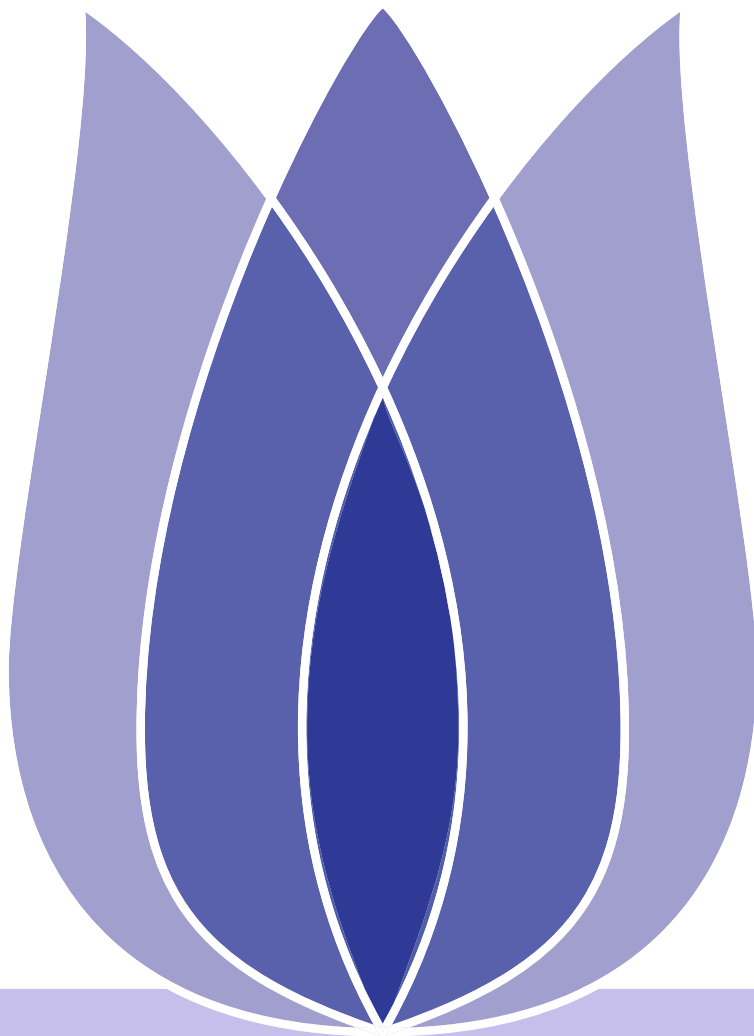


PUBG Game Data Analysis and prediction

Yang Cao

Deakin University

July 5, 2021





Overview

- [Problem Definition](#)
- [Data Preprocess](#)
- [Data Visualization](#)
- [Feature selection](#)
- [Modeling and Forecasting](#)
- [Comparison and Conclusion](#)

Problem Definition

Problem introduction

Data Preprocess

Data description

Dataset Description

Miss vlaue and NaN value

Data Visualization

Feature selection

Modeling and Forecasting

Model Selection

Comparison and Conclusion

Best Parameters

Comparison

Conclusion



Problem Definition

Problem introduction

Data Preprocess

Data Visualization

Feature selection

Modeling and Forecasting

Comparison and Conclusion

Problem Definition



Problem introduction

Problem Definition
Problem introduction
Data Preprocess
Data Visualization
Feature selection
Modeling and Forecasting
Comparison and Conclusion

Defn

In a PUBG game, up to 100 players start in each match (matchId). Players can be on teams (groupId) which get ranked at the end of the game (winPlacePerc) based on how many other teams are still alive when they are eliminated. In game, players can pick up different munitions, revive downed-but-not-out (knocked) teammates, drive vehicles, swim, run, shoot, and experience all of the consequences – such as falling too far or running themselves over and eliminating themselves. Different game behaviors will lead to different final rankings, so the main purpose is to build a model to predicts players' finishing placement based on their final stats, on a scale from 1 (first place) to 0 (last place).

- A game team data analyst may be interested in the **game actions** that make **game teams get higher rank** than others.
- Players can also estimate their final ranking based on the current situation and make strategic decisions in advance (such as running away or fighting).



