**NBA Time-series Forecasting**

**Vignesh Tallam**  [**Shreyas Ramani**](mailto:sraman30@asu.edu)  **Anant Varma**

**ASU ID**: 1233433775  **ASU ID**: 1233983441  **ASU ID**: 1225746899

**Background and Dataset**

As passionate NBA enthusiasts, we wanted to analyze the **league's scoring evolution** through time series analysis. Our dataset spans from 1947 to 2024, capturing the average points scored per game and other metrics of the sport across all NBA teams. The original time series reveals fascinating patterns, with a mean of 102.98 points and a standard deviation of 10.71, showcasing the dramatic shifts in scoring across different eras. From the low-scoring games of the late 1940s (around 65 points) to the explosive offensive era of the mid-1960s (peaking near 118 points), and the recent surge past 110 points, the data tells a compelling story of basketball's evolution.

Compiled from multiple subsets available in Basketball Reference.

**Source**: [**https://www.basketball-reference.com/**](https://www.basketball-reference.com/)

Scraped from NBA sites and culminated multiple predictors and targets into a consolidated dataset

We merged multiple subsets relevant to our problem into one dataset.

**The objective of our forecast**

To build a time series forecasting model based on the historical data of NBA teams collected year-on-year to forecast the average points that will be scored in the upcoming NBA season.

**Features of the Dataset** -

Season (Year)

Total Points scored

Team

3 point attempts

3 PA %

Number of Blocks

Number of Steals

Number of Rebounds

Number of Assists

Overall Efficiency (Field Goal Percentage)

Number of Free throws scored

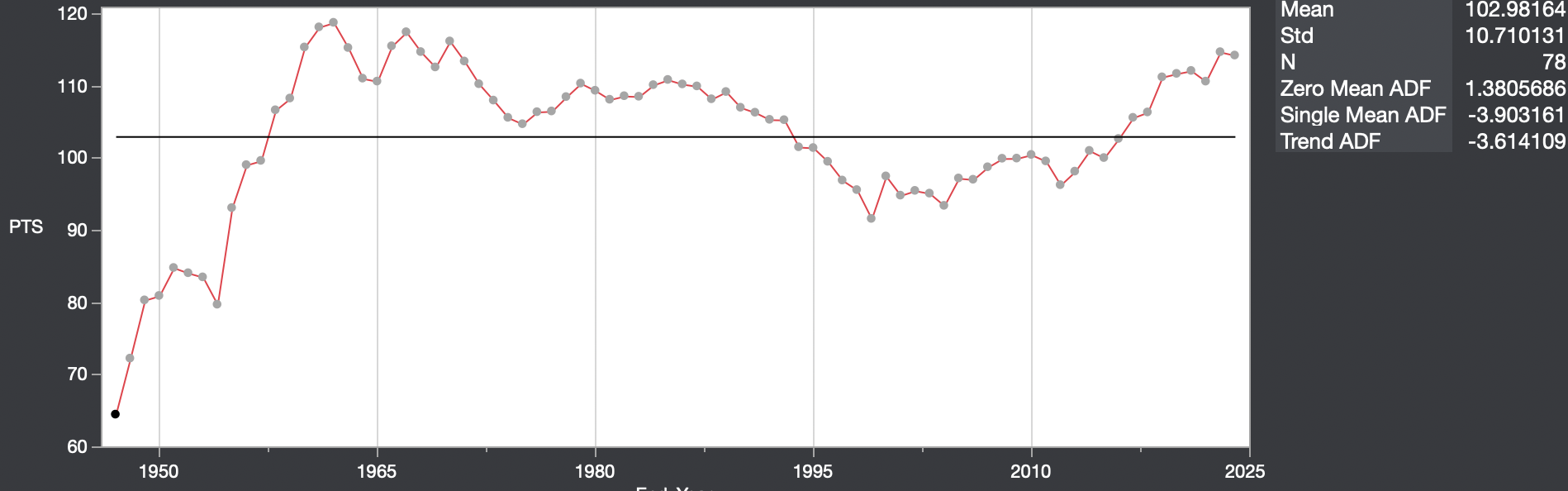
Field Goal Attempt

TurnOver percent

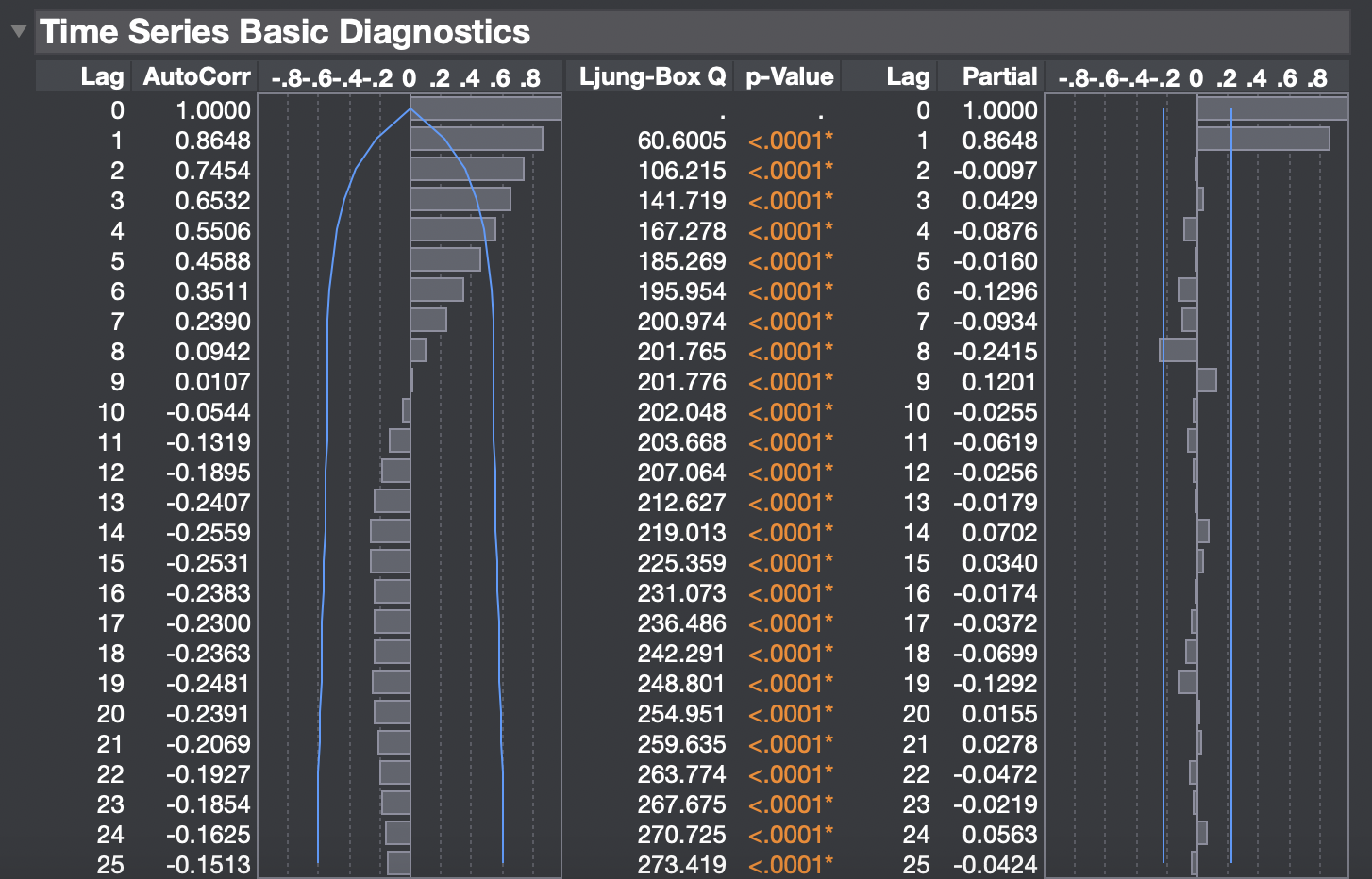
These features were compiled for all the teams for each season from 1947 to 2024.

**JMP Table and Graph**

We plotted a Time series graph for the Average of the total points scored by all the teams per season against each Year.



After observing the ACF and PACF for the initial time series plot, the data seemed non-stationary, and a trend could be observed.



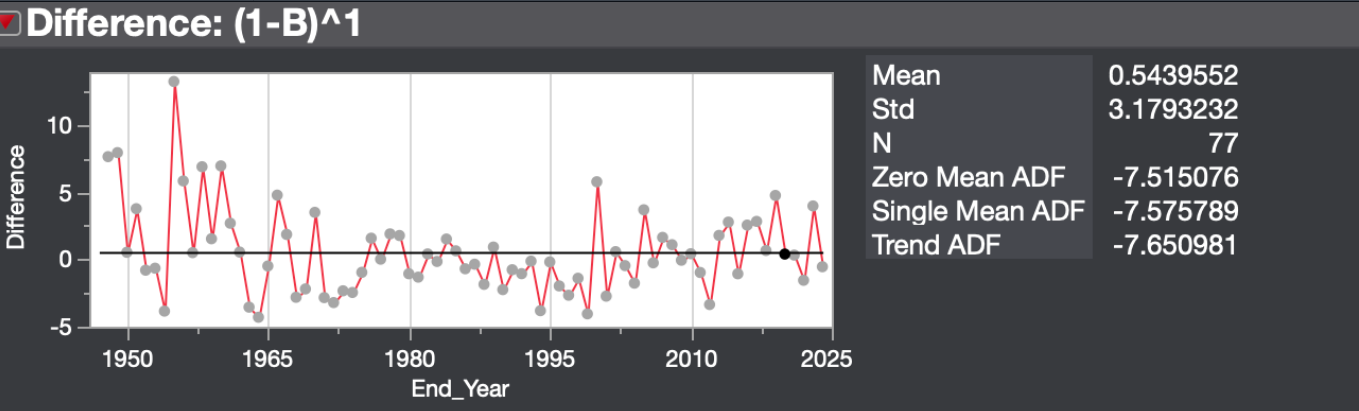
The data shows clear non-stationarity with varying mean levels over time, as evidenced by the fluctuating trend line that doesn't maintain a constant mean.

