National College of Ireland

Data Warehousing and Business Intelligence Project

on

The FIFA world cup

Martin Mohan
x18191339

MSc/PGDip Data Analytics – 2019

Submitted to: Dr. Pierpaolo Dondio

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Martin Mohan
Student ID:	x18191339
Programme:	MSc Data Analytics
Year:	2019
Module:	Data Warehousing and Business Intelligence
Lecturer:	Dr. Pierpaolo Dondio
Submission Due Date:	26/11/2018
Project Title:	The FIFA world cup

I hereby certify that the information contained in this (my submission) is information pertaining to my own individual work that I conducted for this project. All information other than my own contribution is fully and appropriately referenced and listed in the relevant bibliography section. I assert that I have not referred to any work(s) other than those listed. I also include my TurnItIn report with this submission.

ALL materials used must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is an act of plagiarism and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	June 22, 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Table 1: Mark sheet – do not edit

Criteria	Mark Awarded	Comment(s)
Objectives	of 5	
Related Work	of 10	
Data	of 25	
ETL	of 20	
Application	of 30	
Video	of 10	
Presentation	of 10	
Total	of 100	

Project Check List

This section capture the core requirements that the project entails represented as a check list for convenience.

- ☒ Used LaTeX template
- ☐ Three Business Requirements listed in introduction
- ☐ At least one structured data source
- ☐ At least one unstructured data source
- ☐ At least three sources of data
- ☐ Described all sources of data
- ☐ All sources of data are less than one year old, i.e. released after 17/09/2017
- ☐ Inserted and discussed star schema
- ☐ Completed logical data map
- ☐ Discussed the high level ETL strategy
- ☐ Provided 3 BI queries
- ☐ Detailed the sources of data used in each query
- ☐ Discussed the implications of results in each query
- ☐ Reviewed at least 5-10 appropriate papers on topic of your DWBI project

The FIFA world cup

Martin Mohan
x18191339

June 22, 2019

Abstract

The men's soccer World Cup FIFA (2019*a*) is a global football competition contested by the various football-playing nations of the world. It is contested every four years and is the most prestigious and important trophy in the sport of football and the most watched event in the world. It is followed only by the european cup also held every four years but never in the same year as the world cup.
Are the FIFA ranking for teams accurate. Do red cards have an affect on the outcome of international games.This report looks at some statistics based on data collected from the world and european cups.

1 Introduction

Association football is probably the most popular sport in the world as well as a huge industry with millions at stake. Efficacy and robustness of the official ranking system employed by FIFA is of critical importance for fair competition of the involved parties.In general, higher ranked teams are paired with lower ranked ones. In this way, rankings have crucial impact on the competition. The teams at the top of the ranking are less likely to face other strong opponents. This is advantageous as it helps to avoid potential elimination in early stages of the competition Lasek (2016).
Since its introduction in 1992, the FIFA world rankings have been the subject of much debate, particularly regarding the calculation procedure and the resulting disparity between generally perceived quality and world ranking of some teams Paul & Mitra (2008).
Lasek (2016) was critical of the FIFA ranking (which are based on the Elo system for chess ranking) and even suggested methods to game the system because home games are not taken into consideration and friendly games are counted in the ratings.
The Paul & Mitra (2008) report looks at the actual results vs the predicted FIFA rankings but only three world cups had been staged at that time, since the introduction of ranking in 1993. Seven world cups have now been staged since the introduction of the rankings and a comparison of ranking vs actual result will be carried out (Req-1).
Paul & Mitra (2008) defend the FIFA rating system despite the fact that the last world cup before publication was won by Italy (seeded 12) in 2006 7.1.1. They quote statistics like goals scored and red cards e.g. the teams which saw more red cards in 1998 FIFA worldcup were more likely to win the game". A comparison of the results of games aginst red cards awarded will also be examined using data from the world cup and the european cup. (Req-2) (Req-3). The world cup winners are dominated by South America and europe. In fact no country outside these two continents have ever finished in first

or second place wikipedia (2019) as illustrated in 1. To increase the data available for comparison of red cards the european cup winner data (which is held between world cups) was also collected This was used in (Req-2).

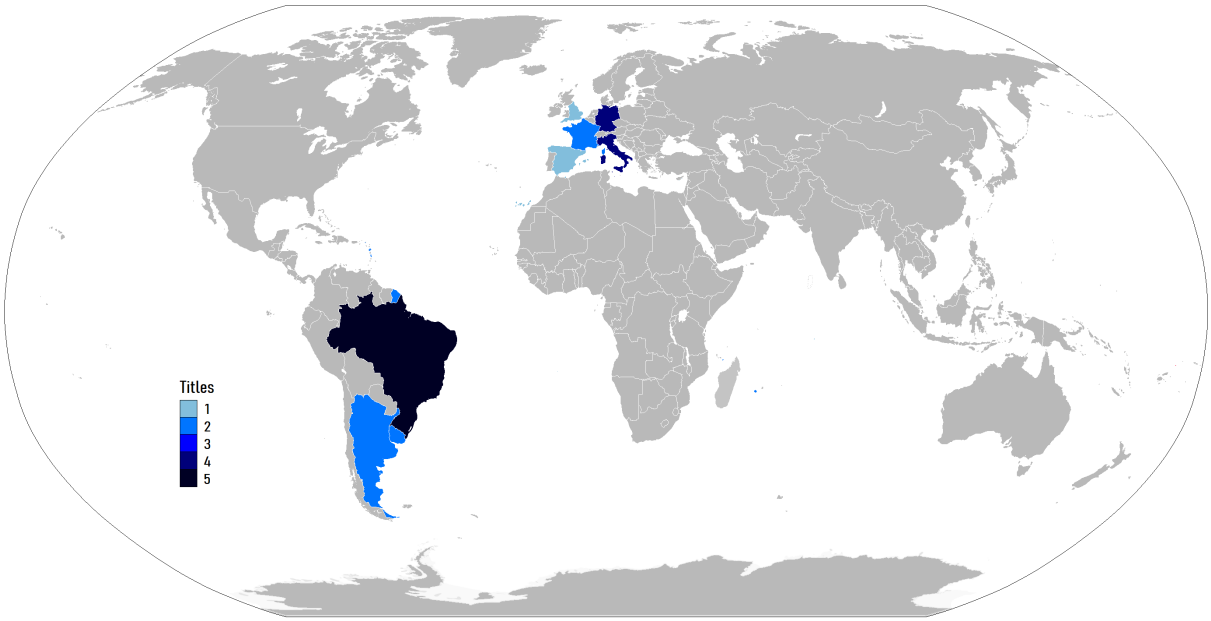


Figure 1: Map of world cup winners

The queries to be addressed are ...

(Req-1) How accurately does FIFA world ranking reflect the actual world cup results.

(Req-2) Is there any correlation between the number of redcards accumulated by a team in a competition and their success.

(Req-3) Which teams have accumulated the most red cards in world cup history

Source	Type	BI query	Brief Summary
Kaggle	Structured	1,2,3	Results from FIFA World Cups and rankings
Statista	Structured	1,2,3	FIFA World Cup World cup data missing from Kaggle (2018)
UN database	Structured	-	World population data GDP,GNI
github	Structured	1,2,3	Fifa country codes
wikipedia	Unstructured	2,3	wiki_euro.py=euro cup champions results 1,2,3, wiki_eurorc.py=euro cup redcards, wiki_redcards.py=world cup red cards

Table 2: Summary of sources of data used in the project

2 Data Sources

The data used in this report were downloaded from the websites shown in figure 2. Table 2 lists the data sources and the queries in which the data was used. The code fcode.py 9 and leftjoin.R 9 were used to clean and merge all the files into 3 CSV files countryDIM.csv, competitionDIM.csv and rankFACT.csv which were used in the DIM and FACT files respectively. A breakdown of the data sources follows.