Description and Evaluation

- Problem Definition
- Problem introduction**
- Data Preprocess
- Data Visualization
- Feature selection
- Modeling and Forecasting
- Comparison and Conclusion

Desc

Mean Square Error: the average squared difference between the estimated values and the actual value

- Train dataset MSE
- Test Dataset MSE



[Problem Definition](#)

[Data Preprocess](#)

[Data description](#)

[Dataset Description](#)

[Miss vlaue and NaN value](#)

[Data Visualization](#)

[Feature selection](#)

[Modeling and Forecasting](#)

[Comparison and Conclusion](#)

Data Preprocess



- [Problem Definition](#)
- [Data Preprocess](#)
- [Data description](#)
- [Dataset Description](#)
- [Miss vlaue and NaN value](#)
- [Data Visualization](#)
- [Feature selection](#)
- [Modeling and Forecasting](#)
- [Comparison and Conclusion](#)

Table 1: Data Field

Attributes	description
DBNOs	Number of enemy players knocked
Assists	Number of enemy players this player damaged that were killed by teammates
Boosts	Number of boost items used
damageDealt	Total damage dealt
headshotKills	Number of enemy players killed with headshots
heals	Number of healing items used
killPlace	Ranking in match of number of enemy players killed
killPoints	Kills-based external ranking of player
killStreaks	Max number of enemy players killed in a short amount of time
kills	Number of enemy players killed
longestKill	Longest distance between player and player killed at time of death
rankPoints	Elo-like ranking of player
revives	Number of times this player revived teammates



Dataset Description

- [Problem Definition](#)
- [Data Preprocess](#)
- [Data description](#)
- [Dataset Description](#)**
- [Miss vlaue and NaN value](#)
- [Data Visualization](#)
- [Feature selection](#)
- [Modeling and Forecasting](#)
- [Comparison and Conclusion](#)

■ Train and Test Dataset

- ◆ train_v2.csv
- ◆ test_v2.csv

train_v2.csv

- ◆ There are 4446966 rows and 29 columns.
- ◆ 4446966 unique ID.
- ◆ 2026745 unique groupId

test_v2.csv

- ◆ There are 1934174 rows and 28 columns.
- ◆ 1934174 unique ID
- ◆ 886238 unique groupId



Miss vlaue and NaN value

- [Problem Definition](#)
- [Data Preprocess](#)
- [Data description](#)
- [Dataset Description](#)
- [Miss vlaue and NaN value](#)**
- [Data Visualization](#)
- [Feature selection](#)
- [Modeling and Forecasting](#)
- [Comparison and Conclusion](#)

Id	0
groupId	0
matchId	0
assists	0
boosts	0
damageDealt	0
DBNOs	0
headshotKills	0
heals	0
killPlace	0
killPoints	0
kills	0
killStreaks	0
longestKill	0
matchDuration	0
matchType	0
maxPlace	0
numGroups	0
rankPoints	0
revives	0
rideDistance	0
roadKills	0
swimDistance	0
teamKills	0
vehicleDestroys	0
walkDistance	0
weaponsAcquired	0
winPoints	0
winPlacePerc	1
dtype: int64	



[Problem Definition](#)

[Data Preprocess](#)

[Data Visualization](#)

[Feature selection](#)

[Modeling and Forecasting](#)

[Comparison and Conclusion](#)

Data Visualization



Game type proportion

- [Problem Definition](#)
- [Data Preprocess](#)
- [Data Visualization](#)
- [Feature selection](#)
- [Modeling and Forecasting](#)
- [Comparison and Conclusion](#)

