Isaac's stuff

Scraping

```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.6.2

##
## 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(rvest)
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.2

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.0     v stringr 1.4.0
## v tidyr   1.1.3     v forcats 0.5.1
## v readr   1.4.0

## Warning: package 'ggplot2' was built under R version 3.6.2

## Warning: package 'tibble' was built under R version 3.6.2

## Warning: package 'tidyr' was built under R version 3.6.2

## Warning: package 'readr' was built under R version 3.6.2

## Warning: package 'purrr' was built under R version 3.6.2

## Warning: package 'forcats' was built under R version 3.6.2

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

```r
library(kableExtra)
```

```
## Warning: package 'kableExtra' was built under R version 3.6.2

##
## Attaching package: 'kableExtra'

## The following object is masked from 'package:dplyr':
##
##     group_rows
```

wnba scraping

```r
wilson <- 'https://www.basketball-reference.com/wnba/players/w/wilsoa01w/gamelog/2022/'
wil_doc <- rvest::read_html(wilson)

wil_doc %>%
  rvest::html_elements(., xpath = "//*[(@id = 'div_wnba_pgl_basic')]") %>%
  rvest::html_table() -> wil
wil <- wil[[1]]
head(wil)
```

```
## # A tibble: 6 x 28
##   Rk    Date    Age   Tm    ``    Opp   ``    GS    MP    FG    FGA   `FG%` `3P`
##   <chr> <chr>   <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>
## 1 1     2022-~ 25-2~ LVA   "@"   PHO   W (+~ 1     28:35 5     8     .625  0
## 2 2     2022-~ 25-2~ LVA   ""    SEA   W (+~ 1     35:06 8     14    .571  1
## 3 3     2022-~ 25-2~ LVA   "@"   WAS   L (-~ 1     29:56 4     11    .364  0
## 4 4     2022-~ 25-2~ LVA   "@"   ATL   W (+~ 1     29:08 6     11    .545  0
## 5 5     2022-~ 25-2~ LVA   ""    PHO   W (+~ 1     33:45 4     8     .500  0
## 6 6     2022-~ 25-2~ LVA   ""    MIN   W (+~ 1     31:16 5     9     .556  1
## # ... with 15 more variables: 3PA <chr>, 3P% <chr>, FT <chr>, FTA <chr>,
## #   FT% <chr>, ORB <chr>, DRB <chr>, TRB <chr>, AST <chr>, STL <chr>,
## #   BLK <chr>, TOV <chr>, PF <chr>, PTS <chr>, GmSc <chr>
```

```r
#wil2 <- mutate_all(wil, function(x) as.numeric(as.character(x)))
#mean(wil2['PTS'])

#wil$eFG<- (wil['FG'] + (0.5*wil['3P']))/wil['FGA']
#wil$eFG ![Screenshot]('~/Google Drive/My Drive/Sports Analytics/SportsAnalyticsBook/images/scraping1')
```

EDA/Probability

Baseball

WAR comparison (Prob)

Link to WAR explaination: https://www.mlb.com/glossary/advanced-stats/wins-above-replacement

Player X has a projected mean WAR of 3 with standard deviation of 2 and player Y has a projected mean WAR of 1.5 with a standard deviation of 3. Assume projected WAR is normally distributed. Q: What is the probability that Player X outperforms Player Y? A: We want Pr(X>Y) or Pr(X-Y>0).
Let Z = X-Y.
E[Z]=1.5 Var(Z)=5 Pr(Z>0)=1-Pr(Z $\leq$ 0)

