An Assignment Report

*On*

# Data Exploration and Analysis

*by*

Aditya Sharma

AU18B1009

*Under the guidance of*

[Satyajit Pangaonkar](https://avantika.instructure.com/courses/345/users/904)



School of Engineering

Avantika University, Ujjain

**2020-2021**

**Content: Page**

1. **Introduction 2**
2. **Title 3**
3. **Statement 3**
4. **Identification of dataset 3 - 4**
5. **Analysis of Attributes (Feature Engineering) 5 - 7**
6. **Identification of Software Dependencies 7**
7. **Key Learning and outcomes 8**

**Introduction:**

We have the Indian Premier League data from 2008 to 2016. We have two files to begin with. Matches - giving us details of each match played. Deliveries - all the deliveries in all the matches put together. We would work on these two files to extract data on each match, batsmen and bowlers.

For all the cricket fans in India IPL is like a festival. During the period of IPL matches, wherever you go, you can hear somebody talking about the matches and player’s skills. Everyone want to analyse the IPL matches to understand the trends and performance of different teams in IPL. Data exploration and visualization has been carried out using Matplotlib, Numpy, Pandas and Seaborn visualization libraries.

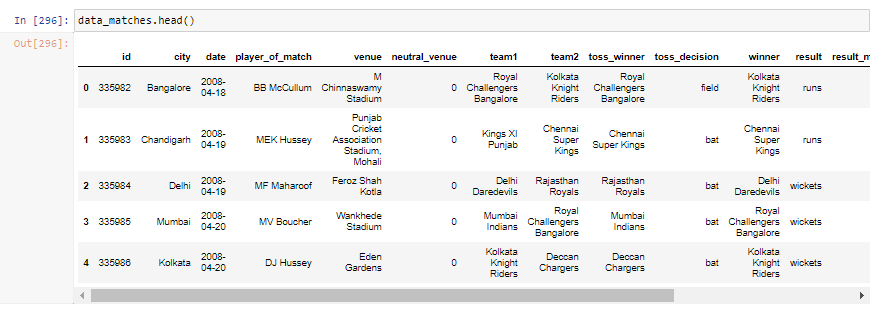
**Title:**

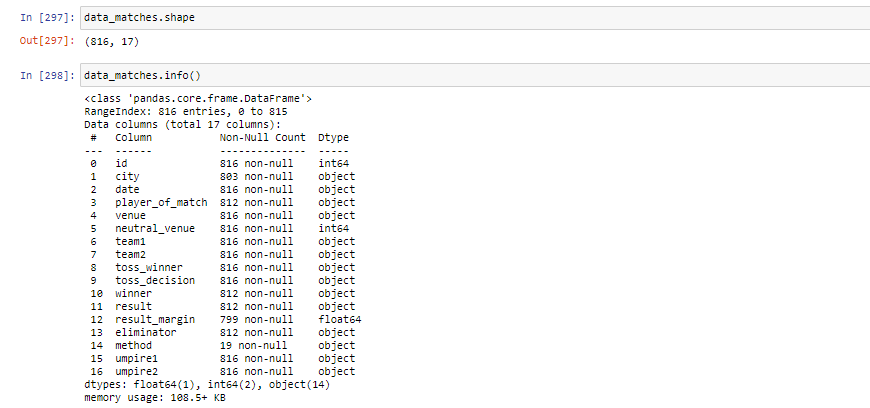
Data Exploration and Analysis

**Statement:**

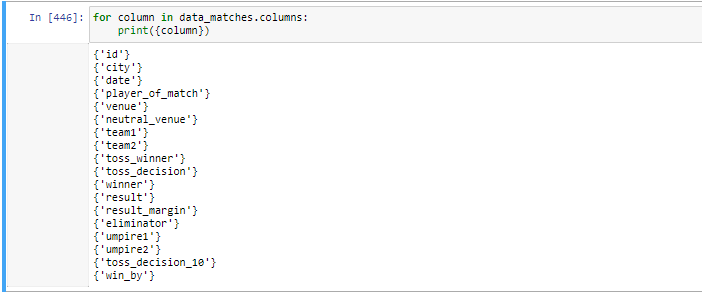
Select the dataset available resources as well as shared by the course advisor. Analyse at least 5-6 attributes of the dataset provided. Identify the learning module. Identify software dependencies. Define the additional requirement and functionality implemented.

1. **Identification of the Dataset:**
2. **Type of the Dataset:** Multivariate.
3. **Data Quality and Analysis:**



