INTERNSHIP REPORT

-Ankita Bhardwaj (ab4685)

EXECUTIVE SUMMARY:

This report is prepared for giving an overview of my 10 week internship program at the United States Tennis Association where I worked in the Data Analytics team of the Pro-Tennis department. I worked on various projects of which some were started by me from the scratch while some were additions and modifications to the existing work. I worked on analyzing the data for the US Open Umpire department and the Player Development Department. Umpire department is for crew allocation and maintenance at the US Open whereas Player Development Department is to train and grow US Players for the US Open.

For conducting all the analysis, the data used were from two different sources. One of the source is called as 'Hawkeye' and the other is called as 'SMT'. Both the sources collect data from the live US Open matches and give it to USTA. There are hundreds of tables in both the data sources which were used by me to extract the relevant information. I will be using the names of these sources multiple times in this report.

As I worked in the Data Analytics Team, I was mainly dealing with data analysis tools such as SQL, Tableau and Python. I was using SQL for extracting relevant data from hundreds of tables, Tableau for creating a visualizations and final products for the Umpire and Player Development Teams and Python for Data Cleaning. I will be elaborating on my projects and the usage of these tools in detail in this report.

The overall experience of working at the USTA was amazing and I got to learn a lot. Everyday was a new challenge but was normally structured in terms of time. My work and Department plays a crucial role in making US Open experience better for everyone and in promoting the growth of US Tennis in particular. I will elaborate on my work relevance further in this report.

To sum up, this report contains detailed information about my 10 weeks journey at the United States Tennis Association and about the relevance of the work done in this duration.

INDEX

INTRODUCTION:

United States Tennis Association is a not-for-profit organization which has a mission to promote and develop the growth of tennis. The organization is also the governing body for Tennis in the United States and it works to promote and develop the sport from the beginner levels to the professional levels. The headquarters of USTA is in the White Plains and there are 17 overall geographical sections where it has its base. This organization is responsible for conducting the US Open which is a hard court tournament and is also a grand slam which is the last grand slam conducted every year, chronologically. US Open is conducted in the New York City and this year, the main draw matches starts from 26th August and will end on September 8th. I worked for 10 weeks for USTA and the coming US Open in their Data Analytics Department as a Graduate Summer Intern.

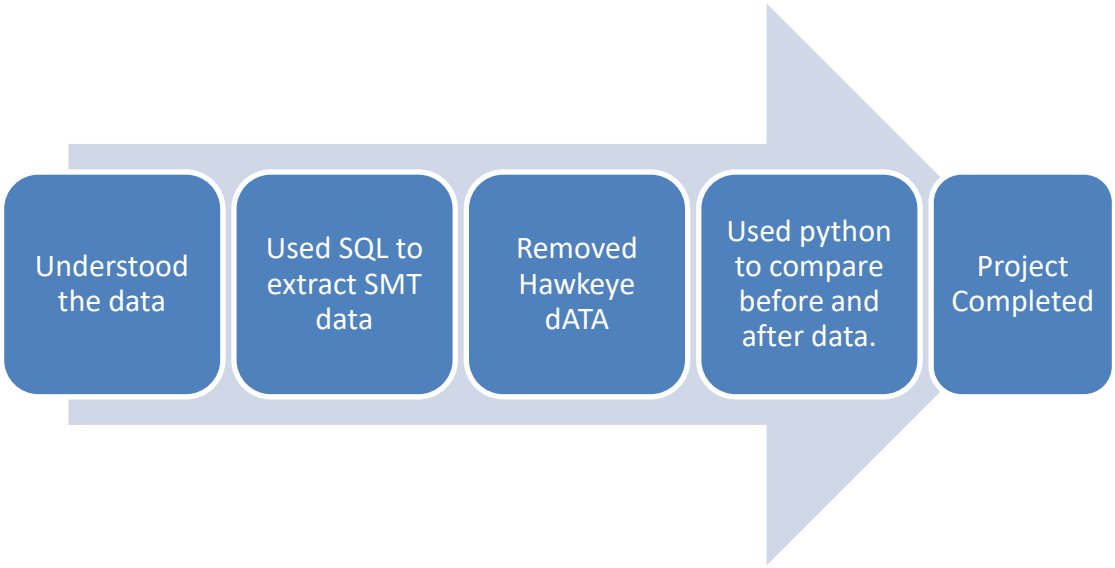
PROJECTS AND RESPONSIBILITIES: (chronologically)

Project 1: Using SQL to replace the Umpire Crew Start time from Hawkeye to SMT. (Week 1 and 2)

Background: USTA Data Analytics team had already created a SQL code which returned the details of Umpire Crew Allocations. This was done for optimizing the crew allotment in the US Open such that the errors made by the line umpires, in particular, could be minimized. The SQL code returned the start time of crew as indicated in the Hawkeye database but USTA wanted to change this time to the one indicated in SMT as SMT is more accurate and populated for this task.

