Station Casinos: Gaming Analytics

Harinarayan Parameswara Sarma

8/1/2021

This is an R Markdown document for Station Casino-PreModule Assignment.

Step1 - Load the necessary packages.

```
library(readr)
library(readxl)
library(openxlsx)
library(ggplot2)
library(priceR)
library(cowplot)
library(dplyr)
library(stats)
library(base)
library(RColorBrewer)
```

Step2 - Data Collection and Data preparation.

```
station_casino <- read_excel("Casino Final_Dec 2013 Raw Data_for test.xlsx", sheet = "Sheet1")
```

```
station_casino_df <- data.frame(station_casino)
```

```
View(station_casino)
station_casino_df
colnames(station_casino_df)
```

```
names(station_casino_df)[1] <- "Players"
names(station_casino_df)[8] <- "Other Games"
names(station_casino_df)[9] <- "Total Spend"
```

```
colnames(station_casino_df)

dim(station_casino_df)
sum(is.na(station_casino_df)) # Ans: 0 - There are no NA values in the data/ data frame.
```

Question1 - How much total dollars were played on each game?

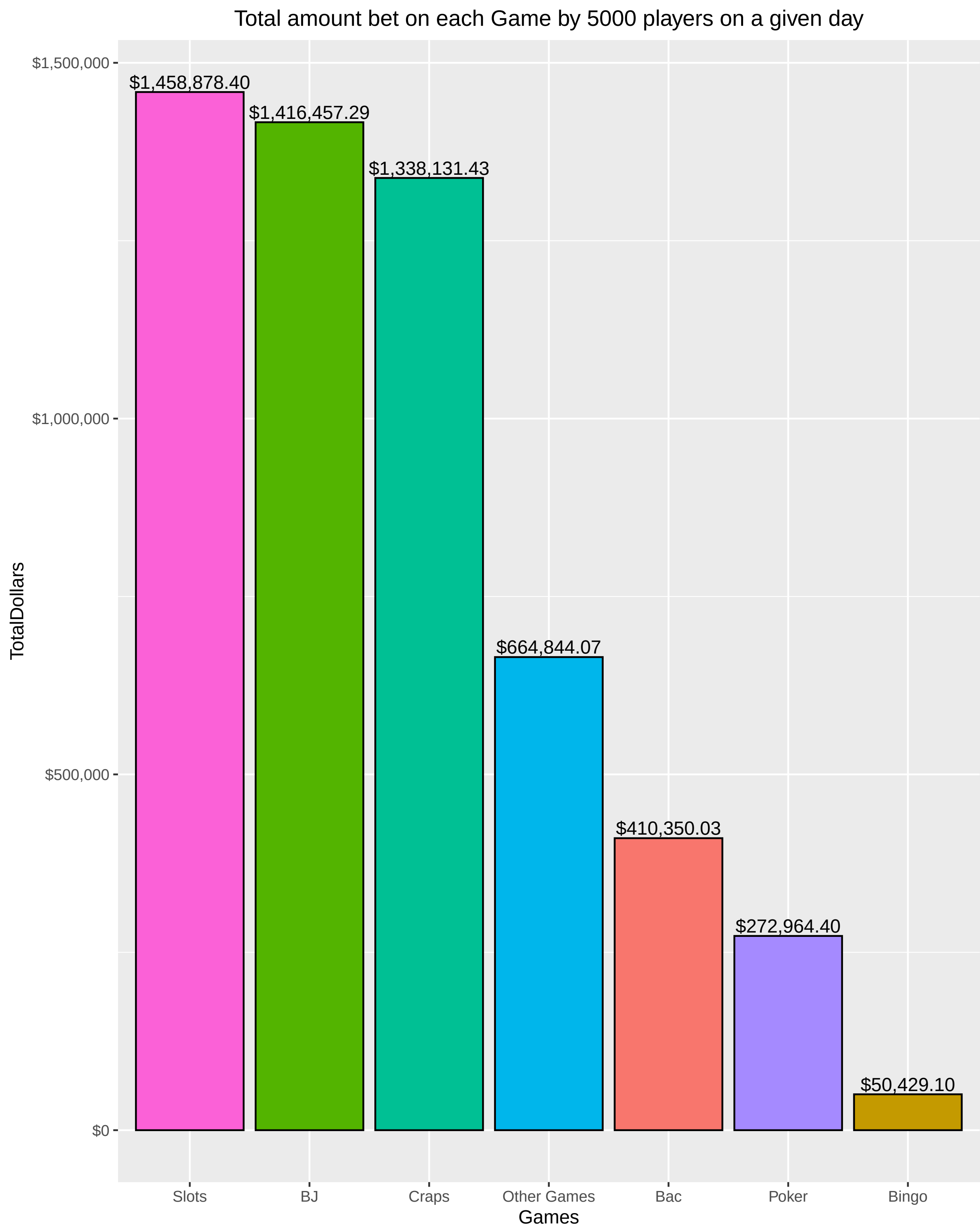
Present the answer to this question as a single bar chart with seven bars, one for each game (six games and 'Other Games' as the seventh), with the length of each bar representing the total amount bet on that game by all 5000 players on that specific day. Since these charts are difficult to read accurately, include the actual correct total dollar value at the top of each bar.

