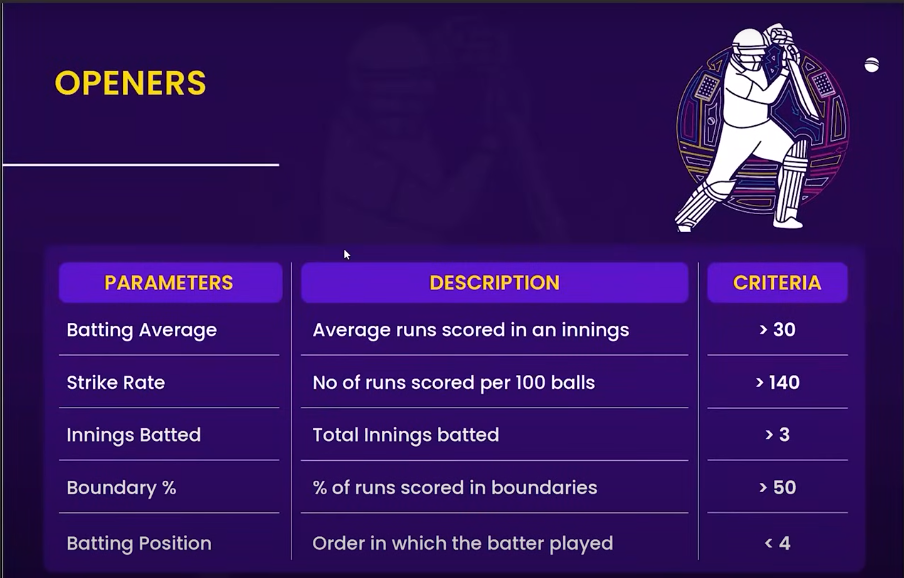
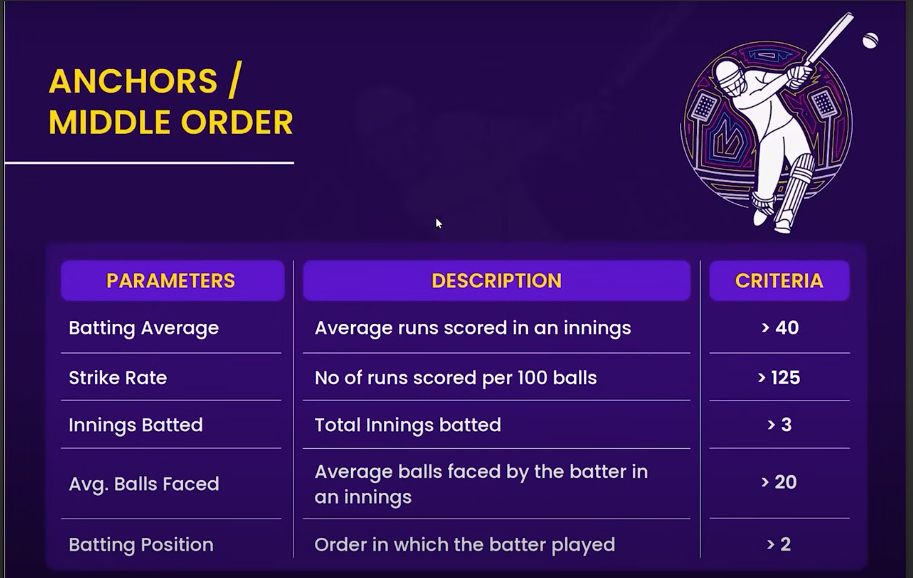
**T20 World Cup Cricket Data Analytics - Cricket Analysis [Python | Web scrapping | Pandas | Power BI]**

Requirement Scoping:

Selecting the parameters to Victory

1. The team should be able to score at-least 180 runs on an average
2. They should be to defend 150 runs on an average











**Web Scrapping**

**Collect the data from Cricket Website**

Best tool to scrape data from a **website is Brightdata.com**

First go to Collector in Brightdata, then click on ‘OpenIDE’. Then click on ‘start from scratch’, then we see the Java script UI, then paste the code given in the GitHub web\_scrapping\_codes for all ‘t20\_wc\_match\_results.js’, ‘t20\_wc\_batting\_summary.js’, ‘t20\_wc\_bowling\_summary.js’. ‘t20\_wc\_player\_info.js’, to get all the mentioned tables from ESPNCrickInfo website.

