Data Preparation

[Kaggle “5 Day Data Cleaning Challenge” by Rachel Tatema 2](#_Toc536763811)

[Play around with a tool 3](#_Toc536763812)

[Tool 3](#_Toc536763813)

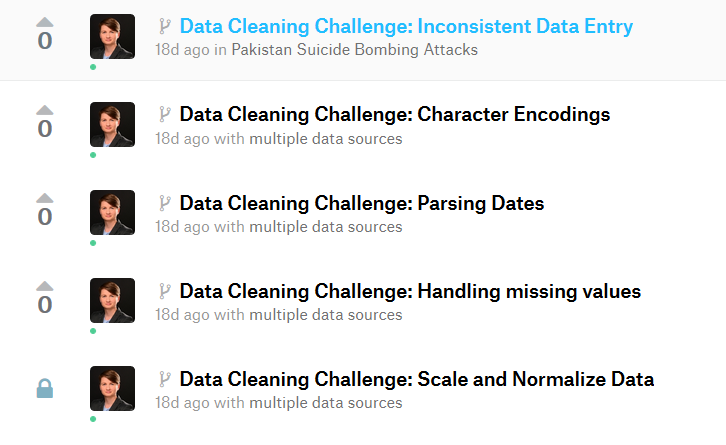
[Dataset 3](#_Toc536763814)

[Cleaning 3](#_Toc536763815)

[Going forward with the flow 11](#_Toc536763816)