Subsetting the Games columns which has the numerical data/monetary columns

```r
station_casino_df_subset <- station_casino_df[,2:8]
station_casino_bar <- data.frame(TotalDollars=apply(station_casino_df_subset,2,sum))
station_casino_bar$Games <- rownames(station_casino_bar)
```

Here we need the Bar chart as the X-axis: Games is categorical.

```r
ggplot(data=station_casino_bar, aes(x=reorder(Games,-TotalDollars), y=TotalDollars, fill=Games)) +
  geom_bar(colour="black", stat="identity")+
  geom_text(label=format_dollars(station_casino_bar$TotalDollars,digits=2),vjust=-0.25)+
  xlab("Games")+
  theme(legend.position="none")+
  ggtitle("Total amount bet on each Game by 5000 players on a given day")+
  theme(plot.title = element_text(hjust = 0.5))+
  scale_y_continuous(labels=scales::dollar_format())
```



From the above barchart, the below percentage of revenue can be assessed

```
Revenue <- data.frame(station_casino_bar$Games,
             station_casino_bar$TotalDollars,
             round(station_casino_bar$TotalDollars/sum(station_casino_bar$TotalDollars)*100,2))
```

```
colnames(Revenue) <- c("Games","TotalRevenue ($)","PercentageRevenue (%)")
```

Percentage of revenue as shown below.

```
Revenue[order(-Revenue$PercentageRevenue),]
```

```
##         Games TotalRevenue ($) PercentageRevenue (%)
## 1       Slots        1458878.4                 26.00
## 2          BJ        1416457.3                 25.24
## 3       Craps        1338131.4                 23.84
## 7 Other Games         664844.1                 11.85
## 4         Bac         410350.0                  7.31
## 6       Poker         272964.4                  4.86
## 5       Bingo          50429.1                  0.90
```

Question2

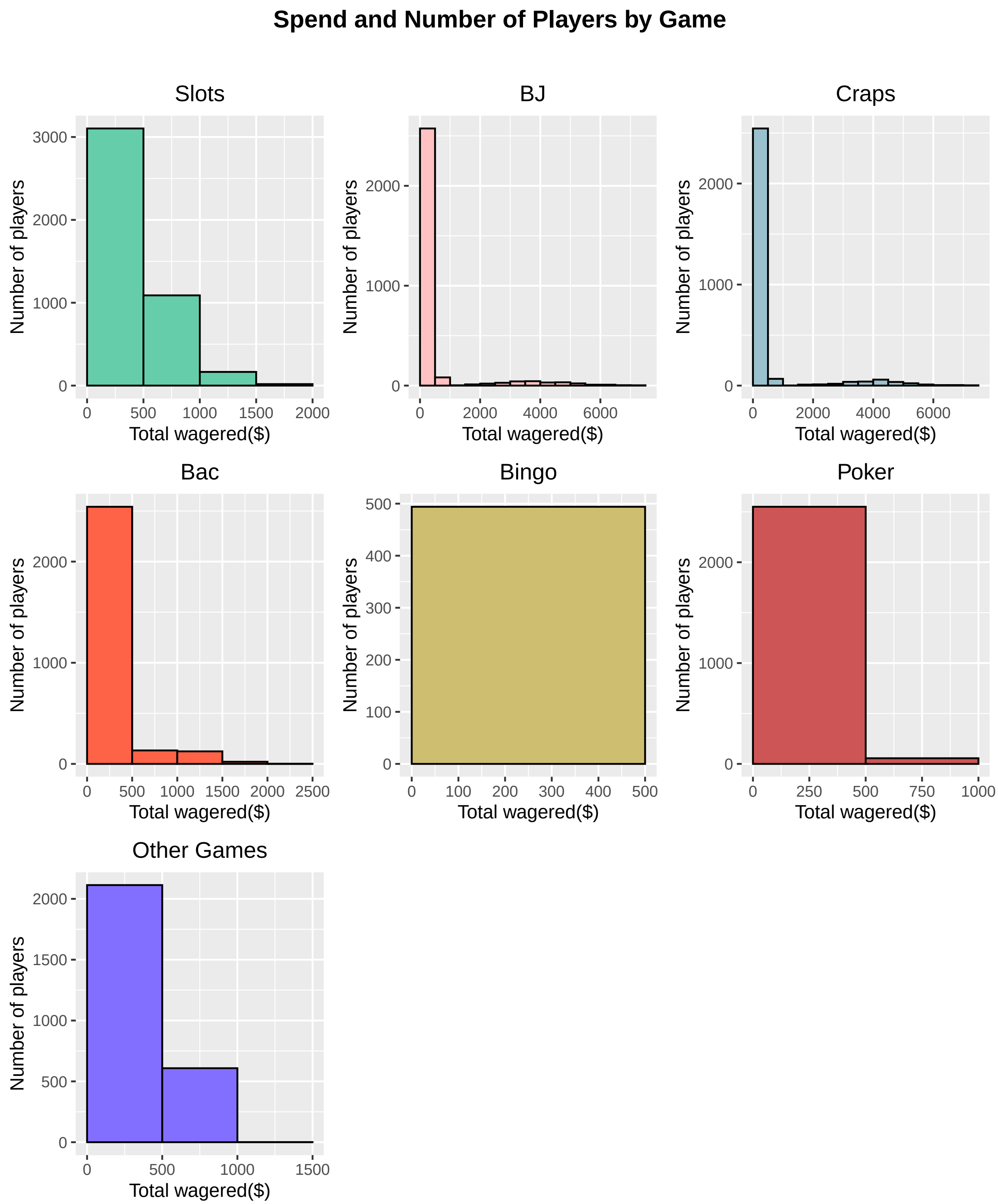
How many people played each of the seven games (six games and 'Other Games" as the seventh) and how much did they wager?

