NBA PLAYER STATISTICAL ANALYSIS AND PREDICTION PROJECT

COLLABORATORS:

JACOB EVANS
KARAN ANAND
MERT OZTOP
PRATIK PUROHIT
PRIYA MARINGANTI

OVERVIEW

This project aims to develop a basketball player statistics analysis and prediction system using machine learning techniques with NBA player datasets. The system utilizes historical player data to provide insights into player performance and predict player stats for the upcoming season, by leveraging the power of machine learning algorithms and the comprehensive NBA player statistics dataset.

Objectives

- Predict basketball player stats for the upcoming season based on historical data
- Leverage player stats from the previous year to forecast player performance
- Identify suitable machine learning algorithms for scalability, accuracy, and interpretability in predicting player performance
- Evaluate the performance of the prediction system and ensure its effectiveness in real-world scenarios, such as team selection and player scouting

METHODOLOGY

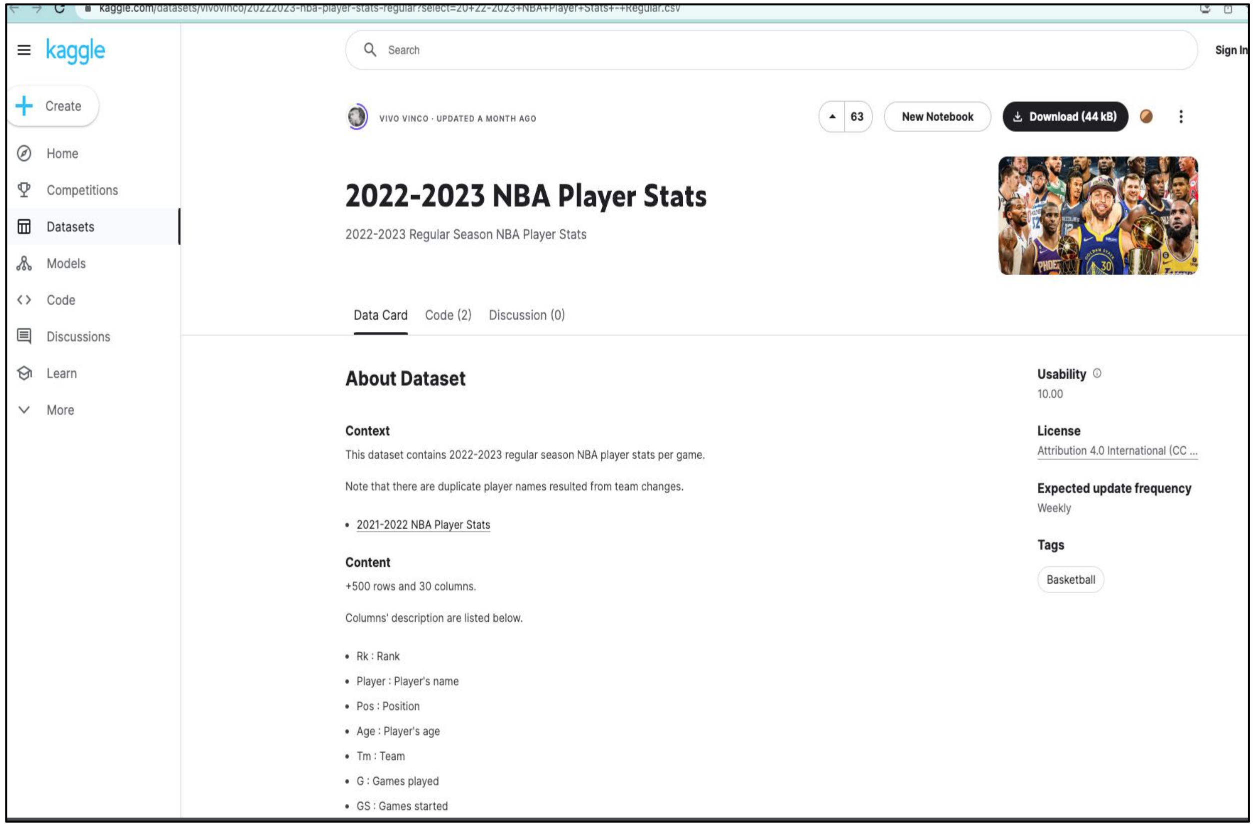
Methodology

1. Data Collection: Gather a large dataset of NBA player statistics from the Kaggle dataset
2. Data Preprocessing: Clean, normalize, encode, and engineer features from the NBA dataset
3. Exploratory Data Analysis: Gain insights, patterns, and analyze feature distributions
4. Model Training and Evaluation: Experiment with various machine learning algorithms, fine-tune models for accurate predictions
5. User Interface Development: Create an intuitive interface for users to input player data and view predictions
6. Testing and Validation: Ensure the accuracy, robustness, and scalability of the prediction system

SOURCE OF DATA

1. [Kaggle (2022-23)]
2. [Kaggle (2021-22)]
3. [Loodibee (Team Logos)]
4. [NBA (Head Shots)]

DATA SOURCE - KAGGLE



VIVO VINCO · UPDATED A MONTH AGO

2022-2023 NBA Player Stats

2022-2023 Regular Season NBA Player Stats

About Dataset

Context

This dataset contains 2022-2023 regular season NBA player stats per game.

Note that there are duplicate player names resulted from team changes.

- 2021-2022 NBA Player Stats

Content

+500 rows and 30 columns.

Columns' description are listed below.

- Rk : Rank
- Player : Player's name
- Pos : Position
- Age : Player's age
- Tm : Team
- G : Games played
- GS : Games started

Usability

10.00

License

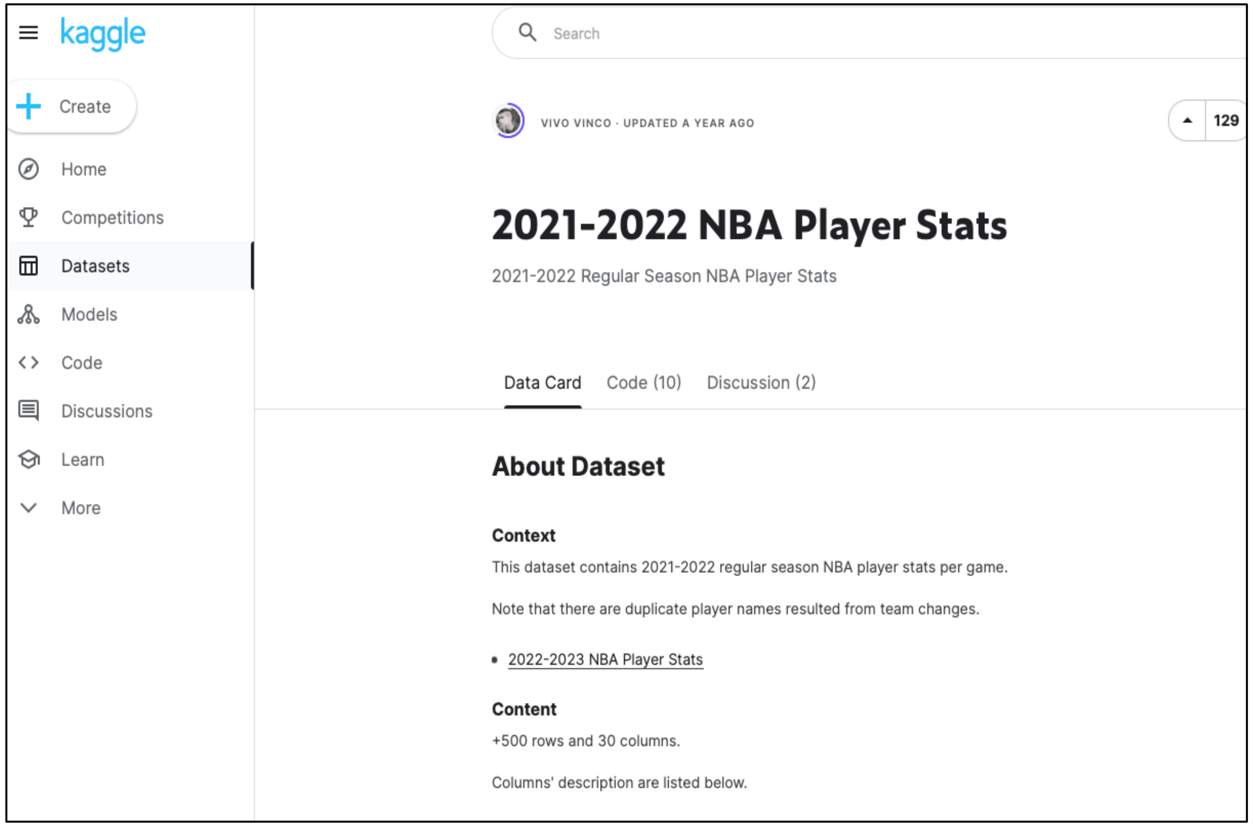
Attribution 4.0 International (CC ...

