**Data Science 1**

Screenshots:

stats\_data

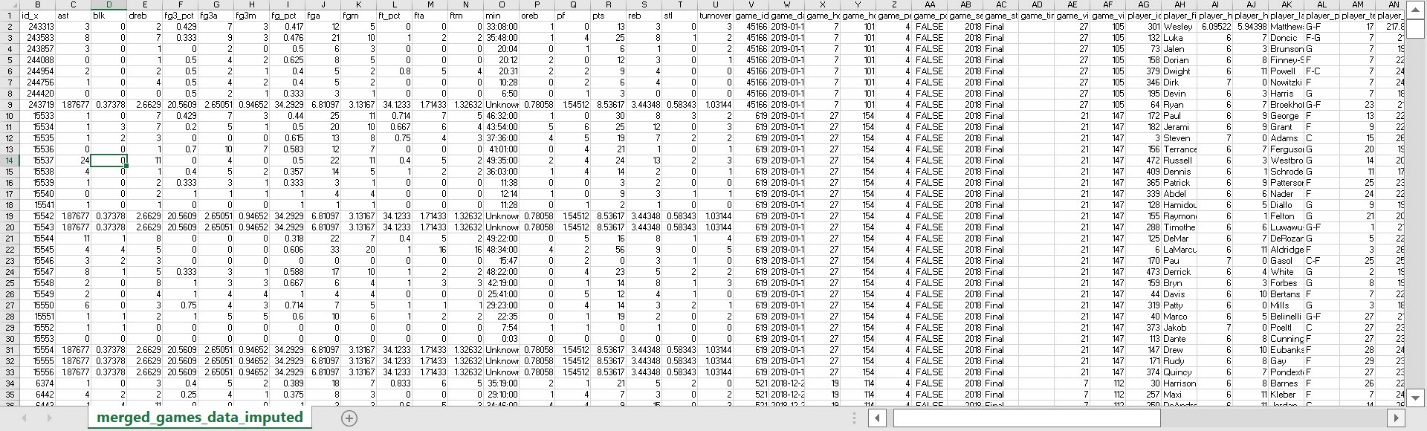


games\_data

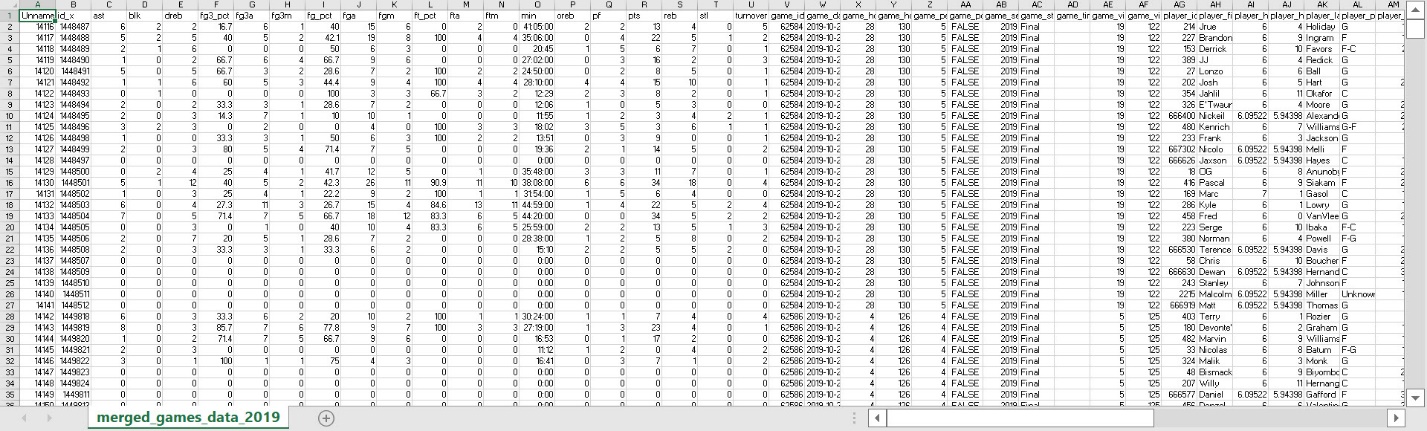
merged\_on\_8782



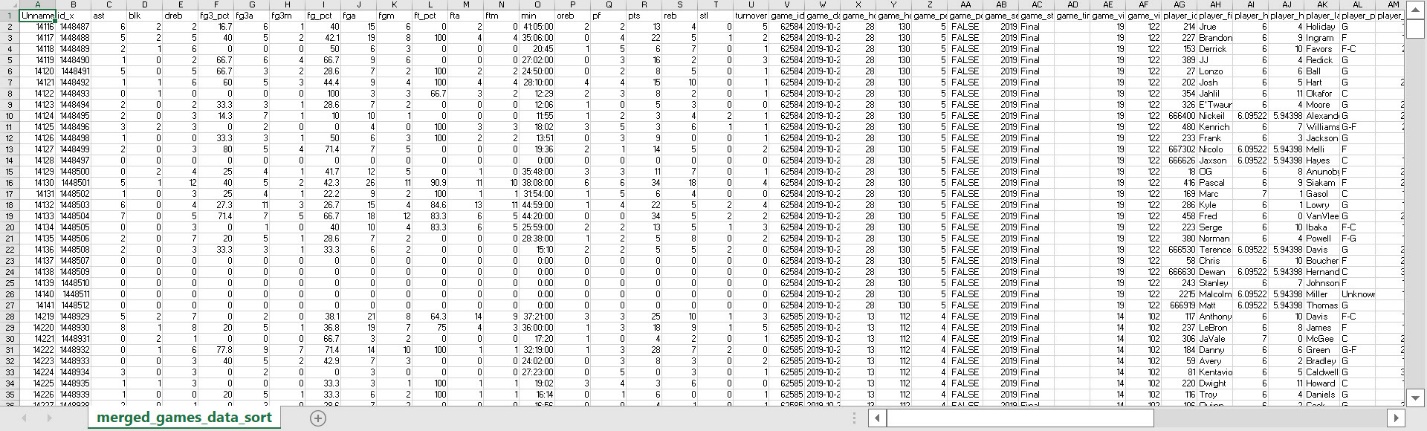
merged\_imputed



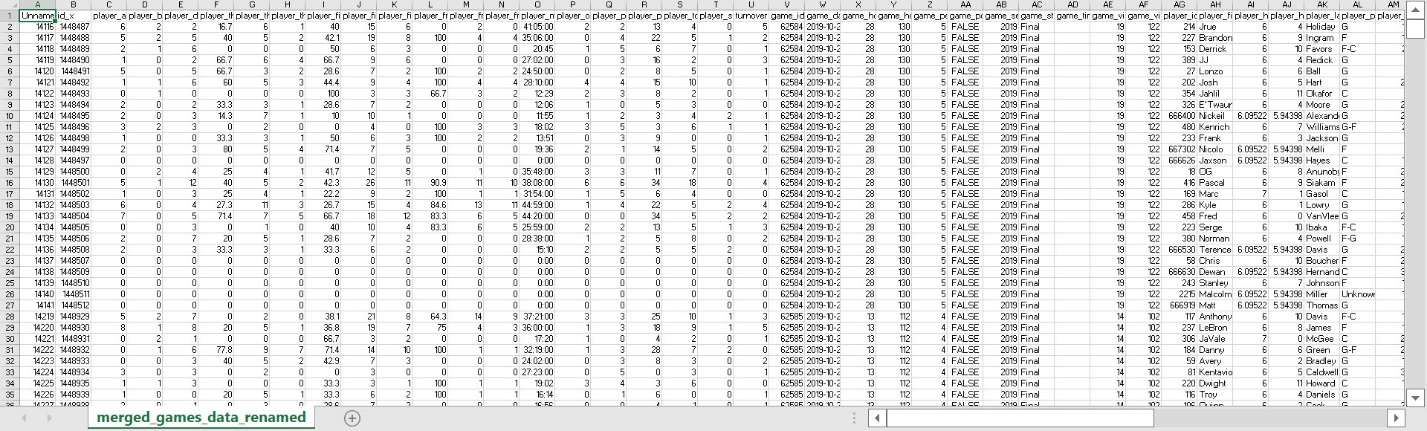
merged\_2019



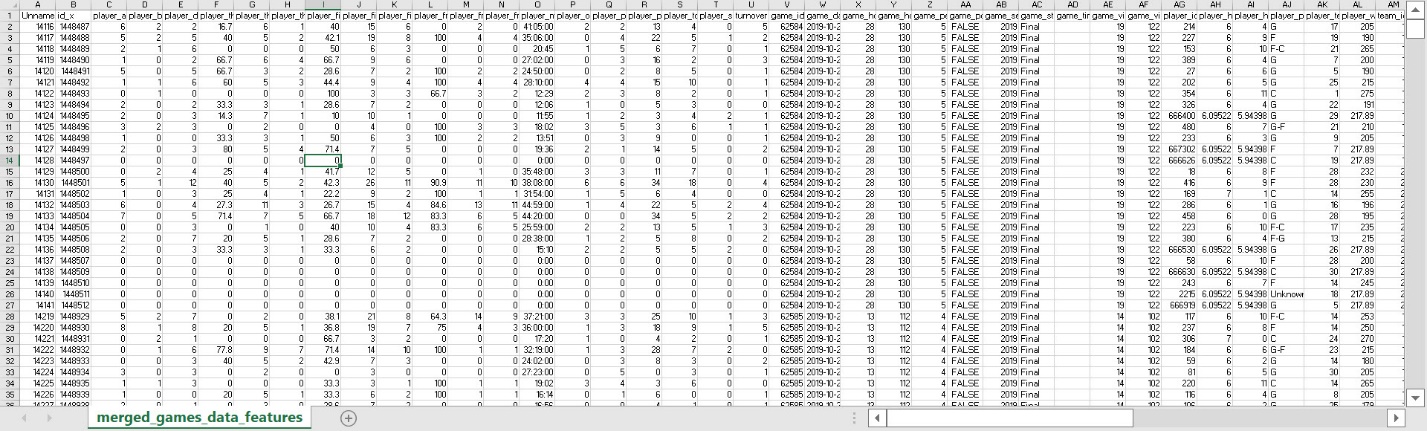
merged\_sorted



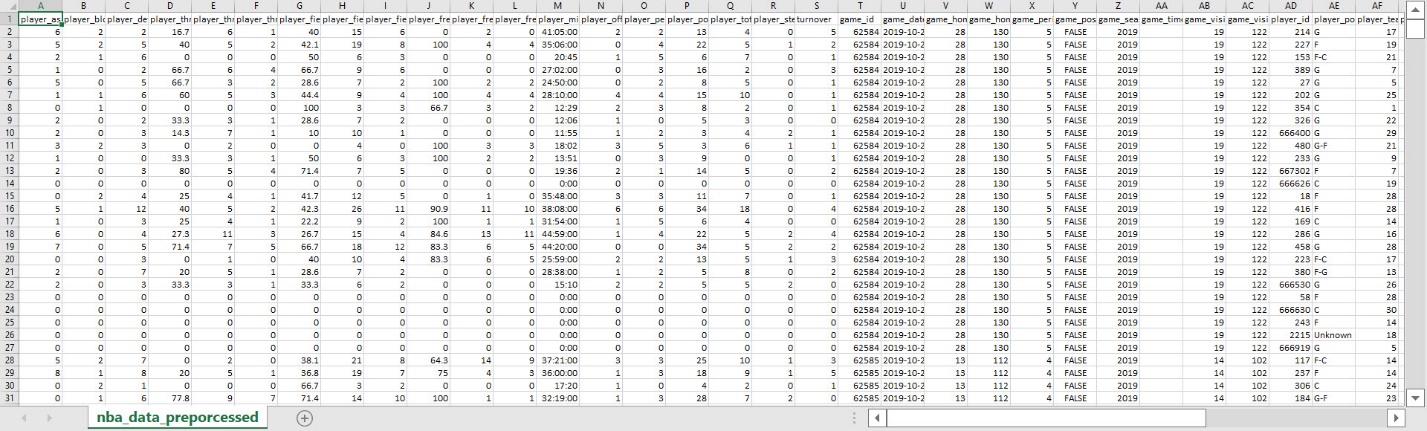
merged\_renamed



merged\_features



nba\_preprocessed



Headers in final data frame:

player\_assists, player\_blocks, player\_defensive\_rebounds, player\_three\_point\_pct, player\_three\_point\_attempts, player\_three\_point\_made, player\_field\_goal\_pct, player\_field\_goal\_attempts, player\_field\_goal\_made, player\_free\_throw\_pct, player\_free\_throw\_attempts, player\_free\_throw\_made, player\_minutes\_played, player\_offensive\_rebounds, player\_personal\_fouls, player\_points, player\_total\_rebounds, player\_steals, turnover, game\_id, game\_date, game\_home\_team\_id, game\_home\_team\_score, game\_period, game\_postseason, game\_season, game\_time, game\_visitor\_team\_id, game\_visitor\_team\_score, player\_id, player\_position, player\_team\_id, player\_weight\_pounds, team\_id, team\_division, team\_full\_name, home\_team\_conference, home\_team\_division, home\_team\_full\_name, visitor\_team\_conference, visitor\_team\_division, visitor\_team\_full\_name, team\_won, abs\_score\_difference, player\_full\_name, player\_height\_cm, home\_player\_stats, player\_efficiency, player\_three\_to\_throw\_attempt\_ratio, player\_three\_to\_throw\_made\_ratio

