NBA Statistical Analysis

Logan T. Donlan

Western Governors University

Table of Contents

A. Project Highlights 3

B. Project Execution 3

C. Data Collection Process 6

C.1 Advantages and Limitations of Data Set 7

D. Data Extraction and Preparation 7

E. Data Analysis Process 9

E.1 Data Analysis Methods 9

E.2 Advantages and Limitations of Tools and Techniques 11

E.3 Application of Analytical Methods 15

F Data Analysis Results 16

F.1 Statistical Significance 16

F.2 Practical Significance 19

F.3 Overall Success 19

G. Conclusion 20

G.1 Summary of Conclusions 20

G.2 Effective Storytelling 21

G.3 Recommended Courses of Action 22

H Panopto Presentation 23

[References 2](#_Toc1410715514)3

# Project Highlights

**Organizational Need:**

This project addresses the organizational need of determining which team statistics influence win percentage, be it positively or negatively. This will be done on the full dataset that includes the regular season data from the 01-24 seasons. Then, the dataset will be split into 2 different eras (01-12 & 13-24) to determine how influences have changed over time.

**Project Scope:**

The project scope includes advanced statistic calculations, primary feature selection, determining influence of primary features to the target, secondary feature selection, determining influence of secondary features on primary features, discovering variance between eras, generating insights, and displaying these insights through a report including visualizations. It is important to note that this is regular season data. To identify features that lead to winning a championship, a similar analysis will need to be done on postseason data.

**Methodology:**

An agile approach allowed this project to be completed iteratively. Once the requirements of determining feature influence on the target of win percentage was set, the design, development, and testing phases were done multiple times to refine the outcomes.

**Tools:**

The coding for this project was done using the python language in a Jupyter notebook environment. The libraries used were pandas, seaborn, matplotlib, numpy, sklearn, statsmodels, math, and sys. The visualizations used in the presentation were generated using Tableau Public.

# B. Project Execution

**Project Plan:**

This project was completed without any changes to the initial plan. All goals, objectives, and deliverables were completed successfully as described in Task 2.

* **Goal:** Explain how features influence win percentage over the entire dataset and identify any variance in the different eras.
* **Objective A:** Calculate advanced statistics and choose which features to use in the model by examining the data through various methods, including a correlation matrix.
  + **Deliverable A1:** This objective will provide us with a dataset that avoids using features with high multicollinearity so that the model can be more accurate.
* **Objective B:** Use Lasso regression and bootstrapping on the selected features to identify regression coefficients with the target of win percentage.
  + **Deliverable B1:** After the model is fitted and refined with the bootstrapping method, a dataframe of primary regression coefficients will be returned.
* **Objective C:** Generate secondary regression coefficients on the primary features to identify which features lead to higher performance of the primary features using the Lasso regression model and bootstrapping.
  + **Deliverable C1:** This objective will deliver a dictionary containing individual dataframes for each primary feature. These dataframes will show the regression coefficients of each primary feature to the other features.
* **Objective D:** Split the dataset into 2 eras and repeat the process to discover variance in the coefficients.
  + **Deliverable D1:** The model will be fitted to the era 1 dataset and return a dataframe of primary regression coefficients.
  + **Deliverable D2:** Using the methods from objective c, secondary feature coefficients will be stored in a dictionary holding dataframes for each primary feature.
  + **Deliverable D3:** The era 2 dataset will be fit with the Lasso model and the bootstrapping method will be applied. This will give me the primary feature regression coefficients for the second era.
  + **Deliverable D4:** Repeating the methods used in objective c, I will gain insight into the secondary coefficients through a dictionary containing dataframes for each individual feature. These dataframes will explain the coefficients of the primary features to the other features.
* **Objective E:** Display these findings using simple, accurate visualizations.
  + **Deliverable E1:** Using Tableau, create multiple visualizations that accurately depict the outcome of the complete analysis.
  + **Deliverable E2:** Create a slideshow to present the findings and visualizations.

