UK ACCIDENTS 2019

Exploratory Data Analysis in R

Created by: Oliver Freimuth

Overview

- Skills used
- Data & Data Wrangling
- Exploratory Data Analysis:
  - *Missing Values*
  - *Filtering & Arranging*
  - *Visualizations*
- Creating a Map

Skills used

- Base R: sapply(), for loop
- Deplyr: select(), filter(), arrange(), rename(), mutate(), glimpse()
- ggplot: geom_bar(), geom_point(), geom_smooth(), geom_col()
- Leaflet

Data & Data Wrangling

- Data source: UK government (Link to be found in the R-file)
- Original dataframe contains 117,536 observations of 32 variables
- Here a subset is used containing all observations but only 11 columns
  - *10 are taken directly from the original dataset*
  - *The column “Month” is created based on the variable “Date”*
- Categorical variables are adjusted to the type “factor”

Exploratory Data Analysis

Missing Values

- There is a total of 119 missing values
  - *Time has 63 missing values*
  - *Longitude has 28 missing values*
  - *Latitude has 28 missing values*
  - *Speed limit has 80 observations with the value "-1", most likely missing values*
- There is no possibility to impute or substitute most of these missing values in a sensible way
- Speed limit is an exception: Based on given latitude and longitude, the speed limit could likely be investigated

Exploratory Data Analysis

Filtering & Arranging

- There are only 9 accidents that involved at least 10 cars
- Investigating the Speed limit:

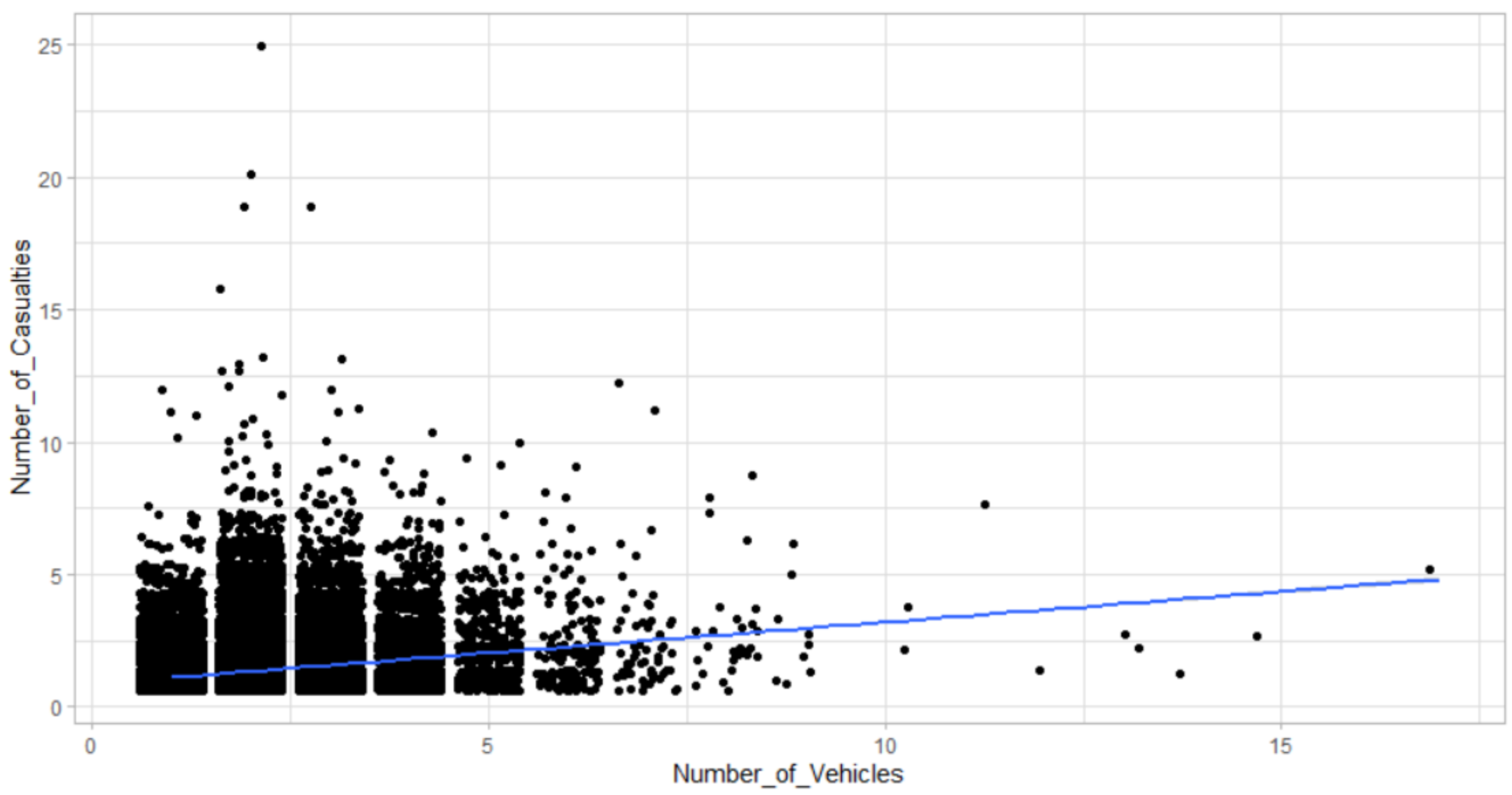
Speed limit	20	30	40	50	60	70
Number of Accidents	11,747	69,305	10,021	4,716	14,514	7,153
Share of Accidents	10.0%	59.0%	8.53%	4.01%	12.35%	6.09%
Average Casualties	1.15	1.23	1.41	1.48	1.51	1.57
Max Casualties	12	16	19	9	52	13

