Data Sources, Datasets, and Features Summary

Prepared for: Player Market & Sentiment Analysis Project

1) Data Sources

Kaggle: Public datasets used for social media sentiment (Twitter training/validation) and football/soccer player information. Kaggle was the primary source for curated CSVs containing tweets with sentiment labels and player/competition metadata.

StatsBomb (Open Data): Used for structured football data such as competitions, players, injuries/impact or related attributes. StatsBomb's open data provides standardized schema for football analytics.

2) Collected Datasets (Files Uploaded)

File	Rows	Columns
twitter_training.csv	74681	4
twitter_validation.csv	999	4
dataset.csv	1301	30
player_injuries_impact.csv	656	42
player_valuations.csv	496606	5
competitions.csv	44	11
players.csv	32601	23

3) Features / Columns by Dataset

twitter_training.csv

`Path: /mnt/data/twitter_training.csv`

Shape: 74681 rows × 4 columns

Features: *2401*, *Borderlands*, *Positive*, *im getting on borderlands and i will murder you all ,*

twitter_validation.csv

`Path: /mnt/data/twitter_validation.csv`

Shape: 999 rows × 4 columns

Features: *3364*, *Facebook*, *Irrelevant*, *I mentioned on Facebook that I was struggling for motivation to go for a run the other day, which has been translated by Tom's great auntie as 'Hayley can't get out of bed' and told to his grandma, who now thinks I'm a lazy, terrible person ■*

dataset.csv

`Path: /mnt/data/dataset.csv`

Shape: 1301 rows × 30 columns

Features: *p_id2*, *start_year*, *season_days_injured*, *total_days_injured*, *season_minutes_played*, *season_games_played*, *season_matches_in_squad*, *total_minutes_played*, *total_games_played*, *dob*, *height_cm*, *weight_kg*, *nationality*, *work_rate*, *pace*, *physic*, *fifa_rating*, *position*, *age*, *cumulative_minutes_played*, *cumulative_games_played*, *minutes_per_game_prev_seasons*, *avg_days_injured_prev_seasons*, *avg_games_per_season_prev_seasons*, *bmi*, *work_rate_numeric*, *position_numeric*, *significant_injury_prev_season*, *cumulative_days_injured*, *season_days_injured_prev_season*

player_injuries_impact.csv

Path: `/mnt/data/player_injuries_impact.csv`

Shape: 656 rows × 42 columns

Features: *Name*, *Team Name*, *Position*, *Age*, *Season*, *FIFA rating*, *Injury*, *Date of Injury*, *Date of return*, *Match1_before_injury_Result*, *Match1_before_injury_Opposition*, *Match1_before_injury_GD*, *Match1_before_injury_Player_rating*, *Match2_before_injury_Result*, *Match2_before_injury_Opposition*, *Match2_before_injury_GD*, *Match2_before_injury_Player_rating*, *Match3_before_injury_Result*, *Match3_before_injury_Opposition*, *Match3_before_injury_GD*, *Match3_before_injury_Player_rating*, *Match1_missed_match_Result*, *Match1_missed_match_Opposition*, *Match1_missed_match_GD*, *Match2_missed_match_Result*, *Match2_missed_match_Opposition*, *Match2_missed_match_GD*, *Match3_missed_match_Result*, *Match3_missed_match_Opposition*, *Match3_missed_match_GD*, *Match1_after_injury_Result*, *Match1_after_injury_Opposition*, *Match1_after_injury_GD*, *Match1_after_injury_Player_rating*, *Match2_after_injury_Result*, *Match2_after_injury_Opposition*, *Match2_after_injury_GD*, *Match2_after_injury_Player_rating*, *Match3_after_injury_Result*, *Match3_after_injury_Opposition* … (+2 more)

player_valuations.csv

Path: `/mnt/data/player_valuations.csv`

Shape: 496606 rows × 5 columns

Features: *player_id*, *date*, *market_value_in_eur*, *current_club_id*, *player_club_domestic_competition_id*

competitions.csv

Path: `/mnt/data/competitions.csv`

Shape: 44 rows × 11 columns

Features: *competition_id*, *competition_code*, *name*, *sub_type*, *type*, *country_id*, *country_name*, *domestic_league_code*, *confederation*, *url*, *is_major_national_league*