My Work: Although the task seems easy but there were several challenges in completion of this project. The first challenge was to understand the databases and everything about USTA. As I did not know about the tables that existed, the form in which data was stored and what type of data does USTA deals with, it was challenging to find the location of the data and to form the join conditions based on that. Also, the level of SQL used was very high and I did not have experience in dealing with such large codes of SQL which posed a challenge for me. Another challenge that I faced was the reduction in number of rows when

SMT time was added and Hawkeye time was removed. This should not have happened as I just replaced one value with the other but on digging deep I found that Hawkeye had duplicate rows with change in just one column and that was an error. I used python to clean the data and to compare both the data sets when the time column was removed. After removing duplicate values and performing the operations again, I got what I needed. The following flow chart summarizes my work for this project.



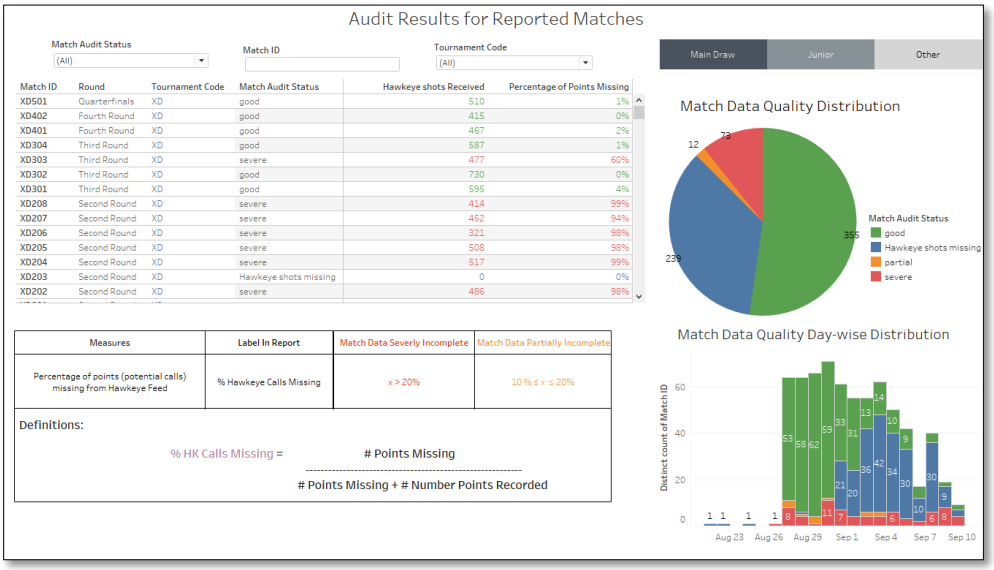
Impact of my work: The crew allotment data is used in several USTA dashboards and by changing the Hawkeye crew start time to SMT's time, I was able to provide more accurate data to the Umpire team for better crew allotment. One of the reasons I say this is because if the crew time is null then the data for any line umpire could be biased as additional data can increase or decrease the accuracy and hence, this potential bias had to be eliminated.

Project 2: Creating an Audit Status Dashboard(Week 2,3,4)

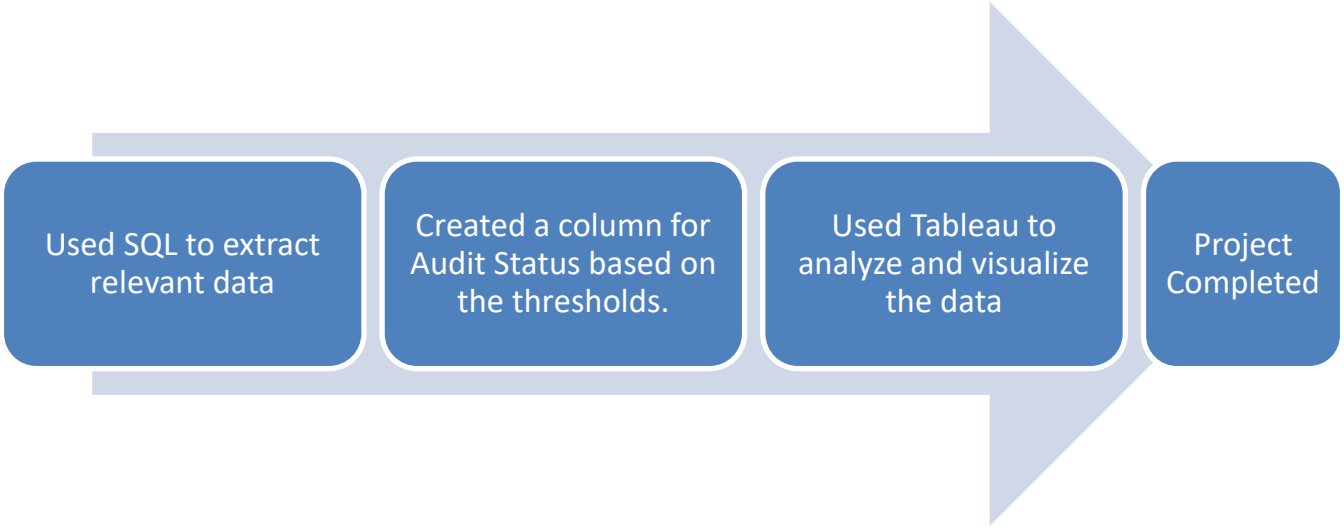
Background: As mentioned previously, USTA receives data from Hawkeye and SMT and all the analysis is based on the data received from them. However, the data sent by them are not 100% complete. There are several matches in which data is missing by a large amount and due to this, the results obtained could be biased. Hence, USTA wanted to evaluate about the incomplete data by these sources.

