# UK Road Safety Data

### Important Information

The UKRoadSafetyData.zip was provided to me to use as test data for analysis practice and demonstration.

#### Tools

* Microsoft SQL Server Management Studio
* Microsoft Visual Studio 2022
* Microsoft Excel (Power Query, CSVs)
* Posit RStudio
* Tableau Public

#### Languages

* SQL
* R

### Dataset First Impressions

* Upon receiving and extracting the ‘UKRoadSafetyData.zip’ file, I notice there are additional zip files.
* Within each zip file is a single .csv or .xls file.
* Several of these files are missing naming conventions that match other similar files.
  + Renamed files for consistency if needed later.
* The 2015 data is in a single zip, and these are also not following later file naming conventions, and yet more 2015 files exist with apparently duplicate data.
  + Both 2015 Accidents files begin with ‘201501BS70001’ in cell A2 and contain 140057 rows.
  + But without looking further, there is not guarantee these have exactly duplicate data. This will require review.
  + Checked through Power Query, matching both sets of data on for each duplicate (Accidents, Casualties, and Vehicles for 2015).
  + Both sets are a perfect Match. Only need to import one set.
* The Variable Lookup is a set of codes tables. One question is whether I will want to keep these as separate tables, or perform a single codes\_table combination.
  + The codes tables can remain separate, but it would require more write scripts.
  + If I combine them into a single table, they could be more manageable on the import.
  + However, I may not need to import these into my database. They may only be needed while visualizing the findings, in which case I can wait until I import them to Tableau. TBD.

### Extract, Transform, Load (ETL)

#### Visual Studio 2022 + SQL

* Created three tables uk\_accidents uk\_vehicles uk\_casualties

uk\_accidents

/\*   
I had a lot of trouble with the Date column   
but finally manged to run through steps in Power Query and SQL  
that worked.  
\*/  
--DROP TABLE uk\_accidents  
--Date VARCHAR(15),  
--Date DATE,  
  
CREATE TABLE uk\_accidents (  
 Accident\_Index VARCHAR(60),  
 Location\_Easting\_OSGR INT,  
 Location\_Northing\_OSGR INT,  
 Longitude FLOAT,  
 Latitude FLOAT,  
 Police\_Force INT,  
 Accident\_Severity INT,  
 Number\_of\_Vehicles INT,  
 Number\_of\_Casualties INT,  
 Date VARCHAR(15),   
 Day\_of\_Week INT,  
 Time TIME,  
 Local\_Authority\_District INT,  
 Local\_Authority\_Highway VARCHAR(20),  
 First\_Road\_Class INT,  
 First\_Road\_Number INT,  
 Road\_Type INT,  
 Speed\_limit INT,  
 Junction\_Detail INT,  
 Junction\_Control INT,  
 Second\_Road\_Class INT,  
 Second\_Road\_Number INT,  
 Pedestrian\_Crossing\_Human\_Control INT,  
 Pedestrian\_Crossing\_Physical\_Facilities INT,  
 Light\_Conditions INT,  
 Weather\_Conditions INT,  
 Road\_Surface\_Conditions INT,  
 Special\_Conditions\_at\_Site INT,  
 Carriageway\_Hazards INT,  
 Urban\_or\_Rural\_Area INT,  
 Did\_Police\_Officer\_Attend\_Scene\_of\_Accident INT,  
 LSOA\_of\_Accident\_Location VARCHAR(60)  
);

uk\_casualties

--DROP TABLE uk\_vehicles  
  
CREATE TABLE uk\_casualties (  
 Accident\_Index VARCHAR(20),  
 Vehicle\_Reference INT,  
 Casualty\_Reference INT,  
 Casualty\_Class INT,  
 Sex\_of\_Casualty INT,  
 Age\_of\_Casualty INT,  
 Age\_Band\_of\_Casualty INT,  
 Casualty\_Severity INT,  
 Pedestrian\_Location INT,  
 Pedestrian\_Movement INT,  
 Car\_Passenger INT,  
 Bus\_or\_Coach\_Passenger INT,  
 Pedestrian\_Road\_Maintenance\_Worker INT,  
 Casualty\_Type INT,  
 Casualty\_Home\_Area\_Type INT,  
 Casualty\_IMD\_Decile INT,  
);

uk\_vehicles

--DROP TABLE uk\_casualties  
  
CREATE TABLE uk\_vehicles (  
 Accident\_Index VARCHAR(20),  
 Vehicle\_Reference INT,  
 Vehicle\_Type INT,  
 Towing\_and\_Articulation INT,  
 Vehicle\_Manoeuvre INT,  
 Vehicle\_Location\_Restricted\_Lane INT,  
 Junction\_Location INT,  
 Skidding\_and\_Overturning INT,  
 Hit\_Object\_in\_Carriageway INT,  
 Vehicle\_Leaving\_Carriageway INT,  
 Hit\_Object\_off\_Carriageway INT,  
 First\_Point\_of\_Impact INT,  
 Was\_Vehicle\_Left\_Hand\_Drive INT,  
 Journey\_Purpose\_of\_Driver INT,  
 Sex\_of\_Driver INT,  
 Age\_of\_Driver INT,  
 Age\_Band\_of\_Driver INT,  
 Engine\_Capacity\_CC INT,  
 Propulsion\_Code INT,  
 Age\_of\_Vehicle INT,  
 Driver\_IMD\_Decile INT,  
 Driver\_Home\_Area\_Type INT,  
 Vehicle\_IMD\_Decile INT,  
);

