INVESTIGATION OF AVIATION ACCIDENTS

(WWW.AVIATIONCV.COM, 2016)

28-Apr-2019

Data Exploration Project

Name: Arunava Munshi	Organization: Monash University

TABLE OF CONTENTS

INTRODUCTION

Project Description

The project aims to explore NTSB Data for Aviation Accidents and tries to answer few questions related to the dataset through visualization. The dataset sourced from ntsb.gov contains potential aviation accidents through 2017. The data set has a long list of different data files consisting of aviation accident data of different time. **So, it is assumed that the timeframe in the answers for different questions may not match with each other.** The whole project is done in R and the snippets for R codes are provided wherever required. The questions for this project are mentioned below.

Project Questions

This project aims to answer the below questions –

Q1: What are the causes of aviation accidents? Is there any visible change in causes over time?

The question seeks to answer the main causes of air accidents. The idea is also to dig down any additional reason for accidents behind the main reasons. The project will explore the potential changes in causes of air accidents over time and try to look for any visible changes in causes over the years due to the advancement in technology and piloting skills.

Q2: Trend in accidents with high casualty rates over time?

The question seeks to track down the trends in accidents with high casualty rates over time. High casualty in aviation accidents particularly traces the accidents with medium or high fatality rates, so the main concentration here will be into the fatal accidents over the years.

Q3: What are major accident prone zones?

This data contains the accident information occurred in different locations in USA. So the project will aim to find out notorious zones in USA accounting for fatal accidents. The question will also answer if there are any visible changes (increase or decrease) in number of accidents in these zones over time.

Data Files

Among all data files, the below files are important and considered for this project.

File Name	Brief Description	Size
aircraft.txt	Have the details of the air transports (aircrafts, helicopters, balloons etc.) involved in the accident.	80982 rows x 93 columns
eADMSPUB_DataDictionary.txt	Have the descriptions of all the columns of all other data files.	4574 rows x 13 columns
Findings.txt	Have data related to NTSB findings on different causes of air accidents.	37315 rows x 13 columns
events.txt	Have the causes of the events of air accidents.	79805 rows x 71 columns
injury.txt	Have the casualty information of	450174 rows x 7 columns

	the accidents.	
dt_events.txt	Also have the causes of the events of air accidents.	327706 rows x 5 columns
Occurrences.txt	Have occurrence of events during the flight phase.	140334 rows x 8 columns
states.txt	Have the names and zip codes of the states of USA.	51 rows x 3 columns

DATA WRANGLING

The data wrangling has been carried out in respect to each question i.e data sets are specifically created and wrangled for every question.

Data Wrangling for Question 1

In order to answer this question, we need to understand the relevant dataset having related information about air accident causes. Accident causes are listed in **Findings.txt** file and the events dates are listed in **events.txt** file. Since we are trying to explore the causes of accidents and trends in particular cause over time, we need information from the above two files. We also need to read the **aircraft.txt** only to filter the data for events related to air accidents, not anything else. With this knowledge, we perform below data wrangling steps to prepare data for exploratory visualization.

a) First we read the data the two files **events.txt** and **Findings.txt**. Read the data from **aircraft.txt** and filter the data only for type **'AIR'** that stands for aircraft. The **events.txt** file is filtered for event type **'ACC'** that stands for accidents, which we are only interested in.

```
# Reading events.txt
events_df <- read.delim("events.txt",header = TRUE, sep = ",", dec = ".")
events_df <- events_df[events_df$ev_type == 'ACC', ]
# Findings.txt
Findings_df <- read.delim("Findings.txt",header = TRUE, sep = ",", dec = ".")
# Reading aircraft.txt
aircraft_data_df <- read.delim("aircraft.txt",header = TRUE, sep = ",", dec =
".")
# Filtering data for Aircraft category
aircraft_df <- aircraft_data_df[aircraft_data_df$acft_category == 'AIR ', ]
```

b) Then we extract the accident dates and months required for our analysis from **events.txt** file and accident cause description from **Findings.txt** file.

```
Getting only years and months columns from events.txt
events_reqd_df <- subset(events_df, select = c(ev_id, ev_year, ev_month))
# Getting findings description from Findings.txt
Findings_reqd_df <- subset(Findings_df, select = c(ev_id,
finding_description))
```

c) After looking into data into the description of the findings we can see the descriptions including different levels are clubbed together. For Example. An accident occurred for Environmental issue including bad weather condition leading to thunder storm is clubbed into one string separated with a '-'. So, we need to separate the main cause from the sub-causes of different level.

	ev_id	finding_description
1	20080109X00036	Environmental issues-Conditions/weather/phenomena-Win...
2	20080109X00036	Aircraft-Aircraft oper/perf/capability-Performance/control p...
3	20080116X00063	Aircraft-Aircraft structures-Doors-Cargo/baggage doors-Inc...
4	20080116X00063	Personnel issues-Task performance-Maintenance-Repair-M...
5	20080116X00063	Organizational issues-Support/oversight/monitoring-Trainin...
6	20080116X00063	Organizational issues-Management-Policy/procedure-Availa...

