Project 1

Allison Howe

Title and Introduction

I chose the datasets from RStudio “USMortality” and “USRegionalMortality”, both coming from the ‘lattice’ packaage. I think they are interesting to me because I have always wondered about the contributing factors to deaths - whether being a part of urban or rural regions altered your fate or whether one cause was more prevalent than another. Each row represents different individuals and are clustered and sorted by cause of death and type of region in both packages, allowing the two packages to be joined by those two keys. They contain both numeric variables (region status, gender, cause) and categorical variables (rate and SE). I could expect there to be a potential relationship between the region status and rate due to the likelihood that rural regions are not as developed and advanced in healthcare. I could also expect there to be a trend in females having lower rates due to women being more active in seeking healthcare. In this exploratory data analysis, I will answer the question of which gender has a higher rate for the cause of death of “Heart disease” in HHS Region 01.

```
# If working on your own computer, remember to install new packages with install.packages("name") in the console
library(lattice)
```

Let’s take a quick look at the datasets:

```
# Take a quick look
head(USMortality)
```

```
##    Status    Sex         Cause  Rate  SE
## 1   Urban   Male Heart disease 210.2 0.2
## 2   Rural   Male Heart disease 242.7 0.6
## 3   Urban Female Heart disease 132.5 0.2
## 4   Rural Female Heart disease 154.9 0.4
## 53  Urban   Male        Cancer 195.9 0.2
## 54  Rural   Male        Cancer 219.3 0.5
```

```
head(USRegionalMortality)
```

```
##           Region Status    Sex         Cause  Rate  SE
## 5  HHS Region 01  Urban   Male Heart disease 188.2 1.0
## 6  HHS Region 01  Rural   Male Heart disease 199.1 2.6
## 7  HHS Region 01  Urban Female Heart disease 115.1 0.6
## 8  HHS Region 01  Rural Female Heart disease 124.5 1.7
## 9  HHS Region 02  Urban   Male Heart disease 226.8 0.8
## 10 HHS Region 02  Rural   Male Heart disease 248.8 3.3
```

The `USMortality` dataset contains information about causes of deaths per male/female in both rural and urban regions.

The goal of the project is to compare causes of deaths across regions and gender, first joining the data.

Joining/Merging

Since the two datasets were already tidy, I decided to join the two datasets into a single one using the “Status” and “Cause” ID variables that are common for both. Because I wanted to keep all observations except unmatching rows, I decided to use the inner_join function. There are a total of 40 observations in ‘USMortality’ and a total of 400 observations in ‘USRegionalMortality’ before joining the two. There are 3 IDs the two datasets have in common and there is one ID that is found in ‘USRegionalMortality’ but not ‘USMortality’.

```
#combining the two datasets by "Status" and "Cause"
mortality <- inner_join(USRegionalMortality, USMortality, by = c("Status", "Cause"))
```

```
## Error in inner_join(USRegionalMortality, USMortality, by = c("Status", : could not fi
nd function "inner_join"
```

```
mortality
```

```
## Error in eval(expr, envir, enclos): object 'mortality' not found
```

After joining, there were 689 rows omitted, but no IDs were left out. This could cause issues depending on which cause I was wanting to investigate because I would not have access to viewing the correlating data. It would also create problems if I were wanting to compare/contrast data across causes.

Tidying

Next, I chose to tidy the data using the pivot_longer function.

```
#tidying dataframe from wide to long
mortality2 <- mortality %>% pivot_longer(cols = c('Rate.x', 'Rate.y'), names_to = mortal
ity$Rate, values_to = 'Rate') %>% pivot_longer(cols = c('SE.x', 'SE.y'), names_to = mort
ality$SE, values_to = 'SE') %>% pivot_longer(cols = c('Sex.x', 'Sex.y'), names_to = mort
ality$Sex, values_to = 'Sex')
```

```
## Error in mortality %>% pivot_longer(cols = c("Rate.x", "Rate.y"), names_to = mortalit
y$Rate, : could not find function "%>%"
```

```
mortality2
```

```
## Error in eval(expr, envir, enclos): object 'mortality2' not found
```

This function was necessary in order to join the values of “Rate.x” and “Rate.y” into one column labeled “Rate” and the values of “SE.x” and “SE.y” into one column labeled “SE”. I also wanted to clean up the data by getting rid of the other duplicate columns “Sex.x” and “Sex.y” and joined them together under

“Sex” using the same tidyr function.

