Data201

# Group Project Diary

# Take Five

*GIT Repo : [https://eng-git.canterbury.ac.nz/jle147/data201-group-project](https://eng-git.canterbury.ac.nz/jle147/data201-group-project)*

## Group Tasks Done Together

17/09/2020 DATA201 Tutorial

- We gathered for the first time and thought up of these ideas for the project
- Ideas:
    1) Covid 19 dataset combined with boxing event occurrences in affected countries
    2) Relationship between profitability of games and their tags and genres
        a) Three different platforms
    3) Weather situations with Hospital data
        a) Data from different hospitals(?)
    4) StatsNZ combined any two

24/09/2020 DATA201 Tutorial

- Discussed thoroughly on what ideas would be good for the project and realised that the previous ideas we had were either too simple or too clean to make a relation out of.
- As such, we spent the lab session googling for datasets which are broad in general, that is, it has many columns to pick from.
- We decided on the Crash Analysis Systems dataset which had many useful fields like crash location,  crash year, light, accident severity etc.
- Using that dataset, we planned to use the weekend to find other datasets, which are hopefully dirty, to complement with this dataset that we just found

26/09/2020 Own locations

- All five of us gathered in a zoom meeting to discuss all the datasets we found to sift through the best dataset to work with.
- Each of us gave 2 datasets each, and we explained our ideas on how to use them with the main dataset to produce some good wrangling.
- We finalised with the Traffic Service dataset, which was about the general traffic in all of NZ, with fields such as road name, average daily traffic, percentage of cars, peak hour etc.
- We planned to meet again 3 days later, to discuss specifically what we are going to do with both datasets, so that we can get started on wrangling.

29/09/2020 Own locations

- We gathered again in another zoom meeting, to discuss our ideas on what has to be done specifically to get the desired difficulty and challenge for the group project.
- We concluded that we would work in two groups, one group working on one aspect of the combined datasets, while another working on another aspect.
- We intend to combine these two ideas together once it has been completed.
- These two ideas are:
    a) Needed columns from Crash Analysis System:
    1) crashLocation1
    2) crashLocation2(if crashLocation1 is not found on the other dataset)
    3) crashYear

    Needed columns from TrafficService:
    1) road_name
    2) count_date
    3) peak_hour
    4) pccar
    5) pclcv
    6) pcheavy - percentage of pcmcv + pchcvi + pchcvii + pcbus (can use a check in the dataset for this part to ensure the numbers are correct)

    Idea:
    Making a relation between the types of vehicles during the average peak hour in a year with the location of accidents

    Using the TrafficService dataset, get the average peak hour of a road in a year. Ignoring rows without peak hours, some rows would already have been filtered out since they do not have peak hours.

Combine the crashLocation(1 or 2, depending on availability) and crashYear in the main dataset with road_name and count_date(focus only on the year) in the modified TrafficService dataset. We will only focus on the year between 2000 to 2019(since that is the max min that both datasets can comply with). Roads which turn out not to have an accident will be ignored.

Using the combined dataset, we can plot out the trend of what type of vehicles are popular in each year(between 2000 to 2019) during the average peak hour, where the accident occured(road name/crash location)

1/10/2020 DATA201 Tutorial

- Four of us gathered in the tutorial today to share our progress for the project.
- Since we split the group into two, cleaning of the data could be done separately, and if any group was done first, they could figure out more ideas to add onto our existing dataframe
- After consultation with the lecturer, we concluded that what we have planned may be too little to amount to a group project, so we decided to think of some web scraping ideas to add onto what we have.

8/10/2020 DATA201 Tutorial

- Three of us gathered in the tutorial today to discuss in person on our progress for the project. Coincidentally, today was also the day where we had to share our progress/ideas to other groups, which worked in our favour as we needed more ideas and feedback on our project.
- Some of the ideas we had were:
    a) Combining the dataset we did for our assignments with the one we currently have. For example, game releases will be connected with crash accident dates to show how games could affect the accident rates in New Zealand.
    b) Make further in-depth wrangling with our existing dataset by narrowing down on specific regions.
- The tutors also gave valuable feedback about the type of wrangling past year students have done, which gave us insight to the amount of things we could do so long as our end goal was clear.

10/10/2020 Own locations

- All five of us had a zoom meeting today to share our progress on the project.
- This time round, it was just a brief update on what we have done on the project to other team members, so that each of us would be on the same page at all times.

14/10/2020 School Library Rooms

- Met at the school library to discuss the presentation for the next day and also the progress on our project.
- We managed to plan and split the presentation slides evenly amongst all five of us to speak for the presentation the next day.
- Some miscommunications on the end goal of the project were cleared through thorough explanations and discussions between the group.
- On the project progress, Jemin and Chathuranga made huge progress on the project, which means now we are able to consider adding more datasets into the cleaned and modified dataset to add more layers of complexity to the project.

15/10/2020 DATA201 Tutorial

- One hour prior to the tutorials, we gathered together to rehearse our parts for the presentation to ensure all of us knew what to say. The practice is to also ensure that we are staying within the 7 minute mark for the presentation.
- Later in the tutorial, we then presented our parts to the class.
- We got to learn more ways to improve our dataset through the questions that were asked by the audience.