d) After separating them on separator '-', the data looks like below. We also need to take care of any trailing spaces after this separation. For the purpose of our exploration, we will only deal with first and second level of causes, so we select first two reason columns.

```
# Separating levels of air accident causes
library(dplyr)
library(tidyr)
Findings_reqd_df <- Findings_reqd_df %>%
separate(finding_description,        c("main_reason",        "detailed_reason",
"sub_reason_2", "sub_reason_3", "sub_reason_4", "sub_reason_5"), "-")
# Selection only mainreason and deailed reason columns
Findings_reqd_df  <-  subset(Findings_reqd_df,  select  =  c(ev_id,  main_reason,
detailed_reason))
# Removing trailing spaces
Findings_reqd_df$main_reason  <-  trimws(Findings_reqd_df$main_reason,  which  =
c("both", "left", "right"))
Findings_reqd_df$detailed_reason  <-  trimws(Findings_reqd_df$detailed_reason,
which = c("both", "left", "right"))
```

	ev_id	main_reason	detailed_reason
1	20080109X00036	Environmental issues	Conditions/weather/phenomena
2	20080109X00036	Aircraft	Aircraft oper/perf/capability
3	20080116X00063	Aircraft	Aircraft structures
4	20080116X00063	Personnel issues	Task performance
5	20080116X00063	Organizational issues	Support/oversight/monitoring
6	20080116X00063	Organizational issues	Management

e) Now, once the data are ready in the individual dataframes, they are combined into one dataframe in order to make them ready for visualization.

```
# Join Findings_reqd_df, events_reqd_df and aircraft_df
accident_causes_df <- merge(Findings_reqd_df, events_reqd_df, by = "ev_id")
accident_causes_final <- merge(accident_causes_df, aircraft_df, by = "ev_id")
accident_causes_final  <-  subset(accident_causes_final,  select  =  c(ev_id,
main_reason, detailed_reason, ev_year, ev_month))
```

	ev_id	main_reason	detailed_reason	ev_year	ev_month
1	20080109X00036	Environmental issues	Conditions/weather/phenomena	2008	1
2	20080109X00036	Aircraft	Aircraft oper/perf/capability	2008	1
3	20080116X00063	Aircraft	Aircraft oper/perf/capability	2008	1
4	20080116X00063	Organizational issues	Support/oversight/monitoring	2008	1
5	20080116X00063	Aircraft	Aircraft structures	2008	1
6	20080116X00063	Organizational issues	Management	2008	1

Data Wrangling for Question 2

In order to answer this question, we need to understand the relevant dataset having related information about air accident fatality rates. Accident casualty related information is listed in **injury.txt** file and the event dates are listed in **events.txt** file. However, the **fatality rate** for each event **is** not given in the dataset, which needs to be derived from diving **Number of Fatalities** by the **Total Number of Injuries** for each event. To do this, we follow the below steps –

a) We need to consider the **event id, injury level** (signifies if an injury is fatal, significant or minor; we consider only **fatal** and **total** from this column) and **count of injured persons** from **injury.txt** file. Then we divide **Number of Fatalities** by the **Total Number of Injuries** for each event to get the **fatality rate.**

```
#Selecting only event id, jnjury level and number of persons injured
injury_reqd_df   <-   subset(injury_df,   select   =   c(ev_id,   injury_level,
inj_person_count))
library(sqldf)
# Getting only Fatal Injuries into injury_high_df
injury_fatal_df_1 <- aircraft_envt_seq_df <- sqldf("
     SELECT d1.ev_id, d1.injury_level, d1.inj_person_count
     FROM injury_reqd_df d1
     WHERE d1.injury_level LIKE '%FATL%'
")
# Getting the sum of total high injuries
injury_fatal_df <- aircraft_envt_seq_df <- sqldf("
     SELECT d1.ev_id,  SUM(d1.inj_person_count) AS 'total_fatal_injury'
     FROM injury_fatal_df_1 d1
     GROUP BY d1.ev_id
")
# Getting only Total Injuries into injury_high_df
injury_totl_df_1 <- aircraft_envt_seq_df <- sqldf("
      SELECT d1.ev_id, d1.injury_level, d1.inj_person_count
      FROM injury_reqd_df d1
      WHERE d1.injury_level LIKE '%TOTL%'
")
# Getting the sum of total high injuries
injury_totl_df <- aircraft_envt_seq_df <- sqldf("
      SELECT d1.ev_id,  SUM(d1.inj_person_count) AS 'total_injury'
      FROM injury_totl_df_1 d1
       GROUP BY d1.ev_id
")
```

```
#Joining injury_fatal_df and injury_totl_df on event id and finding fatality
rate
fatality_rate_df <- merge(injury_fatal_df, injury_totl_df, by = "ev_id")
fatality_rate_df$fatality_rate <-
round((fatality_rate_df$total_fatal_injury/fatality_rate_df$total_injury)    *
100, digits = 2)
fatality_rate_df<- subset(fatality_rate_df, select = c(ev_id, fatality_rate))
```

b) Once the fatality rates are finalized **fatality_rate_df, events_reqd_df** and **aircraft_df** are joined on event ids in order to get the years and months of fatal accidents of aircraft type. The process is same as discussed previously. **The note to be taken here is that the data is available for fatality rates since 1999.**

	ev_id	fatality_rate	ev_year	ev_month
1	20001204X00006	100.00	1999	1
2	20001204X00007	100.00	1999	1
3	20001204X00008	50.00	1999	1
4	20001204X00016	100.00	1999	1
5	20001204X00017	66.67	1999	1
6	20001204X00018	100.00	1999	1

Data Wrangling for Question 3

In order to answer this question, we need to understand the relevant dataset having related information about the accident prone zone. The dataset **events.txt** has the location information about the accidents, so we extract required data from the same file for event type accident. Here in this dataset, all the accident information has been recorded for **USA**. We select the **event identifier, years of events, months of events, city, state and zipcode of events and geocodes such as latitude and longitude** information from event file. The selection process is same as discussed before, so it has not been shown. Also we noted that state information is in codes such as AK, AR, AZ etc, which can be transformed into the name of the states by joining with **states.txt** file and get the names of the states. This step is also similar to what we did previously and hence, is not shown here.

