MORTALITY RATES IN PLANE DISASTERS: A CASE STUDY OF HISTORIC PLANE CRASHES BETWEEN 1990 AND 2008

Dhruvisha Gosai (JC810547) – James Cook University

Executive Summary

Aviation industry has undergone significant growth over the last decade with low operating costs and fuel-efficient aircrafts (Grupo One Air, 2020). However, that has led to a growth in the number of air disasters regardless of proper safety and regulations in place. Using worldwide safety data for 2000-07, a researcher calculated that passengers who fly in developing countries faced 13 times higher risk of being killed in an air crash than that in developed countries (Institute for Operations Research and the Management Sciences, 2020). This creates a potential to explore the association of the Human Development Index (HDI) to the mortality rates from plane disasters and the year of plane calamity.

International Disaster Database (EM-DAT, 2009) provides the air crash data with the number of affected and mortality rates by individual country for each year. When combined with each country's HDI for the disaster year - attained from UN Development Programme (Human Development Data Center, 2009), creates a basis to explore three objectives:

1. Whether the number of deaths for developing countries and developed countries based on their HDI are equal.
2. If the chance of survival depends on the year of crash or not.
3. If survival of anyone on the plane can be predicted by the year of disaster and HDI score of that country.

Multiple statistical tests and regression analysis are used for hypotheses testing which provides relationship indicators of number of deaths from plane crashes. Results indicate that there is no difference in the mean number of deaths in developing countries with HDI band 2 and developed countries with HDI band 3. Nonetheless, HDI band and the year of crash are dependent on each other for the mean number of deaths. This analysis can then be used as basis to create additional funding by international aviation agencies for countries with lower HDI.

Further research is required to determine how the additional factors like total number of flights, how old is the plane, origin country of the flight, reason of crash, etc. impact on predicting the number of crashes and deaths as the result.

Introduction

Since the advancements in technology, control systems in aviation transitioned to digital services in 1985 (Gunston, 1990), hence civil aviation saw a continued expansion ever since. However, with increasing amount of air travel, there posed a risk of air disasters. In 1985, the world witnessed the worst month of passenger and crew deaths with 4 separate accidents and 720 lives lost on a commercial aircraft in a single month (Kelly, 2015) out of a total of 1,497 deaths that year.

As the years progressed, the Federal Aviation Administration (F.A.A) held a summit in 1995 to re-evaluate the safety measures and regulatory guidelines and strive to have zero accidents by introduction of computerized data collection for analysts to spot abnormalities in speed, climbing and descent rate; signs that stance as a precursor to a problem (Wald, 1995). Despite the advancements, in 1996, there were 18 aviation accidents resulting in death of 2,650, 114% higher than the year before as indicated in Figure 1.

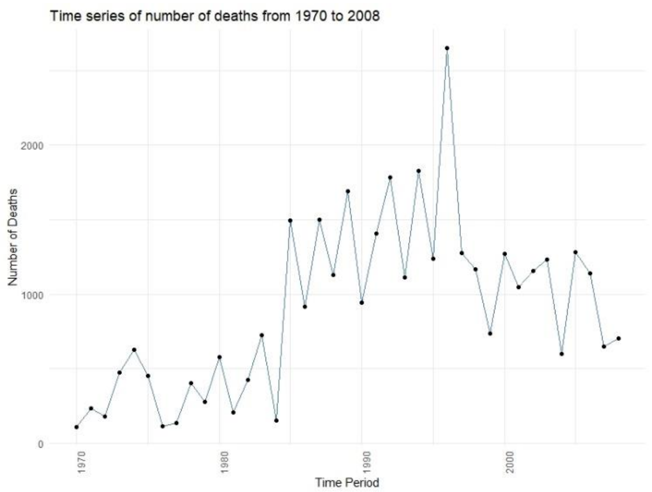

Figure 1: Number of lives lost from air disasters between 1970 and 2008.

These air accident volumes are alarming, but that has not deterred individuals from traveling. With affordable prices and advancement in technology, the total number of passengers carried on commercial flights are as high as 4.5 billion as at 2019 (ICAO Member States, 2019).