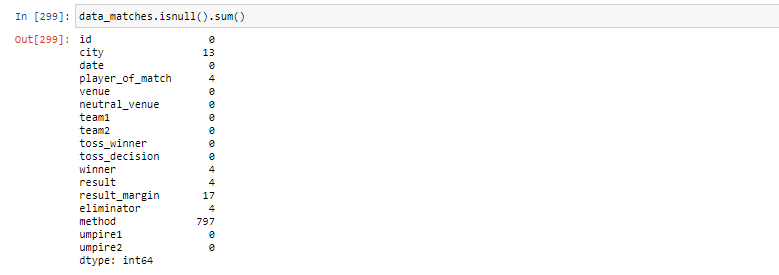


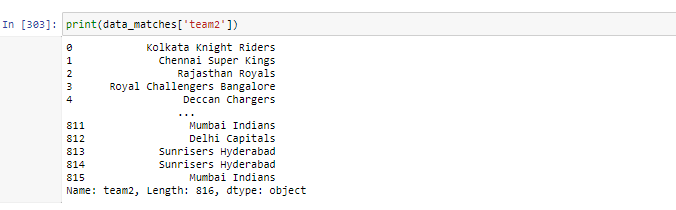
1. **Attributes/ Features:**

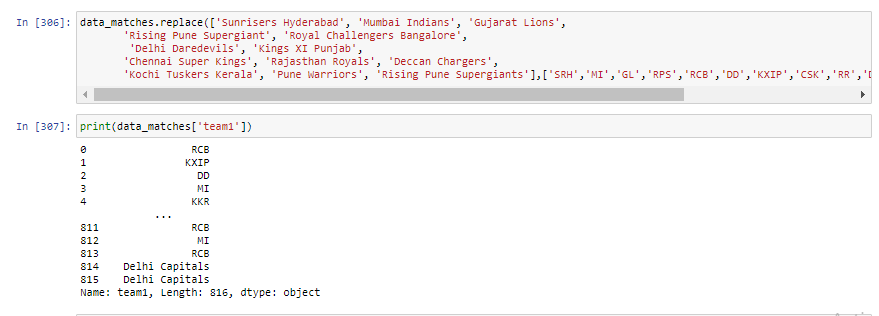


* + id
  + city
  + date
  + player\_of\_match
  + venue
  + neutral\_venue
  + team1
  + team2
  + toss\_winner
  + toss\_decision
  + winner
  + result
  + result\_margin
  + eliminator
  + umpire1
  + umpire2
  + toss\_decision\_10
  + win\_by

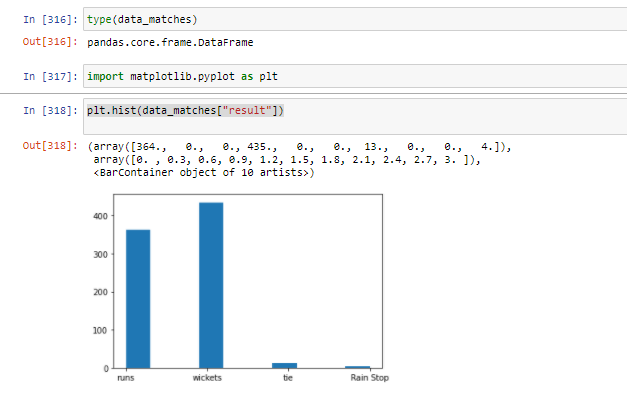
1. **Format of the Dataset:** CSV
2. **Analysis of Attributes (Feature Engineering)**
3. **Feature Pre-Processing:**



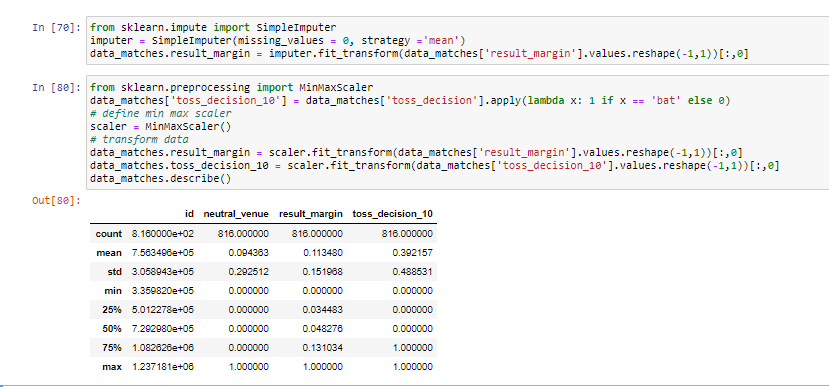








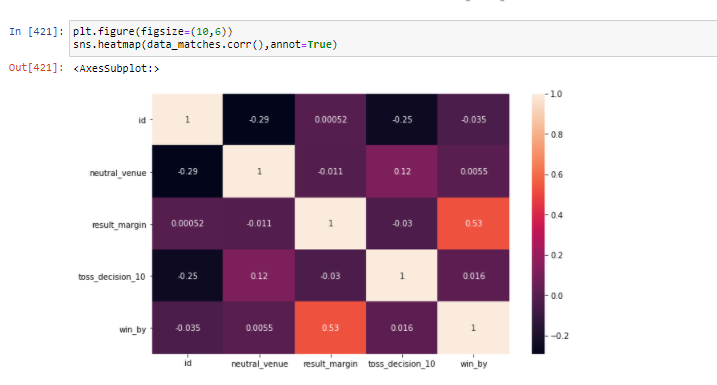




1. **Used method:**

* Interpolate
* Fillna
* Drop
* SKIPROW
* SIMPLE IMPUTE
* STANDARD MINMAX TRANSFORMATION

1. **Format of the Dataset:** CSV
2. **Identification of Software Dependencies**
3. **Features correlations:**



1. **Other software dependencies:**

**Libraries Included:**

os,pandas,numpy,seaborn,matplotlib.pyplot , sklearn.impute , sklearn.preprocessing

1. **Any further requirements :**

**Software Used:**

Anaconda Navigator, Jupyter Notebook

**Model Type Identification:**

Supervised Regression Model

1. **Key Learning Outcomes:**

We philosophised about the variables, analysed ‘toss\_results’ and with the most correlated variables, we dealt with missing data and outliers, we tested some of the fundamental statistical assumptions and we even transformed categorial variables into dummy variables. That's a lot of work that Python helped us make easier.