* Importing Accidents 2016 ran into an issue with Speed Limit column, where there were strings of NULL instead of true null values.
  + Same issue with the 2017 Accidents in the Lat/Long columns.
  + In Power Query, I ran a find/replace NULL for “” on the entire workbook.
* Accidents table: Date
  + Dates were formatted as text in dd/mm/yyyy format. When importing and converting, they were truncated. Niether Excel or SQL were playing nicely with this format.
  + In Power Query, I ran a transformation and re-saved the CSVs.
  + After some difficulty getting dates formatting, and a bit of help from Google and ChatGPT-4, the dates finally worked out.
  + The final CSV format was yyyy-mm-dd, then imported into my database as a VARCHAR
  + Then the final conversion was performed in SQL.

/\*   
I had a lot of trouble with the Date column   
but finally manged to run through steps in Power Query and SQL  
that worked.  
\*/  
--DROP TABLE uk\_accidents  
--Date VARCHAR(15),  
--Date DATE,  
  
SELECT Date   
FROM uk\_accidents   
WHERE ISDATE(Date) = 0  
  
  
-- Update the Date column to swap the day and month  
UPDATE uk\_accidents  
SET Date = CONCAT(  
 SUBSTRING(Date, 1, 4), '-', -- Year  
 SUBSTRING(Date, 9, 2), '-', -- Day  
 SUBSTRING(Date, 6, 2) -- Month  
)  
WHERE ISDATE(Date) = 0;  
  
  
ALTER TABLE uk\_accidents ADD Date\_New DATE;  
  
UPDATE uk\_accidents SET Date\_New = CAST(Date AS DATE);  
  
--UPDATE uk\_accidents SET Date\_New = CONVERT(DATE, Date, 103); -- 103 is for dd/mm/yyyy format  
  
ALTER TABLE uk\_accidents DROP COLUMN Date;  
  
EXEC sp\_rename 'uk\_accidents.Date\_New', 'Date', 'COLUMN';

* Checking for duplicates in the other tables revealed that while there are duplicate Accident\_Index IDs, they are not duplicate observations. Each row is a unique observation.
  + Running the following script is an example of how they are not actually duplicates:

WITH CTE\_CheckDupsC AS (  
 SELECT \*  
-- c.Accident\_Index  
--, c.Casualty\_Reference  
, ROW\_NUMBER() OVER (PARTITION BY Accident\_Index, Casualty\_Reference ORDER BY (SELECT NULL)) AS rn  
 FROM uk\_casualties AS c  
 )  
  
 SELECT \*  
 FROM CTE\_CheckDupsC  
 WHERE Accident\_Index = '201604ED16270'  
-- WHERE rn > 1  
 ORDER BY CTE\_CheckDupsC.Accident\_Index

### Analysis Roadmap

1. Understand the Data: Make notes about each variable, what is it, what signficance does it hold, what potential calculations or analysis can be done on them.
2. Generate Hypotheses: What relationships, correlations, can we expect to find? Are accidents higher in certain regions, road conditions, etc? Are certain vehicles more likely to be in an accident?
3. Exploratory Data Analysis (EDA): Using SQL, R, and Tableau to analyze the datasets, we summarize the data and find patterns. Min, Max, Sum, Count, Mean, Median, etc. will be helpful.
4. Visualization: Once we have some key insights, start creating visualizations (first in R, then in Tableau) to see if anything visually stands out as interesting or if a story begins to emerge.
5. Story Drafting: Draft the narrative around these findings. What story is the data telling? Why is it important? What recommendations or observations can we make?
6. Review and Refine: Review the data story several times, refine it to make sure it’s compelling and understandable. Are the visualizations telling the story on their own?
7. Presentation: Put all these findings into a presentation format. Keep it simple but informative. Use your visualizations to support your story. Keep the script out of the visualizations themselves but in a presentable format for those who are not able to attend the presentation.

Questions to Consider:

Conditions

* Roads: Wet/Dry, Urban/Rural, Maintained/Deferred-Maintenance,
* Cities, Regions, Police Depts, Deprivation Index, do any of these show higher accidents or lower?
* Are any types of crossings or conditions more likely to cause accidents?
* Do Seasons, Days, Times of Day, Light, or Weather play a role?

People

* Does Poverty index, age, sex, or any other human factor play a role in accidents?

### Analysis

#### R & RStudio

Transitioning to RStudio for Analysis

##### Connecting to R RStudio

Loading Libraries

# LOAD LIBRARIES  
library(DBI)

## Warning: package 'DBI' was built under R version 4.3.1

library(odbc)

## Warning: package 'odbc' was built under R version 4.3.1

library(RODBC)

## Warning: package 'RODBC' was built under R version 4.3.1

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.3.1

## Warning: package 'ggplot2' was built under R version 4.3.1

## Warning: package 'lubridate' was built under R version 4.3.1

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.2 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.3 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

#library(dplyr)  
#library(ggplot2)  
#library(scales)  
#library(sqldf)  
#library(ggmap)  
#library(geosphere)  
#library(here)  
#library(skimr)  
#library(janitor)  
#library(Tmisc)

