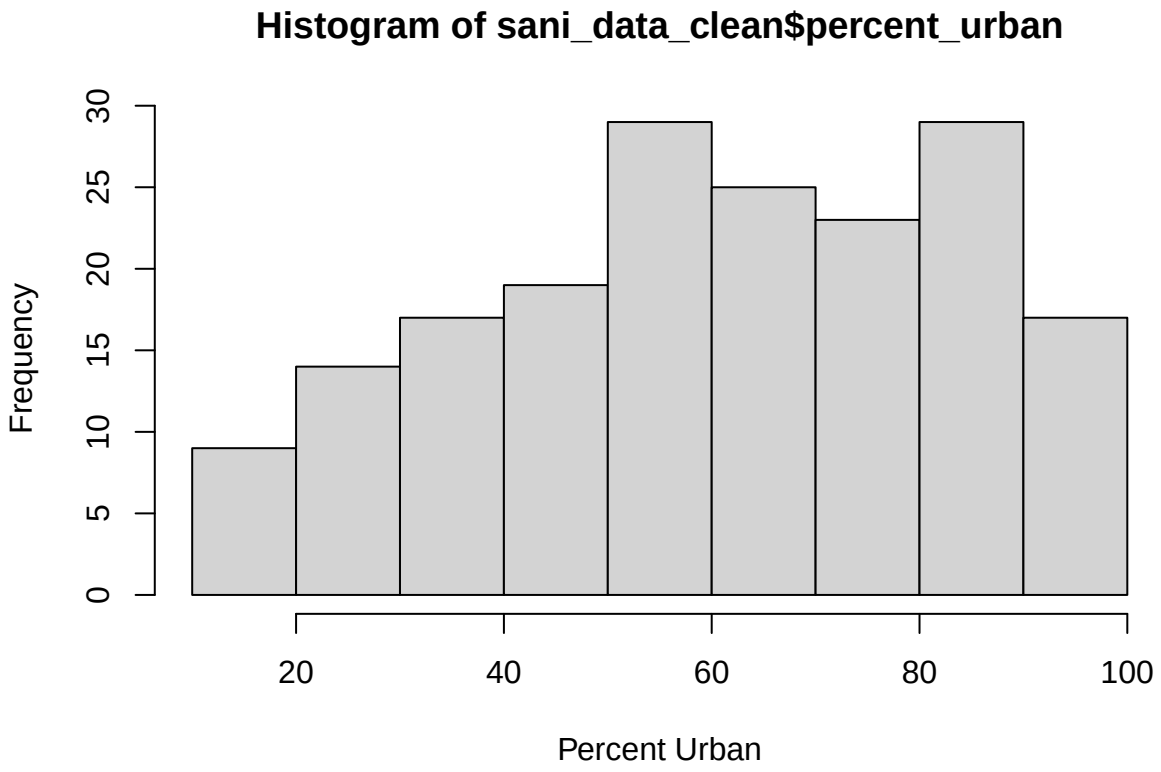
```
sani_normal_QQ_plot <-ggplot(sani_data_clean,aes(sample = percent_urban))+
                    stat_qq()+
                    stat_qq_line(col = "blue")+
                    theme_minimal()
sani_normal_QQ_plot
```



The deviation of the QQ plot from the line on both ends may indicate a distribution with more data points present in the tails, which is not a characteristic of a normal distribution. This means that the data is not normally distributed. The data points are nonlinear on both ends of the line, the data points stray further from the line on the lower end indicating that the distribution is skewed left.

This is also confirmed by the histogram:

```
sani_normal_his_plot <- hist(sani_data_clean$percent_urban, xlab = "Percent Urban")
```




```
sani_normal_his_plot
```

```
## $breaks
##  [1]  10  20  30  40  50  60  70  80  90 100
##
## $counts
## [1]  9 14 17 19 29 25 23 29 17
##
## $density
## [1] 0.004945055 0.007692308 0.009340659 0.010439560 0.015934066 0.013736264
## [7] 0.012637363 0.015934066 0.009340659
##
## $mids
## [1] 15 25 35 45 55 65 75 85 95
##
## $xname
## [1] "sani_data_clean$percent_urban"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

14. [4 marks] Does your dataset contain a binary variable? If so, does this variable meet the criteria to be considered binomially distributed? If so, describe this variable in terms of n and p. Calculate a probability based on this variable, first write the formula for the probability and then using R to calculate the probability (you do not need to calculate the probability by hand). If your data does not contain a binary variable, you can create one based on an underlying continuous variable or a categorical variable with > 2 levels to answer this question.

We have created a binary variable called full_access_to_basic_water, with a yes/no outcome depending on whether the proportion of a country's access to at least basic quality water was 100% or not. This variable cannot be considered binomially distributed, however, because the sanitation data is observed only once for every country, meaning n would be 1.

Using this binomial variable we can calculate the probability that a country will have 100% access to at least basic quality water— P(full access to at least basic water): $P(Full\ access\ to\ at\ least\ basic\ water) = \frac{P(Yes)}{Total\ no.\ of\ countries\ in\ the\ data\ set}$

s

```
# This part was actually done similarly for p12, with some steps being shown
# again.
# We create a new binary variable from our continuous variable with mutate()
# This is created by doing an integer division function on the value, seeing
# if integer dividing by 100 provides 1 or 0.
sanidatap14 <- sani_data_clean %>%
  mutate(full_access_to_basic_water = water_at_least_basic %/% 100)

