Project 1: Exploratory Data Analysis by Jessica Martin

Introduction

In the following document, two separate data sets, both of which were found from https://github.com/rfordatascience/tidytuesday, will be combined and examined.

The first dataset, named "global_mortality" contains mortality data for 228 political entities (mostly independent countries) for all years between 1990-2016. In this dataset, the mortality rate from 32 distinct factors is recorded as a percentage of the total deaths for each year in each country. In the second dataset, named "lifeexp", 241 political entities (which I will refer to as countries, for simplicity), are presented along with the life expectancy of each country for years 1950-2015.

Unfortunately, I did not find any information as to how these datasets were acquired. However, it is fortunate that these datasets were created by the same author; therefore, there is a lot of consistency between the datasets, most notably of which are the variables "country" and "country code". These datasets are so compatible because mortality rates for disease and other factors (found in the global_mortality dataset) will greatly contribute to the life expectancy for each country (found in the lifeexp dataset). It will be interesting to explore how life expectancy fluctuates between countries based upon the factors which have the highest mortality percentage. However, it is necessary to mention that neither dataset contains information on differences in healthcare systems and availability of good hospitals with modern medical technology, which will also have a huge impact on life expectancy and the mortality percentage for each factor.

I picked these datasets because I have a great interest not only in healthcare and knowing which diseases/factors are the most deadly in the world, but also because it is interesting (yet tragic) to see how countries throughout the world differ so greatly in terms of health, mortality, and life expectancy.

Uploading and Tidying

Because both of these datasets were found from an external source, I had to upload them to the R server, one in the form of an imported excel file, the other of which was a .csv file. I then took a glimpse at each of them in order to familiarize myself with the variables in each dataset.

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readxl)
global_mortality <- read_excel("global_mortality.xlsx")
global_mortality%>%glimpse()
```

```
## Observations: 6,156
## Variables: 35
## $ country                                  <chr> "Afghanistan", "Afg...
## $ country_code                             <chr> "AFG", "AFG", "AFG"...
```

```
## $ year                                          <dbl> 1990, 1991, 1992, 1...
## $ `Cardiovascular diseases (%)`                 <dbl> 17.61040, 17.80181,...
## $ `Cancers (%)`                                 <dbl> 4.025975, 4.054145,...
## $ `Respiratory diseases (%)`                    <dbl> 2.106626, 2.134176,...
## $ `Diabetes (%)`                                <dbl> 3.832555, 3.822228,...
## $ `Dementia (%)`                                <dbl> 0.5314287, 0.532497...
## $ `Lower respiratory infections (%)`            <dbl> 10.886362, 10.35696...
## $ `Neonatal deaths (%)`                         <dbl> 9.184653, 8.938897,...
## $ `Diarrheal diseases (%)`                      <dbl> 2.497141, 2.572228,...
## $ `Road accidents (%)`                          <dbl> 3.715944, 3.729142,...
## $ `Liver disease (%)`                           <dbl> 0.8369093, 0.845515...
## $ `Tuberculosis (%)`                            <dbl> 5.877075, 5.891704,...
## $ `Kidney disease (%)`                          <dbl> 1.680611, 1.671115,...
## $ `Digestive diseases (%)`                      <dbl> 1.058771, 1.049322,...
## $ `HIV/AIDS (%)`                                <dbl> 0.01301948, 0.01451...
## $ `Suicide (%)`                                 <dbl> 0.4366105, 0.442280...
## $ `Malaria (%)`                                 <dbl> 0.4488863, 0.455019...
## $ `Homicide (%)`                                <dbl> 1.287020, 1.290991,...
## $ `Nutritional deficiencies (%)`                <dbl> 0.3505045, 0.343212...
## $ `Meningitis (%)`                              <dbl> 3.037603, 2.903202,...
## $ `Protein-energy malnutrition (%)`             <dbl> 0.3297599, 0.322171...
## $ `Drowning (%)`                                <dbl> 0.9838624, 0.954586...
## $ `Maternal deaths (%)`                         <dbl> 1.769213, 1.749264,...
## $ `Parkinson disease (%)`                       <dbl> 0.02515859, 0.02545...
## $ `Alcohol disorders (%)`                       <dbl> 0.02899828, 0.02917...
## $ `Intestinal infectious diseases (%)`          <dbl> 0.1833303, 0.178107...
## $ `Drug disorders (%)`                          <dbl> 0.04120540, 0.04203...
## $ `Hepatitis (%)`                               <dbl> 0.1387378, 0.135008...
## $ `Fire (%)`                                    <dbl> 0.1741567, 0.170671...
## $ `Heat-related (hot and cold exposure) (%)`    <dbl> 0.1378229, 0.134826...
## $ `Natural disasters (%)`                       <dbl> 0.00000000, 0.79760...
## $ `Conflict (%)`                                <dbl> 0.932, 2.044, 2.408...
## $ `Terrorism (%)`                               <dbl> 0.007, 0.040, 0.027...
```

```
lifeexp<-read.csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2018/2018-
lifeexp%>%glimpse()
```

```
## Observations: 17,894
## Variables: 4
## $ country         <fct> Afghanistan, Afghanistan, Afghanistan, Afghani...
## $ code            <fct> AFG, AFG, AFG, AFG, AFG, AFG, AFG, AFG, AFG, A...
## $ year            <int> 1950, 1951, 1952, 1953, 1954, 1955, 1956, 1957...
## $ life_expectancy <dbl> 27.537, 27.810, 28.350, 28.880, 29.399, 29.907...
```

Both datasets appear to be tidy; however, I will untidy and re-tidy each dataset before joining them. Here, I include a glimpse of the final, "tidy" dataset–"global_mortality1" and "longlifeexp". This is to demonstrate that after untidying and tidying the datasets, they are exactly the same as the initial datasets.

For the "global_mortality" dataset, several steps were needed in order to untidy and retidy. First, to untidy the datset, I used pivot wider. I took all years (1990-2015) and combined them with each mortality factor (which came from columns 4-35). I then began to retidy. To pivot longer, I had to pivot all columns besides "country" and "country_code". Then, I had to use "separate()"" to separate each disease from the year it was associated with. Lastly, I had to pivot wider again in order to put all disease names back into their own columns, along with their values. Similar steps were followed to untidy and retidy the "lifeexp" dataset, except I did not have to use separate() and I did not have to pivot wider when re-tidying.

