




# 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>

**Data Cleaning Challenge: Inconsistent Data Entry**  
18d ago in Pakistan Suicide Bombing Attacks

**Data Cleaning Challenge: Character Encodings**  
18d ago with multiple data sources

**Data Cleaning Challenge: Parsing Dates**  
18d ago with multiple data sources

**Data Cleaning Challenge: Handling missing values**  
18d ago with multiple data sources

**Data Cleaning Challenge: Scale and Normalize Data**  
18d ago with multiple data sources

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

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

### CLEANING

**Input**

Text Settings Multiple Files Data Sample Changes (0)

**accidents\_2009\_to\_2011** Fields select

Select the fields to include in your flow.

<input checked="" type="checkbox"/>	Type	Field Name
<input checked="" type="checkbox"/>	Abc	Accident_Index
<input checked="" type="checkbox"/>	#	Location_Eastin...
<input checked="" type="checkbox"/>	#	Location_Northi...
<input checked="" type="checkbox"/>	#	Longitude
<input checked="" type="checkbox"/>	#	Latitude
<input checked="" type="checkbox"/>	#	Police_Force

**Connection**  
Text file  
accidents\_2009\_to\_2011.csv [Edit](#)  
Original Table Name: accidents\_2009\_to\_2011

**Text Options**  
☒ First line contains header  
☐ Generate field names automatically

Field Separator  
Automatic

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

**Input**

Text Settings **Multiple Files** Data Sample Changes (0)

**accidents\_2009\_to\_2011** Fields selected: 33

Select the fields to include in your flow. If you

<input checked="" type="checkbox"/>	Type	Field Name
<input checked="" type="checkbox"/>	Abc	Accident_Index
<input checked="" type="checkbox"/>	#	Location_Eastin...
<input checked="" type="checkbox"/>	#	Location_Northi...
<input checked="" type="checkbox"/>	#	Longitude
<input checked="" type="checkbox"/>	#	Latitude
<input checked="" type="checkbox"/>	#	Police_Force
<input checked="" type="checkbox"/>	#	Accident_Severity
<input checked="" type="checkbox"/>	#	Number_of_Vehi...
<input checked="" type="checkbox"/>	#	Number_of_Cas...
<input checked="" type="checkbox"/>	Calendar	Date
<input checked="" type="checkbox"/>	#	Day_of_Week

☐ Single table  
☒ Wildcard union

Search in  
1-6m-accidents-traffic-flow-over-16-years

☐ Include subfolders

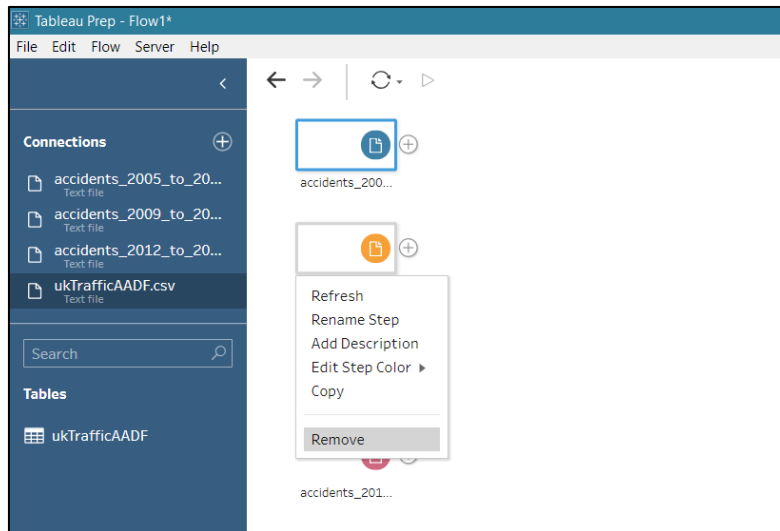
Files  
Include

Matching Pattern (xoc\*)  
accident\*

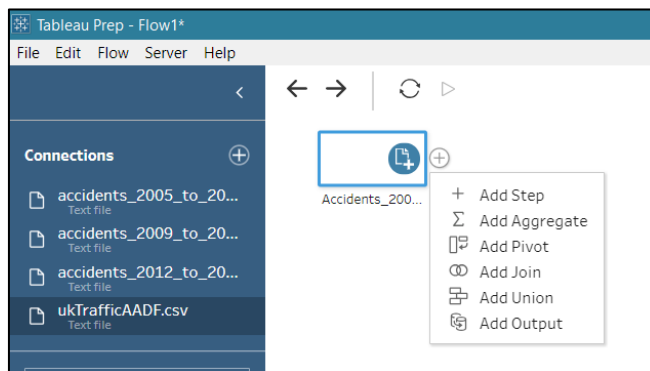
Included files (3)  
accidents\_2005\_to\_2007.csv  
accidents\_2009\_to\_2011.csv  
accidents\_2012\_to\_2014.csv

