hw2-Spandan-Maaheshwari

Spandan Maaheshwari

2022-10-05

Problem 1 ->

1.Resource

Dataset which I am using (https://www.kaggle.com/datasets/heeraldedhia/bike-buyers)

I found this dataset in Kaggle Datasets (https://www.kaggle.com/datasets) which is about has details of 1000 users from different backgrounds i.e. 13 observations and whether or not they buy a bike.

2.Introduction

I am particularly interested in few observations to predict the buying of a bike and for that doing some exploratory data analysis as well as tidying data.

```
library('tidyverse')
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# Reading the dataset to  high level understanding
bike_buyers = read.csv("/Users/SPANDAN/DS 5110/bike_buyers.csv", header=T, na.strings='')
head(bike_buyers)
```

```
##      ID Marital.Status Gender Income Children       Education     Occupation
## 1 12496        Married Female  40000        1       Bachelors Skilled Manual
## 2 24107        Married   Male  30000        3 Partial College       Clerical
## 3 14177        Married   Male  80000        5 Partial College   Professional
## 4 24381         Single   <NA>  70000        0       Bachelors   Professional
## 5 25597         Single   Male  30000        0       Bachelors       Clerical
## 6 13507        Married Female  10000        2 Partial College         Manual
##   Home.Owner Cars Commute.Distance  Region Age Purchased.Bike
## 1        Yes    0        0-1 Miles  Europe  42             No
## 2        Yes    1        0-1 Miles  Europe  43             No
## 3         No    2        2-5 Miles  Europe  60             No
## 4        Yes    1       5-10 Miles Pacific  41            Yes
## 5         No    0        0-1 Miles  Europe  36            Yes
## 6        Yes    0        1-2 Miles  Europe  50             No
```

```
str(bike_buyers)
```

```
## 'data.frame':    1000 obs. of  13 variables:
##  $ ID              : int  12496 24107 14177 24381 25597 13507 27974 19364 22155 19280 ...
##  $ Marital.Status  : chr  "Married" "Married" "Married" "Single" ...
##  $ Gender          : chr  "Female" "Male" "Male" NA ...
##  $ Income          : int  40000 30000 80000 70000 30000 10000 160000 40000 20000 NA ...
##  $ Children        : int  1 3 5 0 0 2 2 1 2 2 ...
##  $ Education       : chr  "Bachelors" "Partial College" "Partial College" "Bachelors" ...
##  $ Occupation      : chr  "Skilled Manual" "Clerical" "Professional" "Professional" ...
##  $ Home.Owner      : chr  "Yes" "Yes" "No" "Yes" ...
##  $ Cars            : int  0 1 2 1 0 0 4 0 2 1 ...
##  $ Commute.Distance: chr  "0-1 Miles" "0-1 Miles" "2-5 Miles" "5-10 Miles" ...
##  $ Region          : chr  "Europe" "Europe" "Europe" "Pacific" ...
##  $ Age             : int  42 43 60 41 36 50 33 43 58 NA ...
##  $ Purchased.Bike  : chr  "No" "No" "No" "Yes" ...
```

```
summary(bike_buyers)
```

```
##        ID        Marital.Status        Gender              Income
##  Min.   :11000   Length:1000        Length:1000        Min.   : 10000
##  1st Qu.:15291   Class :character   Class :character   1st Qu.: 30000
##  Median :19744   Mode  :character   Mode  :character   Median : 60000
##  Mean   :19966                                         Mean   : 56268
##  3rd Qu.:24471                                         3rd Qu.: 70000
##  Max.   :29447                                         Max.   :170000
##                                                        NA's   :6
##     Children     Education          Occupation         Home.Owner
##  Min.   :0.00   Length:1000        Length:1000        Length:1000
##  1st Qu.:0.00   Class :character   Class :character   Class :character
##  Median :2.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :1.91
##  3rd Qu.:3.00
##  Max.   :5.00
##  NA's   :8
##       Cars       Commute.Distance      Region               Age
##  Min.   :0.000   Length:1000        Length:1000        Min.   :25.00
##  1st Qu.:1.000   Class :character   Class :character   1st Qu.:35.00
##  Median :1.000   Mode  :character   Mode  :character   Median :43.00
##  Mean   :1.455                                         Mean   :44.18
##  3rd Qu.:2.000                                         3rd Qu.:52.00
##  Max.   :4.000                                         Max.   :89.00
##  NA's   :9                                             NA's   :8
##  Purchased.Bike
##  Length:1000
##  Class :character
##  Mode  :character
##
##
##
##
```

```r
# Assigning factors to string values

bike_buyers$Marital.Status <- as.factor(bike_buyers$Marital.Status)
bike_buyers$Gender <- as.factor(bike_buyers$Gender)
bike_buyers$Education <- as.factor(bike_buyers$Education)
bike_buyers$Occupation <- as.factor(bike_buyers$Occupation)

bike_buyers$Commute.Distance <- as.factor(bike_buyers$Commute.Distance)
bike_buyers$Region <- as.factor(bike_buyers$Region)
bike_buyers$Home.Owner <- as.factor(bike_buyers$Home.Owner)
bike_buyers$Purchased.Bike <- as.factor(bike_buyers$Purchased.Bike)

# Checking for NA values
colSums(is.na(bike_buyers))
```

```
##               ID   Marital.Status           Gender           Income
##                0                7               11                6
##         Children        Education       Occupation       Home.Owner
##                8                0                0                4
##             Cars Commute.Distance           Region              Age
##                9                0                0                8
##   Purchased.Bike
##                0
```

