EXPLORING THE TOP 5 DEADLIEST LOCATIONS OF ACCIDENTS / INCIDENTS IN AVIATION SINCE 1923

Bienvenu CHOUPO
June 21, 2019

1. INTRODUCTION

Air transport is in a substancial growth nowadays especially in developping countries (Africa). People need to do a lot of things in a small amount of time. The age of speed indeed !
Plane crash is scarce, but once it happens, it can be very serious. Over the years, there have been some serious cases of aircraft incidents / accidents that resulted in loss of many lives. Those accidents happened in various locations around the globe and involved many airlines companies and different types of aircrafts. More often, to keep the souvenir of the victims, memorials for crash victims are built. Those memorial attract many people, especially tourists. For a place to attract tourists, there needs to be some common facilities: hotel, restaurant, museum, historical site...

Let's say a tourist would like to visit one of the top five locations where air crash killed the most people, but don't know which one to choose and ask for suggestions.

In this notebook, we will explore the data of accidents from 1923 to 2019. Analysing the data, we will retrieve the top 5 locations where most people were killed. We will then use the Foursquare API to explore those locations, cluster the neighborhoods of those locations and make suggestions to tourist.

2. DATA

The data have been extracted from the wiki page:
https://en.wikipedia.org/wiki/List_of_aircraft_accidents_and_incidents_resulting_in_at_least_50_fatalities using the BeautifulSoup package from bs4 library in python 3

The data cover a period from 1923-12-21 to 2019-03-10.

Inclusion criteria :

Criteria for inclusion require at least 50 fatalities in a single occurrence involving commercial passenger and cargo flights, military passenger and cargo flights, or general aviation flights that have been involved in a ground or mid-air collision with either a commercial or military passenger or cargo flight.

On the website page, only the names of locations are given. In order to get information on the geographical coordinates of various locations, I will use the geocoder package. It will not be easy due to the fact that many locations' names are not written so as to get the geocoder retrieve their coordinates. So I need to refine the names.
After having all the information, I will save the data in a csv file (Air_Accident.csv) so that the data can be accessible easily.

Foursquare API location will be used for exploring the neighborhood of the top 5 locations were most people were killed in air crash.

2.1. Extraction of Tables from the website¶

We first import all libraries, ping the website and scrape tables from the website using BeautifulSoup package
On the website there are multiple tables, six to be exact. The first two tables are key tables: Key_death and key_location
Table number 3 is the one that interests us.

Here are the tables:

	Abbreviation	Definition
1	C	Crew
2	P	Passenger
3	G	Ground
4	N	Notes
5	†	No survivors
6	1*	Sole survivor
7	COM	Commercial (accident/incident)
8	MIL	Military (accident/incident)
9	INB	Bombing
10	INH	Hijacking
11	EXG	Attacked using ground-based weapons
12	EXS	Attacked by other aircraft

Table 1: Key_death

	Abbreviation	Definition
0	(none)	< 20 km (12.5 mi)
1	"off"	< 20 km (12.5 mi) (water impact)
2	"near"	20 km (12.5 mi) to 50 km (31 mi)
3	"area of"	> 50 km (31 mi)
4	STD	Standing
5	TXI	Taxi
6	TOF	Take off
7	ICL	Initial climb
8	ENR	En route
9	MNV	Maneuvering
10	APR	Approach
11	LDG	Landing
12	UNK	Unknown
13	***	Active or decommissioned military bases; close...

Table 2: Key_location

And our main table:
The shape of our dataframe is (548, 5), that is 548 rows and 15 columns

The first five rows like this:

	Type	Incident	Aircraft	Location	Phase	Airport	Distance	Date	Total	Crew	Passenger	Ground	Notes
0	INH	American Airlines Flight 11	Boeing 767-223ER	usnewyneNew York City, New York, U.S.	ENR[11]			2001-09-11	est. 1,700	11	81	est. 1,600 [nb 2]	†
1	INH	United Airlines Flight 175	Boeing 767-222	usnewyneNew York City, New York, U.S.	ENR[12]			2001-09-11	est. 1,000	9	56	est. 900[nb 2]	†
2	COM	Pan Am Flight 1736 andKLM Flight 4805	Boeing 747-121 and Boeing 747-206B	spctTenerife, Spain	TXI/TOF[10][16][17]	TFN		1977-03-27	583	23	560	0	
3	COM	Japan Airlines Flight 123	Boeing 747SR-46	juUeno, Japan	ENR[18][19]			1985-08-12	520	15	505	0	
4	COM	Saudi Arabian Flight 763 andKazakhstan Airline...	Boeing 747-168B and Ilyushin Il-76TD	indicCharkhi Dadri, India	ENR[20][21]			1996-11-12	349	33	316	0	†

