NBA Games
Course Project - I535
Sonal Fall'23

1. INTRODUCTION

The National Basketball Association is a professional basketball league in North America composed of 30 teams. It is one of the major professional sports leagues in the United States and Canada and is considered the premier professional basketball league in the world.

The National Basketball Association was founded at the Commodore Hotel in New York. Maurice Podoloff was the league's first president, a title later changed to commissioner. Eleven teams were part of that league, originally called the Basketball Association of America. Podoloff's name is now emblazoned on the NBA MVP trophy.

2. BACKGROUND

The dataset, sourced from https://github.com/fivethirtyeight/data/tree/master/nba-forecasts , is named "nba_dataset.csv" and provides valuable insights into NBA games during various seasons. It contains various features, including date, season, neutral venue indicator, playoff status, teams (team1 and team2), pre-game Elo ratings (elo1_pre, elo2_pre), Elo win probabilities (elo_prob1, elo_prob2), post-game Elo ratings (elo1_post, elo2_post), pre-game CARMELO ratings (carm-elo1_pre, carm-elo2_pre), CARMELO win probabilities (carm-elo_prob1, carm-elo_prob2), post-game CARMELO ratings (carm-elo1_post, carm-elo2_post), pre-game RAPTOR ratings (raptor1_pre, raptor2_pre), RAPTOR win probabilities (raptor_prob1, raptor_prob2), final scores (score1, score2), game quality, game importance, and a total rating.

The dataset consists of a detailed game-related information, including team performance assessments, win probabilities based on different rating systems (Elo, CARMELO, RAPTOR), and game-specific metrics like scores, quality, and importance.

Using the NBA dataset I wanted to explore What are the average scores for each team? Which teams have the highest average Elo ratings? What is the distribution of average scores for each team? How have the Elo ratings changed over time for specific teams? Can we predict the score difference between two teams based on pre-game Elo ratings?

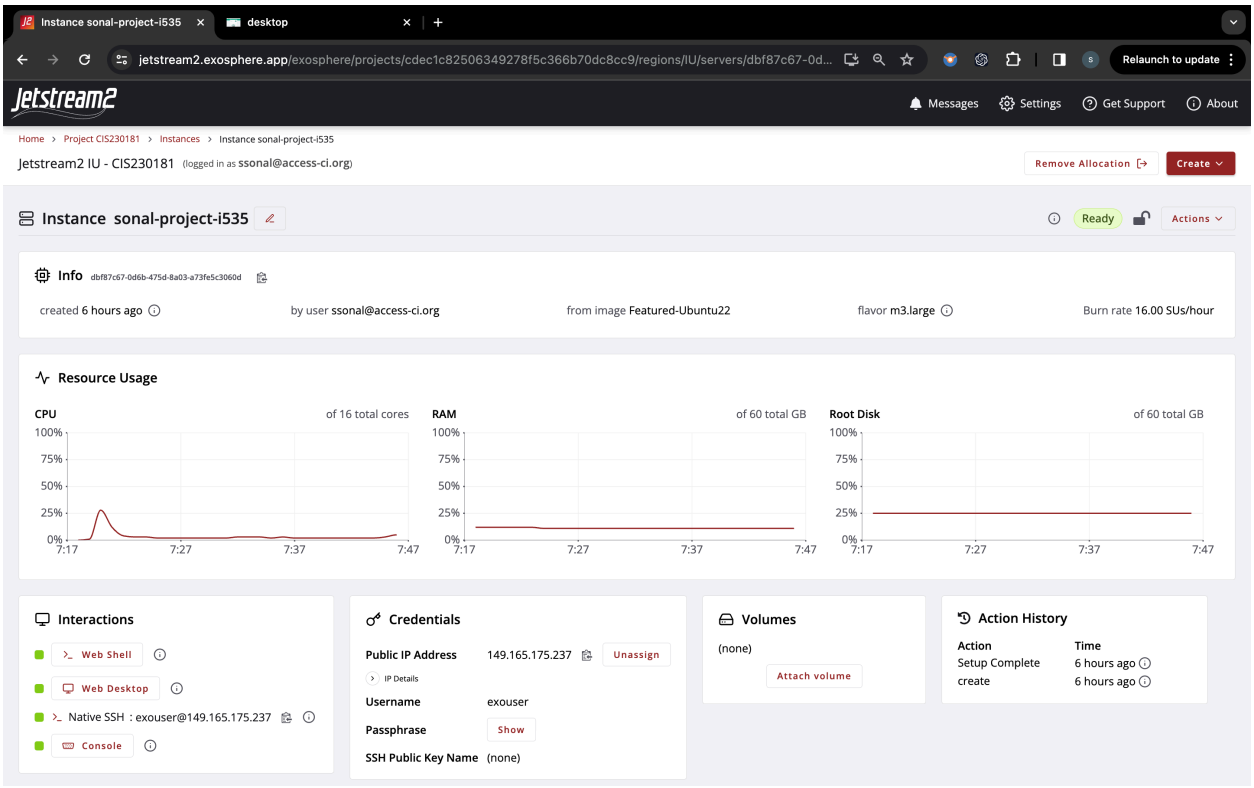
How well does the linear regression model perform in predicting score differences with a limited dataset?