Create a bar chart for each game (so you will have seven bar charts). Each bar chart should show how many people played that game (the vertical axis) and how many dollars were wagered (use dollar range intervals on the horizontal axis). Plot all seven charts together on a single display (i.e. use the plotgrid function from the cowplot package).

Here we need the Histogram as the X-axis: Dollar Wagered is continuous. Also, filtered out each Game with Players with $0(Player wagered no money, hence removed those counts)

```
axes <- labs(x="Total wagered($)",y="Number of players")
Slots <- ggplot(data=station_casino_df%>%filter(Slots >0),aes(x=Slots))+
  #Slots <- ggplot(data=station_casino_df,aes(x=Slots>0))+
  geom_histogram(bins = 10,binwidth = 500,boundary=0,color="black",fill="aquamarine3")+
  axes+
  ggtitle("Slots")+
  theme(plot.title = element_text(hjust = 0.5))

BJ <- ggplot(data=station_casino_df%>%filter(BJ > 0),aes(x=BJ))+
  geom_histogram(bins = 10,binwidth = 500,boundary=0,color="black",fill="rosybrown1")+
  axes+
  ggtitle("BJ")+
  theme(plot.title = element_text(hjust = 0.5))

Craps <- ggplot(data=station_casino_df%>%filter(Craps > 0),aes(x=Craps))+
  geom_histogram(bins = 10,binwidth = 500,boundary=0,color="black",fill="lightblue3")+
  axes+
  ggtitle("Craps")+
  theme(plot.title = element_text(hjust = 0.5))

Bac <- ggplot(data=station_casino_df%>%filter(Bac > 0),aes(x=Bac))+
  geom_histogram(bins = 10,binwidth = 500,boundary=0,color="black",fill="tomato1")+
  axes+
  ggtitle("Bac")+
  theme(plot.title = element_text(hjust = 0.5))
```