Table 3: head of main dataframe exctracted from the website

2.2. Let's search for the location coordinates

In order to get information on the geographical coordinates of various locations, we will use the geocoder package. It will not be easy due to the fact that many locations' names are not written so as to get the geocoder retrieve their coordinates. So we need to refine the names.
This is the process we'll be going through:
First , we create a function to refine the names of the locations, we create a function to retrieve the geographic coordinates. We call the function on our dataframe. There are some coordinates found and other missing. We split our dataframe into two: one made of the correct coordinates and the other made of locations with the missing values.
We repeat the process again and again until almost all coordinates are found. For missing values, we deal with them manually.

2.3. Putting all together

After the above operations, we merge all the splitted dataframes and built our main dataset with coordinates. Then we store the data to a CSV file in order to access them easily.

3. DATA ANALYSIS

We can import the data directly from our csv file. Before going into analysis, we need to do prepare the data.

3.1. Pre-processing the dataframe

Columns 'Airport', 'Distance' and 'Notes' will not be usefull for us.
So are columns 'Crew', 'Passenger' and 'Ground' that are already resumed in the column 'Total'. So let's drop them. Other operations need to be handled on the dataframe like removing some elements in columns, changing the type of certains columns…
All this done, here is our dataframe (the first 5 rows):

	Type	Incident	Aircraft	Location	Phase	Date	Total	Latitude	Longitude
0	INH	American Airlines Flight 11	Boeing 767-223ER	New York City, New York, U.S.	ENR	2001-09-11	1700	40.712728	-74.006015
1	INH	United Airlines Flight 175	Boeing 767-222	New York City, New York, U.S.	ENR	2001-09-11	1000	40.712728	-74.006015
2	COM	Pan Am Flight 1736 andKLM Flight 4805	Boeing 747-121 and Boeing 747-206B	Tenerife, Spain	TXI/TOF	1977-03-27	583	28.293578	-16.621447
3	COM	Japan Airlines Flight 123	Boeing 747SR-46	Ueno, Japan	ENR	1985-08-12	520	35.711788	139.776096
4	COM	Saudi Arabian Flight 763 andKazakhstan Airline...	Boeing 747-168B and Ilyushin Il-76TD	Charkhi Dadri, India	ENR	1996-11-12	349	28.605554	76.147567

Table 4: Our final dataframe

3.2. Descriptive statistics

The incidents/accidents listed in our dataframe cover a period from 1923-12-21 to 2019-03-10, that is 96 years! A brief review of the descriptive statistics of aircraft accidents and incidents since 1923 suggests the following:

3.2.1. Number of Victims

During that period, there have been about 57646 people killed in air accidents/incidents.

By type:

	Incident	Aircraft	Location	Phase	Date	Total	Latitude	Longitude
Type								
COM	439	439	439	439	439	439	439	439
EXG	12	12	12	12	12	12	12	12
EXS	4	4	4	4	4	4	4	4
INB	15	15	15	15	15	15	15	15
INH	10	10	10	10	10	10	10	10
MIL	68	68	68	68	68	68	68	68

Table 5: Number of accidents by type

The two main categories of occurrences were accidents/incidents related (COM + MIL: 508. that is 92.5%) and attacks on aircraft (INH+INB+EXG+EXS: 41 which represents 7.5%).

- Sub-groupings of the first category include commercial (COM, 439; 80.1%) and military (MIL, 68; 12.4%).
- Sub-groupings of the second category include internal attacks with a bomb (INB, 15; 36.6%), internal attacks with hijacking (INH, 10; 24.4%), external attacks from the ground (EXG, 12; 29.3%), and external attacks from the sky (EXS, 4; 9.8%).

