ACCIDENT RISK INDEX

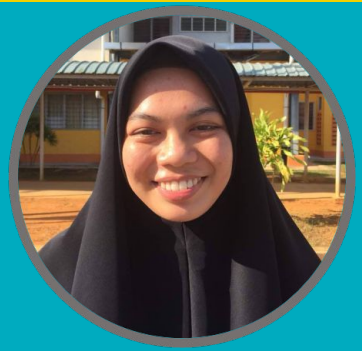
Our Team

Nik

Muhsin

Sarah

Aisyah

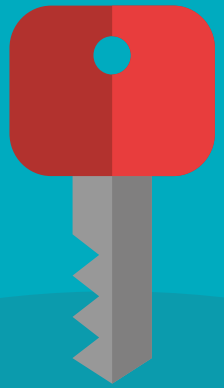


Tacha

Table of Contents

01

Background



Road traffic accidents continue rising despite broad efforts to control and ameliorate the problem. Every year, nearly 1.25 million people are killed in automobile accidents. The cost of deaths and injuries in low- and middle-income countries is roughly 3% of GDP. Southeast Asia continues to outperform Europe in terms of road safety, as measured by the fatality rate (deaths in road crashes per 100,000 persons). Low- and middle-income countries have roughly double the number of fatalities as developed countries.

DATA DESCRIPTION

Variable	Data Type	Description
Accident_ID	integer	The numbering of the datasets.
Police_Force	integer	The number of police forces in the area.
Number_of_Vehicles	integer	The number of vehicles.
Number_of_Casualties	integer	The number of casualties.
Date	date	Date when the accident happened.
Day_of_Week	integer	The day of the accident happened 1- Monday 2- Tuesday 3- Wednesday 4- Thursday 5- Friday 6- Saturday 7- Sunday
Time	time	Time the accident happened.
Local_Authority_(District)	integer	The number of local authorities within the district.

Local_Authority(Highway)	string	The local authority code of the highway.
1st_Road_Class	integer	The first road class.
1st_Road_Number	integer	The first road number.
Road_Type	string	The type of the roads.
Speed_limit	integer	The speed limit of the area.
2nd_Road_Class	integer	The second road class.
2nd_Road_Number	integer	The second road number.
Pedestrian_Crossing-Human_Control	string	Pedestrian crossing without using the facilities for crossing roads.
Pedestrian_Crossing-Physical_Facilities	string	Pedestrian crossing with the usage of facilities for crossing.
Light_Conditions	string	The light condition of the road.
Weather_Conditions	string	Weather conditions that consist of fine without high winds, fine with high winds, raining without high winds, raining with high winds and snowing without high winds.

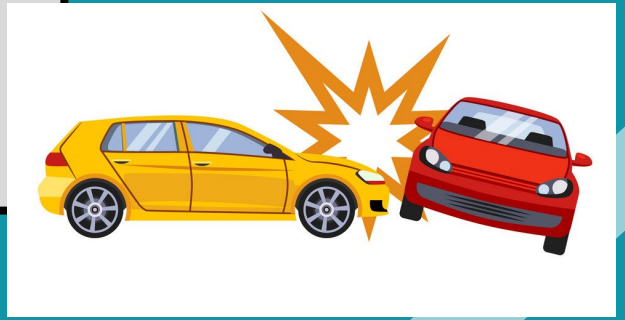
Road_Surface_Conditions	string	The road surface consists of dry, wet/damp, frost/ice, snow and flood that over 3 cm of water.
Special_Conditions_at_Site	string	Whether there is roadworks or road surface defective.
Carriageway_Hazards	string	Carriageway hazards that involve during the accident.
Urban_or_Rural_Area	integer	1 indicates an urban area, 2 indicates a rural area.
Did_Police_Officer_Attend_Scene_of_Accident	string	Whether the police were on site when the accident happened.
state	string	States of the accident area.
postcode	string	Postcode of the accident area.
country	string	The country where the accidents occur which is the United Kingdom.

PROBLEMS TO BE SOLVED

In this case study, we aim to aid in the analysis of factors that may correlate to accident rates. We believed that this study would benefit the authority in wisely organizing the roads and road users in taking steps to avoid accidents.

OBJECTIVES

1. To observe the state that has the highest accident rate.
2. To identify factors that affect the accident rate.
3. To observe the effectiveness of police authority on roads.



DATA SCHEMA

```
population_df = pd.read_csv('population.csv')
population_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8035 entries, 0 to 8034
Data columns (total 10 columns):
 #   Column                                                                                                  Non-Null Count
Dtype
---  ------                                                                                                  --------------
-----
 0   postcode                                                                                                8035 non-null
object
 1   Rural Urban                                                                                             8035 non-null
object
 2   Variable: All usual residents; measures: Value                                                          8035 non-null
int64
 3   Variable: Males; measures: Value                                                                        8035 non-null
int64
 4   Variable: Females; measures: Value                                                                      8035 non-null
int64
 5   Variable: Lives in a household; measures: Value                                                         8035 non-null
int64
 6   Variable: Lives in a communal establishment; measures: Value                                            8035 non-null
int64
 7   Variable: Schoolchild or full-time student aged 4 and over at their non term-time address; measures: Value  8035 non-null
int64
 8   Variable: Area (Hectares); measures: Value                                                              8035 non-null
float64
 9   Variable: Density (number of persons per hectare); measures: Value                                      8035 non-null
float64
dtypes: float64(2), int64(6), object(2)
memory usage: 627.9+ KB
```

```
roads_df = pd.read_csv('roads_network.csv')
roads_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 91566 entries, 0 to 91565
Data columns (total 8 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   WKT                                   91566 non-null  object
 1   roadClassi                            90352 non-null  object
 2   roadFuncti                            90352 non-null  object
 3   formOfWay                             90352 non-null  object
 4   length                                90352 non-null  float64
 5   primaryRou                            90352 non-null  float64
 6   distance to the nearest point on rd   90409 non-null  float64
 7   postcode                              91566 non-null  object
dtypes: float64(3), object(5)
memory usage: 5.6+ MB
```

```
sample_df = pd.read_csv('sample_submission.csv')
sample_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 49772 entries, 0 to 49771
Data columns (total 2 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   postcode             49772 non-null  object
 1   Accident_risk_index  49772 non-null  int64
dtypes: int64(1), object(1)
memory usage: 777.8+ KB
```

