Clean and Link Datasets

Basketball Team

3/2/2020

We will preform some basic cleaning and linking of the primary and secondary datasets in order to

Clean Primary dataset

```
library("readxl")
df_primary <- read_excel('data/raw/primary_dataset_raw.xlsx')
```

```
## Warning in read_fun(path = enc2native(normalizePath(path)), sheet_i = sheet, :
## Expecting numeric in D24626 / R24626C4: got 'z'
```

```
df_primary <- df_primary[,!(names(df_primary)%in%c('#','blanl','blank2'))] # drop empty/non-stat column
names(df_primary) <- c('year','name','salary','position_p','age','team','games','games_started','minutes
                       'player_efficiency_ranking','true_shooting_%','3PA_over_FA','free_throw_%','offen
                       'defensive_rebound_%','total_rebound_%','assist_to_turnover_%','steal_%','block_
                       'usage_%','offensive_wins_shares','defensive_wins_shares','win_shares','win_share
                       'offensive_box_plus_minus','defensive_box_plus_minus','box_plus_minus','value_ove
                       'field_goal','field_goal_attempts','field_goal_%','3_pointer','3_pointer_attempt
                       '2_pointer','2_pointer_attempts','2_point_%','effective_field_goal_%','free_thro
                       'free_throw_%','offensive_rebounds','defensive_rebounds','total_rebounds','assis
                       'turnovers','personal_fouls','points')
df_primary <- df_primary[!is.na(df_primary[['salary']]),] # drop rows with no salaryes
df_primary <- df_primary[df_primary$year%in%c(2016:2020),]
df_primary[is.na(df_primary)] <- 0
head(df_primary)
```

```
## # A tibble: 6 x 51
##    year name  salary position_p   age team  games games_started minutes_played
##   <dbl> <chr>  <dbl> <chr>      <dbl> <chr> <dbl>         <dbl>          <dbl>
## 1  2017 A.J.~ 1.31e6 C             24 DAL      22             0            163
## 2  2016 Aaro~ 2.70e6 PG            31 CHI      69             0           1108
## 3  2017 Aaro~ 2.12e6 PG            32 IND      65             0            894
## 4  2016 Aaro~ 4.35e6 PF            20 ORL      78            37           1863
## 5  2017 Aaro~ 5.50e6 SF            21 ORL      80            72           2298
## 6  2016 Aaro~ 3.76e5 SG            21 CHO      21             0             93
## # ... with 42 more variables: player_efficiency_ranking <dbl>,
## #   `true_shooting_%` <dbl>, `3PA_over_FA` <dbl>, `free_throw_%` <dbl>,
## #   `offensive_rebound_%` <dbl>, `defensive_rebound_%` <dbl>,
## #   `total_rebound_%` <dbl>, `assist_to_turnover_%` <dbl>, `steal_%` <dbl>,
## #   `block_%` <dbl>, `turnover_%` <dbl>, `usage_%` <dbl>,
## #   offensive_wins_shares <dbl>, defensive_wins_shares <dbl>, win_shares <dbl>,
## #   win_shares_over_48 <dbl>, offensive_box_plus_minus <dbl>,
## #   defensive_box_plus_minus <dbl>, box_plus_minus <dbl>,
## #   value_over_replacement_player <dbl>, field_goal <dbl>,
## #   field_goal_attempts <dbl>, `field_goal_%` <dbl>, `3_pointer` <dbl>,
## #   `3_pointer_attempts` <dbl>, `3_point_%` <dbl>, `2_pointer` <dbl>,
## #   `2_pointer_attempts` <dbl>, `2_point_%` <dbl>,
## #   `effective_field_goal_%` <dbl>, free_throws <dbl>,
## #   free_throw_attempts <dbl>, `free_throw_%` <dbl>, offensive_rebounds <dbl>,
## #   defensive_rebounds <dbl>, total_rebounds <dbl>, assists <dbl>,
## #   steals <dbl>, blocks <dbl>, turnovers <dbl>, personal_fouls <dbl>,
## #   points <dbl>
```

