Part A

Importing a data into R, tidying it and performing a simple meaningful visualization.

Chocolate Bar Ratings is the dataset used - https://www.kaggle.com/datasets/rtatman/chocolate-bar-ratings?select=flavors_of_cacao.csv

Variables: 1. Company (Maker-if known) - Name of the company manufacturing the bar.

2. Specific Bean Originor Bar Name - The specific geo-region of origin for the bar.
3. REF - A value linked to when the review was entered in the database. Higher = more recent.
4. ReviewDate - Date of publication of the review.
5. CocoaPercent - Cocoa percentage (darkness) of the chocolate bar being reviewed.
6. CompanyLocation - Manufacturer base country.
7. Rating - Expert rating for the bar. Rating System: 5= Elite (Transcending beyond the ordinary limits) 4= Premium (Superior flavor development, character and style) 3= Satisfactory(3.0) to praiseworthy(3.75) (well made with special qualities) 2= Disappointing (Passable but contains at least one significant flaw) 1= Unpleasant (mostly unpalatable)
8. BeanType - The variety (breed) of bean used, if provided.
9. Broad BeanOrigin - The broad geo-region of origin for the bean.

Preprocessing: 1. Renamed columns to remove white space and shorter names 2. Changed data type - cocoa_percent was string have “%” symbol. Removed the symbol and converted it to numeric.

```r
library(readr)
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
```

```r
library(tidyr)
library(ggplot2)
```

```r
setwd('..')
dir <- getwd()
```

```r
path <- paste(dir, "flavors_of_cacao.csv", sep="/")
cocoa_data <- read_csv(file=path)
```

```
## Rows: 1795 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (6): Company
## (Maker-if known), Specific Bean Origin
## or Bar Name, Cocoa
## ...
## dbl (3): REF, Review
## Date, Rating
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(cocoa_data)
```

```
## # A tibble: 6 x 9
##   Company \n(Make~1 Speci~2   REF Revie~3 Cocoa~4 Compa~5 Rating Bean\~6 Broad~7
##   <chr>             <chr>   <dbl>   <dbl> <chr>   <chr>    <dbl> <chr>   <chr>
## 1 A. Morin          Agua G~  1876    2016 63%     France    3.75         Sao To~
## 2 A. Morin          Kpime    1676    2015 70%     France    2.75         Togo
## 3 A. Morin          Atsane   1676    2015 70%     France    3            Togo
## 4 A. Morin          Akata    1680    2015 70%     France    3.5          Togo
## 5 A. Morin          Quilla   1704    2015 70%     France    3.5          Peru
## 6 A. Morin          Carene~  1315    2014 70%     France    2.75 Criollo Venezu~
## # ... with abbreviated variable names 1: `Company \n(Maker-if known)`,
## #   2: `Specific Bean Origin\nor Bar Name`, 3: `Review\nDate`,
## #   4: `Cocoa\nPercent`, 5: `Company\nLocation`, 6: `Bean\nType`,
## #   7: `Broad Bean\nOrigin`
```

```
colnames(cocoa_data) <- c("company", "bean_orig", "ref", "review_year", "cocoa_perc", "company_loc", "ra

cocoa_data$cocoa_perc <- gsub("%", "", as.character(cocoa_data$cocoa_perc))
cocoa_data <- transform(cocoa_data, cocoa_perc = as.numeric(cocoa_perc))

head(cocoa_data)
```

