FOOTBALL DATA WAREHOUSE

PROJECT

APRIL 24, 2016

ASWIN SIVAM RAVIKUMAR X16134621
MASTERS IN DATA ANALYTICS

INTRODUCTION

Data warehousing it is an efficient way to analyze a large amount of data as well as the statistics and it helps the many organizations, firms, industries, and many more companies to make a decision. By analyzing the data in the date warehousing it helps the users to make a good decision in their future projects.

my project is about football players and their respected clubs and nationalities by analyzing the various source data we are creating the business intelligence query and how well it predict the future lively.

DATA SOURCE

I have taken the three various data sets two are structured data and the other one is unstructured data, first structured data is taken from kaggle and the other structured data is taken by net scrapping with the help of R program, and the last one is unstructured data it is taken from twitter the tweets are derived by using the R program.

STRUCTURED DATA-KAGGLE, NET SCRAPPING

UNSTRUCTURED DATA- TWEETS FROM TWITTER

1)After taking the data sets the first structured data is from kaggle it consists of the every detail of the player related to their clubs, club position, nationality, nationality position, and their personal info such as age, height, weight. It is a structured data but the data is not clean as it consists of some special characters so I used google refine and Microsoft excel to clean that special characters.

Luis SuÃ¡rez-Luis Suarez

2)The second data it is a structured data and it is taken from the Wikipedia by using the net scrapping with the help of R code I derived the international goals and the club goals for the players in the form of a table.

3)Unstructured data, tweets are derived from the twitter with the help of R Programming in the form of positive and negative counts they are converted into scores for out of 100.

Three data source are combined together into a single excel source file so that the implementation can be done easily. Due to some missing cells with the help of previous data, I have randomly generated some data from the SQL data generator and mockaroo random data generator.

R PROGRAMMING CODE TWITTER [3] [moodle source]

“#these are the packages we need for this example - executing this line will install them
install.packages(c("devtools", "rjson", "bit64", "httr", "plyr", "ggplot2", "doBy", "XML", "base64enc"))

#devtools allows us to install from github
library(devtools)

#here we install 2 R packages from github
install_github("geoffjentry/twitteR" , force=TRUE)
install_github('R-package','quandl' , force=TRUE)

#these are various libraries that we use throughout this example
library(plyr)

```
library(httr)
library(doBy)
library(Quandl)
library(twitteR)

#here mz api keys would go, to run this example zou need to input you keys, and secrets
api_key <- "w0u2TyUtAga0dPLAc1huwwGYq"
api_secret <- "sO4kBhyLJEgSB2RaOLIH3qAZbfIjS582vyobM8HVRefLhQMF72"
access_token <- "709310096-v2FkS6uT5cOFbQklVwiH3oeqKP7Hkw8lWjs7WQz5"
access_token_secret <- "MGDyKHNeeZFhi9QL6AdkqRLLJtUd9gMjsCyQHxEnL3src"

#here we setup up twitter and provide authentication details, i.e. the keys and their secrets
setup_twitter_oauth(api_key,api_secret,access_token,access_token_secret)

#here we read in the two dictionaries that have been downloaded -
hu.liu.pos = scan('C:/Users/Aswin_sivam/Downloads/positive-words.txt', what='character',
comment.char=';')
hu.liu.neg = scan('C:/Users/Aswin_sivam/Downloads/negative-words.txt', what='character',
comment.char=';')

#now we can add some domain-specific terminolgy
pos.words = c(hu.liu.pos, 'awesome','good','favorite','great','nice','happy','excellent')
neg.words = c(hu.liu.neg, 'wtf', 'boring', 'douzy',
'hilarious','terrible','fake','glorified','aborted','sad')

#our first function
score.sentence <- function(sentence, pos.words, neg.words) {
  #here some basic cleaning
  sentence = gsub('[[:punct:]]', '', sentence)
  sentence = gsub('[[:cntrl::]]', '', sentence)
  sentence = gsub('\\d+', '', sentence)
  sentence = tolower(sentence)

  #basic data structure construction
  word.list = str_split(sentence, '\\s+')
  words = unlist(word.list)

  #here we count the number of words that are positive and negative
  pos.matches = match(words, pos.words)
  neg.matches = match(words, neg.words)

  #throw away those that didn't match
  pos.matches = !is.na(pos.matches)
  neg.matches = !is.na(neg.matches)

  #compute the sentiment score
  score = sum(pos.matches) - sum(neg.matches)

  return(score)
}

#our second function that takes an array of sentences and sentiment analyses them
score.sentiment <- function(sentences, pos.words, neg.words) {
  require(plyr)
  require(stringr)
```

```
  #here any sentence/tweet that causes an error is given a sentiment score of 0 (neutral)
  scores = laply(sentences, function(sentence, pos.words, neg.words) {
    tryCatch(score.sentence(sentence, pos.words, neg.words ), error=function(e) 0)
  }, pos.words, neg.words)

  #now we construct a data frame
  scores.df = data.frame(score=scores, text=sentences)

  return(scores.df)
}

#our third function, that communicates with twitter and then scores each of the tweets returned
collect.and.score <- function (Aswin, Actor, later, pos.words, neg.words) {


  tweets = searchTwitter(venkat, n=1500, lang="en", since=NULL, retryOnRateLimit=10)
  text = laply(tweets, function(t) t$getText())

  score = score.sentiment(text, pos.words, neg.words)
  score$Player = Player
  score$later = later

  return (score)
}


#here we invoke the function above for each of our players
ronaldo.scores = collect.and.score("@Cristiano","cristiano", "cr7", pos.words, neg.words)
print(ronaldo.scores)
View(ronaldo.scores)

messi.scores = collect.and.score("@messi10stats", "messi", "me", pos.words, neg.words)
print(messi.scores)
View(messi.scores)

manuel.scores = collect.and.score("@Manuel_Neuer", "manuel", "man",pos.words, neg.words)
print(manuel.scores)
View(manuel.scores)

edenhazard.scores = collect.and.score("@hazardeden10", "eden hazard", "eden", pos.words,
neg.words)
print(edenhazard.scores)
View(edenhazard.scores)

zlantan.scores = collect.and.score("@Ibra_official", "zlantan ibramovic", "zla", pos.words,
neg.words)
print(zlantan.scores)
View(zlantan.scores)



# getout.scores = collect.and.score("@getout", "getout", "getout", pos.words, neg.words)
#we combine all data frames into 1
all.scores = rbind(ronaldo.scores,messi.scores,manuel.scores,edenhazard.scores,zlantan.scores)
```

```
all.scores
View(all.scores)
write.csv(all.scores,file='twitter sentimenttop 5.csv')
getwd()

#skim only the most positive or negative tweets to throw away noise near 0
all.scores$very.pos = as.numeric( all.scores$score >= 2)
all.scores$very.neg = as.numeric( all.scores$score <= -2)

#now we construct the twitter data frame and simultaneously compute the pos/neg sentiment
scores for each airline
twitter.df = ddply(all.scores,c('text', 'later'), summarise, pos.count = sum (very.pos), neg.count =
sum(very.neg))

#and here the general sentiment
twitter.df$all.count = twitter.df$pos.count + twitter.df$neg.count

#now in order to be able to compare data sets we normalise the sentiment score to be a
percentage
twitter.df$score = round (100 * twitter.df$pos.count / twitter.df$all.count)

#and to help understand our data, order by our now normalised score
orderBy(~-score, twitter.df)
write.csv(orderBy(~-score, twitter.df),file='aswin.csv')
getwd()"
```

IMPLEMENTATION AND ARCHITECTURE

The most widely used architecture to build a data warehouse is Kimball and Inmon myths the Kimball is the bottom-up approach which consists of the dimensions and the facts and the Inmon is a ton up front design it takes a long time process.

INMON- Relational modeling

In the overview of my dataset, it consists around 20000 rows and it consists of more measures so it can be added as facts in my perspective Kimball approach will be the best way so the EDW duplicate data can be found as long it is controlled by the ETL.

KIMBALL- Dimensional modeling-FOOTBALL DATAWAREHOUSE

The Kimball is the quick way of processing and analyzing the data, the data are integrated via dimensions so the ETL process will be less so there is a less possibility of data duplication it is a business driven model so that users can be more interactive to the process taking place. Inmon is used only by the 30 percentage of companies so most of them will prefer Kimball dimension modeling.

SCHEMA

Two types schema are present they are star and snowflake it's the design used to build the data warehouse.
I'm using the star schema this model uses the denormalized data and converts it into normalized data by grouping so the data will be organized and it will eliminate the redundancy as it is a star structure the center part consists of the fact table surrounded by the dimensions they are connected by the primary and the foreign key relationship

Snowflake also uses the normalized data but due to it the process is slower and the time consumption is more, the dimension tables will be divided into the sub-dimensions.

