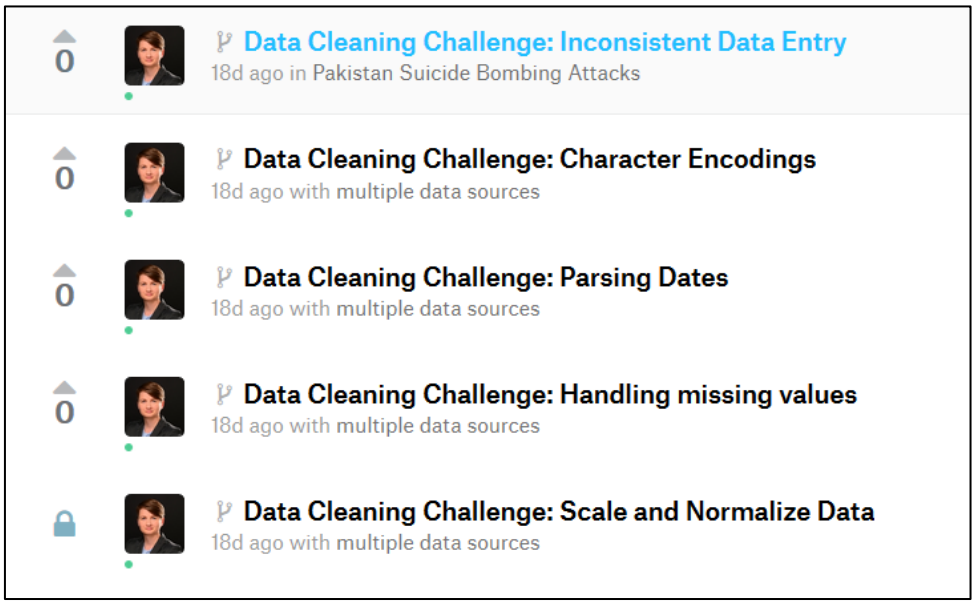
DATA PREPARATION

Kaggle “5 Day Data Cleaning Challenge” by Rachel Tatema 2
Play around with a tool 3
Tool 3
Dataset 3
Cleaning 3
Going forward with the flow 11

KAGGLE "5 DAY DATA CLEANING CHALLENGE" BY RACHEL TATEMA

https://www.kaggle.com/kewagbln/kernels



1 Kaggle kernels forked from the originals

PLAY AROUND WITH A TOOL

TOOL

I have chosen Tableau for merging and cleaning the files instead of Trifacta, because:

- of lot of resources with tutorials
- full free license for students
- widely used tool

