SU22: DATA VISUALIZATION: 8101

Final Project Paper

Analysis of Cricket

By Uday Sankar Boni

Abstract

Cricket is considered the 2nd most popular sport after soccer with an estimated fan following of approx 2 billion people around the world. As part of this project, I am going to analyze the data which is from different formats of the game. Means cricket has 3 main formats: Test Cricket, One Day, and Leagues (specifically IPL), so I'll try to perform a comparison between the countries and players across different teams around the world. Also, will try to show some correlation between the teams and players.

Introduction

Cricket as a sport is widely followed all over the globe. As part of this project, I am trying to visualize how different formats of the cricket.

Motivation

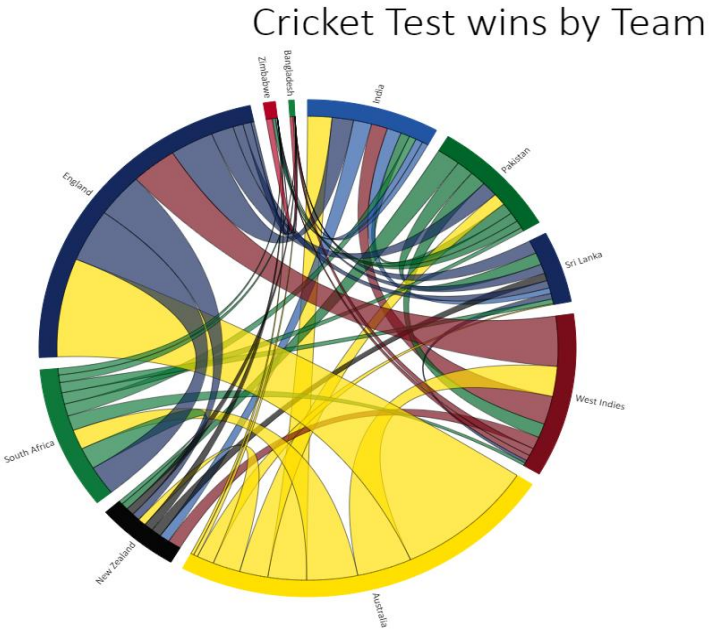
Cricket is one of the more evolving games. Since the first cricket game, which happened back in 1878, from then the game has changed and it would be interesting to see some data insights using data visualization

Background/Existing Work

With time, the game of cricket has evolved over the years with initial days cricket being played over 5 days (Test Cricket), to 1 day cricket (ODI Cricket), to the latest format where cricket is played just for 3 hours (T20 Cricket). With the growing popularity of the sport, several visualizations have been developed that help cricket fans and experts explore cricket matches and track the performances of the team as well as the player performances.

Below visualization provides a detailed information on how countries have fared in the longest format (Test Cricket). The thickness of links between countries encodes the relative frequency

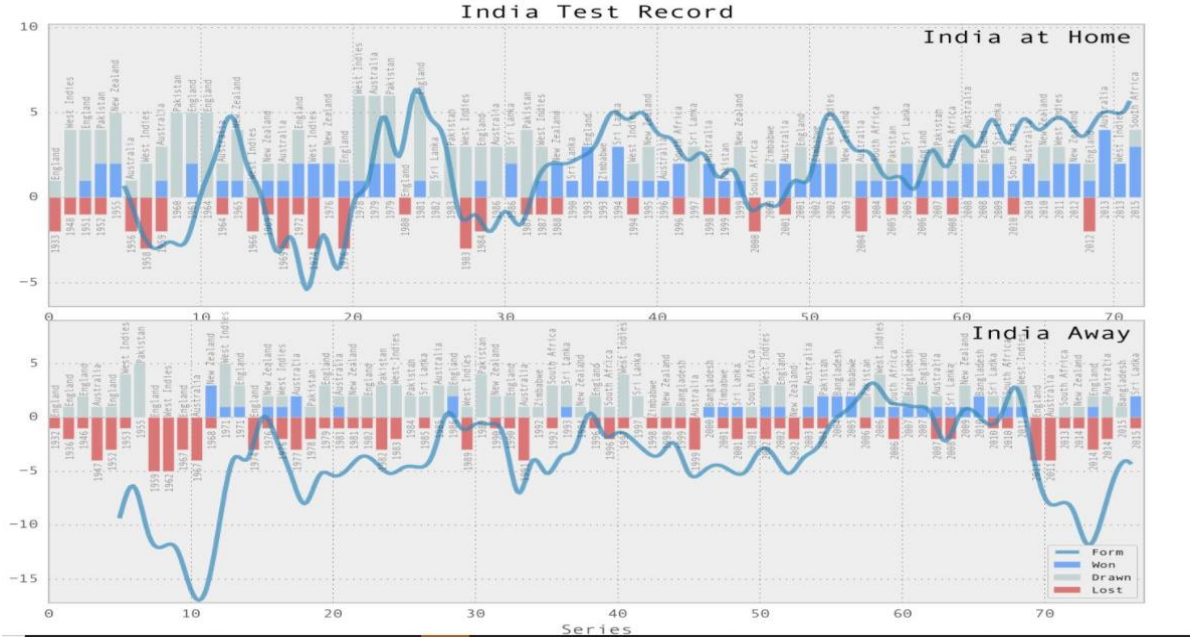
of matches between two countries: thicker links represent more matches which result in more wins.



Reference: https://codepen.io/veereshai/full/pmvwC

Winning in a Cricket Match depends on many key factors like a home ground advantage, past performances on that ground, records at the same venue, the overall experience of the players, record with a particular opposition, the overall current form of the team and also the individual player.

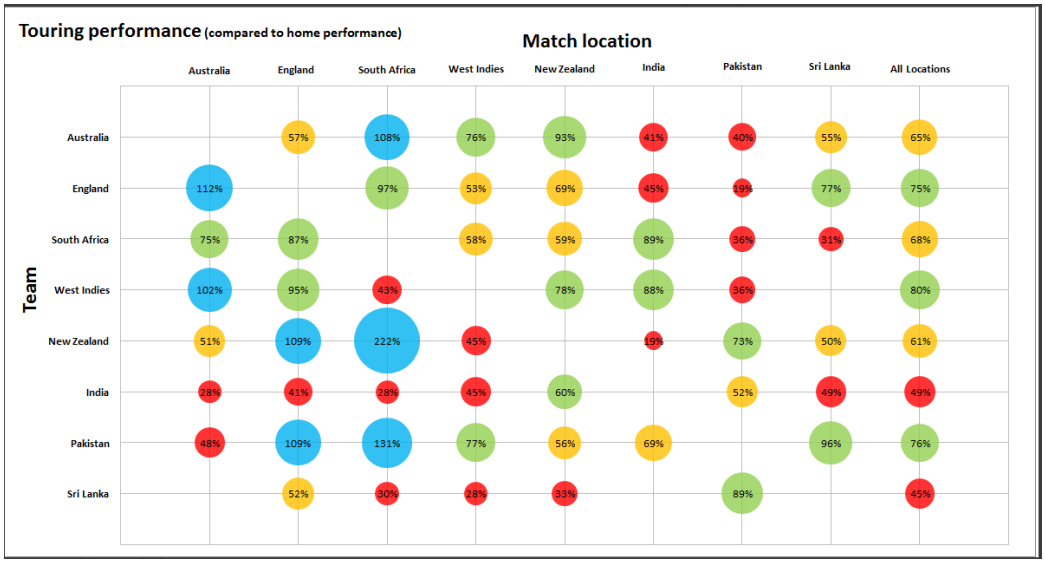
Below is the visualization that tells how a country India has performed at home and abroad:



Reference:

https://www.reddit.com/r/Cricket/comments/499uhs/indias_test_record_data_visualization

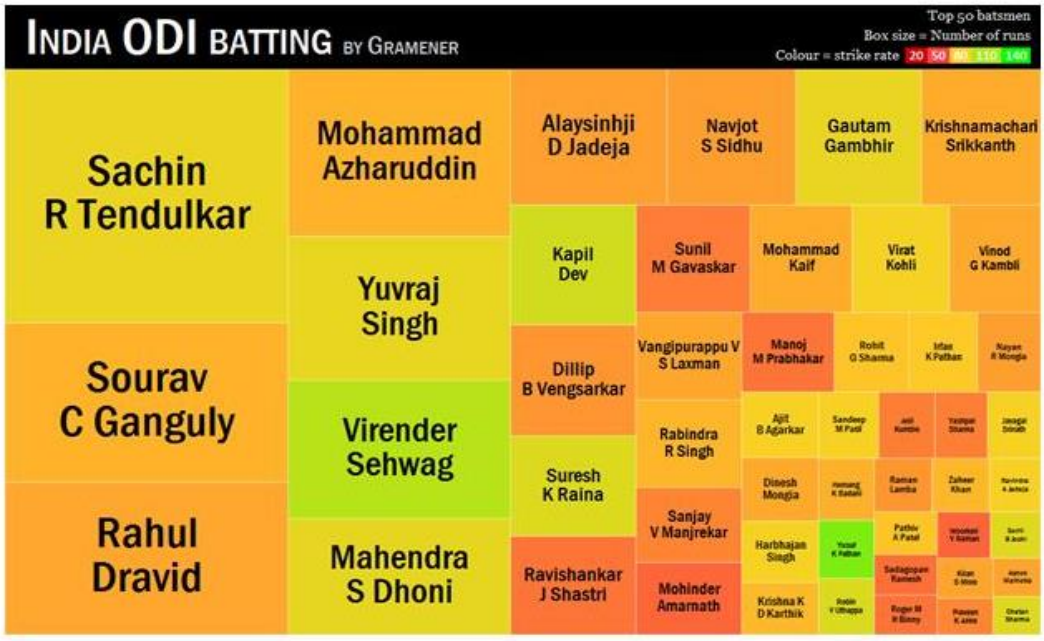
It could have been better if it had comparison with how other countries have done in Test cricket.
Below is one of the visualizations that depicts how countries have performed while playing in abroad tours:



Reference:
https://truii.com/data-curio-blog/sports-statistics/cricket-hardest-place-tour/

As we can see from above visual, Asian countries have struggled to win outside of Asia, whereas non subcontinent teams have performed poorly in Asian conditions. But it can be attributed to lesser total matches played between the countries.

