

Report

by David Robin James

FILE	4073_DAVID_ROBIN_JAMES_REPORT_309231_1009431102.DOCX (850.14K)		
TIME SUBMITTED	01-FEB-2021 11:46PM (UTC+0000)	WORD COUNT	3624
SUBMISSION ID	143976739	CHARACTER COUNT	19361

Machine Learning 1

Student Name: David James (1704050)

Module: Application of Machine Learning

Lecturer: Gordon Dickers

Table of Contents

Table of Figures.....	3
1. Abstract.....	4
2. Introduction	4
2.1. Scope.....	4
2.2. Acquiring a Dataset	5
2.2.1. Data Scraping	5
2.3. Expectations.....	6
2.3.1. Considerations	6
2.3.2. Model Selection	6
2.4. Pre-processing & Cleaning	8
2.5. Polynomial Regression	9
2.6. Logistic Regression	10
3. References	11
4. Appendix	12
4.1. Appendix A – Data Scraping	12
4.2. Appendix B – Pre-processing	12
4.3. Appendix C – Exploratory Data Analysis	13
4.3.1. Appendix C1 – Visualizing the Data.....	14

Table of Figures

Figure 1 - Accident and fatality trend (commercial aviation, worldwide)	4
Figure 2 - Plane crash database homepage	5
Figure 3 - Percentage of military and commercial operators (15% (red) commercial, 85% (blue) military)	7
Figure 4 - Miscellaneous values within the dataset	8
Figure 5 - Date column from raw scraped dataset	8
Figure 6 - Classification column of pre-processed data	9
Figure 7 - Polynomial regression fitted to total number of crashes in dataset	9
Figure 8 - RMSE score	10
Figure 9 - Code snippet of polynomial regression model	10
Figure 10 - Logistic regression classifying flight operators based on number of fatalities	10
Figure 11 - Confusion matrix	11
Figure 12 - All entries for the year 1920	12
Figure 13 - Pre-processing functions	13
Figure 14 - Pre-processed sub-sets of original data	13
Figure 15 - Top 20 aircraft with highest number of crashes	14
Figure 16 - Total fatalities in aviation over a century, death ratio ranges from 0% (dark blue), to 100% (bright yellow)	14
Figure 17 - Total passengers per year	15
Figure 18 - Total amount of crashes per year	16

1. Abstract

Data selection and pre-processing are important steps to fulfil when achieving a successful machine learning model [1]. Alongside this, the methods and techniques carried out when analysing and exploring obtained data play a heavy role when ensuring suitable features are extracted to maximize efficiency and performance of a machine learning model [2]. To achieve effective models each of the discussed procedures are required to be executed to discover which machine learning algorithms provide the best results to solving the desired task. A relatively small dataset is acquired with the focus of predicting aviation accidents. Pre-processing and exploratory techniques are applied for feature selection and data cleaning with the goal of applying two machine learning algorithms.

2. Introduction

To meet the requirements of the assignment a research area was selected. Further investigation on available datasets was then undertaken to obtain data pertinent to the area of research with the goal of applying machine learning algorithms that would yield models that can be used for predicting a target based on features chosen from the dataset collected.

The following report documents brief findings from research relevant to aviation accidents along with selection of a dataset and the process involved in picking appropriate features and machine learning algorithms best suited towards solving the applied problem.

2.1. Scope

Research shows that flying is a recurrent phobia with common fears such as “*turbulence, noise and terrorist attacks*” causing anxiety in passengers [3]. This shows a negative influence, giving the impression that flying may not be as safe as it seems.

In modern times statistics show that flying has become a safe form of travel with data showing less accidents when compared with other modes of travel [4][5]. According to the ICAO (*International Civil Aviation Organization*) fatal accidents in *commercial* aviation have been as low as 5 instances in 2017 and 6 in 2019.

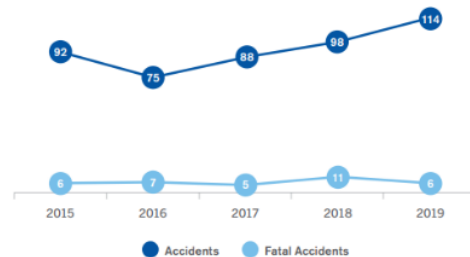


Figure 1 - Accident and fatality trend (commercial aviation, worldwide)

Source: Adapted from [6]

Regarding the scope of the assignment an initial concept of predicting how safe flying will be in the future led to a few immediate issues. With the intentions of applying different machine learning

algorithms to produce two models trained on historical data to predict potential aviation accidents in the future required a different approach for constructing reliable models.