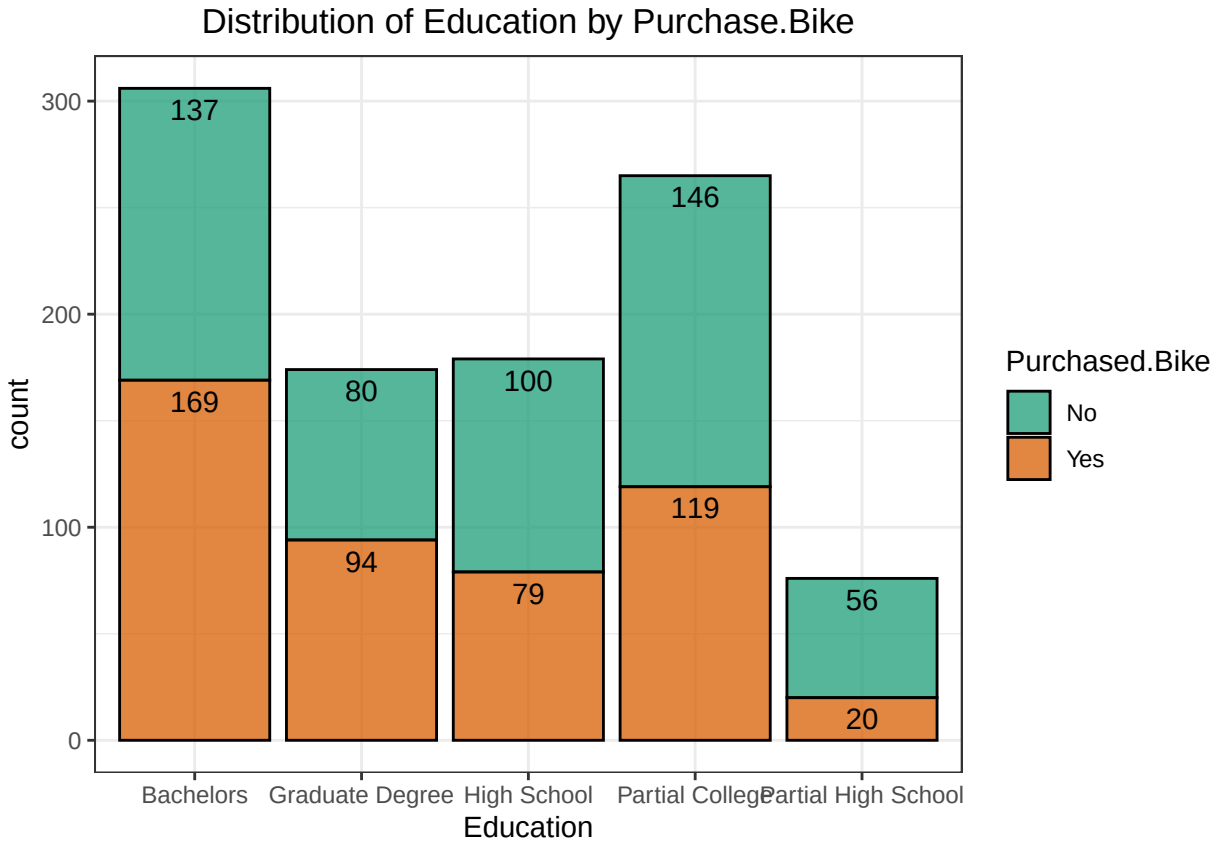
```r
# Dealing with NA values

# Income replaced with Median
bike_buyers$Income[is.na(bike_buyers$Income)] <-
  median(na.omit((bike_buyers$Income)))

# Age replaced with Median
bike_buyers$Age[is.na(bike_buyers$Age)] <-
  median(na.omit((bike_buyers$Age)))

# Creating mode function to calculate the frequency
get_mode <- function(x) {
  unique_x <- unique(x)
  tabulate_x <- tabulate(match(x, unique_x))
  unique_x[tabulate_x == max(tabulate_x)]
}

# Marital Status replaced with Mode
bike_buyers$Marital.Status[is.na(bike_buyers$Marital.Status)] <-
  get_mode(bike_buyers$Marital.Status)

# Gender replaced with Mode
bike_buyers$Gender[is.na(bike_buyers$Gender)] <-
  get_mode(bike_buyers$Gender)

# Children replaced with Mode
bike_buyers$Children[is.na(bike_buyers$Children)] <-
  get_mode(bike_buyers$Children)
```

```r
# Home Owner replaced with Mode
bike_buyers$Home.Owner[is.na(bike_buyers$Home.Owner)] <-
  get_mode(bike_buyers$Home.Owner)

# Cars replaced with Mean
bike_buyers$Cars[is.na(bike_buyers$Cars)] <-
  mean(bike_buyers$Cars, na.rm = TRUE)

colSums(is.na(bike_buyers))
```

```
##                ID   Marital.Status           Gender           Income
##                 0                0                0                0
##          Children        Education       Occupation       Home.Owner
##                 0                0                0                0
##             Cars Commute.Distance           Region              Age
##                 0                0                0                0
##   Purchased.Bike
##                 0
```