Below visualization show the runs scored by Indians in ODI format. It shows bulk of the runs are scored by Sachin Tendulkar, Sourav Ganguly and Rahul Dravid. Though it shows % of runs, it could have been better to have a different color coding that can show the differences in a better way.



Reference:

https://analyticsindiamag.com/data-visualization-a-pov-from-gramener/

The below visualizations show the players who have scored 0 runs most times.

This information could have been better presented if there were color coding of players from different countries, whether the format was Test or ODI or T20.

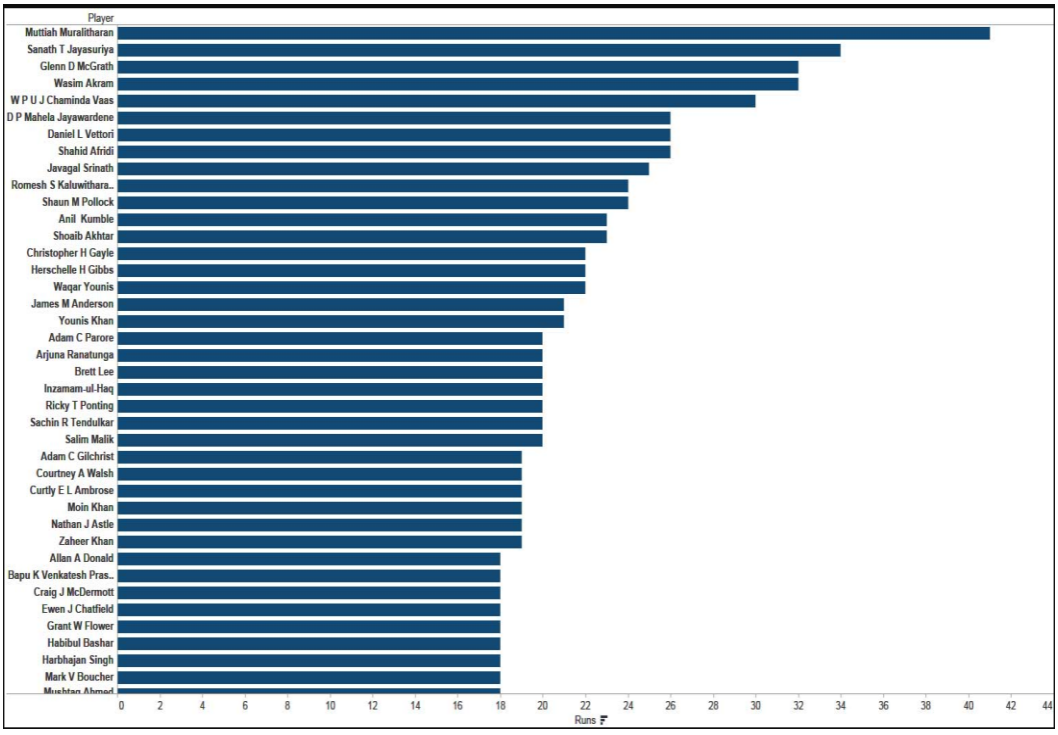


Image Reference:
https://www.edupristine.com/blog/cricket-data-visualization-with-tableau

Objectives

The objective is to create some visualizations which will help the user to understand how the countries have adapted to the different formats in this sport. I'll try to answer three categories of questions through visualization as follows:

Trends

- Basic Trends
  - Changes over time, considering the x-axis to be YEAR. For example, the changes in the number of matches won across the years for each of participating countries as well as how players have fared over the years. These visualizations show audiences how a specific variable has changed over time.
- Advanced Trends
  - Visualization of more variables in the same plot. For example, show the impact of formats on the performance of countries

Comparisons

- Ranking
  - List the top 15 countries in terms of a certain variable (such as the number of matches won, number of players participating etc.)
  - List the top 15 most runs scored by the players or most wickets taken by players.
- Wins between two nations in each of the three formats
- Home Vs Away number of win or loss
- Win/loss/Tied match results for each country in each format

Stats

- Total matches played in each of different formats
- Number of players that have played for each country
- Number of wins for each country

Datasets and Methods

For showing the visualizations, analyzed data from different sources and decided to use certain methods/techniques to show the visualizations.

Data

After analyzing the datasets from different sources, it narrow downed on the extracts from Cricinfo Statsguru which provided information regarding different teams, players as well as

match statistics. Following datasets are used in this project with certain cleanup, merges and additional parameters.

- Cricinfo Statsguru Data (All formats) - From Kaggle
- International Players Data - From Kaggle

The above datasets provided details about the matches being played between the countries, the results for each match and how the players have performed in each of the matches, for each of the different formats.

Note: I have mainly used men's teams' statistics because it has long historical data from beginning of cricket.

The next step was to join (used the python method merge() here) the different datasets into one. Added one of the column "Plat_Format" into each merged file to indicate play format.

Also renamed column headers to replace space with '_' character so that it will be easy to read in program. As the dataset is already very clean, there is no data cleaning activity is needed except for the fact that how to deal with missing values (denoted by null).

Here are the details of merged files and the applicable data variables used from each file:

- Men_All_Team_Match_Results.csv: The result of each match played by team in each play format

    - Play_Format - The match play format Test, ODI or T20I
    - Result - The result of match for country Won or Lost
    - Match_Year - Year when match held
    - Country - The name of country who Won or Lost

    The data for Match results dataset looks like:

Play_Format	Result	Match_Year	Country
T20I	Won	2005	England
T20I	Lost	2006	England
T20I	Lost	2006	England
T20I	Lost	2007	England
T20I	Lost	2007	England

- Men_All_Team_Batting_Stats.csv: The batting stats of each team

    - Play_Format - The match play format Test, ODI or T20I
    - Country - The name of Team's country
    - Avg_Runs_Per_Wicket_Batting - The average number of runs partnership between wickets
    - Avg_Runs_Per_Six_Balls_Batting - The average number of runs per six bowls over
    - Matches_Won - Number of matches won so far
    - Matches_Lost - Number of matches lost so far

- Matches_Tied - Number of matches tied so far
- Matches_With_No_Result - Number of matches with no result
- Win/Loss_Ratio - The ratio between number of matches wins and lost

The data for Batting results dataset looks like:

Play_Format	Country	Avg_Runs_Per_Wicket_Batting	Avg_Runs_Per_Six_Balls_Batting	Matches_Won	Matches_Lost	Matches_Tied	Matches_With_No_Result
ODI	England	31.10	4.91	375	334	9	28
ODI	Australia	34.14	5.01	575	331	9	34
ODI	South Africa	35.76	5.11	385	216	6	18
ODI	West Indies	30.20	4.78	401	381	10	30
ODI	New Zealand	29.21	4.82	351	374	7	40

- **Men_All_Player_Innings_Stats.csv:** The innings stats for each player in team
  - Play_Format - The match play format Test, ODI or T20I
  - Innings_Player - The name of player
  - Innings_Runs_Scored_Num - Number runs layer scored in the match
  - Innings_Date - The date when the math inning played
  - Innings_Wickets_Taken - Number of wickets taken by payer in the match

The data for Player Innings Stats dataset looks like:

Play_Format	Country	Innings_Player	Innings_Runs_Scored	Innings_Runs_Scored_Num	Innings_Date	Innings_Wickets_Taken
ODI	England	JJ Roy	180	180	1/14/18	NaN
ODI	England	AD Hales	171	171	8/30/16	NaN
ODI	England	JJ Roy	162	162	6/29/16	NaN
ODI	England	AJ Strauss	158	158	2/27/11	NaN
ODI	England	AJ Strauss	154	154	7/12/10	NaN

- player personal_male.csv: The personal information for each team player
  - name - The name of player
  - Player_country - Player's country

The data for Player dataset looks like:

name	country
Aakash Chopra	India
Aamer Hameed	Pakistan
Aamer Hanif	Pakistan
Aamer Malik	Pakistan
Aamer Nazir	Pakistan

To show the geo information on the map, It needed something like ISO Country code, I used the python library pycountry to get that information.

Plotting Libraries:

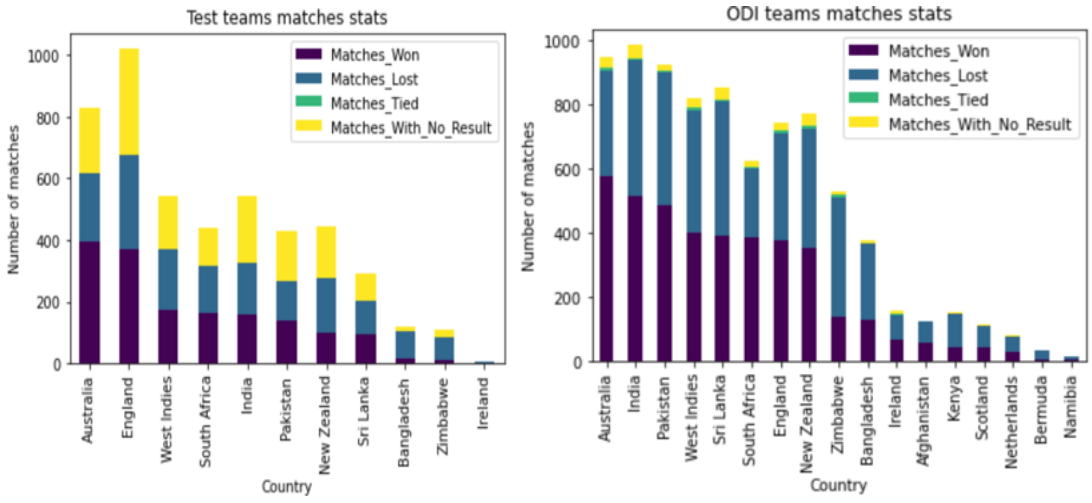
There are multiple python libraries used to achieve project visualization goals.