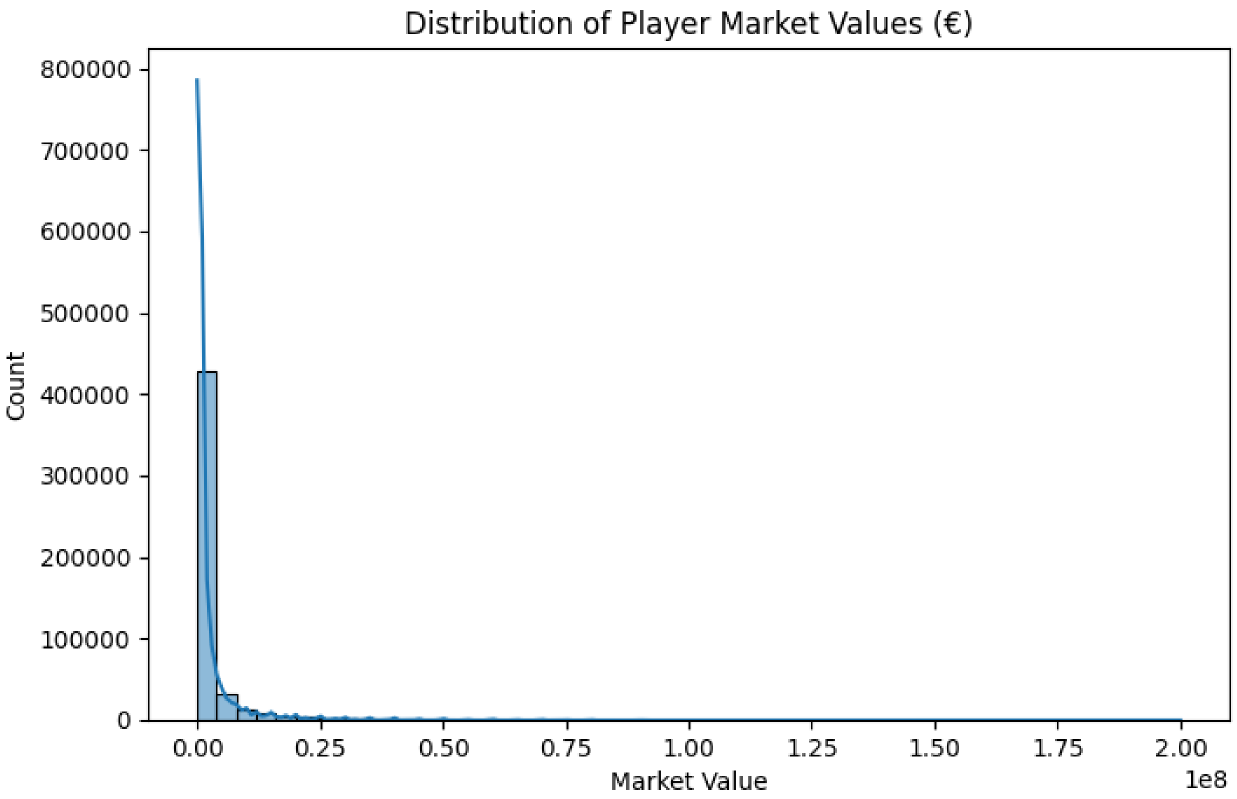
players.csv

Path: `/mnt/data/players.csv`

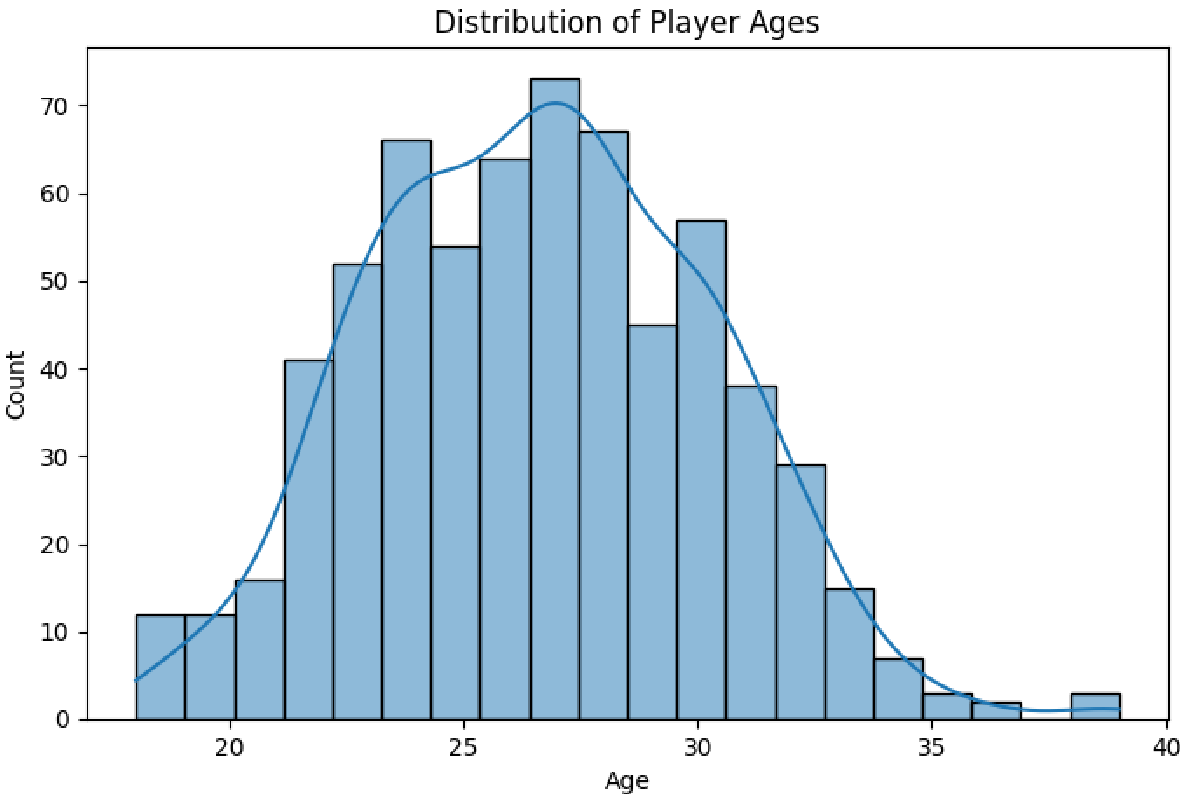
Shape: 32601 rows × 23 columns

Features: *player_id*, *first_name*, *last_name*, *name*, *last_season*, *current_club_id*, *player_code*, *country_of_birth*, *city_of_birth*, *country_of_citizenship*, *date_of_birth*, *sub_position*, *position*, *foot*, *height_in_cm*, *contract_expiration_date*, *agent_name*, *image_url*, *url*, *current_club_domestic_competition_id*, *current_club_name*, *market_value_in_eur*, *highest_market_value_in_eur*