3. METHODOLOGY

For this project, I am using Jetstream in which I have created a new instance named "sonal-project-i535" with the following parameters:

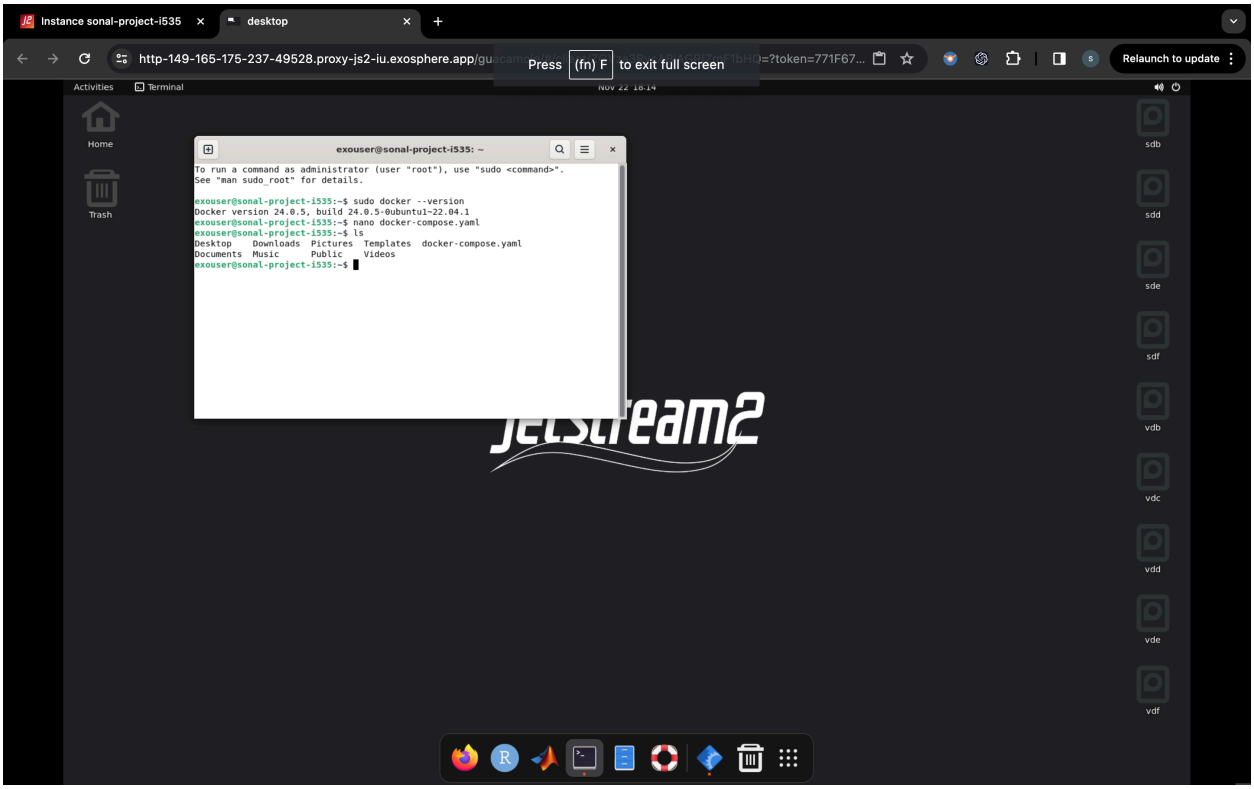
- Ubuntu 22.04(latest)
- m3.large
- Web Desktop enabled

Here I have used my knowledge from **Module 4: Virtualization ("Complete – J2 Virtual Machines")**



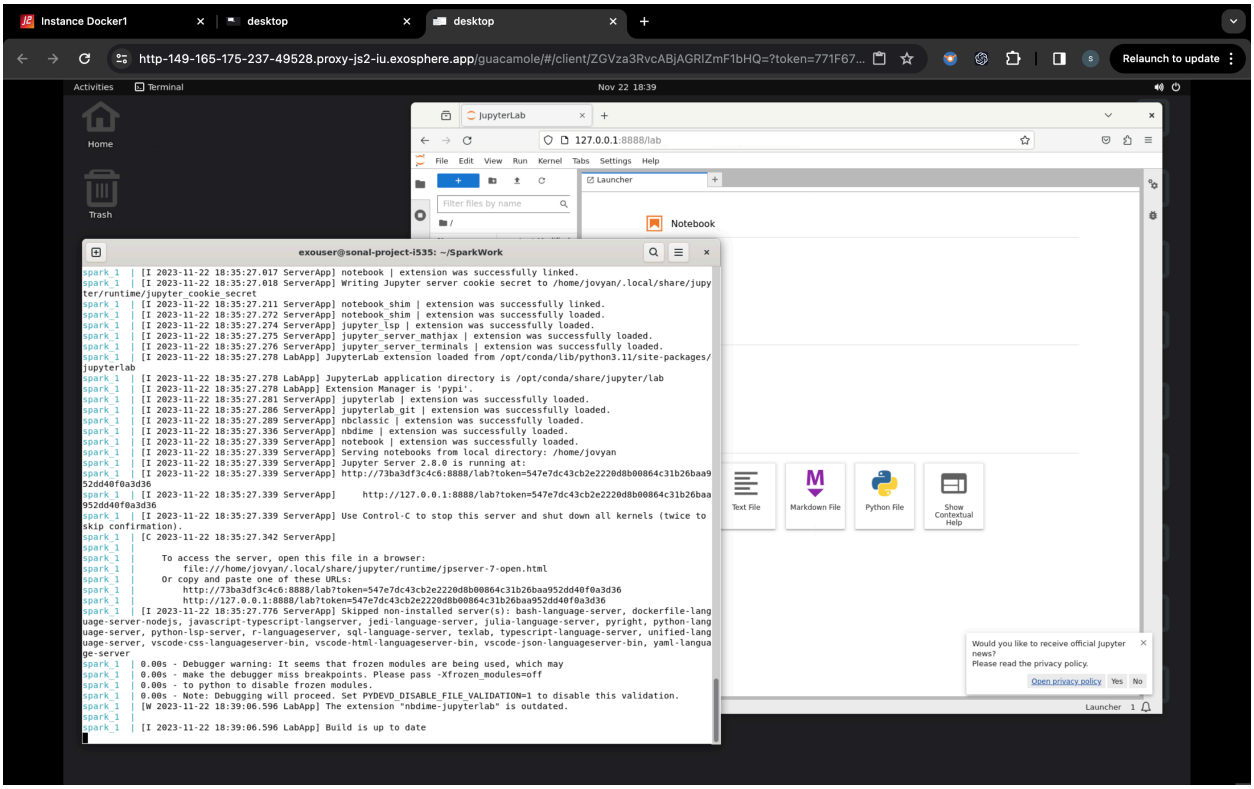
Jetstream is an outstanding choice for your cloud computing needs, especially for researchers and students using the Extreme Science and Engineering Discovery Environment (XSEDE). Designed to provide convenient on-demand access to computing