Or

We can directly **go to below GitHub** and collect the readymade tables

(codebasics data analysis project github) then go to CricketT20Analystics/data\_collection, then download all the Files present in it.

**Data Preparation using Pandas**

All the codes are written in t20\_data\_preprocessing using Jupyter notebook.

**Data Transformation/cleaning using Power Query**

Load the t20\_csv\_files into the Power query by click on ‘Get data’, then go to ‘More’, then select ‘Folder’ and then completely get the required t20\_csv\_folder, then click on ‘Transform’, it will go to Power Query.

In power query

Right click on the t20\_csv\_files & then create duplicates to get 4 required tables

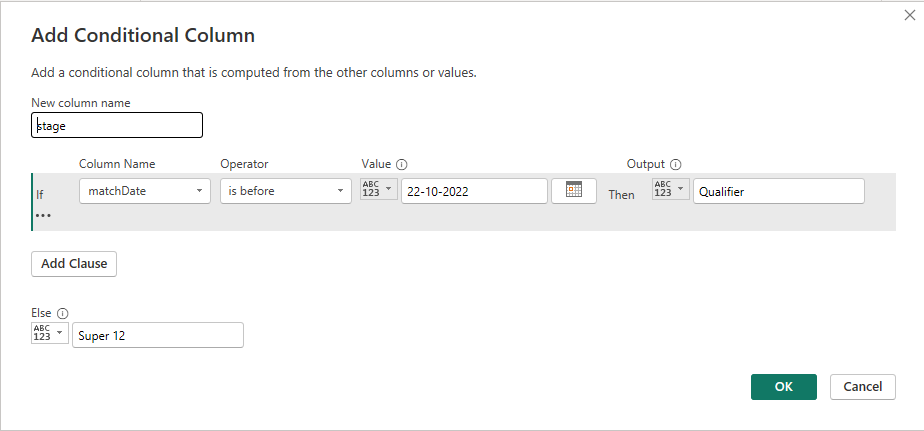
1. Dim\_match\_summary
2. Dim\_players
3. Fact\_batting\_summary
4. Fact\_bowling\_summary

**In dim\_players table we have done following transformations**

1. Use first row as headers
2. Then remove (c) from the player’s name by using ‘Text before delimiter’ and give ‘(’, then click on OK.
3. Then in ‘Format’ in ‘Transform’ tab use ‘Trim’ option to remove white spaces present in the text in the name column.
4. Sort name column from A-Z
5. Then remove the duplicates values by using ‘Remove Duplicates’ by right click on the name column

**In dim\_match\_summary table we have done following transformations**

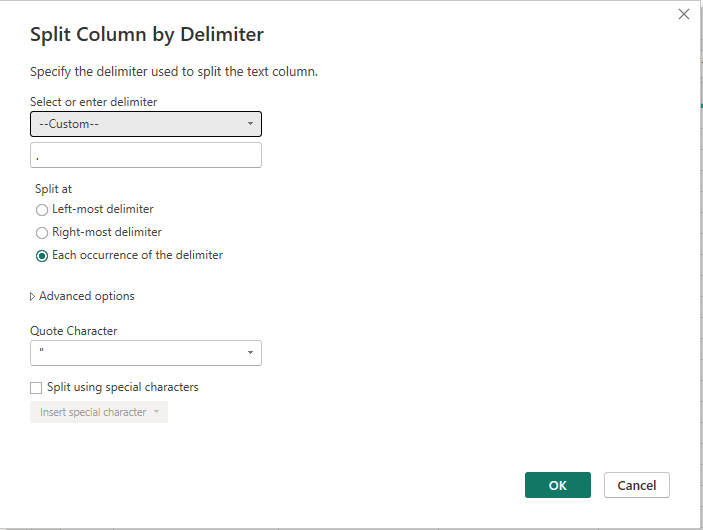
1. Create a New column called ‘stage’ ‘Conditional column’ in ‘Add column’ tab. If the date was before 22-10-2022, then those are ‘Qualifier’ matches, if the date was after 22-10-2022, then those are ‘Super 12’ matches. Shown the same below.



