# Data cleaning for ukVenues dataset

ukVenues dataset is extracted using Python into a comma separated file called ukVenues.csv. First the data is manually observed, and the data errors classified. Then the data is cleaned using Manual Intervention and Automatic Correction.

## Data Error Classification

**(a) Incomplete/missing values**

**(b) Corrupted values**

**(c) Out of range values**

**(d) Wrong data**

**(e) Duplicate data**

**(f) Other errors – comma in the field treated by the database as a delimiter.**

## Cleaning stages

**(1) Using Microsoft Excel remove all double quotes (“) and commas. See examples in figures 1, 2, 3, and 4 below**

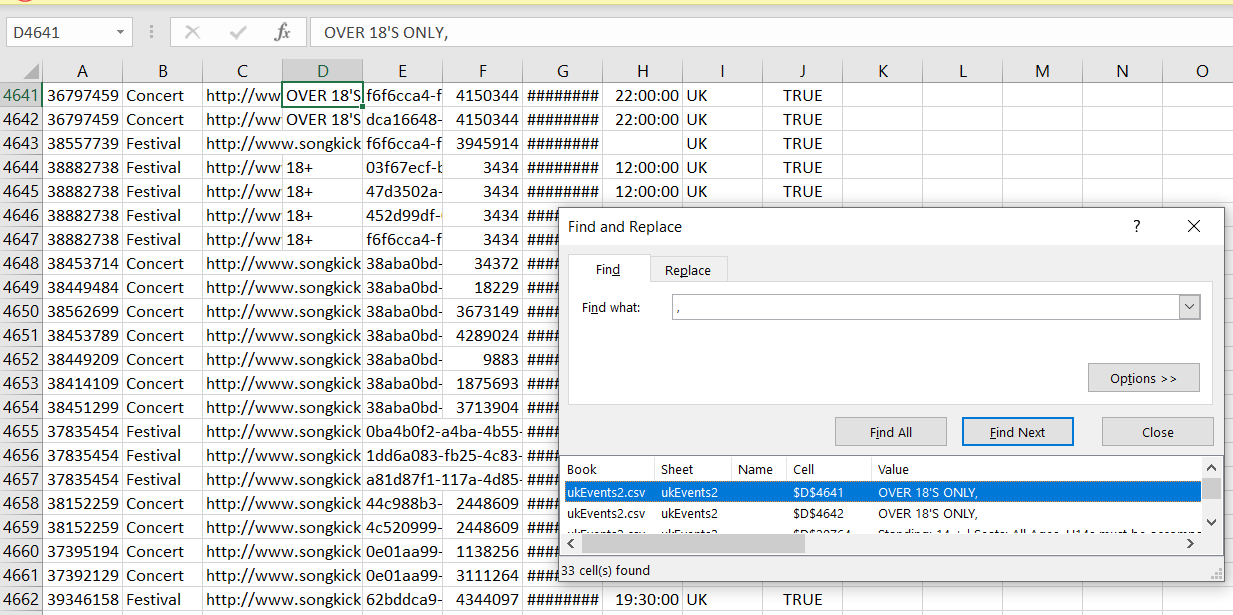
**(2) Using Notepad for each single record that occupies two lines, adjust it to occupy a single line. See an example in figure 5 below.**