**Project Planning Methodology:**

The agile method was used to create this project. This methodology has 6 steps:

* **Requirements:** The requirements never had to be updated.
* **Design:** The design was updated multiple times. Originally everything was done in one notebook. This was updated to using 3 separate notebooks for the full and split datasets to keep everything clean and easy to interpret. Originally the functions were created within the notebooks. Instead of this, there is a separate script to house all the repeatable functions.
* **Development:** This part was iterated over many times. The parameters for the Lasso model had to be refined to work well across all datasets. Originally, the models were scored using Mean Squared Error (MSE), but this was changed to Root MSE for better interpretability.
* **Testing:** This phase was completed multiple times, but the testing methods themselves did not change. The results from this phase drove many of the changes in the design and development phases.
* **Deployment:** Once all other phases were completed successfully, the full analysis was done to generate insights.
* **Review:** The delivery of this project and all aspects leading up to it have been reviewed to create lessons learned to improve future projects.

**Project Timeline and Milestones:**

|  |  |  |  |
| --- | --- | --- | --- |
| Milestone | Projected Time | Start Date | End Date |
| Collect/Import/Study Data | **4 Hours** | **5/27/2024** | **5/28/2024** |
| Clean/Process Data | **8 Hours** | **5/29/2024** | **5/31/2024** |
| Model Training/Testing | **6 Hours** | **6/1/2024** | **6/3/2024** |
| Analyze/Tune Model | **8 Hours** | **6/4/2024** | **6/7/2024** |
| Develop Visualizations | **4 Hours** | **6/6/2024** | **6/8/2024** |
| Create Slideshow | **2 Hours** | **6/9/2024** | **6/9/2024** |
| Present Findings | **1 Hour** | **6/30/2024** | **6/30/2024** |

# C. Data Collection Process

**Data Selection and Collection:**

The dataset is called “NBA Team Stats” and can be found on Kaggle.com. It was authored by user Michael H. A link to this can be found in the “References” section. The dataset was imported into the Jupyter notebook environment using the pandas “read\_csv” method. Nothing was changed from the original plan.

**Data Governance:**

There were no issues concerning data governance with this dataset, as assumed from the original plan.

## C.1 Advantages and Limitations of Data Set

**Advantages:**

* **Clean Data:**  This dataset required minimal cleaning and had little inaccurate data.
* **Comprehensive Statistics:** This dataset contains most basic team statistics.
* **Ease of Access:** Kaggle offers free and easy access to datasets.

**Disadvantages:**

* **Advanced Statistics:** The advanced statistics had to be calculated, as they were not present.
* **Missing Data:** One column (personal\_fouls\_drawn) had to be dropped from the dataset for missing and inaccurate data.

# D. Data Extraction and Preparation

**Data Extraction:**

The data source was identified as Kaggle.com. This was then downloaded and loaded into the proper environment. No transformation was needed to load the data.

**Preparation Process:**

For clarity, here are the advanced statistic feature names and what they represent:

* ast\_to\_ratio: Assist to turnover ratio
* pct\_pts\_fts: Percentage of points from free throws
* ast\_fg\_ratio: Assist to field goal ratio
* eFG%: Effective field goal percentage

Data transformation began with dropping the “teamstatspk” column. This column represented an index of the dataset and was unnecessary for the analysis. Next, the “season” column was changed from a yyyy-yy format to yy, with yy being the last 2 digits. This was done using the pandas “str” and “astype” methods. The “Team”, “games\_played”, “wins”, “losses”, and “Min” columns were dropped, as they were not needed to complete the analysis. Then, the pandas “describe” method was used on each feature to determine data quality. The “field\_goal\_percentage”, “three\_point\_percentage”, and “free\_throw\_percentage” values were changed from whole numbers to decimals for better interpretability. Also, the “personal\_fouls\_drawn” feature had some rows with values ranging from 0-19, which is inaccurate. This feature was dropped from the dataset. After this, advanced statistics of assist/turnover ratio, points from free throws, percentage of points from free throws, assist/field goal ratio, and effective field goal percentage were calculated. The “plus\_minus” feature was dropped because it caused a heavy skew of the data. The “season” column was then dropped because it was not needed for the analysis. From here the target variable is dropped from the dataset. Next, a correlation matrix is generated using the “gen\_corr\_matrix” function to find features that correlate to each other. The first feature from each correlation was then dropped using the “drop\_redundant\_feats” function. From here, features like “assists”, “turnovers”, “points\_from\_free\_throws”, and “field\_goals\_attempted” were dropped, as they can be accounted for with other features. These features would have also caused a high degree of multicollinearity.