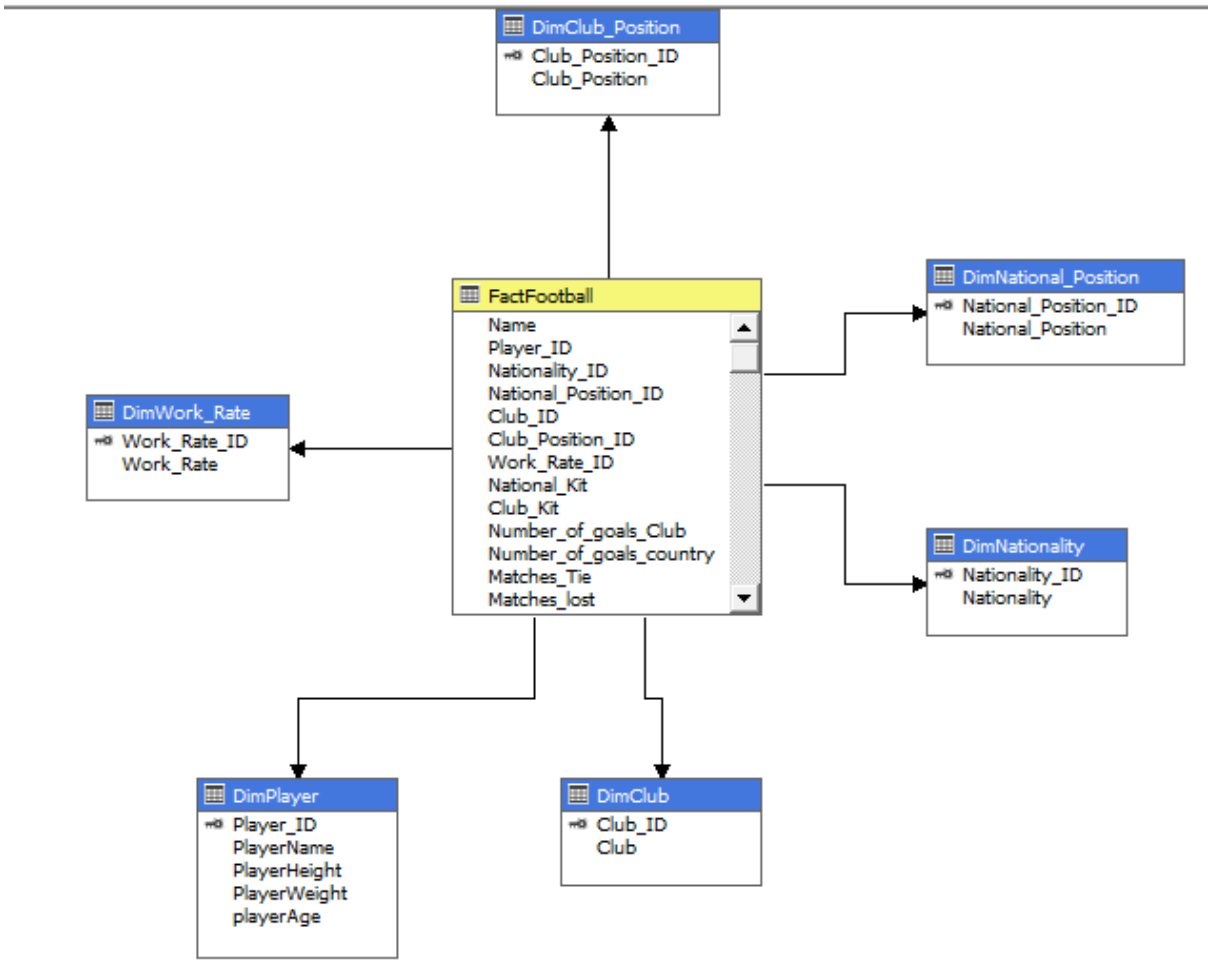


FIGURE REPRESENTATION OF STAR SCHEMA

AUTOMATION PROCESS

The database is created after that By using the SQL query table contents are directly fetched from the excel sheet the primary keys and the foreign keys are assigned to the respected columns and rows the dimension and the fact tables are created using the query and they assigned to the respected tables SQL query used for my process is given below.

SQL QUERY

```
CREATE DATABASE AswinFootball

DROP TABLE DimPlayer
DROP TABLE DimNationality
DROP TABLE DimNational_Position
DROP TABLE DimClub
DROP TABLE DimClub_Position
```

```
DROP TABLE DimWork_Rate
DROP TABLE FactFootball
CREATE TABLE DimPlayer
(
Player_ID INT IDENTITY(0,1) PRIMARY KEY,
PlayerName NVARCHAR(255),
PlayerHeight INT,
PlayerWeight INT,
playerAge INT
)
CREATE TABLE DimNationality
(
Nationality_ID INT IDENTITY(0,1) PRIMARY KEY,
Nationality NVARCHAR(255)
)

CREATE TABLE DimNational_Position
(
National_Position_ID INT IDENTITY(0,1) PRIMARY KEY,
National_Position NVARCHAR(255)
)

CREATE TABLE DimClub
(
Club_ID INT IDENTITY(0,1) PRIMARY KEY,
Club NVARCHAR(255)
)

CREATE TABLE DimClub_Position
(
Club_Position_ID INT IDENTITY(0,1) PRIMARY KEY,
Club_Position NVARCHAR(255)
)

CREATE TABLE DimWork_Rate
(
Work_Rate_ID INT IDENTITY(0,1) PRIMARY KEY,
Work_Rate NVARCHAR(255)
)


CREATE TABLE FactFootball
(
Name NVARCHAR(255),
Player_ID INT FOREIGN KEY REFERENCES dbo.DimPlayer(Player_ID),
Nationality_ID INT FOREIGN KEY REFERENCES dbo.DimNationality(Nationality_ID),
National_Position_ID INT FOREIGN KEY REFERENCES
dbo.DimNational_Position(National_Position_ID),
Club_ID INT FOREIGN KEY REFERENCES dbo.DimClub(Club_ID),
Club_Position_ID INT FOREIGN KEY REFERENCES dbo.DimClub_Position(Club_Position_ID),
Work_Rate_ID INT FOREIGN KEY REFERENCES dbo.DimWork_Rate(Work_Rate_ID),
Preffered_Foot VARCHAR(255),
National_Kit INT,
Club_Kit INT,
Number_of_goals_Club INT,Number_of_goals_country INT,Matches_Tie INT,Matches_lost
```

INT,Matches_wonINT,Weak_footINT,Skill_Moves INT,Ball_Control INT,Dribbling INT,
Marking INT,Sliding_Tackle INT,Standing_Tackle INT,Aggression INT,Reactions INT,
Attacking_Position INT,Interceptions INT,Vision INT,Composure INT,Crossing INT,
Short_Pass INT,Long_Pass INT,Acceleration INT,Speed INT,Stamina INT,Strength INT,
Balance INT,Agility INT,Jumping INT,Heading INT,Shot_Power INT,Finishing INT,
Long_Shots INT,Curve INT,Freekick_Accuracy INT,Penalties INT,Volleys INT,
GK_Positioning INT,GK_Diving INT,GK_Kicking INT,GK_Handling INT,
GK_Reflexes INT,
Rating INT,
Twitter_Rating INT
)

Since I have the dimensional modeling I'm using Kimball architecture for my data warehouse I have mentioned the dimensional process in the following steps

1.)Picking up the process - my selected process for the business is FOOTBALL.

2.)Their level of data for the dimensional model is the goals scored by the players in the nationalities and their respected clubs.

DIMENSIONS

It is a master table composed of individual and nonoverlapping data elements the primary function of the dimension is filtering, grouping and labeling our data.The dimension model used here is slowly changing dimensions it is the most used dimension.the dimensions are the individual tables.

1.) Player- This dimension consists of basic details of the player's name, their respected age, height, and weight.

2.)Nationality- This dimension consists of players nationality respected to their countries.

3.)National Position- It provides the information of national position such as defense, forward, left wing, right wing etc.

4.)Club- It gives the data of club related to their players as the different players will relate to different clubs.

5.)Club Position- The information provide here will be club position related to their clubs and related players.

6.)Work Rate- This dimension consists of the work rate like whether they are high, low or average rate related to their performance in the club and nationality.

FACTS

Fact table usually consists the measurements for the business intelligence process the measurements are listed below.

