Analysis of Ufo Sightings

Anna Bussas bussas@mail.hs-ulm.de
Syndey Nkemakolam
nkemakolam@mail.hs-ulm.de
Kasparas Gudzius gudzius@mail.hs-ulm.de
Zhiwen Lian lian@mail.hs-ulm.de
Tim Hutttlestone
huttlestone@mail.hs-ulm.de
University of Applied Sciences Ulm

***Abstract*—Based on the data resourse from kaggle.com and the data processing by CRISP-DM, we can observe that most people saw UFOs as a light shape in the US during Summer in 2011.**

I. Introduction

We want to analyse the database ufo sightings from the website kaggle.com. For this we want to figure out when and where the UFOs show up mostly and how they look like.

A. Scenario

The data source has in total 11 columns which covers the discovering time, location, shapes, duration and so on. As our goal is to find out the shapes, country and the number of ufo sightings depending on daytime or night-time, we need to create a data warehouse containing at least, the information of shapes, country, time category.

B. Structure of the Paper

The paper is structured as follows: in Section II we present the availability of Open Data we used in this project. Then, in Section III we describe what is our concern to model the original data to our expected data. Finally, we conclude our work in Section VII with (XXXXXXXXXX placeholder –¿ to see if any section we want to add). In Section IV, V, VI, we will introduce how we process the data under CRISP-DM.

II. Open Data: UFO Sightings

As part of the *Data Understanding* phase for our data science project, we first look at the state concerning our problem. The material presented in this section is based on Kaggle [1].

III. Concept for Problem

During the *Data Understanding* phase, we find out a problem that we don't directly have the data we need .Additionally, some entries have NULL values. Accordingly, we need to have our Data prepared before the Analysis.

A. Data Profiling

In the *Data Profiling* step, we realise we do not need to keep some columns. Accordingly, we will only keep the following columns:

- datetime
- city
- state
- country
- shape
- duration(second)
- date posted

Also, for the NULL values, we will fill a necessary value in different columns:

- city / state / country
  If there are NULL values in city or state but not country. We will fill in the country in those two columns because we will mainly focus on the fact that the sighting was in a specific country. For those with both state and country missing, we will try to look at the city name, find the country and manually map as accurately as possible. The rest without enough information were filled with "unknown".
- shape
  We will fill in "unknown" to replace the NULL values.

B. Tools

The list of all tools we use in this project and the purpose:

- SQLiteManager
  Data Cleaning, ETL process, CDWH & DM creation
- DBSchema
  Forward Engineering
- BIRT
  Analytics
- Python
  Data Cleaning

C. DWH layout

After the ELT process, we should have a DHW which looks like this:

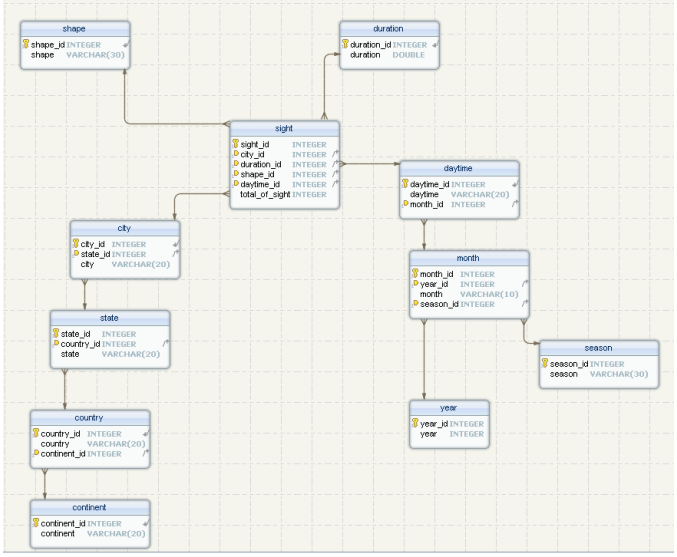

Fig. 1. DWH layout

IV. Data Preparation

This section deals with the implementation of the concept described in Section III following the *Data Preparation* phase of CRISP-DM.

To solve the problem, we have the following steps:

- step 1 Extract
  Download the data from the website then create a database to store all raw data.
- step 2 Filtering
  Fill in values by the rule that was mentioned in III-A.
- step 3 Enrichment
  We would like to see some relationship between the sighting, season, daytime or nighttime, country and so on.In other words, we need different dimentions. So we will create ”new” data by using algorithms:
  - Create new columns for the year, month, day, hour, daytime, season
  - Column year, month, day,hour
    Extract data from column datetime
  - Column daytime
    From 6am to 6pm, it is defined as daytime. The rest is defined as nighttime
  - Column season
    Compute by column month

V. Modeling

This section is to analyse the data we have prepared from Section IV.