# 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

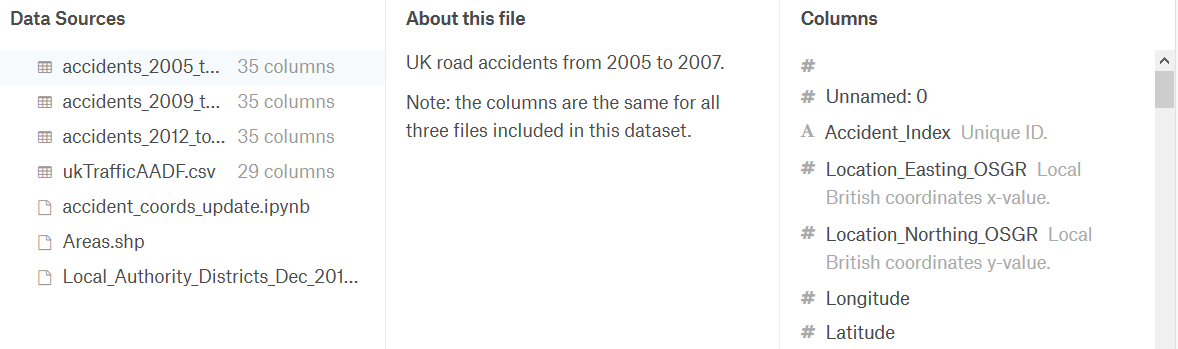
## Tool

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

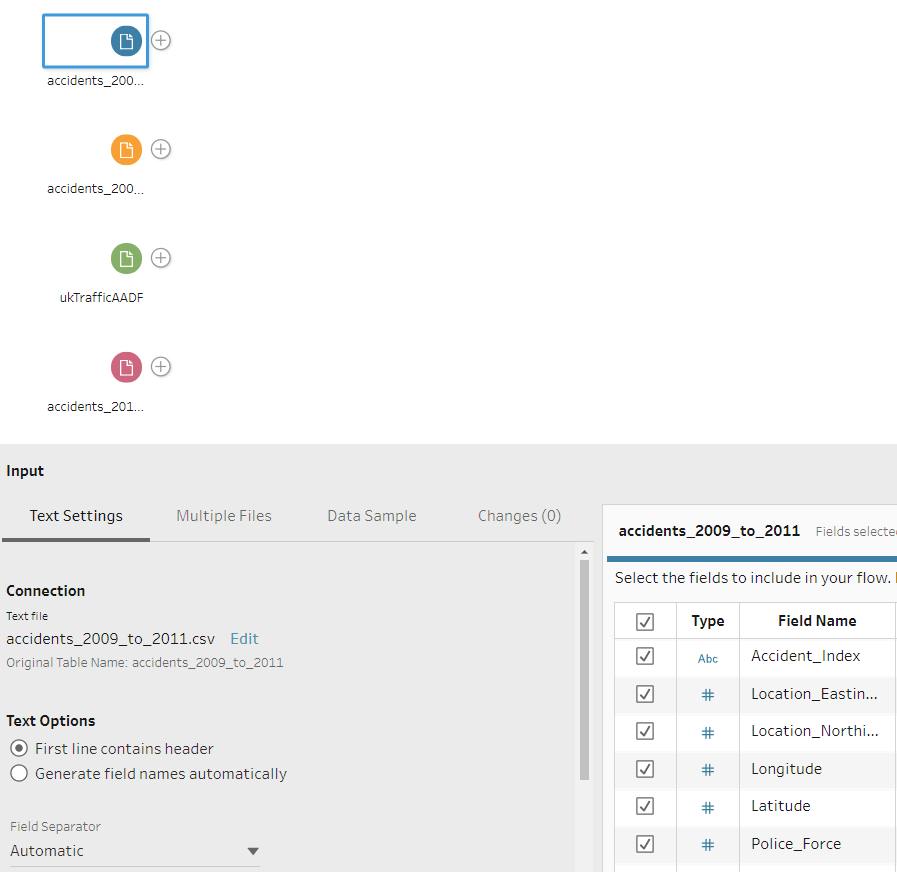
## Dataset

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

