Revised 2020 FIA Formula One Calendar

1. Introduction

1.1. Background

The FIA Formula One World Championship is the highest class of single-seater racing. The 2020 season was intended to run from March to December and comprise of 22 races, in 22 different countries. However, the COVID-19 crisis caused travel and mass gatherings to be severely restricted and, consequently, the first 10 races of the season to be abandoned. With the sport's commercial rights' holders and, more importantly, the teams themselves getting most of their income from broadcast rights, race fees and sponsorship details, this presented the sport with a series problem.

The governing body, the FIA, and teams believe the best way to recover their income – keep their companies in business and employees employed – would be to run an 18-race season over a condensed period of 6 months.

1.2. Business Problem

Condensing the season introduces practical and logistical problems for the teams, who must ship equipment around the world, and team members, who must do so much travelling in such a short space of time. There are two questions that ned answering:

1. Which 4 of the 22 races should be cut from the calendar?
2. How should the season be arranged to minimise the strain on the teams and travelling team members?

The answers to these questions would be of interest to the FIA and the teams who are currently trying to put a new race calendar together

2. Data

2.1. Data Required

To answers the two questions posed in 1.2, we must first decide which 4 races to remove from the calendar. Data pertaining to the relative importance of each race to the sport would offer an impartial metric in which to choose the races to leave out. For this, a metric of interest in each race will be sort such that the races that fans are least interested in can be cut from the calendar. This will likely keep the sponsors happy as their primary reason for their involvement in the sport is advertising. Keeping the races that have the most fans will mean the best advertising.

Once the races have been decided upon, the order in which they are to run will needs deciding such the pressure on logistics and travelling team members is minimised. For this, clusters of races will be proposed that minimise travel between them and allow a break for team members between clusters.

2.2. Data Sources

The current 2020 calendar will be sourced from scraping data from Wikipedia and parsing with BeautifulSoup. A measure of interest in the races will be found by leveraging FourSquare data. The number of photos and 'likes' at each race circuit will be used as a metric for interest in that particular race. This information will then be used to remove the 4 races with the least interest.

The final list of 18 races will then be clustered by region to provide the proposed final calendar.

3. Methodology

3.1. Global Race Distribution

To first understand what races were on the original calendar and how they are distributed around the world, the 2020 race calendar was sort parsed from Wikipedia tables using beautiful soup

Schedule of events

Round	Grand Prix	Circuit	Race date
1	Austrian Grand Prix	Red Bull Ring, Spielberg	5 July
2	British Grand Prix	Silverstone Circuit, Silverstone	19 July
3	Hungarian Grand Prix	Hungaroring, Mogyoród	2 August
4	Belgian Grand Prix	Circuit de Spa-Francorchamps, Stavelot	30 August[b]
5	Italian Grand Prix	Autodromo Nazionale di Monza, Monza	6 September
6	Singapore Grand Prix	Marina Bay Street Circuit, Singapore	20 September
7	Russian Grand Prix	Sochi Autodrom, Sochi	27 September
8	Japanese Grand Prix	Suzuka International Racing Course, Suzuka	11 October
9	United States Grand Prix	Circuit of the Americas, Austin, Texas	25 October
10	Mexico City Grand Prix	Autódromo Hermanos Rodríguez, Mexico City	1 November
11	Brazilian Grand Prix	Autódromo José Carlos Pace, São Paulo	15 November
12	Abu Dhabi Grand Prix	Yas Marina Circuit, Abu Dhabi	29 November

Source:[27]

Grand Prix	Circuit	Original date	Status	New date
Australian Grand Prix	Albert Park Circuit, Melbourne	15 March	Cancelled	TBA[c]
Bahrain Grand Prix	Bahrain International Circuit, Sakhir	22 March	Postponed	TBA
Vietnamese Grand Prix	Hanoi Street Circuit, Hanoi	5 April	Postponed	TBA
Chinese Grand Prix	Shanghai International Circuit, Shanghai	19 April	Postponed	TBA
Dutch Grand Prix	Circuit Zandvoort, Zandvoort	3 May	Postponed	TBA
Spanish Grand Prix	Circuit de Barcelona-Catalunya, Montmeló	10 May	Postponed	TBA
Monaco Grand Prix	Circuit de Monaco, Monte Carlo	24 May	Cancelled	N/A
Azerbaijan Grand Prix	Baku City Circuit, Baku	7 June	Postponed	TBA
Canadian Grand Prix	Circuit Gilles Villeneuve, Montréal	14 June	Postponed	TBA
French Grand Prix	Circuit Paul Ricard, Le Castellet	28 June	Cancelled	N/A