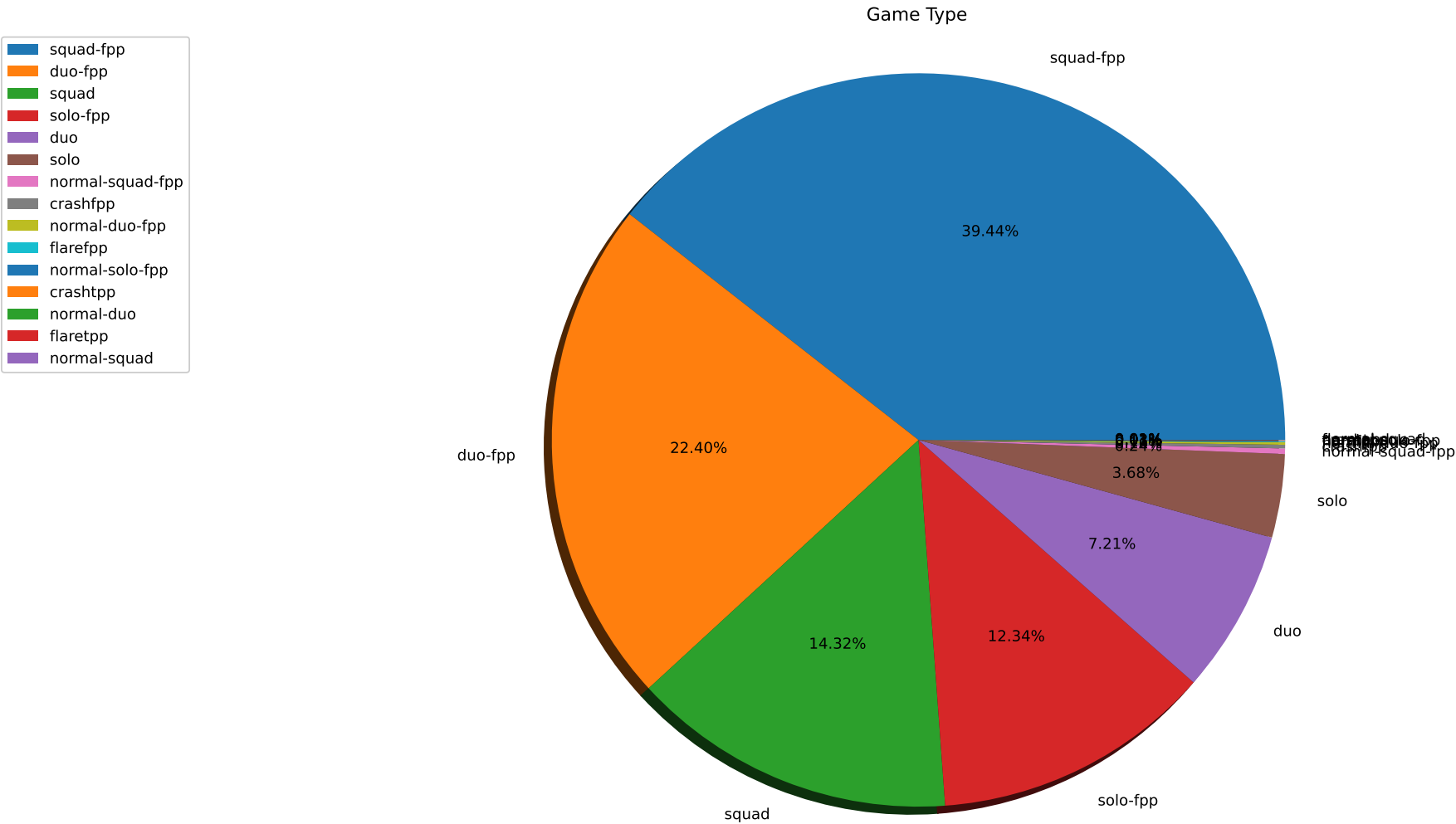


Figure 2: game type proportion



Walk distance with win place

- [Problem Definition](#)
- [Data Preprocess](#)
- [Data Visualization](#)
- [Feature selection](#)
- [Modeling and Forecasting](#)
- [Comparison and Conclusion](#)

