Statistical Analysis of Capstone Project

Carl Larson

2/1/2018

Statistical Analysis of Capstone Data Set

There are plenty of interesting statistics questions we can ask about a dataset such as the gameflow table.

Can We Count Something Interesting?

Sure. We can count a lot of different things in this data set, and to be on the level of this question, let's count the total score of this game for each team.

```
require(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.2
```

```
require(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 3.4.2
```

```
require(ggplot2)
require(reshape2)

df <- read.csv("/Users/EagleFace/Downloads/Gameflow of Bulls76ersApr282012.csv", header = FALSE)

(df$V1[nrow(df)]);(df$V2[nrow(df)])
```

```
## [1] 91
```

```
## [1] 103
```

I think this is interesting, because it shows us the end of the story here, and this code should work for any gameflow table that is put in.

2. Can You Find Some Trends?(high, low, increase, decrease, anomalies)?

We can definitely find some trends in the gameflow chart.

What was each team's largest lead?

```
TeamAlargestLead <- max(df$V1-df$V2)

TeamBlargestLead <- abs(min(df$V1-df$V2))

#Team A's Largest Lead:
print(TeamAlargestLead)
```

```
## [1] 3
```

```
#Team B's Largest Lead:
print(TeamBlargestLead)
```

```
## [1] 20
```

3. Can We Make A Bar Plot Or Histogram?

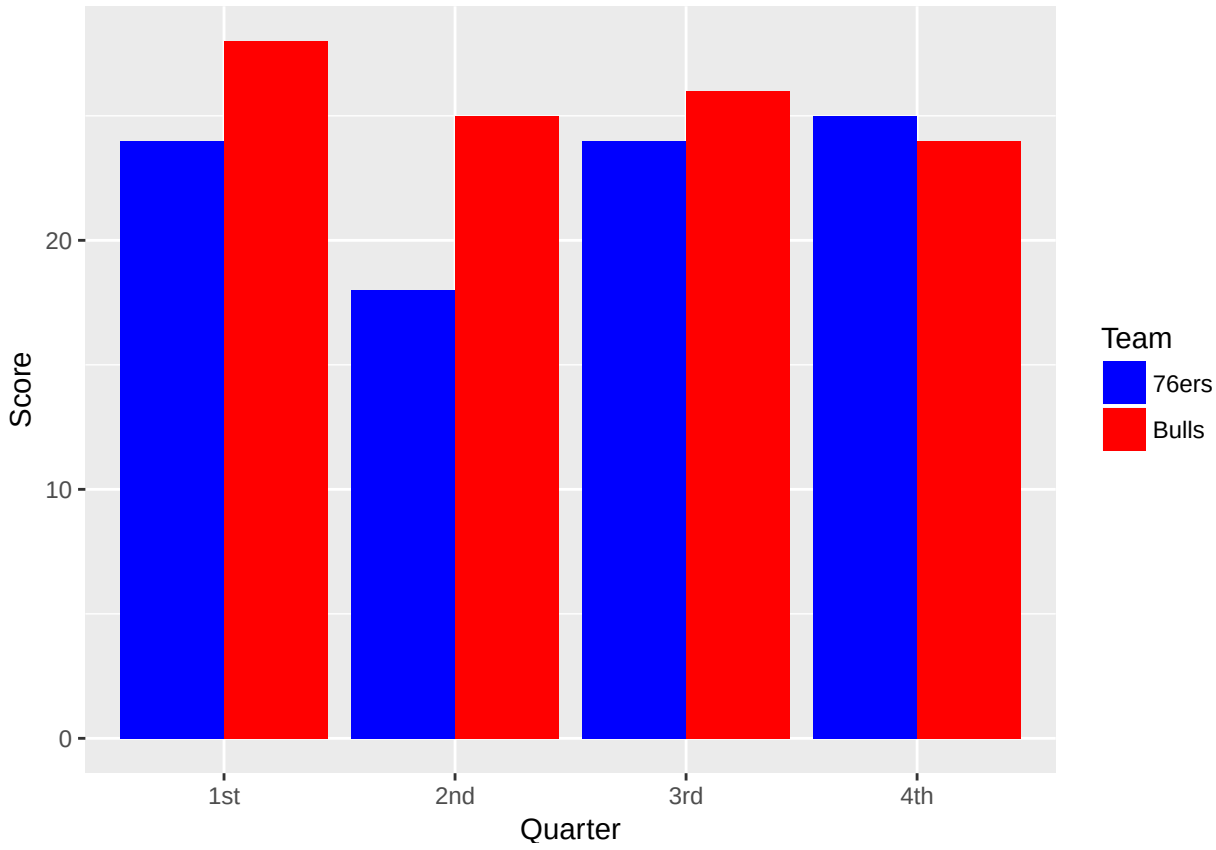
Absolutely. We can make an interesting one.