DATA SCHEMA

```
test_df = pd.read_csv('test.csv')
test_df.info()
```

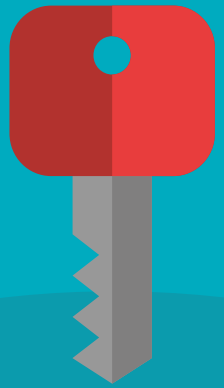
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 121259 entries, 0 to 121258
Data columns (total 27 columns):
 #   Column                                       Non-Null Count   Dtype
---  ------                                       --------------   -----
 0   Accident_ID                                  121259 non-null  int64
 1   Police_Force                                 121259 non-null  int64
 2   Number_of_Vehicles                           121259 non-null  int64
 3   Number_of_Casualties                         121259 non-null  int64
 4   Date                                         121259 non-null  object
 5   Day_of_Week                                  121259 non-null  int64
 6   Time                                         121258 non-null  object
 7   Local_Authority_(District)                   121259 non-null  int64
 8   Local_Authority_(Highway)                    121259 non-null  object
 9   1st_Road_Class                               121259 non-null  int64
 10  1st_Road_Number                              121259 non-null  int64
 11  Road_Type                                    121259 non-null  object
 12  Speed_limit                                  121259 non-null  int64
 13  2nd_Road_Class                               121259 non-null  int64
 14  2nd_Road_Number                              121259 non-null  int64
 15  Pedestrian_Crossing-Human_Control            121259 non-null  object
 16  Pedestrian_Crossing-Physical_Facilities      121259 non-null  object
 17  Light_Conditions                             121259 non-null  object
 18  Weather_Conditions                           121259 non-null  object
 19  Road_Surface_Conditions                      121220 non-null  object
 20  Special_Conditions_at_Site                   121249 non-null  object
 21  Carriageway_Hazards                          121259 non-null  object
 22  Urban_or_Rural_Area                          121259 non-null  int64
 23  Did_Police_Officer_Attend_Scene_of_Accident  121259 non-null  object
 24  state                                        121259 non-null  object
 25  postcode                                     121259 non-null  object
 26  country                                      121259 non-null  object
dtypes: int64(12), object(15)
memory usage: 25.0+ MB
```

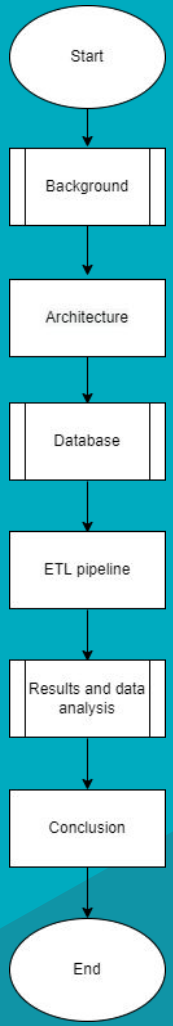
```
train_df = pd.read_csv('train.csv')
train_df.info()
```

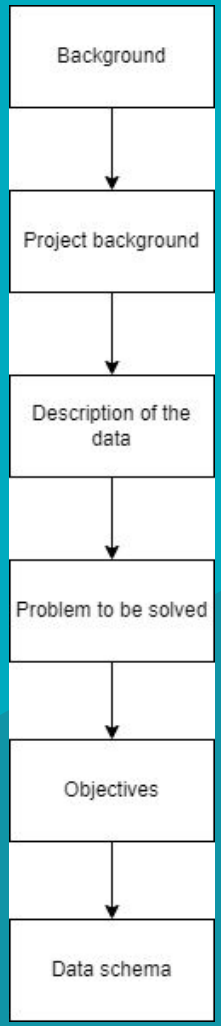
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 478741 entries, 0 to 478740
Data columns (total 27 columns):
 #   Column                                       Non-Null Count   Dtype
---  ------                                       --------------   -----
 0   Accident_ID                                  478741 non-null  int64
 1   Police_Force                                 478741 non-null  int64
 2   Number_of_Vehicles                           478741 non-null  int64
 3   Number_of_Casualties                         478741 non-null  int64
 4   Date                                         478741 non-null  object
 5   Day_of_Week                                  478741 non-null  int64
 6   Time                                         478727 non-null  object
 7   Local_Authority_(District)                   478741 non-null  int64
 8   Local_Authority_(Highway)                    478741 non-null  object
 9   1st_Road_Class                               478741 non-null  int64
 10  1st_Road_Number                              478741 non-null  int64
 11  Road_Type                                    478741 non-null  object
 12  Speed_limit                                  478741 non-null  int64
 13  2nd_Road_Class                               478741 non-null  int64
 14  2nd_Road_Number                              478741 non-null  int64
 15  Pedestrian_Crossing-Human_Control            478741 non-null  object
 16  Pedestrian_Crossing-Physical_Facilities      478741 non-null  object
 17  Light_Conditions                             478741 non-null  object
 18  Weather_Conditions                           478741 non-null  object
 19  Road_Surface_Conditions                      478289 non-null  object
 20  Special_Conditions_at_Site                   478678 non-null  object
 21  Carriageway_Hazards                          478741 non-null  object
 22  Urban_or_Rural_Area                          478741 non-null  int64
 23  Did_Police_Officer_Attend_Scene_of_Accident  478741 non-null  object
 24  state                                        478741 non-null  object
 25  postcode                                     478741 non-null  object
 26  country                                      478741 non-null  object
dtypes: int64(12), object(15)
memory usage: 98.6+ MB
```