```r
#Calculate probability Z<=0
pr <- pnorm(0,1.5,sqrt(5))
print(1-pr)
```

```
## [1] 0.7488325
```

The Probability that Player X outperforms Player Y is 0.7488.

Injured Baserunner (Prob)

A runner on first base with 2 out and nobody else on base will attempt to steal second base on the first pitch 70% of the time if he is fully healthy but only 10% of the time if he is playing through an injury. Assume that 80% of the player population is healthy. You see a randomly selected runner not attempt a steal in this situation. Q: What is the probability that the runner is playing through an injury? A: From Bayes Theorem:

Pr(Injury given No Steal) = Pr(No Steal given Injury)*Pr(Injury)/P(No Steal).

Pr(No Steal given Injury) = 1 - Pr(Steal given Injury) = 0.9.

Pr(Injury) = 1- Pr(Healthy) = 0.2.

Pr(No Steal) = Pr(No Steal given Injury)*Pr(Injury)+Pr(No Steal given Healthy)*Pr(Healthy).

Pr(No Steal) = 0.9*0.2+0.7*0.8 = 0.74.

Therefore Pr(Injury given No Steal) = 0.9*0.2/0.74 = 0.243.

OPS (EDA)

Q: Using the dataset, plot the leagues average OPS from every year in the data to see the progression. A:

```
mlb = read.csv('~/Google Drive/My Drive/Sports Analytics/SportsAnalyticsBook/data/mlb_team_stats_histor
head(mlb)
```

```
##   yearID lgID teamID franchID divID Rank   G Ghome  W  L DivWin WCWin LgWin
## 1   1976   NL    ATL      ATL     W    6 162    81 70 92      N           N
## 2   1976   AL    BAL      BAL     E    2 162    81 88 74      N           N
## 3   1976   AL    BOS      BOS     E    3 162    81 83 79      N           N
## 4   1976   AL    CAL      ANA     W    4 162    81 76 86      N           N
## 5   1976   AL    CHA      CHW     W    6 161    80 64 97      N           N
## 6   1976   NL    CHN      CHC     E    4 162    81 75 87      N           N
##   WSWin   R   AB    H  X1B X2B X3B  HR  BB  SO  SB CS HBP SF  RA    BA  ER  ERA
## 1     N 620 5345 1309 1027 170  30  82 589 811  74 61  19 47 700 0.245 617 3.86
## 2     N 619 5457 1326  966 213  28 119 519 883 150 61  23 35 598 0.243 541 3.32
## 3     N 716 5511 1448 1004 257  53 134 500 832  95 70  29 59 660 0.263 571 3.52
## 4     N 550 5385 1265  969 210  23  63 534 812 126 80  42 48 631 0.235 551 3.36
## 5     N 586 5532 1410 1082 209  46  73 471 739 120 53  34 55 745 0.255 684 4.25
## 6     N 611 5519 1386 1041 216  24 105 490 834  74 74  30 41 728 0.251 643 3.93
##   CG SHO SV IPouts   HA HRA BBA SOA   E  DP    FP              name
## 1 33  13 27   4314 1435  86 564 818 167 151 0.973    Atlanta Braves
## 2 59  16 23   4406 1396  80 489 678 118 157 0.982 Baltimore Orioles
## 3 49  13 27   4374 1495 109 409 673 141 148 0.978    Boston Red Sox
## 4 64  15 17   4432 1323  95 553 992 150 139 0.977 California Angels
## 5 54  10 22   4344 1460  87 600 802 130 155 0.979 Chicago White Sox
## 6 27  12 33   4414 1511 123 490 850 140 145 0.978      Chicago Cubs
##                           park attendance BPF PPF teamIDBR teamIDlahman45
## 1 Atlanta-Fulton County Stadium     818179 106 108      ATL            ATL
## 2              Memorial Stadium    1058609  94  93      BAL            BAL
## 3                Fenway Park II    1895846 113 112      BOS            BOS
## 4               Anaheim Stadium    1006774  93  94      CAL            CAL
## 5                 Comiskey Park     914945 101 102      CHW            CHA
## 6                 Wrigley Field    1026217 108 109      CHC            CHN
##   teamIDretro
## 1         ATL
## 2         BAL
## 3         BOS
## 4         CAL
## 5         CHA
```

```
## 6         CHN
```

```r
# make new variables
mlb=mutate(mlb,SLG=(X1B+2*X2B+3*X3B+4*HR)/(AB))
mlb=mutate(mlb,OBP=(H+BB+HBP)/(AB+BB+HBP+SF))
mlb=mutate(mlb,OPS=OBP+SLG)