1. The ADF (Augmented Dickey-Fuller) test statistics shown in the plot support this conclusion:

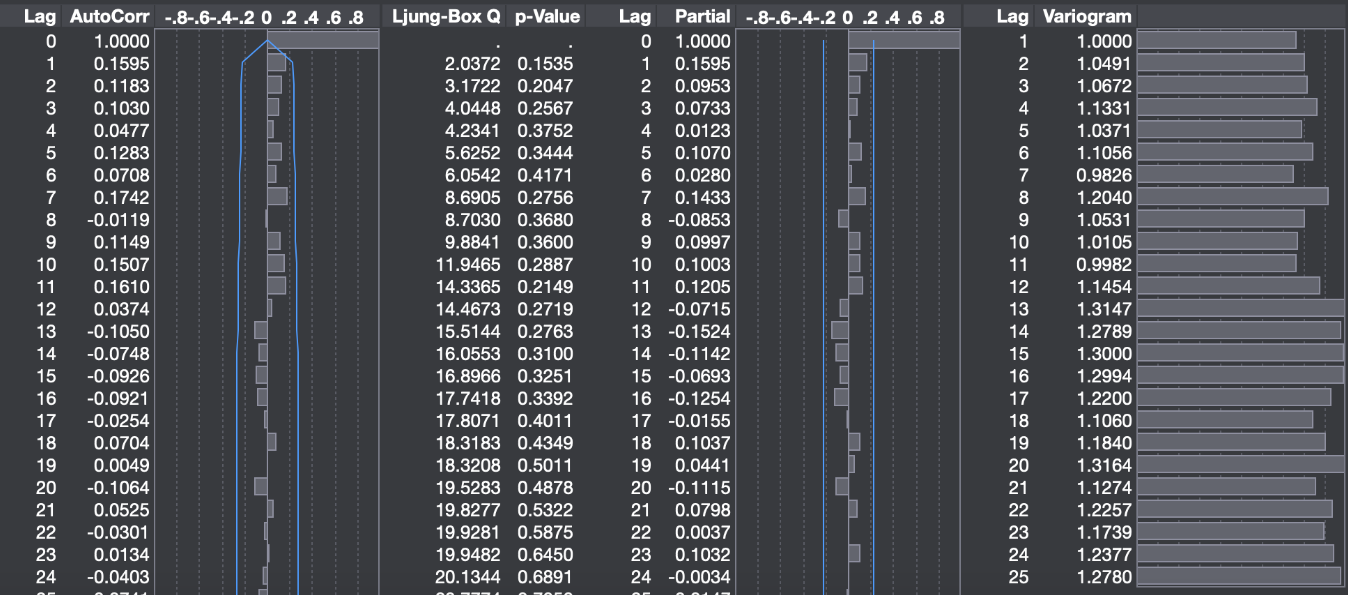
* Zero Mean ADF: 1.3805
* Single Mean ADF: -3.9031
* Trend ADF: -3.6141

The positive Zero Mean ADF particularly indicates non-stationarity. The plot also shows long-term trends and patterns that could be stabilized through differencing, particularly noticeable in the upward trends from 1950-1965 and 2010-2025, as well as the more stable period between 1980-2000.

Therefore, we performed a First-order differencing before applying the time series techniques.



After differencing one time and detrending, the Autocorrelation plot was as follows:



The significantly more negative ADF values after differencing indicate the transformed series is stationary. The differencing has definitely removed the non-stationary components making the time series now suitable for fitting to a model.

**Competitive Study**

We came across a previous study that used a subset of the data we collected from the same source. A researcher in the JMP community presented the Positional Matchup Model in JMP. This presentation was made on May 17, 2024, and it demonstrates how to use a Positional Matchup Model in JMP to predict offensive player performance in the NBA.

**Source**:https://community.jmp.com/t5/Abstracts/Predicting-NBA-Player-Performance-with-a-Positional-Matchup/ev-p/755526

This model aimed to predict offensive player performance in the NBA, focusing on points per game as the primary metric. The researcher chose a Lasso regression for selecting the features for a Multiple Linear Regression model. The final model chosen seemed to be a Multiple Linear regression model with the features chosen by the Lasso method. In the presentation, the metrics like R-squared or Mean Absolute Error were not mentioned. The details that were found related to the model performance were the Parameter estimates that were significant statistically at the alpha 0.05 level.

* Points per game (had a positive relationship)
* Turnover percentage (showed a negative relationship)
* Defensive pace (minimal positive relationship)
* Defensive rebounding percentage

The presenter mentioned performing cross validation whereas we used a train-test split to validate.

On the other hand, our model forecasts for average points scored by all teams in a given season as this study concentrated on forecasting points per player.

Our model found the following parameters to be statistically significant for our use case:

* Field goal attempts
* Field Goal % (FG)
* Assist
* 3 point %

Our approach, which used an ARIMA model to forecast average points for all teams in the upcoming season, differs from the Positional Matchup Model.

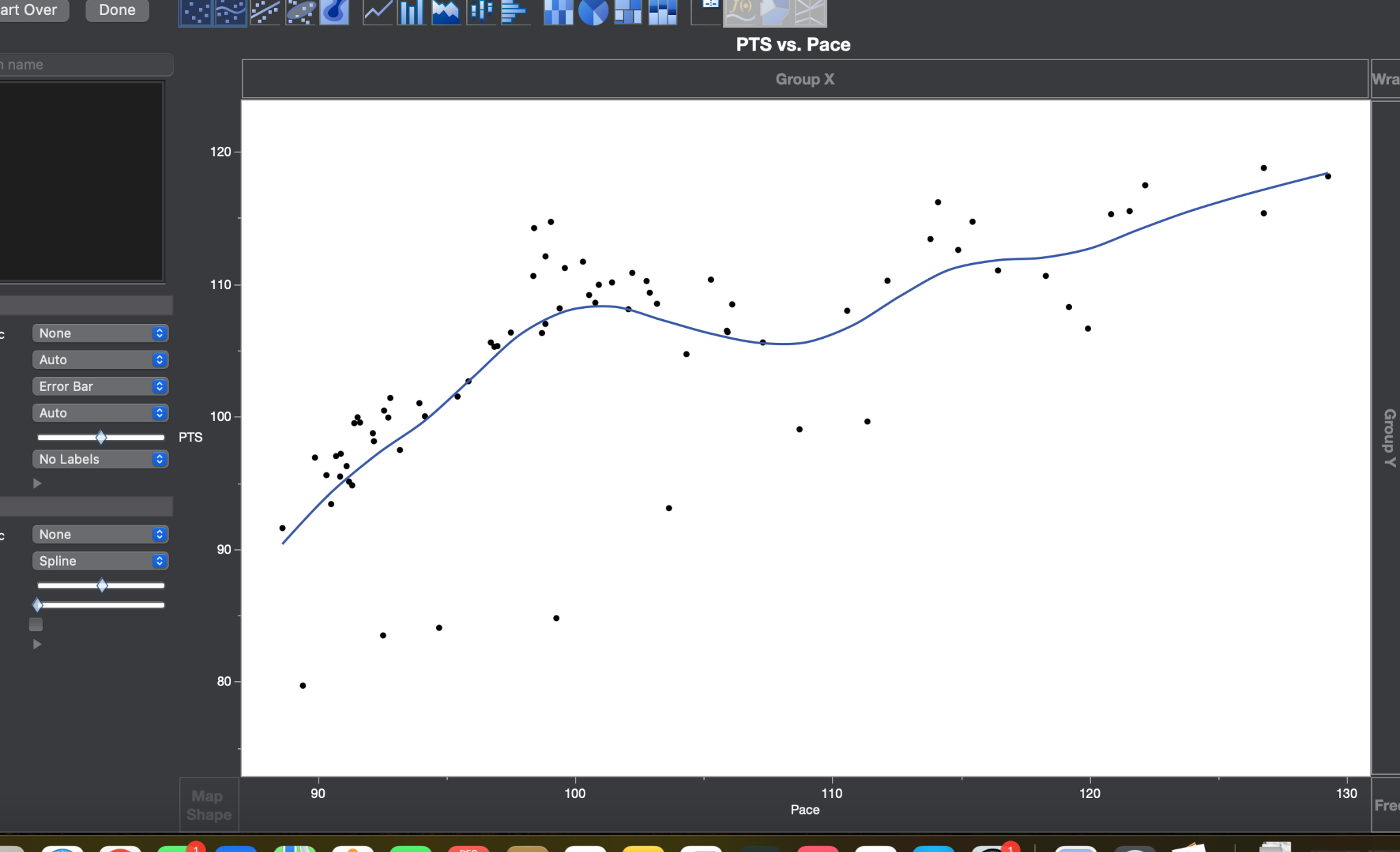
* Our ARIMA model focuses on team-level predictions, while the JMP model predicts individual player performance.
* Time Series vs. Cross-sectional: ARIMA is specifically designed for time series data, whereas the JMP model appears to use cross-sectional data with multiple predictors.
* Predictors: The JMP model (Positional Matchup Model) incorporates various player and team statistics, defensive metrics, and matchup data, while ARIMA primarily relies on historical point data and its own lagged values.