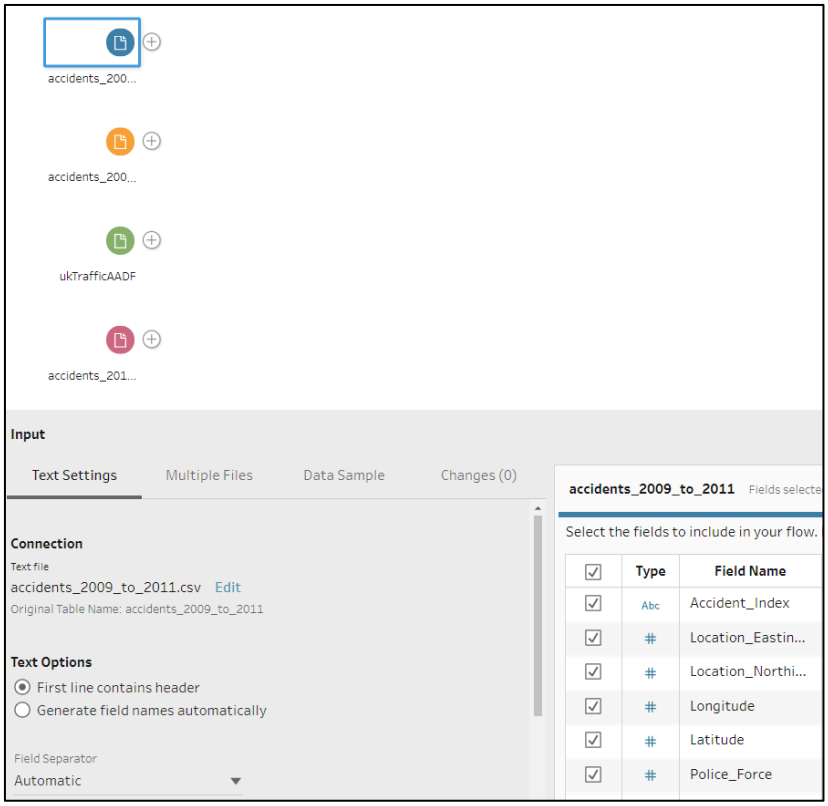
DATASET

https://www.kaggle.com/daveianhickey/2000-16-traffic-flow-england-scotland-wales

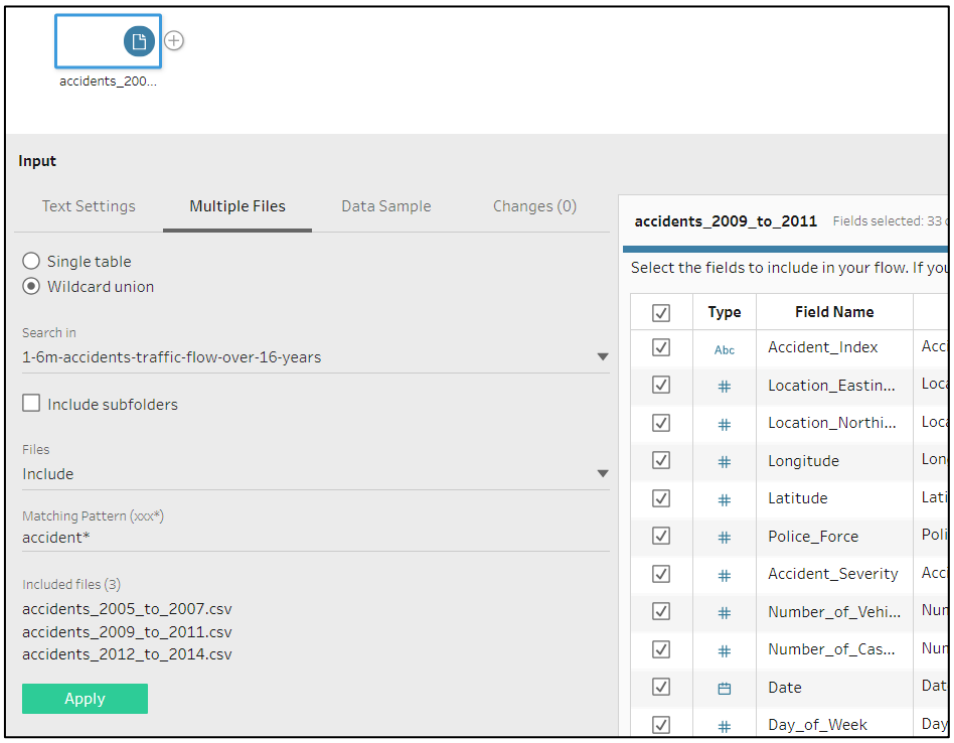
Data Sources	About this file	Columns
accidents_2005_t... 35 columns accidents_2009_t... 35 columns accidents_2012_to... 35 columns ukTrafficAADF.csv 29 columns accident_coords_update.ipynb Areas.shp Local_Authority_Districts_Dec_201...	UK road accidents from 2005 to 2007. Note: the columns are the same for all three files included in this dataset.	# # Unnamed: 0 A Accident_Index Unique ID. # Location_Easting_OSGR Local British coordinates x-value. # Location_Northing_OSGR Local British coordinates y-value. # Longitude # Latitude