This creates a need to explore three main objectives for this research:

- Whether countries that fall in HDI band 2 and 3 (developing countries and developed countries) have the same mean number of deaths due to plane accidents or not.
- If chances of surviving a plane crash depends on the year of crash or not.
- If survival of any individuals in a plane accident can be predicted by HDI score and the year of accident.

Data

Data for this research was collated form Gapminder.org for plane crashes between 1970 and 2008, published by EM-DAT International Disaster Database; and combined with HDI data published by UN Development programme for 1990 to 2018.

The annual crash data for affected and deaths detailed the number of deaths and the number of individuals affected (injured or dead) by each country. Including records of 120 countries with 40 years of volumes of affected and dead, this data was then merged with HDI data for 188 countries over a period of 30 years till 2018. Having observations for different timeframe, it posed a challenge when combining these datasets where only the overlapping data between 1990 and 2008 could be used for further analysis. A list of variables used in this research is as below in Figure 2:

Field Name	Derived	Description	Format
Country		Country of the plane crash	Character
Year		Year of crash. (Result of transpose)	Numeric
HDI		HDI score for each country for a particular year	Numeric
Affected_Per_Year		Number of affected individuals from plane crash. (Result of transpose)	Numeric
Deaths_Per_Year		Number of affected individuals from plane crash. (Result of transpose)	Numeric
HDI_Band	Derived	HDI score grouped in 3 categories – 1, 2, 3	Character
Year_Band	Derived	5-year band	Character
Decade_Band	Derived	10-year band	Character
Injured_Per_Year	Derived	Number of affected – Number of dead	Numeric
Survived_flag	Derived	If Injured_Per_Year is greater than 0 then survived, else not survived	Character

Figure 2: List of variables used for analysis.

As the data was a wide table of volumes of affected, dead and HDI for each year, it had to be transposed using pivot_longer() function to have the countries by each year with volumes of affected, alongside. Year being a character when imported, it was converted to numeric to make it easier when for further computations. All 3 datasets were merged using inner_join() function by 'year' and 'country' to discard any observations weren't common to all 3 datasets. Any observations that had a missing HDI score were dropped along with any that didn't have any affected individuals or deaths from a crash. This brought down the size of dataset to mere 339 observations.

New variable **HDI_band** was created using HDI score of a country, categorized into 3 buckets – 1, 2 and 3 where 1 represents 'under-developed country' with score less than or equal to 0.333, 2 represents 'developing country' with score between 0.334 and 0.666, and 3 'developed country' with score between 0.667 and 1. This variable would further assist in conducting hypotheses testing and understand its association to Year_Band. **Year_Band** and **Decade_Band** were created to group the 5 year and 10-year periods to study the impacts of crashes over longer period.

To conduct hypotheses testing, a random sample of 150 observations was taken out of the population of 339 using set.seed() and sample(1:nrow(),150) function.

A summary of the sampled dataset is as below in Figure 3 –

```
   country                year              HDI
 Length:150         Min.   :1990   Min.   :0.2980
 Class :character   1st Qu.:1995   1st Qu.:0.4788
 Mode  :character   Median :2000   Median :0.6385
                    Mean   :1999   Mean   :0.6182
                    3rd Qu.:2004   3rd Qu.:0.7140
                    Max.   :2008   Max.   :0.9070
 Affected_Per_Year Deaths_Per_Year    HDI_Band
 Min.   : 10.00    Min.   :  3.00   Length:150
 1st Qu.: 22.00    1st Qu.: 19.25   Class :character
 Median : 40.50    Median : 37.50   Mode  :character
 Mean   : 83.12    Mean   : 70.09
 3rd Qu.:109.50    3rd Qu.: 91.75
 Max.   :902.00    Max.   :432.00
  year_band          decade_band         Injured_Per_Year
 Length:150         Length:150         Min.   :  0.00
 Class :character   Class :character   1st Qu.:  0.00
 Mode  :character   Mode  :character   Median :  0.00
                                       Mean   : 13.03
                                       3rd Qu.:  3.00
                                       Max.   :470.00
 Survived_flag      Log_Deaths_Per_Year Log_Affected_Per_Year
 Length:150         Min.   :1.099        Min.   :2.303
 Class :character   1st Qu.:2.957        1st Qu.:3.091
 Mode  :character   Median :3.624        Median :3.701
                    Mean   :3.711        Mean   :3.844
                    3rd Qu.:4.519        3rd Qu.:4.696
                    Max.   :6.068        Max.   :6.805
```

Figure 3: Summary of the variables for final dataset

Methods

All analysis and statistical procedures were performed in R studio version 1.3.959 with a list of libraries referenced in Appendix 1.3 (RStudio Team, 2020).

