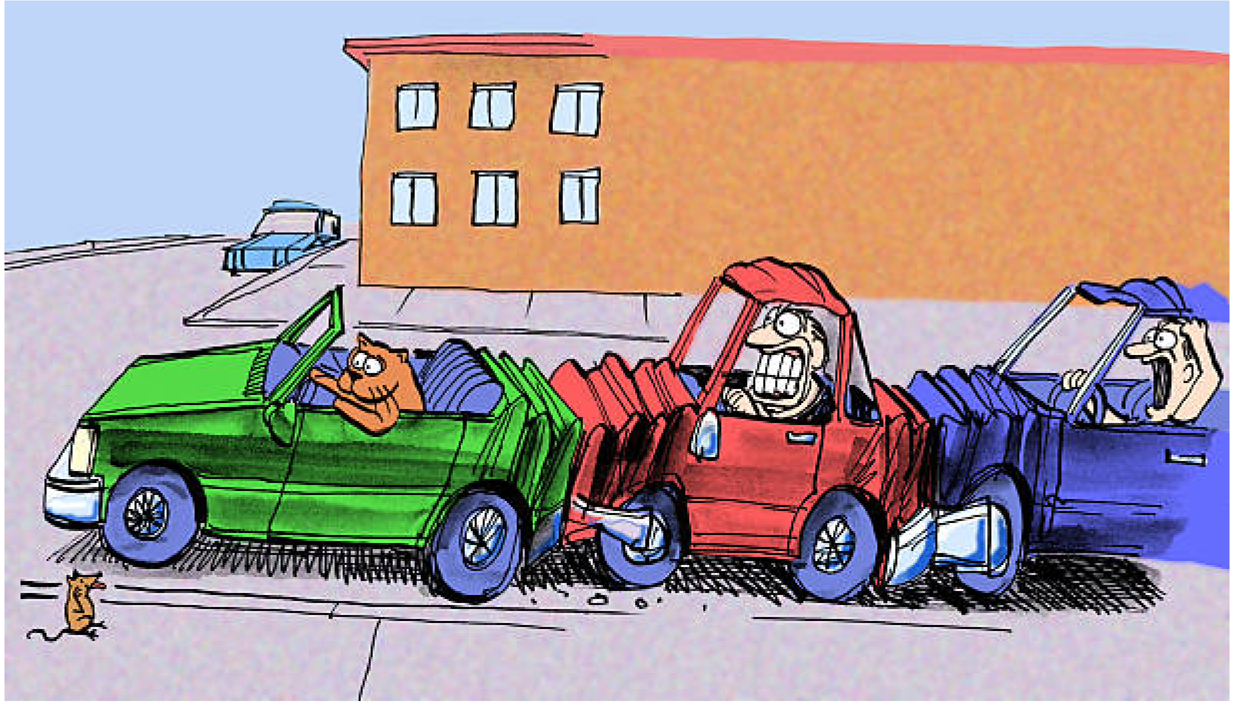
Project Report

ROAD SAFETY



Parkavi Jayachandran
Balvinder Rajbans

Sources of Data

The data is sourced from The South Australian Government Data Directory: https://data.sa.gov.au/data/dataset/road-crash-data/resource/1057e9ae-4672-4123-9c1d-1877483da401 and using the csv files from 2012_DATA_SA_Crash to 2021_DATA_SA_Crash

No	Source	Link	File
1	The South Australian Government Data Directory	https://data.sa.gov.au/data/dataset/road-crash-data/resource/1057e9ae-4672-4123-9c1d-1877483da401	2012_DATA_SA_Crash to 2021_DATA_SA_Crash

Modifications

The csv files were combined to get a larger data set to ensure accuracy of the Data Model.

1.1 MachineLearning Model

A Deep Learning Machine Learning model called Neural Network Model is used to train the system to predict the accuracy in classifying if a Crash would end up Fatal, Severe Injury, Minor Injury or Property damage only.

What is a Neural Network?
A neural network is a method in artificial intelligence that teaches computers to process data in a way that is inspired by the human brain. It is a type of machine learning process, called deep learning, that uses interconnected nodes or neurons in a layered structure that resembles the human brain.

Why Neural Networks?
Neural networks can help computers make intelligent decisions with limited human assistance. This is because they can learn and model the relationships between input and output data that are nonlinear and complex.

1.2 Database

An embedded database has been selected, SQLite3.Python SQLite3 module is used to integrate the SQLite database with Python. It is a standardised Python DBI API 2.0 and provides a straightforward and simple-to-use interface for interacting with SQLite database.

Database Identification

Database Name: crash_db

Table Name: crash_table
Primary Key: REPORT_ID

1.3 Table Schema Information

```
CREATE TABLE crash(
        REPORT_ID VARCHAR(15) PRIMARY KEY,
        Stats_Area VARCHAR(50),
        Suburb VARCHAR(100),
        Postcode INT,
        LGA_Name VARCHAR(100),
        Total_Units INT,
        Total_Cas INT,
        Total_Fats INT,
        Total_SI INT,
        Total_MI INT,
        Year  INT,
        Month VARCHAR(20),
        Day VARCHAR(20),
        Time VARCHAR(20),
        Area_Speed INT,
        Position_Type VARCHAR(20),
        Horizontal_Align VARCHAR(20),
        Vertical_Align VARCHAR(20),
        Other_Feat VARCHAR(20),
        Road_Surface VARCHAR(20),
        Moisture_Cond VARCHAR(20),
        Weather_Cond VARCHAR(20),
        DayNight VARCHAR(20),
        Crash_Type VARCHAR(20),
        Unit_Resp INT,
        Entity_Code VARCHAR(20),
        CSEF_Severity VARCHAR(20),
        Traffic_Ctrls VARCHAR(20),
        DUI_Involved  VARCHAR(20),
        Drugs_Involved VARCHAR(20),
        ACCLOC_X float8,
        ACCLOC_Y float8,
        UNIQUE_LOC float8
);
```

2.0 Data Extraction, Transformation and Analysis

2.1 Data Extraction

The data used is in csv format.

In order to extract the data from the csv files, Google CoLab Notebook is used and the library is Pandas.