2 Four csv files: 3 with accidents + 1 with additional data

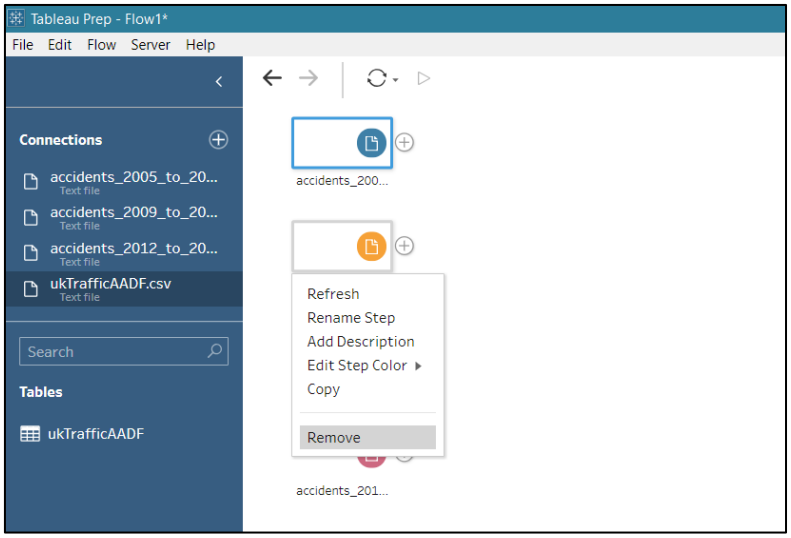
CLEANING



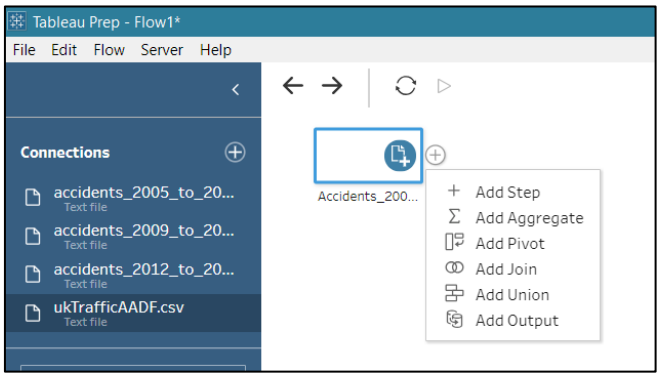
3 After loading the files: 4 inputs created – 3x accident + the 1 with traffic data*



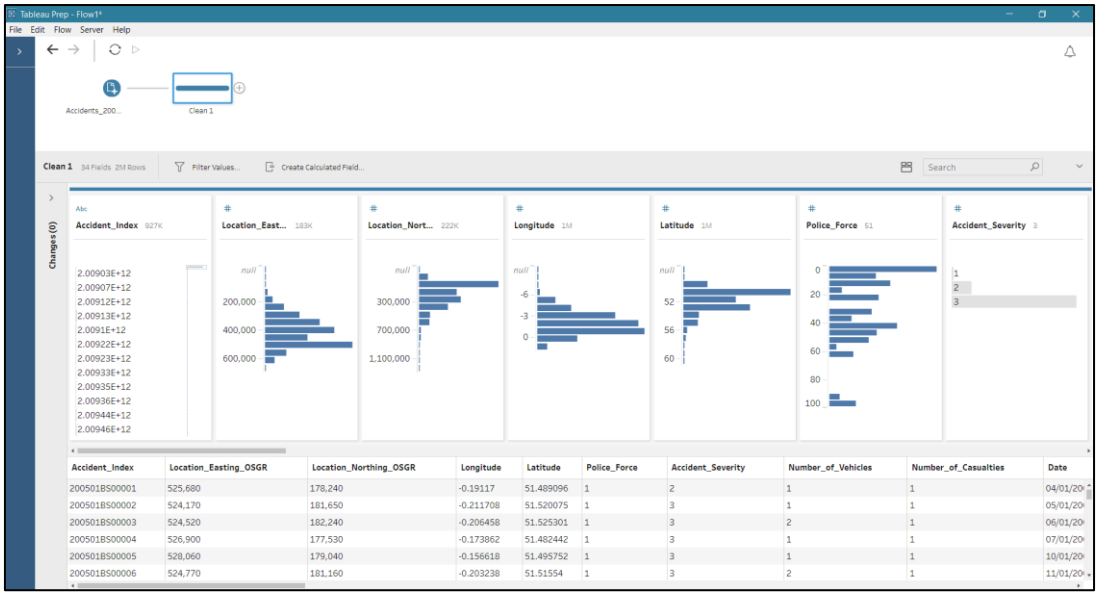
4 Accident inputs can be merge via "multiple files" and "wildcard union"