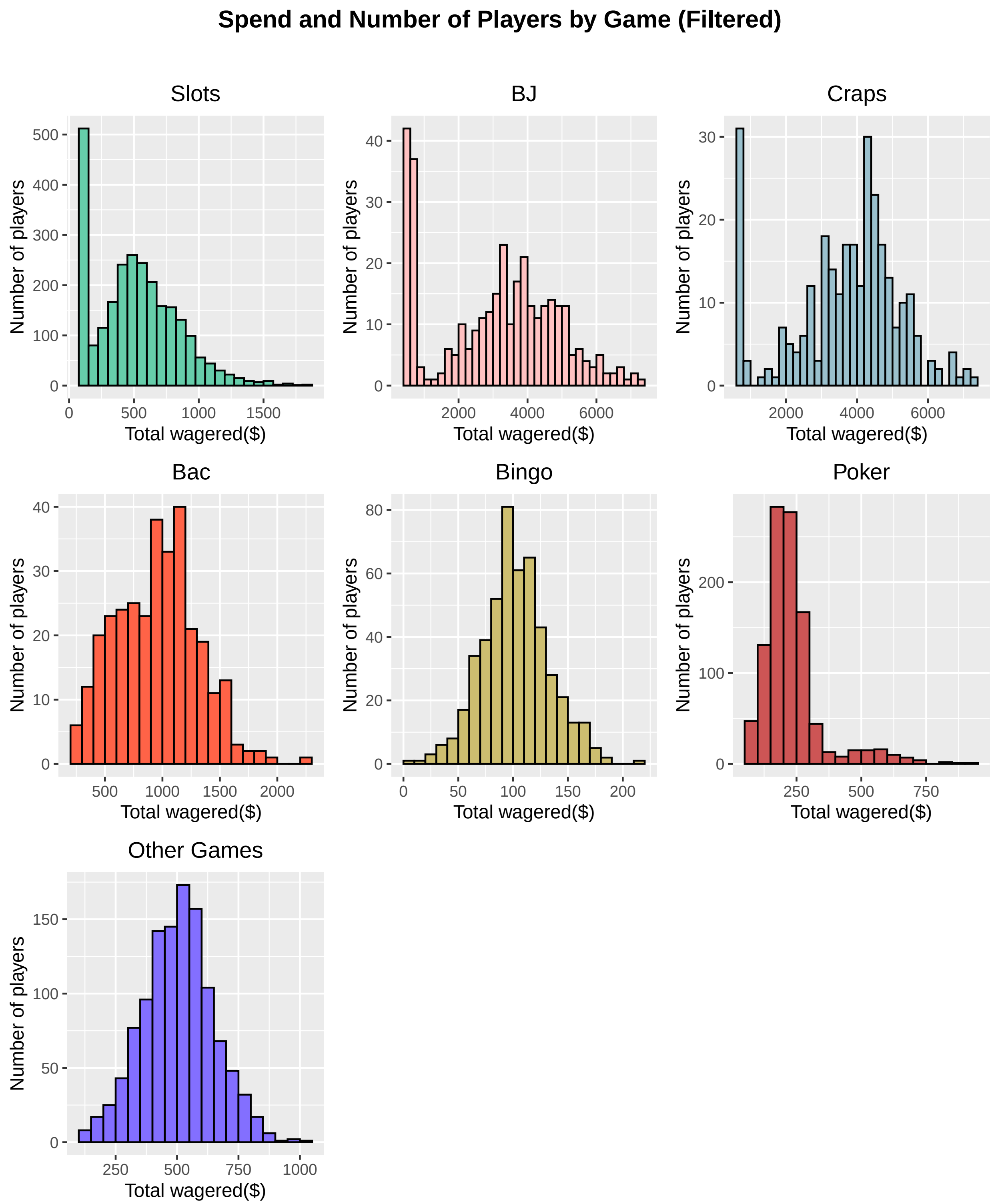
```r
Bingo <- ggplot(data=station_casino_df%>%filter(Bingo > 0),aes(x=Bingo))+
  geom_histogram(bins = 10,binwidth = 500,boundary=0,color="black",fill="lightgoldenrod3")+
  axes+
  ggtitle("Bingo")+
  theme(plot.title = element_text(hjust = 0.5))

Poker <- ggplot(data=station_casino_df%>%filter(Poker > 0),aes(x=Poker))+
  geom_histogram(bins = 10,binwidth = 500,boundary=0,color="black",fill="indianred3")+
  axes+
  ggtitle("Poker")+
  theme(plot.title = element_text(hjust = 0.5))

OtherGames <- ggplot(data=station_casino_df%>%filter(`Other Games` > 0),aes(x=`Other Games`))+
  geom_histogram(bins = 10,binwidth = 500,boundary=0,color="black",fill="slateblue1")+
  axes+
  ggtitle("Other Games")+
  theme(plot.title = element_text(hjust = 0.5))

title <- ggdraw() + draw_label("Spend and Number of Players by Game", fontface='bold')
plot_games <- plot_grid(Slots,BJ,Craps,Bac,Bingo,Poker,OtherGames)
plot_grid(title,plot_games,ncol = 1,
          # rel_heights values control vertical title margins
          rel_heights = c(0.1, 1))
```



Filtered the Plot for a meaningful insight.

```
axes <- labs(x="Total wagered($)",y="Number of players")
Slots <- ggplot(data=station_casino_df%>%filter(Slots >= 100),aes(x=Slots))+
```