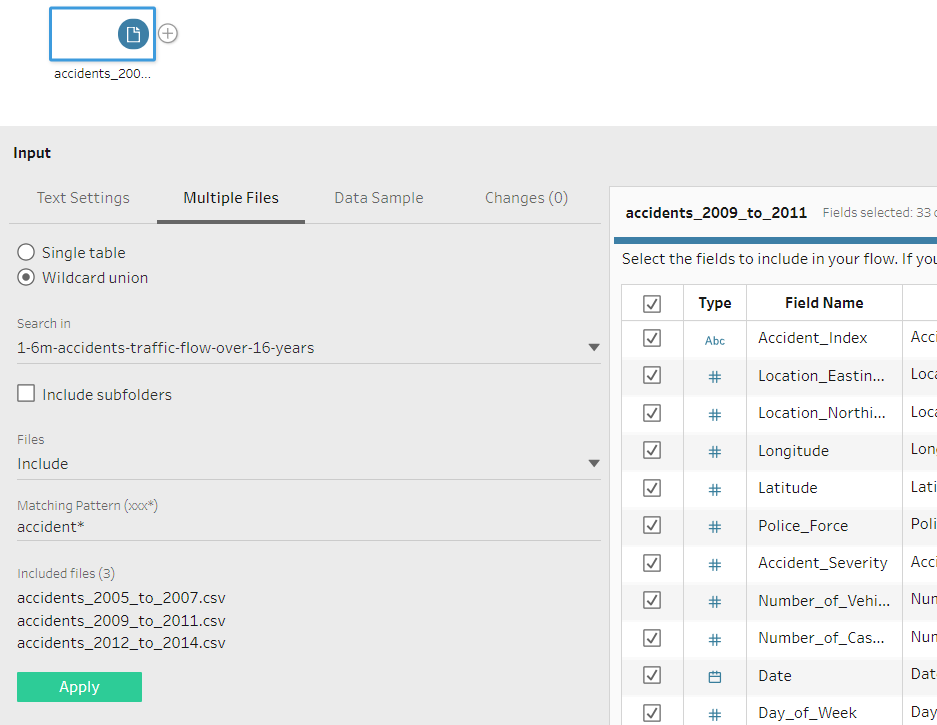


2 Four csv files: 3 with accidents + 1 with additional data

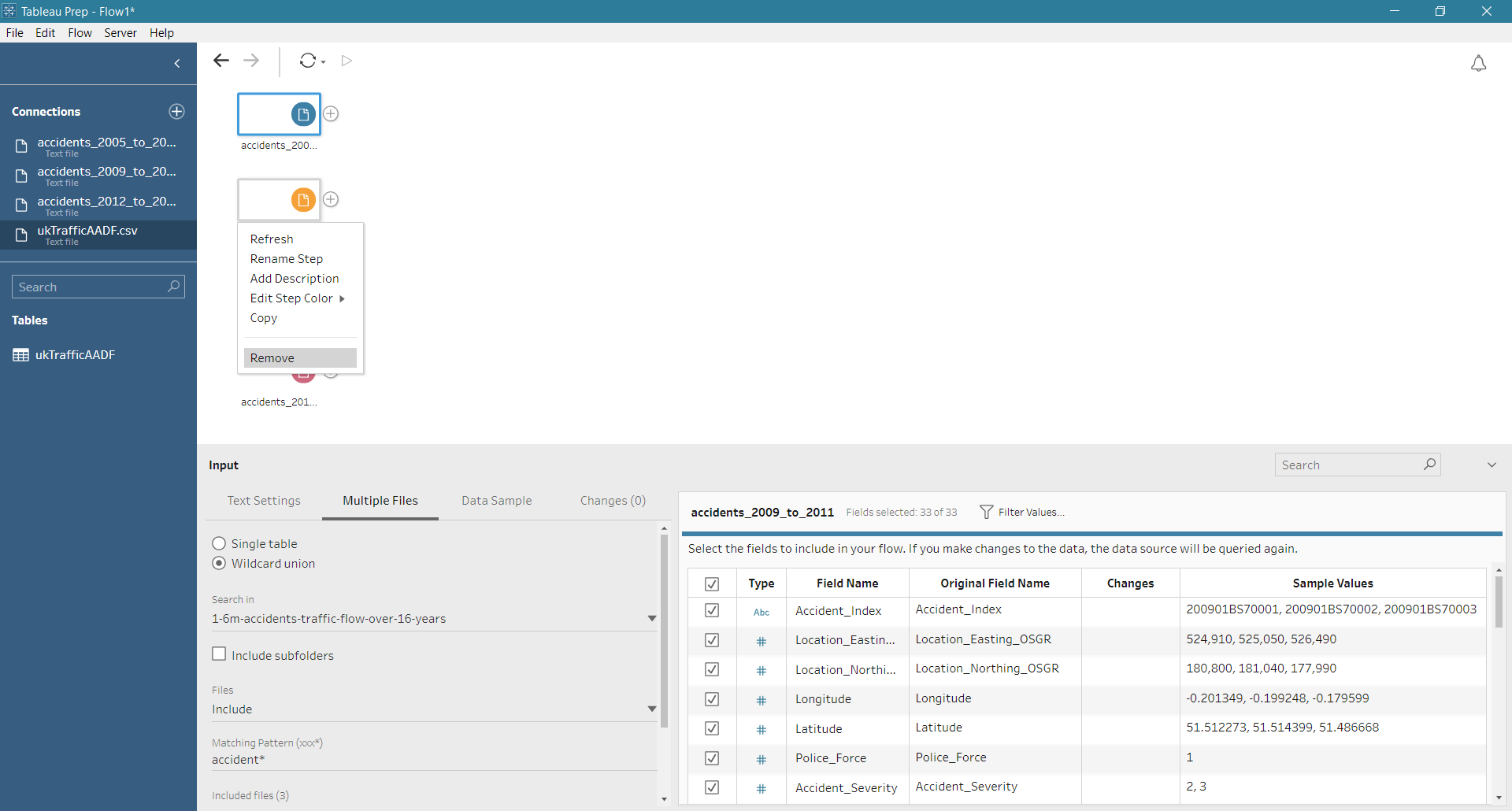
## Cleaning



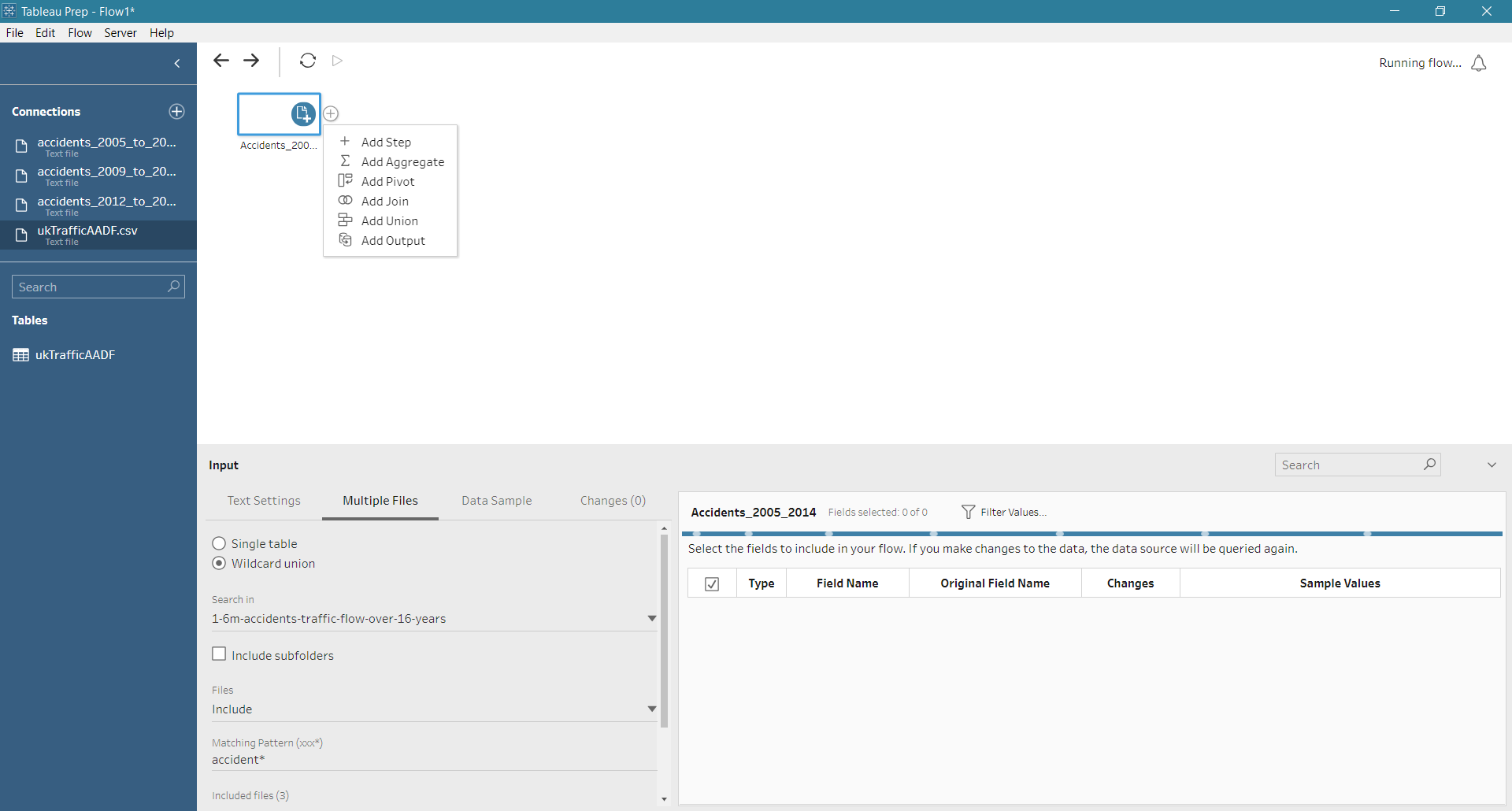
3 After loading the files: 4 inputs created – 3x accident\* + the 1 with traffic data



