README: NHL Prediction Data Files

Overview

This README describes two flat files prepared for the NHL prediction assignment. Both datasets are restricted to **regular season** games and contain only regulation-time outcomes.

File 1: `playergamedata.csv`

Unit of observation: One row per player–team–game combination. **Primary keys:** `gameId`, `playerId`

Table 1: Column dictionary: `playergamedata.csv`

Column	Type	Description / Units
gameId	string	NHL game identifier (e.g., 2011020001).
date	YYYY-MM-DD	Game date.
season	int	Season start year (e.g., 2011 for 2011–12).
gametype	int	NHL game type; all rows are 2 (regular season).
teamname	string	Team abbreviation (e.g., BOS).
teamid	int	Team numeric ID.
home	int (0/1)	1 if player's team was home, else 0.
playerId	string	NHL player ID.
player_name	string	Player name for the game.
position	string	Player position: C, L, R, D, G.
sweaterNumber	int	Jersey number.
goals	int	Player goals in the game.
assists	int	Player assists in the game.
shots	int	Shots on goal.
plusMinus	int	Plus/minus rating.
toi_raw	string	Time on ice as given (e.g., 13:27 or 1:02:15).
toi_seconds	int	Time on ice in seconds, parsed from `toi_raw`.
team_goals_reg	int	Team's regulation goals for this player's team.
home_goals_reg	int	Home team regulation goals.
away_goals_reg	int	Away team regulation goals.
result_reg	int	Regulation result: 0=away win, 1=tie, 2=home win.

File 2: `gamedata.csv`

Unit of observation: One row per game. **Primary key:** `gameId`

Table 2: Column dictionary: `gamedata.csv`

Column	Type	Description / Units
season	int	Season start year (e.g., 2011).
gameId	string	NHL game identifier.
date	YYYY-MM-DD	Game date.
result_reg	int	Regulation result: 0=away win, 1=tie, 2=home win.
team_goals_reg_home	int	Home team regulation goals.
team_goals_reg_away	int	Away team regulation goals.
teamname_home	string	Home team abbreviation.
teamname_away	string	Away team abbreviation.
odd_winhome	numeric	Decimal odds for home win after 60:00 (regulation only).
odd_draw	numeric	Decimal odds for draw after 60:00.
odd_awaywin	numeric	Decimal odds for away win after 60:00.

Notes

- All odds refer to **regulation-time results only**, excluding overtime and shootouts.

- Bookmaker probabilities can be derived as:

$$p_k^* = \frac{1/\text{odds}_k}{\sum_j 1/\text{odds}_j}, \quad k \in \{\text{home, draw, away}\}.$$

- The two files can be linked using the variable `gameId`.