**Inference**: Our ARIMA model is particularly suited for capturing trends and seasonality in time series data, while the Positional Matchup Model allows for more granular, player-specific predictions based on a wide range of factors.

**Analysis**

The NBA time-series forecasting experiment is a comprehensive approach to predicting average points scored per season in the NBA based on historical data and advanced time-series methods. Our dataset, which spans almost 80 years, offers a solid basis for long-term trend analysis. The dataset is extensive and biases may be introduced if NBA regulations, play styles, and season durations are not sufficiently taken into consideration. Thorough pretreatment procedures improve the dataset's dependability, and using the ACF, PACF, and ADF tests is in line with industry best practices for time-series data preparation.

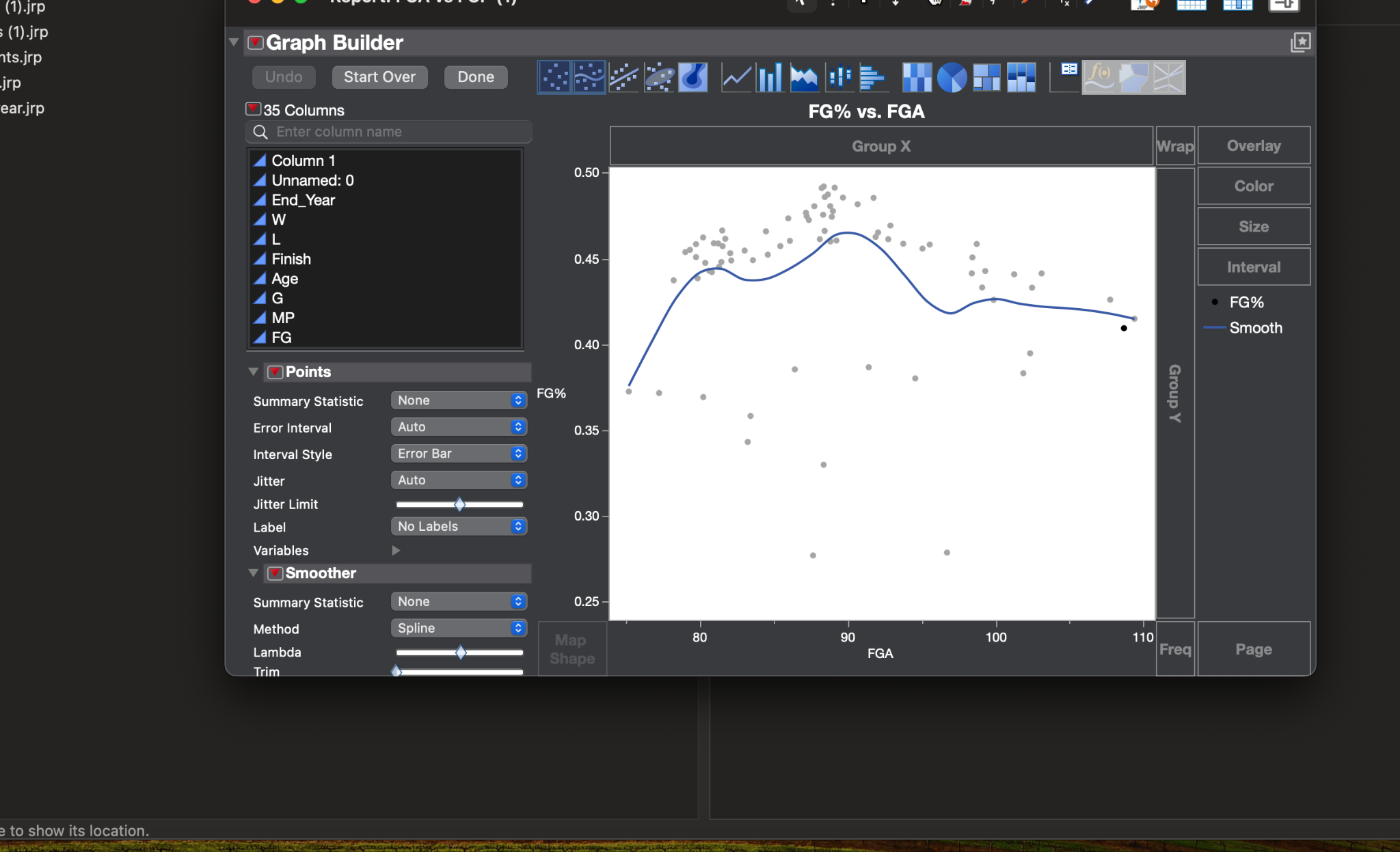
**PTS vs Pace**

****

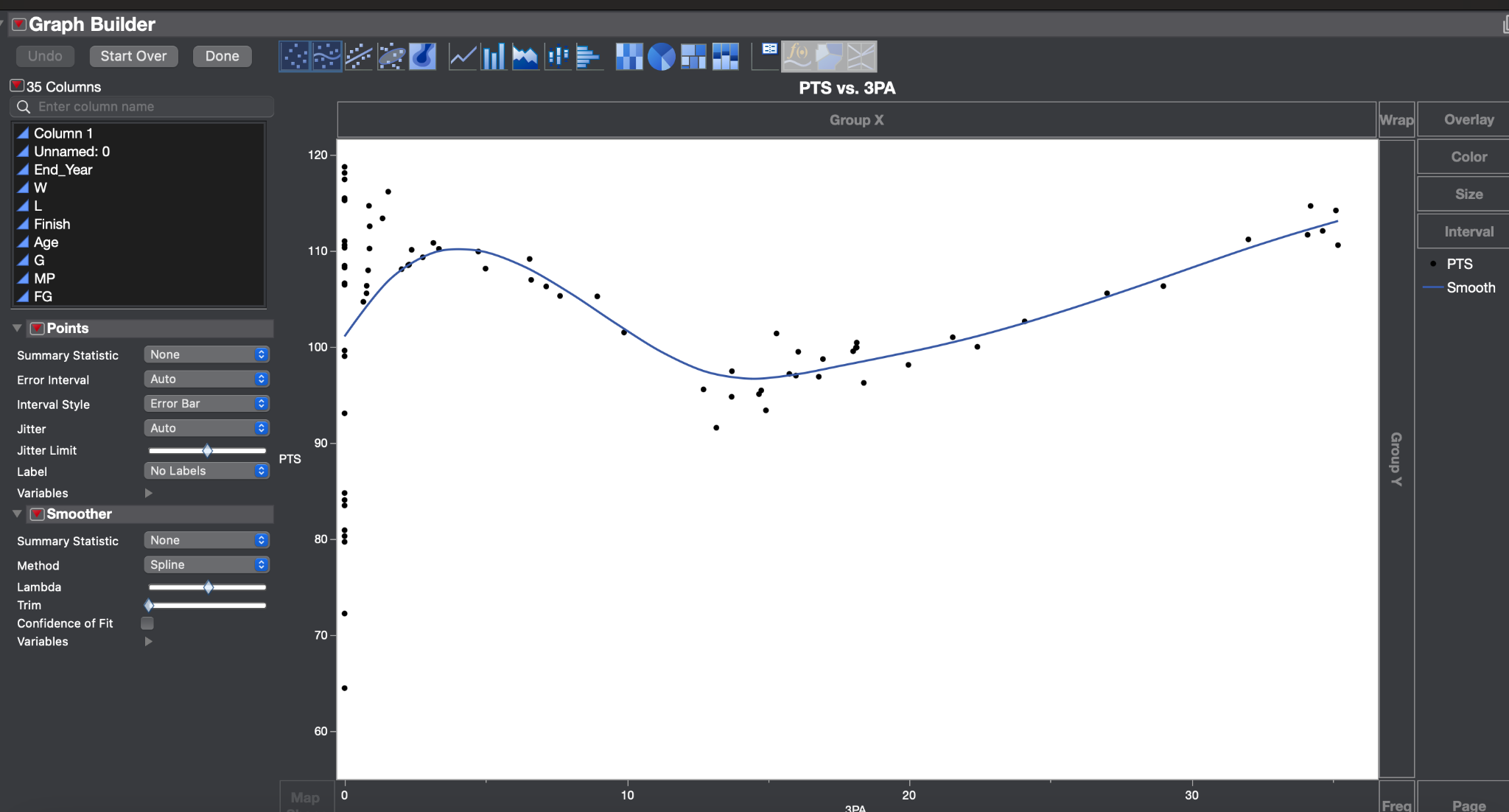
There is a general upward trend in the data as evidenced by the spline smoother.