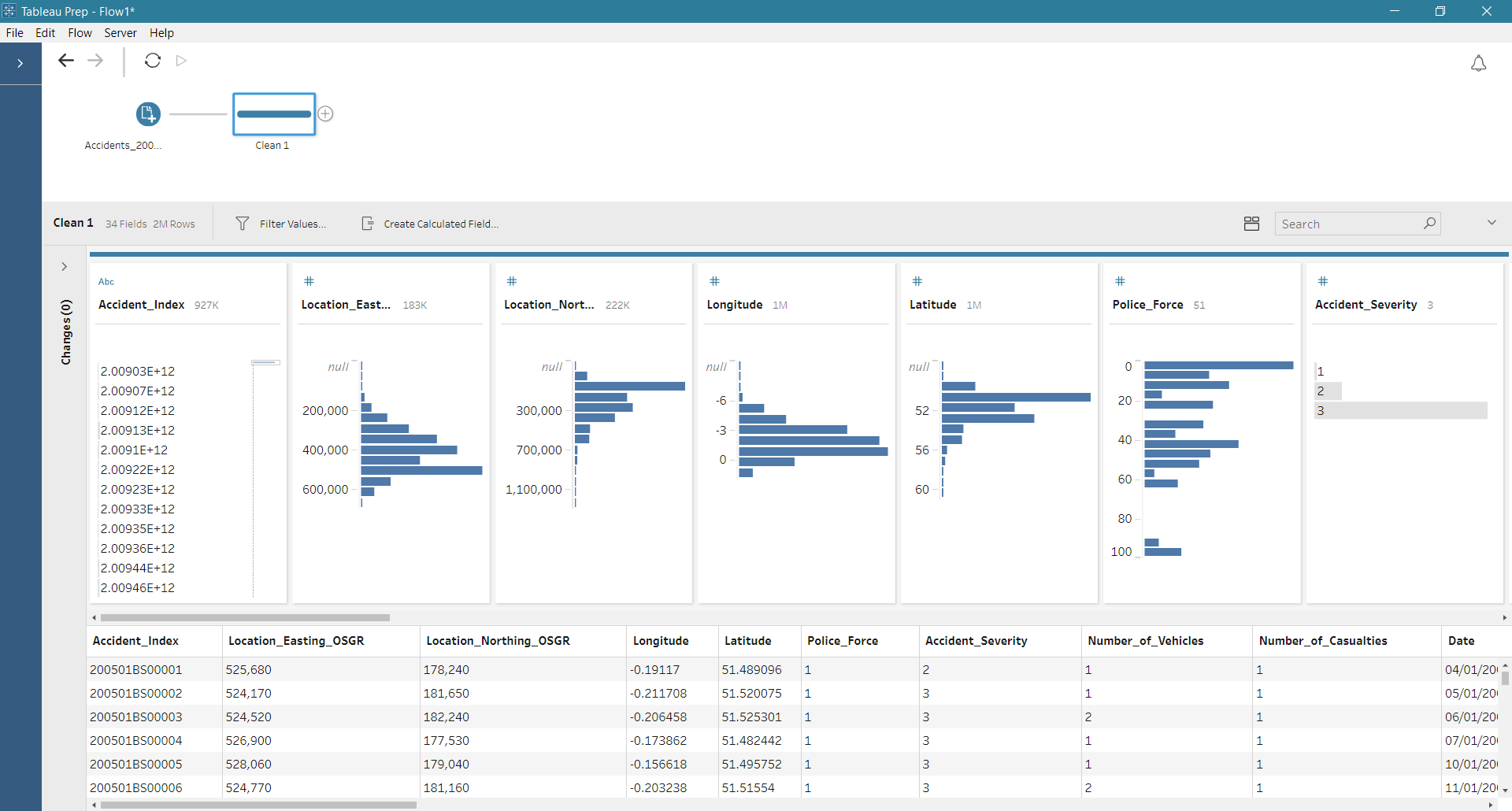
4 Accident inputs can be merge via “multiple files” and “wildcard union”