Research led to a common method known as *time series analysis* [7][8][9]. With the selected dataset, applying regression or classification techniques alone wouldn't lead to accurate predictions as the data available didn't supply enough information to formulate any trends that account for *why* all accidents happened. Additional data such as economic and employment state, weather conditions, etc, would be required which would provide some fundamental understanding of the potential factors that produced the data.

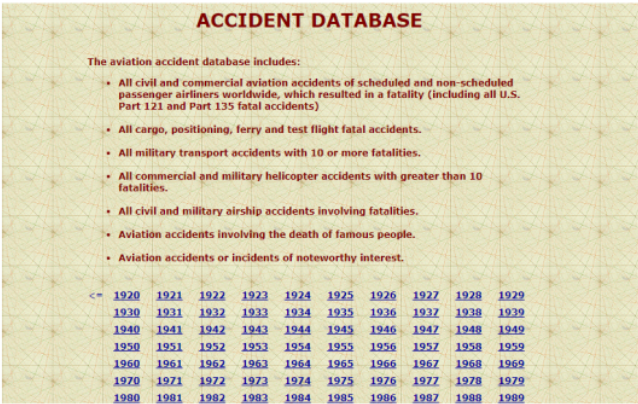
With the complexity of applying time series analysis to the context of supervised learning algorithms along with the constraints of time and expectations of the assignment requirements, scope of the project shifted towards a more simplified problem of finding a predictor within the dataset and applying supervised learning algorithms to produce machine learning models.

2.2. Acquiring a Dataset

Many available datasets provided information on accidents in aviation and sourcing a dataset that offered enough information proved a difficult task. Most datasets only provided data over a short span of time or didn't offer enough consistent data to be useful.

The dataset used for the project provides over 100 years' worth of recorded aviation accidents in both commercial and military aviation, also documenting fatalities, aircraft type and total number of passengers. All entries in the database are exclusive to plane crashes, this offered a strict limitation which directly corresponded towards the interests of the project.

Appropriately named the '*Plane Crash Database*', selection of the dataset provided a wide span of historic information that ranges from 1908 to 2020 offering fairly persistent entries throughout.



ACCIDENT DATABASE

The aviation accident database includes:

- All civil and commercial aviation accidents of scheduled and non-scheduled passenger airliners worldwide, which resulted in a fatality (including all U.S. Part 121 and Part 135 fatal accidents)
- All cargo, positioning, ferry and test flight fatal accidents.
- All military transport accidents with 10 or more fatalities.
- All commercial and military helicopter accidents with greater than 10 fatalities.
- All civil and military airship accidents involving fatalities.
- Aviation accidents involving the death of famous people.
- Aviation accidents or incidents of noteworthy interest.

<=	1920	1921	1922	1923	1924	1925	1926	1927	1928	1929
	1930	1931	1932	1933	1934	1935	1936	1937	1938	1939
	1940	1941	1942	1943	1944	1945	1946	1947	1948	1949
	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959
	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969
	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979
	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989

Figure 2 - Plane crash database homepage

2.2.1. Data Scraping

In order to retrieve the data a technique known as data or web scraping [10] was required. Time was spent obtaining knowledge on how to perform the data scraping procedure which involved observing the layout of the website and how the data was formatted and stored.

A Python script was developed (refer to Appendix A – Data Scraping) for scraping the entire available database.

Due to the database residing on an online server many connection requests were needed to obtain the information for each entry. This resulted in a long waiting process to accrue all the data which hindered testing, leaving not much room for programming-related errors.

After successfully retrieving all information from the database the raw, scraped data was written to a comma separated values file format for ease of access. This was to accommodate an offline solution to working with the dataset.  3

2.3. Expectations

Gaining further insight into the data to answer questions pertinent to the project aim, visualizing the data delivered interesting observations (see Appendix C1 – Visualizing the Data). Some few questions were imposed to deduce further insights to discovering features and selecting effective machine learning algorithms that provided best fit models.

The following questions were asked to gauge further knowledge into what the data had to offer:

- Which aircrafts had the most accidents?
- Which years saw the most crashes?
- What was the total fatalities across all years?
- What percentage of passengers survived each crash?

Examining the data to reveal the statistics of each question broadened understanding of potential pre-processing techniques that needed to be applied and also uncovered what information the data provided.

Visualization was also an opportunity to reveal potential patterns and trends which produced suggestions towards applicable algorithms and also gaining independent variables that would make up the feature set used to train both models.

2.3.1. Considerations

Whilst the selected dataset offers a wide array of information it was noticed that the data was fairly limited when applied to the intended use. All entries in the dataset relate to plane crashes exclusively, this was felt to be a logical choice when sourcing a dataset but after exploring the data it was revealed that not enough information was available.