2.1 Kaggle

Data sets from the FIFA World Cups from 1930 to 2014. `https://www.kaggle.com/abecklas/fifa-world-cup` This data contained several csv files with data on world cups from 1930 to 2014 including year,winner,second,third and fourth place. This was put in file competition.csv
FIFA International Men's Ranking were listed from August 1993 until June 2018. FIFA carries out ranking several times during year but the csv file was modified to take only one set of ranking at the start of the year (usually taken in January or February). This simplified the star schema as no date dimension was needed. This was put in rankFACT.csv.

2.2 Statista

The world cup data from Kaggle was missing data for 2018 so I got this from Statista. Several databases were needed goals scored FIFA (2019*b*) ,attendance (sport.de) and ranking FIFA (2019*d*).

2.3 UN database

Information on data.un.org on countries GDP,GNI and population. The latest value for 2017 was downloaded.. The UN database provied information on Country, Population, GDP Some countries in UN database were not in world cup rankings. E.g. UN measures GDP for Great Britain. We needed to add GDP for England, Ireland, Scotland and Wales separately using google search.
This data was foreseen to be used in a query on world cup winners but this data was later found on the internet FIFA (2019*c*) 1
The data was put in countryDIM.csv

Program	List scraped	url
wiki_euro.py	Euro cup champions results 1,2,3,4	https://en.wikipedia.org/wiki/List_of_FIFA_World_Cup_red_cards
wiki_eurorc.py	European cup redcards	https://en.wikipedia.org/wiki/List_of_UEFA_European_Championship_red_cards
wiki_worldrc.py	World cup red cards	https://en.wikipedia.org/wiki/List_of_FIFA_World_Cup_red_cards

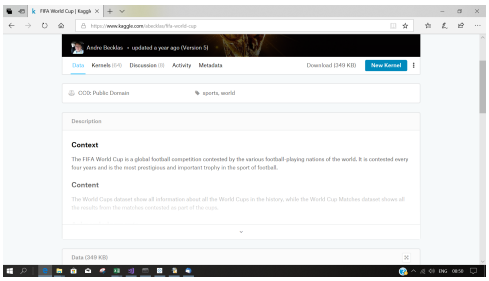
Table 3: web scraping programs

2.4 github

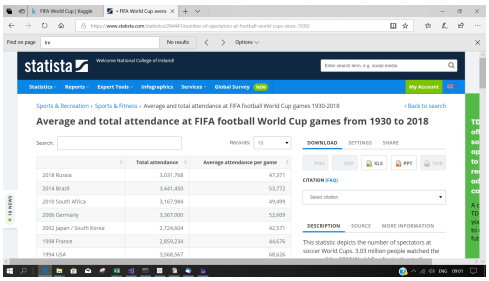
Country names on all sites was slightly different. In order to assign standardised fifa codes to different files it was necessary to download standardised FIFA codes and use these a reference. These codes were downloaded into a file fifa_codes.csv from https://github.com/openmundi/world.csv
A crude but effective python code fcode.py 9 was created to compare country names in a file against standardised FIFA country names.This code then generated a report showing the differences and suggesting solutions.
For red cards some historical data was changed for simplicity. Redcards for Yugoslavia and Serbia and Montenegro were assigned to Serbia. Redcards for West Germany and East Germany were assigned to Germany. Redcards for the soviet Union were assigned to Russia. The selection of lookups could be made more sophisticated in future. The data was merged into file countryDIM.csv.

2.5 wikipedia

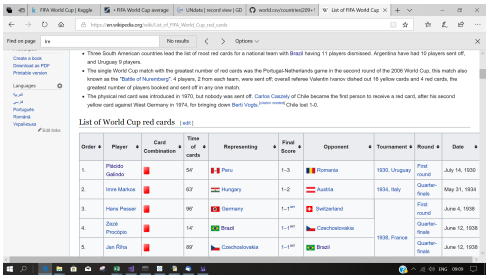
Several python programs were written to scrape unstructured data from wikipedia and are listed in table 2.5 . The programs are very similar so only one program example was put in the appendix wiki_redcards.py 9.
Redcard data and rank data was cleaned and merged into the file rankFACT.csv. The list of european champion results downloaded using wiki_euro.py was put into the file competitionDIM.csv.



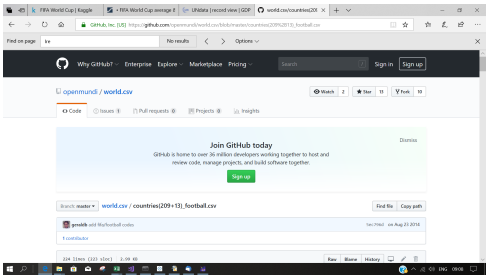
(a)



(b)

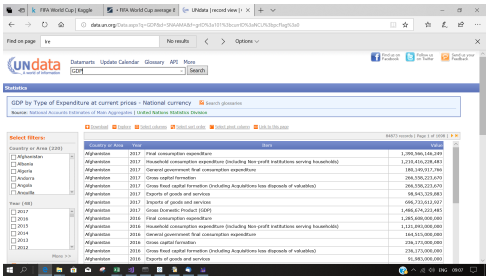


(c)



(d)

(e)



(f)

Figure 2: websites used as data sources

3 Related Work

Much has been written on the World Cup, which has become the most watched sports event in the world, with viewers numbering in the billions, surpassing the Olympic Games Deena (2018).
Much of the world cup literature concentrates on the economic impact of the world cup Baade & Matheson (2004) although some also mention that despite becoming "big business in a globalized world" it is passion which drives it e.g. "In the Biafra-Nigeria civil war, at one point in 1967, fighting was halted so the combatants could watch the Brazilian star Pelé." Miller (2010)
This report mainly concentrates on the accuracy of the world cup ranking and delves deeper into some of the statistics. The report refers to Lasek (2016) which indicates that world cup rankings are inaccurate. This report also further investigates if red cards have an influence on the winner in follow up to paper Paul & Mitra (2008) which suggests there may be some correlation.
The report shows it is difficult to predict a winner. This makes a paper based on another world sport held every 4 years interesting (cricket). The paper Abdurazzag et al. (2018) says in the introduction "This research work aims to predict the winner of the 12th version of ICC world cup using Business Intelligent (BI) and K Nearest Neighbors KNN bigdata approach". They conclude either England or India teams will win.

4 Data Model

The star schema 3 has 2 dimensions and one FACT table. The table 4 shows how attributes in the star schema map to each of the three BI queries ((Req-1),(Req-2), or (Req-3)) The FACT table contains the grain data rank,redcards and the two dimension tables contain country and competition data. Data was modified using python programs which assigned fifa codes and merged tables into 3 csv countryDIM.csv,competitionDIM.csv and rankFACT.csv. The relationship of each BI query to webiste is shown in table 2 and bI query to star map is in table 4. A detailed logical data map for each column is shown in table4

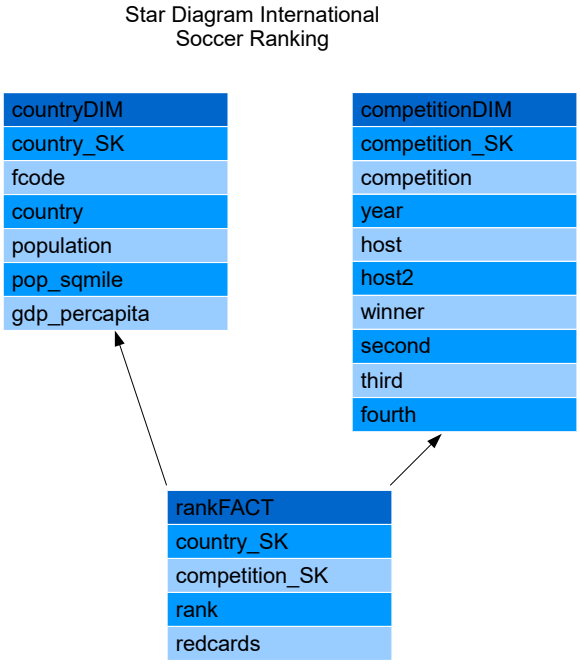


Figure 3: Star schema

rankFACT (BI)	competitionDIM (BI)	countryDIM (BI)
rank (1)	competition (1,2)	fcode (1 2 3)
redcards (2,3)	year (1 2)	country (1 2)
	host	population
	host2	pop_sqmile
	winner (1 2 3)	gdp_percapita
	second (1 2)	
	third (1 2)	
	fourth (1 2)	

Table 4: Star schema and which BI query uses it ((Req-1),(Req-2), or (Req-3)) see also table 2 column BI query

5 Logical Data Map

This section describes the logical data map, i.e. This source numbers refer to the original data source kaggle, statista etc ...