By phase of flight:

	Type	Incident	Aircraft	Location	Date	Total	Latitude	Longitude
Phase								
	2	2	2	2	2	2	2	2
APR	184	184	184	184	184	184	184	184
APR/ENR	2	2	2	2	2	2	2	2
APR/TOF	1	1	1	1	1	1	1	1
ENR	249	249	249	249	249	249	249	249
ENR/LDG	1	1	1	1	1	1	1	1
ICL	51	51	51	51	51	51	51	51
LDG	21	21	21	21	21	21	21	21
LDG/STD	1	1	1	1	1	1	1	1
MNV	3	3	3	3	3	3	3	3
STD	1	1	1	1	1	1	1	1
TOF	21	21	21	21	21	21	21	21
TOF/TXI	2	2	2	2	2	2	2	2
TXI/TOF	1	1	1	1	1	1	1	1
UNK	8	8	8	8	8	8	8	8

Table 6: Number of accidents by phase of flight

- The highest number of occurrences took place while en route (ENR + ENR/APR + ENR/LDG 252; 45.98%)
- 186 (34%) accidents happened during the Approach
- 51 (9,3%) accidents happened during the Initial climb
- Almost 24 (4,4%) took place during the take-off
- 22 (4%) accidents occured during the landing

3.2.2. Which Aircrafts are involved in the deadliest accidents?

The top 5 Aircrafts involved in the deadliest accidents are:

- Boeing 767-223ER with 1700 killed
- Tupolev Tu-154M with 1218 killed
- Boeing 767-222 with 1000 killed
- McDonnell Douglas DC-9-32 with 836 killed
- Ilyushin Il-18V with 833 killed

3.2.3. Which Aircrafts are involved the most in accidents?

	Type	Incident	Aircraft	Location	Phase	Date	Total	Latitude	Longitude
Aircraft									
Tupolev Tu-154M	3	10	1	10	3	10	10	10	10
Ilyushin Il-18V	1	9	1	4	3	9	8	4	4
Douglas DC-4	1	9	1	9	2	9	8	9	9
McDonnell Douglas DC-9-32	1	8	1	8	4	8	7	8	8
Douglas DC-6B	2	8	1	8	3	8	6	8	8

Table 7: Number of accidents by Aircraft

The top 5 aircrafts most involved in accidents are:

- Tupolev Tu-154M : 10 times
- Ilyushin Il-18V : 9 times
- Douglas DC-4 : 9 times
- McDonnell Douglas DC-9-32 : 8 times
- Douglas DC-6B : 8 times

3.2.4. What are the top 5 deadliest flights?

American Airlines Flight 11	1700 deaths
United Airlines Flight 175	1000 deaths
Pan Am Flight 1736 and KLM Flight 4805	583 deaths
Japan Airlines Flight 123	520 deaths
Saudi Arabian Flight 763 and Kazakhstan Airlines Flight 1907	349 deaths

3.2.5. What are the top 5 darkest days in aviation?

The top5 darkest days are:

- 2001-09-11 2889 deaths
- 1977-03-27 583 deaths
- 1985-08-12 520 deaths
- 1996-11-12 349 deaths
- 1974-03-03 346 deaths

