U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

DATA MINING PROJECT

T01G15 - COMPUTATIONAL LEARNING (2023/2024)

JOÃO PEREIRA
UP202007145

NUNO PEREIRA
UP202007865

SOFIA COSTA
UP202300565

AGENDA

01 Business Understanding

02 Domain Description

03 Data Understanding

04 Data Preparation

05 Data Classification

06 Conclusions

BUSINESS UNDERSTANDING

GENERAL OVERVIEW

- Basketball tournaments are split into two parts.
- First, all teams play each other aiming to achieve the greatest number of wins possible.
- Then, at the end of the first part of the season, a pre-determined number of teams that won the most games are qualified for the playoff season.
- In the playoff season, they play a series of knock-out matches for the trophy.

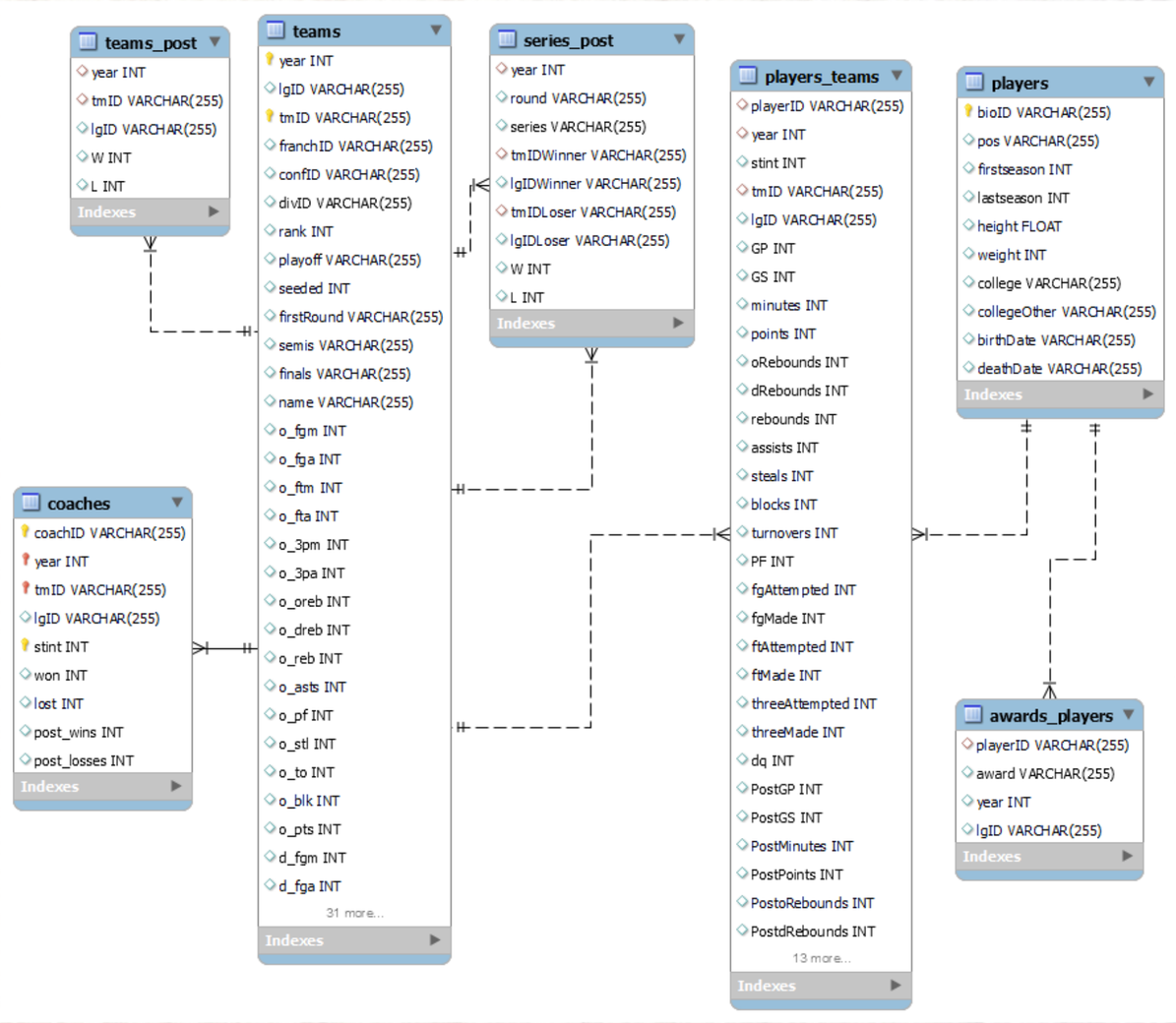
GOAL

The goal is to use 10 years worth of data, from players, teams, coaches, games, and several other metrics to predict which teams will qualify for the playoffs in the next season.