Table 5: Logical Data Map describing all transformations, sources and destinations for all components of the data model illustrated in Figure 3

Source 2	Column	Destination	Column	Type	Transformation
none	none	countryDIM	country_sk	integer	primary key $
4	fcode	countryDIM	fcode	varchar(100)	Use fcode.py to convert
1,4	country	countryDIM	country	varchar(100)	Use fcode.py to convert
3	population	countryDIM	population	integer	Use fcode.py to convert. Join using R code merge
3	pop_sqmile	countryDIM	pop_sqmile	integer	Use fcode.py to convert. Join using R code merge
3	gdp_percapita	countryDIM	gdp_percapita	integer	Use fcode.py to convert. Join using R code merge
none	none	competitionDIM	competition_sk	integer	primary key
1,2(wc), 4(euro)	competition	competitionDIM	competition	integer	Use fcode.py to convert
1,2,4	host	competitionDIM	host	varchar(100)	fcode.py convert country to fifa code
1,2,4	host2	competitionDIM	host2	varchar(100))	fcode.py convert country to fifa code
1,2,4	year	competitionDIM	year	varchar(100))	fcode.py convert country to fifa code
1,2,4	winner	competitionDIM	winner	varchar(100))	fcode.py convert country to fifa code
1,2,4	second	competitionDIM	second	varchar(100))	fcode.py convert country to fifa code
1,2,4	third	competitionDIM	third	varchar(100))	fcode.py convert country to fifa code
1,2,4	fourth	competitionDIM	fourth	varchar(100))	fcode.py convert country to fifa code
countryDIM	country_sk	rankFACT	country_sk	integer	surrogage key points to primary key
competitionDIM	competition_sk	rankFACT	competition_sk	integer	surrogage key points to primary key

Continued on next page

Table 5 – *Continued from previous page*

Source	**Column**	**Destination**	**Column**	**Type**	**Transformation**
1,2	rank	rankFACT	rank	integer	fcode.py and merge using leftjoin.R
4	redcards	rankFACT	redcards	integer	fcode.py and merge using leftjoin.R

6 ETL Process

All data was cleaned and merged before it was input to SSIS using the model shown in figure 4 After several attempts at SSIS it proved more effective to write python or R programs to manipulate data downloaded from the web. The files were modified and then one csv file was created for each DIMENSION and FACT file namely countryDiM.csv,competitionDIM.cs and rankFACT.csv

The most used program was fcode.py 9

This program does various checks. e.g. A test file ctest.csv 5 was used to test the program (see video attached). After running fcode.py against ctest.csv a new file called out_test.csv was created. The example output 5 reports anomalies such as "United States of America" and "Polonia" are unknown (they should be "United States" and Poland).The user correct the anomalies and re-runs the program until no more anomalies are found. Another R code was run from R commander to merge files leftjoin.R 9. After creating proper fifa codes and fifa names and then merging files 3 csv files were created. The files competitionDIM.csv, countryDIM.csv and rankFACT.csv correspond to the dimensions and FACT table used by SSIS see 6

The R and python codes produced are crude, however they proved effective. It raises the question of whether better tools for ETL analysis can be created.

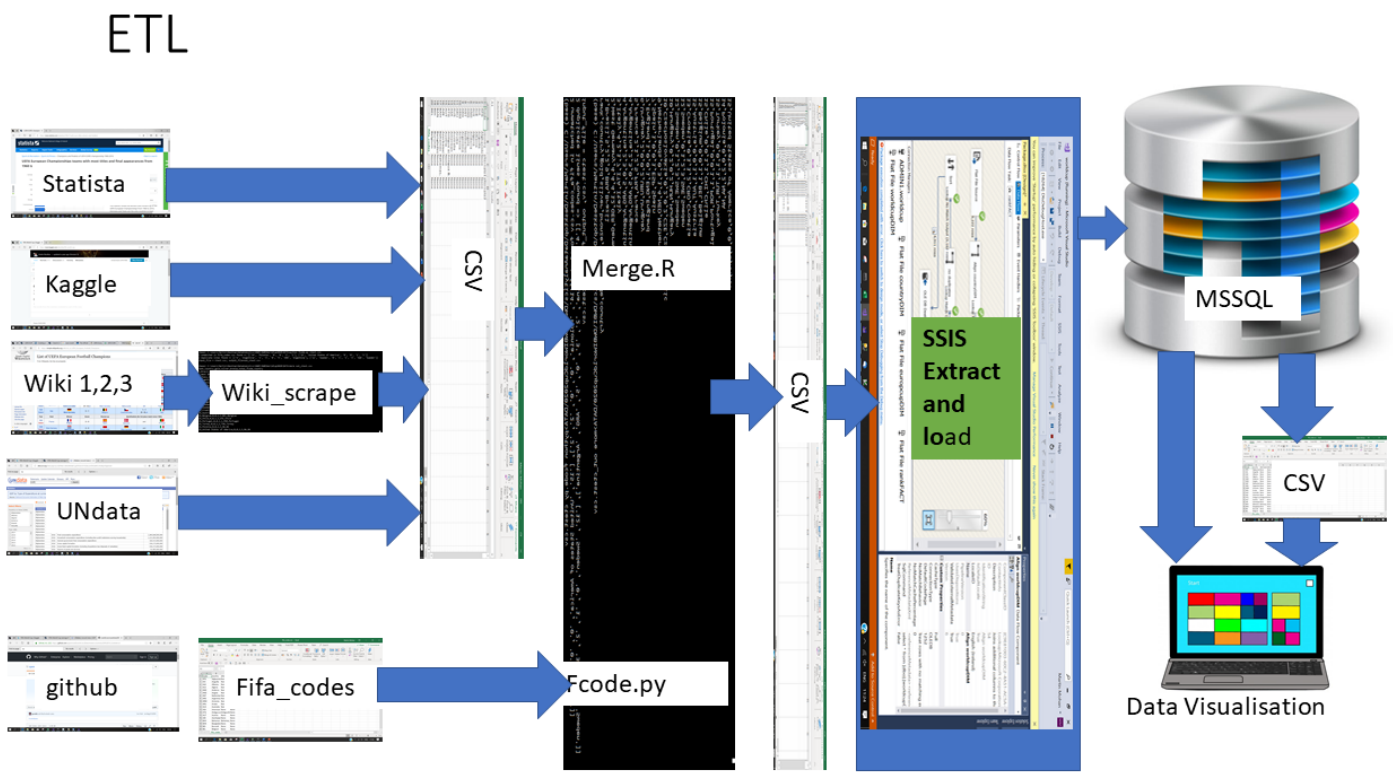
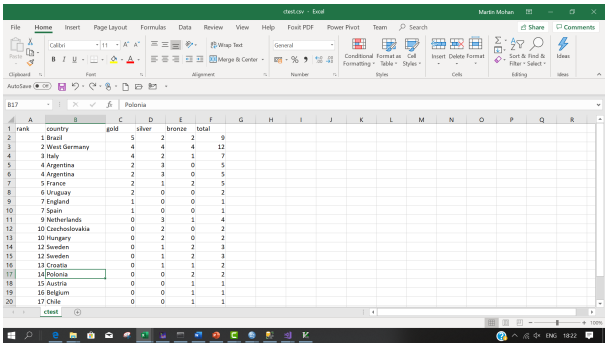
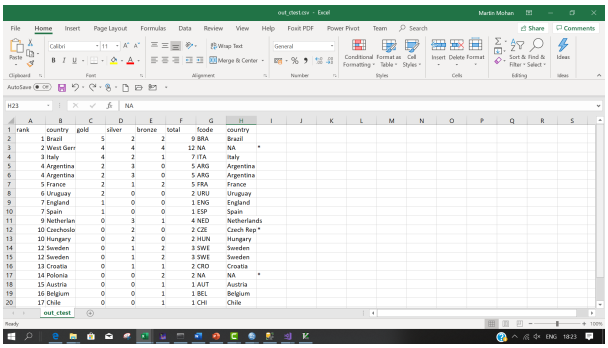