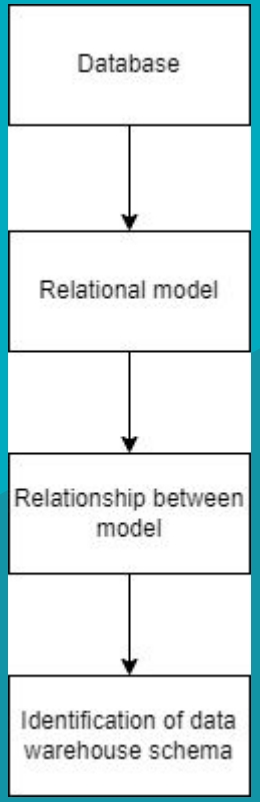
02

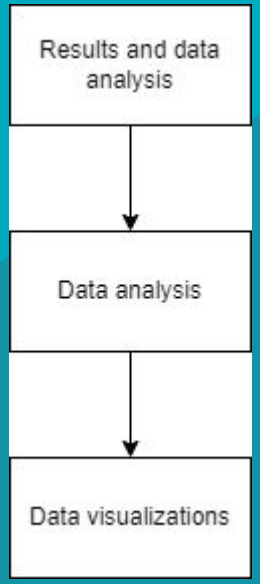
Architecture

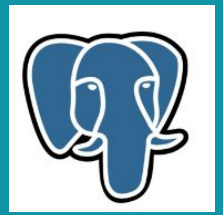



Start
Background
Architecture
Database
ETL pipeline
Results and data analysis
Conclusion
End


Background
Project background
Description of the data
Problem to be solved
Objectives
Data schema

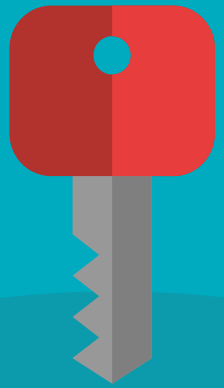

Database
Relational model
Relationship between model
Identification of data warehouse schema