# E. Data Analysis Process

## E.1 Data Analysis Methods

**Lasso Regression Model:** A Lasso regression model was applied to the data to generate regression coefficients that were used to determine feature influence on the target. The models were generated using a custom function. This function used the train\_test\_split, GridSearchCV, Lasso, mean\_squared\_error, r2\_score, and StandardScaler methods from the sklearn library. The metrics used to verify the model will be described below. Once these metrics are verified, this model was used to accept or reject each null hypothesis. This type of model is appropriate because it performs automatic feature selection by shrinking less important coefficients to zero, thereby highlighting the most significant predictors while managing multicollinearity and enhancing model interpretability. These hypotheses were:

1. **Full Dataset Hypothesis:**
   1. Null Hypothesis- No features will significantly influence the target feature.
   2. Alternative Hypothesis- Certain features will influence the target feature.
2. **Split Era Datasets Hypothesis:**
   1. Null Hypothesis- There will be no variance in the regression coefficients between the 2 different eras.
   2. Alternative Hypothesis- There will be some variance in the regression coefficients between the 2 different eras.

**GridSearchCV:** The alpha parameter in a Lasso regression model determines regularization strength. The optimal alpha can differ based on different datasets and target features. To find the best alpha value for each model, the GridSearchCV method from the sklearn library was used. A grid of alpha values was created and then applied to each dataset before the model was generated. This was appropriate because of it’s ability to ensure model optimization across different datasets with differing targets.

**Variance Inflation Factor (VIF):** Although Lasso regression handles multicollinearity well, it is still important to mitigate this as much as possible. To address this, VIF scores were applied to each feature using a custom function. This function uses the variance\_inflation\_factor method from the statsmodels library. This was appropriate as it allowed for proper feature selection based on eliminating those variables with high multicollinearity.

**Bootstrapping with 95% Confidence Intervals:** The model was bootstrapped using a custom function that utilizes the train\_test\_split, resample, mean\_squared\_error, and r2\_score methods from the sklearn library. The regression coefficients were also calculated with confidence intervals of 95%, ensuring model interpretability and accuracy. This method is appropriate for it’s ability to determine the reliability of estimates while providing a robust, non-parametric method for assessing the variability and confidence of model parameters without assuming a specific data distribution.

**Root Mean Squared Error (RMSE):** RMSE was used as a metric to determine how the model predictions line up with what happened in the data. The value of this feature was measured against the mean of the target feature to realize model accuracy. This was an appropriate method as it helped determine practical significance of the regression coefficients returned by the model.

**R2:** R2 score was used as a metric to understand how well the selected features explained variance in the target. This was appropriate as a low R2 score suggests that the present variables do not fully explain changes in the target. A high R2 score would suggest that the present features explain target variance well.

## E.2 Advantages and Limitations of Tools and Techniques

**Tools**

**Python:**

* **Advantages:**
  + Integration capabilities
  + Easy readability
  + Built-in libraries for data preparation, analysis, and visualization
  + Reproducibility and automation
* **Disadvantages:**
  + Memory consumption and speed
  + Debugging and error handling
  + Learning curve

**Jupyter Notebook:**

* **Advantages:**
  + Interactive and incremental development
  + Visualization integration
  + Integration with machine learning frameworks
  + Data exploration and analysis
* **Disadvantages:**
  + Resource intensive
  + Version control challenges
  + Learning curve for advanced features