DOMAIN DESCRIPTION

- awards_players (96 objects) - each record describes awards and prizes received by players across 10 seasons,
- coaches (163 objects) - each record describes all coaches who've managed the teams during the time period,
- players (894 objects) - each record contains details of all players,
- players_teams (1877 objects) - each record describes the performance of each player for each team they played,
- series_post (71 objects) - each record describes the series' results,
- teams (143 objects) - each record describes the performance of the teams for each season,
- teams_post (81 objects) - each record describes the results of each team at the post-season.

- There is data for 20 teams, 57 coaches and 893 players.



PROCESS

STEP 1

Data Understanding

- Explore data
- Verify data quality

STEP 2

Data Preparation

- Completeness
- Consistency
- Conformity
- Accuracy
- Integrity
- Timeliness

STEP 3

Modelling

- Select Models
- Generate test design
- Build Model
- Assess Model

STEP 4

Evaluation

- Evaluate results
- Review process

EXPLORATORY DATA ANALYSIS - TEAMS

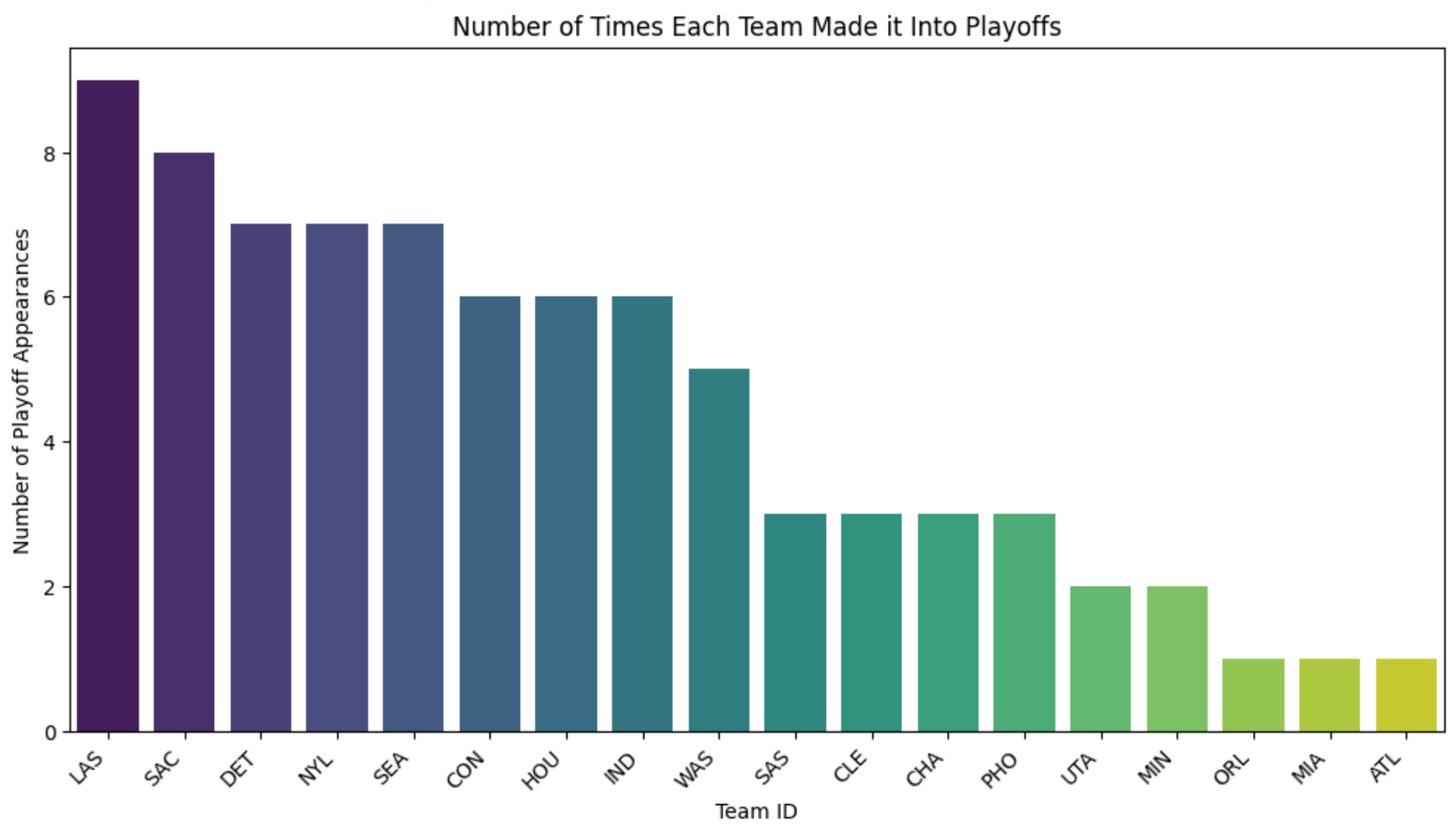
PLAYOFF TEAMS AND CHAMPIONS OF EACH YEAR