Let's look at each team's scoring, broken down by quarter. It might be interesting to see what teams know how to turn it on in the 2nd half or 4th quarter, for example.

```
#First to subset out the quarters into their own matrices.
dfq1 <- subset(df, df$V4==1)
dfq2 <- subset(df, df$V4==2)
dfq3 <- subset(df, df$V4==3)
dfq4 <- subset(df, df$V4==4)
```

This helps us avoid having to use for-while loops.

```
#Here we set up clean variables to populate
#the histogram based on the raw data

Q1scoreA <- max(dfq1$V1)
Q1scoreB <- max(dfq1$V2)

Q2scoreA <- (max(dfq2$V1)-Q1scoreA)
Q2scoreB <- (max(dfq2$V2)-Q1scoreB)

Q3scoreA <- (max(dfq3$V1))-(Q2scoreA+Q1scoreA)
Q3scoreB <- (max(dfq3$V2))-(Q2scoreB+Q1scoreB)

Q4scoreA <- (max(dfq4$V1))-(Q3scoreA+Q2scoreA+Q1scoreA)
Q4scoreB <- (max(dfq4$V2))-(Q3scoreB+Q2scoreB+Q1scoreB)


dat1 <- data.frame(
  Team = factor(c("76ers", "Bulls")),
  Quarter = factor(c("1st", "1st", "2nd", "2nd", "3rd", "3rd", "4th", "4th")), levels = c("Team A", "Tea
  Score = c(Q1scoreA, Q1scoreB, Q2scoreA, Q2scoreB, Q3scoreA, Q3scoreB, Q4scoreA, Q4scoreB)
)

ggplot(data=dat1, aes(x=Quarter, y=Score, fill = Team)) +
  geom_bar(stat="identity", position="dodge") +
  scale_fill_manual(values = c("blue","red"))
```




```
#This histogram is coming out great, and the bars are even color-coded for the appropriate team colors
```

As we can see here, the home team, the Bulls, did win each of the first three quarters. It's noteworthy that the Bulls had their biggest lead halfway through the 4th, and the 76ers did come back to win that quarter. Indeed, the Bulls' coach wasn't technically wrong when he said "The score was going the other way," late in the game, the mistake was a lack of dynamic in-game analytical tools for the coach like this garbage time formula.

4. Can You Compare Two Related Quantities?

Certainly. We can compare the team scores and determine which team won, or for games that are in progress, we can determine who is in the lead. Not a huge feat, but still a nice one to show programatically.

```
#Note that for games that are over, this only determines who won. For games that #are in progress, this
#final score team, in the context of team A or B

fstA <- as.numeric(df$V1[nrow(df)])
fstB <- as.numeric(df$V2[nrow(df)])

WhoWon <- function(fstA, fstB){
  if (fstA>fstB){
    return("Team A won or is in the lead.")
      }else{
    return("Team B won or is in the lead.")}}

WhoWon(fstA, fstB)
```