and data analysis resources, Jetstream shares similarities with Google Cloud Platform but stands out with its user-friendly interface. It has an easy to understand graphical user interface (GUI), which simplifies complex tasks, making them more accessible for users.

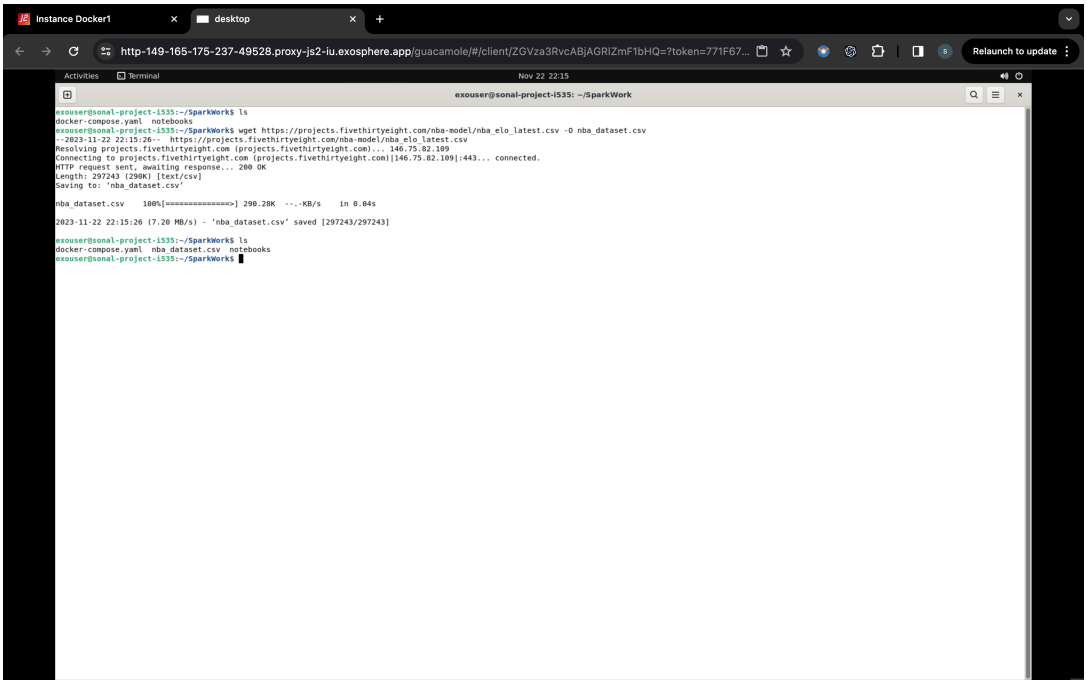


This emphasis on user-friendliness makes it an excellent option for interactive research on a smaller scale.

And then, after creating the instance for my project I have used the web desktop mode and utilizing my knowledge from **Module 9: Processing and Analytics (Complete - Analyzing data with PySpark),** I have created a separate folder “SparkWork” and further created a docker-compose.yaml file and updated it using nano editor to create and start the container.