Figure 4: Overview of ETL process



rank	country	gold	silver	bronze	total
1	Brazil	5	2	2	9
2	West Germany	4	4	4	12
3	Italy	4	2	1	7
4	Argentina	2	3	0	5
4	Argentina	2	3	0	5
5	France	2	1	2	5
6	Uruguay	2	0	0	2
7	England	1	0	0	1
7	Spain	1	0	0	1
9	Netherlands	0	3	1	4
10	Czechoslovakia	0	2	0	2
10	Hungary	0	2	0	2
12	Sweden	0	1	2	3
12	Sweden	0	1	2	3
13	Croatia	0	1	1	2
14	Polonia	0	0	2	2
15	Austria	0	0	1	1
16	Belgium	0	0	1	1
17	Chile	0	0	1	1

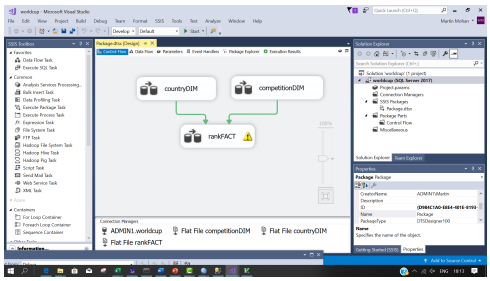
(a)



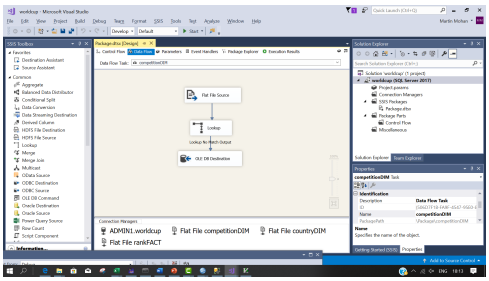
rank	country	gold	silver	bronze	total	fcode	country
1	Brazil	5	2	2	9	BRA	Brazil
2	West Germ	4	4	4	12	NA	NA
3	Italy	4	2	1	7	ITA	Italy
4	Argentina	2	3	0	5	ARG	Argentina
4	Argentina	2	3	0	5	ARG	Argentina
5	France	2	1	2	5	FRA	France
6	Uruguay	2	0	0	2	URU	Uruguay
7	England	1	0	0	1	ENG	England
7	Spain	1	0	0	1	ESP	Spain
9	Netherlan	0	3	1	4	NED	Netherlands
10	Czechoslo	0	2	0	2	CZE	Czech Rep*
10	Hungary	0	2	0	2	HUN	Hungary
12	Sweden	0	1	2	3	SWE	Sweden
12	Sweden	0	1	2	3	SWE	Sweden
13	Croatia	0	1	1	2	CRO	Croatia
14	Polonia	0	0	2	2	NA	NA
15	Austria	0	0	1	1	AUT	Austria
16	Belgium	0	0	1	1	BEL	Belgium
17	Chile	0	0	1	1	CHI	Chile

(b)

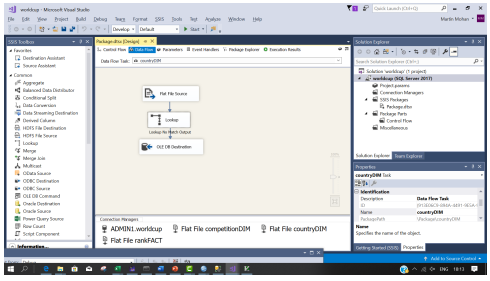
Figure 5: fcode.py ctest -> out_ctest.csv (2 extra columns with fifa code + proper fifa country)



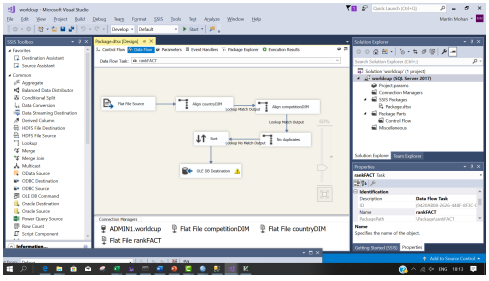
(a)



(b)



(c)



(d)

Figure 6: SSIS views

7 Application

The following 3 subsections show how the 3 questions (Req-1), (Req-2) and (Req-3) raised in Section 1 were addressed.

7.1 BI Query 1: How accurately does FIFA world ranking reflect the actual world cup results.(Req-1)

For this query, the contributing sources of data are marked as 1 in the tables 2 and 4
To compare rank we compared the FIFA rankings taken at the start of the world cup year against the actual result of the world cup for the winning 4 teams. An sql query shown in 7.1.1 was run to obtain the 1st,2nd,3rd and 4th places in the world cup. The values are are summarised in table 7.1.1 and Figure 7. The sql query data was analysed directly from mssql database using tableau 7. The result of the sql query was copied to a csv file and the data was analysed using R plots 8 and IBM SPSS 9.
The normal distribution 8 (a) shows a positive skew (skew=1.59)9 (b) and the qq plot is not quite in line 8 (b) but a kolmogrov-smirnoff test of normality 9 (c) was significant for rank p>.05 (but not result p<.05 which has only 4 values)
Tests were also made for outliers. South Korea was ranked as 42 but they came fourth in the 2002 worldcup. This could be partly explained that South Korea hosted the world cup that year (along with Japan) giving home advantage, which is not considered when FIFA rank the countries Lasek (2016) Cooks distance was calculated using IBM SPSS to test whether South Korea could be considered an outlier. Cooks distance was calculated as .26 which is below 1 so it was not removed. Field (2009) page 269
The scatter plot in 8 (c) shows the postive linear regression line. The 4th plot (d) is a summary of the linear regression which confirms the SPSS output 9 (d)

7.1.1 BI Query 1: Summary

Simple linear regression was carried out to assess the whether the FIFA ranking system could be used to predict the final result in a world cup competition. Preliminary analysis was carried out to insure no violation due to normality and outliers. The results were positively skewed as we only had four results winner, second, third and fourth. Despite the non normality, the results of the linear analysis 9 (c) showed rank was statistically significant p<.05. The linear regression equation is

$$Y = \beta_0 + \beta_1 rank \tag{1}$$

$$Y = 1.710 + 0.075 rank \tag{2}$$

The current data suggests that FIFA ranking systematically ranks teams 1.71 places lower than they should be for the four top places. This is evident in table 7.1.1 and 7 where it can be seen that of the 28 results in world cups since ranking began only one country was ranked higher than the expected result Brazil won in 1998 but was ranked 2. Three teams ranking were correct Brazil 1994 (1), Spain 2010 (1) and Sweden 1994 (3). The other 24 countries were ranked too low. The results are however skewed by the fact that only the top 4 results are considered

year	Country	rank	result	rank-result
1994	Brazil	1	1	0
1998	France	6	1	5
2002	Brazil	3	1	2
2006	Italy	12	1	11
2010	Spain	1	1	0
2014	Germany	2	1	1
2018	France	9	1	8
1994	Italy	4	2	2
1998	Brazil	1	2	-1
2002	Germany	12	2	10
2006	France	5	2	3
2010	Netherlands	3	2	1
2014	Argentina	3	2	1
2018	Croatia	15	2	13
1994	Sweden	3	3	0
1998	Croatia	16	3	13
2002	Turkey	23	3	20
2006	Germany	17	3	14
2010	Germany	6	3	3
2014	Netherlands	10	3	7
2018	Belgium	5	3	2
1994	Bulgaria	16	4	12
1998	Netherlands	25	4	21
2002	South Korea	42	4	38
2006	Portugal	10	4	6
2010	Uruguay	21	4	17
2014	Brazil	9	4	5
2018	England	16	4	12