Starting off by unzipping the master zip file that contains all the crash, causality and unit CSV files from the year 2012 to 2021 as separate zip files for each year. Each year's zip file is unzipped only for the crash CSV files and all other zip files are removed by using the !rm *.zip function. The CSVs are then read into a Pandas Dataframe using the 'read_csv' function inside Pandas, the relevant data can be extracted specific to our needs. The file crash_CSV has been used.

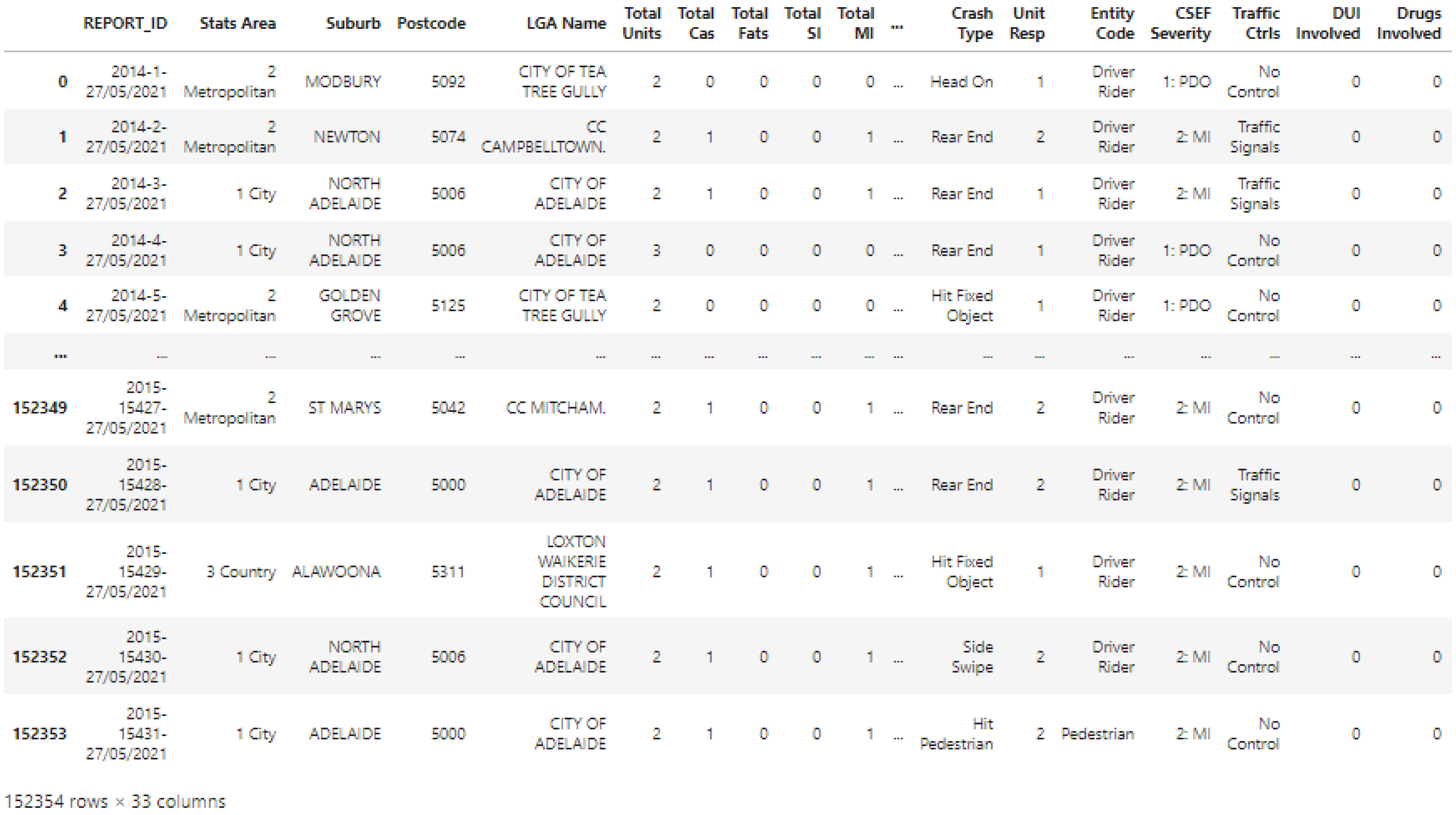
The extracted dataframe is then housed in a relational database, SQLite3 and retrieved from the database for further analysis.

2.2 Data Transformation

2.2.1 Crash Data Filtering

- Narrow Down Dataset

The crash dataframe is created by reading into the crash csv file and the extraction of data is processed and additional columns not required are dropped ('REPORT_ID', 'Suburb', 'Postcode', 'LGA Name', 'Total Units', 'Total Cas', 'Total Fats', 'Total SI' ,'Total MI', 'Year','ACCLOC_X','ACCLOC_Y','UNIQUE_LOC').
The extracted data is stored in a new Dataframe 'crash_df_filtered_1' and will be used to load into the database.



