KAGGLE-UFO Sightings

Project B3

Link to github: https://github.com/KristinaMumm/IDS_Project_UFO_Sightings

Members:

Karl-Erik Kanal

Kristina Mumm

Business understanding

When searching for projects to do for this course, we came across an interesting dataset on Kaggle that contains data about UFO Sightings from an extended period of time. Since we both have never really researched this topic ourselves, we decided that this would be an interesting topic for the project. We hope to find out new interesting facts for both ourselves and for other students in this course. Besides producing visually enticing charts and presenting facts, we also hope to find interesting correlations between real life events at the time of the UFO sightings. Should we find an interesting rule or pattern in our data, we will look to see if such an occurrence can be somehow explained or coincides with events that have happened in that area. For example, we will see if UFOs have been spotted more near Area 51 or coincide with areas where meteors have frequently been seen.

For this project, we have set out three primary goals, which aim to find answers to the following questions and tasks at hand:

1) Find out where and what kind of UFO-s are most frequently seen
2) Find which US states report the most UFO sightings and how frequently they happen
3) Have UFO sightings increased over time?

The project can be considered a success if we manage to give informative and interesting insight into the world of UFOs to the people who will look at our poster. In addition, we should have fun doing it as well.

Personnel wise, it will be just the two of us doing this project and the lab assistant who will

give us aid if need be. Data wise we have one large dataset which is described in the data understanding segment below. For hardware we have our personal laptops which should do the trick.

The project should be aimed to be completed at least 2 days before the poster presentation session on Thursday the 16th of December. This is to allow buffer time in case there are issues with the project. Likely sources of delay would be other courses and tasks involved with them. To make sure the project is completed on time, dedicated segments from our schedules should be committed to the project at hand.

To accomplish our data-mining goals, we should produce different graphs which best illustrate the data and the tasks we have set. Furthermore, we should write down interesting facts we have found while mining the data. This should all be combined into one informative poster. If an onlooker finds the poster compelling then our data-mining goals can be considered a success.

Data understanding

The data comes from the Kaggle competition. The data is in one file. There are a total of 80332 records with 11 columns. The records include UFO sightings from 1949-2014.

Information in the columns:

- the date something was seen. (DD:MM:YYYY hh:mm)
- city
- state / province
- country
- UFO shape
- duration in seconds
- duration in hours / minutes
- seer comments
- date added to database (DD:MM:YYYY)
- latitude
- longitude

The data needs to be thoroughly cleaned. The only fields that do not require cleaning are fields with dates. There are columns with incomplete values and columns with incorrect data. For example, when importing data in a jupyter notebook, it was found that the duration (seconds) field has a value of "2`" and latitude has value "33q.200088".

Only Australia, Canada, USA, UK, Germany are listed in the database as countries, some rows have empty value there. Examining some rows with empty country values reveals that there are

also others countries represented like Norway and Kuwait. So there are certainly other countries. Also, not all rows have values in state / county columns. Since all rows have both longitude and latitude, one way is to find the missing data by using them. As a result of a quick search, we found two options: Reverse Geocoder and Geocoder. They both can be installed in the Anaconda environment. The second option is the most preferred, as it seems to handle all locations.

The city column is filled quite freely. Although there is no missing data in this column, it is quite difficult to determine the location. For example there is a value like “purcellville and lovettsville (near)”. We probably will not use this column.

There are columns for duration (hours / min) that are probably filled with the same words that the seer used. Values can be simply "60 minutes" and "16 minutes", but can also be "20 sec +/-" or "2-3 minutes". Since there is also a duration (seconds) column, where there are no missing values, and at first glance it seems that the values correspond to the duration (hours / min) field, there is no need for a column (hours / min).

The comments column is free text. It's impossible to clean easily, and in fact we don't need this column for our project. Although it would be interesting to know which words are most common.

The column “shape” is quite clean. There are few different values, so the cleaning will not be too hard. For example there are values "change" and "changing". So one value would be replaced by another. The remaining 2,000 values with empty values could be replaced as "light" because this value is represented about 16,000 times. The second most represented value count is about 7,000.

Planning your project

1. Cleaning the data. (Kristina. Probably around 2 hours.)
2. Finding out where and what kind of UFO-s are most frequently seen. (Karl-Erik. About 2 hours)
3. Creating plots to visualise found information. (Karl-Erik. Likely 2 hours.)
4. Researching which US states report the most UFO sightings. (Kristina. About an hour.)
5. Creating plots to visualise found information. (Kristina. Around an hour.)
6. Answering the question set for task 3 and finding other interesting correlations from our data. (Both of us. Around 2 hours.)
7. Designing the poster for our project. (Both of us. At least 2 hours.)

8. Creating a video presentation of our project. (Both of us. Around an hour.)

Tools we plan to use are the general data science modules for Python like numpy, pandas, matplotlib etc. Also a good idea would be to use the GeoPandas module which will allow us to create cool graphs depicting the world by country and more. The project will be done preferably using Jupyter notebook like we have done so far in our course.

The most important tasks are visualising the data and designing an attractive poster since our project aims to provide interesting information to others. The estimated time taken for the US segment is decreased since we'll have done the info for the world by then. Therefore we will have an understanding of how to achieve those tasks.