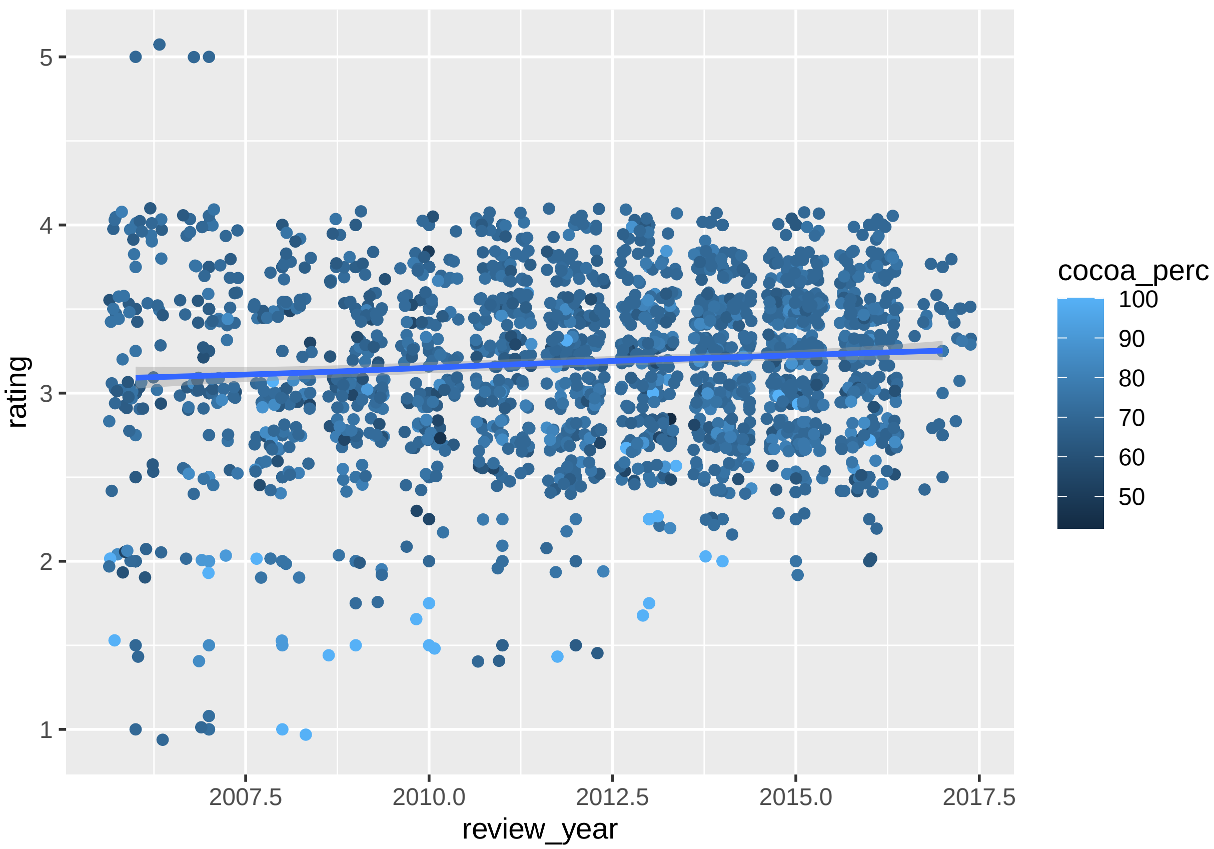
```
##    company   bean_orig  ref review_year cocoa_perc company_loc rating bean_type
## 1 A. Morin Agua Grande 1876        2016         63      France   3.75
## 2 A. Morin       Kpime 1676        2015         70      France   2.75
## 3 A. Morin      Atsane 1676        2015         70      France   3.00
## 4 A. Morin       Akata 1680        2015         70      France   3.50
## 5 A. Morin      Quilla 1704        2015         70      France   3.50
## 6 A. Morin    Carenero 1315        2014         70      France   2.75   Criollo
##   broad_bean_orig
## 1        Sao Tome
## 2            Togo
## 3            Togo
## 4            Togo
## 5            Peru
## 6       Venezuela
```

How the cocoa percentage of chocolate bars change over time? How does that affect ratings?

```
ggplot(cocoa_data, aes(x= review_year, y = rating, color = cocoa_perc)) +
    geom_point() +
    geom_jitter() +
    geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

```
## Warning: The following aesthetics were dropped during statistical transformation: colour
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



There are more reviews each year. It looks like chocolate bars with very high cocoa percents tend to get lower ratings.