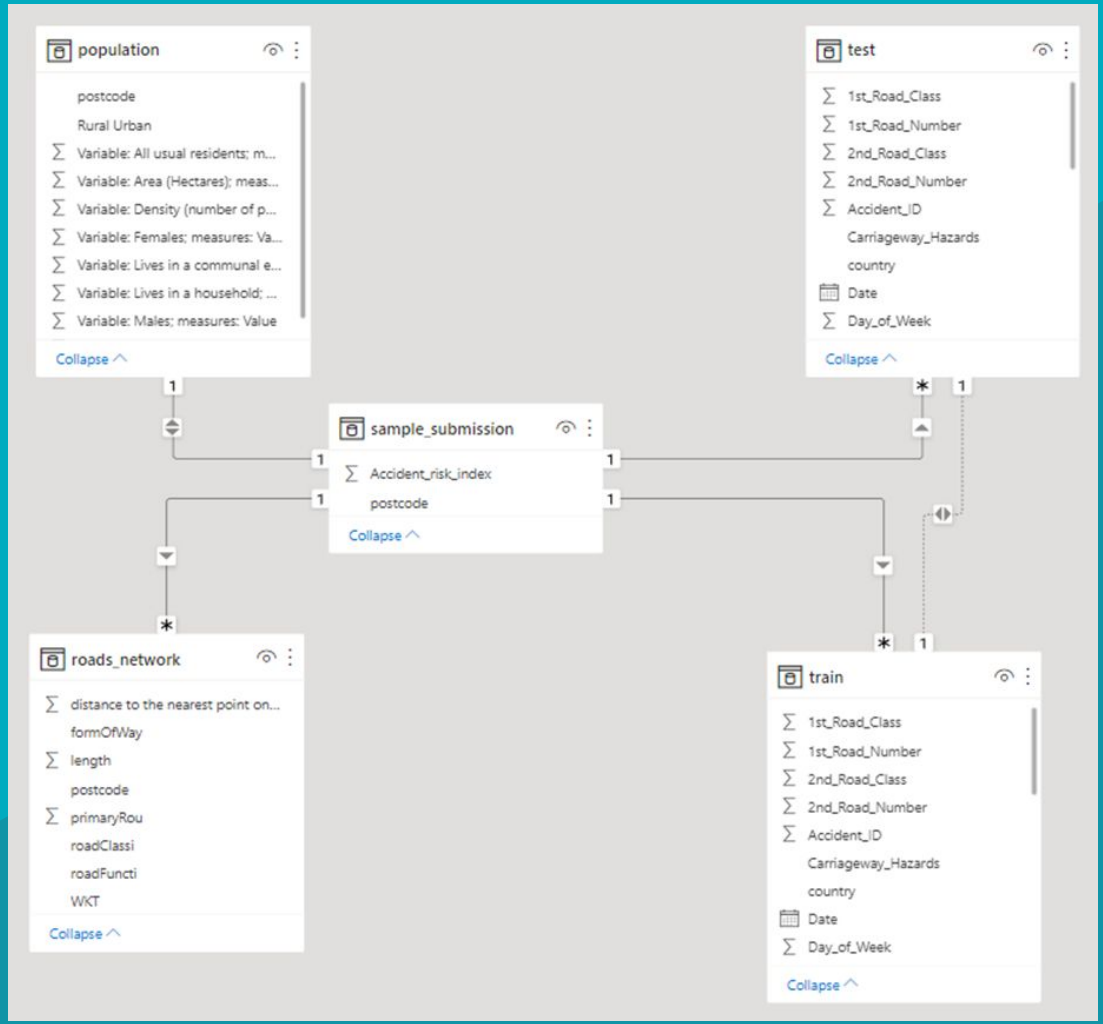

Results and data analysis
Data analysis
Data visualizations



03

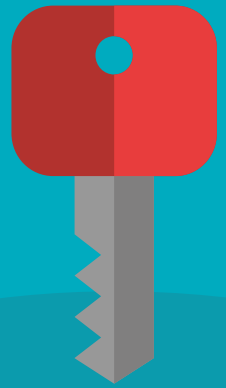
Database

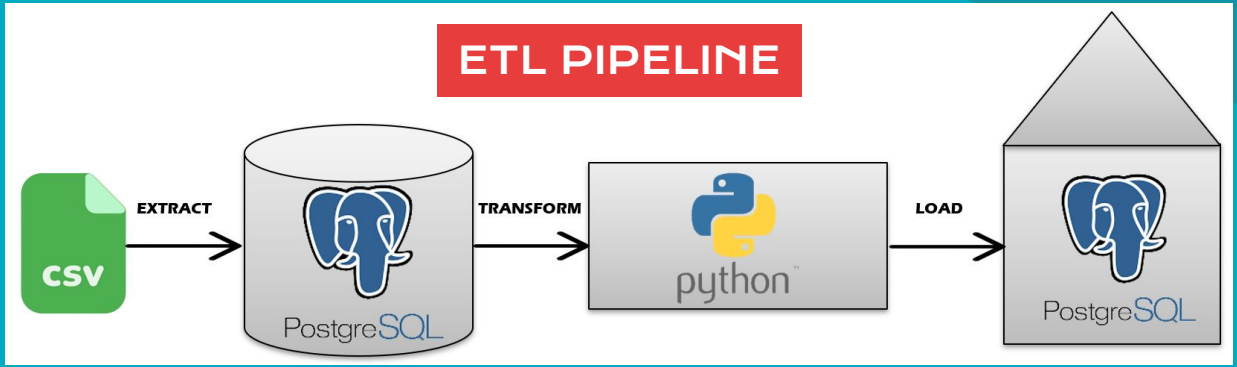



population
postcode
Rural Urban
Variable: All usual residents; m...
Variable: Area (Hectares); meas...
Variable: Density (number of p...
Variable: Females; measures: Va...
Variable: Lives in a communal e...
Variable: Lives in a household; ...
Variable: Males; measures: Value
Collapse
test
1st_Road_Class
1st_Road_Number
2nd_Road_Class
2nd_Road_Number
Accident_ID
Carriageway_Hazards
country
Date
Day_of_Week
Collapse
sample_submission
Accident_risk_index
postcode
Collapse
roads_network
distance to the nearest point on...
formOfWay
length
postcode
primaryRou
roadClassi
roadFuncti
WKT
Collapse
train
1st_Road_Class
1st_Road_Number
2nd_Road_Class
2nd_Road_Number
Accident_ID
Carriageway_Hazards
country
Date
Day_of_Week
Collapse

04

ETL pipeline





Extract

CSV file is imported to PostgreSQL and then connected to Jupyter Notebook

Transform

1. Check and drop the null values
2. Drop unknown weather

```
newTrain.drop(newTrain[newTrain['weathercondition'] == 'Unknown'].index, inplace = True)
newTest.drop(newTest[newTest['weathercondition'] == 'Unknown'].index, inplace = True)
```

Figure 4.2.2.5 Remove Unknown Weather

3. Drop unnecessary column

```
newPopulation = populationDF.drop(columns = ['ruralurban'])
newRoads2 = newRoads.drop(columns = ['roadclassi', 'roadfunction'])
```

Figure 4.2.2.6 Drop Columns

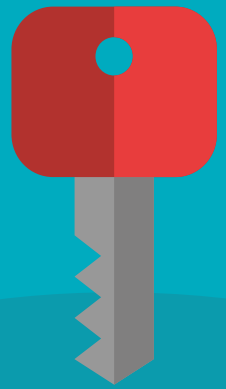
4. Generate new csv to load the data

Load

Data that has been cleaned being load back into PostgreSQL directly from Jupyter Notebook

05

Results and data analysis



DATA ANALYSIS

SLICING

Query Editor | Query History

```
1  select states, sum(num_casualties)
2  from train
3  where roadtype = 'Dual carriageway' and speedlimit = '20'
4  group by states;
```