# We find the no. of total rows in the original dataset, and also
# no. of total rows with full_access_to_basic_water = 1.
# no. of countries with "1" as the value over the total no. of rows
p14filtered <-sanidatap14 %>%
  filter(full_access_to_basic_water == 1)
p14totalrows <- count(sanidatap14)
p14fullaccess <- count(p14filtered)

# P(full access to at least basic water) is calculated by dividing
# no. of countries with "1" as the value over the total no. of rows
p14 <- p14fullaccess / p14totalrows
p14
```

```
##           n
## 1 0.3351648
```

Part III

16. [1 mark] Include your work for Part I.

Parts I + II are attached.

17. [2 marks] Identify a statistical test to apply to your data. This must be a statistical test that we cover in part III of the course. Name the statistical test you have chosen and explain why this is the appropriate test for these data. For example, if I have pre- and post-intervention measurements of morning sleepiness recorded as a quantitative variable, I might choose a paired t test, because the paired t-test is appropriate for continuous outcome data in 2 groups that are inherently related.

We created a new variable name (basicaccess) to divide the countries into four different groups: full access to basic water (full), moderate access to basic water (moderate), less than moderate access to basic water (less than moderate), and low access to basic water (low). In order to evaluate whether or not there are statistically significant differences between each group in relation to the dependent variable, diarrheal death rates, we can perform ANOVA testing. The null hypothesis would indicate that all groups respond equally under different sanitation treatments, the alternative would suggest that at least one group differs. We can use the ANOVA test with our continuous ratio data to determine whether one or more groups contain statistically significant differences. If significance is found further tests can be applied, such as Tukey's test to determine pairwise differences.

An important note is that this test is to be observed for mainly three of the four groups (moderate, less than moderate, and low access) populations, but with another population of interest (full access). The standard deviation of the latter population will be very narrow, given that their basic access has to be 100%, but we feel that the drastic difference to be observed in the means demonstrates the strength of the impact of even slight changes in access to basic quality water. The main focus of analysis will be between "moderate access", "less than moderate access", and "low access".

