Final Project

Mark Bryant

6/14/2021

Question: What age group do MLB players have a productive season as a hitter

Data Initiation Baseball Stats

```
 ##install.packages("ISLR")
 ##install.packages("dplyr")
 ##install.packages("ggplot2")
 ##install.packages("tidyverse")
 ##install.packages(ggpubr)
library(ISLR)
library(dplyr)

##
## 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)
library(tidyverse)

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

## v tibble  3.1.2     v purrr   0.3.4
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(readr)
library(ggpubr)
```

```
stats <- read_csv("~/115/Final Project/stats (1).csv") ## Installed all
packages and data I needed for this project.
```

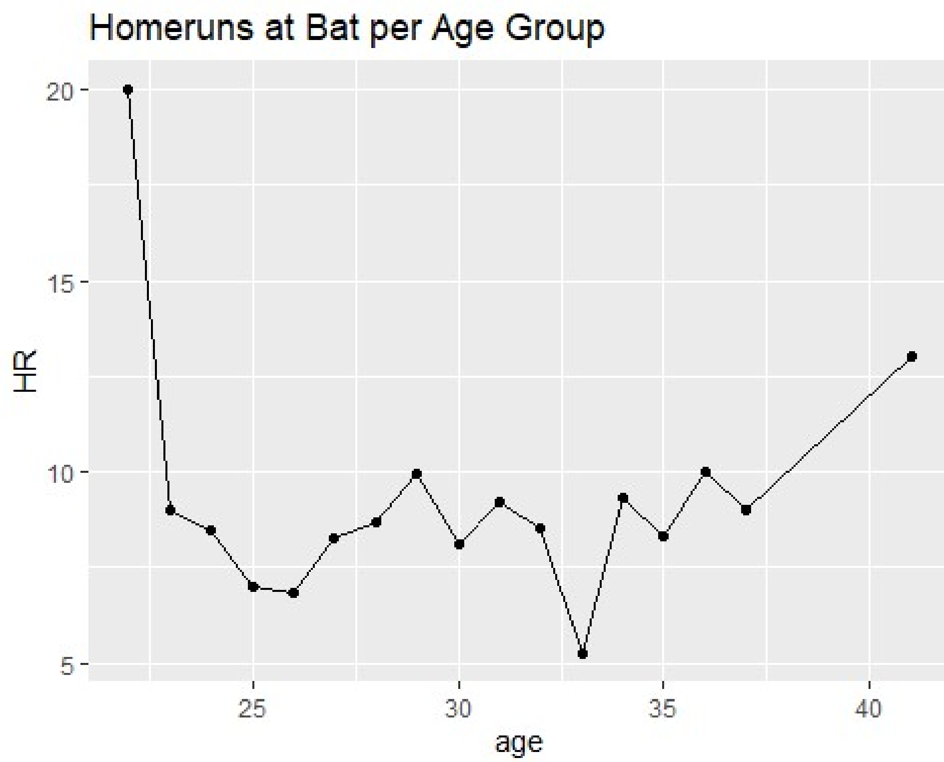
```
## Warning: Missing column names filled in: 'X12' [12]

##
## -- Column specification --------------------------------------------------
------
## cols(
##   last_name = col_character(),
##   first_name = col_character(),
##   player_id = col_double(),
##   year = col_double(),
##   player_age = col_double(),
##   b_ab = col_double(),
##   b_total_hits = col_double(),
##   b_home_run = col_double(),
##   b_strikeout = col_double(),
##   b_walk = col_double(),
##   on_base_percent = col_double(),
##   X12 = col_logical()
## )
```

Data Cleaning

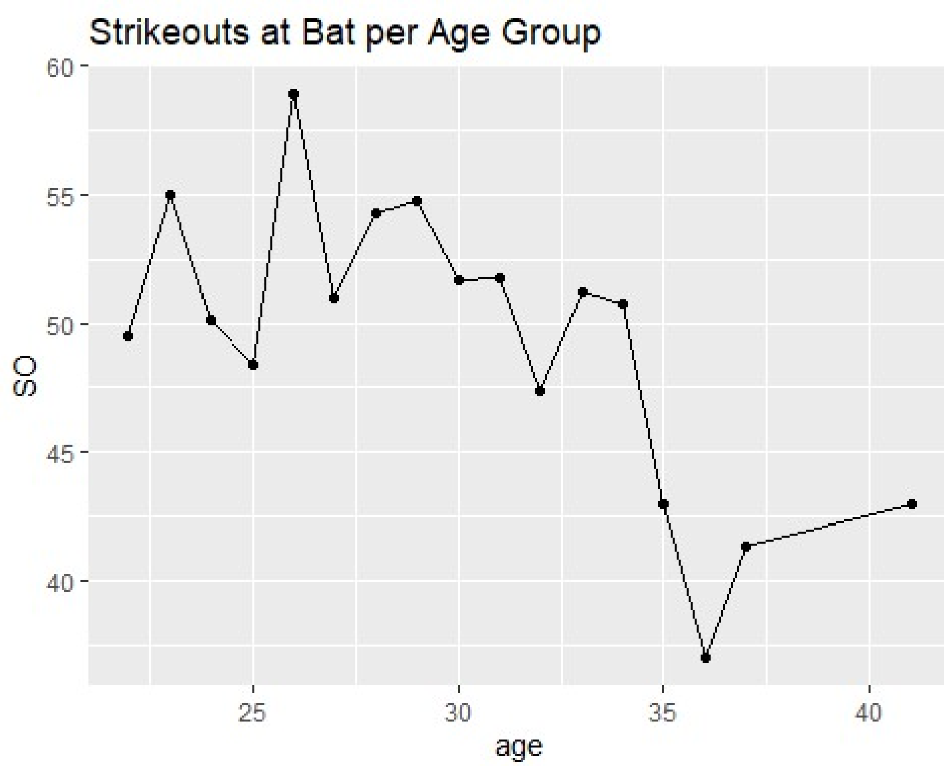
```
statslm <- stats %>% select("age" = player_age, "SO" = b_strikeout, "HR" =
b_home_run, "walk" = b_walk, "hits" = b_total_hits, "obp" = on_base_percent)



HRAge <- statslm %>%
  group_by(age) %>% summarize(HR = mean(HR))

SOAge <- statslm %>%
  group_by(age) %>% summarize(SO = mean(SO))

HitAge <-statslm %>%
  group_by(age) %>% summarize(HT = mean(hits))

OBPAge <-statslm %>%
  group_by(age) %>% summarize(OB = mean(obp))
```