[Apply](#)

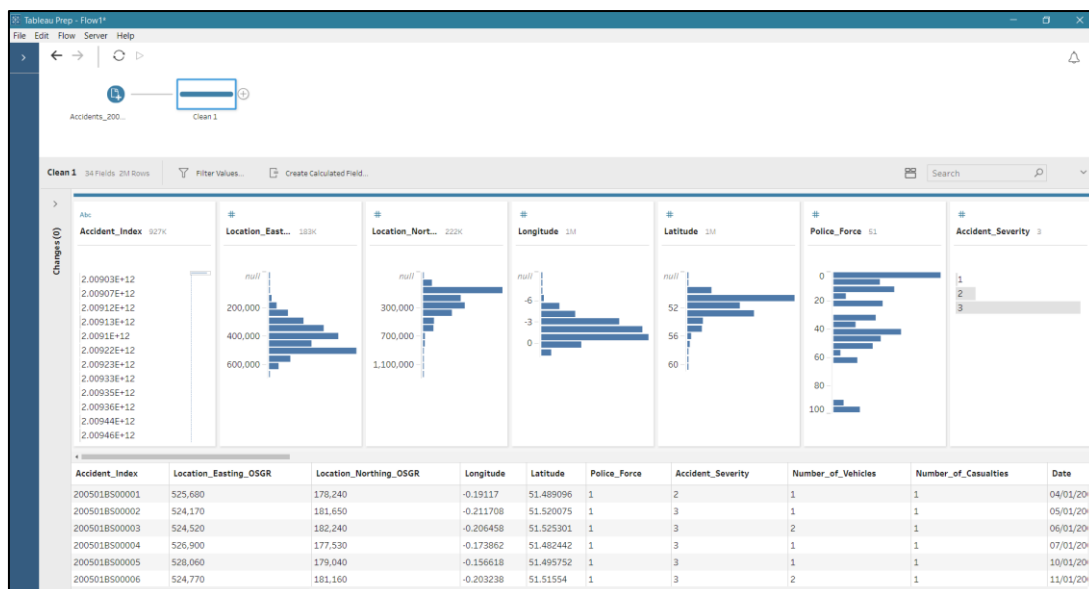
4 Accident inputs can be merge via “multiple files” and “wildcard union”



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

Field Name

Weekday\_Name

DATENAME('day', [Date])

Reference

All

Search

ABS  
ACOS  
AND  
ASCII  
ASIN  
ATAN  
ATAN2  
CASE  
CEILING  
CHAD

DATENAME(date\_part, date, [start\_of\_week])

Returns a part of the given date as a string, where the part is defined by date\_part. If start\_of\_week is omitted, the week start day is determined by the start day configured for the data source.

Example: DATENAME('month', #2004-04-15#) = "April"

### 8 Add a new column for the weekday name calculation

Weekday_Name	31
27	
28	
29	
3	
30	
31	
4	
5	
6	
7	
8	
9	

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

Clean 1

35 Fields 2M Rows

Filter Values...

Changes (2)

Change Type  
[Accident\_Index]  
To String type

Calculated Field  
[Weekday\_Name]  
DATENAME('weekday', [Date])

Edit Field

Field Name  
Weekday\_Name

DATENAME('weekday', [Date])

10 Calculation can be easy refreshed

Abc

**Weekday\_Name** 7

Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday

11 Expected result

Time
30/12/1899, 14:10:00
30/12/1899, 13:20:00
30/12/1899, 14:20:00

12 The time column looks stranges

```
04/01/2005, 3, 17:42, 12, E0
, 05/01/2005, 4, 17:36, 12, E
, 06/01/2005, 5, 00:15, 12, E
, 07/01/2005, 6, 10:35, 12, E
, 10/01/2005, 2, 21:13, 12, E
```