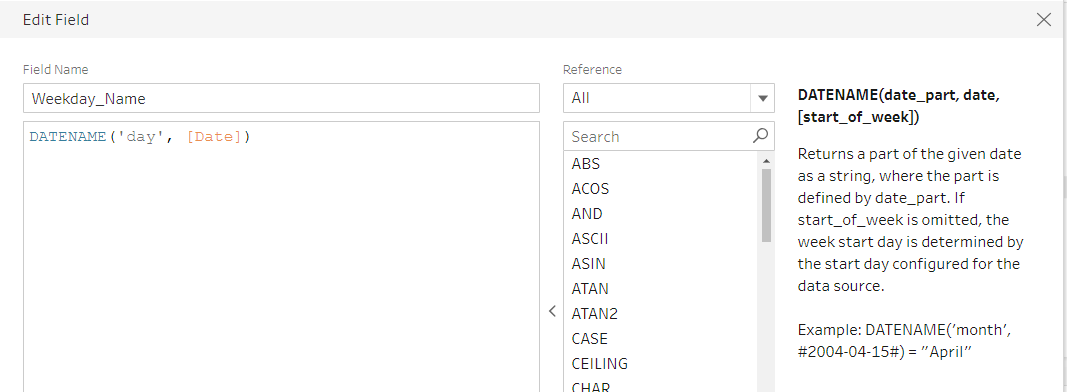
5 The unnecessary files can be easily removed



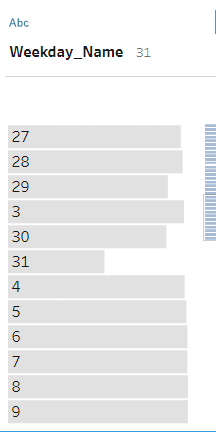
6 “add step” adds a by default “clean step”



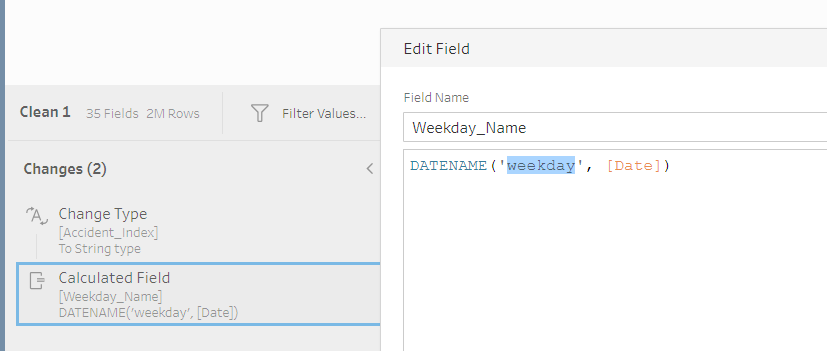
7 Clicking on the step shows data and allows to change them



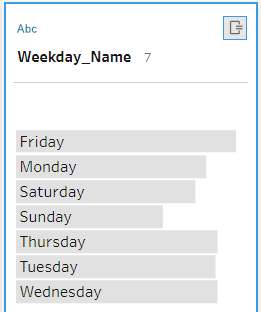
8 Add a new column for the weekday name calculation



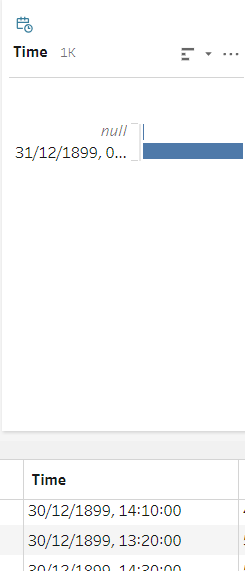
9 Result, instead of "Monday, Tuesday, ..."



