Data wrangling

Load Packages

```
require(mosaic)
require(tidyverse)
require(lubridate)
require(rvest)
require(lme4)
```

Load Datasets

```
wd <- getwd()
regions <- read_csv(paste0(wd,"/Data/NOC_Region.csv"))
athletes <- read_csv(paste0(wd,"/Data/Olympics_Athletes.csv"))
countrycode <- read_csv(paste0(wd,"/Data/ISOCountryCode.csv"))
temphistory <- readxl::read_xlsx(paste0(wd,"/Data/Historical_Temp_Data.xlsx"), sheet = 2)
olymweat <- readxl::read_xlsx(paste0(wd,"/Data/Olympics_Weather_Data.xlsx"))
```

Data Previews

```
head(regions)
```

```
## # A tibble: 6 x 3
##   NOC   region      notes
##   <chr> <chr>       <chr>
## 1 AFG   Afghanistan <NA>
## 2 AHO   Curacao     Netherlands Antilles
## 3 ALB   Albania     <NA>
## 4 ALG   Algeria     <NA>
## 5 AND   Andorra     <NA>
## 6 ANG   Angola      <NA>
```

```
head(athletes)
```

```
## # A tibble: 6 x 15
##      ID Name  Sex     Age Height Weight Team  NOC   Games  Year Season
##   <int> <chr> <chr> <int>  <int>  <int> <chr> <chr> <chr> <int> <chr>
## 1    22 Andr~ F        22    170    125 Roma~ ROU   2016~  2016 Summer
## 2    51 Nsto~ M        23    167     64 Spain ESP   2016~  2016 Summer
## 3    51 Nsto~ M        23    167     64 Spain ESP   2016~  2016 Summer
## 4    51 Nsto~ M        23    167     64 Spain ESP   2016~  2016 Summer
## 5    51 Nsto~ M        23    167     64 Spain ESP   2016~  2016 Summer
## 6    51 Nsto~ M        23    167     64 Spain ESP   2016~  2016 Summer
## # ... with 4 more variables: City <chr>, Sport <chr>, Event <chr>,
## #   Medal <chr>
```

```
head(countrycode)
```

```
## # A tibble: 6 x 11
```

```
##   name  `alpha-2` `alpha-3` `country-code` `iso_3166-2` region `sub-region`
##   <chr> <chr>     <chr>     <chr>          <chr>        <chr>  <chr>
## 1 Afgh~ AF        AFG       004            ISO 3166-2:~ Asia   Southern As~
## 2 "\xc~ AX        ALA       248            ISO 3166-2:~ Europe Northern Eu~
## 3 Alba~ AL        ALB       008            ISO 3166-2:~ Europe Southern Eu~
## 4 Alge~ DZ        DZA       012            ISO 3166-2:~ Africa Northern Af~
## 5 Amer~ AS        ASM       016            ISO 3166-2:~ Ocean~ Polynesia
## 6 Ando~ AD        AND       020            ISO 3166-2:~ Europe Southern Eu~
## # ... with 4 more variables: `intermediate-region` <chr>,
## #   `region-code` <chr>, `sub-region-code` <chr>,
## #   `intermediate-region-code` <chr>
```

```
head(temphistory)
```

```
## # A tibble: 6 x 14
##   ISO_3DIGIT Jan_Temp Feb_temp Mar_temp Apr_Temp May_temp Jun_Temp
##   <chr>         <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
## 1 AFG          0.0731     2.11    7.60     13.4      18.2    23.2
## 2 AGO         22.6       22.7    22.8      22.4      20.7    18.4
## 3 ALB          2.02       3.22    6.04      9.92     14.4    17.9
## 4 ARE         18.4       19.4    22.6      26.6      30.6    32.5
## 5 ARG         20.8       19.9    17.5      14.0      10.6     7.66
## 6 ARM         -8.66      -6.65   -0.566     6.62     11.4    15.6
## # ... with 7 more variables: July_Temp <dbl>, Aug_Temp <dbl>,
## #   Sept_temp <dbl>, Oct_temp <dbl>, Nov_Temp <dbl>, Dec_temp <dbl>,
## #   Annual_temp <dbl>
```

```
head(olymweat, 20)
```

```
## # A tibble: 20 x 11
##     Year City  Season `Start Date`        `End Date`
##    <dbl> <chr> <chr>  <dttm>              <dttm>
##  1  2016 Rio ~ Summer 2016-08-05 00:00:00 2016-08-21 00:00:00
##  2  2012 Lond~ Summer 2012-07-27 00:00:00 2012-08-12 00:00:00
##  3  2008 Beij~ Summer 2008-08-08 00:00:00 2008-08-24 00:00:00
##  4  2004 Athe~ Summer 2004-08-13 00:00:00 2004-08-29 00:00:00
##  5  2000 Sydn~ Summer 2000-09-15 00:00:00 2000-10-01 00:00:00
##  6  1996 Atla~ Summer 1996-07-19 00:00:00 1996-08-09 00:00:00
##  7  1992 Barc~ Summer 2018-07-25 00:00:00 2018-08-09 00:00:00
##  8  1988 Seoul Summer 2018-09-17 00:00:00 2018-10-02 00:00:00
##  9  1984 Los ~ Summer 2018-07-28 00:00:00 2018-08-12 00:00:00
## 10  1980 Mosc~ Summer 2018-07-19 00:00:00 2018-08-03 00:00:00
## 11  1976 Mont~ Summer 2018-07-17 00:00:00 2018-08-01 00:00:00
## 12  1972 Muni~ Summer 2018-08-26 00:00:00 2018-09-10 00:00:00
## 13  1968 Mexi~ Summer 2018-10-12 00:00:00 2018-10-27 00:00:00
## 14  1964 Tokyo Summer 2018-10-10 00:00:00 2018-10-24 00:00:00
## 15  1960 Rome  Summer 2018-08-25 00:00:00 2018-09-11 00:00:00
## 16  1956 Melb~ Summer 2018-11-22 00:00:00 2018-12-08 00:00:00
## 17  1952 Hels~ Summer 2018-07-19 00:00:00 2018-08-03 00:00:00
## 18  1948 Lond~ Summer 2018-07-29 00:00:00 2018-08-14 00:00:00
## 19  1944 Lond~ Summer NA                  NA
## 20  1940 Tokyo Summer NA                  NA
## # ... with 6 more variables: `Number of Days` <dbl>, Altitude <dbl>, `Hist
## #   Avg Temp Mon1 [C]` <dbl>, `Hist Avg Temp Mon2 [C]` <dbl>, `Diff Avg
## #   Temp Mon1 [C]` <dbl>, `Diff Avg Temp Mon2 [C]` <dbl>
```