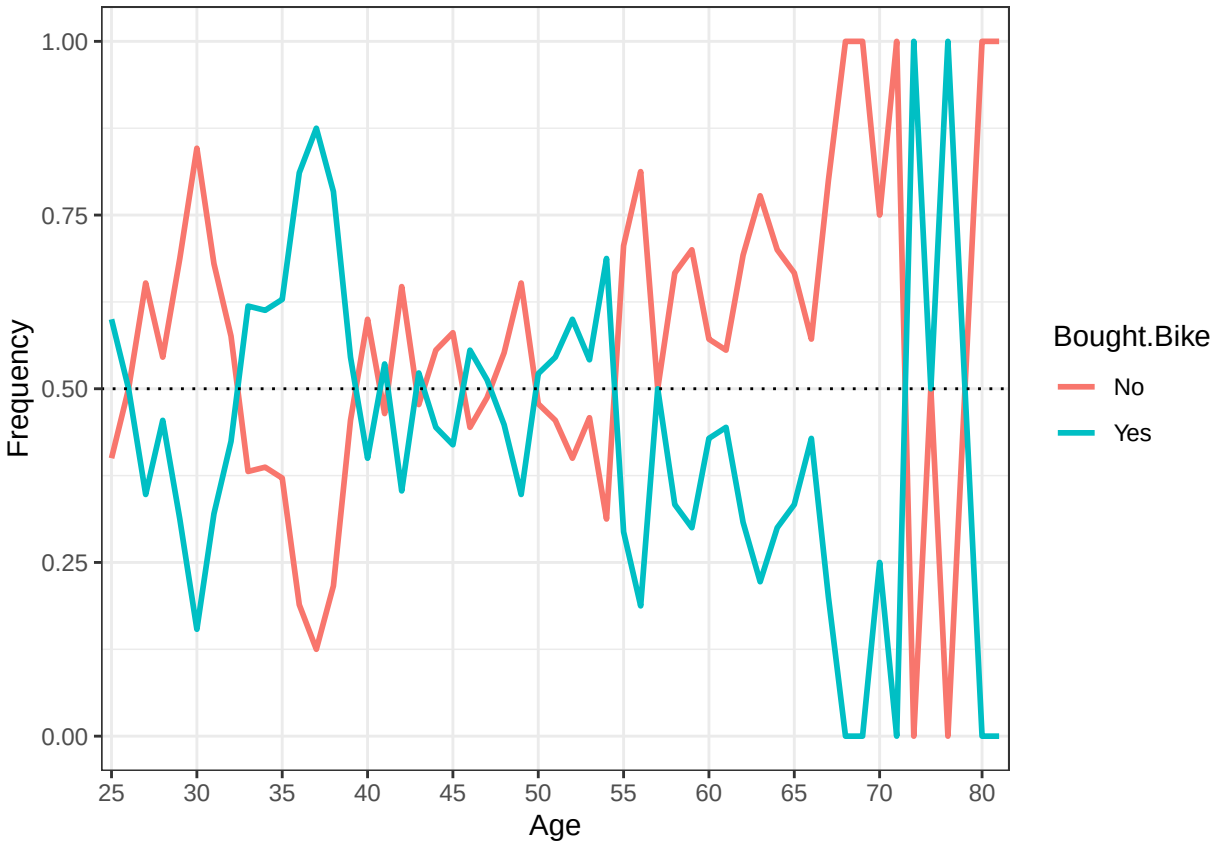
Problem 2 ->

```r
library('ggplot2')
ggplot(data = bike_buyers, aes(x = Education, fill = Purchased.Bike,group = Purchased.Bike)) +
  geom_bar(color = "black",alpha = 0.75) +
  geom_text(stat = "count", aes(label = ..count..),
            inherit.aes = TRUE,position = position_stack(), vjust = 1.5) +
  theme_bw() + scale_fill_brewer(palette="Dark2") +
  labs(title = "Distribution of Education by Purchase.Bike") + theme(plot.title = element_text(hjust = 
```