Plots

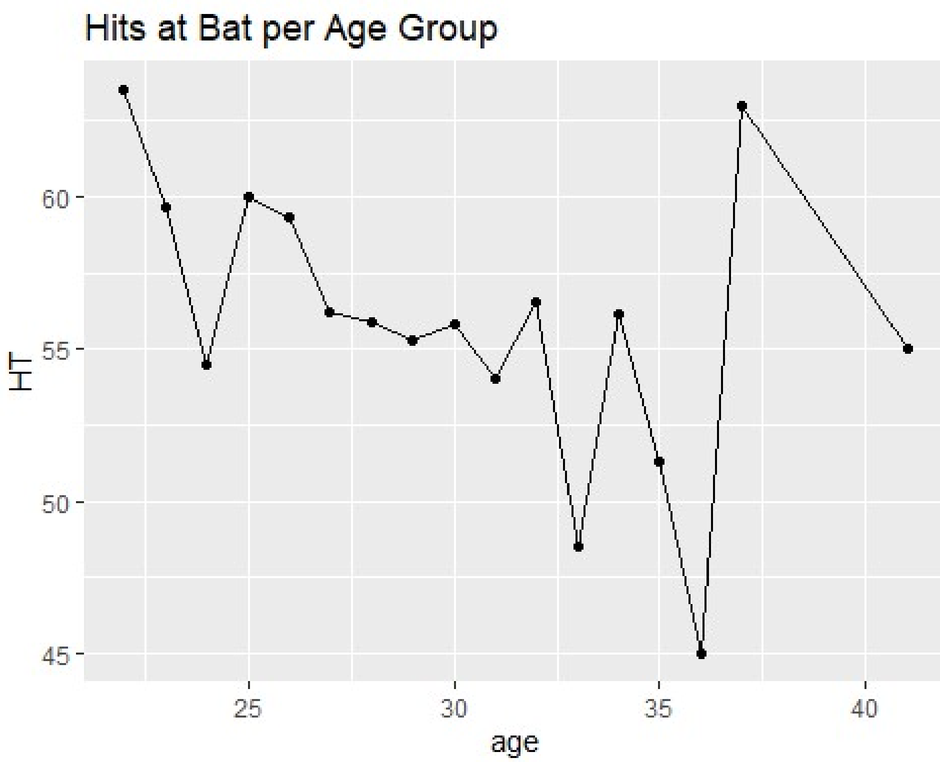
```
HRAge %>%
ggplot(aes(x=age, y=HR)) +geom_line()+geom_point()+ ggtitle("Homeruns at Bat
per Age Group")
```



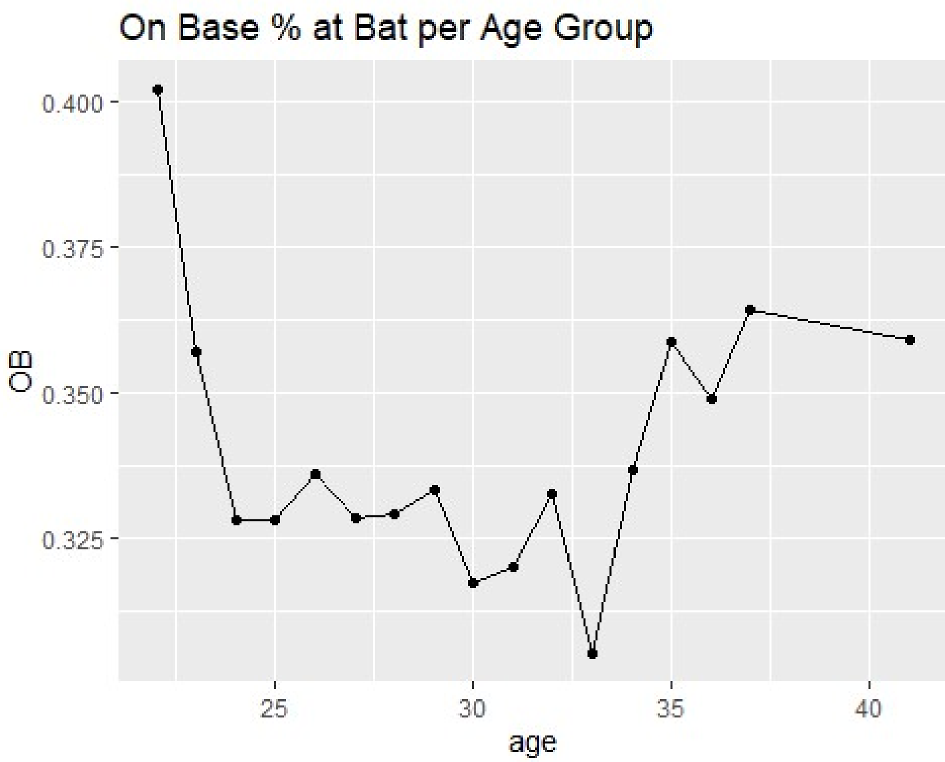

```
SOAge %>%
ggplot(aes(x=age, y=SO)) +geom_line()+geom_point()+ ggtitle("Strikeouts at
Bat per Age Group")
```



```
HitAge %>%
ggplot(aes(x=age, y=HT)) +geom_line()+geom_point()+ ggtitle("Hits at Bat per
Age Group")
```

