# 

Olympic host cities

Charles Fung

[charles.fung.87@gmail.com](mailto:charles.fung.87@gmail.com)

Published: Nov, 2018

Table of Contents

[Abstract 3](#__RefHeading___Toc1308_3407726222)

[Introduction 3](#__RefHeading___Toc1293_3407726222)

[Methodology 4](#__RefHeading___Toc1310_3407726222)

[Data 5](#__RefHeading___Toc1295_3407726222)

[Past host cities 5](#__RefHeading___Toc1422_3407726222)

[City coordinates 6](#__RefHeading___Toc1620_3407726222)

[Venues in the cities 6](#__RefHeading___Toc1622_3407726222)

[Data Exploration on host cities 7](#__RefHeading___Toc1624_3407726222)

[KMeans clustering by venue categories 7](#__RefHeading___Toc1626_3407726222)

[Best cities in the world 9](#__RefHeading___Toc1424_3407726222)

[Observations: 11](#__RefHeading___Toc1628_3407726222)

[Model evaluation 13](#__RefHeading___Toc1303_3407726222)

[Logistic Regression 13](#__RefHeading___Toc1630_3407726222)

[Decision Tree 15](#__RefHeading___Toc1632_3407726222)

[Results 18](#__RefHeading___Toc1458_3407726222)

[Discussion 18](#__RefHeading___Toc1634_3407726222)

[Conclusion 19](#__RefHeading___Toc1316_3407726222)

[References 20](#__RefHeading___Toc1460_3407726222)

Olympic host city – who is next?

This is part of capstone project for Coursera course “Applied Data Science Capstone”

## Abstract

It costs a city several millions just to enter Olympic Games host candidature process, and 10+ billions to host the Games. This report used data on popular venues in a city trying to provide some indication if a city is competitive to be a host.

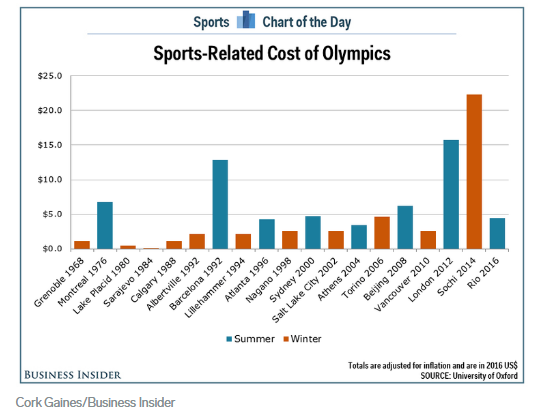
It used data publicly available from the web, and used machine learning technique to build a prediction model, and made a prediction on potential host cities.

## Introduction

As the International Olympic Committee (IOC) stated about the host city selection process:

Hosting the Olympic Games offers manifold benefits and opportunities to a Candidate City and the host region and country. Many years of careful and precise planning are required to host successful Olympic Games, with all of the relevant organizations, authorities and stakeholders working together as one united team, to ensure that the Games leave a positive, long term and sustainable legacy.

(<https://www.olympic.org/all-about-the-candidature-process>)

There are a lot of fiscal and financial commitments and it is a multi-year endeavour by a city to host the Olympic Games. Business Insider has charted the cost (in billions US$) of past Games (<https://www.businessinsider.com/summer-winter-olympics-sports-cost-2016-8>) with data compiled by University of Oxford.

As shown, many cities have host the games but not all of them were successful in making money. It is highly critical for a city to evaluate and decide if it should get involved into such an undertaking.

Hosting an Olympic Game involves many areas: sports facilities, transportation, communication, accommodation, etc. To be a successful and memorable event for the visitors, it also requires support facilities like restaurants and entertainment, etc.

In this report, I looked at the popular venues in different cities, and tried to see if the number and type of venues were related to the determination of a city becoming a host of Olympic Game.

Target audiences for this report would include:

1. IOC, NOC(National Olympic Committees) involved in the election process
2. City governments to evaluate their chance of being selected
3. Businesses and event sponsors that are interested in investing into upcoming Olympic Games host city

Winter Games require cooler weather conditions, so their host cities are more likely to be at higher latitudes. I would restrict this report on Summer Games.

## Methodology

I used only data publicly available on the web so there was no special survey of data collection required.

An Olympic Game is an important international tourist event to attract travelers. FourSquare allows its users to use this social media to document the popular venues that they frequent. Their input will represent a significant opinion on the characteristic venues of a city.

