LOG	***DA-DV***
TA	*Willemijn Becks*
GROUP	*15*
Participants	*Sarah Yeguez*
	River Vaudrin
	Jelle Bosscher
	Serge van Haag

Date	Topic of Interest	Indication of Timespan	Planning
DAY ONE 06/04/18	The four of us thought of questions to ask. We met with our TA - Asked how we need to start. - Questions about the project in general. Created a repository on GitHub. Made a journal on Google Docs.	2 HOURS	Think about questions to ask our TA Meeting with our TA Create a repository Create a journal Think about which dataset to choose
DAY TWO 06/05/18	Jelle taught us how to work with Github. Everyone made branches on Github. We choose for the dataset: Gun Violence in the USA. River found out that Panda fully supports our dataset (CSV). Updated the journal.	2 HOURS	Serge, Sarah and River need to understand github. Choose a dataset Check if Panda supports the format Think about questions to ask Willemijn
DAY TREE 06/06/18	We met with our TA. We did a few things with the dataframe: 1.Counted the missing classes. 2.We are changing address to street. 3. There is a problem with age of participants because it can contain more values. We are thinking about it how to handle this as a team. 4.Thought about what to do with missing values: ● If long/lat is not missing, but street is missing; then we can fill in the street value by making use of the long/lat	3 HOURS	Cleaning data - Missing values - Correcting inconsistent data Meet with Willemijn

	● Other missing data will be changed to ‘Unknown’. Updated the journal.		
DAY FOUR 06/07/18	Sarah found Health Care data. Sarah found Population data. Jelle cleaned the data for 80%. River and Serge started to work in a Jupyter notebook.	2 HOURS	Clean the data Search for datasets regarding: ● Health Care per state ● Population per state ● Gun licenses distributed per state Understand Pandas and create a working environment for pandas.
DAY FIVE 06/08/18	Visualize an overview of the importance and completeness of the dataset. Sarah scraped Health Care data. Jelle started on a script that should clean the rest. River and Serge got a few charts from pandas in Matplotlib. Updated the journal.	4 HOURS	Clean the final part of the data. Reduce the data if possible Keep searching for datasets regarding: ● Health Care per state ● Population per state ● Gun licenses distributed per state Create a few graphs using pandas and matplotlib.
DAY SIX 06/09/18	Jelle cleaned the data. Sarah cleaned the Health Care data. Updated the journal.	1.5 HOURS	Clean the final part of the data since it took more work than expected.
DAY SEVEN	Sarah scraped and cleaned Population data.	1 HOUR	Make sure all data is consistent and clean.

06/10/18	Serge checked if no data was missing. Updated the journal.		

DAY EIGHT 06/11/18	Meeting Willemijn - Willemijn said that mental health care might be interesting as well. Serge and River read the EDA. Made an extra question about mental health.	3 HOURS	Clean the data further. Read the EDA Think of extra questions.
DAY NINE 06/12/18	Every member of the team read the EDA. Serge and River created some explorative questions. Updated the journal.	3 HOURS	Read the EDA Creating explorative questions for the EDA and categorize them in one of the four categories. (univariate, multivariate, graphical, non-graphical)
DAY TEN 06/13/18	Applied techniques discussed in the EDA to the three main questions. Updated the journal.	2 HOURS	Apply knowledge from the EDA to the three main questions of our dataset. Make sure we can show our TA more in-depth graphs.
DAY ELEVEN 06/14/18	Met with Willemijn. - Displayed a few graphs - We need to make some values more clear Sarah read about clustering.	3 HOURS	Meet with our TA. Read about various machine learning techniques. Sarah started doing research on clustering.

DAY TWELVE 06/15/18	Updated the journal. Sarah clustered a couple of columns using k-means techniques.		Sarah is going to apply clustering techniques.
DAY THIRTEEN 06/16/18	Q1 is answered for states and cities, Serge still need to answer the years-part of the question. River found out that there is a peak in violence on the 1th of January and 4th of July. Also found some other notable data needed to answer Q2. Sarah tried to implement clustering techniques in Python.	4 HOURS	Serge answers Q1. River answers Q2. Jelle and Sarah work further on clustering.
DAY FOURTEEN 06/17/18	Sarah reduced the clustering implementation code. Updated the journal. Jelle is working on getting geo-locations from the dataframe.	2 HOURS	Jelle will cluster geo-locations in order to display an interactive map of the USA.

DAY FIFTEEN 06/18/18	Plotted graphs for question 4 Jelle created a map that is working by using the Google Maps api.	2 HOURS	Work on plotting gun violence incidents per capita for question 4. Work on clustering.
DAY SIXTEEN 06/19/18	Meeting Willemijn Serge and River read more about: - Linear Regression	6 HOURS	Meet with Willemijn. Ask our TA questions about regression and clustering.