1.)Rating,
2.)Twitter Rating
3.)Number_of_goals_Club

4.)Number_of_goals_country
5.)Matches_Tie
6.)Matches_lost
7.)Matches_won

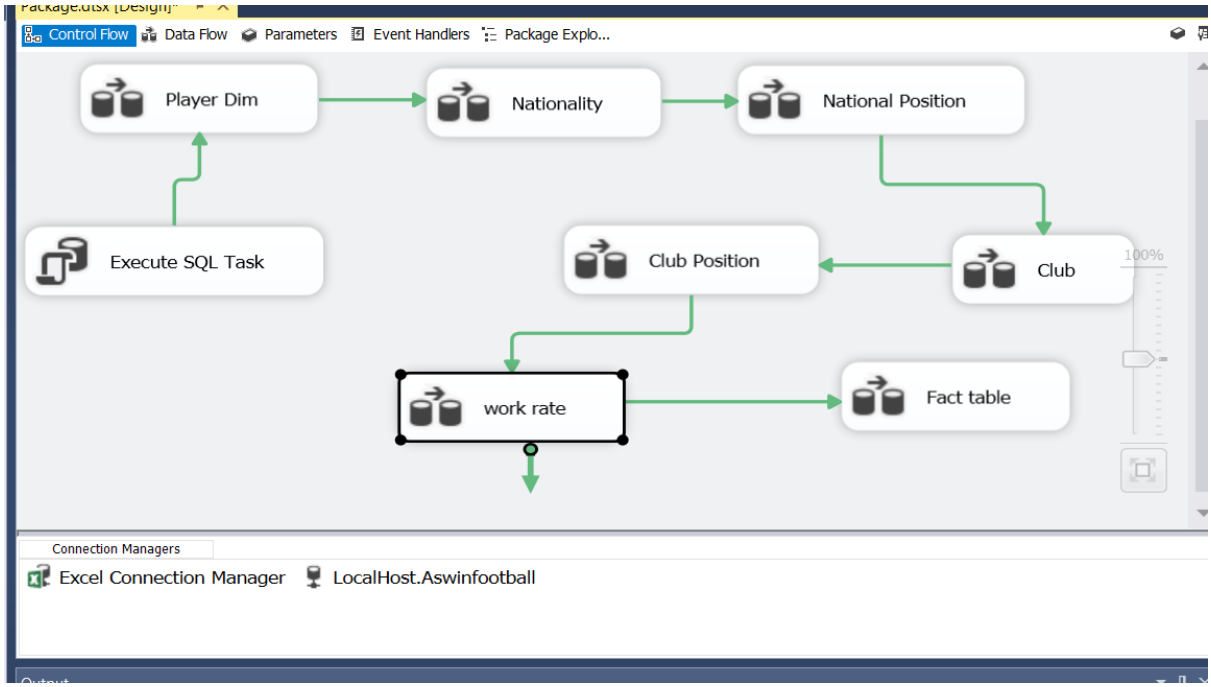
8.)Weak_foot ,Skill_Moves ,Ball_Control ,Dribbling ,Marking ,Sliding_Tackle ,Standing_Tackle ,Aggression ,Reactions ,attacking_Position ,Interceptions ,Vision ,Composure ,Crossing ,Short_Pass ,Long_Pass ,Acceleration ,Speed ,Stamina ,Strength ,Balance ,Agility ,Jumping ,Heading ,Shot_Power ,Finishing ,Long_Shots ,Curve ,Freekick_Accuracy ,Penalties ,Volleys,GK_Positioning ,GK_Diving ,GK_Kicking ,GK_Handling ,GK_Reflexes .

I have added the player's skills as a fact so its help to compare between the players that who has the best ability to perform.

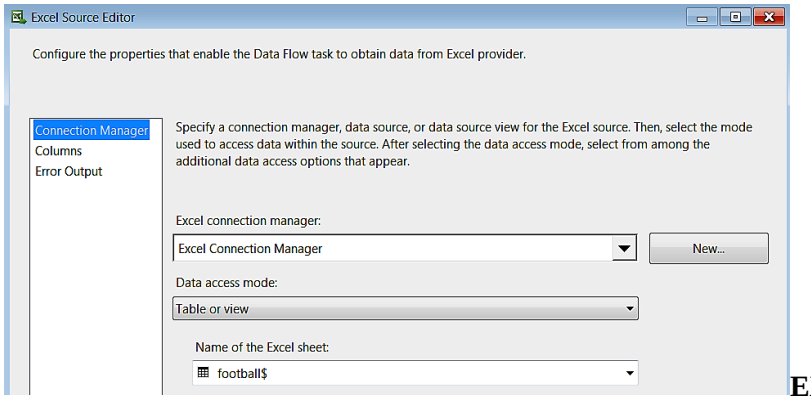
ETL EXTRACT TRANSFORM AND LOAD

Data is fetched from the heterogeneous source from the source systems and moved into the staging area.
The following process will explain how the data extraction is done in visual studio server data tools. Open and click on new project and choose integration services. Choose the data flow task and rename it to our dimensions

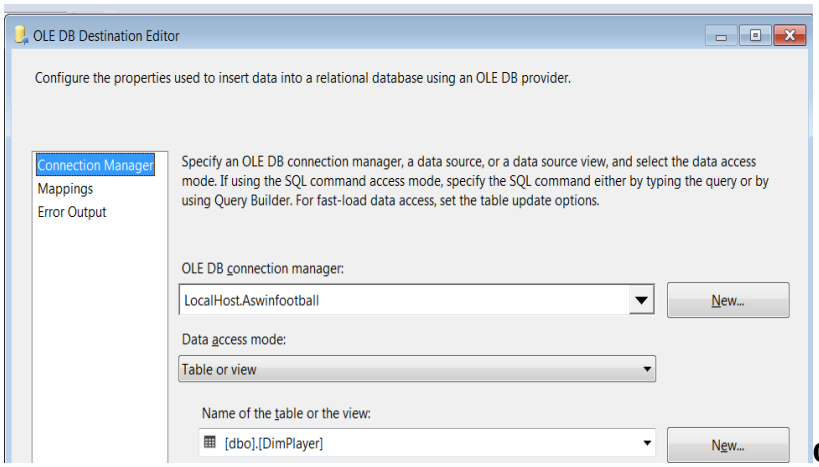
1.)The control flow describes the data flow task as my data warehouse have six dimensions all the dimensions are created the data are fetched and loaded into the dimension table.



2.)Every data flow task will have the excel connection manager by using the excel file is loaded as the input and they are connected to the OLEDB destination which is connected to the database football.

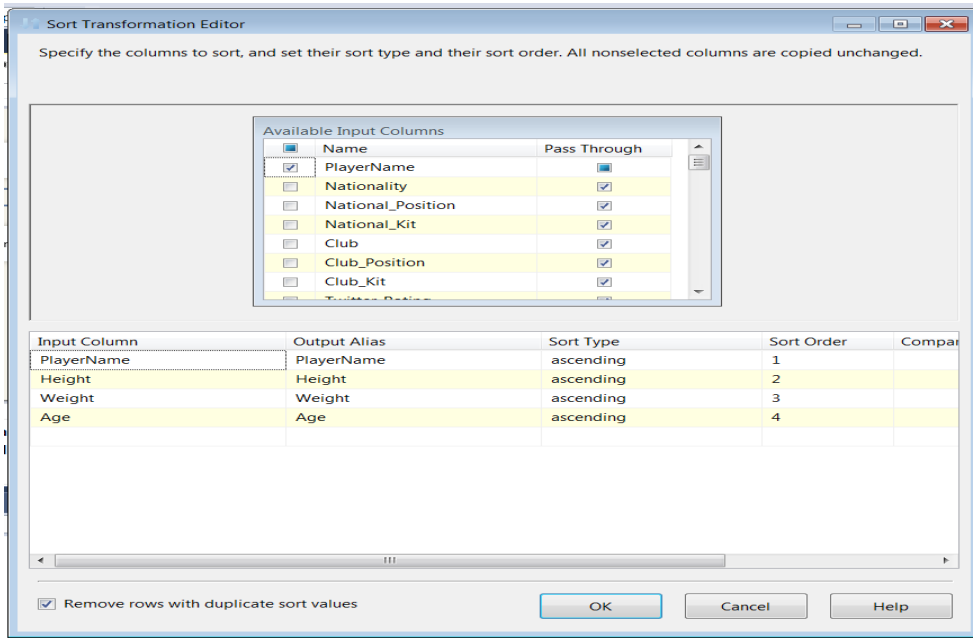


EXCEL CONNECTION

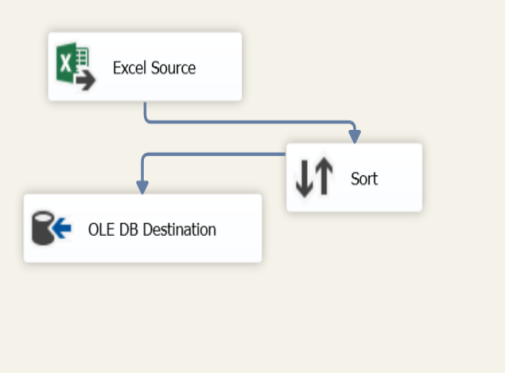


OLEDB CONNECTION

3.)The excel connection is the input and the OLEDB connection is the dimension table we are sending the inputs from excel to a destination I have used the sort function so it helps to remove the duplicate rows and sends to the destination.if any error occurs in the datatype it can be converted into the advanced options of the excel source.once the name of the table is specified in OLEDB destination editor check out the mappings are made correctly.