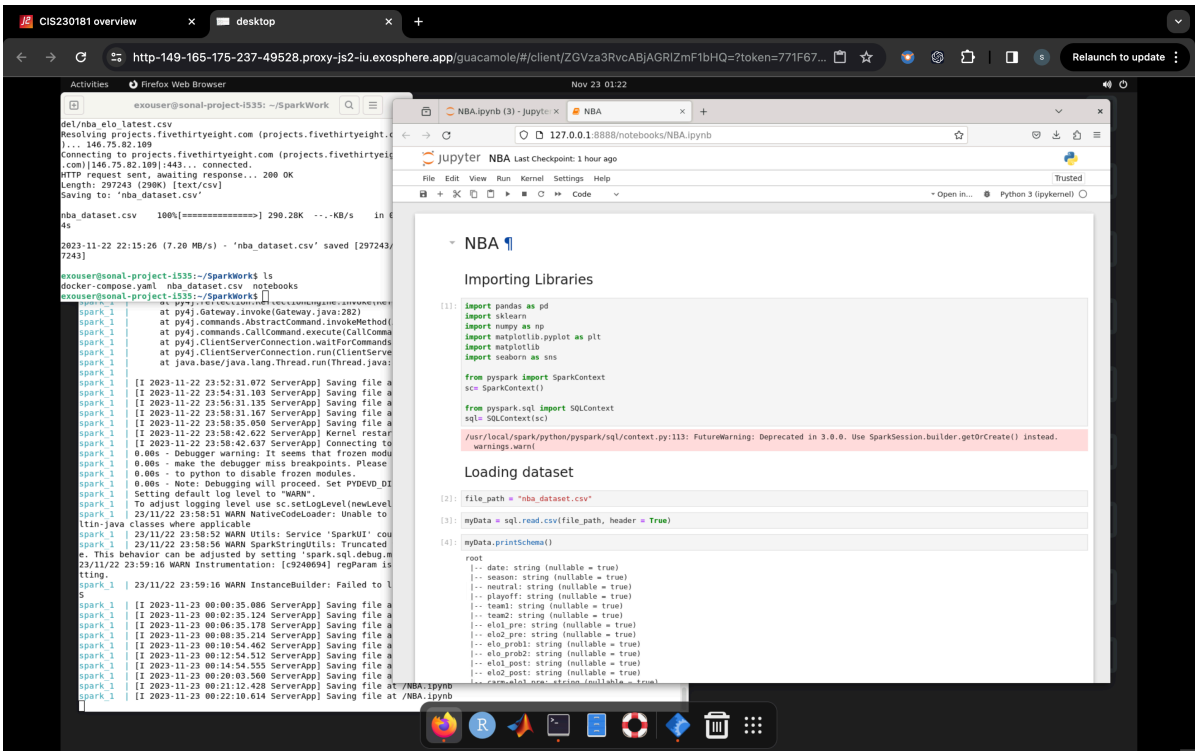
Next, I have run the following command to extract the css version of the NBA dataset provided by “FiveThirtyEight”:
“wget https://projects.fivethirtyeight.com/nba-model/nba_elo_latest.csv -O nba_dataset.csv”



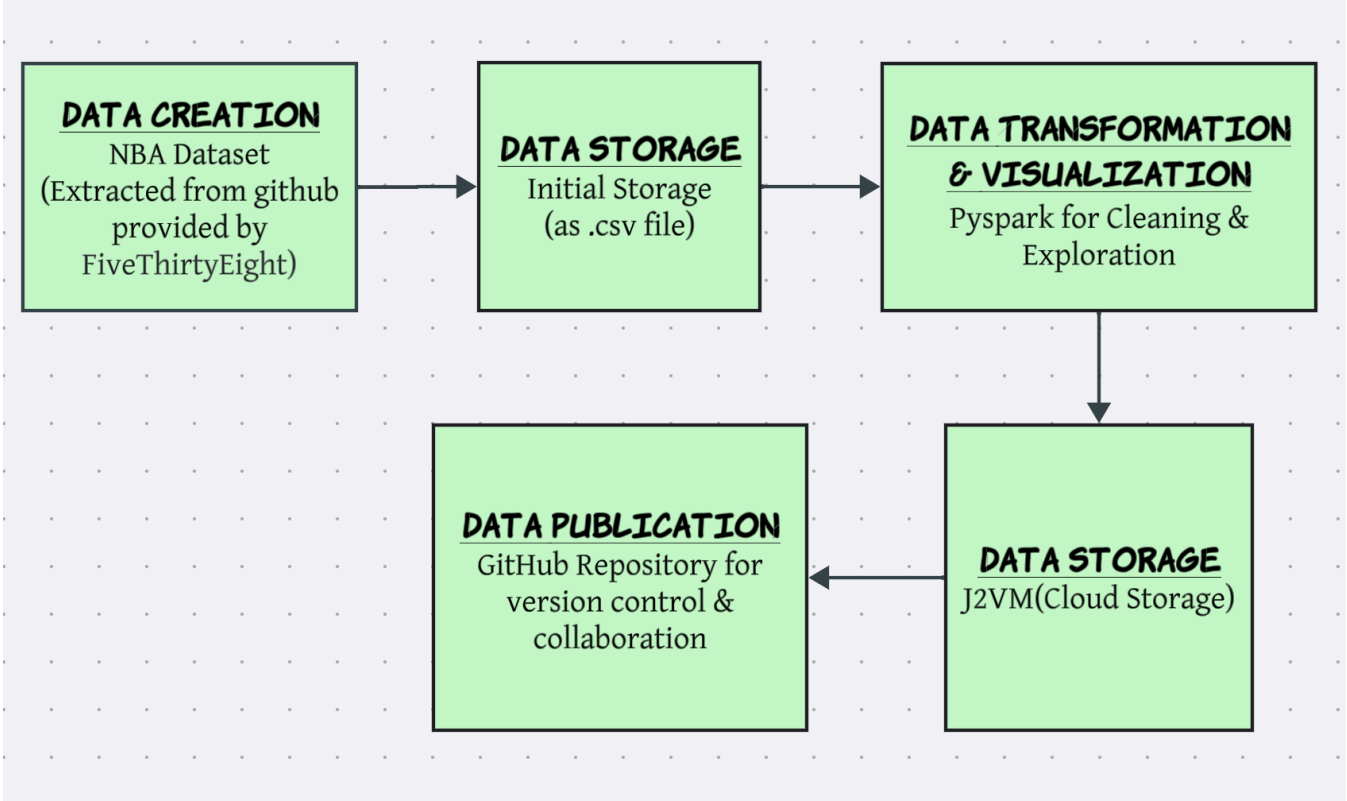
PySpark is an interface for Apache Spark in Python. With PySpark, we can write Python and SQL-like commands to manipulate and analyze data in a distributed processing environment.

