Project Group Diary

30/09/2018

Group sat down to look at different data sets from WHO, UN and OurWorldInData.
It was decided that Google Colab would be used for all document reports and managing version controls. RStudio will be used for exploratory analysis of the data sets. Jupyter Lab will be used for the Julia part of the project.

Topics of data that were discussed and questions asked included:

- HIV/AIDS - how the proportion of HIV/AIDS relates to literacy rates and access to sexual health information for males and females in different countries.
- Drinking water - how access to safe drinking water, hygiene and sanitation facilities is related to GDP for various countries (especially in Africa).
- Child Health - how malnutrition and child mortality rates are related between genders and different countries with socioeconomic classes.
- **Suicide Rates** - how suicide rates are related to various features such as age, sex, employment, happiness, mental health funding, socioeconomic class, burden of disease, etc. for different countries.
- The relationship between environmental change and human health, like:deaths attributable to the environment;burden of disease attributable to the environment and so on.

Jaibir Singh Batth

30/09/2018

- Discussion about Scraping the data from different websites to show some relationship of Suicidal Rates with some other attributes.
- The Health Expenditure, the Happiness Score and the amount of money spent on HealthCare as a percentage of the GDP, were decided to be the attributes for which we need to discuss the correlation with the Suicidal Rates.

04/10/2018

- The data sources selected from which the data was to be fetched:
    1. https://en.wikipedia.org/wiki/List_of_countries_by_total_health_expenditure_per_capita
    2. https://en.wikipedia.org/wiki/List_of_countries_by_suicide_rate
    3. https://en.wikipedia.org/wiki/World_Happiness_Report#2016_World_Happiness_Report
    4. https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal)_per_capita

06/10/2018

- Exploring the rvest package for scraping a web page.
- Converting the HTML tables to data frames to perform further operations which is not possible with the HTML tables.
- Saving these data frames in the form of .csv files.

10/10/2018

- The data sets were merged and only the relevant columns were selected to make the merged data set concise and meaningful.
- The columns were renamed to give some readable names to the columns of the merged data sets.

12/10/2018

- Each of the above mentioned attribute was plotted against the suicidal rates in the different countries.
- Similarly the correlation was found for each attribute with the suicidal rates and represented graphically using a Scatter plot.
- The functions of rvest package make it super-easy to scrape the data from the websites.

14/10/2018

- Creating bar charts for showing the relation of Suicidal rates with Health Expenditure and Happiness score using the data frames merged earlier.

15/10/2018

- Finding the correlation of the different quantitative attributes with suicidal rates in different countries.

- Visualizing these correlations by plotting them with a Scatterplot and analysing the pattern which the observations follow.

16/10/2018

- Scraping the GDP data and selecting the relevant columns after removing the commas from the figures.
- Creating a data frame using this wrangled data.
- Merging this data frame with the suicidal rate data frame to perform further analysis

17/10/2018

- Calculating the percentage of GDP spent for Healthcare purposes by governments of different countries using mutate().
- Plotting the percentage of GDP spent for Healthcare against the suicidal rate for those countries using a bar chart and a point plot on the same visual.
- Collaborating the work done till date to prepare slides for the presentation.

20/10/2018

- Working on the project documentation as per the prescribed format.
- Adding comments to the code to explain what is the function of each piece of code.
- Adding observations and inferences made after visualizing the data in the form of graphs and plots in the form of Markdown in Jupyter lab.

21/10/2018

- Adding a brief about the Scraping that we did in the project into the project report.
- Synchronising the work that I have done, with the Github repository created by Blake.
- Reviewing the documents and code to be submitted.

Main Issues:

- The main issue was to find the right and authentic source of data to be scraped.
- It was not easy to find the data for all the attributes, that we were going to discuss, for the year 2016. The data was given either for a five-year range or for some year other than 2016.
- Moreover, some websites did not allow to scrape and we were not able to fetch the data frames out of it, even if the data were present on that website.
- The figures for some data sets contain commas (,) in them. So it was not possible to perform any computations on these figures as the numeric values, without any special characters, were required.
- For some data sources, the figures contain with the Unicode characters, and it was not possible for us to figure out how to deal with such data.