Connecting R & RStudio to the database

# ESTABLISH CONNECTION TO MY LOCAL DATABASE  
  
connection <- odbcDriverConnect("driver={SQL Server};server=LAPTOP-76LHVPRQ\\SQLEXPRESS;database=UK\_RoadSafety;trusted\_connection=true")

##### Assign Dataframes

Assigning the new cleaned database tables as R dataframes:

* UK\_AccidentData
* UK\_CasualtyData
* UK\_VehicleData

# ASSIGN THE DATABASE TABLE AS A DATAFRAME (df) VARIABLE FOR EASIER RECALL  
  
UK\_AccidentData <- sqlFetch(connection, "dbo.uk\_accidents")  
  
UK\_CasualtyData <- sqlFetch(connection, "dbo.uk\_casualties")  
  
UK\_VehicleData <- sqlFetch(connection, "dbo.uk\_vehicles")

##### Column Names (Variables)

* What variables do each dataset contain?
* How are they related?
* What might they tell us?

The Accident Index acts as a key. The Accident Table is the key table, there is one Accident Index for each observation (row) in this table. The Casualty and Vehicle tables refer back to this Accident Index key to tie their data into specific accidents. The Number of Vehicles and Number of Casualties are essentially a count of the other related data from the other two tables.

The Vehicle Reference builds on the Accident Index. It is contained in both the Vehicle Table and Casualty Table. For each accident, individual vehicles are numbered (Vehicle 1, Vehicle 2, etc.). This allows us to see how many vehicles were involved in any given accident (from the Vehicle Table) while maintaining a simple Accident Index in the origin table. The Casualty Table can then assign the casualty to which vehicle was involved in that injury.

The Casualty Reference only exists on the Casualty Table, and serves as an index for the individual who was injured.

There are also Key Tables/Codes Tables. During my work in R, I will reference them via CSV/Excel to understand what each one is, as I do not need them for determining averages or means or counts. However, when visualizing this data, I will import these into Tableau in a star schema for more accurate labeling.

colnames(UK\_AccidentData)

## [1] "Accident\_Index"   
## [2] "Location\_Easting\_OSGR"   
## [3] "Location\_Northing\_OSGR"   
## [4] "Longitude"   
## [5] "Latitude"   
## [6] "Police\_Force"   
## [7] "Accident\_Severity"   
## [8] "Number\_of\_Vehicles"   
## [9] "Number\_of\_Casualties"   
## [10] "Day\_of\_Week"   
## [11] "Time"   
## [12] "Local\_Authority\_District"   
## [13] "Local\_Authority\_Highway"   
## [14] "First\_Road\_Class"   
## [15] "First\_Road\_Number"   
## [16] "Road\_Type"   
## [17] "Speed\_limit"   
## [18] "Junction\_Detail"   
## [19] "Junction\_Control"   
## [20] "Second\_Road\_Class"   
## [21] "Second\_Road\_Number"   
## [22] "Pedestrian\_Crossing\_Human\_Control"   
## [23] "Pedestrian\_Crossing\_Physical\_Facilities"   
## [24] "Light\_Conditions"   
## [25] "Weather\_Conditions"   
## [26] "Road\_Surface\_Conditions"   
## [27] "Special\_Conditions\_at\_Site"   
## [28] "Carriageway\_Hazards"   
## [29] "Urban\_or\_Rural\_Area"   
## [30] "Did\_Police\_Officer\_Attend\_Scene\_of\_Accident"  
## [31] "LSOA\_of\_Accident\_Location"   
## [32] "Date"   
## [33] "TimeGroup"

colnames(UK\_CasualtyData)

## [1] "Accident\_Index" "Vehicle\_Reference"   
## [3] "Casualty\_Reference" "Casualty\_Class"   
## [5] "Sex\_of\_Casualty" "Age\_of\_Casualty"   
## [7] "Age\_Band\_of\_Casualty" "Casualty\_Severity"   
## [9] "Pedestrian\_Location" "Pedestrian\_Movement"   
## [11] "Car\_Passenger" "Bus\_or\_Coach\_Passenger"   
## [13] "Pedestrian\_Road\_Maintenance\_Worker" "Casualty\_Type"   
## [15] "Casualty\_Home\_Area\_Type" "Casualty\_IMD\_Decile"

colnames(UK\_VehicleData)

## [1] "Accident\_Index" "Vehicle\_Reference"   
## [3] "Vehicle\_Type" "Towing\_and\_Articulation"   
## [5] "Vehicle\_Manoeuvre" "Vehicle\_Location\_Restricted\_Lane"  
## [7] "Junction\_Location" "Skidding\_and\_Overturning"   
## [9] "Hit\_Object\_in\_Carriageway" "Vehicle\_Leaving\_Carriageway"   
## [11] "Hit\_Object\_off\_Carriageway" "First\_Point\_of\_Impact"   
## [13] "Was\_Vehicle\_Left\_Hand\_Drive" "Journey\_Purpose\_of\_Driver"   
## [15] "Sex\_of\_Driver" "Age\_of\_Driver"   
## [17] "Age\_Band\_of\_Driver" "Engine\_Capacity\_CC"   
## [19] "Propulsion\_Code" "Age\_of\_Vehicle"   
## [21] "Driver\_IMD\_Decile" "Driver\_Home\_Area\_Type"   
## [23] "Vehicle\_IMD\_Decile"

##### UK\_Road Safety Data Summary

In comparing certain variable’s mean and median, we can make some preliminary statements.