	REPORT_ID	Stats Area	Suburb	Postcode	LGA Name	Total Units	Total Cas	Total Fats	Total SI	Total MI	...	Crash Type	Unit Resp	Entity Code	CSEF Severity	Traffic Ctrls	DUI Involved	Drugs Involved
0	2014-1-27/05/2021	2 Metropolitan	MODBURY	5092	CITY OF TEA TREE GULLY	2	0	0	0	0	...	Head On	1	Driver Rider	1: PDO	No Control	0	0
1	2014-2-27/05/2021	2 Metropolitan	NEWTON	5074	CC CAMPBELLTOWN.	2	1	0	0	1	...	Rear End	2	Driver Rider	2: MI	Traffic Signals	0	0
2	2014-3-27/05/2021	1 City	NORTH ADELAIDE	5006	CITY OF ADELAIDE	2	1	0	0	1	...	Rear End	1	Driver Rider	2: MI	Traffic Signals	0	0
3	2014-4-27/05/2021	1 City	NORTH ADELAIDE	5006	CITY OF ADELAIDE	3	0	0	0	0	...	Rear End	1	Driver Rider	1: PDO	No Control	0	0
4	2014-5-27/05/2021	2 Metropolitan	GOLDEN GROVE	5125	CITY OF TEA TREE GULLY	2	0	0	0	0	...	Hit Fixed Object	1	Driver Rider	1: PDO	No Control	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
152349	2015-15427-27/05/2021	2 Metropolitan	ST MARYS	5042	CC MITCHAM.	2	1	0	0	1	...	Rear End	2	Driver Rider	2: MI	No Control	0	0
152350	2015-15428-27/05/2021	1 City	ADELAIDE	5000	CITY OF ADELAIDE	2	1	0	0	1	...	Rear End	2	Driver Rider	2: MI	Traffic Signals	0	0
152351	2015-15429-27/05/2021	3 Country	ALAWOONA	5311	LOXTON WAIKERIE DISTRICT COUNCIL	2	1	0	0	1	...	Hit Fixed Object	1	Driver Rider	2: MI	No Control	0	0
152352	2015-15430-27/05/2021	1 City	NORTH ADELAIDE	5006	CITY OF ADELAIDE	2	1	0	0	1	...	Side Swipe	2	Driver Rider	2: MI	No Control	0	0
152353	2015-15431-27/05/2021	1 City	ADELAIDE	5000	CITY OF ADELAIDE	2	1	0	0	1	...	Hit Pedestrian	2	Pedestrian	2: MI	No Control	0	0

152354 rows × 33 columns

2.2.2 Crash Data Preprocessing

The Dtype of the column variables are found and converted into appropriate Dtype as per requirement.
The ‘Time’ column which is an Object(String) is converted into DateTime and then converted into 24 Hours format from 12 Hours so that the data can be binned into 24 bins, one for each hour.

The Target column ‘CSEF Severity’ is converted into Numerical values.

2.3 Data Analysis

As the Dataframe is now cleaned and transformed ready to be loaded into the machine learning model, the Deep Neural Network Model is used to predict the accuracy of the Target variable classification.

The columns with Object Dtype are One-Hot Encoded and a new dataframe is created with the one_hot_encoded(ohe) columns using the pd.get_dummies function.

The preprocessed data is then Split into Target and Features.
The ‘CSEF Severity’ column is taken as ‘y’ for the Target Variable and all other columns are taken as ‘x’ for the Feature Variables.

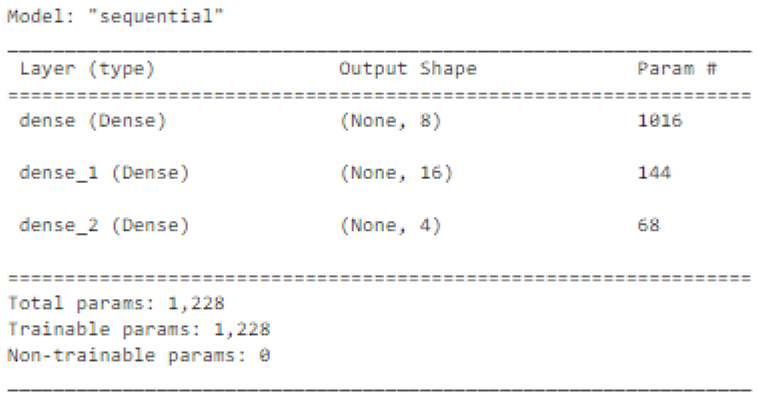
The data is then split into a Training and Testing dataset using the train_test_split function imported from the Scikit-learn Library.

The train and test data is then scaled using the Standardscaler instance.

The scaled and transformed data is then passed through the neural network model, the number of input features and the hidden node in each layer is defined.

Two hidden layers are added to the Neural network with the ‘ReLu’ activation function.

‘Softmax’ is used as the activation function for the Output Layer with 4 output units as per the unique values in our Target column.

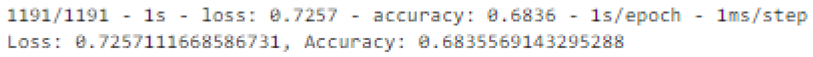


```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 8)                 1016

 dense_1 (Dense)             (None, 16)                144

 dense_2 (Dense)             (None, 4)                 68

=================================================================
Total params: 1,228
Trainable params: 1,228
Non-trainable params: 0
_________________________________________________________________
```

The neural network model is then compiled with CategoricalCrossEntroy for the Loss and ‘Adam’ as Optimizer. The learning rate is set at 0.001 and the metrics is set for ‘Accuracy’.

The model is then trained using the fit-model function with the number of epochs set to 10.

The model_loss and model_accuracy is then calculated using the nn.evaluate function.



```
1191/1191 - 1s - loss: 0.7257 - accuracy: 0.6836 - 1s/epoch - 1ms/step
Loss: 0.7257111668586731, Accuracy: 0.6835569143295288
```