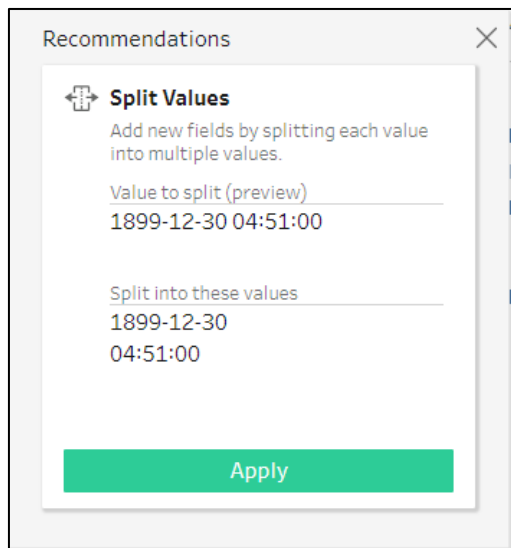
13 A look into the csv shows, that there should be only the time, not the date. But Tableau knows only "date + time"

Abc

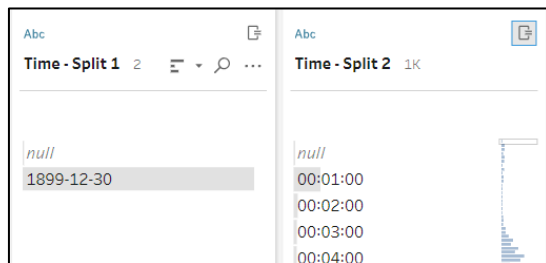
**Time** 1K

null
1899-12-30 00:01:00
1899-12-30 00:02:00
1899-12-30 00:03:00
1899-12-30 00:04:00
1899-12-30 00:05:00
1899-12-30 00:06:00
1899-12-30 00:07:00
1899-12-30 00:08:00

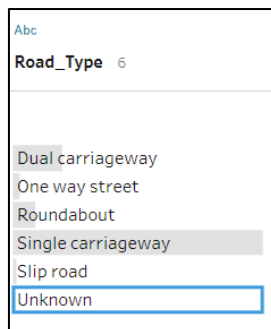
14 First step to clean the column: change to string



15 Tableau recommends a split into two columns

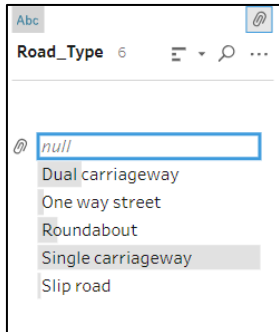


16 The result are two new columns- the left one and the old one can be removed



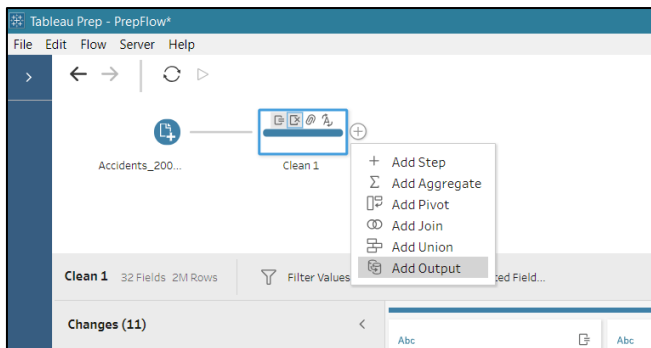
17 "Roadtype" also uses an "unknown" – let's make it empty



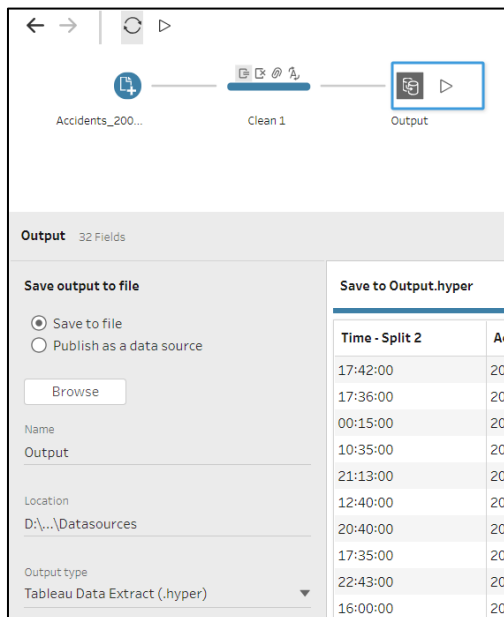


18 Done by right click -> “replace with null”

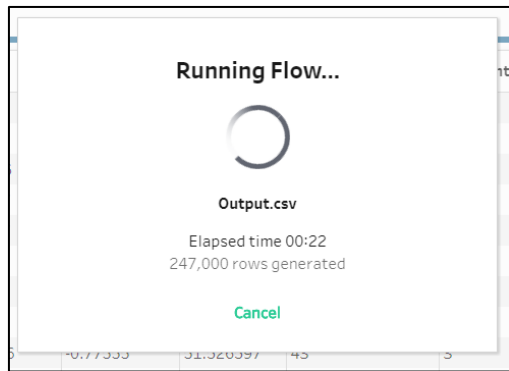
Remove unnecessary columns likes filepath, location\_easting/northing (we have also latitude and longitude)