It is evident from the above bar plot that there is a higher chance of a customer buying a bike when he is more educated. Here clients having Bachelors and Graduate degree as their level of education has the highest purchase of bikes than one's with High School and Partial College completion.

```
df <- as.data.frame(prop.table(table(bike_buyers$Age,bike_buyers$Purchased.Bike),margin = 1))
names(df)<- c("Age","Bought.Bike","Frequency")
ggplot(data = df, aes(x = Age, y = Frequency, group = Bought.Bike)) +
  geom_line(aes(color = Bought.Bike),size = 1) +
    scale_x_discrete(breaks = seq(25 , 89 , by = 5)) + theme_bw() +
      geom_hline(yintercept = 0.50, color = "black", linetype = "dotted")
```



Analyzing the above line plot it seems that most bike buyers are around 30 to 55 and the period where a client is most likely to make a purchase is between 32 and 39.

Problem 3 ->

```
library(readr)
library(tidyr)
library(tidyverse)
ncaa = read_tsv("C:/Users/SPANDAN/DS 5110/NCAA-D1-APR-2003-14/26801-0001-Data.tsv")
```

```
## Rows: 6511 Columns: 76
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr  (4): SCL_NAME, SPORT_NAME, CONFNAME_14, D1_FB_CONF_14
## dbl (69): SCL_UNITID, SPORT_CODE, ACADEMIC_YEAR, SCL_DIV_14, SCL_SUB_14, SCL...
## lgl  (3): DATA_TAB_GENERALINFO, DATA_TAB_MULTIYRRATE, DATA_TAB_ANNUALRATE
##
## 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.
```

```
# change missing data codes from -99 as NAs.
ncaa[ncaa == -99] <- NA
head(ncaa, 10)
```

```
## # A tibble: 10 x 76
##    DATA_TAB_GE~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~
##  7 NA             100654 Alabam~       2 Men's ~    2014       1       2 Southw~
##  8 NA             100654 Alabam~      34 Women'~    2014       1       2 Southw~
##  9 NA             100654 Alabam~      35 Women'~    2014       1       2 Southw~
## 10 NA             100654 Alabam~      31 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>, ...
```