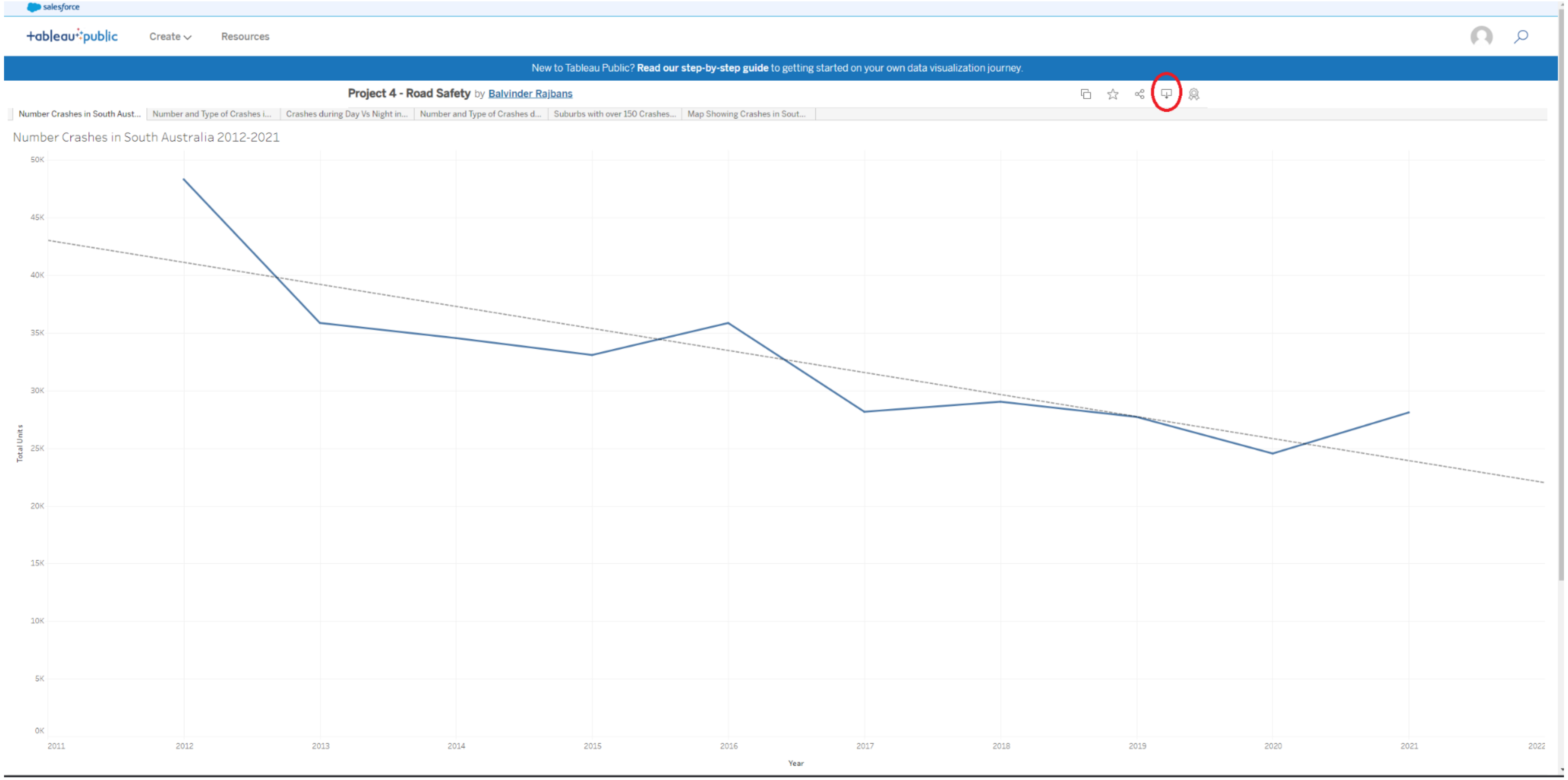
The model was then trained by tuning the hyperparameters to see if the accuracy can be increased but no further improvement was found with the given dataset.

3.0 Visualisations

Public Tableau has been used to generate all the visualisations.

https://public.tableau.com/views/Project4-RoadSafety/NumberCrashesinSouthAustralia2012-2021?:language=en-US&:display_count=n&:origin=viz_share_link

You can access the Viz using the link above and click the download button to access the file. You are able to edit and follow the steps to modify some of the visualisations where the year filter is used.



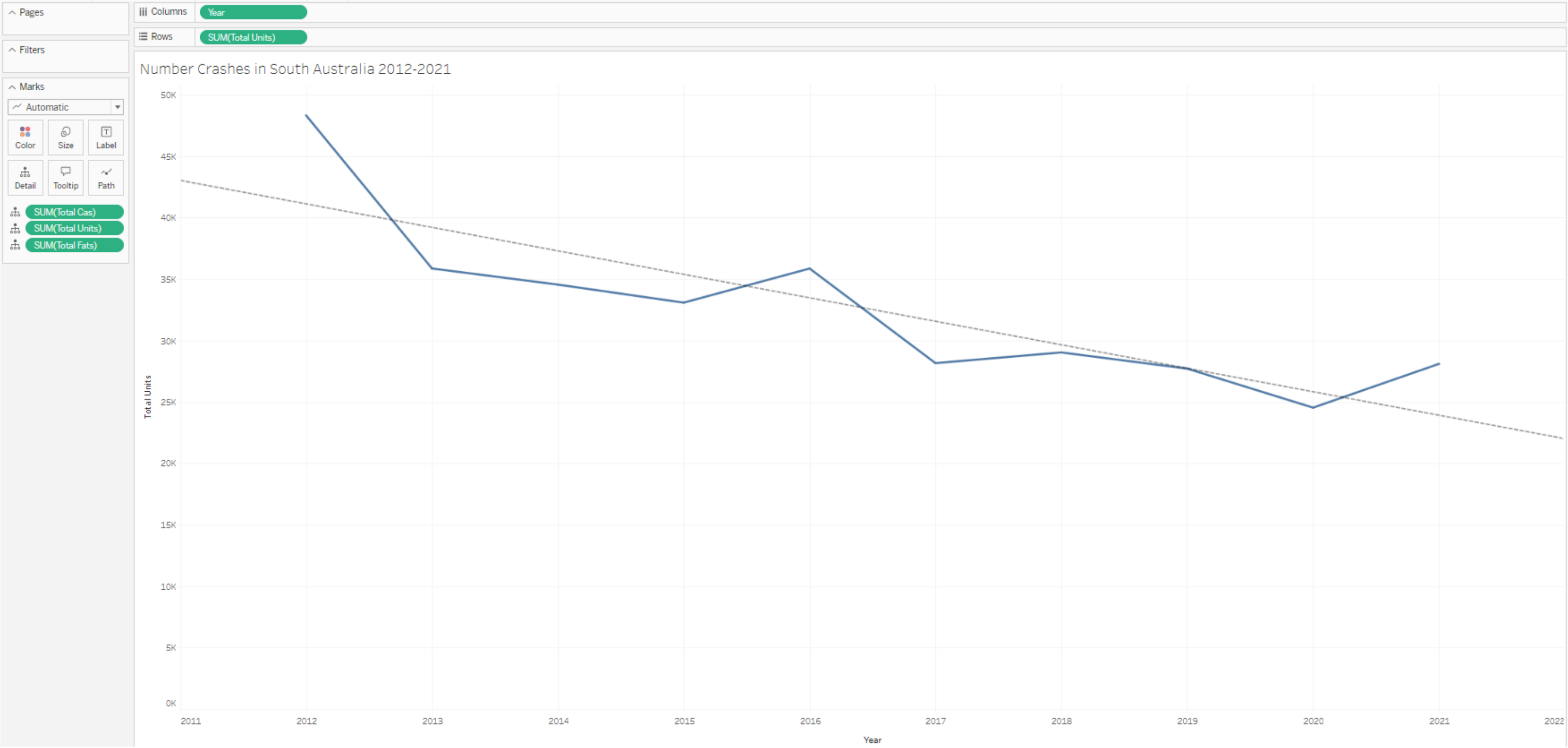
New to Tableau Public? **Read our step-by-step guide** to getting started on your own data visualization journey.