  - *Most accidents happened at a speed limit of 13*
  - *The average number of casualties increases with the speed limit*
  - *The maximum number of casualties happened at 60 but seems to be an outlier*

Visualizations

Scatter Plot: Number of Casualties vs Number of Vehicles

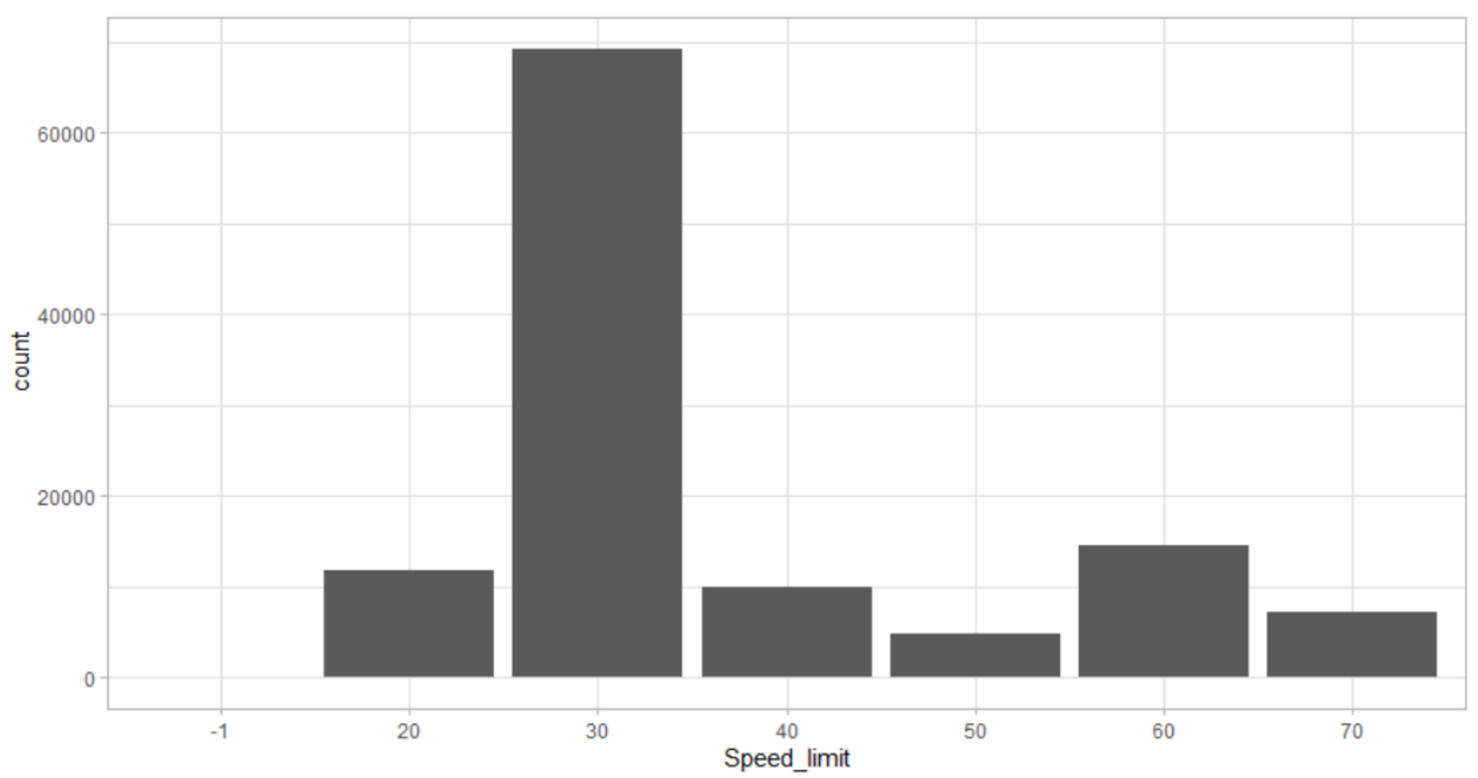
- Points are somewhat displaced to provide an idea how many observations there are
- Most observations involve up to five casualties and vehicles
- The regression line shows that number of vehicles positively correlates with the number of casualties



Visualizations

Bar Plot: How many accidents occurred at a given speed limit?