Part B

Used data on NCAA student-athlete academic performance. The files include the codebook and tab-delimited data for team-level Academic Progress Rates (APRs) of Division I student-athletes from 2003-2014.

```
path <- paste(dir, "NCAA-D1-APR-2003-14/DS0001/26801-0001-Data.tsv", sep="/")
apr_df_raw <- read_tsv(path, na="-99")
```

```
## Warning: One or more parsing issues, call `problems()` on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```
## Rows: 6511 Columns: 76
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr  (4): SCL_NAME, SPORT_NAME, CONFNAME_14, D1_FB_CONF_14
## dbl (68): SCL_UNITID, SPORT_CODE, ACADEMIC_YEAR, SCL_DIV_14, SCL_SUB_14, SCL...
## lgl  (4): DATA_TAB_GENERALINFO, DATA_TAB_MULTIYRRATE, DATA_TAB_ANNUALRATE, D...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(apr_df_raw)
```

```
## # A tibble: 6 x 76
##   DATA_TAB_GEN~1 SCL_U~2 SCL_N~3 SPORT~4 SPORT~5 ACADE~6 SCL_D~7 SCL_S~8 CONFN~9
##   <lgl>            <dbl> <chr>     <dbl> <chr>     <dbl>   <dbl>   <dbl> <chr>
## 1 NA              100654 Alabam~      20 Women'~    2014       1       2 Southw~
## 2 NA              100654 Alabam~      14 Men's ~    2014       1       2 Southw~
## 3 NA              100654 Alabam~       4 Footba~    2014       1       2 Southw~
## 4 NA              100654 Alabam~       1 Baseba~    2014       1       2 Southw~
## 5 NA              100654 Alabam~      19 Women'~    2014       1       2 Southw~
## 6 NA              100654 Alabam~      33 Women'~    2014       1       2 Southw~
## # ... with 67 more variables: D1_FB_CONF_14 <chr>, SCL_HBCU <dbl>,
## #   SCL_PRIVATE <dbl>, DATA_TAB_MULTIYRRATE <lgl>,
## #   MULTIYR_APR_RATE_1000_RAW <dbl>, MULTIYR_APR_RATE_1000_CI <dbl>,
## #   MULTIYR_APR_RATE_1000_OFFICIAL <dbl>, MULTIYR_ELIG_RATE <dbl>,
## #   MULTIYR_RET_RATE <dbl>, MULTIYR_SQUAD_SIZE <dbl>,
## #   DATA_TAB_ANNUALRATE <lgl>, APR_RATE_2014_1000 <dbl>, ELIG_RATE_2014 <dbl>,
## #   RET_RATE_2014 <dbl>, NUM_OF_ATHLETES_2014 <dbl>, ...
```

