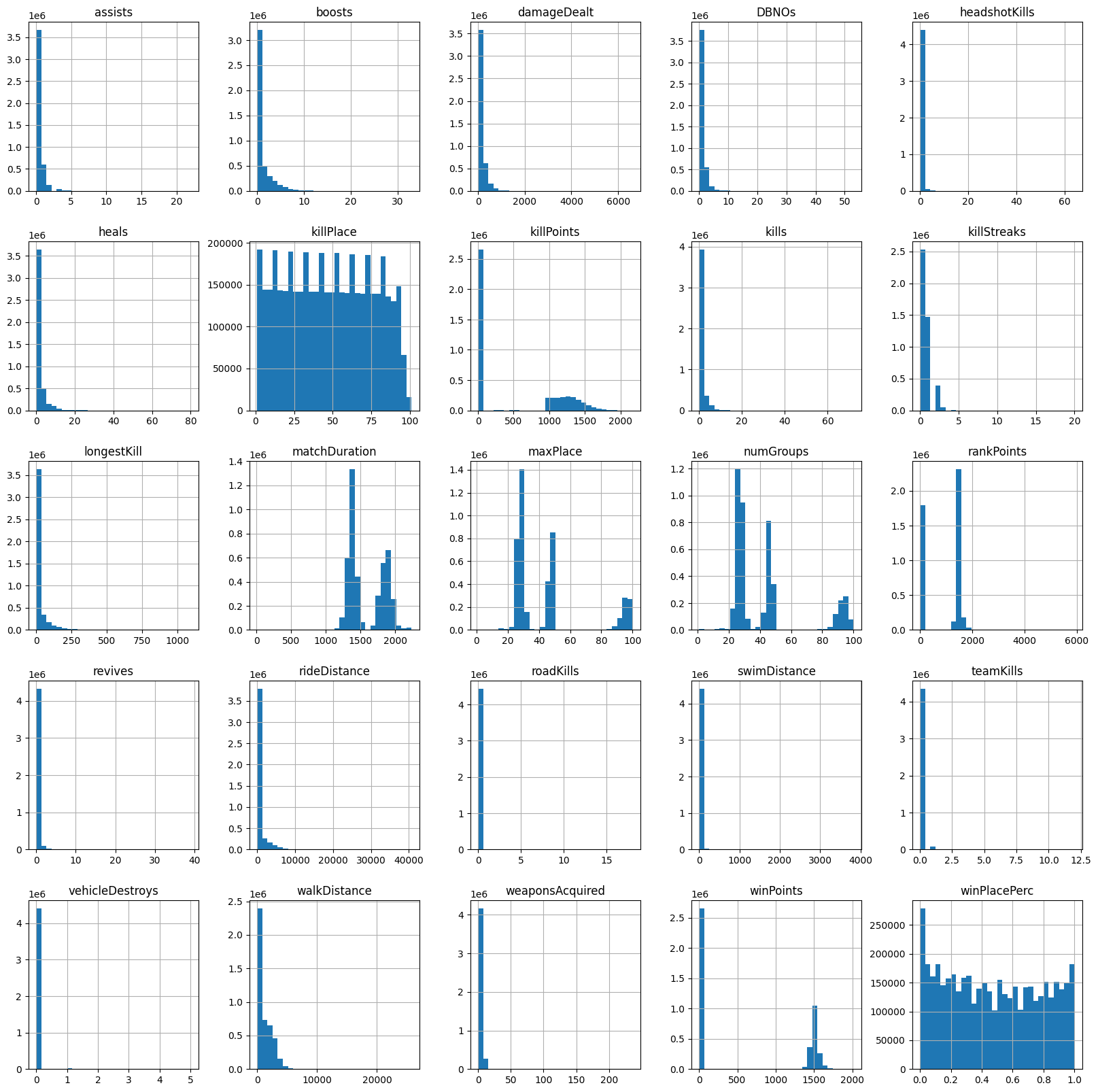
### **Problem Statement**

The objective of this project is to analyze PUBG gameplay data to:

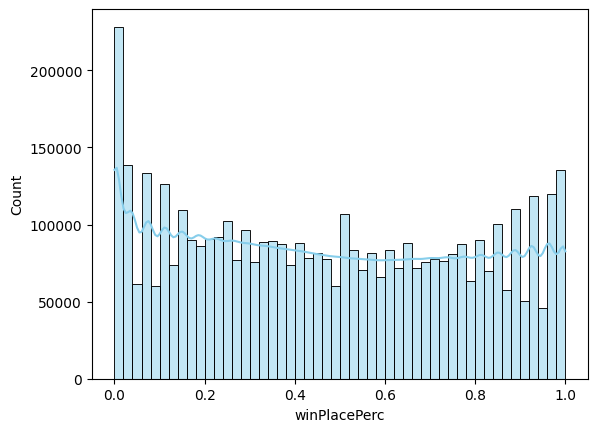
* Prepare a detailed data analysis report.
* Develop a predictive model to estimate the **win probability (**winPlacePerc**)** for a player or group in a PUBG match.
* Identify **important features** affecting the likelihood of winning.

