# Data Joining and Data Visualization

Jun Yu Chen

10/8/2022

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.2
```

```
##
## 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
```

```
library(ggplot2)

##Filter NBA seasons after the year 2005
raw_nba_set<-read.csv("PER.csv")
new_nba_set<-raw_nba_set%>%filter(Year>="2005")%>%select(Year,Player,Pos,G,MP,PER)


##Naming the raw injury dataset
raw_injury_set<-read.csv("injuries_2010-2020.csv")

##Selecting players who had out of season injuries
OutSeason<-raw_injury_set%>%filter(grepl("out for season",Notes))
Years<-substr(OutSeason$Date,1,4)
OutSeason$Years<-as.integer(substr(OutSeason$Date,1,4))
OutSeason<-OutSeason[,-c(1,3)]


##Joining datasets together
Injured_joined<-new_nba_set%>%inner_join(OutSeason,by=c("Player"="Relinquished","Year"="Years"))%>%dist
head(Injured_joined)
```

```
##   Year         Player Pos  G   MP  PER      Team
## 1 2010  Udonis Haslem  PF 78 2177 14.6      Heat
## 2 2010  Jonas Jerebko  PF 80 2232 13.9   Pistons
```

```
## 3 2010      Greg Oden   C 21  502 23.1   Blazers
## 4 2011  Ryan Anderson  PF 64 1424 19.0     Magic
## 5 2011 Darrell Arthur  PF 80 1609 15.7 Grizzlies
## 6 2011      Omer Asik   C 82  989 11.8     Bulls
##                                                                                 Notes
## 1                              placed on IL with torn ligament in left foot (out for season)
## 2 placed on IL recovering from surgery to repair torn right Achilles tendon (out for season)
## 3                                    placed on IL with left knee injury (out for season)
## 4                                                        placed on IL (out for season)
## 5                                        ruptured right Achilles tendon (out for season)
## 6                                        stress fracture in left fibula (out for season)
```

```
library(showtext)
```

```
## Warning: package 'showtext' was built under R version 4.1.2
```

```
## Loading required package: sysfonts
```

```
## Warning: package 'sysfonts' was built under R version 4.1.2
```

```
## Loading required package: showtextdb
```

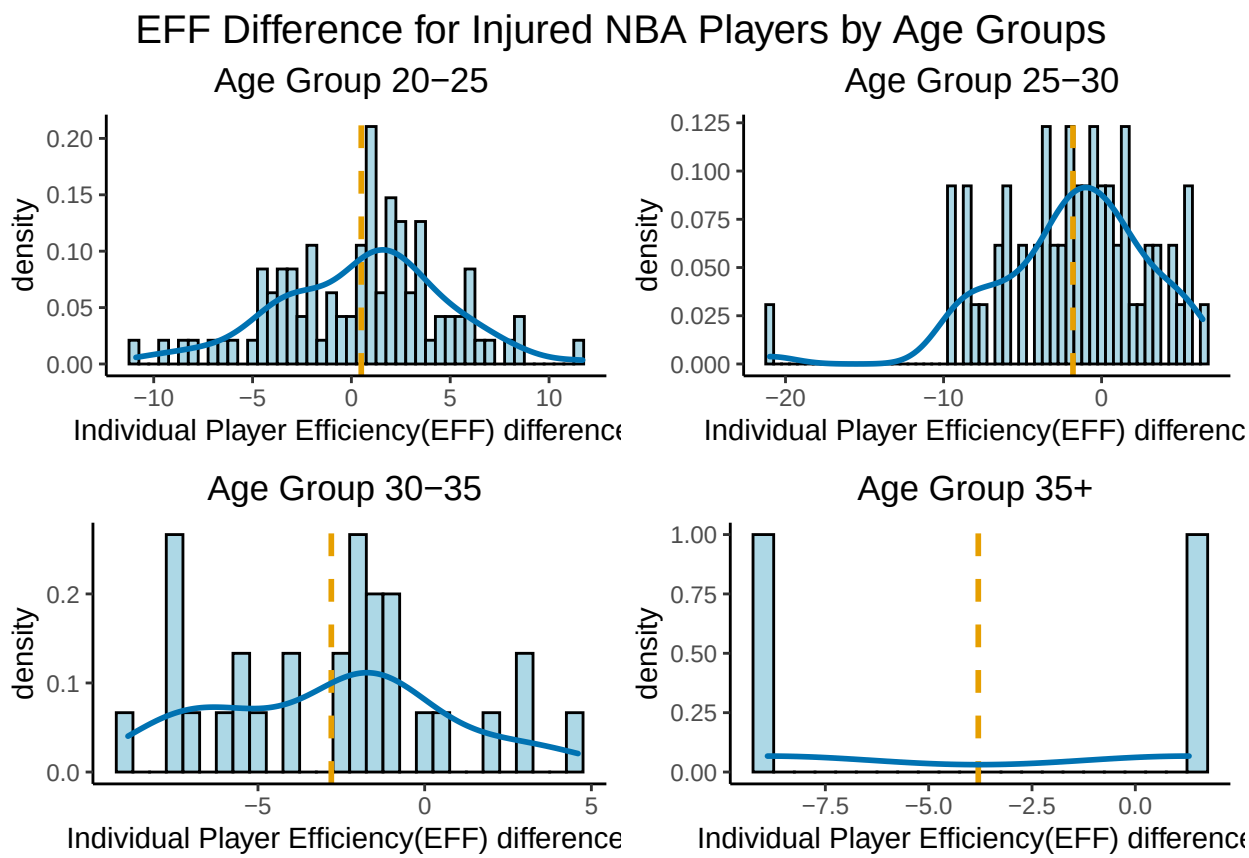
```
library(dplyr)
library(ggplot2)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

