# Data cleaning for ukEvents dataset

ukEvents dataset is extracted using Python into a comma separated file called ukEvents.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

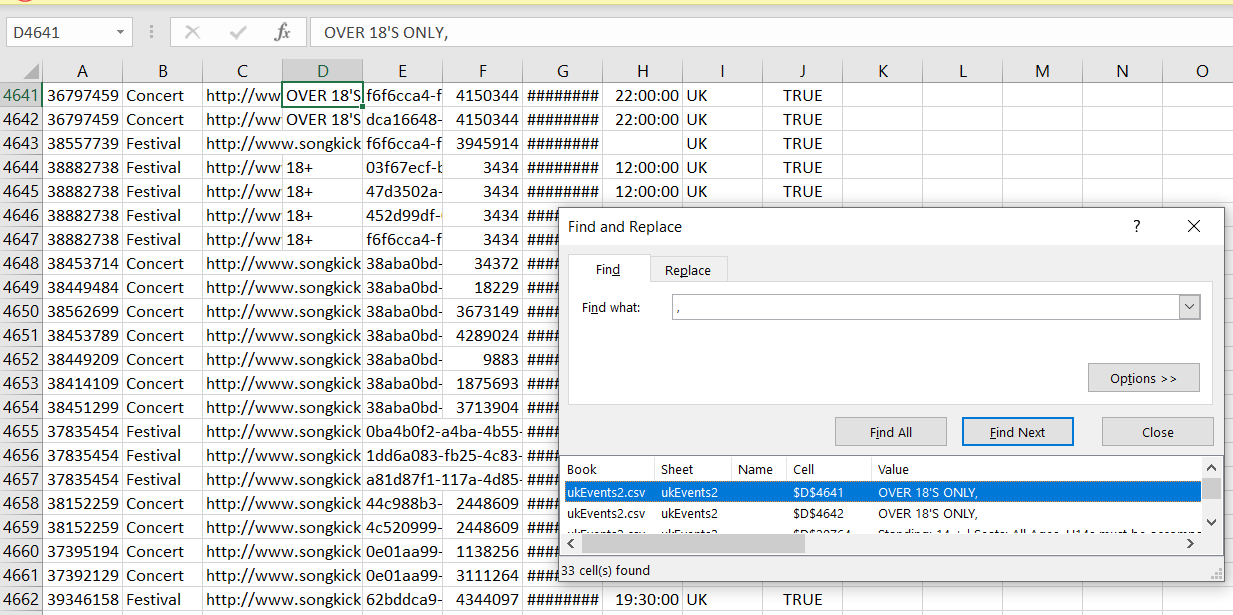
**(1) Using Microsoft Excel remove all double quotation marks (“) and commas as shown in figures 1, 2, 3, and 4 below.**

