CRASHED AIRPLANES

Alanoud Almutairi ,Zahra Sameer and Alnirah Alqahtani

INTRODUCTION..

- Over the last centuries, the world has seen several airplane crashes both in the sky and on the ground. The development of technology aims to decrease technological difficulties and human errors. However, fatalities and aircraft damages caused by crashes do not cease to exist.
- Our results found in this project will benefit the ongoing investigations into this important topic. Understanding what factors cause airplane crashes helps aviation industries make continuous improvement in flight safety.

OBJECTIVES:

- Our objective is to perform an Exploratory Data Analysis (EDA) to determine the common reason of airplane crash, countries with maximum/minimum airplane crashes, fatalities vs survived ratio and any other interesting trend. In addition, to develop a Model determining the following;
  - NMF Model
  - LDA Model

DESCRIPTION OF THE DATA:

- This dataset presents number of crashed airplanes since 1908, where we have 5268 record, Also we have 13 features in this dataset.

Out[7]:

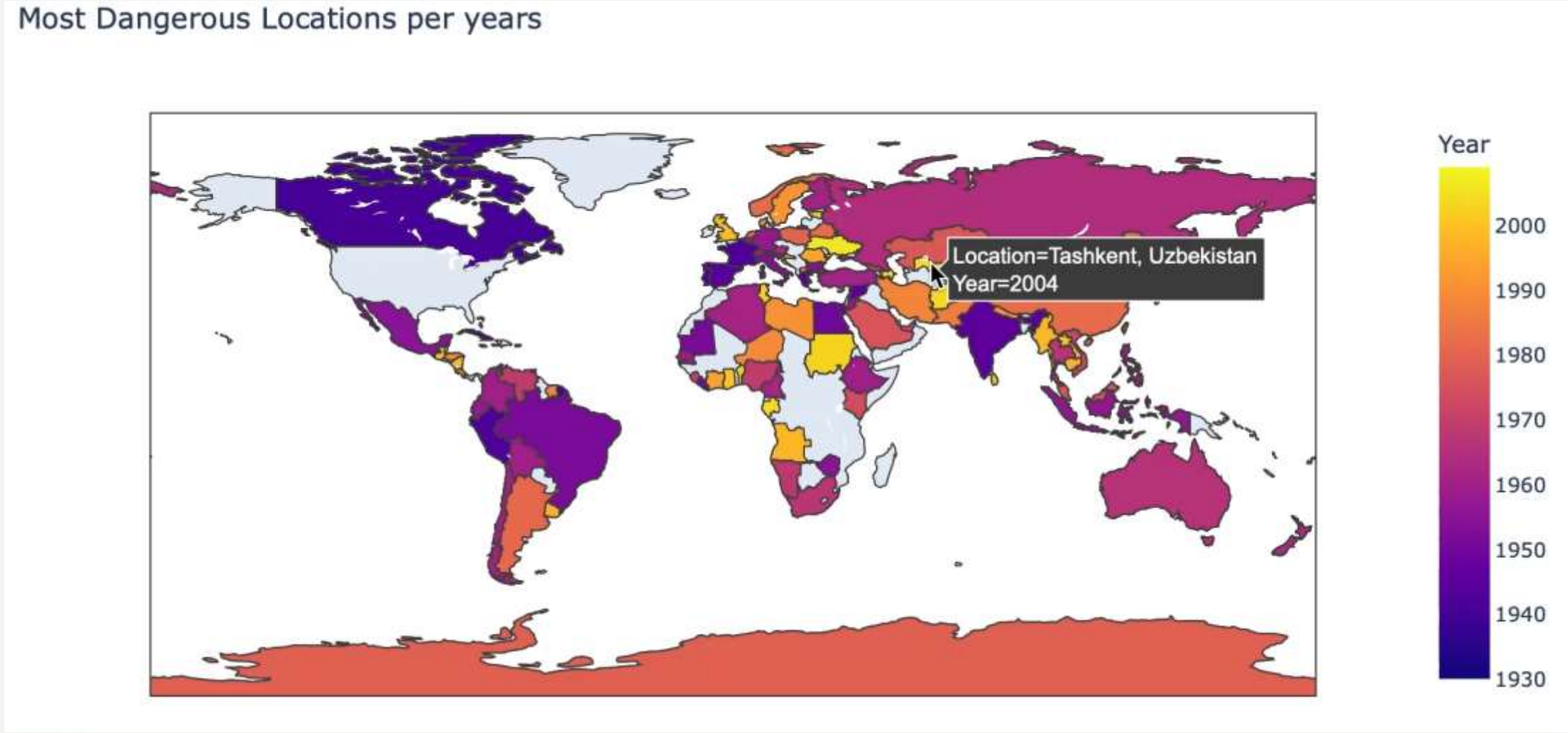
	Date	Time	Location	Operator	Flight #	Route	Type	Registration	cn/ln	Aboard	Fatalities	Ground	Summary	Year
208	01/19/1930	18:23	Oceanside, California	Maddux Airlines	7	Aqua Caliente, Mexico - Los Angeles	Ford 5-AT-C Tri Motor	NC9689	5-AT-046	16.0	16.0	0.0	While en route to Los Angeles, the pilot, flyi...	1930
236	03/31/1931	10:45	Bazaar, Kansas	Trans Continental and Western Air	599	Kansas City - Wichita - Los Angeles	Fokker F10A Trimotor	NC-999	1063	8.0	8.0	0.0	Shortly after taking off from Kansas City, one...	1931
334	08/31/1934	23:42	Amazonia, Missouri	Rapid Air Transport	6	Omaha - St. Joseph	Stinson SM-6000B	NC10809	5004	5.0	5.0	0.0	The plane crashed about 11 miles from St. Jose...	1934