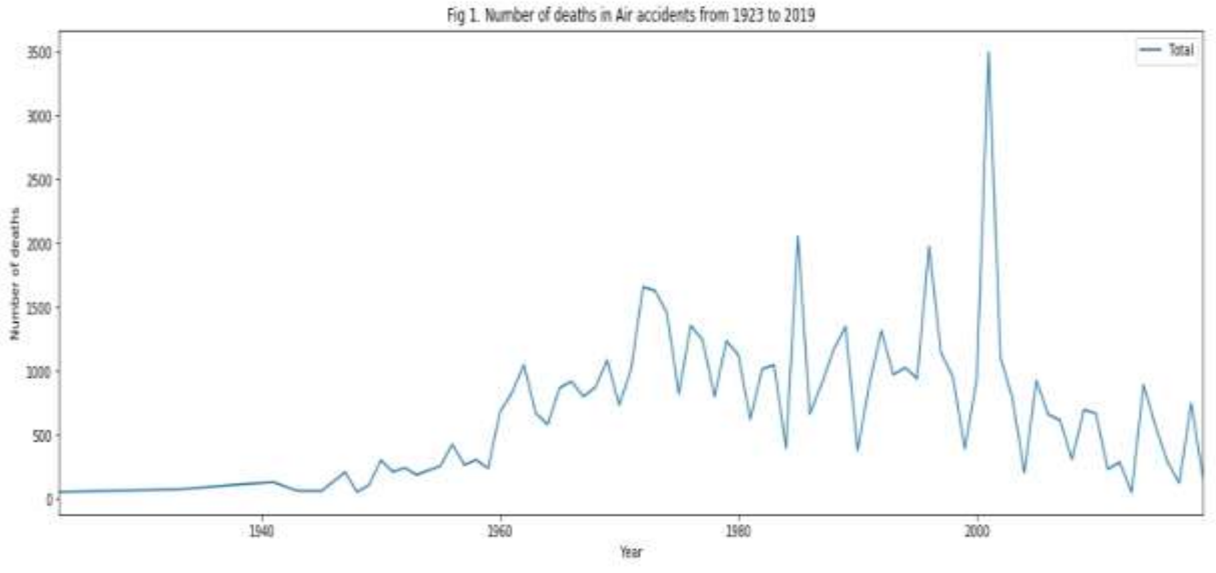
We clearly remember 2001-09-11 when U.S.A. faced the deadly terrorist attack of history.

3.2.6. What are the top 5 darkest years in aviation?

In terms of number of death, the dealiest years are :

- 2001: 3495 killed
- 1985: 2052 killed
- 1996: 1975 killed
- 1972: 1655 killed
- 1973: 1627 killed

Let's view it on the line plot below.



And in terms of occurrence of accidents:

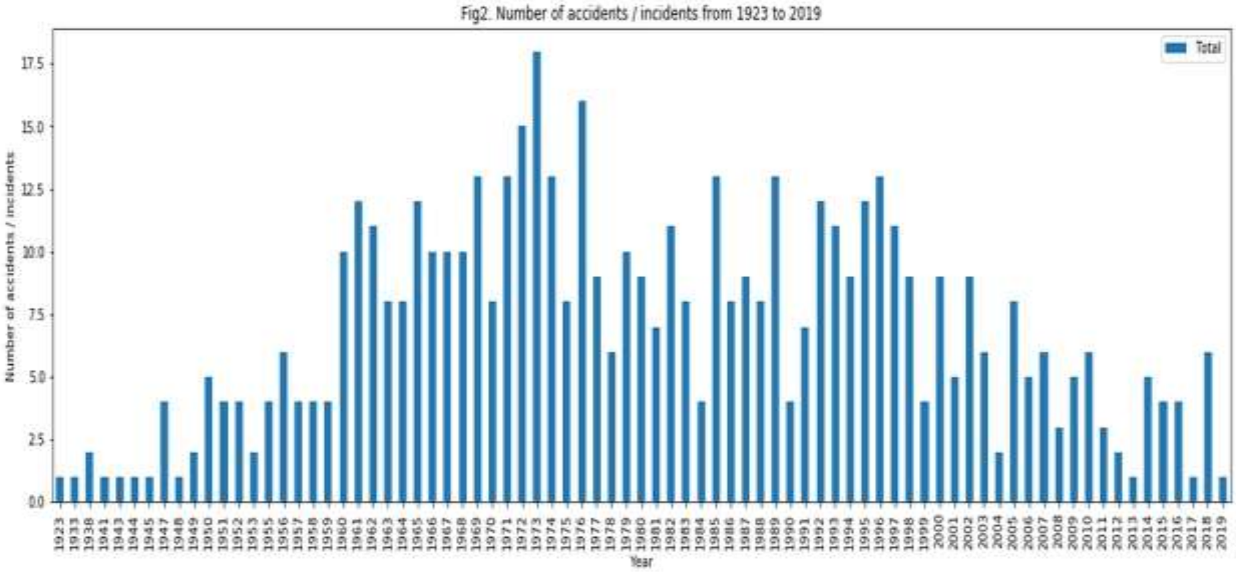
The top 5 years where accidents were frequent are:

- 1973 with 18 accidents
- 1976 with 16 accidents
- 1972 with 15 accidents
- 1974 with 13 accidents