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

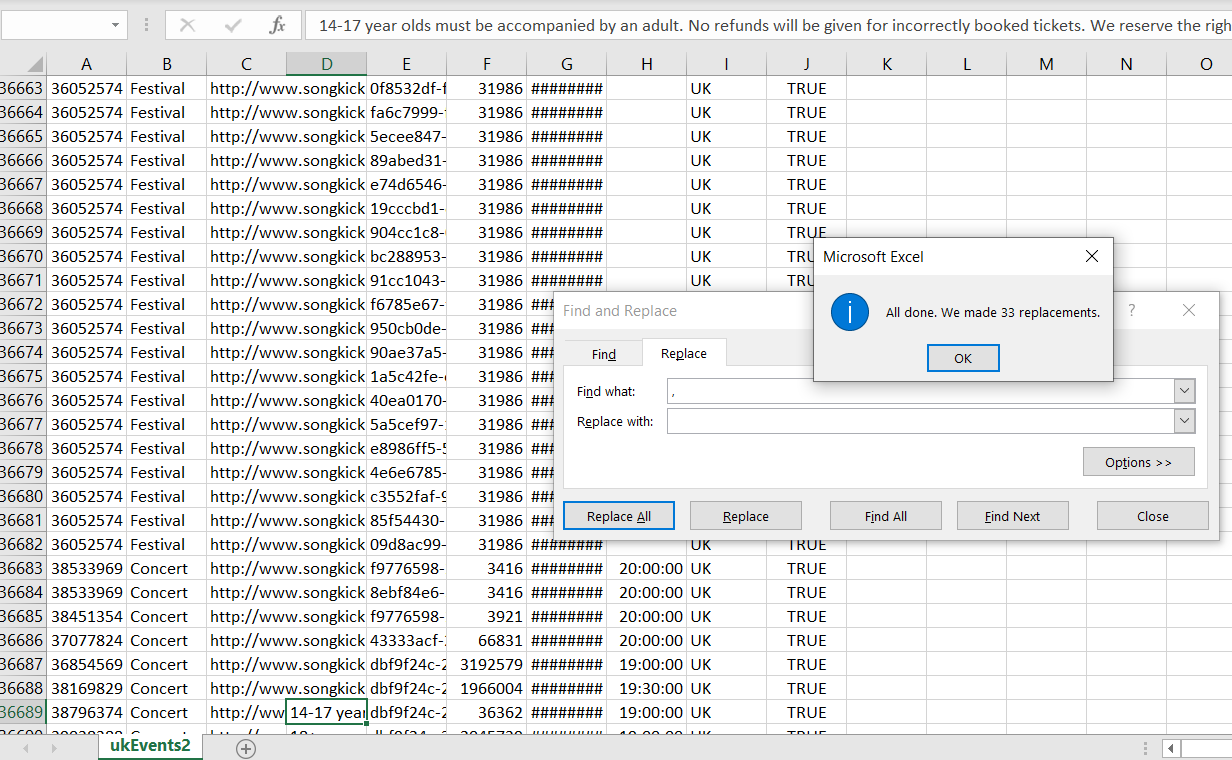
## Data quality analysis

|  |  |  |  |
| --- | --- | --- | --- |
| **S/No.** | **Field** | **Data Error** | **Cleaning stage (Remedy)** |
| 1 | eventID |  |  |
| 2 | eventType |  |  |
| 3 | eventURI |  |  |
| 4 | eventAgeRestriction | a, d, f | 1, 2 |
| 5 | artistMBID |  |  |
| 6 | venueID | a | 2 |
| 7 | eventStartDate |  |  |
| 8 | eventStartTime | a | 2 |
| 9 | venueCountry |  |  |
| 10 | eventFlaggedAsEnded |  |  |

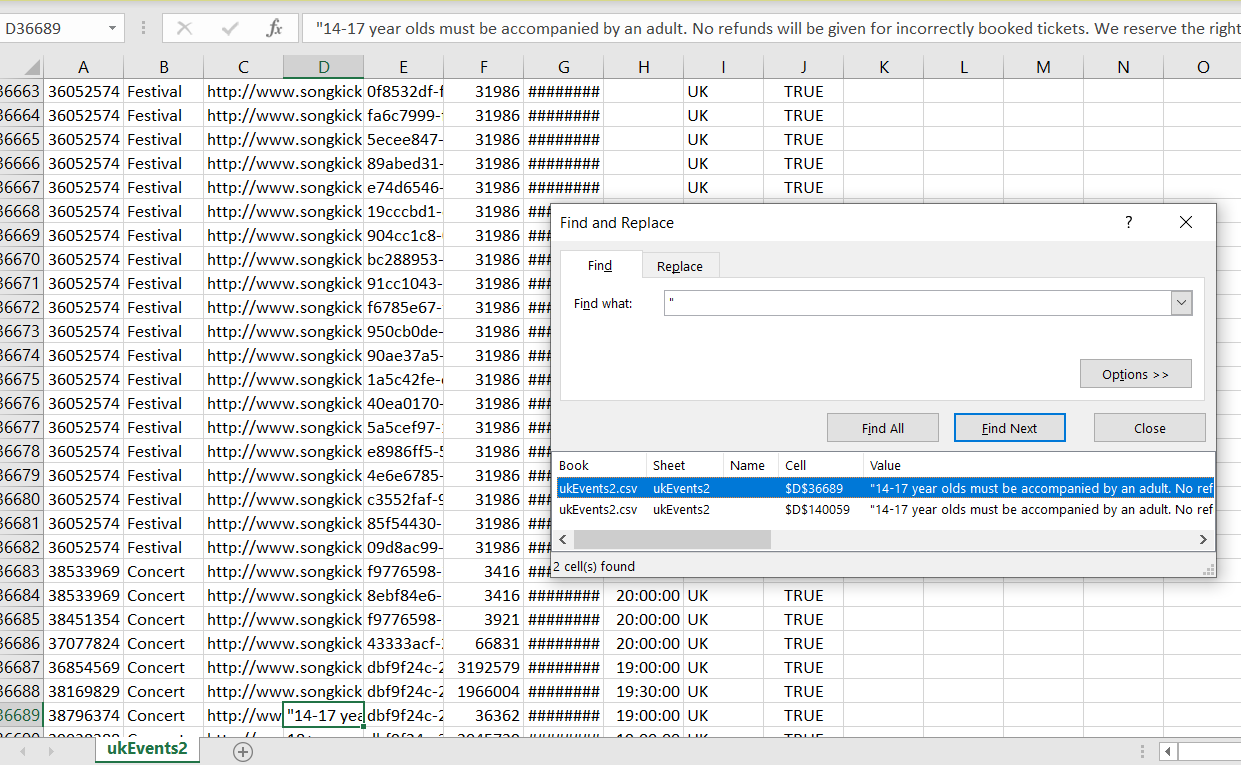
*Figure 1: Microsoft Excel output showing eventAgeRestriction record with a comma and showing 33 cells with commas.*



*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 (“)*