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

```
library(grid)

##import cleaned data
injury_data_cleaned<-read.csv("injury_data_cleaned.csv")
injury_data_cleaned<-injury_data_cleaned%>%mutate(Avg_EFF_diff=After_EFF-Prev_EFF)

##EFF difference
EFF_difference<-injury_data_cleaned%>%ggplot(aes(x=Avg_EFF_diff,y=..density..))+
  geom_histogram(binwidth=.5, colour="black", fill="lightblue") +
  geom_vline(aes(xintercept=mean(Avg_EFF_diff, na.rm=TRUE)),
               color="darkblue", linetype="dashed", size=1)+
  geom_density(color = "pink",size=1)+
  xlab("Individual Player Efficiency(EFF) difference ") +
  ggtitle("Individual Player Efficiency(EFF) Difference for Injured NBA Players ")+
  theme(plot.title = element_text(hjust = 0.5))+theme(panel.grid.major = element_blank(), panel.grid.mi
```

```r
#EFF difference by Age groups

Age_18_25<-injury_data_cleaned%>%filter(Age>=18&Age<=25)%>%ggplot(aes(x=Avg_EFF_diff,y=..density..))+
  geom_histogram(binwidth=.5, colour="black", fill="lightblue")+
  geom_vline(aes(xintercept=mean(Avg_EFF_diff, na.rm=TRUE)),  color="#E69F00", linetype="dashed", size=
  geom_density(color = "#0072B2" ,size=1)+
  theme(panel.grid.major = element_blank(), panel.grid.minor =   element_blank(),panel.background =
  xlab("Individual Player Efficiency(EFF) difference ") +
  ggtitle("Age Group 20-25 ")+
  theme(plot.title = element_text(hjust = 0.5))

Age_25_30<-injury_data_cleaned%>%filter(Age>25&Age<=30)%>%ggplot(aes(x=Avg_EFF_diff,y=..density..))+
  geom_histogram(binwidth=.5, colour="black", fill="lightblue")+
  geom_vline(aes(xintercept=mean(Avg_EFF_diff, na.rm=TRUE)),  color="#E69F00", linetype="dashed", size=
  geom_density(color = "#0072B2" ,size=1)+
  theme(panel.grid.major = element_blank(), panel.grid.minor =   element_blank(),panel.background =
  xlab("Individual Player Efficiency(EFF) difference ") +
  ggtitle("Age Group 25-30 ")+
  theme(plot.title = element_text(hjust = 0.5))

Age_30_35<-injury_data_cleaned%>%filter(Age>30&Age<=35)%>%ggplot(aes(x=Avg_EFF_diff,y=..density..))+
  geom_histogram(binwidth=.5, colour="black", fill="lightblue")+
  geom_vline(aes(xintercept=mean(Avg_EFF_diff, na.rm=TRUE)),  color="#E69F00", linetype="dashed", size=
  geom_density(color = "#0072B2" ,size=1)+
  theme(panel.grid.major = element_blank(), panel.grid.minor =   element_blank(),panel.background =
  xlab("Individual Player Efficiency(EFF) difference ") +
  ggtitle("Age Group 30-35 ")+
  theme(plot.title = element_text(hjust = 0.5))

Age_35_40<-injury_data_cleaned%>%filter(Age>35)%>%ggplot(aes(x=Avg_EFF_diff,y=..density..))+
  geom_histogram(binwidth=.5, colour="black", fill="lightblue")+
  geom_vline(aes(xintercept=mean(Avg_EFF_diff, na.rm=TRUE)),  color="#E69F00", linetype="dashed", size=
  geom_density(color = "black" )+
  geom_density(color = "#0072B2" ,size=1)+
  theme(panel.grid.major = element_blank(), panel.grid.minor =element_blank(),panel.background =
  xlab("Individual Player Efficiency(EFF) difference ") +
  ggtitle("Age Group 35+ ")+
  theme(plot.title = element_text(hjust = 0.5))

grid.arrange(Age_18_25,Age_25_30,Age_30_35,Age_35_40,ncol=2,top = textGrob("EFF Difference for Injured
```




```
plot
```

```
## function (x, y, ...)
## UseMethod("plot")
## <bytecode: 0x7fe969ff1898>
## <environment: namespace:base>
```

```
##Quartiles of Weight
with(injury_data_cleaned,summary(Weight))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   76.66   92.65  102.06  100.51  108.86  131.09
```