```
wideglobal<-global_mortality%>%pivot_wider(names_from="year",values_from=c(4:35))
longglobal1<-wideglobal%>%pivot_longer(cols=-c("country","country_code"))
longglobal2<-longglobal1%>%separate(name,into=c("disease","year"),sep="_")
global_mortality1<-longglobal2%>%pivot_wider(names_from="disease",values_from="value")

global_mortality1%>%glimpse()
```

```
## Observations: 6,156
## Variables: 35
## $ country                                    <chr> "Afghanistan", "Afg...
## $ country_code                               <chr> "AFG", "AFG", "AFG"...
## $ year                                       <chr> "1990", "1991", "19...
## $ `Cardiovascular diseases (%)`              <dbl> 17.61040, 17.80181,...
## $ `Cancers (%)`                              <dbl> 4.025975, 4.054145,...
## $ `Respiratory diseases (%)`                 <dbl> 2.106626, 2.134176,...
## $ `Diabetes (%)`                             <dbl> 3.832555, 3.822228,...
## $ `Dementia (%)`                             <dbl> 0.5314287, 0.532497...
## $ `Lower respiratory infections (%)`         <dbl> 10.886362, 10.35696...
## $ `Neonatal deaths (%)`                      <dbl> 9.184653, 8.938897,...
## $ `Diarrheal diseases (%)`                   <dbl> 2.497141, 2.572228,...
## $ `Road accidents (%)`                       <dbl> 3.715944, 3.729142,...
## $ `Liver disease (%)`                        <dbl> 0.8369093, 0.845515...
## $ `Tuberculosis (%)`                         <dbl> 5.877075, 5.891704,...
## $ `Kidney disease (%)`                       <dbl> 1.680611, 1.671115,...
## $ `Digestive diseases (%)`                   <dbl> 1.058771, 1.049322,...
## $ `HIV/AIDS (%)`                             <dbl> 0.01301948, 0.01451...
## $ `Suicide (%)`                              <dbl> 0.4366105, 0.442280...
## $ `Malaria (%)`                              <dbl> 0.4488863, 0.455019...
## $ `Homicide (%)`                             <dbl> 1.287020, 1.290991,...
## $ `Nutritional deficiencies (%)`             <dbl> 0.3505045, 0.343212...
## $ `Meningitis (%)`                           <dbl> 3.037603, 2.903202,...
## $ `Protein-energy malnutrition (%)`          <dbl> 0.3297599, 0.322171...
## $ `Drowning (%)`                             <dbl> 0.9838624, 0.954586...
## $ `Maternal deaths (%)`                      <dbl> 1.769213, 1.749264,...
## $ `Parkinson disease (%)`                    <dbl> 0.02515859, 0.02545...
## $ `Alcohol disorders (%)`                    <dbl> 0.02899828, 0.02917...
## $ `Intestinal infectious diseases (%)`       <dbl> 0.1833303, 0.178107...
## $ `Drug disorders (%)`                       <dbl> 0.04120540, 0.04203...
## $ `Hepatitis (%)`                            <dbl> 0.1387378, 0.135008...
## $ `Fire (%)`                                 <dbl> 0.1741567, 0.170671...
## $ `Heat-related (hot and cold exposure) (%)` <dbl> 0.1378229, 0.134826...
## $ `Natural disasters (%)`                    <dbl> 0.00000000, 0.79760...
## $ `Conflict (%)`                             <dbl> 0.932, 2.044, 2.408...
## $ `Terrorism (%)`                            <dbl> 0.007, 0.040, 0.027...
```

```
widelife<-lifeexp%>%pivot_wider(names_from="year",values_from="life_expectancy")
longlifeexp<-widelife%>%pivot_longer(cols=-c("country","code"),names_to="year",values_to="life_expectan

longlifeexp%>%glimpse()
```

```
## Observations: 73,987
## Variables: 4
## $ country         <fct> Afghanistan, Afghanistan, Afghanistan, Afghani...
## $ code            <fct> AFG, AFG, AFG, AFG, AFG, AFG, AFG, AFG, AFG, A...
```

```
## $ year            <chr> "1950", "1951", "1952", "1953", "1954", "1955"...
## $ life_expectancy <dbl> 27.537, 27.810, 28.350, 28.880, 29.399, 29.907...
```

Joining

The next step is to join the datasets into one and remove the NAs. I used a full join for this because I want to keep as much information as I can from each data set; by using full_join, no rows of data will be dropped. However, there will inevitably be a lot of "NA" observations from the lifeexp dataset because it includes data for 40 previous years than the global_mortality dataset. Thus, the final joined dataset dropped 14,731 rows from the initially joined dataset. This seems like a significant loss of data; however, it just means that the joined dataset focuses solely on years 1990-2015, which is still relevant for this project.

```
joined<-full_join(lifeexp,global_mortality)
```

```
## Joining, by = c("country", "year")
```

```
## Warning: Column `country` joining factor and character vector, coercing
## into character vector
```

```
noNA<-joined%>%na.omit()
final_joined<-noNA%>%select(-"country_code")
final_joined%>%glimpse()
```

```
## Observations: 4,275
## Variables: 36
## $ country                                 <chr> "Afghanistan", "Afg...
## $ code                                    <fct> AFG, AFG, AFG, AFG,...
## $ year                                    <dbl> 1990, 1991, 1992, 1...
## $ life_expectancy                         <dbl> 49.856, 50.627, 51....
## $ `Cardiovascular diseases (%)`           <dbl> 17.61040, 17.80181,...
## $ `Cancers (%)`                           <dbl> 4.025975, 4.054145,...
## $ `Respiratory diseases (%)`              <dbl> 2.106626, 2.134176,...
## $ `Diabetes (%)`                          <dbl> 3.832555, 3.822228,...
## $ `Dementia (%)`                          <dbl> 0.5314287, 0.532497...
## $ `Lower respiratory infections (%)`      <dbl> 10.886362, 10.35696...
## $ `Neonatal deaths (%)`                   <dbl> 9.184653, 8.938897,...
## $ `Diarrheal diseases (%)`                <dbl> 2.497141, 2.572228,...
## $ `Road accidents (%)`                    <dbl> 3.715944, 3.729142,...
## $ `Liver disease (%)`                     <dbl> 0.8369093, 0.845515...
## $ `Tuberculosis (%)`                      <dbl> 5.877075, 5.891704,...
## $ `Kidney disease (%)`                    <dbl> 1.680611, 1.671115,...
## $ `Digestive diseases (%)`                <dbl> 1.058771, 1.049322,...
## $ `HIV/AIDS (%)`                          <dbl> 0.01301948, 0.01451...
## $ `Suicide (%)`                           <dbl> 0.4366105, 0.442280...
## $ `Malaria (%)`                           <dbl> 0.4488863, 0.455019...
## $ `Homicide (%)`                          <dbl> 1.287020, 1.290991,...
## $ `Nutritional deficiencies (%)`          <dbl> 0.3505045, 0.343212...
## $ `Meningitis (%)`                        <dbl> 3.037603, 2.903202,...
## $ `Protein-energy malnutrition (%)`       <dbl> 0.3297599, 0.322171...
## $ `Drowning (%)`                          <dbl> 0.9838624, 0.954586...
## $ `Maternal deaths (%)`                   <dbl> 1.769213, 1.749264,...
## $ `Parkinson disease (%)`                 <dbl> 0.02515859, 0.02545...
## $ `Alcohol disorders (%)`                 <dbl> 0.02899828, 0.02917...
## $ `Intestinal infectious diseases (%)`    <dbl> 0.1833303, 0.178107...
```