Yu Sun (Nina)

30/09/2018

- Group had the first meeting and discussed about the potential topics like HIV, drinking water and suicide rate.
- Searched for different data sources using WHO, Our world in data
- Collected several datasets and had a brief look of the missing values, the number of row and columns
- Discussed several potential topics.
- Decided to combine some datasets

03/10/2018

- Searched on the website again for suicide rate data and mental health
- Found the overall suicide rate for each country from 1995 to 2016
- Found some other datasets that could have some relationship with the suicide rate like employment, perception of corruption for government spending on employee compensation, but the countries in those data are a bit different from the suicide dataset we decide to use and years also differs
- links:https://ourworldindata.org/suicide

05/10/2018

- Found another data for mental health situation in different countries, but realized the year and country was not consistent with our suicide rate dataset
- Started reading CSV into R and have done a bit wrangling on the missing values on suicide dataset.

08/10/2018

- Started wrangling on suicide rate for male and female
- Selected different columns from the dataset created which contains year, country, male and female suicide rate and other variables like healthy life expectancy, social support, log GDP
- Decided to use bar chart to show the correlation between suicide rate and other variables like social support rate etc.

10/10/2018

- Done the wrangling on suicide rate for male and female in all continents and have created bar chart regarding different continents and have found that male tend to have a higher suicide rate than female.
- Done the wrangling on suicide rate & Log GDP in all the continents and have done bar charts to see the correlation.

11/10/2018

- Done the wrangling on suicide rate and social support rate as well as healthy life expectancy with bar plot.
- Had a look of all the bar chart and after discussion realized bar chart may not be a good way for visualizing the correlation then have started to change the plot into scatterplot

- Have redid some of the wrangling because the scatterplot requires a bit different data format.

14/10/2018

- Finished wrangling on suicide rate with social support rate, log GDP and have done scatterplot for those.

16/10/2018

- Finished wrangling on suicide rate regarding different gender and healthy life expectancy.
- Finished the scatterplot for the above
- Uploaded the csv file for dataset on Google file.
- Upload the plots to show the linearity into the presentation slides

20/10/2018

- Worked with Jaibir and Blake on the report
- Added some conclusions from the plot, as well as the success part and failure part of the plot onto the report
- Transferred the code from R into Jupyter notebook
- Created a Github account to upload our code and work etc

21/10/2018

- Worked with Jaibir and Blake together regarding the report
- Did a final check of the jupyter notebook and those files that will be uploaded

Main Issues:

- The wrangling was a bit time consuming because different columns was selected each time when analyzing correlation on different variables. Each time a new datatable was created.
- When decided to use scatter plot instead of bar chart. I realized those separate datatables actually need a different format for scatter plot.
- It also takes a bit time to choose the right plot to show the comparison between different variables with suicide rate. Because there is a large number of countries in the dataset, bar chart may not be a good way to show the comparison. But we have decided to use scatter plot later since that could present the linearity more clearly and the it works well.
- There are some missing values existed for some specific countries, those missing values will have to be omitted because a lack of information. But omitting the missing values will make the size of the data even smaller.

Blake List

30/09/2018

- Discussed potential topics of research for the group project. Investigated many different datasets from WHO, UN, Our World in Data, Google Datasets, and Kaggle.
- Set out the structure and environment for the group project. Google Colab would be used for all documents, reports and version control. RStudio would be used for analysis and visualisation of data and then converted into Jupyter Lab Notebooks for submission. Jupyter Lab would be used for the Julia part.
- Suggested a range of topics including:
    - HIV/AIDS - how the proportion of HIV/AIDS relates to literacy rates and access to sexual health information for males and females in different countries.
    - Drinking water - how access to safe drinking water, hygiene and sanitation facilities is related to GDP for various countries (especially in Africa).
    - Suicide Rates - how suicide rates are related to various features such as age, sex, employment, happiness, mental health funding, socioeconomic class, burden of disease, etc. for different countries.
- Explored datasets related to the above topics and decided on the topic of World Suicide Statistics.

