**Final Report: ETL Project**

Project Members: Max Patterson, Anastasios Tomazos, Erin Wills

Abstract

The objective of this project is to initially *extract* National Football League (“NFL”) data from at least two different sources, then *transform* such data into a comprehensible and usable format, followed by *loading* the data into a central database. Although there seems to be numerous sources for NFL data, the dearth of free and publicly accessible data sources is a major stumbling block for the interested fanbase, like us, in performing significant football analytics. Be that as it may, we did find NFL data from two different sources with which we created useful databases comprised of player and gameday data for the 2009 – 2013 seasons.

Breakdown of Tasks

The overall ETL process is comprised of *Extract*, *Transform*, and *Load* subprocesses.

*Extract*

Our data sources are extracted from the **nflscrapR-data** repository and the **NFLsavant.com** website. **nflscrapR-data** was created by a group of [Carnegie Mellon University statistical researchers](http://www.stat.cmu.edu/) who publicly released their data**,** which is an R package that “uses an API maintained by the NFL to scrape, clean, parse, and output clean datasets at the individual play, player, game, and season levels.”[[1]](#footnote-1) NFLsavant.com is a website that obtains publicly-available NFL play-by-play data on the internet.

Our plan is to create the following two databases, which includes:

* Player data from the **nflscrapR-data** and append to this dataset the NFLsavant.com player information, which includes the college of each player
* Game data from **nflscrapR-data** and append to this dataset the NFLsavant.com weather data for each game

The data we are interested in are formatted in CSV files. We use 2009 to 2013 data because this is the time period both data sources overlap. In accordance with our plan, we intend to use the following CSV files:

* **nflscrapR-data:** 
  + NFL gameday information for 2009, 2010, 2011, 2012 and 2013 (5 CSV files)
  + NFL player rosters for 2009, 2010, 2011, 2012 and 2013 (5 CSV files)
* **NFLsavant.com**:
  + All NFL Weather from 1960 to 2013 (1 CSV file)
  + NFL player information from 2009, 2010, 2011, 2012 and 2013 (5 CSV file)

*Transform*

What data cleaning or transformation was required?

*Load*

The final database, tables/collections, and why this was chosen?

1. See, https://www.kaggle.com/maxhorowitz/nflplaybyplay2009to2016 [↑](#footnote-ref-1)