- 1969 with 13 accidents

Let's plot those information on a bar plot

Fig2. Number of accidents / incidents from 1923 to 2019



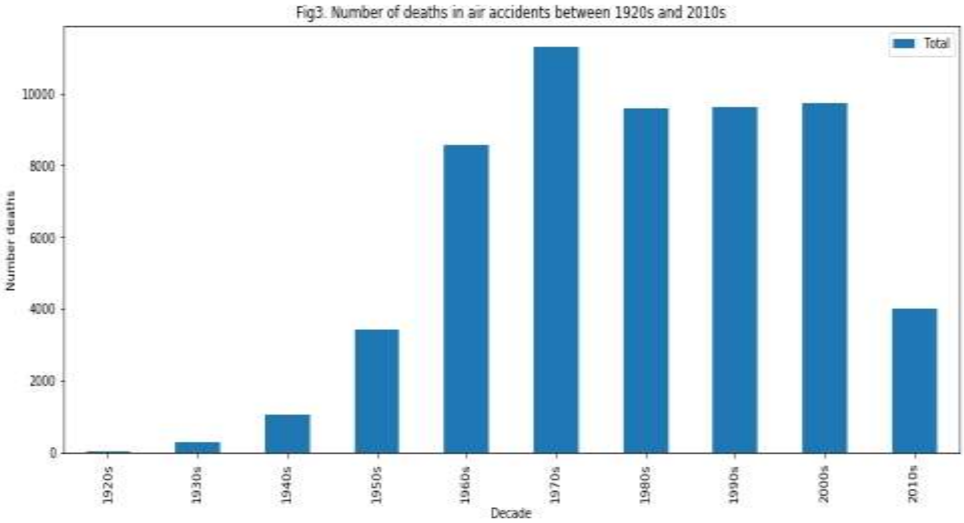
3.2.7. How about the decades?

- **In terms of number of people killed:**

The darkest decade is 1970s with about 11309 people killed

	Total
Decade	
1970s	11309
2000s	9713
1990s	9609
1980s	9578
1960s	8580

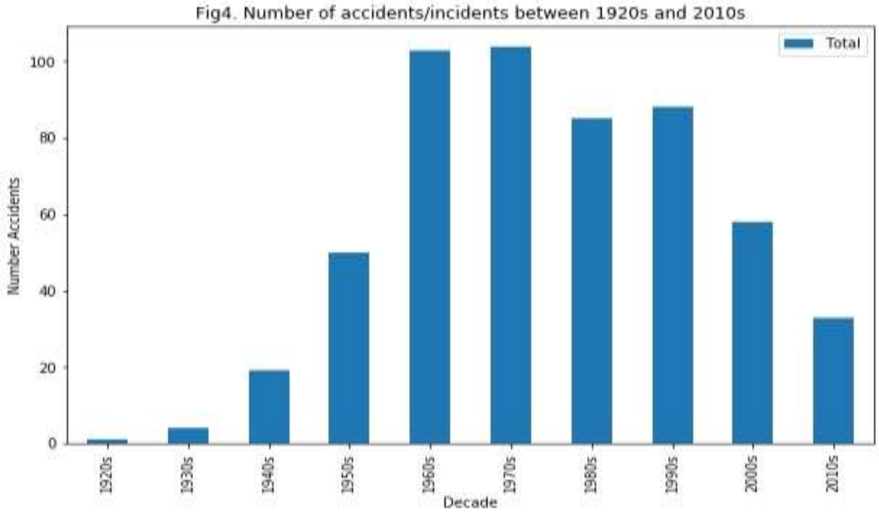
Fig3. Number of deaths in air accidents between 1920s and 2010s



- **And in terms of occurrence of accidents:**

Decade	Total
1920s	1
1930s	4
1940s	19
1950s	50
1960s	103
1970s	104
1980s	85
1990s	88
2000s	58
2010s	33

Fig4. Number of accidents/incidents between 1920s and 2010s



The 1970s and 1960s have seen the greatest number of occurrences of accidents

3.2.8. What are the top 5 locations that recorded the highest number of deaths?

The top 5 locations are:

- **USSR** with 3451 killed
- **New York City, New York, U.S.** with 2700 killed
- **Tenerife, Spain** with 583 killed
- **Ueno, Japan** with 520 killed
- **Iran** with 495

Here's the dataframe made of the Top5:

	Location	Latitude	Longitude	Total
0	USSR	55.750446	37.617494	3451
1	New York City, New York, U.S.	40.712728	-74.006015	2700
2	Tenerife, Spain	28.293578	-16.621447	583
3	Ueno, Japan	35.711788	139.776096	520
4	Iran	32.940750	52.947134	495

Table 8: The top 5 locations

3.2.9. Visualization

First let's visualize the different locations where accidents happened:

Fig 5: World map with the various accidents locations.