From the records for 120 countries over 40 years (4,800 observations), there were only 339 observations with aviation disasters affecting people where they were either injured or dead, which makes up only 7.06% chance of randomly selecting a country for any particular year to have had a plane crash. This created the data for number of deaths and affected extremely skewed as witnessed in Figure 4 and verified by skewness test by skewness() function that returned with **2.386191**. Meaning the variables Deaths_Per_Year and Affected_Per_Year had to be transformed before further analysis. Log Transformation was used on these two variables to normalize them. A comparison of before and after a log transformed for number of deaths could be observed in Figure 4. These graphs were created using ggplot(), geom_histogram() and geom_density() and presented side-by-side using plot_grid()

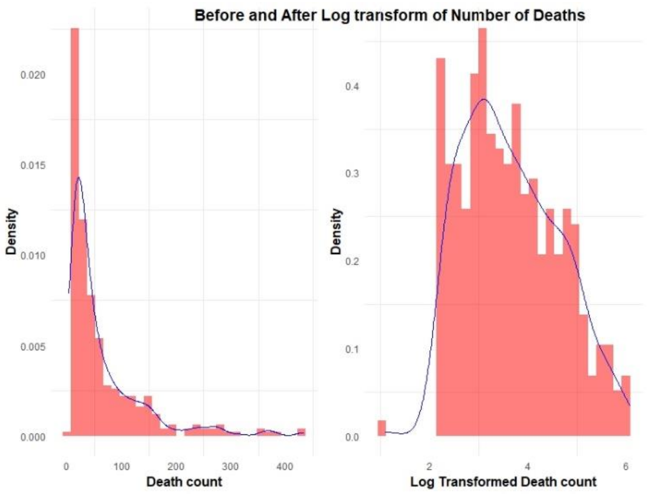

Figure 4: Histogram for Deaths_Per_Year before and after log transformation

Objective 1:

Null hypothesis for this objective is that there is no difference in the number of deaths for HDI group 2 and 3. That is, the mean of the number of deaths is the same for both HDI groups.

Alternative hypothesis is that the mean number of deaths is higher in HDI 2 compared to HDI 3.

This creates a hypothesis as follows:

$$H_0: \mu_2 = \mu_3$$
$$H_A: \mu_2 \neq \mu_3$$

It is of interest to understand how the mean number of deaths are different for HDI 2 and 3. The reason for selecting only groups 2 and 3 is because of extremely low number of air crashes for HDI group 1. This could potentially be due to low number of flights scheduled to fly into under-developed countries. As visible in Figure 5, HDI group 1 with minimum number of crashes is insignificant for testing the mean difference between the groups.

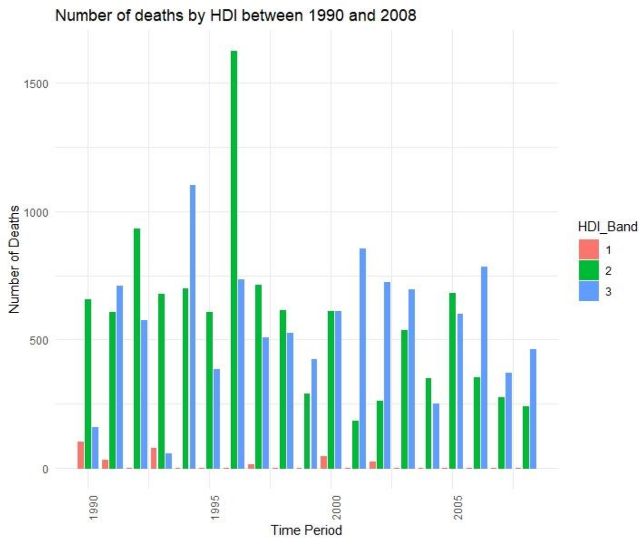

Figure 5: Number of deaths by HDI groups between 1990 and 2008

Using the Q-Q plot (qqnorm() function), log transformed number of deaths for HDI 2 and 3 are plotted to determine whether the data satisfies normal distribution before conducting the test. Figure 6 denotes HDI group 2 on left and group 3 on right. Group 2 seems normally distributed by group 3 requires confirmation.

