Analysis for Top Soccer Player’s Wage

615 Final Project for Samuel Luo

Background

In the current world, many famous soccer players are fully appreciated by soccer clubs and fans. Since they are in top-level clubs, they are paid an attractive salary that many people dream of it. If we take their nationalities into account, which countries’ players would make more money compared to other? Although we know that a player’s wage highly depends on their abilities and which club does he join, for the top players who their abilities are nearly the same, is there a relationship between their nationalities and their wages? What about their height and weight? This report is going to reveal to what extent and how does height, weight and nationality affect top soccer player’s wages and how is the accuracy of their wages in this data?

Data introduce

The data source I used is originally scraped from sofifa.com, which is now under the folder ‘Fifa18 more complete player dataset’ in Kaggle.com.

Data import

This file includes nearly all soccer players in the current world, including many famous soccer stars such as C.Ronaldo and L.Messi. First, I selected the potential variables that might be useful for later’s analysis. Then, since the subjects I am focusing on are the top soccer players, I removed the players whose official evaluations are under 70/100.

```
library(tidyverse)
library(dplyr)

fifa1data = read.csv(file = "complete.csv")
fifaldata <- dplyr::select(fifaldata, name, overall, club, age, league, height_cm, weight_kg, nationality, eur_
wage)
colnames(fifa1data) <- c("name", "evaluation", "club", "age", "league", "height","weight", "nationality", "wage
")
fifa1data = fifa1data[-which(fifa1data$wage == 0),]
fifa1data = fifa1data[-which(fifa1data$evaluation < 70),]
```

Brief observation

Before fitting a model, I am willing to have a basic understanding of the average weight, height and average wage for top players from different countries. First, I counted the countries in this dataset, I found that the number of some countries’ soccer players is very small. In order to improve the accuracy of analysis, I removed the rows contain the countries that the number of such countries’ soccer players is less than 10.

```
countrycount <- fifa1data %>%
  group_by(nationality) %>%
  count()
countrycount
```

nationality <fctr>	n <int>
Albania	16
Algeria	33
Angola	6
Argentina	382
Armenia	5
Australia	23
Austria	52
Azerbaijan	1
Belarus	5
Belgium	114

1-10 of 127 rows   Previous **1** 2 3 4 5 6 ... 13 Next

```
fifa1data <- fifa1data %>%
  group_by(nationality) %>%
  mutate(count = n()) %>%
  filter(count>10)
```

After that, I calculated the average weight, height and average wage of top soccer players from different countries. And then I would put these results into leaflet which can help the audience clearly see the average weight, height and wage in different countries.

```
weightcountry <- fifa1data %>%
  group_by(nationality) %>%
  summarise(avg.weight = mean(weight))
weightcountry
```

nationality <fctr>	avg.weight <dbl>
Albania	73.81250
Algeria	75.18182
Argentina	75.96859
Australia	77.39130
Austria	78.76923
Belgium	78.03509
Bosnia Herzegovina	79.10000
Brazil	76.11594
Cameroon	78.36667
Cape Verde	77.83333

1-10 of 62 rows Previous **1** 2 3 4 5 6 7 Next

```
heightcountry <- fifaldata %>%
  group_by(nationality) %>%
  summarise(avg.height = mean(height))
heightcountry
```

nationality <fctr>	avg.height <dbl>
Albania	179.7500
Algeria	181.3636
Argentina	179.5393
Australia	182.2174
Austria	183.8654
Belgium	183.3947
Bosnia Herzegovina	185.2333
Brazil	180.7723
Cameroon	182.1000
Cape Verde	181.6667

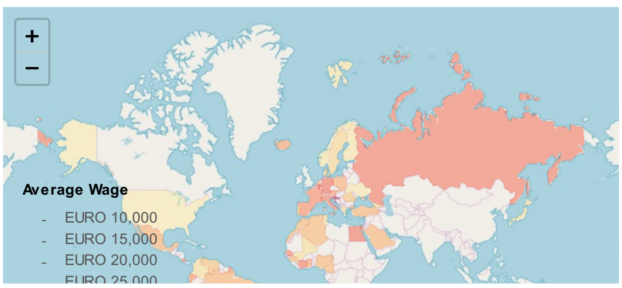
1-10 of 62 rows Previous **1** 2 3 4 5 6 7 Next

