**PROJECT DIARY – THE WRANGLERS**

**DATA422 - GROUP PROJECT – DATA WRANGLING**

**WEEK - 8**

**Group Project Official Launch**

**Team Name**: The Wranglers (Updated in Excel Sheet Provided by Guilio)

**Team Members:** Raghuraman, Kartik, Gajalakshmi, Rishil and Oscar

**Topic Discussion:**

Each one of us suggested out a random topic of their interest.

1. Cricket (Suggested by Raghuraman)
2. Transportation (Suggested by Gajalakshmi)
3. Weather Events (Suggested by Kartik)
4. Accidents (Suggested by Oscar)
5. Politics (Suggested by Rishil)

We have kept in mind to search for relevant data source keeping the following criteria in mind

1. Data source needs to be reliable
2. Data source has provision for web scraping
3. Terms and Conditions of the provider allows us to scrape and use the data
4. Data is not in data frame format
5. Data is untidy and has scope for wrangling
6. Availability of common variables for joining the dataset

Created a WhatsApp group to stay connected on meet ups.

**WEEK - 9**

**Fast Prototyping**

* Each one of us found relevant data sources on their suggested topic.
* Based on joining criteria and interesting association, we have finalized on taking Weather Events and Compare that with ACC Accident Claims.
* Discussed our finalization with Phil and obtained his feedback
* Created a Slack group to share findings, documents and progress
* Met at Puaka James Height to discuss on variables/fields to be extracted, their naming conventions and data types.

**WEEK - 10**

**Group Work**

* Discussed on web scraping ideas based on lab work and assignment 2.
* Divided the work among us
* Discussed and found a couple of research questions to keep ourselves right on track.

**WEEK – 10,11 and 12**

**Task Tracking – (Wrangling, Presentation and Report)**

|  |  |  |
| --- | --- | --- |
| **Contributor** | **Task** | **Contribution** |
| **Raghuraman Srinivasan** | **Organization** | Drawn an overall outline on how to proceed to the group project.  Maintained and updated the project diary. |
| **Scraping/Wrangling** | Automated the web scraping process of Weather Events data in R using paging logic.  Eliminated duplicates and combined all the data frames into a single data frame using “rbind”.  Refined the active claims Julia code on ACC claims data and added comments it. |
| **Visualization** | Analysed the relationship between some factors such as Mode of Transport, Ethnicity and Weather Event Type.  Created graphs for visualization addressing research questions in R. |
| **Presentation** | Offered ideas and guidelines on what to analyze and present.  Prepared the conclusion and intended future use of our project work. |
| **Report** | Wrote report part related to introduction, web scrapping, and wrangling related to the Weather Events data.  Edited the group project report and added supplementary explanation. |
| **Kartik Mahajan** | **Organization** | Raised questions about necessity of discussion related to project on topic searching and problem solving. |
| **Scraping/Wrangling** | Scraped the Weather Events data by utilizing the rvest package using R.  Utilized Regular Expression logic to filter data from text present on the headlines.  Handled all the Null values through effective replacement from summary containing textual information. |
| **Visualization** | Analysed the relationship between some factors such as Gender, Age Category and Weather Event Type.  Sorted and filtered the data from joined data frame to plot for the research questions. |
| **Presentation** | Prepared the slides for web scraping of Weather Events data. |
| **Report** | Uploaded relevant files of the group project to Slack, and shared with other members on steps to work.  Added supplementary information to the report on techniques used. |
| **Gajalakshmi Sakthivel** | **Organization** | Coordinated with the team and arranged for regular meetings. |
| **Scraping/Wrangling** | Wrangled new claims part of the untidy ACC claims data into independent data frames suiting to research questions using R.  Converted new claims part of the ACC claims data from wide to long format so as to structure the joining field without duplicates using R. |
| **Visualization** | Analysed the relationship between some factors such as, Month, region and Weather Event Type.  Added comments to the R code steps on data wrangling. |
| **Presentation** | Prepared the slides for wrangling of ACC claims data.  Read through the terms of use and conditions and added it to the presentation. |
| **Report** | Reviewed codes regarding the wrangling and plotting in Julia part.  Added supplementary information to the report on challenges faced. |
| **Oscar Vanhanen** | **Organization** | Created a slack group for sharing ideas.  Created and maintained GitHub project repository for handling code versions and accessibility. |
| **Scraping/Wrangling** | Wrangled active claims part of the untidy ACC claims data into independent data frames using Julia.  Converted active claims part of the ACC claims data from wide to long format so as to structure the joining field without duplicates using Julia. |
| **Visualization** | Sorted and filtered the data from joined data frame to plot for the research questions |
| **Presentation** | Prepared the introduction slides.  Designed the relational data model diagram of the project to suit the research needs. |
| **Report** | Consolidated final codes in R and Julia by collecting each parts from group members.  Summarised the whole group project into a brief txt file and published the group project to GitHub. |
| **Rishil Parab** | **Organization** | Booked meeting rooms for project discussions. |
| **Scraping/Wrangling** | Joined the individual data frames based on “year” of occurrence to answer the research questions using R.  Addressed the duplicate values on the merged data frames. |
| **Visualization** | Created graphs for visualization addressing research question in R.  Refined the codes and added comments to the R code steps on ACC claims data. |
| **Presentation** | Maintained the presentation slides, edited and added all the supplementary information. |
| **Report** | Wrote report part related to goals and data sources.  Added supplementary information to the report on failures that occurred. |