9951666670
INNOMATICS
RESEARCH LABS
DATA SCIENCE | MACHINE LEARNING | ARTIFICIAL INTELLIGENCE
BIG DATA | DIGITAL MARKETING | AWS | DEVOPS

INNOMATICS
RESEARCH LABS

Project: Analysis on IPL T20

INNOMATICS RESEARCH LABS

About me:

- I, Rachakatla C S S Ramakrishna, hold a graduate degree in Mining Engineering.
- I am interested in data. So, I did research on data science and data analytics. There are many opportunities for both Data scientist and Data analyst which made me to learn this course.
- Limited members in a class and special mentoring sessions made me join Innomatics Research labs. I found classroom lectures, recording sessions, mentoring sessions are good.
- Assignment and projects with real time examples are giving me good knowledge and experience.

Use case & Domain:

Domain: Sports(cricket)

Use case:

- IPL is a professional T20 cricket league in India contested during April and may of every year by teams representing Indian cities.
- Data of 5 IPL seasons from 2015-2020
- Match details
- Player details
- Season wise highest score
- Average

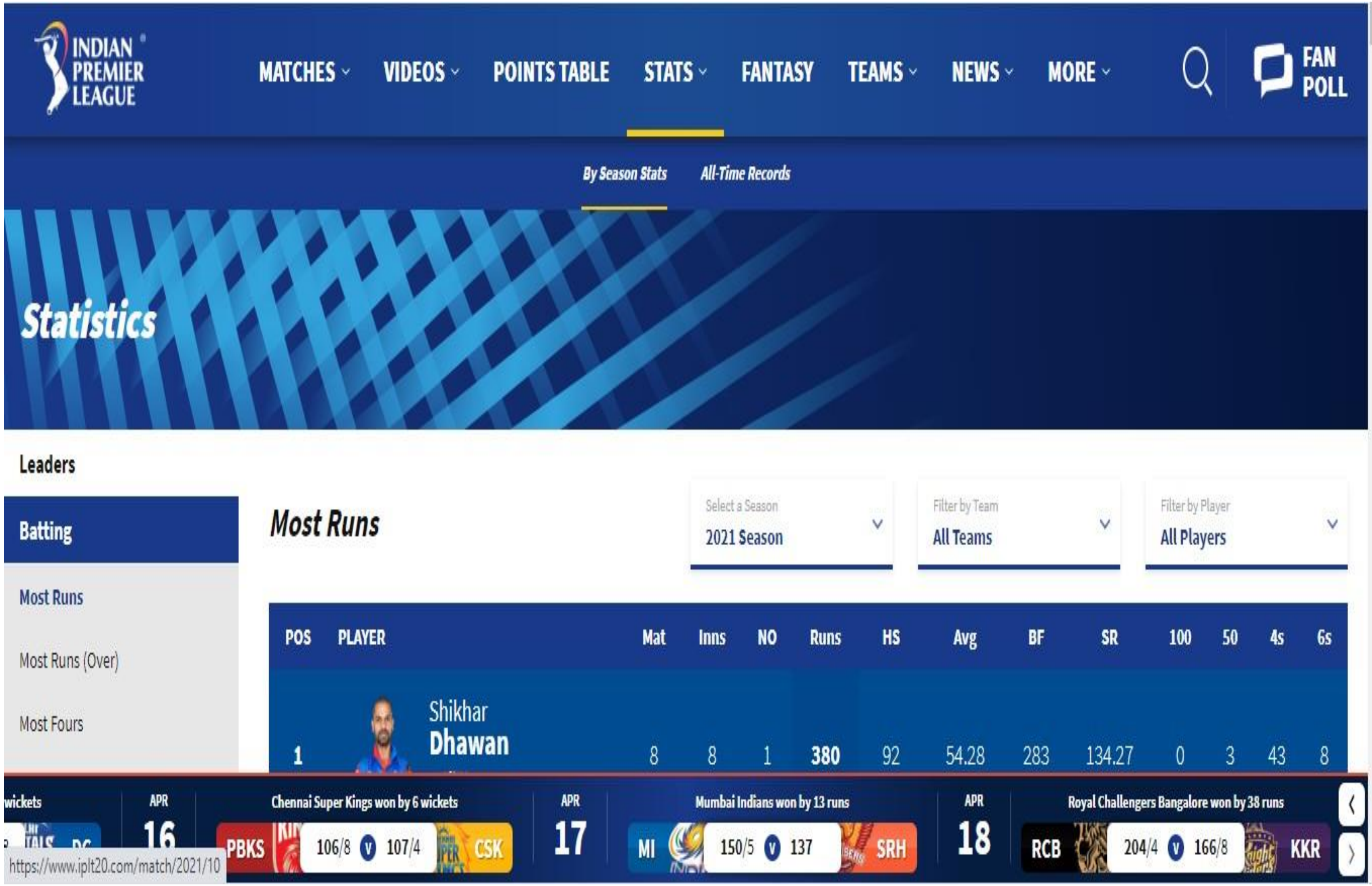
Business understanding of Use case:

Analysis of the use case provides below insights to business:

- List of batsman
- List of bowlers
- Best bowler
- Consistency of player
- Types of batsman & bowlers

Possible website for scraping data:

IPLT20.COM



Most Runs

Select a Season: 2021 Season

Filter by Team: All Teams

Filter by Player: All Players

POS	PLAYER	Mat	Inns	NO	Runs	HS	Avg	BF	SR	100	50	4s	6s
1	Shikhar Dhawan	8	8	1	380	92	54.28	283	134.27	0	3	43	8