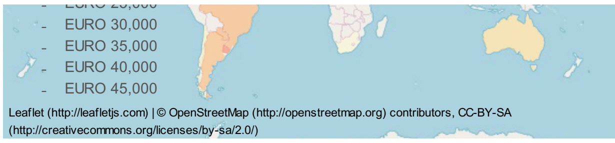
```
bodycountry <- cbind(weightcountry,heightcountry)
colnames(bodycountry) <- c("nationality", "avg.weight", "nationality2", "avg.height")
bodycountry <- dplyr::select(bodycountry, nationality, avg.weight, avg.height)

wagecountry <- fifaldata %>%
  group_by(nationality) %>%
  summarise(avg.wage = mean(wage))
wagecountry
```

nationality <fctr>	avg.wage <dbl>
Albania	17437.500
Algeria	25848.485
Argentina	25780.105
Australia	18739.130
Austria	30653.846
Belgium	39122.807
Bosnia Herzegovina	33933.333
Brazil	26993.789
Cameroon	25633.333
Cape Verde	20166.667

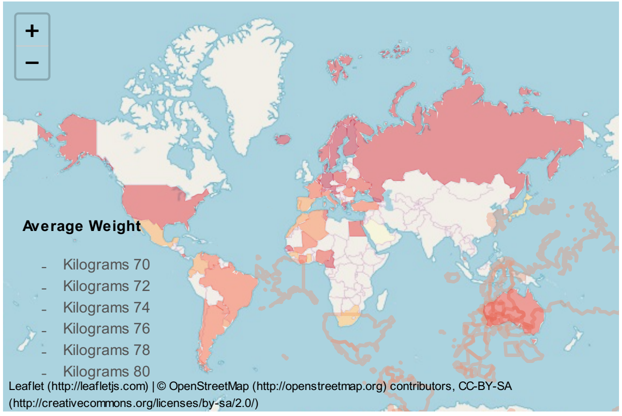
1-10 of 62 rows Previous **1** 2 3 4 5 6 7 Next

```r
library(leaflet)
library(rgdal)
library(tidyverse)

#Import Map Data
url <- "https://www.naturalearthdata.com/http//www.naturalearthdata.com/download/50m/cultural/ne_50m_admin_0_co
untries.zip"

tmp <- tempdir()
file <- basename(url)
download.file(url, file)
unzip(file, exdir = tmp)

#Merge Data
countries <- readOGR(dsn=tmp,
                     layer = "ne_50m_admin_0_countries",
                     encoding = "UTF-8", verbose=FALSE)
country_name<-intersect(countries$SUBUNIT,wagecountry$nationality)
fifa2018 <-sp::merge(countries, wagecountry %>% filter(nationality%in%country_name),
                  by.y="nationality",by.x="SUBUNIT",sort=FALSE,duplicateGeoms =
                  TRUE,all.x=FALSE)

#map1
pal <- colorNumeric("YlOrRd", domain = fifa2018$avg.wage)
labels<- sprintf(
  "<strong>%s</strong><br/>%s EURO",
  fifa2018$SUBUNIT, round(fifa2018$avg.wage,2)
) %>% lapply(htmltools::HTML)

map1 <- leaflet() %>%
  addTiles() %>%
  addPolygons(data=fifa2018,
  fillColor = ~pal(fifa2018$avg.wage),
  weight = 4,opacity = 0.4, color = "darksalmon",
  dashArray = "1",fillOpacity = 0.4,

   highlight = highlightOptions(
    weight = 5,color = "#e9967a",
    dashArray = "",fillOpacity = 0.7,bringToFront = TRUE),

  label = labels)%>%