Sources:[28][29][30][31][32][33]

Source: https://en.wikipedia.org/wiki/2020_Formula_One_World_Championship

Find the original 2020 F1 Calender

```
# scrape calender information from Wikipedia
url2020 = 'https://en.wikipedia.org/wiki/2020_Formula_One_World_Championship'
html = urllib.request.urlopen(url2020).read()

# parse using BeautifulSoup
phtml = bs(html,'html.parser')

# Initialise features
races = []
circuits = []
cities = []
countries = []
schedule = []
```

Parse table of races already postponed in 2020 - assign a 'schedule' value of 0

```
table_postponed = phtml.find_all('table')[3]
rows = table_postponed.find_all('tr')

for tr in rows:
    td = tr.find_all('td')
    row = [i.text for i in td]
    if len(row) == 5:
        temp = row[1].strip().split(',')
        races.append(row[0].strip())
        circuits.append(temp[0])
        cities.append(temp[1])
        schedule.append(0)
```

The Geolocator package was used to find Longitude and Latitude coordinates a for all race tracks. This information was put in a dataframe (df), and each race was given a binary coded depending on whether it was scheduled or postponed by COVID-19

Country	Grand Prix	Circuit	Location	Schedule	Latitude	Longitude
Australia	Australian Grand Prix	Albert Park Circuit	Melbourne	0	-37.841353	144.963743
Vietnam	Vietnamese Grand Prix	Hanoi Street Circuit	Hanoi	0	21.017009	105.764044
China	Chinese Grand Prix	Shanghai International Circuit	Shanghai	0	31.339979	121.219598
Netherlands	Dutch Grand Prix	Circuit Zandvoort	Zandvoort	0	52.382273	4.533707
Monaco	Monaco Grand Prix	Circuit de Monaco	Monte Carlo	0	48.264099	4.616128

The distribution of the races around the world was then investigated by two methods:

- Data Binning
- K-means Clustering

The data was split into two bins using the Latitude coordinate data, centred on the equator, to ascertain how the races are distributed between the Northern and Southern Hemispheres. Separately, the data was interrogated using the k-means clustering method to help decide how the clusters of races asked for by the FIA could be arranged.

3.2. Race Attendance

Obtaining race attendance figures was quite difficult – for some circuits, impossible. Google was used to find what figures were publicly available and this information was added to the df. Those races where attendance figures could not be found were assigned an NaN value.

Which races had the highest attendance in 2019?

```
total_att = [324100,97000,160428,85000,307000,160000,203000,351000,
             230000,251864,200000,202146,268000,345694]
countries_att = ['Australia', 'Bahrain', 'Spain', 'Azerbaijan', 'Canada', 'France', 'Austria',
 'United Kingdom', 'Hungary', 'Belgium', 'Italy', 'Japan', 'United States', 'Mexico']

total_att_append = []
```

```
for country in df['Country']:
    count = 0
    for att in countries_att:
        if att == country:
            C = country
            T = total_att[count]
            count = count + 1
            break
        else:
            C = country
            T = np.nan
            count = count + 1
    total_att_append.append(T)
```

```
df['Attendance'] = list(map(float,total_att_append))
```

3.3. Fan Engagement

Since race attendance figures are not all publicly available, they could not be used reliably as a means of deciding if a race was popular with fans or not. Therefore, an additional measure was sort, Fan Engagement. This was obtained by making used of FourSquare's API to leverage data for each Grand Prix Circuit.

A list of URLs was created to located each circuit by its Longitude and Latitude coordinates:

```
search_query = 'Formula 1'
radius = 15000
urls = [];
count = 0
for circuit in circuits:
    urls.append('https://api.foursquare.com
    count = count + 1
```

```
}&radius={}&limit={}'.format(CLIENT_ID, CLIENT_SECRET, df.iloc[count,5], df.iloc[count,6], VERSION, search_query, radius, LI
```

If FourSquare returned a category called 'Racetrack,' then the corresponding venue ID was interrogated further for the number of photos and 'likes.' The sum of these was termed 'Engagement.'