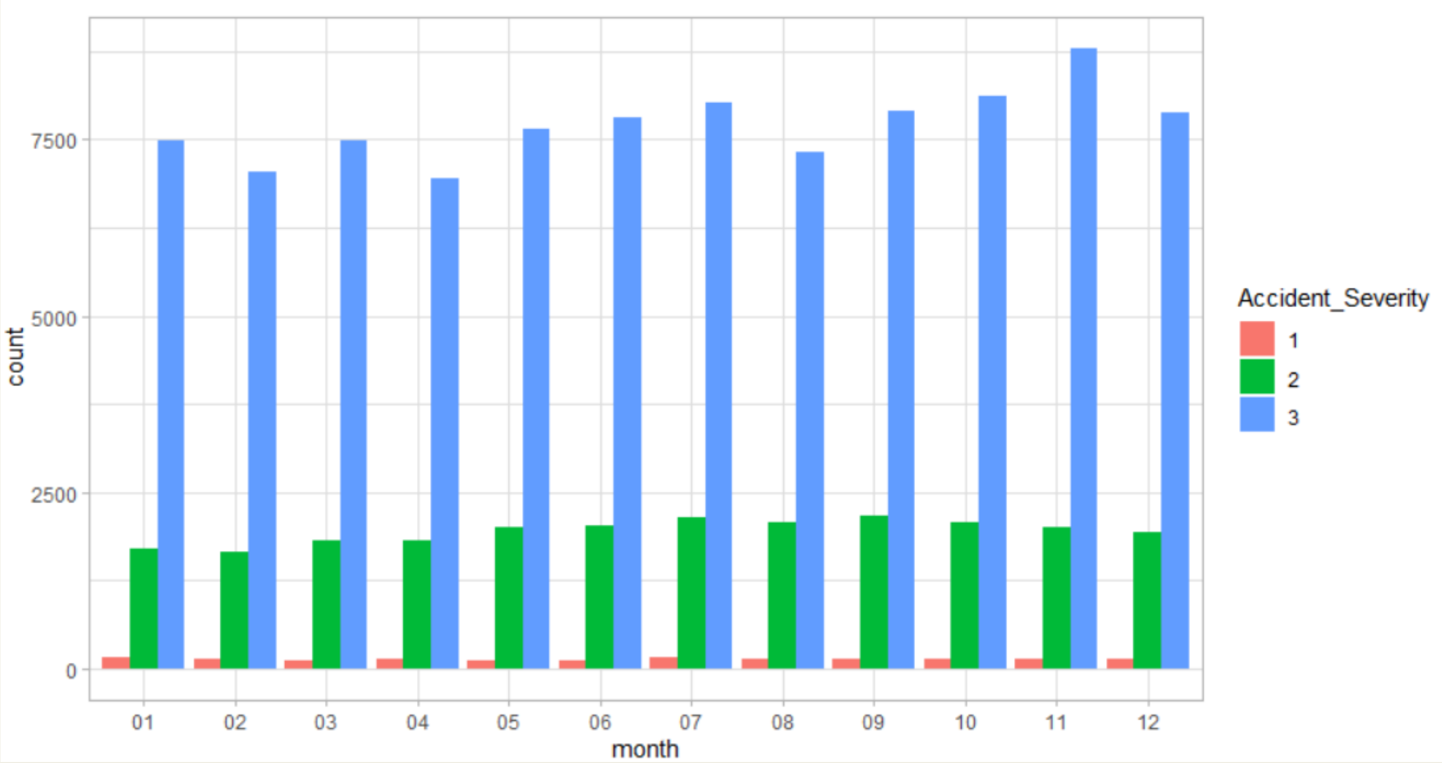
- Most accidents occur at a speed limit of 30
- The least accidents happened at a speed limit of 50
- The value -1 most likely indicates a missing value



Visualizations

Bar Plot: Accident month and Accident Severity (Absolute Numbers)