```
# Creating the column
part3data <- sani_data_clean %>%
  mutate(basicaccess = case_when(
    water_at_least_basic == 100 ~ "Full Access",
    water_at_least_basic >= 85 ~ "Moderate Access",
    water_at_least_basic >= 60 ~ "Less than Moderate Access",
    water_at_least_basic < 60 ~ "Low Access"
  ))
```

18. [2 marks] What assumptions are required by the testing method you chose? Are these assumptions met by your data? How did you assess this? For example, one of the assumptions of the t-test is that the data are normally distributed, so you might choose to assess this with a histogram, or a q-q plot.

Three conditions must be met to perform an ANOVA test.

Condition 1: An independent random sample was taken for each group/population from k populations.

Our data meets this assumption, because each country has its own access to water level that is a result of factors we cannot control. We assessed this by looking at the data and seeing that there is no particular pattern that places one country in either category.

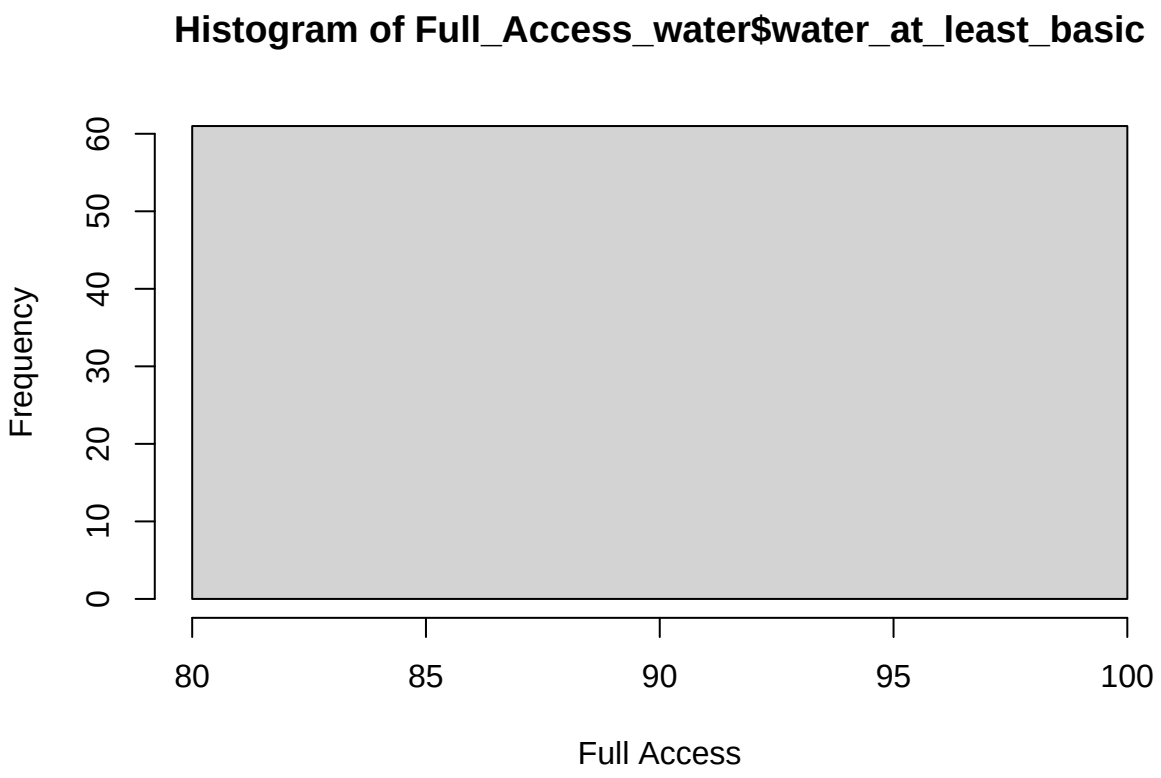
Condition 2: Each of the k populations has an approximately Normal distribution with an unknown mean μi.

This assumption is not met by our data, as only one group—the Low access to water group— came out to have a normal distribution. We assessed this by plotting histograms for the data in each group. Even though most of our data does not follow a normal distribution, we can still use the NOVA test because the test is robust to non-normality.

Condition 3: All the populations have the same standard deviation, whose value is unknown.

We assessed this by using the group_by() and summarize() functions to calculate the standard deviations of the data for each group, and see how closely they align with each other. In doing so, we found that the Full Access to water had a standard deviation of 0, the Moderate Access to water had a standard deviation of 4.9, and the Less Moderate Access to water had a SD of 10.7. Given that we are mainly focusing on the Less Moderate Access to water and the Moderate Access to water groups, this still roughly satisfies the assumption, because as a rule of thumb, the largest sample standard deviation is to be less than twice as large as the smallest one.

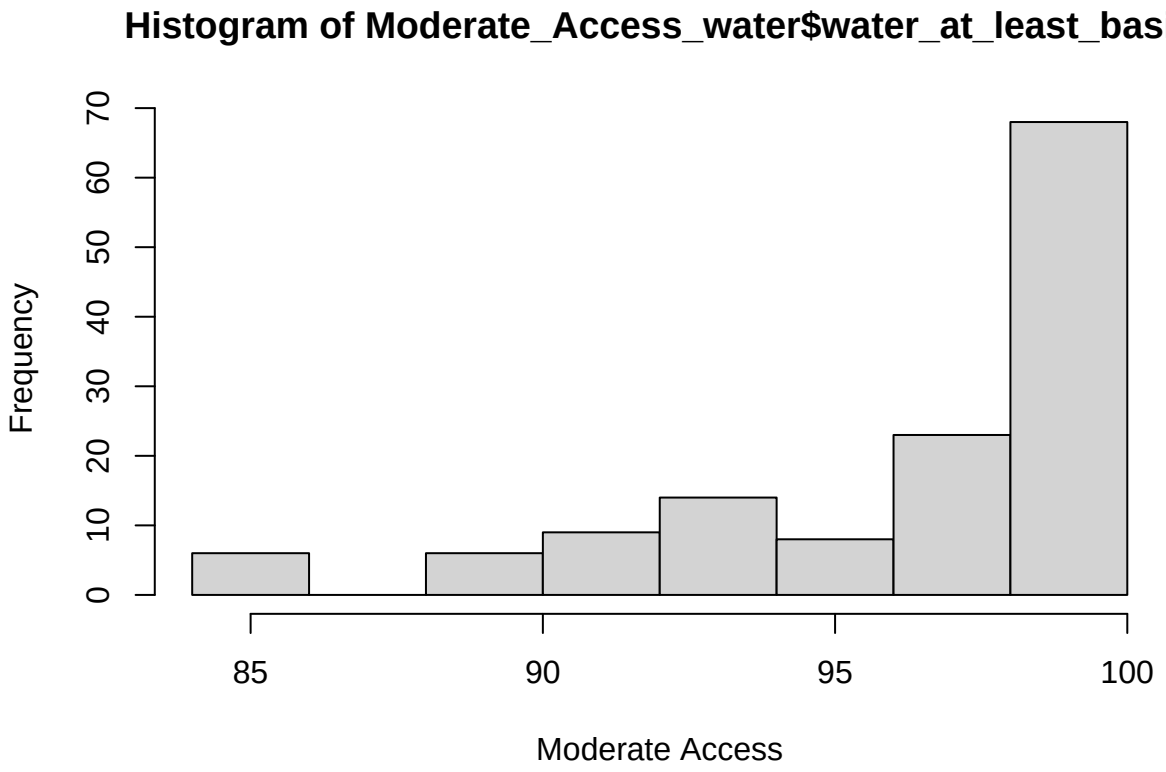
```
#In order to find the distribution of each group, we filtered the data into three different tables base

#The first group looks at Full Access to basic quality water, which is equal to 100
Full_Access_water<-part3data %>% select(basicaccess,water_at_least_basic)%>%filter(water_at_least_basic=

#The next group looks at Moderate Access to basic quality water, and its values are greater than or equ
Moderate_Access_water<-part3data%>%select(basicaccess,water_at_least_basic)%>%filter(water_at_least_bas

#The third group looks at Less than Moderate Access to basic quality water. It's values are less than 8
Less_Moderate_water<-part3data%>%select(basicaccess,water_at_least_basic)%>%filter(water_at_least_basic<

#The fourth group looks at Low Access to basic quality water. It's values are less than 60.
Low_water<-part3data%>%select(basicaccess,water_at_least_basic)%>%filter(water_at_least_basic<60)