3.4. Correlation

A regression analysis was performed on the Engagement and attendance figures to understand if there was a relationship between the two. This would help decide the races of least importance. This analysis was only performed on the available data and outliers (Mexico, which had extremely high engagement) were removed.

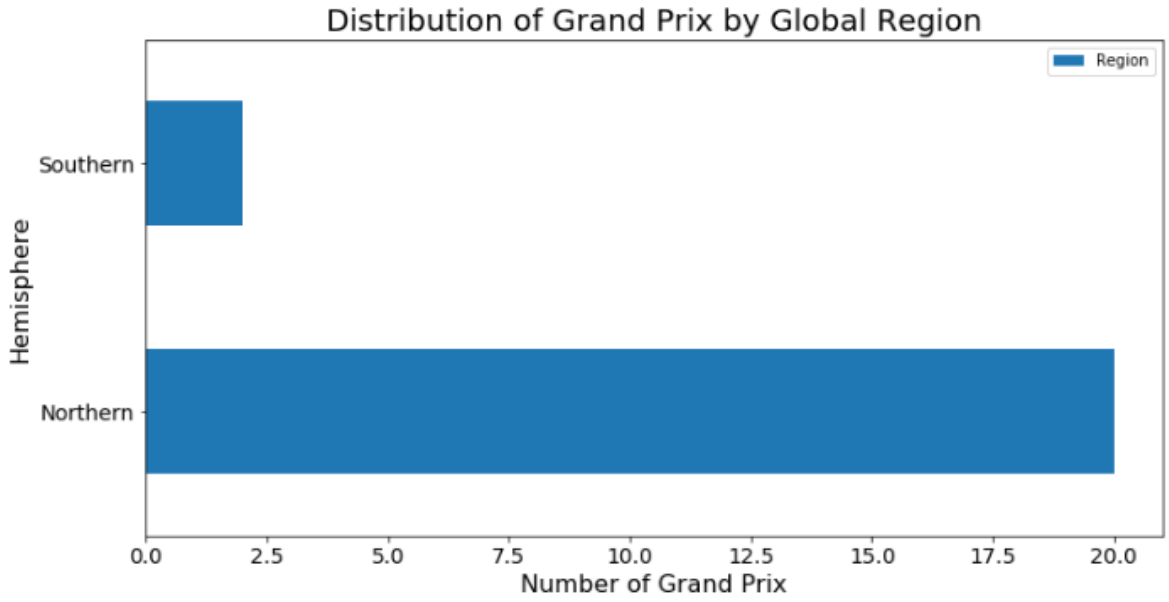
4. Results

4.1. Global Race Distribution

The Folium plot below shows how the races are distributed around the world and identifies those which have already been postponed. The race events were also clustered using the k-means method. This produced 4 distinct clusters.