wickets

DC

APR 16

Chennai Super Kings won by 6 wickets

PBKS 106/8 V 107/4 CSK

APR 17

Mumbai Indians won by 13 runs

MI 150/5 V 137 SRH

APR 18

Royal Challengers Bangalore won by 38 runs

RCB 204/4 V 166/8 KKR

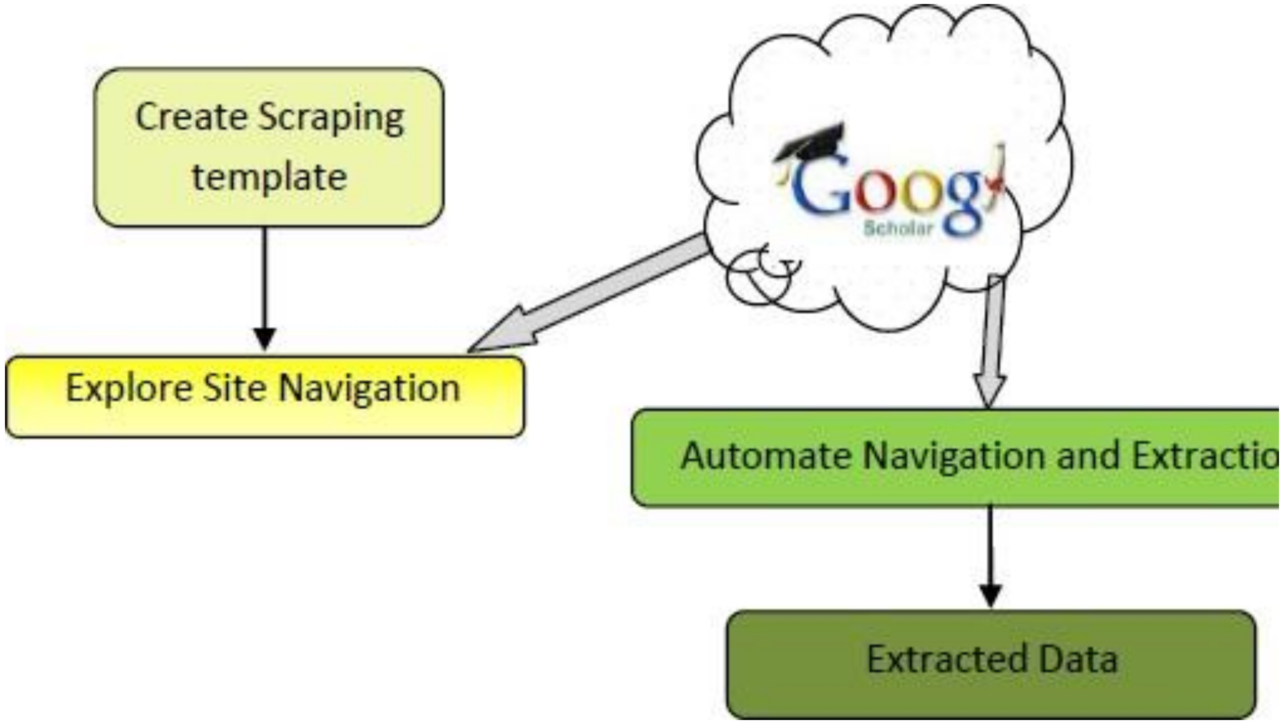
https://www.iplt20.com/match/2021/10

INNOMATICS RESEARCH LABS

Data important for scraping and why?

- Ipl data is to get the hands-on match data.
- To slice and dice the data with an intent to gain insights into the teams/players.
- The more ambitious is to predict the results of these games.
- The ability to collect data from a source all by ourselves and in the format that we would like.
- we get to decide how and what data we scrape from the data available at the source.
- www.iplt20.com to scrape the match and scores data

Scraping methodology:

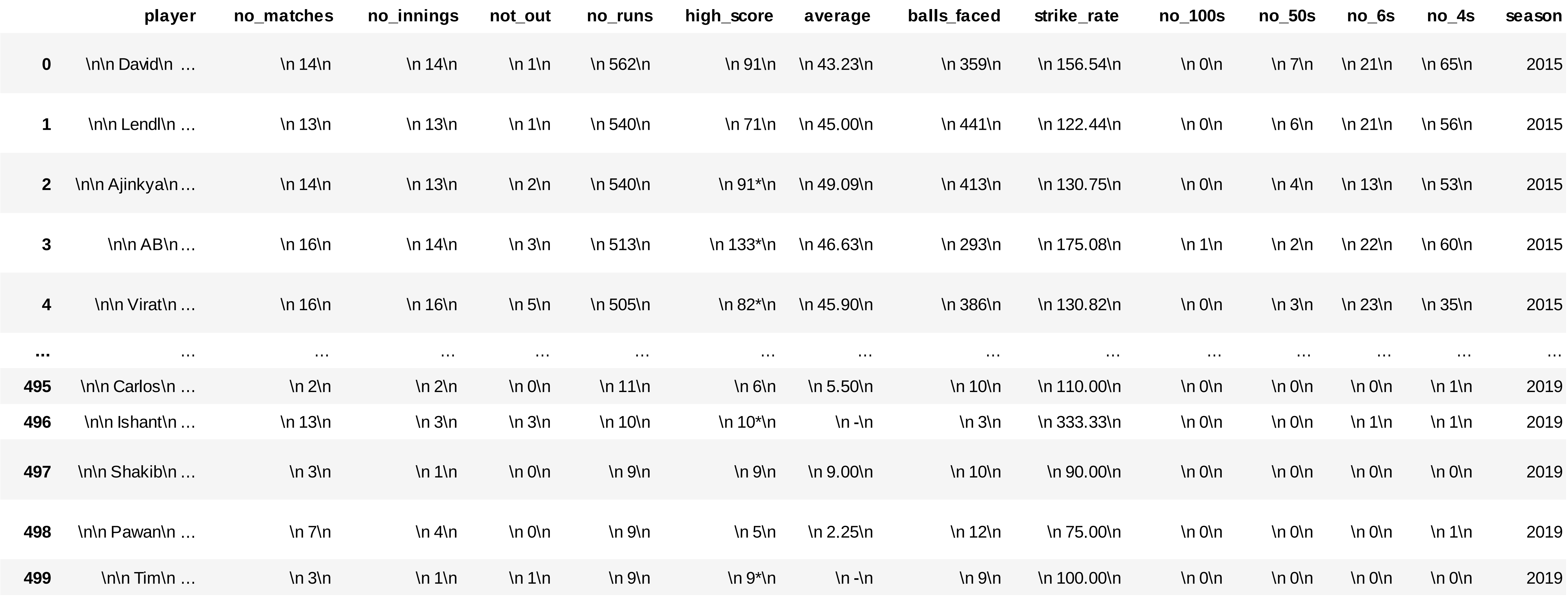


- Explore the site www.iplt20.com
- Create scraping template
- Extract all match links

•Navigate match links and scrape the match details one by one.