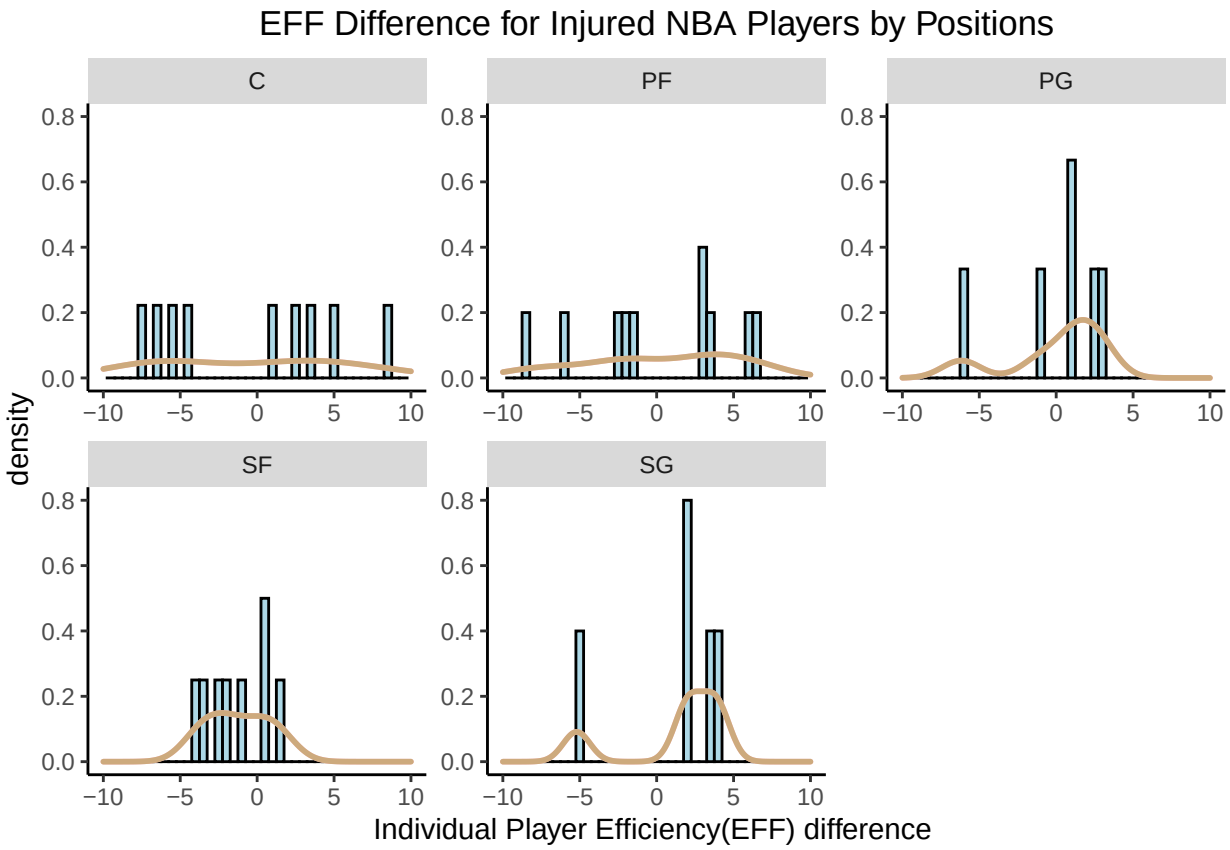
```
#EFF difference by Position
injury_data_cleaned%>%filter(Pos==c("C","PF","SF","SG","PG"))%>%ggplot(aes(x=Avg_EFF_diff,y=..density..)
  geom_histogram(binwidth=.5, colour="black", fill="lightblue")+
  geom_density(color = "pink",size=1)+
  geom_density(color = "burlywood3" ,size=1)+
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),panel.background =
  xlab("Individual Player Efficiency(EFF) difference ") +
  ggtitle("EFF Difference for Injured NBA Players by Positions")+
  theme(plot.title = element_text(hjust = 0.5))+
  facet_wrap(~Pos,scales='free',ncol=3)+
  scale_x_continuous(limits=c(-10,10)) + scale_y_continuous(limits=c(0,0.8))
```

```
## Warning in Pos == c("C", "PF", "SF", "SG", "PG"): longer object length is not a
## multiple of shorter object length
```