DATA CHECKING

The data checking is done on top of data wrangling in order to check if there is any data error or not and to fix the error if possible and required for the purpose of exploratory visualization.

Data Checking for Question 1

During a check on wrangled data, we found that the column **main_reason** of **accident_causes_final** dataframe has a value '**Not determined**' that implies that there are no recorded reasons for these accidents. So, for the purpose of this visualization, we discarded the rows of **accident_causes_final** dataframe with the '**Not determined**' value in the **main reason** column.

```
# Remove the rows with causes not determined
accident_causes_final <- accident_causes_final[accident_causes_final$main_reason !=
```

```
'Not determined', ]
```

```
> unique(accident_causes_final$main_reason)
[1] "Environmental issues"  "Aircraft"              "Organizational issues" "Personnel issues"
>
```

Data Checking for Question 2

During a check on wrangled data, we found that the column **fatality rate** for **fatality_rate_df** dataframe has huge number of null values. In order to correct this data we need to follow some sophisticated statistical technique to impute the null values with best values that it would have held in reality. Because fatality rate is a continuous data within the range [0-100], we follow **Linear Regression** to impute the null values.

	ev_id	fatality_rate	ev_year	ev_month
9962	20050208X00153	*NA*	2005	2
10199	20050921X01501	*NA*	2005	8
10505	20060808X01115	*NA*	2006	7
10770	20070520X00598	*NA*	2007	4
10783	20070601X00676	*NA*	2007	5
10796	20070616X00749	*NA*	2007	6

Selection of Predictors

In order to select the potential predictors for linear regression to predict the fatality rate we need to consider two things. 1. **The aircraft type and its structure or model** and 2. **The situation/circumstances of the aircraft during accident** (Exp. Ground collision, involved in storm, water landing etc.). The aircraft structure data can be found from **aircraft.txt**. The situation related data can be found from the datasets **dt_events.txt, Occurrences.txt** and **eADMSPUB_DataDictionary.txt**. The wrangling of this part is same as done previously, so, we are not getting this anymore, but showing the end result of the wrangling. After careful consideration, we choose the below potential predictors that might have impacted the fatality rate.

Occurrence_Code	code_iaids	damage	acft_make	acft_model	acft_series	acft_serial_no	cert_max_gr_wt	fatality_rate
230	IC	DEST	Beech	300	300	FA-70	14100	100
230	IPW	DEST	Beech	300	300	FA-70	14100	100
230	SQAL	DEST	Beech	300	300	FA-70	14100	100
230	L	DEST	Beech	300	300	FA-70	14100	100
230	WIND	DEST	Beech	300	300	FA-70	14100	100
230	WHIR	DEST	Beech	300	300	FA-70	14100	100

Now because we are doing linear regression, we need to transform the string type data into numeric type. For that, we do the below transformation. We do it like below for column **code_iaids** and for the rest we follow the same technique and after transformation, we have the below dataframe.

```
fatality_predict_df_final$code_iaids<-
as.numeric(fatality_predict_df_final$code_iaids)
```

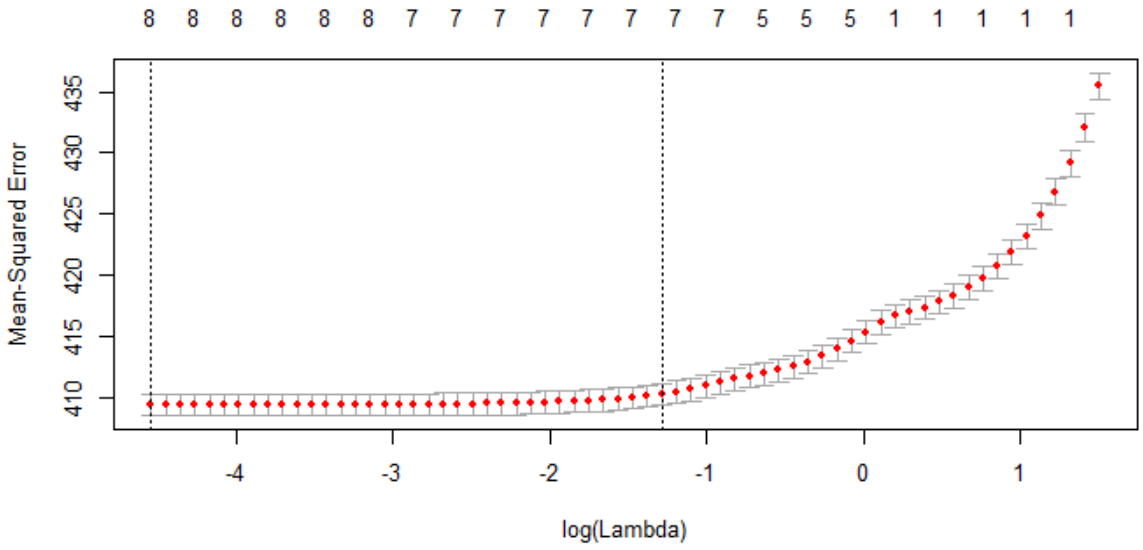
Occurrence_Code	code_iaids	damage	acft_make	acft_model	acft_series	acft_serial_no	cert_max_gr_wt	fatality_rate
230	18	2	742	1120	822	59622	14100	100
230	21	2	742	1120	822	59622	14100	100
230	40	2	742	1120	822	59622	14100	100
230	22	2	742	1120	822	59622	14100	100
230	49	2	742	1120	822	59622	14100	100
230	48	2	742	1120	822	59622	14100	100