Crashes in aviation aren't a frequent occurrence, where one year may see many crashes, another may not. Because of this, when visualizing the distribution of the data irregularities can be seen that could affect the accuracy and performance of the model. This could lead to the assumption of many outliers being present in the data. As each entry is a true event, describing the spread visualized in the data as anomalies or outliers may not be the correct choice and could lead to further issues that affect the quantity of data, as well as the integrity. For this reason, combining the current dataset with similarly formatted data for flights that didn't result in a crash would have provided a more diverse and complete dataset with much richer information.

2.3.2. Model Selection

With the observations made from visualizing the data clear concepts formed which led to targets and features that were usable. This provided context for which algorithms would be suitable to apply to the sub-set of pre-processed data.

Initial thoughts on a linear regression model where the sub-set of data provided a single feature of time and a target of accumulated crashes per-year was deemed acceptable. When visualizing the data (as seen in Figure 18 - Total amount of crashes per year) a clear non-linear trend can be observed where a peak in crashes occurs with a sudden fall off. With this observed a model that would better fit the data such as polynomial regression would provide a more accurate measure of prediction.

Inspecting the data further, visualization of which operators each crash inherited provided data which suited for the basis of a classification problem. Reviewing the data exposed clear categories for each flight logged in the dataset, visualizing the spread of data between both commercial and military operators (see

Figure 3).

Percentage of Military and Commercial Operators

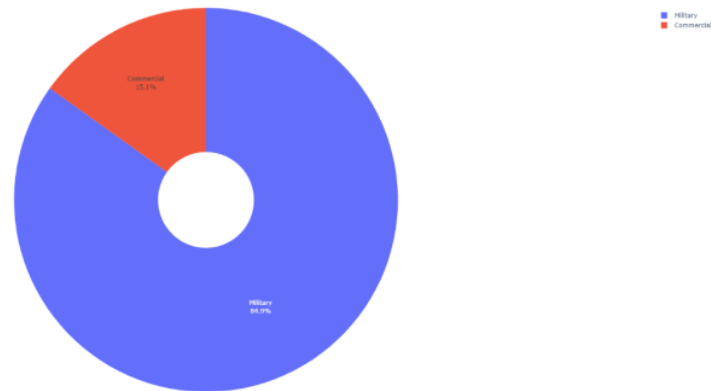


Figure 3 - Percentage of military and commercial operators (15% (red) commercial, 85% (blue) military)

With each flight correlating to either a commercial or military operator a classification algorithm such as KNN would provide a suitable way to predict which flights in the sub-set of data belong to which operators. A clear association between the amount of deaths of commercial and military flights revealed a pattern where the amount of passengers onboard commercial flights was higher than military flights.

Utilizing this correlation could lend a transparent solution towards classifying commercial and military flights based on the amount of passengers. KNN provides a suitable model for solving this form of classification problem and fits the requirements of the data as it is supervised learning algorithm. Logistic regression is felt to be the better fit as the sub-set of data provides two classes suiting a binary classifier which logistic regression provides.

2.4. Pre-processing & Cleaning

On first inspection of the dataset many fields held miscellaneous and missing values. Simple cleaning for removing each row of information resolved to a dataset with over 4000 entries. The process involved iterating over the entire dataset and checking each column of each row for nullified values (see Functions within the data scraping script were created for common operations that needed to be completed to scrap data from each year. The process involved selecting a year, scraping all entries of each year for their corresponding links, accessing every link for every entry in a

I	J	K	L
cn / ln:	Aboard:	Fatalities:	Ground: Summary:
1 2	(passengers:1 crew:1)	1 (passengers:1 crew:0)	0 During a demonstration flig
?	1 (passengers:0 crew:1)	1 (passengers:0 crew:0)	0 Eugene Lefebvre was the fi
?	5 (passengers:0 crew:5)	5 (passengers:0 crew:5)	0 First U.S. dirigible Akron ex
?	1 (passengers:0 crew:1)	1 (passengers:0 crew:1)	0 The first fatal airplane acci
?	20 (passengers:? crew:?)	14 (passengers:? crew:?)	0 The airship flew into a thur
?	30 (passengers:? crew:?)	30 (passengers:? crew:?)	0 Hydrogen gas which was be
?	41 (passengers:? crew:?)	21 (passengers:? crew:?)	0 Crashed into trees while at
?	19 (passengers:? crew:?)	19 (passengers:? crew:?)	0 Exploded and burned near
?	20 (passengers:? crew:?)	20 (passengers:? crew:?)	0 Crashed near the Black Sea,
?	22 (passengers:? crew:?)	22 (passengers:? crew:?)	0 Shot down by British aircraf

year and further scraping the required data.

Appendix B – Pre-processing). Due to the columns in each row relating to a single instance of data within the dataset, dropping values within a row would have caused further anomalies within the dataset.