Summary of UK Accidents

* Accident Severity Code 3 is Slight. The Median (3) and Mean (2.81) indicate that the average accident only presents Slight injuries.
* Most accidents involve two vehicles and do involve at least one injury.
* Day of Week: I’d like to see a visual count on this for the “difference” but most accidents occur on a Wednesday (Median 4, Mean 4.104, Code 4 is Wednesday).
* The Mean (37.64) and Median (30) speed limits may indicate the “average” accident occurs at these lower limits. However, a count graphic may indicate whether there was a slight difference or major difference in accidents counts for each limit range.
* Light Conditions. Code 1 is light, Codes 4-7 indicate various lighting conditions in the dark. The Median (1) and Mean (1.993) may indicate the average accident occurs during the day or with good lighting. Graphing this data may help illuminate this key.
* Weather Conditions. Code 1 is Fine. Codes 2-8 are other than fine. 9 is unknown. Median (1) and Mean (1.579) indicate weather is not a major influence “on average”, but graphing this will show any trends or difference in the data.
* Police Attendance would be an after-the-fact variable, so while it may indicate the seriousness of the accident, it would not be predictive in any way for preventing accidents. That being said, it does appear the police were involved in the average accident, Median (1), Mean (1.251), with Code 1 being Yes, and Code 2 being No.

Summary of UK Acc*ident Casualty*

*Note: By “Casualty”, this dataset means “Injury” not necessarily “Death”. They are segregated into Fatality, Serious, and Slight.*

* Sex of Casualty Code 1, Male, Code 2, Female. Median (1), Mean (1.406); this does not tell me much right now. I’d like to see the data visualized and see if this acts as a component against or with other variables.
* Age of Casualty: Median 33, Mean 36.48. This does seem to indicate something. Although, this is the prime working age, so it’s possible this has to do with the number of humans in this age range in the vicinity and not anything to do with their driving. High probability with higher representation. As with the other table, when a casualty is involved, it appears to be serious or slight, but Fatalities are not the average. Although, with the Insurance rates higher for “teens” I would have expected this to trend that direction. I would like to see the full variable visualized.
* Pedestrian Location & Movement reflect Median of 0, and Mean of 0.714/0.5418. Pedestrians are rarely involved.
* Road Maintenance workers are rarely involved. Although, if a road working company wanted to analyze the data for situations in which they could reduce their workers being hurt, they could slice the data for only these situations and see what trends emerge.
* Home Area (1 Urban, 2 Small Town, 3 Rural) - Urban cities seem to be the average place for accidents.
* IMD 4 / 3.714 may indicate a trend toward more casualties in mid-lower income regions? Needs further visualization analysis.

Summary of UK Vehicles in Accidents

* Vehicle Type 9/9.804 is Car. Most people on most trips are in a car, so this holds to the expected pattern. Could be interesting to visualize this data variable to see what the second or third most common types are. Motorcycle, Van, Taxi, etc.
* !!! Manoeuvre (Reversing, Parked, Waiting, Slowing, U-Turn, etc.). Anecdotally, when I was a young driver most of my accidents were in parking lots while reversing. I later learned this is common for ADHD neurotypes. This may be interesting to follow.
* Vehicle Location Median 0 means the central data point is on the main roads. Not surprising.
* Junction Median of 1 (Approaching or waiting at a junction) may indicate this is a common place for accidents. I was rear-ended more than once while waiting at a stoplight.
* Journey or Purpose Unknown (6) may make this a useless variable?
* Several variables show no indication of relevance, all being zero median.
* Age of Driver (35/35.49) matches the Casualty table. Interesting.
* Propulsion Code - Mostly Gasoline/Petrol cars. Not shocking.
* Age of Vehicle - Average car was 5.569 years old. That’s may be a reflection of the average age of cars in general, and may only be relevant to the accident data if this was compared to the general average age of cars in the UK. If they were divergent, then it would be relevant. Otherwise it is a reflection and not indicative of anything useful.
* Engine Capacity CC - Median (1390), Mean (1422) actually mean nothing to me because I don’t know anything about Engine CCs or how that’s relevant. And as with Age of Vehicle, may be a reflection and not an indication.

summary(UK\_AccidentData)