Project 4 - Road Safety by Balvinder Rajbans

salesforce
+ableau public
Create
Resources
Number Crashes in South Aust...
Number and Type of Crashes i...
Crashes during Day Vs Night in...
Number and Type of Crashes d...
Suburbs with over 150 Crashes...
Map Showing Crashes in Sout...
Number Crashes in South Australia 2012-2021
50K
45K
40K
35K
30K
25K
20K
15K
10K
5K
0K
Total Units
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
Year

1. <u>Number of crashes in South Australia from 2012-2021</u>

For this we have used the Year in the columns and Total Units in rows with the Detail of Total Casualties and Total Fatalities to show when hovering over.
There is a trend line to show the trend of the crashes over the years.



Columns: Year

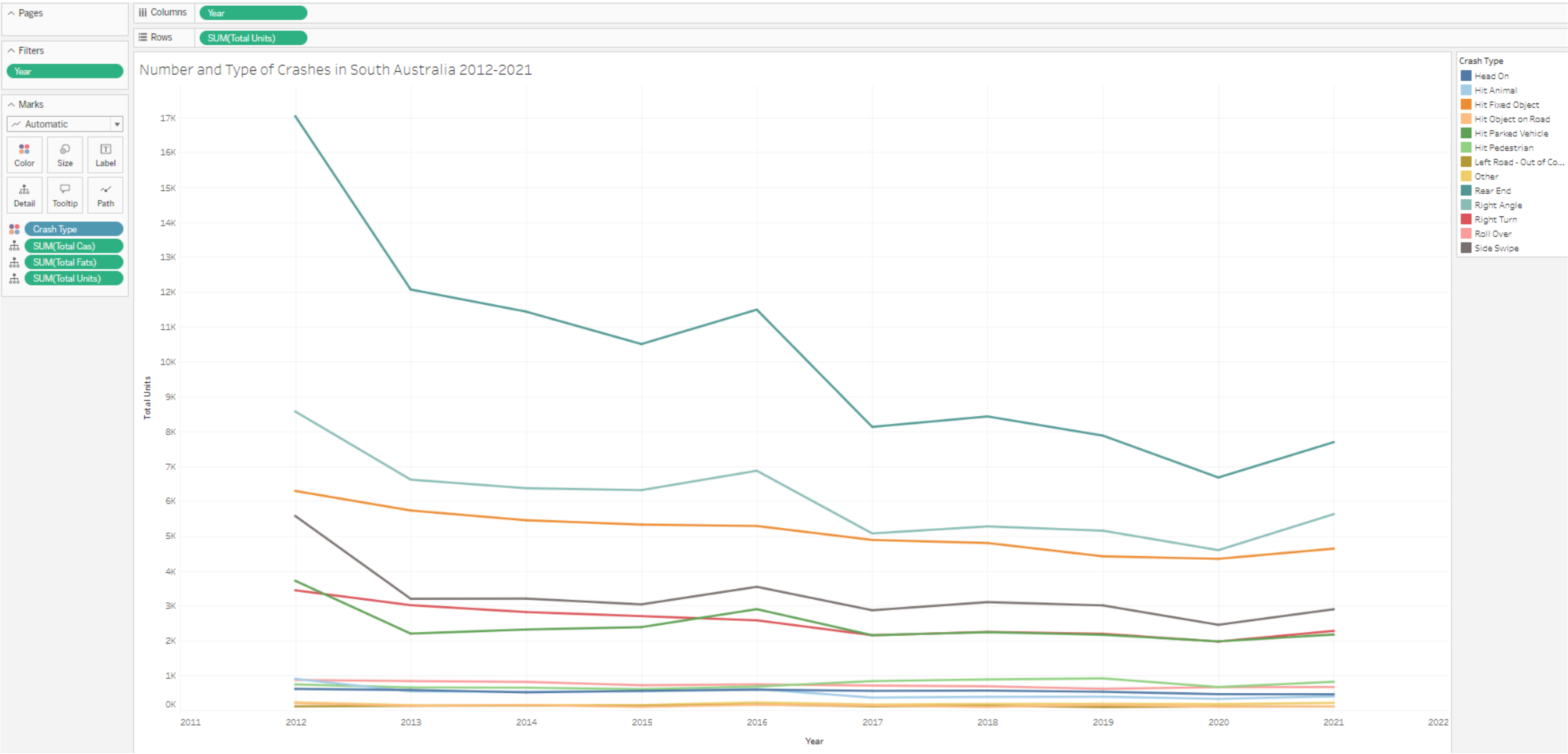
Rows: SUM(Total Units)

Marks: Automatic

SUM(Total Cas)
SUM(Total Units)
SUM(Total Fats)