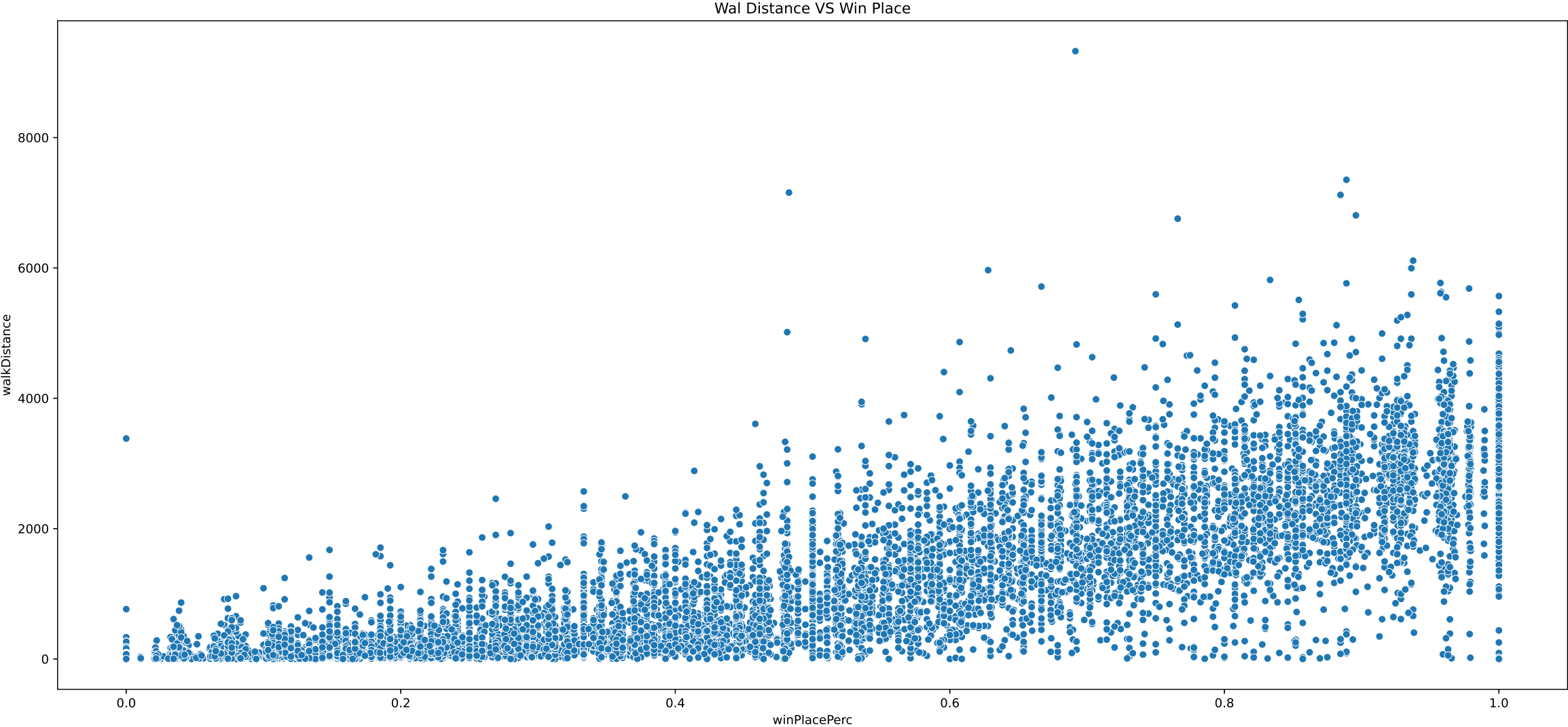


Figure 3: walking distance VS Win Place



- [Problem Definition](#)
- [Data Preprocess](#)
- [Data Visualization](#)
- [Feature selection](#)**
- [Modeling and Forecasting](#)
- [Comparison and Conclusion](#)

Feature selection



Correlation

- Problem Definition
- Data Preprocess
- Data Visualization
- Feature selection
- Modeling and Forecasting
- Comparison and Conclusion