Expected update frequency

Weekly

Tags

Basketball



VIVO VINCO · UPDATED A YEAR AGO

2021-2022 NBA Player Stats

2021-2022 Regular Season NBA Player Stats

About Dataset

Context

This dataset contains 2021-2022 regular season NBA player stats per game.

Note that there are duplicate player names resulted from team changes.

- 2022-2023 NBA Player Stats

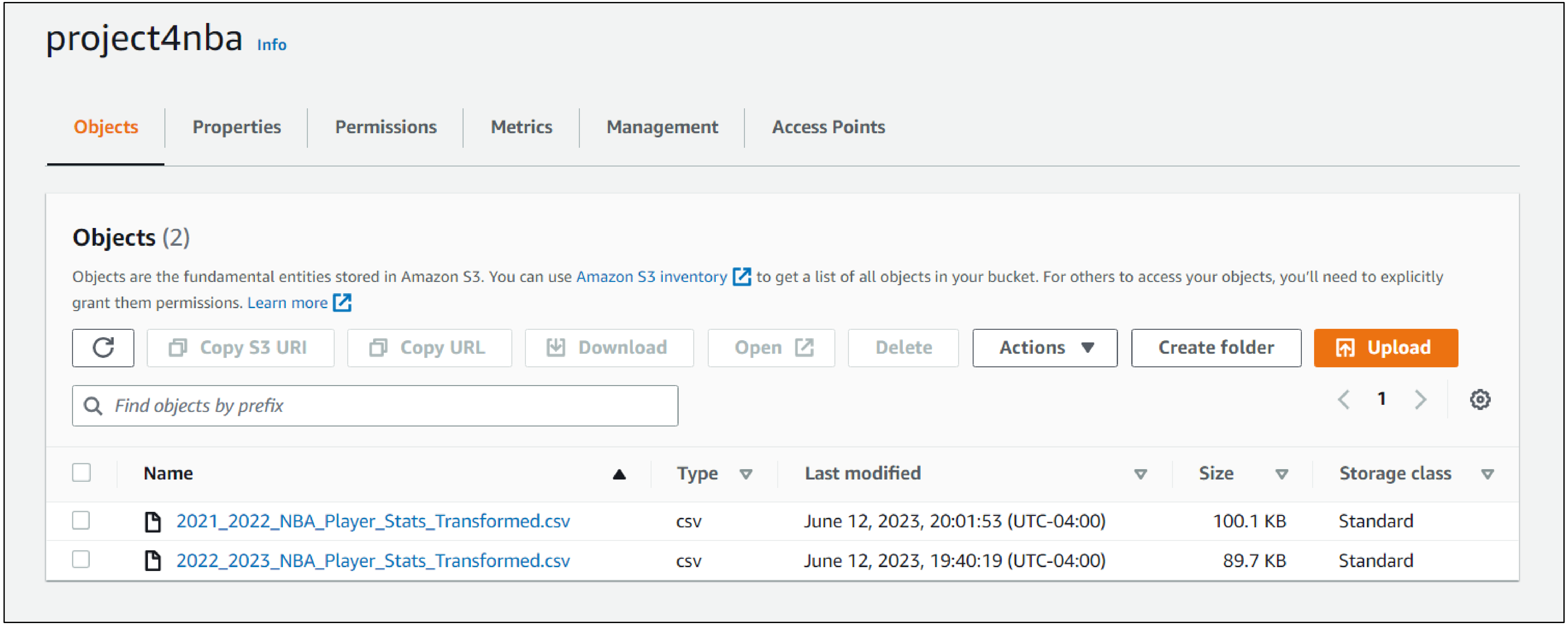
Content

+500 rows and 30 columns.

Columns' description are listed below.

Retrieving last two seasons’ data from Kaggle

DATA ON AWS



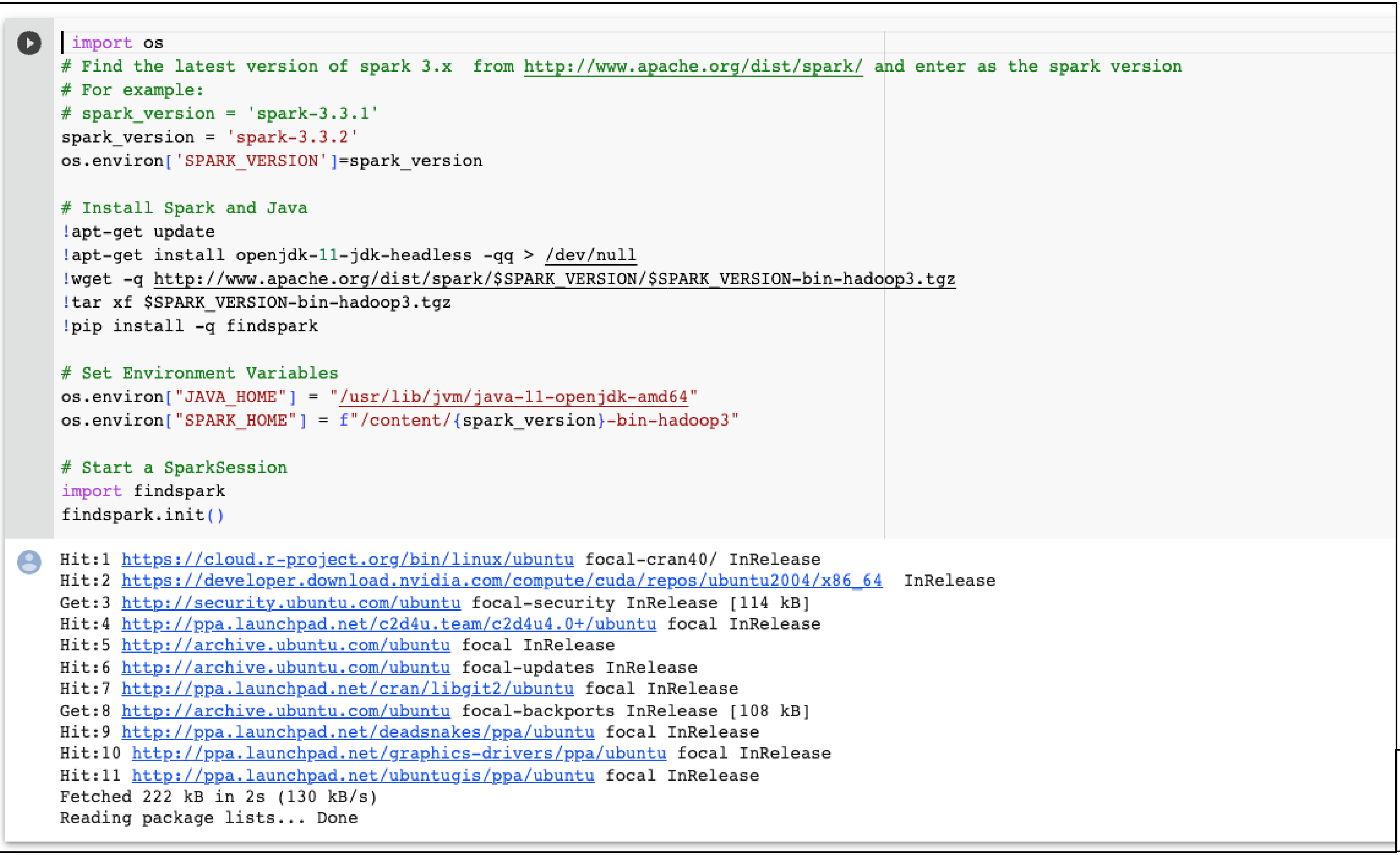
Objects (2)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

