# Data Science Executive Summary

This project investigated the question “What measurable factors contribute to success among ATP (Association of Tennis Professionals) tournament champions spanning 2014 to mid-2024”. The datasets, which were sourced from Kaggle, were manually compiled and appended to each other in Excel, before being imported into SAS for cleaning and analysis. This project centres around three null hypotheses:

* Hypothesis 1: Player’s heights have no impact on the numbers of aces achieved
* Hypothesis 2: Match length does not differ across different tournaments
* Hypothesis 3: Player handedness does not impact number of aces achieved

To solve these hypotheses, data was filtered to only include final round matches and ensure only the winners’ data was included in analysis, and missing values were dealt with by imputing mean values.

To answer the first hypothesis regarding player height and number of aces, Pearson Correlation was used. A moderate positive relationship was found (r=0.486, p < 0.0001) which suggests that taller players tend to serve more aces.

For the second hypothesis about match length and tournament, a one-way ANOVA was conducted which compared match length across the tournaments. The results ( F = 3.23, p < 0.0001) were statistically significant, indicating that tournament structure or conditions may influence match duration.

For the third hypothesis which focussed on the handedness of a player, and the number of aces achieved, an independent T-Test was used. This also revealed statistically significant results ( p = 0.0009), suggesting that handedness plays a role in the number of aces won.

Overall, the findings highlight how statistical methods can uncover meaningful trends in sports data. These insights could inform coaching strategies, player development, and tournaments planning. The project also demonstrates the importance of careful data preparation and the use of appropriate tools to support analysis.

# Data Infrastructure and tools

For this project, Excel, SAS, and Power BI (PBI) were used to combine, transform, analyse and visualise ATP data from Kaggle (www.kaggle.com/datasets/guillemservera/tennis). This data was available on an annual basis, meaning there were eleven spreadsheets, one for each year from 2014 to 2024.

Excel was chosen to append the spreadsheets together into a master sheet. This choice was based on its ease of use and familiarity. The master sheet was small enough for Excel to handle comfortably, but it is important to note that Excel can struggle with much larger datasets, which can lead to performance issues and data loss (Mehar, 2023). Other tools, such as SAS, could have been used for the consolidation stage, as it can handle much larger datasets, however Excel was sufficient and allowed for quick consolidation. Using Excel also allowed data quality checks to be carried out at this stage, ensuring that all datasets contained the same number of columns in the same order.

Once the data was combined, it was imported into SAS for cleaning and analysis. SAS was chosen for its strong analytical capabilities (Geeks for geeks, 2025). Its simple syntax, (Pawar, 2022), made tasks like filtering, and handling missing values easy to deal with. Other tools, such as R could have been used as it also has analytical abilities and strong visualisation capabilities (Geeks for geeks, 2025), however, as this tool isn’t as familiar to the developer, it was decided that it would be quicker and easier to continue the project with SAS.

After analysis, the results were exported into Excel then uploaded into PBI. PBI was chosen for its user-friendly interface and strong visualisation features (Patnaik, 2025). While Tableau can handle larger datasets and has advanced visuals, its steeper learning curve made PBI a more practical choice (geeks, 2025).

Together, these tools allowed the data to be combined, cleansed, transformed, analysed, and displayed in a visual dashboard quickly, and easily.

# Data Engineering

The project followed a structured ETL (Extract, Transform, Load) pipeline, which was designed to prepare ATP data for statistical analysis and visualisation (See Appendix A for ETL Diagram). Planning began by identifying the relevant data sources. The data was taken from Kaggle and the years that were needed were extracted. Key variables were identified once the projects hypotheses were decided.

Extraction of the data was completed by opening each dataset in Excel and copying and pasting the data into one sheet. During this stage quality checks were completed, ensuring data from one year to the next contained the same necessary data points, and the formatting was consistent. This was suitable for the scale of the project, however it would be inefficient if there were more datasets, or if the datasets themselves contained more entries.

This data was imported into SAS where transformation was completed (See Appendix B for the SAS code). This involved filtering the data to include only ‘final round’ data and removing unnecessary columns. Filtering ensured focus on the right data. The remaining columns were renamed for clarity throughout the code. Once this step was complete the remaining data was checked for nulls and dealt with by using mean imputation. This decision was based on descriptive statistics which showed the mean was a reasonable estimate for the missing values in player height, match length, and number of aces. Median imputation could have been used if the dataset had any anomalies, however, as this wasn’t the case, mean imputation was selected as this is the option that considers all the available data (Vaj, 2024).

The data was then loaded into Excel, before being loaded into PBI for visualisations.

Each step in the ETL process contributed to the overall success of the project by ensuring the data was clean, consistent, and ready for analysis and visualisation. This approach meant the project is repeatable and scalable.

# Data visualisation

PBI was used to present the results of the hypotheses in a simple, clear, and accessible way. The dashboard which was titled “What makes a great tennis player”, was designed to summarise the findings from each hypothesis test, making it easy for users to interpret the outcomes without needing to review the raw statistical outputs.