Wrangling

I explored my data with summary tables and statistics using all 6 core dplyr functions.

```
#slicing function
mortality3 <- mortality2 %>% filter(Cause == "Heart disease" & Region == "HHS Region 0
1") %>%
  slice(1:50)
```

```
## Error in mortality2 %>% filter(Cause == "Heart disease" & Region == "HHS Region 01")
%>% : could not find function "%>%"
```

```
#arranging in descending Rate order
mortality4 <- mortality3 %>% arrange(desc(Rate))
```

```
## Error in mortality3 %>% arrange(desc(Rate)): could not find function "%>%"
```

```
#selecting order of variables
mortality5 <- mortality4 %>% select(Region, Status, Sex, Cause, Rate, SE)
```

```
## Error in mortality4 %>% select(Region, Status, Sex, Cause, Rate, SE): could not find
function "%>%"
```

```
#mutating categorical variable into numerical
mortality5 <- mortality5 %>% mutate(Sex2 = recode(Sex, Male = 1, Female = 2))
```

```
## Error in mortality5 %>% mutate(Sex2 = recode(Sex, Male = 1, Female = 2)): could not f
ind function "%>%"
```

```
#summarizing variables
mortality5 %>% summarize(Sex)
```

```
## Error in mortality5 %>% summarize(Sex): could not find function "%>%"
```

```
mortality5 %>% summarize(mean(Rate, na.rm=T), n(), n_distinct(Region))
```

```
## Error in mortality5 %>% summarize(mean(Rate, na.rm = T), n(), n_distinct(Region)): co
uld not find function "%>%"
```

```
#grouping variables by status and different summaries
mortality5 %>% group_by(Status) %>% summarize(mean(Rate, na.rm=T), n(), n_distinct(Regio
n))
```

```
## Error in mortality5 %>% group_by(Status) %>% summarize(mean(Rate, na.rm = T), : could
not find function "%>%"
```

```
mortality5 %>% group_by(Status) %>% summarize(mean(SE, na.rm=T), n(), n_distinct(Regio
n))
```

```
## Error in mortality5 %>% group_by(Status) %>% summarize(mean(SE, na.rm = T), : could n
ot find function "%>%"
```

When using the mutate function, I created a numeric variable from a categorical variable. I summarized the categorical variable “Sex” as well as the two numerical variables means “Rate” and “SE”. I found for rural regions the mean rate was 191 deaths per year while the mean rate for urban regions was 162 deaths per year. I also found rural SE mean was 1.57 and urban SE mean was 0.5.

Visualizing

What about missing years for some countries? Those would not appear explicitly in the dataset, they just would not be there. Using `dplyr` functions, find the total number of distinct years for each country in `tidy_who`. Also report the minimum and maximum year contained in the dataset for each country. Which countries had less than the expected 34 years (1980 to 2013)? Why do you think these years are missing? *Note: To understand why we have missing years, look at Serbia & Montenegro. What happened to this country in 2005?*

```
#visualizing numerical variable Rate
ggplot(data = mortality5, mapping = aes(x = Rate)) +
  geom_bar(fill = "blue") +
  labs(title = "Mortality Rate for Region 1", x = "Rate")
```

```
## Error in ggplot(data = mortality5, mapping = aes(x = Rate)): could not find function
"ggplot"
```

```
#visualizing numerical variable Rate and categorical variable Sex
ggplot(data = mortality5, mapping = aes(x = Sex, y = Rate)) +
  geom_point(color = "blue") +
  labs(title = "Mortality Rate for Region 1 based on Sex", x = "Sex", y = "Rate")
```

```
## Error in ggplot(data = mortality5, mapping = aes(x = Sex, y = Rate)): could not find
function "ggplot"
```

```
#visualizing Rates in Region 1 by Sex
gplot(mortality5, aes(fill = Sex, x = Region, y = Rate)) +
  geom_boxplot(position = "dodge") +
  labs(title = "mortality rates in region 1 by sex")
```

```
## Error in gplot(mortality5, aes(fill = Sex, x = Region, y = Rate)): could not find fun
ction "gplot"
```

Plot 1: This plot shows the mortality rates for Region 1.It is evident that there is much variation within it and there is not much distribution.

Plot 2: This plot shows the mortality rates based upon the two Sexes.It is evident that there is much variation within it and there is only a few similarities between the two values.

Plot 3: This plot shows the mortality rates in region 1 between the two sexes.It is evident that there is much variation within it and there is not a lot of similarities between the variables.

Discussion

I found that Males have a higher rate of mortality than females do within Region 1. This is highly supported by Plot 3 and the summarized findings from the wrangling section that put male statistics higher than females. This project was challenging doing it alone because I had to rely on only myself, but very rewarding as it allowed me to understand more in-depth the variations you can create with RStudio. I learned from this process that it is okay to ask for help, but it is also important to push yourself and try to accomplish what you may not have thought you previously could have done. It also allowed me to b