Figure 4 - Miscellaneous values within the dataset

Prior to visualizing and exploring the data, intuition led to an opportunity for feature engineering [11]. With available data for total passengers and fatalities, condensing the data into a single value whilst maintaining the information resulted in creating a ratio between both variables. This was done to further decrease the amount of features in order to decimate dimensions in an attempt to relieve complexity.

Further pre-processing was conducted before visualizing the data in order to clean fields that held unnecessary information such as the date column.

A	B
1 Date:	Time: Location:
2 September 17, 1908	1718 Fort Myer, Virginia
3 September 07, 1909	? Juvisy-sur-Orge, France
4 July 12, 1912	630 Atlantic City, New Jersey
5 August 06, 1913	? Victoria, British Columbia
6 September 09, 1913	1830 Over the North Sea
7 October 17, 1913	1030 Near Johannisthal, Germi

Figure 5 - Date column from raw scraped dataset

Inspecting the format of the date column shows values for month, day, and year (see Figure 5). To keep the essence of the data, a compact representation which maintains just the year provided a solution which can be used as an approximate measure when visualizing the data.

With the additions of new features created within the dataset such as the survival and death ratio, a new feature to encode which class each entry belonged to was made. Two operators within the dataset provided information for commercial and military flights, leveraging the current format of

the data, all military flight entries were prefixed with the *Military* keyword, this provided a way to encode a binary classifier where military flights were marked as a 0 and commercial flights were marked with a 1 (see Figure 6).

	A	B	C	D	E	F
1	Year	Aboard	Fatalities	Survival Ratio	Death Ratio	Classification
2	1908	2	1	50	50	1
3	1912	5	5	0	100	1
4	1913	1	1	0	100	0
5	1913	20	14	30	70	1
6	1913	30	30	0	100	1
7	1915	41	21	49	51	1
8	1915	19	19	0	100	1
9	1916	20	20	0	100	1
10	1916	22	22	0	100	1
11	1916	19	19	0	100	1
12	1916	28	27	4	96	1
13	1916	20	20	0	100	1
14	1917	20	20	0	100	1

Figure 6 - Classification column of pre-processed data

With a collection of features extracted from the dataset all other fields were dropped to reduce disorder. Writing the data to a new .csv file kept a working copy for the desired state of the dataset.

Throughout the feature extraction phase various new .csv files were written in order to keep a record of changes over the course of extracting features from the dataset.

2.5. Polynomial Regression

Loading of the pre-processed dataset resulted in storing the loaded data into a data-frame object. With all pre-processing and cleaning complete, writing the data to a new .csv file eased the process when performing polynomial regression.

```
# Open and read the training .csv file
data_frame = pandas.read_csv('data\\pre-processed\\total_number_of_crashes_yearly.csv')

X = data_frame.Year.values.reshape(-1, 1)
y = data_frame.Crashes

# Create polynomial regresson model
poly = PolynomialFeatures(degree = 2)

# Create feature matrix with polynomial terms
X_poly = poly.fit_transform(X)

# Fit polynomial regression model
polynomial_regression = LinearRegression()
polynomial_regression.fit(X_poly, y)

y_prediction = polynomial_regression.predict(poly.fit_transform(X))
```

Figure 7 - Polynomial regression fitted to total number of crashes in dataset

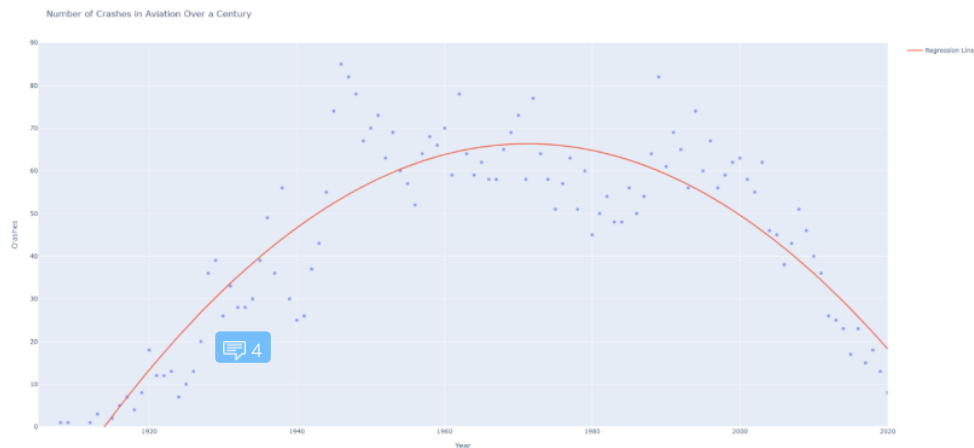


Figure 9 - Code snippet of polynomial regression model

Evaluating the model through RMSE shows a score of 10.46 (see Figure 8 - RMSE score).

```
D:\Dropbox\Python Projects\acca7025-
ssion.py"
RMSE: 10.463081440340389
```