	Name	Type	Last modified	Size	Storage class
□	2021_2022_NBA_Player_Stats_Transformed.csv	csv	June 12, 2023, 20:01:53 (UTC-04:00)	100.1 KB	Standard
□	2022_2023_NBA_Player_Stats_Transformed.csv	csv	June 12, 2023, 19:40:19 (UTC-04:00)	89.7 KB	Standard

Uploading the data sets on AWS for the main code

SETTING UP ENVIRONMENT

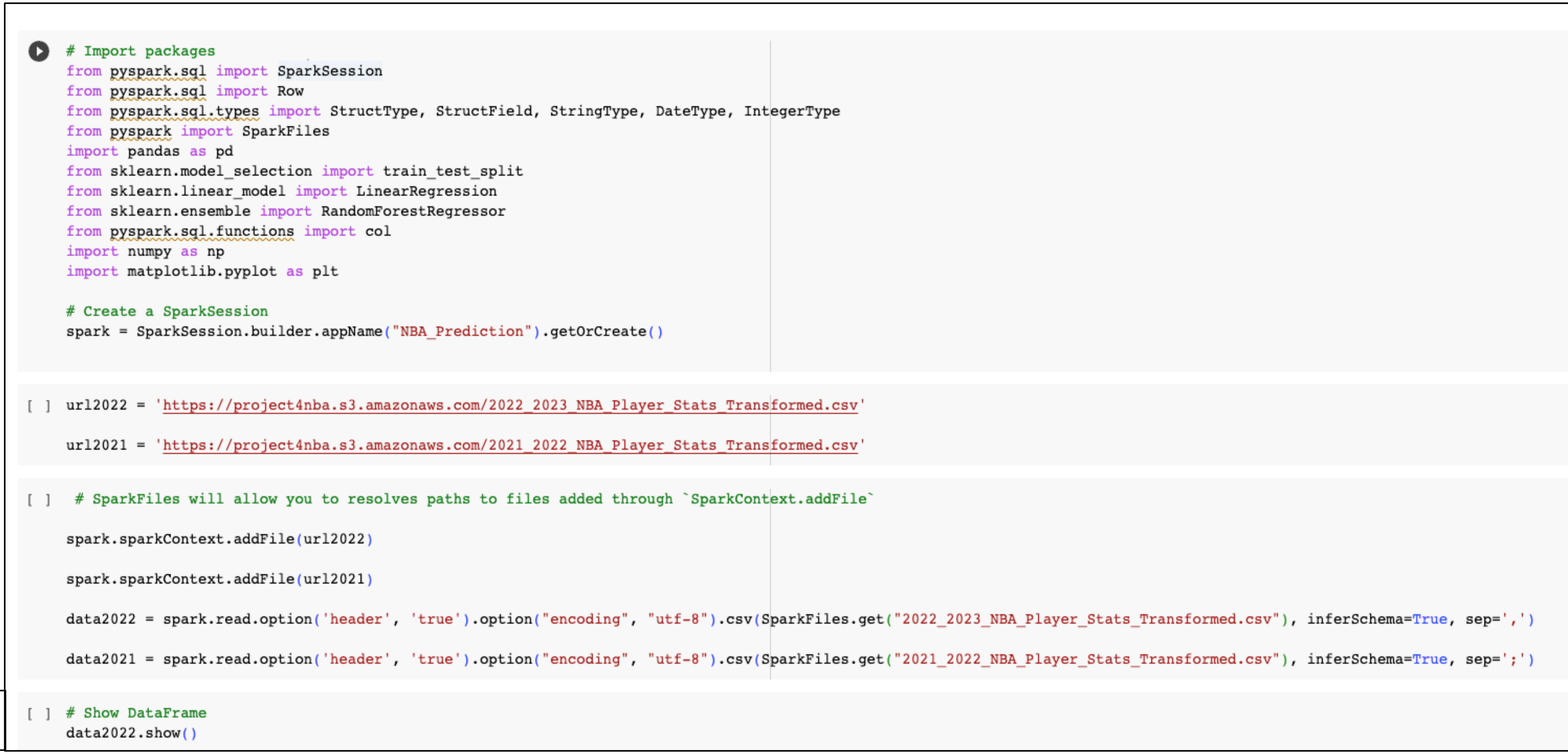


```
import os
# Find the latest version of spark 3.x  from http://www.apache.org/dist/spark/ and enter as the spark version
# For example:
# spark_version = 'spark-3.3.1'
spark_version = 'spark-3.3.2'
os.environ['SPARK_VERSION']=spark_version

# Install Spark and Java
!apt-get update
!apt-get install openjdk-11-jdk-headless -qq > /dev/null
!wget -q http://www.apache.org/dist/spark/$SPARK_VERSION/$SPARK_VERSION-bin-hadoop3.tgz
!tar xf $SPARK_VERSION-bin-hadoop3.tgz
!pip install -q findspark

# Set Environment Variables
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-11-openjdk-amd64"
os.environ["SPARK_HOME"] = f"/content/{spark_version}-bin-hadoop3"

# Start a SparkSession
import findspark
findspark.init()
```

```
Hit:1 https://cloud.r-project.org/bin/linux/ubuntu focal-cran40/ InRelease
Hit:2 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2004/x86_64  InRelease
Get:3 http://security.ubuntu.com/ubuntu focal-security InRelease [114 kB]
Hit:4 http://ppa.launchpad.net/c2d4u.team/c2d4u4.0+/ubuntu focal InRelease
Hit:5 http://archive.ubuntu.com/ubuntu focal InRelease
Hit:6 http://archive.ubuntu.com/ubuntu focal-updates InRelease
Hit:7 http://ppa.launchpad.net/cran/libgit2/ubuntu focal InRelease
Get:8 http://archive.ubuntu.com/ubuntu focal-backports InRelease [108 kB]
Hit:9 http://ppa.launchpad.net/deadsnakes/ppa/ubuntu focal InRelease
Hit:10 http://ppa.launchpad.net/graphics-drivers/ppa/ubuntu focal InRelease
Hit:11 http://ppa.launchpad.net/ubuntugis/ppa/ubuntu focal InRelease
Fetched 222 kB in 2s (130 kB/s)
Reading package lists... Done
```

Installed the necessary tools and environment of Spark and Java

CREATING SPARK SESSION



```
# Import packages
from pyspark.sql import SparkSession
from pyspark.sql import Row
from pyspark.sql.types import StructType, StructField, StringType, DateType, IntegerType
from pyspark import SparkFiles
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
from pyspark.sql.functions import col
import numpy as np
import matplotlib.pyplot as plt

# Create a SparkSession
spark = SparkSession.builder.appName("NBA_Prediction").getOrCreate()
```