#Below are the histograms for the data in each group

#Full Access to Water Histogram:
sani_Full_Access_plot <- hist(Full_Access_water$water_at_least_basic, xlab = "Full Access")
```




```
sani_Full_Access_plot
```

```
## $breaks
## [1]  80 100
##
## $counts
## [1] 61
##
## $density
## [1] 0.05
##
## $mids
## [1] 90
##
## $xname
## [1] "Full_Access_water$water_at_least_basic"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

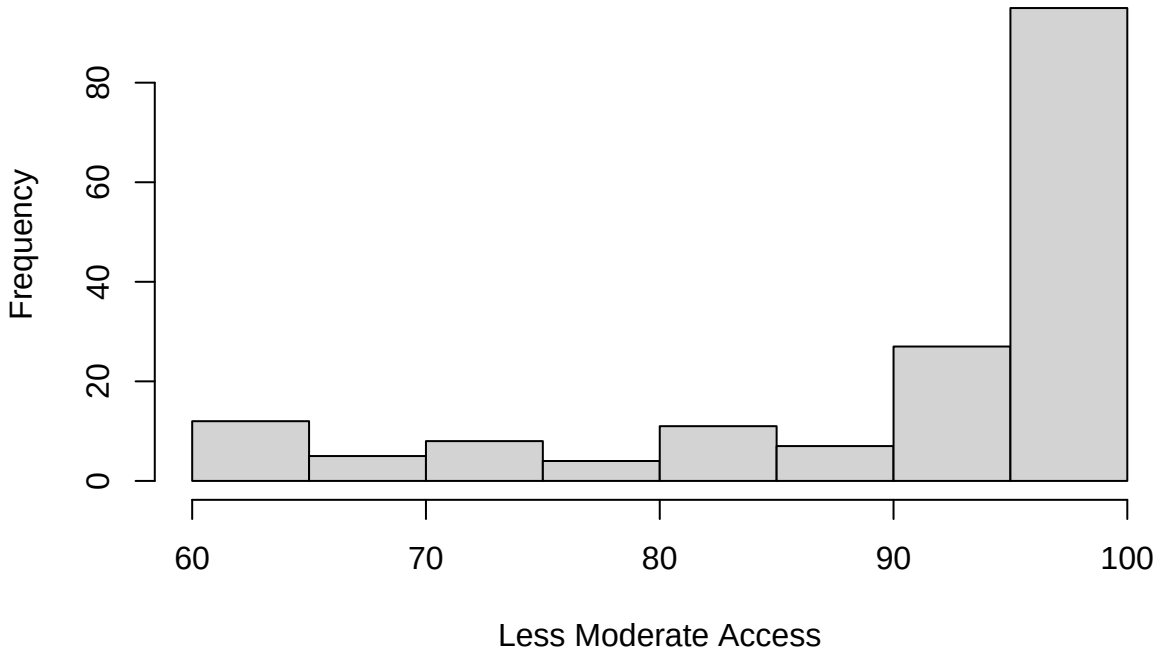
```
#Moderate Access to Water Histogram:
sani_Moderate_Access_plot <- hist(Moderate_Access_water$water_at_least_basic, xlab = "Moderate Access")
```




```
sani_Moderate_Access_plot
```

```
## $breaks
## [1]  84  86  88  90  92  94  96  98 100
##
## $counts
## [1]  6  0  6  9 14  8 23 68
##
## $density
## [1] 0.02238806 0.00000000 0.02238806 0.03358209 0.05223881 0.02985075 0.08582090
## [8] 0.25373134
##
## $mids
## [1] 85 87 89 91 93 95 97 99
##
## $xname
## [1] "Moderate_Access_water$water_at_least_basic"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

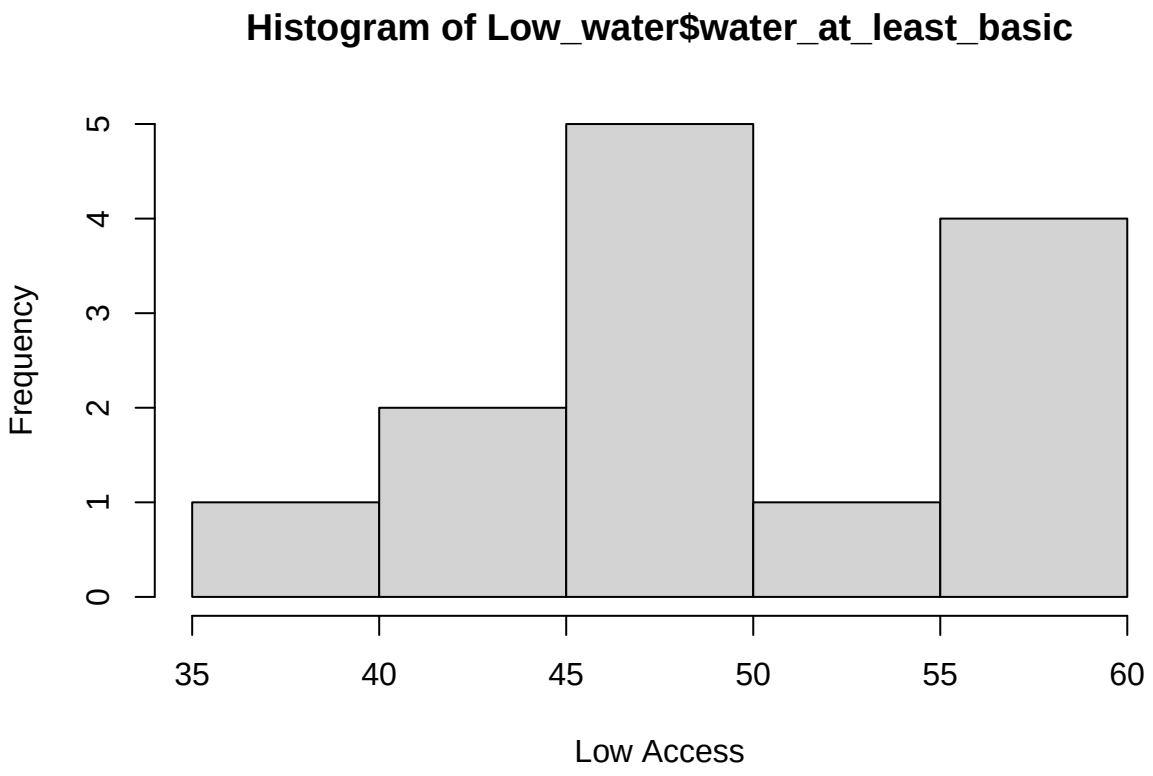
```
#Less than Moderate Access to Water Histogram:
sani_Less_Moderate_plot <- hist(Less_Moderate_water$water_at_least_basic, xlab = "Less Moderate Access"
```