Now we check for the data errors in the predictors and found that column **cert_max_gr_wt** has null values tha t we impute with column mean using the statistical mean imputation technique.

```
fatality_predict_df_final[is.na(fatality_predict_df_final[,9]),          9]          <-
mean(fatality_predict_df_final[,9], na.rm = TRUE)
```

Once we are done with mean imputation and any other data correction, we then go for selecting the best subset of predictors with one of the superior statistical technique, also called shrinkage method, **Lasso**. But before that, we need to separate the train set on which we will train our model. We do the following.

```
# Selecting Training and Test dataset
Training_df                                                                    <-
fatality_predict_df_final[!is.na(fatality_predict_df_final$fatality_rate),]
Test_df                                                                        <-
fatality_predict_df_final[is.na(fatality_predict_df_final$fatality_rate),]
#Applying Lasso on Training_df
library(glmnet)
xmat <- model.matrix(fatality_rate ~ ., data = subset(Training_df, select = -
c(ev_id)))[, -1]
lasso_plot <- cv.glmnet(xmat, Training_df$fatality_rate, alpha = 1)
plot(lasso_plot)
```



So Lasso selects 7 out of 8 predictors and they are as follows. So lasso rejected **code_iaids** and the rest are selected.

```
#Name of the best predictors
Best_lamda <- lasso_plot$lambda.1se
Lasso_fit <- glmnet(xmat, Training_df$fatality_rate, alpha = 1)
predict(Lasso_fit, s = Best_lamda, type = "coefficients")[1:7, ]
```

```
  (Intercept) Occurrence_Code      code_iaids          damage       acft_make       acft_model
1.025849e+02   -1.731838e-02    0.000000e+00   -3.813733e+00    5.787127e-05    2.369919e-04
 acft_series
-2.179190e-04
```

Model Building and Predict the Fatality Rate on Test Data

The next step is quite straight forward. Now we build the linear model on Training Data and do the prediction on Test Data. After the prediction is done, newly predicted Fatality Rates are imputed into the Null values. The imputation part is quire straight forward, so is not shown here. The linear model and prediction technique is shown below.

```
#Building Linear Model and Doing the Prediction
model <- lm(fatality_rate ~ ., data = subset(Training_df, select = -c(ev_id, code_iaids)))
predicted_fatality_rate <- predict(model, Test_df)
```

Data Checking for Question 2

During a check on wrangled data, it has been found that the geocodes such as latitude and longitude have null values.

	ev_id	ev_year	ev_month	ev_time	ev_city	ev_state	ev_site_zipcode	latitude	longitude
1	20001204X00000	1999	1	1330	QUINHAGAK	ALASKA	99655		
2	20001204X00002	1999	1	1425	ANCHORAGE	ALASKA	99502		
3	20001204X00003	1999	1	1351	BETHEL	ALASKA	99559		
4	20001204X00004	1999	1	1050	CHEVAK	ALASKA	99563		
5	20001204X00005	1999	1	1435	ANCHORAGE	ALASKA	99502		
6	20001204X00006	1999	1	918	CULLMAN	ALABAMA	35057		

In order to correct this we take the help of library **zipcode** that has full information about the geocodes for all zip codes in USA.

```
library(zipcode)
data(zipcode)
View(zipcode)
```

	zip	city	state	latitude	longitude
1	00210	Portsmouth	NH	43.00590	-71.01320
2	00211	Portsmouth	NH	43.00590	-71.01320
3	00212	Portsmouth	NH	43.00590	-71.01320
4	00213	Portsmouth	NH	43.00590	-71.01320
5	00214	Portsmouth	NH	43.00590	-71.01320
6	00215	Portsmouth	NH	43.00590	-71.01320

Now we join these two dataframes on their zip codes, similar to what we did before, to resolve this issue with null values of geocodes.

	ev_id	ev_year	ev_month	ev_time	ev_city	ev_state	ev_zip	latitude	longitude
1	20001204X00018	1999	1	1212	Salem	MISSOURI	65560	37.63090	-91.51423
2	20001204X00031	1999	1	2156	Rockford	ILLINOIS	61109	42.21344	-89.05595
3	20001205X00179	1999	2	1306	Van Nuys	CALIFORNIA	91406	34.20149	-118.49376
4	20001205X00245	1999	2	1441	Joshua Tree	CALIFORNIA	92252	34.17593	-116.29137
5	20001205X00486	1999	4	1931	Warner Springs	CALIFORNIA	92086	33.33715	-116.69355
6	20001205X00490	1999	4	2130	Trona	CALIFORNIA	93562	35.76443	-117.38202