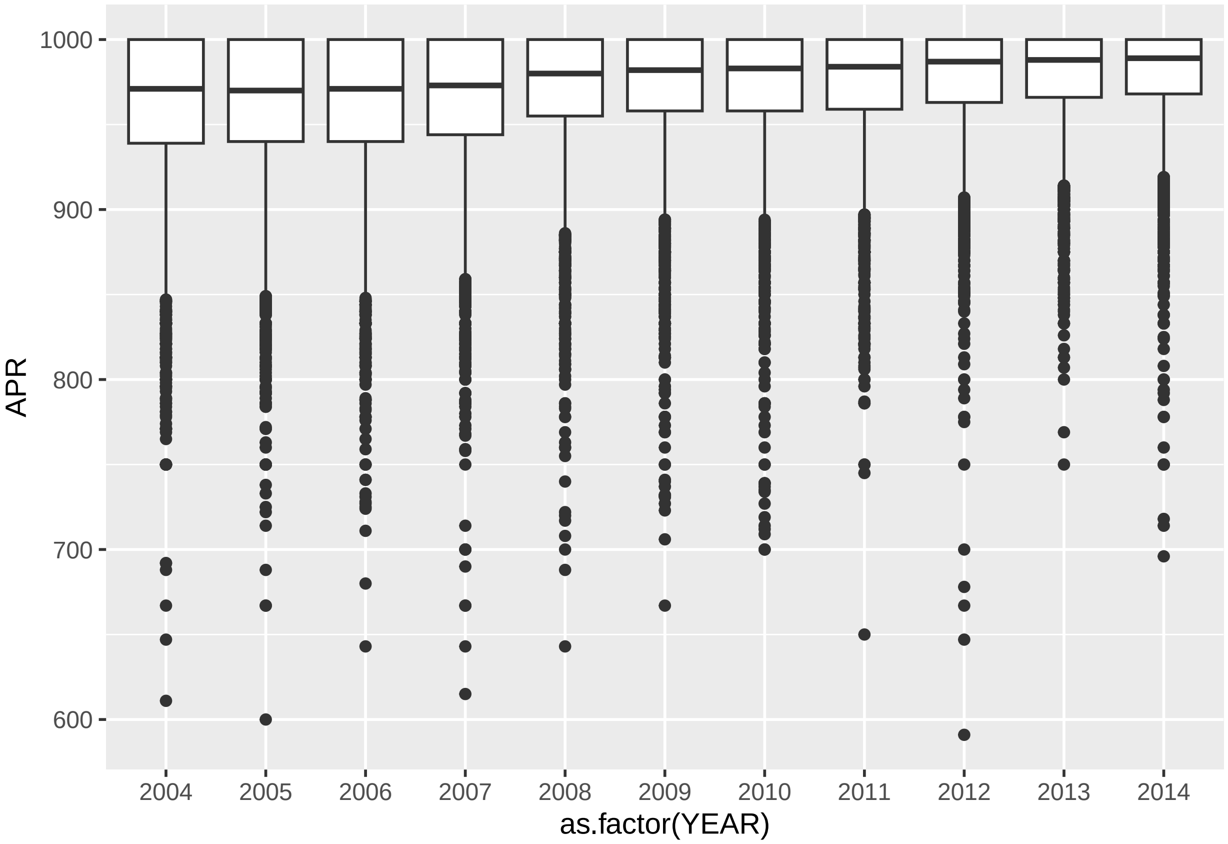
1. Visualizing the distributions of APRs over time.

```
apr_df <- apr_df_raw %>%
  pivot_longer(cols=starts_with("APR_RATE"), names_to="YEAR", values_to="APR") %>%
  select(SCL_UNITID, SCL_NAME, SPORT_CODE, SPORT_NAME, YEAR, APR) %>%
  mutate(YEAR=as.numeric(stringr::str_sub(YEAR, start=10, 13)))
head(apr_df)
```

```
## # A tibble: 6 x 6
##   SCL_UNITID SCL_NAME               SPORT_CODE SPORT_NAME       YEAR   APR
##        <dbl> <chr>                       <dbl> <chr>           <dbl> <dbl>
## 1     100654 Alabama A&M University         20 Women's Bowling  2014  1000
## 2     100654 Alabama A&M University         20 Women's Bowling  2013  1000
## 3     100654 Alabama A&M University         20 Women's Bowling  2012  1000
## 4     100654 Alabama A&M University         20 Women's Bowling  2011  1000
## 5     100654 Alabama A&M University         20 Women's Bowling  2010   950
## 6     100654 Alabama A&M University         20 Women's Bowling  2009  1000
```

```
ggplot(apr_df) + geom_boxplot(aes(x=as.factor(YEAR), y=APR))
```

```
## Warning: Removed 4732 rows containing non-finite values (`stat_boxplot()`).
```



It looks like APR is increasing over time from 2004 to 2014.