Table 6: BI1.csv: Rank vs actual result in world cup result(Winner=1, Second=2,Third=3,Fourth=4) rank-result=7

SQL code to rank countries. To put in a csv file it needs to be run four times replacing Winner with second, third and fourth but in Tableau four custom queries were linked.. . .

```
use worldcup
SELECT competitionDIM.[Year] as year, countryDIM.[country] as Winner,
rankFACT.rank as rank
FROM competitionDIM, countryDIM
JOIN rankFACT ON countryDIM.country_SK = rankFACT.country_SK
Where countryDIM.fcode=competitionDIM.Winner
AND rankFACT.competition_SK=competitionDIM.competition_SK
AND rankFACT.country_SK=countryDIM.country_SK
AND competitionDIM.competition = 'WorldCup'
order by year
```

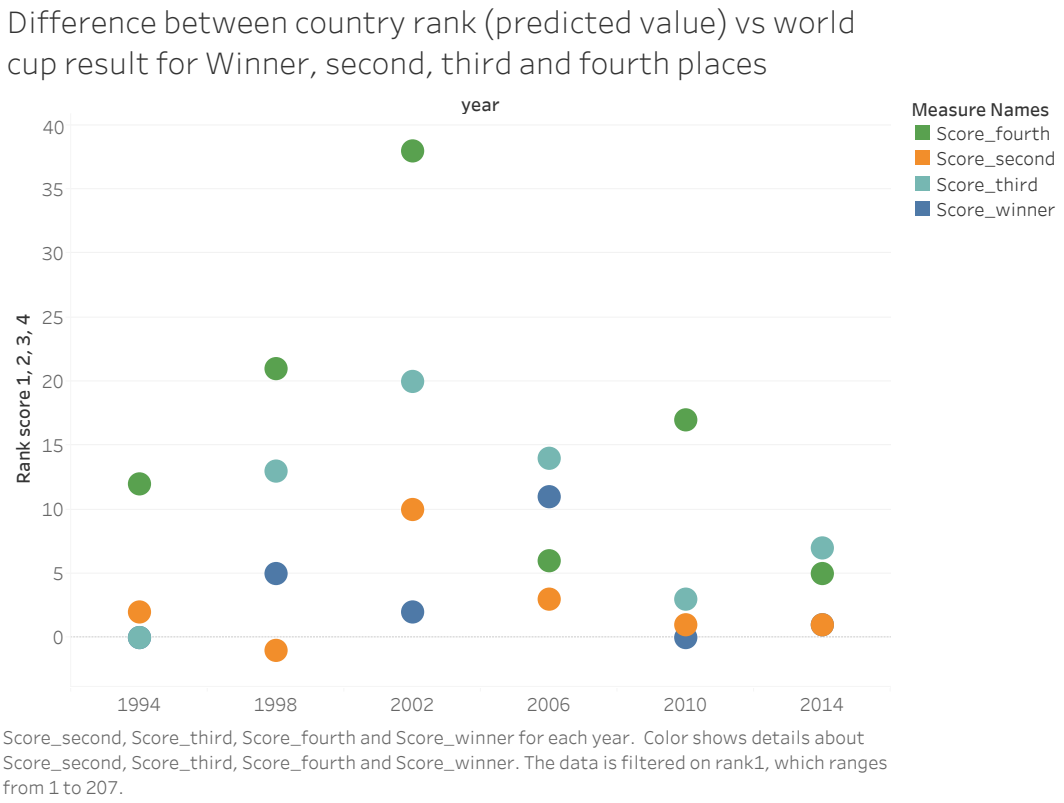
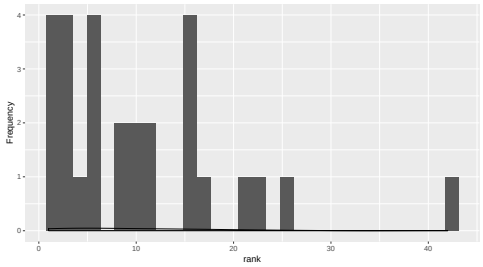
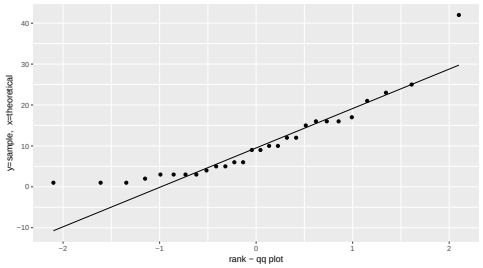


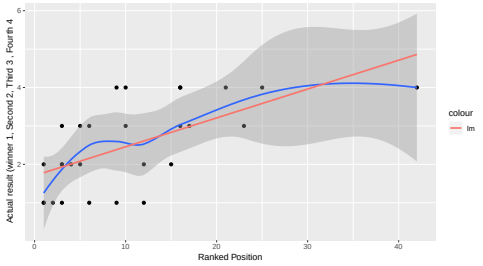
Figure 7: How FIFA rank compared to world cup final results 7.1.1 col=rank-result



(a)



(b)



(c)

```
> lmrank <- lm(result~rank, data=rank)
> summary(lmrank)

Call:
lm(formula = result ~ rank, data = rank)

Residuals:
     Min       1Q   Median       3Q      Max
-1.60732 -0.79390  0.04293  0.74829  1.61805

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.70585    0.26433   6.453 7.72e-07 ***
rank         0.07512    0.01891   3.973 0.000502 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9152 on 26 degrees of freedom
Multiple R-squared:  0.3778,    Adjusted R-squared:  0.3538
F-statistic: 15.78 on 1 and 26 DF,  p-value: 0.0005017

>
```

(d)

Figure 8: histograms / qqplots, scatter plot and summary of worldcup rank

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
rank	.152	28	.094	.856	28	.001
result	.170	28	.038	.859	28	.001

a. Lilliefors Significance Correction

(a)

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
rank	28	1	42	10.57	9.315	1.585	.441	3.443	.858
result	28	1	4	2.50	1.139	.000	.441	-1.389	.858
Valid N (listwise)	28								

(b)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	1.706	.264		6.453	.000	1.163	2.249
	rank	.075	.019	.615	3.973	.001	.036	.114

a. Dependent Variable: result

(c)

Figure 9: SPSS statistics skew, kolmogrov-smirnoff test of normality and Coefficients

7.2 BI Query 2: Is there any correlation between the number of redcards accumulated by a team in a competition and their success (Req-2)

For this query, the contributing sources of data are marked as 2 in the tables 2 and 4. The total number of redcards awarded to Winner,second,third and fourth place finalists at the European and World Cup are shown in 7. The number of redcards received by a country in a competition were requested using 7.2 and the results are presented in table 7.2
A one way ANOVA between groups was performed using IBM SPSS with following null hypothesis
h0: The number of redcards issued to winner, second, third or fourth has no affect on the final postion of the first 4 teams.
h1: The number of redcards affects the outcome.

The single categorical variable was number of redcards in a competition and this was compared against four equal groups namely Winnner,second, third and fourth. A csv file was prepared from the SQL output BI2_anova.csv with 2 columns red-cards and results (1,2,3,4). The one-way ANOVA is used to tell if there are significant differences in the mean scores of the dependent variable across the four groups.
Assumptions for one-way ANOVA are Homogeneity of variance and a normal distribution of values although the graphs. A graphical view 11 shows the values are skewed because I am only looking at the first 4 results.
The output of the IBM SPSS is shown if figure 12. The levene's statistic for homogeneity of variance showed p>.05 which means that homogeneity of variance had _not_ been violated.The ANOVA table however gave a Sig p-value of .345 (p>.05). Therefore we can conclude that the null hypothesis is accepted i.e. the number of red-cards issued to a team in the world cup and European cup had no statistical significance on the outcome of the first 4 places.