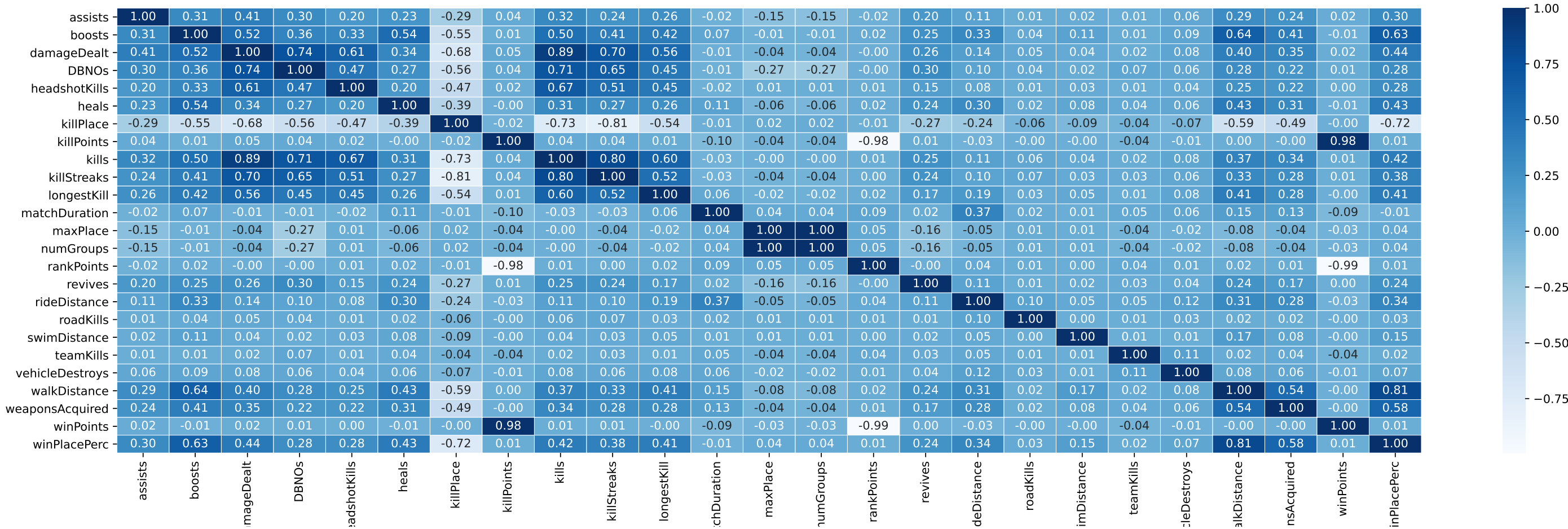


Figure 4: Correlation



High Correlation

- Problem Definition
- Data Preprocess
- Data Visualization
- Feature selection
- Modeling and Forecasting
- Comparison and Conclusion

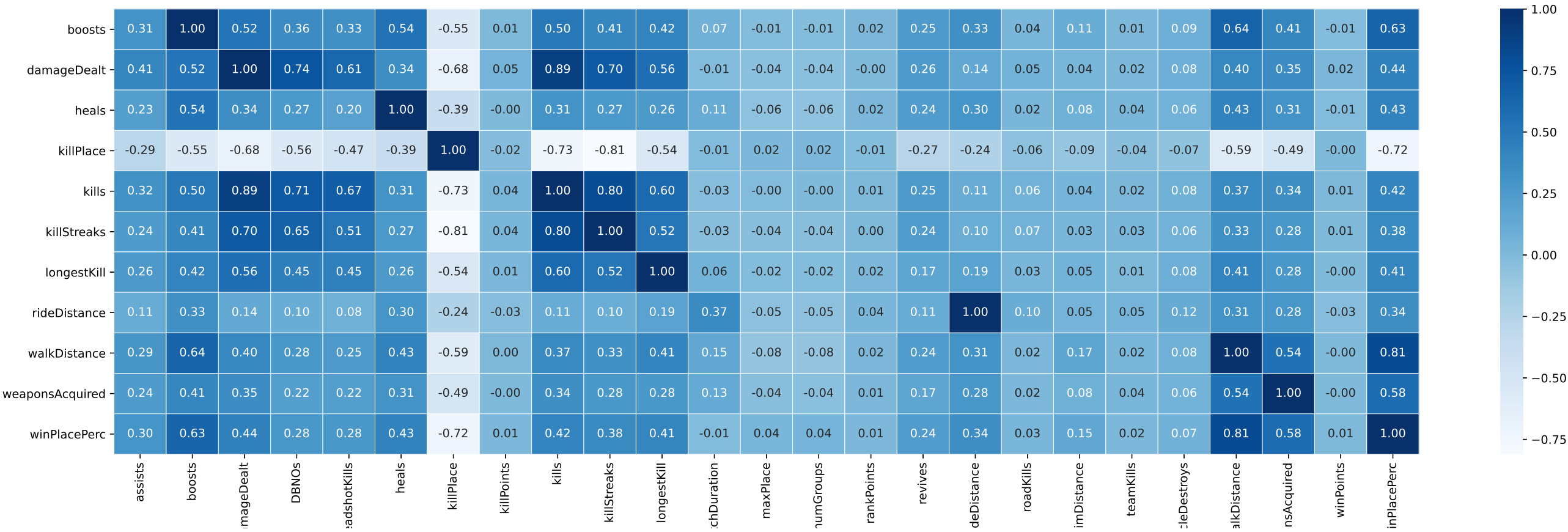


Figure 5: High Correlation



- [Problem Definition](#)
- [Data Preprocess](#)
- [Data Visualization](#)
- [Feature selection](#)
- [Modeling and Forecasting](#)**
- [Model Selection](#)
- [Comparison and Conclusion](#)