```
## $ `Drug disorders (%)`                          <dbl> 0.04120540, 0.04203...
## $ `Hepatitis (%)`                               <dbl> 0.1387378, 0.135008...
## $ `Fire (%)`                                    <dbl> 0.1741567, 0.170671...
## $ `Heat-related (hot and cold exposure) (%)`    <dbl> 0.1378229, 0.134826...
## $ `Natural disasters (%)`                       <dbl> 0.00000000, 0.79760...
## $ `Conflict (%)`                                <dbl> 0.932, 2.044, 2.408...
## $ `Terrorism (%)`                               <dbl> 0.007, 0.040, 0.027...
```

```
nrow(joined)-nrow(final_joined)
```

```
## [1] 14731
```

Wrangling

In this section, I will use all dplyr functions to explore and summarize statistics in my newly joined dataset.

Filter, Select, Arrange

```
final_joined%>%filter(year==max(year)|year==min(year))%>%select(1,3,4)%>%arrange(life_expectancy)%>%gli
```

```
## Observations: 342
## Variables: 3
## $ country         <chr> "Rwanda", "Sierra Leone", "Angola", "Mozambiqu...
## $ year            <dbl> 1990, 1990, 1990, 1990, 1990, 1990, 1990, 1990...
## $ life_expectancy <dbl> 34.217, 37.348, 41.696, 42.915, 43.525, 43.540...
```

This code is useful for observing the life expectancy for each country in the first year of the dataset and for the last year of the dataset and lists the life expectancies from least to greatest.

Mutate

```
mutated<-final_joined%>%group_by(country)%>%mutate(meanlifeexp=mean(life_expectancy,na.rm=T))%>%glimpse
```

```
## Observations: 4,275
## Variables: 37
## Groups: country [171]
## $ country                                  <chr> "Afghanistan", "Afg...
## $ code                                     <fct> AFG, AFG, AFG, AFG,...
## $ year                                     <dbl> 1990, 1991, 1992, 1...
## $ life_expectancy                          <dbl> 49.856, 50.627, 51....
## $ `Cardiovascular diseases (%)`            <dbl> 17.61040, 17.80181,...
## $ `Cancers (%)`                            <dbl> 4.025975, 4.054145,...
## $ `Respiratory diseases (%)`               <dbl> 2.106626, 2.134176,...
## $ `Diabetes (%)`                           <dbl> 3.832555, 3.822228,...
## $ `Dementia (%)`                           <dbl> 0.5314287, 0.532497...
## $ `Lower respiratory infections (%)`       <dbl> 10.886362, 10.35696...
## $ `Neonatal deaths (%)`                    <dbl> 9.184653, 8.938897,...
## $ `Diarrheal diseases (%)`                 <dbl> 2.497141, 2.572228,...
## $ `Road accidents (%)`                     <dbl> 3.715944, 3.729142,...
## $ `Liver disease (%)`                      <dbl> 0.8369093, 0.845515...
## $ `Tuberculosis (%)`                       <dbl> 5.877075, 5.891704,...
## $ `Kidney disease (%)`                     <dbl> 1.680611, 1.671115,...
```

```
## $ `Digestive diseases (%)`                    <dbl> 1.058771, 1.049322,...
## $ `HIV/AIDS (%)`                              <dbl> 0.01301948, 0.01451...
## $ `Suicide (%)`                               <dbl> 0.4366105, 0.442280...
## $ `Malaria (%)`                               <dbl> 0.4488863, 0.455019...
## $ `Homicide (%)`                              <dbl> 1.287020, 1.290991,...
## $ `Nutritional deficiencies (%)`              <dbl> 0.3505045, 0.343212...
## $ `Meningitis (%)`                            <dbl> 3.037603, 2.903202,...
## $ `Protein-energy malnutrition (%)`           <dbl> 0.3297599, 0.322171...
## $ `Drowning (%)`                              <dbl> 0.9838624, 0.954586...
## $ `Maternal deaths (%)`                       <dbl> 1.769213, 1.749264,...
## $ `Parkinson disease (%)`                     <dbl> 0.02515859, 0.02545...
## $ `Alcohol disorders (%)`                     <dbl> 0.02899828, 0.02917...
## $ `Intestinal infectious diseases (%)`        <dbl> 0.1833303, 0.178107...
## $ `Drug disorders (%)`                        <dbl> 0.04120540, 0.04203...
## $ `Hepatitis (%)`                             <dbl> 0.1387378, 0.135008...
## $ `Fire (%)`                                  <dbl> 0.1741567, 0.170671...
## $ `Heat-related (hot and cold exposure) (%)` <dbl> 0.1378229, 0.134826...
## $ `Natural disasters (%)`                     <dbl> 0.00000000, 0.79760...
## $ `Conflict (%)`                              <dbl> 0.932, 2.044, 2.408...
## $ `Terrorism (%)`                             <dbl> 0.007, 0.040, 0.027...
## $ meanlifeexp                                 <dbl> 57.18436, 57.18436,...
```

This code creates a new column in the dataset which presents the total mean life expectancy for each country in combined years 1990-2015.

Summarize and Group_by

For the following ten summarizing statistics, I used pivot_longer on the final_joined dataset and named it "longdata".

Finding the mean mortality percentage of cardiovascular disease in years 1990-2015 for each country:

```
longdata<-final_joined%>%pivot_longer(cols=c(5:36),names_to="diseases",values_to="values")
longdata%>%group_by(country)%>%filter(diseases=="Cardiovascular diseases (%)")%>%summarize(meanmortalit
```

```
## Observations: 171
## Variables: 2
## $ country       <chr> "Georgia", "Bulgaria", "Ukraine", "Belarus", "Ma...
## $ meanmortality <dbl> 64.11682, 64.08298, 61.93131, 59.50497, 59.33890...
```

Mean mortality percentage of cancers in all countries from years 1990-2015:

```
longdata%>%group_by(country)%>%filter(diseases=="Cancers (%)")%>%summarize(meanmortality=mean(values))%
```

```
## Observations: 171
## Variables: 2
## $ country       <chr> "Canada", "Netherlands", "Japan", "France", "Den...
## $ meanmortality <dbl> 30.20443, 30.14955, 29.68774, 29.42927, 28.60376...
```

Mean mortality percentage of respiratory diseases in all countries from years 1990-2015:

```
longdata%>%group_by(country)%>%filter(diseases=="Respiratory diseases (%)")%>%summarize(meanmortality=m
```

```
## Observations: 171
## Variables: 2
## $ country       <chr> "China", "Papua New Guinea", "North Korea", "Ind...
## $ meanmortality <dbl> 13.125347, 12.746989, 11.035083, 10.136311, 9.17...
```

Mean mortality percentage of diabetes in all countries from years 1990-2015:

```
longdata%>%group_by(country)%>%filter(diseases=="Diabetes (%)")%>%summarize(meanmortality=mean(values))%
```

```
## Observations: 171
## Variables: 2
## $ country       <chr> "Fiji", "Bahrain", "Trinidad and Tobago", "Mauri...
## $ meanmortality <dbl> 25.816062, 18.272696, 18.109252, 17.719682, 16.1...
```

Mean US mortality percentage of cardiovascular disease from years 1990-2015:

```
longdata%>%filter(country=="United States" & diseases=="Cardiovascular diseases (%)")%>%summarize(meanUS
```

```
## # A tibble: 1 x 1
##   meanUSmortality
##             <dbl>
## 1            36.3
```

Mean US mortality percentage of cancer from years 1990-2015:

```
longdata%>%filter(country=="United States" & diseases=="Cancers (%)")%>%summarize(meanUSmortality=mean(v
```

```
## # A tibble: 1 x 1
##   meanUSmortality
##             <dbl>
## 1            24.2
```

Mean US mortality percentage of respiratory diseases from years 1990-2015:

```
longdata%>%filter(country=="United States" & diseases=="Respiratory diseases (%)")%>%summarize(meanUSmo
```

```
## # A tibble: 1 x 1
##   meanUSmortality
##             <dbl>
## 1            6.10
```

Mean US mortality percentage of diabetes from years 1990-2015:

```
longdata%>%filter(country=="United States" & diseases=="Diabetes (%)")%>%summarize(meanUSmortality=mean
```

```
## # A tibble: 1 x 1
##   meanUSmortality
##             <dbl>
## 1            6.51
```

The standard deviation of cardiovascular diseases in all countries from years 1990-2015:

```
longdata%>%group_by(country)%>%filter(diseases=="Cardiovascular diseases (%)")%>%summarize(sdmortality=s
```

```
## Observations: 171
## Variables: 2
## $ country     <chr> "Syria", "Bangladesh", "Albania", "North Korea", "...
## $ sdmortality <dbl> 8.164897, 7.603258, 6.940401, 6.935442, 6.468381, ...
```

The standard deviation of diarrheal diseases in all countries from years 1990-2015:

```
longdata%>%group_by(country)%>%filter(diseases=="Diarrheal diseases (%)")%>%summarize(sdmortality=sd(val
```

```
## Observations: 171
## Variables: 2
## $ country     <chr> "Angola", "Mauritania", "Nicaragua", "Nepal", "Swa...
## $ sdmortality <dbl> 5.922003, 4.792131, 4.331496, 4.251117, 3.613355, ...
```

The first four summary statistic codes explore the mean mortality prevalence of cardiovascular disease, cancers, respiratory diseases, and diabetes in all countries and ranks them from highest mortality percentage to lowest. This is a useful statistic because for each specified disease, we can see where the mortality prevalence of that disease is the highest. The next four summary statistic codes explore the mean mortality prevalence of the same four diseases in the United States. These statistics allow for relative comparison between the country that we are most familiar with and the other countries of the world. These statistics revealed unexpected results. For example, it is commonly thought that the US has extremely high mortality rates due to cardiovascular disease. However, upon comparing the US cardiovascular disease mortality with that of Georgia (the country with the highest mean mortality from cardiovascular disease), the US prevalence does not seem nearly as high.

The last two summary statistic codes explore the standard deviations of cardiovascular diseases and diarrheal diseases in years 1990-2015 for all countries, and ranks the standard deviations from high to low. I chose to examine the standard deviation of cardiovascular disease because it has the highest overall mortality in the world. I chose diarrheal diseases because it is more common in third-world countries (countries with low life expectancy); by summarizing the standard deviation, it can be determined which countries have experienced the most fluctuation in mortality prevalence for each disease.

Visualization

Correlation Heat Map

```
numeric<-final_joined%>%select(4:15)
cormap<-round(cor(numeric),2)
library(reshape2)
```

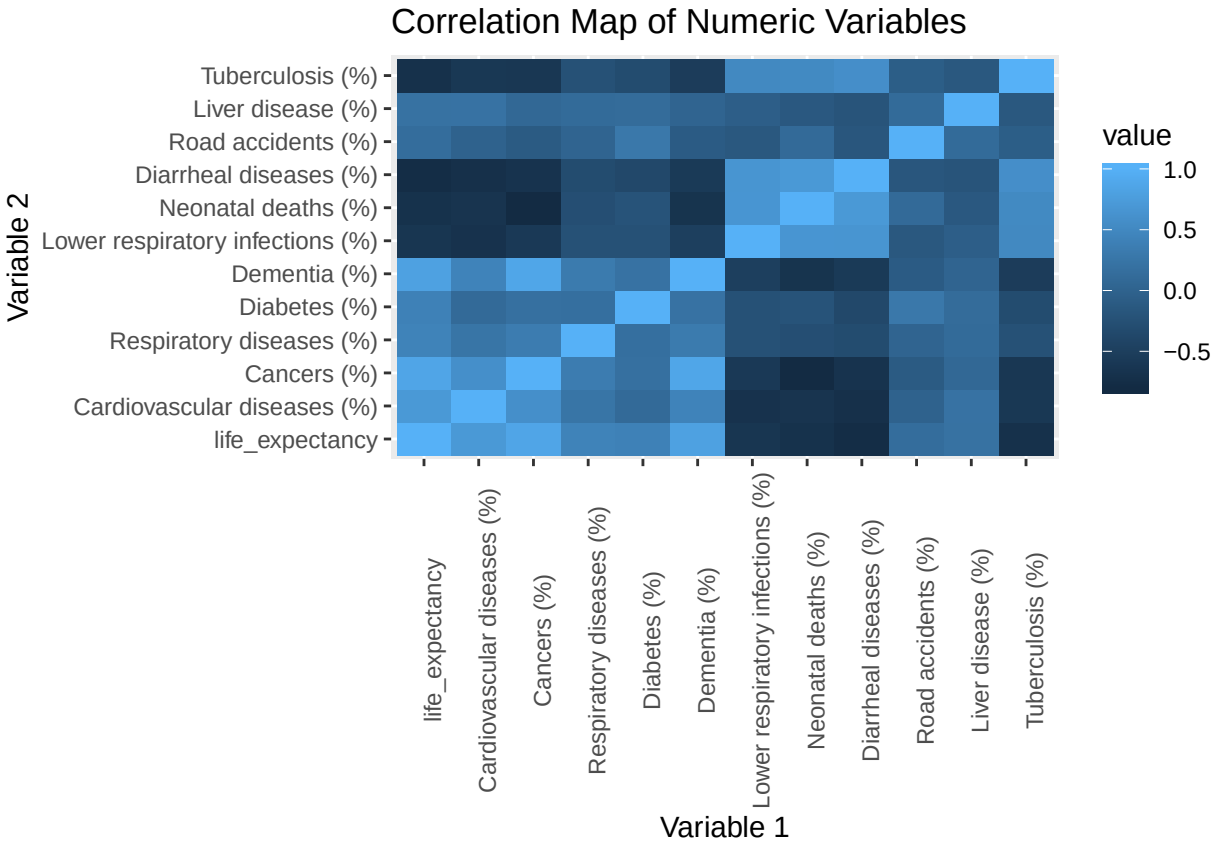
```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
melted_cormap<-melt(cormap)
head(melted_cormap)
```