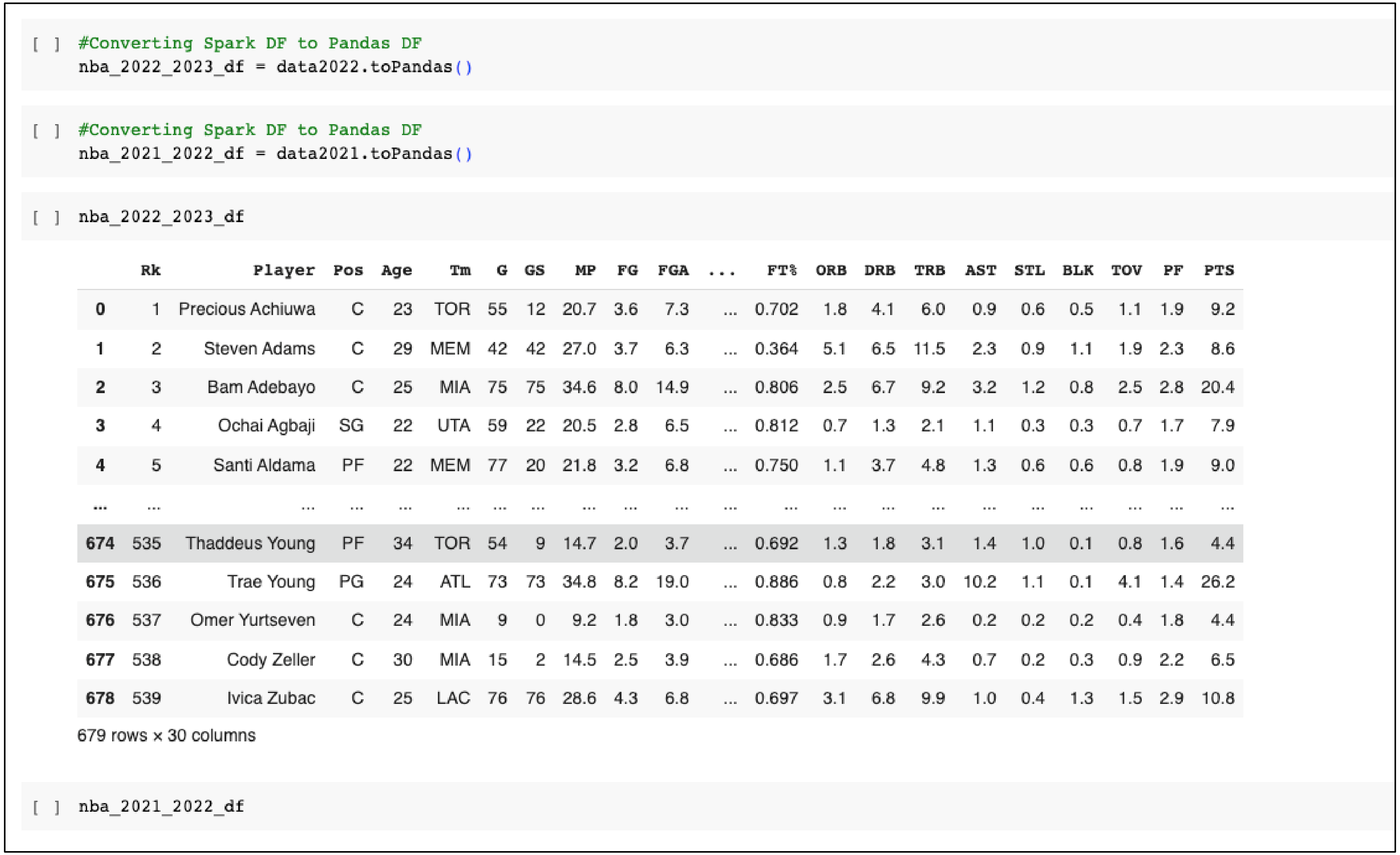
```
url2022 = 'https://project4nba.s3.amazonaws.com/2022_2023_NBA_Player_Stats_Transformed.csv'

url2021 = 'https://project4nba.s3.amazonaws.com/2021_2022_NBA_Player_Stats_Transformed.csv'
```

```
 # SparkFiles will allow you to resolves paths to files added through `SparkContext.addFile`

spark.sparkContext.addFile(url2022)

spark.sparkContext.addFile(url2021)

data2022 = spark.read.option('header', 'true').option("encoding", "utf-8").csv(SparkFiles.get("2022_2023_NBA_Player_Stats_Transformed.csv"), inferSchema=True, sep=',')

data2021 = spark.read.option('header', 'true').option("encoding", "utf-8").csv(SparkFiles.get("2021_2022_NBA_Player_Stats_Transformed.csv"), inferSchema=True, sep=';')
```

```
# Show DataFrame
data2022.show()
```

Created a Spark Session

SPARK DF TO PANDA DF



```
[ ] #Converting Spark DF to Pandas DF
    nba_2022_2023_df = data2022.toPandas()
```

```
[ ] #Converting Spark DF to Pandas DF
    nba_2021_2022_df = data2021.toPandas()
```

```
[ ] nba_2022_2023_df
```

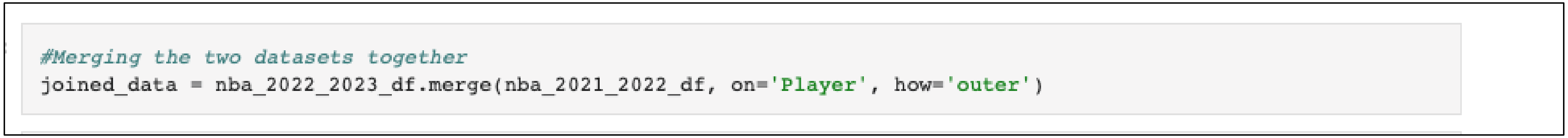
	Rk	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	...	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
0	1	Precious Achiuwa	C	23	TOR	55	12	20.7	3.6	7.3	...	0.702	1.8	4.1	6.0	0.9	0.6	0.5	1.1	1.9	9.2
1	2	Steven Adams	C	29	MEM	42	42	27.0	3.7	6.3	...	0.364	5.1	6.5	11.5	2.3	0.9	1.1	1.9	2.3	8.6
2	3	Bam Adebayo	C	25	MIA	75	75	34.6	8.0	14.9	...	0.806	2.5	6.7	9.2	3.2	1.2	0.8	2.5	2.8	20.4
3	4	Ochai Agbaji	SG	22	UTA	59	22	20.5	2.8	6.5	...	0.812	0.7	1.3	2.1	1.1	0.3	0.3	0.7	1.7	7.9
4	5	Santi Aldama	PF	22	MEM	77	20	21.8	3.2	6.8	...	0.750	1.1	3.7	4.8	1.3	0.6	0.6	0.8	1.9	9.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
674	535	Thaddeus Young	PF	34	TOR	54	9	14.7	2.0	3.7	...	0.692	1.3	1.8	3.1	1.4	1.0	0.1	0.8	1.6	4.4
675	536	Trae Young	PG	24	ATL	73	73	34.8	8.2	19.0	...	0.886	0.8	2.2	3.0	10.2	1.1	0.1	4.1	1.4	26.2
676	537	Omer Yurtseven	C	24	MIA	9	0	9.2	1.8	3.0	...	0.833	0.9	1.7	2.6	0.2	0.2	0.2	0.4	1.8	4.4
677	538	Cody Zeller	C	30	MIA	15	2	14.5	2.5	3.9	...	0.686	1.7	2.6	4.3	0.7	0.2	0.3	0.9	2.2	6.5
678	539	Ivica Zubac	C	25	LAC	76	76	28.6	4.3	6.8	...	0.697	3.1	6.8	9.9	1.0	0.4	1.3	1.5	2.9	10.8