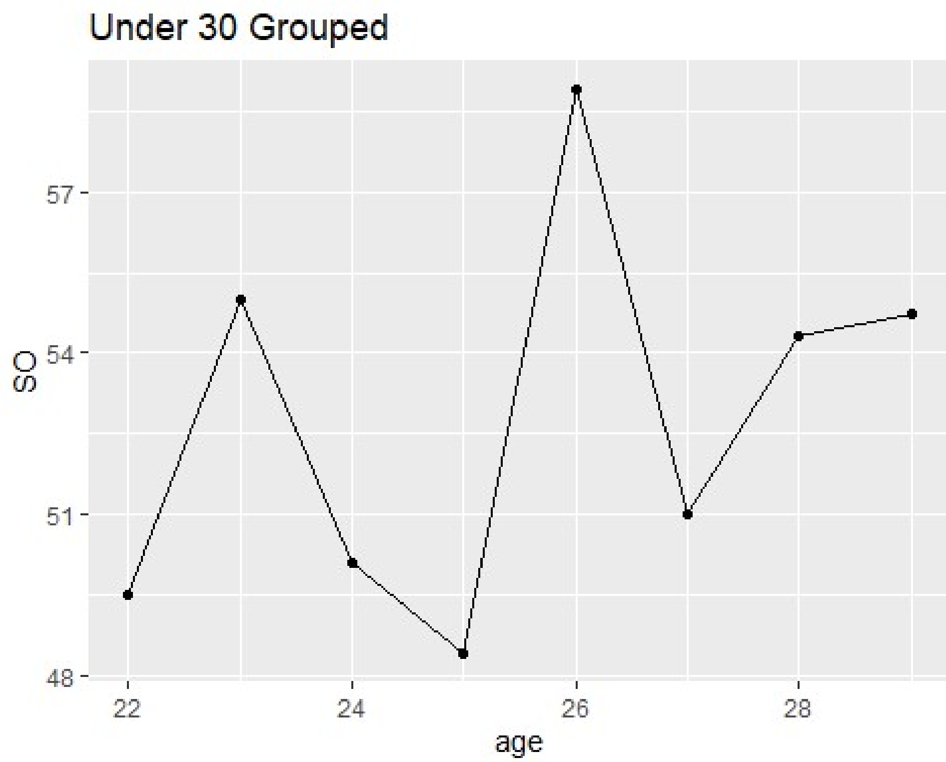



```
OBPAge %>%
ggplot(aes(x=age, y=OB)) +geom_line()+geom_point()+ ggtitle("On Base % at Bat
per Age Group")
```