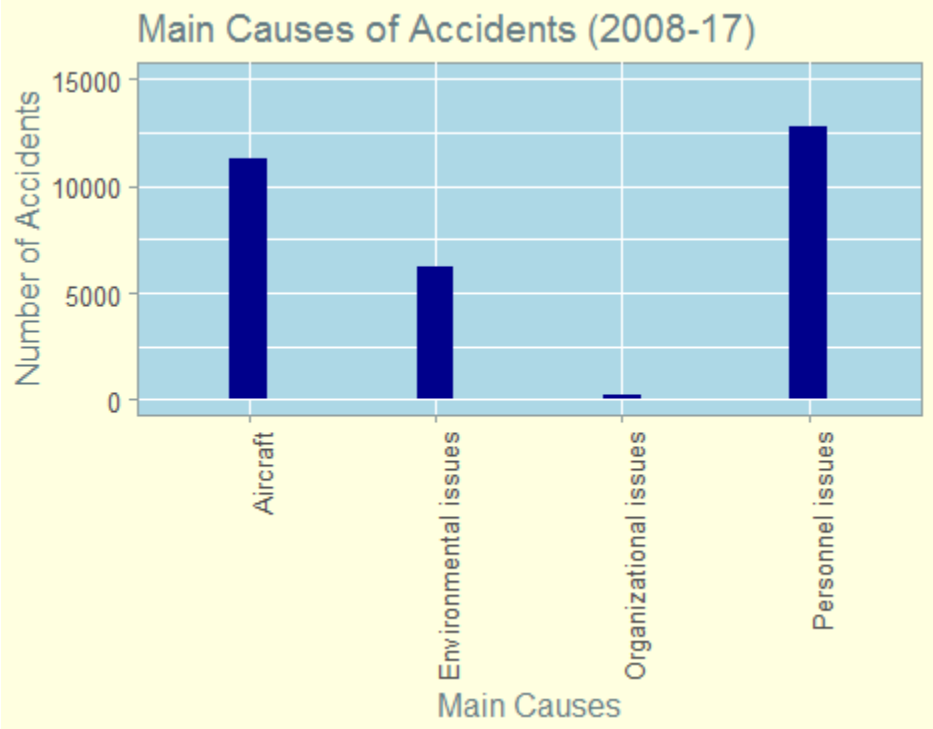
DATA EXPLORATION

Data Exploration aims to explore the corrected and wrangled data in order to visualize the trend or patterns in data and to generate the insights. Here, as part of this project, the data exploration is done through R ggplot2 visualization.

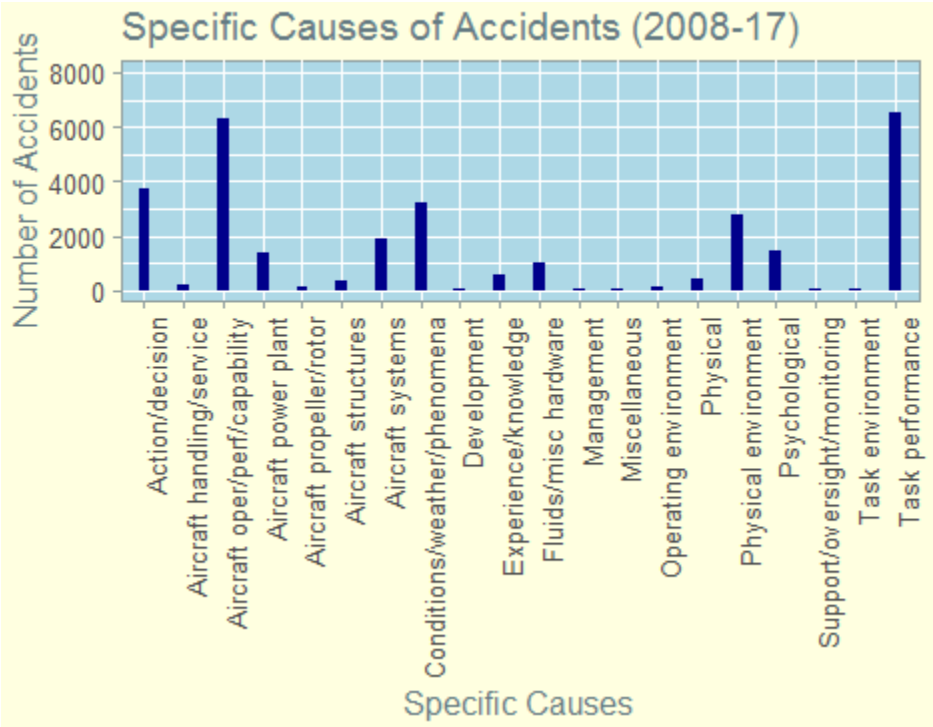
Q1: What are the causes of aviation accidents? Is there any visible change in causes over time?