679 rows × 30 columns

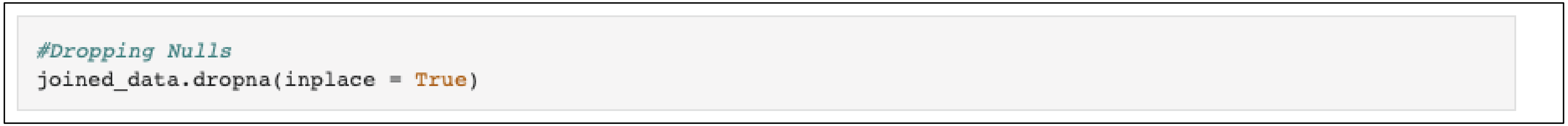
```
[ ] nba_2021_2022_df
```

Converted Spark DF to Panda DF

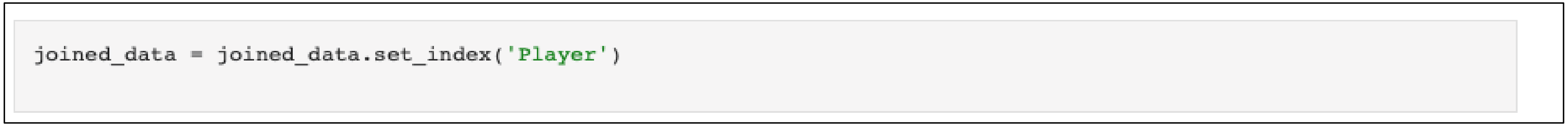
DATA CLEANING



```
#Merging the two datasets together
joined_data = nba_2022_2023_df.merge(nba_2021_2022_df, on='Player', how='outer')
```



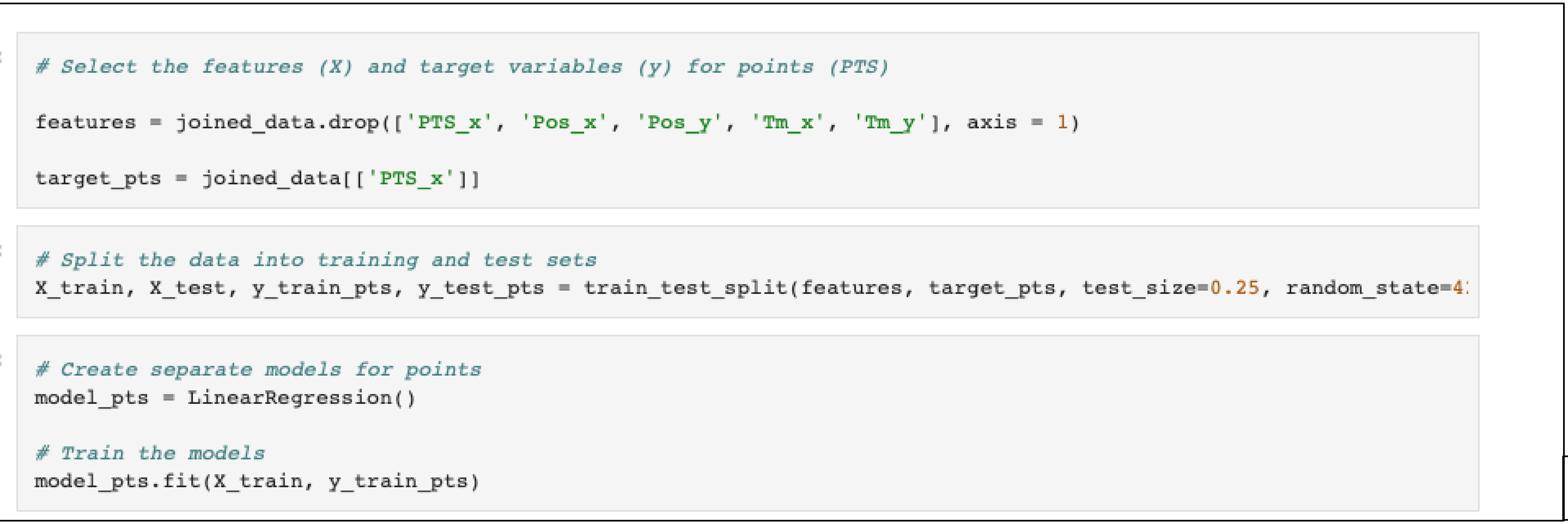
```
#Dropping Nulls
joined_data.dropna(inplace = True)
```



```
joined_data = joined_data.set_index('Player')
```

Cleaned the data

DATA MODELLING



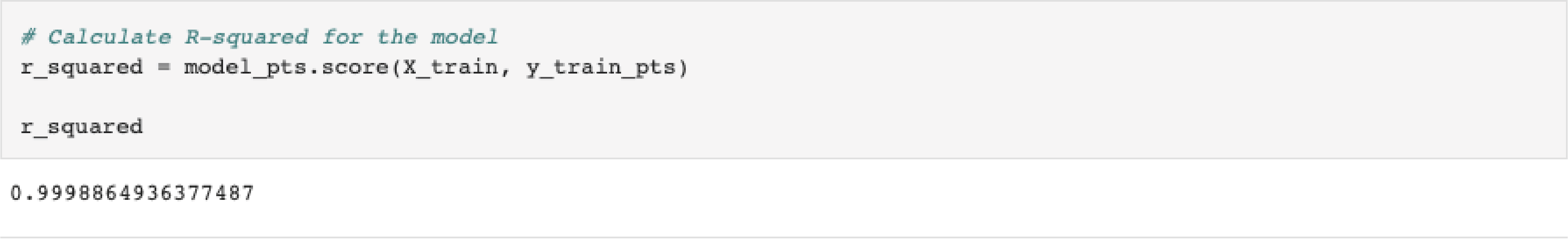
```
# Select the features (X) and target variables (y) for points (PTS)

features = joined_data.drop(['PTS_x', 'Pos_x', 'Pos_y', 'Tm_x', 'Tm_y'], axis = 1)

target_pts = joined_data[['PTS_x']]
```

```
# Split the data into training and test sets
X_train, X_test, y_train_pts, y_test_pts = train_test_split(features, target_pts, test_size=0.25, random_state=4
```

```
# Create separate models for points
model_pts = LinearRegression()

# Train the models
model_pts.fit(X_train, y_train_pts)
```

Modeled the data

DATA ACCURACY

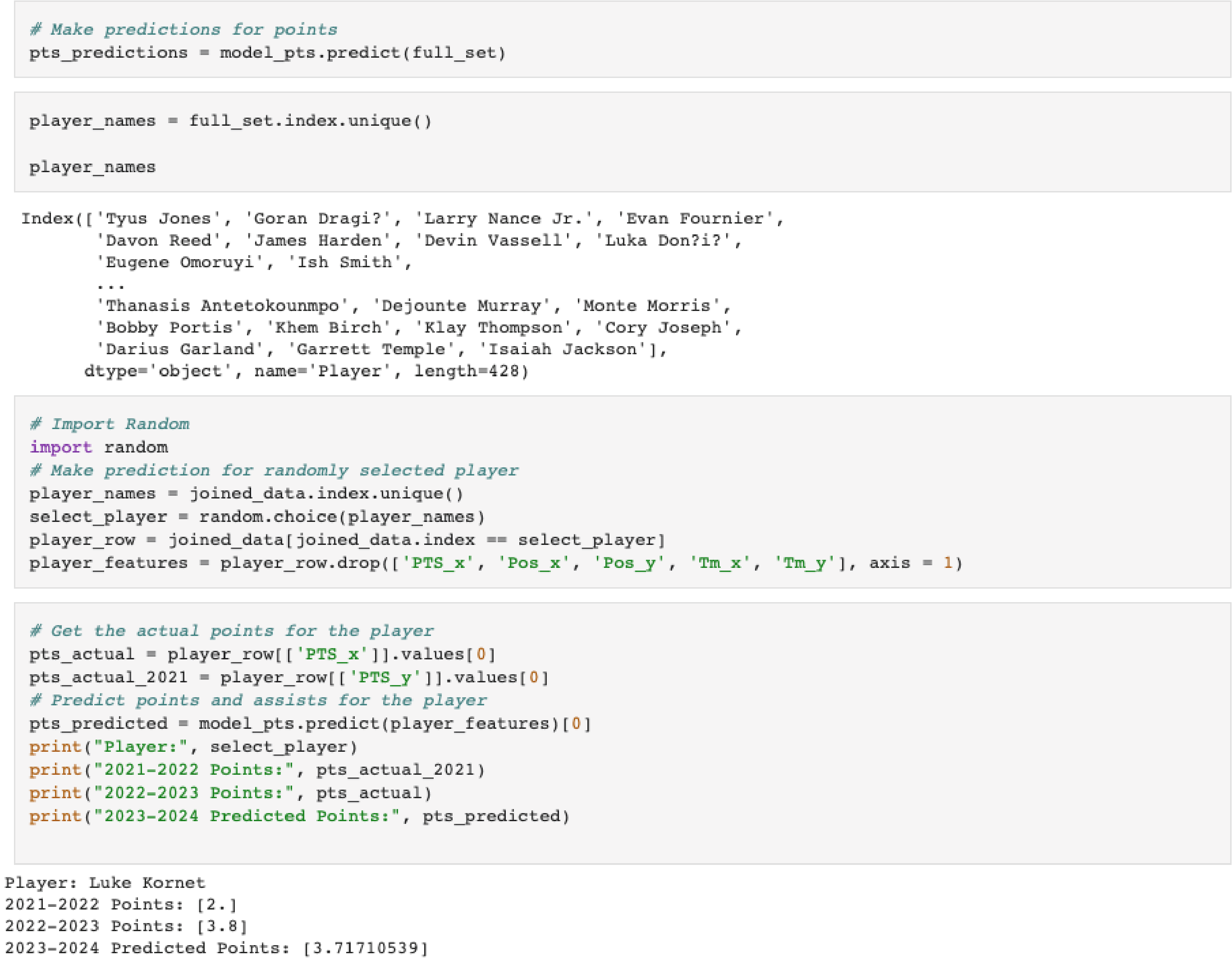


```
# Calculate R-squared for the model
r_squared = model_pts.score(X_train, y_train_pts)

r_squared
```

```
0.9998864936377487
```