To analyse, we have the following steps:

- step 1
  Import Cleaning Data in DWH (see DWH Layout in III-C)
- step 2
  For the performance and speed, we created four data marts for our goal (See Fig. 2, 3, 4, 5)

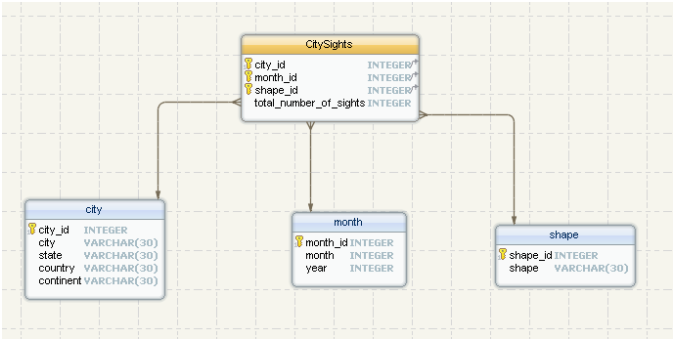

Fig. 2. Data mart - city/month/shape

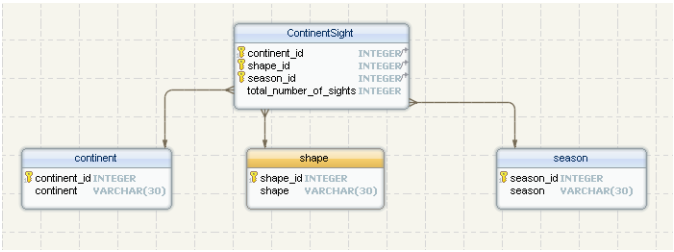

Fig. 3. Data mart - continent/shape/season

- step 3 Using BIRT, we generated reports and found the following facts:
  - Number of sightings in North American is extremely higher than in other continents. The next continent with the highest number of sightings is Europe.(See Fig.6)
    - Data mart - Shape/Country/Year
    - Data cube - continent
  - Nearly all sightings were in the US.(See Fig. 7)

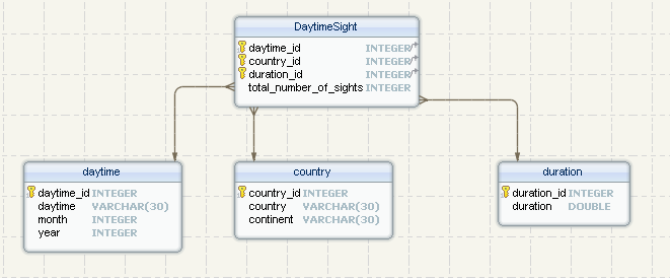

Fig. 4. Data mart - daytime/country/duration

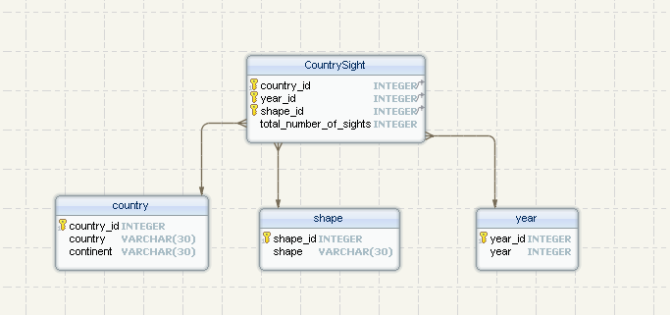

Fig. 5. Data mart - shape/country/year

- Data mart - Shape/Country/Year
- Data cube - contry

– The most Sightings were seen in 2011. We can see that it obviously increased between the years 1995 and 2011.(See Fig. 8)
  - Data mart - Shape/Country/Year
  - Data cube - year
– UFOs are most seen during the Summer. (See Fig. 9)
  - Data set - data mart - Continent/Shape/Season
  - Data cube - season
– Different UFO shapes are seen and the most common was light.(See Fig. 10)
  - Data set - data mart - Continent/Shape/Season
  - Data cube - shape
– In different continents, the most seen shape was also light. (See Fig. 11)
  - Data set - data mart - Continent/Shape/Season
  - Data cube - shape,continent
– In each season, light shape was also the most common. (See Fig. 12)
  - Data set - data mart - Continent/shape/season
  - Data cube - season, shape

VI. Evaluation

From the reports, we can find out the number of shapes, country and the number of ufo sightings depending on daytime or nighttime. Also the reports in different