Figure 8 - RMSE score

2.6. Logistic Regression

With pre-processing complete after feature engineering a classifier within the dataset, logistic regression could be performed on the data to classify which entries in the dataset belonged to which operators based on the number of fatalities.

```
# Open and read the training .csv file
data_frame = pandas.read_csv('data\pre-processed\total_commercial_military_operators.csv')

X = data_frame.Fatalities.values.reshape(-1, 1)
y = data_frame.Classification

log_reg = LogisticRegression()
log_reg.fit(X, y)

y_prediction = log_reg.predict(X)
```

Figure 10 - Logistic regression classifying flight operators based on number of fatalities

To evaluate the performance of the model a confusion matrix is created to reveal false positives and negatives and true positives and negatives.

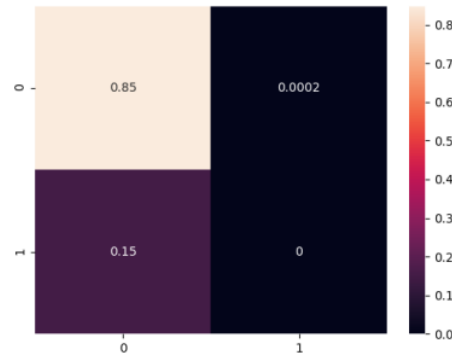


Figure 11 - Confusion matrix

3. References

- [1] S. B. Kotsiantis, P. E. Pintelas, and D. Kanellopoulos, "Data Preprocessing for Supervised Learning," *Int. J. Comput. Inf. Eng.*, vol. 1, no. 12, pp. 4104–4109, 2007.
- [2] Y. Roh, G. Heo, and S. E. Whang, "A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective," *IEEE Trans. Knowl. Data Eng.*, pp. 1–1, Nov. 2019.
- [3] T. K. Grimholt *et al.*, "Flight Anxiety Reported from 1986 to 2015," *Aerosp. Med. Hum. Perform.*, vol. 90, no. 4, pp. 384–388, Apr. 2019.
- [4] European Transport Safety Council (ETSC), "Transport Safety Performance in the EU - a Statistical Overview," 2003.
- [5] I. Savage, "Comparing the fatality risks in United States transportation across modes and over time," *Res. Transp. Econ.*, vol. 43, no. 1, pp. 9–22, Jul. 2013.
- [6] International Civil Aviation Organization, "Safety Report 2020," 2020.
- [7] V. Olsavszky, M. Dosius, C. Vladescu, and J. Benecke, "Time Series Analysis and Forecasting with Automated Machine Learning on a National ICD-10 Database," *Int. J. Environ. Res. Public Health*, vol. 17, no. 14, p. 4979, Jul. 2020.
- [8] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and Machine Learning forecasting methods: Concerns and ways forward," *PLoS One*, vol. 13, no. 3, p. e0194889, Mar. 2018.
- [9] B. Pavlyshenko, "Machine-Learning Models for Sales Time Series Forecasting," *Data*, vol. 4, no. 1, p. 15, Jan. 2019.
- [10] M. Dogucu and M. Çetinkaya-Rundel, "Web Scraping in the Statistics and Data Science Curriculum: Challenges and Opportunities," *J. Stat. Educ.*, pp. 1–11, Aug. 2020.
- [11] M. Uddin, J. Lee, S. Rizvi, and S. Hamada, "Proposing Enhanced Feature Engineering and a

Selection Model for Machine Learning Processes,” *Appl. Sci.*, vol. 8, no. 4, p. 646, Apr. 2018.

- [12] “The immortal DC-3,” *Royal Aeronautical Society*. [Online]. Available: <https://www.aerosociety.com/news/the-immortal-dc-3/>. [Accessed: 10-Nov-2020].
- [13] “Aerospace History Timeline,” *American Institute of Aeronautics and Astronautics*. [Online]. Available: <https://www.aiaa.org/about/History-and-Heritage/History-Timeline>. [Accessed: 11-Nov-2020].
- [14] “The Golden Age of Flight,” *Wall Street Journal*. [Online]. Available: <https://www.wsj.com/articles/SB10001424052748704684604575380992283473182>. [Accessed: 11-Nov-2020].

4. Appendix

4.1. Appendix A – Data Scraping

Scraping the data from the plane crash database involved parsing the source HTML code of the plane crash database website. Various HTML objects were gathered and stored to extract the relevant information from the website.

The home page of the website contained all links for every year of the database, which when accessed revealed all entries of plane crashes for the year being observed. The plane crash database stored all entries within HTML table objects, by parsing the code of the table objects source code exposed links and other relevant information for each entry.