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 1 rows containing non-finite values (stat_density).
## Removed 1 rows containing non-finite values (stat_density).
```

```
## Warning: Removed 10 rows containing missing values (geom_bar).
```



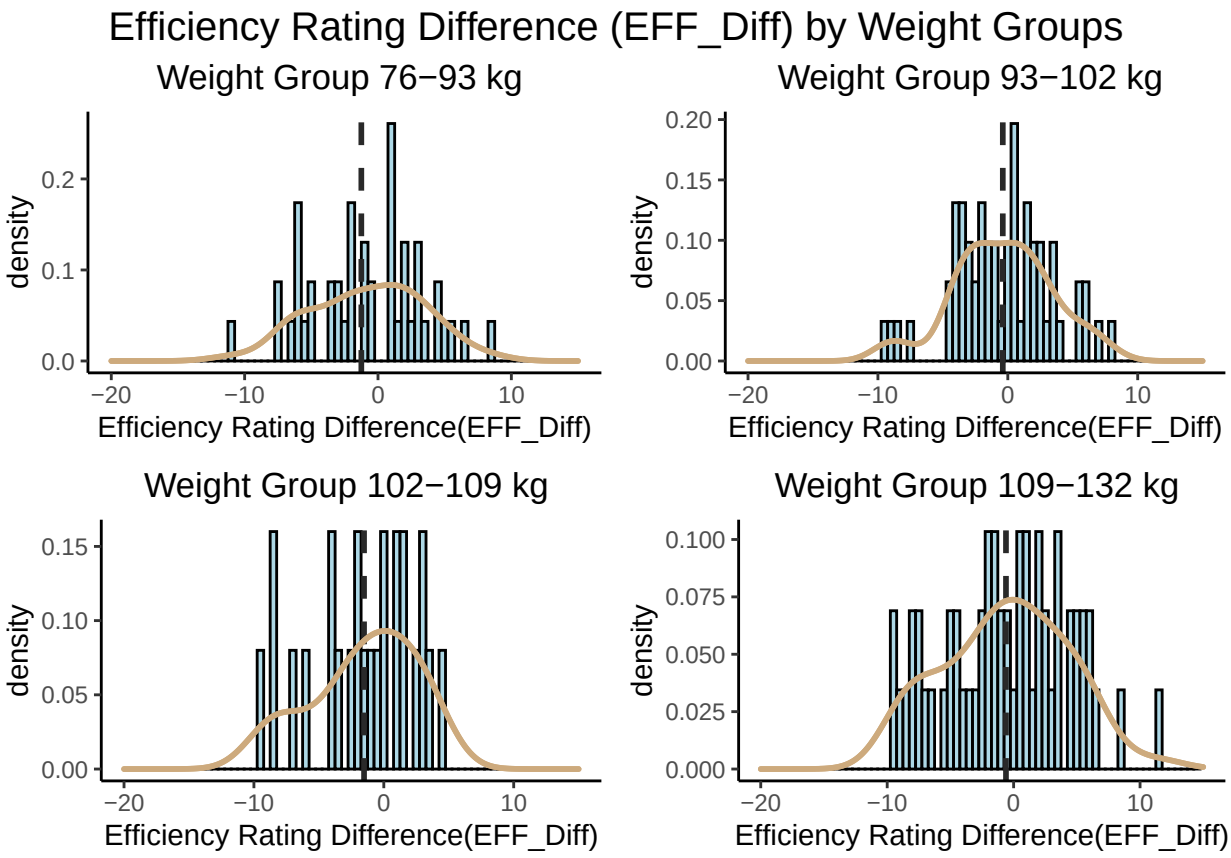