```
sani_Less_Moderate_plot
```

```
## $breaks
## [1]  60  65  70  75  80  85  90  95 100
##
## $counts
## [1] 12  5  8  4 11  7 27 95
##
## $density
## [1] 0.014201183 0.005917160 0.009467456 0.004733728 0.013017751 0.008284024
## [7] 0.031952663 0.112426036
##
## $mids
## [1] 62.5 67.5 72.5 77.5 82.5 87.5 92.5 97.5
##
## $xname
## [1] "Less_Moderate_water$water_at_least_basic"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

```
#Low Access to Water Histogram:
sani_Low_plot <- hist(Low_water$water_at_least_basic, xlab = "Low Access")
```




```
sani_Low_plot
```

```
## $breaks
## [1] 35 40 45 50 55 60
##
## $counts
## [1] 1 2 5 1 4
##
## $density
## [1] 0.01538462 0.03076923 0.07692308 0.01538462 0.06153846
##
## $mids
## [1] 37.5 42.5 47.5 52.5 57.5
##
## $xname
## [1] "Low_water$water_at_least_basic"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

```
#Below are the standard deviations for the data in each group

#Standard Deviation of Full Access to Water:
SD_Full_Access<-Full_Access_water %>% group_by()%>%summarize(Full_Access_SD=sd(water_at_least_basic))
SD_Full_Access
```

```
## # A tibble: 1 x 1
##   Full_Access_SD
##            <dbl>
## 1              0
```

```
#Standard Deviation of Moderate Access to Water:
SD_Moderate_Access<-Moderate_Access_water%>%group_by()%>%summarize(Moderate_Access_SD=sd(water_at_least_
SD_Moderate_Access
```

```
## # A tibble: 1 x 1
##   Moderate_Access_SD
##                <dbl>
## 1               4.04
```

```
#Standard Deviation of Less than Moderate Access to Water:
SD_Less_Moderate<-Less_Moderate_water %>% group_by()%>%summarize(Less_Moderate_SD=sd(water_at_least_basi
SD_Less_Moderate
```

```
## # A tibble: 1 x 1
##   Less_Moderate_SD
##              <dbl>
## 1             11.7
```

```
#Standard Deviation of Low Access to Water:
SD_Low<-Low_water %>% group_by() %>% summarize(Low_SD=sd(water_at_least_basic))
SD_Low
```