**(3) Using Python Pandas Dataframe missing data handling, automate the replacement of missing data in each field.**

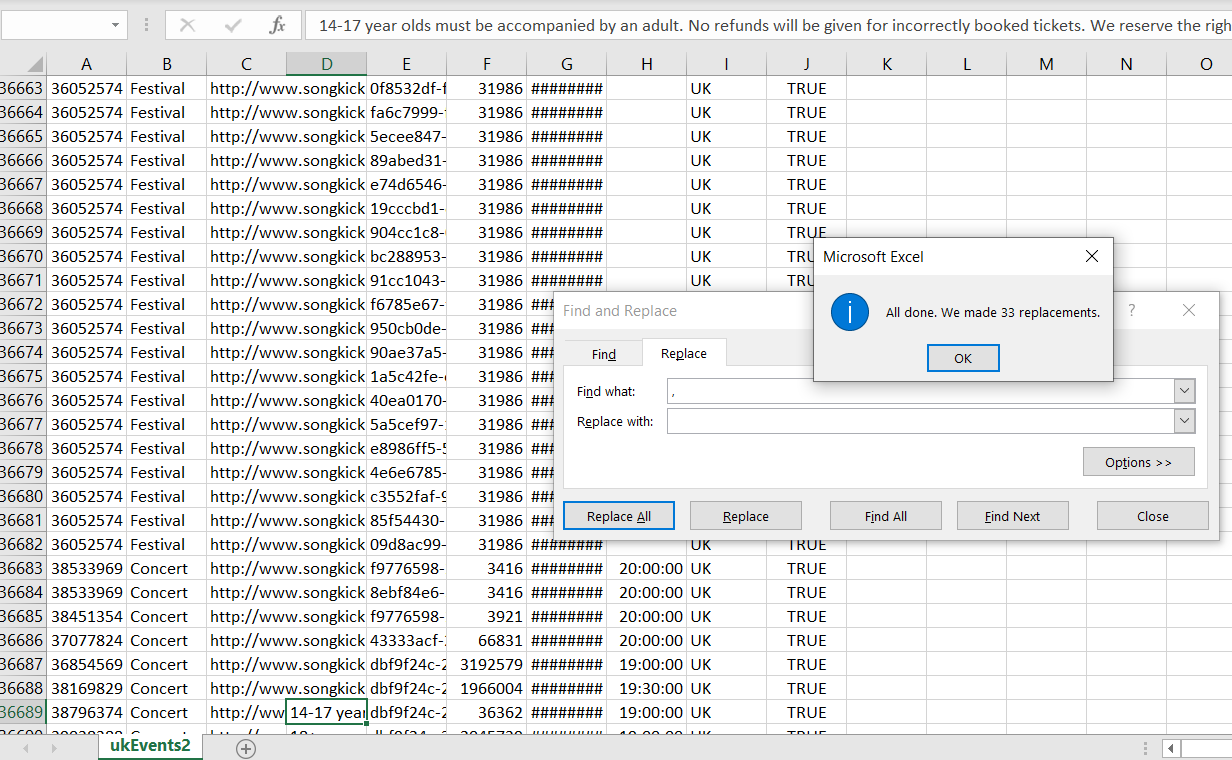
## Data quality analysis

|  |  |  |  |
| --- | --- | --- | --- |
| **S/No.** | **Field** | **Data Error** | **Cleaning stage (Remedy)** |
| 1 | venueID | a, d, f | 2 |
| 2 | venueName | a | 1, 3 |
| 3 | venueStreet | a, d | 1, 2, 3 |
| 4 | venuePostcode | d, f | 2, 3 |
| 5 | venueCity | d, f | 2, 3 |
| 6 | venueCountry | d, f | 2, 3 |
| 7 | venueCapacity | a, d, f | 2, 3 |
| 8 | venueWebsite | a, d, f | 2, 3 |
| 9 | venuePhone | a, d, f | 2, 3 |

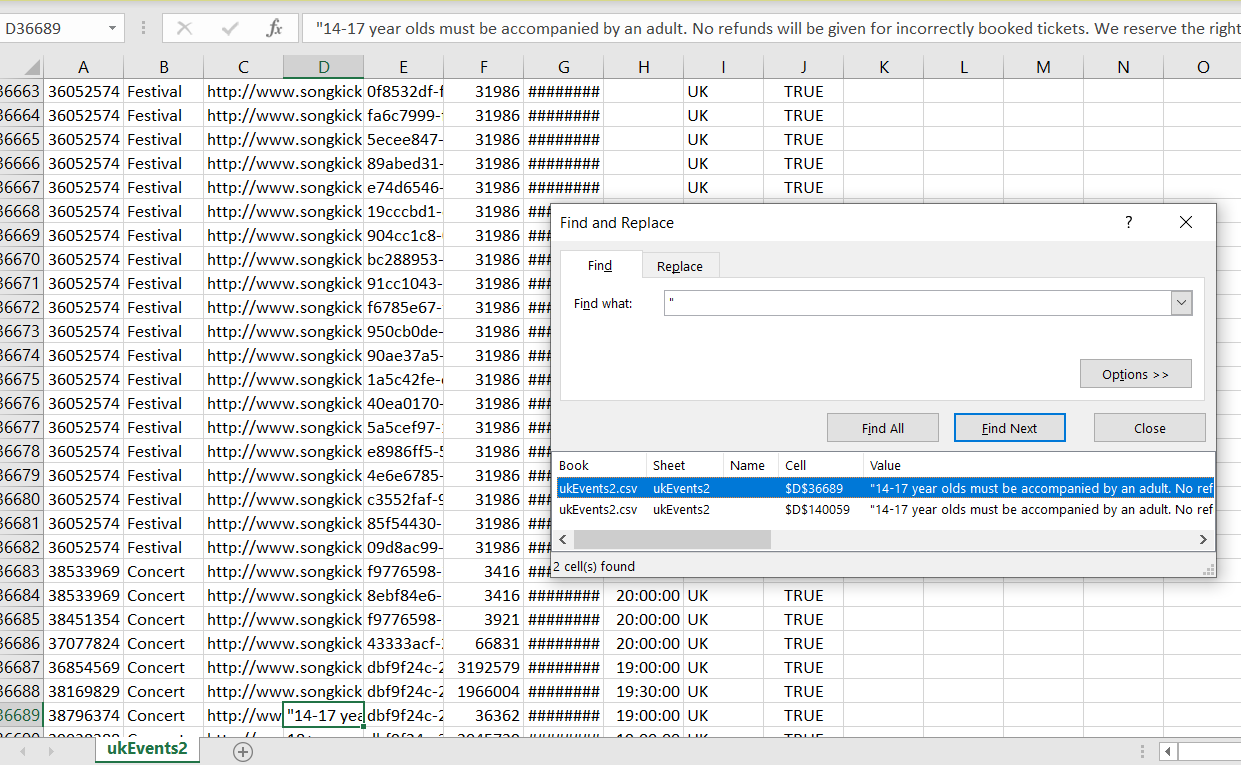
*Figure 1: Microsoft Excel output showing eventAgeRestriction record with a comma and showing 33 cells with commas.*