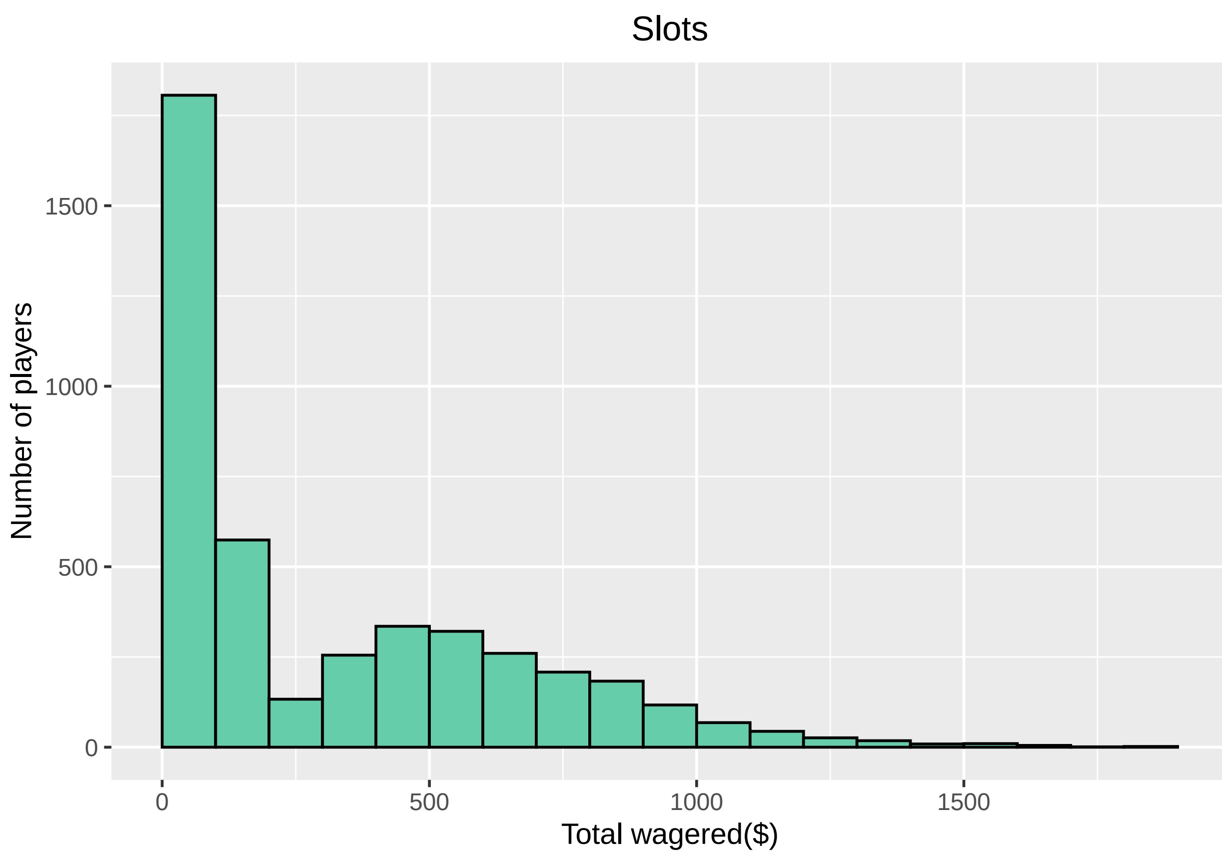
```r
  #Slots <- ggplot(data=station_casino_df,aes(x=Slots>0))+
  geom_histogram(bins = 10,binwidth = 75,boundary=0,color="black",fill="aquamarine3")+
  axes+
  ggtitle("Slots")+
  theme(plot.title = element_text(hjust = 0.5))

BJ <- ggplot(data=station_casino_df%>%filter(BJ > 500),aes(x=BJ))+
  geom_histogram(bins = 10,binwidth = 200,boundary=0,color="black",fill="rosybrown1")+
  axes+
  ggtitle("BJ")+
  theme(plot.title = element_text(hjust = 0.5))

Craps <- ggplot(data=station_casino_df%>%filter(Craps > 600),aes(x=Craps))+
  geom_histogram(bins = 25,binwidth = 200,boundary=0,color="black",fill="lightblue3")+
  axes+
  ggtitle("Craps")+
  theme(plot.title = element_text(hjust = 0.5))

Bac <- ggplot(data=station_casino_df%>%filter(Bac > 200),aes(x=Bac))+
  geom_histogram(bins = 25,binwidth = 100,boundary=0,color="black",fill="tomato1")+
  axes+
  ggtitle("Bac")+
  theme(plot.title = element_text(hjust = 0.5))

Bingo <- ggplot(data=station_casino_df%>%filter(Bingo > 0),aes(x=Bingo))+
  geom_histogram(bins = 25,binwidth = 10,boundary=0,color="black",fill="lightgoldenrod3")+
  axes+
  ggtitle("Bingo")+
  theme(plot.title = element_text(hjust = 0.5))

Poker <- ggplot(data=station_casino_df%>%filter(Poker > 50),aes(x=Poker))+
  geom_histogram(bins = 25,binwidth = 50,boundary=0,color="black",fill="indianred3")+
  axes+
  ggtitle("Poker")+
  theme(plot.title = element_text(hjust = 0.5))

OtherGames <- ggplot(data=station_casino_df%>%filter(`Other Games` > 100),aes(x=`Other Games`))+
  geom_histogram(bins = 25,binwidth = 50,boundary=0,color="black",fill="slateblue1")+
  axes+
  ggtitle("Other Games")+
  theme(plot.title = element_text(hjust = 0.5))

title <- ggdraw() + draw_label("Spend and Number of Players by Game (Filtered)", fontface='bold')
plot_games <- plot_grid(Slots,BJ,Craps,Bac,Bingo,Poker,OtherGames)
plot_grid(title,plot_games,ncol = 1,
          # rel_heights values control vertical title margins
          rel_heights = c(0.1, 1))
```



Interpretation of the Gaming charts.

Take 'Slots' bar chart as an example, and re-distribute the frequency. A similar interpretation can be done for all other Game charts.

```
ggplot(data=station_casino_df%>%filter(Slots > 0),aes(x=Slots))+
  geom_histogram(bins = 10,binwidth = 100,boundary=0,color="black",fill="aquamarine3")+
  axes+
  ggtitle("Slots")+
  theme(plot.title = element_text(hjust = 0.5))
```



Some of the below observations can be inferred for the management.

Observation1: \$0-\$500 - spent by adding up the bars - 1800+600+125+250+300 ~ 3075 players

Observation2: \$500-\$1000 - spent by adding up the bars - 275+250+225+200+150 ~ 1100 players

Question3

Management wants to offer comps to their players based on which games they play, how much they wager and their value to the casino. Clearly the 5000 players have strong differences in terms of what games they play and the amount of their bankrolls. (i.e. a player who wagers \$5000-\$10000 per day should be offered much more than a 25-cent slot machine player who wagers \$40 per day.)