Calculated the R-squared for the model

MAKING PREDICTIONS



```
# Make predictions for points
pts_predictions = model_pts.predict(full_set)
```

```
player_names = full_set.index.unique()

player_names
```

```
Index(['Tyus Jones', 'Goran Dragi?', 'Larry Nance Jr.', 'Evan Fournier',
       'Davon Reed', 'James Harden', 'Devin Vassell', 'Luka Don?i?',
       'Eugene Omoruyi', 'Ish Smith',
       ...
       'Thanasis Antetokounmpo', 'Dejounte Murray', 'Monte Morris',
       'Bobby Portis', 'Khem Birch', 'Klay Thompson', 'Cory Joseph',
       'Darius Garland', 'Garrett Temple', 'Isaiah Jackson'],
      dtype='object', name='Player', length=428)
```

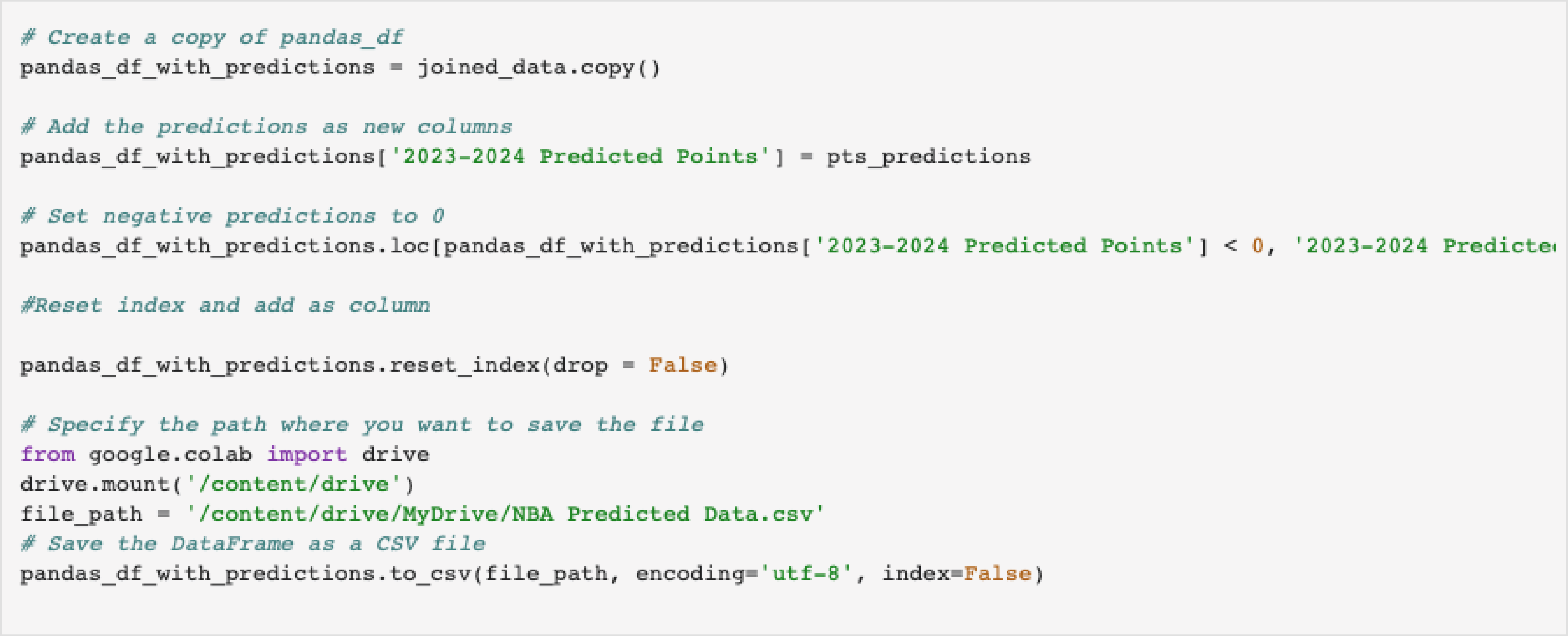
```
# Import Random
import random
# Make prediction for randomly selected player
player_names = joined_data.index.unique()
select_player = random.choice(player_names)
player_row = joined_data[joined_data.index == select_player]
player_features = player_row.drop(['PTS_x', 'Pos_x', 'Pos_y', 'Tm_x', 'Tm_y'], axis = 1)
```

```
# Get the actual points for the player
pts_actual = player_row[['PTS_x']].values[0]
pts_actual_2021 = player_row[['PTS_y']].values[0]
# Predict points and assists for the player
pts_predicted = model_pts.predict(player_features)[0]
print("Player:", select_player)
print("2021-2022 Points:", pts_actual_2021)
print("2022-2023 Points:", pts_actual)
print("2023-2024 Predicted Points:", pts_predicted)
```

```
Player: Luke Kornet
2021-2022 Points: [2.]
2022-2023 Points: [3.8]
2023-2024 Predicted Points: [3.71710539]
```