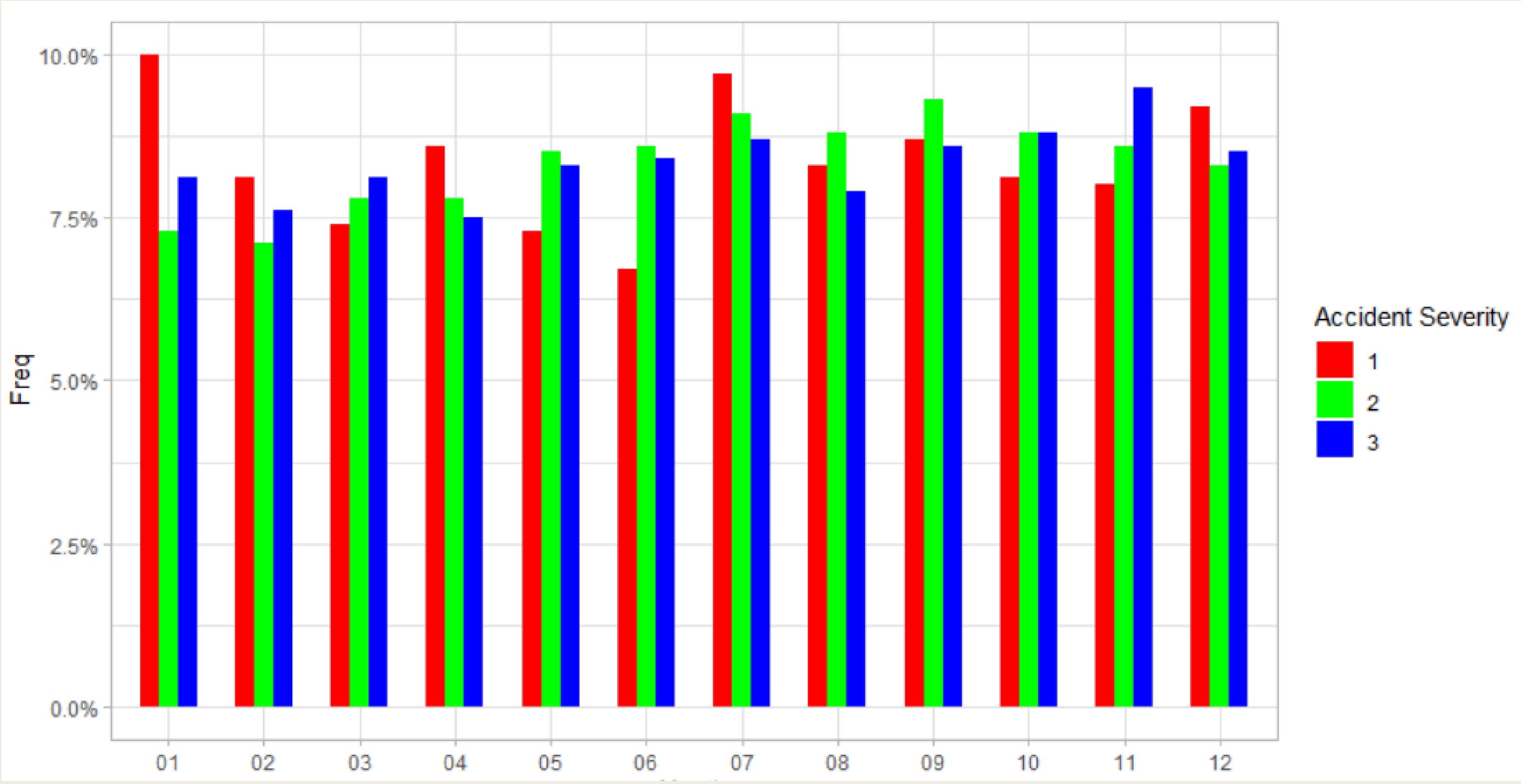
- Most accidents are of severity level 3
- Each accident severity peaks at a different month
- Accidents of severity level 1 are difficult to investigate from plotting absolute numbers



Visualizations

Bar Plot: Accident month and Accident Severity (Relative Numbers)

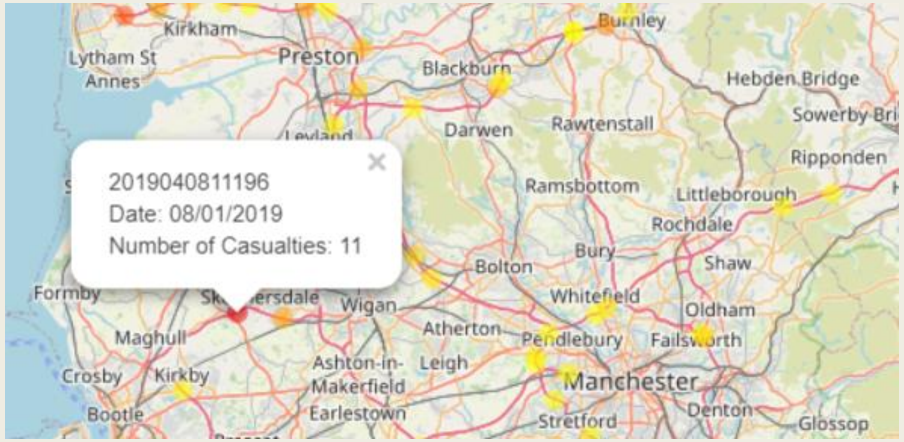
- 10% of accidents with severity 1 happen in January, which is the maximum
- We see that instances of accidents with different severity levels follow different patterns over the year