Combine countrycode and temphistory data

```
# combine countrycode and temphistory
countrytemp <- left_join(temphistory, countrycode, by = c("ISO_3DIGIT" = "alpha-3")) %>%
  select(-c(16:24), -`ISO_3DIGIT`)

# change name of each month column so that wrangling is easier later on
names(countrytemp)[c(1:12)] <- c(1:12)
```

Clean up olymweat dates

```
# clean up dates column
olymweat2 <- olymweat %>%
  rename("StartDate" = `Start Date`,
         "EndDate" = `End Date`) %>%
  mutate(StartDate = as.character(StartDate),
         EndDate = as.character(EndDate),
         StartMonth = strsplit(StartDate, split = "-")[[1]][2],
         EndMonth = strsplit(EndDate, split = "-")[[1]][2],
         StartDay = strsplit(StartDate, split = "-")[[1]][3],
         EndDay = strsplit(EndDate, split = "-")[[1]][3]) %>%
  select(-StartDate, -EndDate)

head(olymweat2)
```

```
## # A tibble: 6 x 13
##    Year City  Season `Number of Days` Altitude `Hist Avg Temp ~
##   <dbl> <chr> <chr>             <dbl>    <dbl>            <dbl>
## 1  2016 Rio ~ Summer               16        7             22.5
## 2  2012 Lond~ Summer               16       19             16.5
## 3  2008 Beij~ Summer               16       46             25.1
## 4  2004 Athe~ Summer               16      231             24.3
## 5  2000 Sydn~ Summer               16        0             15.2
## 6  1996 Atla~ Summer               21      312             24.3
## # ... with 7 more variables: `Hist Avg Temp Mon2 [C]` <dbl>, `Diff Avg
## #   Temp Mon1 [C]` <dbl>, `Diff Avg Temp Mon2 [C]` <dbl>,
## #   StartMonth <chr>, EndMonth <chr>, StartDay <chr>, EndDay <chr>
```