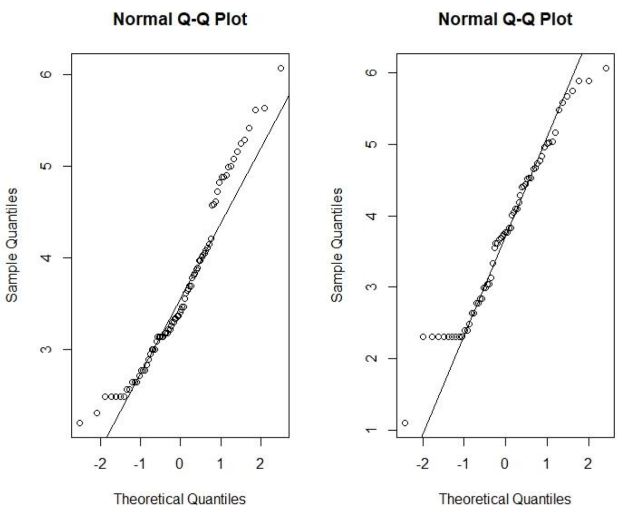

Figure 6: Left is number of deaths for HDI group 2, right is number of deaths for group 3.

Shapiro's test is done using shapiro.test() for both groups to identify whether the data satisfies the normality. With the probability of committing a type I error, common significance level of 5% is used here. Therefore, at α=0.05, p-value of 0.05769 is rejected for HDI group 3.

Given the assumption of normality wasn't fulfilled for t-test, non-parametric test of Wilcoxon signed-rank test for two independent samples was used. Wilcoxon test assumes that the data are two independent simple random sample (SRS) and each sample has 11 or more cases. Both these assumptions are met.

Objective 2:

H_0: Chance of survival doesn't dependon year of crash
H_A: Chance of survival depends on crash year

To test this hypothesis, chi-squared test of independence is used as this would provide a relationship analysis and understand the dependency of year of crash and the survival. This test requires most of the expected counts to be greater than 5, none less than 1; data represents actual counts; and observations occur in 1 and only 1 of several distinct categories.

As the variable survived_flag was created to have a single value for each crash and meeting all the other assumptions.

Objective 3:

Predicting the chances of survival from a crash provides a significant breakthrough in aviation industry. Getting anywhere close to gathering relevant factors which are least expected would provide a significantly strong base for future research.

This objective investigates if HDI score and year of crash predicts the survival from a plane crash or not. Logistic regression is used for this purpose as an outcome could only be binary.

Random sample of 150 variables is used for this test and survived_flag is converted to have 1 for 'survived' and 0 for 'none survived'. Glm() function from r is used with family = 'binomial'.

Results and Discussion

Objective 1:

With p-value of 0.5689, at significance of 5%, there is not enough evidence to reject null hypothesis. Concluding that mean number of deaths for HDI group 2 countries is the same as HDI group 3. Meaning that both developing countries and developed countries have equal chance of having a plane crash at any given year.

Looking at the top 10 countries that had the maximum number of deaths in last decade from 1999 to 2008 shown in Figure 7, 6 of those countries belonged to HDI 3 and 4 belonged to HDI 2. This extends on the findings that HDI 2 and 3 have equal mean number of deaths from plane crashes.

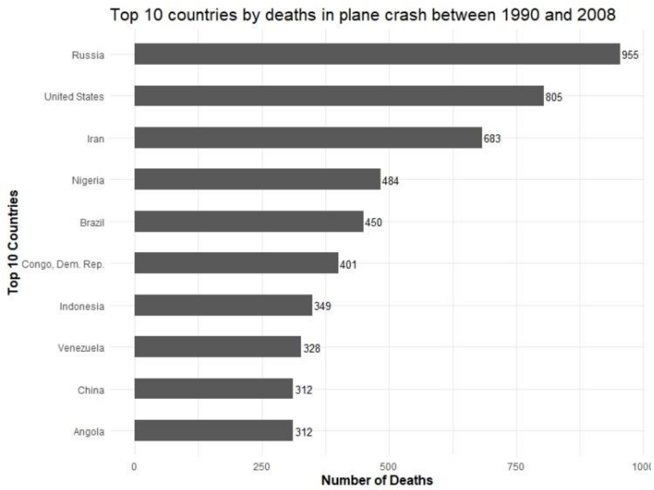

Figure 7: Top 10 countries by number of deaths in last decade as of 2008.