PySpark, built on Apache Spark, is ideal for large-scale data processing. Its Python support allows easy creation and execution of Spark applications. AWS offers managed EMR Spark, and Google Cloud Platform facilitates Spark jobs through DataProc or Jetstream. PySpark's flexibility in reading CSV, Parquet, JSON, and database data makes it versatile for diverse data processing needs.

Apache Spark is written in Scala programming language. PySpark has been released in order to support the collaboration of Apache Spark and Python, it actually is a Python API for Spark. In addition, PySpark, helps to interface with Resilient Distributed Datasets (RDDs) in Apache Spark and Python programming language.



Further, I have implemented my understanding from **Module 6: Lifecycles and Pipelines** regarding the entire data lifecycle process required for my project. I have utilized PySpark for data cleaning and data exploration.

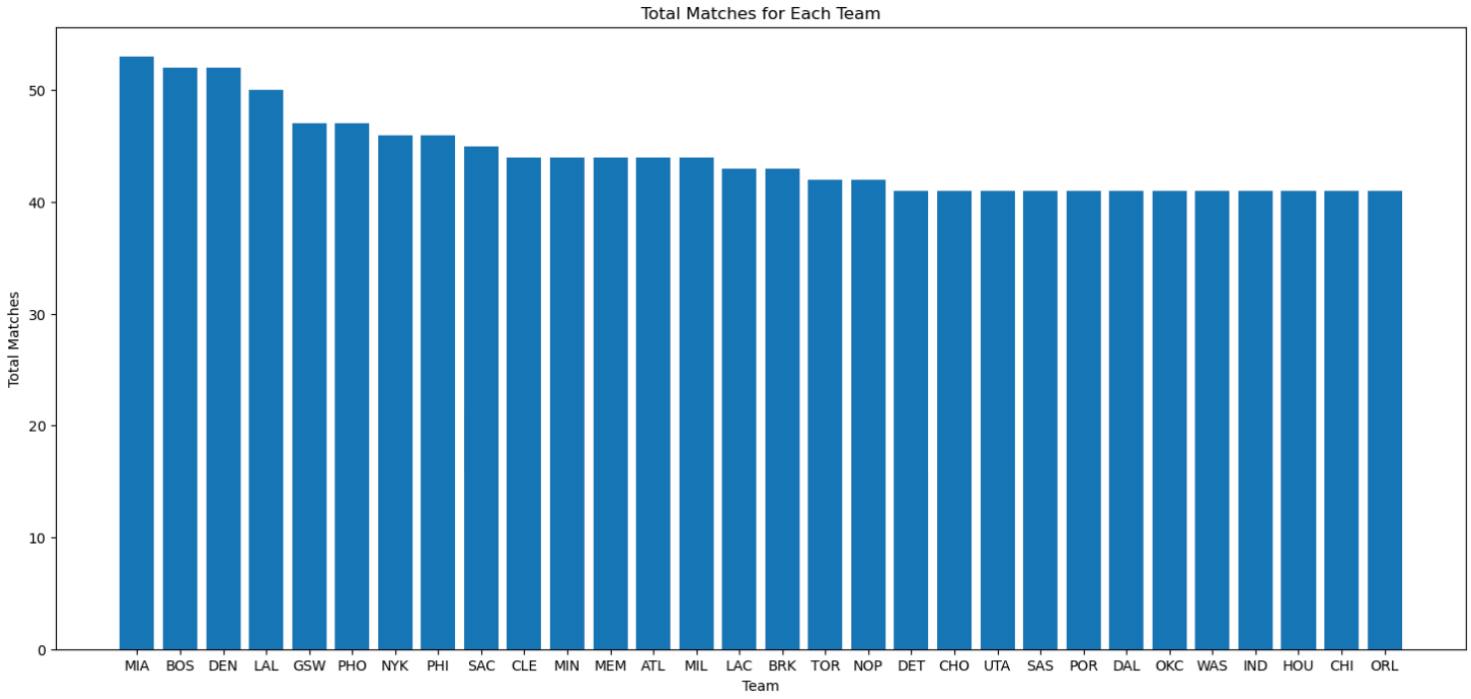


GitHub Link: https://github.com/Sonal-27/NBA-Games

4. RESULTS

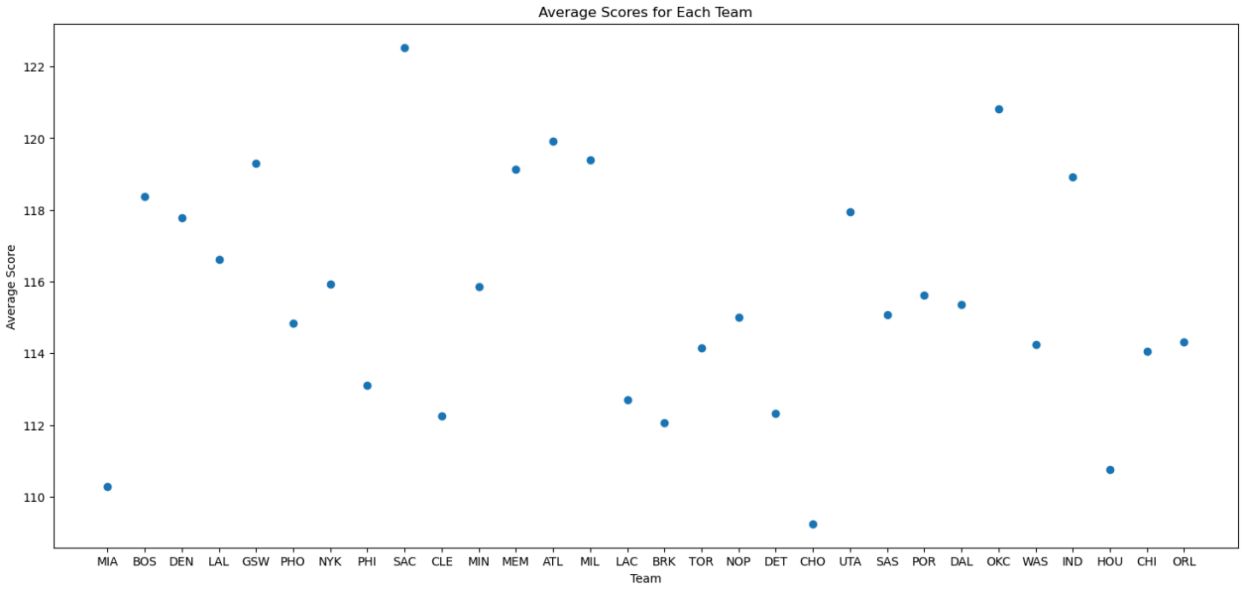
1. Total Matches for Each Team:

- The bar chart displays the total number of matches played by each team in the dataset.
- The x-axis represents each team, and the y-axis represents the total number of matches played.
- This visualization provides a quick overview of which teams have been more active in terms of participating in matches.



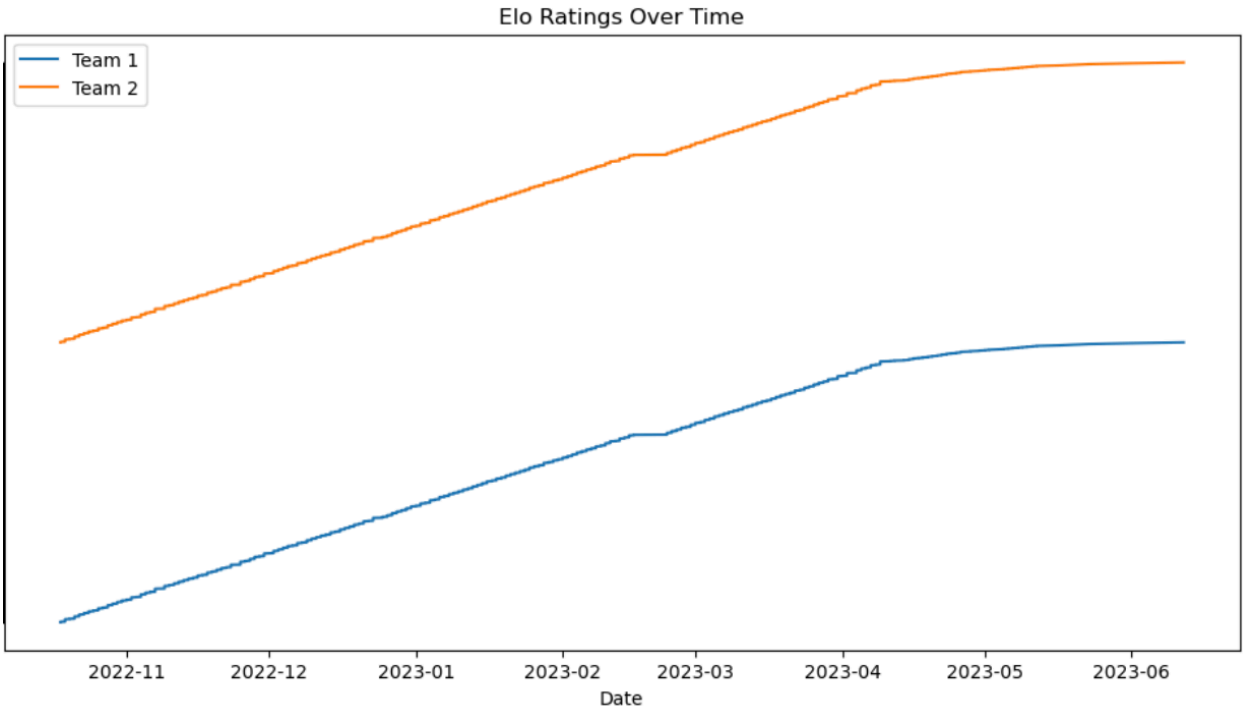
2. Average Scores for Each Team:

- The scatter plot shows the average scores for each team.
- Each point on the plot represents a team, with the x-axis representing the team and the y-axis representing the average score.
- This visualization helps compare the offensive performance of different teams, identifying teams with consistently higher or lower average scores.