Data Output | Explain | Messages | Notifications

	states text	sum bigint
1	Alba / Scotland	392
2	Cymru / Wales	42
3	England	1400

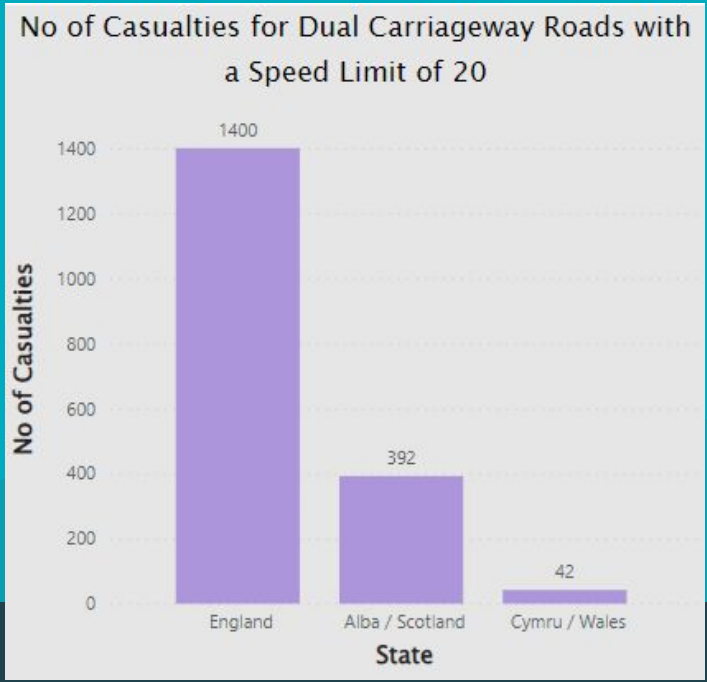
ROLL-UP

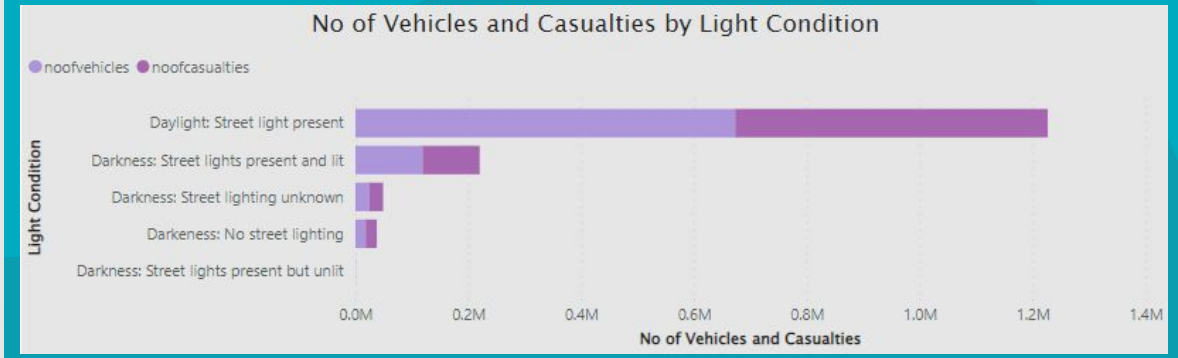
Query Editor | Query History

```
1  select num_vehicles, lightcond, sum(num_casualties)
2  from train
3  group by num_vehicles, lightcond
4  order by sum(num_casualties);
```

Data Output | Explain | Messages | Notifications

	num_vehicles integer	lightcond text	sum bigint
1	3	Darkness: Street lights present but unlit	21
2	2	Darkness: Street lights present but unlit	96
3	1	Darkness: Street lights present but unlit	147
4	4	Darkness: Street lighting unknown	433
5	4	Darkeness: No street lighting	451
6	3	Darkeness: No street lighting	984
7	3	Darkness: Street lighting unknown	1173
8	4	Darkness: Street lights present and lit	1674
9	2	Darkeness: No street lighting	4784
10	3	Darkness: Street lights present and lit	6058
11	2	Darkness: Street lighting unknown	8095
12	4	Daylight: Street light present	8399
13	1	Darkeness: No street lighting	13454
14	1	Darkness: Street lighting unknown	14991
15	3	Daylight: Street light present	32078
16	1	Darkness: Street lights present and lit	40446
17	2	Darkness: Street lights present and lit	53016
18	1	Daylight: Street light present	163103
19	2	Daylight: Street light present	350290


No of Casualties for Dual Carriageway Roads with a Speed Limit of 20
No of Casualties
1400
1200
1000
800
600
400
200
0
1400
392
42
England
Alba / Scotland
Cymru / Wales
State


No of Vehicles and Casualties by Light Condition
noofvehicles
noofcasualties
Light Condition
Daylight: Street light present
Darkness: Street lights present and lit
Darkness: Street lighting unknown
Darkeness: No street lighting
Darkness: Street lights present but unlit
0.0M
0.2M
0.4M
0.6M
0.8M
1.0M
1.2M
1.4M
No of Vehicles and Casualties

DATA ANALYSIS

DICING

Query Editor | Query History

```
1  select accidenttime, dayofweek, speedlimit, sum(num_casualties)
2  from train
3  where dayofweek = 6
4  group by accidenttime, dayofweek, speedlimit
5  order by accidenttime;
```

Data Output | Explain | Messages | Notifications

	accidenttime time without time zone	dayofweek integer	speedlimit integer	sum bigint
1	00:01:00	6	30	42
2	00:01:00	6	40	3
3	00:01:00	6	50	4
4	00:01:00	6	60	13
5	00:02:00	6	30	25
6	00:02:00	6	40	2
7	00:02:00	6	60	8
8	00:02:00	6	70	1
9	00:03:00	6	30	16
10	00:03:00	6	40	5