Combine athletes and region

```
athletes2 <- athletes %>%
  left_join(regions, by = c('NOC' = 'NOC')) %>%
  select(-notes, -NOC, -Games)
```

Wrangling elevation data

```
wikiTable <- function(source) {
  read_html(source)%>%
    html_nodes("table.wikitable") %>%
```

```
    html_table(fill=T)%>%
    magrittr::extract2(1)
}
# Read in wikipedia table for average elevation per country
elevation <- wikiTable("https://en.wikipedia.org/wiki/List_of_countries_by_average_elevation#cite_note-

# Clean up elevation data
elelist <- strsplit(elevation$Elevation, "m")
vec <- c()
for (i in 1:length(elelist)){
  vec[i] <- elelist[[i]][1]
}

# Clean up dataframe
elevation$Elevation <- vec
elevation <- elevation %>%
  mutate(Elevation = parse_number(Elevation))
```

Combine olymweat2 and athletes

```
olympics1 <- left_join(athletes2, olymweat2, by = c("Year" = "Year", "Season" = "Season")) %>%
  select(-City.y) %>%
  rename(City = City.x) %>%
  filter(!(Year %in% c(1896,1900,1904,1906,1908,1912,1920,1924,1928) & Season == "Summer")) %>%
  mutate(CityTemp1 = `Hist Avg Temp Mon1 [C]` + `Diff Avg Temp Mon1 [C]`,
         CityTemp2 = `Hist Avg Temp Mon2 [C]` + `Diff Avg Temp Mon2 [C]`) %>%
  select(-`Hist Avg Temp Mon1 [C]`, -`Hist Avg Temp Mon2 [C]`, -`Diff Avg Temp Mon1 [C]`, -`Diff Avg Ter
```

Combine olympics1 with countrytemp

```
olympics2 <- left_join(olympics1, countrytemp, by = c("region" = "name"))

# create vector of the starting and ending months of each olympic
startmon <- as.character(as.integer(olympics1$StartMonth))
endmon <- as.character(as.integer(olympics1$EndMonth))

# loop through olympics2 to gather the temperature of months that corresponds to the olympic months for
# store this data in histtemp1 and histtemp2

histtemp1 <- c()
histtemp2 <- c()
for (i in 1:nrow(olympics2)){
  histtemp1[i] <- olympics2[[i,startmon[i]]]
  histtemp2[i] <- olympics2[[i,endmon[i]]]
}

# add create
olympics2$histtemp1 <- histtemp1
olympics2$histtemp2 <- histtemp2
```

```r
# get rid of temperature data of every month
olympics2 <- olympics2 %>%
  select(-`1`,-`2`,-`3`,-`4`,-`5`,-`6`,-`7`,-`8`,-`9`,-`10`,-`11`,-`12`, -Annual_temp)

olympics2
```

```
## # A tibble: 244,108 x 24
##        ID Name  Sex     Age Height Weight Team   Year Season City  Sport
##     <int> <chr> <chr> <int>  <int>  <int> <chr> <dbl> <chr>  <chr> <chr>
##  1    22 Andr~ F        22    170    125 Roma~  2016 Summer Rio ~ Weig~
##  2    51 Nsto~ M        23    167     64 Spain  2016 Summer Rio ~ Gymn~
##  3    51 Nsto~ M        23    167     64 Spain  2016 Summer Rio ~ Gymn~
##  4    51 Nsto~ M        23    167     64 Spain  2016 Summer Rio ~ Gymn~
##  5    51 Nsto~ M        23    167     64 Spain  2016 Summer Rio ~ Gymn~
##  6    51 Nsto~ M        23    167     64 Spain  2016 Summer Rio ~ Gymn~
##  7    51 Nsto~ M        23    167     64 Spain  2016 Summer Rio ~ Gymn~
##  8    55 Anto~ M        26    170     65 Spain  2016 Summer Rio ~ Athl~
##  9    62 Giov~ M        21    198     90 Italy  2016 Summer Rio ~ Rowi~
## 10    65 Pati~ F        21    165     49 Azer~  2016 Summer Rio ~ Taek~
## # ... with 244,098 more rows, and 13 more variables: Event <chr>,
## #   Medal <chr>, region <chr>, `Number of Days` <dbl>, Altitude <dbl>,
## #   StartMonth <chr>, EndMonth <chr>, StartDay <chr>, EndDay <chr>,
## #   CityTemp1 <dbl>, CityTemp2 <dbl>, histtemp1 <dbl>, histtemp2 <dbl>
```