```
##                           Var1            Var2 value
## 1             life_expectancy life_expectancy  1.00
## 2 Cardiovascular diseases (%) life_expectancy  0.70
## 3                 Cancers (%) life_expectancy  0.85
## 4    Respiratory diseases (%) life_expectancy  0.42
## 5                Diabetes (%) life_expectancy  0.40
## 6                Dementia (%) life_expectancy  0.80
```

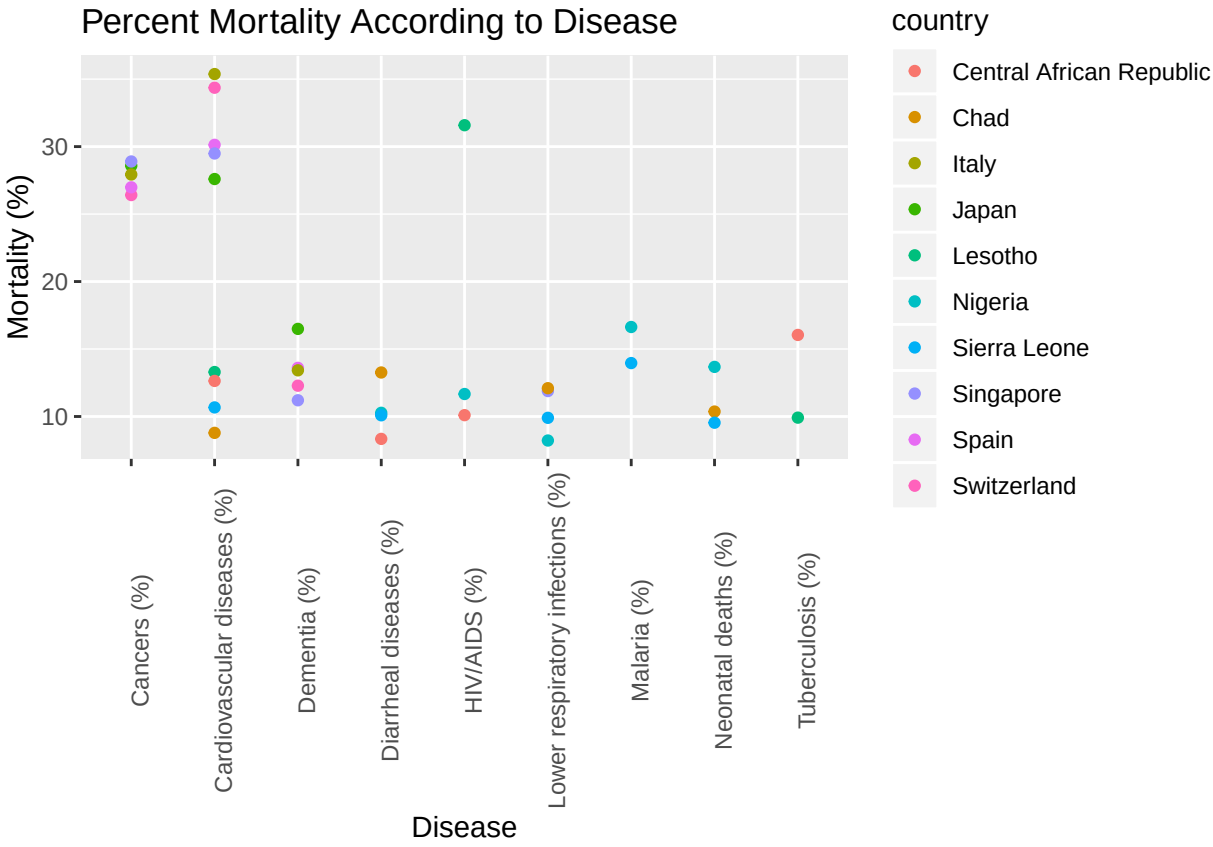
```
ggplot(data=melted_cormap, aes(x=Var1,y=Var2,fill=value))+geom_tile()+ggtitle("Correlation Map of Numeri
```



GGPlots

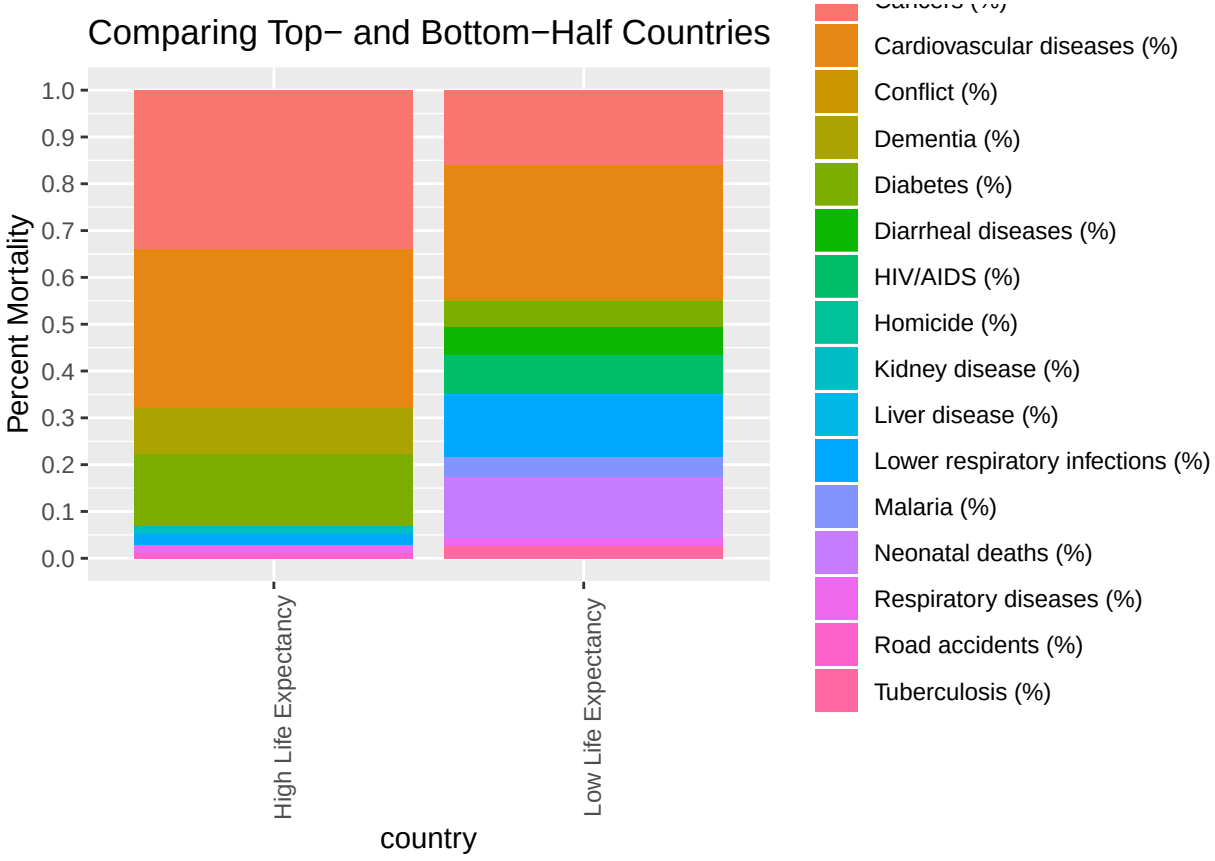
In order to graph with three variables (disease, prevalence, and country), it was necessary for me to pivot the dataset longer. Because there are so many different countries in the dataset, I decided to use only 10 total–the five countries with highest life expectancies and the five countries with lowest life expectancy. Additionally, due to the large number of mortality factors in this dataset, I removed all factors which had lower than 8% mortality rate in order to prevent a cluttered graph.