	- Polynomial Regression Jelle en Sarah read in-depth about: - Machine learning - Clustering Updated the journal.		Read on regression and clustering as well.
DAY SEVENTEEN 06/20/18	Jelle, River and Serge met because Sarah had physiotherapy. We wrote down a few todo's which we shared with Sarah as well.	3 HOURS	Meet with the team to discuss the progress and todo's of this week.
DAY EIGHTEEN 06/21/18	Meeting Willemijn Everyone met at Sarah's house. Applied techniques: - Polynomial Regression - Linear Regression - K-means (clustering) Updated the journal.	5 HOURS	Meet with our TA Go to Sarah's house Work on making a regression lines. Clustering map is done.
DAY NINETEEN 06/22/18	River made a report in ShareLatex. River and Serge added labels to graphs. Everyone searched for interactive alternatives for Matplotlib. - Bokeh - Plotly - ggPlot	3 HOURS	Add labels to graphs. Find out how we are going to make graphs interactive in HTML. Read about data visualisation for the website. Added labels to graphs.

DAY TWENTY 06/23/18	We choose to use plotly instead of Matplotlib. Tried to implement Plotly in jupyter. Updated the journal.	1.5 HOURS	Choose an alternative to matplotlib for interactive plotting.
DAY TWENTY ONE 06/24/18	River and Jelle wrote a concept for the introduction and method. Serge plotted charts and dates in plotly. Sarah finished clustering characteristics of each incident.	4 HOURS	Create a basis for the report and start with introduction and method. Make charts that support the answers to questions. Work on clustering.

DAY TWENTY TWO 06/25/18	Met at Jelle's house. Report layout is completely done by River. Serge created graphs using Plotly. River plotted a regression line in plotly. Jelle is working on question three. Website is now accessible via Github. Updated the journal.	4 HOURS	Meeting at Jelle's house. Work on the Report. Work on answering the questions. Host the website on github.

DAY TWENTY THREE 06/26/18	Meeting Willemijn - She told us that we should write down mathematical methods in the method part of the report. Also that results are purely factual and that the discussion is needed for interpretation. Jelle did not feel well so we met at Sarah’s place. Worked on graphs. River made a website	5 HOURS	Meet willemijn - Ask about report Meet at jelle Work on graphs. Make a website.
DAY TWENTY FOUR 06/27/18	Uploaded graphs to our website. Added descriptive text to the site. Updated the journal.	3 HOURS	Make sure that we got all the graphs for the website. Put text on website
DAY TWENTY FIVE 06/28/18	Meeting Willemijn - Asked about our report and our website - We got advice on what to do this day. Updated the journal one last time. Serge is working on the website. Everyone is writing in the report. Sarah puts her data in the report and on the website.	10 HOURS	Finish the website Finish the report

DAY TWENTY SIX 06/29/18			Presenting our study.

Questions:

1. Are there any notable differences between states/cities/years? Visualize the differences (or similarities) that you can find.

2. What are the patterns you discovered that you suspect could be interesting? Does the data contain any unusual patterns that you did not expect?
    - Cases of gunviolence are more frequent through the years
    - Almost 50/50 suspect/victim ratio (maybe unexpected)
    - Chicago has the most cases
    - Compton has the highest murder rate / incident
    - 1 januari 2017 was most violent, 1 januari 2016 in violent top 10.
        - 1. 2017-01-01: 342 incidents
          2. 2017-07-04: 248 incidents
          3. 2017-05-28: 242 incidents
          4. 2016-08-28: 230 incidents
          5. 2017-04-16: 229 incidents
          6. 2016-08-21: 227 incidents
          7. 2016-07-04: 224 incidents
          8. 2017-04-22: 222 incidents
          9. 2016-01-01: 221 incidents
          10. 2016-07-17: 221 incidents
          11. 2014-09-06: 220 incidents
          12. 2017-10-21: 220 incidents
          13. 2016-10-16: 217 incidents
          14. 2016-11-27: 217 incidents
          15. 2017-05-27: 217 incidents

3. How does the pattern between killer/victim look between various incidents over the years. Is it mostly family/relation or do we see more reported gang violence for instance?

Extra questions:

1. Is there a correlation between mental health and gun violence?

   -See whatsapp group for the picture of the graph.

2. Which state is more dangerous if you take the population/cases ratio into consideration?

- Look at the population dataset

3. Factcheck this quote

> In Chicago, which has the toughest gun laws in the United States, probably you could say by far, they have more gun violence than any other city. So we have the toughest laws, and you have tremendous gun violence.

Donald Trump
Transcript of the Third Presidential Debate, www.nytimes.com. October 20, 2016.

Variables to look at:

- The gunviolence cases per state
- The number of gun laws per state

Serge's Notes:
Regression
Clustering
Testset

Training : 70%
Test : 10%

Polynomial fitting.

Introduction:

Gun violence has been a heavily covered topic in the United States for the past years. With the rising amount of mass-shootings, the debate about gun-control is a recurring theme across all of the media outlets.

The main purpose of this study is to answer six questions about gun violence in the United States, the.

This

Data:

The data was taken from the Gun Violence Archive, this is a corporation formed in 2013, that provides free access to accurate information about gun related incidents in the United States. The data provides a detailed overview of all the gun related incidents, such as numbers injured, numbers killed, type of gun used, amount of guns used, participant types, participant ages and the location details. The results found in the following pages are comprised of these 239.677 cases of gun violence in the United States from January 2013 to March 2018

.