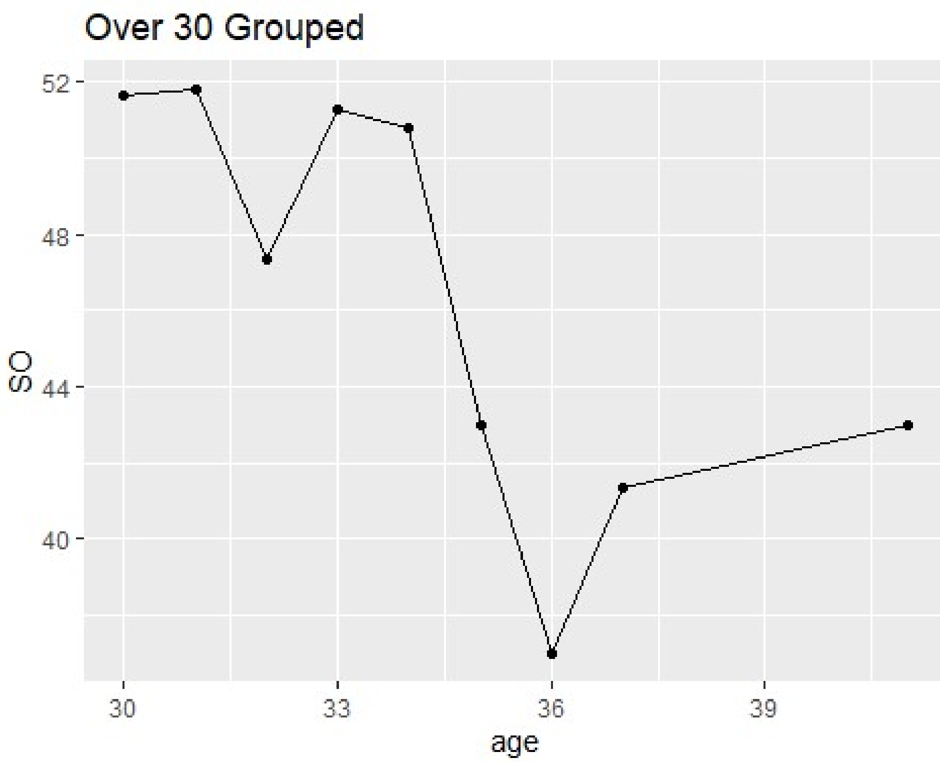


Exploration

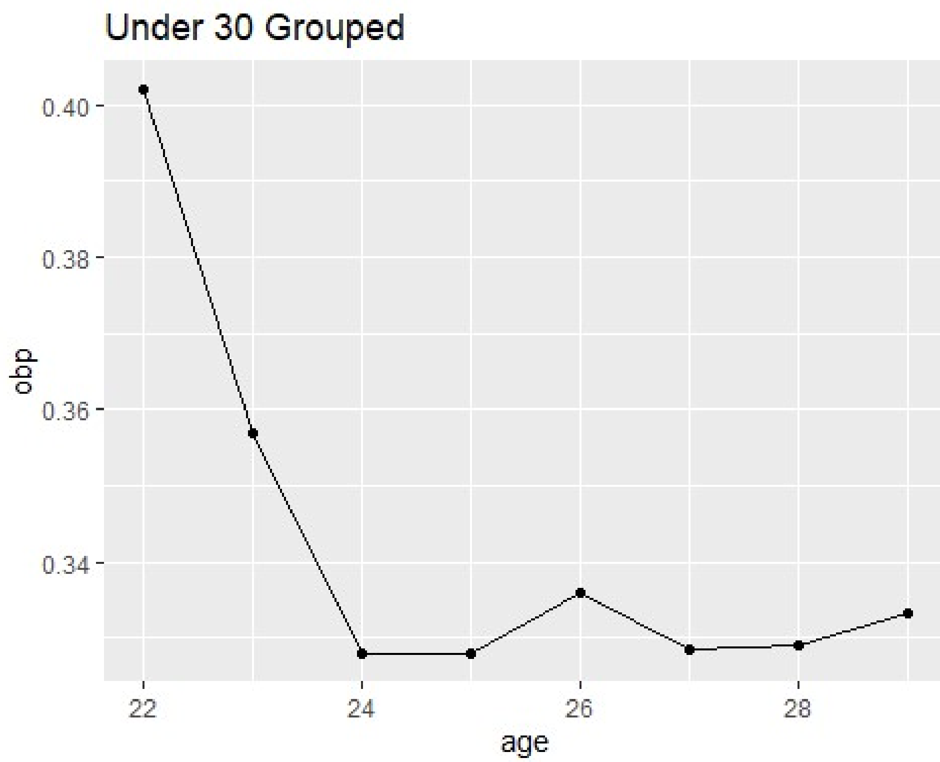
```
Under30 <- statslm%>% filter( between(age, 20, 29) )## Filter for age 20-29

Over30 <- statslm%>% filter( between(age, 30, 45) )## Filter for age 30-44


PlayerAge2 <- Under30 %>%
  group_by(age) %>% summarize(SO = mean(SO)) ## Mean for players under 30
with under 30
PlayerAge3 <- Over30 %>%
  group_by(age) %>% summarize(SO = mean(SO))  ## Mean for players under 30
with over 30

PlayerAge2OB <- Under30 %>%
  group_by(age) %>% summarize(obp = mean(obp)) ## Mean for players under 30
with under 30
PlayerAge3OB <- Over30 %>%
  group_by(age) %>% summarize(obp = mean(obp))  ## Mean for players under 30
with over 30

PlayerAge2 %>%
ggplot(aes(x=age, y=SO)) +geom_line()+geom_point()+ ggtitle("Under 30
Grouped")
```




```
## Mean age for under 30 for strike outs

PlayerAge3 %>%
ggplot(aes(x=age, y=SO)) +geom_line()+geom_point()+ ggtitle("Over 30 
Grouped")
```




```r
## Mean age for under 30 for strike outs

PlayerAge2OB %>%
ggplot(aes(x=age, y=obp)) +geom_line()+geom_point()+ ggtitle("Under 30
Grouped")
```




```
## Mean age for under 30 for On base Percentage

PlayerAge30B %>%
ggplot(aes(x=age, y=obp)) +geom_line()+geom_point()+ ggtitle("Over 30
Grouped")
```


```
## Mean age for under 30 for On base Percentage



resU <- cor.test(Under30$age, Under30$SO, method = "pearson")
resU

##
##  Pearson's product-moment correlation
##
## data:  Under30$age and Under30$SO
## t = 0.6519, df = 83, p-value = 0.5163
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.143941  0.280234
## sample estimates:
##        cor
## 0.07137264

resO <- cor.test(Over30$age, Over30$SO, method = "pearson")
resO

##
##  Pearson's product-moment correlation
##
## data:  Over30$age and Over30$SO
## t = -1.5359, df = 55, p-value = 0.1303
```

```
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.44010940  0.06099236
## sample estimates:
##        cor
## -0.2027983
```

```r
resUOB <- cor.test(Under30$age, Under30$obp, method = "pearson")
resUOB
```

```
##
##  Pearson's product-moment correlation
##
## data:  Under30$age and Under30$obp
## t = -1.2375, df = 83, p-value = 0.2194
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.33802777  0.08084262
## sample estimates:
##        cor
## -0.1346006
```

```r
resOOB <- cor.test(Over30$age, Over30$obp, method = "pearson")
resOOB
```

```
##
##  Pearson's product-moment correlation
##
## data:  Over30$age and Over30$obp
## t = 2.3856, df = 55, p-value = 0.02052
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.04961345 0.52490684
## sample estimates:
##       cor
## 0.3062224
```