To answer the above question we have explored the data in the following ways –

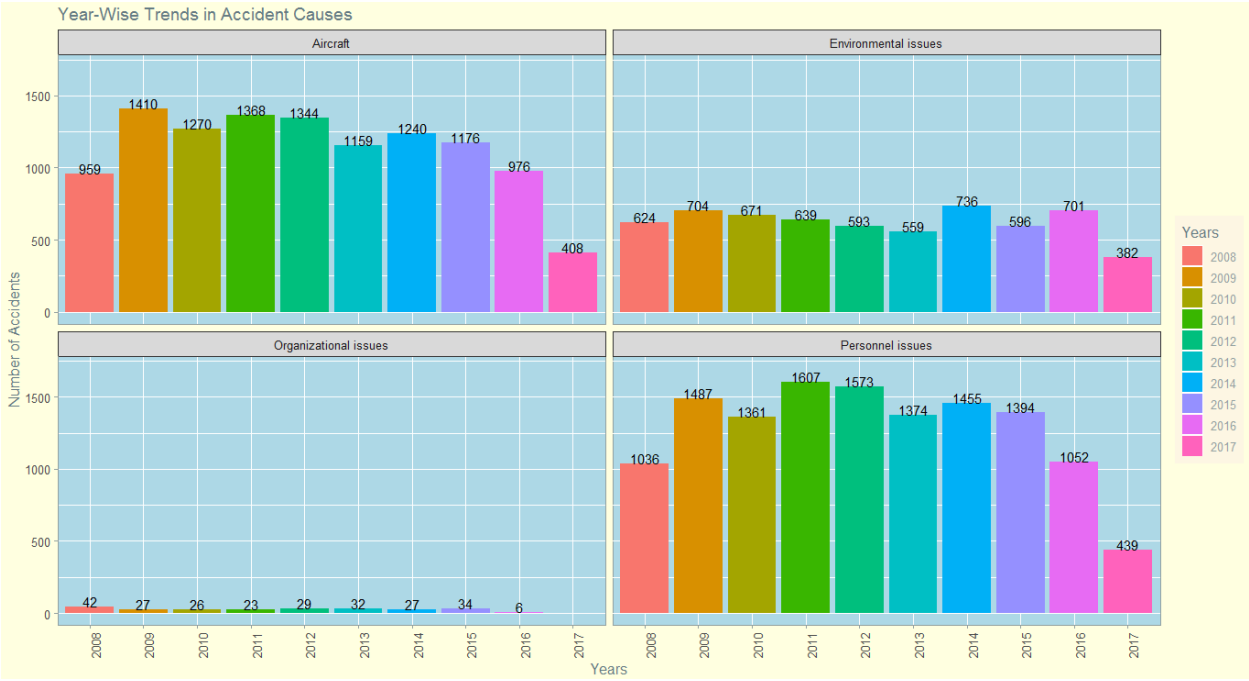
a) Trying to find the main causes, we explored four main reasons why aviation accidents took place. These are issues related to '**Aircraft**', '**Environment**', '**Organization**', and '**Personnel/Crew**'. It can be seen from the below bar chart that fault from crew members accounted for highest number of accidents within 2008 to 2017, whereas organizational issues accounted for the lowest ever number of accidents. Aircraft related issue also stood significant for the accidents in given period.



b) Going deeper, we have found even more specific causes to influence the accidents during the same period. From the below bar chart we can see that highest number accidents occurred because crew members could not perform their tasks efficiently, confirming the personnel issues in the previous chart. Aircraft operations/performance/capability accounted second most important reason in this case, where wrong decision from crew members held third responsible for most number of accidents. We saw that Aircraft maintenance, aircraft structures, organizational management or development, Aircraft support etc. became insignificant for the accidents.



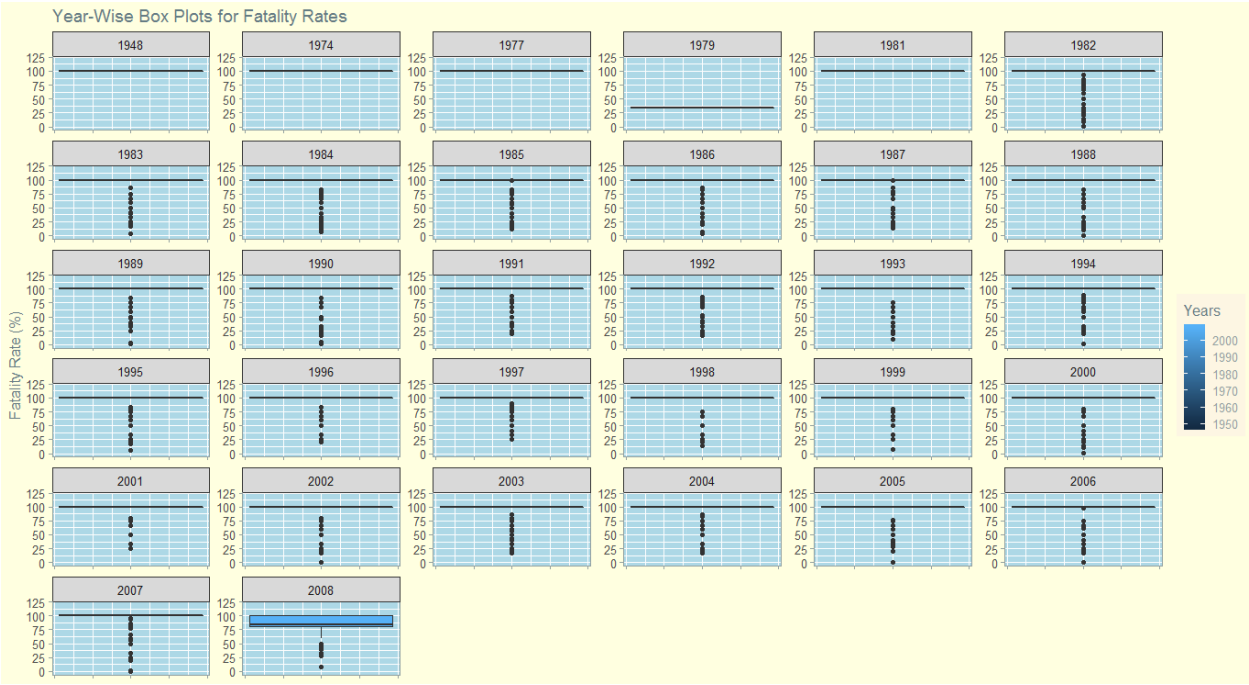
c) Now we will see the trends in accidents related to above said four causes over the years. The below pair bar plot coveys the required information. We can see that though aircraft related issue stood a big concern for aviation experts, after 2015, it the concern decreased significantly with lower number of accidents falling under this category, possibly because of some significant technological improvement. The same trend is found for crew related issues, indicating that there may be a significant training or operational improvements between 2015 and 2017. However, Environmental issues, though not too significant, stood a concern for the experts, possibly because no improvements in this area could be achieved other than just a lower number of incidents under this category in 2017. Experts could possibly be relieved with the fact that no organizational level transformation is required to prevent air accidents.