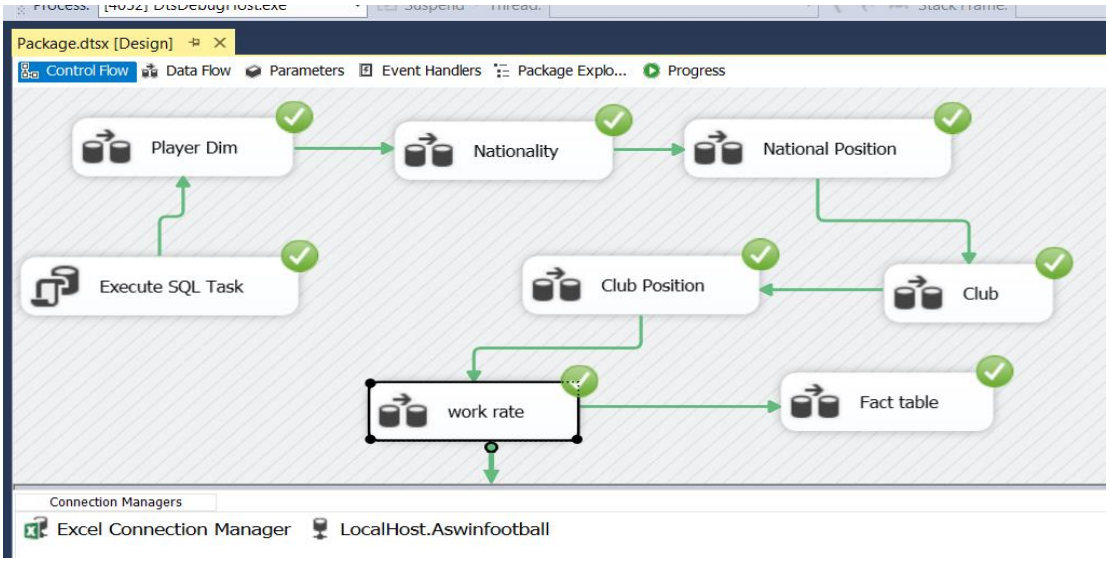


SORT FUNCTION



DATA FLOW

4.)The rest of dimensions of created and then execution process takes place so the data are obtained from the respected source.

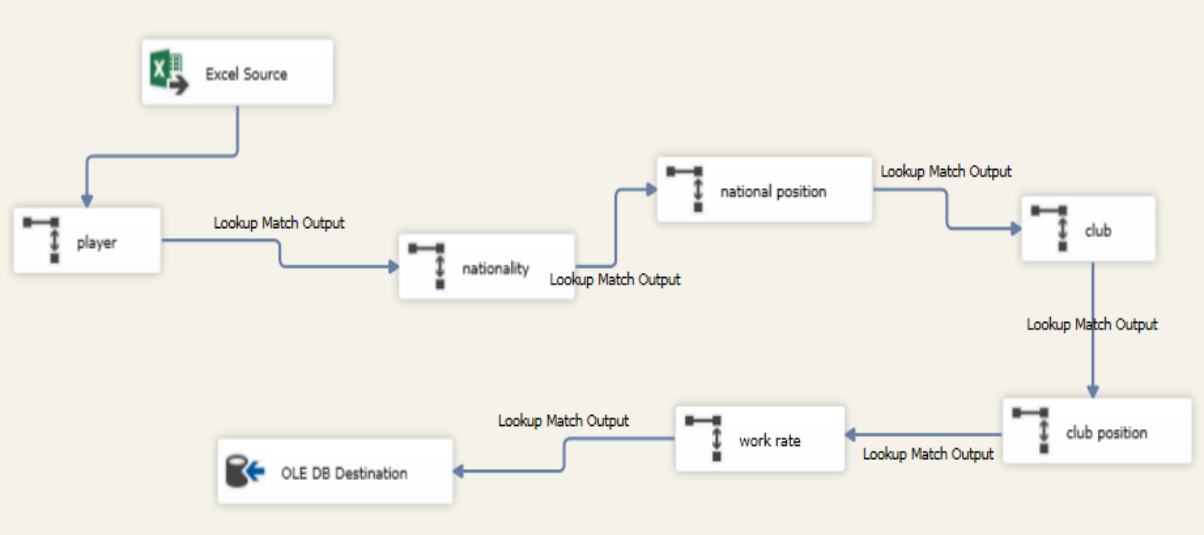


EXECUTION PROCESS

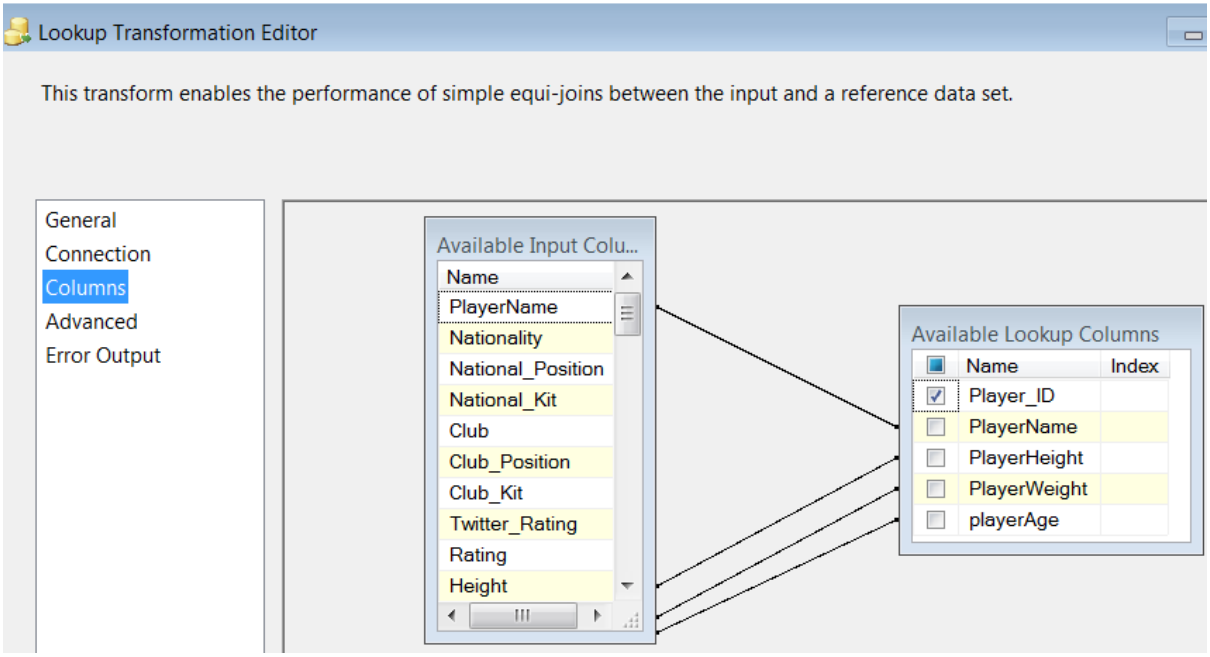
The queries for creating the dimension table and the fact table are written in the SQL server management studio 2014.

5.)The fact table consists of the source file and the other lookup files each lookup file will be dimension if there is any error data it can be converted by using the data converter if the data types are accurate then there is no need of using the data converters.

FACT TABLE CONTROL FLOW

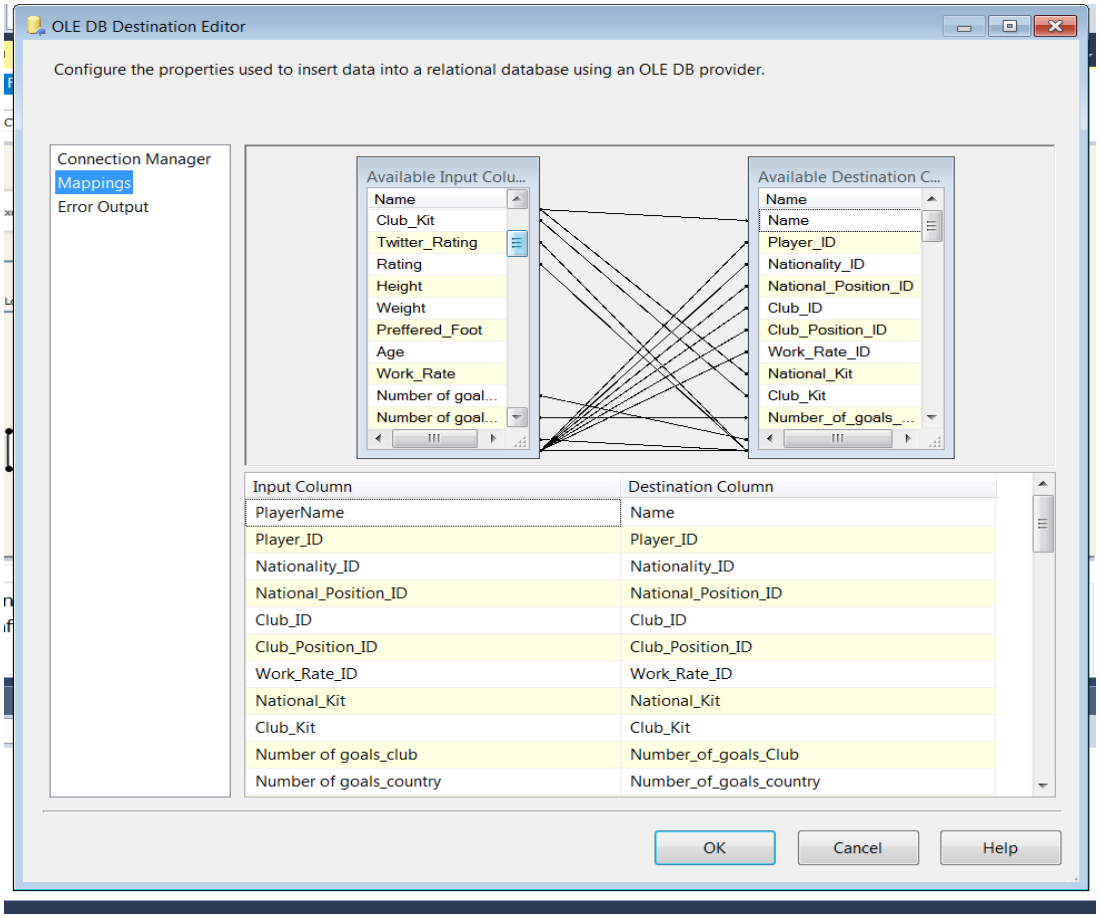


6.)Every lookup is a dimension the facts from the fact table are connected using the look transformation editor and check that the input columns are mapped into the lookup columns so when we map the next lookup the first lookup column id will be seen in the input column.



