

Wrangle report

I collected data from several sources (CSV , API , and json)

I have used Panda to read these various files

I have checked these files to see if there is a deficiency or through in this data in two different ways. The first is to use excel to know the structure of these files and what they contain and the other using the methods provided by a panda

After the validation process I found some problems like timestamp colum
It is assumed to be dateTime instead of string

"None" >should be> Nan

Because in this way a panda does not recognize the presence of empty cells it is count "None" cell as cells has content

Some dog names contain the value "a", "None

There were about 10 quality problems and two related to structuring

For the rest of the issues I have included them in jupyter notebook code file

After the inspection process, I cleaned these problems one by one in three steps, the first is defining the problem, the second is the code used to solve this problem, and the third is testing the validity of the code

I merged some tables with each other and reconfigured some tables using
pd.melt
pd.merge

In order to facilitate the analysis process ,
replace NaN by the mean of the column to can convert this column type into integer

I stored the clean file to do the processing and ask some questions

?What are the 10 most famous dog types

?what is the most favourit type of dogs

? What is the averge rating for African_hunting_dog

create a summery for a pug dog type

During the process of asking questions, I revealed some minor errors, such as that there are names for different types of dogs

desktop_computer , microphone , coffee_mug , pencil in image_predictions table -