# get avg ops
summarize(mlb, Average = mean(OPS,na.rm=T))
```

```
##     Average
## 1 0.7330384
```

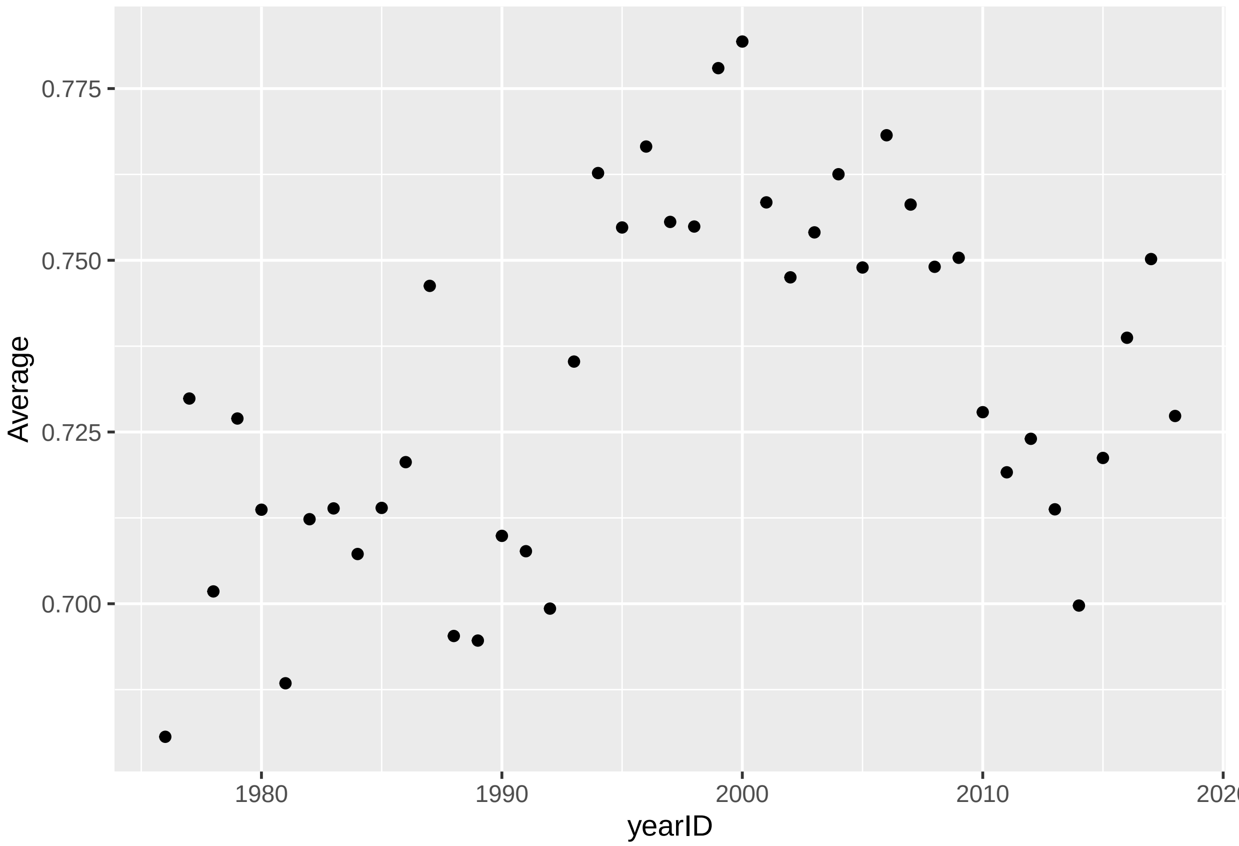
```r
# get avg ops by year
group_by(mlb, yearID)%>%
summarize(Average = mean(OPS, na.rm=T))
```

```
## # A tibble: 43 x 2
##    yearID Average
##     <int>   <dbl>
##  1   1976   0.681
##  2   1977   0.730
##  3   1978   0.702
##  4   1979   0.727
##  5   1980   0.714
##  6   1981   0.688
##  7   1982   0.712
##  8   1983   0.714
##  9   1984   0.707
## 10   1985   0.714
## # ... with 33 more rows
```