Modeling and Forecasting



Model Selection

- [Problem Definition](#)
- [Data Preprocess](#)
- [Data Visualization](#)
- [Feature selection](#)
- [Modeling and Forecasting](#)
- [Model Selection](#)**
- [Comparison and Conclusion](#)

- Linear Regression
- Decision Tree



- [Problem Definition](#)
- [Data Preprocess](#)
- [Data Visualization](#)
- [Feature selection](#)
- [Modeling and Forecasting](#)
- [Model Selection](#)
- [Comparison and Conclusion](#)

- Sklearn linear regression parameters on grid search and cross validation.

Table 2: linear regression parameters and cross validation

Parameters	Values	CV
fit_intercept	True/False	3
normalize	True/False	3



- [Problem Definition](#)
- [Data Preprocess](#)
- [Data Visualization](#)
- [Feature selection](#)
- [Modeling and Forecasting](#)
- [Model Selection](#)
- [Comparison and Conclusion](#)

- Sklearn Decision Tree parameters on grid search and cross validation.

Table 3: Decision Tree parameters and cross validation

Parameters	Values	CV
criterion	"mse", "friedman_mse", "mae"	3
min_samples_leaf	1,2	3



[Problem Definition](#)

[Data Preprocess](#)

[Data Visualization](#)

[Feature selection](#)

[Modeling and Forecasting](#)

[Comparison and Conclusion](#)

[Best Parameters](#)

[Comparison](#)

[Conclusion](#)

Comparison and Conclusion



Best Parameters

- Problem Definition
- Data Preprocess
- Data Visualization
- Feature selection
- Modeling and Forecasting
- Comparison and Conclusion
- Best Parameters**
- Comparison
- Conclusion

- Best parameters by grid search
- Linear Regression: `copy_X=True`, `fit_intercept=True`, `n_jobs=None`, `normalize=True`
- Decision Tree: `ccp_alpha=0.0`, `criterion='mse'`, `max_depth=None`, `max_features=None`, `max_leaf_nodes=None`, `min_impurity_decrease=0.0`, `min_impurity_split=None`, `min_samples_leaf=2`, `min_samples_split=2`, `min_weight_fraction_leaf=0.0`, `presort='deprecated'`, `random_state=None`, `splitter='best'`



- [Problem Definition](#)
- [Data Preprocess](#)
- [Data Visualization](#)
- [Feature selection](#)
- [Modeling and Forecasting](#)
- [Comparison and Conclusion](#)
- [Best Parameters](#)
- [Comparison](#)**
- [Conclusion](#)

■ Compare MSE result between Linear regression and Decision Tree

Table 4: Linear regression VS Decision Tree

	train MSE	test MSE
linear regression	0.01564124116618947	0.015303007019988265
decision tree	0.01859312767771217	0.17048691321544765



Conclusion

- [Problem Definition](#)
- [Data Preprocess](#)
- [Data Visualization](#)
- [Feature selection](#)
- [Modeling and Forecasting](#)
- [Comparison and Conclusion](#)
- [Best Parameters](#)
- [Comparison](#)
- [Conclusion](#)

- Both training and testing data shows that linear model get lower mean square error value.
- Most players choose to play squad-fpp and duo-fpp
- More walking distance always can bring higher win place.



Questions?

- [Problem Definition](#)
- [Data Preprocess](#)
- [Data Visualization](#)
- [Feature selection](#)
- [Modeling and Forecasting](#)
- [Comparison and Conclusion](#)
- [Best Parameters](#)
- [Comparison](#)
- [Conclusion](#)



Contact Information

YANG CAO

Deakin University, Australia



CAOYANG@DEAKIN.EDU.AU



TEAM FOR UNIVERSAL LEARNING AND INTELLIGENT PROCESSING