1920

Date	Location / Operator	Aircraft Type / Registration	Fatalities
17 Sep 1908	Fort Myer, Virginia Military - U.S. Army	Wright Flyer III ?	1/2(0)
07 Sep 1909	Juvisy-sur-Orge, France ?	Wright Byplane SC1	1/1(0)
12 Jul 1912	Atlantic City, New Jersey Military - U.S. Navy	Dirigible ?	5/5(0)
06 Aug 1913	Victoria, British Columbia, Canada Private	Curtiss seaplane ?	1/1(0)
09 Sep 1913	Over the North Sea Military - German Navy	Zeppelin L-1 (airship) ?	14/20(0)
17 Oct 1913	Near Johannisthal, Germany Military - German Navy	Zeppelin L-2 (airship) ?	30/30(0)
05 Mar 1915	Tienen, Belgium Military - German Navy	Zeppelin L-8 (airship) ?	21/41(0)
03 Sep 1915	Off Cuxhaven, Germany Military - German Navy	Zeppelin L-10 (airship) ?	19/19(0)
28 Jul 1916	Near Jambol, Bulgaria Military - German Army	Schutte-Lanz S-L-10 (airship) ?	20/20(0)
24 Sep 1916	Billerica, England Military - German Navy	Zeppelin L-32 (airship) ?	22/22(0)

Figure 12 - All entries for the year 1920

Functions within the data scraping script were created for common operations that needed to be completed to scrap data from each year. The process involved selecting a year, scraping all entries of each year for their corresponding links, accessing every link for every entry in a year and further scraping the required data.

4.2. Appendix B – Pre-processing

In order to maintain data for further use, each operation performed when cleaning and pre-processing the data was written to a new .csv file. This was for ease of access as various parts of the scraped data was pre-processed.

The main operations within the pre-processing script are wrapped in 3 functions which defines the scope of the code for each step of pre-processing.

```
# calculate_survival_ratio(pandas.DataFrame)
# Calculates the percentage of survival based on the passengers aboard and amount of fatalities
> def calculate_survival_ratio(row):...

# calculate_death_ratio(pandas.DataFrame)
# Calculates the percentage of deaths based on the passengers aboard and amount of fatalities
> def calculate_death_ratio(row):...

# clean_column(str):
# Cleans all fields in a column by removing unnessecary strings and characters
> def clean_column(column_name):...

# classify_data():
# Pre-processes data by assigning 0 for Military operators and 1 for commercial operators, writes results to a new .csv file
> def classify_data(dataset_directory):...
```

Figure 13 - Pre-processing functions

Various data-frames were created to manipulate and store the data for pre-processing to take place. As all data in each column related to a single instance of data per-row, pre-processing focused on cleaning the data of miscellaneous values alongside extracting the data that would be the features used in each algorithm.

Sub-sets of the scraped data were written to new .csv files resulting stored for later use.

```
▼ data
  ▼ backup
    plane_crashes_dataset_pre-processed.csv
    plane_crashes_dataset_scraped.csv
  ▼ pre-processed
    highest_fatalities_yearly.csv
    plane_crashes_dataset_pre-processed.csv
    total_commercial_military_operators.csv
    total_number_of_crashes_yearly.csv
    plane_crashes_dataset_scraped.csv
```

Figure 14 - Pre-processed sub-sets of original data

A simple file structure within the code project was created for organizing the new sub-sets of pre-processed data (see Figure 14 - Pre-processed sub-sets of original data).

4.3. Appendix C – Exploratory Data Analysis

Exploration of the dataset was conducted to extract various quantities of figures for further insight between the information the dataset offered along with potential cleaning opportunities that may have been missed when inspecting the dataset without visualisation.

Further research has been applied to provide context to the observations made which ultimately led to the selection of machine learning algorithms used. Throughout visualizing the data, the primary goal of selecting features and targets was kept in mind.

4.3.1. Appendix C1 – Visualizing the Data

The first observation focused on which aircraft were involved in the most crashes, this provided clear insight into what aircraft had the most occurrences within the dataset.