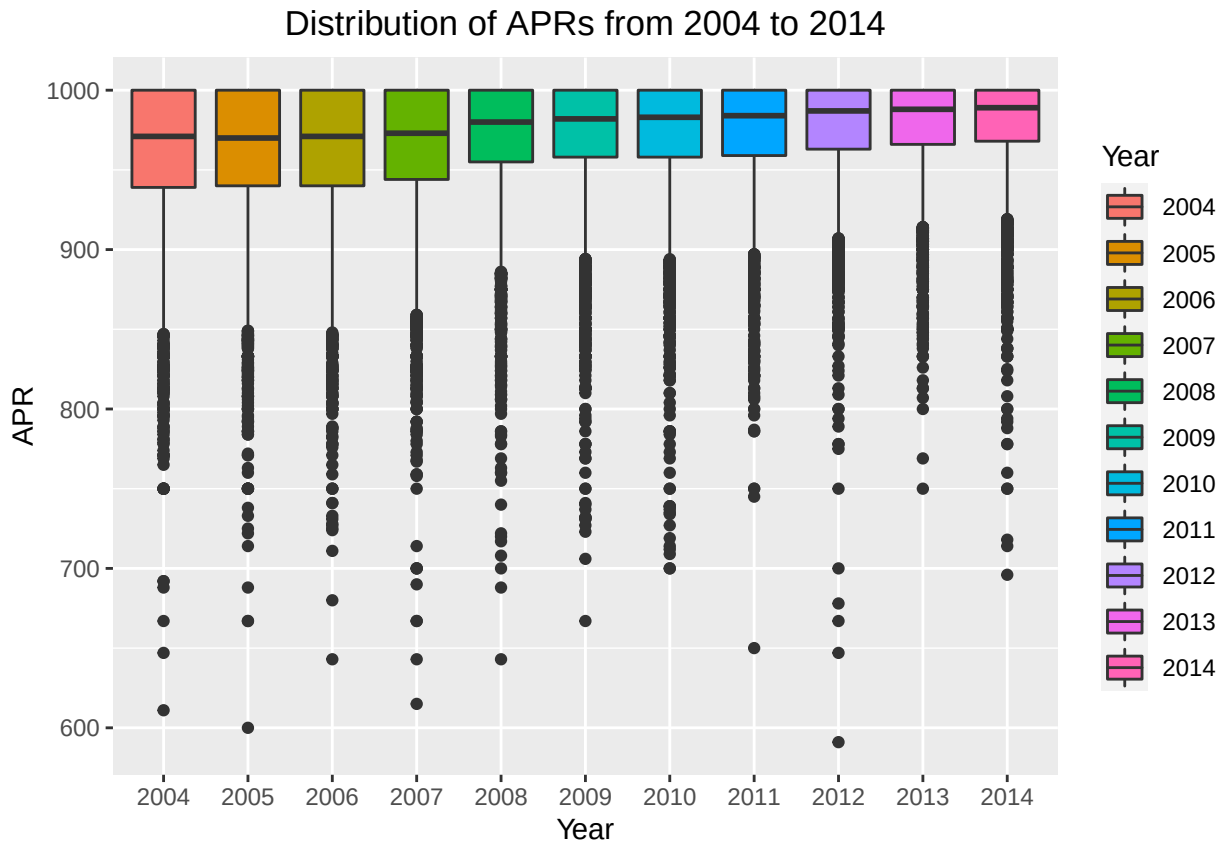
```r
#selecting needed columns
ncaa <- ncaa %>% select(starts_with(c("SCL_UNITID", "SCL_NAME", "SPORT_CODE","SPORT_NAME","APR_RATE")))

#changing column names
ncaa<- ncaa %>%
rename(
School_ID = SCL_UNITID,
School_name = SCL_NAME,
Sport_code = SPORT_CODE,
Sport_name = SPORT_NAME,
"2004" = APR_RATE_2004_1000,
"2005" = APR_RATE_2005_1000,
"2006" = APR_RATE_2006_1000,
"2007" = APR_RATE_2007_1000,
"2008" = APR_RATE_2008_1000,
"2009" = APR_RATE_2009_1000,
"2010" = APR_RATE_2010_1000,
"2011" = APR_RATE_2011_1000,
"2012" = APR_RATE_2012_1000,
"2013" = APR_RATE_2013_1000,
"2014" = APR_RATE_2014_1000
)

#pivot_longer to make final dataset.
df= pivot_longer(ncaa, cols=`2014`: `2004`, names_to = "Year", values_to = "APR")
head(df, 10)
```

```
## # A tibble: 10 x 6
##    School_ID School_name            Sport_code Sport_name      Year    APR
##        <dbl> <chr>                       <dbl> <chr>           <chr> <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
##  7    100654 Alabama A&M University          20 Women's Bowling 2008   1000
##  8    100654 Alabama A&M University          20 Women's Bowling 2007    958
##  9    100654 Alabama A&M University          20 Women's Bowling 2006    875
## 10    100654 Alabama A&M University          20 Women's Bowling 2005   1000
```

```
#Visualing the distributions of APR's over time
library('ggplot2')
ggplot(data = df, mapping = aes(x=Year, y=APR, fill=Year)) +
   geom_boxplot()+labs(x="Year", y = "APR", title = "Distribution of APRs from 2004 to 2014") +
      theme(plot.title = element_text(hjust = 0.5) )
```