LOOKUP MAPPING

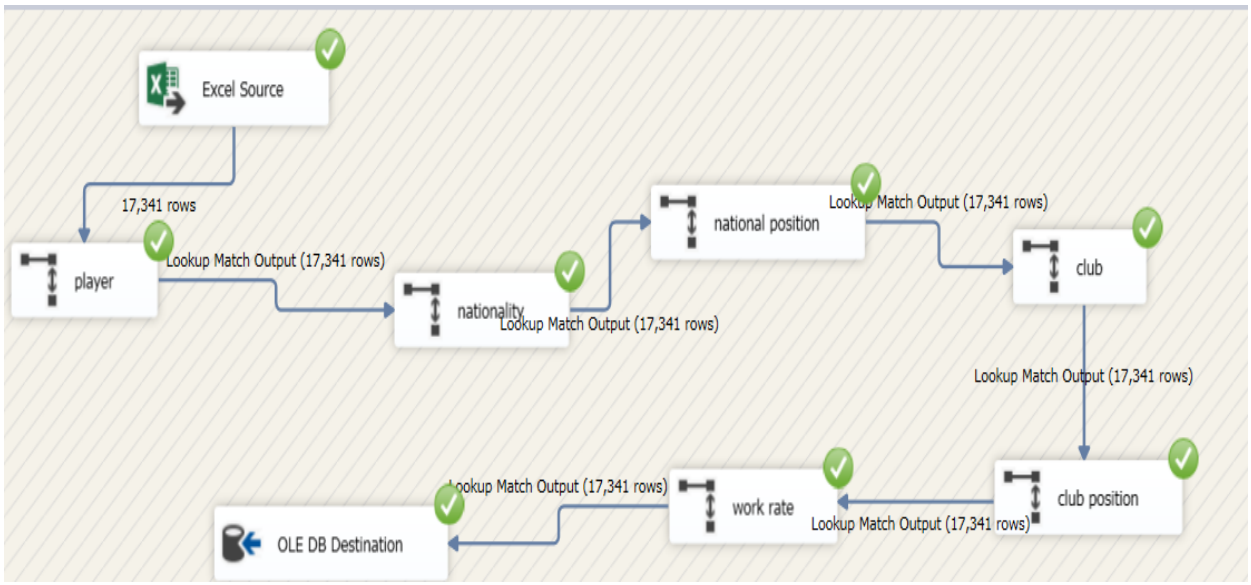
7.) After mapping all lookups they are connected to the OLEDB destination to the fact table so that the input columns will be directly mapped to the fact tables if some measures are missed map yourself to the equivalent and same measures.



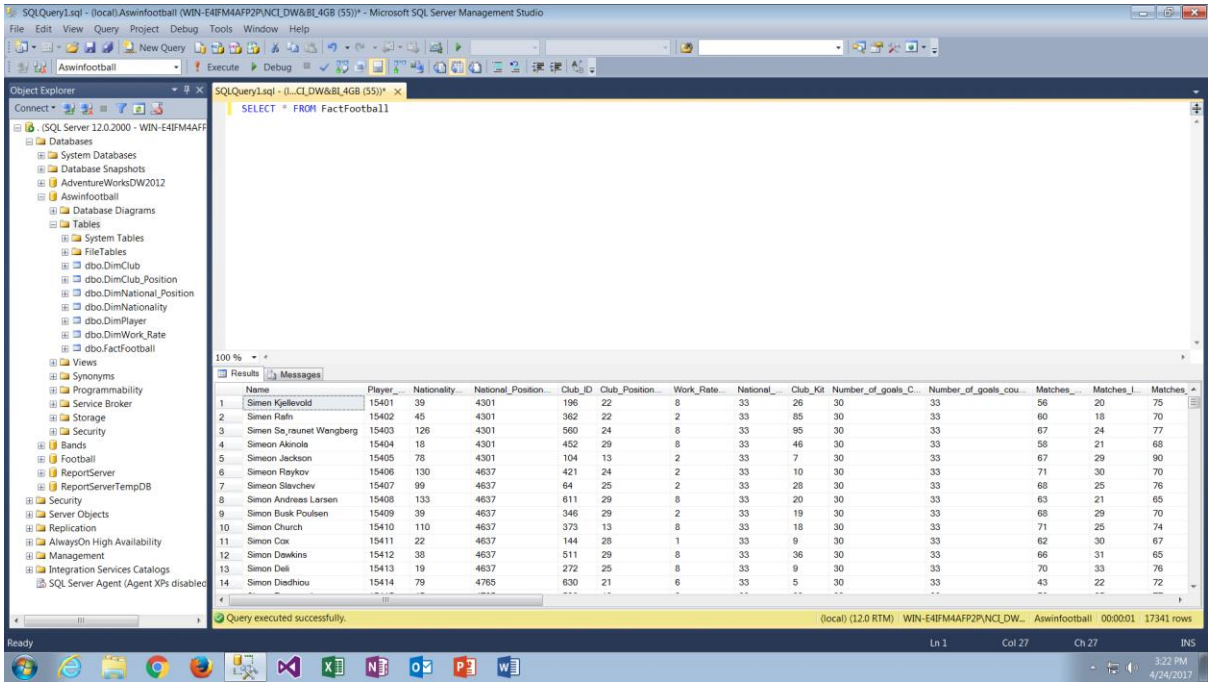
INPUT COLUMNS AND DESTINATION MAPPING

The figure represents the mapping of available input columns to the destination columns.

8.)Execute the process



9.)Once the executing process is successful check the SQL server management studio by using the query SELECT * FROM FactFootball.



DISPLAYING FACT TABLE

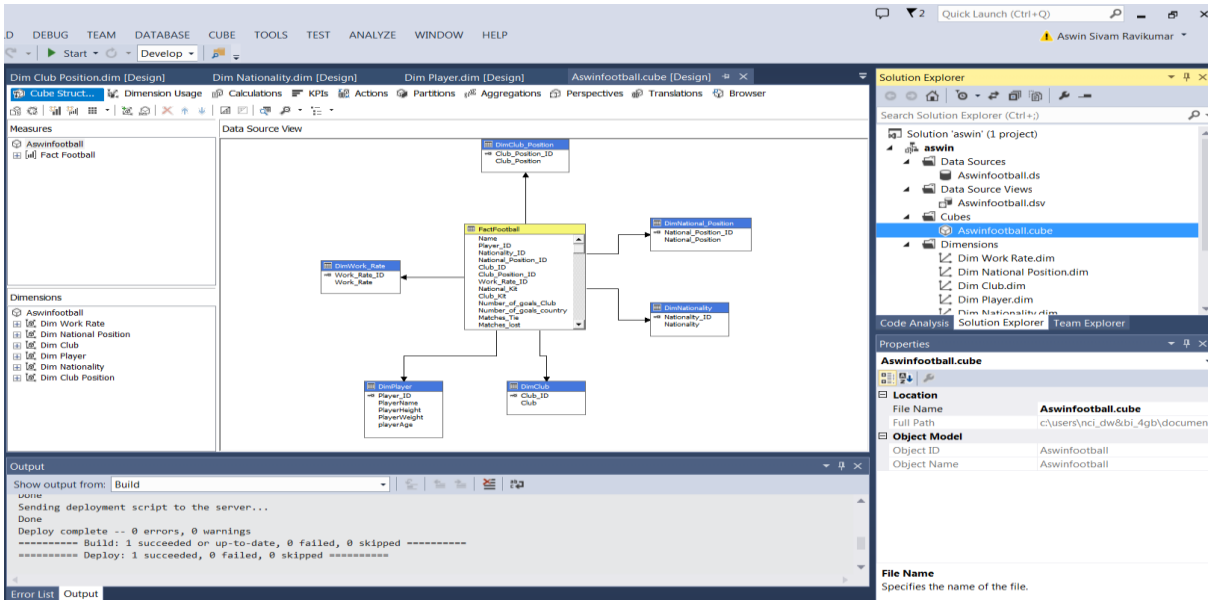
CUBE DEPLOYMENT PROCESS

1.)Open the visual studio 2013 create a new project select the analysis for business intelligence.

2.)Select the data source to connect to the local host database football the data source will be created and it will connect to the dimensions and the facts.Data source view is created using the data source football.