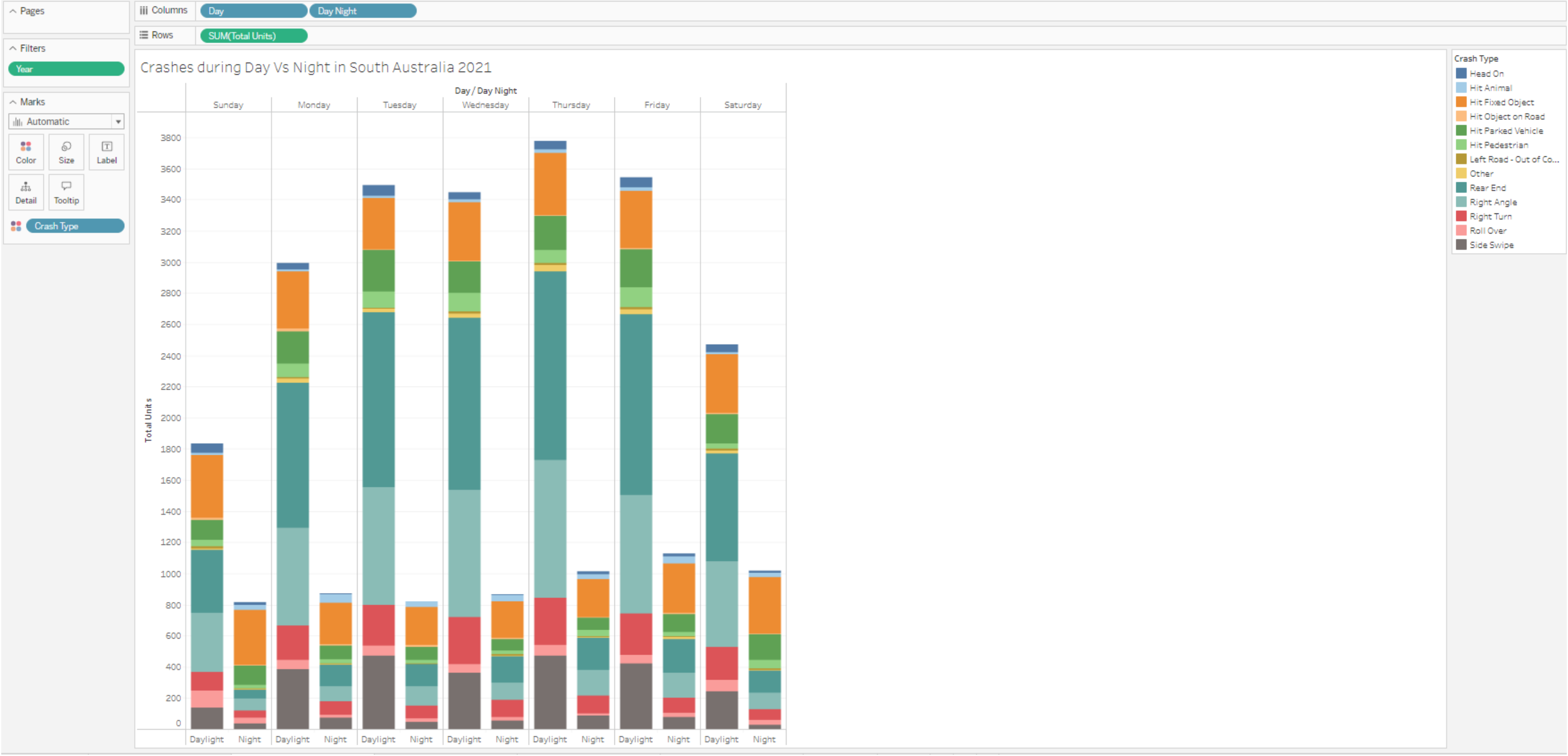
2. Number and Type of crashes in South Australia from 2012-2021

For this we have used the Year in the columns and Total Units in rows with the Detail of Total Units, Total Causalities and Total Fatalities to show when hovering over.
The crash type is colour coded.



3. Crashes during Day Vs Night in South Australia 2021

For this we have used the Day & Day Night in the columns and Total Units in rows with the Detail of Total Units to show when hovering over.
The crash type is colour coded.
There is a filter for the year to be able to view other years as the current data used is for 2021.



Columns: Day, Day Night

Rows: SUM(Total Units)