```
## # A tibble: 1 x 1
##   Low_SD
##    <dbl>
## 1   6.64
```

19. [2 marks] Clearly state the null and alternative hypotheses for your test.

Null hypothesis: There is no statistically significant difference in means between diarrheal death rates across regions with basic water sanitation, limited water sanitation, and unimproved water sanitation levels.

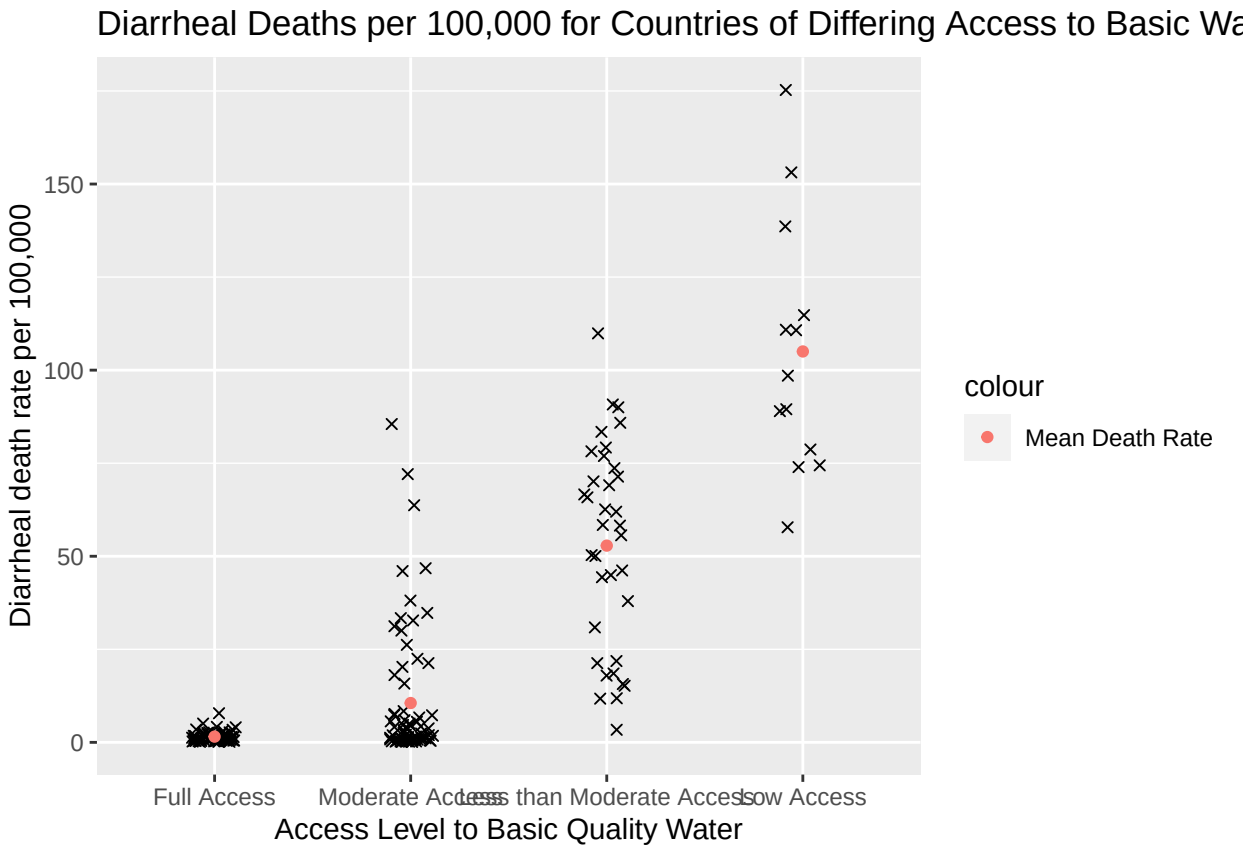
Alternative hypothesis: There is at least one group with a statistically significant difference in mean regarding diarrheal death rates across regions with basic water sanitation, limited water sanitation, and unimproved water sanitation levels.

20. [2 marks] Conduct the statistical test. Include the R code you used to generate your results. Annotate your code to help us follow your reasoning.

```
# Selecting only relevant columns
P20 <- part3data %>%
  select(basicaccess, diarrheal_death_2017)

# Reordering for Groups
library(forcats)
P20 <- P20 %>%
  mutate(accessorder = fct_relevel(basicaccess, c("Full Access", "Moderate Access", "Less than Moderate

# Calculating means and SD for each of the three groups
P20Summary <- P20 %>%
  group_by(accessorder) %>%
  summarise(mean_deathrate = mean(diarrheal_death_2017),
            sd_deathrate = sd(diarrheal_death_2017),
            sample_size = n())
P20Summary <- P20Summary[order(P20Summary$mean_deathrate),]

# Summary data
P20Summary
```

```
## # A tibble: 4 x 4
##   accessorder               mean_deathrate sd_deathrate sample_size
##   <fct>                              <dbl>        <dbl>       <int>
## 1 Full Access                         1.54         1.38          61
## 2 Moderate Access                    10.6          17.7          73
## 3 Less than Moderate Access          52.9          27.6          35
## 4 Low Access                        105.           34.0          13
```

```
# Graphing each access level with diarrheal death rates
ggplot(P20, aes(x = accessorder, y = diarrheal_death_2017)) +
  geom_jitter(pch = 4, width = 0.12) +
  geom_point(data = P20Summary, aes(y = mean_deathrate, col = "Mean Death Rate"), pch = 19) +
  labs(x = "Access Level to Basic Quality Water", y = "Diarrheal death rate per 100,000",
       title = "Diarrheal Deaths per 100,000 for Countries of Differing Access to Basic Water")
```



The mean diarrheal data for countries with differing access to basic quality water were: - Full Access: 1.539078 deaths per 100,000 - Moderate Access: 10.567127 deaths per 100,000 - Less than Moderate Access: 52.855425 deaths per 100,000 - Low Access: 105.016000 deaths per 100,000

We can also see from the visualization that the data for each group varies significantly. Countries with full access have an average diarrheal death rate near 0. Countries with moderate access have around 7 times that rate, while on the other hand, countries that had less than 85% access to basic quality water are observed to vary much greater than the two aforementioned group, with significantly higher average diarrheal death rates (~34 times and ~70 times as that of full access countries).

