Alexandros Gkalkin

a.gkalkin@student.fontys.nl

Abstract

This document presents the process followed in order to create the visualization for the Data Visualization 2 course of the ERP&BI minor of Fontys University of Applied Sciences.

DAVI 2

Data Visualization procedure

Contents

[Introduction 2](#_Toc454151402)

[Process 2](#_Toc454151403)

[Selecting a data set. 2](#_Toc454151404)

[ETL 2](#_Toc454151405)

[Visualizing 3](#_Toc454151406)

[Results 5](#_Toc454151407)

[Tools 5](#_Toc454151408)

[Feedback 5](#_Toc454151409)

# Introduction

The scope of this document is to describe the process of the assignment for data visualization. The assignment is to visualize a data set in order to present it to other people. The topic of the visualization is battles of the American Civil War. The visualization itself can be found [here](https://app.powerbi.com/view?r=eyJrIjoiZTA1ZTJhYmItNDQ2NC00YTNlLWJkZDItZTMxYzUwN2ZiYmQ2IiwidCI6ImM2NmI2NzY1LWI3OTQtNGEyYi04NGVkLTg0NWIzNDFjMDg2YSIsImMiOjh9).

# Process

## Selecting a data set.

During the process a lot of data sets were tried out. Since one of the objectives of the assignment was to represent data in a way that it tells a story. Therefore, for this purpose the data set selected represents the battles of the American Civil war. The data came from a website dedicated to the American Civil war events in the form of a spreadsheet file[[1]](#footnote-1).

## ETL

Before explaining the operations performed on the data set it, we need to describe the actual content of the data set. The data set itself is not a very large one containing 371 rows of data. Some of the rows in the data set are not used (yet) in the actual visualization. In figure 1, we can see the content of the data set as it is categorized.

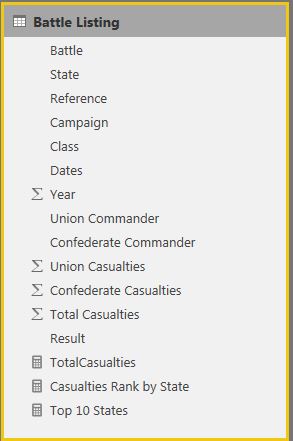


Figure - Dataset content.

**Battle**: This entry represents the name of the battle that took place in the American Civil War.

**State**: This entry represents the abbreviation of the state that the battle took place at.

**Reference**: This is the NPS reference which assigns a unique id number to a civil war battle.

**Campaign**: This entry represents the campaign of which the battle was part of.

**Class**: Class groups battles according to their importance during the war. The battles are assigned from A to D, with the following meaning:

* A = having a decisive influence on a campaign and a direct impact on the course of the war.
* B = having a direct and decisive influence on their campaign
* C = having observable influence on the outcome of a campaign
* D = having a limited influence on the outcome of their campaign or operation but achieving or affecting important local objectives

**Dates**: This represents the days that a battle took place.

**Year**: The year a battle took place on.

**Union** **commander**: The last name of the commander(s) of the United States of America (USA) army.

**Confederate** **commander**: The last name of the commander(s) of the Confederate States of America (CSA) army.

**Union** **Casualties**: The number of casualties of the USA army.

**Confederate** **Casualties**: The number of casualties of the CSA army.

**Total** **Casualties**: The number of the combined USA and CSA army casualties.

**Result**: The winner of the battle. This has the following entries:

* C = Confederate Victory
* I = Inconclusive
* U = Union Victory

Furthermore, some measures were created in order to combine some data for additional information (the measures use DAX, a sort of query language). Originally the data set contained additional information such as the Indian commander, Indian Casualties, number of images in the photo gallery of the website, the status of the author’s visit to the battlefields, references and links to civil war Wikipedia pages. But since the main scope of this assignment was to create a visualization of battles between USA and CSA, all that content was removed during processing. Also, the data set was missing some casualty information, so in some cases, the information was manually filled, and in the cases that the information was not found, it was substituted with a zero (0). This done in order to not complicate the calculation of the casualties in the visualization process. Finally, in the result category, the data was replaced from C,I,U to Confederate, Inconclusive, Union in order to make the data easily understandable.

## Visualizing

In order to have a certain sense of organization in the visualization of the data set, it was decided that the visualization should be separated into the following categories:

1. Theatre of War.
2. Generals
3. States
4. Years
5. Statistics

Theatre of War: Gives a visual map representation of the states that the battles took place. It provides a map, a selectable list of battles, a selectable year, a selectable result, and a selectable class to control the information on the map. Furthermore, it contains a few statistics that are related to the battle.

Generals: This tab is used to show a few statistics of battles between selected generals. There is a selectable list of the Union commanders, Confederate commanders, so you can easily choose one from one side and one from the opposing side and see some statistics of the battles between them.

States: This tab is here to present information over the casualties in different states. It has a Top 10 of the deadliest battles, with ability to filter the data by year, result, class.

Years: This represents the casualties and the results, divided by years, and grouped by casualty category and result category.

Statistics: represents some interesting statistics that are able to be filtered by certain groups.

# Results

The result of this assignment is an interactive story book, that can be flipped from page to page. Some of the pages give an insight of the followed data visualization, while others actually contain the visualizations.

# Tools

The main tools used for this project is Microsoft Power Bi for the creation of the visualizations, Power Bi service for hosting of the visualization, and finally Microsoft Excel for data transformation. Due to the use of the Power Bi hosting service I deemed it unnecessary to create a specific Git repository.

# Feedback

For feedback I asked

Jeffry Mozalbat, Student of Software Science at Technical University of Eindhoven:

In general terms the visualization is good. Once you read the titles of the charts, it is clear what is shown on it. To improve I would suggest the following:

* Make clear what each section is. (done)
* Be more specific about the class filter. If you don’t explain that A is important and D is the least important, it is not that clear. (done)
* Change the numbering in the top 10 to regular numbers, instead of 50k to show 50000. (done)
* In the Years section make the titles clearer of what is shown. (done)

Dimitrios Giotis, BSc in Statistics, financial and actuarial mathematics, University of Aegean, Greece

* Add calculation for average casualties per year in the year page. (not done)
* In the year page again, add data labels on the second chart. (done)
* Change the labels to give more understandable information. (done)

Aleksios Gialamas, Bachelor of Industrial Management, University of Pireaus, Greeece & Bachelor of Arts, Kraków Academy of Fine Arts, Poland.

* I don't understand what some of the titles mean. (Done)
* Add some instructions above the filters, for example Filter by year. (Done)
* The colour of the background is horrible. ( I decided not to change it)
* Add some slides before each page to show what the next page is about. Maybe add some explanation as well. (Done)

1. <http://www.civil-war-journeys.org/spreadsheet_of_civil_war_battles.htm> [↑](#footnote-ref-1)