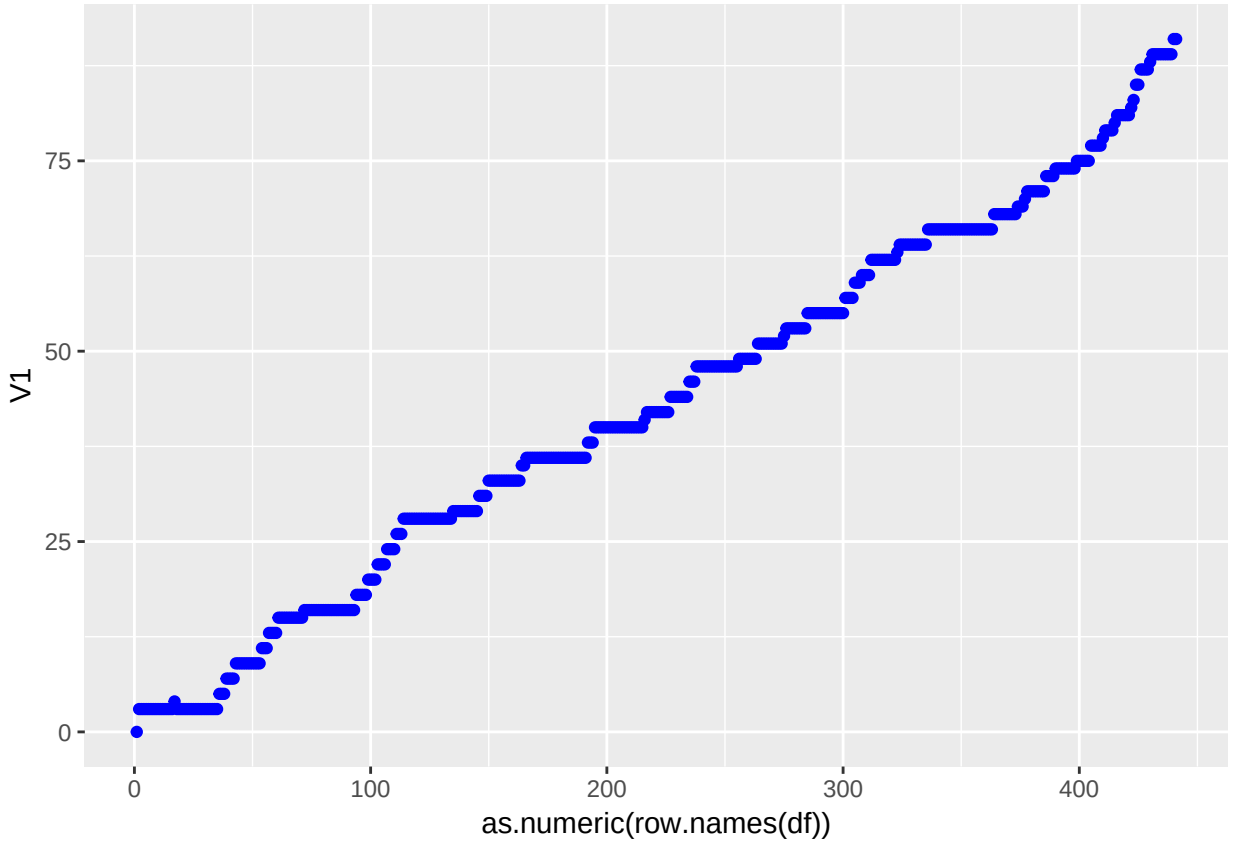
```
## [1] "Team B won or is in the lead."
```

So with this function above, all we need is a gameflow table, and this function will tell us who is in the lead or who won.