The curve is not perfectly linear, there are regions where the rate of increase in PTS with respect to pace seems to slow down or accelerate.

**FG% vs FGA**

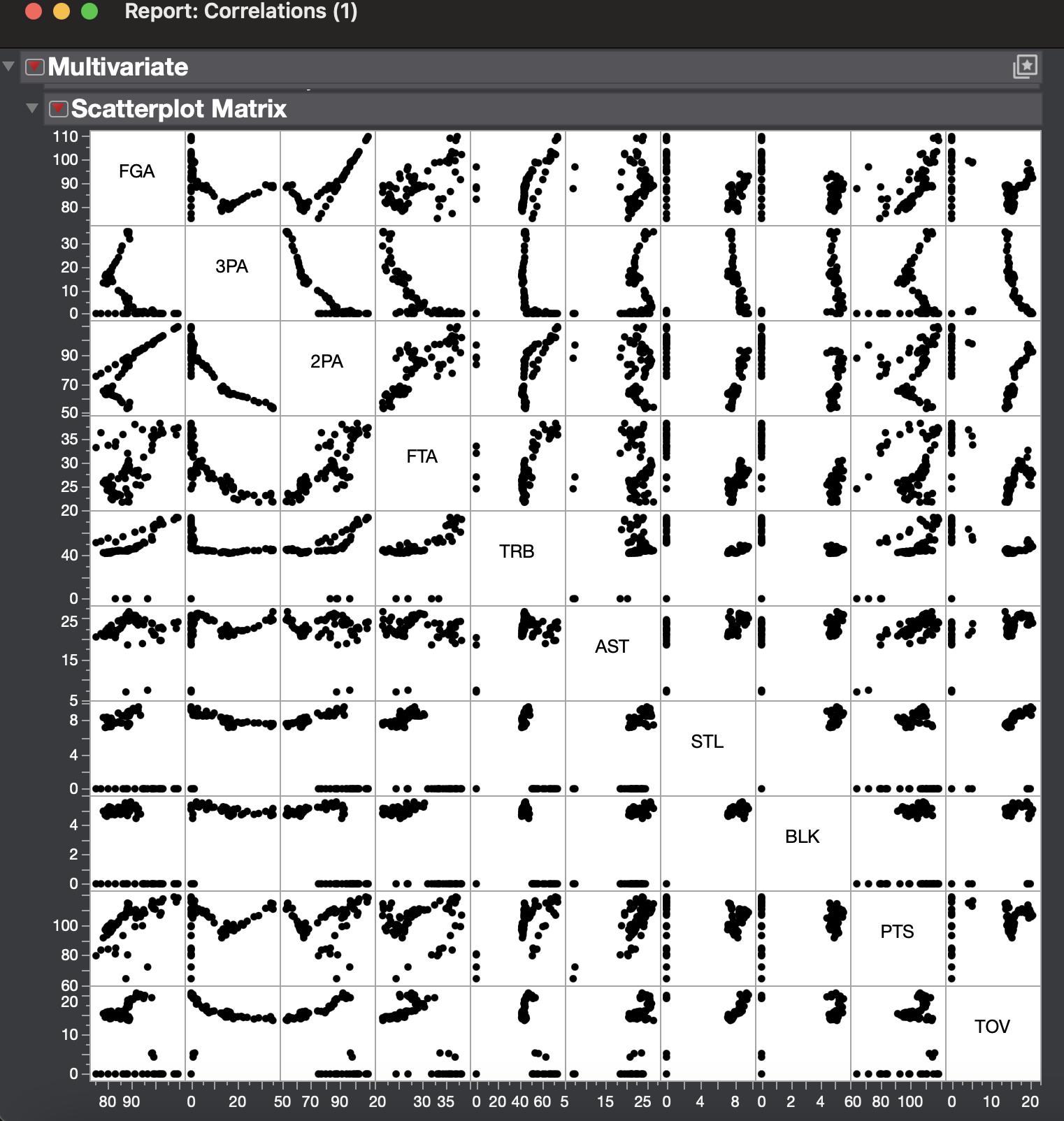
****

There is an overall non-linear trend between FG% and FGA. The data points are spread widely, with variability in FG% being substantial across all levels of FGA. This spline smoother effectively captures the non-linear trend, showing an initial increase followed by a gradual decrease and plateauing at high FHA levels.

**PTS vs 3PA  
**

The relationship between 3PA and PTS appears nonlinear. This spline smoother highlights the dip and subsequent upward trend effectively  
This level of smoothing provides a clear depiction of the nonlinear relationship.

**Correlations**

****

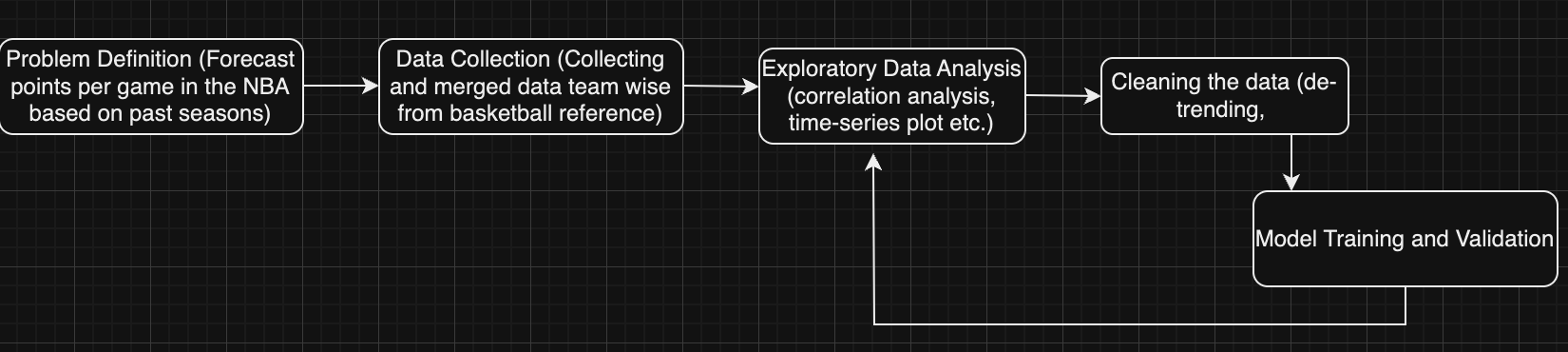
Some scatter plots exhibit clustering, indicating sub-groups or specific playing styles.

Variables like 3PA, 2PA, and FTA strongly influence scores (PTS).

Future research on this subject and model experiments could involve adding derived measures like possessions per game or offensive efficiency to improve prediction accuracy, as well as expanding the analysis to player-level data or team-specific patterns for more detailed insights.

**Model Pipeline and Validation**

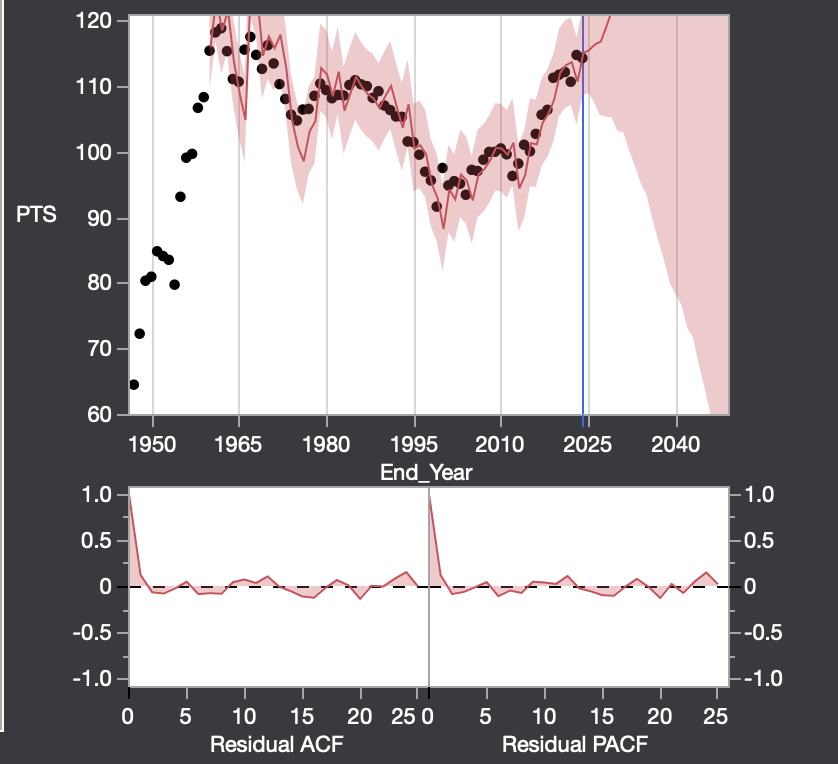
Below is the flowchart for the analysis process:

****

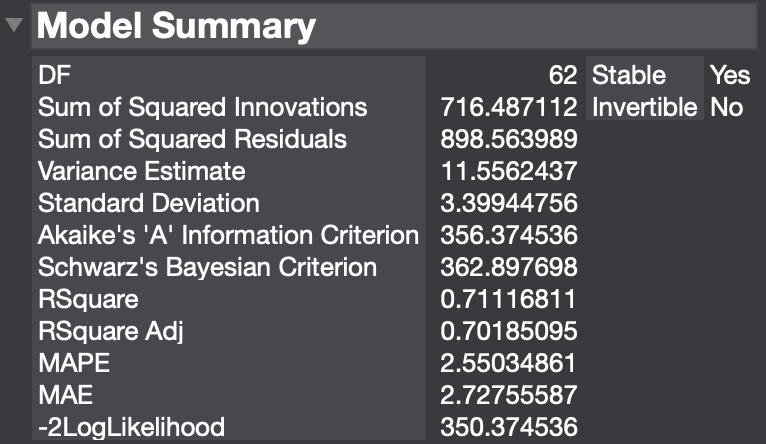
1. **Problem Definition:** Scoring in the NBA has surged significantly in recent years. This increase can be attributed to several factors, including rule changes that favor offensive play and a growing emphasis on three-point shooting. These shifts have fundamentally altered the way teams approach scoring, making it an intriguing challenge to predict future trends in points per game. Understanding these trends could provide valuable insights into how teams are adapting their strategies and how scoring dynamics are likely to evolve in the coming years.
2. **Data Collection:** Data for this analysis was collected from the web pages of all 30 NBA teams on Basketball Reference. The relevant data tables were extracted from hyperlinks, cleaned, and merged into a single dataset. To analyze league-wide trends, the yearly averages for all 30 teams were calculated, covering the period from 1947 to 2024. This comprehensive dataset provided a historical view of scoring patterns across nearly eight decades, enabling a detailed analysis of the evolution of points per game in the NBA.
3. **Exploratory Data Analysis:** The dataset was explored to identify correlations between points per game and other features, such as three-point attempts, assists, and pace. This analysis provided insights into how these variables influence scoring trends. Additionally, a time series plot of points per game was created to visualize trends over the years. Autocorrelation analysis was also performed to examine the relationship between scoring in successive years, helping to identify patterns and potential season-to-season dependencies in the data.
4. **Cleaning Data**: The dataset underwent several preprocessing steps to ensure accuracy and usability for analysis. First, any missing or incomplete data points were imputed or removed based on the extent of missingness. Variables were normalized to account for differences in scales across features like pace and three-point attempts. Outliers were detected and treated using statistical thresholds to prevent distortion of results. To prepare the data for time series modeling, the trend component was removed, and first-order differencing was applied to reduce autocorrelation and stabilize the data. Additionally, duplicate entries were removed, and feature engineering was performed to create meaningful variables, such as three-point attempts per possession, to enhance the model’s predictive power. These steps ensured the dataset was clean, consistent, and ready for robust analysis
5. **Model training and validation:** Several models were trained and validated to forecast points per game. A multiple linear regression model using ordinary least squares was the starting point, incorporating features like three-point attempts and pace as predictors. Time series-specific techniques were then applied, including first-order and second-order smoothing, to capture trends and reduce noise. Advanced models like ARIMA and AR(2) were also implemented to account for season-to-season dependencies and long-term patterns in the data. Model performance was evaluated using metrics such as RMSE and AIC to ensure the best fit. This multi-model approach allowed for a comprehensive understanding of scoring trends and provided robust predictions for future points per game.

**Different Models and Comparison of Performances**

1. **Winters method Additive**

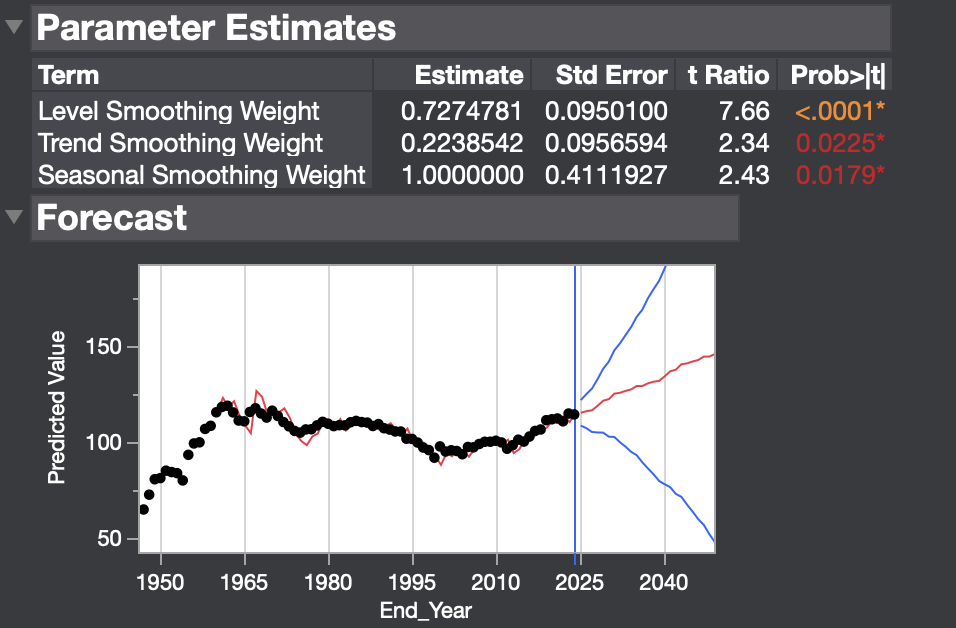


**Model summary:**

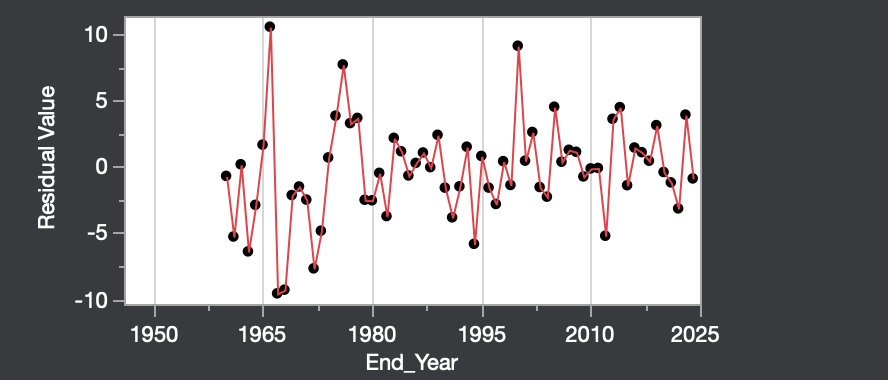


**Take away:** Low AIC and BIC indicates a relatively efficient model, higher SSR and low R2 could mean less accurate fit