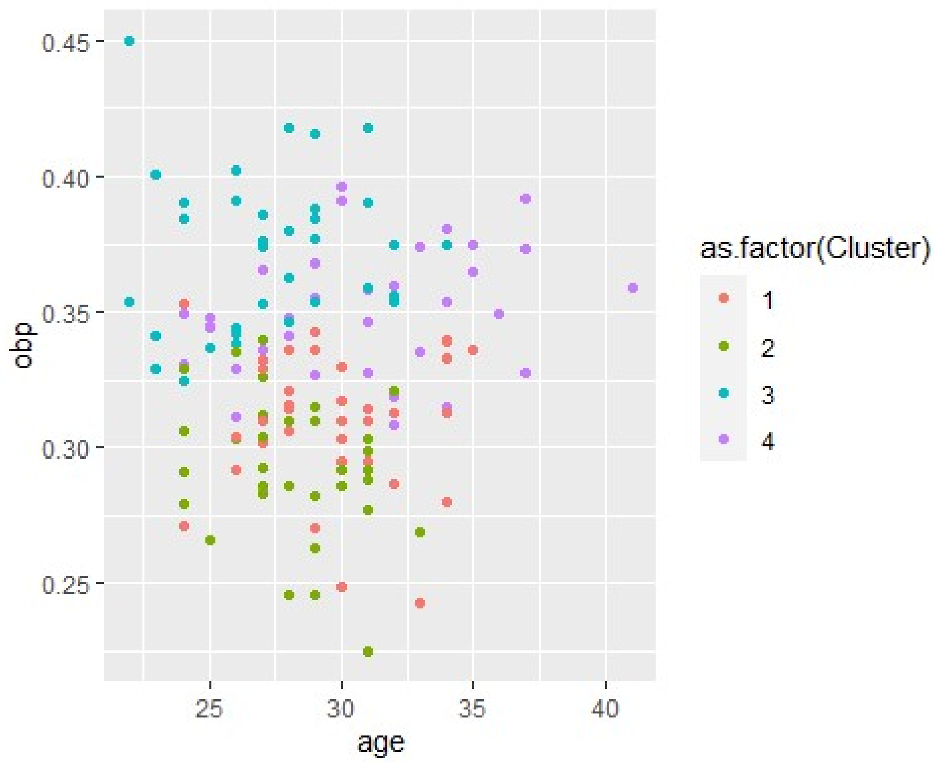
```r
## Correlation test for the age group and strikeouts
```

