Download as PDF

URL to view Results	[Click Here]

Response Summary:

Acquire Worksheet

Goal: Identify appropriate data sources, analyze the data, identify data types, variables, list assumptions about the data
Objectives: Students will identify and acquire data from appropriate data sources
Outcomes: Data for the current visualization challenge

1. Student Information *

First Name	Ivan
Last Name	Voitov
Course ***(e.g. CGT 270-001)***	CGT 270-009
Term ***(e.g. F2019)***	F2021

2. Email Address *
ivoitov@purdue.edu

3. Visualization Assignment *
- Training Data

Generate

4. Identify appropriate data sources: is the data publicly available? What search methods were used? *

Data source 1	https://www.kaggle.com/rgupta09/world-cup-2018-tweets
Data source 2	https://www.kaggle.com/rezaghari/fifa-worldcup-2018
Data source 3	world_cup_2018_squads

5. Data format: what format is the data in? Structured vs instructed? All text, a combination, multiple sources? Is it primary or secondary data? *

Data set one is unstructured since it is taken from twitter a social media site. The data is mostly text with some numbers so it would be a combination. It is taken from one main source twitter, but 520k tweets were used to make the data set. This data is primary since it takes ideas and data from twitter and puts them into a file .csv file for the first time, this data isnt based on any other data collected by anybody or anyone elses ideas. The second two data sets are both structured since they are just raw values and information taken from data sections of websites. The data in the second two sets is secondary since they are taking info from public Fifa websites that give them info on the teams and values and such.

6. Data types: what types of data are in the data? How are they stored? What is the access to the data (API, JSON, txt, csv, etc.)? What structure holds the data (data base, spreadsheet, etc.)? *

For all 3 of the files there are Strings, Integers, dates, Floating-point numbers, and Characters. Like dates of matches, names of players, goals scored by said players, and initials of teams. The access to the data for the first set is a csv file, but it was created via a Twitter API, and for the other two they are txt files with simple tables. The data is held by a spreadsheet for all 3 of the files.

Evaluate

7. Variables: list the data variables? What are the parameters? Give them names. What are the dependent variables and independent variables? *

There is a Type which is a String being Age or Caps, there is a team with a String which is the country, there is a group with a Character indicating what group the team was in, there is a name which is a String with each players name, there is a DOB which is a Date, there is Caps which is an Integer, there is goals which is an Integer, and there is a Country and Club which is a String. All of the data is independent since it is all raw and taken from biographies of each of the players listed.

8. Audience & Assumptions: list any assumptions you have about the data. Who is your audience? *

I would assume that European teams would do better than African or Asian teams since there is way more goals scored by those players, and I would assume that a European team would win the whole thing since they again seem to have the best players. I think that the intended audience is a combination of soccer enthusiasts, soccer analysts, and wikipedia or similar website page creators. These people seem like they would use and be interested in data like this.

Generate

9. What real life behavior does the data reflect? Does it show patterns of activity, regularity of events, a timeline, population data, etc? Explain. *

This is mostly a population type data of all the soccer players present at the 2018 World Cup.

11. What are the weaknesses of the data source? Is it likely that the source will be available in the future? Is the data complete? What is the quality of the data? Is it specific to your needs for. the current project? Is the data in the format you need? Are there missing data? Explain. *

I think that the way that this data set is structured it will not be complete in the near future, since the Goals section is a combination of all of the goals the player has scored across World Cups so it will be changing every 4 years until every single one of these players retire, assuming that no new players are added to the data set every 4 years. The data is good quality since I assume that all of this info is correct. It satisfies my needs for the current project, there is no important data missing from the past.

12. What information is emphasized? What is the central focus of the data? Explain. *

The way the data is presented I fell like the country/club and goals are the most important, since that its what most people look at along with the name of the player.

13. At what level of granularity is the data provided? Is the data summarized, or do you have access to the raw data? Is the data categorized or is the data in a format that allows you to create your own categories, etc. Explain. *

The data is raw and in sections, I cannot create my own categories since they are already made, the categories are the different variables described in a previous question.

14. What is the scope of the data? What topics can be covered using the data? Is there a time range/frame? Is the data for a specific area/discipline/demographic etc.? Explain. *

Topics surrounding stats of different players in comparison to one another and teams in comparison to one another can be made, I feel like the demographic is the same as the intended audience in this case being soccer enthusiasts and an analyst or two. The time frame seems to be the duration of the year that this world cup was held.