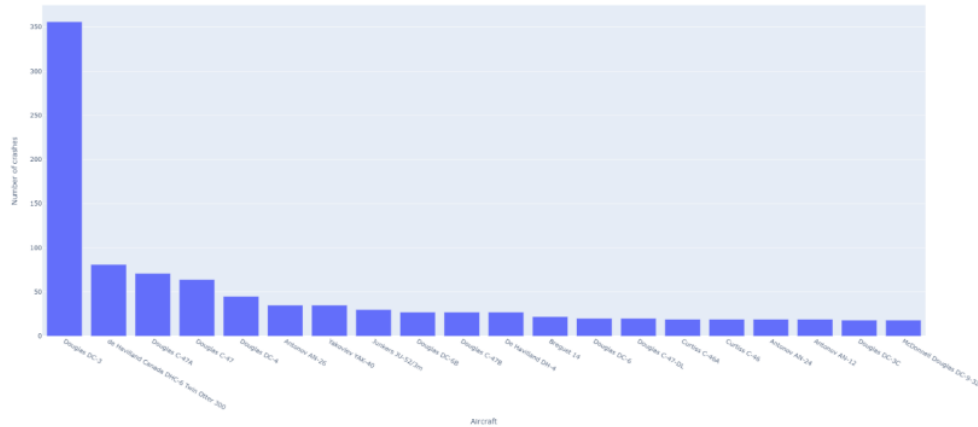


Figure 15 - Top 20 aircraft with highest number of crashes

Reviewing the results in Figure 15 **Error! Reference source not found.** showed that the *Douglas DC-3* sat at first place with over 350 crashes. Further research the era of the aircraft where the *Douglas DC-3* first took flight during the 1930's and was in operation until 2018 [12].

Most aircraft plotted within the graph dated from the early years of aviation. Assumptions can be made on the results shown to conclude that the aircraft with the most occurrences within the dataset would most certainly be involved in a high amount of crashes as aviation was at its infancy.

Inspecting the total number of *fatalities* per year (see Figure 16) provided interesting observations towards the increase of aviation safety.

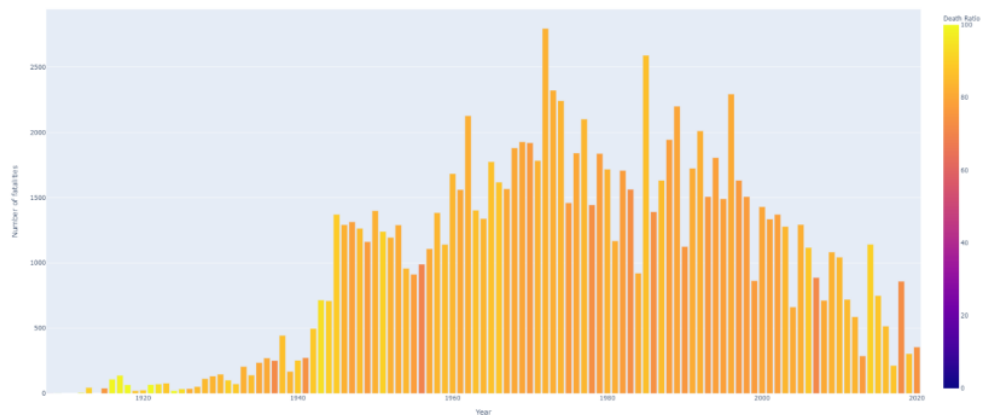


Figure 16 - Total fatalities in aviation over a century, death ratio ranges from 0% (dark blue), to 100% (bright yellow)

The start of aviation can be seen with minimal crashes, in the context of the early days of flight the results are relatable as aviation began in 1903 with the *Wright Brothers* constructing the first self-propelled aircraft [13]. An increase in fatalities can be viewed around the 1940's, which marries well with the birth of commercial flights becoming an available service [14]. Considering that aviation wasn't as advanced then as it is now, the results hold reasonable meaning and formally demonstrate an approximation of the progression aviation has seen over its years of development.

One particular area of interest appears as a slightly sporadic uprising of fatalities that occurred between the 1940's up to the 1990's, this implies that aviation became much more accessible and popular at this time resulting in more flights and unfortunate accidents.

The highest combined amount of fatalities can be seen in 1972, the data shows 2,796 total fatalities from all recorded crashes that year with 1985 following second place with a combined fatality rate of 2,590 and 1996 with a total of 2,294 fatalities.

With the above results showing a high fatality rate between the previously mentioned dates, this may lead to the slightly nuanced opinion of assuming most crashes potentially happened between the proposed dates of 1972 to 1996.

To accompany this observation, further investigation was applied to visualize the total amount of passengers each year according to the data (see Figure 17). The total passengers per year is the *accumulation of all passengers aboard each flight*. Due to the data only offering flights which resulted in a crash, observations on the results of total passengers didn't provide a true measure of total passengers travelling by air throughout the entire time period.