SQL query to get red cards awarded per competition ...

```
use worldcup
--- Find red cards issued to a country and add column showing FinalResult 1,
SELECT competitionDIM.[Year] as year, competitionDIM.[competition] ,
countryDIM.[country] as country,
rankFACT.red-cards as red-cards,
CASE countryDIM.fcode WHEN competitionDIM.winner THEN '1'
WHEN competitionDIM.second THEN '2'
WHEN competitionDIM.third THEN '3'
WHEN competitionDIM.fourth THEN '4'
ELSE countryDIM.fcode
END "FinalResult"
FROM competitionDIM, countryDIM
JOIN rankFACT ON countryDIM.country_SK = rankFACT.country_SK
Where countryDIM.fcode in (competitionDIM.winner,competitionDIM.Second,
competitionDIM.Third,competitionDIM.Fourth) AND
rankFACT.competition_SK=competitionDIM.competition_SK
order by FinalResult,year
```

Year	competition	**Winner**	redcards1	**Second**	redcards2
1954	World Cup	Germany	0	Hungary	1
1962	World Cup	Brazil	1	Czech Republic	0
1976	EuropeanCup	Czech Republic	1	Germany	0
1978	WorldCup	Argentina	0	Netherlands	1
1984	EuropeanCup	France	2	Spain	0
1986	WorldCup	Argentina	0	Germany	1
1990	WorldCup	Germany	1	Argentina	3
1994	WorldCup	Brazil	1	Italy	2
1996	EuropeanCup	Germany	1	Czech Republic	1
1998	WorldCup	France	3	Brazil	0
2000	EuropeanCup	France	0	Italy	1
2002	WorldCup	Brazil	1	Germany	1
2004	EuropeanCup	Greece	0	Portugal	0
2006	WorldCup	Italy	2	France	1
2008	EuropeanCup	Spain	0	Germany	1
2010	WorldCup	Spain	0	Netherlands	1
2012	EuropeanCup	Spain	0	Italy	0
2014	WorldCup	Germany	0	Argentina	0
2016	EuropeanCup	Portugal	0	France	0
2018	WorldCup	France	0	Croatia	0
Year	competition	**Third**	redcards3	**Fourth**	redcards4
1938	WorldCup	Brazil	2	Sweden	0
1958	WorldCup	France	0	Germany	1
1962	WorldCup	Chile	1	Serbia	1
1974	WorldCup	Poland	0	Brazil	1
1976	EuropeanCup	Netherlands	2	Serbia	0
1984	EuropeanCup	Denmark	1	Portugal	0
1994	WorldCup	Sweden	2	Bulgaria	2
1996	EuropeanCup	England	0	France	0
1998	WorldCup	Croatia	0	Netherlands	2
2000	EuropeanCup	Netherlands	0	Portugal	1
2002	WorldCup	Turkey	2	South Korea	0
2004	EuropeanCup	Czech Republic	0	Netherlands	1
2006	WorldCup	Germany	0	Portugal	2
2008	EuropeanCup	Russia	0	Turkey	1
2010	WorldCup	Germany	1	Uruguay	2
2012	EuropeanCup	Germany	0	Portugal	0
2014	WorldCup	Netherlands	0	Brazil	0
2016	EuropeanCup	Wales	0	Germany	0
2018	WorldCup	Belgium	0	England	0

Table 7: BI2.csv and BI_anova.csv: Total red-cards awarded at World and European Cup for 1st,2nd,3rd,4th place

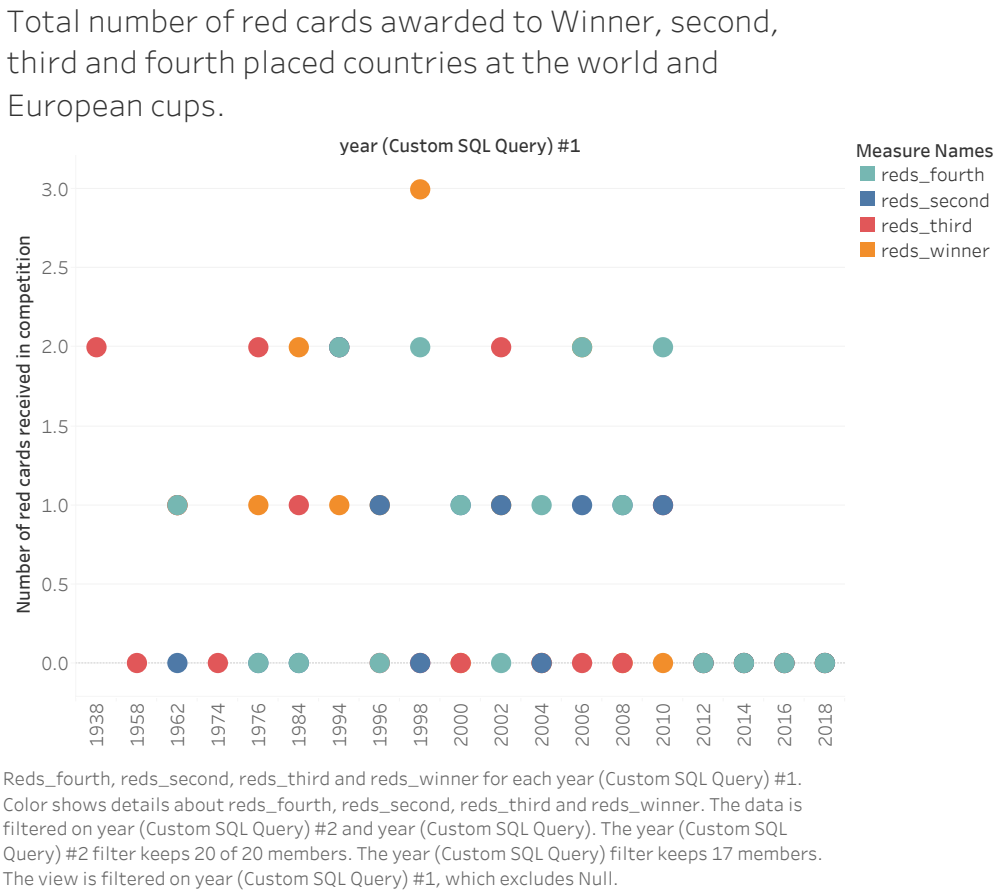


Figure 10: Total red cards awarded to winner, second, third and fourth place at World and European Cups

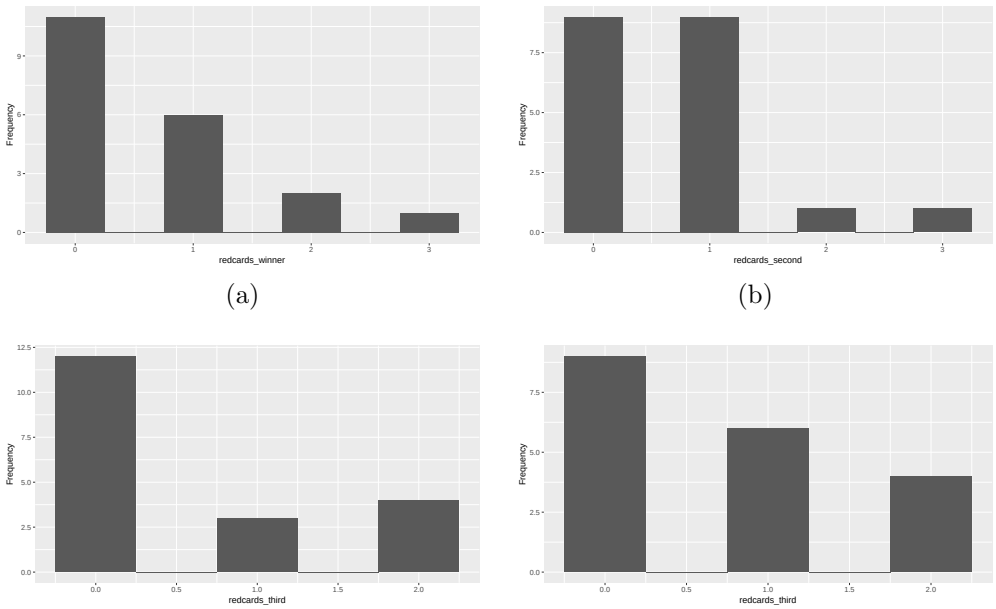


Figure 11: histograms of number of redcards issued to winner (a) second(b) third(c) and fourth placed(d) in world and european cup competitions since 1934