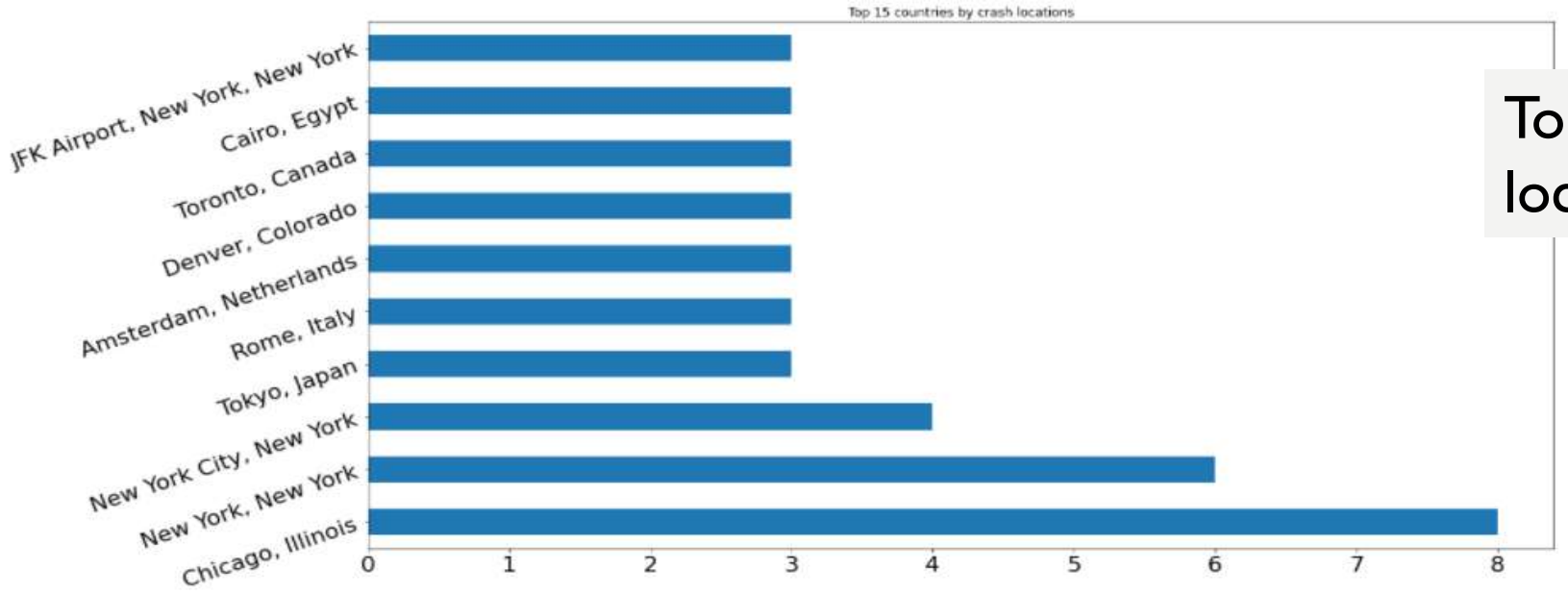
DATA PREPARATION:

- First, we check for duplicates and null data and drop it we get 944 cleaned record
- Second, add year column to the dataset.
- Third add Survivor column by subtract number of Aboard and Fatalities on each airplane.
- Forth, add 'summary clean' column that contain summary after text cleaning using StopWords, Stemmer.

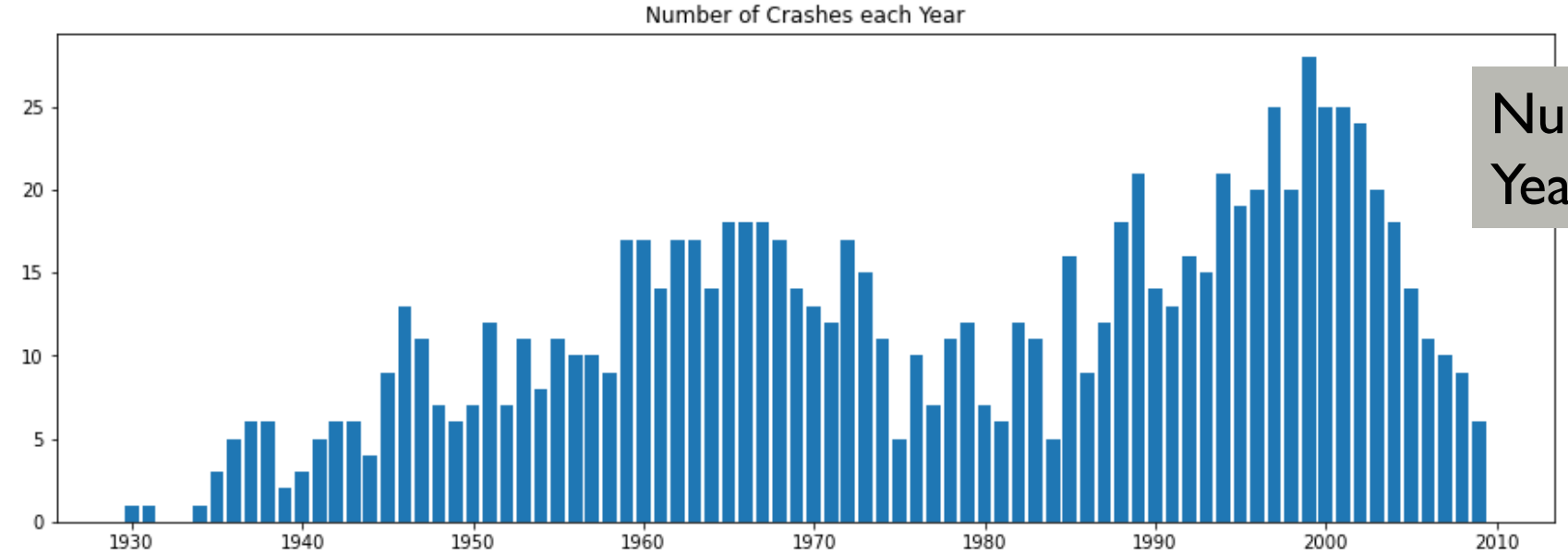
	Date	Time	Location	Operator	Flight #	Route	Type	Registration	cn/ln	Aboard	Fatalities	Ground	Summary	Year	Survivor	Summary_clean
3	01/19/1930	18:23	Oceanside, California	Maddux Airlines	7	Aqua Caliente, Mexico - Los Angeles	Ford 5-AT-C Tri Motor	NC9689	5-AT-046	16.0	16.0	0.0	While en route to Los Angeles, the pilot, flyi...	1930	0.0	en route los angeles pilot flying low altitude...
5	03/31/1931	10:45	Bazaar, Kansas	Trans Continental and Western Air	599	Kansas City - Wichita - Los Angeles	Fokker F10A Trimotor	NC-999	1063	8.0	8.0	0.0	Shortly after taking off from Kansas City, one...	1931	0.0	shortly taking kansas city one aircrafts wings...