```
##EFF by Weight
Weight_77_93<-injury_data_cleaned%>%filter(Weight>76.66&Weight<=92.65)%>%ggplot(aes(x=Avg_EFF_diff,y=..c
  geom_histogram(binwidth=.5, colour="black", fill="lightblue")+
  geom_vline(aes(xintercept=mean(Avg_EFF_diff, na.rm=TRUE)),  color="gray16", linetype="dashed", size=1
  geom_density(color = "burlywood3" ,size=1)+
  theme(panel.grid.major = element_blank(), panel.grid.minor =   element_blank(),panel.background =
  xlab("Efficiency Rating Difference(EFF_Diff)") +
  ggtitle("Weight Group 76-93 kg ")+
  theme(plot.title = element_text(hjust = 0.5))+
  geom_density(color = "burlywood3",size=1 )+
  scale_x_continuous(limits=c(-20,15))

Weight_93_102<-injury_data_cleaned%>%filter(Weight>92.65&Weight<=102.06)%>%ggplot(aes(x=Avg_EFF_diff,y=
  geom_histogram(binwidth=.5, colour="black", fill="lightblue")+
  geom_vline(aes(xintercept=mean(Avg_EFF_diff, na.rm=TRUE)),  color="gray16", linetype="dashed", size=1
```

```
  geom_density(color = "burlywood3" ,size=1)+
  theme(panel.grid.major = element_blank(), panel.grid.minor =   element_blank(),panel.background =
  xlab("Efficiency Rating Difference(EFF_Diff)") +
  ggtitle("Weight Group 93-102 kg ")+
  theme(plot.title = element_text(hjust = 0.5))+
  geom_density(color = "burlywood3",size=1 )+
  scale_x_continuous(limits=c(-20,15))

Weight_102_109<-injury_data_cleaned%>%filter(Weight>102.06&Weight<=108.86)%>%ggplot(aes(x=Avg_EFF_diff,
  geom_histogram(binwidth=.5, colour="black", fill="lightblue")+
  geom_vline(aes(xintercept=mean(Avg_EFF_diff, na.rm=TRUE)),  color="gray16", linetype="dashed", size=1
  geom_density(color = "burlywood3" ,size=1)+
  theme(panel.grid.major = element_blank(), panel.grid.minor =   element_blank(),panel.background =
  xlab("Efficiency Rating Difference(EFF_Diff)") +
  ggtitle("Weight Group 102-109 kg ")+
  theme(plot.title = element_text(hjust = 0.5))+
  geom_density(color = "burlywood3",size=1 )+
  scale_x_continuous(limits=c(-20,15))


Weight_109_132<-injury_data_cleaned%>%filter(Weight>108.86&Weight<=131.09)%>%ggplot(aes(x=Avg_EFF_diff,
  geom_histogram(binwidth=.5, colour="black", fill="lightblue")+
  geom_vline(aes(xintercept=mean(Avg_EFF_diff, na.rm=TRUE)),  color="gray16", linetype="dashed", size=1
  geom_density(color = "burlywood3" ,size=1)+
  theme(panel.grid.major = element_blank(), panel.grid.minor =   element_blank(),panel.background =
  xlab("Efficiency Rating Difference(EFF_Diff)") +
  ggtitle("Weight Group 109-132 kg ")+
  theme(plot.title = element_text(hjust = 0.5))+
  geom_density(color = "burlywood3",size=1 )+
  scale_x_continuous(limits=c(-20,15))

grid.arrange(Weight_77_93,Weight_93_102,Weight_102_109,Weight_109_132,ncol=2,top = textGrob("Efficiency
```

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 1 rows containing non-finite values (stat_density).
## Removed 1 rows containing non-finite values (stat_density).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
## Removed 2 rows containing missing values (geom_bar).
## Removed 2 rows containing missing values (geom_bar).
## Removed 2 rows containing missing values (geom_bar).
```


Efficiency Rating Difference (EFF_Diff) by Weight Groups
Weight Group 76−93 kg
Weight Group 93−102 kg
Weight Group 102−109 kg
Weight Group 109−132 kg
density
Efficiency Rating Difference(EFF_Diff)