*Project Name*: ETL Technical Report

*Team 3*: Braxton Van Cleave, Brittany Johnson, Josh Hopkins, Collin Alston

*Date*: (6/1/21)

*Context*: Team #3 decided to explore historical sports betting data grouped by sport, in order to practice the ETL steps. As a group, we extracted several CSV files online and created a shared “main” python file to run our data frames we each created separately.

*STEP 1 Extract*: We first pulled CSV’s from online using a for loop and attaching their respected URL links. Then using the pd.read excel function, we appended each sports data frame for each year (2010-2020).

Text

Description automatically generated with medium confidence

*STEP 2 Individual Transforms*:

Author: Braxton

Transform:

The following was completed in order to transform the raw extracted data:

* Append each year of Odds Data for NHL seasons 2010-2021
* Drop unnecessary Columns as to standardize my table’s column with the rest of the team
* Transform the Date column to become a DateTime data type
* Map Team Abbreviation to reflect Team Location and Name
* Removed outlier where a final score or MoneyLine (ML) was not present

Author: Josh

Transform:

* Identified the nhl records from the past 11 years
* appended the records into a single data frame
* updated the date to datetime
* Copied desired columns
* identify unified ML data
* rename column for standardization throughout db
* Remapped teams to include the team and mascot
* Removed outlier where a final score was not present

Author: Brittany

Transform:

* Identified the nba records from the past 11 years
* Concatenated them into a single data frame
* Altered the dates to match the source between seasons
* Removed various columns to distill data and make it usable across sports
* Remapped team names to be more descriptive and eliminate data entry issues
* Conditional added to the date loop because the season spans two years

Author: Collin

Transform: Our group decided ahead of time on column names and types in order to standardize our data frames. I then cleaned up the data, first by checking unique team name values and then replacing incorrect values with their correct values. There are 32 teams in the NFL + 3 teams that have changed locations over the last 10 years. That means a total of 35 team values. Lastly, I changed the date column values into date type: (‘<M8[ns]’). So instead of reading “908” it was converted to “2010-09-08”.

STEP 3 Load: We connected an engine to interact with PostgreSQL. We then made empty tables/columns in PgAdmin 4. Finally, we filled the data using df.to\_sql to complete the load process.