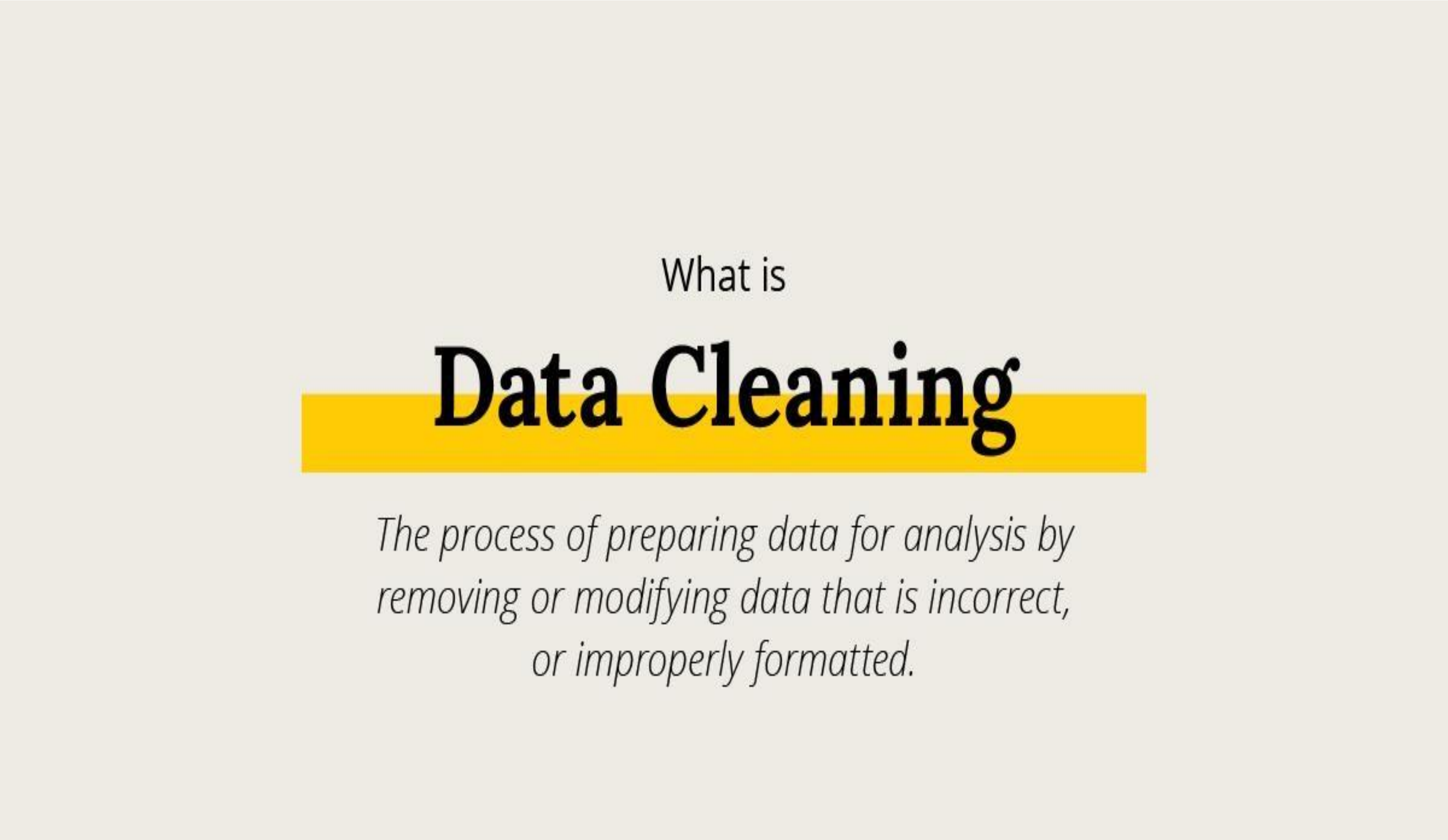
- Extract the data .
- Save it to a data frame and finally export it to a csv file.

IPL data frame after web scraping



	player	no_matches	no_innings	not_out	no_runs	high_score	average	balls_faced	strike_rate	no_100s	no_50s	no_6s	no_4s	season
0	\n\n David\n ...	\n 14\n	\n 14\n	\n 1\n	\n 562\n	\n 91\n	\n 43.23\n	\n 359\n	\n 156.54\n	\n 0\n	\n 7\n	\n 21\n	\n 65\n	2015
1	\n\n Lendl\n ...	\n 13\n	\n 13\n	\n 1\n	\n 540\n	\n 71\n	\n 45.00\n	\n 441\n	\n 122.44\n	\n 0\n	\n 6\n	\n 21\n	\n 56\n	2015
2	\n\n Ajinkya\n ...	\n 14\n	\n 13\n	\n 2\n	\n 540\n	\n 91*\n	\n 49.09\n	\n 413\n	\n 130.75\n	\n 0\n	\n 4\n	\n 13\n	\n 53\n	2015
3	\n\n AB\n ...	\n 16\n	\n 14\n	\n 3\n	\n 513\n	\n 133*\n	\n 46.63\n	\n 293\n	\n 175.08\n	\n 1\n	\n 2\n	\n 22\n	\n 60\n	2015
4	\n\n Virat\n ...	\n 16\n	\n 16\n	\n 5\n	\n 505\n	\n 82*\n	\n 45.90\n	\n 386\n	\n 130.82\n	\n 0\n	\n 3\n	\n 23\n	\n 35\n	2015
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
495	\n\n Carlos\n ...	\n 2\n	\n 2\n	\n 0\n	\n 11\n	\n 6\n	\n 5.50\n	\n 10\n	\n 110.00\n	\n 0\n	\n 0\n	\n 0\n	\n 1\n	2019
496	\n\n Ishant\n ...	\n 13\n	\n 3\n	\n 3\n	\n 10\n	\n 10*\n	\n -\n	\n 3\n	\n 333.33\n	\n 0\n	\n 0\n	\n 1\n	\n 1\n	2019
497	\n\n Shakib\n ...	\n 3\n	\n 1\n	\n 0\n	\n 9\n	\n 9\n	\n 9.00\n	\n 10\n	\n 90.00\n	\n 0\n	\n 0\n	\n 0\n	\n 0\n	2019
498	\n\n Pawan\n ...	\n 7\n	\n 4\n	\n 0\n	\n 9\n	\n 5\n	\n 2.25\n	\n 12\n	\n 75.00\n	\n 0\n	\n 0\n	\n 0\n	\n 1\n	2019
499	\n\n Tim\n ...	\n 3\n	\n 1\n	\n 1\n	\n 9\n	\n 9*\n	\n -\n	\n 9\n	\n 100.00\n	\n 0\n	\n 0\n	\n 0\n	\n 0\n	2019

INNOMATICS
RESEARCH LABS



What is

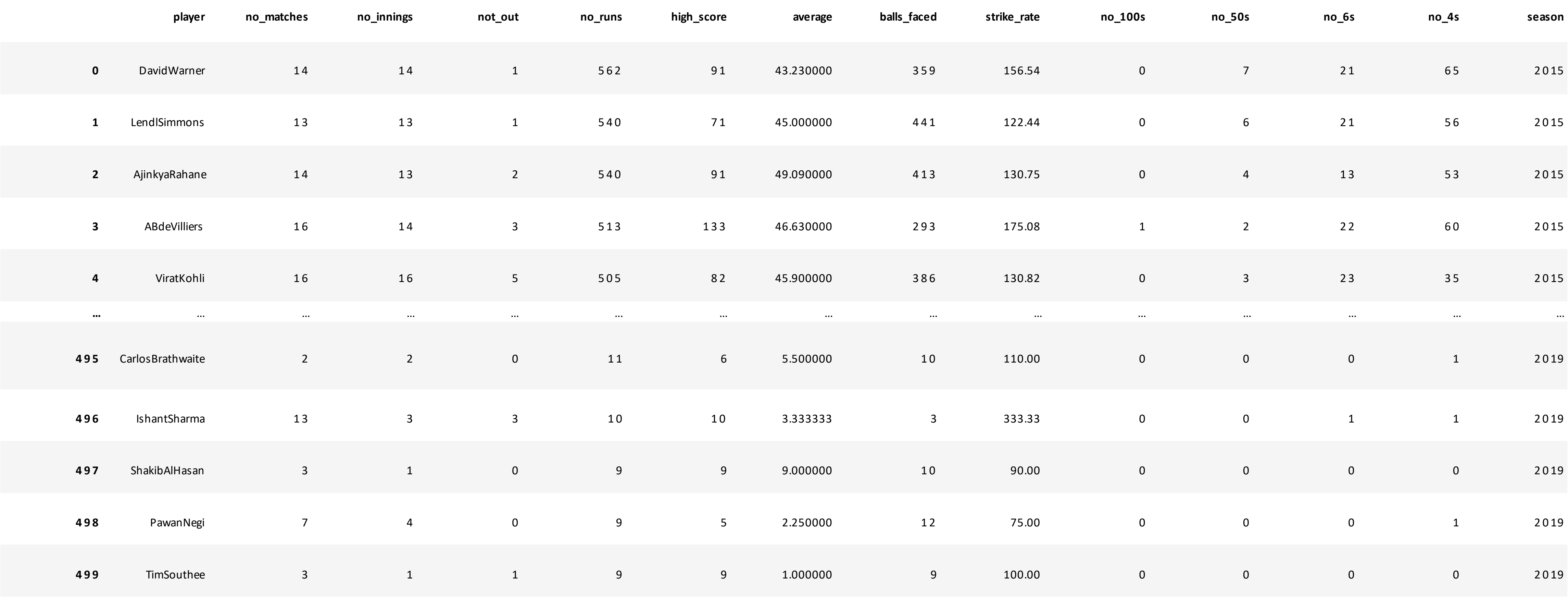
Data Cleaning

The process of preparing data for analysis by removing or modifying data that is incorrect, or improperly formatted.