Made predictions for the points

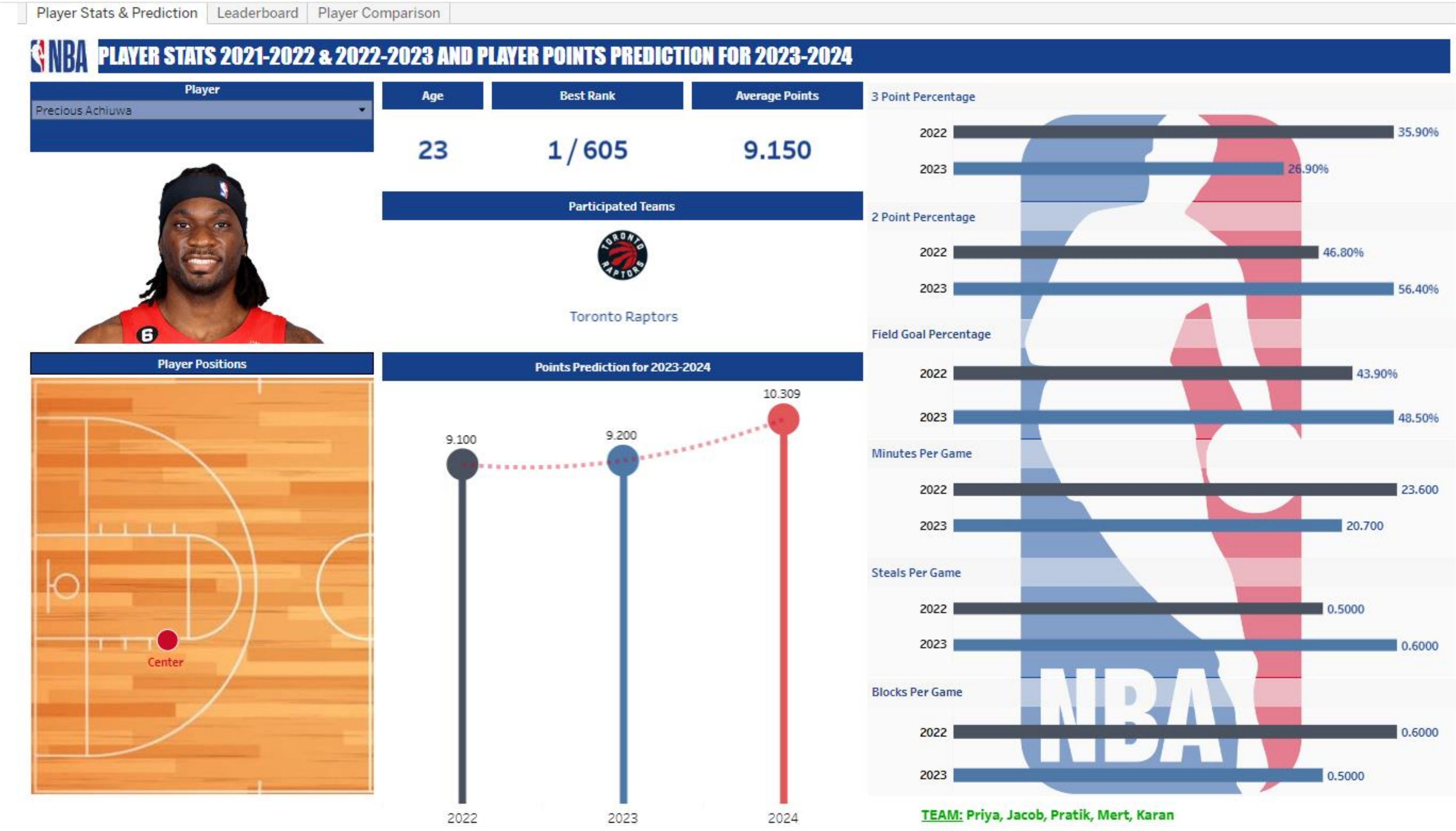
CREATING A FILE FOR TABLEAU



```
# Create a copy of pandas_df
pandas_df_with_predictions = joined_data.copy()

# Add the predictions as new columns
pandas_df_with_predictions['2023-2024 Predicted Points'] = pts_predictions

# Set negative predictions to 0
pandas_df_with_predictions.loc[pandas_df_with_predictions['2023-2024 Predicted Points'] < 0, '2023-2024 Predicte

#Reset index and add as column

pandas_df_with_predictions.reset_index(drop = False)

# Specify the path where you want to save the file
from google.colab import drive
drive.mount('/content/drive')
file_path = '/content/drive/MyDrive/NBA Predicted Data.csv'
# Save the DataFrame as a CSV file
pandas_df_with_predictions.to_csv(file_path, encoding='utf-8', index=False)
```

Downloaded the final data set for Tableau

TABLEAU DASHBOARD – PLAYER STATS AND PREDICTION



PLAYER STATS 2021-2022 & 2022-2023 AND PLAYER POINTS PREDICTION FOR 2023-2024

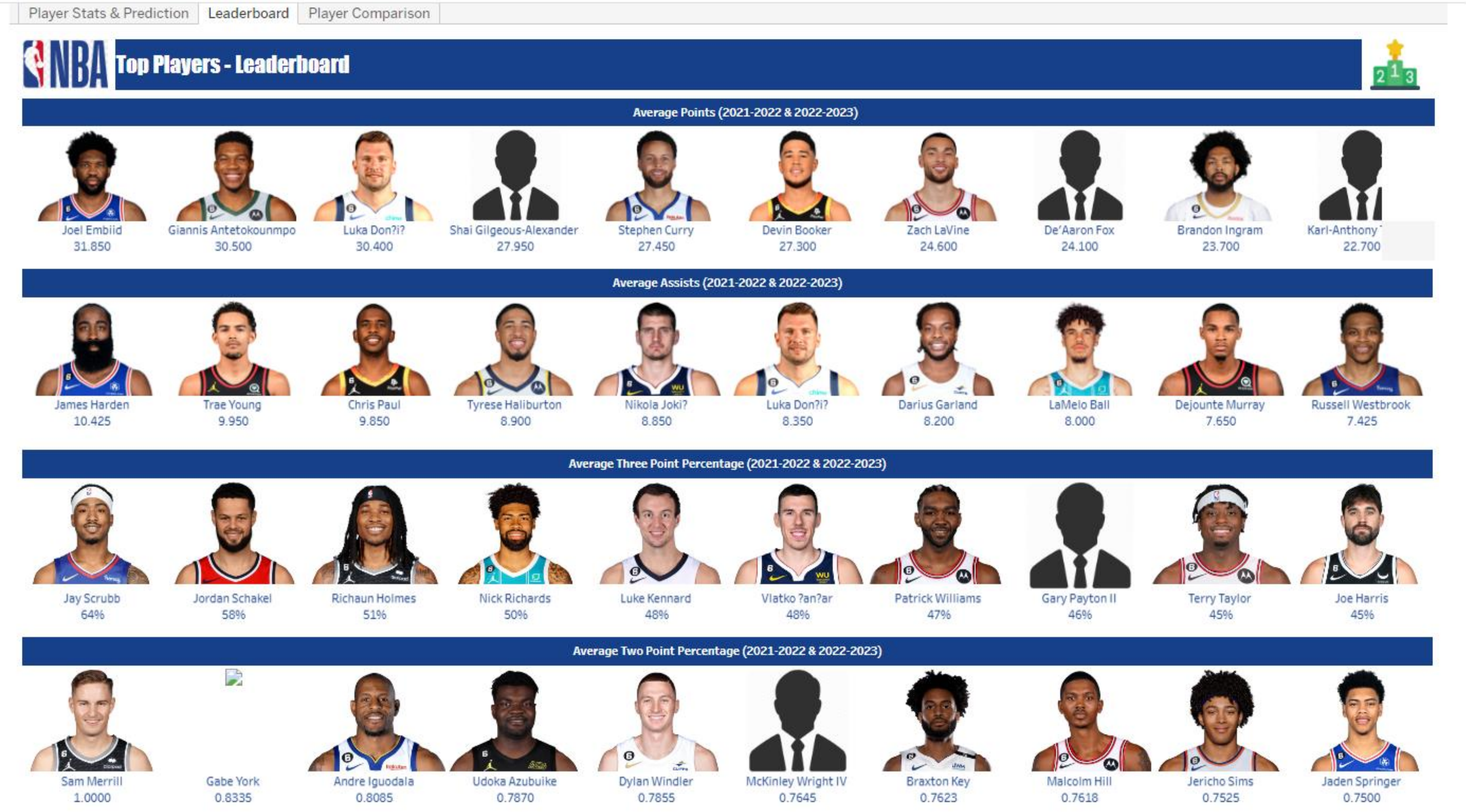
Player: Precious Achiuwa

Age	Best Rank	Average Points
23	1 / 605	9.150

Participated Teams: Toronto Raptors

TEAM: Priya, Jacob, Pratik, Mert, Karan

TABLEAU DASHBOARD – LEADERBOARD



Player Stats & Prediction | Leaderboard | Player Comparison

Top Players - Leaderboard

Average Points (2021-2022 & 2022-2023)