```
longdata<-final_joined%>%pivot_longer(cols=c(5:36),names_to="diseases",values_to="values")
highnlow<-longdata%>%group_by(country)%>%filter(year==2015 & values>8)%>%arrange(desc(life_expectancy))
topbottom5<-highnlow%>%filter(life_expectancy>82.8 | life_expectancy<54)
ggplot(data=topbottom5, aes(x=diseases,y=values,color=country))+geom_point()+theme(axis.text.x = elemen
```



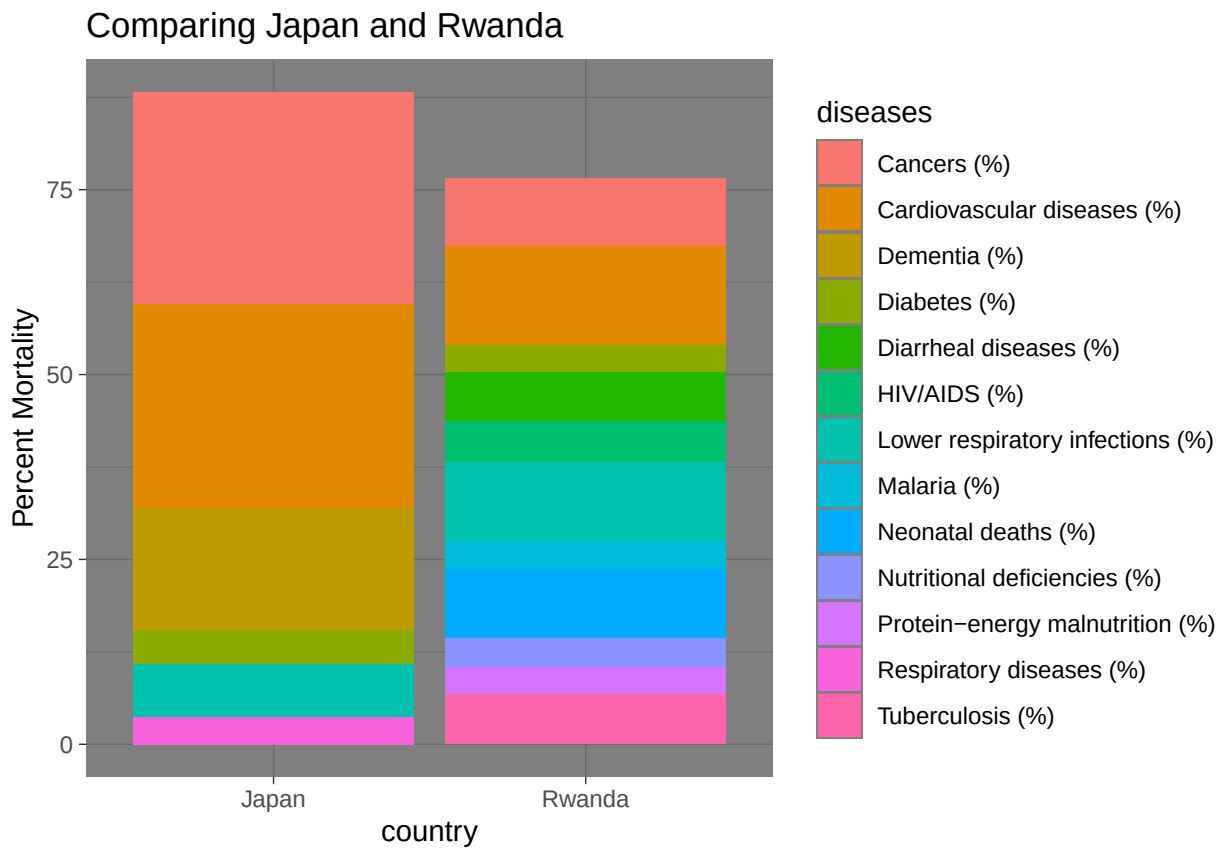
In the graph above, the five countries with the highest life expectancy in the year 2015 was plotted along with the five countries with the lowest life expectancy in the year 2015. Because I only selected values which are higher than 8% mortality, not all countries are represented for each disease. However, there are still some interesting findings; for example, Lesotho's mortality prevalence for HIV/AIDS is significantly higher than other countries who have similar life expectancy values. Additionally, there is an exactly even split between the top five and bottom five countries in regards to cardiovascular disease. This will be explored further in subsequent graphs.

```
highnlow$country <- ifelse(highnlow$life_expectancy<median(highnlow$life_expectancy),"Low Life Expectan
ggplot(data=highnlow, aes(x=country,fill=diseases))+geom_bar(position="fill")+theme(axis.text.x = eleme
```



The plot above places all countries above the median life expectancy in one group (High Life Expectancy) and those below the median life expectancy in another group (Low Life Expectancy). It then shows the relative mortality prevalence in each of the selected diseases. This graph is interesting because it shows that many diseases are much more prevalent in the low life expectancy countries–such as neonatal deaths and lower respiratory infections. Conversely, it shows that the mortality prevalence of cardiovascular disease (the top color on the bars) is much higher in the high life expectancy countries.

