Ashkan Dehghani Zahedani and Steven Ortiz

ETL Project

UCI Bootcamp

**Introduction:**

We are using NBA players data to find if the performance of NBA players determines their salaries. Our goal is to create a database that has three tables. One table will illustrate each players statistic. The other table shows each players biography (age, position, etc). The last table shows each player’s salary. Using this database one can conclude whether the performance of players has affected their salaries.

**Extract:**

We have used two datasets from Kaggle to gather necessary data to create a database about NBA players in 17-18 season. Both datasets are in CSV format. However, it needs some transformation which we’ll talk about it in the next section.

The sources of data that we have extracted from:

NBA Players Stats 17-18:

<https://www.kaggle.com/mcamli/nba17-18>

A close up of a map

Description automatically generated

The above dataset includes two CSV files.

NBA Player Salary 17-18:

<https://www.kaggle.com/koki25ando/salary>

A screenshot of a social media post with text and people in background

Description automatically generated

**Transformation:**

1. The first step is to import our datasets to python:

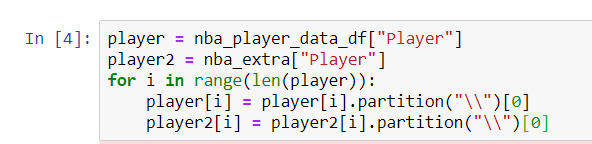
A screenshot of a computer

Description automatically generated

We have imported the other dataset as well.

1. As it can be seen the player column has to be cleaned.

We have written the following code to remove strings after “//”



1. The next step is to identify the columns that we need.

A screenshot of a cell phone

Description automatically generated

1. Next, we are going to merge data frames above to create a data set called “nba\_stat”
2. We have removed players who played less than 40 minutes during the season.
3. We have also renamed columns to lowercase letters. (to be matched with pgadmin)
4. We have indexed player columns.

A screenshot of a social media post

Description automatically generated

1. We have done the same to create a data frame called “nba\_bio”

A screenshot of a social media post

Description automatically generated

1. We have done steps 4 to 7 to the nba salary dataset too.

A screenshot of a social media post

Description automatically generated

**Load:**

We have created a database in “pgadmin” called “nba\_db”. We have created three tables with proper column names. The tables are called “nba\_stat”, “nba\_bio” and “nba\_salary”. Now we need to load results from python to pgadmin. We have created a connection to database and confirmed the connection by reading the table names.

A screenshot of a social media post

Description automatically generated

Finally, we have loaded our results to tables in pgadmin.

**A screenshot of a cell phone

Description automatically generated**

The following is the screen shot of three tables that was created and filled.

Table 1:

A screenshot of a cell phone

Description automatically generated

Table 2:

A screenshot of a cell phone

Description automatically generated

Table 3:

A screenshot of a cell phone

Description automatically generated

**Conclusion**

For this project, we have extracted data from two separate Kaggle datasets; The NBA players stats and their respective salary. For the transformation, we used pandas and cleaned up all the data to find certain columns and then rename them. We have also removed players who have played less than 400 minutes during the season. Lastly, to load, we have connected and loaded the data to SQL pgAdmin and create tables on pgAdmin demonstrating the two datasets into three different data frames. Using this database, one can analyze the relation between players performances to their salaries.