The dashboard was designed with storytelling in mind, designed to guide the viewer through the findings. At the top is a summary table which presents each hypothesis, test type, test statistic, p-value, and conclusion (see Figure 1).

A screenshot of a graph

AI-generated content may be incorrect.

Figure 1: Statistical Summary Table

This allows users to quickly understand which relationships were significant. For example, the Pearson correlation between player height and number of aces shows a moderate positive (r=0.486, p<0.0001) while the ANOVA test confirms that match length varies significantly across tournaments (F-3.23, p<0.0001). The t-test comparing aces by handedness also shows a statistically significant difference. Presenting these results in a structured format supports transparency and help communicate finding to both technical and non-technical audiences.

As well as having the summary table, a column-chart titled “Aces by handedness” (see Figure 2) was used to compare the average number of aces served by right- and left-handed players.

A blue rectangular object with text

AI-generated content may be incorrect.  
Figure 2: Column Chart Showing Aces by Handedness

A scatter plot (see Figure 3) titled “Height vs Aces” showed the relationship between player height and the average number of aces served. A trend line was added to further highlight the correlation. Another column-chart (see Figure 4) titled “Match length vs tournament” was used to display the average length of play at each tournament.

A graph of height versus aces

AI-generated content may be incorrect.*Figure 3: Scatter Plot with Trend line showing Aces by Height*

A graph of a match

AI-generated content may be incorrect.

Figure 4: Column Chart showing Match Length vs Tournament

Column charts are ideal for comparing categorical data, such as player hand, or tournament, and so was the perfect choice for showing “Aces by Handedness”, however they can take up a lot of room , and when there are many categories to display, it may not be possible to compare all of the data together (ZoomCharts, 2023). This was the case for “Match length vs tournament”. Scatter charts are effective at showing relationships between two continuous variables; however, caution must be taken if there are many data points being displayed as viewers may not be able to draw insightful conclusions (ZoomCharts, 2023).

On the left side, slicers were added to allow users to filter the results depending on specific interests and to make the dashboard dynamic.

The dashboard was intentionally kept simple to focus on clarity. For the purposes of this project, colour was intended to match the official ATP colours which are “Cool Black”, and “Vivid Cerulean” (SchemeColor, No date). In this dashboard, as all the conclusions were “Reject Null Hypothesis” having different colours wasn’t important. However, different colours could be used to highlight different outcomes and could have improved visual impact. (See figure 5).

A screenshot of a computer

AI-generated content may be incorrect.Figure 5: Complete PBI Dashboard

# Data Analysis

The central question of this project was: What measurable factors contribute to success among ATP tournament champions? To explore this, three hypotheses were developed, each targeting a different performance-related variable.

For Hypothesis 1, a Pearson correlation test was used. This was chosen because both variables are continuous and normally distributed. The result (r = 0.486, p < 0.0001 (See Figures 6&7)) showed a moderate positive correlation, suggesting that taller players tend to serve more aces. A Spearman correlation could have been used if the data had shown non-linear patterns or outliers (Kumar, 2023).

A screenshot of a number

AI-generated content may be incorrect.

Figure 6: Screenshot of Pearson\_Stats Output

A screenshot of a computer

AI-generated content may be incorrect.

Figure 7: Screenshot of Pearson\_Results Output

For Hypothesis 2, a one-way ANOVA was performed. This test is suitable for comparing means across multiple groups. The result (F = 3.23, p < 0.0001 (See Figures 8&9)) indicated that match duration varies significantly depending on the tournament. A Kruskal-Wallis test could have been considered if ANOVA assumptions were violated (Bobbit, 2022).



Figure 8: Screenshot of ANOVA\_Stats Output

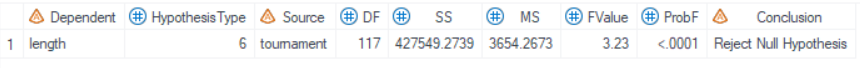


Figure 9: Screenshot of ANOVA\_Results Output

For Hypothesis 3, an independent sample t-test was conducted. The results (t = -3.34 / -3.58, p = 0.00089 / 0.00050 (See Figures 10&11)) showed a significant difference, suggesting that handedness affects serving performance. Welch’s t-test could have been used if variances were unequal (Lee, 2025).

A screenshot of a computer

AI-generated content may be incorrect.  
Figure 10: Screenshot T-Test\_Stats Output

A screenshot of a calculator

AI-generated content may be incorrect.

Figure 11: Screenshot of T-Test\_Results Output

Ethical and legal considerations were addressed throughout the project. The dataset was publicly available and did not contain sensitive information. No data points were removed, and mean imputation was used only where it was justified by the distribution. Bias was minimised by using standardised methods and treating all variables consistently.The structured methodology ensured reproducibility, while the findings offer practical value for player development and match strategy.

With such a rich dataset, there’s strong potential to expand the dashboard further—such as analysing match-by-match data rather than just finals. Future iterations could also benefit from automation and AI integration, which would streamline analysis and uncover deeper patterns. Overall, the project demonstrated how good planning, tool selection, and statistical rigour can lead to meaningful outcomes in sports analytics.