Challenges faced for collection and cleaning:

- Removing \n for each and every element in data frame.
- Remove * in column high_score.
- These above data cleaning is done using lambda functions.
- Using Regular expression, player column is made into sequence of characters.
- Casting pandas objects to a specified datatype using astype() method.
- Average.isnull().sum() -to get the count of null values in Average column
- Fill the null values using formula- ipl_df['average'].fillna(ipl_df["no_runs"]/ipl_df["balls_faced"]

IPL data frame after data cleaning:



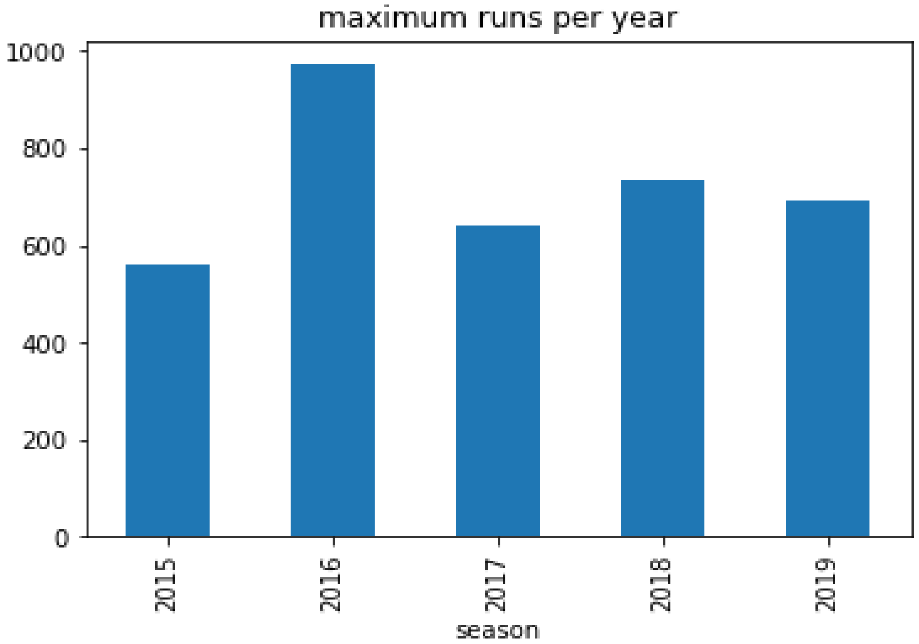
	player	no_matches	no_innings	not_out	no_runs	high_score	average	balls_faced	strike_rate	no_100s	no_50s	no_6s	no_4s	season
0	DavidWarner	14	14	1	562	91	43.230000	359	156.54	0	7	21	65	2015
1	LendlSimmons	13	13	1	540	71	45.000000	441	122.44	0	6	21	56	2015
2	AjinkyaRahane	14	13	2	540	91	49.090000	413	130.75	0	4	13	53	2015
3	ABdeVilliers	16	14	3	513	133	46.630000	293	175.08	1	2	22	60	2015
4	ViratKohli	16	16	5	505	82	45.900000	386	130.82	0	3	23	35	2015
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
495	CarlosBrathwaite	2	2	0	11	6	5.500000	10	110.00	0	0	0	1	2019
496	IshantSharma	13	3	3	10	10	3.333333	3	333.33	0	0	1	1	2019
497	ShakibAlHasan	3	1	0	9	9	9.000000	10	90.00	0	0	0	0	2019
498	PawanNegi	7	4	0	9	5	2.250000	12	75.00	0	0	0	1	2019
499	TimSouthee	3	1	1	9	9	1.000000	9	100.00	0	0	0	0	2019

Data visualization:

- The process of displaying data in graphical charts, figures and bars.
- Understands the trends and patterns of data.
- Know the distribution of the variables in data.
- Visualize the relationship that exists between different variables.
- The number of variables of interest featured by the data classifies it as:
    - Univariate
    - Bivariate
    - Multivariate

Univariate data visualization

The data features only one variable of interest.



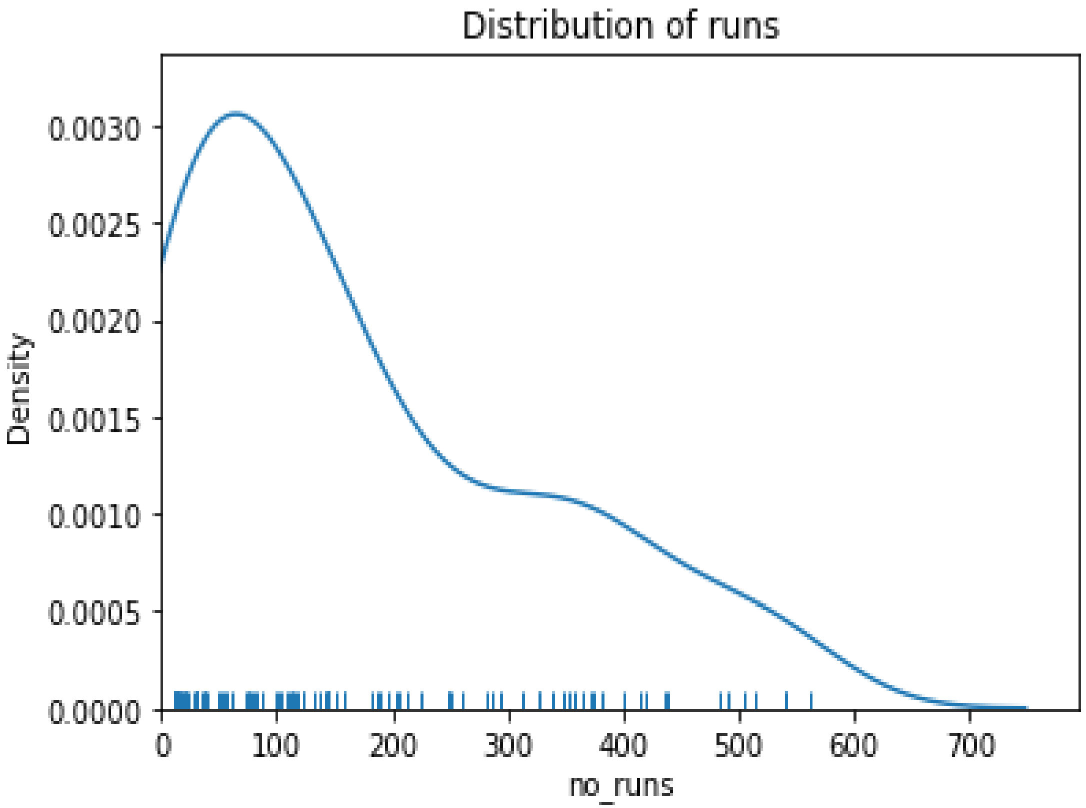
```
max_runs=ipl_df.groupby(["season"])["no_runs"].max().plot.bar()
plt.title("maximum runs per year")
max_runs
```