List of new features:

player\_efficiency: This feature represents the player's efficiency in the game. It is calculated by summing the player's points, total rebounds, assists, steals, and blocks, and then subtracting the missed field goal attempts, missed free throw attempts, and turnovers. A higher value indicates a more efficient performance by the player.

player\_three\_to\_throw\_attempt\_ratio: This feature represents the ratio of three-point attempts to total field goal attempts by the player. It is calculated by dividing the number of three-point attempts by the number of total field goal attempts. It provides insight into the player's preference for shooting three-pointers compared to other types of field goals.

player\_three\_to\_throw\_made\_ratio: This feature represents the ratio of three-point made shots to total field goals made by the player. It is calculated by dividing the number of three-point made shots by the number of total field goals made. It gives an indication of the player's accuracy and success rate in making three-point shots compared to other types of field goals.

These features provide additional insights into the player's overall performance, shooting tendencies, and efficiency on the court.

Answers to questions:

The team with the most wins is: ('Los Angeles Lakers', 46)

The team with the least wins is: ('Charlotte Hornets', 31)

Total LA Lakers points: 10297

Total Charlotte Hornets points: 6687

a. Supervised learning refers to the type of machine learning where the model is trained on labeled data, where the input data is paired with corresponding target labels. The goal is to learn a mapping function that can predict the labels for new, unseen data. Some use cases for supervised learning include email spam classification, sentiment analysis, and image recognition.

Unsupervised learning, on the other hand, involves training a model on unlabeled data without any predefined target labels. The model learns patterns, structures, or relationships within the data without specific guidance. Use cases for unsupervised learning include clustering similar documents, customer segmentation, and anomaly detection.

b. Structured data refers to data that is organized and well-defined, typically stored in databases or structured file formats. It has a fixed schema and follows a consistent format, making it easy to analyze using traditional database operations. Use cases for structured data include financial transactions, sales records, and sensor data from IoT devices.

Unstructured data, on the other hand, does not adhere to a specific schema or format. It includes text documents, images, audio files, videos, social media posts, etc. Analyzing unstructured data requires specialized techniques such as natural language processing, computer vision, or audio processing. Use cases for unstructured data include social media sentiment analysis, image recognition, and speech-to-text conversion.

c. The Cross Industry Standard Process for Data Mining (CRISP-DM) is a widely used data mining process model. The main steps of CRISP-DM are:

Business Understanding: Understand the project objectives, requirements, and constraints from a business perspective. Define the data mining problem and set goals.

Data Understanding: Gather and explore the available data to gain insights into its structure, content, and quality. Identify any data issues or limitations.

Data Preparation: Select, clean, and transform the data to create a suitable dataset for modeling. This involves tasks like handling missing values, encoding categorical variables, and normalizing data.

Modeling: Select appropriate modeling techniques and build a predictive or descriptive model based on the prepared dataset. This step includes model training, validation, and evaluation.

Evaluation: Assess the quality and effectiveness of the developed model. Evaluate its performance against the project objectives and requirements.

Deployment: Deploy the model into a production environment, integrating it with existing systems or processes. Monitor the model's performance and make necessary adjustments.

Some use cases include customer churn prediction, fraud detection, and market basket analysis.

d. A feature vector is a representation of an object or data point in a machine learning model. It is a numerical representation that captures the relevant characteristics or features of the data. Feature vectors are used as input to machine learning algorithms for training and making predictions.

In general, feature vectors can be used to represent various types of data, such as text documents (bag-of-words representation), images (pixel values or extracted features), or audio signals (MFCC coefficients). In the preprocessed data, feature vectors can be derived from the extracted features of the basketball game data, such as player statistics, team performance metrics, or game-related attributes.

Some use cases of feature vectors include sentiment analysis of text, image classification, and speech recognition.

e. Feature engineering is the process of creating new features or transforming existing features to improve the performance of machine learning models. It involves selecting, combining, or deriving features that capture the most relevant information from the data.

In general, feature engineering aims to enhance the predictive power of the model, reduce overfitting, and improve generalization. It can involve techniques like feature scaling, one-hot encoding, feature extraction, or feature selection. In the preprocessed data, feature engineering could include creating derived features like player efficiency, three-point-to-field-goal attempt ratio, or any other