**DATA422 Group Project**

**Effects of Weather Events on ACC Accident Claims (2011 – 2017)**

**Contributors**

**Raghuraman Srinivasan #59231264**

**Gajalakshmi Sakthivel**

**Oscar Vanhanen**

**Kartik Mahajan**

**Rishil Parab**

**Introduction**

The purpose of the project is about the wrangling and insights of ACC Accident Claims Data due to the effects caused by Weather Events in New Zealand. The ACC has been a great boon touching the lives of a large number of New Zealanders, improving their quality of life by minimising the incidence and impact of injury. The causes of accidents is contributed by multiple factors and one such is due to weather events. Insurers paid out $167.6 million in weather-related [losses](https://www.insurancebusinessmag.com/nz/news/breaking-news/icnz-releases-total-weatherrelated-losses-for-2019-223522.aspx) over 2019, according to new figures released by the Insurance Council of New Zealand ([ICNZ](https://www.insurancebusinessmag.com/nz/companies/icnz/61657/)). However, there is not much precise evidence on different weather conditions and how they affect factors like age groups and gender of people, modes of transport, . In this project, we collated data and analysed the impacts of different weather events on ACC accident claims and found some interesting statistical rules. These findings are not only the results of the project but also have particular reference significance for ACC in devising new plans and NZTA in acting in advance and preventing road accidents during different weather events.

**Data Source**

1. We used ACC – Accident Claims Data 2011 – 2017 obtained from Stats NZ. The dataset obtained is in the form of an excel file. The dataset contains 8 tabs of data and information on Active Claims and New Claims. This data is untidy and needs to be wrangled to suit our project needs.

**Source:**

<https://catalogue.data.govt.nz/dataset/e9767749-01e5-4b9f-88de-10b73ae73b19/resource/f53f15d8-fb9c-474b-9fc7-d05907070262/download/road-data-publication.xlsx>

**Reason for Choosing the Data:**

This data is trustable as it is obtained from a government source. The data has apt parameters that can be utilised to associate with Weather data and establish a data model to address the research questions. The data is feature rich and could help to gain good perspective on different kinds of associations. The data provides a good scope for data wrangling which is the primary requirement of this project. The result of wrangling provides 12 different dataframe, 6 each in Active claims and New claims holding information on factors such as Age, Gender, Ethnicity, Mode of Transport, Month and Region.

**Terms and Conditions:**

Image

Link

1. We used Historical Weather Events Data 2011 – 2017 obtained from NIWA. The dataset obtained is in the form of an webpage. This data needs to be scraped from the source, and needs to be wrangled to suit our project needs.

Source: <https://hwe.niwa.co.nz/search/summary/Startdate/01-01-2011/Enddate/31-12-2018/Regions/all/Hazards/all/Impacts/all/Keywords/none/numberOfEvents/20/page/1>

**Reason for Choosing the Data:**

This data is trustable as it is obtained from a government source. The data has apt parameters that can be utilised to associate with the ACC claims data and establish a data model to address the research questions. The data has accurate information on weather events and location information that can be used for different kinds of associations. The data provides a good scope for web scraping, data wrangling which forms the primary requirements of this project. The result of wrangling provides a single data frame holding information on different Weather Events.

**Terms and Conditions:**

Image

Link

**Target Data**

The target data is obtained by merging each of the 12 ACC data frames with Weather data frame. Each merged data frame can be used to address research questions pertaining to corresponding

**Intended use of the Target Data**

One of our goals is to discover and visualise some interesting patterns in the ACC data in terms of the box office. We explored some factors like Number of Weather Events, Number of ACC Claims on different features to conclude our findings.

General Stats on Target data

1. –
2. –
3. –
4. –

RQ 1

RQ 1 – Diagram (visualization plot)

RQ 1 – Comments

RQ 2

RQ 2 – Diagram (visualization plot)

RQ 2 – Comments

**How organizations can use this data ?**

ACC can use our data to devise plans and provide advice to customers on their plans.

NZTA can use our data to alert New Zealanders, manipulate traffic and create policies to avoid loss incurred during different weather events.

**Challenges Faced**

1. Though we have successfully obtained data from two sources, unfortunately, the datasets couldn’t be able to join by “region” as the “region” is more generalised in many cases and matches specifically only in few cases. This creates a significantly high mismatch and therefore, it was impossible to join these completely different datasets by “region”. Later, “year” has been successfully used to join these datasets.
2. After scraping weather events data, it is found that for some of the fields the “region” field is missing. This is due to the absence of the region information in the corresponding headlines. Then we did a detailed exploration on the summary, applied regular expressions and fetched the affected regions by processing the summary text.
3. And still there were few cases left where the “region” is not extractable form the summary. In those cases, we found out that the summary indirectly referred to weather events that occurred throughout New Zealand. In such cases, the “region” is directly assigned as New Zealand.

We can draw a flow diagram and highlight the point of challenge

**Techniques Applied**

In our project, the following data wrangling techniques was implemented:

* Data were scrapped by using \_\_\_\_\_\_
* ACC Data is wrangled using wide to long format to suit the research questions and add more dimension of use.
* Regular expressions were used to classify the data extract them into appropriate fields.
* Two lists containing various weather events and regions is created to map (using purrr library) and classify different weather events and regions since maintaining an appropriate naming convention is essential.
* Missing data and NA values in headline were handled well by replacing them using the information present in summary.
* Paging logic was implemented on Weather events data to automate the scraping process and rbind was used to unite all the individual data frames.
* Converted data types of all the fields to the appropriate measures.
* Data was checked by using the R function glimpse and head.

#We need to add further points… I have added whatever stuck my mind… Please included if I have missed something

Add Final Data Model diagram

**Achievements & Failures**

After data scraping, wrangling and analysis, we finally built a reasonable data model. By retrieving “year”, we can connect different datasets to be analysed.

Mention about how we split into different data frames based on features in ACC data to address various RQs. Mention about automating the scraping process.

Also please add any other Achievements

However, the project also has some limitations. Due to time constraints and availability of truthful data we were only able to extract the data corresponding to years 2011 to 2017.To better analyse the weather factors that affect ACC accident claims and to provide a more reasonable future forecast, the data from an even wider timeframe might be useful.

Also, in an event of failure to finding reliable sources to extract, trustable news articles pertaining to the search terms involving “ACC” and “Weather” “Events” can be obtained. A corpus analysis can be made using this information can be utilised to make reasonable decisions and answer research questions better.

Also please add any other Failures