And now let’s focus on the Top5: The radius is proportional to the number of deaths.

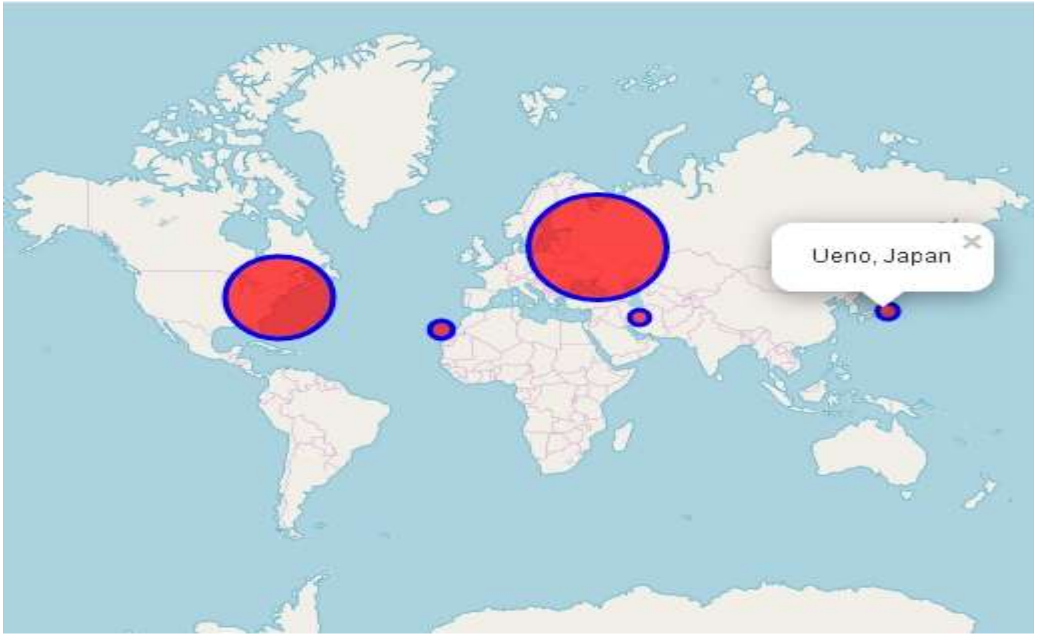


Fig 6: World map with the top5 accidents locations.

4. EXPLORING THE TOP5 DEADLIEST LOCATIONS

We are going to utilize the Foursquare API to explore the neighborhoods and segment them.
Let's check how many venues were returned for each neighborhood

	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
New York City, New York, U.S.	100	100	100	100	100	100
Tenerife, Spain	2	2	2	2	2	2
USSR	21	21	21	21	21	21
Ueno, Japan	100	100	100	100	100	100

Table 9: Venues returned for each of the Top5

No information as far as Iran, the 5th location is concerned.
There is a total of 109 unique categories that can be curated from all the returned venues

Now let's create the new dataframe and display the top 10 venues for each neighborhood.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	New York City, New York, U.S.	Coffee Shop	Sandwich Place	Café	Falafel Restaurant	Park	Italian Restaurant	Hotel	Gym	Plaza	Bakery
1	Tenerife, Spain	Restaurant	Waterfront	Furniture / Home Store	Coworking Space	Cuban Restaurant	Dance Studio	Discount Store	Donburi Restaurant	Electronics Store	Event Space
2	USSR	History Museum	Plaza	Palace	Historic Site	Boutique	Concert Hall	Hotel	Event Space	Government Building	Garden
3	Ueno, Japan	Sake Bar	Japanese Restaurant	Café	Chinese Restaurant	Bed & Breakfast	BBQ Joint	Tonkatsu Restaurant	Wagashi Place	Ramen Restaurant	Yakitori Restaurant

Table 10: Top 10 Venues returned for each of the Top5

5. CLUSTERING THE TOP5 DEADLIEST LOCATIONS

5.1. Building clusters

We run *k*-means to cluster the neighborhood into 2 clusters

Let's create a new dataframe that includes the cluster as well as the top 10 venues for each neighborhood.