    addLegend(pal = pal, values = fifa2018$avg.wage, opacity = 0.4, title = "Average Wage", labFormat = labelFo
rmat(prefix = "EURO "), position = "bottomleft")

#map23 formulate
country_name2<-intersect(countries$SUBUNIT,bodycountry$nationality)
fifa2018_2 <-sp::merge(countries, bodycountry %>% filter(nationality%in%country_name2),
                  by.y="nationality",by.x="SUBUNIT",sort=FALSE,duplicateGeoms =
                  TRUE,all.x=FALSE)


#map2
pal2 <- colorNumeric("YlOrRd", domain = fifa2018_2$avg.weight)
labels2<- sprintf(
  "<strong>%s</strong><br/>%s Kilogram",
  fifa2018_2$SUBUNIT, round(fifa2018_2$avg.weight,2)
) %>% lapply(htmltools::HTML)

map2 <- leaflet() %>%
  addTiles() %>%
  addPolygons(data=fifa2018_2,
  fillColor = ~pal2(fifa2018_2$avg.weight),
  weight = 4, opacity = 0.4, color = "darksalmon",
  dashArray = "1", fillOpacity = 0.4,

   highlight = highlightOptions(
    weight = 5, color = "#e9967a",
    dashArray = "", fillOpacity = 0.7, bringToFront = TRUE),

  label = labels2)%>%

    addLegend(pal2, values = fifa2018_2$avg.weight, opacity = 0.4, title = "Average Weight", labFormat = labelF
ormat(prefix = "Kilograms "), position = "bottomleft")

#map2
pal3 <- colorNumeric("YlOrRd", domain = fifa2018_2$avg.height)
labels3<- sprintf(
  "<strong>%s</strong><br/>%s centimeter",
  fifa2018_2$SUBUNIT, round(fifa2018_2$avg.height,2)
) %>% lapply(htmltools::HTML)

map3 <- leaflet() %>%
  addTiles() %>%
  addPolygons(data=fifa2018_2,
  fillColor = ~pal3(fifa2018_2$avg.height),
  weight = 4, opacity = 0.4, color = "darksalmon",
  dashArray = "1", fillOpacity = 0.4,
   highlight = highlightOptions(
    weight = 5, color = "#e9967a",
    dashArray = "", fillOpacity = 0.7, bringToFront = TRUE),

  label = labels3)%>%

    addLegend(pal3, values = fifa2018_2$avg.height, opacity = 0.4, title = "Average Height", labFormat = labelF
ormat(prefix = "Centimeters "), position = "bottomleft")

map1
```