## Task 1: Data Analysis Report

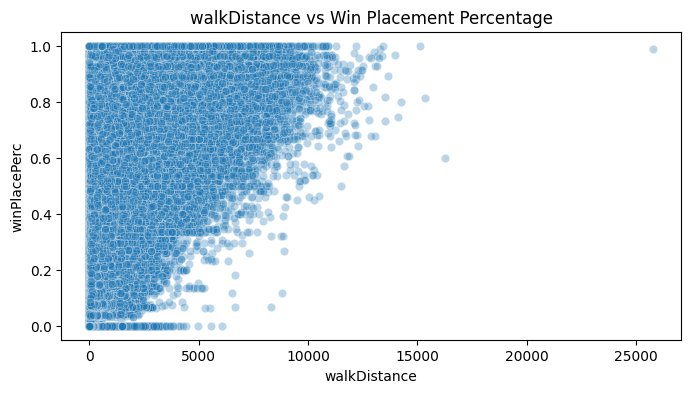
* **Total Observations:** ~4.4 million (before cleaning)
* **Total Features:** 29
* **Target Variable:** **winPlacePerc** (continuous, range 0 to 1)
* Pairplot

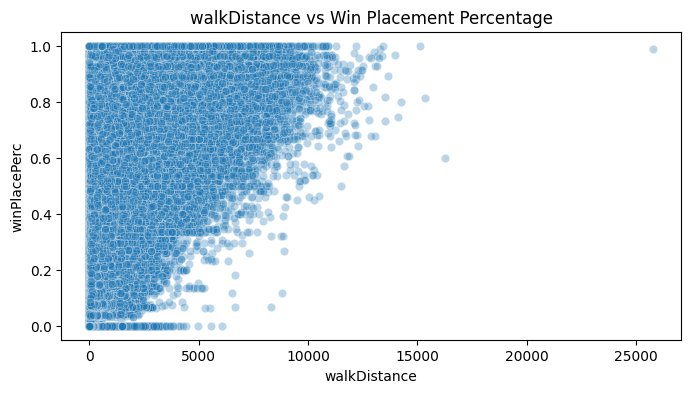


* Univariate Analysis



* Bi variate Analysis





| **Feature Type** | **Examples** | **Action Taken** |
| --- | --- | --- |
| Identifier Columns | Id, groupId, matchId | Dropped (not predictive) |
| Categorical Column | matchType | One-hot encoded |
| Numerical Features | kills, damageDealt, walkDistance, etc. | Retained & scaled where needed |

## Target Variable Insights

* winPlacePerc distribution was **bimodal**, with peaks at 0 and 1.
* Indicates most players either die early or survive till the end.
* Bottom 1% of values were optionally filtered to reduce noise from inactive players.

### **Feature Relationships**

* **Highly Correlated with winPlacePerc**:
  + walkDistance, damageDealt, boosts, heals, kills
* **Low/No Correlation**:
  + roadKills, vehicleDestroys, rankPoints (in some cases)

### **Data Cleaning Highlights**

* Handled missing values in winPlacePerc, rankPoints, and other sparse features.
* Removed outliers using the **IQR method** for features like kills, damageDealt, and walkDistance.
* Dropped non-predictive IDs and duplicates.

| **Model** | **R² Score** |
| --- | --- |
| Linear Regression | 0.84 |
| Decision Tree Regressor | 0.85 |
| Random Forest Regressor | 0.92 |
| XGBoost Regressor | 0.92 |
| Gradient Boosting | 0.89 |

**Best Model:** XGBoost Regressor  
Highest accuracy and lowest error, with strong handling of non-linearity and outliers.

### **Feature Importance from XGBoost**

Top 5 Features:

1. walkDistance
2. boosts
3. damageDealt
4. heals
5. kills

| **Challenge** | **Description** | **Solution Used** |
| --- | --- | --- |
| Redundant/ID columns | Id, groupId, matchId are not predictive | Dropped to reduce noise |
| Feature encoding | matchType is categorical | Applied one-hot encoding |
| Outliers in key metrics | damageDealt, kills, longestKill had extreme outliers | Used IQR to remove |
| Overfitting in decision trees | Single trees performed well on train but poorly on test | Shifted to ensemble models like RF & XGBoost |