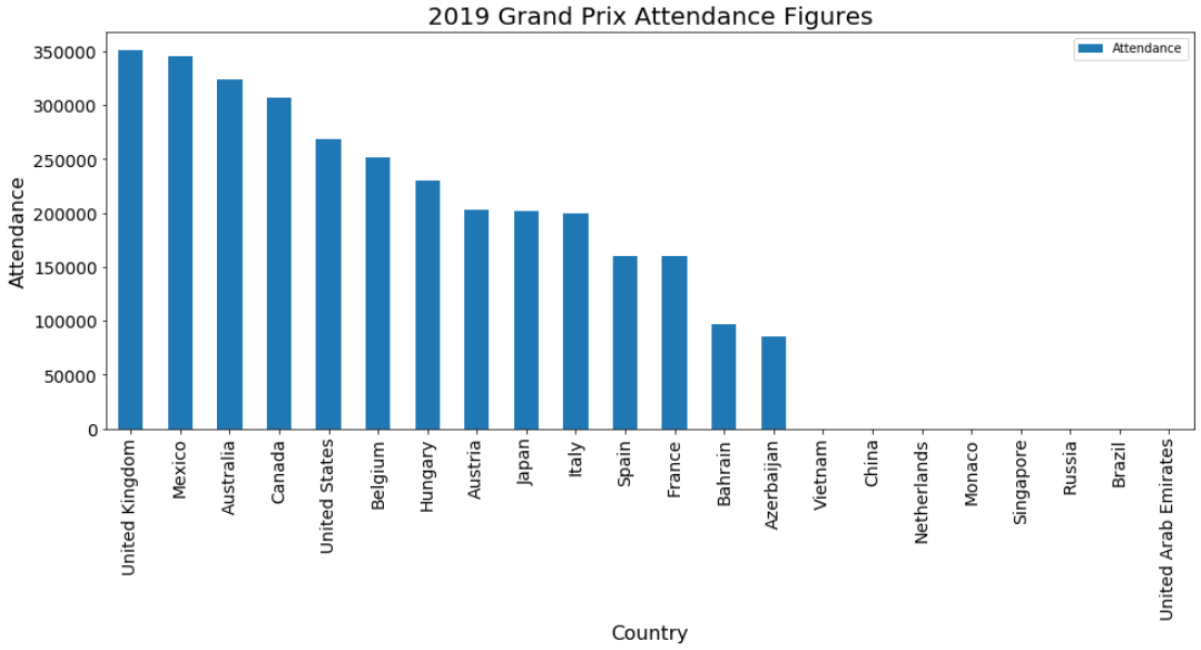
Folium plot for Formula One race calender showing postponsed and scheduled races

To see how the races are distributed by hemisphere they were put into 2 bins based on their latitude coordinate. It can be seen from the figure below that the races are strongly biased towards the Northern Hemisphere, with only 2 Southern Hemisphere races.



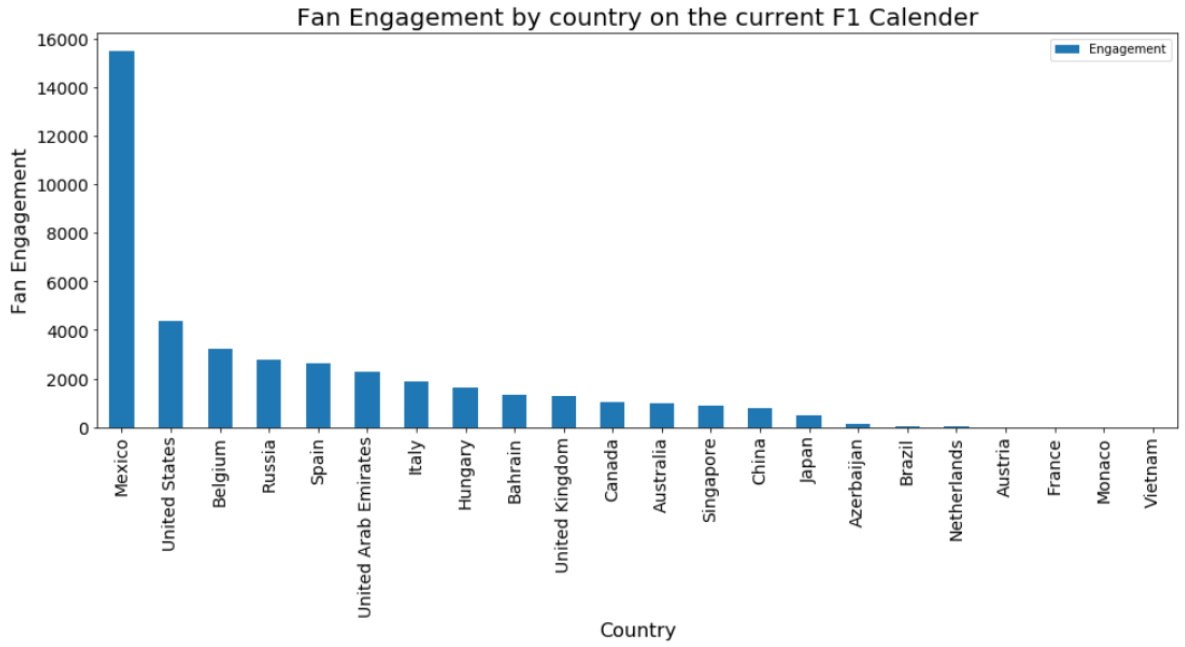
4.2. Race Attendance

The figure below shows the attendance of 14 races in 2019. Data was not available for 8 races (it is worth pointing out that 2 of those were not I the 2019 calendar). The British Grand Prix had the highest attendance, while the Bahrain and Azerbaijan Grand Prix had low attendance figures.