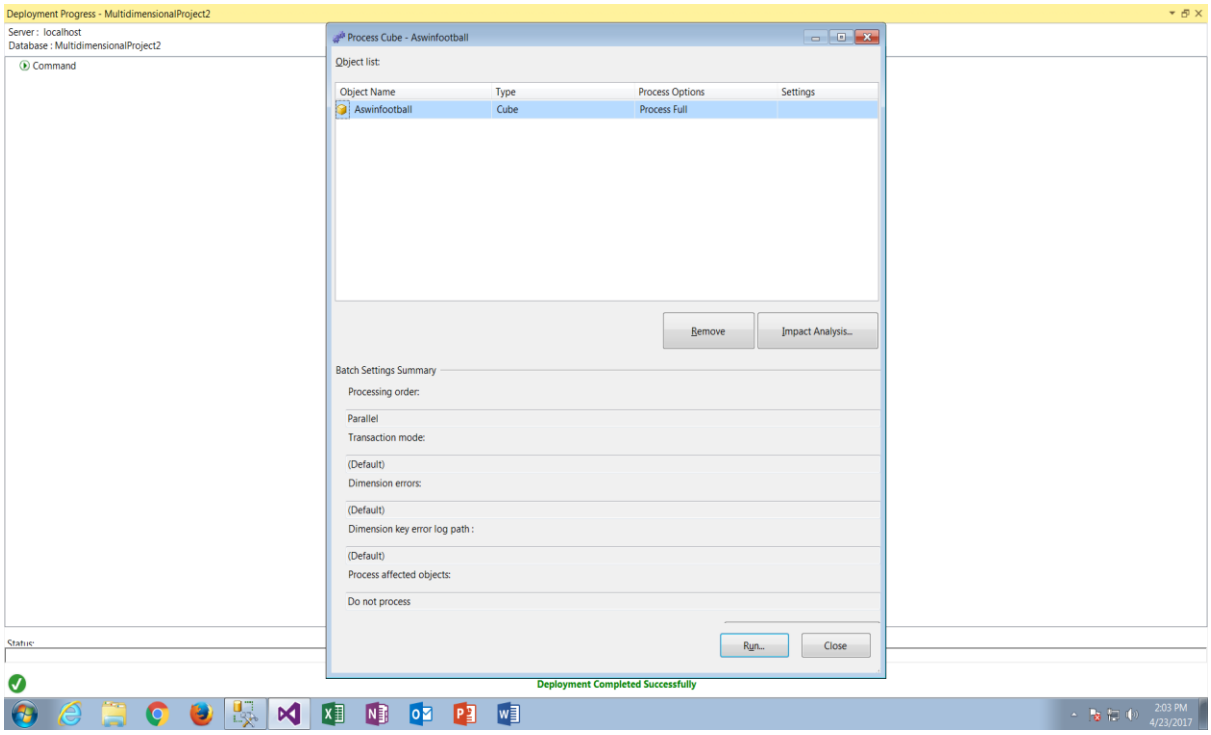
3.)create a cube a dialog box appears to drag the facts and dimensions and click next there will be a suggest button mark it the cube will be created.

5.)Automatically dimensions will be created drag them from data source view to the measures group and deploy the cube. **GENERAL VIEW OF CUBE**

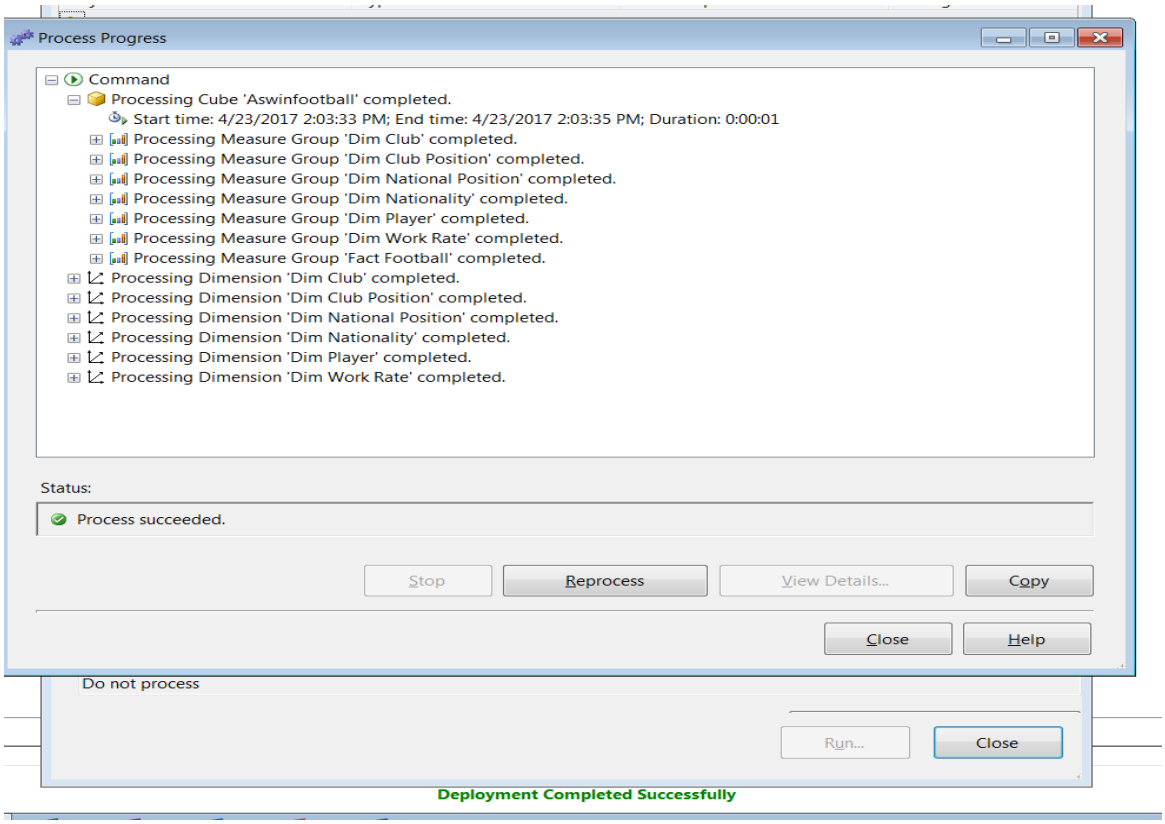


6.)Dechecckploy the cube and check whether the deployment succeeds.

DEPLOYMENT



7.)Run the process. **PROCESS**

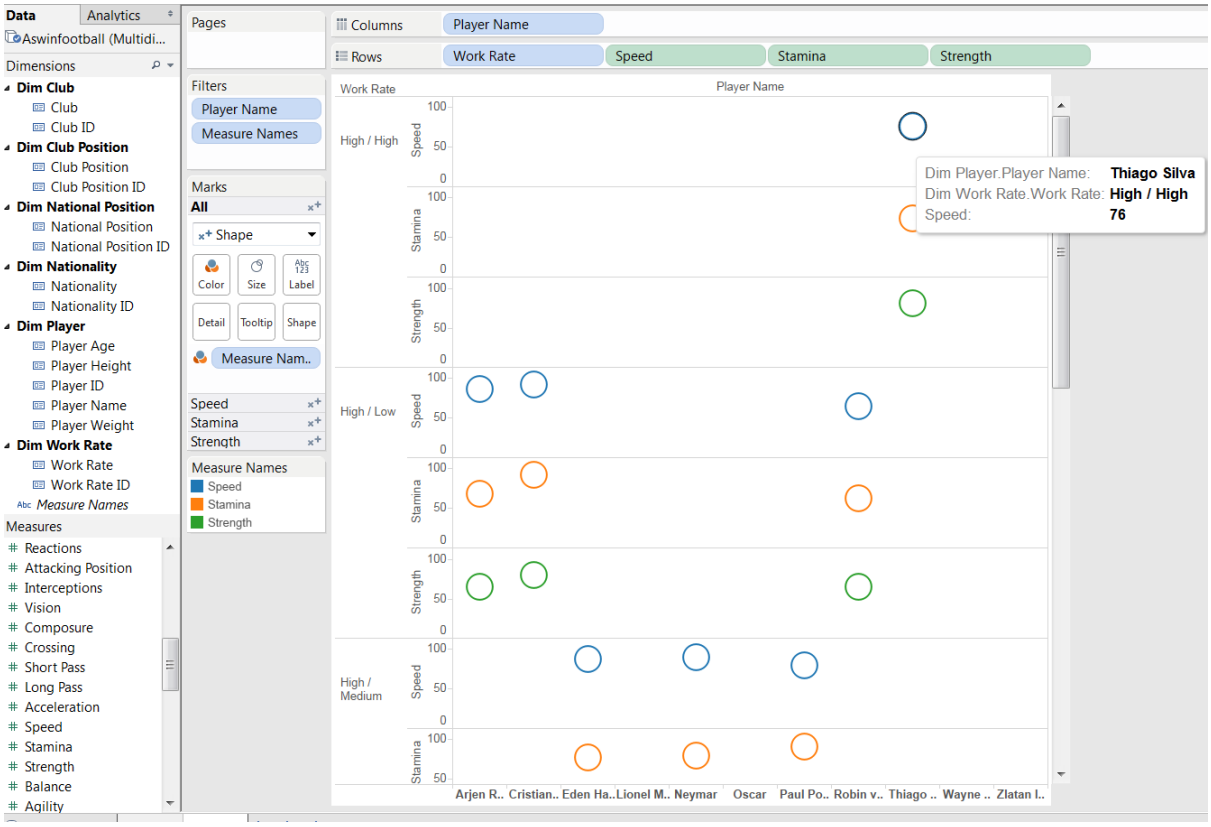


BUSSINESS INTELLIGENCE

DOES ANALYTICS CHIPS REQURIMENT IS NECESSARY

1.)Will analytics chips need to be used by all players? Will it be helpful for the clubs and nationalities are they are wasting the money chips?