RESULT:

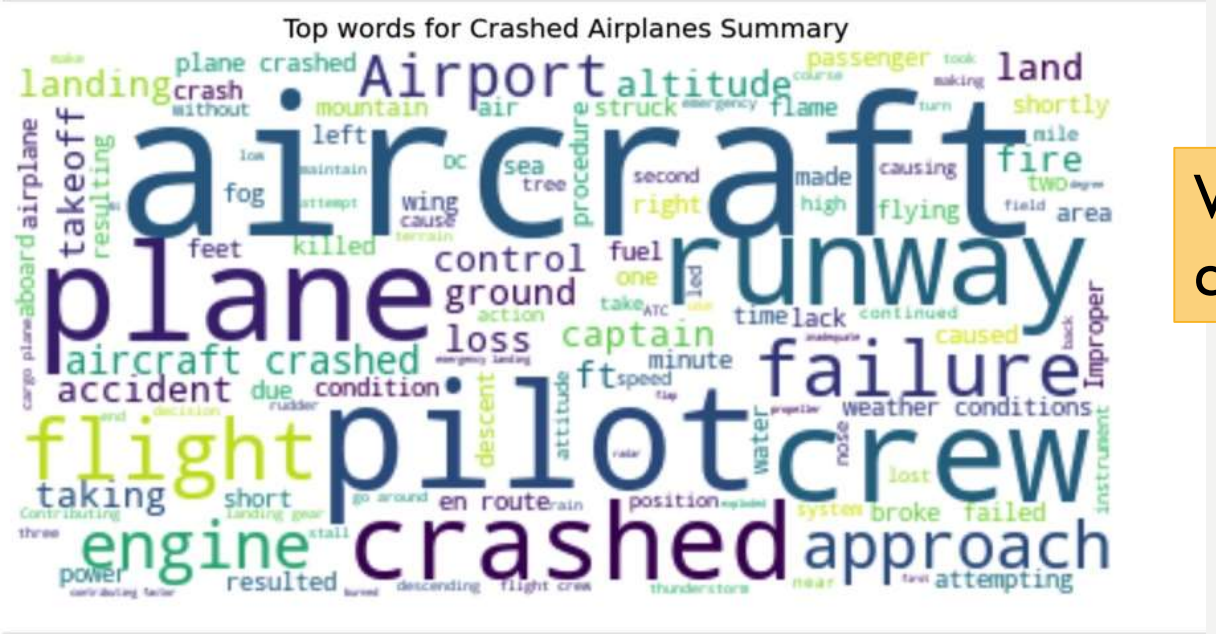

Most Dangerous Locations per years
Location=Tashkent, Uzbekistan
Year=2004
Year
2000
1990
1980
1970
1960
1950
1940
1930



Top 15 countries by crash locations..