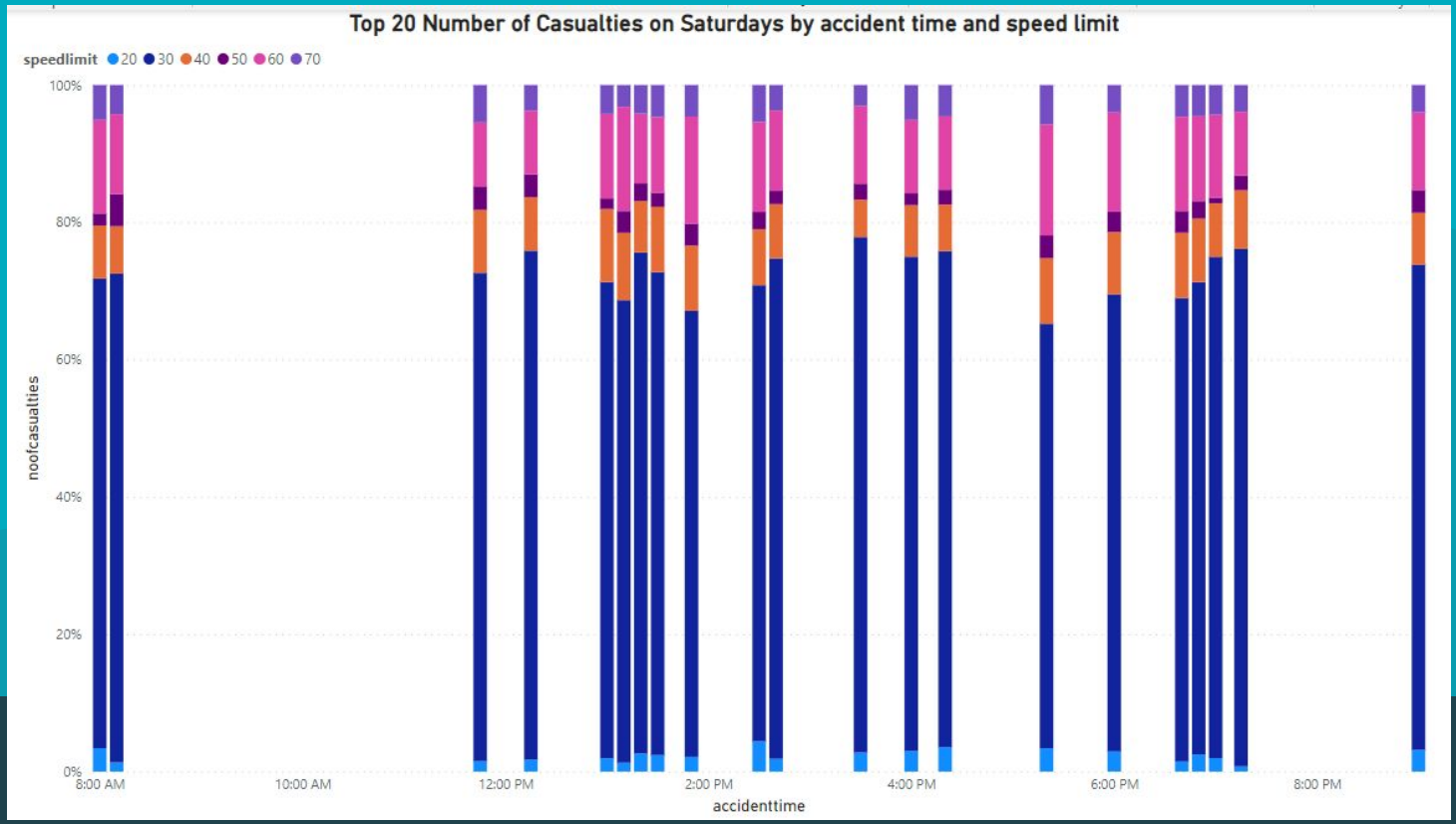
ROLL-UP

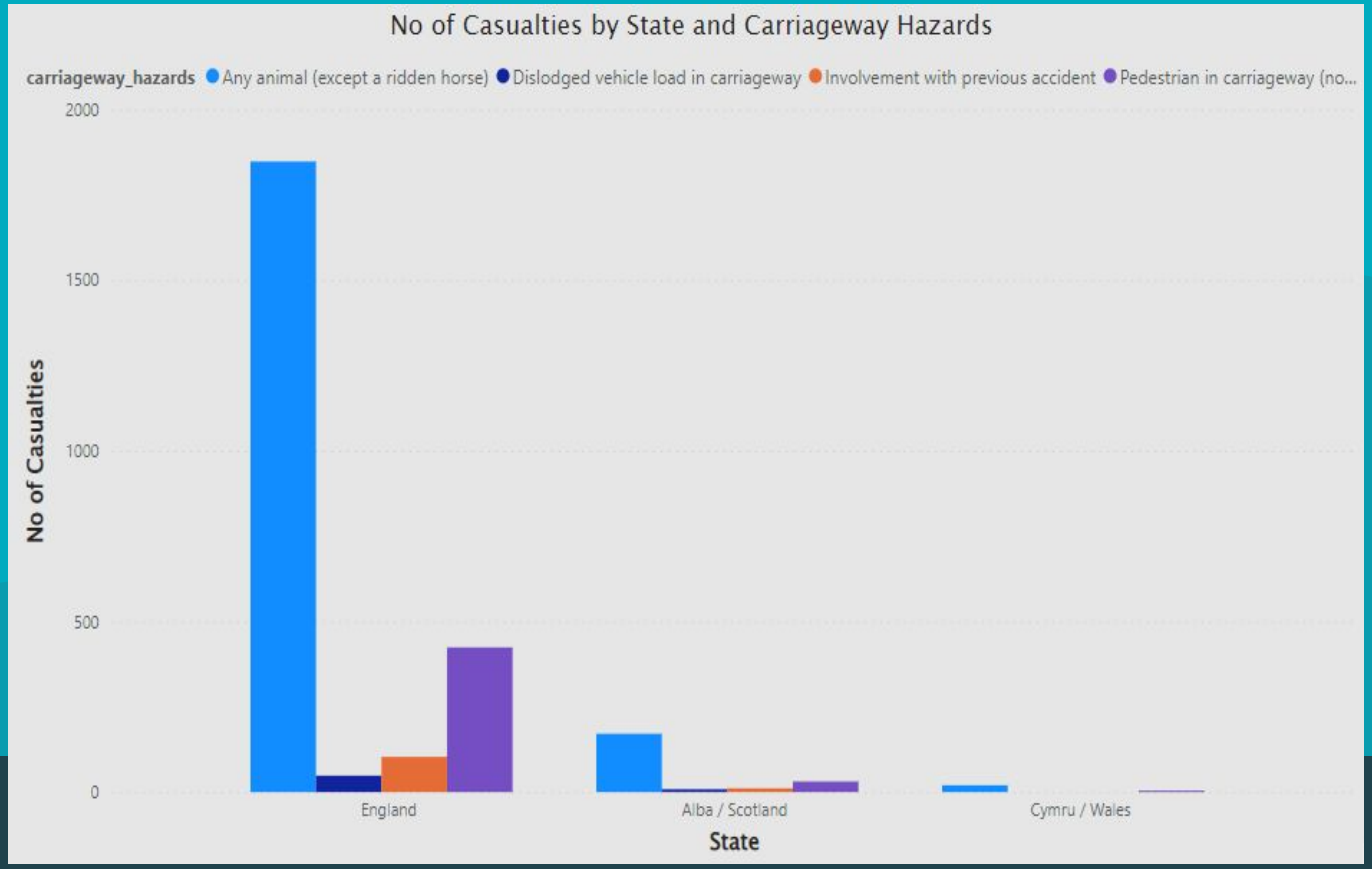
Query Editor | Query History

```
1  select states, carriageway_hazards, sum(num_casualties)
2  from train
3  group by states, carriageway_hazards
4  order by sum(num_casualties);
```