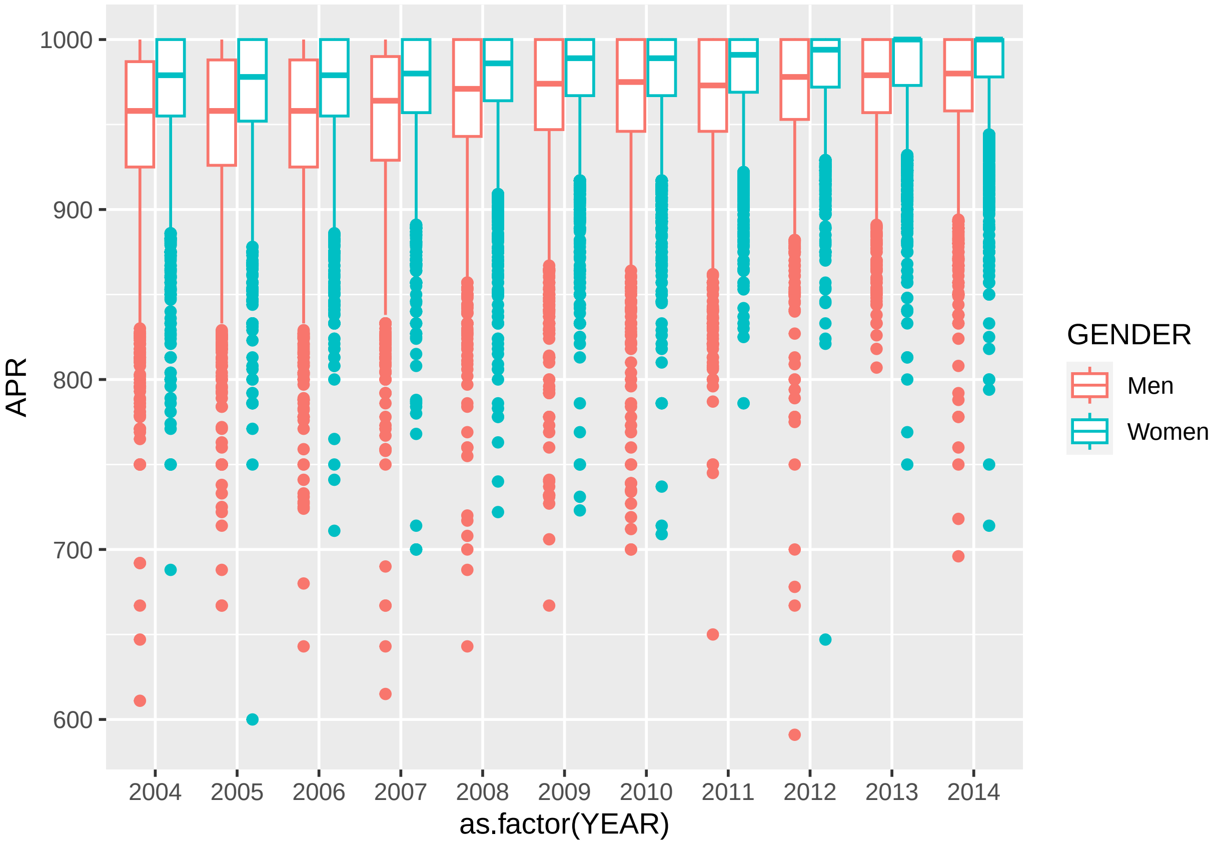
2. Visualizing the distribution of APR over time broken down by gender division:

```
gender_df <- apr_df %>% filter(SPORT_CODE != 38)
gender_df$GENDER <- ifelse(gender_df$SPORT_CODE < 19, "Men", "Women")
head(gender_df)
```

```
## # A tibble: 6 x 7
##   SCL_UNITID SCL_NAME               SPORT_CODE SPORT_NAME      YEAR   APR GENDER
##        <dbl> <chr>                       <dbl> <chr>          <dbl> <dbl> <chr>
## 1     100654 Alabama A&M University         20 Women's Bowli~  2014  1000 Women
## 2     100654 Alabama A&M University         20 Women's Bowli~  2013  1000 Women
## 3     100654 Alabama A&M University         20 Women's Bowli~  2012  1000 Women
## 4     100654 Alabama A&M University         20 Women's Bowli~  2011  1000 Women
## 5     100654 Alabama A&M University         20 Women's Bowli~  2010   950 Women
## 6     100654 Alabama A&M University         20 Women's Bowli~  2009  1000 Women
```

```
ggplot(gender_df) + geom_boxplot(aes(x=as.factor(YEAR), y=APR, color=GENDER))
```

```
## Warning: Removed 4696 rows containing non-finite values (`stat_boxplot()`).
```