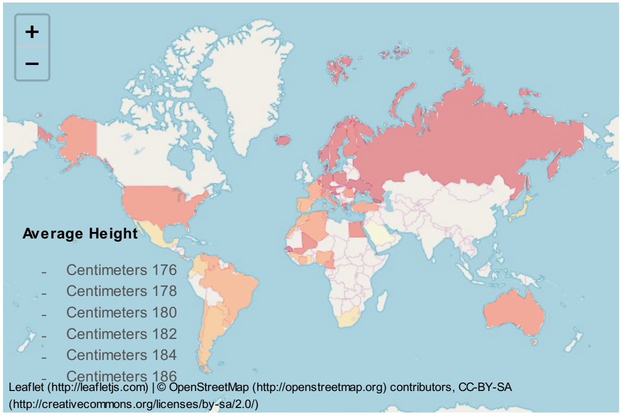





```
map2
```



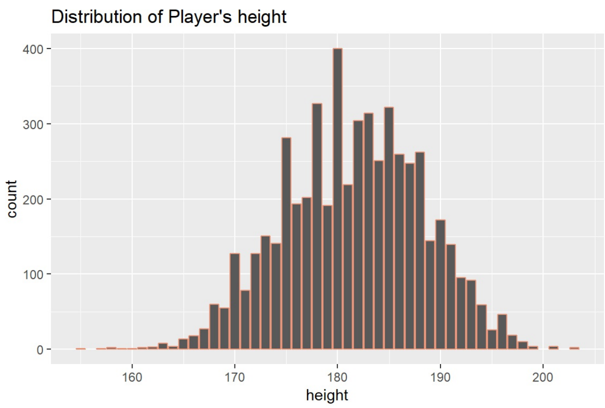

```
map3
```



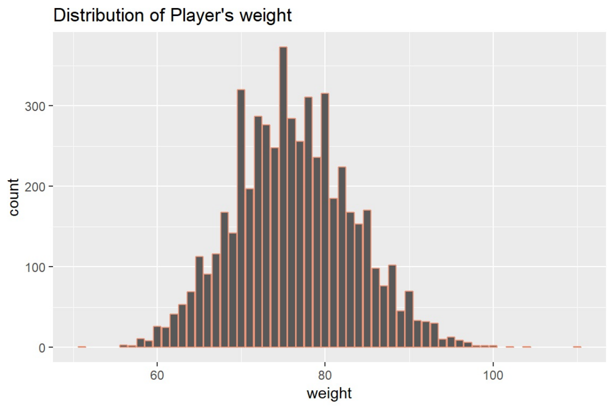
I thought that the taller a player is, the higher wage he will get. But from the maps we can clearly figure out that at least for some countries, it is the opposite.

EDA

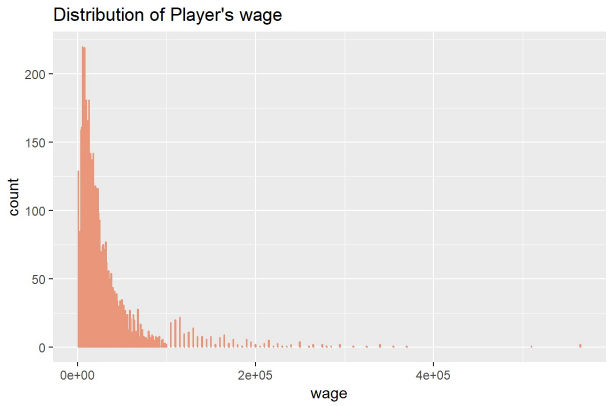
```
ggplot(data = fifaldata, aes(x = height))+
geom_bar(color = "darksalmon") + ggtitle("Distribution of Player's height")
```




```
ggplot(data = fifaldata, aes(x = weight))+
geom_bar(color = "darksalmon") + ggtitle("Distribution of Player's weight")
```



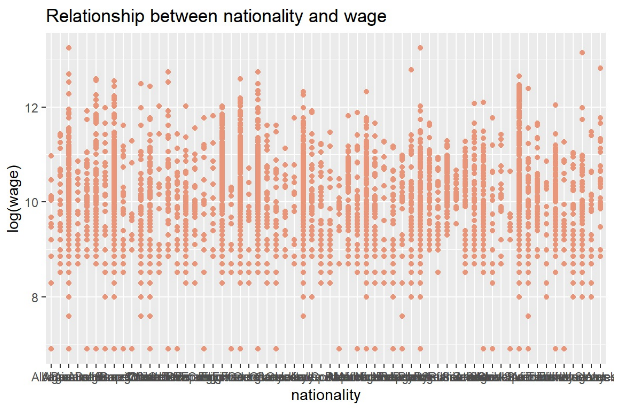

```
ggplot(data = fifaldata, aes(x = wage))+
geom_bar(color = "darksalmon") + ggtitle("Distribution of Player's wage")
```



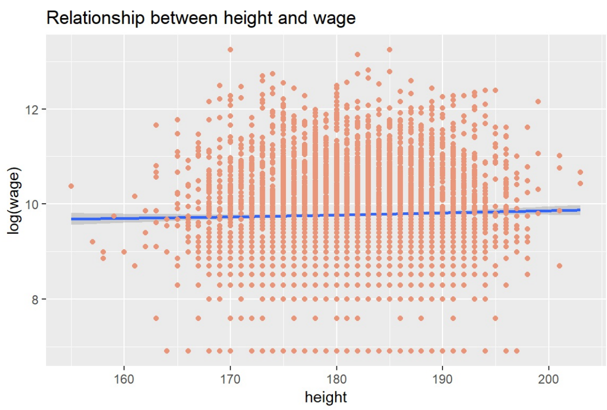
From the distributions of players' weight and height, which seems like normal distributions, we can see most of the players are around 180 cm height and 75 kg weight. But for the distribution of players' wage, as the amount of wage increases, the number of players who can get this amount decreases rapidly.