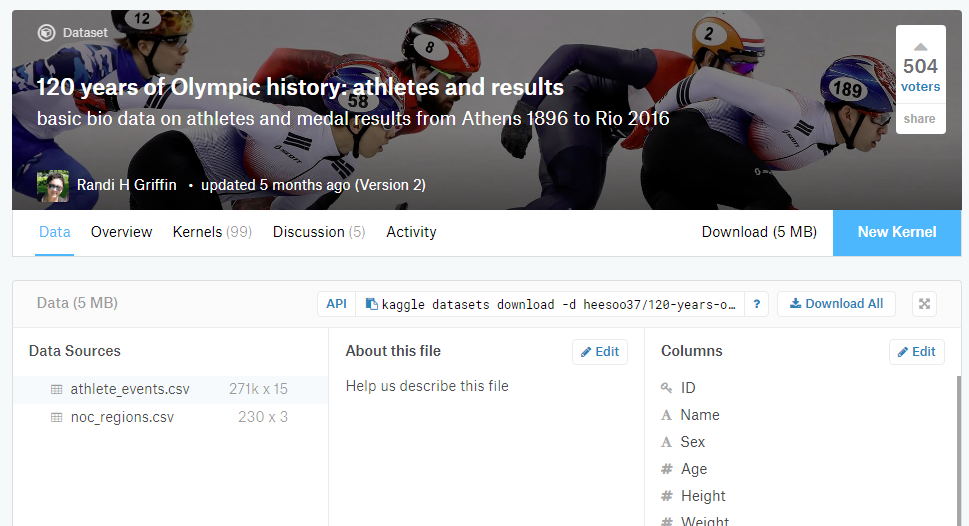
## Data

For this report, I compiled a list of the past host cities and a list of non-host cities for comparison. I gathered data on popular venues in these cities. FourSquare API is a social media that provides the venues with their categories, but the API requires the coordinates of the cities which I would get using geopy.

### Past host cities

There is a data set from Kaggle on past Olympic athletic events, which lists for each medal the athlete, year, city, etc.

("120 years of Olympic history: athletes and results" [https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results#athlete\_events.csv](https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results" \l "athlete_events.csv))

This list contains the medals for events in each Olympic Games, the city where the event was played. It has data for both Summer and Winter Games, so I extracted the unique city names for the Summer Games. Some cities have hosted the Games more than once. By extracting the unique names, this removed the duplicates.

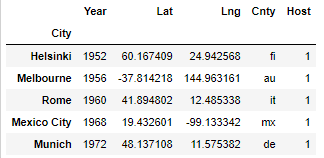
The spelling of some European city names followed their spelling in their native languages, and geopy provided more accurate data with their native spelling. (For example, there are multiple Athens around the world, but geopy identified Athina correctly.) After getting the coordinates from geopy, I modified the names to their English spelling so that they will match properly with the best cities list.

Similar to the data compiled by Oxford University, I used only the cities that host the Games after WWII. Stockholm was co-host for only one sport, and I chose to exclude it as well. 2020, 2024 and 2028 Games will be hosted in Tokyo, Paris and Los Angeles respectively. These cities have already host the Games before.

The past host cities are: Helsinki, Melbourne, Rome, Mexico City, Munich, Montreal, Moscow, Seoul, Barcelona, Atlanta, Sydney, Athens, Beijing, London, Rio de Janeiro, Tokyo, Paris, Los Angeles.

### City coordinates

I used geopy API to retrieve the coordinates for the cities:

This shows a snap shot of the coordinates.

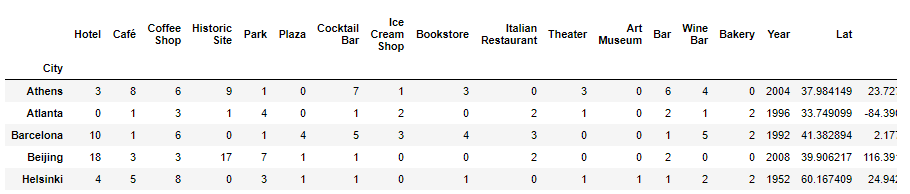
### Venues in the cities

I used FourSquare Explore API to retrieve 100 top-pick venues within 10km around the cities, and used the frequencies of venue categories as data points to evaluate the cities.

This is a sample of the venue data.