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



Explanation for Problem 3:

For the boxplot shown above for distribution of APRs over time, median of APR over time is an upward trend so APRs over year from 2004 to 2014 is growing up.

Problem 4 ->

```
# Filtering to remove Mixed sports
df1 <- filter(df, Sport_code <= 37)
head(df1,10)
```

```
## # A tibble: 10 x 6
```

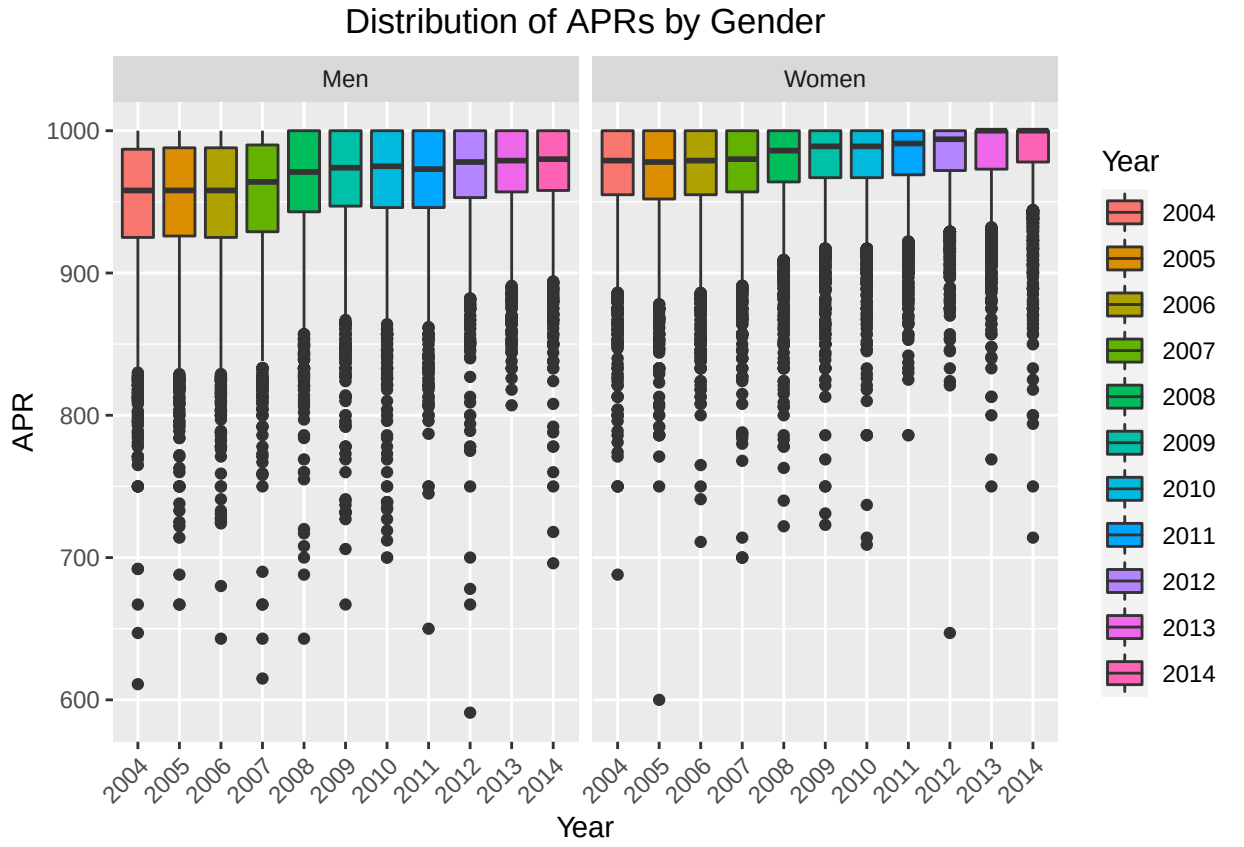
```
##    School_ID School_name            Sport_code Sport_name      Year    APR
##        <dbl> <chr>                       <dbl> <chr>           <chr> <dbl>
##  1    100654 Alabama A&M University         20 Women's Bowling 2014   1000
##  2    100654 Alabama A&M University         20 Women's Bowling 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
##  7    100654 Alabama A&M University         20 Women's Bowling 2008   1000
##  8    100654 Alabama A&M University         20 Women's Bowling 2007    958
##  9    100654 Alabama A&M University         20 Women's Bowling 2006    875
## 10    100654 Alabama A&M University         20 Women's Bowling 2005   1000
```

```
#to make sure it is Men or Women.
x = ifelse(df1[,"Sport_code"] <= 18,"Men","Women")

# adding a new column indicating gender division
df1$Gender <- c(x)
head(df1,10)
```

```
## # A tibble: 10 x 7
##    School_ID School_name            Sport_code Sport_name     Year    APR Gender
##        <dbl> <chr>                       <dbl> <chr>          <chr> <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
##  7    100654 Alabama A&M University         20 Women's Bowli~ 2008   1000 Women
##  8    100654 Alabama A&M University         20 Women's Bowli~ 2007    958 Women
##  9    100654 Alabama A&M University         20 Women's Bowli~ 2006    875 Women
## 10    100654 Alabama A&M University         20 Women's Bowli~ 2005   1000 Women
```

```
#Visualizing the distributions of APR by Gender over time
library('ggplot2')
ggplot(data = df1, mapping = aes(x=Year, y=APR, fill=Year)) +
   geom_boxplot() + facet_grid(~Gender)+ labs(x="Year", y = "APR", title = "Distribution of APRs by Gend
      theme(plot.title = element_text(hjust = 0.5) ) + theme(axis.text.x = element_text(angle = 45, hjus
```