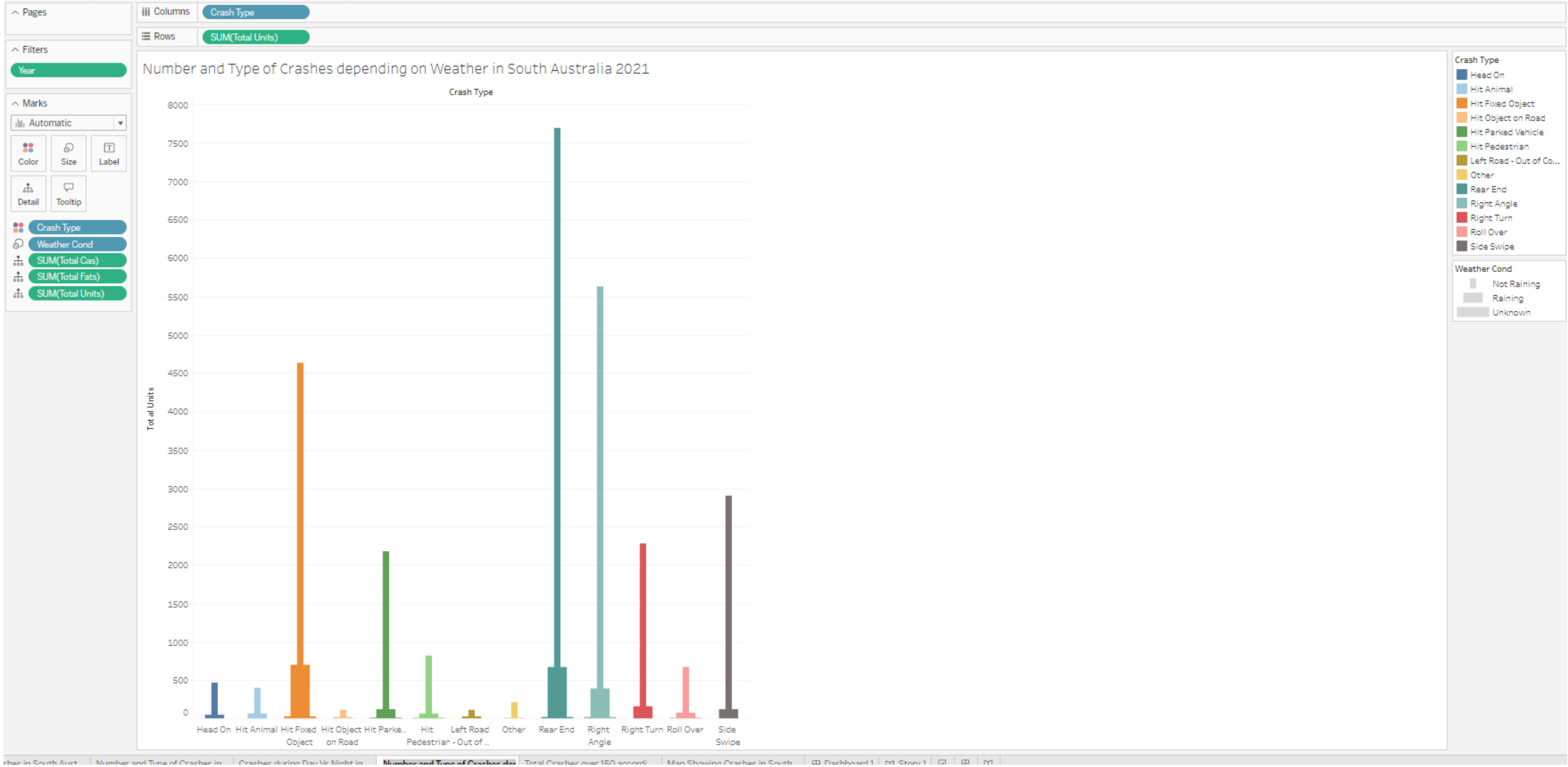
Filters: Year

Marks: Automatic

Crash Type

4. Number and Type of Crashes depending on Weather in South Australia 2021

For this we have used the Crash Type in the columns and Suburb in rows with the Detail of Max Area Speed, Total Units, Total Casualties and Total Fatalities to show when hovering over.
There is a filter for the year and total units to be able to view other years as the current data used is for 2021.

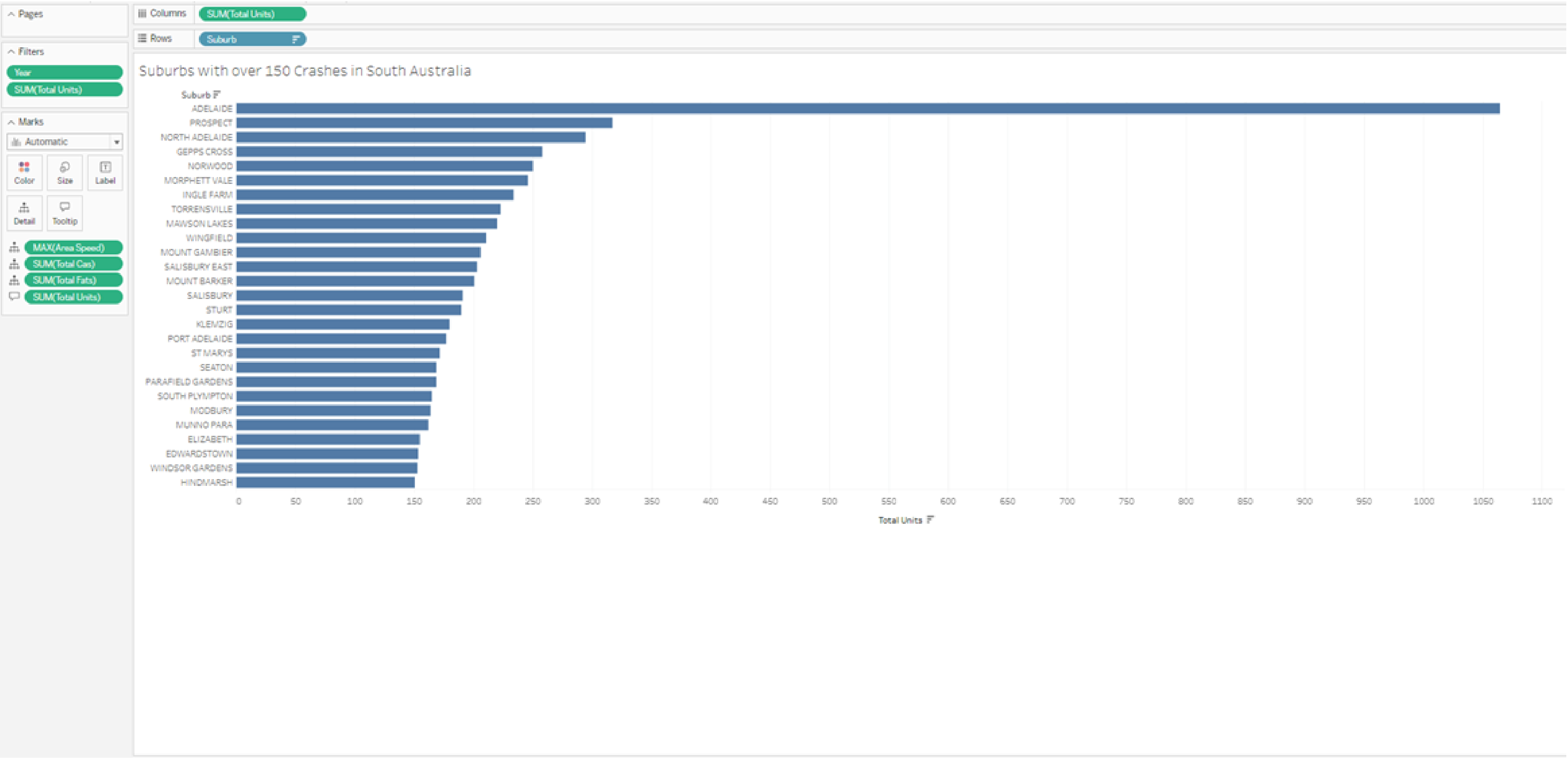


Columns: Crash Type
Rows: SUM(Total Units)
Filters: Year
Marks: Automatic
Crash Type
Weather Cond
SUM(Total Cas)
SUM(Total Fats)
SUM(Total Units)

Number and Type of Crashes depending on Weather in South Australia 2021

5. Suburbs with over 150 Crashes in South Australia 2021

For this we have used the Total Units in the columns and Total Units with the Detail of Total Units, Total Casualties and Total Fatalities to show when hovering over.
The crash type is colour coded.
There is a filter for the year to be able to view other years as the current data used is for 2021. The filter for the number of crashes is over 150.