**Tableau Public:**

* **Advantages:**
  + Rich visualization capabilities
  + Data integration
  + Interactivity and exploration
  + Public sharing and collaboration
* **Disadvantages:**
  + Functionality constraints
  + Learning curve
  + Lack of advanced analytics

**Analytical Methods**

**Lasso Regression Model:**

* **Advantages:**
  + Feature selection
  + Handling multicollinearity
  + Improved prediction accuracy
  + Computational efficiency
* **Disadvantages:**
  + Model interpretability
  + Absence of confidence intervals
  + Loss of some feature information

**GridSearchCV:**

* **Advantages:**
  + Optimal hyperparameter tuning
  + Automation and efficiency
  + Reproducibility and documentation
* **Disadvantages:**
  + Computational intensity
  + Dependency on parameter grid
  + Not addressing data issues

**VIF:**

* **Advantages:**
  + Identification of redundant variables
  + Facilitation of model refinement
  + Guidance for data transformation
* **Disadvantages:**
  + Computational complexity
  + Threshold ambiguity
  + Not indicative of other issues

**Bootstrapping with 95% Confidence Intervals:**

* **Advantages:**
  + Robustness
  + Accuracy of estimates
  + Flexibility in model evaluation
  + Resilience to violations of assumptions
* **Disadvantages:**
  + Computational intensity
  + Lack of theoretical foundation
  + Potential misuse

**RMSE:**

* **Advantages:**
  + Sensitivity to large errors
  + Widely used and recognized
  + Error distribution insights
  + Comprehensive measure
* **Disadvantages:**
  + Sensitivity to outliers
  + Interpretation complexity
  + Scale dependence

**R2:**

* **Advantages:**
  + Interpretability
  + Model comparison
  + Supplementary information
  + Detection of overfitting
  + Performance benchmark
* **Disadvantages:**
  + Insensitive to overfitting
  + Does not indicate prediction accuracy
  + Cannot detect model bias

## E.3 Application of Analytical Methods

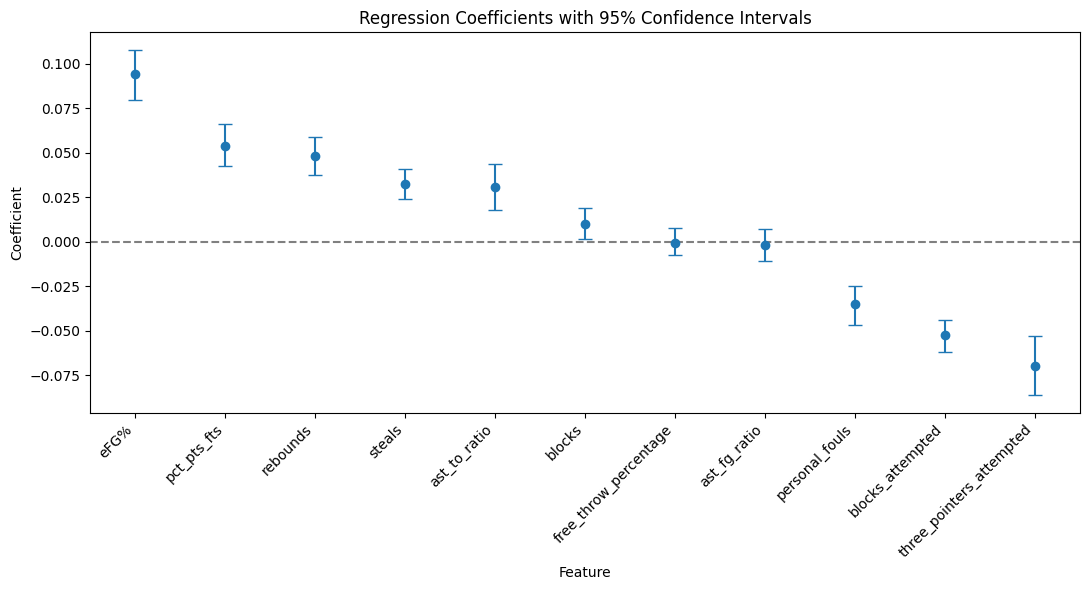
Once the data was preprocessed and the advanced statistics were calculated, a correlation matrix with a heatmap was created. A dataframe of features was created from this that had correlations of greater than 0.8. The base features present in the dataframe were then dropped from the dataset. Next, base features that could be accounted for with the advanced statistics were dropped. Based on the RMSE and VIF scores, supplementary features were dropped, and the model was regenerated. Once RMSE was below 20% of the target mean and VIF scores were less than 5, a horizontal bar chart was created to display the weight of each selected feature. The data is then bootstrapped 1000 times. This returned the average RMSE and R2 scores across all iterations, as well as the average regression coefficients. Confidence intervals of 95% allowed for the determination of which features were significant by eliminating those that contained 0 within their lower and upper ranges. An errorbar chart was then created to visualize these features and their intervals. To finish the analysis, a description of expected target change for each practically significant feature was given. To calculate this, a realistic increase of 2% was used. Once this was completed, the same steps were taken on each feature selected from the original target analysis to discover deeper influences. The only changes were to the analysis of the rebounds and three\_pointers\_attempted features. For rebounds, the offensive\_rebounds feature was dropped. The eFG% feature was dropped from the three\_pointers\_attempted analysis. These were dropped because they heavily skewed the distribution of regression coefficients. They are also known to contribute heavily to their perspective targets, and the objective was to find less obvious connections between variables.