Objective 2:

Probability of committing a type I error, common significance level of 5% is used here. Therefore α=0.05.

Conducting the chi-squared test, with p-value of 0.259, there is not enough evidence to reject null hypothesis and conclude that year of plane crash does not influence if anyone survives from a particular plane crash or not.

This concludes that over the years with increased safety standards and regulatory measures, there is still very minimal chance of survival from a plane crash.

Objective 3:

From regression model, none of the variables were statistically significant but year band of '2000 –

2004' and '2005 +' would have been significant at significance level of 10%. However, because type I error is of interest, no variables satisfy the cut.

Concluding Remarks

Retrospection of this research provides case study between unique variables that have no research done on prior to this. HDI provides a country's position compared to the world was explored to have mean number of deaths from air disasters the same for both developing and developed countries.

With progression in aviation industry over the years, it would have been assumed that the chances of survival would be higher in a plane crash in recent years compared to 1990s', but it was contrary with not enough evidence to suggest that survival in a plane crash would depend on year of crash.

Lastly, to study the relationship of HDI score and year band to predict the survival of any individuals from a crash suggested that neither of them were statistically significant predictors.

Although with data constraints with lack of significant variables such as flight model, flight operator, reason for the crash, time of the crash, etc. to conduct sufficient analysis, this research provides basis to explore different avenues other than HDI score and the year of crash and focus more on the other variables.

Future research in this field could provide ground-breaking solution for aviation industry to reduce the number of deaths from air disasters significantly.

Appendix

References

EM-DAT. (2009). *The International Disaster Database*. Retrieved from Emergency Events Database: https://www.emdat.be/

Grupo One Air. (2020, July 19). *Commercial aviation growth and forecasts 2020-2038*. Retrieved from Grupo One Air: https://www.grupooneair.com/analysis-global-growth-commercial-aviation/

Gunston, B. (1990). *Avionics: The story and technology of aviation electronics.* Wellingborough, UK: Patrick Stephens Ltd.

Human Development Data Center. (2009). Retrieved from Human Development Reports: UNITED NATIONS DEVELOPMENT PROGRAMME: http://www.hdr.undp.org/en/data

ICAO Member States. (2019). *The World of Air Transport in 2019*. Retrieved from ICAO: https://www.icao.int/annual-report-2019/Pages/the-world-of-air-transport-in-2019.aspx

Institute for Operations Research and the Management Sciences. (2020, September 1). *Airline passengers in developing countries face 13 times crash risk as US*. Retrieved from ScienceDaily: https://www.sciencedaily.com/releases/2010/09/100901132235.htm

Kelly, J. (2015, August 18). *August 1985: The worst month for air disasters*. Retrieved from BBC News: https://www.bbc.com/news/magazine-33931693

RStudio Team. (2020). *RStudio: Integrated Development Environment for R.* Retrieved from RStudio, PBC: http://www.rstudio.com/

Wald, M. L. (1995, December 8). *Aviation Meeting Analyzes Slow, Steady Progress on Safety*. Retrieved from The New York Times Archives: https://www.nytimes.com/1995/12/08/us/a