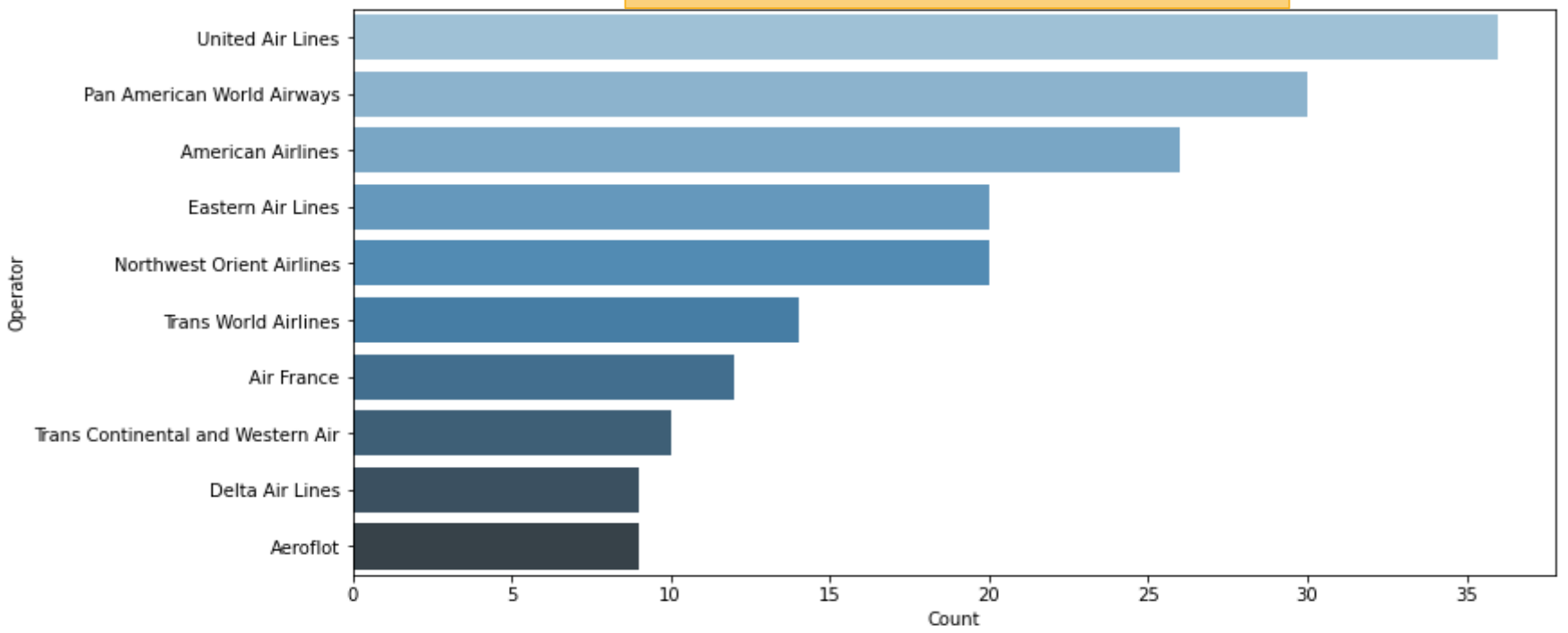


Number of Crashes each Year

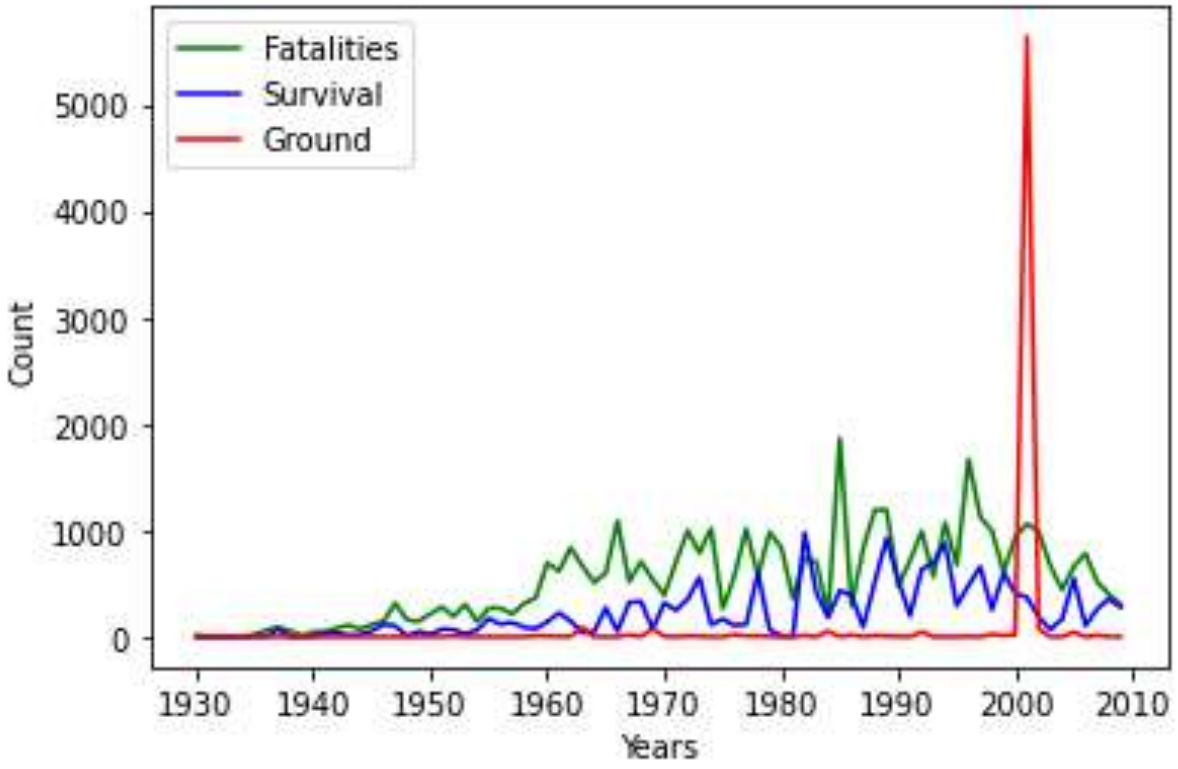