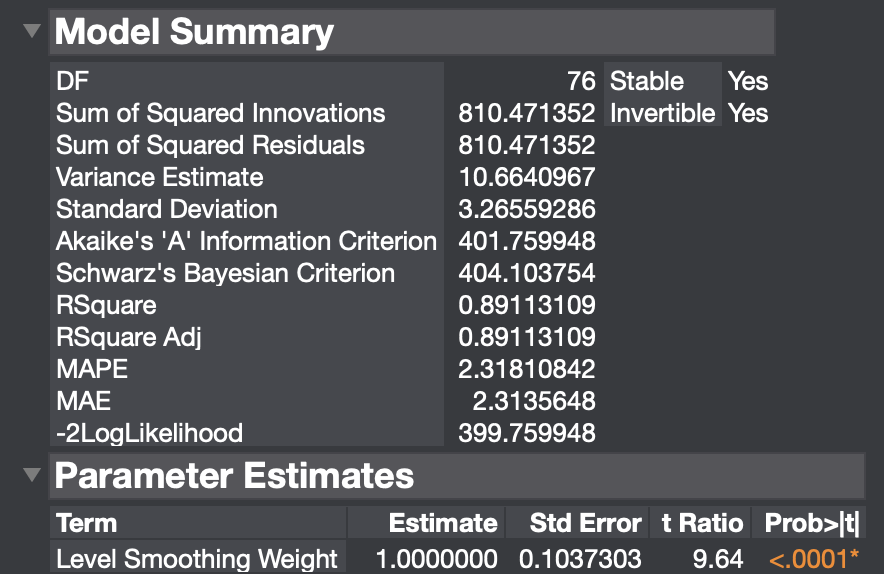
**Forecast:**



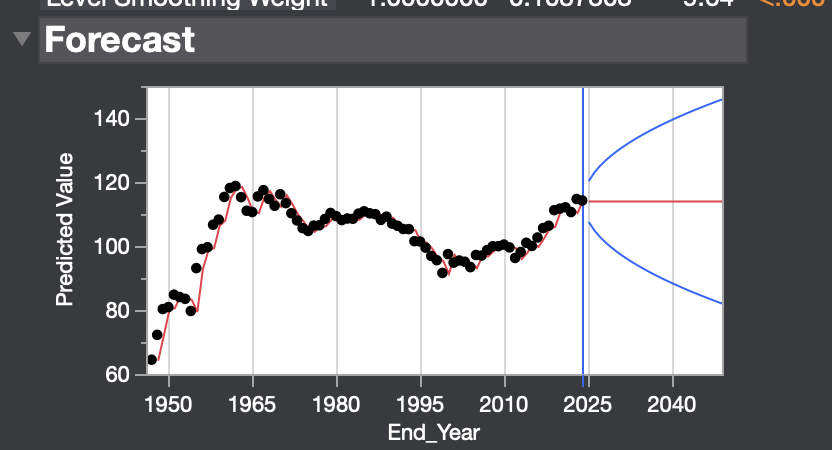
**Residual Value:**



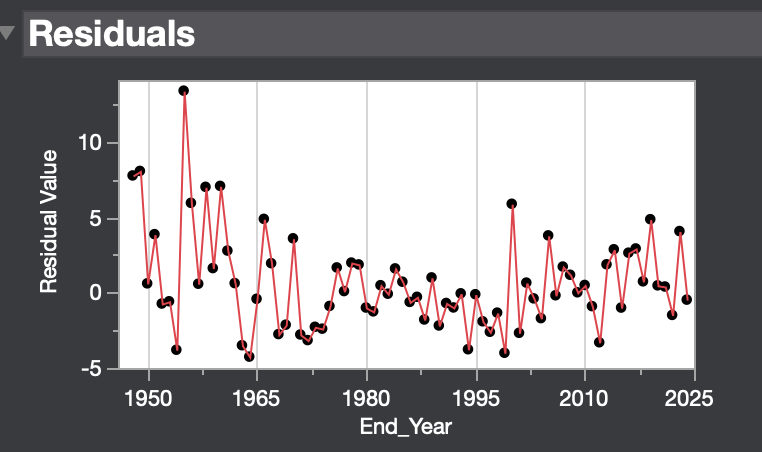
1. **Simple moving averages**

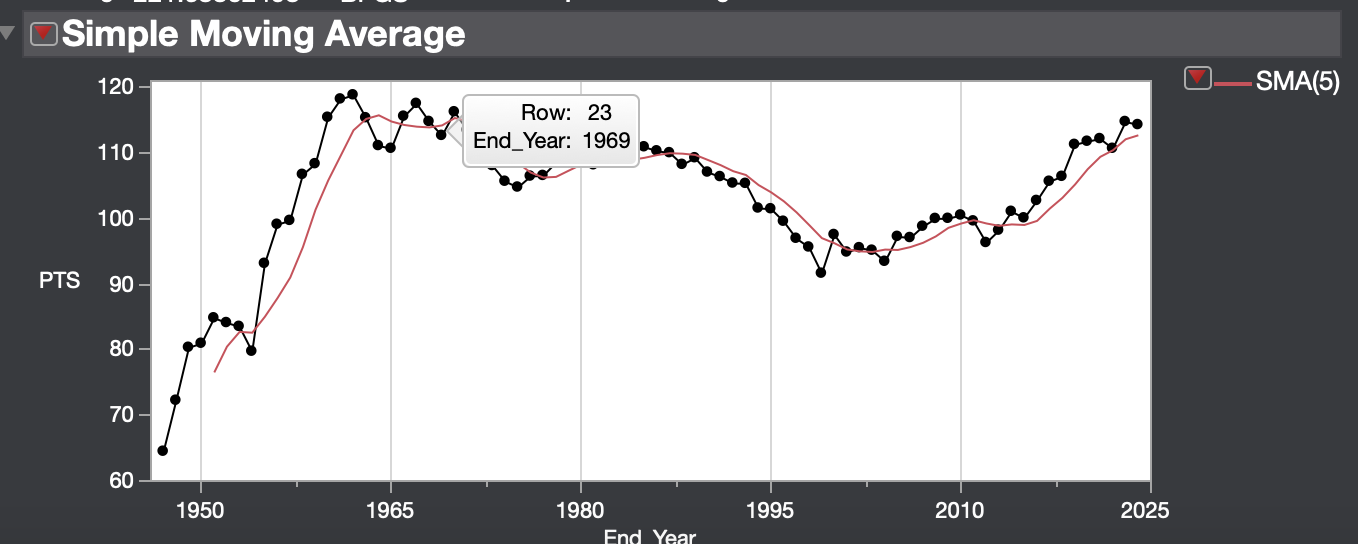


**Take away:** Higher AIC and BIC indicates a relatively less efficient model, lower SSR compared to Winters method and high R2 could mean more accurate fit