```
# Calculating the F statistic
library(broom)
anova <- aov(formula = diarrheal_death_2017 ~ basicaccess, data = P20)
tidy(anova)
```

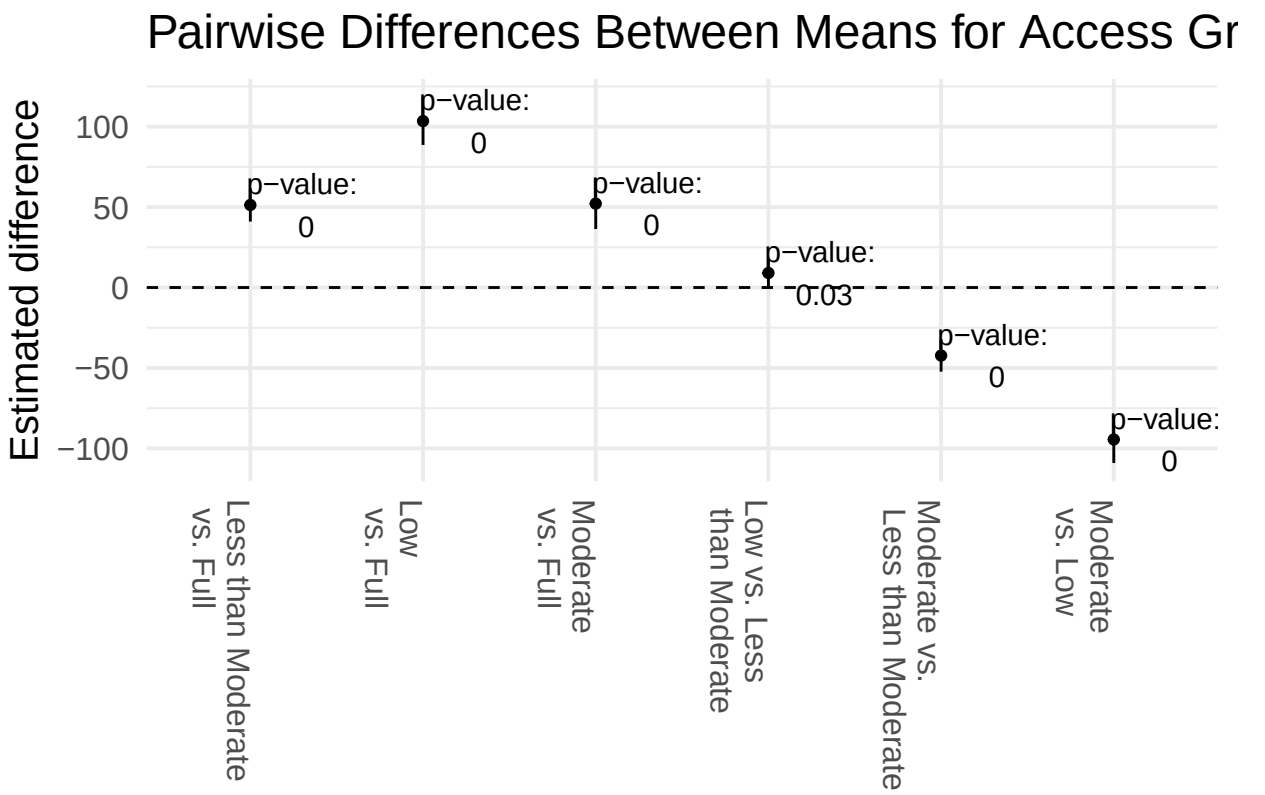
```
## # A tibble: 2 x 6
##   term           df   sumsq meansq statistic   p.value
##   <chr>       <dbl>   <dbl>  <dbl>     <dbl>     <dbl>
## 1 basicaccess     3 157961. 52654.      150.  1.75e-48
## 2 Residuals     178  62519.   351.       NA  NA
```

The calculated F statistic comes out to be 149.9119 , with p-value of 1.74884e-48.

```
# Conducting Tukey's test
differences <- TukeyHSD(anova, conf.level = 0.95) %>%
  tidy()
differences
```

```
## # A tibble: 6 x 7
##   term    contrast            null.value estimate conf.low conf.high adj.p.value
##   <chr>   <chr>                    <dbl>    <dbl>    <dbl>     <dbl>       <dbl>
## 1 basica~ Less than Moderate~          0    51.3    41.0        61.6    3.33e-15
## 2 basica~ Low Access-Full Ac~          0   103.     88.6       118.     3.11e-15
## 3 basica~ Moderate Access-Fu~          0     9.03    0.597      17.5    3.06e- 2
## 4 basica~ Low Access-Less th~          0    52.2    36.4        67.9    7.11e-14
## 5 basica~ Moderate Access-Le~          0   -42.3   -52.3       -32.3    2.60e-14
## 6 basica~ Moderate Access-Lo~          0   -94.4  -109.        -79.8    3.11e-15
```

```r
# Visualizing the differences from Tukey's test
ggplot(differences, aes(x = contrast, y = estimate)) +
  geom_point() +
  geom_segment(aes(y = conf.low, yend = conf.high, xend = contrast)) +
  theme_minimal(base_size = 15) +
  geom_hline(aes(yintercept = 0), lty =2) +
  geom_text(aes(label = paste0("p-value:\n ", round(adj.p.value, 2))), nudge_x = 0.3) +
  labs(y = "Estimated difference", x = "", title = "Pairwise Differences Between Means for Access Groups
  scale_x_discrete(labels = c("Less than Moderate \n vs. Full",
                              "Low \n vs. Full",
                              "Moderate \n vs. Full",
                              "Low vs. Less \n than Moderate",
                              "Moderate vs. \n Less than Moderate",
                              "Moderate \n vs. Low")) +
  theme(axis.text.x=element_text(angle = -90, hjust = 0))
```



Using Tukey's HSD, we can conclude that the means between all pairs are different with the exception of Low vs. Less than Moderate. For this group, the confidence interval appears to include the null value of 0, so we cannot say for sure if that particular group pairing has significantly different means.

21. [4 marks] Present your results in a clear summary. This should include both a text summary and a table or figure with appropriate labeling. For example, if your outcome and predictor/exposure variables are both binary, this might be a 2x2 table. If your method was regression, you might present your regression line graphically. Include your code and annotations.