Most of the teams are using the analytics chips to get acceleration, speed, and their movement they are wasting the huge amounts on them by using it for every player. Is it necessary to use for every player?



FINDINGS

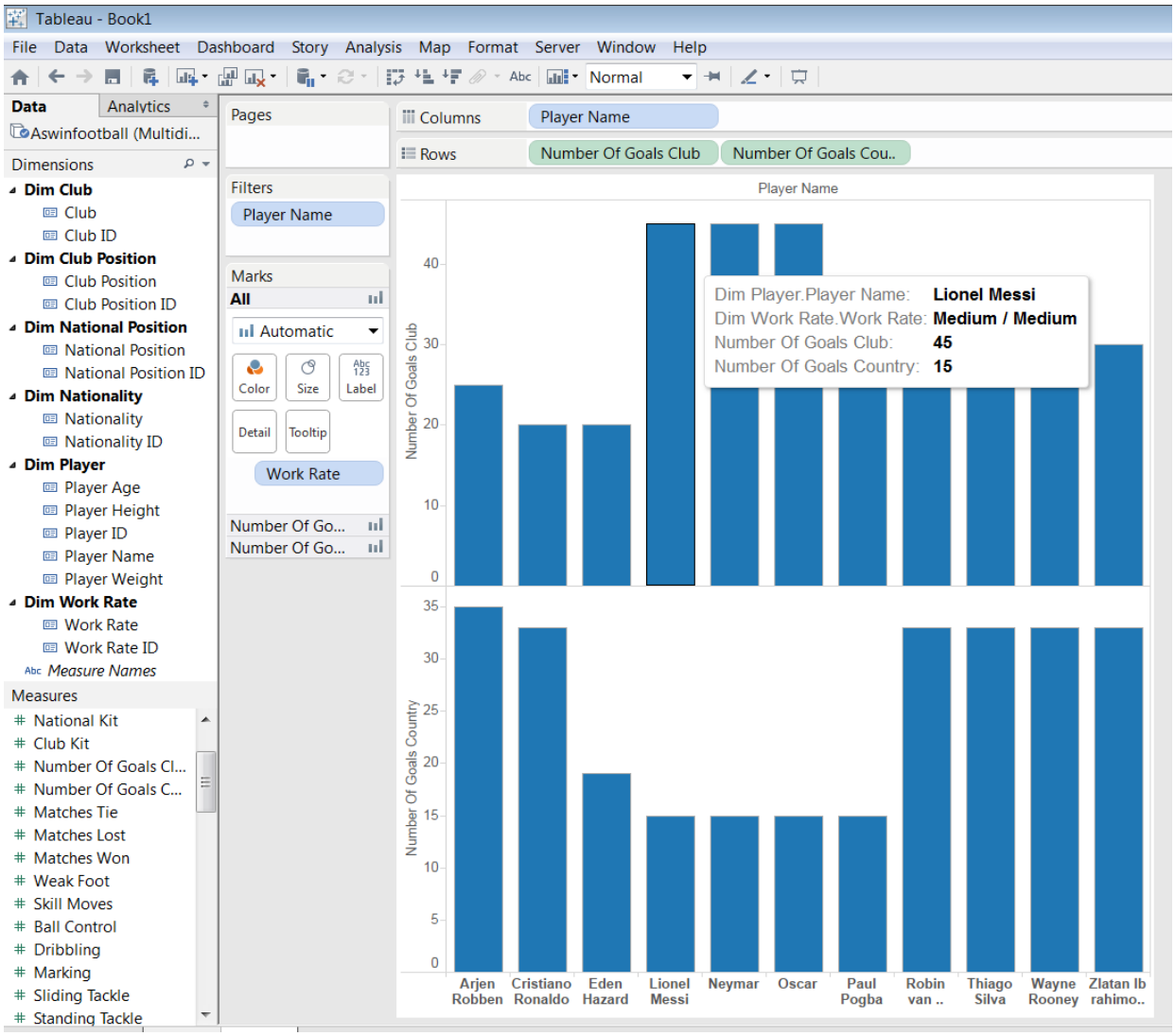
1.) I have compared the players with their work rate, speed, stamina, and strength these three facts are some basic requirements for every player by comparing it after analyzing some of the players like Thiago silva, Ronaldo, Messi, Wayne Rooney they have the high rate with the consistency so these types of players are not in the need of analytics chip.

2.) So the clubs doesn't want to invest more money on some players so it would help to reduce the business cost.

BEST PLAYER

2.)Who is the best player and who will have more money offer in the next auction?

Usually, the players who performed well will have the more money offer in the next auction so I have compared the players with their national goals and their club goals.



FINDINGS

1.) Players like lionel Messi, Neymar, Ronaldo, even hazard are performing well so they will be offered more.

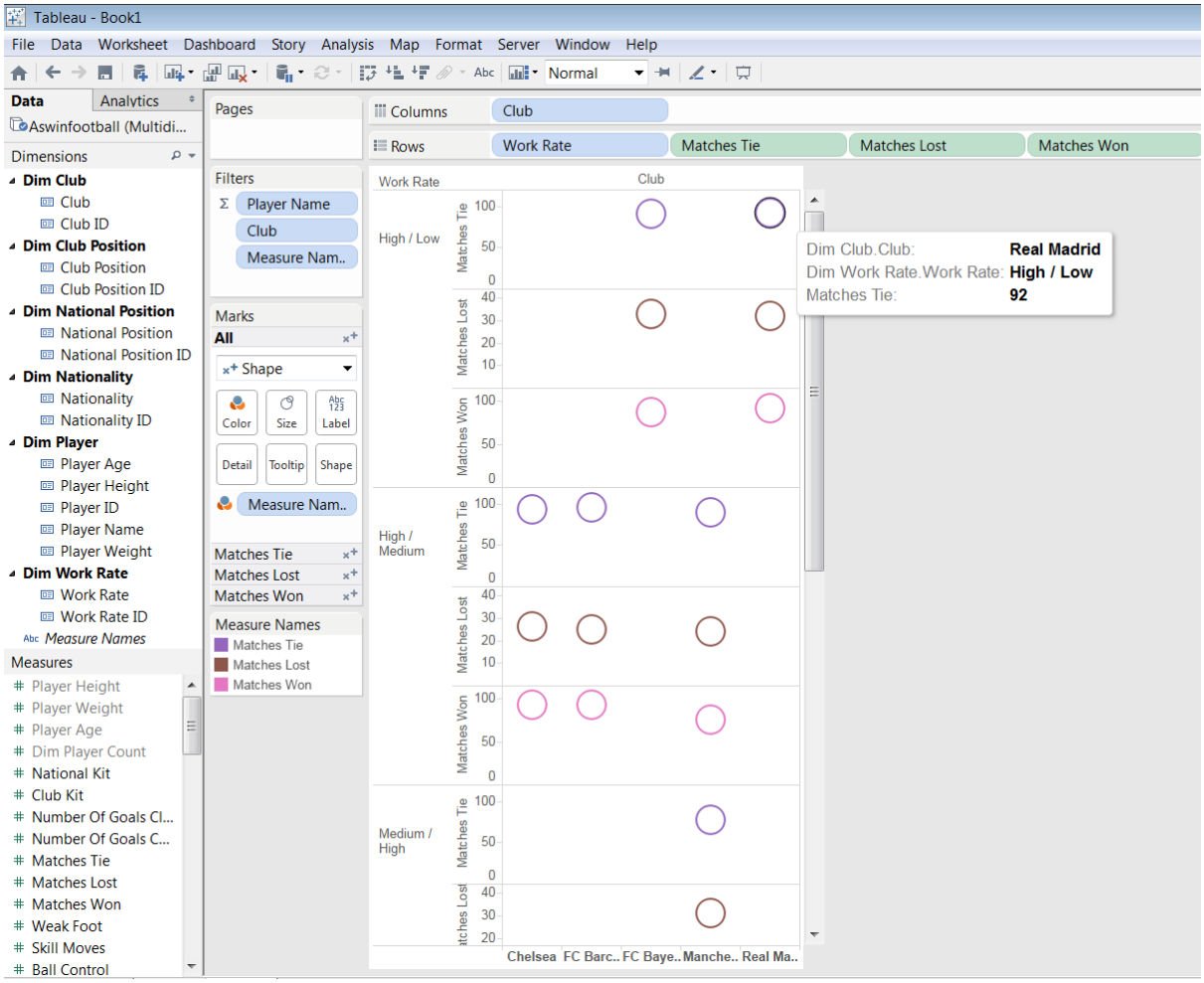
2.) If the club buys these types players the winning percentage will be more them if the clubs win more business interaction will be automatically increased.

3.) If the business companies used these type of players in advertisement there is a chance of firm growth.