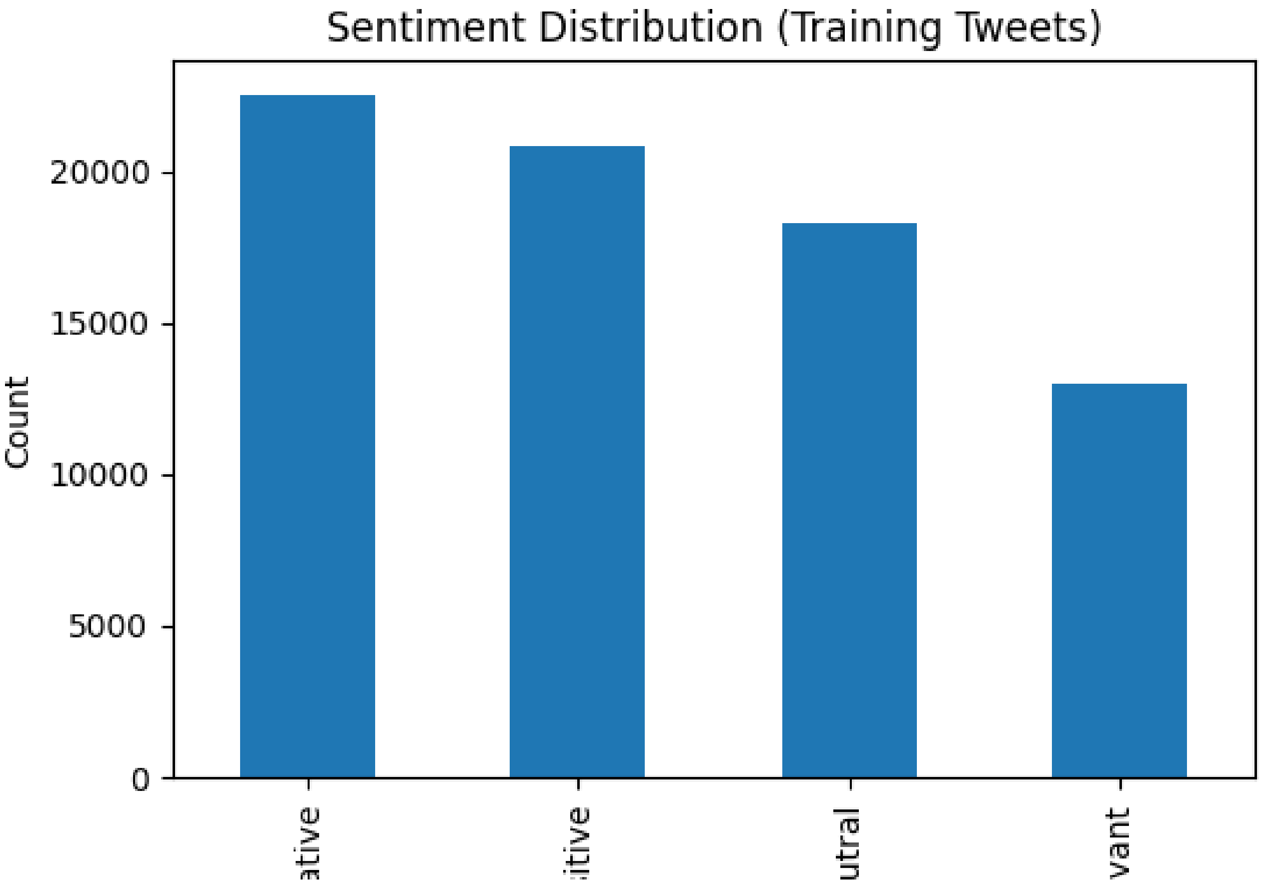
4) Visualizations Generated



Distribution of Player Market Values (€)



Distribution of Player Ages



Sentiment Distribution (Training Tweets)

5) Notes on Preparation

- CSV files were obtained from Kaggle and StatsBomb open data repositories and combined locally for analysis.
- Player-related tables (e.g., players, player_valuations, injuries/impact, competitions) provide demographics, market values, and contextual metadata used for distribution plots.
- Twitter datasets (training and validation) include tweet text and sentiment labels, enabling the computation of sentiment distributions.
- Plots included: (a) Player Market Values distribution, (b) Player Ages distribution, and (c) Sentiment label counts for training tweets.
- Standard cleaning steps typically include handling missing values, type casting (e.g., ensuring numeric market values), and de-duplicating records before visualization.