# TransferIQ — Project Progress Report

## Milestone 1: Week 1 – Data Collection and Initial Exploration

Day 1 – StatsBomb Data Exploration  
Explored the StatsBomb Open Data repository. This dataset provides rich, event-level football analytics with detailed player information such as name, ID, team, rank, gender, and nationality. It forms the foundation for performance profiling, player comparison, and match-level event analysis across seasons.  
  
Day 2 – Transfermarkt Data (Kaggle Source)  
Collected and explored Transfermarkt data from Kaggle. The dataset includes player demographics (age, nationality, position), club data, and historical transfer details (fees, dates, transfer types). It also tracks market value trends, appearances, goals, and assists — useful for analyzing player value evolution, club spending patterns, and transfer strategies.  
  
Day 3 – Combined Dataset Creation  
Merged multiple data sources to form a unified player dataset integrating:  
- Performance & Match Data (Futbin / StatsBomb)  
- Market Value & Financial Data (Top 5 Players dataset)  
- Placeholders for Sentiment, Injury, and Advanced Analytics  
  
Day 4 – Sentiment Analysis via Twitter API  
Used Tweepy to fetch player-related tweets and TextBlob for basic sentiment classification. Created an initial dataset containing columns such as player, tweet, created\_at, and sentiment label.  
  
Day 5 – Injury History Data Integration  
Began collecting and analyzing injury history data for key players, linking them to the master dataset to understand injury impact on player valuation.

## Milestone 2: Week 2 – Data Cleaning and Preprocessing

Day 6  
Initiated data cleaning and feature engineering. Verified data integrity across all sources and finalized data collection.  
  
Day 7–8  
Completed cleaning of all datasets — main player data, sentiment data, and injury data. Began compilation of the Master Player List.  
  
Day 9  
Finalized the Master Player List and connected it to SQL for structured storage and querying.

## Milestone 3: Weeks 3–4 – Advanced Feature Engineering & Sentiment Analysis

Day 10  
Performed feature engineering on master\_list\_final.csv. Generated analytical features such as age\_experience, contract\_risk, and total\_days\_missed.  
  
Day 11  
Finalized feature engineering and saved master\_list\_final\_features.csv. Ensured placeholders for sentiment, injury, and performance metrics were filled.  
  
Day 12  
Implemented a Random Forest Regression model to predict market\_value\_in\_eur. Calculated MSE and R². Saved rf\_model.pkl, rf\_predictions.csv, and feature\_importances.csv.  
  
Day 13  
Conducted Exploratory Data Analysis (EDA) — generated correlation heatmaps, checked multicollinearity, and visualized feature distributions.  
  
Day 14  
Applied PCA and LDA for dimensionality reduction to simplify the dataset while retaining key information.  
  
Day 15  
Benchmarked Linear Regression and XGBoost models against Random Forest using RMSE, MAE, and R² metrics.

## Milestone 4: Week 5 – LSTM Model Development

Day 16  
Built a prediction pipeline handling preprocessing, feature selection, and transfer value prediction for new data.  
  
Day 17  
Developed a Univariate LSTM model using TensorFlow/Keras to predict future transfer trends.  
  
Day 18  
Expanded to a Multivariate LSTM incorporating performance, injury, and sentiment data. Defined key hyperparameters.  
  
Day 19  
Performed hyperparameter tuning for the LSTM — experimented with learning rates, hidden units, and dropout values.  
  
Day 20  
Integrated Early Stopping and Model Checkpointing to improve stability and prevent overfitting.

## Milestone 5–7: Weeks 6–8 – Model Integration, Evaluation & Visualization

Day 21  
Conducted model evaluation and comparison between LSTM, Random Forest, Linear Regression, and XGBoost. LSTM captured time-based trends most effectively.  
  
Day 22  
Finalized the Model Development Phase. Exported the final Multivariate LSTM model, documented hyperparameters, and prepared for deployment and interpretability phase.

## Deliverables Achieved

- Multi-source integrated dataset (performance, injury, sentiment, market)  
- Cleaned and feature-engineered datasets ready for modeling  
- Trained LSTM, XGBoost, and Random Forest models  
- Comparative evaluation and visualization of player value trends  
- Ready transition to deployment and interpretability phase