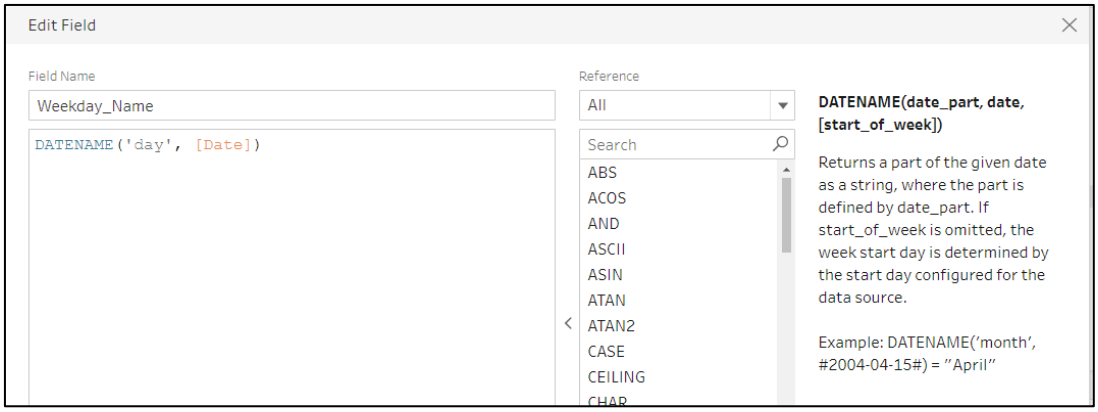
5 The unnecessary files can be easily removed



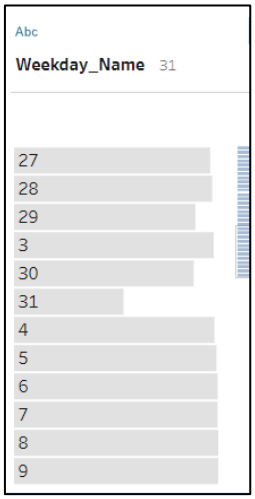
6 "add step" adds a by default "clean step"



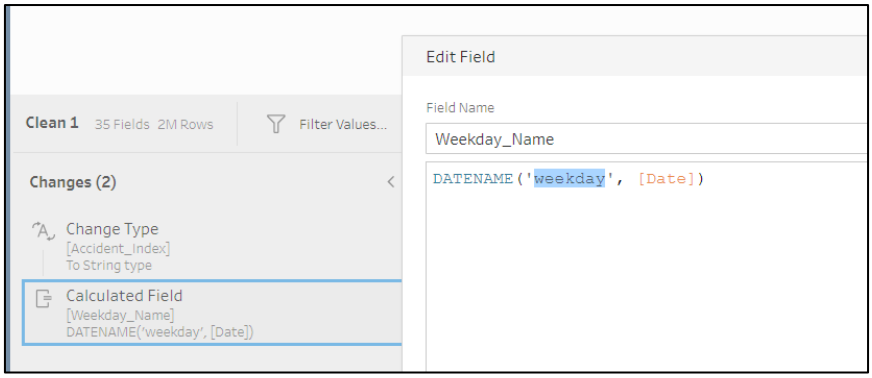
7 Clicking on the step shows data and allows to change them



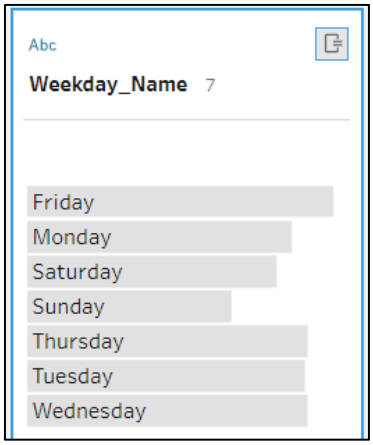
8 Add a new column for the weekday name calculation



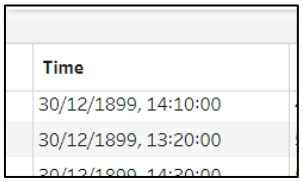
9 Result, instead of "Monday, Tuesday, ..."