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

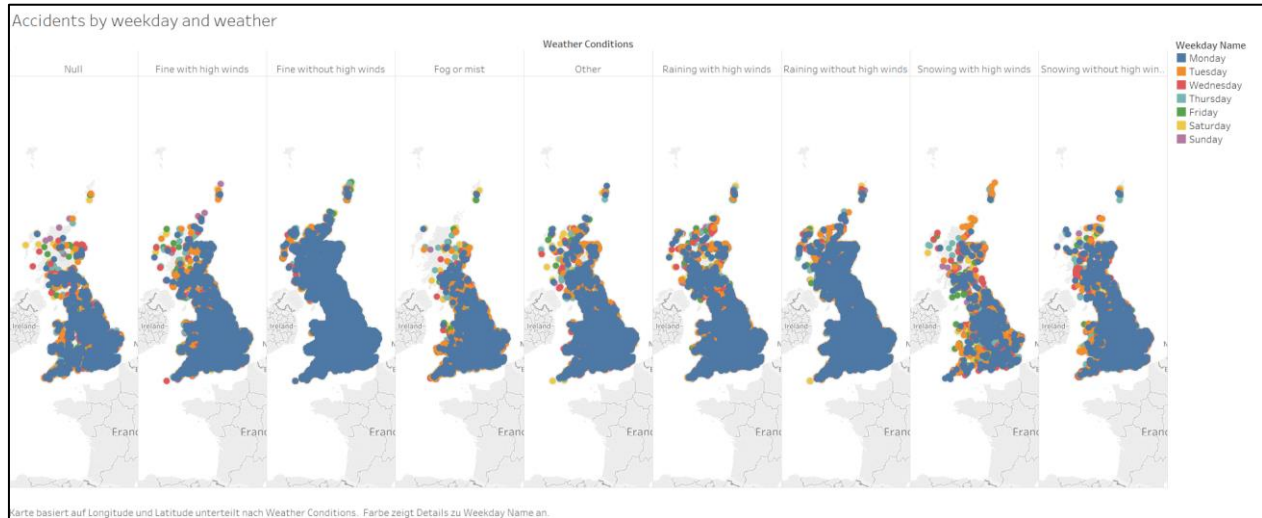
22 Final csv has 1.504.151 rows

The output will be created and the files saved.

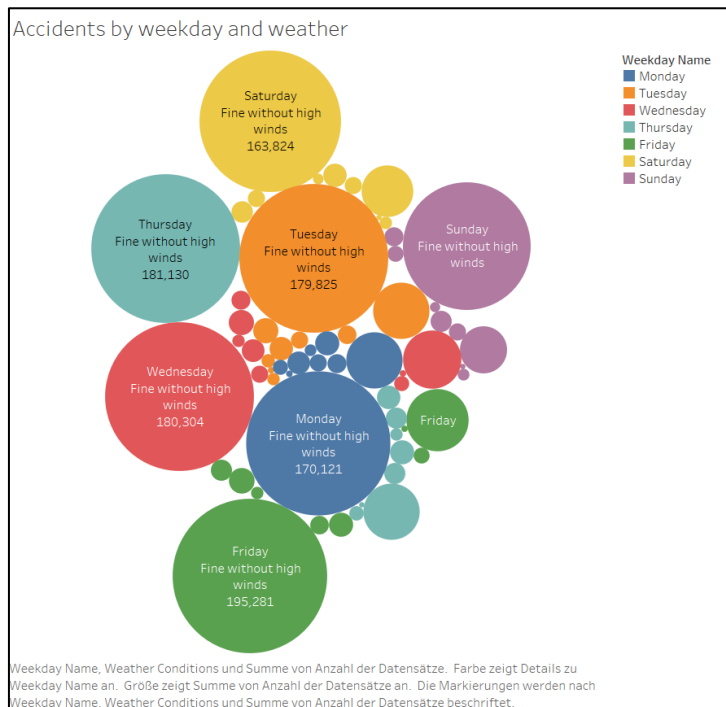
## GOING FORWARD WITH THE FLOW

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