Table of Contents

1.0 Summary

The main objective in this project is to dig through large datasets from NBA.com and gain meaningful insights that will drive my curiosity as basketball fan, Production Analyst (Sports data), and most importantly, someone who wants to learn and add Python as a skill for Data Analytics.

1.1 Questions

1. How are basketball statistical categories related to each other? Highlight 10 key insights.
2. How the game has changed since we entered the 21th century (2000-2022)
3. How are player minutes and scoring distributed (Regular Season vs Post-season)

1.2 Deliverables

1. A clear summary of the business task
2. A description of all data sources used
3. Documentation of any cleaning or manipulation of data
4. A summary of your analysis
5. Supporting visualizations and key findings
6. Your top high-level content recommendations based on your analysis

2.0 Data Preparation

In this step, we need to prepare data for processing and analyzing. We will need to take a close look at the datasets, summarize them and discover some data quality characteristics. The dataset that we will be using is classified as primary data which is from NBA.com - The official site of NBA. We will be performing web scraping on NBA API using Python via Jupyter notebooks.

2.1 Data Summary

The extracted and cleaned dataset from NBA.com contains historical player data from 2000-2001 to 2021-2022 regular and post-seasons which sums up to 435,203 data points (15007 rows × 29 columns).

2.2 Data Limitations

Since this is a case study, and we can live these limitations for the purpose of gaining experience and skills. But had this been a real-life project, it will be highly recommended to use larger historical data for more accurate and interesting analysis.

2.3 Data Privacy

The NBA website provides a thorough data privacy section which can be seen here.

3.0 Data Cleaning and Processing

3.1 Steps taken for data cleaning:

The detailed steps taken during the cleaning process can be seen in the GitHub repository of this project. In summary here are the steps:

- Dropped unnecessary columns
- Replaced old team names with updated values
- Checked if 30 team names are unique and updated
- Updated column names for uniformity

3.2 Columns and Descriptions

Figure 01: Metadata

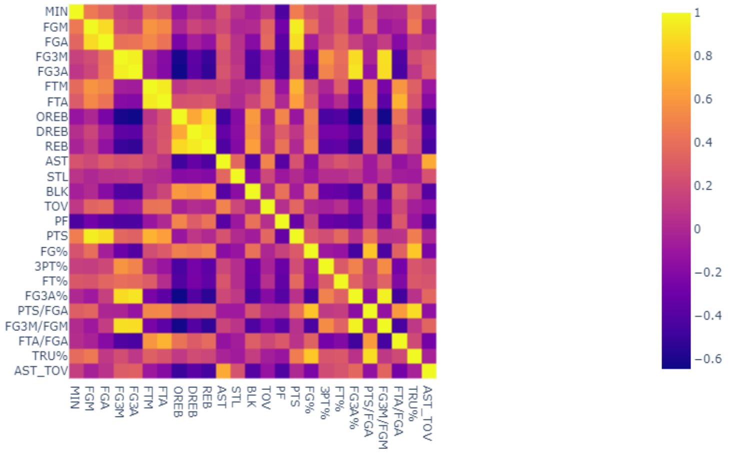
Column Name	Description
YEAR	NBA calendar year
SEASON_TYPE	Regular season or Post-season
PLAYER_ID	Unique Player Identifier
PLAYER	Player name
TEAM_ID	Unique Team Identifier
TEAM	Team name
GP	Sum of games played
MIN	Sum of minutes played
FGM	Sum of field goals made
FGA	Sum of field goal attempts
FG_PCT	Field goal percentage
FG3M	Sum of 3-pt field goals made
FG3A	Sum of 3-pt field goal attempts
FG3_PCT	3-pt Field goal percentage
FTM	Sum of free-throws made
FTA	Sum of free-throw attempts
FT_PCT	Free-throw percentage
OREB	Sum of Offensive rebounds
DREB	Sum of Defensive rebounds
REB	Sum of OREB and DREB
AST	Sum of Assists
STL	Sum of Steals
BLK	Sum of Blocks
TOV	Sum of Turnovers
PF	Sum of Personal fouls
PTS	Sum of Points
EFF	Sum of Efficiency
AST_TOV	Assist-to-turnover ratio
STL_TOV	Steal-to-turnover ratio

4.0 Analysis and Visualization

4.1 Correlations

- How are basketball statistical categories related to each other? Highlight 10 key insights.