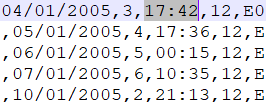
10 Calculation can be easy refreshed



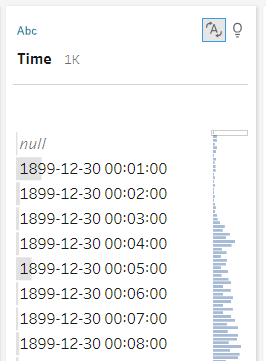
11 Expected result



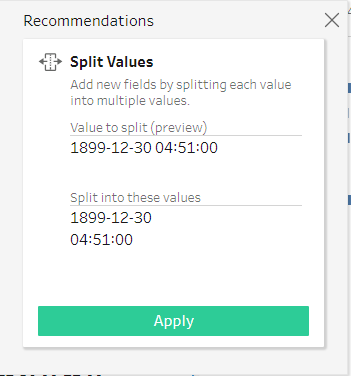
12 The time column looks stranges



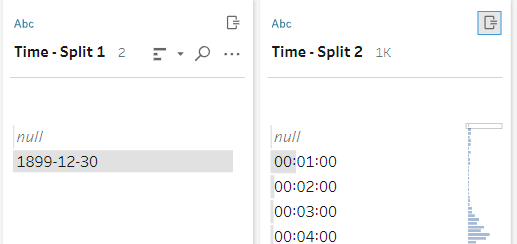
13 A look into the csv shows, that there should be only the time, not the date. But Tableau knows only "date + time"



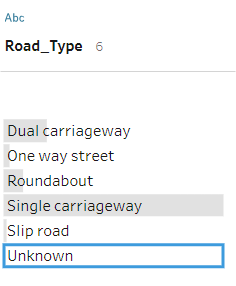
14 First step to clean the column: change to string



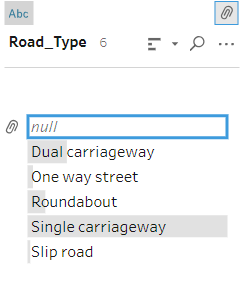
15 Tableau recommends a split into two colums