YEAR	PLAYOFF TEAMS
1	CLE, HOU, LAS, NYL, ORL, PHO, SAC, WAS
2	CHA, CLE, HOU, LAS, MIA, NYL, SAC, UTA
3	CHA, HOU, IND, LAS, NYL, SEA, UTA, WAS
4	CHA, CLE, CON, DET, HOU, LAS, MIN, SAC
5	CON, DET, LAS, MIN, NYL, SAC, SEA, WAS
6	CON, DET, HOU, IND, LAS, NYL, SAC, SEA
7	CON, DET, HOU, IND, LAS, SAC, SEA, WAS
8	CON, DET, IND, NYL, PHO, SAC, SAS, SEA
9	CON, DET, IND, LAS, NYL, SAC, SAS, SEA
10	ATL, DET, IND, LAS, PHO, SAS, SEA, WAS

YEAR	WINNER
1	HOU
2	LAS
3	LAS
4	DET
5	SEA
6	CON/SAC
7	DET
8	PHO
9	DET
10	PHO

EXPLORATORY DATA ANALYSIS - TEAMS

TIMES EACH TEAM MADE IT INTO PLAYOFFS



EXPLORATORY DATA ANALYSIS - TEAMS

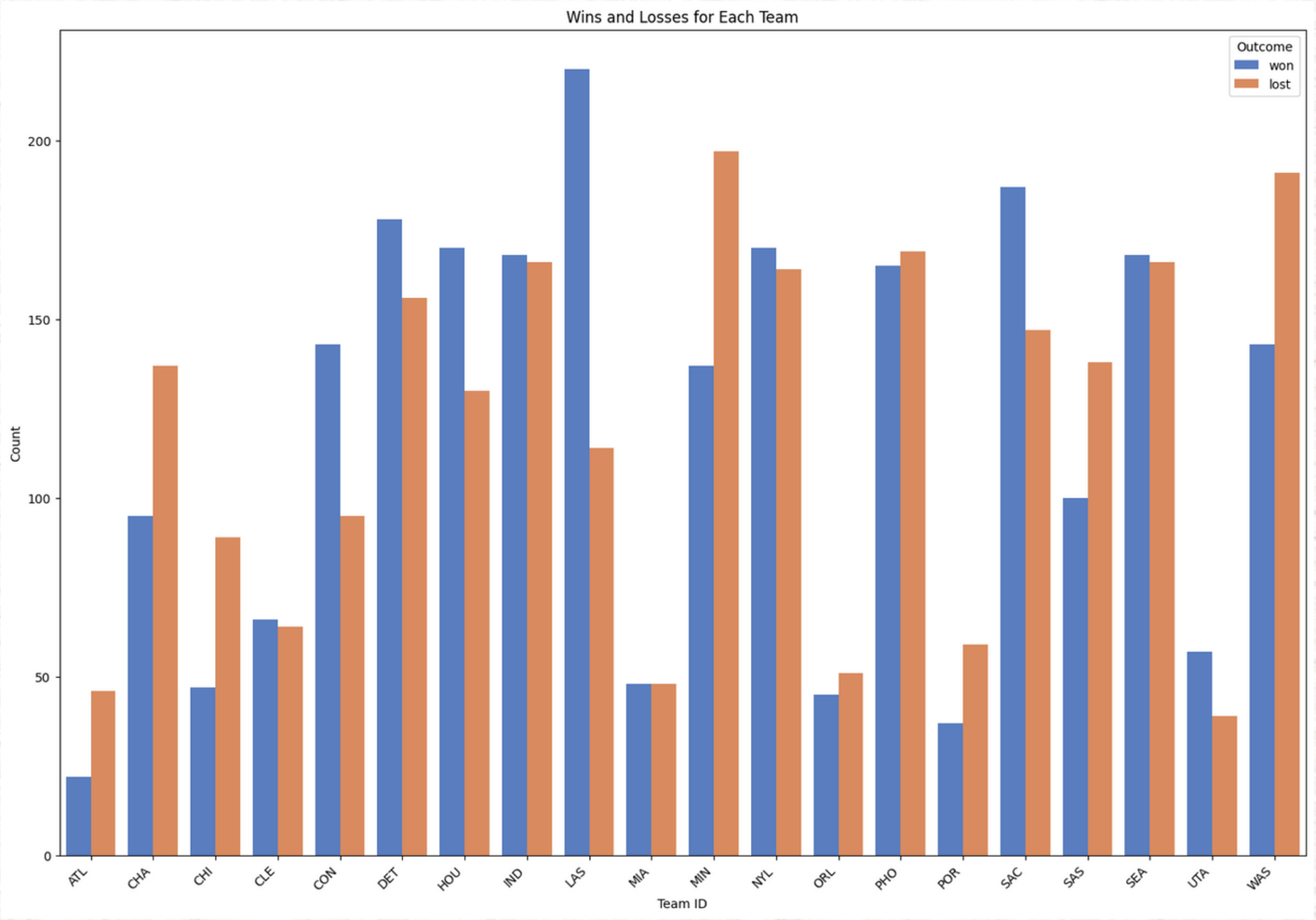
TEAM SEASON WINS & LOSSES

Top 3 Teams with the most season wins:

1. **LAS**
2. **SAC**
3. **DET**

Top 3 teams with the most season losses:

1. **MIN**
2. **WAS**
3. **PHO**



EXPLORATORY DATA ANALYSIS - TEAMS POST

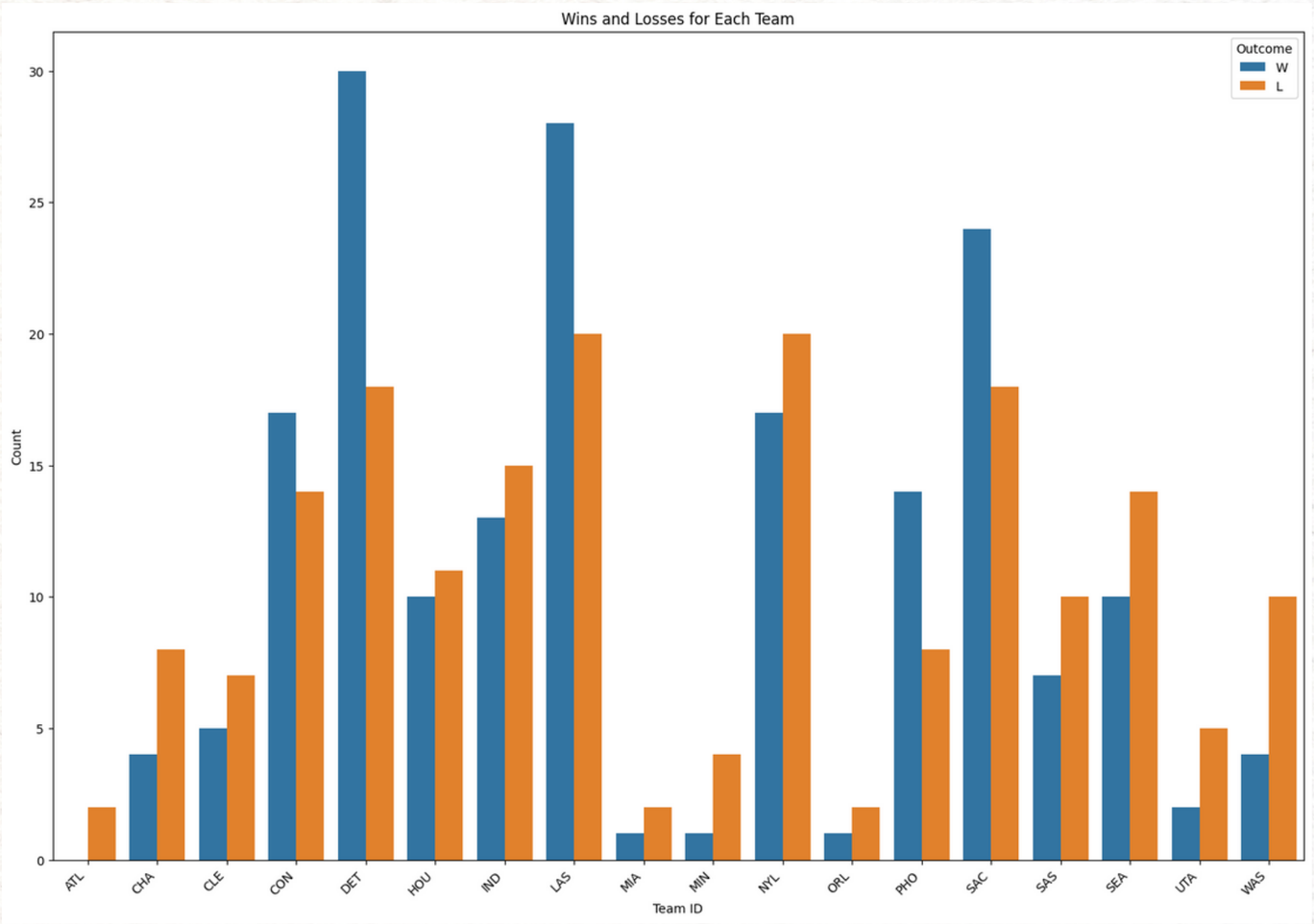
TEAM POST SEASON WINS & LOSSES

Top 3 Teams with the most season wins:

1. DET
2. LAS
3. SAC

Top 3 teams with the most season losses:

1. LAS
2. NYL
3. DET



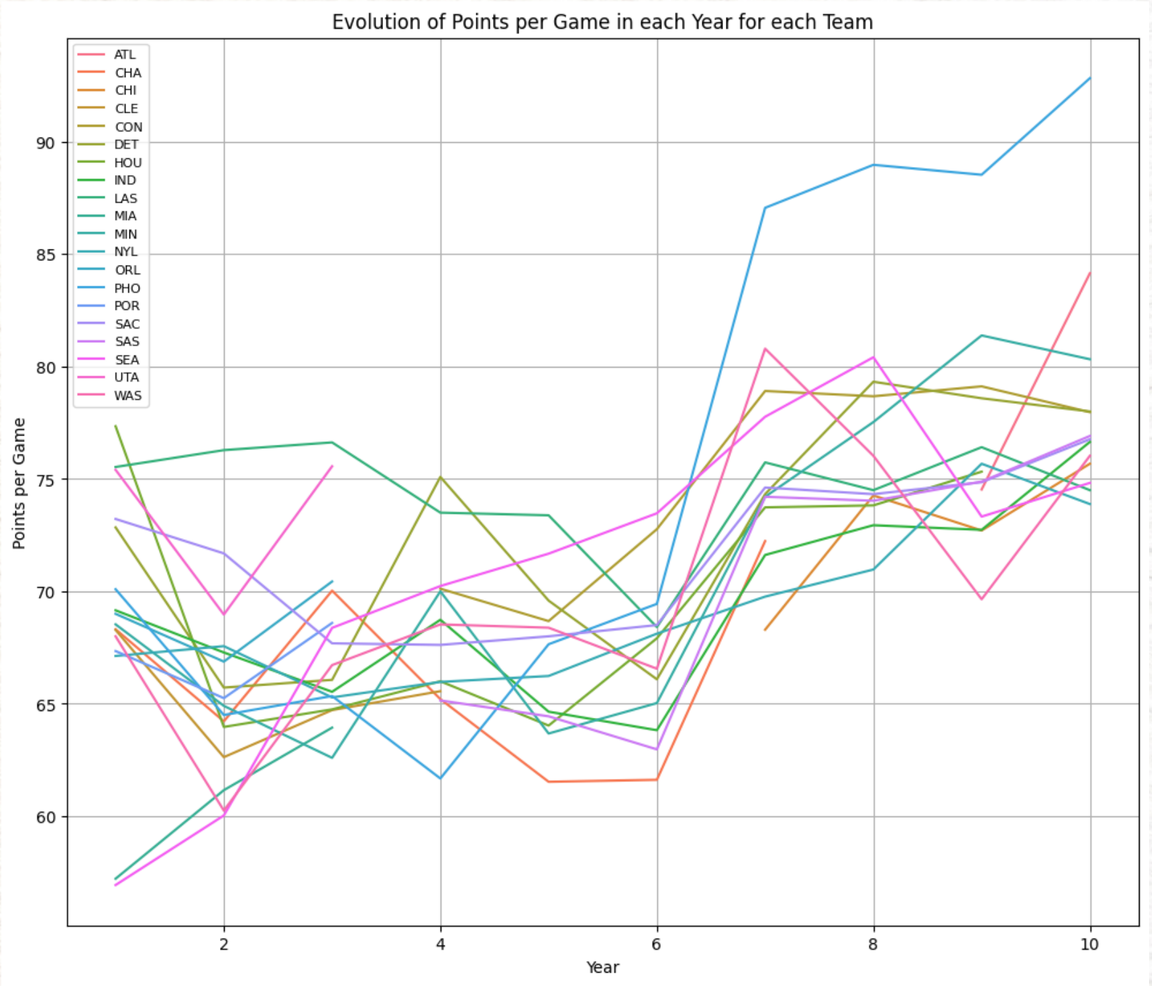
EXPLORATORY DATA ANALYSIS - TEAMS

POINTS PER GAME

Team PHO, Phoenix Mercury, showed the biggest increase in points per game, over the years, compared to any other team.

In year 7, all teams showed an increase in the amount of points per game relative to the previous year.

None of the teams showed consistency in the number of points per game as the years went by.