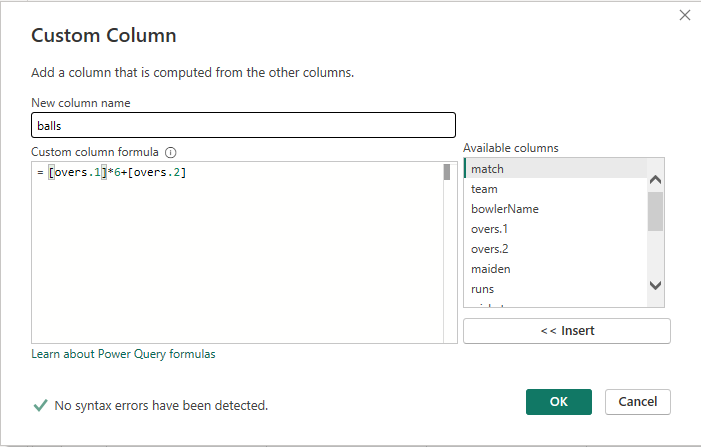
1. Change the ‘stage’ column data type from number to text data type.

**In dim\_bowling\_summary table we have done following transformations**

1. Rename columns from ‘bowlingTeam’ to ‘team’, ‘0s’ to ‘zeros’, ‘4s’ to ‘fours’, and ‘6s’ to ‘sixes’
2. Split the ‘overs’ column values for example if value has 2.5, it will split 2 into one column and 5 into another new column. By using ‘Split Column by Delimiter’ by giving ‘.’



1. Then Replace null values with zeros by using ‘Replace values’ in ‘overs.2’ column.
2. Add the new column ‘balls’ by using ‘Custom column’.



1. Change the text data type into ‘whole number’.

**In dim\_batting\_summary table we have done following transformations**

1. Rename columns from ‘teamInnings to ‘team’, ‘0s’ to ‘zeros’, ‘4s’ to ‘fours’, ‘6s’ to ‘sixes’ and ‘out/not\_out’ to ‘out’
2. In the ‘out’ column replace ‘out’ with value 1 and not\_out with 0 by using ‘Replace values’ operation.
3. Then remove (c) from the player’s name by using ‘Text before delimiter’ and give ‘(’, then click on OK.

Finally click on ‘Close & Apply’ to load the transformed data from Power query to Power BI desktop.

**Data Modelling and building parameters using DAX**

Build the relationship between the tables in ‘Model View’

1. By default, Power BI built the relationship between ‘dim\_match\_summary’, ‘fact\_bowling\_summary’ and ‘fact\_bowling\_summary’ by using ‘match\_id’ columns.
2. Newly we have to create relationship between ‘fact\_bowling\_summary’ and ‘dim\_players’ by using ‘bowlerName’ and ‘name’ columns respectively.
3. Newly we have to create relationship between ‘fact\_batting\_summary’ and ‘dim\_players’ by using ‘batsmanName’ and ‘name’ columns respectively.

**Create DAX measures**

To keep all the measures at one place will use ‘Enter data’, then rename as ‘Key\_measures’, then click on Load.

How to group the list of measures in Power BI Desktop.

First go to Model view, then click on the ‘Key\_measures’ table, then select any one measure, then in the properties, go to ‘Display folder’, give the folder name for example ‘Batting’, then press Enter. Now new folder is created, then drag and drop the other measures into the folder to keep it in single folder.