```
highnlow1<-longdata%>%group_by(country)%>%filter(year==2015 & values>3)%>%arrange(desc(life_expectancy)
JR<-highnlow1%>%filter(country=="Japan"|country=="Rwanda" & values>3)
ggplot(data=JR, aes(x=country, y=values, fill=diseases))+geom_bar(stat="summary",fun.y="mean")+theme_da
```



The graph above compares two specific countries–one with the highest life expectancy (Japan) and one with the lowest life expectancy (Rwanda). It is important to note that neither country's bar reaches 100%; this is because I excluded all mortality factors that had values lower than 8%. Just as the previous graph, this one shows relative mortality prevalence of disease in a high life expectancy country along with a low life expectancy country. This graph is relevant because the two bars look nothing alike, meaning that each respective countries faces mortality from completely different factors. Most of Japan's mortality comes from cancers, cardiovascular disease, and dementia, and suffers insignificantly from all other factors (which is why its bar is closer to 100%). Conversely, Rwanda suffers mortality much more evenly from all the diseases/factors listed (which is why a larger percentage of data is missing).

Dimensionality Reduction

For clustering, I wanted to examine the relationship between cardiovascular disease (a high mortality factor in all countries) and life expectancy. I chose cardiovascular disease from all other mortality factors because of the interesting implications found in the plots above. First, I first had to access the "final_joined" dataset and change the inconvenient column name "Cardiovascular diseases (%)" to "cardio_diseases" because R is not able to run column names with spaces in them, apparently. After this, I used the silhouette method to determine how many clusters would be sufficient for this data. The highest silhouette width was 2. Typically, I would want more clusters; however, I want to continue making the distinction between low life expectancy countries and high life expectancy countries, as I did in the plots above.

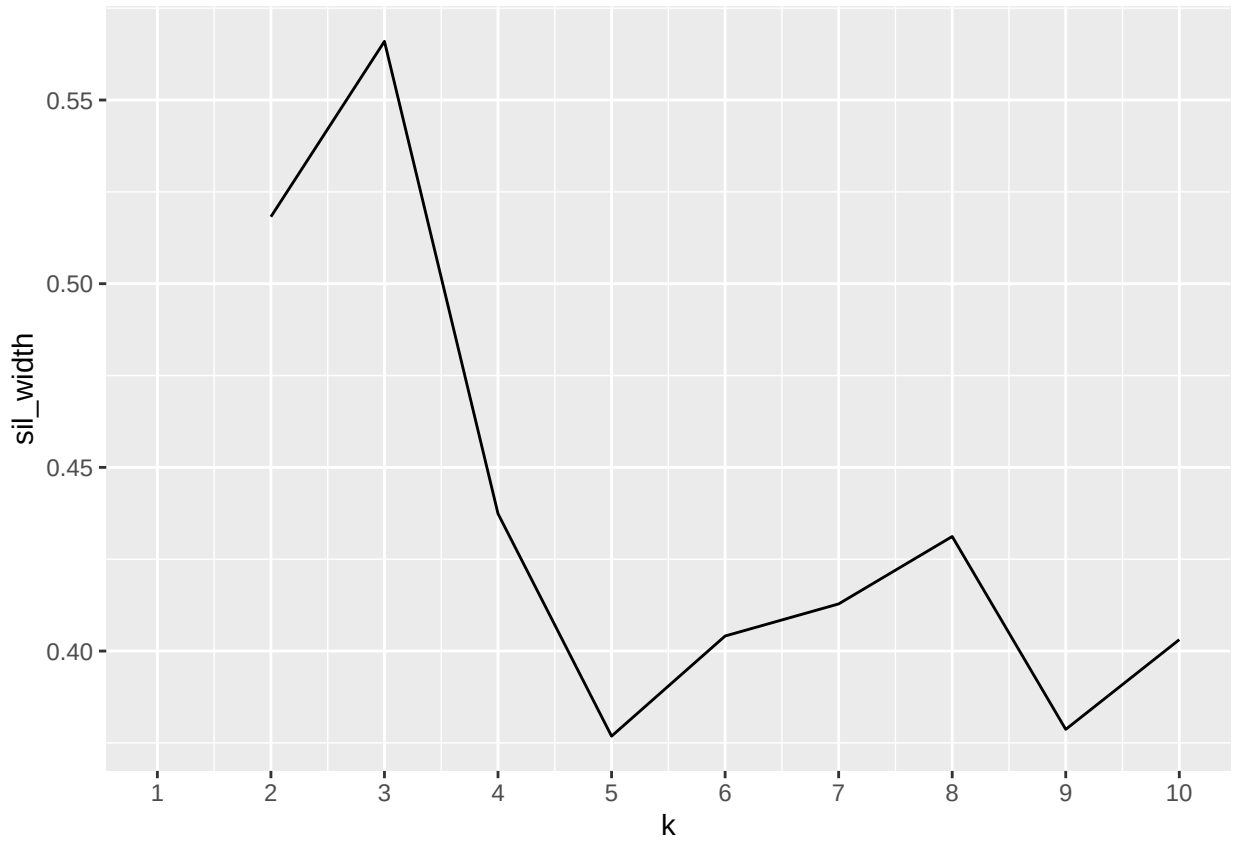
```
names(final_joined)[5]<-"cardio_disease"
names(final_joined)
```

```
##  [1] "country"
```

```
##  [2] "code"
##  [3] "year"
##  [4] "life_expectancy"
##  [5] "cardio_disease"
##  [6] "Cancers (%)"
##  [7] "Respiratory diseases (%)"
##  [8] "Diabetes (%)"
##  [9] "Dementia (%)"
## [10] "Lower respiratory infections (%)"
## [11] "Neonatal deaths (%)"
## [12] "Diarrheal diseases (%)"
## [13] "Road accidents (%)"
## [14] "Liver disease (%)"
## [15] "Tuberculosis (%)"
## [16] "Kidney disease (%)"
## [17] "Digestive diseases (%)"
## [18] "HIV/AIDS (%)"
## [19] "Suicide (%)"
## [20] "Malaria (%)"
## [21] "Homicide (%)"
## [22] "Nutritional deficiencies (%)"
## [23] "Meningitis (%)"
## [24] "Protein-energy malnutrition (%)"
## [25] "Drowning (%)"
## [26] "Maternal deaths (%)"
## [27] "Parkinson disease (%)"
## [28] "Alcohol disorders (%)"
## [29] "Intestinal infectious diseases (%)"
## [30] "Drug disorders (%)"
## [31] "Hepatitis (%)"
## [32] "Fire (%)"
## [33] "Heat-related (hot and cold exposure) (%)"
## [34] "Natural disasters (%)"
## [35] "Conflict (%)"
## [36] "Terrorism (%)"
```

```r
library(cluster)
clust_dat<-final_joined%>%filter(year==2015)%>%select(cardio_disease,life_expectancy)
sil_width<-vector()
for(i in 2:10){
  kms <- kmeans(clust_dat,centers=i)
  sil <- silhouette(kms$cluster,dist(clust_dat))
  sil_width[i]<-mean(sil[,3])
}
ggplot()+geom_line(aes(x=1:10,y=sil_width))+scale_x_continuous(name="k",breaks=1:10)
```

