Contents

[Introduction 1](#_Toc116241896)

[Proposed Webservice 1](#_Toc116241897)

[Data Considerations 1](#_Toc116241898)

[Web service format 3](#_Toc116241899)

[Proposed front end 3](#_Toc116241900)

# Introduction

Student ID: 19000170

This report explains the process of selecting and creating a new web service as required for the UHI Web Services Module.

# Proposed Webservice

The city of New York is one of the most populated cities in the world, with over 8.5 million residents. As such, traffic accidents are more common than in other cities and road safety is a concern for many residents. Many factors can lead to the likelihood of an accident, including weather.

The data for this web service is available as part of Google’s Big Query offering for free. This data was used, in addition to NOAA data, to establish correlations between weather and accidents. The resulting data set showed accidents per day (and their location) and the number of deaths and injuries to motorists, cyclists, and pedestrians. As this data was for the entire of New York City (and conditions can vary from area to area) the data was then collated by borough in New York.

Therefore the new web service will:

1. Allow a user to view the historic accidents and any weather conditions which have occurred in New York. The result will also contain details of any injuries and deaths.
2. The service could be modified to bring in additional data to show the historical accidents at specific points in the city.
3. Finally, the Machine Learning model could be made available to the web service to allow a user to plug in values and to get predictions for accidents, injuries, and deaths.

The audience for this web service will vary greatly depending on which part of the service is being used.

Professional users could be those who are involved in planning the levels of resourcing for emergency vehicles, and finding if any specific times and conditions would require additional resources to be allocated. Non-professional users could be homeowners, or those looking to purchase in an area, who would like to see, historically, just how “safe” the area is. These users would also benefit from the prediction feature, but not to the same extent as professional users.

To develop the web service, only data for 2019 and 2020 will be loaded in.

## Data Considerations

As previously noted, accident data is freely available from Google as part of their BigQuery offering. This data is part of the NYPD dataset so, in theory, should be reliable. The data is part of the “**new\_york\_mv\_collisions.nypd\_mv\_collisions**” data set.

The data provides the following information:

* borough
* contributing\_factor\_vehicle\_1
* contributing\_factor\_vehicle\_2
* contributing\_factor\_vehicle\_3
* contributing\_factor\_vehicle\_4
* contributing\_factor\_vehicle\_5
* cross\_street\_name
* timestamp
* latitude
* longitude
* location
* number\_of\_cyclist\_injured
* number\_of\_cyclist\_killed
* number\_of\_motorist\_injured
* number\_of\_motorist\_killed
* number\_of\_pedestrians\_injured
* number\_of\_pedestrians\_killed
* number\_of\_persons\_injured
* number\_of\_persons\_killed
* off\_street\_name
* on\_street\_name
* unique\_key
* vehicle\_type\_code1
* vehicle\_type\_code2
* vehicle\_type\_code\_3
* vehicle\_type\_code\_4
* vehicle\_type\_code\_5
* zip\_code

It should be noted that the data is not always complete. Sometimes, only the time of the accident has been recorded, sometimes the location is missing – this will be covered later.

The proposed service won’t care about a number of these fields, specifically, it **will** care about:

* Borough
* Timestamp
* Latitude
* Longitude
* number\_of\_cyclist\_injured
* number\_of\_cyclist\_killed
* number\_of\_motorist\_injured
* number\_of\_motorist\_killed
* number\_of\_pedestrians\_injured
* number\_of\_pedestrians\_killed
* number\_of\_persons\_injured
* number\_of\_persons\_killed

It will also care about the weather conditions. Again, this is available from Google and is part of the NOAA dataset.

Putting the data and the weather together identifies a problem. Some collision data has a borough, some data has only lat and long and some have neither of these. Using some SQL with another dataset, the borough can be found based on the latitude and longitude. Those which have no location data simply have to be discarded. This is approximately 10% of all the data (though this web service is using only 2 years of data, the figure will be smaller.

With this data, we can create a data source for the web service. The data source will have 5 records for each day of the 2 years. Why five records? New York City has five boroughs. This means our web service will be able to provide borough-specific data. The JSON files used (2019.json and 2020.json) are included in the “data” directory.

This data would be enough to create a basic MVP which would be consumed by a front end, the MVP would cover point 1 of the requirements.

## Web service format

With only 2 years of data in this web service, there is sill over 3,700 records. And with 28 fields that could be returned, GraphQL is the technology to use. This is because GraphQL will return only what is requested, unlike with a REST API where the code on the front end would have to disregard information at the point of consumption.

# Proposed front end