```
TestResults <- tidy(anova) %>%
  select(term, statistic, p.value)
TestResults
```

```
## # A tibble: 2 x 3
##   term        statistic   p.value
##   <chr>           <dbl>     <dbl>
## 1 basicaccess      150.  1.75e-48
## 2 Residuals         NA  NA
```

The p-value from our ANOVA test is approximately 1.74884e-48. Under the null hypothesis of no difference between group means, there is an extremely small chance of observing the F-statistic of 149.9119 which we calculated. The F-statistic furthermore shows that the variation between means is nearly 150 times as large as the variation among groups. Thus, this provides evidence against the null in favor of the alternative hypothesis that at least one mean is different from the others. This conclusion is also clearly supported by the graphs, where we can visually see significant differences in means and variations across countries of differing basic water access.

22. [4 marks] Interpret your findings. Include a statement about the evidence, your conclusions, and the generalizability of your findings. Our analyses and conclusions depend on the quality of our study design and the methods of data collection. Any missteps or oversights during the data collection process could potentially change the outcome of what we are trying to find. Consider the methods used to collect the data you analyzed. Was there any potential issue in how the participants were selected/recruited, retained, or assessed that may have impacted the outcome of your analysis/visualization? Were there any potential biases that you might be concerned about? Were there factors that were not measured or considered that you think could be important to the interpretation of these data?

As expected, the mean and variation for the "full access" group is significantly smaller than that of any other group. While this may have complicated the ANOVA test and analysis slightly, we felt it was important to point out that every single first world country had classified as a country with full access to basic quality water by the data that we found, and hence it was a worthwhile endeavor to isolate such countries in their own country to see how they compare with diarrheal death rates for countries with anything less than 100%. The mean for diarrheal death rates of the next group, moderate access, had around 7 times as more deaths per 100,000 than for the full access group. At worst, the lowest group with less than 60% access to basic quality water, had up to 70 times that of the full access group.

According to the World Health Organization, approximately 830,000 are estimated to die each year from diarrhea as a result of unsafe drinking water, sanitation, and hand hygiene. Our statistical analysis has explored the consequence of varying levels of access to clean water in regions across the world. We chose to further analyze the impact access to clean water has by diarrhoeal death rates and the percentage of populations with access to clean, basic quality water. Diarrheal death is largely preventable by increasing access to clean water and improving water sanitization methods, that is why it proves to be an important measure in understanding the impact of widespread access to clean water. Through ANOVA testing we are able to establish that there is a statistically significant difference across access levels: Full access to basic clean water (100% of the population has access), moderate access to basic clean water (85% to 100% of the population has access), less than moderate access to basic clean water (60% to 85% of the population has access), and low access (less than 60% of the population has access). Regions with full access to basic clean water have significantly lower— almost nonexistent— diarrheal death rates as indicated by the mean diarrheal death rates near 1 per 100,000 as opposed to those countries with moderate and below access to basic clean water.

Regarding some issues with data cleaning, the main dataset that we worked with combines data from the WHO regarding countries' proportional access to different qualities of water in 2020, and diarrheal data from the Global Health Data Exchange from 2017. The diarrheal data is 3 years older than the sanitation data we worked with, which could have led to analysis based on outdated information. In addition, we had to filter out some countries from the initial dataset because they did not have proper data in regards to proportions of water access, which in combination with the fact that the diarrheal dataset did not contain data for every country in the WHO dataset, could mean that the analysis was done without factoring in data from other countries that could have revealed a more telling correlation between water quality access and diarrheal death rates.

The methods used to collect the sanitation data involved household surveys and censuses combined with data provided by administrative sources and regulators. A combination of different sources for data, as well as survey methods in collaboration with UNICEF MICS, means that thankfully we can reliably say that chances of biases or missteps are low.

23. [1 mark] Create a statement of contribution. This is now common in journal articles. For example, the American Journal of Epidemiology provides the following instructions to authors: "Authorship credit should be based on criteria developed by the International Committee for Medical Journal Editors (ICMJE): 1) substantial contributions to conception and design, or acquisition of data, or analysis and interpretation of data; 2) drafting the article or reviewing it and, if appropriate, revising it critically for important intellectual content; 3) final approval of the version to be published. Authors should meet all conditions. In addition, each author must certify that he or she has participated sufficiently in the work to believe in its overall validity and to take public responsibility for appropriate portions of its content. Author names should be listed in ScholarOne and author contributions should be detailed in the cover letter (e.g., "Author A designed the study and directed its implementation, including quality assurance and control. Author B helped supervise the field activities and designed the study's analytic strategy. Author C helped conduct the literature review and prepare the Methods and the Discussion sections of the text.").

We discussed with each other and chose to combine the datasets of sanitation and water found by Grace Rajan, and diarrheal death found by Phoenix Ding. Phoenix Ding cleaned and organized the data into csv and Stephen Hwang helped to find external packages to import data in R studio. We equally distribute our responsibilities in the project and work collectively towards the coding.