```r
group_by(mlb, yearID)%>%
summarize(Average = mean(OPS, na.rm=T))%>%View

#create new dataset
mlbYr=group_by(mlb, yearID)%>%
summarize(Average = mean(OPS, na.rm=T))

#plot it
ggplot(mlbYr, aes(x=yearID, y= Average))+geom_point()
```



Followup Q: What would cause the data to peak around the year 2000? A: PED's

Run Variance (Probability)

Runs Scored	Probability
0	0.55
1	0.25
2	0.15
3	0.05

```
col1 = c('Runs Scored','Probability')
col2 = c('0',0.55)
col3 = c('1',0.25)
col4=c('2',0.15)
col5=c('3',0.05)
runs <- data.frame(col1,col2,col3,col4,col5)
colnames(runs) <- NULL
runs
```

```
##
## 1 Runs Scored    0    1    2    3
## 2 Probability 0.55 0.25 0.15 0.05
```

```
kbl(runs)
```

Runs Scored	0	1	2	3
Probability	0.55	0.25	0.15	0.05

Tennis

Link for brief explanation of tennis scoring: https://www.sportingnews.com/us/tennis/news/tennis-scoring-explained-rules-syste
7uzp2evdhbd11obdd59p3p1cx

Probability of Winning a Game (Prob)

The formula for the probability of a tennis player winning a game (from Analyzing Wimbledon) is given by $\frac{p^4*(-8*p^3+28*p^2-34*p+15)}{p^2+(1-p)^2}$ where p is the probability of a player winning their service point. Q: If a player wins their service points 62% of the time, what is the probability they win the game? A:

```
p <- 0.62
pr_game <- (p^4*(-8*p^3+28*p^2-34*p+15))/(p^2+(1-p)^2)
pr_game
```

```
## [1] 0.7758627
```

Graph Example of Probability of Winning Point vs Probability of Winning Game (Prob)

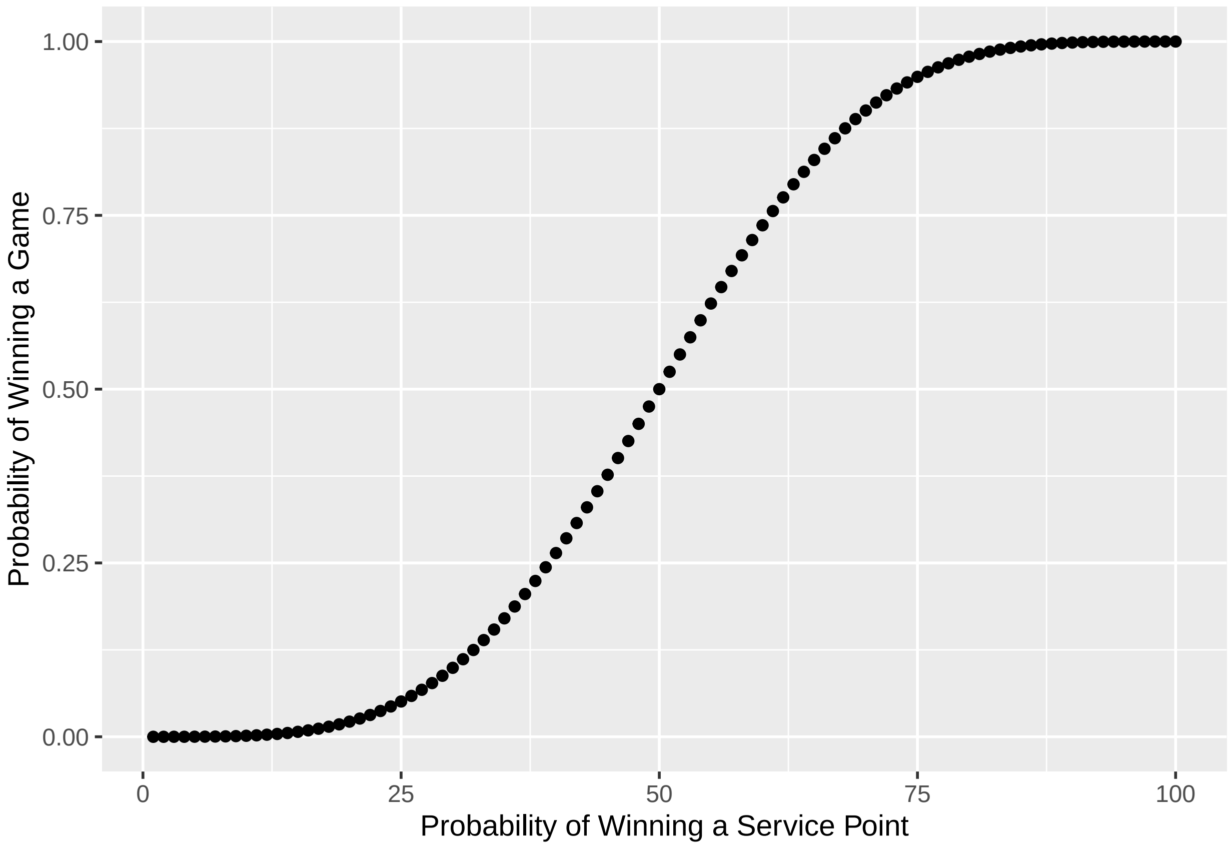
```
game <- c(0)
pr <- 1:100
for(x in pr) {
  p <- pr/100
  pr_game <- (p^4*(-8*p^3+28*p^2-34*p+15))/(p^2+(1-p)^2)
  game <- c(game,pr_game)
}
game[1]
```