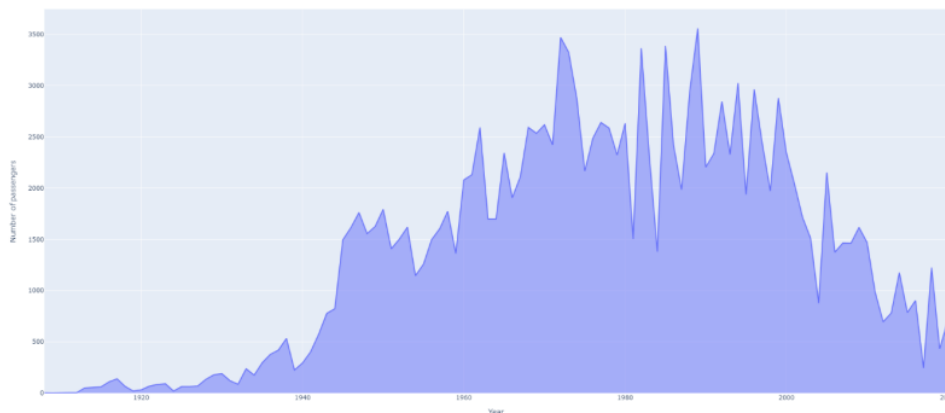


Figure 17 - Total passengers per year

With this in mind, the total amount of passengers per year still provides insight into how popular flying became and can also be claimed as one of the potential driving factors of why a high amount of crashes can be observed between the above-mentioned dates.

It can be determined that flying has become safer over the years as a clear decline in fatalities can be seen in Figure 16 showing a fairly linear falloff in the amount of total fatalities each year.

This trend is also reflected in the total amount of crashes according to the data.

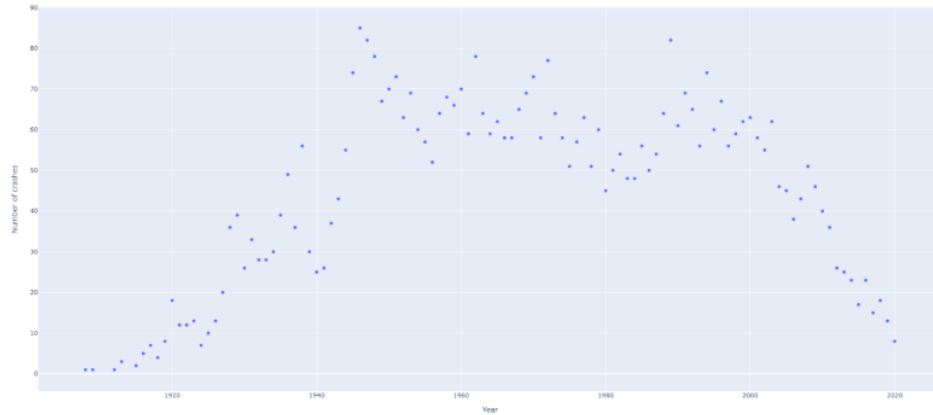


Figure 18 - Total amount of crashes per year

A rapid decline from 1996 onwards shows a much lower amount of crashes. It was realised that with a larger scope of data further investigation could have been conducted to provide a clear contrast between flights that ended in a crash and flights that didn't. Having this data available may have provided a more accurate scale when visualizing the data of all crashes, leaving fewer potential outliers as well as providing extra context towards aviation accidents.

It can be deduced that around 1996 aviation saw a clear turning-point in safety, this could be correlated with the advancement of communications and flight assist technology that can be found onboard most modern aircraft.

Report

GRADEMARK REPORT

FINAL GRADE

GENERAL COMMENTS

Instructor

54/100

PAGE 1

PAGE 2

PAGE 3

PAGE 4



Comment 1

Summarize results here.

PAGE 5



Comment 2

Be careful using subjective terms.

PAGE 6



Comment 3

Good work. Tedious, I know.

PAGE 7

PAGE 8

PAGE 9

PAGE 10



Comment 4

Explain the meaning - any units here?

GRADING FORM: SMU GENERIC_GJ

DAVID ROBIN JAMES

53.50

RESEARCH-INFORMED LITERATURE



Refs good and well used. Key topics identified and reviewed.

6.5

KNOWLEDGE AND UNDERSTANDING



A good understanding of essential concepts demonstrated. Good data scraping and pre-processing in evidence. Only one model covered in detail (polynomial regression).

11

ANALYSIS



Some very good work done here to create the dataset. Well done. Some nice relationships

18

extracted and a suitable polynomial regression fitted. Good. You produce a confusion matrix but don't attempt to discuss it ? These numbers need further investigation, I think. You mention KNN but do not go further with it? Report ends abruptly. No results/conclusion sections.

PRACTICAL APPLICATION AND DEP.

 Excellent data creation. Scraping, etc. Well done. Further pre-processing in evidence. Only polynomial regression attempted. You mention KNN but do not go further with it? Confusion matrix generated.

15

SKILLS FOR PROFESSIONAL PRAC.

 Structured. No conclusions or results. Refs well used. Diagrams helpful.

3