	Location	Latitude	Longitude	Total	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	USSR	55.750446	37.617494	3451	0	History Museum	Plaza	Palace	Historic Site	Boutique	Concert Hall	Hotel	Event Space	Government Building	Garden
1	New York City, New York, U.S.	40.712728	-74.006015	2700	0	Coffee Shop	Sandwich Place	Café	Falafel Restaurant	Park	Italian Restaurant	Hotel	Gym	Plaza	Bakery
2	Tenerife, Spain	28.293578	-16.621447	583	1	Restaurant	Waterfront	Furniture / Home Store	Coworking Space	Cuban Restaurant	Dance Studio	Discount Store	Donburi Restaurant	Electronics Store	Event Space
3	Ueno, Japan	35.711788	139.776096	520	0	Sake Bar	Japanese Restaurant	Café	Chinese Restaurant	Bed & Breakfast	BBQ Joint	Tonkatsu Restaurant	Wagashi Place	Ramen Restaurant	Yakitori Restaurant

Table 11: Top 10 Venues returned for each of the Top5 with clusters

Let's visualize the resulting clusters

Fig 7: Locations clustered.

5.2. Examining Clusters

Now, we examine each cluster and determine the discriminating venue categories that distinguish each cluster. Based on the defining categories.

Cluster 1 (red on fig. 7**)**

	Location	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	USSR	History Museum	Plaza	Palace	Historic Site	Boutique	Concert Hall	Hotel	Event Space	Government Building	Garden
1	New York City, U.S.	Coffee Shop	Sandwich Place	Café	Falafel Restaurant	Park	Italian Restaurant	Hotel	Gym	Plaza	Bakery
3	Ueno, Japan	Sake Bar	Japanese Restaurant	BBQ Joint	Café	Chinese Restaurant	Bed & Breakfast	Wagashi Place	Tonkatsu Restaurant	Ramen Restaurant	Yakitori Restaurant

Table 12: Cluster1

Cluster 2 (blue on fig. 7**)**

	Location	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
2	Tenerife, Spain	Restaurant	Waterfront	Furniture / Home Store	Coworking Space	Cuban Restaurant	Dance Studio	Discount Store	Donburi Restaurant	Electronics Store	Event Space

Table 13: Cluster2

6. CONCLUSION

Our question in the beginning was: **Let's say a tourist would like to visit one of the top five locations where air crash killed the most people, but don't know which one to choose. What would you suggest?**

In order to answer the question, we first imported and pre-processed the data, we made some analysis and found that the top5 most deadly locations in air crash from 1923 to 2019 are in order: USSR, New-York (USA), Tenerife (Spain), ueno (Japan) and Iran.
Using the Foursquare API, we determined the top 5 most common venues for each location and finally we clustered the locations (Unfortunately, we didn't get any information on Iran).

So for a tourist interested in Museum or historical sites, we would suggest he visits USSR. Whereas if he is more interested in restaurant, hotel or park, we would suggest he visits New-York City. But if he wants to visit the very unique site among the top 5, we would suggest he visits Tenerife in Spain.
In any case, Table 12 and Table 13 will provide directions to the tourist.

7. TO GO FURTHER

As we mentioned, retrieving the exact coordinates of different locations where accidents happened was not easy using the geocoder. For example USSR is very generic, it groups so many locations (with the name USSR) just because we couldn't get the precise geographic coordinates based on the names in the initial database.

For better results, we suggest to redo the study by better refining the names of locations or using another tool apart from the geocoder (if one exists) so as to have the exact locations.

As we have the data, way too far from the question of this document, couldn't we push the research further by seeing for example whether any link exists between accident and a particular type of aircraft, a particular airline company. Is there a relationship between accidents and a particular day of the year? Based on the data we have, can we predict the next air crash?