**List of measures**

**Batting**

1. Average Balls Faced = AVERAGE(fact\_batting\_summary[balls])
2. Batting Average = DIVIDE([Total Runs],[Total Innings Dismmissed],0)
3. Batting Position = ROUNDUP(AVERAGE(fact\_batting\_summary[battingPos]),0)
4. Boundary % = DIVIDE(SUM(fact\_batting\_summary[Boundary runs]),[Total Runs],0)\*100/100
5. Strike Rate = DIVIDE([Total Runs],[total balls faced],0)\*100
6. Total Balls Faced = SUM(fact\_batting\_summary[balls])
7. Total Innings Batted = COUNT(fact\_batting\_summary[match\_id])
8. Total Innings Dismmissed = SUM(fact\_batting\_summary[out])
9. Total Runs = SUM(fact\_batting\_summary[runs])

**Bowling**

1. Balls Bowled = SUM(fact\_bowling\_summary[balls])
2. Bowling Average = DIVIDE([Runs Conceded],[wickets],0)
3. Bowling Strike Rate = DIVIDE([balls Bowled], [wickets],0)
4. Dot ball % = DIVIDE(SUM(fact\_bowling\_summary[zeros]), SUM(fact\_bowling\_summary[balls]),0)\*100
5. Economy = DIVIDE( [Runs Conceded], ([balls Bowled]/6),0)
6. Runs Conceded = SUM(fact\_bowling\_summary[runs])
7. Total Innings Bowled = DISTINCTCOUNT(fact\_bowling\_summary[match\_id])
8. Wickets = SUM(fact\_bowling\_summary[wickets])

**Others**

1. Color Callout Value = if([Player Selection]="0", "#D0CF1D","#1D1D2E")
2. Display Text = if([Player Selection] = "1", " " ,"Select Player(s) by clicking

the player’s name to see their individual or combined strength.")

1. Player Selection = if(ISFILTERED(dim\_players[name]),"1","0")

List of calculated columns

Dim\_match\_summary

1. T1\_CODE = SWITCH(dim\_match\_summary[team1],

            "Namibia", "NMB",

            "Netherlands","NT",

            "Scotland","SCT",

            "Ireland","IRE",

            "Sri Lanka","SL",

            "West Indies","WI",

            "Australia","AUS",

            "Afghanistan","AFG",

            "India","IND",

            "Bangladesh","BAN",

            "South Africa", "SAF",

            "England", "ENG",

            "Pakistan", "PAK",

            "New Zealand", "NZ",

            "Zimbabwe", "ZIM",

            "U.A.E.","UAE"

            )

1. T2\_CODE = SWITCH(dim\_match\_summary[team2],

            "Namibia", "NMB",

            "Netherlands","NT",

            "Scotland","SCT",

            "Ireland","IRE",

            "Sri Lanka","SL",

            "West Indies","WI",

            "Australia","AUS",

            "Afghanistan","AFG",

            "India","IND",

            "Bangladesh","BAN",

            "South Africa", "SAF",

            "England", "ENG",

            "Pakistan", "PAK",

            "New Zealand", "NZ",

            "Zimbabwe", "ZIM",

            "U.A.E.","UAE"

            )

1. Match = dim\_match\_summary[T1\_CODE] & " vs " & dim\_match\_summary[T2\_code]

Dim\_players

1. Custom Batting Order = SWITCH(

    TRUE(),

dim\_players[name] = "Jos Buttler",1,

dim\_players[name] = "Rilee Rossouw",2,

dim\_players[name] = "Alex Hales",2,

dim\_players[name]  = "Virat Kohli",3,

dim\_players[name] = "Suryakumar Yadav" ,4,

dim\_players[name] = "Glenn Phillips" ,5,

dim\_players[name] = "Marcus Stoinis" ,6,

dim\_players[name] = "Glenn Maxwell" ,6,

dim\_players[name] = "Sikandar Raza" ,7,

dim\_players[name] = "Rashid Khan" ,8,

dim\_players[name] = "Shadab Khan" ,8,

dim\_players[name] = "Sam Curran" ,9,

dim\_players[name] = "Shaheen Shah Afridi" ,10,

dim\_players[name] = "Anrich Nortje" ,11

)

Fact\_batting\_summary

1. boundary runs = fact\_batting\_summary[fours]\*4 + fact\_batting\_summary[sixes]\*6

Fact\_bowling\_summary

1. Boundary runs = fact\_bowling\_summary[fours]\*4 + fact\_bowling\_summary[Sixes]\*6

**Build the visuals and dashboard in Power BI**

