

|  |  |
| --- | --- |
| Data analytics bootcamp  ETL PROJECT PRPOSAL |  |
| *Team members:*   * *Chadi* * *Hamza* * *Minnie* * *Ruchi* |  |

|  |  |
| --- | --- |
| *Proposal* Our two data sets are from Maven Analytics and Kaggle and are concerning Airbnb listings across the world.  We need to clean and filter the datasets so that they only have listings on New York. To do this we will be using Python and the panda’s library to clean and filter the data on the two csv files.  We will load the data onto three SQL tables corresponding to each dataset using SQL Alchemy into PostgreSQL as a final production database.  The reason we decided to use SQL in this case is because our data is in the form of structured datasets and is strictly text.  This means that there would be no advantage of using a NoSQL database such as mongo DB as they excel in storing a multitude of different data types such as text, image audio etc.  To summarise our findings, we removed unnecessary columns, removed all entries outside of New York, made sure the neighbourhoods matched before renaming the columns and loading the data into PostgreSQL. |  |

Graphical user interface

Description automatically generated