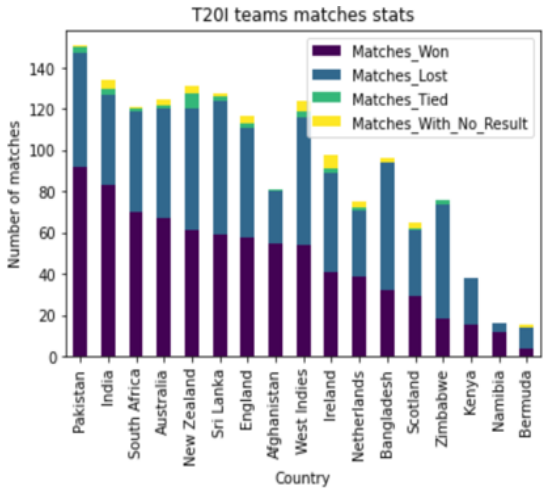
- Matplotlib
- Seaborn
- Plotly
- Altair
- Pandas Plotting Parallel Coordinates
- Bar chart race
- Bar chart race - Flourish

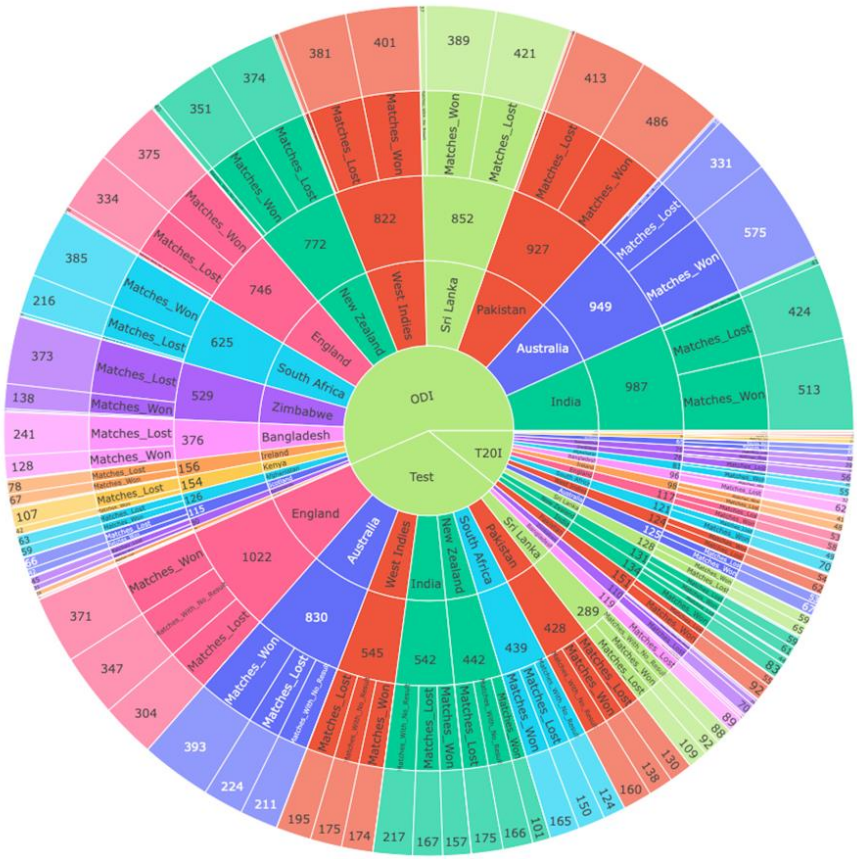
Visualization Method Selections:

There are various data visualization methods available with python stack that can be used to visualize the data. When going through the process of visualizing the sports statistics, one has to make sure to study some of the most basic and fundamental visuals in order to select a visual that best fills the requirements. I used the following visualizations to depict the statistics for cricket for different formats.

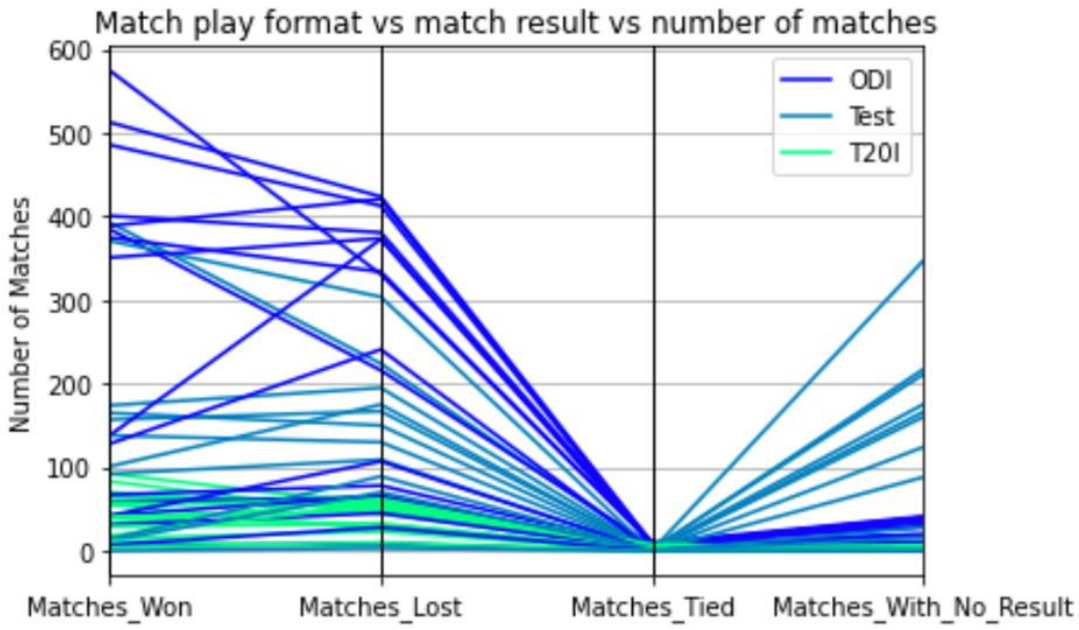
- **Stacked Bar/Sunburst chart** – The idea is to visualize the country wise number of matched wins, losses, ties and no results. This type of visualization will provide the consolidated view of multi-dimensional data like play format for each country and countries numbers associated with each format match outcomes. This will potentially help us to glance each format's match results those are associated with each country.







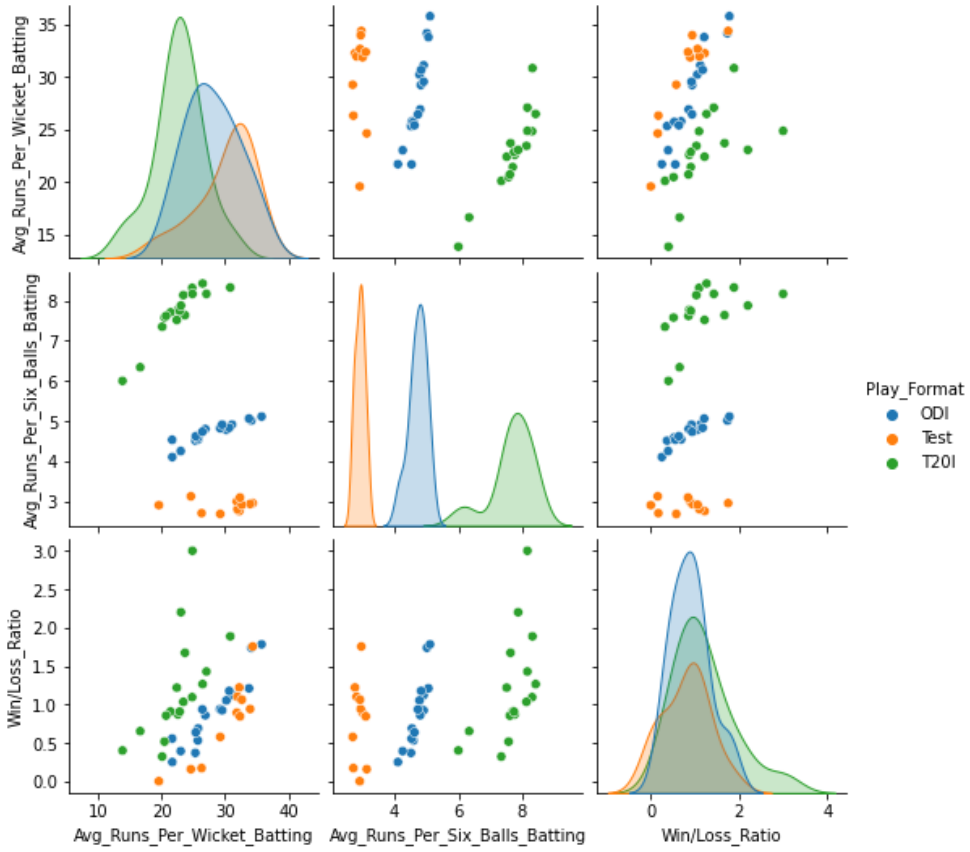
- **Parallel Coordinates plot:** The idea is to visualize win, loss, ties and no results for each match play format. With help of parallel coordinate plot the multi-dimensional data like countries, match format and number of win/loss/ties/no results can be plotted efficiently to recognize the outcome of each match format.