My Work: This work had to be done in tableau by connecting it to Microsoft SQL Server. The first task was to gather the scattered data at one place and I obtained that by using SQL and created a final SQL code containing all the code snippets. After gathering the data, I

used tableau to visualize them and make some sense of them. This was my first experience in building a full-fledged dashboard on Tableau on my own and hence, felt a little challenging. One more challenge that I faced was percentages more than 100 which is not practically possible and on digging deep, I found that there was error from Hawkeye part in entering the number of shots and I had to include it under 'severe' points directly. The dashboard looks something like this:



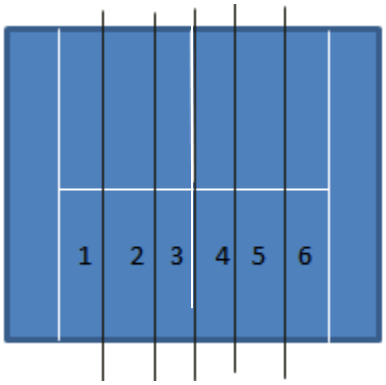
As it can be seen in the dashboard, depending on the % of calls missing (missing + Incomplete –common), the match audit status is of 4 types: Complete (<10% missing values), Partial (missing values between 10% and 20%), Severe (>20% missing values) and Hawkeye shots missing where there are no shots entered by Hawkeye. The following flow chart summarizes my work for this project.



Impact of my work: This dashboard has multiple purposes. One purpose is general assessment of the quality of data received by USTA. Now, USTA will know about which matches have what amount of data and can used the information to talk to Hawkeye and SMT about the same. Another use is that it can be incorporated in the other dashboards (almost all player development dashboards) to eliminate bias. I have included it on one of the dashboards and I will elaborate about it in project 4.

Project 3: Adding details to Serve % dashboard.(Week 4)

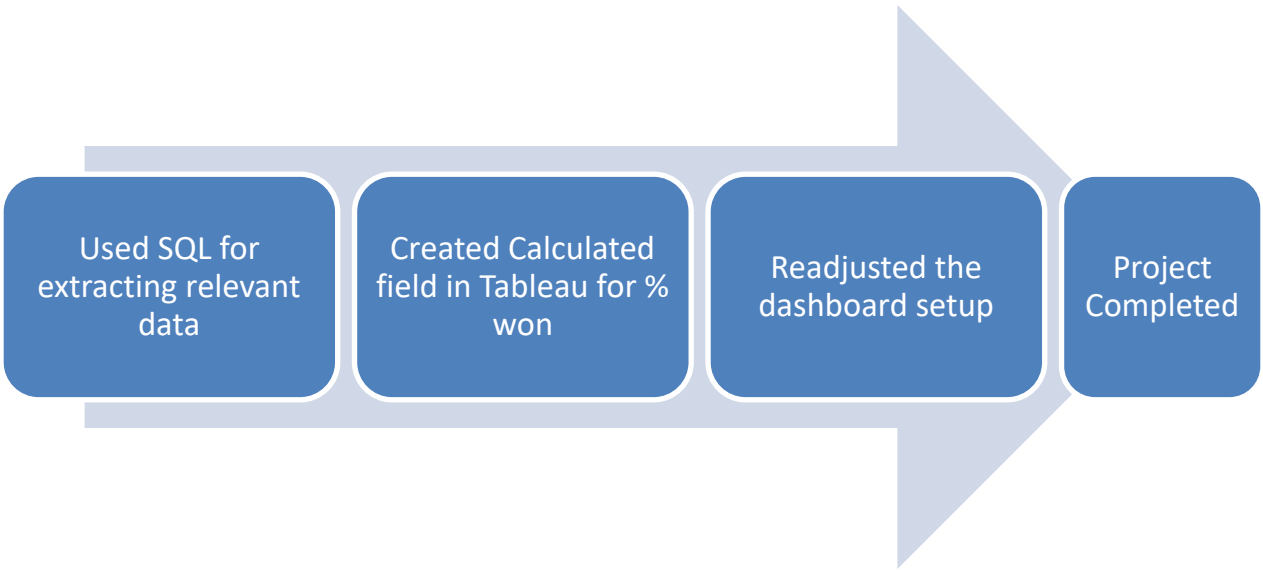
Background: USTA have created a dashboard for analyzing the serves that falls in a particular block and the blocks looks something as follows:



There were several calculations such as number of serves landed in a box (1 to 6) and % of serves were already present but USTA wanted me to add one new calculation : % points won for serves in that box.