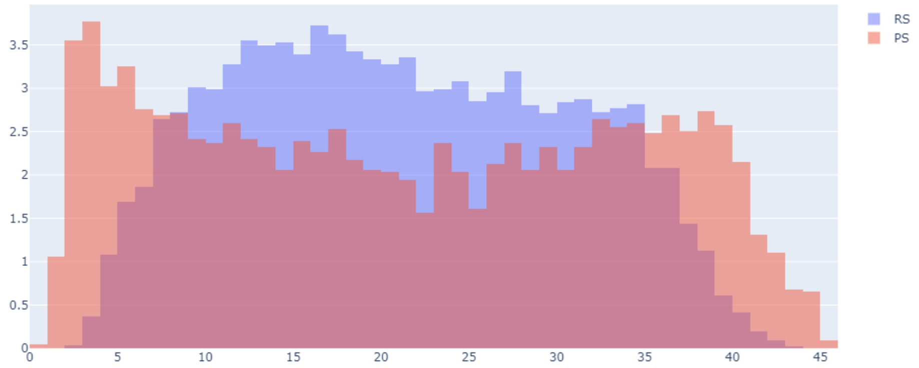
Figure 02: heat_map_corr



1. FG3M - OREB are negatively correlated
2. MIN - PTS are positively correlated
3. FGA - TOV are positively correlated
4. AST - TOV are positively correlated
5. FG% - BLK are positively correlated
6. 3PT% - BLK are negatively correlated
7. PTS - TOV are positively correlated
8. FG% - OREB are positively correlated
9. PF - REB are positively correlated
10. FT%- OREB are negatively correlated

- How are player minutes distributed during the regular season and post-season.

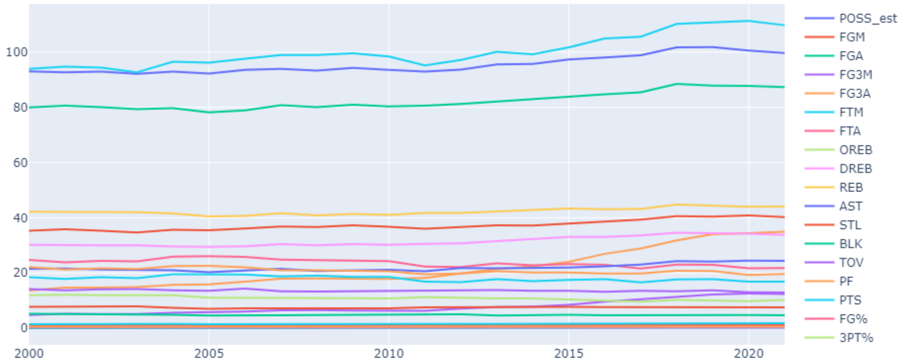
Figure 03: histogram_ps_rs_df_MIN



1. The data suggests that fewer players are being played at higher minutes during post-seasons
    - Player rotations are shallower during post-seasons compared to regular seasons
    - All-star players and starters are getting more playing time during post-seasons
    - Most Bench players are getting less minutes during the post-seasons

- How the game has changed since 2000. Highlight 5 key insights.

Figure 04: fig_change_per48_df



  - An upward trend can be seen in PTS, FGA, FG3A, FG3M, and POSS_est categories. This trend can be supported by the POSS_est which suggests that Points and Field goal attempts per game are highly impacted by the number of possessions per game.

5.0 Discussion

In this section, we will discuss the key insights from the analysis above. As an NBA fan and someone who want to learn Python, this project has been fun and exciting. Although, I wasn’t surprised with the insights because I have a good understanding with how the game is played and its rich history. Nonetheless, I will be looking forward to expanding this project in the future for the sake of being a fan of the game.

Key Takeaways

- Notable correlations:
  - Positively correlated
    - MIN - PTS
    - FGA - TOV
    - FG% - OREB
  - Negatively correlated
    - FT% - OREB
    - 3PT% - BLK
    - FG3M - REB
- Player rotations are shallower during post-seasons compared to regular seasons
- The upward trend in Points and Field goal attempts per game are highly impacted by the number of possessions per game. More possessions mean higher statistical numbers.

6.0 Recommendations

The main goal of this project is to hone my skills in Python programming language in a way that I’m familiar with - NBA Data. At the end, I came up with some interesting insights that are very evident in real-world. NBA basketball has become more fast-paced and for few reasons: Rule changes, evolution of player skillsets, and simple math (two 3-pt shot made = three 2-pt shots made).

This is a great starting point for a much bigger project. Here are some of my ideas for further and deeper analysis in the future:

1. Extract, prepare, and analyze individual team data from NBA.com as well to gain even more interesting insights
2. Create a script that will extract real-time data from NBA.com and make a real-time dynamic dashboard
3. Make a real-time dynamic dashboard in BI Tools such as Power BI and Tableau for more interactive experience

7.0 References

Web scraping URL:

https://stats.nba.com/stats/leagueLeaders?LeagueID=00&PerMode=Totals&Scope=S&Season=2021-22&SeasonType=Regular Season&StatCategory=PTS

GitHub Repository:

https://github.com/notoriousclay

Google Drive Files:

https://drive.google.com/drive/folders/1L_1DBbEOSS4JlD3_pKDAxvMaYzccBFV1?usp=share_link