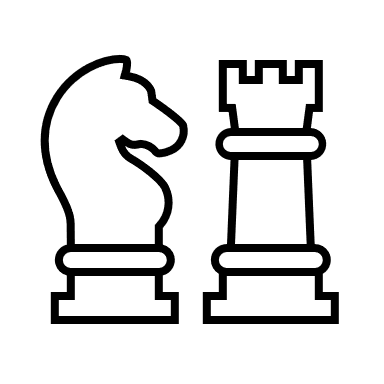
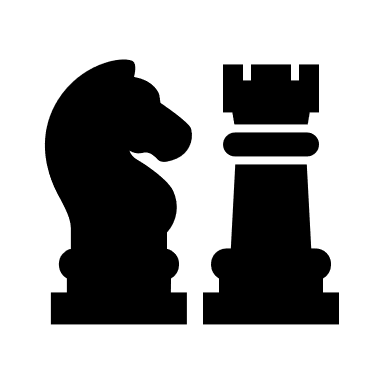
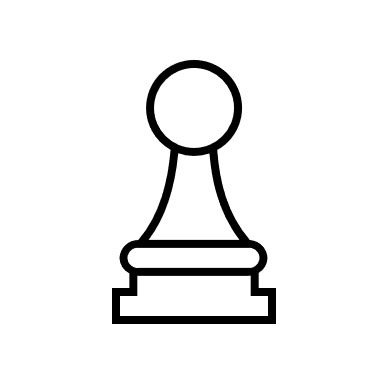
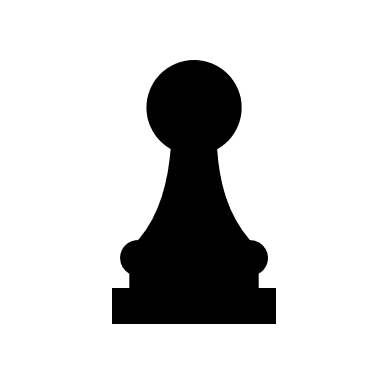
**A dog and a cat looking at a keyboard

Description automatically generated with low confidenceA picture containing cat, floor, indoor, shoes

Description automatically generatedMembers: Erin Zheng, Viera Garcia, Tina Huang, Ting Jennings, Ricardo Ramon**

**ETL Project**

**Plan**

Our goal for this project was to extract, transform and load a dataset. We wanted to cover a topic that we were all interested in. Fortunately, we were able to come to an agreement on chess.  
  
FIDE (International Chess Federation) is the governing body of the sport of chess. They regulate all international chess competitions. We found a Kaggle data set with all the players and their rankings from 2015-2020.

The following table illustrates our data issues and the actions taken to clean the data set.

| **Step** | **Data** | **Action** |
| --- | --- | --- |
|  | We found a Kaggle data set which shows all players in the FIDE organization and their ratings for each year. | First, we had to merge all our data together. We picked out the largest data sets we could find which were years 2019-2020. We started by merging data based on fide\_id which is a unique identifier for each player. |
|  | The data set was out of order. | We had to sort values by the rating and the name. |
|  | After merging, we realized our data set had over 1 millions rows, many being duplicates | Because FIDE has a tournament every month, there our multiple duplicates of each player (up to 12). Our solution was to get the highest rating of each player then drop the rest. |
|  | There were a lot of extra rows we did not need. | We dropped all the rows except fide\_id, name, YoB, and the rating standard. |
|  | The data set gave the year of birth, but no age. | Since only the year of birth was listed, we had to set up a formula on jupyter notebooks where we would subtract the YoB from the year of the data set. |
|  | Cleaned data, uncleaned labels | Once we had the cleaned data, we had to create bins to place the values and labels for each of the bins. |
|  | Cleaned finished data | Exported our finalized data into a csv file |

**Extraction**

We used 3 different CSVs from a Kaggle data set that was published by FIDE. The three data sets we found were:

* Players.csv
  + This csv has the players information including their name, fide\_id, federation, chess title, gender, and YoB.
* Rating\_2019.csv
  + This csv displays fide\_id, year, month, rating standard, rating rapid, and rating blitz for 2019.
* Rating\_2020.csv
  + This csv displays fide\_id, year, month, rating standard, rating rapid, and rating blitz for 2020.

The fields of interest include the following:

* Name
* Fide Id
* Rating Standard
* Federation
* Year of Birth

The source of the dataset:

<https://www.kaggle.com/rohanrao/chess-fide-ratings>

**Transformation**

To transform the data, we had to do the following:

* Used Pandas functions in jupyter notebook to load all three CSV files.
* Merged players.csv to both ratings\_2019 and ratings\_2020.
* Reviewed the files and transformed into data frames.
* Removed the unnecessary columns.
* Identified and removed duplicates by pulling the highest rating for each player for the particular year.
* Created bins for each of the values pulled and labeled the bins based on age.

Text

Description automatically generated with medium confidence

Table

Description automatically generated

Graphical user interface, text

Description automatically generated

**Load**

After exporting and cleaning all the csvs, we began running the queries and created the new tables with only the relevant information that we reconnected to the database and generated the tables for the data frames.

**Table

Description automatically generated2019:**

**Application, table

Description automatically generated2020:**

**Summary**

We had some limitations in our findings. We were originally planning of finding the average player rating in each country, but with all 195 countries included in FIDE, many countries had less than 10 players total. The data was inconsistent and just not good information to use.

**Chart, bar chart

Description automatically generated**We took additional time to plot out the results, so the data is more visually appealing.

Chart, bar chart

Description automatically generatedWe found that majority of players were under the age of 20, yet there were still over 20,000 people over the age of 70 in both 2019 and 2020.

**Chart, bar chart

Description automatically generated**Here is the average rating for each age groups:

Chart, bar chart

Description automatically generatedAs expected, the ratings increase with age. There is a large jump from the age range under 20 to 30-40 but the rating evens out between 30-70. There is also a drop after the age of 70. These results remain constant in both years.