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
0	55	2.51	1.120	.151	2.21	2.81	1	4
1	24	2.38	1.135	.232	1.90	2.85	1	4
2	11	2.91	1.136	.343	2.15	3.67	1	4
3	2	1.50	.707	.500	-4.85	7.85	1	2
Total	92	2.50	1.124	.117	2.27	2.73	1	4

Test of Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
Result	Based on Mean	.670	3	88	.573
	Based on Median	.360	3	88	.782
	Based on Median and with adjusted df	.360	3	86.672	.782
	Based on trimmed mean	.708	3	88	.550

ANOVA

Result

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	4.220	3	1.407	1.118	.346
Within Groups	110.780	88	1.259		
Total	115.000	91			

Figure 12: ANOVA dependent variable= number of red-cards, independent groups are winner,second,third and fourth place

7.3 BI Query 3: Which teams have accumulated the most red cards in world cup history(Req-3)

For this query, the contributing sources of data are marked as 3 in the tables 2 and 4. The SQL query is shown in 7.3. The results are summarised in table 7.3. It must be noted that teams that do not qualify or get knocked out of the world cup early will not have the chance to accumulate many cards. Although Brazil has the most red cards (11), it has competed in every world cup since the competition started and have won the most times five.

SQL query to find the teams which have accumulated the most red cards ...

```
order by Total desc
SELECT countryDIM.country,SUM(red-cards) as 'Total_Redcards'
FROM competitionDIM, countryDIM
JOIN rankFACT ON countryDIM.country_SK = rankFACT.country_SK
Where rankFACT.competition_SK=competitionDIM.competition_SK AND
rankFACT.country_SK=countryDIM.country_SK AND competitionDIM.competition = 'WorldCup'
group by countryDim.country
order by Total_Redcards desc
```

County	Total_Redcards
Brazil	11
Argentina	10
Uruguay	9
Cameroon	8
Italy	8
Germany	8
Netherlands	7
France	6
Mexico	6
Portugal	6
Czech Republic	6
Hungary	5
United States	4
Australia	4
Croatia	4
Sweden	3
Serbia	3
Bulgaria	3
England	3
Belgium	3
Denmark	3
Chile	3
South Africa	2
Turkey	2
Paraguay	2
Algeria	2
Bolivia	2
South Korea	2
Honduras	2
Switzerland	2
Russia	2

Table 8: BI3.csv: Countries awarded 2 or more red cards at world cups

8 Discussion

Three queries regarding the world cup were addressed in this report. The main query (Req-1) looked at the ranking and the other two queries (Req-2) and (Req-3) looked at whether the number of red-cards affects the outcome result of the competition and which team had collected the most redcards. Although (Req-1) gave a significant result only the top 4 competition results had been examined and a more accurate linear regression or ranking would be obtained with more result data.
More data could be obtained by including the results of different competitions such as the Copa América and the qualifying matches for the world cup and Lasek (2016) suggest that friendly matches should not be included in the ranking. Many competitions such as the world cup do not have a fifth place so a method would have to be found to mark the position gained in the competition e.g.if a team is one of 4 who failed to qualify from the last 8 teams it will be put in 6th place.
Lasek (2016) also considers home advantage should be considered in the ranking and home advantage does seem to play a part as demonstrated by South Korea who were the host nation in 2002 and came fourth despite being ranked 42 in world 7.1.1. After their success in the world cup they climbed to 22nd place. A future investigation could be done to see how significant home advantage is to the ranking.
Improving the data is probably the best way to improve the statistical analysis however non-parametric statistical analysis could be another solution (Spearman rho instead of linear regression or Kruskal-Wallis test instead of one-way ANOVA).
Although this report was aimed at ranking in international soccer the tools and methods could be applied in future (with tweaks) to other ranked sports. The same tools developed could be used for the cricket world cup although they would need to be generalised and improved.
A date dimension was not necessary for any of the queries proposed but it may be useful to see how ranking vary on a monthly basis in future.
The next two queries (Req-2) and (Req-3) were to check if red-cards had any significance on the output following up on a suggestions there may be a link Paul & Mitra (2008). Statistical analysis was difficult because there are few red cards handed out at each competition which is reflected in the graph in figure 10. There were no red cards handed to the top 4 teams at the 2014,2016 and 2018 world and European cups. The ANOVA test indicated that the number of red cards a team received had no impact on the top 4 places. However the statistics take in the red cards over the period of a competition, not individual matches. If a team was reduced to 10 men at the start of a match it would have more impact on the game than at the end. Did the famous sending off of Zidane in extra time of the 2006 world cup affect the final 5-3 penalty win by Italy over France?
The third query (Req-3) showed Brazil to have received the most red cards but they had also played the most matches of any team.
It was necessary to develop several tools in this project although this may be to lack of familiarity with SSIS. SSIS is a major go-to solution and has it's devotees itcentralstation (2019).ITProToday (2019) calls SSIS an "insanely beneficial tool" but warns of the "hidden costs and expensive potential gotcha s" associated with SSIS. SSIS does seem to have room for improvement. The tool is linked strongly to windows and it difficult to link with analysis tools such as R and IBM SPSS (not everyone wants to use windows cubes and OLAP). In this project linking between the various stages was done with a mix of python, R and csv. Although the scripts I wrote were useful I think it would be

possible to create a more generalised solution with less glue logic using techniques such as those used in the EDA industry by successful companies like cadence Cadence (2019) and synopsys Synopsys (2019).

9 Conclusion and Future Work

This paper looked at the ranking for the world cup and whether red cards had an effect on the outcome. A linear regression showed a correlation between the result of the first 4 teams and the ranking, although teams seemed to be systematically ranked lower than the actual results obtained. The analysis could be improved by taking more data from other competitive competitions and qualifying matches or by using non-parametric tests on the data already available.
The effect of red cards on the result of the competition was investigated. An ANOVA test showed that red cards had no affect on the outcome but this applies to the outcome over a whole competition and the number of red cards issued since the start of the world cup in 1934 is very small making statical analysis difficult. Brazil who have competed in all competitions obtained the highest number of red cards which is only 11 in total.
Several tools were developed during the project because of difficulties using SSIS software and it was noted that it may be possible to create a more user friendly tool which links better with non-windows software.

References

Abdurazzag, A., Muhammed, M. & Yusuf, M. (2018), 'Icc world cup prediction based data analytics and business intelligent (bi) techniques', *(2018) 2018 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC),* p. p.273.

Baade, R. A. & Matheson, V. A. (2004), 'The quest for the cup: Assessing the economic impact of the world cup', *Regional Studies, , 38(4)* **27**(5), 343–353.

Cadence (2019), 'Cadence design systems'. Online; accessed 11/05/2019.
URL: *https://en.wikipedia.org/wiki/Cadence_Design_Systems*

Deena, W. (2018), 'Fifa world cup', *Salem Press Encyclopedia* .

Field, A. (2009), *Discovering statistics using R*, Sage publications.

FIFA (2019*a*), 'About fifa'. Online; accessed 09/04/2019.
URL: *https://www.fifa.com/about-fifa/index.html*

FIFA (2019*b*), 'Fifa, total number of goals scored at each fifa world cup from 1930 to 2018'. Online; accessed 09/04/2019.
URL: *https://www.statista.com/statistics/269029/number-of-goals-scored-at-fifa-world-cups-since-1930/*

FIFA (2019*c*), 'Fifa world cup winners'. Online; accessed 12/05/2019.
URL: *https://en.wikipedia.org/wiki/List_of_FIFA_World_Cup_winners*

FIFA (2019*d*), 'Fifa world ranking of men's national soccer teams as of april 4, 2019, statista'. Online; accessed 10/04/2019.
URL: *https://www.statista.com/statistics/262862/world-ranking-of-national-soccer-teams/*

itcentralstation (2019), 'itcentralstation'. Online; accessed 06/05/2019.
URL: *https://www.itcentralstation.com/product_reviews/ssis-review-30205-by-garym*

ITProToday (2019), 'Itprotoday'. Online; accessed 06/05/2019.
URL: *https://www.itprotoday.com/sql-server/hidden-costs-ssis-how-avoid-sql-server-integration-services-gotchas*

Lasek, J. e. a. (2016), 'How to improve a team's position in the fifa ranking? a simulation study", *Journal of Applied Statistics* **43**(7), 1349–1368.