Word cloud for the most common terms

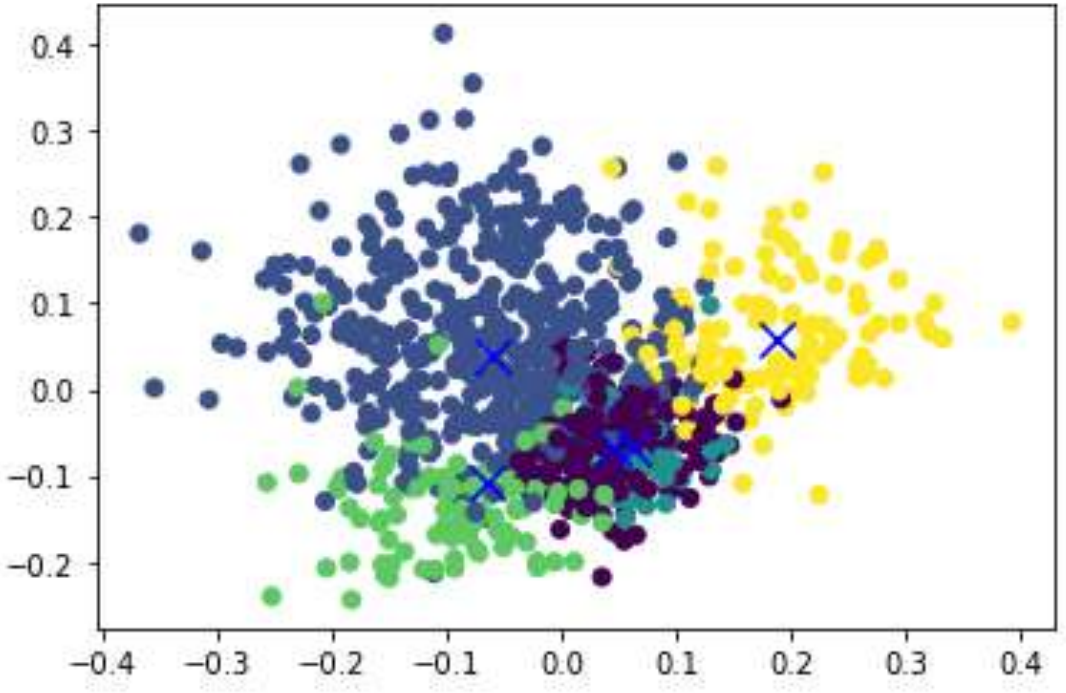
Total Count by Operator



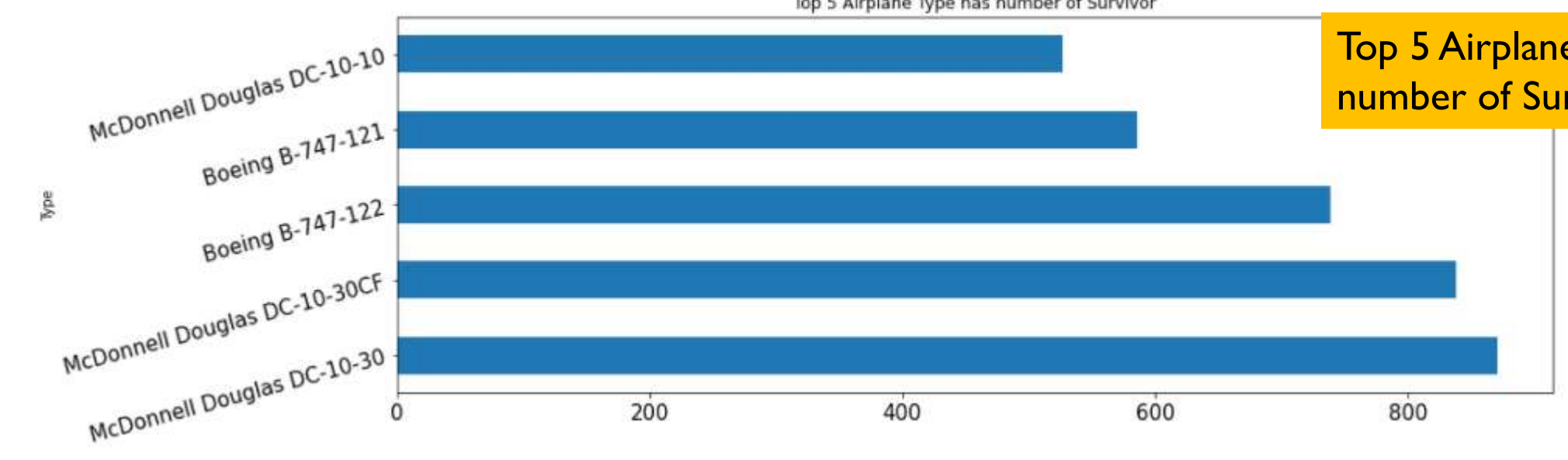