Cluster Analysis

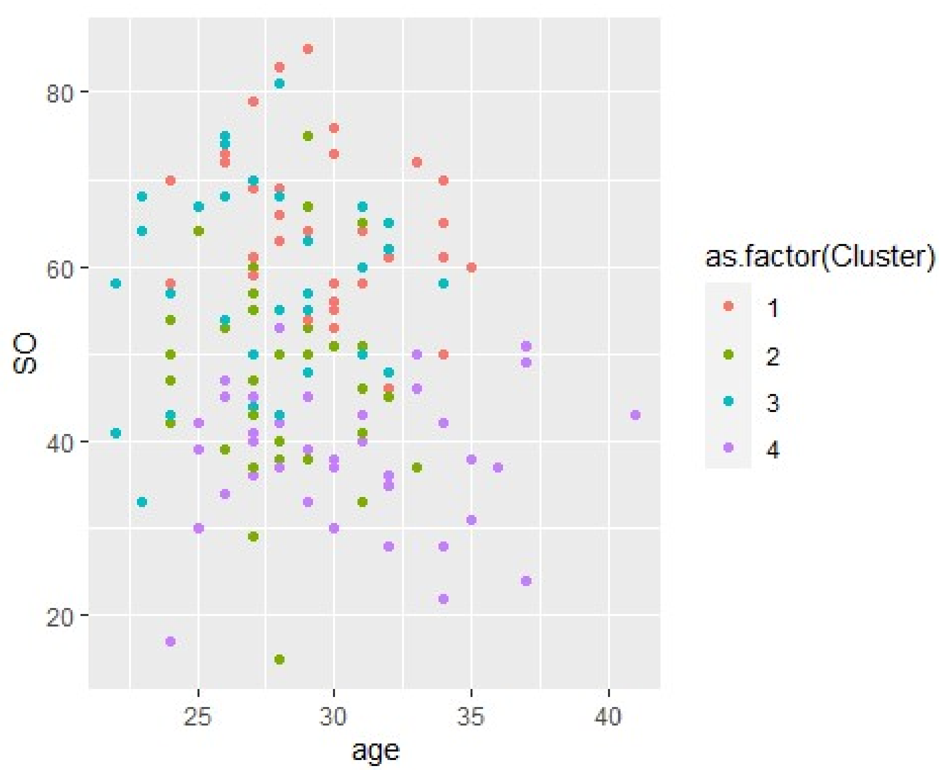
```r
X <- statslm[,1:6]
str(X)
```

```
## tibble [142 x 6] (S3: tbl_df/tbl/data.frame)
##  $ age : num [1:142] 41 34 35 35 34 37 33 35 37 34 ...
##  $ SO  : num [1:142] 43 42 31 60 65 49 37 38 51 22 ...
##  $ HR  : num [1:142] 13 4 4 10 14 12 0 11 5 3 ...
##  $ walk: num [1:142] 22 19 25 35 28 32 15 44 25 12 ...
##  $ hits: num [1:142] 55 56 56 46 46 63 45 52 55 63 ...
##  $ obp : num [1:142] 0.359 0.315 0.375 0.336 0.333 0.373 0.269 0.365 0.328
0.381 ...
```

```r
scale_X <- scale(X)

kmeans_X <- kmeans(scale_X,4) # Kmeans function takes two arguements -
dataset, number of clusters
kmeans_X

## K-means clustering with 4 clusters of sizes 36, 33, 34, 39
##
## Cluster means:
##          age         SO         HR       walk       hits        obp
## 1  0.2002485  0.8640545  0.5162648 -0.3751755 -0.2949670 -0.5508535
## 2 -0.2194364 -0.3133682 -0.9580055 -0.6303485 -0.9279897 -0.9486608
## 3 -0.4751237  0.4598528  0.8625580  0.8773223  0.7072528  1.0318646
## 4  0.4150425 -0.9333283 -0.4179057  0.1148426  0.4409199  0.4116188
##
## Clustering vector:
##   [1] 4 4 4 1 1 4 2 4 4 4 4 1 4 3 1 4 1 2 4 3 4 4 1 1 1 2 4 1 3 4 4 3 2 2
1 1 1
##  [38] 1 1 4 1 2 3 1 1 1 2 1 2 3 3 3 2 3 4 2 2 4 3 4 1 2 1 2 2 2 2 1 4 4 1
4 3 4
##  [75] 4 3 3 1 2 3 3 4 3 2 4 2 4 3 4 4 2 3 1 2 2 1 4 4 4 3 2 2 4 2 4 1 1 3
1 4 4
## [112] 1 3 3 3 3 4 1 3 3 2 2 1 2 2 2 3 3 3 3 2 3 3 1 2 3 3 1 2 4 1 1
##
## Within cluster sum of squares by cluster:
## [1] 107.90274  88.76898 146.05541 158.14270
##  (between_SS / total_SS =  40.8 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
"tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"

# Plot

X$Cluster <-kmeans_X$cluster # Creating a new column in dataset X with
cluster information obtained through kmeans clustering



ggplot(X,aes(x=age,y=obp,color=as.factor(Cluster))) +geom_point()
```




```
ggplot(X,aes(x=age,y=SO,color=as.factor(Cluster))) +geom_point()
```



```
X30 <- Over30[,1:6]
str(X30)
```

```
## tibble [57 x 6] (S3: tbl_df/tbl/data.frame)
##  $ age : num [1:57] 41 34 35 35 34 37 33 35 37 34 ...
##  $ SO  : num [1:57] 43 42 31 60 65 49 37 38 51 22 ...
##  $ HR  : num [1:57] 13 4 4 10 14 12 0 11 5 3 ...
##  $ walk: num [1:57] 22 19 25 35 28 32 15 44 25 12 ...
##  $ hits: num [1:57] 55 56 56 46 46 63 45 52 55 63 ...
##  $ obp : num [1:57] 0.359 0.315 0.375 0.336 0.333 0.373 0.269 0.365 0.328
0.381 ...
```

```
scale_X30 <- scale(X30)
kmeans_X <- kmeans(scale_X30,3) # Kmeans function takes two arguements -
dataset, number of clusters
kmeans_X
```