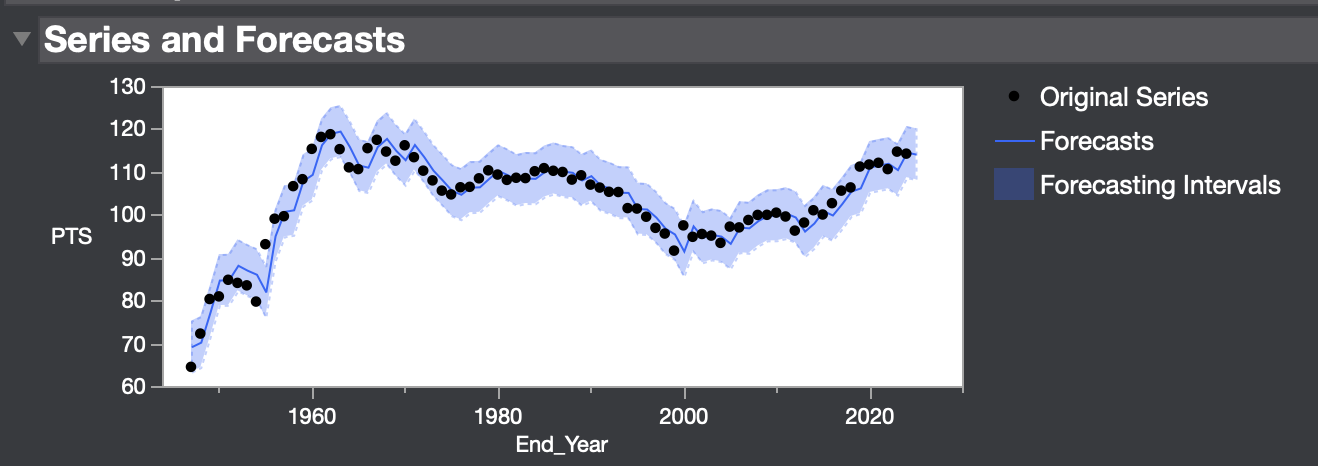
**Residuals:**





**c. Exponential Smoothing**

**Plot:**

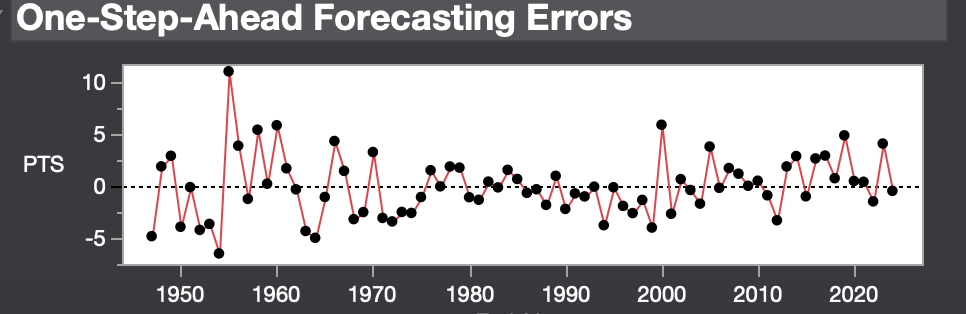


**Model Summary:**

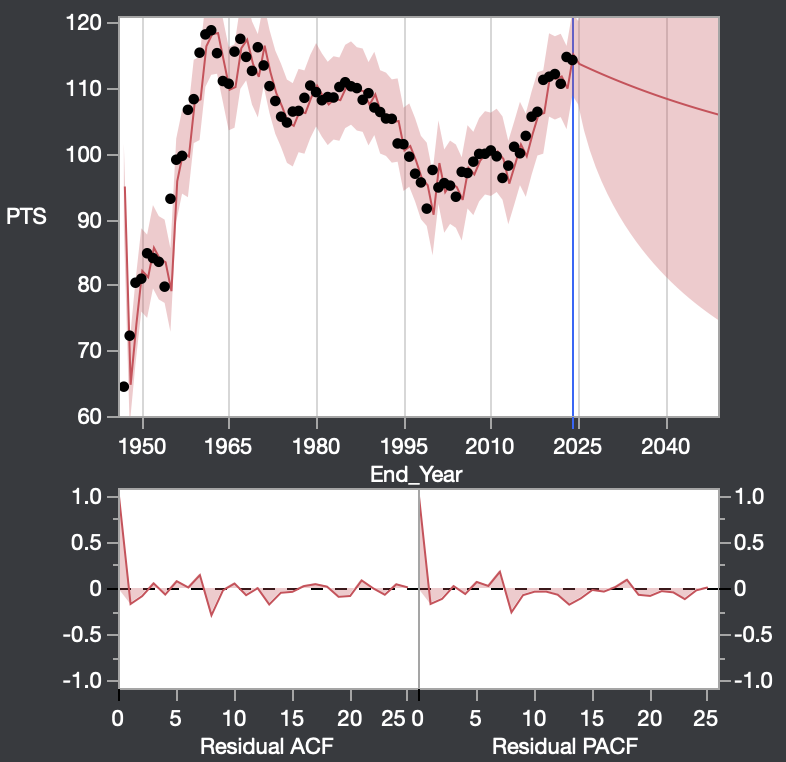


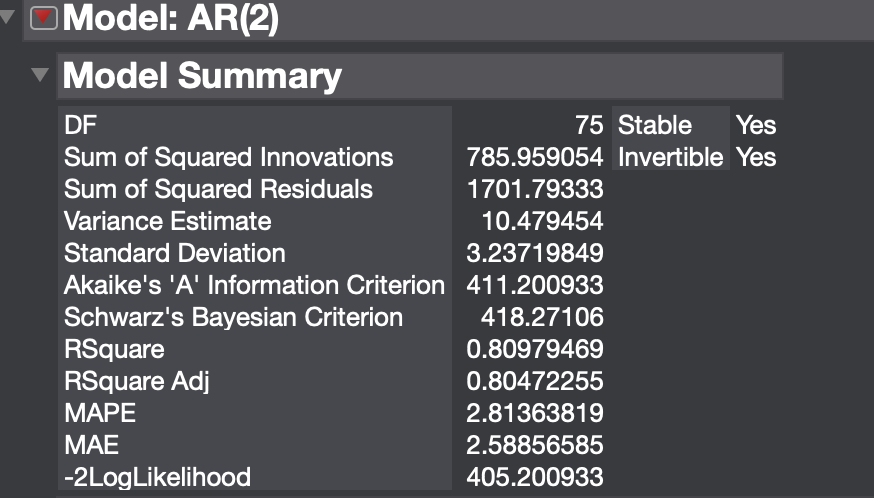
**Take away**: Due to low beta value, trend effect decreases over time, and high alpha value indicates higher weightage to more recent values, phi value indicates trend effect decreases over time. AIC and BIC are a little bit high making model less efficient