## Accident\_Index Location\_Easting\_OSGR Location\_Northing\_OSGR  
## Length:529294 Min. : 70860 Min. : 10235   
## Class :character 1st Qu.:386552 1st Qu.: 176240   
## Mode :character Median :455233 Median : 231908   
## Mean :449456 Mean : 286296   
## 3rd Qu.:528110 3rd Qu.: 389580   
## Max. :655391 Max. :1209512   
## NA's :108 NA's :108   
## Longitude Latitude Police\_Force Accident\_Severity  
## Min. :-7.4229 Min. :49.91 Min. : 1.00 Min. :1.00   
## 1st Qu.:-2.2016 1st Qu.:51.47 1st Qu.: 6.00 1st Qu.:3.00   
## Median :-1.1849 Median :51.97 Median :30.00 Median :3.00   
## Mean :-1.2934 Mean :52.46 Mean :29.19 Mean :2.81   
## 3rd Qu.:-0.1533 3rd Qu.:53.40 3rd Qu.:45.00 3rd Qu.:3.00   
## Max. : 1.7596 Max. :60.76 Max. :98.00 Max. :3.00   
## NA's :118 NA's :118   
## Number\_of\_Vehicles Number\_of\_Casualties Day\_of\_Week Time   
## Min. : 1.000 Min. : 1.000 Min. :1.000 Length:529294   
## 1st Qu.: 1.000 1st Qu.: 1.000 1st Qu.:2.000 Class :character   
## Median : 2.000 Median : 1.000 Median :4.000 Mode :character   
## Mean : 1.843 Mean : 1.321 Mean :4.104   
## 3rd Qu.: 2.000 3rd Qu.: 1.000 3rd Qu.:6.000   
## Max. :37.000 Max. :59.000 Max. :7.000   
##   
## Local\_Authority\_District Local\_Authority\_Highway First\_Road\_Class  
## Min. : 1 Length:529294 Min. :1.00   
## 1st Qu.: 95 Class :character 1st Qu.:3.00   
## Median :307 Mode :character Median :4.00   
## Mean :336 Mean :4.17   
## 3rd Qu.:514 3rd Qu.:6.00   
## Max. :941 Max. :6.00   
##   
## First\_Road\_Number Road\_Type Speed\_limit Junction\_Detail   
## Min. : 0.0 Min. :-1.000 Min. : 0.00 Min. :-1.000   
## 1st Qu.: 0.0 1st Qu.: 6.000 1st Qu.:30.00 1st Qu.: 0.000   
## Median : 60.0 Median : 6.000 Median :30.00 Median : 1.000   
## Mean : 900.5 Mean : 5.175 Mean :37.64 Mean : 2.252   
## 3rd Qu.: 630.0 3rd Qu.: 6.000 3rd Qu.:40.00 3rd Qu.: 3.000   
## Max. :9918.0 Max. : 9.000 Max. :70.00 Max. : 9.000   
## NA's :37   
## Junction\_Control Second\_Road\_Class Second\_Road\_Number  
## Min. :-1.000 Min. :-1.000 Min. : -1.0   
## 1st Qu.:-1.000 1st Qu.:-1.000 1st Qu.: 0.0   
## Median : 2.000 Median : 3.000 Median : 0.0   
## Mean : 1.657 Mean : 2.647 Mean : 320.6   
## 3rd Qu.: 4.000 3rd Qu.: 6.000 3rd Qu.: 0.0   
## Max. : 4.000 Max. : 6.000 Max. :9999.0   
##   
## Pedestrian\_Crossing\_Human\_Control Pedestrian\_Crossing\_Physical\_Facilities  
## Min. :-1.000000 Min. :-1.0000   
## 1st Qu.: 0.000000 1st Qu.: 0.0000   
## Median : 0.000000 Median : 0.0000   
## Mean : 0.004485 Mean : 0.8375   
## 3rd Qu.: 0.000000 3rd Qu.: 0.0000   
## Max. : 2.000000 Max. : 8.0000   
##   
## Light\_Conditions Weather\_Conditions Road\_Surface\_Conditions  
## Min. :-1.000 Min. :-1.000 Min. :-1.000   
## 1st Qu.: 1.000 1st Qu.: 1.000 1st Qu.: 1.000   
## Median : 1.000 Median : 1.000 Median : 1.000   
## Mean : 1.993 Mean : 1.579 Mean : 1.286   
## 3rd Qu.: 4.000 3rd Qu.: 1.000 3rd Qu.: 2.000   
## Max. : 7.000 Max. : 9.000 Max. : 5.000   
##   
## Special\_Conditions\_at\_Site Carriageway\_Hazards Urban\_or\_Rural\_Area  
## Min. :-1.00000 Min. :-1.00000 Min. :-1.000   
## 1st Qu.: 0.00000 1st Qu.: 0.00000 1st Qu.: 1.000   
## Median : 0.00000 Median : 0.00000 Median : 1.000   
## Mean : 0.08348 Mean : 0.05296 Mean : 1.338   
## 3rd Qu.: 0.00000 3rd Qu.: 0.00000 3rd Qu.: 2.000   
## Max. : 7.00000 Max. : 7.00000 Max. : 3.000   
##   
## Did\_Police\_Officer\_Attend\_Scene\_of\_Accident LSOA\_of\_Accident\_Location  
## Min. :-1.000 Length:529294   
## 1st Qu.: 1.000 Class :character   
## Median : 1.000 Mode :character   
## Mean : 1.251   
## 3rd Qu.: 1.000   
## Max. : 3.000   
##   
## Date TimeGroup   
## Length:529294 Length:529294   
## Class :character Class :character   
## Mode :character Mode :character   
##   
##   
##   
##

summary(UK\_CasualtyData)