viation-meeting-analyzes-slow-steady-progress-on-safety.html

R Code

```
#-------------------------------#
# MA5820 Assignment 3: Capstone
# Author: Dhruvisha Gosai
#-------------------------------#

RStudio.Version()

#Loading relevant R packages

library(car)
library(datasets)
library(ggplot2)
library(dplyr)
library(tidyverse)
library(qqplotr)
library(ggfortify)
library(ggmap)
library(ggthemes)
library(hrbrthemes)
library(viridis)
library(tidyr)
library(MASS)
library(lmtest)
library(boot)
library(plyr)
library(readr)
library(data.table)
library(crosstalk)
library(plotly)
library(shiny)
library(leaflet)
library(zoo)
library(cluster)
library(UsingR)
library(cowplot)
library(sqldf)
library(moments)
library(ggpubr)

#========================================
========================================#

#--------------------------------#
#          IMPORT DATA           #
#--------------------------------#

HDI <- read.csv("D:/Dhru Folder/JCU -
Master of Data science/MA5820 -
Statistical Methods and data
analysis/Assignment
3/hdi_human_development_index.csv"
                ,header = TRUE
                ,sep=",")

plane_affected <- read.csv("D:/Dhru
Folder/JCU - Master of Data
science/MA5820 - Statistical Methods and
data analysis/Assignment
3/plane_crash_affected_annual_number.csv
"
                ,header = TRUE
                ,sep=",")

plane_death <- read.csv("D:/Dhru
Folder/JCU - Master of Data
science/MA5820 - Statistical Methods and
data analysis/Assignment
3/plane_crash_deaths_annual_number.csv"
                ,header = TRUE
                ,sep=",")

#========================================
========================================#

#----------------------------------------
------------------#
#  Data Transformation: Clean and
transform as required   #
#----------------------------------------
------------------#

# View dataset
view(HDI)
view(plane_affected)
View(plane_death)

# Summary
str(HDI)
str(plane_affected)
str(plane_death)

summary(HDI)
summary(plane_affected)
summary(plane_death)

describe(HDI)
describe(plane_affected)
describe(plane_death)

#----------------------------
# Transpose the data to join
HDI_transpose              <- HDI
%>% pivot_longer(-country, names_to =
"year", values_to = "HDI") %>%

transform(year_clean = substr(year,2,5))
plane_affected_transpose  <-
plane_affected %>% pivot_longer(-
country, names_to = "year", values_to =
"Affected_Per_Year") %>%

transform(year_clean = substr(year,2,5))
plane_death_transpose     <- plane_death
%>% pivot_longer(-country, names_to =
"year", values_to = "Deaths_Per_Year")
%>%

transform(year_clean = substr(year,2,5))

# Creating year variable as numeric and
dropping year_clean column
HDI_transpose$year <-
as.numeric(HDI_transpose$year_clean,repl
ace = T)
HDI_transpose      <-
subset(HDI_transpose, select = -
c(year_clean)) #Drop column
```

```
plane_affected_transpose$year <-
as.numeric(plane_affected_transpose$year
_clean,replace = T)
plane_affected_transpose      <-
subset(plane_affected_transpose, select
= -c(year_clean)) #Drop column

plane_death_transpose$year <-
as.numeric(plane_death_transpose$year_cl
ean,replace = T)
plane_death_transpose      <-
subset(plane_death_transpose, select = -
c(year_clean)) #Drop column

#---------------------------
# Creating a single table with data for
HDI, deaths and affected -> cleaning the
year to remove x
full_data_x1 <-
inner_join(HDI_transpose,
plane_affected_transpose, by=c("year" =
"year","country"="country")) %>%

inner_join(.,plane_death_transpose,
by=c("year" =
"year","country"="country"))


# Checking to see if year is numeric and
all variables
summary(full_data_x1)

#---------------------------
# Creating Char HDI variable
full_data_x2 <- full_data_x1 %>%
                  filter(!is.na(HDI))
%>%
                  mutate(HDI_Band    =