Right now, let me plot the relationships between wage and some potential factors.

```
ggplot(fifaldata, aes(x = nationality, y= log(wage))) + geom_smooth() + geom_point(color="darksalmon")  + ggtit
le("Relationship between nationality and wage")
```




```
ggplot(fifaldata, aes(x = height, y= log(wage))) + geom_smooth() + geom_point(color="darksalmon")  + ggtitle("R
elationship between height and wage")
```




```
ggplot(fifaldata, aes(x = weight, y= log(wage))) + geom_smooth() + geom_point(color="darksalmon")  + ggtitle("R
elationship between weight and wage")
```

We can see that players from some countries make more money compared to others. For the relationship between height and wage, I thought that players tend to gain more money if they are taller, but the fact is that height does not have any influence on their wages. By looking at the relationship between weight and wage, what is interesting is that although the influence of players' weight is really tiny, there is still a small tendency that the more weight results in the more wage they get.

Model formulating and Interpretation

```
library(arm)
relationshipmodel <- lm(data = fifaldata, log(wage) ~ nationality + weight)
summary(relationshipmodel)
```

```
##
## Call:
## lm(formula = log(wage) ~ nationality + weight, data = fifaldata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2325 -0.5519 -0.0263  0.5288  3.7205
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     9.138396   0.266997  34.227  < 2e-16 ***
## nationalityAlgeria              0.396231   0.281092   1.410 0.158712
## nationalityArgentina            0.294114   0.235484   1.249 0.211728
## nationalityAustralia            0.186841   0.300442   0.622 0.534042
## nationalityAustria              0.640431   0.263935   2.426 0.015279 *
## nationalityBelgium              0.532568   0.246446   2.161 0.030740 *
## nationalityBosnia Herzegovina   0.434155   0.285795   1.519 0.128794
## nationalityBrazil               0.403209   0.234497   1.719 0.085587 .
## nationalityCameroon             0.324117   0.285754   1.134 0.256739
## nationalityCape Verde           0.038727   0.352430   0.110 0.912504
## nationalityChile               -0.562197   0.238398  -2.358 0.018398 *
## nationalityColombia            -0.432833   0.242508  -1.785 0.074347 .
## nationalityCosta Rica           0.305084   0.337718   0.903 0.366371
## nationalityCroatia              0.693504   0.266485   2.602 0.009283 **
## nationalityCzech Republic      -0.079599   0.276174  -0.288 0.773189
## nationalityDenmark              0.279258   0.256536   1.089 0.276393
## nationalityDR Congo             0.467475   0.293176   1.595 0.110878
## nationalityEcuador              0.483164   0.309477   1.561 0.118530
## nationalityEgypt                0.705275   0.344592   2.047 0.040737 *
## nationalityEngland              1.047681   0.237377   4.414 1.04e-05 ***
## nationalityFinland             -0.086355   0.344672  -0.251 0.802177
## nationalityFrance               0.570268   0.235886   2.418 0.015658 *
## nationalityGeorgia              0.121820   0.352529   0.346 0.729687
## nationalityGermany              0.677229   0.237258   2.854 0.004328 **
## nationalityGhana                0.505035   0.277241   1.822 0.068564 .