Data Observations:

- The ODI and T20I match format has higher chances of win or loss results vs match tied or no results
- The Test match format has less win or loss chances compare to no results

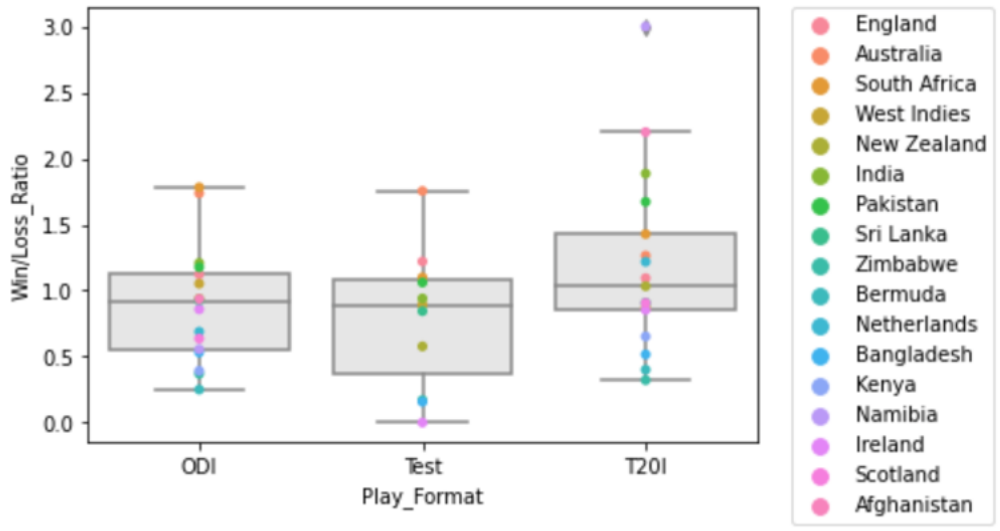
Pair Plot – The idea is to visualize relationships between main factors of maximizing win/loss ratio. There two main batting factors are average runs between wickets partnership and average number of runs scored per over (six bowls). So, there are multiple dimensions to decide win/loss and it is important to find the relationship between these variables. It clearly indicates that these three variables are dependent on each other. It means that more the average runs per six ball or partnership between wickets, there would be high chances of wins.



Data Observations:

- The average runs per six balls is in order - T20I > ODI > Test
- The average runs per wicket batting is in the order - Test > ODI > T20I
- The win and loss ratios are ODI > T20I > Test
- The win loss ratio is definitely impacted by the runs per six over or runs per wicket
- The formats Test, ODI and T20I is played with 90 overs per day (max 5 Days), 50 overs per team and 20 overs per team. And looking the above stats, the if we look for the runs per wicket or runs per six balls are ordered higher to lower for each format is T20I > ODI > Test.

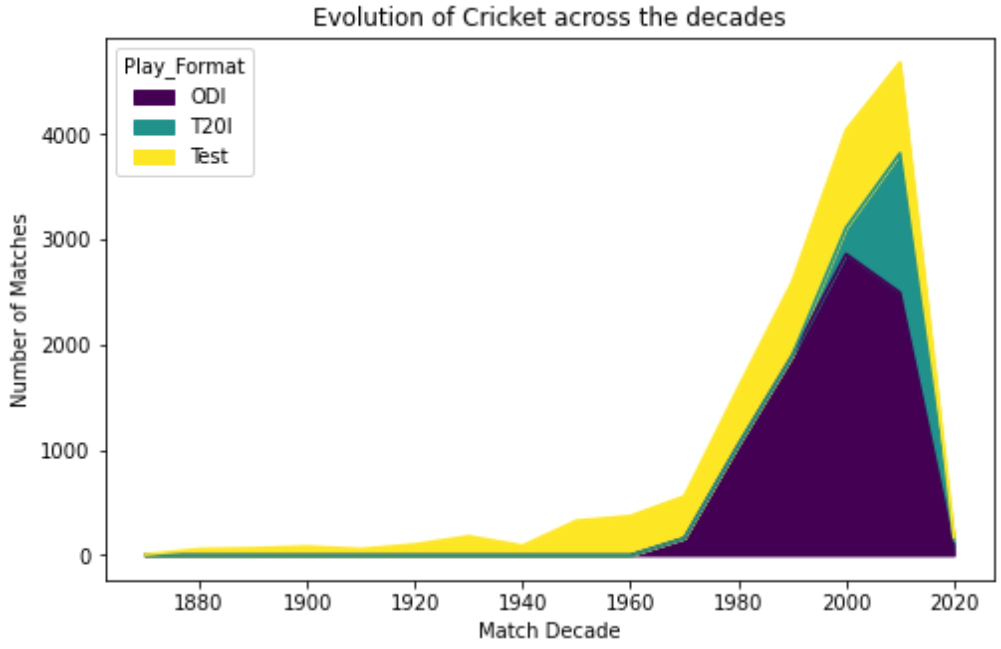
- **Strip Plot with box plot –** The idea is to visualize relationships between win/loss ratio for each of the country in each format.

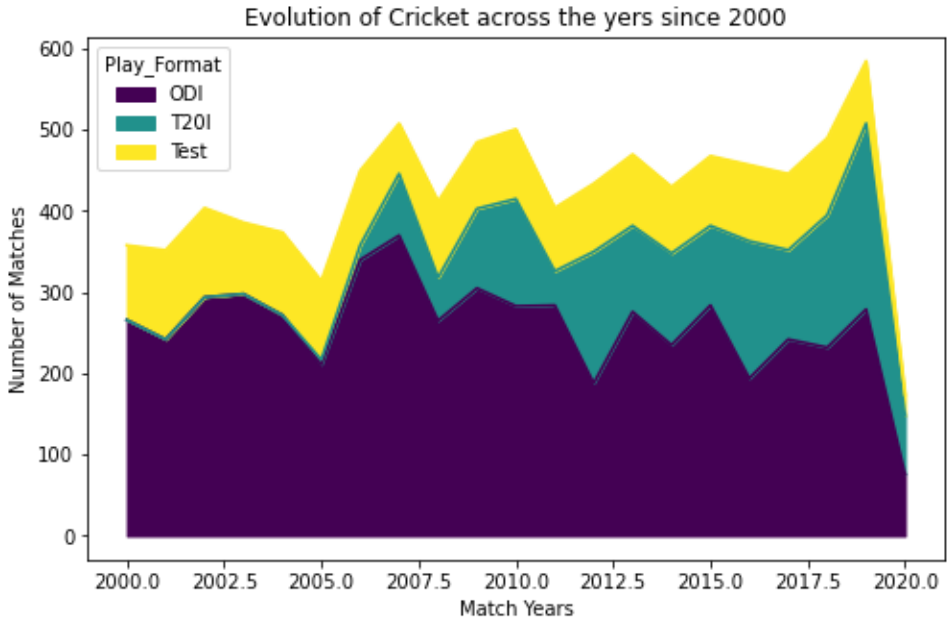


Data Observations:

- The popularity of play format among countries is in order higher to lower is T20I, ODI and Test
- The countries have win/loss greater that mean is consistently good performers for each format.
- Not all countries play each format.

- **Area Chart** – The idea is to compare the team's performance over the decade and past 20 years under each format. Where it will plot the area of number of matches played in each decade under each format and size of area will help to define its birth with the game and how it became popular over the period of time.