```
summary(df_primary)
```

```
##       year          name               salary          position_p
##  Min.   :2016   Length:965         Min.   :   11534   Length:965
##  1st Qu.:2016   Class :character   1st Qu.: 1551659   Class :character
##  Median :2017   Mode  :character   Median : 4000000   Mode  :character
##  Mean   :2017                      Mean   : 6789399
##  3rd Qu.:2017                      3rd Qu.:10500000
##  Max.   :2017                      Max.   :34682550
##       age           team              games      games_started
##  Min.   :19.00   Length:965         Min.   : 1.00   Min.   : 0.00
##  1st Qu.:23.00   Class :character   1st Qu.:32.00   1st Qu.: 1.00
##  Median :26.00   Mode  :character   Median :61.00   Median :12.00
##  Mean   :26.48                      Mean   :53.41   Mean   :25.99
##  3rd Qu.:29.00                      3rd Qu.:75.00   3rd Qu.:52.00
##  Max.   :40.00                      Max.   :82.00   Max.   :82.00
##  minutes_played player_efficiency_ranking true_shooting_%   3PA_over_FA
##  Min.   :   1   Min.   :-35.30            Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 496   1st Qu.: 10.50            1st Qu.:0.5040   1st Qu.:0.1360
##  Median :1197   Median : 13.30            Median :0.5380   Median :0.3110
##  Mean   :1247   Mean   : 13.61            Mean   :0.5324   Mean   :0.3045
##  3rd Qu.:1954   3rd Qu.: 16.30            3rd Qu.:0.5710   3rd Qu.:0.4470
##  Max.   :3125   Max.   : 39.30            Max.   :1.0000   Max.   :1.0000
##   free_throw_%    offensive_rebound_% defensive_rebound_% total_rebound_%
##  Min.   :0.0000   Min.   : 0.000      Min.   : 0.00       Min.   : 0.000
##  1st Qu.:0.1670   1st Qu.: 1.900      1st Qu.:10.30       1st Qu.: 6.200
##  Median :0.2400   Median : 3.300      Median :14.00       Median : 8.800
##  Mean   :0.2682   Mean   : 4.868      Mean   :15.13       Mean   : 9.992
##  3rd Qu.:0.3380   3rd Qu.: 7.100      3rd Qu.:19.20       3rd Qu.:13.100
##  Max.   :2.0000   Max.   :27.300      Max.   :39.20       Max.   :30.300
##  assist_to_turnover_%    steal_%          block_%        turnover_%
##  Min.   : 0.00        Min.   : 0.000   Min.   : 0.000   Min.   : 0.00
##  1st Qu.: 7.00        1st Qu.: 1.100   1st Qu.: 0.500   1st Qu.: 9.90
##  Median :10.40        Median : 1.500   Median : 1.200   Median :12.50
##  Mean   :13.38        Mean   : 1.583   Mean   : 1.652   Mean   :12.82
##  3rd Qu.:17.80        3rd Qu.: 1.900   3rd Qu.: 2.300   3rd Qu.:15.20
##  Max.   :72.30        Max.   :11.100   Max.   :15.100   Max.   :43.60
##     usage_%      offensive_wins_shares defensive_wins_shares   win_shares
##  Min.   : 0.00   Min.   :-3.300        Min.   :0.000         Min.   :-2.10
##  1st Qu.:15.30   1st Qu.: 0.100        1st Qu.:0.400         1st Qu.: 0.50
##  Median :18.40   Median : 0.800        Median :1.000         Median : 1.80
##  Mean   :18.85   Mean   : 1.387        Mean   :1.272         Mean   : 2.66
##  3rd Qu.:21.80   3rd Qu.: 2.100        3rd Qu.:1.900         3rd Qu.: 3.80
##  Max.   :41.70   Max.   :13.800        Max.   :6.000         Max.   :17.90
##  win_shares_over_48 offensive_box_plus_minus defensive_box_plus_minus
##  Min.   :-0.28300   Min.   :-17.3000         Min.   :-8.5000
##  1st Qu.: 0.05000   1st Qu.: -2.4000         1st Qu.:-1.5000
##  Median : 0.08700   Median : -0.9000         Median :-0.3000
##  Mean   : 0.08683   Mean   : -0.9566         Mean   :-0.2671
##  3rd Qu.: 0.12100   3rd Qu.:  0.4000         3rd Qu.: 1.0000
##  Max.   : 0.63400   Max.   : 15.3000         Max.   :12.0000
##  box_plus_minus    value_over_replacement_player   field_goal
##  Min.   :-24.100   Min.   :-1.4000               Min.   :  0.0
##  1st Qu.: -3.100   1st Qu.:-0.1000               1st Qu.: 62.0
##  Median : -1.200   Median : 0.2000               Median :166.0
##  Mean   : -1.225   Mean   : 0.6493               Mean   :200.8
##  3rd Qu.:  0.700   3rd Qu.: 1.0000               3rd Qu.:294.0
##  Max.   : 15.600   Max.   :12.4000               Max.   :824.0
##  field_goal_attempts  field_goal_%      3_pointer      3_pointer_attempts
##  Min.   :   0.0      Min.   :0.0000   Min.   :  0.00   Min.   :  0.0
##  1st Qu.: 146.0      1st Qu.:0.4050   1st Qu.:  3.00   1st Qu.: 12.0
##  Median : 368.0      Median :0.4410   Median : 30.00   Median : 92.0
##  Mean   : 441.5      Mean   :0.4463   Mean   : 47.83   Mean   :133.8
##  3rd Qu.: 644.0      3rd Qu.:0.4810   3rd Qu.: 77.00   3rd Qu.:215.0
##  Max.   :1941.0      Max.   :1.0000   Max.   :402.00   Max.   :886.0
##    3_point_%        2_pointer   2_pointer_attempts   2_point_%
##  Min.   :0.0000   Min.   :  0   Min.   :   0.0     Min.   :0.0000
##  1st Qu.:0.2450   1st Qu.: 43   1st Qu.:  93.0     1st Qu.:0.4460
##  Median :0.3330   Median :113   Median : 235.0     Median :0.4830
##  Mean   :0.2846   Mean   :153   Mean   : 307.8     Mean   :0.4837
##  3rd Qu.:0.3750   3rd Qu.:219   3rd Qu.: 444.0     3rd Qu.:0.5290
##  Max.   :1.0000   Max.   :730   Max.   :1421.0     Max.   :1.0000
##  effective_field_goal_%  free_throws     free_throw_attempts  free_throw_%
##  Min.   :0.0000         Min.   :  0.00   Min.   :  0.0       Min.   :0.0000
##  1st Qu.:0.4670         1st Qu.: 23.00   1st Qu.: 33.0       1st Qu.:0.6740
##  Median :0.5010         Median : 59.00   Median : 78.0       Median :0.7640
##  Mean   :0.4986         Mean   : 92.23   Mean   :120.3       Mean   :0.7305
##  3rd Qu.:0.5360         3rd Qu.:120.00   3rd Qu.:161.0       3rd Qu.:0.8310
##  Max.   :1.0000         Max.   :746.00   Max.   :881.0       Max.   :1.0000
##  offensive_rebounds defensive_rebounds total_rebounds      assists
##  Min.   :  0.00     Min.   :  0        Min.   :   0.0   Min.   :  0.0
##  1st Qu.: 13.00     1st Qu.: 62        1st Qu.:  79.0   1st Qu.: 30.0
##  Median : 33.00     Median :143        Median : 178.0   Median : 74.0
##  Mean   : 52.69     Mean   :173        Mean   : 225.7   Mean   :115.5
##  3rd Qu.: 70.00     3rd Qu.:243        3rd Qu.: 307.0   3rd Qu.:151.0
##  Max.   :395.00     Max.   :817        Max.   :1198.0   Max.   :906.0
##      steals           blocks         turnovers      personal_fouls
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.0