Joel Embiid	Giannis Antetokounmpo	Luka Don?i?	Shai Gilgeous-Alexander	Stephen Curry	Devin Booker	Zach LaVine	De'Aaron Fox	Brandon Ingram	Karl-Anthony
31.850	30.500	30.400	27.950	27.450	27.300	24.600	24.100	23.700	22.700

Average Assists (2021-2022 & 2022-2023)

James Harden	Trae Young	Chris Paul	Tyrese Haliburton	Nikola Joki?	Luka Don?i?	Darius Garland	LaMelo Ball	Dejounte Murray	Russell Westbrook
10.425	9.950	9.850	8.900	8.850	8.350	8.200	8.000	7.650	7.425

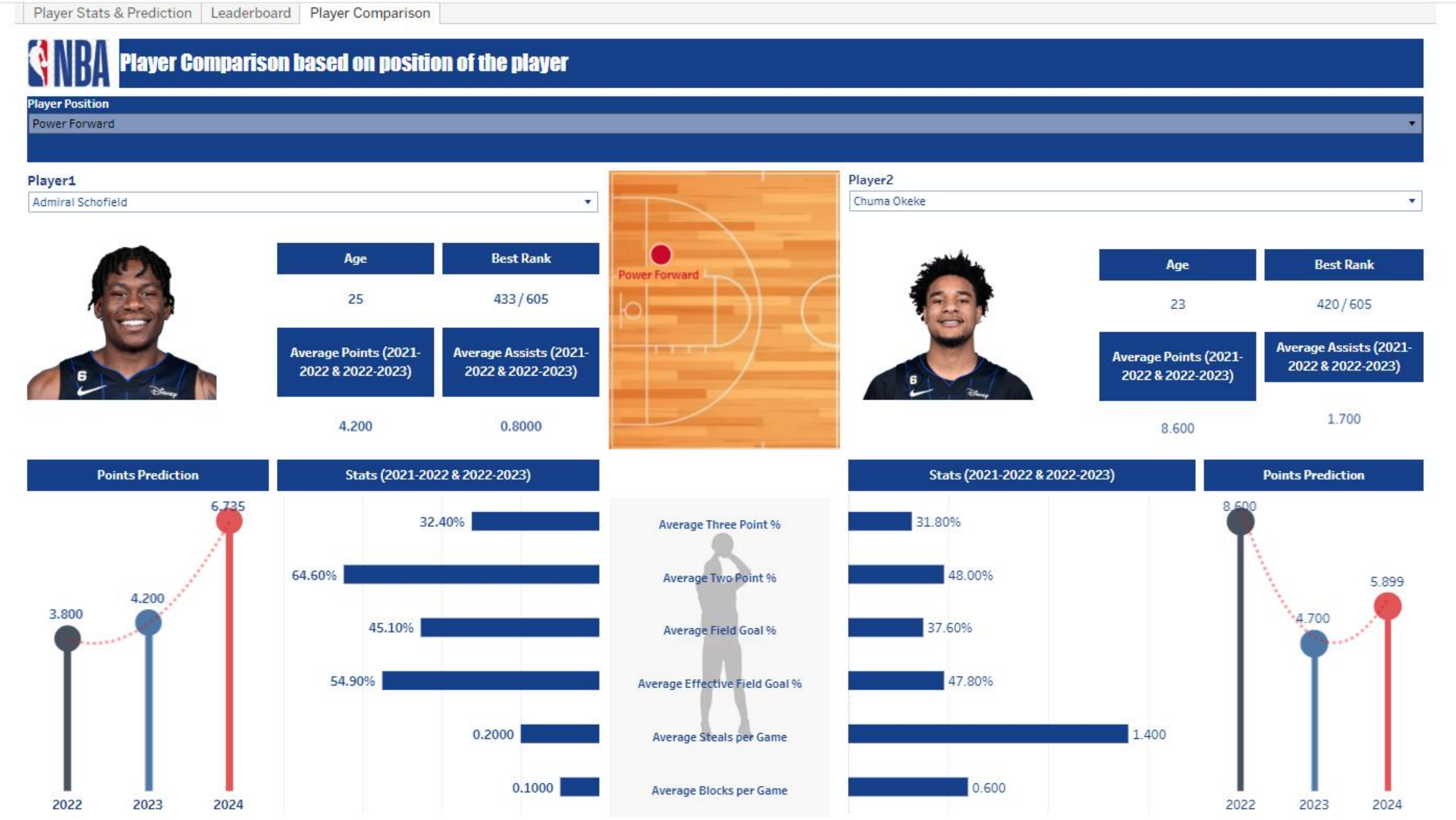
Average Three Point Percentage (2021-2022 & 2022-2023)

Jay Scrubb	Jordan Schakel	Richaun Holmes	Nick Richards	Luke Kennard	Vlatko ?an?ar	Patrick Williams	Gary Payton II	Terry Taylor	Joe Harris
64%	58%	51%	50%	48%	48%	47%	46%	45%	45%

Average Two Point Percentage (2021-2022 & 2022-2023)

Sam Merrill	Gabe York	Andre Iguodala	Udoka Azubuike	Dylan Windler	McKinley Wright IV	Braxton Key	Malcolm Hill	Jericho Sims	Jaden Springer
1.0000	0.8335	0.8085	0.7870	0.7855	0.7645	0.7623	0.7618	0.7525	0.7500

TABLEAU DASHBOARD – PLAYER COMPARISION



Player Comparison based on position of the player

Player Position: Power Forward

Player1: Admiral Schofield

Age	Best Rank
25	433 / 605

Average Points (2021-2022 & 2022-2023)	Average Assists (2021-2022 & 2022-2023)
4.200	0.8000

Player2: Chuma Okeke

Age	Best Rank
23	420 / 605

Average Points (2021-2022 & 2022-2023)	Average Assists (2021-2022 & 2022-2023)
8.600	1.700

Stats (2021-2022 & 2022-2023)		Stats (2021-2022 & 2022-2023)
32.40%	Average Three Point %	31.80%
64.60%	Average Two Point %	48.00%
45.10%	Average Field Goal %	37.60%
54.90%	Average Effective Field Goal %	47.80%
0.2000	Average Steals per Game	1.400
0.1000	Average Blocks per Game	0.600

CHALLENGES

- Bad Encoding: The original data had encoding issues that we struggled to handle. We had to apply encoding techniques to ensure proper handling and interpretation of the data.

- External Factors: While player statistics provide valuable insights, it's important to note that other factors can influence a player's performance on the court. Factors such as injuries, team dynamics, coaching strategies, and external circumstances were not included in our analysis. Considering these external factors could further enhance the accuracy and predictive power of the model.

- Outliers: To handle these outliers, we implemented a post-processing step where we replaced any negative predicted values with zeros. This approach allowed us to address the outliers and ensure that the predicted statistics remain within a valid range. By zeroing out the negative values, we mitigated the impact of outliers on the model's performance and ensured that the predicted player statistics align with the expectations of NBA player performance.

RESULT

The prediction system achieved **R-squared value of 0.9998**, indicating a high level of accuracy in predicting player statistics based on the historical data.

The NBA Player Statistics Analysis and Prediction System leverages historical player data, applies machine learning techniques, and provides valuable insights and predictions on player performance. The system can assist with team selection, player scouting, and forecasting player statistics for the upcoming season.

THANK YOU!

QUESTIONS ARE WELCOME!

COLLABORATORS:

JACOB EVANS
KARAN ANAND
MERT OZTOP
PRATIK PUROHIT
PRIYA MARINGANTI