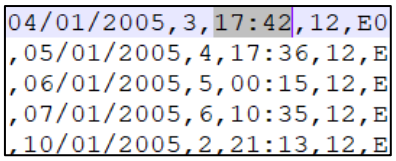
10 Calculation can be easy refreshed



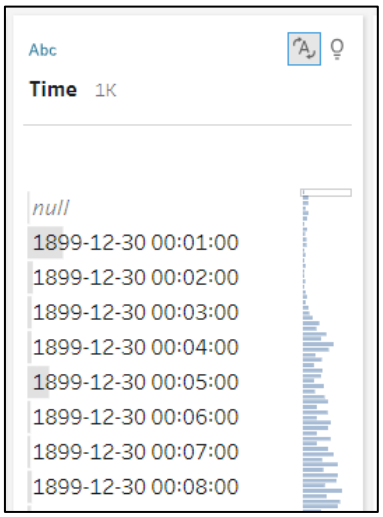
11 Expected result



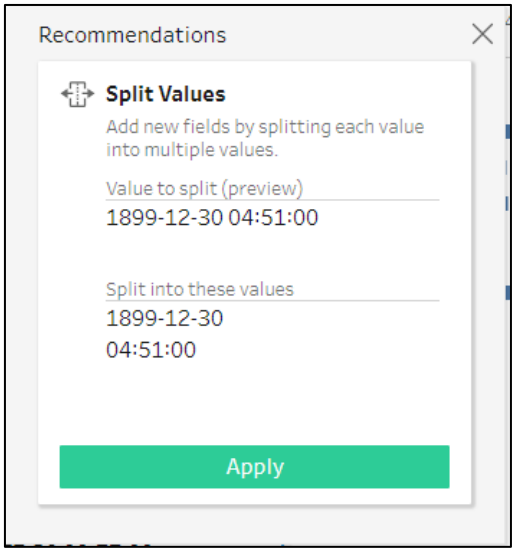
12 The time column looks stranges



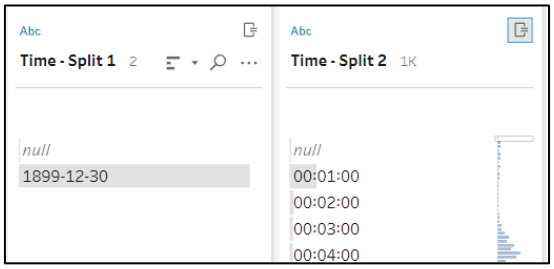
13 A look into the csv shows, that there should be only the time, not the date. But Tableau knows only "date + time"



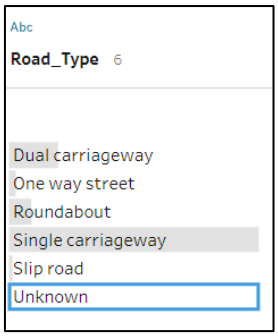
14 First step to clean the column: change to string



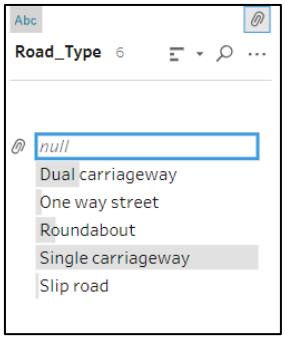
15 Tableau recommends a split into two colums