Bar plot:

- X-axis represents season from 2015-2019.
- Y- axis represents the number of runs
- This bar plot says 2016 has the highest value number of runs. It appeared at 972

Distribution plot:

Visually access the sample data comparing empirical distribution of data with theoretical values.

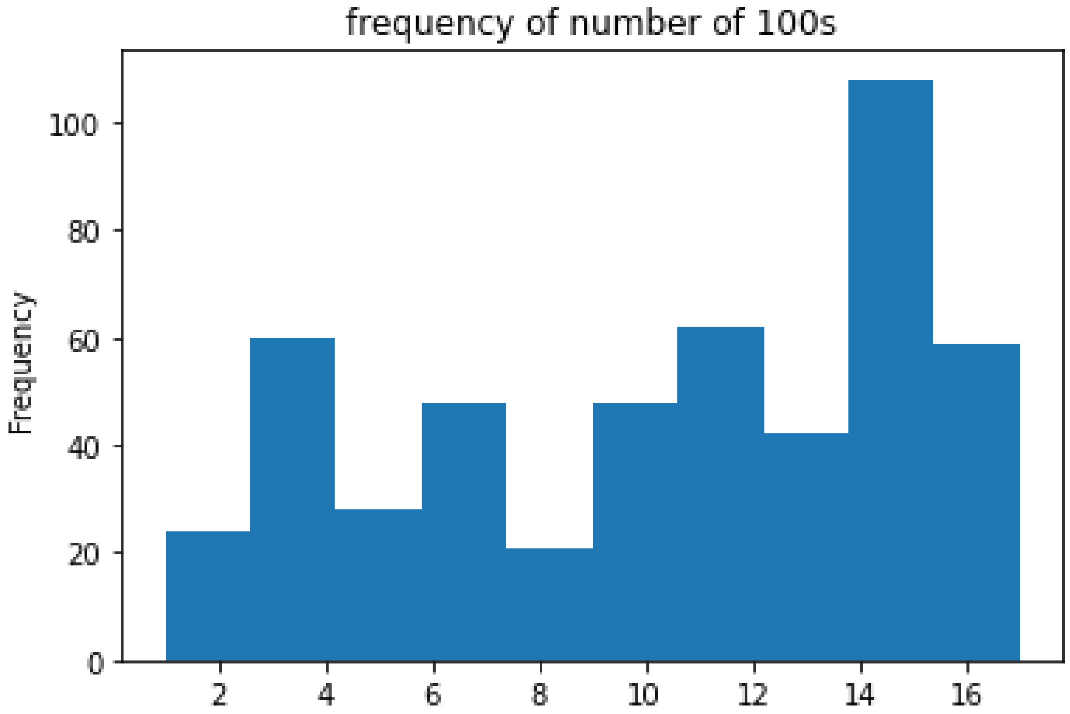


sns.distplot(ipl_df[ipl_df["season"]==2015]['no_runs'],rug=True, hist=False).set(xlim=(0),title="Distribution of runs")

- The relation between number of runs and density has been depicted via distribution plot.
- The X axis indicate the number of run in a single season – 2015
- Y axis indicate the density of no_runs
- The current plot shows that there are more number of players who scored between 0-150
- Only 4 players scored above 500 runs in that season.

Histogram:

Representation of data that buckets a range of outcomes into columns along x-axis

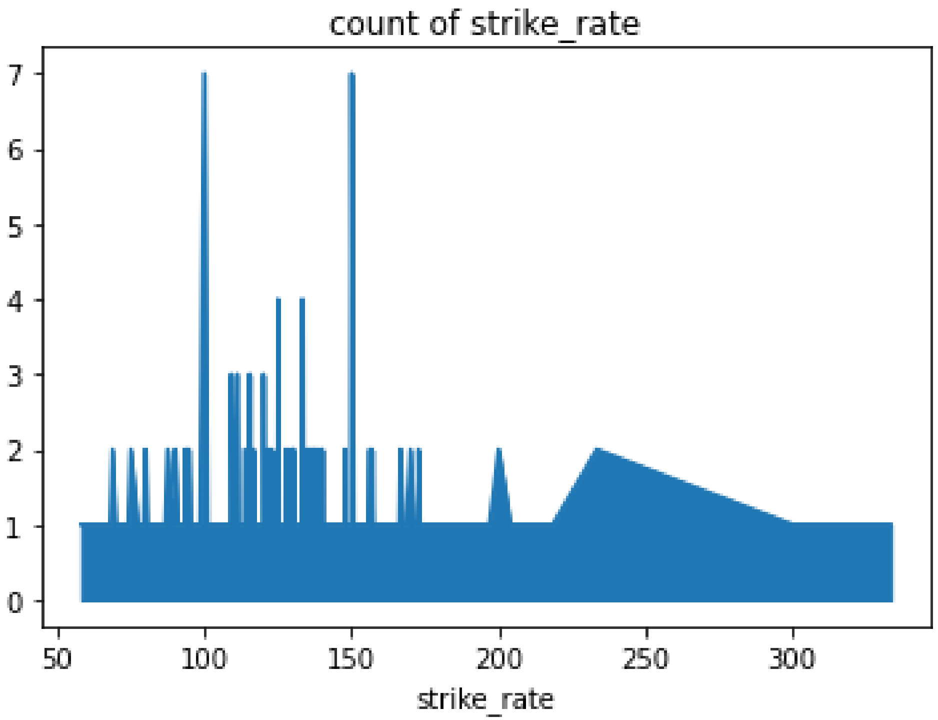


ipl_df['no_matches'].plot.hist(xlabel="no_matches",title="frequency of number of 100s")

- This plot shows the number of 100s and its frequency in Ipl
- Frequency for intervals of data are high in between 14-16 at 100
- Frequency for intervals of data number of 100s are low for 8 is in between 18-20

Area plot:

