Lab 3

Lab Overview

For this question, a subset of the tables contained in the History of Baseball database are available. Additional details are available here: https://www.kaggle.com/seanlahman/the-history-of-baseball. The following tables will be used for these questions:

- player
- all_star
- salary

```
library(readr)
library(dplyr)
library(knitr)
library(tidyr)
player <- read_csv("http://math.montana.edu/ahoegh/teaching/stat408/datasets/player.csv")
all_star <- read_csv("http://math.montana.edu/ahoegh/teaching/stat408/datasets/all_star.csv")
salary <- read_csv("http://math.montana.edu/ahoegh/teaching/stat408/datasets/salary.csv")
```

1. (2 points)

How many players were born in Montana.

```
num_MT <- player %>% filter(birth_state == 'MT') %>% tally() %>% pull()
```

There are 24 players in the database that were born in Montana.

2. (4 points)

Print a table that contains each player born in Montana. The table should contain the player_id as well as given name and their total salary across all years. If salary is not available (pre-1985), include the player but have an NA for salary.

```
total_salary <- salary %>% group_by(player_id) %>% summarize( total_salary = sum(salary))

MT_player <- player %>% filter(birth_state == 'MT') %>% left_join(total_salary, by = "player_

kable(MT_player)
```

player_id	name_first	name_last	total_salary
ballaje01	Jeff	Ballard	937000
bouched01	Ed	Bouchee	NA
browsc01	Scott	Brow	564000
couchjo01	Johnny	Couch	NA
doyleje02	Jeff	Doyle	NA
duffla01	Larry	Duff	NA
gibbojo02	John	Gibbons	140000
grahaty01	Tyler	Graham	NA
himslve99	Vedie	Himsl	NA
johnsro07	Rob	Johnson	1234200
lowenjo01	John	Lowenstein	NA
mcintjo01	Joe	McIntosh	NA
mcnalda01	Dave	McNally	NA
meierda01	Dave	Meier	NA
mickoka01	Kam	Mickolio	417000
neibaga01	Gary	Neibauer	NA
ottenji01	Jim	Otten	NA
plewshe01	Herb	Plews	NA
ryanro02	Rob	Ryan	200000
schmicu01	Curt	Schmidt	NA
tanketa01	Taylor	Tankersley	772500
tobinma01	Mason	Tobin	414000
tyackji01	Jim	Tyack	NA
willist01	Steamboat	Williams	NA

3. (2 points)

Create a thin dataset for that contains the yearly salaries of David Ortiz, Derek Jeter, and Troy Tulowitzki.

```
thin_salaries <- player %>% filter(name_last %in% c('Tulowitzki', 'Jeter', 'Ortiz')) %>% filt

kable(thin_salaries)
```

name_first	name_last	year	salary
Derek	Jeter	1996	130000
Derek	Jeter	1997	550000
Derek	Jeter	1998	750000
David	Ortiz	1998	170000
Derek	Jeter	1999	5000000
Derek	Jeter	2000	10000000
David	Ortiz	2000	220000
Derek	Jeter	2001	12600000
David	Ortiz	2001	260000
Derek	Jeter	2002	14600000
David	Ortiz	2002	950000
Derek	Jeter	2003	15600000
David	Ortiz	2003	1250000
Derek	Jeter	2004	18600000
David	Ortiz	2004	4587500
Derek	Jeter	2005	19600000
David	Ortiz	2005	5250000
Derek	Jeter	2006	20600000
David	Ortiz	2006	6500000
Derek	Jeter	2007	21600000
David	Ortiz	2007	13250000
Troy	Tulowitzki	2007	381000
Derek	Jeter	2008	21600000
David	Ortiz	2008	13000000
Troy	Tulowitzki	2008	750000
Derek	Jeter	2009	21600000
David	Ortiz	2009	13000000
Troy	Tulowitzki	2009	1000000
Derek	Jeter	2010	22600000
David	Ortiz	2010	13000000
Troy	Tulowitzki	2010	3500000

name_first	name_last	year	salary
Derek	Jeter	2011	14729364
David	Ortiz	2011	12500000
Troy	Tulowitzki	2011	5500000
Derek	Jeter	2012	15729364
David	Ortiz	2012	14575000
Troy	Tulowitzki	2012	8250000
Derek	Jeter	2013	16729365
David	Ortiz	2013	14500000
Troy	Tulowitzki	2013	10000000
Derek	Jeter	2014	12000000
David	Ortiz	2014	15000000
Troy	Tulowitzki	2014	16000000
David	Ortiz	2015	16000000
Troy	Tulowitzki	2015	20000000

4. (4 points)

Create a wide dataset for that contains the salaries of David Ortiz, Derek Jeter, and Troy Tulowitzki.

```
wide_salaries <- thin_salaries %>% mutate(year = paste('year',year, sep = '')) %>%
  pivot_wider(names_from = c(year), values_from = salary)
wide_salaries
```

```
# A tibble: 3 x 22
  name_first name_last  year1996 year1997 year1998 year1999 year2000 year2001
  <chr>      <chr>         <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
1 Derek      Jeter        130000   550000   750000  5000000 10000000 12600000
2 David      Ortiz            NA       NA   170000       NA   220000   260000
3 Troy       Tulowitzki       NA       NA       NA       NA       NA       NA
# i 14 more variables: year2002 <dbl>, year2003 <dbl>, year2004 <dbl>,
#   year2005 <dbl>, year2006 <dbl>, year2007 <dbl>, year2008 <dbl>,
#   year2009 <dbl>, year2010 <dbl>, year2011 <dbl>, year2012 <dbl>,
#   year2013 <dbl>, year2014 <dbl>, year2015 <dbl>
```

5. (4 points)

Which player(s) made the most appearances as an all star representing the National League (NL)?

```
all_star %>% filter(league_id == 'NL') %>% group_by(player_id) %>% tally() %>% arrange(desc(n
```

name_first	name_last	num_allstars
Hank	Aaron	24
Willie	Mays	24
Stan	Musial	24