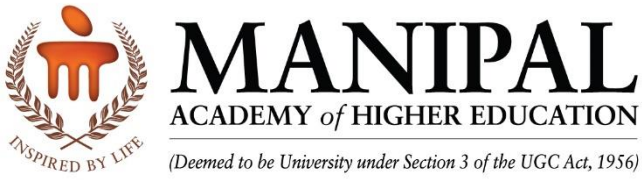
Analysis of Olympic Data

Mini-Project Report

Submitted to



By

Srinivas K
181046001
Big Data and Data Analytics

UDAY T
181046016
Big Data and Data Analytics

Under the guidance of
Deepak Rao
Assistant professor
SOIS, MAHE, Manipal

Abstract

The purpose of this project to do an efficient Data Analysis using the raw dataset of the Olympic events and results from the web. This involves transforming the extracted data by preprocessing, cleaning and categorizing the obtained data and then using data analysis, try to find the major patterns which affected the results and visualize the obtained patterns.

Contents

Introduction

Project Design

Methodology

Requirement Specification

Proposed Analysis

Future Analysis

References

Introduction

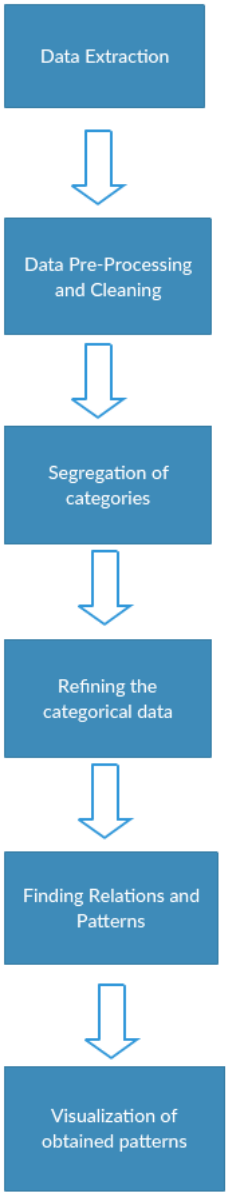
Olympic Games have been the biggest sporting event hosted for over 100 - 150 years, hence there is a large amount of data that is generated over these years. Here we have a dataset that comprises all the information about all the athletes who have participated for every Olympic Games since the inaugural games of 1896.

The 'Olympic Games' dataset comprises of 15 columns:

- ID – Unique no for each athlete
- Name – Athlete Name
- Sex – Male/Female
- Age
- Height – In Centimeters
- Weight – In Kilograms
- Team – Name of the Team
- NOC – National Olympic Committee as abbreviation in 3 letters
- Season – Winter/Summer
- Games – Year and Season
- Year – Year
- City – Host City
- Sport – Category of Sport
- Event – Event respective to the Sport
- Medal – Gold/Silver/Bronze/NA

Using this data we perform data analysis and plot major patterns and visualize them using python programming language which consists of major data and mathematical analysis modules like Numpy, Pandas and Matplotlib.

Project Design



Requirement Specification

Hardware Requirements:

- Processor : Intel core i5-6500 3.20 GHz
- RAM: 4.00 GB
- Storage : 20 GB

Software Requirements:

- Operating System : Ubuntu 18.04 LTS, Windows 10
- Jupyter Notebook
- Visual Studio Code
- Language : Python 3.6
    - Modules:
        - Numpy
        - Pandas
        - Matplotlib

Proposed Analysis

Test Procedures:

- Fetching of the data from web
- Data pre-processing and Analysis
- Extraction of required data from the obtained data
- Finding and plotting various patterns and relations using extracted data
- Visualizing the obtained data

Test Environment:

- Programming Language : Python 3.6
- Web Open-Source : Jupyter Notebook 4.4.0
- Programming Tool : Visual Studio Code

Future Analysis

- Extracting further/more data that's related to current dataset
    - [GDP, Population and Budget]
- Integrating the new data to current dataset
- Finding relations between the current data and integrated data
- Analyzing the patterns affecting the obtained relations
- Visualizing the major patterns of current data and relations with respect integrated data

References

[1] Python for Data Analysis: Wrangling with Pandas, Numpy and IPython *by Wes Mckinney*

[2] Designing Data Visualizations *by Noab Illinky and Julie Steele*

[3] *https://www.kaggle.com/*

[4] *https://github.com/chrisalbon/notes/tree/master/content/python/*