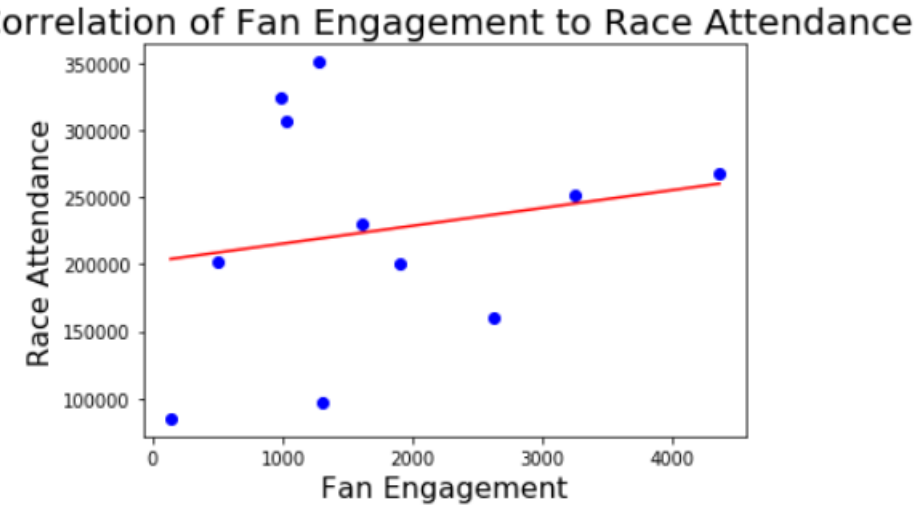
4.3. Fan Engagement

Fan engagement data was found based on the likes and photos posted form each racetrack on FourSquare. This data reveals that the Mexican racetrack is by far the most engaging, with the Circuit of the Americas next. Data could not be found for 4 racetracks (Austria, Vietnam, Monaco, France) and 3 (Netherlands, Brazil, Azerbaijan) has very low engagement



4.4. Correlation

A linear regression was performance to ascertain whether there was a correlation between the attendance at the Grand Prix and the Fan Engagement with the circuits. This analysis was only performed on the available data and outliers (Mexico, which had extremely high engagement) were removed



There is a weak positive correlation, suggesting the race attendance is some what related to how much engagement there is from fans with the circuit.

5. Discussion

5.1. Which races should be dropped?

We must drop 4 races form the 22. These must be clustered in order to facilitate easy travel between races.