```
## [1] 0
```

```
game <- game[2:101]
game[1]
```

```
## [1] 1.495898e-07
```

```
df <- do.call(rbind, Map(data.frame, point_pr=pr, game_pr=game))
ggplot(df, aes(x=point_pr, y=game_pr)) +
  geom_point()+xlab('Probability of Winning a Service Point')+ylab('Probability of Winning a Game')
```



WNBA Scores (EDA)

Q: What is the difference in PPG for a winning team at home vs a winning team away? A:

```
wnba=read.csv('~/Google Drive/My Drive/Sports Analytics/SportsAnalyticsBook/data/WNBA_Games2019_Scores.c
head(wnba)
```

```
##   Game          HomeTeam           AwayTeam Winner PTSwin PTSlose
## 1    1     Atlanta Dream       Dallas Wings   Home     76      72
## 2    2 New York Liberty      Indiana Fever   Away     81      80
## 3    3  Connecticut Sun Washington Mystics   Home     84      69
## 4    4   Minnesota Lynx         Chicago Sky   Home     89      71
## 5    5     Seattle Storm     Phoenix Mercury   Home     77      68
## 6    6   Las Vegas Aces Los Angeles Sparks   Home     83      70
##        WinningTeam
## 1    Atlanta Dream
## 2    Indiana Fever
## 3 Connecticut Sun
## 4  Minnesota Lynx
## 5    Seattle Storm
## 6   Las Vegas Aces
```

```
group_by(wnba, Winner)%>%
  summarize(Count=n())%>%
  mutate(Percent=Count/sum(Count))
```

```
## # A tibble: 2 x 3
##   Winner Count Percent
```

```
##   <fct>  <int>   <dbl>
## 1 Away      80   0.392
## 2 Home     124   0.608
```

```
group_by(wnba, Winner)%>%
  summarize(Average=mean(PTSwin,na.rm=T),sd=sd(PTSwin,na.rm=T))
```

```
## # A tibble: 2 x 3
##   Winner Average    sd
##   <fct>    <dbl> <dbl>
## 1 Away      83.8  9.20
## 2 Home      84.8 10.8
```

```
84.822-83.787
```

```
## [1] 1.035
```

A home team winner scores on average 1.035 PPG more than an away team winner.

NFL

```
nfl=read.csv('~/Google Drive/My Drive/Sports Analytics/SportsAnalyticsBook/data/nfl_pbp.csv')
nfl2 <- select(nfl, c('Date','GameID','qtr','down','time','yrdline100','ydstogo','Yards.Gained','Touchd
head(nfl2)
```