```
## Warning: Removed 1 rows containing missing values (geom_path).
```




```
kmeans1<-clust_dat%>%kmeans(2)
kmeans1
```

```
## K-means clustering with 2 clusters of sizes 110, 61
##
## Cluster means:
##   cardio_disease life_expectancy
## 1       37.92568        76.10282
## 2       16.63499        63.02254
##
## Clustering vector:
##   [1] 2 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 2 2 1 1 1 2 2 2 2 1 2 2 1 1 1 2
##  [36] 2 1 1 1 1 1 2 1 2 1 2 1 2 2 2 1 2 1 1 1 2 2 1 1 2 1 1 2 2 2 1 2 1 1 1
##  [71] 2 1 1 1 1 1 1 1 1 1 1 2 1 1 2 1 1 2 2 1 1 1 1 2 2 1 1 2 1 2 1 2 1 1 1
## [106] 2 2 2 2 1 1 1 2 2 1 1 1 1 1 2 1 2 1 1 1 1 1 1 2 1 2 1 2 1 1 1 2 2 1 2
## [141] 1 1 2 1 2 1 1 1 1 1 2 2 2 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 2 2 2
##
## Within cluster sum of squares by cluster:
## [1] 12854.581  4362.551
##  (between_SS / total_SS =  58.7 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

```
kmeans1$size
```

```
## [1] 110  61
```

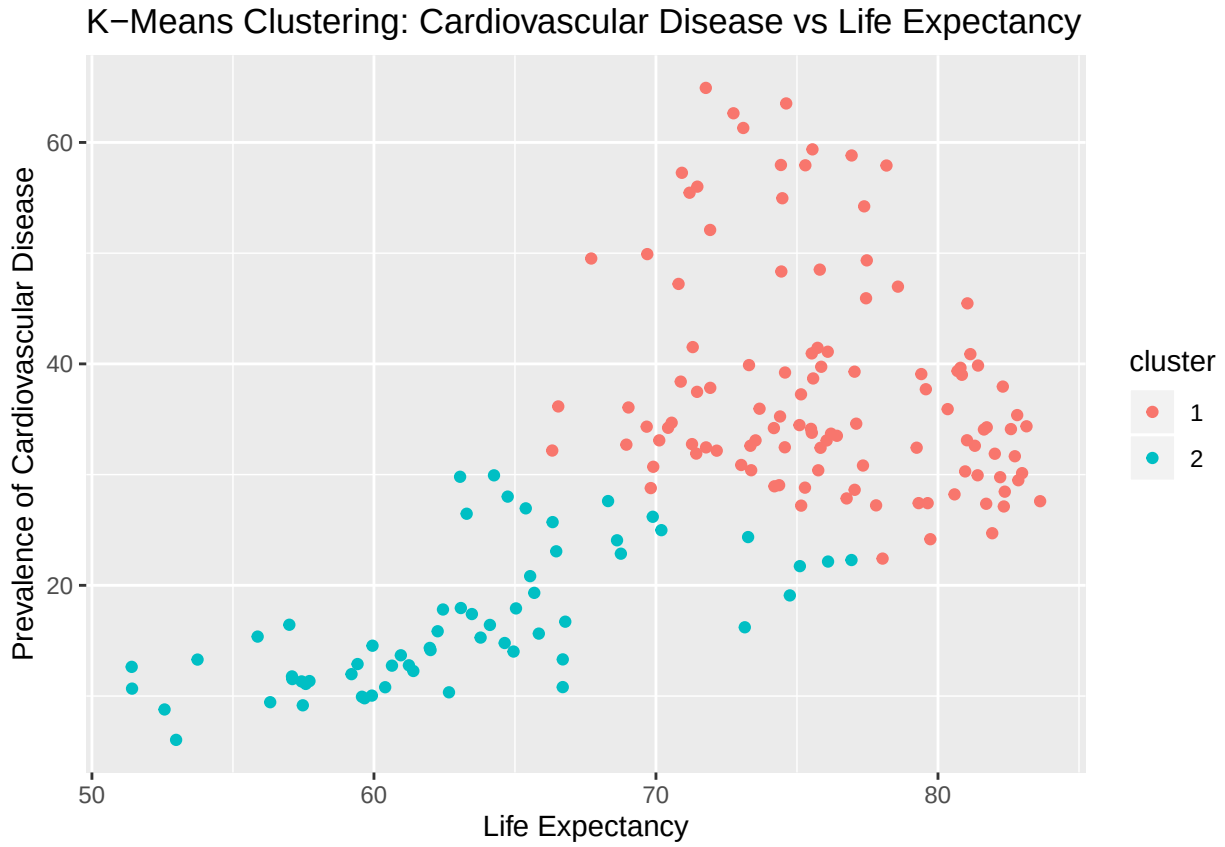
```
kmeans1$centers
```

```
##   cardio_disease life_expectancy
## 1       37.92568        76.10282
## 2       16.63499        63.02254
```

```
kmeans1$cluster
```

```
##   [1] 2 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 2 2 1 1 1 2 2 2 2 1 2 2 1 1 1 2
##  [36] 2 1 1 1 1 1 2 1 2 1 2 1 2 2 2 1 2 1 1 1 2 2 1 1 2 1 1 2 2 2 1 2 1 1 1
##  [71] 2 1 1 1 1 1 1 1 1 1 1 2 1 1 2 1 1 2 2 1 1 1 1 2 2 1 1 2 1 2 1 2 1 1 1
## [106] 2 2 2 2 1 1 1 2 2 1 1 1 1 1 2 1 2 1 1 1 1 1 1 2 1 2 1 2 1 1 1 2 2 1 2
## [141] 1 1 2 1 2 1 1 1 1 1 2 2 2 1 1 1 1 2 1 1 1 1 1 1 1 1 1 2 2 2
```

```
kmeans1$cluster<-as.factor(kmeans1$cluster)
clust_dat<-as.data.frame(clust_dat)
kmeansclust <- clust_dat %>% mutate(cluster=kmeans1$cluster)
ggplot(data=kmeansclust,(aes(x=life_expectancy, y=cardio_disease, color=cluster))) + geom_point()+ggtitl
```



As can be deduced from the graph above, a high mortality rate for cardiovascular disease is mostly associated with countries who have higher life expectancies (this is the blue cluster). The opposite is also true: low percent mortality in regards to cardiovascular disease is associated with countries who have low life expectancies (this is the red cluster). Ultimately, this makes sense because the countries who have low life expectancies suffer a greater overall mortality percentage to a variety of different factors; this can be observed in the bar charts above. For countries who have high life expectancy, there are a smaller number of mortality factors which

significantly affect life expectancy. This alludes to the healthcare quality that exists in low life expectancy countries and high life expectancy countries. In low life expectancy countries, there are many more factors that can cause mortality, and at higher percentages, than in a high life expectancy country. This conclusion supports the cluster graph above.