## nationalityGreece              -0.672192   0.271220  -2.478 0.013228 *
## nationalityGuinea               0.536950   0.337667   1.590 0.111854
## nationalityIceland              0.520149   0.344668   1.509 0.131325
## nationalityItaly                0.545956   0.237346   2.300 0.021472 *
## nationalityIvory Coast          0.712733   0.272015   2.620 0.008813 **
## nationalityJapan                0.049069   0.261585   0.188 0.851209
## nationalityKorea Republic      -0.373761   0.264402  -1.414 0.157536
## nationalityMali                 0.109615   0.317050   0.346 0.729557
## nationalityMexico               0.646163   0.249152   2.593 0.009528 **
## nationalityMorocco              0.284575   0.269368   1.056 0.290809
## nationalityNetherlands          0.402326   0.241992   1.663 0.096459 .
## nationalityNigeria              0.405141   0.269431   1.504 0.132720
## nationalityNorthern Ireland     0.497371   0.344581   1.443 0.148965
## nationalityNorway              -0.164301   0.269488  -0.610 0.542100
## nationalityParaguay             0.486743   0.281119   1.731 0.083430 .
## nationalityPoland               0.262786   0.259273   1.014 0.310844
## nationalityPortugal             0.102144   0.239460   0.427 0.669718
## nationalityRepublic of Ireland  0.816307   0.269416   3.030 0.002458 **
## nationalityRomania              0.449944   0.300411   1.498 0.134254
## nationalityRussia               0.949496   0.248402   3.822 0.000134 ***
## nationalitySaudi Arabia         0.784190   0.285820   2.744 0.006096 **
## nationalityScotland             0.810510   0.261556   3.099 0.001953 **
## nationalitySenegal              0.604046   0.260668   2.317 0.020525 *
## nationalitySerbia               0.327903   0.254226   1.290 0.197173
## nationalitySlovakia             0.296750   0.321468   0.923 0.355993
## nationalitySlovenia             0.631409   0.303421   2.081 0.037484 *
## nationalitySouth Africa        -1.335414   0.313082  -4.265 2.03e-05 ***
## nationalitySpain                0.447785   0.234214   1.912 0.055947 .
## nationalitySweden              -0.158551   0.252110  -0.629 0.529444
## nationalitySwitzerland          0.691297   0.261693   2.642 0.008275 **
## nationalityTunisia              0.118695   0.344587   0.344 0.730517
## nationalityTurkey               0.546838   0.247201   2.212 0.027001 *
## nationalityUkraine             -0.768194   0.285705  -2.689 0.007194 **
## nationalityUnited States       -0.235012   0.256488  -0.916 0.359567
## nationalityUruguay              0.478427   0.256749   1.863 0.062460 .
## nationalityVenezuela            0.221090   0.306213   0.722 0.470317
## nationalityWales                1.117668   0.295451   3.783 0.000157 ***
## weight                          0.003544   0.001822   1.945 0.051770 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9227 on 5342 degrees of freedom
## Multiple R-squared:  0.1614, Adjusted R-squared:  0.1517
## F-statistic: 16.58 on 62 and 5342 DF,  p-value: < 2.2e-16
```

```
anova(relationshipmodel)
```