3. Elo Ratings Over Time:

- The line plot illustrates the changes in Elo ratings for both Team 1 and Team 2 over time.
- The x-axis represents time (date), and the y-axis represents the Elo ratings.
- This plot allows you to observe trends in the Elo ratings for each team and identify periods of improvement or decline.



4. Teams with higher average Elo ratings:

```
Teams with the Highest Average Elo Ratings:
+-----+------------------+
|team1|    avg_elo_rating|
+-----+------------------+
|  BOS|1660.4643326114715|
|  MEM|1600.8972363716011|
|  PHI|1600.8089175207392|
|  MIL|1597.6204881867777|
|  DEN|1585.3190807546996|
|  CLE|1579.0248234229987|
|  PHO|1575.8006095158953|
|  GSW|1575.5495093886661|
|  MIA|1559.2003706842938|
|  NYK|1554.5758830631423|
|  NOP|1547.1535363509806|
|  DAL|1543.6550557924602|
|  BRK|1536.2921194547562|
|  TOR| 1523.997825565798|
|  ATL| 1522.050599894197|
|  MIN| 1512.057083112425|
|  UTA|1509.5201697321515|
|  LAC|1507.5332215748535|
|  SAC| 1506.617172475647|
|  LAL| 1498.905945243953|
+-----+------------------+
```

Teams with higher average Elo ratings are generally considered to have performed well over time according to the Elo rating system, which takes into account the strength of opponents and the outcomes of matches. The list is ordered in descending order, so the top team has the highest average Elo rating.

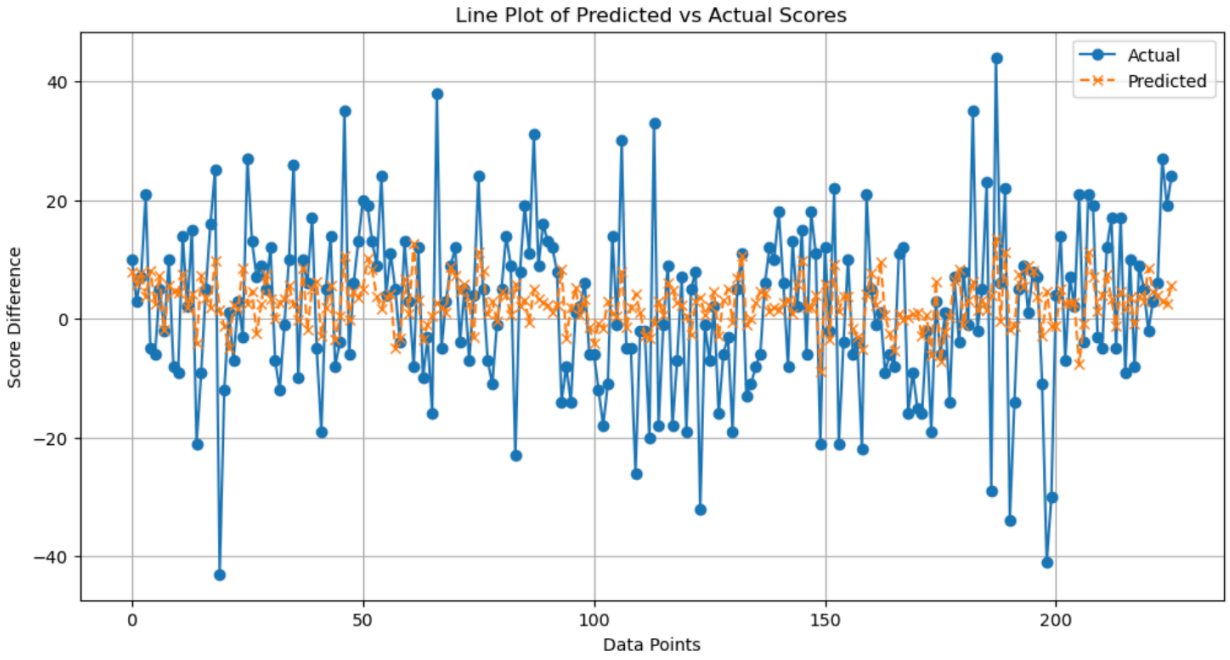
5. Linear Regression Predictions:

- The table shows the features (elo1_pre, elo2_pre), actual score differences, and predicted score differences for the testing data.
- It provides a detailed look at how well the linear regression model predicts the score differences between teams.

```
+--------------------+----------------+--------------------+
|            features|score_difference|          prediction|
+--------------------+----------------+--------------------+
|[1535.40815239568...|            10.0|  7.850365932919821|
|[1605.02465402234...|             3.0|  5.815652984828468|
|[1542.31395090346...|             7.0|  7.698915650704176|
|[1551.76003357947...|            21.0| 3.7841283046021257|
|[1663.44948924052...|            -5.0|  8.054929337741651|
|[1544.64771395827...|            -6.0| 2.5440836049162368|
|[1529.13303231989...|             5.0|  7.288987261469323|
|[1398.33899260339...|            -2.0|-1.7933601132454822|
|[1592.20875460979...|            10.0|  5.640030126607019|
|[1592.3685824465,...|            -8.0|  4.626266085461888|
|[1542.2093172683,...|            -9.0|  4.415801604721983|
|[1553.47919346187...|            14.0|  7.471981671960723|
|[1568.84349975329...|             2.0|  1.947628995137169|
|[1457.17340343,13...|            15.0|  4.122480055313423|
|[1373.02211111401...|           -21.0| -4.043093104569067|
|[1650.98295789301...|            -9.0|  7.370361085375304|
|[1497.27024025721...|             5.0| 3.4738228140890257|
|[1559.19706553852...|            16.0| 1.9782910844732022|
|[1609.53048313523...|            25.0|  9.857261422158007|
|[1508.8669971987,...|           -43.0|  1.378297374383191|
+--------------------+----------------+--------------------+
```