02/10/2018

- Downloaded datasets regarding the topic identified.
- Performed small exploratory analysis on certain datasets to see what sort of condition it was in, any missing values, relational columns, etc.

04/10/2018

- Formulated some questions that the project could be centred on:
    - How suicide rates differ between males and females per country.
    - How GDP and unemployment could affect suicide.
    - Whether the ‘happiness’ of a country correlates with its suicide.
    - How the amount of money spend on mental health relates to suicide rates per country.
- Discussed questions with the group about ideas regarding the questions.
- Got feedback from other groups about potential roadblocks and methods.
- Performed some wrangling on the suicide rate per 100,000 population by gender dataset.

07/10/2018

- Continued wrangling the datasets and achieved a format that would be suitable for plotting.
- Visualised the gender suicide rate data using bar plots.
- Investigated packages that would allow for a map to be produced, potentially with animation.
- Wrangled the world happiness report dataset to a form that could also be used for plotting or joining.

09/10/2018

- Created dataframes for individually exploring suicide statistics and world happiness indicators.
- Created a single relational dataframe containing intersecting data from both suicide rates per 100,000 population by gender and world happiness report variables, merged by country and year. These could then be worked on by the group to individually and collaboratively answer the questions at the centre of our project.

11/10/2018

- Worked on the datasets with the group to investigate different relationships between suicide statistics and the world happiness indicators.
- Worked on the suicide rates per 100,000 population for different years by creating bar plots and map plots.
- Wrote functions to filter different years and plot their corresponding map plots for males and females.
- Investigated other forms of maps that could be used, e.g. Europe, United States, Africa, etc. Packages including ggmaps, worldmap, maps, etc.

14/10/2018

- Worked on the Julia component of the project. Had many issues with updating Julia packages.
- Loaded in and wrangled the data for suicide statistics per 100,000 population and world happiness report information.
- Encountered an issue when trying to merge these dataframes together as the class type of year was a symbol after being gathered and would not merge.
- Began working on the project presentation, creating the slides and template to be filled with content.
- Animated the world map of the suicide rates by gender for a given year using gganimate. Unfortunately, the transformation was not as pleasing as first hoped.

15/10/2018

- Finished the Julia part of the project with the help of Giulio in solving the dataframe merging issue.
- Added content to the presentation such as background, questions/hypothesis, and methods.
- Found a more comprehensive dataset featuring total suicide rates per 100,000 population for the years 1950 – 2016 from Gapminder.
- Created an animated bar plot of the total suicide rates for countries that had adequate numbers of observations, transitioning between 1951 and 2015.

16/10/2018

- Completed the presentation slides including graphs, discussion and conclusion, and practiced presentation.
- Tidied and commented both R and Julia code.

- Created CSV's of the final dataframes featuring the format used for visualising data that could be used in the future.

18/10/2018

- Practiced and checked presentation. Performed it during the lab.
- Discussed report format, ideas and final data analysis with Nina and Jaibir.

20/10/2018

- Worked on report with Nina and Jaibir.
- Wrote the introduction and discussion and added results to the report.
- Converted R code into a Jupyter Lab Notebook.
- Created a document linking Jupyter Lab Notebooks together, explaining their relevance to the project aim.
- Created a document with the content of each final CSV.
- Tidied up Google Colab folder.
- Created an initial GitHub repository to upload the project report and code.

21/10/2018

- Finished and checked the report with Nina and Jaibir.
- Completed a final check of the code notebooks and Google Colab files together.
- Committed all the documents and code to GitHub.

Main Issues:

- Implementing map and animated plots using new packages that had not been previously seen in class. There was little documentation or examples online to help so trial and error of parameters and testing solved the problems.
- Julia has massive package issues. The need to manually delete and rebuild packages costed a lot of time.
- Merging dataframes in Julia after gathering caused issues as one column was of type symbol and therefore could not be merged with a string column.
- Lack of participation from Shun added stress and more work to the group. Overall, the active members worked very well together.