Q2: Trend in accidents with high casualty rates over time?

To answer the above question we have explored the data in the following ways –

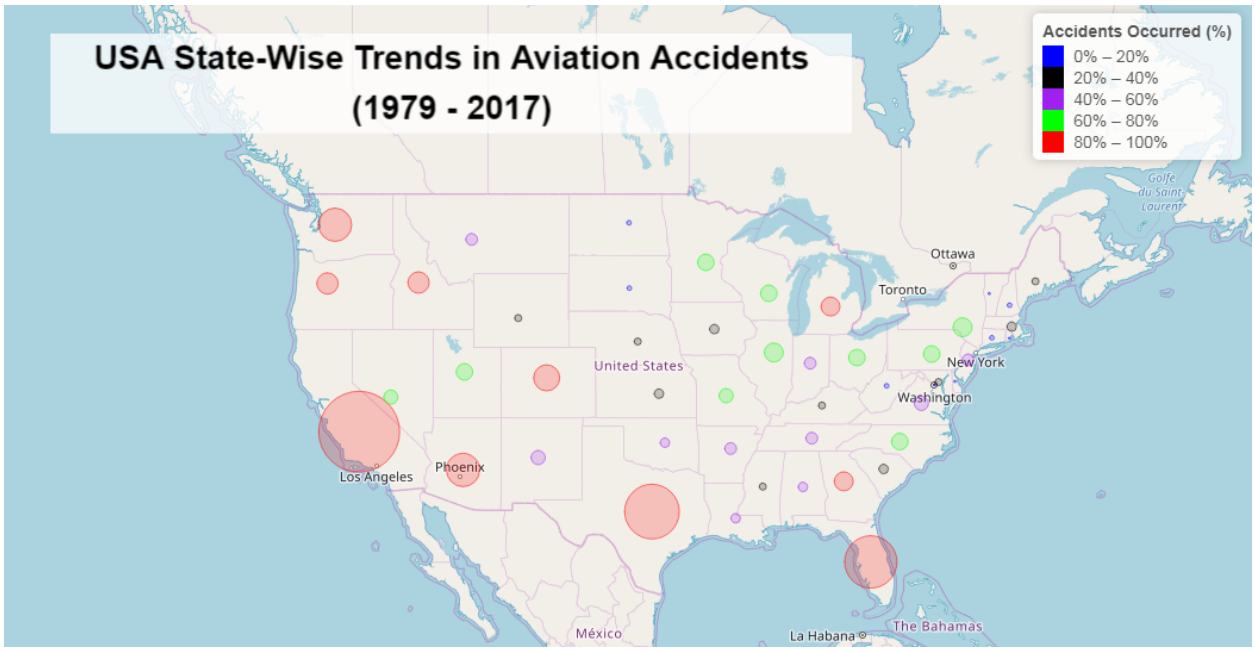
Here we have the data since 1948 to 2008 and we did a year-wise set of boxplots on the fatality rate through y axis for the same time period. We can see all accidents between 1948 and 1981 were completely fatal (100% death rate) except for 1979 in which the fatality rate was surprisingly low, around 30%. We looked into the data and saw that only one recorded fatal accident was there in the year 1979. Between 1982 and 2007, we saw some accidents with reduced fatality rates with some recoded minimum fatality rates (near 0%) in the years 1986, 1988, 1989, 1990, 1994, 1995, 2000, 2006 etc. However, compared to number of complete fatal accidents during the same period, these numbers are significantly low because these low fatal accidents came as outliers on the box plots, and hence normally discarded in statistical analysis. But, surprisingly year 2008 has shown some ray of hopes as the median fatality rate of the year has come down to around 80%, indicating some revolutionary change happened in aviation technology in that year.