Use a K-Means Cluster Analysis to create groupings among the players that will help management to determine what type (and value) of comps should be offered to players in each grouping. Keep the number of groups in the 3-6 range (K=3 to 6). Show the K=6 results as a Scatterplot Matrix showing every combination of two games (i.e. use the pairs function from the cars package).

Data Exploration and Measures of Central Tendency.

```
str(station_casino_df)
```

Subsetting the Games features

```r
station_casino_games <- station_casino_df[2:8]
```

As the summary function denotes some features come to dominate solely because they have a larger range of values than the others. The Summary function confirms again that there no NA values. So, no Imputation required

```r
summary(station_casino_games$Slots)
summary(station_casino_games$BJ)
summary(station_casino_games$Craps)
summary(station_casino_games$Bac)
summary(station_casino_games$Bingo)
summary(station_casino_games$Poker)
summary(station_casino_games$`Other Games`)
```

Some features come to dominate solely because they have a larger range of values than the others. The process of z-score standardization rescales features such that they have a mean of zero and a standard deviation of one. This transformation changes the interpretation of the data in a way that may be useful here.

To apply z-score standardization to the station_casino_games data frame, we can use the scale() function with lapply(). Since lapply() returns a matrix, it must be coerced back to data frame form using the as.data.frame() function, as follows.

```r
station_casino_games_z <- as.data.frame(lapply(station_casino_games, scale))
names(station_casino_games_z)[7] <- "Other Games"
```

To confirm that the transformation worked correctly, we can compare the summary statistics.

```r
summary(station_casino_games$Slots)
summary(station_casino_games_z$Slots)
```

Because the k-means algorithm utilizes random starting points, we use the set.seed() function for any later point analysis.

```r
RNGversion("3.5.2")
set.seed(2345)
# Set k=3
station_casino_clusters3 <- kmeans(station_casino_games_z, 3)
station_casino_clusters3$size
```

```
## [1] 4258  479  263
```

```r
station_casino_clusters3$centers
```

```
##        Slots         BJ      Craps        Bac      Bingo       Poker
## 1 -0.1675407 -0.2087613 -0.2168432 -0.2023394 -0.3122847  0.08987179
## 2  0.3124436 -0.3150581 -0.2865742 -0.3320875  2.9493842 -0.51573328
## 3  2.1434518  3.9536824  4.0326522  3.8807263 -0.3157667 -0.51573328
##   Other Games
## 1   0.1065286
## 2  -0.6113191
## 3  -0.6113191
```

```r
# Set k=4
station_casino_clusters4 <- kmeans(station_casino_games_z, 4)
station_casino_clusters4$size
```

```
## [1]  263 1040 3220  477
```

```r
station_casino_clusters4$centers
```

```
##        Slots          BJ       Craps         Bac      Bingo      Poker
## 1  2.1434518  3.95368238  4.03265217  3.88072633 -0.3157667 -0.5157333
## 2  1.0054805 -0.01013876 -0.09067313  0.02631186 -0.3157667  1.6527692
## 3 -0.5462426 -0.27297871 -0.25763713 -0.27627007 -0.3102252 -0.4152911
## 4  0.3133618 -0.31505810 -0.28657423 -0.33208753  2.9567489 -0.5157333
##   Other Games
## 1  -0.6113191
## 2   1.5225688
## 3  -0.3512719
## 4  -0.6113191
```

```r
# Set k=5
station_casino_clusters5 <- kmeans(station_casino_games_z, 5)
station_casino_clusters5$size
```

```
## [1]  263   99 3201  477  960
```

```r
station_casino_clusters5$centers
```

```
##        Slots          BJ       Craps         Bac      Bingo      Poker
## 1  2.1434518  3.95368238  4.03265217  3.88072633 -0.3157667 -0.5157333
## 2 -0.8926563 -0.31505810 -0.28657423 -0.33208753 -0.3157667  4.3996100
## 3 -0.5435643 -0.27261721 -0.25741024 -0.27581028 -0.3101923 -0.4205621
## 4  0.3133618 -0.31505810 -0.28657423 -0.33208753  2.9567489 -0.5157333
## 5  1.1615845  0.01489864 -0.07453187  0.05575009 -0.3157667  1.3461464
##   Other Games
## 1  -0.6113191
## 2  -0.6113191
## 3  -0.3630478
## 4  -0.6113191
## 5   1.7448049
```

```r
#Finally, k=6 for the detailed analysis
station_casino_clusters <- kmeans(station_casino_games_z, 6)
```

The goal of this analysis was to identify clusters of players with similar interests for compensation purposes. we will largely measure our success in qualitative terms.