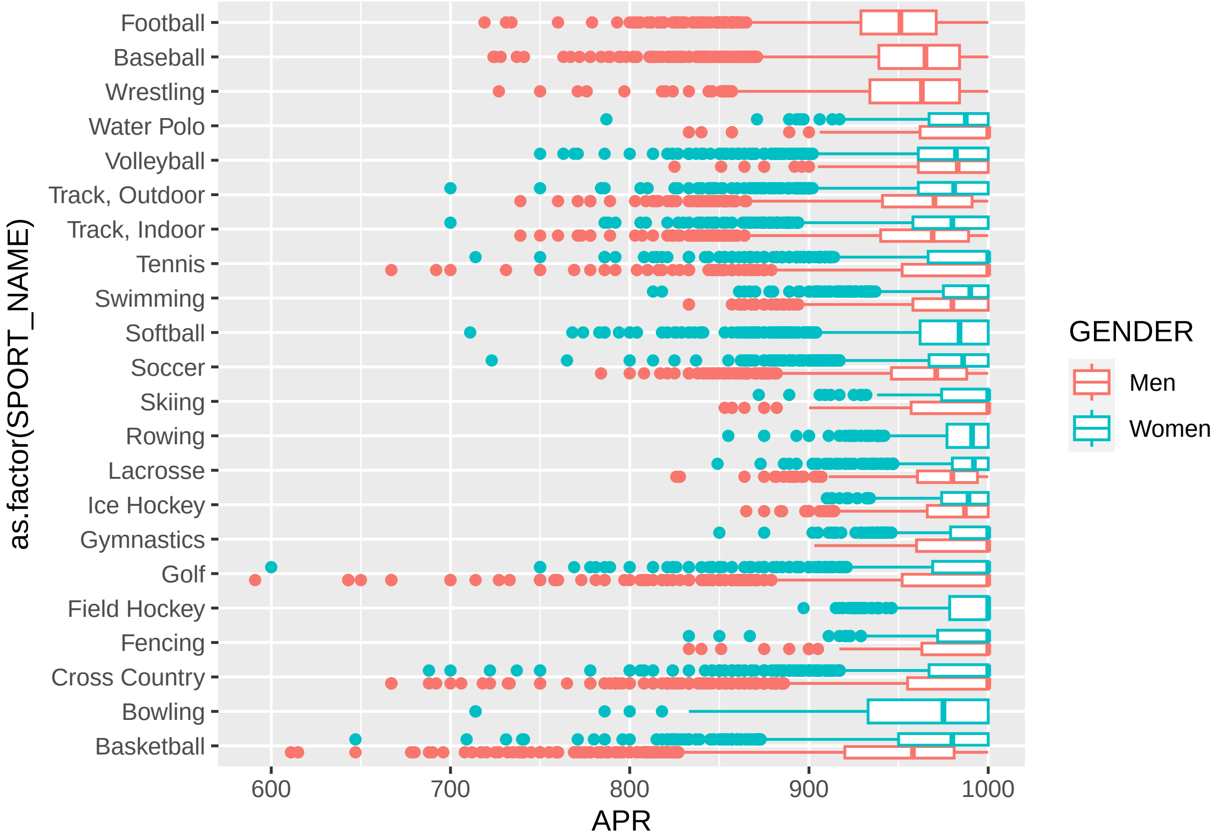


It seems that over the years 2004 to 2014, women's teams have a more APR than men's teams on an average.

3. Visualizing the distribution of APR for both men's and women's teams for each sport:

```
df <- gender_df %>% mutate(SPORT_NAME = stringr::str_remove(SPORT_NAME, "Men's")) %>% mutate(SPORT_NAME
ggplot(df) + geom_boxplot(aes(x=as.factor(SPORT_NAME), y=APR, color=GENDER)) + coord_flip()
```

```
## Warning: Removed 4696 rows containing non-finite values (`stat_boxplot()`).
```



The sports - Voleyball and Fencing have similar APR for Men and Women.