##  1st Qu.: 14.00   1st Qu.:  5.00   1st Qu.: 25.00   1st Qu.: 47.0
##  Median : 33.00   Median : 15.00   Median : 57.00   Median :102.0
##  Mean   : 40.02   Mean   : 25.03   Mean   : 70.13   Mean   :103.4
##  3rd Qu.: 58.00   3rd Qu.: 33.00   3rd Qu.: 99.00   3rd Qu.:152.0
##  Max.   :169.00   Max.   :269.00   Max.   :464.00   Max.   :278.0
##      points
##  Min.   :   0.0
##  1st Qu.: 166.0
##  Median : 437.0
##  Mean   : 541.8
##  3rd Qu.: 780.0
## Max.   :2558.0
```

Pool 2k Data

Pull nba 2k ratings

```
secondary_attriutes <- c('name','position_s','overall','outside','inside','playmaking','athleticism','de
df_secondary <- vector('list',9)
names(df_secondary) <- secondary_attriutes
path_f = 'data/raw/nba2k/nba2k_%d.csv'
for (year in c(16:20)){
  df_year <- read.csv(sprintf(path_f,year))
  headers <- names(df_year)
  names(df_year) <- c('drop1',headers[1:length(headers)-1])
  df_year <- df_year[,c('name','position','ovr','out','ins','pla','ath','def','reb')]
  names(df_year) <- secondary_attriutes
  df_year[,'year'] <- 2000+year
  df_secondary <- rbind(df_secondary,df_year)}
df_secondary <- df_secondary[df_secondary$year%in%c(2016,2017),]
df_secondary[is.na(df_secondary)] <- 0
head(df_secondary)
```

```
##                      name position_s overall outside inside playmaking
## 1      '96 Michael Jordan         SG      99      95     88         91
## 2         '15 Kobe Bryant         SG      99      97     79         95
## 3           Stephen Curry         PG      99      98     66         98
## 4            LeBron James         SF      99      94     89         91
## 5 '71 Kareem Adbul-Jabbar          C      99      75     93         56
## 6            Kyrie Irving         PG      98      98     70         95
##   athleticism defending rebounding year
## 1          93        92         75 2016
## 2          84        88         65 2016
## 3          89        78         54 2016
## 4          92        91         91 2016
## 5          89        86         98 2016
## 6          91        74         49 2016
```

```
summary(df_secondary)
```

```
##                  name        position_s     overall         outside
##  Jimmy Butler     :  10   PG     :812   Min.   :40.00   Min.   :25.0
##  Kyrie Irving     :  10   SF     :782   1st Qu.:71.00   1st Qu.:62.0
##  Russell Westbrook:  10   SG     :749   Median :78.00   Median :73.0
##  Damian Lillard   :   9   PF     :710   Mean   :78.89   Mean   :71.3
##  Demar Derozan    :   9   C      :708   3rd Qu.:86.00   3rd Qu.:82.0
##  James Harden     :   9   C/PF   :  0   Max.   :99.00   Max.   :99.0
##  (Other)          :3704   (Other):  0
##      inside        playmaking     athleticism      defending
##  Min.   :25.00   Min.   :25.00   Min.   :25.00   Min.   :25.00
##  1st Qu.:58.00   1st Qu.:48.00   1st Qu.:68.00   1st Qu.:58.00
##  Median :64.00   Median :61.00   Median :74.00   Median :65.00
##  Mean   :65.43   Mean   :62.04   Mean   :73.68   Mean   :66.28
##  3rd Qu.:72.00   3rd Qu.:76.00   3rd Qu.:80.00   3rd Qu.:73.00
##  Max.   :98.00   Max.   :99.00   Max.   :98.00   Max.   :98.00
##
```

```
##    rebounding         year     
##  Min.   :25.00   Min.   :2016  
##  1st Qu.:43.00   1st Qu.:2016  
##  Median :57.00   Median :2016  
##  Mean   :59.62   Mean   :2016  
##  3rd Qu.:75.00   3rd Qu.:2017  
##  Max.   :99.00   Max.   :2017  
##
```