16 The result are two new columns- the left one and the old one can be removed

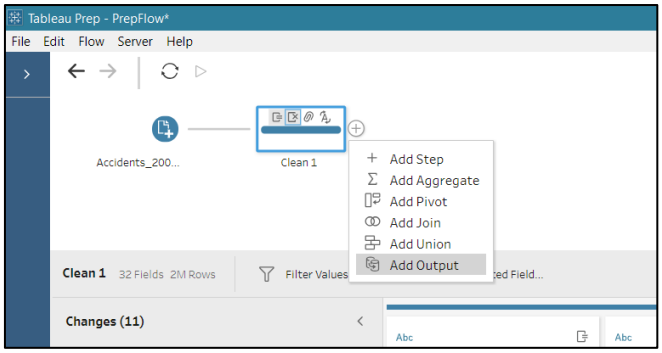


17 "Roadtype" also uses an "unknown" – let's make it empty

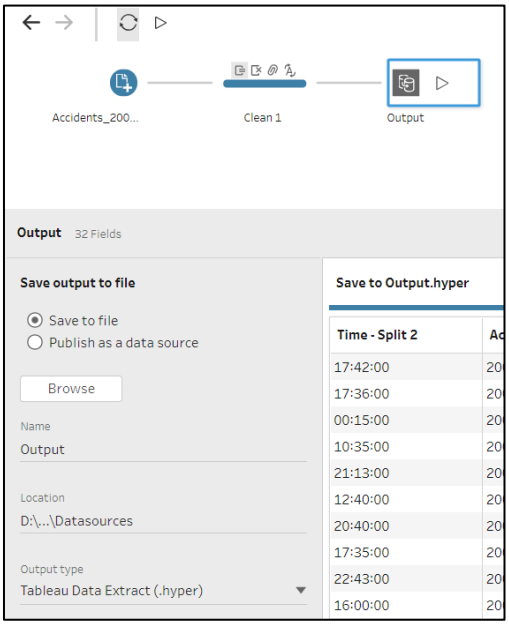


18 Done by right click -> "replace with null"

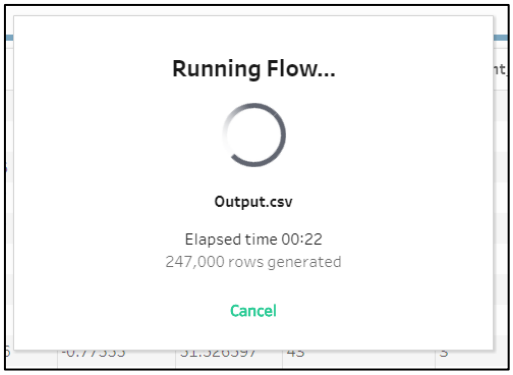
Remove unnecessary columns likes filepath, location_easting/northing (we have also latitude and longtitude)



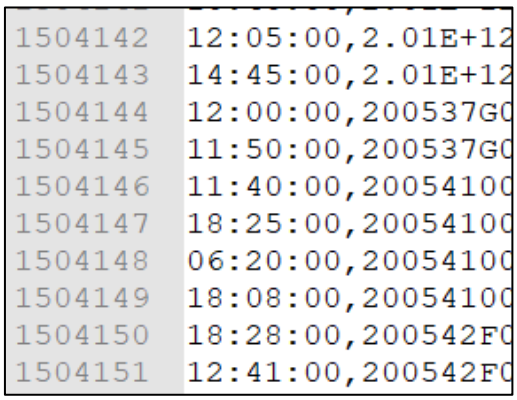
19 Last step in Tableau Prep: save the output



20 Select the output file name and file type



21 Click on "run flow"



1504142	12:05:00,2.01E+12
1504143	14:45:00,2.01E+12
1504144	12:00:00,200537G0
1504145	11:50:00,200537G0
1504146	11:40:00,20054100
1504147	18:25:00,20054100
1504148	06:20:00,20054100
1504149	18:08:00,20054100
1504150	18:28:00,200542F0
1504151	12:41:00,200542F0