Next, the dataset was split into 2 eras, with era 1 consisting of the 2001-2012 seasons, and era 2 containing the 2013-2024 seasons. There were 356 and 360 rows of data in each era, respectively. The analysis of both eras was completed in the same way as stated above on the full dataset. Again, the offensive\_rebounds feature was dropped from the rebounds feature analysis in both eras. In both eras, the eFG% variable was dropped from the three\_point\_percentage feature analysis because of the heavy skew it caused when left in. In era 2, the eFG% was dropped from the three\_pointers\_attempted feature analysis for the same reason.

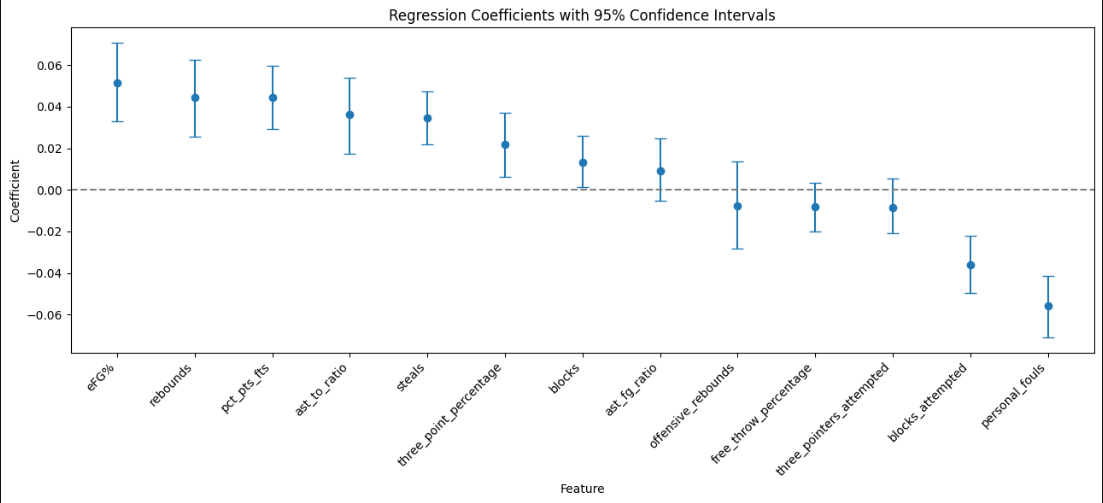
# F Data Analysis Results

## F.1 Statistical Significance

The model used for the analysis was a supervised regression technique known as Least Absolute Shrinkage and Selection Operator, or Lasso. These coefficients explained the influence of the specified feature on the target variable. To determine their statistical significance, the model was bootstrapped 1000 times while generating confidence intervals of 95%. If a feature contained a 0 value within the range of the confidence intervals it was deemed statistically insignificant. The conclusions drawn from the determination of statistical significance of each feature allowed me to reject both null hypotheses. The results are shown in the figures below:



The figure above displays the regression coefficients after bootstrapping has occurred with confidence intervals of 95%. The free\_throw\_percentage and ast\_fg\_ratio features are both statistically insignificant, however the rest of the features show clear signs of influence on the target of win percentage. This allowed the first null hypothesis, which stated that no features will significantly influence the target feature, to be rejected.



The chart overhead shows the regression coefficients from era 1 once they have been bootstrapped and assigned confidence intervals of 95%. The ast\_fg\_ratio, offensive\_rebounds, free\_throw\_percentage, and three\_pointers\_attempted features were all statistically insignificant. The remaining features were statistically significant.

A graph with blue and white dots

Description automatically generated

This chart was from era 2, again showing the regression coefficients after bootstrapping took place and confidence intervals of 95% were generated. Here, the blocks, free\_throw\_percentage, offensive\_rebounds, and ast\_fg\_ratio were all thrown out because of their statistical insignificance. Still, there was a clear difference in the features and their influence on the target. This allowed the rejection of the null hypothesis that there would be no variance between the two different eras.

## F.2 Practical Significance

The practical significance of the Lasso model was determined by the RMSE after the data went through the bootstrapping process. The average of the training and testing RMSE was calculated and measured against the target mean. This value needed to be less than 20% of the target mean to prove the coefficients derived from the model as practically significant. Each model generated throughout the project followed these guidelines.

## F.3 Overall Success

The first factor in gauging success for this project was the accurate and informative calculation of advanced statistics. These stats were not only generated correctly but provided valuable insight, as well. Next, the bootstrapped model needed to display significant accuracy in predicting target influence. To measure this, the average RMSE was taken from the training and testing sets and compared to the target mean. Every model had an average RMSE of less than 20% of the target mean, leading to successful model deployment. The visualizations generated throughout the analysis and through Tableau were all considered successful, as they depict accurate and informative data. Overall, this project generates valuable insights for all types of businesses. This demonstrates a high level of success.

# G. Conclusion

## G.1 Summary of Conclusions

**Full Dataset Analysis:**

The positive influences to win\_percentage in descending order were eFG%, pct\_pts\_fts, rebounds, steals, ast\_to\_ratio, and blocks. The negative factors from least to greatest were personal\_fouls, blocks\_attempted, and three\_pointers\_attempted. Here is an example of how each feature effects the target:

* A 0.010 unit increase in eFG% results in an approximate 0.0335 (or 3.35 percentage points) change in the target variable.
* A 0.004 unit increase in pct\_pts\_fts results in an approximate 0.0085 (or 0.85 percentage points) change in the target variable.
* A 68.781 unit increase in rebounds results in an approximate 0.0143 (or 1.43 percentage points) change in the target variable.
* A 12.188 unit increase in steals results in an approximate 0.0054 (or 0.54 percentage points) change in the target variable.
* A 0.032 unit increase in ast\_to\_ratio results in an approximate 0.0043 (or 0.43 percentage points) change in the target variable.
* A 7.869 unit increase in blocks results in an approximate 0.0012 (or 0.12 percentage points) change in the target variable.
* A 33.376 unit increase in personal\_fouls results in an approximate -0.0067 (or -0.67 percentage points) change in the target variable.
* A 7.869 unit increase in blocks\_attempted results in an approximate -0.0070 (or -0.70 percentage points) change in the target variable.
* A 36.573 unit increase in three\_pointers\_attempted results in an approximate -0.0039 (or -0.39 percentage points) change in the target variable.