# References

BI Connector Team, 2025. *BI Connector.* [Online]   
Available at: https://www.biconnector.com/blog/top-10-pros-cons-of-power-bi/  
[Accessed 17 August 2025].

Bobbit, Z., 2022. *Statology.* [Online]   
Available at: https://www.statology.org/kruskal-wallis-test/  
[Accessed 15 August 2025].

Geeks for geeks, 2025. *geeks for geeks.* [Online]   
Available at: https://www.geeksforgeeks.org/blogs/sas-vs-r-vs-python/  
[Accessed 18 August 2025].

geeks, G. f., 2025. *Geeks for Geeks.* [Online]   
Available at: https://www.geeksforgeeks.org/blogs/power-bi-vs-tableau/  
[Accessed 18 August 2025].

Kumar, A., 2023. *Analytics Yogi.* [Online]   
Available at: https://vitalflux.com/pearson-vs-spearman-choosing-the-right-correlation-coefficient/  
[Accessed 14 August 2025].

Lee, S., 2025. *Number Analytics.* [Online]   
Available at: https://www.numberanalytics.com/blog/guide-using-welchs-t-test-stats  
[Accessed 15 August 2025].

Mehar, A., 2023. *Alma Better Bytes.* [Online]   
Available at: https://www.almabetter.com/bytes/articles/advantages-and-disadvantages-of-ms-excel  
[Accessed 17 August 2025].

Patnaik, M., 2025. *Alma Better Bytes.* [Online]   
Available at: https://www.almabetter.com/bytes/articles/advantages-and-disadvantages-of-power-bi  
[Accessed 25 August 2025].

Pawar, M., 2022. *Tecklearn.* [Online]   
Available at: https://www.tecklearn.com/advantages-and-disadvantages-of-sas-programming-language/  
[Accessed 18 August 2025].

SchemeColor, No date. *SchemeColor.* [Online]   
Available at: https://www.schemecolor.com/association-of-tennis-professionals-atp-world-tour-logo-colors.php  
[Accessed 16 August 2025].

Vaj, T., 2024. *Medium.* [Online]   
Available at: https://vtiya.medium.com/when-to-use-mean-median-mode-imputation-b0fd6be247db  
[Accessed 17 August 2025].

ZoomCharts, 2023. *Power BI Chart Types: A Comprehensive Guide.* [Online]   
Available at: https://zoomcharts.com/en/microsoft-power-bi-custom-visuals/blog/power-bi-chart-types-choosing-the-right-visuals-for-your-data  
[Accessed 16 August 2025].

# Appendix

**Appendix A: ETL Pipeline Diagram**

A diagram of a software system

AI-generated content may be incorrect.

**Appendix B: SAS Code for Data Transformation**

**data** initial;

set work.ATP;

**run**;

**proc** **contents**

data=work.ATP;

**run**;

**data** columns;

set work.ATP(keep=tourney\_name round winner\_hand winner\_ht w\_ace minutes

rename=(

tourney\_name = tournament

round = round

winner\_hand = hand

winner\_ht = height

w\_ace = ace

minutes = length));;

where round = "F";

**run**;

**Data** check\_nulls;

set columns;

if missing(tournament) or missing(hand) or missing(height) or missing(ace) or missing(length) then output;

**run**;

**proc** **means** data= columns n mean median std min max;

var height length ace;

output out=mean\_data mean=mean\_height mean\_length mean\_ace;

**run**;

**data** cleansed;

if \_n\_ = **1** then set mean\_data;

set columns;

if missing(height) then height = mean\_height;

if missing(length) then length = mean\_length;

if missing(ace) then ace = mean\_ace;

keep tournament hand height length ace;

**run**;

/\*ANOVA Analysis\*/

ods output ModelANOVA=Anova\_Stats;

**proc** **ANOVA**

data=cleansed;

class tournament;

model length = tournament;

**run**;

ods output close;

**data** ANOVA\_Results;

set Anova\_Stats;

length Conclusion $**50**;

if Probf < **0.05** then Conclusion = "Reject Null Hypothesis";

else Conclusion = "Fail to Reject Null Hypothesis";

**run**;

**proc** **print** data=ANOVA\_Results noobs;

title "ANOVA Results";

**run**;

/\*Pearson Analysis\*/

ods output PearsonCorr=Pearson\_Stats;

**proc** **corr** data=cleansed;

var height ace;

**run**;

**data** Pearson\_Results;

set Pearson\_Stats;

length Conclusion $**50**;

if Prob < **0.05** then Conclusion = "Reject Null Hypothesis";

else Conclusion = "Fail to Reject Null Hypothesis";

**run**;

/\*T-Test Analysis\*/

ods output TTests=TTest\_Stats Equality=TTest\_VarTest;

**proc** **ttest**

data=cleansed

alpha=**0.05**;

class hand;

var ace;

**run**;

**data** TTest\_Results;

set TTest\_Stats;

length Conclusion $**50**;

if Probt < **0.05** then Conclusion = "Reject Null Hypothesis";

else Conclusion = "Fail to Reject Null Hypothesis";

**run**;