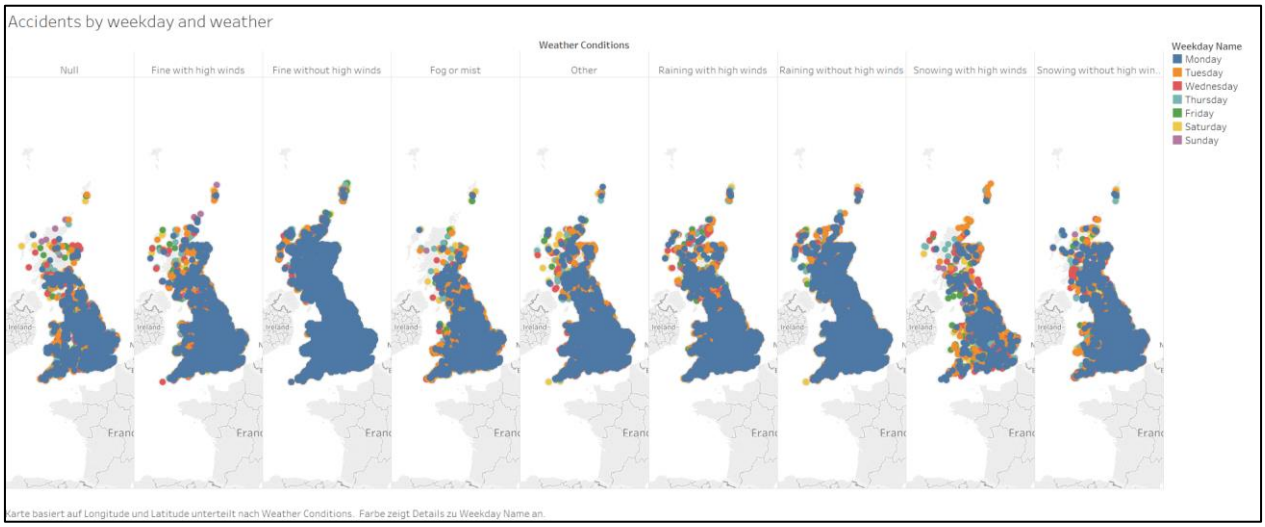
22 Final csv has 1.504.151 rows

The output will be created and the files saved.

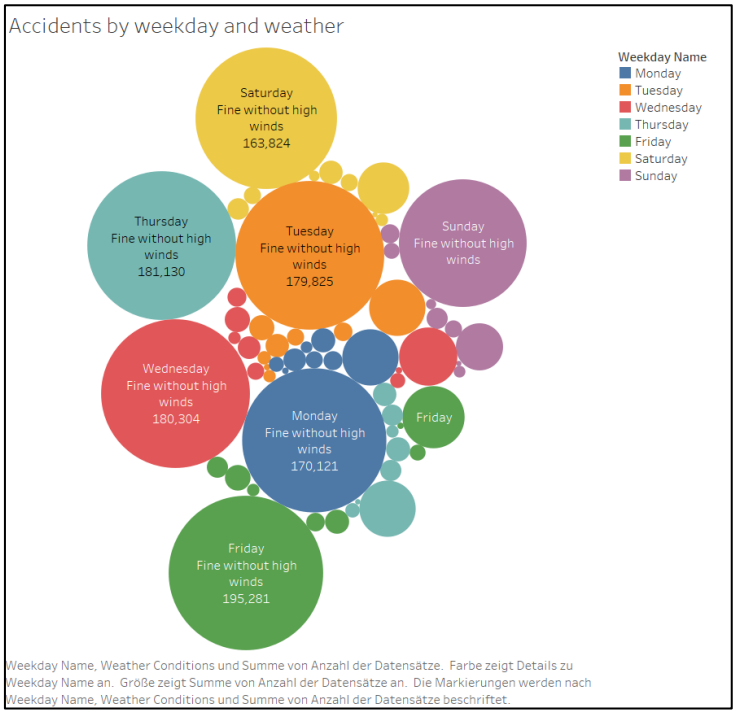
GOING FORWARD WITH THE FLOW

I have also saved the output as *.hyper, and then I have gone some steps forward and played with Tableau Desktop.

Just some clicks – et voilà:



23 Accidents by weekday and weather situation on a map



24 Same data as a bubble chart