```
##         Date     GameID qtr down  time yrdline100 ydstogo Yards.Gained
## 1 2009-09-10 2009091000   1   NA 15:00         30       0           39
## 2 2009-09-10 2009091000   1    1 14:53         58      10            5
## 3 2009-09-10 2009091000   1    2 14:16         53       5           -3
## 4 2009-09-10 2009091000   1    3 13:35         56       8            0
## 5 2009-09-10 2009091000   1    4 13:27         56       8            0
## 6 2009-09-10 2009091000   1    1 13:16         98      10            0
##   Touchdown PlayType FieldGoalResult FieldGoalDistance ScoreDiff Season
## 1         0  Kickoff            <NA>                NA         0   2009
## 2         0     Pass            <NA>                NA         0   2009
## 3         0      Run            <NA>                NA         0   2009
## 4         0     Pass            <NA>                NA         0   2009
## 5         0     Punt            <NA>                NA         0   2009
## 6         0      Run            <NA>                NA         0   2009
```

4th Down Analysis (EDA)

Q: Using NFL Play by Play data, what percentage of the time do coaches choose to go for it on 4th down? And what percentage of 4th down attempts are successful? A:

```
# add indicator column for successful first down attempt
nfl2 <- nfl2 %>%
  mutate(FirstDown = case_when(
    ydstogo < Yards.Gained ~ 1,
    ydstogo > Yards.Gained ~ 0
    ))
# filter by only plays on 4th down
down4 = filter(nfl2, nfl2['down']==4)

#see what play types are run on first down and remove the noise
group_by(down4,PlayType) %>%
```

```
  summarize(Count=n())%>%
  mutate(Percentage=Count/sum(Count))
```

```
## # A tibble: 8 x 3
##   PlayType   Count Percentage
##   <fct>      <int>      <dbl>
## 1 Field Goal  7265  0.226
## 2 No Play     1433  0.0446
## 3 Pass        2239  0.0698
## 4 Punt       19551  0.609
## 5 QB Kneel      22  0.000685
## 6 Run         1424  0.0444
## 7 Sack         164  0.00511
## 8 Timeout        1  0.0000312
```

```
down4 = filter(down4, down4['PlayType']!='No Play' || down4['PlayType']!= 'QB Kneel' || down4['PlayType

# add indicator column for going for it on 4th
down4 <- down4 %>%
  mutate(GoForIt = case_when(
    PlayType == 'Pass' ~ 1,
    PlayType == 'Run' ~ 1,
    PlayType == 'Sack' ~ 1,
    PlayType == 'Field Goal' ~ 0,
    PlayType == 'Punt' ~ 0
  ))
# get percentage of 4th downs are gone for
group_by(down4,GoForIt) %>%
  summarize(Count=n())%>%
  mutate(Percentage=Count/sum(Count))
```

```
## # A tibble: 3 x 3
##   GoForIt Count Percentage
##     <dbl> <int>      <dbl>
## 1       0 26816     0.835
## 2       1  3827     0.119
## 3      NA  1456     0.0454
```

```
# get percentage of successful attempted 4th downs
down4 %>%
  filter(down4['GoForIt']==1) %>%
  group_by(FirstDown) %>%
    summarize(Count=n())%>%
    mutate(Percentage=Count/sum(Count))
```

```
## # A tibble: 3 x 3
##   FirstDown Count Percentage
##       <dbl> <int>      <dbl>
## 1         0  1971      0.515
## 2         1  1553      0.406
## 3        NA   303      0.0792
```

11% of 4th downs are gone for and 40% of those are successful, regardless of how many yards to go there are

Football Sample Space (Probability)

A sample space contains all possible outcomes. An american football game can either end with a win (W), loss (L) or a tie (T) which means our sample space is $\Omega = \{W, L, T\}$ and an event, E would be one of the possible outcomes. If a team wins the game, the event for that game would be $E = \{W\}$ or if we want the event of the 2021 CSU football season, it would be $E = \{L, L, W, L, W, W, L, L, L, L, L, L\}$.