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



Explanation for Problem 4:

Comparing Men's and Women's sport using box plot to find the relationship between APR over time. According to the box plot, the median of Women's APR over time is always higher than the Men's APR. Another observation is that both their APR's are showing an upward trend over year from 2004 to 2014 with median of Women being maximum in year 2014.

Problem 5 ->

```
df2 <- filter(df1, Sport_code <= 18)
df2 <- df2 %>% select(starts_with(c("Sport_name", "APR")))
head(df2,10)
```

```
## # A tibble: 10 x 2
##    Sport_name            APR
##    <chr>               <dbl>
##  1 Men's Track, Indoor   910
##  2 Men's Track, Indoor   932
##  3 Men's Track, Indoor   945
##  4 Men's Track, Indoor   946
##  5 Men's Track, Indoor   922
##  6 Men's Track, Indoor  1000
##  7 Men's Track, Indoor    NA
##  8 Men's Track, Indoor    NA
##  9 Men's Track, Indoor   903
## 10 Men's Track, Indoor   926
```

```
#install package data.table
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

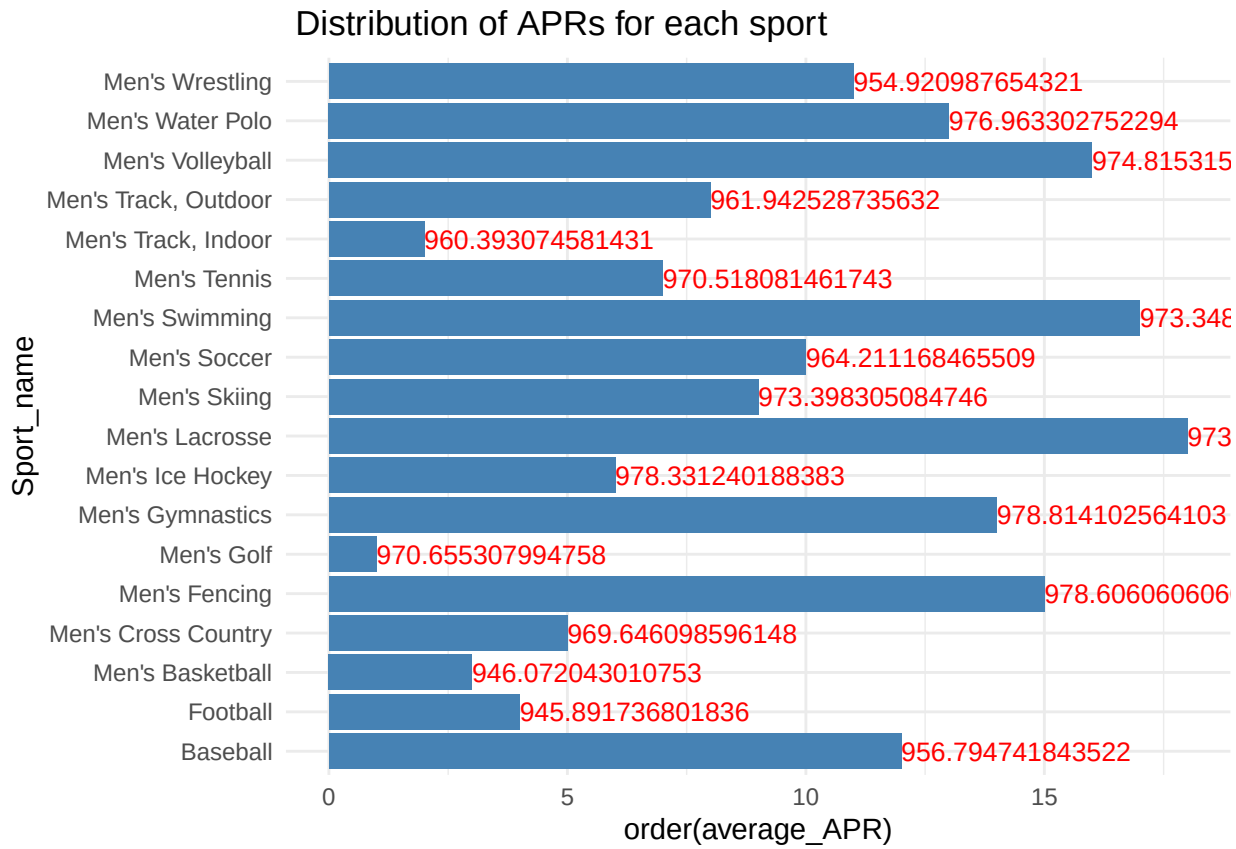
```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
keys <- colnames(df2)[!grepl('APR',colnames(df2))]
X <- as.data.table(df2)
df2 <- X[,list(average_APR = mean(APR, na.rm = TRUE)),keys]
df2
```

```