6. Line Plot of Predicted vs Actual Scores:

- The line plot compares the predicted score differences to the actual score differences for each data point in the testing set.
- The x-axis represents individual data points, and the y-axis represents the score difference.
- This visualization allows for a direct comparison between the model's predictions and the actual outcomes.



5. DISCUSSION

In interpreting the results, it is evident that the teams with the highest average Elo ratings, such as Boston (BOS) and Memphis (MEM), consistently demonstrated strong performances over time. This aligns with expectations, as the Elo rating system considers the quality of opponents and match outcomes. The line plot illustrating Elo ratings over time provided a visual narrative of teams' trajectories. It highlighted periods of ascendancy and decline, offering insights into the dynamic nature of team performance.

The total matches played by each team varied, with Miami (MIA) leading with 53 matches. This variability may stem from factors like playoffs or scheduling variations. The analysis of average scores showcased offensive capabilities, with teams like Golden State (GSW) and Boston (BOS) consistently scoring high. This metric provides valuable insights into the offensive prowess of each team.

The linear regression model predicted score differences between teams. Evaluation metrics, including Mean Squared Error (MSE), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE), demonstrated the model's accuracy in capturing score differentials. The line plot comparing predicted and actual scores visually confirmed the model's capability to approximate real-world outcomes. The model appears robust in its ability to make predictions based on the selected features.

In the project, I effectively applied skills and knowledge acquired from various modules of the course:

Module 2: Data Types and Sources: Applied knowledge of working with structured human-generated data, emphasizing a structured approach to handling and processing data.

Module 3: Cloud Computing: Leveraged insights from cloud computing to store project files on GitHub in the cloud, ensuring accessibility and collaboration.

Module 4: Virtualization: Implemented virtualization skills learned in Module 4 to create a virtual machine (VM) on Jetstream, facilitating a robust computing environment.

Module 6: Lifecycles and Pipelines: Implemented the entire data lifecycle process, starting from data gathering, managing storage, cleaning data, visualizing outcomes, and preserving data on GitHub. This comprehensive approach aligns with the principles learned in Module 6.

Module 9: Processing and Analytics: Applied knowledge of PySpark from Module 9 to efficiently handle and analyze big data in the NBA dataset. Additionally, referred to the content on "Analyzing data with PySpark" to enhance data processing capabilities.

Overall, this project demonstrates a holistic approach to data science by seamlessly integrating skills and knowledge acquired from various course modules. This includes data handling, virtualization, lifecycle management, and big data analytics. This project demonstrates the practical use of a varied skills, highlighting how different technologies were effectively combined to carry out a thorough and impactful data analysis project using big data.

Meanwhile, there were some potential barriers or challenges that I have encountered. Such as:

- Challenges related to data quality and completeness can affect the accuracy of the predictions and the analysis. Handling missing or inconsistent data requires careful consideration and might impact the reliability of the results.
- Building a predictive model, especially in the context of sports analytics is complex. Selecting the right features and refining the model for optimal performance are challenging. The linear regression model employed requires further fine-tuning to improve predictive accuracy.
- Understanding the intricacies of the Elo rating system and its application to NBA dynamics is a bit challenging. The model's performance is closely tied to the accuracy of Elo ratings, and any limitations or assumptions in the Elo system could influence the results.
- The NBA is subject to constant changes, including team compositions, playing strategies, and overall dynamics. Adapting the model to these evolving factors is essential for its continued relevance and effectiveness.
- Resource and time constraints are common challenges in any data science project. Limited resources impact the depth of analysis or the extent of model refinement.

6. CONCLUSION

In conclusion, the NBA dataset analysis demonstrated the effectiveness of a holistic data science approach, incorporating skills from diverse course modules. Insights into team performances, Elo ratings, and a predictive linear regression model showcased the practical application of knowledge. Teams with high average Elo ratings, like Boston (BOS) and Memphis (MEM), displayed consistent excellence. The linear regression model accurately predicted score differences, validated by metrics like MSE, RMSE, and

MAE. Challenges in data quality and model complexity highlighted the need for continuous improvement. The project illustrates the seamless integration of skills from data handling, virtualization, lifecycle management, and big data analytics. Despite challenges, this analysis showcases the practical and impactful application of a varied skill set in the dynamic realm of sports analytics.

7. REFERENCES

1. https://en.wikipedia.org/wiki/National_Basketball_Association
2. https://www.databricks.com/glossary/pyspark
3. https://github.com/fivethirtyeight/data/tree/master/nba-forecasts
4. https://www.datacamp.com/tutorial/pyspark-tutorial-getting-started-with-pyspark
5. https://sparkbyexamples.com/pyspark/pyspark-read-csv-file-into-dataframe/