Columns: SUM(Total Units)

Rows: Suburb

Filters: Year, SUM(Total Units)

Marks: Automatic

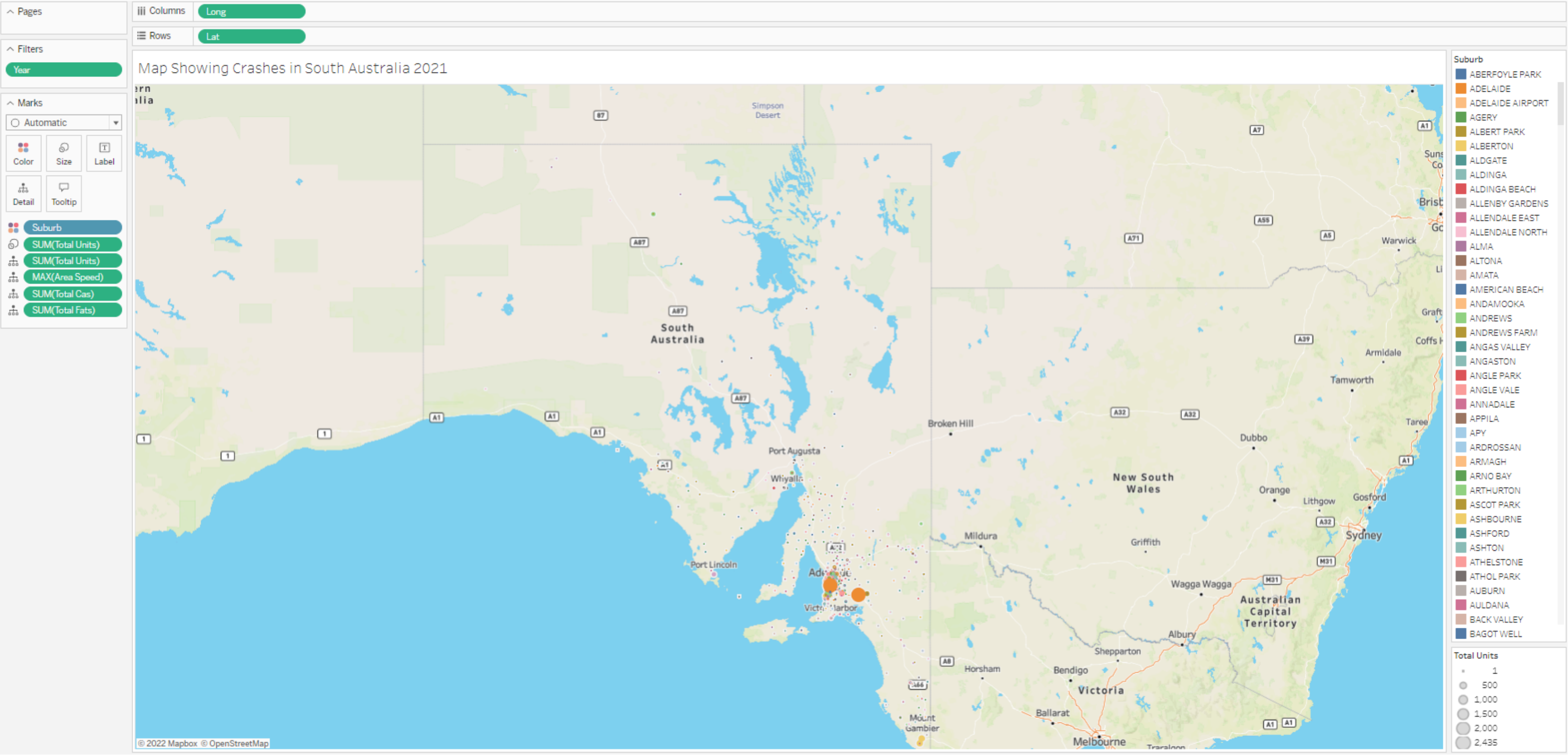
MAX(Area Speed)
SUM(Total Cas)
SUM(Total Fats)
SUM(Total Units)

Suburbs with over 150 Crashes in South Australia

6. Map Showing Crashes in South Australia 2021

To get the accurate Longitude and Latitude we merged the combined_crash.csv file and the postcodes_SA.csv file using Suburb and Locality respectively.

For this we have used the Long from postcodes_SA.csv in the columns and Lat from postcodes_SA.csv in rows. Change the items to dimensions to get accurate results and from the map tab select the Background Maps ass Streets.
We have used the Detail of Total Units, Total Casualties and Total Fatalities to show when hovering over.
The Suburb is colour coded and Total Units increases in size as the number of crashes in a certain suburb increase.
There is a filter for the year to be able to view other years as the current data used is for 2021.



Pages

Filters
Year

Marks
○ Automatic

Suburb
SUM(Total Units)
SUM(Total Units)
MAX(Area Speed)
SUM(Total Cas)
SUM(Total Fats)

Columns: Long
Rows: Lat

Map Showing Crashes in South Australia 2021

Suburb
ABERFOYLE PARK
ADELAIDE
ADELAIDE AIRPORT
AGERY
ALBERT PARK
ALBERTON
ALDGATE
ALDINGA
ALDINGA BEACH
ALLENBY GARDENS
ALLENDALE EAST
ALLENDALE NORTH
ALMA
ALTONA
AMATA
AMERICAN BEACH
ANDAMOOKA
ANDREWS
ANDREWS FARM
ANGAS VALLEY
ANGASTON
ANGLE PARK
ANGLE VALE
ANNADALE
APPILA
APY
ARDROSSAN
ARMAGH
ARNO BAY
ARTHURTON
ASCOT PARK
ASHBOURNE
ASHFORD
ASHTON
ATHELSTONE
ATHOL PARK
AUBURN
AULDANA
BACK VALLEY
BAGOT WELL

Total Units
1
500
1,000
1,500
2,000
2,435