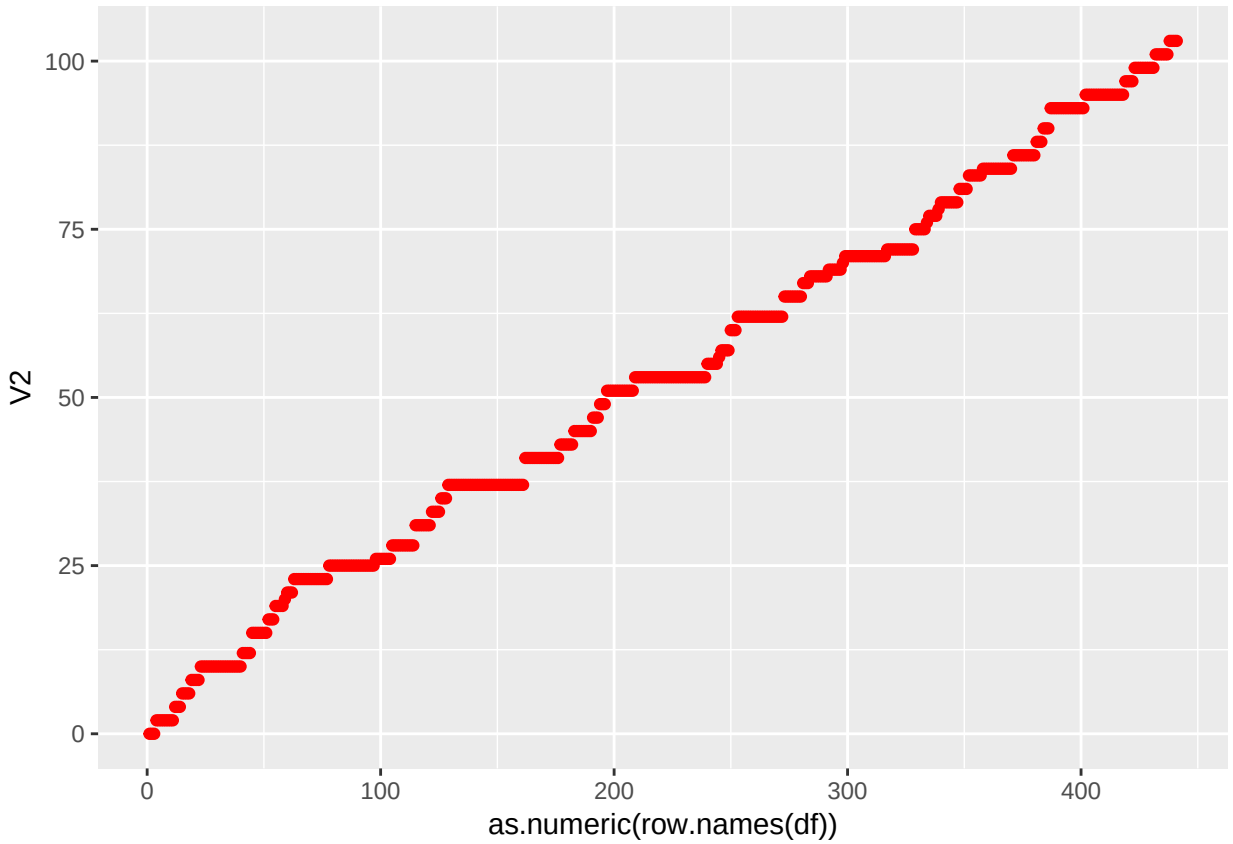
5. Can You Make A Scatterplot?

Sure. We can make scatterplots of the points for each team, individually, or on the same chart.

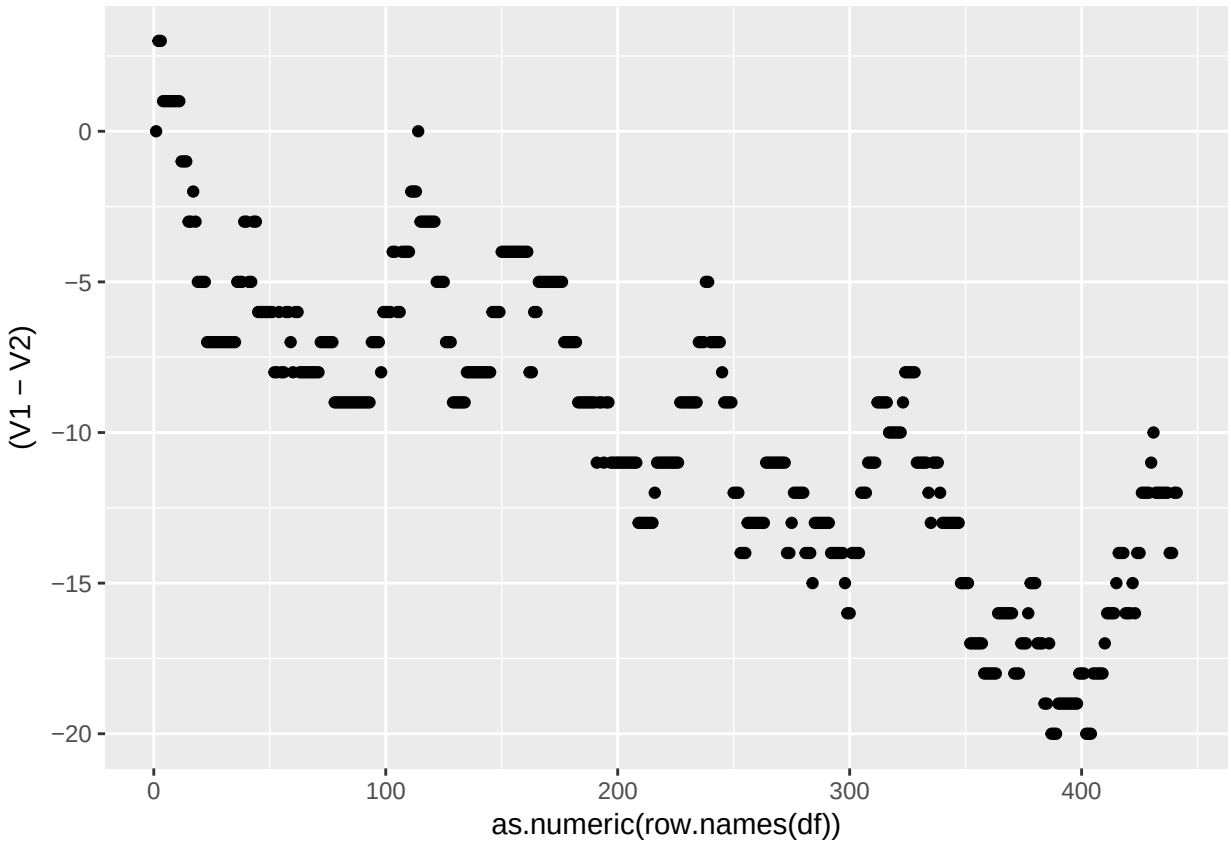
```
ggplot(df, aes(x= as.numeric(row.names(df)), y=V1))+
  geom_point(color="blue")
```




```
ggplot(df, aes(x=as.numeric(row.names(df)), y=V2))+
  geom_point(color="red")
```




```
ggplot(df, aes(x=as.numeric(row.names(df)), y=(V1-V2)))+
  geom_point()
```