EXPLORATORY DATA ANALYSIS - COACHES

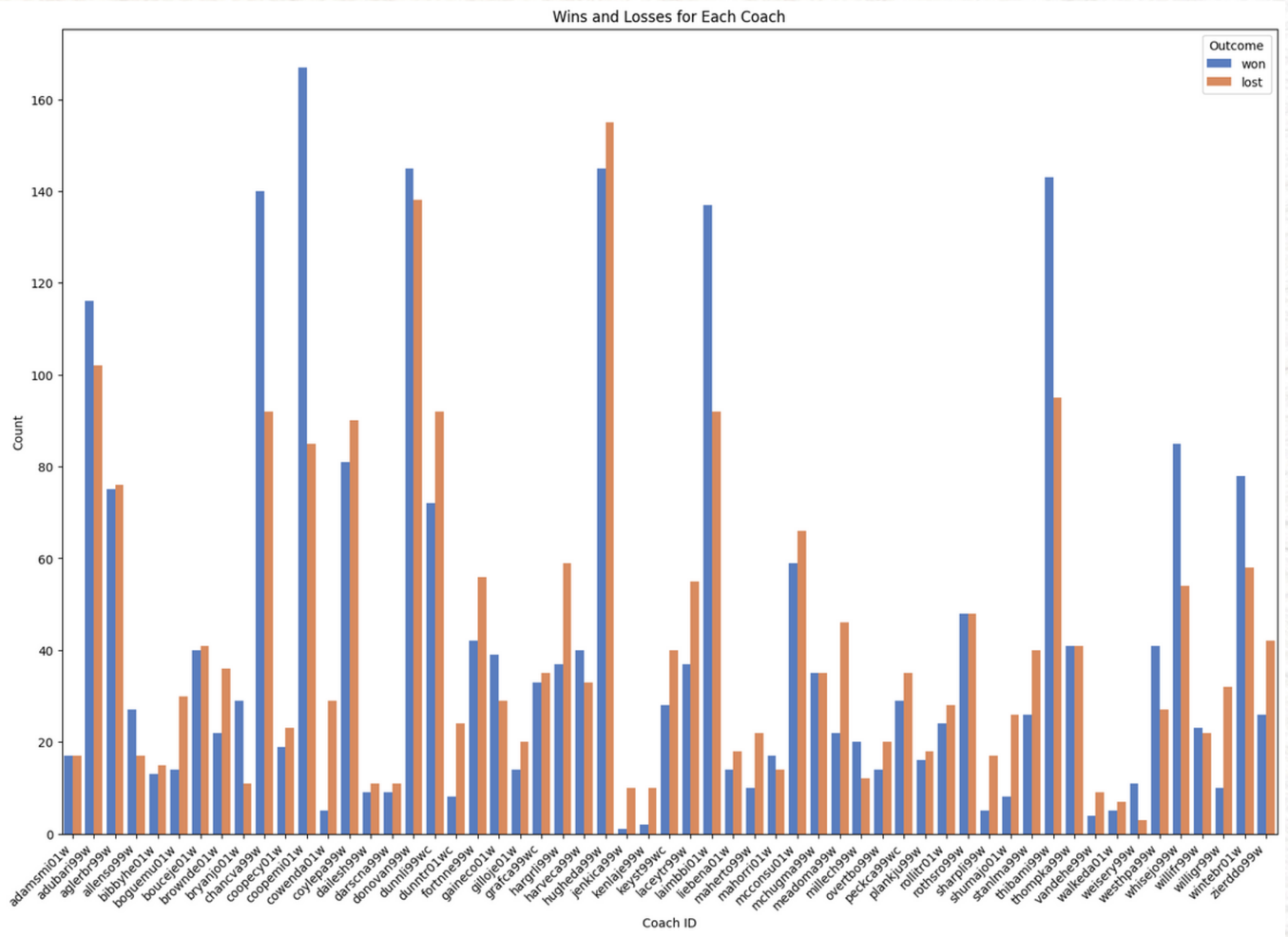
COACH SEASON WINS & LOSSES

Top 3 Coaches with the most season wins:

1. coopemi01w
2. hugheda99w
3. donovan99w

Top 3 Coaches with the most season losses:

1. hugheda99w
2. donovan99w
3. adubari99w



EXPLORATORY DATA ANALYSIS - TEAMS & COACHES

COACHES EACH TEAM HAS EVER HAD

ATL	[meadoma99w
CHA	[boguemu01w, donovan99w, dunntr01wc, laceytr99w
CHI	[cowenda01w, keyst99wc, overtbo99w
CLE	[hugheda99w
CON	[thibami99w
DET	[laimbbi01w, liebena01w, mahorri01w, willigr99w
HOU	[chancva99w, thompka99w
IND	[donovan99w, dunnli99wc, fortnne99w, wintebr01w
LAS	[bibbyhe01w, bryanjo01w, coopemi01w, thompka99w, weisery99w
MIA	[rothsro99w
MIN	[aglerbr99w, gillolje01w, jenkica99w, mcconsu01w, vandehe99w, zierddo99w
NYL	[adubari99w, coylepa99w, donovan99w
ORL	[browndeo1w, peckca99wc
PHO	[coopecy01w, gainecoo1w, grafca99wc, millech99w, sharpli99w, shumajoo1w, westhpa99w
POR	[hargrli99w
SAC	[allenso99w, boucejeo1w, mchugma99w, whisejo99w
SAS	[browndeo1w, dailesh99w, harveca99w, hugheda99w
SEA	[aglerbr99w, donovan99w, dunnli99wc
UTA	[harveca99w, willifr99w
WAS	[adamsmio1w, adubari99w, darscna99w, kenlaje99w, maherto99w, plankju99w, rollitro1w, stanlma99w, walkeda01w