	Df <int>	Sum Sq <dbl>	Mean Sq <dbl>	F value <dbl>	Pr(>F) <dbl>
nationality	61	871.971272	14.2946110	16.79082	4.168933e-158
weight	1	3.222209	3.2222092	3.78489	5.176969e-02
Residuals	5342	4547.830966	0.8513349	NA	NA

3 rows

From the summary, we can clearly find that for a players' weight increases by 1 unit, his wage would increase by 0.003544, which is expected by comparing it with the former EDA. Compare to weight, nationality has more impact on soccer players' wage.

Since different nationalities would have different impacts on players' wage, I am going to use the method K-means clustering to make these nationalities' coefficients into several clusters.

```
nationalcoef <- relationshipmodel$coefficients[2:62]

set.seed(1)
nationalkmean <- kmeans(nationalcoef, 4, nstart = 25)
nationalkmean
```

```
## K-means clustering with 4 clusters of sizes 28, 15, 12, 6
##
## Cluster means:
##            [,1]
## 1  0.4180504318
## 2  0.7685292812
## 3  0.0002577608
## 4 -0.6907652112
##
## Clustering vector:
##             nationalityAlgeria           nationalityArgentina
##                              1                              1
##           nationalityAustralia             nationalityAustria
##                              3                              2
##             nationalityBelgium  nationalityBosnia Herzegovina
##                              1                              1
##              nationalityBrazil            nationalityCameroon
##                              1                              1
##          nationalityCape Verde               nationalityChile
##                              3                              4
##            nationalityColombia          nationalityCosta Rica
##                              4                              1
##             nationalityCroatia      nationalityCzech Republic
##                              2                              3
##             nationalityDenmark            nationalityDR Congo
##                              1                              1
##             nationalityEcuador               nationalityEgypt
##                              1                              2
##             nationalityEngland             nationalityFinland
##                              2                              3
##              nationalityFrance             nationalityGeorgia
##                              1                              3
##             nationalityGermany               nationalityGhana
##                              2                              1
##              nationalityGreece              nationalityGuinea
##                              4                              1
##             nationalityIceland               nationalityItaly
##                              1                              1
##         nationalityIvory Coast               nationalityJapan
##                              2                              3
##      nationalityKorea Republic                nationalityMali
##                              4                              3
##              nationalityMexico             nationalityMorocco
##                              2                              1
##         nationalityNetherlands             nationalityNigeria
##                              1                              1
##    nationalityNorthern Ireland              nationalityNorway
##                              1                              3
##            nationalityParaguay              nationalityPoland
##                              1                              1
##            nationalityPortugal nationalityRepublic of Ireland
##                              3                              2
##             nationalityRomania              nationalityRussia
##                              1                              2
##        nationalitySaudi Arabia            nationalityScotland
##                              2                              2
##             nationalitySenegal              nationalitySerbia
##                              2                              1
##            nationalitySlovakia            nationalitySlovenia
##                              1                              2
##        nationalitySouth Africa               nationalitySpain
##                              4                              1
##              nationalitySweden         nationalitySwitzerland
##                              3                              2
##             nationalityTunisia              nationalityTurkey
##                              3                              1
##             nationalityUkraine       nationalityUnited States
##                              4                              3
##             nationalityUruguay           nationalityVenezuela
##                              1                              1
##               nationalityWales
##                              2
##
## Within cluster sum of squares by cluster:
## [1] 0.2857326 0.3411461 0.2113508 0.6054636
##  (between_SS / total_SS =  88.1 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

By using this K-means clustering method, nationalities' coefficients are divided into four clusters, which means of coefficients are 0.41805 (cluster#1),0.76852 (cluster#2), 0.00025 (cluster#3), -0.69076 (cluster#4).

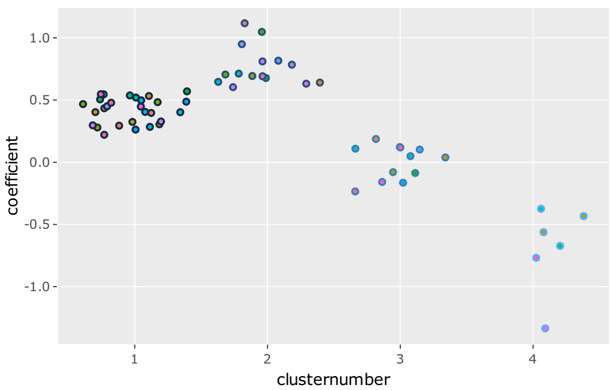
```
library(plotly)
nationalname = names(relationshipmodel$coefficients)[2:62]
nationalcluster <- as.data.frame(nationalkmean$cluster)
nationalcluster <- cbind(nationalname, nationalcluster,nationalcoef)
colnames(nationalcluster) <- c("nationality","clusternumber","coefficient")
nationalcluster
```

	nationality <fctr>
nationalityAlgeria	nationalityAlgeria
nationalityArgentina	nationalityArgentina
nationalityAustralia	nationalityAustralia
nationalityAustria	nationalityAustria
nationalityBelgium	nationalityBelgium
nationalityBosnia Herzegovina	nationalityBosnia Herzegovina
nationalityBrazil	nationalityBrazil
nationalityCameroon	nationalityCameroon

nationalityCape Verde	nationalityCape Verde
nationalityChile	nationalityChile

1-10 of 61 rows | 1-2 of 4 columns   Previous 1 2 3 4 5 6 7 Next