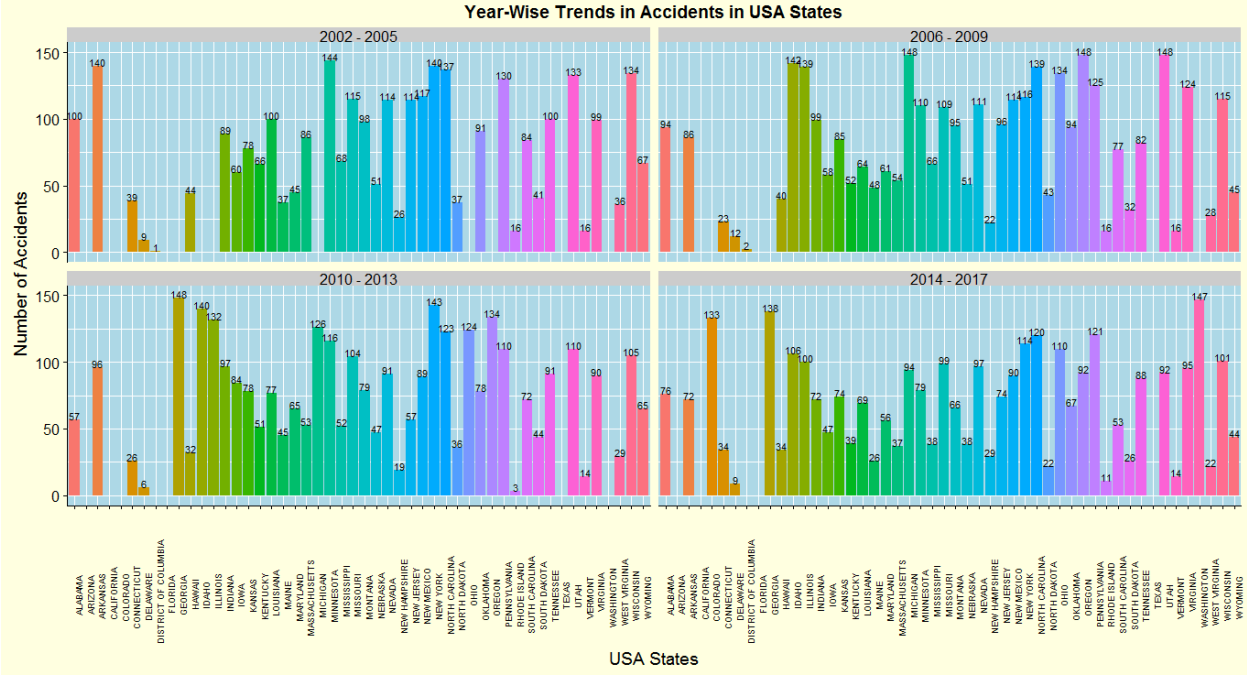
Q3: What are major accident prone zones?

To answer the above question we have explored the data in the following ways –

a) Here we have the geocoding information for the accidents took place in USA. We tried to divide the accident zones on US states and tried to visualize which US states accounted for notorious air crash zones. We divided these zones depending on the range of accident rates and inside each zone we also compared the accident rates with the radius of the circles. The higher is the radius, the more is the accident rates. We can see the states such as California, Florida, and Washington etc. are notorious accident zones with over 80% accident rates. But these notorious zones have **high variance** in accident rate. For example, even though both California and Georgia had more than 80% accident rate, California stood highest with significantly more accidents than those occurred in Georgia, which stands lowest in 80%-100% range. The zones with accident rates between 80% and 60% did not have high variance of accident rates among themselves. The zones with 40%-60% accident rates again encountered high variance. We can also see that overall Eastern states have lower accident rates in general than the Western states, with Vermont being the state with lowest ever accident rate.



b) We have further plotted the data of the number of accidents occurring in 53 states of USA, grouping 4 years together starting from 2002 to 2017. It can be observed that the number of accidents in the states such as Arkansas, Nebraska, Massachusetts, and Wisconsin etc. gradually decreased over the time period whereas it can be observed that the states such as Colorado, Georgia, and Washington etc. have seen a dramatic increase in the number of accidents. There is not much change in the number of accidents in states such as Vermont and Delaware. States such as Alabama, Kansas, Kentucky, Wyoming etc. have also shown fluctuation in the accident counts over time.



CONCLUSION

Finally after all these we can safely make the following conclusions –

a) 2008 has been the significant year in the history of aviation because of the significant rate of decrease in fatality rates in air accidents took place in that year, indicating some significant transformation in aviation technology.
b) Year 2017 is significant with respect to its previous years because of the steep decrease in the number of accidents happened in that year, allowing to believe huge improvement in the human controlled factors including aircraft technology, pilot training, etc. affecting the number accidents.
c) There is an overall reduction in number of accidents over the years, implying that aviation technology is gradually improving to mitigate the main causes and factors affecting the air accidents and we can hope for a day, when technology will be too sophisticated to handle any kind of unprecedented situation.

REFLECTION

This project helped me to learn visualization with R using **ggplot2** package. It also helped me learn how to do exploratory data analysis on random and complex dataset and do necessary data wrangling in R. I now feel confident to progress with an unknown data in order to generate insights through data visualization and conclusions form it. This project also helped me briefly touch upon Machine Learning, giving a good reason to further explore it in detail.

BIBLIOGRAPHY

Clark, C. (2017). *NTSB Aviation Accidents Through 2017*. Retrieved 2019, from www.kaggle.com: https://www.kaggle.com/cwclark/ntsb-aviation-accidents-through-2017

www.aviationcv.com. (2016, Jan 25). *Deadliest Accident in Aviation History*. Retrieved Apr 28, 2019, from www.aviationcv.com: https://www.aviationcv.com/aviation-blog/2016/deadliest-accident-in-aviation-history