Extract:

We got our data from Kaggle from the link below:

https://www.kaggle.com/martinellis/undefined

The link contained 9 datasets formatted in CSV, describing statistics of the players and teams and we chose one for the players' information and the second for the players' game statistics.

We also looked at this link: https://puckpedia.com/players

We used this link to find out the players' salaries and we scraped the data using the pandas function (pd.read_html(url))

Transform:

We first dropped columns that is not needed in our research to reduce the size of the Dataframe. One of the tables (player_info.csv) have 22 columns so we had to reduce that and end up using just 4

We realized that there were duplicate values in the playerinfo dataset where some players appeared more than once so we dropped the duplicate rows and set the index of the dataframes to the "player_id"

For the data we scrapped from puckpedia.com, we only wanted the names of the players and their salaries. The player names were originally displayed as, for example, "Connor McDavid", but we had the players' first and last names in separate columns in our other two dataframes so we needed to use the split function and created a separate dataframe called "puckpedia_playersdf". The "puckpedia_playersdf" dataframe then has the first name and last names in separate columns, as well as a column named "current salary" which was taken from the original table we scrapped from puckpedia.com

Load:

We loaded the dataframes from jupyter notebook to pgAdmin