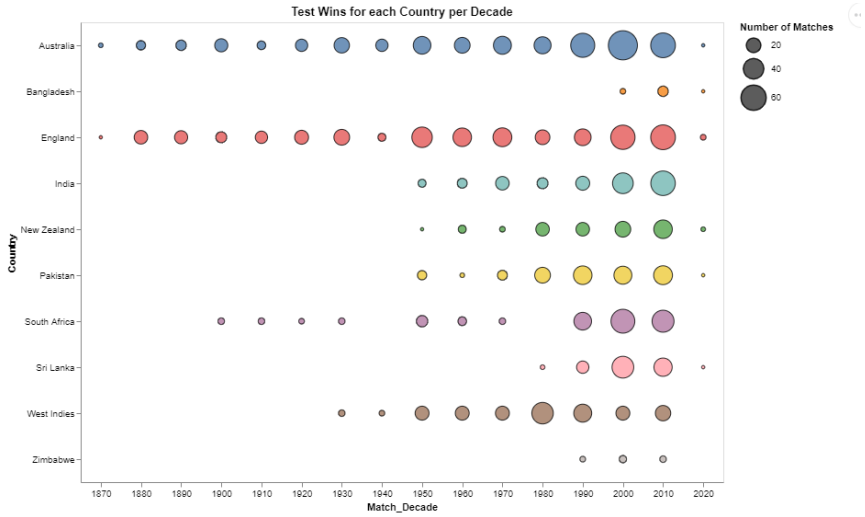


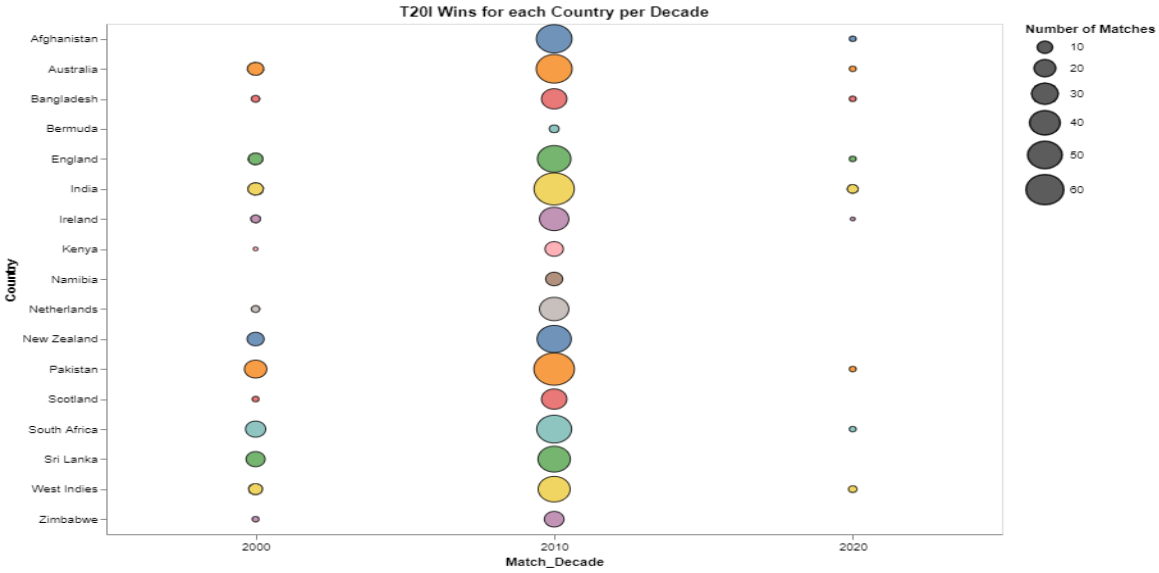


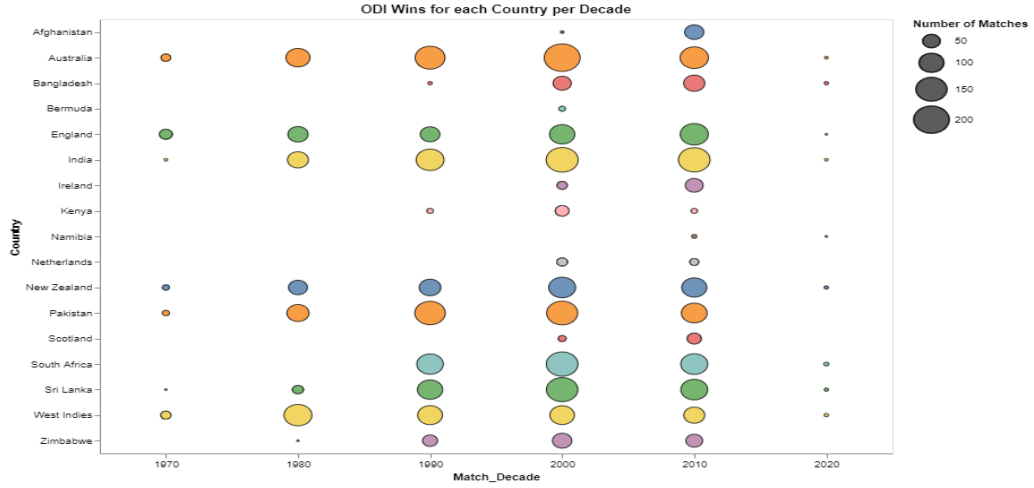
Data Observations:

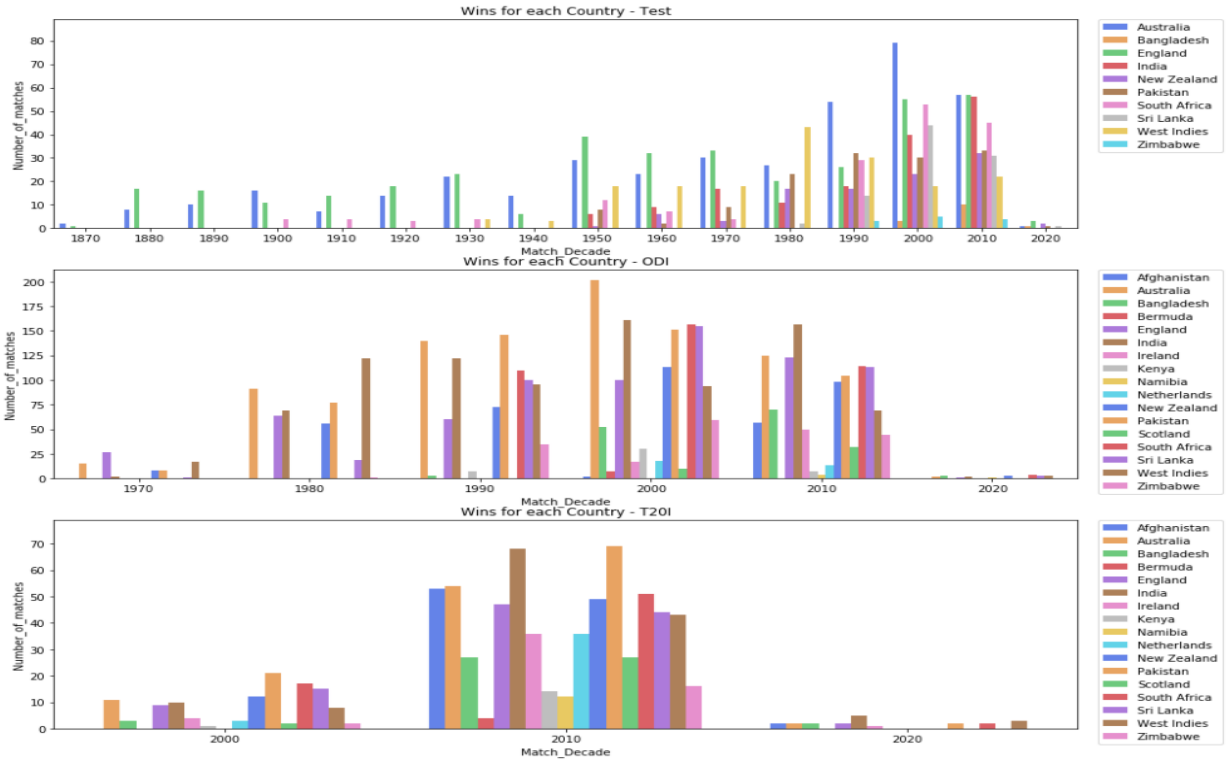
- The Test play format is the oldest format which was started to play in 1870's and increased number of matches played every decade gradually.
- The ODI format was introduced in decade of 1960 to 1970 and got popular rapidly in every decade afterward.
- The T20I format was introduced in decade of 1990 to 2000 got rapid popularity in subsequent decade, as result the ODI and Test matches numbers reduced comparatively.
- Even T20I got rapid popularity the ODI format is still most popular format in last 20 years and Test format is running with consistent popularity.
- From last 4-5 years the ODI and T20I formats are competing each other in popularity.

- **Bubble Chart/Bar Chart** – The idea is to show Number of matches won by each country per format across the decades. This would help us understand which country was the best in that decade.




T20I Wins for each Country per Decade
Afghanistan
Australia
Bangladesh
Bermuda
England
India
Ireland
Kenya
Namibia
Netherlands
New Zealand
Pakistan
Scotland
South Africa
Sri Lanka
West Indies
Zimbabwe
Country
2000
2010
2020
Match_Decade
Number of Matches
10
20
30
40
50
60


ODI Wins for each Country per Decade
Afghanistan
Australia
Bangladesh
Bermuda
England
India
Ireland
Kenya
Namibia
Netherlands
New Zealand
Pakistan
Scotland
South Africa
Sri Lanka
West Indies
Zimbabwe
Country
1970
1980
1990
2000
2010
2020
Match_Decade
Number of Matches
50
100
150
200



Bubble chart VS Grouped Bar chart:

- Bubble chart shows the counts of matches won for each country per decade by means on different size bubbles, but it is little difficult to gauge the relative performances of each side per decade
- Grouped Bar chart shows the relative performance of each side per decade.

Data Observations:

Test:

- England and Australia are the only countries which started playing test cricket in 18th century.
- Australia is on top with huge number of wins every decade

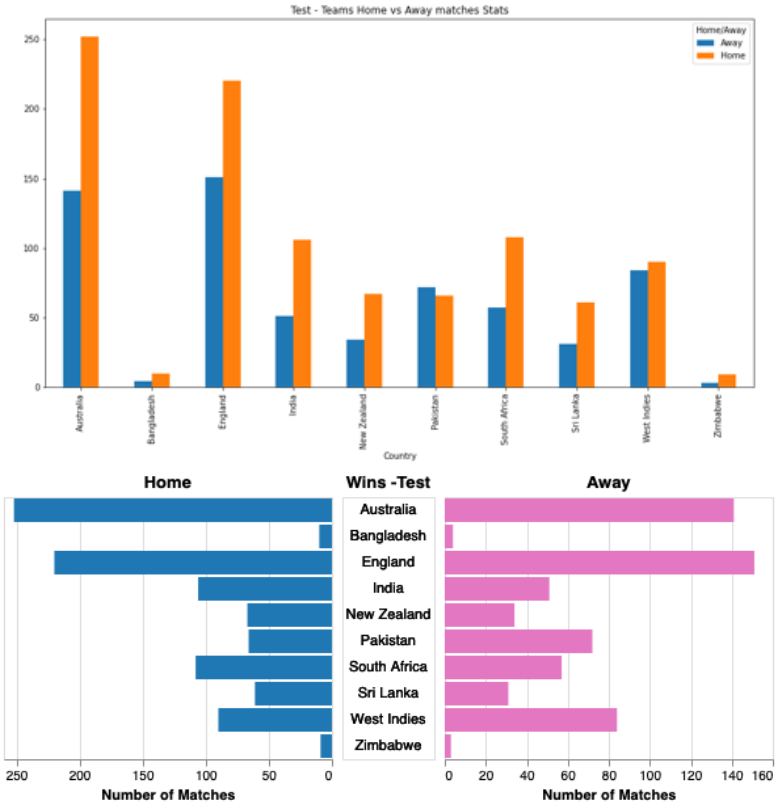
Test:

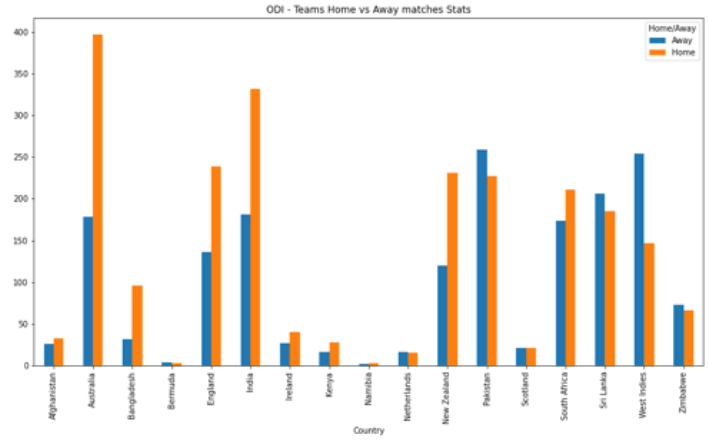
- Australia is on top with huge number of wins every decade

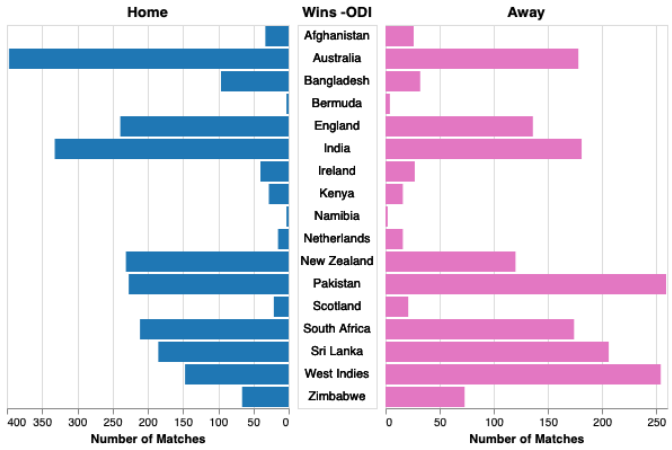
Test:

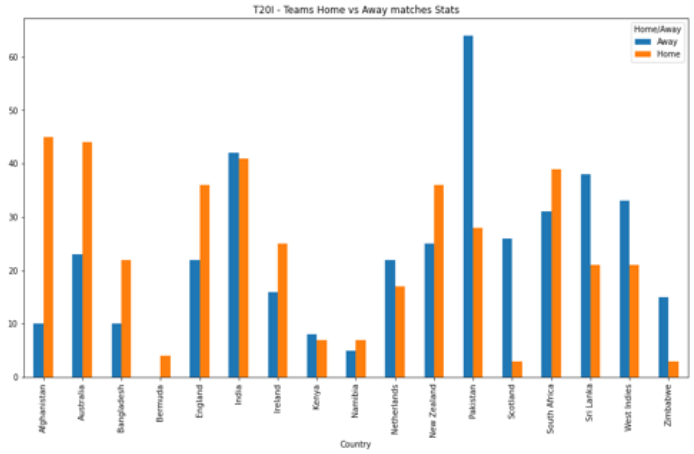
- Most of the countries are on par with number of wins in the 2000's.

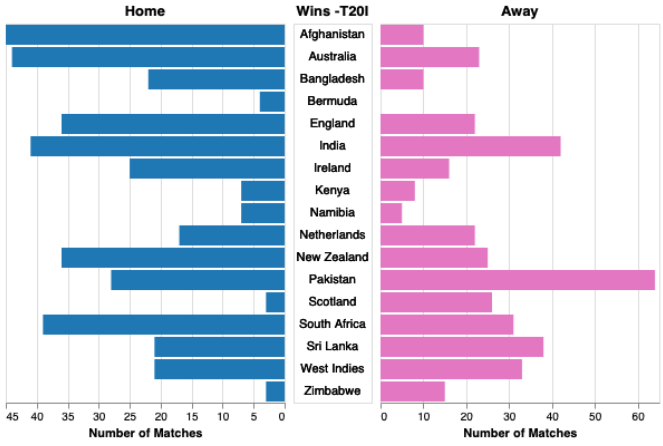
- **Bar chart/Pyramid graph**–The idea is to visualize comparison in teams winning performance between home grounds and foreign grounds. This would help to judge which team will perform better while playing in home country and foreign country to future matches. It is one of the important factors with big events like world cups or multi team series to predict winning team.







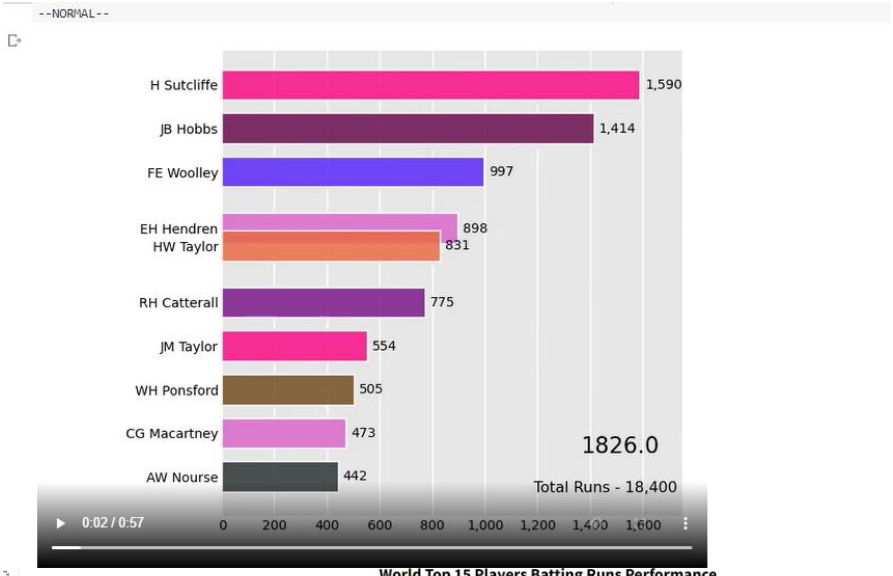


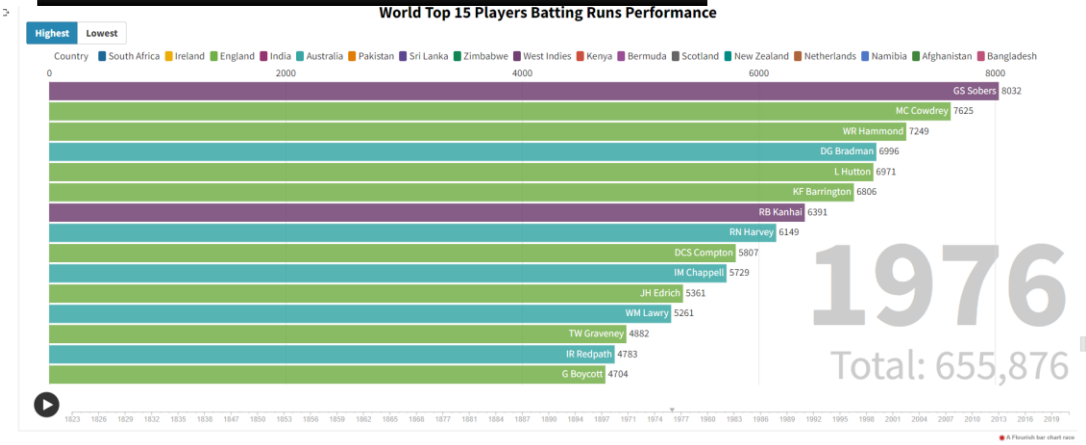


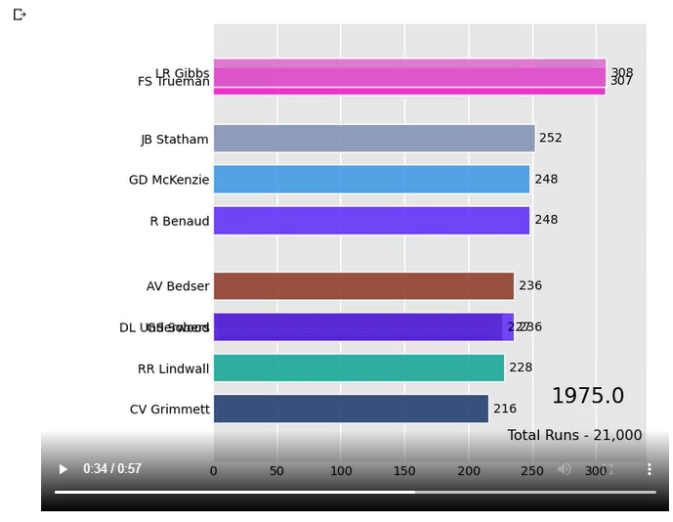
Bar chart VS Altair population pyramid methods:

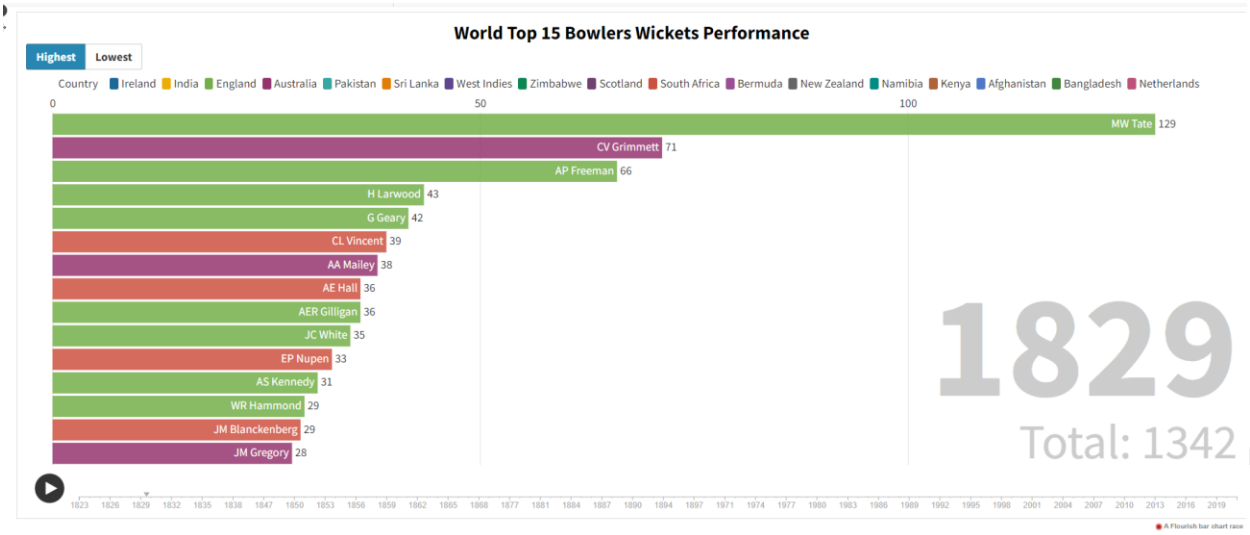
- The bar chart has x axis as countries and y axis as number of matched won with two adjacent bars for each country home and away counts, whereas with the population pyramid the x axis is count of won matches and y axis is countries with two separate left and right horizontal bars for home and away counts.
- Both methods are serving same purpose, but I found population pyramid chart is better over bar chart that uses less horizontal space and provide effective comparison vertically with rational utilization of space.

- **Bar Chart Race** – The idea is to visualize top 15 scored batsman so far and top 15 wicket taking bowlers so far. This would be animated visualization to showcase all top international players batting or bowling performance trend over past 200+ years of time since cricket was born.







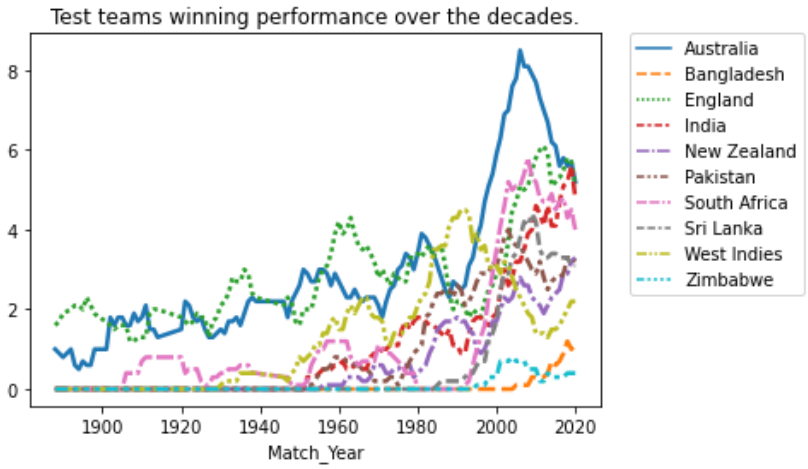


Plotting more effective race chart (right side ones from above) using flourish app which enables to upload data on their website and prepare visualizations with various properties. This also facilitates to add more dimensions to the chart compared to bar race chart.

Data Observations:

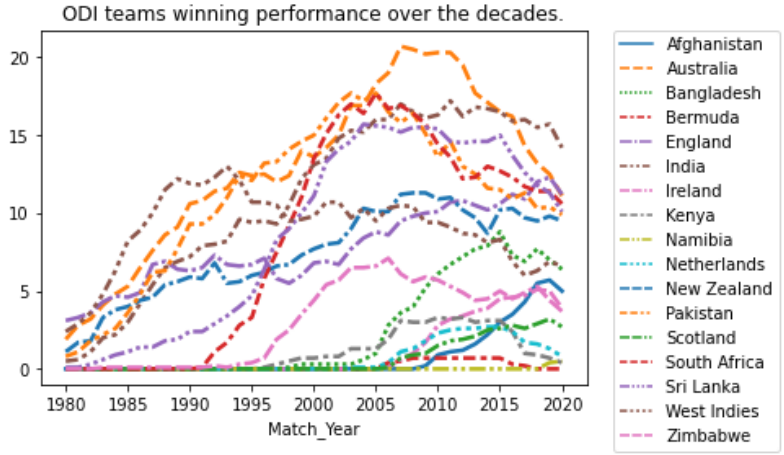
- The topmost player's score/wickets are pretty high to reach other players at this position and the topmost player would be unbeatable legend in future too.
- It seems these top 15 player played many years than average players and they become at this position.
- May be most of these players are retired from cricket.
- As cricket has evaluated over past 200+ years, the historical players performances were taken over by new players over the time.

- **Line Chart** – The idea is to visualize the teams winning history over the decades and possibility of future top teams in each play format. Where x-axis would be the decade of years and the y-axis would be each teams rolling mean of wins in each decade. It will plot separate line for each team that will show teams winning performance trend and user can judge the future top teams for couple of years.



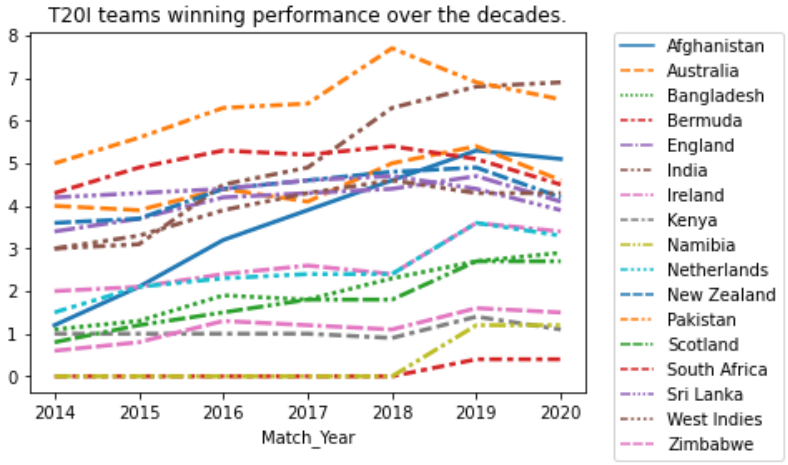
Data Observations:

- Australia and England are the longest player in test history and those are still in race of top teams
- India, New Zealand and South Africa joined the top performers race over the time.
- Pakistan, West Indies and Sri Lanka are lost their performance over the time in this format.



Data Observations:

- There are multiple countries playing this format since last 50 years and consistent with their performance.
- Australia, India, South Africa and England would be strong top 4 teams in this format.
- Next world cup 2 out of above 4 teams could reach up to finals.