Data Output | Explain | Messages | Notifications

	states text	carriageway_hazards text	sum bigint
1	Cymru / Wales	Pedestrian in carriageway (not injured)	4
2	Alba / Scotland	Dislodged vehicle load in carriageway	8
3	Alba / Scotland	Involvement with previous accident	11
4	Cymru / Wales	Any animal (except a ridden horse)	20
5	Alba / Scotland	Pedestrian in carriageway (not injured)	31
6	England	Dislodged vehicle load in carriageway	48
7	England	Involvement with previous accident	103
8	Alba / Scotland	Any animal (except a ridden horse)	171
9	Cymru / Wales	Other object in carriageway	397
10	England	Pedestrian in carriageway (not injured)	424
11	England	Any animal (except a ridden horse)	1848
12	Alba / Scotland	Other object in carriageway	2370
13	Cymru / Wales	None	15606
14	England	Other object in carriageway	22842
15	Alba / Scotland	None	85853
16	England	None	569957


Top 20 Number of Casualties on Saturdays by accident time and speed limit
speedlimit 20 30 40 50 60 70
100%
80%
60%
40%
20%
0%
noofcasualties
8:00 AM
10:00 AM
12:00 PM
2:00 PM
4:00 PM
6:00 PM
8:00 PM
accidenttime


No of Casualties by State and Carriageway Hazards
carriageway_hazards
Any animal (except a ridden horse)
Dislodged vehicle load in carriageway
Involvement with previous accident
Pedestrian in carriageway (no...
2000
1500
1000
500
0
No of Casualties
England
Alba / Scotland
Cymru / Wales
State

DATA ANALYSIS

ROLL-UP

Query Editor | Query History

```
1  select postcode, max(policeforce)
2  from test
3  group by postcode, policeforce
4  order by max(policeforce) desc;
```

Data Output | Explain | Messages | Notifications

	postcode text	max integer
1	SN4 8HN	98
2	LA14 3HZ	98
3	CA7 2HS	98
4	BL2 6QP	98
5	E11 1GA	98
6	EC1Y 4XY	98
7	RG17 8YY	98
8	TS5 6EU	98
9	FY5 3SU	98
10	UB5 6JR	98

SLICING

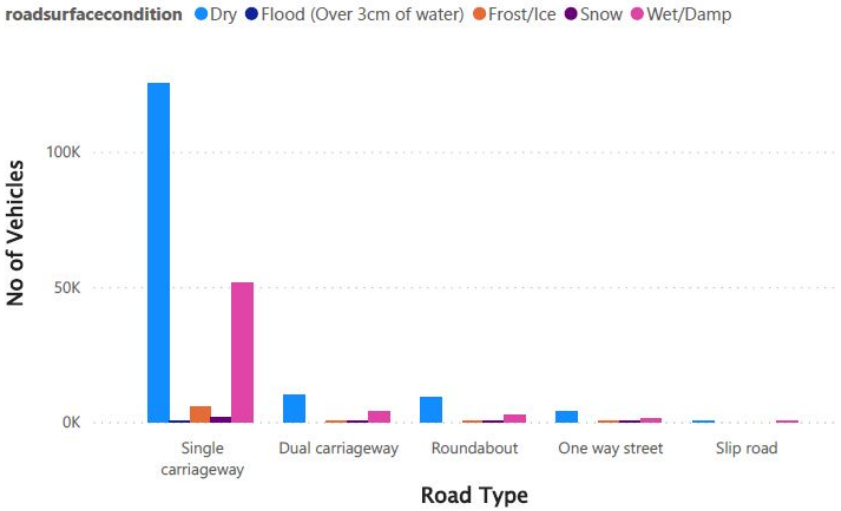
Query Editor | Query History

```
1  select postcode, sum(num_casualties)
2  from train
3  where postcode = 'SN4 8HN' or postcode='LA14 3HZ'
4  group by postcode;
```

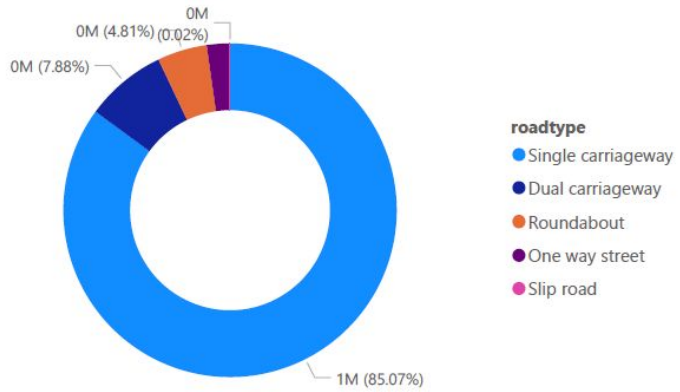
Data Output | Explain | Messages | Notifications