#### Data Exploration on host cities

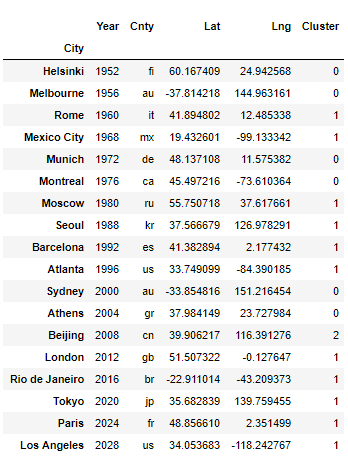
The above showed the counts of venue categories in the host cities.

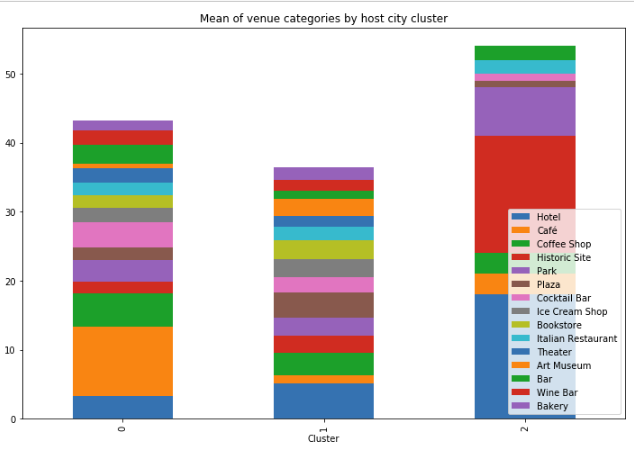


The above summed the number of venues by their categories in each city, and found the top 15 most common categories, which were: Hotel, Café, Coffee Shop, Historic Site, Park, Plaza, Cocktail Bar, Ice Cream Shop, Bookstore, Italian Restaurant, Theater, Art Museum, Bar, Wine Bar, Bakery.

#### KMeans clustering by venue categories

I wondered if the host cities were all similar in their venue counts, or if there was any grouping among them. I did a one-hot coding on the venue categories and ran them through KMeans classifier.

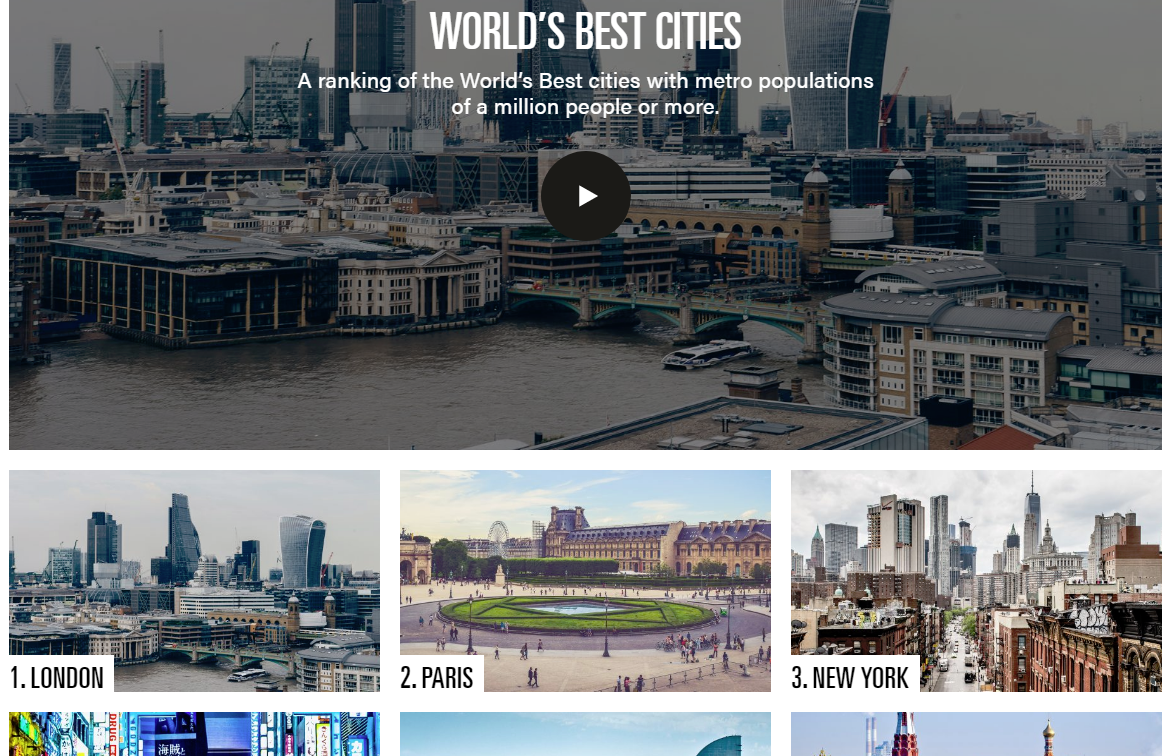
KMeans resulted in three clusters:

The above showed that cities in cluster 1 had a more even distribution of the venue categories and they were the more recent host cities. Cluster 0 represented the ‘older’ host cities. Cluster 2 had Beijing only. To allow for more accurate model evaluation, for the ongoing study, I limited to host cities in Cluster 1, where were: Rome, Mexico City, Moscow, Seoul, Barcelona, Atlanta, London, Rio de Janeiro, Tokyo, Paris, Los Angeles

There were over 250 venue categories for the cities, so the number of features far exceeded that of data points. I picked only the 15 most frequent categories for further analysis.

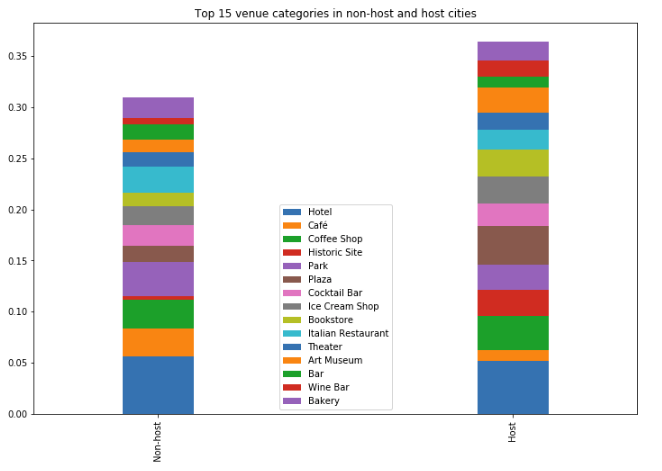
### Best cities in the world

There are many lists of best cities on the web. I picked "World's best cities" which listed 100 cities (<https://www.bestcities.org/rankings/worlds-best-cities/>).

I did web scraping to extract the city names. Since there were only a small number of past host cities, I used a matching number of best cities to maintain a balance of positive (past host) and negative samples (non-past-host) in the data set.

The non-host cities were: Amsterdam, Boston, Chicago, Dubai, Hong Kong, Madrid, New York, San Francisco, San Jose, Singapore, Toronto, Washington, Zurich

I followed similar steps as for the host cities: get city coordinates and the popular venues with their categories.

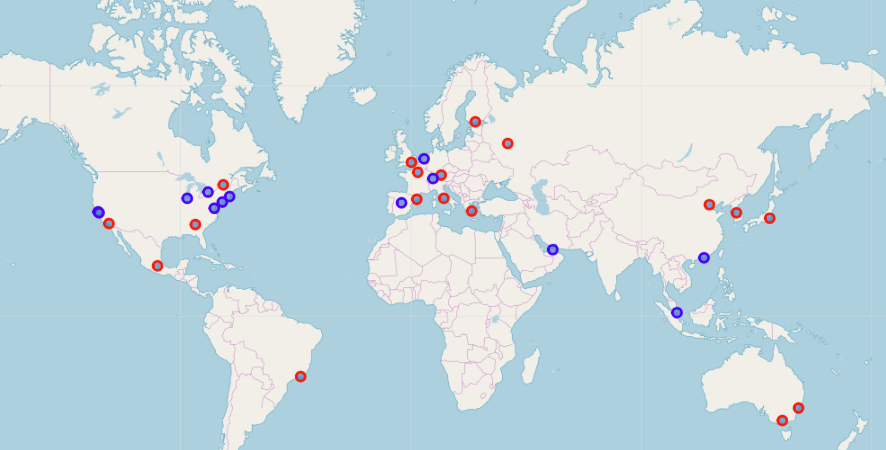
The above showed the common venue categories between non-host and host cities (cluster 1 only).