# Solo Contributions

## Phua Sheng-He

1) General Tasks:
    - Maintaining of group diary (every meeting)
    - 6/10/2020

      Started on the group report, added details on:
        a) Reasons why we chose that data source
        b) Intended Use of that data(Final goal for the wrangling)
        c) Difficulties section of the report
2) DATA201 specific Tasks:
    - 2/10/2020

      Converted the TrafficeService.csv file into a dataframe in R. Extracted out the columns we needed for the project into a dataframe, and cleaned these columns as some of the data were inconsistent. The main columns which had more issues were:
        a) Peak hour, as this column had inconsistent time formats in them.
        b) Count date, as this column had day, month and year in them, of which we only wanted the year part of it.
    - 17/10/2020

      Added functions like skim(), summary(), str() to explain the datasets throughout the notebook and tidied the notebook. It also acted like a check, as it can show us if NA values are present in the dataset, of which, it enabled me to detect NA values as listed below.

      Also added a function to solve NA values for the Merged_to_Crash dataset, which had a NA value in it.

## Jemin Lee

1) General Tasks:
2) Data201 Specific Tasks:
    - 05/10/2020

      Jack and I converted TrafficService.csv, and Crash_Analysis_System.csv files to dataframes. We cleaned and filtered the dataframes, like removing NA values, removing unwanted columns, and changing variables using regex.

    - 08/10/2020

      Used concat.split function to combine street names into lists.

      Came up with a function called location_match() to return a logical result if the location strings match or not.

Chathuranga Alwis

1) General Tasks:
2) Data201 Specific Tasks:
    - 4/10/2020

      After Phua and Ziling was done with the basic cleaning and filtering on the TrafficService.csv file, I logged in and did some minor changes to the tibbles and functions. (I went through the tibble and did some changes to the data types and created a copy of the tibble with renamed column names so that it is easier to read and understand.)

      Currently waiting on Jemin and Jack to finish cleaning their dataset in order to proceed and do further cleaning and filtering.

      Looking for ways to improve the project. i.e. by introducing more data/ideas. I suggested that we include a web scraping part into the project as well.
    - 11/10/2020 & 12/10/2020

      After Jemin and Jack were done with the cleaning part, they handed it over to me to combine the two datasets. I went through their work, and tried to combine them using lists. Unfortunately, as combining the two datasets by using lists did not work and seemed to be too complicated, I ended up using SET functions and INNER-JOIN to combine the two datasets. Afterwards I did some commenting on the file, and also some structuring. Pushed it back on GIT so that someone else could further work on it.
    - 20/10/2020

      After Phua was done editing and structuring the project file, I logged in. I added the codes of the attempts we did in order to merge the original two datasets using lists. Afterwards, I restructured the file, commented, and did some minor cleaning. As a student asked about it during the presentation, I added a few more columns into our final dataset and also merged our dataset with another excel file which contains road indexes for each Region. This is so that our final dataset could be used by someone to make decisions for a given road segment. Then I pushed all my changes onto the GITHUB Repo and handed it over to Ziling to do some plots.
    - 20/10/2020 TO 23/10/2020

      Did some final touches to the file. Added some more comments, cleaned up the final notebook by removing all unnecessary stuff.

      Worked on the report, restructured it, did the final touches on the report. Finally cleaned up this document, restructured it and sent all the updated files onto our group chat. Looking forward to submitting the document today.

Ziling Huang

1) General Tasks:
2) DATA201 specific tasks:
    - 2/10/2020

      After Phua converted TrafficService.csv file into a dataframe in R. I started to clean the columns. First, I deleted rows with NAs in 'pccar', 'pclcv', 'pcheavy', since I thought they were not useful and I could not find a reasonable value to replace it. And then I used the average value of the 'pccar' column to replace the value 0 in that column. Meanwhile, I kept adding comments that help my teammates understand.
    - 3/10/2020

      After I push my version of code into gitlab. Phua found an error, and I fixed that part. I also did a check, whereby for each row, you sum up the columns, pccar, pclcv and pcheavy. Because these are percentages for the road traffic, it has to add up to 100%. I did a check and deleted all the rows that did not obey the principle. But there are only 300+ rows left which is not good for the rest of tasks.
    - 15/10/2020

      Since I cannot be in the presentation in person. I recorded my part to fit into the team. And what I recorded is to explain carefully the difficulty I faced during the wrangling. So the biggest difficulty I have is converting inconsistent format into consistent time format. And I explained each step I have done to make sure the audience could get it. which is converting inconsistent format into consistent time format. And I explained carefully to make sure the audience could get it.
    - 20/10/2020

      After all the wangling parts are finished. I am in charge of plotting parts. I plot 8 graphs to show the relationship between each column. And giving the conclusion that most of the crashes happened in mid day and in Auckland and in sealed road surfaces. And the traffic flow is related to accidents but it is not that related. After that, I also finished my part in the report to give a conclusion of what we achieved. And pasting all the graphs onto the report.

Jack Walsh

1) General Tasks:
    - Designed and managed the presentation to make sure it flowed well and we met the criteria.
3) Data201 Specific Tasks:
    - 05/10/2020

      Along with Jemin converted TrafficService.csv, and Crash_Analysis_System.csv files to dataframes, and cleaned the dataframes by removing NA values, removing unwanted columns, and changing variables using regex.