	postcode text	sum bigint
1	LA14 3HZ	5
2	SN4 8HN	1

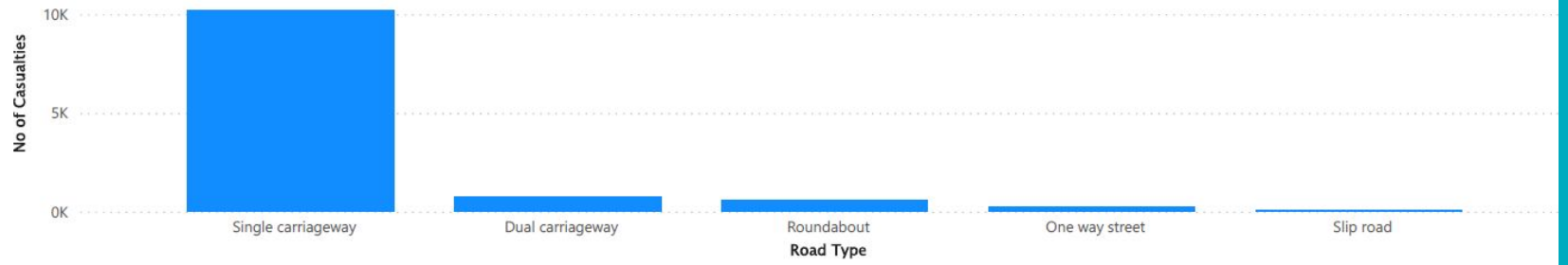
No of Vehicles by Road Type and Road Surface Condition



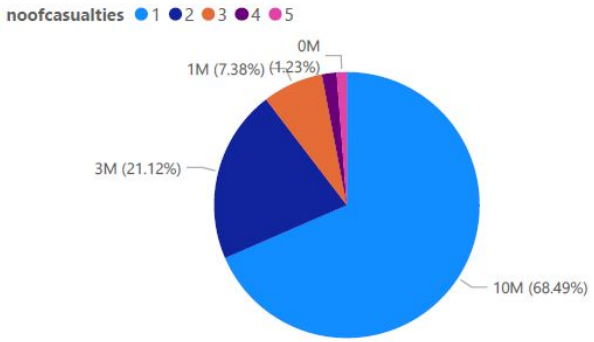
Police Force by Road Type



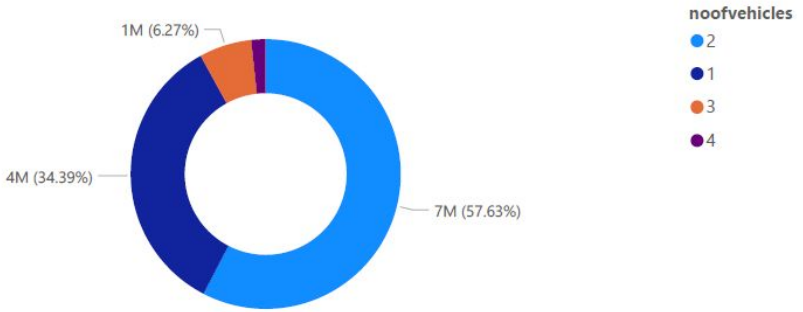
No of Casualties by Road Type



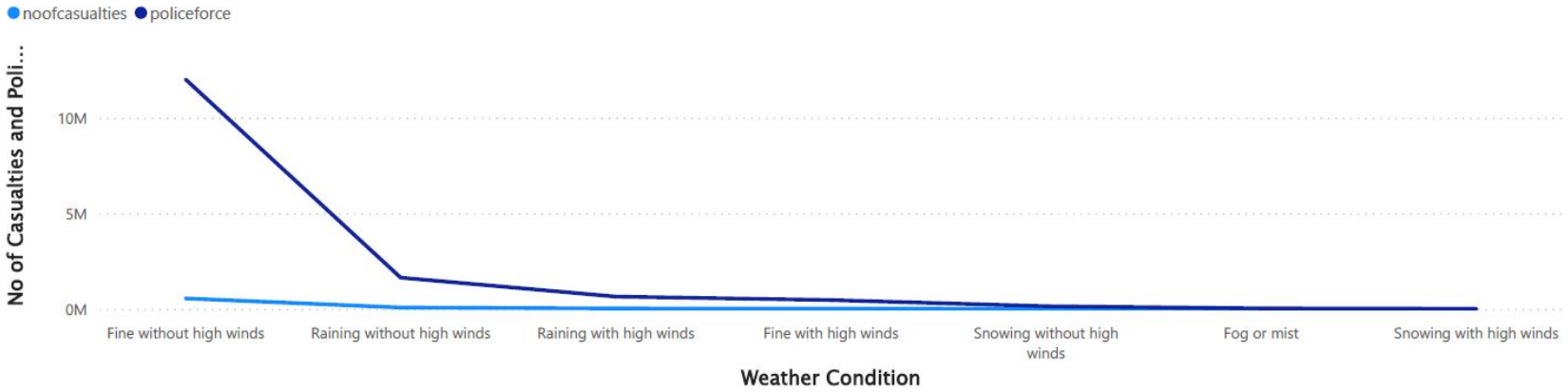
Police Force at 11:59PM and Number of Casualties



Police Force during the Daylight with the Present of Street Light and Number of Cars Involved

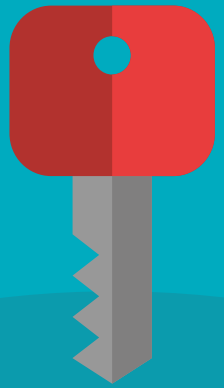


No of Casualties and Police Force by Weather Condition



06

Conclusion



Conclusion

We can conclude that:

England has the most accident rate in United Kingdom compared to Wales and Scotland

Factors that causes accidents are:

1. Light condition
2. Weather condition
3. Drivers' speed limit
4. The time of accidents

The authority of police is proportional to the number of casualties


THANK YOU