# Year-Long Data Engineering Project: Comprehensive Storyline and Problem Statements

### Month 1-3: Introduction and Data Familiarization  
\*\*Context:\*\*   
Welcome to your new role as a Data Engineer at Sports Analytics Inc. You have joined a dynamic team focused on leveraging data to gain insights into player performance and team strategies. Your first task involves getting familiar with the dataset, which includes player statistics from various teams.  
  
#### Tasks:  
1. \*\*Data Cleaning and Augmentation:\*\*  
 - \*\*Problem Statement 1:\*\* Identify and handle missing values using advanced imputation techniques. Correct anomalies by identifying outliers using statistical methods and domain knowledge. Standardize data formats and ensure consistency across the dataset. Augment the dataset by generating synthetic data using data augmentation techniques and collecting additional data from public sports databases. Integrate this data into a unified dataset.  
 - \*\*Additional Complexity:\*\* Implement data validation scripts to automate the data cleaning process.  
  
2. \*\*Position Analysis:\*\*  
 - \*\*Problem Statement 2:\*\* Analyze player positions to identify the highest and lowest number of players. Use statistical analysis to determine if the distribution of players across positions is significantly different from a uniform distribution. Create a plot showing the count of players for each position and a pie chart for distribution.  
 - \*\*Additional Complexity:\*\* Perform a chi-square test to validate your findings.  
  
### Month 4-6: Data Ingestion and Initial Analysis  
\*\*Context:\*\*   
With a clean and robust dataset, it's time to implement data ingestion strategies. You will design and implement data pipelines using Python, pandas, and SQL to ensure efficient data storage and management.  
  
#### Tasks:  
1. \*\*Data Ingestion Strategies:\*\*  
 - \*\*Problem Statement 3:\*\* Design and implement a data ingestion pipeline that supports incremental data loading. Optimize storage by using data partitioning and indexing strategies. Implement logging and monitoring to track the performance and reliability of the ingestion process. Utilize Python, pandas, and SQL for implementation.  
 - \*\*Additional Complexity:\*\* Use parallel processing to enhance the data ingestion performance.  
  
2. \*\*Pass Completion Rate vs. Assists:\*\*  
 - \*\*Problem Statement 4:\*\* Analyze the relationship between pass completion rate and assists. Create a scatter plot and identify outliers using advanced outlier detection methods like DBSCAN or Isolation Forest. Plot a line of best fit and use regression analysis to model the relationship. Evaluate the model using appropriate metrics.  
 - \*\*Additional Complexity:\*\* Implement cross-validation to ensure the robustness of your regression model.  
  
### Month 7-9: Advanced Data Transformation and Warehousing  
\*\*Context:\*\*   
Now that you understand the business requirements better, it's time to work on advanced data transformations and create a data warehouse to store the transformed data.  
  
#### Tasks:  
1. \*\*Advanced Data Transformations:\*\*  
 - \*\*Problem Statement 5:\*\* Perform complex transformations on the dataset, including feature engineering to create new meaningful features. Implement additional strategies for data optimization, such as data normalization and dimensionality reduction.  
 - \*\*Additional Complexity:\*\* Use machine learning techniques for feature selection and extraction.  
  
2. \*\*Data Warehousing:\*\*  
 - \*\*Problem Statement 6:\*\* Design and implement a data warehouse schema using advanced SQL features like window functions and CTEs (Common Table Expressions). Store the transformed data efficiently and ensure it supports complex analytical queries. Implement data security and access control mechanisms.  
 - \*\*Additional Complexity:\*\* Optimize the data warehouse for performance using techniques like indexing, partitioning, and materialized views.  
  
### Month 10-12: Reporting and Visualization  
\*\*Context:\*\*   
With transformed data stored in the warehouse, your final task is to create meaningful reports and visualizations that will drive business decisions.  
  
#### Tasks:  
1. \*\*Team Goals Analysis:\*\*  
 - \*\*Problem Statement 7:\*\* Identify the team with the highest number of goals. Create a horizontal bar plot and a stacked bar chart. Perform a time series analysis to understand trends in goal scoring over the season. Identify the top goal scorer in that team and analyze their performance metrics over time.  
 - \*\*Additional Complexity:\*\* Use advanced visualization tools like Plotly or D3.js for interactive visualizations.  
  
2. \*\*Reporting and Visualization:\*\*  
 - \*\*Problem Statement 8:\*\* Develop interactive dashboards and visualizations using tools like Power BI, Tableau, or custom web applications using Dash or Streamlit. Create reports that provide insights into player performance, team strategies, and potential areas for improvement. Incorporate advanced analytics like clustering and predictive modeling to forecast future performance.  
 - \*\*Additional Complexity:\*\* Integrate real-time data feeds to update the dashboards dynamically.  
  
### Evaluation Criteria  
- Completeness and accuracy of the dataset.  
- Efficiency and performance of data pipelines.  
- Appropriateness and correctness of data transformations.  
- Insightfulness and relevance of the analysis.  
- Clarity and impact of the visualizations.