My Work: Inserting the % of points won for serves in that box is not that easy as it looks as new data has to be added using SQL to an already existing complex code in addition to understanding the complex code. Also, the worksheets were manually adjusted in a way that the dashboard looks coherent but even a slight change in data would change the entire set up. For this short project, I used SQL to extract necessary data, calculated the required field and readjusted the dashboard to match it to the original set up. The following flow chart summarizes my work for this project:

(Dashboard image cannot be inserted due to sensitive data))



Impact of my work: This dashboard is used by the player development team to guide the players on where to serve and the additional metric of % of points won is extremely helpful for the team to decide on where to serve more and hence, practice to serve more to win more points.

Project 4: Adding Audit Status to Officiating (Umpire) Dashboard. (week 5)

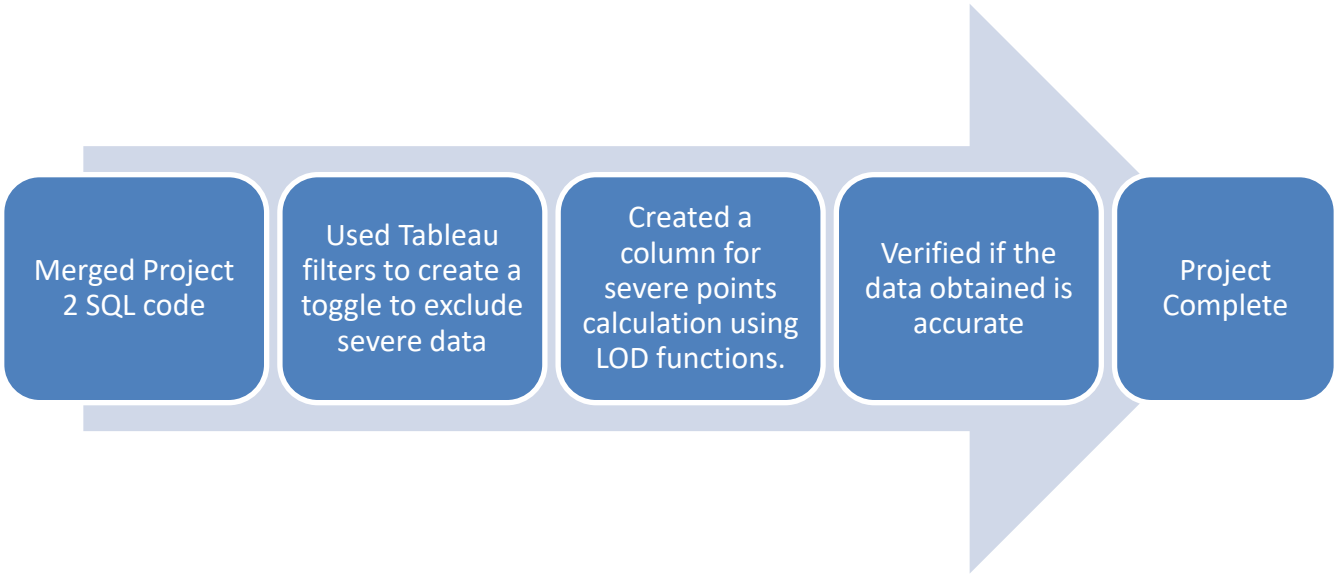
Background: The officiating dashboard is used in the US Open by the lead Referee for crew allotment in all the matches. It contains various analyses of data, one of which is the accuracy of any line umpire for different positions. However, the accuracy calculation also includes the matches where there were a lot of missing data ('severe' as said in project 2). Hence, USTA wanted me to include a column stating how many points, for the individual line umpires, are from the matches where majority of data is missing and to also create a toggle bar which when toggled would exclude all the severe match data.