#### Observations:

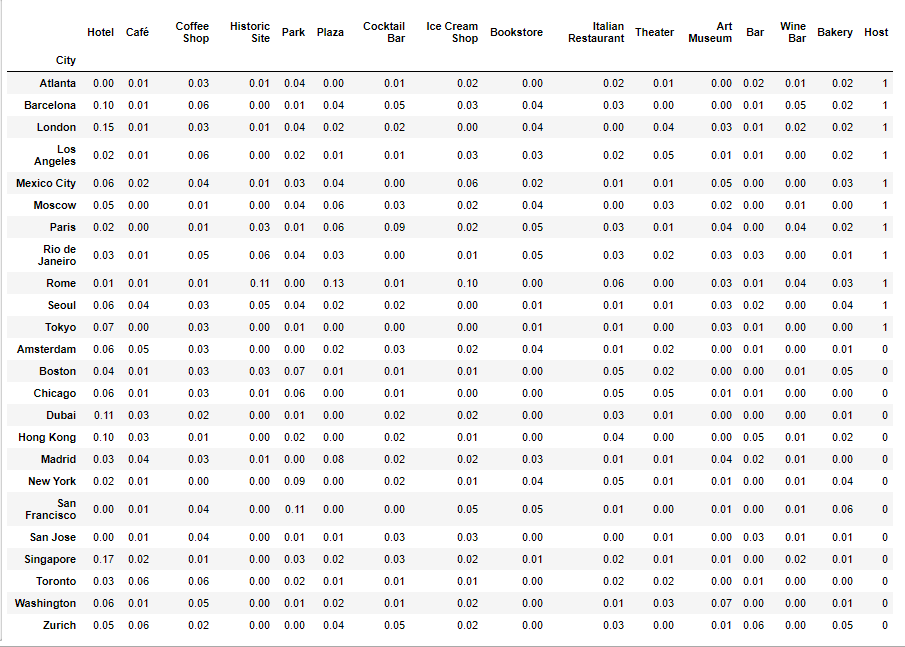
These big cities have a similar amount of hotels.

In host cities, the top 15 categories added up to over 35%, and the amount of each category was evenly distributed.

In non-host cities, the top 15 categories were not as evenly distributed. They were also low in Historic Sites.

The map showed the city locations. The host cities are in red, and the non-host cities are in blue.

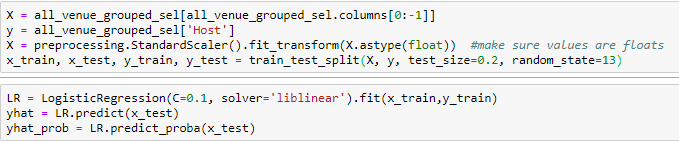
This is the data from the selected cities, with the Host label attached to them (1 for host cities).

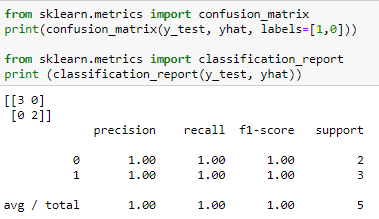


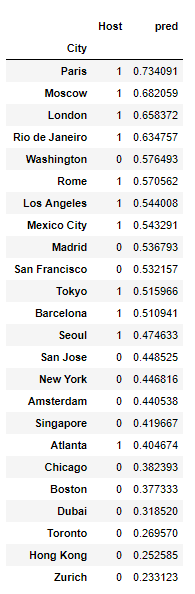
## Model evaluation

### Logistic Regression

The cities were known/labeled as host or non-host, so I used Logistic Regression as the supervised learning model because it provided the prediction probabilities.





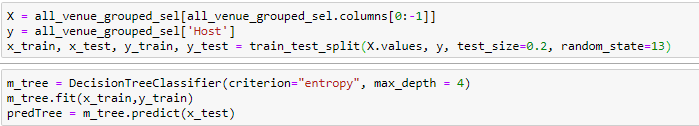
The table on the right showed the predictions made by the model.

Washington, Madrid and San Francisco appeared to be potential candidates for the future Games.

The prediction seemed to be better with the more recent host cities.

### Decision Tree

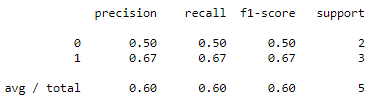
For comparison, I did a Decision Tree on the same data set.

I used the data set (venues only, without coordinates) to train a Decision Tree with maximum depth set to 4. Setting a higher depth caused the model to over-fit the training data and predicted poorly on test data.