Calculate difference between host city temperature and home country temperature

```r
# calculate the ratio of days in the first and second month
olympics3 <- olympics2 %>%
  mutate(NumDayMon1 = ifelse(as.integer(EndDay) > `Number of Days`, `Number of Days`, `Number of Days` -
         NumDayMon2 = `Number of Days` - NumDayMon1,
         RatioMon1 = NumDayMon1/`Number of Days`,
         RatioMon2 = 1 - RatioMon1,
         CityTemp2 = ifelse(is.na(CityTemp2), 0, CityTemp2)) %>%
# calculate the total difference
  mutate(tempdiff = (CityTemp1 - histtemp1)*RatioMon1 + (CityTemp2 - histtemp2)*RatioMon2) %>%

# get rid of columns that were used in the calculation for temperature difference
  select(-CityTemp1, -CityTemp2, -histtemp1, -histtemp2, -NumDayMon1, -NumDayMon2, -RatioMon1, -RatioMo
```

Combine elevation data with olympics3 and compute elevation difference

```r
olympics4 <- olympics3 %>%
  left_join(elevation, by = c("region" = "Country")) %>%
  mutate(elevdiff = Altitude - Elevation) %>%
  select(-Elevation, -Altitude)
```

Clean up medal data

```
olympics5 <- olympics4 %>%
  mutate(Medal = ifelse(is.na(Medal), "0", Medal),
         Medal = ifelse(Medal == "Bronze", "1", Medal),
         Medal = ifelse(Medal == "Silver", "2", Medal),
         Medal = ifelse(Medal == "Gold", "3", Medal),
         Medal = as.integer(Medal))


totalmedals <- olympics5 %>%
  group_by(Year,Season) %>%
  summarise(TotMed = sum(Medal))
totalmedals
```

```
## # A tibble: 42 x 3
## # Groups:   Year [?]
##     Year Season TotMed
##    <dbl> <chr>   <int>
##  1  1924 Winter    278
##  2  1928 Winter    177
##  3  1932 Summer   1319
##  4  1932 Winter    188
##  5  1936 Summer   1851
##  6  1936 Winter    217
##  7  1948 Summer   1714
##  8  1948 Winter    265
##  9  1952 Summer   1800
## 10  1952 Winter    270
## # ... with 32 more rows
```

Calculate percentage of medals won by each country

```
olympics6 <- olympics5 %>%
  group_by(Year, Season, region) %>%
  summarise(medalswon = sum(Medal)) %>%
  left_join(totalmedals, by = c("Year" = "Year", "Season" = "Season")) %>%
  mutate(medratio = medalswon / TotMed)
```

Grouped Countries by getting rid of distinctions between each athletes

```
olympics7 <- olympics4 %>%
  select(Year, Season, region, tempdiff, elevdiff)

duplicated <- olympics4 %>%
  select(Year, Season, region, tempdiff, elevdiff) %>%
  duplicated()

olympics8 <- olympics7[!duplicated,]
```

Combined temperature/elevation data with medal data

```
olympics9 <- olympics6 %>%
  left_join(olympics8, by = c("Year" = "Year", "Season" = "Season", "region" = "region"))
```

save olympics9 as R.data

```
save(olympics9, "olympics9", file = "olympics9.Rdata")
```