My Work: I merged the SQL code created in project 4 with the already existing code. Then, I created a filter using tableau filters in slider format which would exclude all the severe data when toggled. Then I created a column with a calculated field which would count the number of points in severe matches per umpire. However, this was a different kind of calculation as Level Of Detail functions had to be used. Level of Detail calculation is used when we fix certain columns and then perform the needed calculation. The code looked something as follows:

{fixed UmpireName,matchnumber,sessionnumber: sum(points)}

After creating the column, I manually verified the output and found that it was working. I also learned a new concept of context filter which has a power of superseding any other filter and I used it for filtering the day and night sessions. The following flow chart summarizes my work for this project:

(Dashboard image cannot be inserted due to presence of sensitive data))



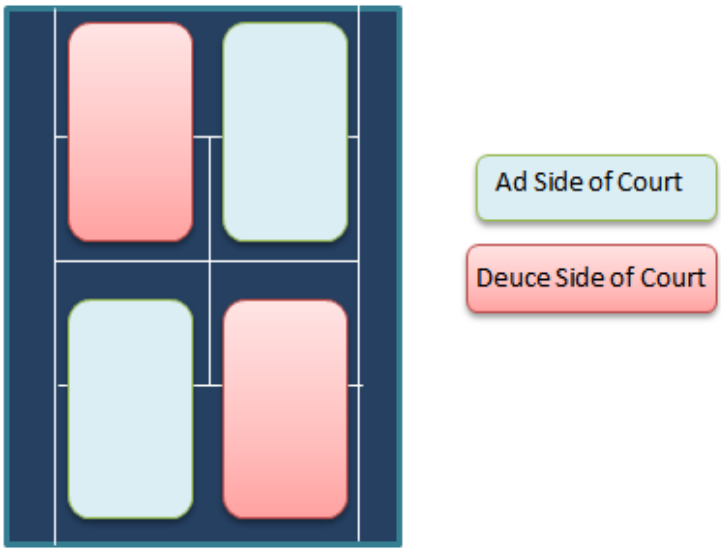
Impact of my work: The addition of audit status in the officiating dashboard is very important for the career of line umpires as it shows how accurate they are at different lines. If the accuracy is low, the head referee may not take that umpire for future matches. Hence, proper evaluation of accuracy was important and my work does exactly that.

Project 6: Creating a Rally Count Dashboard (Week 6,7)

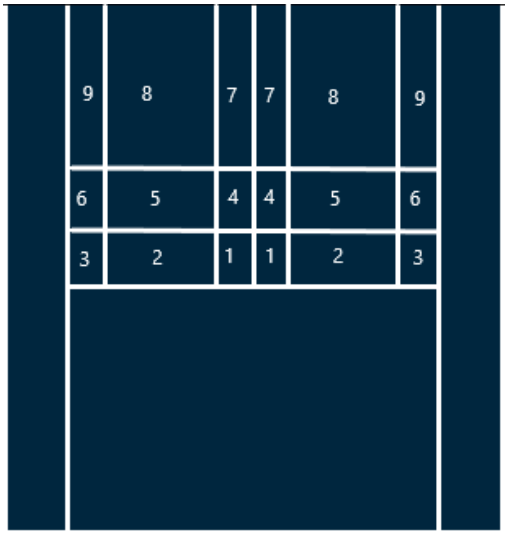
Background: This dashboard was requested by the player development team as they wanted to see how the serve location impacts the rally length. Lesser the rally length, better it is for the players as they can conserve more energy.

My Work: This project had to be built from the scratch and had several steps. The first step was the grid formation. Forming grids i.e. certain areas for serve analysis was important because analyzing the individual points makes no sense as mostly the points would be different and the players always aim for an area when they serve and not any particular point. So as an initial analysis, I looked at the databases to find the coordinates of the ball

for that serve, rally length for that particular point and also the linking factors between the two as there are hundreds of tables available. After extracting the relevant data, I realized that I can show the data on just one side of the court by merging the data of deuce side of court together and of ad side of court together. Ad and Deuce court side are as follows:



Then I looked at only one side of the court(Deuce side) and saw which regions have the most number of aces. According to that I created 9 grids per side of the court which looks as follows:

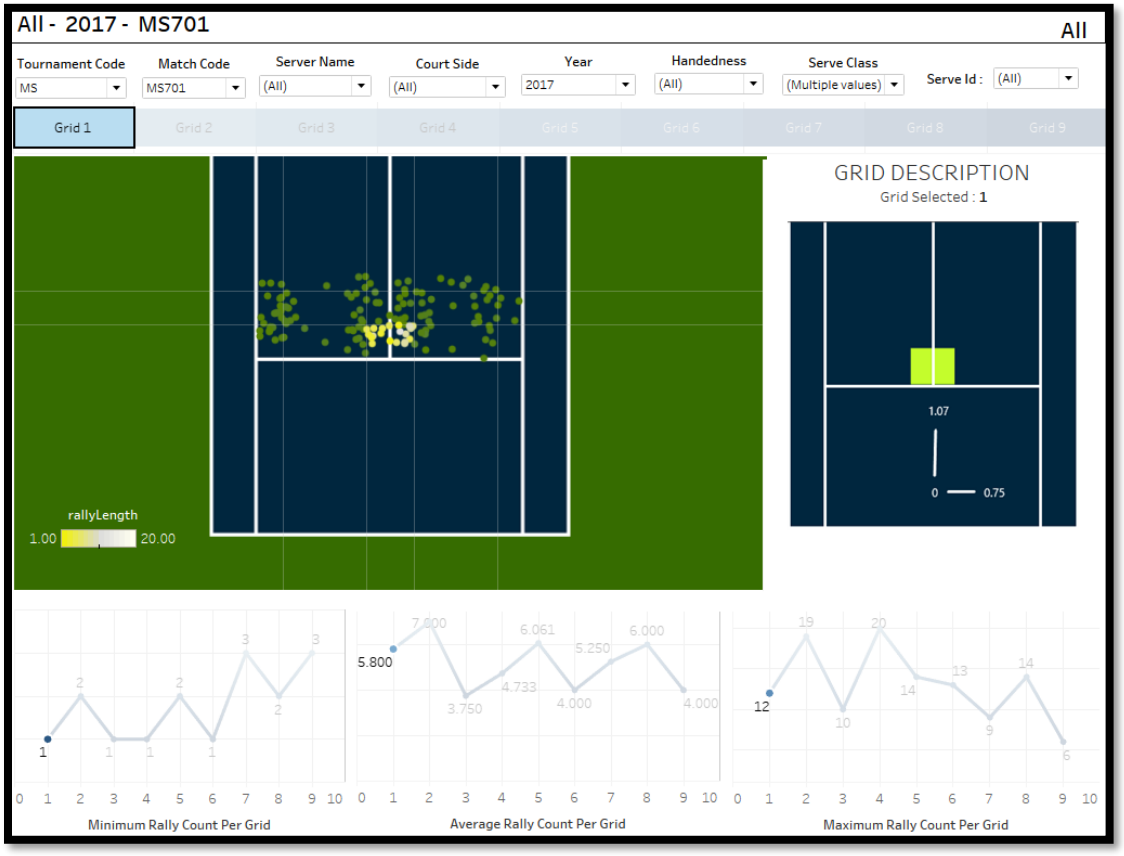


The decision was made on the basis of coordinates which are able to differentiate between the rally lengths to better define the areas where players must aim to serve. However, due to private data being present, I cannot disclose the coordinates decided.

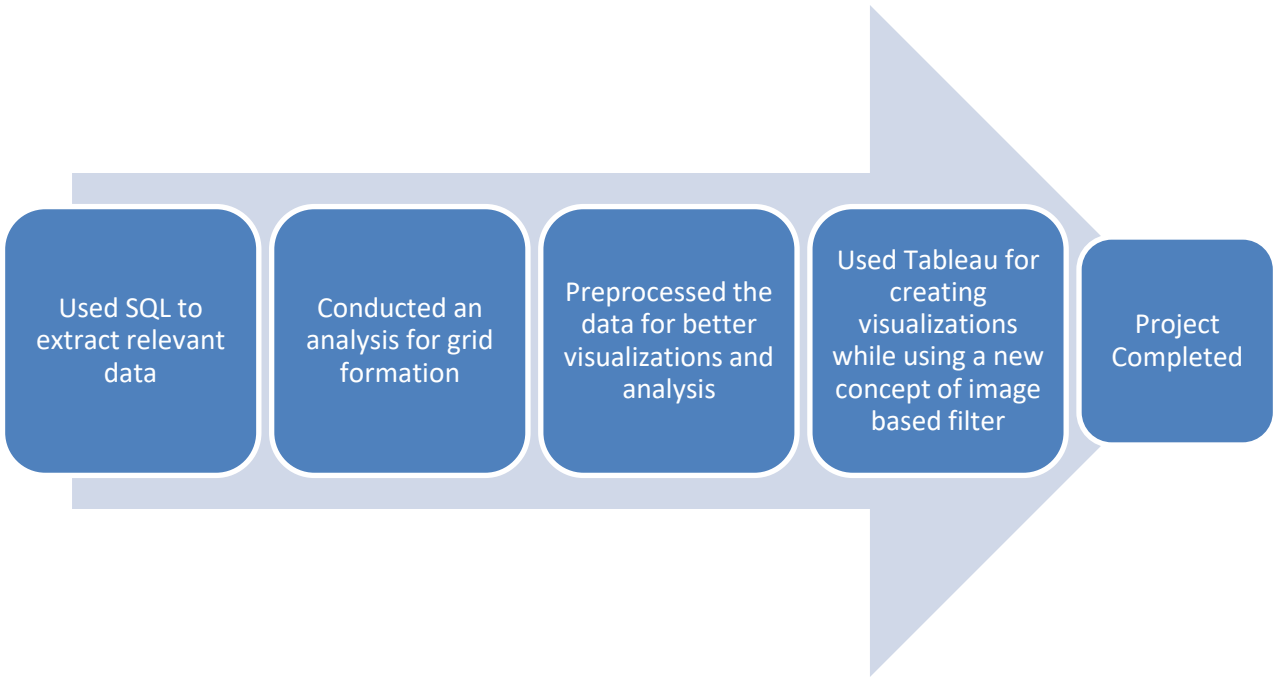
After forming the grids, I used tableau for the visuals and along with using the regular functions and calculated fields, I used a concept of using changing the images based on filter. So, if a person clicks on grid 1, then following items would be displayed on the dashboard:

1. Image showing the grid description
2. Points where serves landed
3. Average, minimum and maximum rally length for that grid.