BEST CLUB-SPONSORSHIP

FOOTBALL IS MOSTLY USED FOR ADVERTISEMENT PURPOSES EVERY CLUB AND NATIONALITY WILL HAVE SPONSORS FOR THEIR BUSINESS GROWTH.WHICH CLUB IS THE BEST?

I have compared the clubs and their win, loss and tie percentage by comparing and analyzing them we can have the clear view that which is the best club



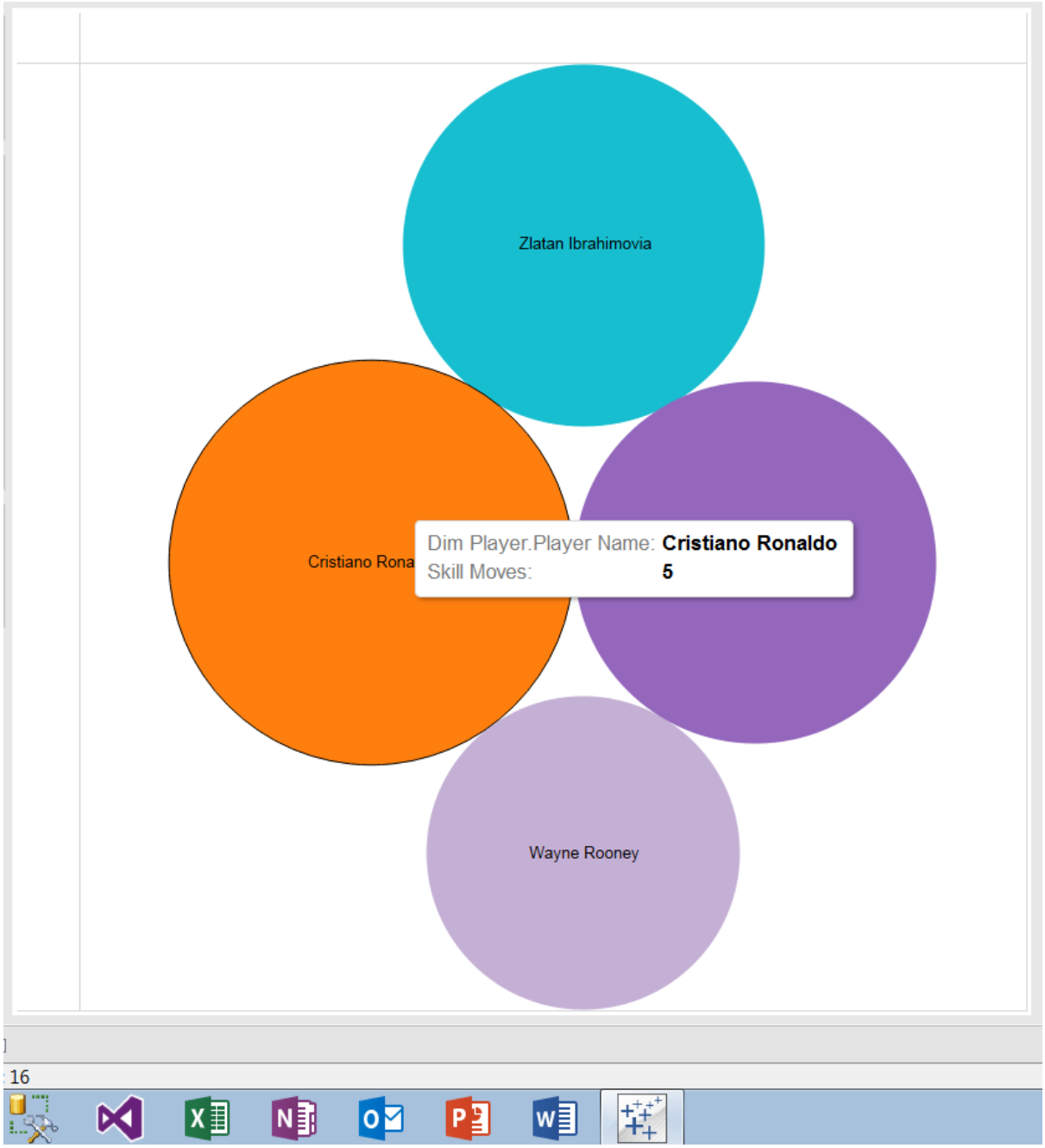
FINDINGS

1.) Clubs like real Madrid, manchester united have the more win percentage so they will have the more sponsors.

2.) If the business and corporate companies target these types of clubs and if they sponsor them there will more chance of an increase in profit.

BEST SKILLED PLAYER- BUSINESS ADVETISTEMENTS

Who will be the best-skilled player



I have compared the top four players with their skills list

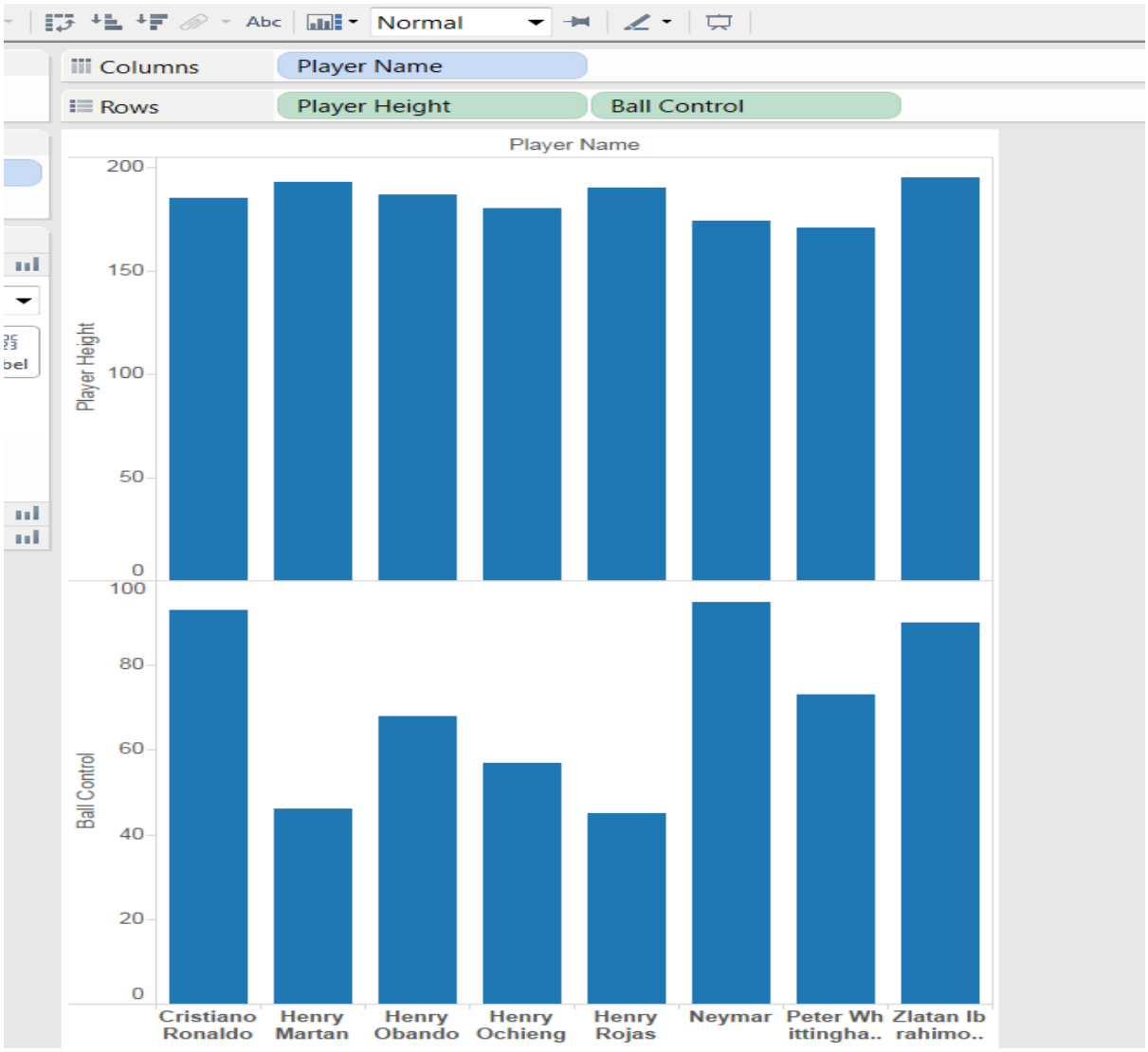
FINDINGS

1.) Comparing these players Christiano Ronaldo tends to have the more skills, so there is a chance of having more fans.if business companies use him as a brand ambassador there is a possibility of companies growth.

2.) Usually, football players will have the more fans when compared to the other games so the product advertised by this kind of people will reach around the world easily.

PLAYERS HEIGHT VS SKILLS

I have compared the player's height with their skills is there is a possibility of tall players having more skills.



FINDINGS

1.) According to the analysis graph, the taller people have the more skills when compared to the short players

REFERENCE

1.)Kimball, R,2013. The data warehouse toolkit.3rd ed. Indiana: John wiley&sons.

2.)data sets -kaggle, Wikipedia, twitter.

3.) Matloff, N., 2011. The art of R programming.

4.)Mockraoo,sql random data generator.