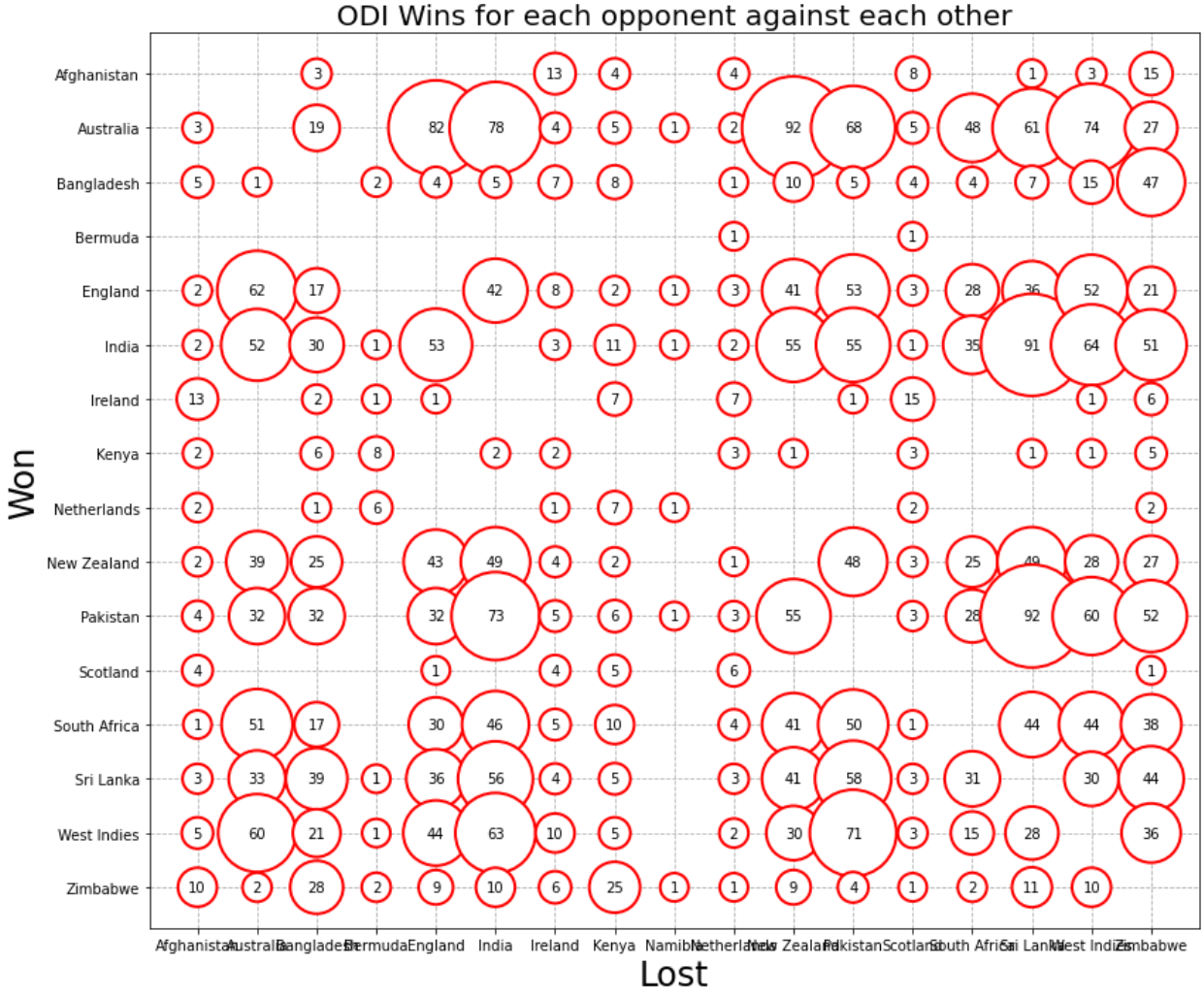


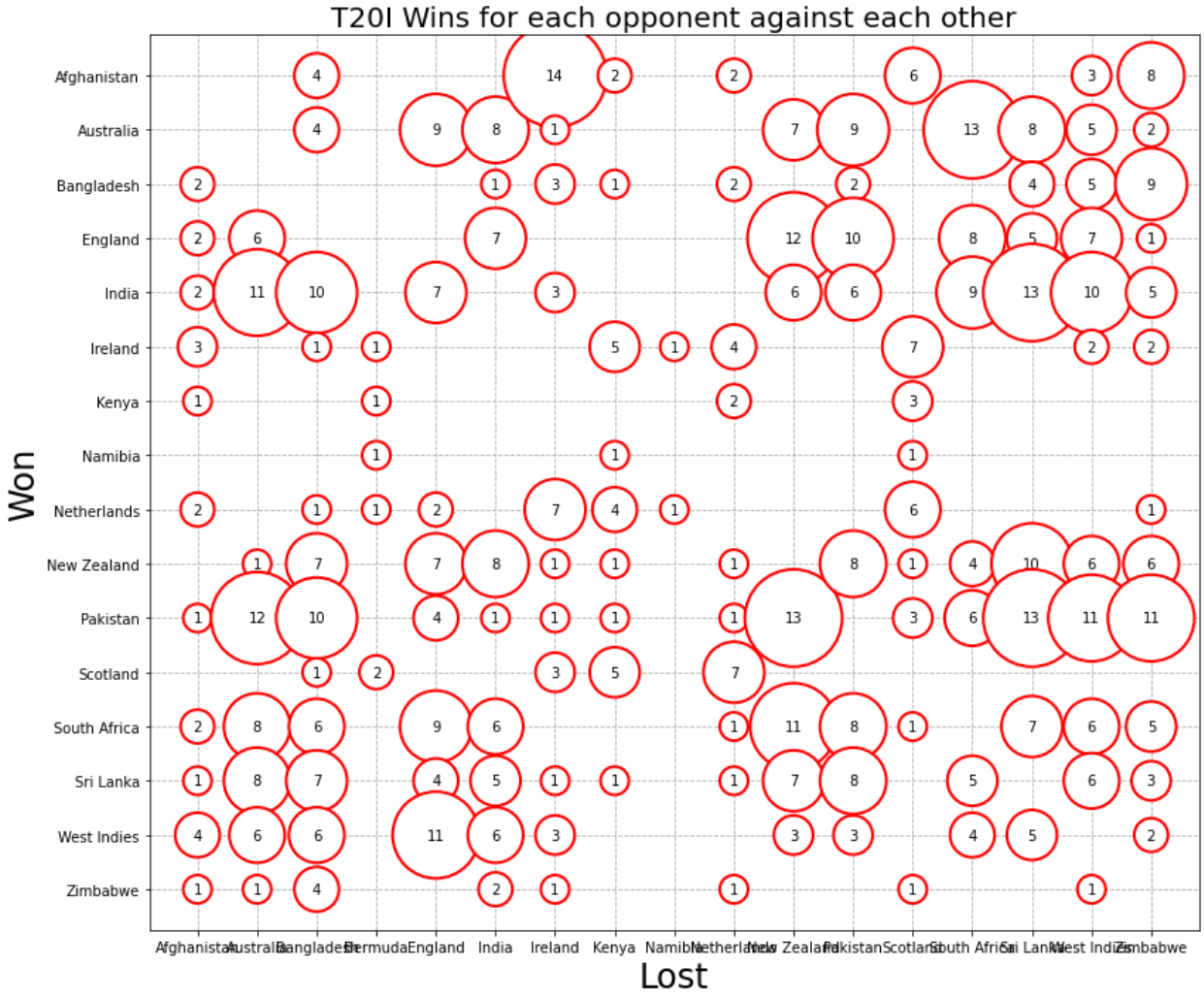
Data Observations:

- There are multiple countries playing this format since last 20 years and consistent with their performance.
- Australia, India and Pakistan would be strong top 3 teams in this format.
- Next world cup 2 out of above 3 teams could reach up to finals.

- **Bubble Plot** – The idea is to consolidate the win verses loss matrix between all international teams and observe which team is stronger or weaker than opponent country, so that it can provide guess that future match happens between two countries then which country could be possible winner of match. This is used to show correlation between two variables. Here, it is showing the number of wins achieved by each country against other countries.


Test Wins for each opponent against each other
Won
Lost
Australia
Bangladesh
England
India
New Zealand
Pakistan
South Africa
Sri Lanka
West Indies
Zimbabwe
Ireland
5 146 42 34 33 52 19 58 3
1 1 1 4 7
110 9 47 1 48 25 64 15 49 3
28 9 26 21 9 14 20 22 7
8 12 11 12 12 4 16 15 11
15 10 21 12 1 25 4 20 20 10
26 10 34 15 25 15 14 18 8
4 16 8 7 9 16 9 9 14
32 10 57 30 13 17 3 4 7
7 2 3


ODI Wins for each opponent against each other
Won
Lost
Afghanistan
Australia
Bangladesh
Bermuda
England
India
Ireland
Kenya
Netherlands
New Zealand
Pakistan
Scotland
South Africa
Sri Lanka
West Indies
Zimbabwe



Data Observations:

Test:

- Above bubble plot if read horizontally then it's showing number of wins of country on y-axis against the country on x-axis.
- The visualization tuned more effective as with bubble size the number also present within the circle, so it's easy to render the chart with human eyes.
- As we can clearly see England and Australia are the top two countries having won most matches against each other, since they started playing cricket since a long time, as depicted by their number of wins, followed by South Africa, India, and West Indies

ODI:

- As we can clearly see Australia, India and Pakistan are the top countries having won most matches against each other, as depicted by their number of wins
- This is followed by South Africa, England, and West Indies teams

T20I:

- Most of the countries have won against each other.

Results

Insights from Visualizations

- There are three popular formats in International Cricket played across world and those are Test, ODI and T20I
- Cricket started with test Matches in 1870's but with the start of ODI's cricket became more popular.
- With the advent of T20I format, cricket got fast popularity and more and more countries started playing cricket.
- Players who plays longer time in international cricket have high chances to get themselves added into top players list
- Most of countries perform better at home grounds rather than foreign grounds.
- Not all countries are consistent performer across all formats of the game
- In Test cricket, there are high chances of No results, but with ODI and T20I cricket, almost every match has a result.
- Few countries like India, Australia, England, South Africa, New Zealand, Pakistan are top performers across different formats of the game
- In T20I, even new countries are competing with stalwarts.

Conclusion

With this project, I tried to show the way game of Cricket has moved across the years with the help of various visualizations as shown above. When trying to come up with visually intriguing and engaging plots, usually to tell a story, one must be sure they are still able to convey the message the story was created to convey.

Visual storytelling can be an extremely powerful tool, with appropriate application, as the decisions of many stakeholders may rely on it. While doing analysis on the data and creating visualizations, learned how graphs and visuals can be used in cricket to discover hidden but crucial insights. With these kinds of visualizations, I can collectively make better, data-driven decisions, which will, eventually get us back to a connected world; one global community linked together once again. Using the techniques and knowledge gained in the article, I can perform network analysis for other sports as well.

Future Work

As a next step to further visualize the data, I can:

- Predict the match result based on previous outcomes

Appendix:

Code URL:

DataVisualzation/Final_Project_Evolution_of_Cricket_uday.ipynb at main · udayIU/DataVisualzation (github.com)

udayIU/DataVisualzation: DataVisualzation (github.com)

References

FusionCharts, 2018 May 17, The Best Python Data Visualization Libraries
URL: https://www.fusioncharts.com/blog/best-python-data-visualization-libraries/

Kneoma 2013 Oct 10, Test – Cricket Player Performance
URL: https://knoema.com/lgzuseg/test-cricket-player-performance-card

Sakshi Srivastava, 2020 Mar 31, IPL Data Visualization
URL: https://github.com/Sakshisrivastava413/IPL_Data_Visualization

Nick Marsh, 2016 Feb 18, Cricket Performance on Touring
https://truii.com/data-curio-blog/sports-statistics/cricket-hardest-place-tour/

Dhilip Subramanian, 2019, Exploring IPL with Visualizations
https://towardsdatascience.com/exploring-indian-premier-league-with-interactive-visualizations-7a6ae053449

Andreas Pogiatzis, 2019, Step by Step Tutorial — Create a bar chart race animation
https://towardsdatascience.com/step-by-step-tutorial-create-a-bar-chart-race-animation-da7d5fcd7079

Gramener, 2012, Data Visualization

https://analyticsindiamag.com/data-visualization-a-pov-from-gramener/