**Forecast:**

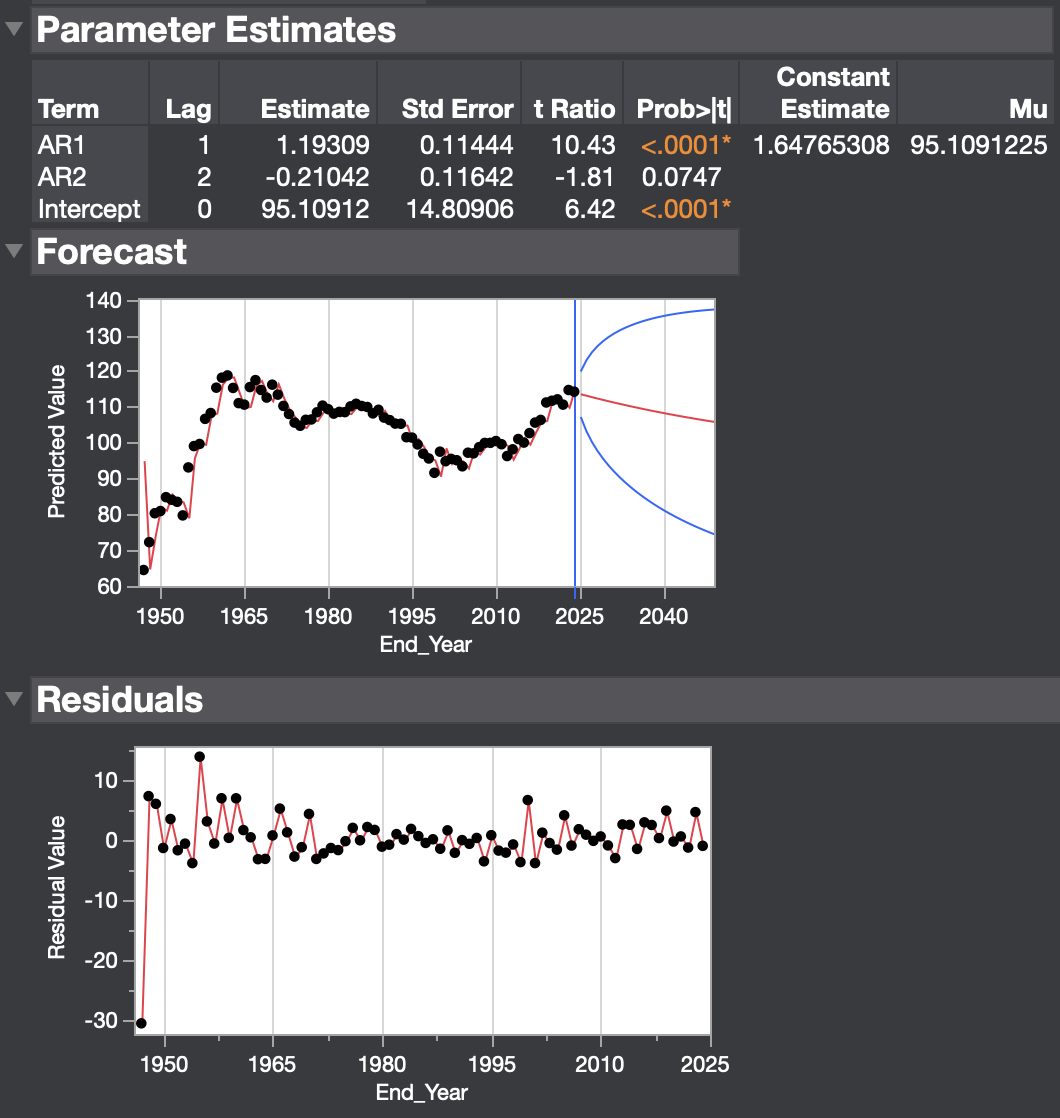


**d. ARIMA**

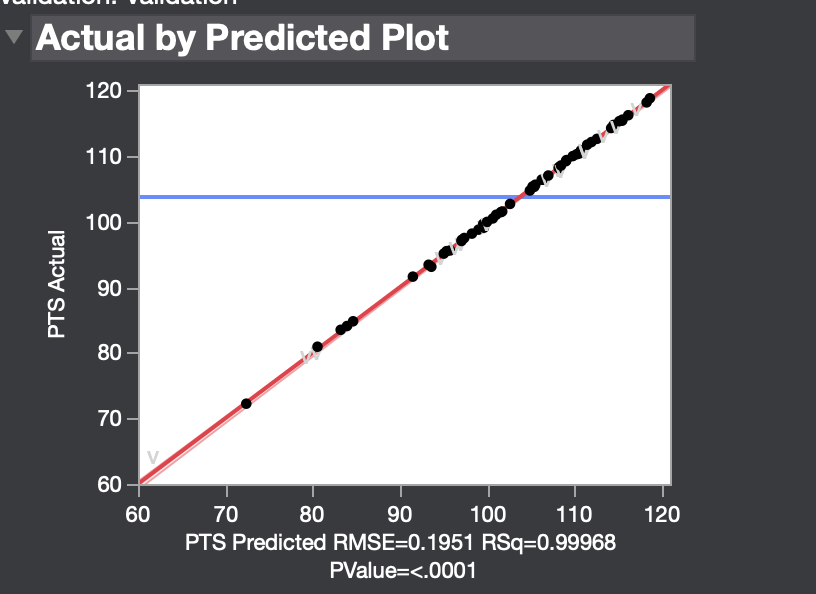


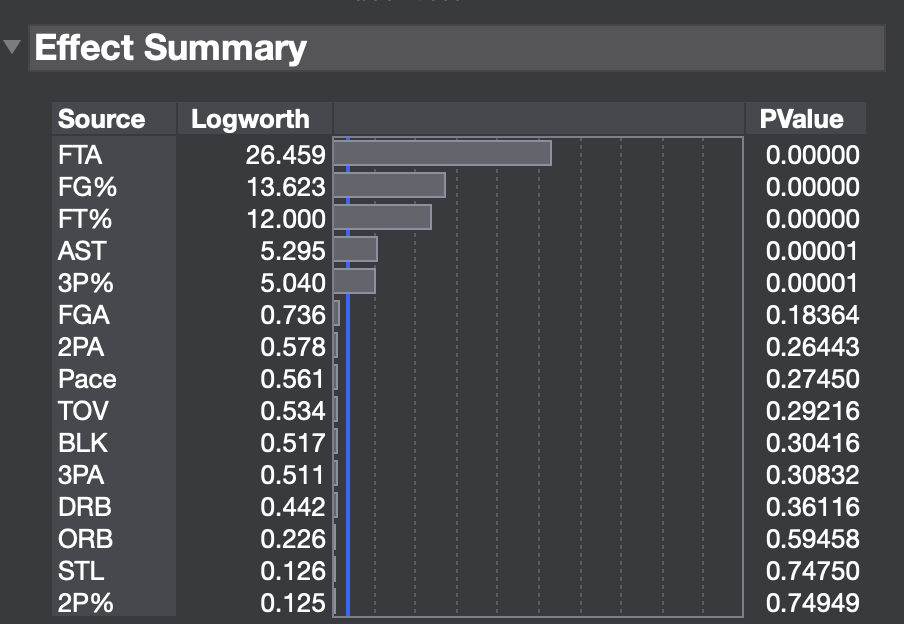


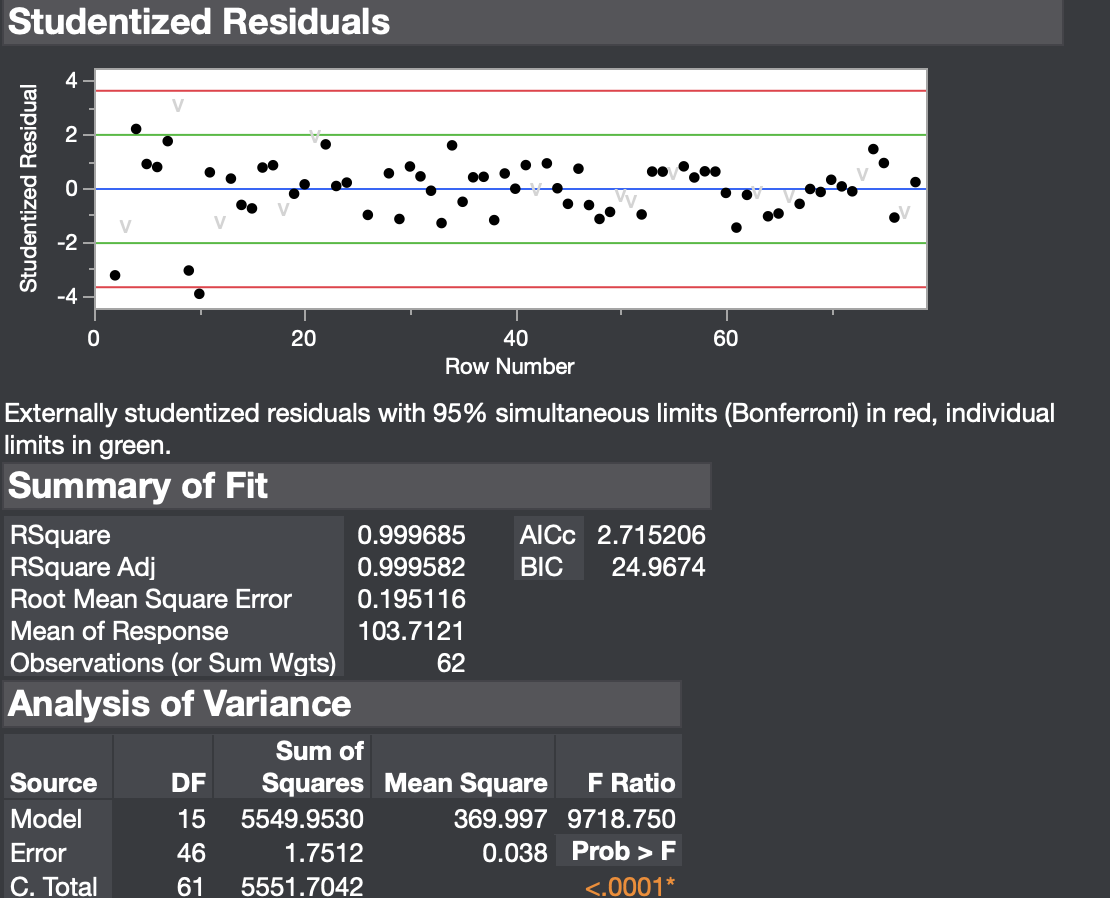
AIC and BIC are a little lower indicating higher efficiency and moderate R2 compared to simple moving average

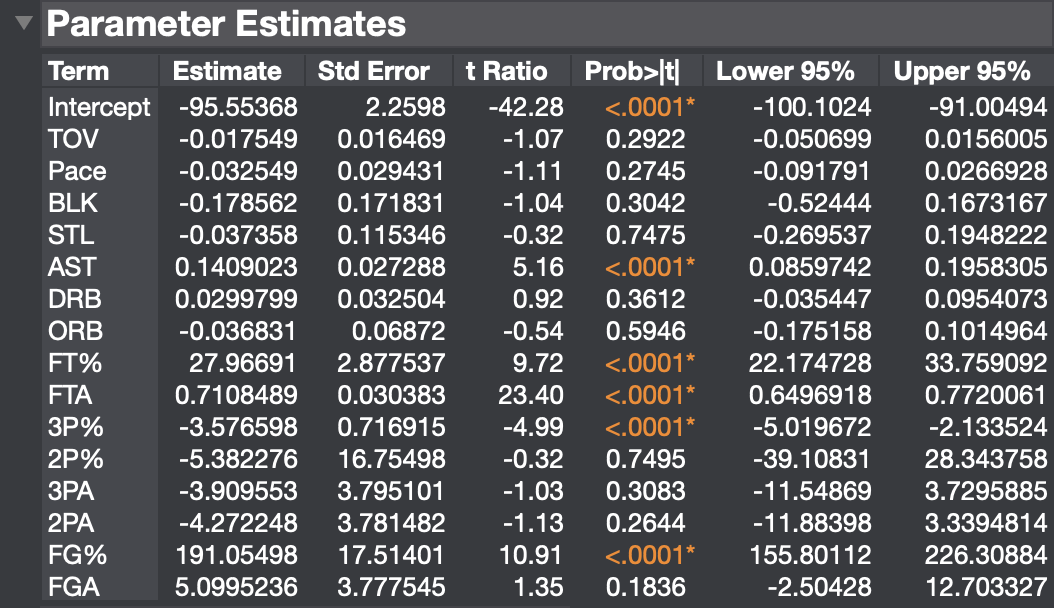


**e. Fitted Least Squares**

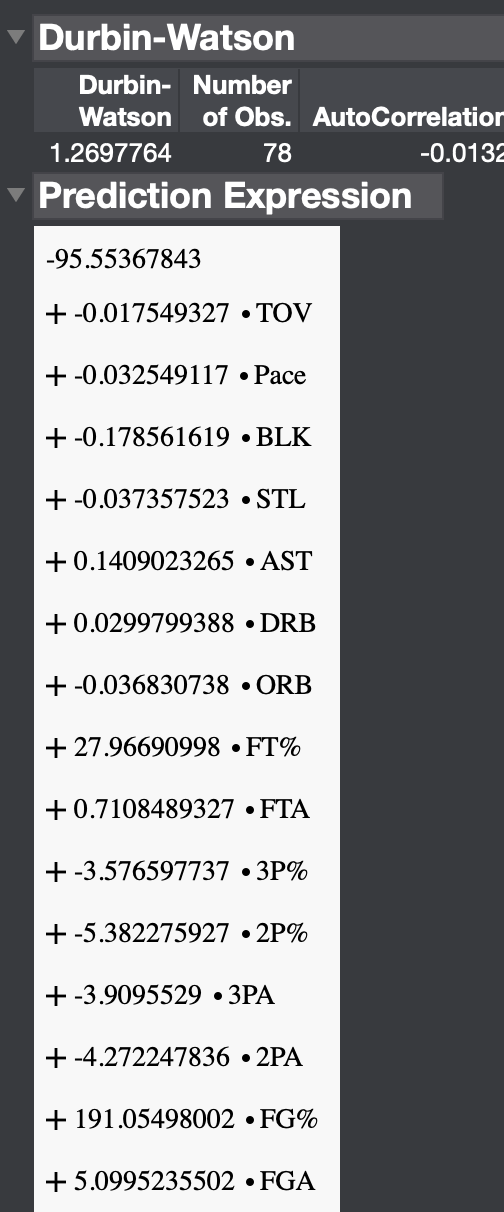


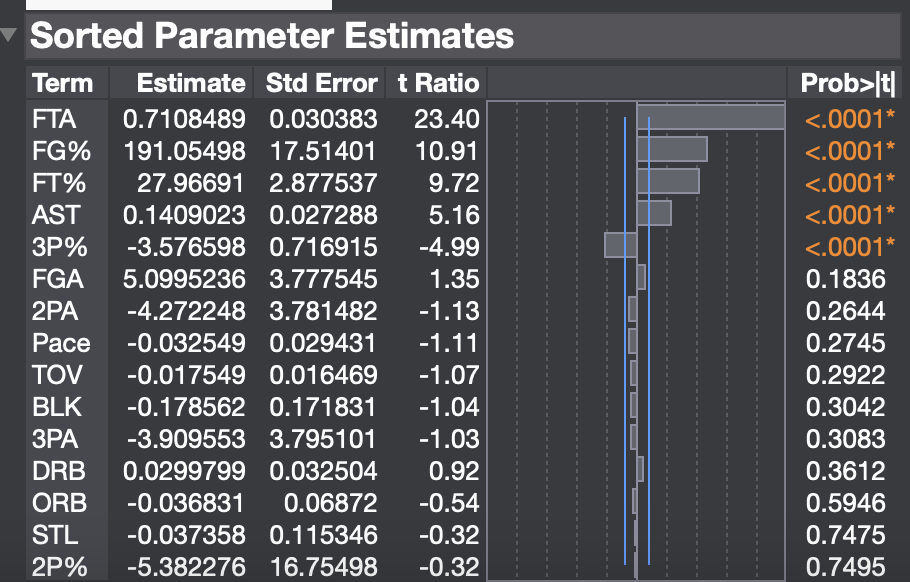






**Durbin-Watson Statistic and Prediction Expression**





Model chosen: **Simple Moving Averages** as it has the best balance between goodness of fit (SSR), predictive accuracy (low MAPE and MAE) and efficiency (moderate AIC and BIC)

Repository to all our JMP outputs and Scripts:

**https://github.com/ShreyasR26/Forecasting-Points-per-game/**