The number of players in each cluster

```r
station_casino_clusters$size
```

```
## [1] 3010  219  932  263   99  477
```

Smallest cluster is 99 with 2% and largest is 3010 with 60%. To confirm confirm the percentage gap is real or random, we will look into cluster's homogeneity For a more in-depth look at the clusters, we can examine the coordinates of the cluster centroids.

```
station_casino_clusters$centers
```

```
##        Slots          BJ       Craps         Bac      Bingo      Poker
## 1 -0.5220901 -0.27012797 -0.25569514 -0.27253814 -0.3098386 -0.4150741
## 2 -0.8606686 -0.30146691 -0.27897349 -0.31721126 -0.3157667 -0.4857651
## 3  1.2179712  0.02227622 -0.06950988  0.06487173 -0.3157667  1.3968205
## 4  2.1434518  3.95368238  4.03265217  3.88072633 -0.3157667 -0.5157333
## 5 -0.8926563 -0.31505810 -0.28657423 -0.33208753 -0.3157667  4.3996100
## 6  0.3133618 -0.31505810 -0.28657423 -0.33208753  2.9567489 -0.5157333
##   Other Games
## 1  -0.4882678
## 2   1.7498181
## 3   1.7160652
## 4  -0.6113191
## 5  -0.6113191
## 6  -0.6113191
```

The rows of the output (labeled 1 to 6) refer to the six clusters, while the numbers across each row indicate the cluster's average value for the dollar wager listed at the top of the column. Because the values are z-score standardized, positive values are above the overall mean level and negative values are below the overall mean.

For example, the fourth row has the highest value in the Bac column, which means that cluster 4 has the highest average dollar wager in Bac among all the players.

Cluster 4 is substantially above the mean level on Slots, BJ, Craps and Bac. Cluster 1 players have lower-than average levels of dollar wager. One potential explanation is that these players will come to Casino but do not wager much.

Cluster1 = less revenue cluster Cluster2 = Other Games Cluster3 = Slots, BJ, Bac, Poker, Other Games Cluster4 = Slots, BJ, Craps, Bac Cluster5 = Poker Cluster6 = Slots, Bingo

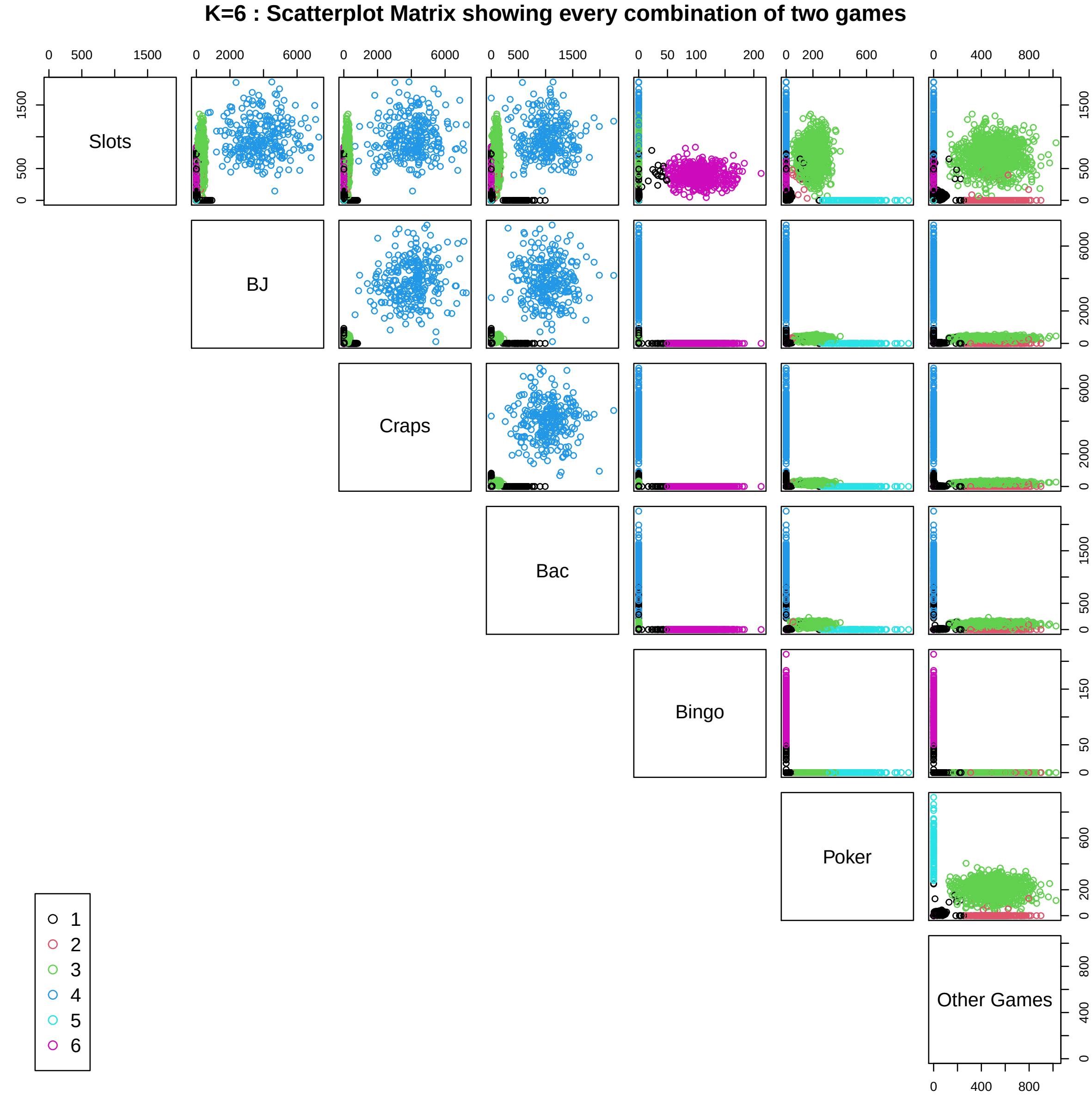
Given the above high-level inference, the Station Casino management would have a clear depiction of six types of players. Based on these profiles, the executive could provide appropriate comps.