##               Sport_name average_APR
##  1:  Men's Track, Indoor    960.3931
##  2:             Football    945.8917
##  3:             Baseball    956.7947
##  4:     Men's Basketball    946.0720
##  5:           Men's Golf    970.6553
##  6:         Men's Tennis    970.5181
##  7: Men's Track, Outdoor    961.9425
##  8:         Men's Soccer    964.2112
##  9:     Men's Ice Hockey    978.3312
## 10:  Men's Cross Country    969.6461
## 11:       Men's Swimming    973.3488
## 12:      Men's Wrestling    954.9210
## 13:     Men's Volleyball    974.8153
## 14:     Men's Water Polo    976.9633
## 15:     Men's Gymnastics    978.8141
## 16:         Men's Skiing    973.3983
## 17:       Men's Lacrosse    973.3455
## 18:        Men's Fencing    978.6061
```

```
# Using y = order(average_APR) for making difference in bar more obvious.
ggplot(df2, aes(x=Sport_name, y= order(average_APR))) +
  geom_bar(stat="identity", fill="steelblue") +
     geom_text(aes(label=average_APR), vjust=0.5,hjust=0, color="red", size=3.5) +
       theme_minimal()+coord_flip()+labs(title=" Distribution of APRs for each sport")
```



Explanation for Problem 5:

In order to further investigate APR for different Men's team with distribution for each plot bar plot is used for visualization. The ARP for Men's sport don't have much difference, so order is used to easily identify the difference in the plot. Also we can observe that, the Men's Gymnastics and Men's Fencing has the highest APR while Men's Basketball and Football with the lowest.