case_when(HDI <= 0.333 ~ "1",

HDI > 0.333 & HDI <= 0.666  ~ "2",

HDI > 0.666 & HDI <= 1      ~ "3",

HDI > 1 ~ "4"),
                         year_band   =
case_when(year >= 1990 & year <= 1994 ~
"1990 - 1994",

year >= 1995 & year <= 1999 ~ "1995 -
1999",

year >= 2000 & year <= 2004 ~ "2000 -
2004",

year >= 2005                ~ "2005 +"),
                         decade_band =
case_when(year >= 1990 & year <= 1999 ~
"1990 - 1999",

year >= 2000 & year <= 2008 ~ "2000 -
2008"))

full_data_x2$Injured_Per_Year <-
full_data_x2$Affected_Per_Year -
full_data_x2$Deaths_Per_Year

full_data_x2$Survived_flag <-
ifelse((full_data_x2$Affected_Per_Year -
full_data_x2$Deaths_Per_Year) > 0,
"Survived", "None Survived")


#========================================
========================================#

#---------------------#
# Data Visualisation  #
#---------------------#

#------ Time series to visualise
datasets: Plane deaths ------#

TimeSeries <- plane_death_transpose %>%
  group_by(year) %>%
  summarise(Deaths_Per_Year =
sum(Deaths_Per_Year)) %>%
  select(year,Deaths_Per_Year)

ggplot(data=TimeSeries, aes(x=year,
y=Deaths_Per_Year)) +
  geom_line( color="steelblue") +
  geom_point() +
  xlab("Time Period") +
  ylab("Number of Deaths") +
  ggtitle("Time series of number of
deaths from 1970 to 2008") +
  theme_minimal() +

theme(axis.text.x=element_text(angle=90,
hjust=1))

#------ Visualise datasets: Full merge 2
by HDI ------#
TimeSeries2 <- full_data_x2 %>%
  group_by(year, HDI_Band) %>%
  summarise(Deaths_Per_Year =
sum(Deaths_Per_Year)) %>%
  select(year, HDI_Band
,Deaths_Per_Year)

ggplot(data=TimeSeries2, aes(x=year,
y=Deaths_Per_Year, fill=HDI_Band,
color=HDI_Band)) +
  geom_bar(stat="identity", width=.65,
position = position_dodge(width=0.9))  +
  xlab("Time Period") +
  ylab("Number of Deaths") +
  ggtitle("Number of deaths by HDI
between 1990 and 2008") +
  theme_minimal() +

theme(axis.text.x=element_text(angle=90,
hjust=1))

#-------- Top 20 Countries with highest
deaths ------#

Top20_Countries_death <-
plane_death_transpose %>%
  filter(between(year, 1999, 2008)) %>%
  group_by(country) %>%
```

```r
  summarise(Deaths_Per_Year =
sum(Deaths_Per_Year)) %>%
  select(country,Deaths_Per_Year) %>%
  arrange(desc(Deaths_Per_Year))

Top20_Countries_death <-
data.frame(head(Top20_Countries_death, n
= 10))

ggplot(data=Top20_Countries_death, aes(x
= reorder(country, Deaths_Per_Year), y =
Deaths_Per_Year)) +
  geom_bar(stat = 'identity', width =
0.5) +
  geom_text(aes(label =
Deaths_Per_Year), stat = 'identity',
data = Top20_Countries_death, hjust = -
0.1, size = 3.5) +
  coord_flip() +
  xlab('Top 10 Countries') +
  ylab('Number of Deaths') +
  ggtitle('Top 10 countries by deaths in
plane crash between 1990 and 2008') +
  theme_minimal() +
  theme(plot.title = element_text(size =
16),
        axis.title = element_text(size =
12, face = "bold"))

HDI_2008 <- full_data_x2 %>%
filter(year=="2008") %>%
select(HDI_Band, country)
Top20_Countries_death_HDI <-
inner_join(Top20_Countries_death,
HDI_2008, by="country")
#========================================
========================================#
#-------------------#
# Data Exploration #
#-------------------#

#Need to remove observations with deaths
per year as 0
full_data_x3 <- full_data_x2 %>%
  filter(Deaths_Per_Year > 0 |
Affected_Per_Year > 0)

#Check for distribution for number of
deaths
ggplot1 <- ggplot(full_data_x3,
aes(x=Deaths_Per_Year)) +
  geom_histogram(aes(y = ..density..),
fill = 'red', alpha = 0.5) +
  geom_density(colour = 'blue')+
  xlab("Death count") +
  ylab("Density") +
  theme_minimal() +
  theme(plot.title = element_text(size =
16),
        panel.grid.major =
element_blank(),
        axis.title = element_text(size =
12, face = "bold"))


#****** Log Transformation *****