Miller, T. (2010), 'Soccer conquers the world', *Chronicle of Higher Education* pp. B6–B9.

Paul, S. & Mitra, R. (2008), 'How predictable are the fifa worldcup football outcomes? an empirical analysis", *Applied Economics Letters* **15**(15), 1171–1176.

(sport.de), R. (2018), 'Average and total attendance at fifa football world cup games from 1930 to 2018'. accessed 10/04/2019.
URL: *https://www.statista.com/statistics/264441/number-of-spectators-at-football-world-cups-since-1930/*

Synopsys (2019), 'Synopsys'. Online; accessed 11/05/2019.
URL: *https://en.wikipedia.org/wiki/Synopsys*

wikipedia (2019), 'wikipedia'. Online; accessed 07/05/2019.
URL: *https://en.wikipedia.org/wiki/List_of_FIFA_World_Cup_finals*

Appendix

Python program to scrape for redcards

```
from bs4 import BeautifulSoup
import requests
import csv
from collections import Counter

# this is the url that we've already determined is safe and legal to scrape
page_link='https://en.wikipedia.org/wiki/List_of_FIFA_World_Cup_red_cards'

#get the content from the url, using the requests library
page_response = requests.get(page_link, timeout=5)

#parsing the page
page_content = BeautifulSoup(page_response.content, "html.parser")

#searching for the tag "table"
tables = page_content.find_all("table")

#we are interested in the second table (from the html)
t_heritage = tables[1]
#print(t_heritage.prettify())   #note how to access a subtag , with the "."
tr = t_heritage.find_all("tr")

csvfile = open('redcards.csv', 'w',encoding='utf-8')
writer = csv.writer(csvfile, delimiter=',',lineterminator='\n',quotechar = '
#writer.writerow( ["representing","year","host"]  )

redcards=list()
for rows in tr[1:]:   #start from 1, and for each rows...
    tag = rows.findAll('td') #all "td" tag in a list
#    print(len(tag))
    nr_tags=len(tag)
    if nr_tags==10:
        tournament=tag[7].text.strip().split(",␣") # only updated for 10 col
        year=tournament[0].replace("␣","")
#        host=tournament[1].replace(" ","")
        host=tournament[1]
        representing=tag[4].text.strip()
        rc=representing+"_"+year+"_"+host
        redcards.append(rc)
    elif nr_tags==9:
        representing=tag[4].text.strip()
        rc=representing+"_"+year+"_"+host
        redcards.append(rc)
    # Put everything in one list

# Count frequency of red cards
freq=Counter(redcards)
```

```
#print("freq",freq)
writer.writerow( ["country","year","host","nr_cards"]  )
for x in freq:
#    print('{0}: {1}'.format(x, freq[x]))
    cards=x.strip().split("_") # Split string
#    print('{0} {1} {2} {3}'.format(cards[0],cards[1],cards[2],freq[x]))
    writer.writerow ([cards[0],cards[1],cards[2],freq[x]])
#writer.writerow ([freq])
#print("File stored in redcards.csv")
print("output=redcards.csv")
csvfile.close()
```

Python program to find FIFA codes

```
import csv,sys
from collections import Counter

def check_duplicates(lrows):
nodups = []
dups = []
for row in lrows:
if row in nodups:
dups.append(row)
continue
else:
nodups.append(row)
print('{0} duplicates found => {1}'.format(len(dups),dups))

input="inputfile.csv"
output="out_"+input

# INPUT
colname="country"
argc=len(sys.argv)
if argc < 2:
print('Usage {0} {1} "colname"'.format(sys.argv[0],input))
sys.exit()
elif (argc==2):
input=sys.argv[1]
else:
input=sys.argv[1]
colname=sys.argv[2]
output="out_"+sys.argv[1]

# Read input file and check it has valid country header
validFile=False
myindex=0
ccol=0
with open(input, 'r') as f:
reader = csv.reader(f)
inlist = list(reader)
validFile=False
```

```
# Does file have a header called country
i=0
for header in inlist[0]:
i=i+1
if header == colname:
ccol=i-1
validFile=True

if(not validFile):
print('exit - invalidFile country not found in {0}'.format(inlist[0]),header
sys.exit()

# Read fifa_codes
fifacodes="fifa_codes.csv"
with open(fifacodes, 'r') as f:
reader = csv.reader(f)
fifa_list = list(reader)

csvfile = open(output, 'w',encoding='utf-8')
writer = csv.writer(csvfile, delimiter=',',lineterminator='\n',quotechar = '

# match and send to output
inplist=list()

for rows in inlist:
inplist.append(rows) # list of rows
found=False
for f in fifa_list:
if(rows[ccol]==f[0] or rows[ccol]==f[1] or rows[ccol]==f[2] or rows[ccol]==f
found=True
fcode=f[0]
fcountry=f[1]
rows.append(fcode)
rows.append(fcountry)
if(found!=True):
rows.append("NA")
rows.append("NA")
i=0
for rows in inplist: # skip line 0
#    if(rows[ccol] != rows[len(rows)-1]):
if((rows[ccol] == rows[len(rows)-2]) or (rows[ccol] == rows[len(rows)-1])):
pass
else:
#        print('{0} {1} {2} {3} '.format(rows[ccol],rows[len(rows)-2],rows[c
rows.append("*")
i=i+1
writer.writerow(rows)
print('{0} '.format(rows))

# Fifa countries with not match in input
```

```
nofifal=list()
for f in fifa_list[1:]: # Skip first line
found=False
for rows in inplist:
#        print('{0}.{1} -> {2}.{3} {4}'.format(fifacodes,f[0],input,rows[cco
if(f[0]==rows[ccol] or f[0]==rows[ccol-1]): # Find a matching FIFA entry
found=True
if (found!=True):
nofifal.append(f[0])

print('In␣{0}␣not␣in␣{1}␣=␣{2}'.format(fifacodes,input,nofifal))
check_duplicates(inplist)
print('in={0},out={1}␣counted␣{2}␣anomalies␣(*)␣'.format(input,output,i))
csvfile.close()
```

R program to plot qq,normal and regression curves

```
rank<- read_csv("VideoBI/BI1.csv")

ggplot(rank, aes(sample=rank))+stat_qq() + stat_qq_line() +
labs(x="WHS2_138␣-␣qq␣plot",y="y=sample,␣␣x=theoretical")
ggsave("figures/BI1qq.pdf")

ggplot(rank,aes(rank)) + geom_histogram() + labs(x="rank)",y="Frequency") +
ggsave("figures/BI1histogram.pdf")

ggplot(rank, aes(x=rank, y=result)) + geom_point() +
geom_smooth(mode=lm) +
geom_smooth(method="lm", aes(colour="lm"), se=FALSE) +
labs(x="Ranked␣Position",y="Actual␣result␣(winner␣1,␣Second␣2,␣Third␣3␣,␣Fou
ggsave("figures/BI1scatter.pdf")

# Calculate regression model
lmrank <- lm(result~rank, data=rank)
summary(lmrank)
```

R program to merge csv files

```
setwd("C:/Users/Martin/Desktop/DataAnalytics/DWBI/DWBIWorldCup2020")
infile1<-"DATA/rankFACT.csv" # All countries ranked for all years
infile2<-"DATA/redcardFACT.csv"  # Only some countries ranked
outfile<-"DATA/wcFACT.csv"
df1 <- read_csv(infile1)
df2 <- read_csv(infile2)

# Join FIFA country codes
df <- df1 %>% left_join(df2, by=c("fcode","year"))
#df<-na.omit(df) # Only used after manual check
write.csv(df,file=outfile)
```