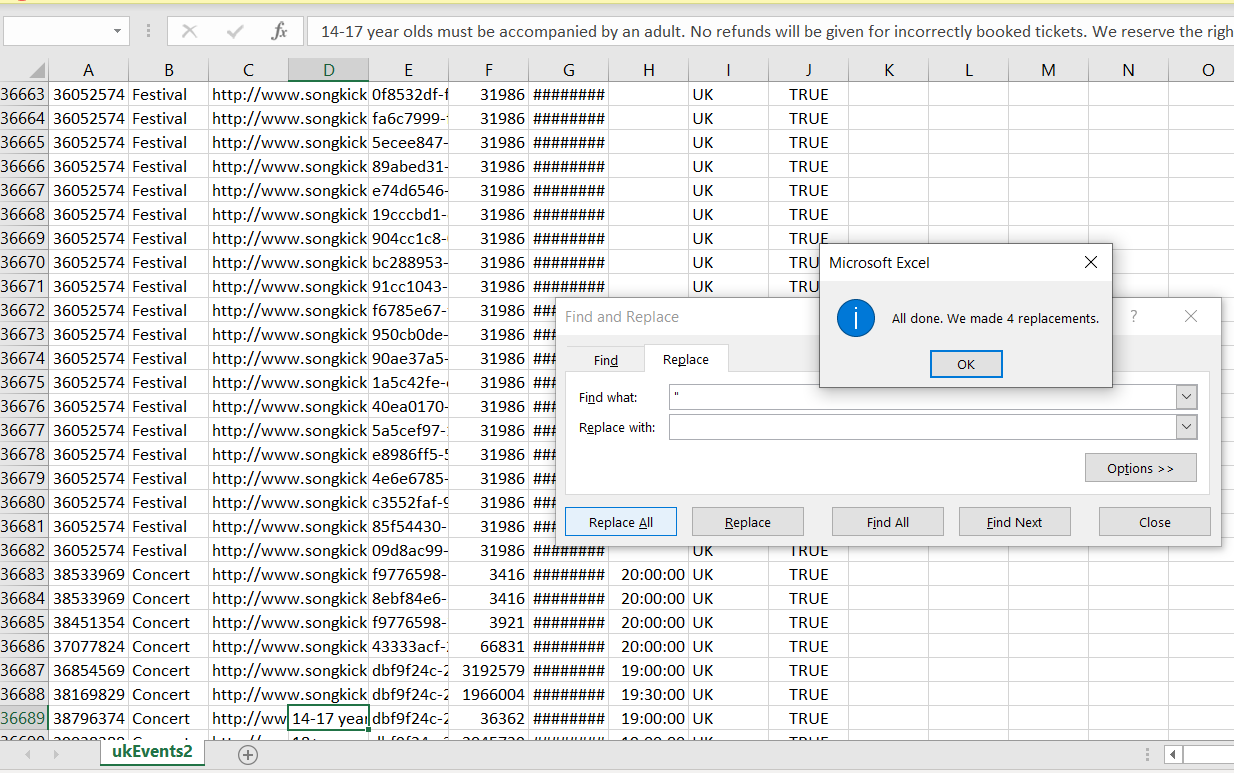
*Figure 2: Microsoft Excel output showing replacement of 33 cells with commas.*



*Figure 3: Microsoft Excel output showing eventAgeRestriction record with a double quotation mark (“) and showing 2 cells with double quotation marks.*



*Figure 4: Microsoft Excel output showing replacement of 4 cells with double quotation marks (“)*



*Figure 5: Notpad output showing a record occupying two lines instead of one*