Train set Accuracy: 1.0

Test set Accuracy: 0.6

The classification report on comparing test data with prediction is:

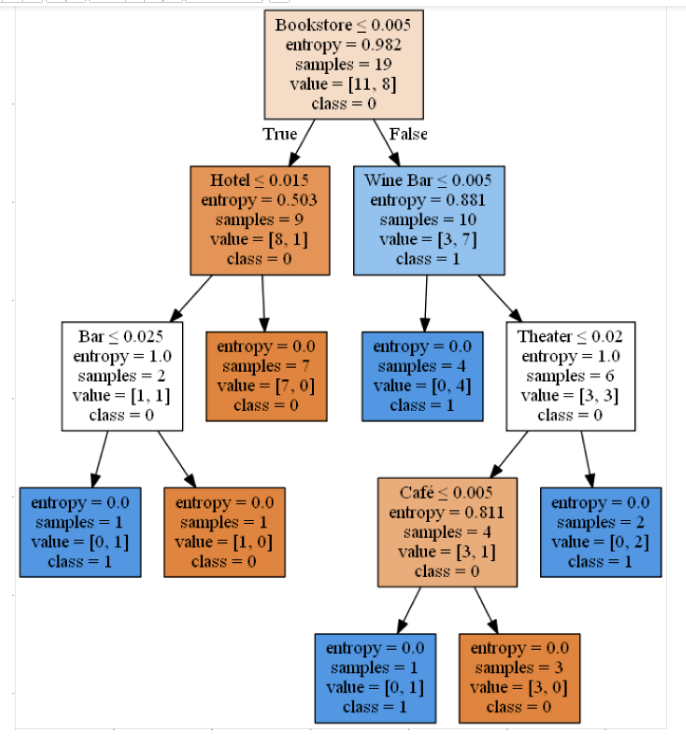


with F1-score = 0.6.

Decision Tree predicted the probability of each host city except London to be a host city.

In this model. Amsterdam was the most promising candidate.

This shows the Decision Tree model.



## Results

Washington, Madrid and San Francisco showed up as having prediction over 0.5 in Logistic Regression.

Amsterdam was the only non-host city that was suggested by Decision Tree. Therefore, the two models showed conflicting results.

Conventionally, Olympic host cities alternate between regions. Since Los Angeles will host the 2028 Games, it is less likely another US city will be selected for 2032. No city in the Middle East has ever host the Games, so there is a chance that Madrid or Amsterdam may clinch it. Or, as recent history shows, a past host city may be re-selected.

## Discussion

It appeared that the peripheral venues were more important than the sports stadium when it came to Olympic Games!

In the modern world, it is more important that a city is receptive and attractive to tourists for it to be selected to be a host city.

With several runs using different test size and random state when splitting the data set, I found that the test results can vary.

Possible reason were:

* There were not enough data points, so whether any outlier got into the training data set had significant impact. It was difficult to find the outliers with the large number of features.
* The number of venue categories far exceeded the number of data points. I had to limit the data to a small number of features for the models to work properly.
* FourSquare Explore API returned live data that varied with the time of the day.

The other factors that influence the predictions were:

* This venue data for the current time had a lesser correlation with the host city selected decades ago.
* Host city selection had many other factors involved: financial, political, geographical, etc.
* FourSquare users tended to input food related venues. This was typical for social-media users who like to share their food experience with others.

Despite the above factors, the predictions still provide a quick means of comparing among large cities.

## Conclusion

Predictions of future Olympic Games using venues provided a quick means for evaluation and comparison amount modern large cities.

## References

<https://eureka.sbs.ox.ac.uk/6195/1/2016-20.pdf> The Oxford Olympics Study 2016: Cost and Cost Overrun at the Games

<https://www.olympic.org/all-about-the-candidature-process>

<https://www.businessinsider.com/summer-winter-olympics-sports-cost-2016-8>

[https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results#athlete\_events.csv](https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results" \l "athlete_events.csv)

<https://www.bestcities.org/rankings/worlds-best-cities/>

[https://geopy.readthedocs.io/en/stable/#module-geopy.geocoders](https://geopy.readthedocs.io/en/stable/" \l "module-geopy.geocoders) – geopy geocoder

<https://developer.foursquare.com/> - FourSquare developer

<https://www.coursera.org/learn/applied-data-science-capstone/home/info> – Coursera “Applied Data Science Capstone”