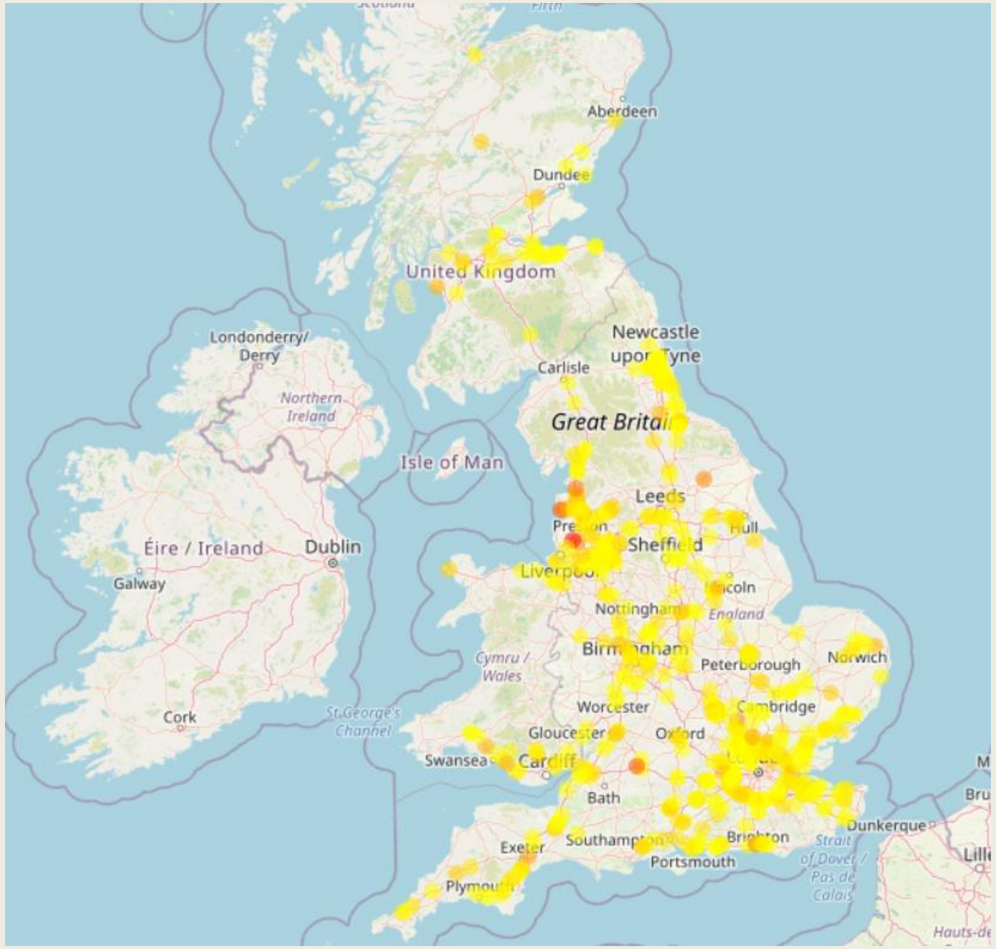
Visualizations

Accidents in January at speed limit 70 (n = 576)

- Darker markers indicate a higher number of casualties
- There seem to be certain hot spots
- Some areas are (mostly) free of accidents meeting the parameters



2019040811196
Date: 08/01/2019
Number of Casualties: 11



Aberdeen
Dundee
United Kingdom
Londonderry/ Derry
Newcastle upon Tyne
Carlisle
Northern Ireland
Great Britain
Isle of Man
Leeds
Éire / Ireland
Dublin
Preston
Sheffield
Hull
Galway
Liverpool
Lincoln
Nottingham
England
Birmingham
Peterborough
Norwich
Cymru / Wales
Cork
St George's Channel
Worcester
Cambridge
Gloucester
Oxford
Swansea
Cardiff
Bath
Exeter
Southampton
Brighton
Dunkerque
Portsmouth
Plymouth
Strait of Dover / Pas de Calais
Lille
Hauts-de