We will begin by applying the clusters back onto the full dataset - The cluster assignments for all 5000 players

```
station_casino_df$cluster <- station_casino_clusters$cluster
```

Show the K=6 results as a Scatterplot Matrix showing every combination of two games (i.e. use the pairs function from the cars package)

```
pairs(station_casino_df[2:8],col=station_casino_df$cluster, lower.panel = NULL,
      main=" K=6 : Scatterplot Matrix showing every combination of two games")
par(xpd = TRUE)
legend("bottomleft", legend = sort(unique(station_casino_df$cluster)), pch=1,
col = sort(unique(station_casino_df$cluster)))
```



Inference:

Deriving the insights from the clusters

```
station_casino_Interpret <- data.frame(aggregate(data = station_casino_df, Slots ~ cluster, mean),
aggregate(data = station_casino_df, BJ ~ cluster, mean),
aggregate(data = station_casino_df, Craps ~ cluster, mean),
aggregate(data = station_casino_df, Bac ~ cluster, mean),
aggregate(data = station_casino_df, Bingo ~ cluster, mean),
aggregate(data = station_casino_df, Poker ~ cluster, mean),
aggregate(data = station_casino_df, `Other Games` ~ cluster, mean),
aggregate(data = station_casino_df, `Total Spend` ~ cluster, mean))

station_casino_Interpret <- select(station_casino_Interpret,-cluster.1, -cluster.2, -cluster.3, -cluste
```

```r
station_casino_Interpret
```

```
##   cluster     Slots         BJ       Craps         Bac       Bingo      Poker
## 1       1 121.12412   40.39992   28.837402   14.716658   0.1893486  10.655269
## 2       2  10.45559   12.22082    7.098185    3.676425   0.0000000   3.172277
## 3       3 689.88454  303.32162  202.712318   98.101993   0.0000000 202.453132
## 4       4 992.38922 3838.33239 4033.644353 1041.128117   0.0000000   0.000000
## 5       5   0.00000    0.00000    0.000000    0.000000   0.0000000 520.313015
## 6       6 394.20182    0.00000    0.000000    0.000000 104.5265448   0.000000
##   Other.Games Total.Spend
## 1    26.76505    242.6878
## 2   513.57400    550.1973
## 3   506.23237   2002.7059
## 4     0.00000   9905.4941
## 5     0.00000    520.3130
## 6     0.00000    498.7284
```

Conclusions and Recommendations

The objective of the Station Casino management team is to offer comps to their players based on which games they play, how much they wager and their value to the casino by identifying and matching each player to their specific group, defined by the games they play and how much they wager. These groupings should aim to differentiate the type and value of the comps that are offered to the players. This report has deployed segmentation analysis, where the high-value players were identified using the k -means cluster analysis alogorithm.

The main insights of the clustering are:

- Cluster 4 is the group which represents the players whose wagering budget could be around $10,000 per day. The characteristic of this group is they tend to play Slots, BJ, Craps and Bac which are the games where the management can control. These players do not wager much on Bingo, Poker and Other Games.

- Cluster 3 is another important group as it is formed by players who wager $2000 in a day. These players appears to bet majorly on Slots, BJ, Bac, Poker, Other Games.

- Cluster 5 players concentration on Poker and having a daily wager of $520 per day.

- Cluster 6 players are wagered on Slots and Bingo and having a total spent of $500 per day

- Cluster 2 is segmented mainly on Other Games, however having a wagering of $500 per day

Note: Clusters 2, 5, 6 gambling budget is not significantly high, but it is not considered low either.

- Clusters 1 is characterized as the group of players that have the lowest average spend in all the games with average amounts less than $250 in a day.

As per these insights, players in cluster 4 should be offered much more comps than players in cluster1. Generally, players in clusters 4 and 3 should be provided with high-value comps to reward their loyalty and retain them as clients.

Players in 2, 5, 6 can be offered some comps to retain them and increase the wager.

Players in 1 should not be offered high-value comps as their wager is significantly low.

In order to obtain a greater overview and insights to decide the value of the comps, it is necessary to use training data that includes information for a longer time period.