Displays graphically quantitative data



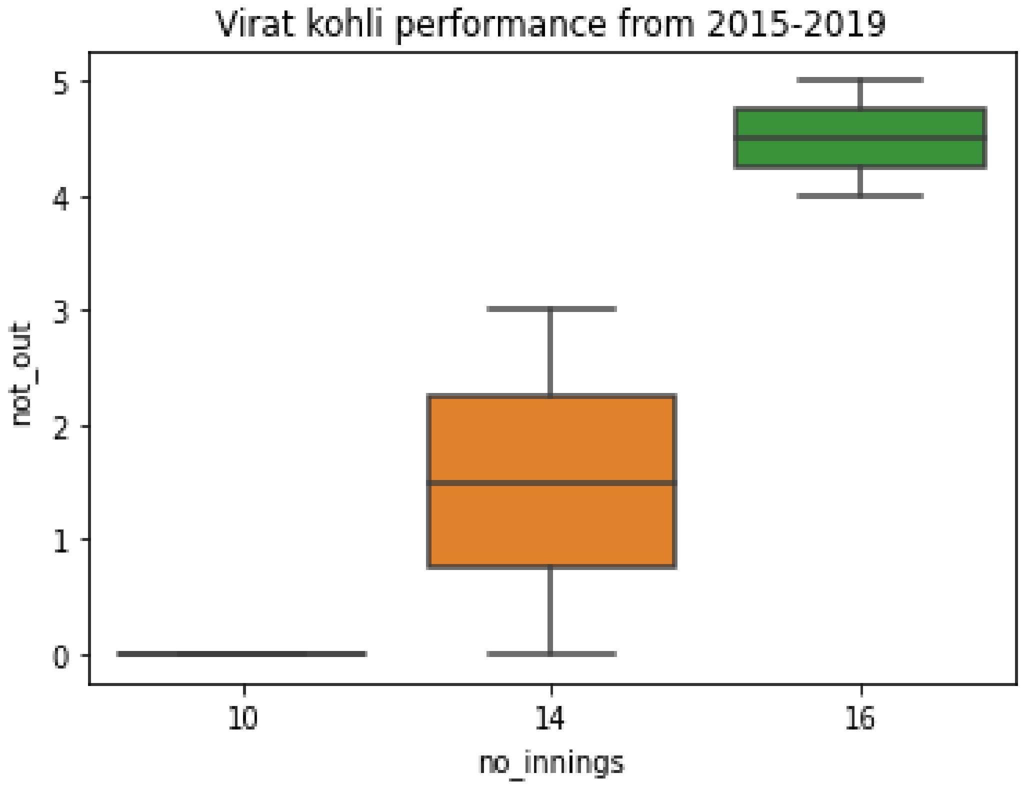
From this plot we can refer :

- Quantitative data of strike rate data counts varies between 0-7
- For 100 & 150 strike rate the count is very high
- After 300 the strike rate is too low
- Most of the strike rate is in between 1-2
- Strike rate between 100-150 is little high up to 4

```
ipl_df['strike_rate'].value_counts().sort_index().plot.area(xlabel
="strike_rate",title="count of strike_rate")
```

Bivariate data analysis:

Purpose for determining the empirical relation between two variables



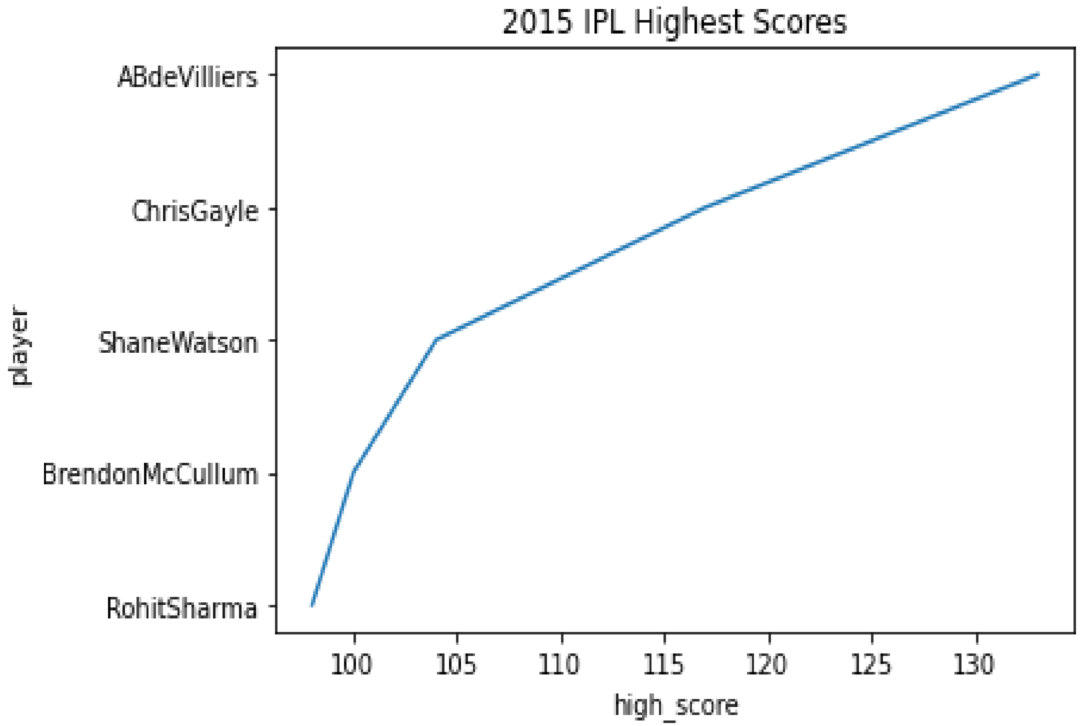
Box plot:

- The current box plot indicates the following details – Minimum value, Maximum value, Median, 25th percentile and 75th percentile values
- In innings where Virat Kohli's performance is poor, there were NO not outs.
- In Innings where Virat performance is good, the number of not outs varied between 0-3
- In innings where Virat performance is best, the number of not outs varied between 4-5

sns.boxplot(x='no_innings',y='not_out',data=player3).set(title=" Virat kohli performance from 2015-2019")

Line plot:

Graph that represents data using a number line

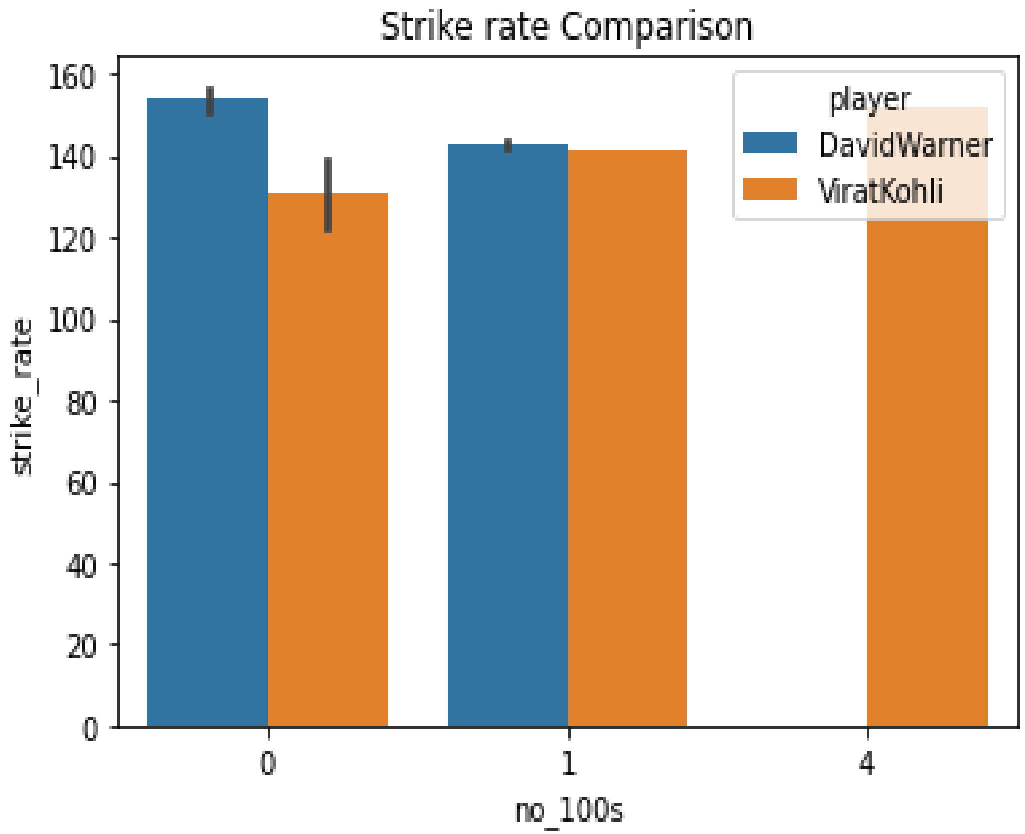


- The Line plot indicates the highest scores of the IPL season 2015.
- The highest score was 130+ scored by AB deVilliers
- The next highest score was 115+ score by Chris Gayle and so on

sns.lineplot(data=player19,x="high_score",y="player")
.set(title="2019 IPL Highest Scores")

Bar plot:

Represents the estimate of central tendency.

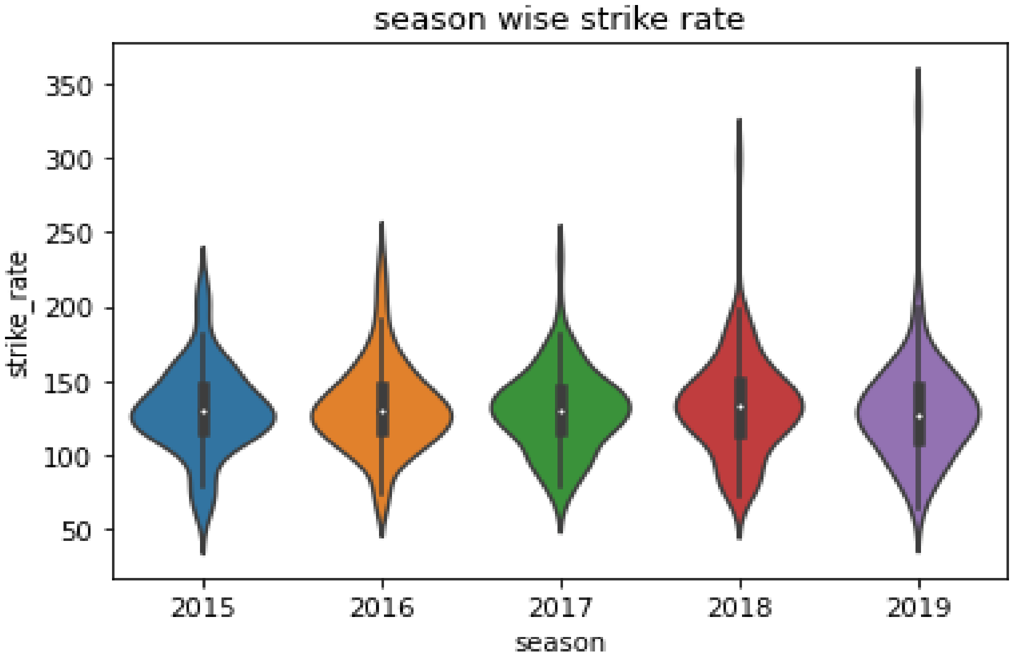


sns.barplot(x='no_100s',y="strike_rate",hue="player",data=player4).set(title="Strike rate Comparison")

- Strike rate vs Number of hundreds ; comparison between David warner and Virat Kohli over 5 seasons of IPL
- The Strike rate of David warner is good compared to Kohli
- However, the number of centuries scored by Virat Kohli are more compared to Warner

Violin plot:

Shows the distribution of quantitative data across several levels of one or more

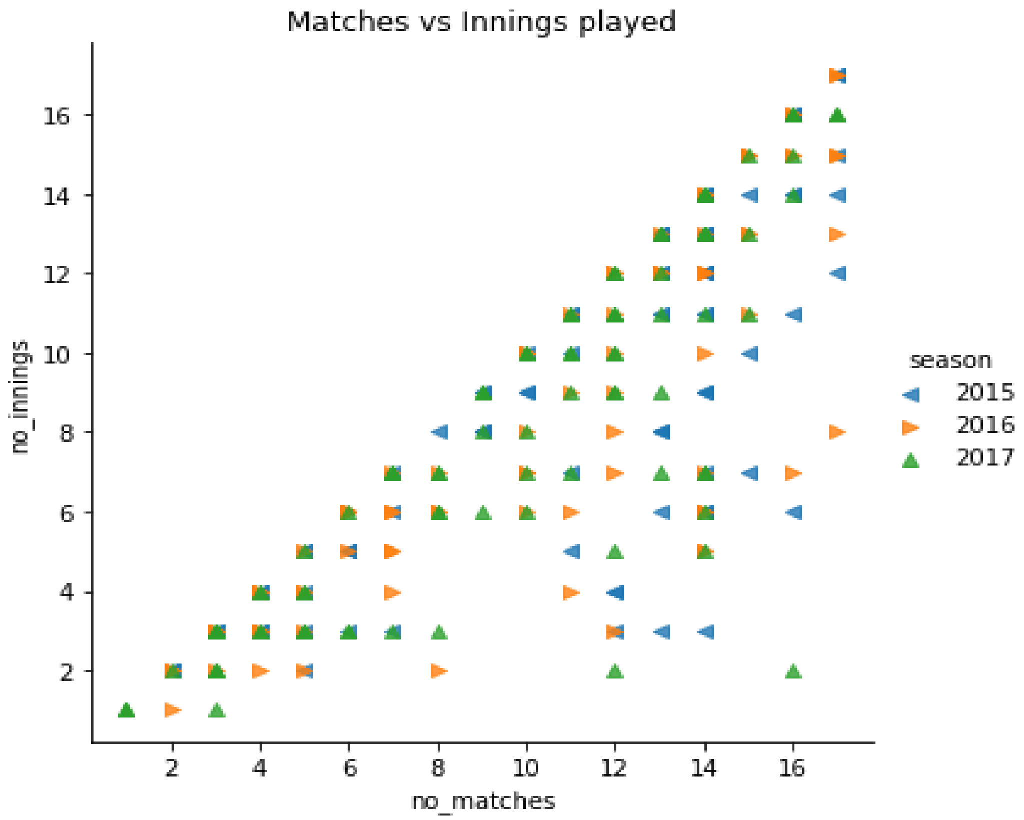


sns.violinplot(x="season", y="strike_rate", data=ipl_df).set_title("season wise strike rate")

