ETL Project

Extract:

* Data for this project was extracted from Kaggle, data.world and StatCrunch
* Using 1 csv (arrest\_data) and 1 json file (player\_profile)

Transform:

CSV (arrest\_data)

* + Created new data with selected columns
  + Filtered dataframe from specificed columns
  + Renamed column headers
    - NAME : player\_name
    - DATE: date\_of\_arrest
    - CATEGORY: category
    - DESCRIPTION: description
  + Dropped all NuN
  + Set index to date\_of\_arrest

JSON (player\_profile)

* Stored json data into DataFrame
* Created a filtered dataframe from specific columns
* Renamed column headers
  + Player\_id: id
  + name: player\_name
  + current\_team: team
* Dropped all Nun
* Set index to id

Joined 2 Datasets (arrest\_data + player\_profile) Into 1 Dataset (combined\_ds)

* Merged both datasets using player\_name column

Load:

* Database: Postgres (Relational): We decided to use this since we converted all files into datasets and created player\_name as the relationship
* Within jupyter connected to local database (postgres)
* Checked if there were existing tables
* Created 3 Tables:
  1. arrest\_table
  2. player\_profile
  3. combine\_table
* Transfer data into each table
* Confirm data was added by querying the tables (Select \* From TableName) in jupyter

Analysis From 2000-2017 NFL Players Arrest Records

* Discovered that Cincinnati Bangels team had the most arrest from the years 2000- 2017
* 5 teams that had the least amount of arrest were Bears, Panthers, Buccaneers, Texans and Seahawks
* The top 3 players with the most arrest (in order): Adam Jones, Kenny Britt, Brandon Marshall
* The top 5 arrest charges (in order): DUI, Drugs, Assault, Domestic Violence, License

(See Graphs Next Page)

Graphs