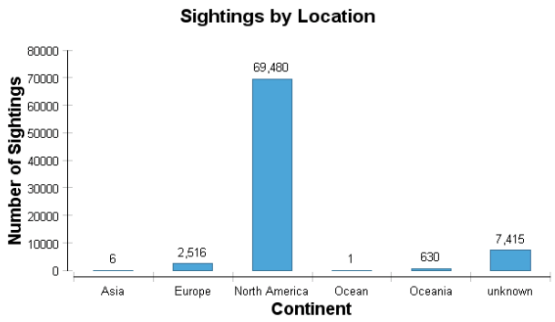

Fig. 6. Sightings by Location - Continent

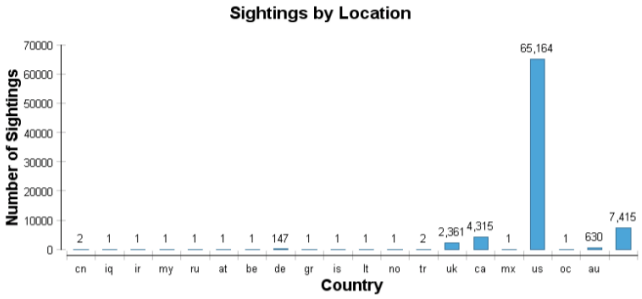

Fig. 7. Sightings by Location - country

dimentions. We can safely say that we have reached the goal after evaluation.

VII. Conclusion

We have shown in this paper how we processed the data under CRISP-DM and it shows different dimentions of shape, continents, country, seasons.

References

[1] *UFO Sightings*. Kaggle. https://www.kaggle.com/NUFORC/ufo-sightings

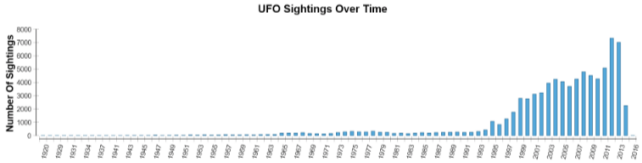

Fig. 8. Sightings Over Time

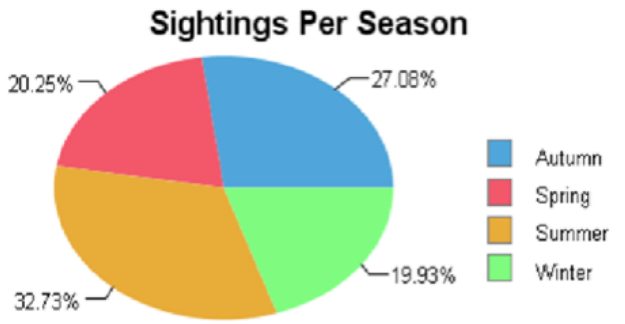

Fig. 9. Sightings Per Season

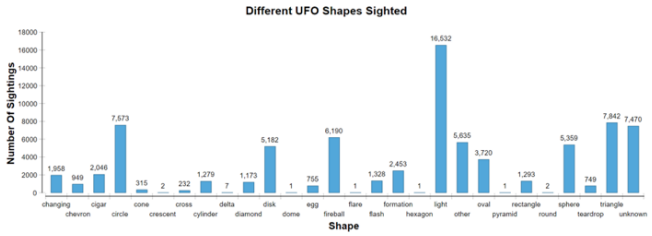

Fig. 10. Different UFO Shapes Sighted

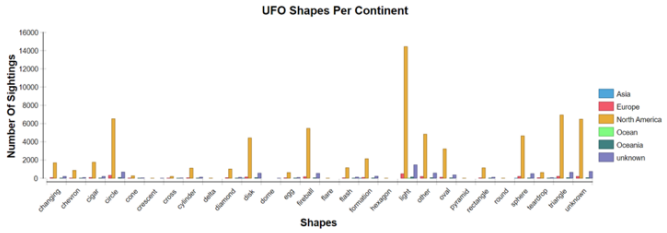

Fig. 11. Shapes Per Continent

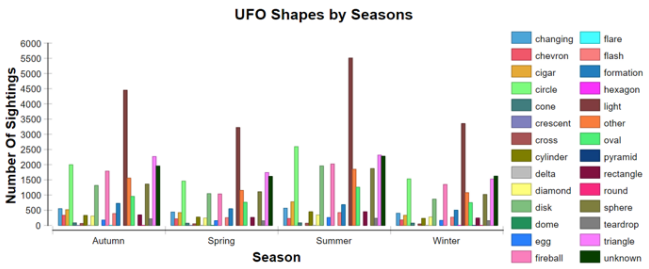

Fig. 12. Shapes Per Seasons