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

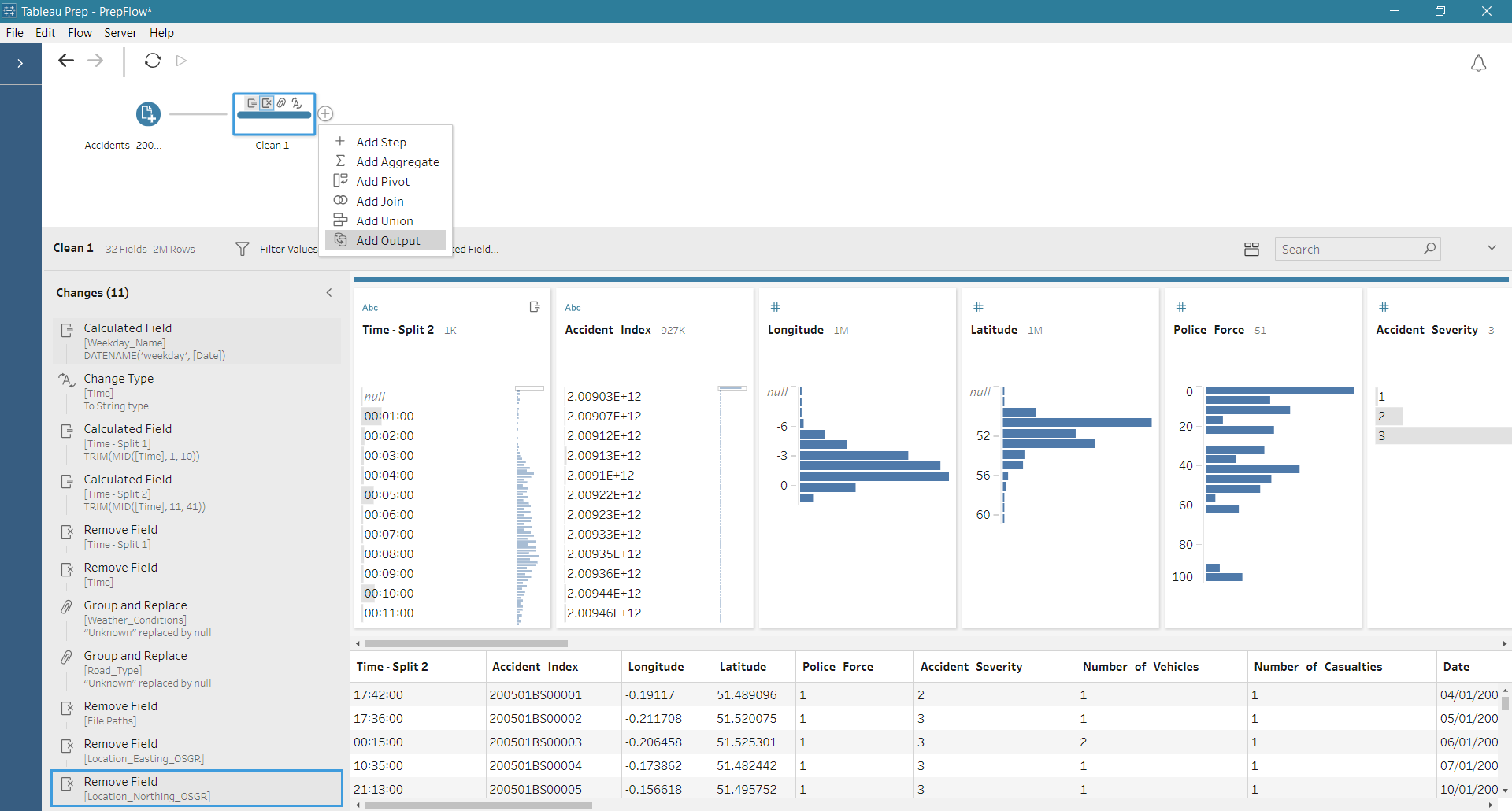


17 “Roadtype” also uses an “unknown” – let’s make it empty

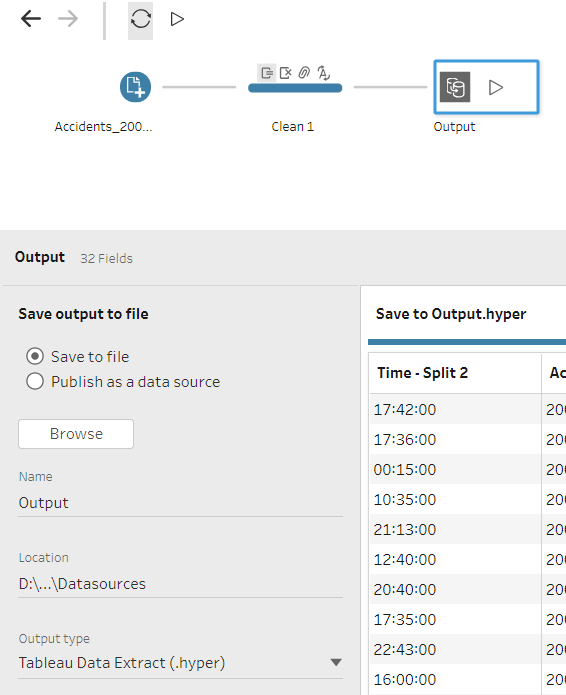


18 Done by right click -> “replace with null”

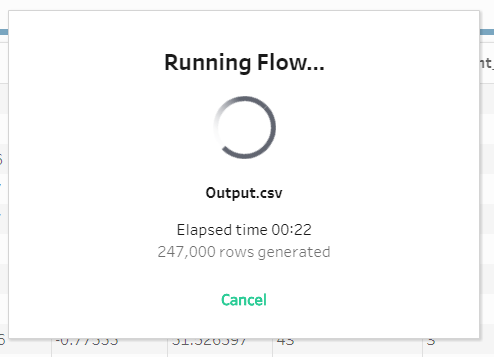
Remove unnecessary columns likes filepath, location\_easting/northing (we have also latitude and longtitude)



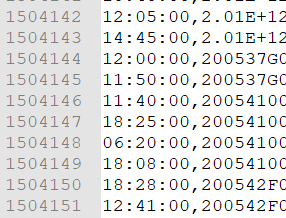
19 Last step in Tableau Prep: save the output