- Density is mirrored and flipped over resulting violin shape.
- Lower strike up to 240 is in season 2015
- 2019 is highest season for strike rate 350
- Second highest season is 2018 with 300 strike rate

Multivariate analysis:

To reveal the relationship among various variables simultaneously



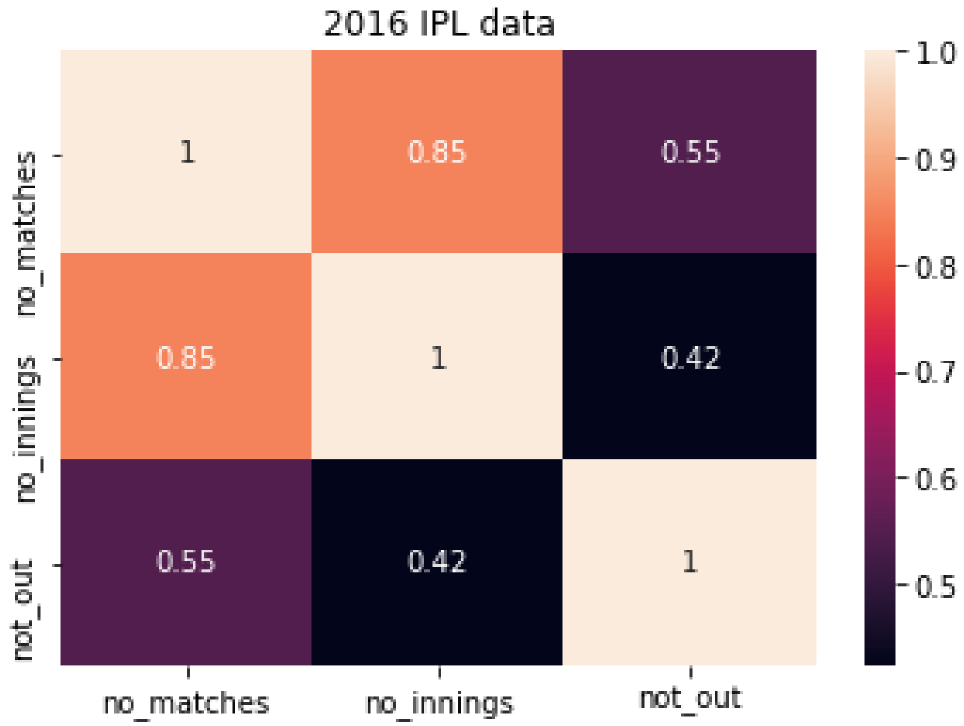
sns.lmplot(x='no_matches', y='no_innings', hue='season', markers=['<','>','^'], data=ipl_df.loc[ipl_df['season'].isin([2015, 2016, 2017])],fit_reg=False).set(title="Matches vs Innings played")

Scatter plot:

- The comparison is made between Matches and Innings over a period of 3 seasons
- Season 2015 is indicated by <
- Season 2016 is indicated by >
- Season 2017 is indicated by ^
- The distribution shows the number of innings played by each player among all the matches that they have played.

Heat map:

Plots rectangular data as a colour encoded matrix



Heat map shows the correlation between data.

From this plot we can find:

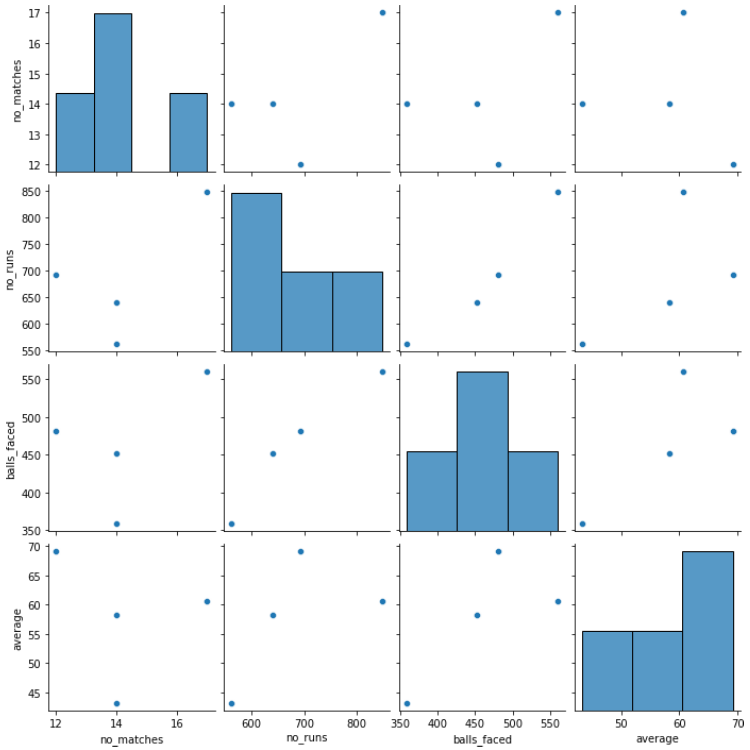
- In this plot relationship varies from 0 to 1 and represented in color boxes.
- No_matches, no innings & no_runs are highly correlated
- 1 perfect correlation

```
p=(player21.loc[:, ['no_matches', 'no_innings',
'not_out']].corr())
sns.heatmap(p,annot=True).set(title="2016 IPL data")
```

INNOMATICS RESEARCH LABS

Pair plot:

Pair wise relationship in a dataset



From this plot we can refer:

- Number of runs of player are positively correlated to no_matches.
- balls faced are corelated to runs, average, matches
- Average run rate of players are related to runs and balls faced

```
sns.pairplot( player2, x_vars=["no_matches","no_runs", “balls faced"], y_vars=["no_matches","no_runs",“balls faced"],diag_kind="hist")
```

Interpretation and insights:

- Here are some of examples from my project.
- From the analysis of visualization we observe that Virat Kohli has highest number of centuries throughout IPL seasons.
- In innings Virat's performance Is poor there were no not outs, where Virat's performance is good no of not outs varied between 0-3, where Virat's performance is best no of not outs varied between 4-5.
- I conclude that Virat Kohli are having more chance to win in upcoming matches.
- To continue this success sponsors has to select him as the best batting players as top most players which leads the team win the success.
- From the analysis of visualisation, David Warner has a good strike rate through out seasons compared to others.
- I conclude that David Warner are having more chances of winning. So, there may be chance for next seasons David performance will be high.

Experience after the project:

- Initially this was very challenging for me because the subject is very new and hard to understand.
- Lecturer and mentor were extremely helpful & approachable which made me feel comfortable working on this project.
- Although it was less than 2 months I have learnt a lot starting from basics.
- My knowledge on different aspects like data scraping, data visualization has expanded tremendously.
- Assignments are very helpful to practise & clear my doubts which helped me to do this project successful.

Future scope of project:

- We can know individual player performances
- Runs scored by each batsman
- The number of wickets taken by each bowler
- Matches won by individual teams etc.
- This data play a significant role in how teams operate, pick their players, how they play a game etc.
- The teams and individual players can dig deep into data and find areas of improvement.
- It can also be useful to assess an opponents strengths and weaknesses

THANK YOU