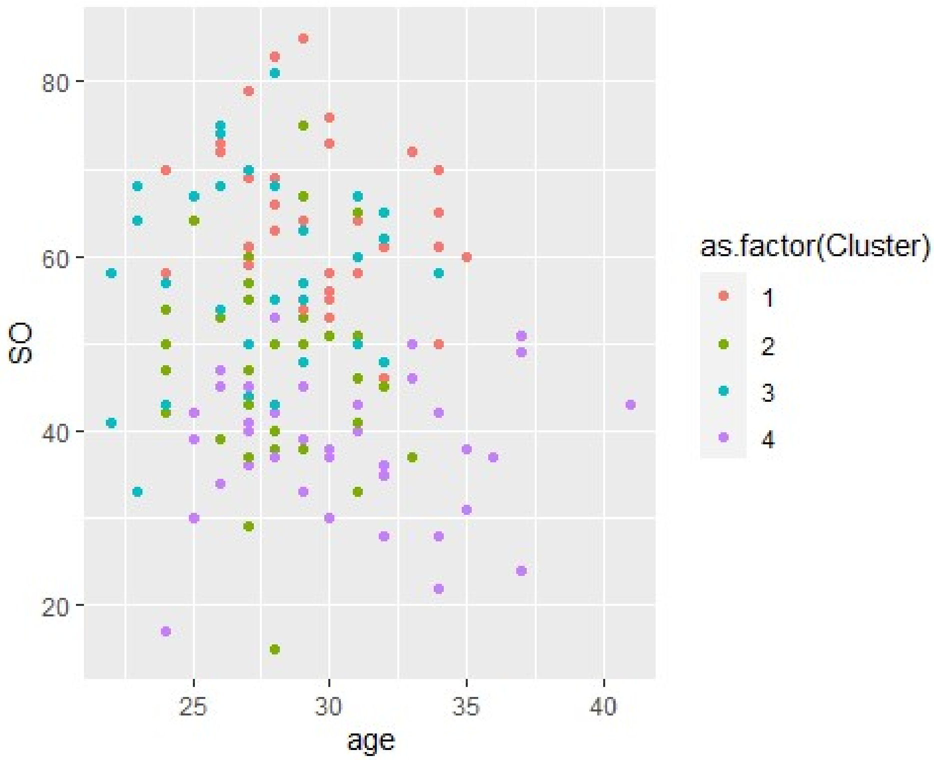
```
## K-means clustering with 3 clusters of sizes 20, 17, 20
##
## Cluster means:
##          age         SO          HR       walk       hits        obp
## 1 -0.3707343  0.8060072  0.77509183 -0.3995020  0.1148861 -0.4742776
## 2 -0.2553998 -0.5845969 -0.94346187 -0.7490731 -0.4609020 -0.5739591
## 3  0.5878241 -0.3090998  0.02685076  1.0362142  0.2768806  0.9621428
##
## Clustering vector:
##  [1] 3 2 3 3 1 3 2 3 3 2 3 1 3 3 1 3 1 2 3 3 3 3 2 1 1 2 1 1 3 2 2 1 2 2 1
1 1 1
## [39] 1 1 1 2 3 1 1 1 2 2 3 2 2 1 2 3 2 3 3
##
## Within cluster sum of squares by cluster:
## [1] 63.99327 45.40736 99.72770
##  (between_SS / total_SS =  37.8 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
"tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
summary(kmeans_X)
```

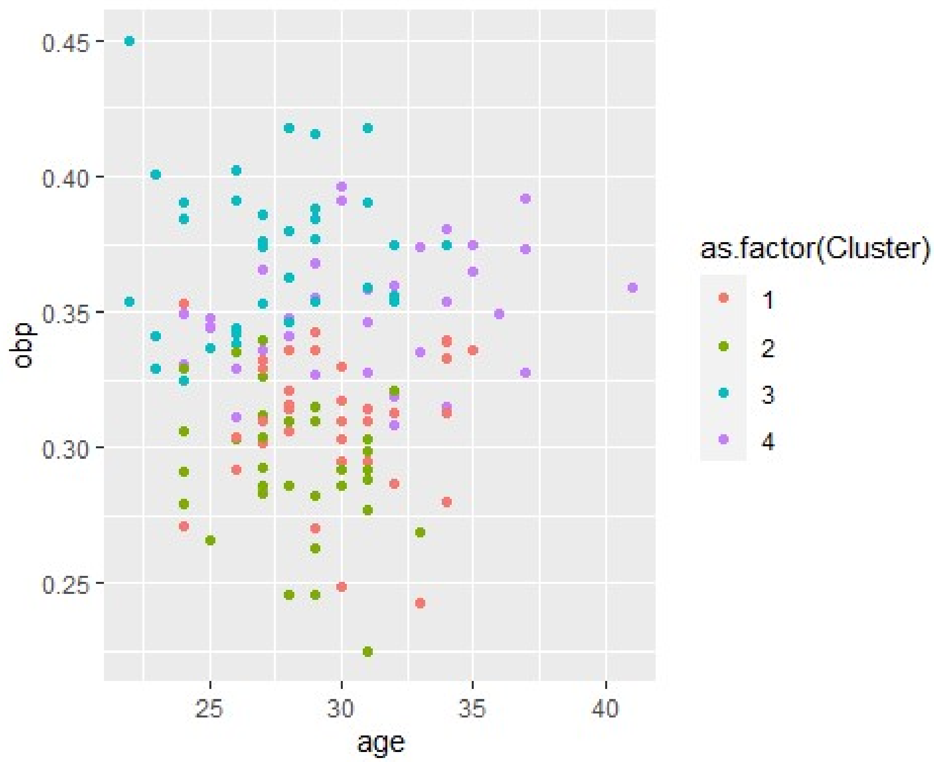
```
##              Length Class  Mode
## cluster      57     -none- numeric
## centers      18     -none- numeric
## totss         1     -none- numeric
## withinss      3     -none- numeric
## tot.withinss  1     -none- numeric
## betweenss     1     -none- numeric
## size          3     -none- numeric
```

```
## iter          1     -none- numeric
## ifault        1     -none- numeric
```