```
a <- ggplot(data= nationalcluster, aes(x=clusternumber, y=coefficient, color=clusternumber, fill= nationality))
+ geom_jitter() + theme(legend.position="none")
ggplotly(a)
```

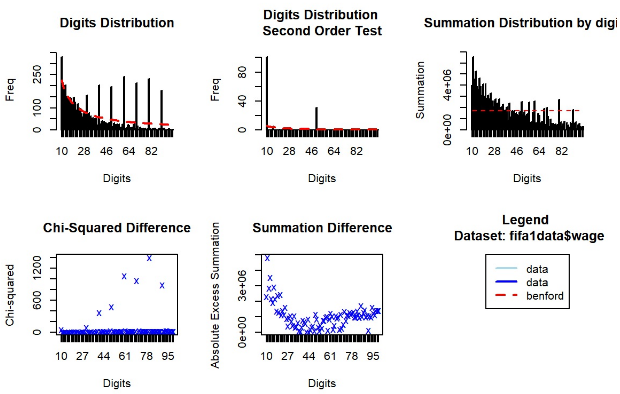


After transferring the data of cluster number into a data frame and a plot, it is easy for us to find that the top soccer players whose nationalities in cluster#2 have more chance to get more wages, such as Germany, Croatia, Egypt, England and so on. Also, the top soccer players whose nationalities in cluster#4 might get relatively fewer wages, such as Chile, Colombia and so on.

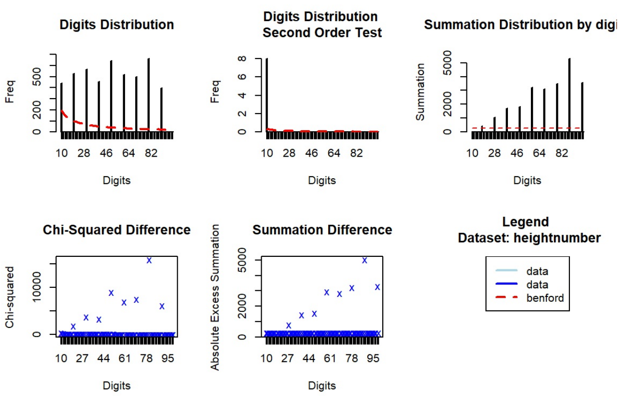
Benford Law Test

Now, one question comes to us. Is this dataset reliable for our analysis?

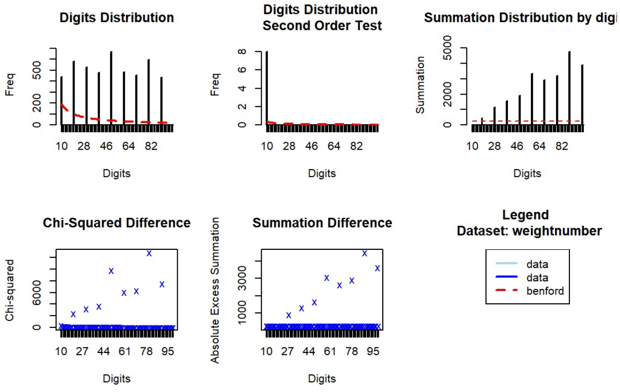
```
library(benford.analysis)
bf_wage <- benford(fifaldata$wage)

plot(bf_wage)
```




```
a <- as.character(fifaldata$height)
heightnumber <- as.numeric(substr(a, 3,3))
bf_height <- benford(heightnumber)

plot(bf_height)
```




```
b <- as.character(fifaldata$weight)
weightnumber <- as.numeric(substr(b,2,2))
bf_weight <- benford(weightnumber)

plot(bf_weight)
```



According to the digit distribution of soccer players' wage, the data somehow have a tendency to follow Benford's law, but discrepancies are also clear at around 30,40,50,60,70 and 80. In additon, for the Benford Analysis Plot of height and weight, we can clearly find that height And weight does not follow the tendency of Benford Law.

Conclusion and Future

In short, a player's height and weight do not have any obvious effect on their wage. Players from some countries would have a higher wage, maybe because their countries pay more attention to the development of sports so that they were better trained and had better physical qualities. Furthermore, the Benford Analysis shows that the data might not authentic enough. In the future, to do a more solid analysis for top soccer players' wage, additional variables and methods are needed.