20 Select the output file name and file type



21 Click on “run flow”



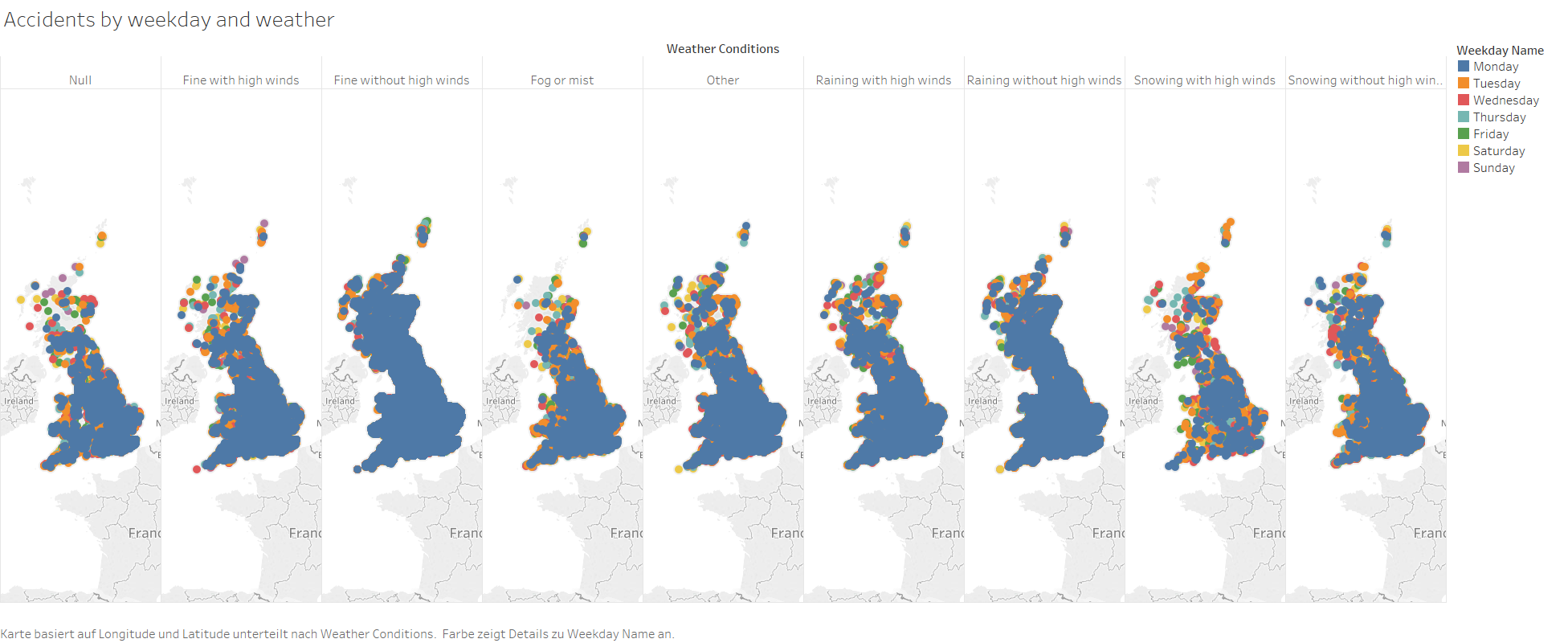
22 Final csv has 1.504.151 rows

The output will be created and the files saved.

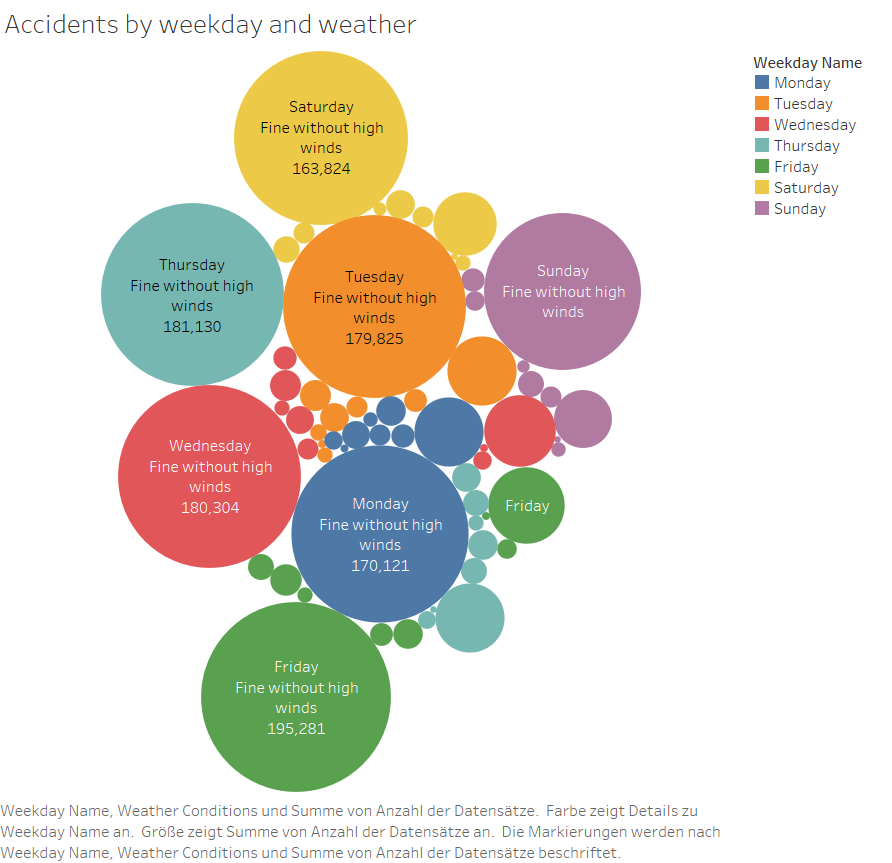
# Going forward with the flow

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