## Accident\_Index Vehicle\_Reference Casualty\_Reference Casualty\_Class  
## Length:699163 Min. : 1.000 Min. : 1.000 Min. :1.00   
## Class :character 1st Qu.: 1.000 1st Qu.: 1.000 1st Qu.:1.00   
## Mode :character Median : 1.000 Median : 1.000 Median :1.00   
## Mean : 1.488 Mean : 1.406 Mean :1.49   
## 3rd Qu.: 2.000 3rd Qu.: 2.000 3rd Qu.:2.00   
## Max. :999.000 Max. :991.000 Max. :3.00   
## Sex\_of\_Casualty Age\_of\_Casualty Age\_Band\_of\_Casualty Casualty\_Severity  
## Min. :-1.000 Min. : -1.00 Min. :-1.000 Min. :1.000   
## 1st Qu.: 1.000 1st Qu.: 22.00 1st Qu.: 5.000 1st Qu.:3.000   
## Median : 1.000 Median : 33.00 Median : 6.000 Median :3.000   
## Mean : 1.406 Mean : 36.48 Mean : 6.289 Mean :2.842   
## 3rd Qu.: 2.000 3rd Qu.: 50.00 3rd Qu.: 8.000 3rd Qu.:3.000   
## Max. : 2.000 Max. :104.00 Max. :11.000 Max. :3.000   
## Pedestrian\_Location Pedestrian\_Movement Car\_Passenger   
## Min. :-1.000 Min. :-1.0000 Min. :-1.0000   
## 1st Qu.: 0.000 1st Qu.: 0.0000 1st Qu.: 0.0000   
## Median : 0.000 Median : 0.0000 Median : 0.0000   
## Mean : 0.714 Mean : 0.5418 Mean : 0.2523   
## 3rd Qu.: 0.000 3rd Qu.: 0.0000 3rd Qu.: 0.0000   
## Max. :10.000 Max. : 9.0000 Max. : 2.0000   
## Bus\_or\_Coach\_Passenger Pedestrian\_Road\_Maintenance\_Worker Casualty\_Type   
## Min. :-1.00000 Min. :-1.00000 Min. :-1.000   
## 1st Qu.: 0.00000 1st Qu.: 0.00000 1st Qu.: 3.000   
## Median : 0.00000 Median : 0.00000 Median : 9.000   
## Mean : 0.07744 Mean : 0.07626 Mean : 7.281   
## 3rd Qu.: 0.00000 3rd Qu.: 0.00000 3rd Qu.: 9.000   
## Max. : 4.00000 Max. : 2.00000 Max. :98.000   
## Casualty\_Home\_Area\_Type Casualty\_IMD\_Decile  
## Min. :-1.000 Min. :-1.000   
## 1st Qu.: 1.000 1st Qu.: 1.000   
## Median : 1.000 Median : 4.000   
## Mean : 0.986 Mean : 3.714   
## 3rd Qu.: 1.000 3rd Qu.: 7.000   
## Max. : 3.000 Max. :10.000

summary(UK\_VehicleData)

## Accident\_Index Vehicle\_Reference Vehicle\_Type Towing\_and\_Articulation  
## Length:975680 Min. : 1.000 Min. :-1.000 Min. :-1.00000   
## Class :character 1st Qu.: 1.000 1st Qu.: 9.000 1st Qu.: 0.00000   
## Mode :character Median : 1.000 Median : 9.000 Median : 0.00000   
## Mean : 1.562 Mean : 9.804 Mean : 0.02147   
## 3rd Qu.: 2.000 3rd Qu.: 9.000 3rd Qu.: 0.00000   
## Max. :999.000 Max. :98.000 Max. : 5.00000   
## Vehicle\_Manoeuvre Vehicle\_Location\_Restricted\_Lane Junction\_Location  
## Min. :-1.00 Min. :-1.0000 Min. :-1.000   
## 1st Qu.: 6.00 1st Qu.: 0.0000 1st Qu.: 0.000   
## Median :17.00 Median : 0.0000 Median : 1.000   
## Mean :12.43 Mean : 0.1404 Mean : 2.441   
## 3rd Qu.:18.00 3rd Qu.: 0.0000 3rd Qu.: 5.000   
## Max. :18.00 Max. : 9.0000 Max. : 8.000   
## Skidding\_and\_Overturning Hit\_Object\_in\_Carriageway Vehicle\_Leaving\_Carriageway  
## Min. :-1.0000 Min. :-1.0000 Min. :-1.0000   
## 1st Qu.: 0.0000 1st Qu.: 0.0000 1st Qu.: 0.0000   
## Median : 0.0000 Median : 0.0000 Median : 0.0000   
## Mean : 0.1827 Mean : 0.2983 Mean : 0.3094   
## 3rd Qu.: 0.0000 3rd Qu.: 0.0000 3rd Qu.: 0.0000   
## Max. : 5.0000 Max. :12.0000 Max. : 8.0000   
## Hit\_Object\_off\_Carriageway First\_Point\_of\_Impact Was\_Vehicle\_Left\_Hand\_Drive  
## Min. :-1.0000 Min. :-1.000 Min. :-1.0000   
## 1st Qu.: 0.0000 1st Qu.: 1.000 1st Qu.: 1.0000   
## Median : 0.0000 Median : 1.000 Median : 1.0000   
## Mean : 0.4375 Mean : 1.754 Mean : 0.9695   
## 3rd Qu.: 0.0000 3rd Qu.: 3.000 3rd Qu.: 1.0000   
## Max. :11.0000 Max. : 4.000 Max. : 2.0000   
## Journey\_Purpose\_of\_Driver Sex\_of\_Driver Age\_of\_Driver Age\_Band\_of\_Driver  
## Min. :-1.000 Min. :-1.000 Min. : -1.00 Min. :-1.000   
## 1st Qu.: 2.000 1st Qu.: 1.000 1st Qu.: 23.00 1st Qu.: 5.000   
## Median : 6.000 Median : 1.000 Median : 35.00 Median : 6.000   
## Mean : 4.714 Mean : 1.432 Mean : 35.49 Mean : 5.959   
## 3rd Qu.: 6.000 3rd Qu.: 2.000 3rd Qu.: 50.00 3rd Qu.: 8.000   
## Max. : 6.000 Max. : 3.000 Max. :101.00 Max. :11.000   
## Engine\_Capacity\_CC Propulsion\_Code Age\_of\_Vehicle Driver\_IMD\_Decile  
## Min. : -1 Min. :-1.000 Min. : -1.000 Min. :-1.000   
## 1st Qu.: 113 1st Qu.: 1.000 1st Qu.: -1.000 1st Qu.:-1.000   
## Median : 1390 Median : 1.000 Median : 5.000 Median : 1.000   
## Mean : 1422 Mean : 0.936 Mean : 5.569 Mean : 2.206   
## 3rd Qu.: 1910 3rd Qu.: 2.000 3rd Qu.: 10.000 3rd Qu.: 5.000   
## Max. :99999 Max. :12.000 Max. :105.000 Max. :10.000   
## Driver\_Home\_Area\_Type Vehicle\_IMD\_Decile  
## Min. :-1.0000 Min. :-1.000   
## 1st Qu.: 1.0000 1st Qu.:-1.000   
## Median : 1.0000 Median : 1.000   
## Mean : 0.8256 Mean : 2.206   
## 3rd Qu.: 1.0000 3rd Qu.: 5.000   
## Max. : 3.0000 Max. :10.000