Image of the dashboard when grid 1 is selected looks something as follows:



The following flow chart summarizes my work for this project:



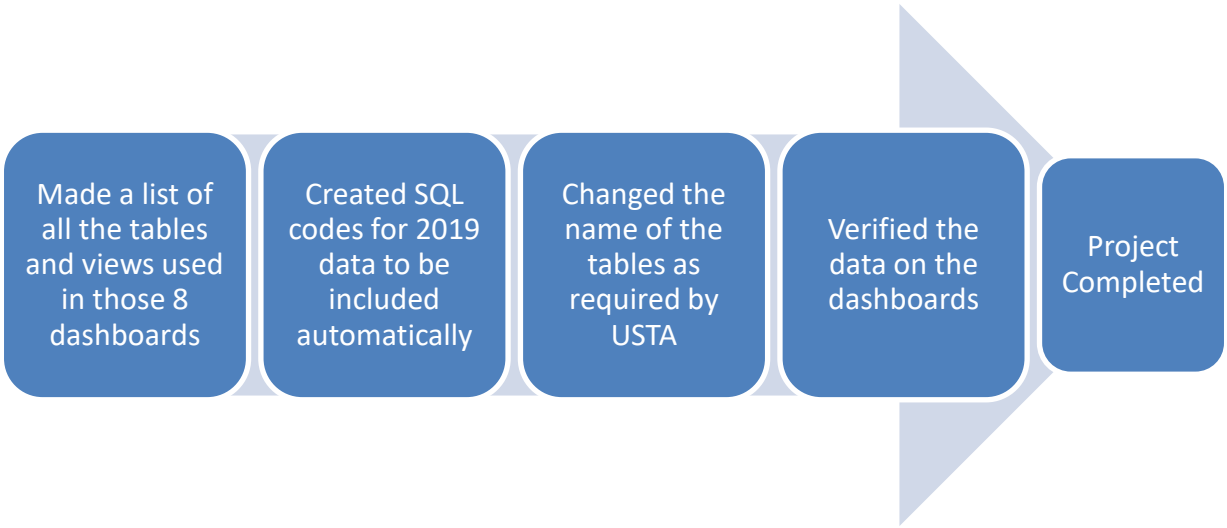
Impact of my work: This dashboard will help the player development team to train the US Players for serving at points where the rally length can be minimized and their energy can be conserved, something that is beneficial in a long-run.

Project 7: Player Development KPI Updates (Week 8,9)

Background: USTA have changed its server and have introduced a staging server due to which the names of all the tables and databases have been changed. Also, as US Open 2019 is coming up, USTA wanted me to create SQL codes such that when the US Open will start and they refresh the source, 2019 data has to be added to 8 dashboards along with making the changes in the name of the changed tables and databases.

My work: I went through all 8 dashboards and wrote down the list of all the tables and views being used in each. Then I went through each of the tables and views and made the necessary SQL changes to include 2019 data as well and also changed the names of the tables wherever needed. The process was really time and patience demanding but I could make the changes and then, could verify if the changes are making sense in all the 8 dashboards.

The following flow chart summarizes my work for this project:

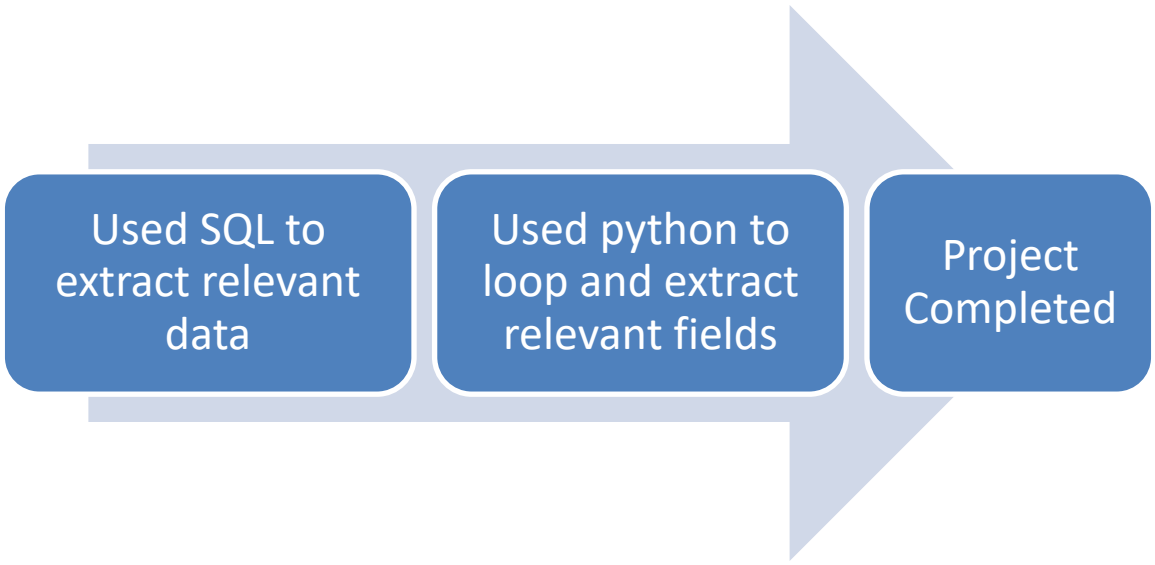


Impact of my work: The changes made by me would be very useful for the coming US Open as all the data obtained will be automatically structured in the required form and the dashboard can be seen by people without any errors being present.

Project 8: Calculation of Player Data received per feed (week 10)

Background: USTA obtains all the data in XML format from the 2 sources and they wanted to know that per each XML file, data of how many players is sent.

My work: For getting the data in the format : Date, Timestamp, XML content, and number of players, I first extracted the date and XML content column and then I exported it to csv format. Then, I imported the csv file in python (I tried to pull Microsoft SQL server data in python but could not do so due to access problems and hence, had to follow this longer procedure) and used python to loop through each XML file and counted the unique number of players. Finally, I presented my outcome in the format needed. The following flow chart summarizes my work for this project:



Impact of my work: From the code developed and the data obtained, USTA will now be able to keep a track of data sent by Hawkeye and SMT.

A TYPICAL DAY OF MY INTERNSHIP:

Every day of my internship was very typical based on time allocation of my activities but was very varied in terms of the work that I did. As my internship was in White Plains, I had to wake up at 5:30 in the morning and catch a train at 7:20. I used to reach office by 8:30 every morning and used to start working from that time. Every day my supervisor had allocated 3:30-4:00 pm as a slot for daily meeting and hence, every morning when I entered my office, I knew exactly what had to be done on that day. I used to leave my office by 5 and reach home by 6:45. Although, it looks like I had a routine but in terms of the work every day was a new challenge.

WORK RELEVANCE:

1. **How my work fits within the organization?**
   The mission of USTA is to promote and develop the growth of tennis in the USA and to achieve this mission, it is important to train the players well and to keep the game fair. To achieve the same, USTA has Player Development and Umpire Teams. My

work is to help these teams achieve their goals and hence, to help USTA achieve its mission.

2. **How did my work contribute to the business of the organization?**
   As mentioned above, my work is to help the player development and umpire teams to achieve their goals which in turn contributes to the mission of the organization. The main business of USTA is US Open and by having high quality players and accurate umpires, more audience would like to come to watch the match, thus increasing the revenue of USTA.

3. **Who do I work with?**
   I work in the Data Analytics Department of US Open and it is a small team of around 5 members.

HOW THIS INTERNSHIP CONTRIBUTES TO:

1. **My career goals:**

I want to go in the field of Data Science and had taken Data related courses at Columbia which taught me Machine Learning and Data Analytics in Pyhton and R. But, I never got to learn SQL and Tableau which are essential tools for a Data Scientist. I got to learn both, SQL and Tableau, of high level very well in this internship. The knowledge and experience gained till now along with a course (Big Data Analytics) that I am taking in this fall, will make me completely prepared for the role of Data Scientist.

2. **Coursework at Columbia:**

Management Science and Engineering is a diverse course and I am focusing on the data science/analytics track. Hence, I took subjects such as Data Analytics, Business Analytics and Tools for Analytics in my spring semester and learned Python, R and Machine Learning algorithms. I believe that the additional skills that I have acquired during my internship i.e. SQL and Tableau compliments them well and I will be able to improve the level of projects in my next semester (especially for Big Data Analytics course). Along with the technical aspect, I have learned a lot of things about the professional world and discipline which will help me do well in whatever I take up at Columbia University.

CONCLUSION:

To sum up, my 10 weeks internship experience at the United States Tennis Association has been an amazing experience that I will remember throughout my life. I got a chance to meet new people and learn about the work culture in the US, something that was crucial for me as an international student. I got to know behind the scenes of the US Open and the feeling to be a part of such an amazing tournament in a field I love (Data science) cannot be expressed in words. I could add new technical skills of Tableau and SQL to my skills list which is a perfect complement to the skills I already possess to get in the Data Science field.

Also, I got to know the importance of academics more when I saw how the knowledge that I gained at Columbia helped me to grasp everything quickly. I got to figure out my strengths and weaknesses and I also learned how to work on my weaknesses to turn them into my strengths slowly. Hence, in the 10 week duration, I got to learn a lot professionally and personally and I believe that it was a once in a lifetime opportunity.

THANK YOU!