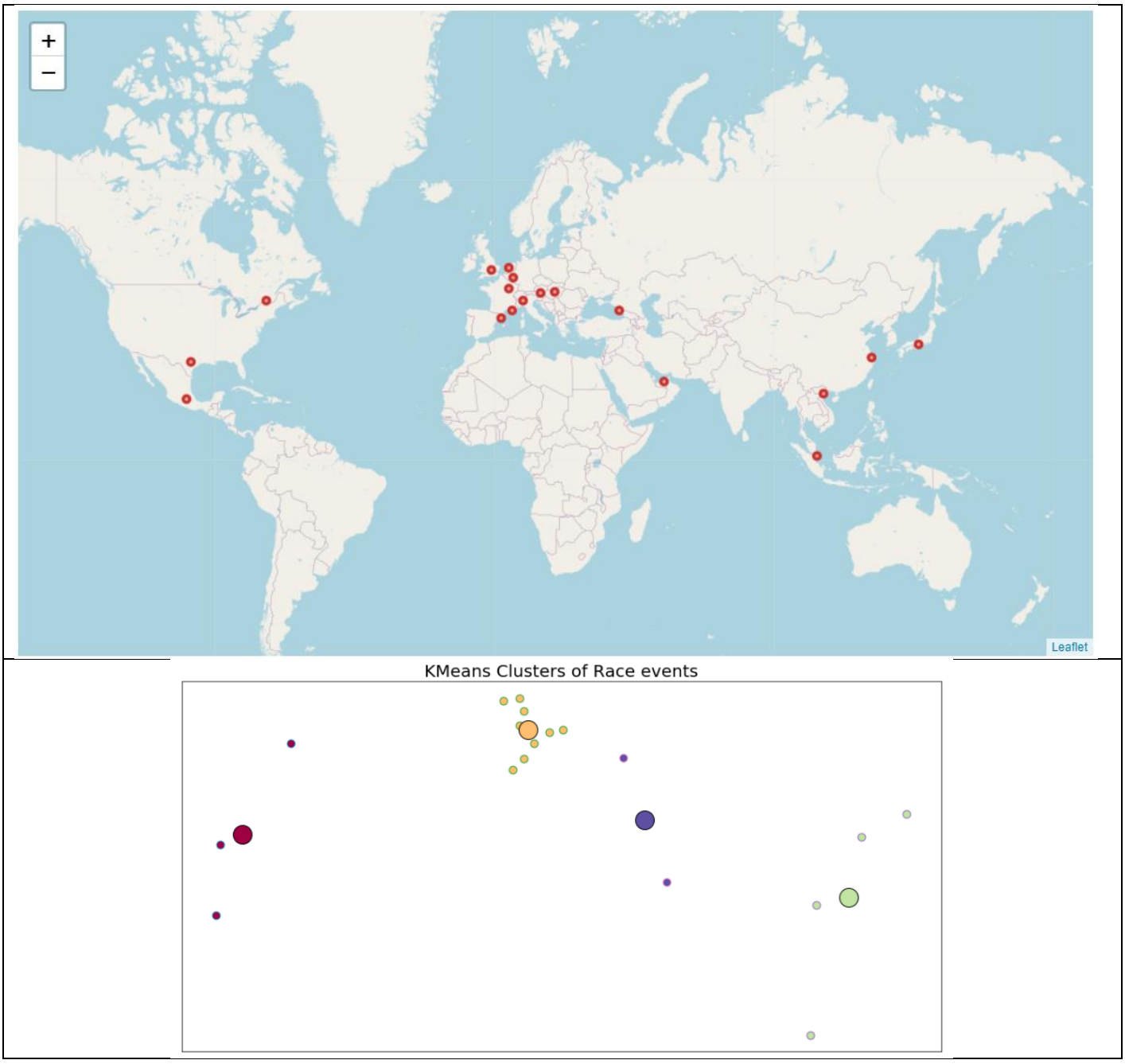
Recalling the global distribution, the races were overwhelmingly biased to the Northern Hemisphere, with only 2 Southern Hemisphere races. region had only 1 race. These geographical outliers do not make sense if we are to reduce travelling and it is therefore recommended that **Brazil and Australia** are removed from the new calendar.

From reviewing the attendance and engagement figure we recall that **Azerbaijan** scored poorly on both and should therefore also be left off the calendar. Finally, was the second worst attended Grand Prix was **Bahrain**. This also was also only ninth in the list of Fan Engagement and should be the final race removed from the calendar.

5.2. How should the races be clustered?

With those 4 races (Australia, Bahrain, Azerbaijan, Brazil) removed from the calendar, the 18 races season is distributed entirely in the the Northern Hemisphere as shown on the Folium plot below. The k-means clustering algorithm suggests a 4-cluster season as shown.

The season would start with an intense 9-race cluster in Europe, close to the teams' factories. Following a short break, a 2-race cluster in the Middle East would be followed by a 4-race cluster in the Far East. The season would end with a race cluster in the Americas.



6. Conclusion

The truncated Formula One season has been reduced from 22 races to 18 by investigating the distribution of races around the globe and their respective fan engagement. The former was found be extracting race location information from Wikipedia and using this with geopy to obtain geolocation data. Data binning and K-Means Clustering was used to analyse the data. It was found that most of the races (20) were in the northern hemisphere, while only 2 were in the south. To help with the travelling burden, the 2 Southern Hemisphere races were removed form the calendar.

The remaining two races removed (Bahrain and Azerbaijan) from the calendar were chosen based on a Fan Engagement metric. This was found from Foursquare data at each individual circuit. The number of photos and 'likes' where summed together to determine the measure