|  |
| --- |
|  |
| UWA Data Analytics Boot Camp | |
| ETL Project Report  Ernest Bondi and Alysha Snowden | |
| Document No.: 0000-GP-UWA-00002  Security Classification: Unrestricted | |
|  | |
|  | |
|  | |

**TABLE OF CONTENTS**

[1 project proposal 3](#_Toc74256209)

[2 Extract 4](#_Toc74256210)

[3 Transform 5](#_Toc74256211)

[4 Load 6](#_Toc74256212)

# project proposal

​This Project aims to Utilise EPL Player stats to build an EPL Dream team. By gathering multiple dataset statistics, the objective is to analyse, dissect and extract the important data to be readily available for useful application.

This required the creation of a database to store the EPL Soccer Player stats data. The end product data view will be based on each individual field position (Goalkeeper, Defender, Midfield and Forward). To complete this objective, We utilised our ETL skills by extracting data from multiple sources, transforming that data using Pandas and SQLAlchemy, and loading the data into Postgres (our Database).

Due to the project being completed within a short time period, one week. Team effort and Team work proved to be imperative to our success.

# Extract

***E****xtract: your original data sources and how the data was formatted (CSV, JSON, pgAdmin 4, etc).*

For this first step we retrieved structured and unstructured data from Two sources, Kaggle and fbref (sports website). The data sources retrieved were:

* Kaggle:(Kaggle|https://www.kaggle.com/krishanthbarkav/english-premier-leagueepl-player-statistics)
  + pl\_15-16.csv
  + pl\_16-17.csv
  + pl\_17-18.csv
  + pl\_18-19.csv
* fbref: (https://fbref.com/en/comps/9/stats/Premier-League-Stats)
  + Player\_complete\_nationality.csv
  + Player\_Stats.csv

The first 4 data sources were retrieved straight as csv files, whereas the Player stats were retrieved utilising panda request.

Table

Description automatically generated

Graphical user interface, text, application

Description automatically generated

# Transform

**T**ransform: what data cleaning or transformation was required.

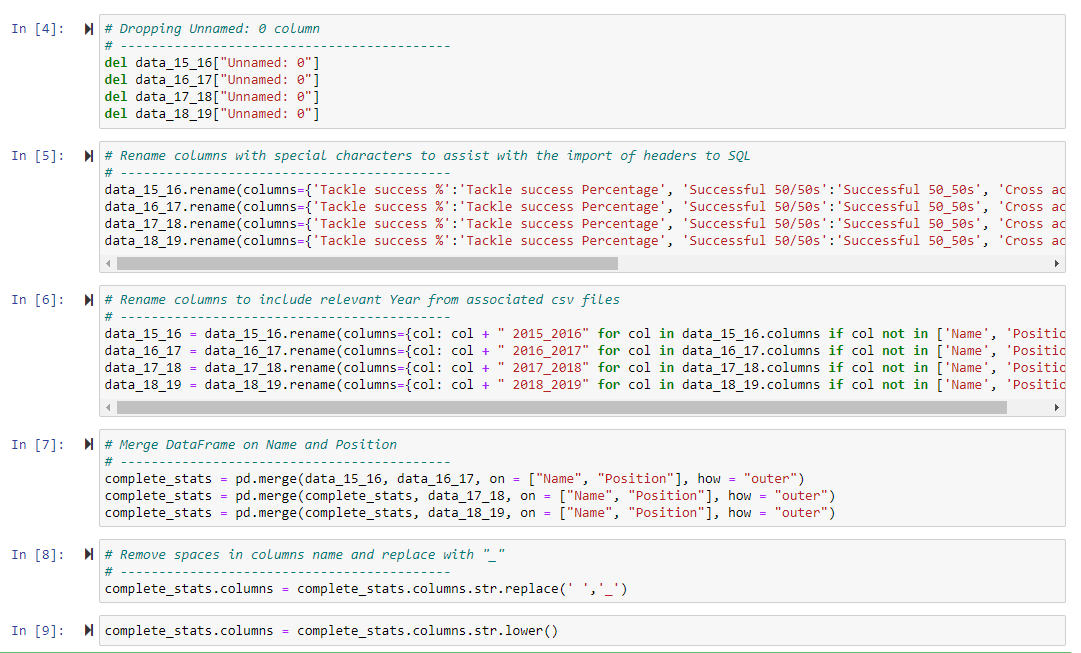
To transform our data we utilised two programs Jupyter notebook and SQl admin.

Within Jupyter notebook (Pandas) we cleaned the data:

* Cleaned
  + We replaced all NaN value within the csv files with zero (0). Then went to removed special characters from the player names and stripped all commas with spaces.
* Joining
  + We joined the Four separate EPL year stats for the seasons into one data frame
  + We merged the player nationality dataset from the multiple API pulls we called from the website into one data frame
* Aggregate
  + We created an Averages column on multiple data frames and totalled each players average across the Four seasons (each csv file)
  + Created a data frame utilising the “Grouby” function for each position (Goalkeeper, Defender, Midfield and Forward)
    - With the columns being the Averaged values of each attribute (i.e. Goals scored per game).

Whereas within SQL admin:

* We created the SQL database



# Load

***L****oad: the final database, tables/collections, and why this was chosen.*

We utilised the SQL Alchemy on Jupyter Notebook to import and load the multiple data frames created.