This means that a 1% increase in eFG% leads to a 3.35% increase to win\_percentage. Conversely, 37 more three-point attempts diminish win percentage by -.39%. This would appear to say that shooting less from 3 would lead to more winning, but it’s important to look deeper into the data. When running the model to analyze feature influence of eFG%, three\_pointers\_attempted was the largest positive influence. It seems counterintuitive, but it points to a need for balance. Shooting more 3’s for the sake of attempts is the wrong answer. Quality looks and efficiency are vital to benefit from the extra attempts. There are many examples like this one throughout the secondary feature analysis. The key takeaway is not looking at surface data and basing decisions solely from that. It’s always worth taking a deeper look.

**Split Dataset Analysis:**

The point here was to determine what differences occur when splitting the dataset into 2 different eras. eFG% and rebounds were the top two positive influences in both eras. The evolution of the game starts to show when three\_point\_percentage takes the place of pct\_pts\_fts from era 1 to the era 2. Clearly, much of the league has turned to a 3-point driven offense. The influence of steals has also risen from era 1 to era 2, showing the increased pace at which teams play. Although there wasn’t much variation between the split datasets, there was variation, nonetheless. Even when looking at the secondary feature analyses, there are slight differences in feature influences between eras.

## G.2 Effective Storytelling

**Correlation Matrix with Heatmap:**

The correlation matrix helps determine which features have strong correlations. The heatmap does a great job of displaying the connections. These were generated using the matplotlib.pyplot library in python while in a Jupyter notebook environment.

**Horizontal Bar Chart:**

The horizontal bar charts found in each feature analysis effectively presents the weights of each feature and their corresponding regression coefficient. The charts were created in a Jupyter notebook using the matplotlib.pyplot library within python.

**Error Bar Chart:**

The error bar charts help to visualize the statistical significance of each feature selected in every analysis. The matplotlib.pyplot library in python was used to create these in the Jupyter notebook.

**Tableau Story:**

The story put together using Tableau contains numerous bar, line, and shaded area graphs to depict findings and trends throughout the analysis.

## G.3 Recommended Courses of Action

**Recommendations based on the analyses:**

1. Efficient shooters are vital to team success. Effective field goal percentage was the highest positive correlation in all 3 datasets. Looking into this feature, 3-point attempts and percentage are the highest positive influences on eFG%, except in era 1 where the assist to turnover ratio is above attempts. We can also see through 3 point and scoring trends that all metrics point to the evolution of the game. This evolution displays the growing value of the 3. The key thing here is to be efficient, as 3-point attempts was a negative influence across all datasets. Look for players that take good shots from beyond the ark. The attempted emulation of Stephen Curry is inefficient and typically leads away from success. He must be considered the exception to the rule.
2. Another thing to look for is players with high assist to turnover ratios. This feature seems to display a high level of basketball IQ and court vision. Although it didn’t appear as a primary feature in era 2, it is present in most positive influences across the datasets. This indicates that players and teams that maintain higher ratios are good decision makers and will help not only to create efficient offense, but also make smart decisions on the defensive end.

# H. Panopto Presentation

[**https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=be3f5b65-b29c-466c-8d34-b1ae0172eb43**](https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=be3f5b65-b29c-466c-8d34-b1ae0172eb43)

# References

H, M. (2024, April). NBA Team Stats. Retrieved from Kaggle: https://www.kaggle.com/datasets/mharvnek/nba-team-stats-00-to-18

Haefner, J. (2016). 9 Stats That Every Serious Basketball Coach Should Track. Retrieved from

Basketball Breakthrough: https://www.breakthroughbasketball.com/stats/9\_stats\_basketball\_coach\_should\_track.html

Sullivan, R. (2020, April). NBA’s Most Valuable Statistic Discovered: How To Predict Team

Wins With 95% Accuracy. Retrieved from SPGN: https://www.sportsgamblingpodcast.com/2020/04/20/nba-most-valuable-statistic/

Peterson, D. (2020, May). How Different Metrics Correlate with Winning in the NBA over 30

Years. Retrieved from Medium: https://towardsdatascience.com/how-different-metrics-correlate-with-winning-in-the-nba-over-30-years-57219d3d1c8