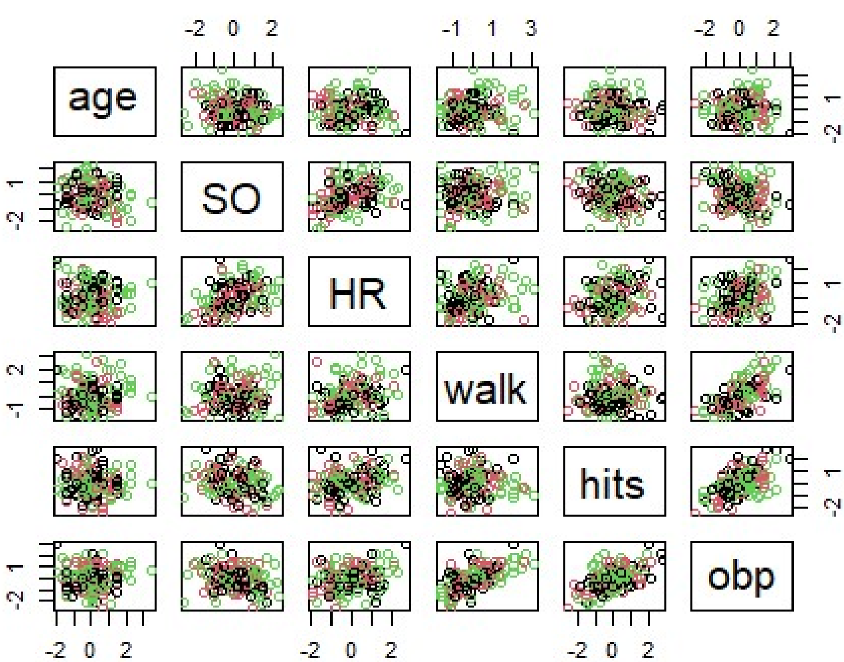
```
# Plot

X30$Cluster <-kmeans_X$cluster # Creating a new column in dataset X with
cluster information obtained through kmeans clustering


X30$Cluster <-kmeans_X$cluster


ggplot(X,aes(x=age,y=SO,color=as.factor(Cluster))) +geom_point()
```



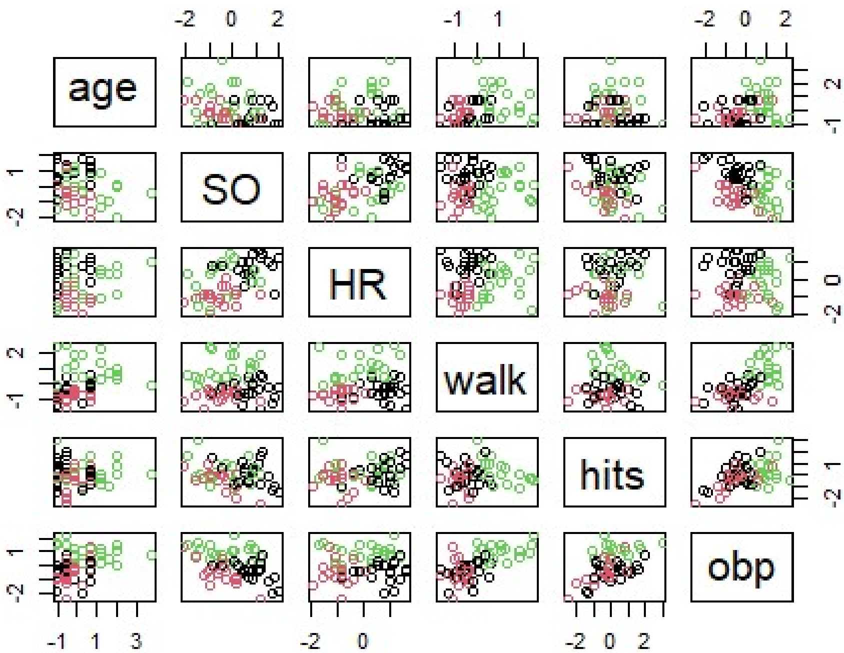

```
ggplot(X,aes(x=age,y=obp,color=as.factor(Cluster))) +geom_point()
```




```r
# Plot Clusters over all the varibles

pairs(scale_X,col=as.factor(kmeans_X$cluster))
```



```r
pairs(scale_X30,col=as.factor(kmeans_X$cluster))
```




```r
# Try yourself "ggparis" function from "ggally package for a better
visualization
```

Summary of Final

Describe the dataset and why you selected it for this project.

The data set I chose was baseball data from the MLB for the 2021 current season. The data contained player statistics for hitting and strikeouts.

Describe any processing problems you identified with the data and how you overcame those issues

THe data has a few issues. I had to clean up the data because some of the naming headers were very long. I also had a lot of data that was not relavent to my research to support my question.

Describe your 'Big Question' and why the data is a good choice to answer it.

My big question was "At which age can you expect a productive season?" This means that I was searching for the player age group that had a chance to hit, get on base, or score a homerun.

Describe the results of your exploratory analysis and what preliminary conclusions you were able to draw based on this analysis

I unfortunately did not have a great way to prove my question with the current methods I have learned from class. I was able to correlative reseach but that came up inconclusive. I also did a cluster analysis, but that also came up inconclusive. I was able to plot data, but unfortunately, it still did not represent strong evidence to my question. The question I was trying to answer was out of the scope of the methods I used to analyze the data.

Describe how you selected the methodology for your analysis of the big question and the pros and cons of that method and any alternative methods you considered

I used correlation, the pro was it was able to tell me there was no immediate relationships with the analysis I was trying to research, given the variables I chose to correlate. The con, it only correlated the values, but it was not representative overall of the truth of my data.

Cluster Analysis and KMeans, I use this method, but it unfortunately did not provide sufficient evidence ot proof to answer my question.

Describe your final conclusions based on your analysis and support them with analytics on your dataset

One conclusion that I was able to get out of the analysis was the age group was not the important factor, but the individual age's were a possible factor to MLB players being productive.

Describe any additional analyses that you would have liked to carry out and any additional data that would have been needed in order to extend your analysis

I think KMEANs would have been better if I had a bigger data set that represented multi year player information. This would then give me a year over year review of how the age groups did during thier season.