Fatalities, Survival and on Ground per year:



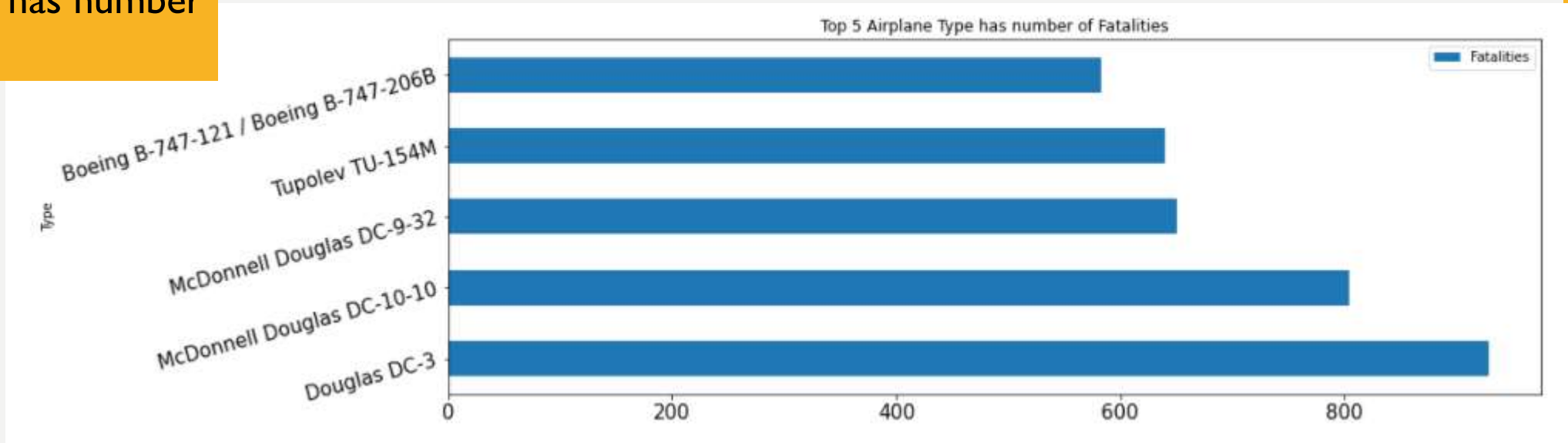
Clustering the most common terms:



Top 5 Airplane Type has highest number of Survivor:



Top 5 Airplane Type has number highest of Fatalities:



NMF - Model

	pilot error	On Fire	crash	poor weather	shot down	dominant_topic
Doc0	0.126	0.000	0.145	0.004	0.045	2
Doc1	0.041	0.025	0.075	0.094	0.000	3
Doc2	0.153	0.000	0.000	0.000	0.023	0
Doc3	0.007	0.000	0.004	0.155	0.000	3
Doc4	0.000	0.042	0.199	0.000	0.000	2
Doc5	0.091	0.248	0.031	0.000	0.003	1
Doc6	0.045	0.113	0.000	0.033	0.064	1
Doc7	0.052	0.000	0.048	0.000	0.054	4
Doc8	0.037	0.012	0.178	0.072	0.000	2
Doc9	0.001	0.006	0.000	0.204	0.000	3

	pilot error	On Fire	shot down	poor weather	crash	dominant_topic
Doc0	0.04	0.01	0.00	0.02	0.11	4
Doc1	0.04	0.01	0.08	0.10	0.01	3
Doc2	0.00	0.01	0.01	0.18	0.00	3
Doc3	0.00	0.15	0.00	0.00	0.00	1
Doc4	0.16	0.00	0.04	0.02	0.01	0
...	...	...	...	...	...	...
Doc655	0.00	0.00	0.05	0.09	0.11	4
Doc656	0.02	0.02	0.22	0.00	0.03	2
Doc657	0.00	0.00	0.01	0.03	0.10	4
Doc658	0.06	0.00	0.03	0.08	0.04	3
Doc659	0.00	0.23	0.00	0.00	0.04	1

660 rows × 6 columns

LDA-Model

	pilot error	On Fire	crash	dominant_topic
Doc0	0.07	0.07	0.87	2
Doc1	0.06	0.06	0.88	2
Doc2	0.09	0.09	0.82	2
Doc3	0.09	0.09	0.83	2
Doc4	0.06	0.06	0.88	2

	pilot error	On Fire	crash	dominant_topic
Doc0	0.07	0.07	0.87	2
Doc1	0.06	0.06	0.88	2
Doc2	0.06	0.06	0.89	2
Doc3	0.08	0.09	0.83	2
Doc4	0.06	0.06	0.89	2
Doc5	0.08	0.08	0.84	2
Doc6	0.06	0.06	0.88	2
Doc7	0.06	0.06	0.89	2
Doc8	0.05	0.05	0.90	2
Doc9	0.06	0.06	0.87	2

PROJECT LIMITATIONS:

- Although dataset have a large amount of data, they have quite several unrecorded values, therefore some columns had to be either removed or filled with correct values.

Tools:

- Technologies : Python, Jupyter Notebook
- Libraires : Pandas, Numpy , Seaborn , Sklearn, Matplot ,WordCloud and number of python libraries.

CONCLUSION..

The main objective of our project is to raise awareness of flight safety and better understand its problems and progress, so that aviation industries can continue to improve. We hope that more information understanding will lead to industry changes that save lives.

THANK YOU ..