### New Questions

Based on the summary review, here are some questions to pursue.

* Accidents Table
  + How many accidents involve 0, 1, 2, or more vehicles?
  + How many accidents involve a casualty/injury of any kind?
  + How many accidents occur at various speed limit ranges?
  + How many accidents occur on which days of the week?
  + How many accidents occur at various light conditions?
  + How many accidents occur at various weather conditions?
  + How many accidents for each IMD?
* Casualty Table
  + How many casualties for each sex/gender?
  + How many casualties for each age band?
  + How many casualties for each Home Areas?
  + How many casualties for each IMD?
* Vehicles Table
  + How many of each vehicle type? We know cars are most common, but which are 2nd or 3rd?
  + Which Manoeuvres are most common?
  + Besides main roads, which other Locations are common?
  + Explore variables, see what else sticks out.

### UK\_AccidentData Variable Analysis

The following represents a look at the counts for occurrences of variables in the UK\_AccidentData dataset. Comments above each represent findings.

# As expected, most accidents involved two vehicles, then one, then three.  
# Numbers drop off significantly after this.  
UK\_AccidentData %>%   
 count(Number\_of\_Vehicles)

## Number\_of\_Vehicles n  
## 1 1 155544  
## 2 2 319053  
## 3 3 42056  
## 4 4 9372  
## 5 5 2182  
## 6 6 637  
## 7 7 256  
## 8 8 87  
## 9 9 51  
## 10 10 27  
## 11 11 9  
## 12 12 4  
## 13 13 3  
## 14 14 3  
## 15 15 1  
## 16 16 5  
## 17 18 1  
## 18 23 1  
## 19 24 1  
## 20 37 1

# At first glance, this doesn't appear to say much. Weekends are fewer.  
UK\_AccidentData %>%   
 count(Day\_of\_Week)

## Day\_of\_Week n  
## 1 1 59060  
## 2 2 74984  
## 3 3 78967  
## 4 4 79955  
## 5 5 80881  
## 6 6 86219  
## 7 7 69228

# There are a LOT more accidents at 30 MPH than at any other speed limit.  
UK\_AccidentData %>%   
 count(Speed\_limit)

## Speed\_limit n  
## 1 0 1  
## 2 10 2  
## 3 20 31333  
## 4 30 327671  
## 5 40 44098  
## 6 50 21652  
## 7 60 69574  
## 8 70 34926  
## 9 NA 37

# The vast majority of accidents occur during the day (when most are driving)  
# Next is darkness in well lit aready (again, where most people are driving)  
UK\_AccidentData %>%   
 count(Light\_Conditions)

## Light\_Conditions n  
## 1 -1 14  
## 2 1 382690  
## 3 4 106713  
## 4 5 3362  
## 5 6 27075  
## 6 7 9440

# The lion share of accidents occur when weather is fine (when most people are driving)  
# Of the remaining, the most occur in rain.  
UK\_AccidentData %>%   
 count(Weather\_Conditions)

## Weather\_Conditions n  
## 1 -1 33  
## 2 1 430121  
## 3 2 57140  
## 4 3 2602  
## 5 4 5990  
## 6 5 6130  
## 7 6 755  
## 8 7 2480  
## 9 8 9507  
## 10 9 14536

# Not useful, needs better grouping.   
#UK\_AccidentData %>%   
# count(Time)  
  
# This will be very useful for Visualization in Tableau,   
# added this new column to table in SQL  
  
UK\_AccidentData %>%   
 mutate(TimeGroup = case\_when(  
 Time >= "06:00:00.0000000" & Time < "9:00:00.0000000" ~ "Morning\_Commute",  
 Time >= "09:00:00.0000000" & Time < "11:00:00.0000000" ~ "Morning\_Late",  
 Time >= "11:00:00.0000000" & Time < "13:00:00.0000000" ~ "Lunch\_Hours",  
 Time >= "13:00:00.0000000" & Time < "17:00:00.0000000" ~ "Afternoon",  
 Time >= "17:00:00.0000000" & Time < "20:00:00.0000000" ~ "Evening\_Commute",  
 Time >= "20:00:00.0000000" & Time < "21:00:00.0000000" ~ "Evening\_Late",  
 TRUE ~ "Late\_Night"  
 )) %>%   
 count(TimeGroup)

