Data Visualization. Milestone 1.

Roxane Burri, Jean-Luc Stoupy, Mark Tropin

March 2022

Contents

1 Dataset

1.1 Characters

This is a dataset which describe the main biographical and physical data about Star Wars characters. There are ten different columns, like height, mass, hair color, ... for 87 characters. This dataset well describe specificity of each characters in Star Wars. For the preprocessing part, the main goal is to choose the most famous characters/or the one with the must fun facts.

1.2 Movie script

This dataset contains the entirety of the movie scripts of the original Star Wars trilogy. It consists of only two columns (excluding the index): the character and their dialogue line. Unlike the other datasets, it does not contain any numerical values, so it will require some statistical processing. This dataset can be used for simple statistical measures (which character is the most/least talkative) as well as for more advanced analysis (sentiment analysis).

1.3 Planets

This dataset contains a description of each planets in the Star Wars universe. There are 9 different attributes for each of the 61 planets. As for the characters dataset, the main goal is to show the most famous planets/or the one with the craziest values.

1.4 Star-ships

This dataset contains description of the 37 star-ships. As previous, the goal is to show the most famous star-ships/or the one with the craziest values.

1.5 Worldwide box office movie

This dataset contains the top movies based on the cumulative worldwide box office. This dataset describe the 50 top movies, so for the preprocessing part, all the other movies must be removed.

2 Problematic

In a current context where most of the information is sad, we decided to choose a joyful universe allowing to have fun and so we chose the cinematographic universe of Star Wars.
Through the different datasets detailed above, we want to show some facts about the Star Wars universe. These visualizations should be simple to capture the interest of any kind of knowledge about Star Wars. This universe is vast and complex, so the goal of all the work is to bring fun facts that can bring new

people into the universe or that can entertain people who already know Star Wars. We are primarily targeting Star Wars fans that would like to learn more about the Star Wars universe as well as newcomers that want to get to know the Star Wars saga, in a visual and entertaining manner.

3 Exploratory Data Analysis

In order to explore our datasets and to prepare the data for further processing and visualization, we have conducted the following steps:

- Remove columns/rows with missing values (not all NAs represent missing values)
- Reformat the data, remove the units (starships.speed, $1000km \rightarrow 1000$)
- Choose relevant data : we will not describe all the attributes in the datasets
- Collect basic statistics: number of attributes, number of datapoints, number (ratio) of missing values, range of values for each attribute, histograms and/or distributions of the values

In the figures below we provide some examples of the plots we produced to explore the data.

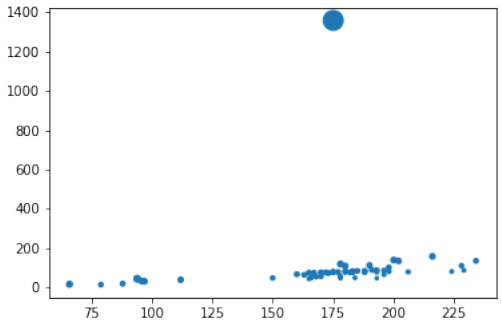

Figure 1: A plot of the weight of all characters against their height. The size of each point corresponds to the BMI (body mass index). One noticeable outlier is Jabba the Hutt.

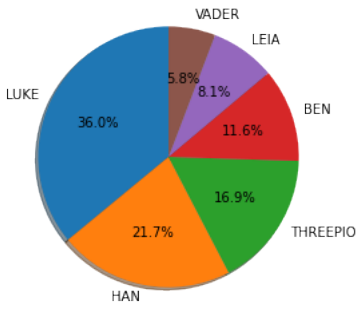

Figure 2: A pie chart depicting the ratio of dialogue lines said by a given character in Star Wars: Episode IV. This chart includes the top 6 most talkative characters.

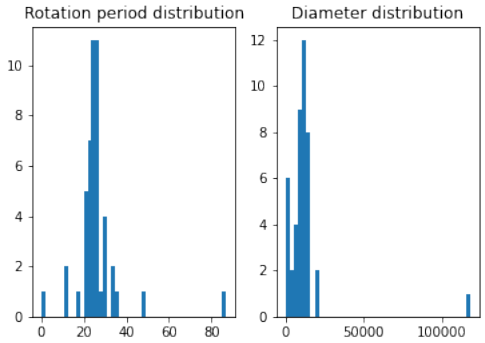

Figure 3: Histogram plots for the `planets` dataset. On the left, it shows the rotation period distribution and on the right the diameter distribution. For this last one, we have one outlier : Bespin

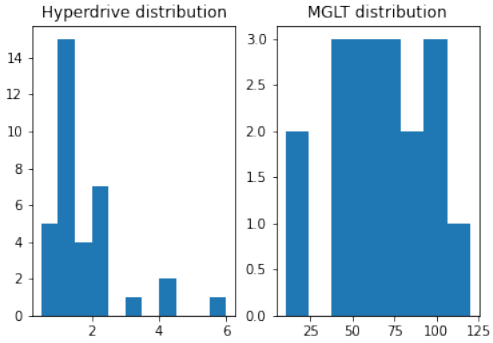

Figure 4: Histogram plots for the `starships` dataset. Left: hyperdrive rating distribution. Right: MGLT (megalight per hour) distribution.

4 Related work

- Most of the code that uses these datasets has been focused on EDA (histograms, descriptive statistics, etc.)
- The difference between the code that we have found and our project is twofold. First, the goal of our project is not statistical analysis, but visualization. This means that we don't aim to provide insight into the mathematical properties of the data, but rather attract the viewer with eye-catching graphics and present the data in a clear, visual manner.

  Second, as we are focused on providing our audience with the most captivating trivia about the Star Wars universe, we will not be using the

entirety of the dataset, just the parts that we personally find most relevant (and fun!).

- Our main sources of inspiration were the projects we have seen in class and the code examples of D3.js. For example, the tree-like links seen in this Beatles infographic have inspired us to mimic this visualization for associating planets/starships found in the Star Wars universe with the characters that we see in the movies.

  We have also used the D3.js website to find code examples that may be relevant to visualizing our data. For instance, this way of visualizing graphs may be very helpful for displaying relationships between Star Wars characters (allies/enemies).

- Some examples of blog about Star Wars : Wookieepedia, The force, Starwars-holonet these websites are really complete on the universe, but visually very annoying.