EXPLORATORY DATA ANALYSIS - PLAYERS

AVERAGE HEIGHT

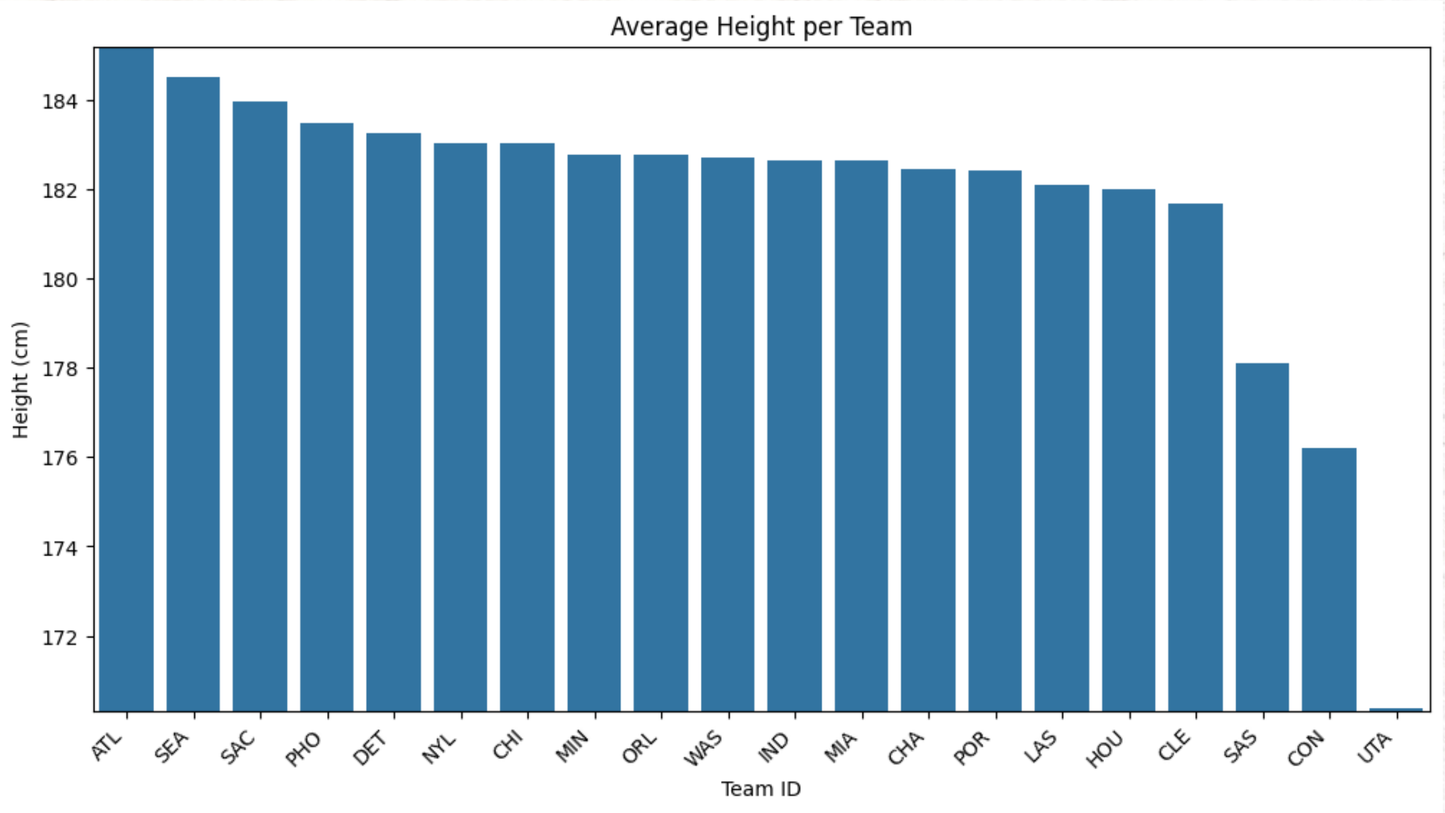
inches: 71.9

centimeters: 182.6

AVERAGE WEIGHT

pounds: 167.7

kilograms: 76.1



DATA PREPARATION

- Drop noisy data
- Drop unnecessary features
- Drop highly correlated features
- Drop single value features
- Merge dataframes
- Fill N/A values
- Some value replacement for consistency

DATA CLASSIFICATION

MODELLING RESULTS

MODEL	Last Year Accuracy	Last Year AUC-ROC
DecisionTree	0.62	0.54
SVM	0.54	0.51
LogisticRegression (sag)	0.62	0.61
RandomForest	0.62	0.58
GradientBoostingClassifier	0.54	0.48
KNeighbors	0.69	0.64

BEST RESULT

K-NEAREST NEIGHBOURS

Accuracy: 0.6923

Params: leaf_size=8, n_neightbors=8, p=1, weights=distance

CONCLUSIONS

- After performing a Randomized Grid Search to tune the parameters for each model, the most accurate prediction was made using K-Nearest Neighbors, with an accuracy of 69.23%, and it is the model we will be focusing on.
- Although the dataset has a lot of noisy and outright bad data, we managed to neatly clean it and use it to make accurate predictions. More work is needed in creating more useful features and further improving the data that is fed to the models.

THANK YOU