## TimeGroup n  
## 1 Late\_Night 28984  
## 2 Morning\_Commute 500310

# Most accidents occur in the Afternoon (1p-5p)  
# Second most occur during the evening commute (5p-8p)  
  
UK\_AccidentData %>%   
 count(TimeGroup)

## TimeGroup n  
## 1 Afternoon 147299  
## 2 Evening\_Commute 112939  
## 3 Evening\_Late 20148  
## 4 Late\_Night 69267  
## 5 Lunch\_Hours 57138  
## 6 Morning\_Commute 73060  
## 7 Morning\_Late 49443

# The following three seem evenly dispersed.   
# Maybe let's get a dashboard with these?  
# See if anything pops out visually?  
  
# Seems fairly evenly dispersed?  
UK\_AccidentData %>%   
 count(Police\_Force) %>%   
 arrange(-n) %>%   
 head(10)

## Police\_Force n  
## 1 1 101812  
## 2 20 23173  
## 3 46 19116  
## 4 13 18757  
## 5 43 18022  
## 6 44 16770  
## 7 47 16216  
## 8 50 14458  
## 9 45 14215  
## 10 4 14120

# Seems fairly evenly dispersed?  
  
UK\_AccidentData %>%   
 count(Local\_Authority\_District) %>%  
 arrange(-n) %>%  
 head(10)

## Local\_Authority\_District n  
## 1 300 11095  
## 2 204 7164  
## 3 1 6319  
## 4 9 5114  
## 5 596 4759  
## 6 926 4456  
## 7 5 4405  
## 8 200 4375  
## 9 215 4375  
## 10 8 4160

# Seems fairly evenly dispersed?  
UK\_AccidentData %>%   
 count(Local\_Authority\_Highway) %>%  
 arrange(-n) %>%  
 head(10)

## Local\_Authority\_Highway n  
## 1 E10000016 16625  
## 2 E10000030 14215  
## 3 E10000012 11190  
## 4 E10000014 11130  
## 5 E08000025 11095  
## 6 E10000017 11089  
## 7 E10000015 9447  
## 8 E10000019 7912  
## 9 E10000032 7862  
## 10 E08000035 7164

# No Intersection & T Intersections are most common accident types.  
# That is surprising, given the MPH 30 thing? Let's look at MPH, Junctions, and Time.   
  
UK\_AccidentData %>%   
 count(Junction\_Detail)

## Junction\_Detail n  
## 1 -1 1488  
## 2 0 218682  
## 3 1 44888  
## 4 2 6698  
## 5 3 162272  
## 6 5 7828  
## 7 6 50729  
## 8 7 5077  
## 9 8 15585  
## 10 9 16047

# Mostly missing data (-1)  
UK\_AccidentData %>%   
 count(Junction\_Control)

## Junction\_Control n  
## 1 -1 223831  
## 2 0 388  
## 3 1 1375  
## 4 2 56285  
## 5 3 2945  
## 6 4 244470

# Mostly missing data (-1), filter these out.  
# With (-1, missing data) filtered out, these show a strong preference for 4.  
# Uncontroled intersections lead the way.  
UK\_AccidentData %>%   
 filter(Junction\_Control != -1) %>%   
 count(Junction\_Control)

## Junction\_Control n  
## 1 0 388  
## 2 1 1375  
## 3 2 56285  
## 4 3 2945  
## 5 4 244470

# One lane roads dominate  
UK\_AccidentData %>%   
 count(Road\_Type)

## Road\_Type n  
## 1 -1 1  
## 2 1 34444  
## 3 2 12742  
## 4 3 80196  
## 5 6 388887  
## 6 7 5875  
## 7 9 7149

# Most accidents (383,735) occurred on dry roads.  
# 2nd Most (131,236) on wet roads.  
# As these are the two most common road conditions, that is only so helpful to know.  
UK\_AccidentData %>%   
 count(Road\_Surface\_Conditions)

## Road\_Surface\_Conditions n  
## 1 -1 4213  
## 2 1 383735  
## 3 2 131236  
## 4 3 2325  
## 5 4 7205  
## 6 5 580

# Roughly 3/4 of accidents had a Police Officer  
# Roughly 1/4 did not  
# A fraction had no police officer but did self-report via a form.  
UK\_AccidentData %>%   
 count(Did\_Police\_Officer\_Attend\_Scene\_of\_Accident)

## Did\_Police\_Officer\_Attend\_Scene\_of\_Accident n  
## 1 -1 10  
## 2 1 399445  
## 3 2 126704  
## 4 3 3135

### UK\_VehicleData Variable Analysis

The following represents a look at the counts for occurrences of variables in the UK\_VehicleData dataset. Comments above each represent findings.

# TBC

### UK\_CasualtyData Variable Analysis

The following represents a look at the counts for occurrences of variables in the UK\_CasualtyData dataset. Comments above each represent findings.

# TBC

### xxxxxx

### xxxxxx

### xxxxxx

### xxxxxx

### xxxxxx

### xxxxxx

### xxxxxx