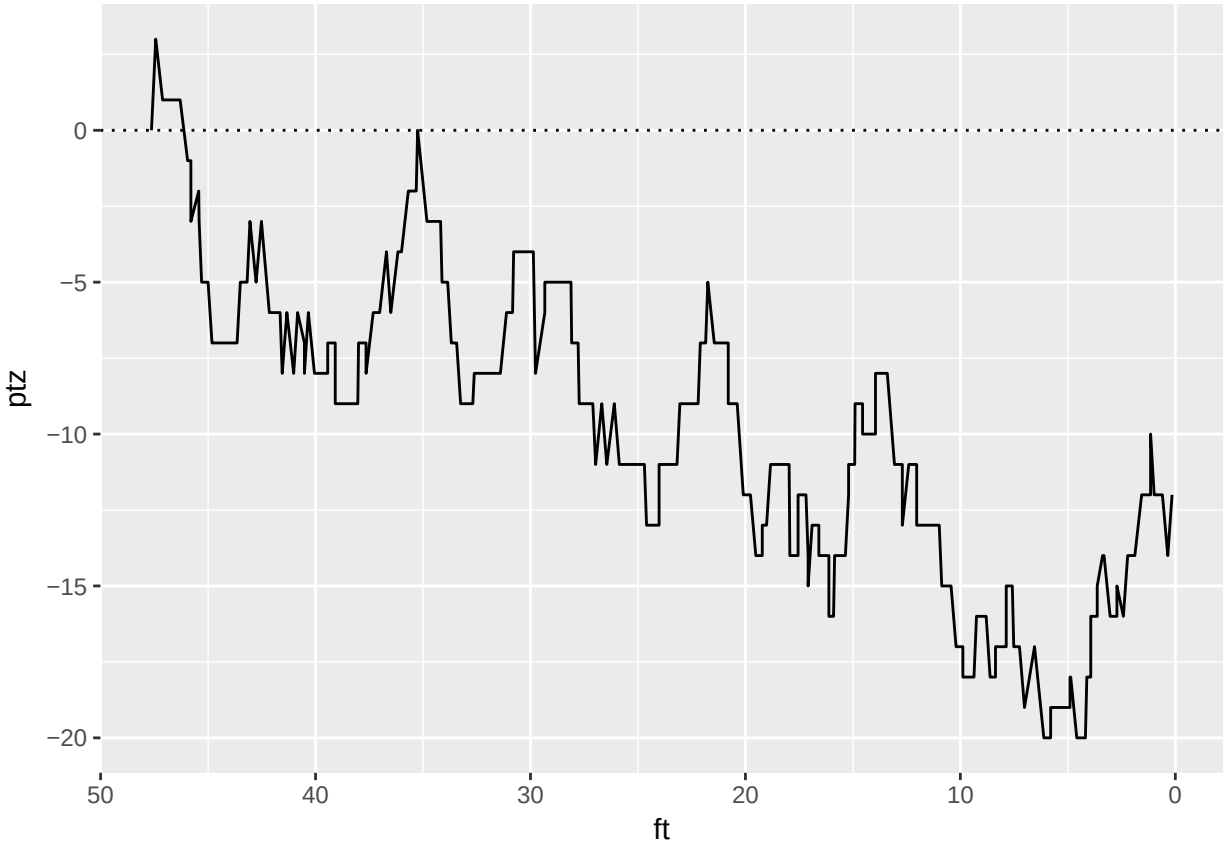


6. Can You Make A Time-Series Plot?

Absolutely. Time-series is the very nature of this type of data, and time-series really is the best way to look at this data. The data naturally lends itself to time-series analysis, so it will be very straightforward to get the actual time-series data, we just need to clean up the time column.

```
df <- separate(df, V3, c("MIN", "SEC"), sep = ":", remove = TRUE)

df$MIN <- as.numeric(df$MIN)
df$SEC <- as.numeric(df$SEC)

#Now that we have a minutes and seconds column and have scrapped
#the colon in the clock variable V3, we can declare our #formatted time variable as "ft" for formatted

ft <- numeric(length = nrow(df))
ptz <- df$V1 - df$V2

for(i in 1:nrow(df)){
  while(df$V4[i] == 1){
    ft[i] <- (df$MIN[i] + (df$SEC[i]/60) + 36)
      (i=i+1)}

  while(df$V4[i] == 2){
    ft[i] <- (df$MIN[i] + (df$SEC[i]/60) + 24)
      (i=i+1)}
```

```
  while(df$V4[i] == 3){
    ft[i] <- (df$MIN[i] + (df$SEC[i]/60) + 12)
      (i=i+1)}

  while(df$V4[i] == 4){
    ft[i] <- (df$MIN[i] + (df$SEC[i]/60))
    if(i==(nrow(df))){break}
    else{
      (i=i+1)}}}

ggplot(df, aes(x=ft, y=ptz))+
  geom_line() +
  scale_x_reverse() +
  geom_hline(yintercept=0, linetype="dotted")
```



Here we have a true time-series graph, showing the story of the 2012 Bulls' fall from greatness.

What Other Statistical Questions Can We Ask?

There are many interesting second-level questions we can ask such as, do coaches who employ this strategy lengthen the careers of their players?

Conversely, do coaches who break the garbage time formula shorten the length of their players' careers?

There is no way to directly measure career expectancy by coach. Coaches also change their styles over time, ocasionally, but some are consistent on other topics.

Some coaches also switch teams often, while others have been tenured for a long time.

The longest-tenured coach in the NBA is Gregg Popovich of the San Antonio Spurs. He started with the Spurs in 1996, and is widely regarded as the best coach in the NBA.

Two interesting future hypotheses would be, does Popovich coach in accordance with the garbage time formula, more than the average for NBA coaches?

Additionally, we can ask, do Spurs' players under Popovich, such as Tim Duncan (who played 19 seasons), Manu Ginobili (whose professional playing career began 23 years ago in 1995), and Tony Parker (who is in his 16th NBA season), have a longer than average career length than the rest of the NBA? Intuitively it seems obvious that this answer would be yes, and I am confident this could be proved with the data.

But how do the Spurs compare in this regard to the average for playoff teams, rather than just the whole NBA?

As you can see, questions like this are as interesting as they are difficult to answer specifically. Personally, in my watching of the NBA, it seems to me that indeed Popovich's Spurs do sub out their star players far earlier in games than the NBA average. I would say this *does* have a positive impact on lengthening his players' careers.

There is little doubt in my mind that the Spurs employ this type of garbage time formula, and that it lengthens their players' carrers, and contributes to their ability to win championships.