full_data_x3$Log_Deaths_Per_Year    <-
log(full_data_x3$Deaths_Per_Year)
full_data_x3$Log_Affected_Per_Year <-
log(full_data_x3$Affected_Per_Year)
full_data_x3$Injured_Per_Year      <-
log(full_data_x3$Injured_Per_Year)


#Check for distribution for number of
deaths
ggplot2 <- ggplot(full_data_x3,
aes(x=Log_Deaths_Per_Year)) +
  geom_histogram(aes(y = ..density..),
fill = 'red', alpha = 0.5) +
  geom_density(colour = 'blue')+
  xlab("Log Transformed Death count") +
  ylab("Density") +
  theme_minimal() +
  theme(plot.title = element_text(size =
16),
        panel.grid.major =
element_blank(),
        axis.title = element_text(size =
12, face = "bold"))

plot_grid(ggplot1, ggplot2, labels =
"Before and After Log transform of
Number of Deaths")

# Need to make sure there are no
outliers - NONE FOUND
boxplot(full_data_x3$Log_Deaths_Per_Year
, main = "Boxplot of log transformed
number of deaths")
outlier <-
boxplot.stats(full_data_x3$Log_Deaths_Pe
r_Year)$out     #Identifying outlier


#------ Confirmation that the
transformation worked
# Skewness test before transformation
skewness(full_data_x3$Deaths_Per_Year)
#Before log transform
skewness(full_data_x3$Log_Deaths_Per_Yea
r) #After log transform
#----------------------------------------
------------#
#----------------------------------------
------------#


# Conduct Shapiro test for normality
shapiro.test(full_data_x3$Log_Deaths_Per
_Year)


# Create a random sample
seed <- set.seed(1122)
Random_Sample <-
sample(1:nrow(full_data_x3), 150)

full_data_x4 <-
full_data_x3[Random_Sample, ]
view(full_data_x4)
summary(full_data_x4)

# Skewness test before transformation
```

```
skewness(full_data_x3$Deaths_Per_Year)
#Before log transform
skewness(full_data_x3$Log_Deaths_Per_Yea
r) #After log transform
#========================================
========================================#
# Hypothesis Testing
#========================================
========================================#

#------- Objective 1: Mean deaths by HDI
group --------#
# H0: Number of deaths by HDI groups are
equal
# H1: Number of deaths by HDI groups are
not equal
#----------------------------------------
---------------#

ggplot3 <-
ggqqplot(full_data_x4$Deaths_Per_Year,
         ylab="Deaths from Plane Crash")
#distribution check

ggplot4 <-
ggqqplot(full_data_x4$Log_Deaths_Per_Yea
r,
         ylab="Log Deaths from Plane
Crash") #distribution check

plot_grid(ggplot3, ggplot4,
          labels = c('Q-Q Plot of Deaths
(Sample Data)', 'Q-Q Plot of Log Deaths
(Sample Data)'),
          ncol = 1)

shapiro.test(full_data_x4$Log_Deaths_Per
_Year) #Normality test - Not normally
distributed

HDI_2 <- subset(full_data_x4, HDI_Band
== "2")
HDI_3 <- subset(full_data_x4, HDI_Band
== "3")

par(mfrow=c(1,2))
qqnorm(HDI_2$Log_Deaths_Per_Year)
qqline(HDI_2$Log_Deaths_Per_Year)
qqnorm(HDI_3$Log_Deaths_Per_Year)
qqline(HDI_3$Log_Deaths_Per_Year)


shapiro.test(HDI_2$Log_Deaths_Per_Year)
#Normality test - Not normally
distributed
shapiro.test(HDI_3$Log_Deaths_Per_Year)
#Normality test - Not normally
distributed

HDI_2_3 <- subset(full_data_x4, HDI_Band
!= "1")

wilcox.test(Log_Deaths_Per_Year ~
HDI_Band,
data=HDI_2_3,alternative="two.sided",mu=
0)

# RESULT: With significance level of
0.05, there is not enough evidence to
reject null hypothesis with p-value of
0.5689
#          and conclude that mean is same
for each HDI banding which is consistent
with histogram we generated (HDI 2, 3
are almost identical)
#--------------------------------------

#========================================
========================================#
# Hypothesis Testing
#========================================
========================================#

#------- Objective 2: Dependency of Year
banding and HDI banding (2 and 3) ------
--#
# H0: Year banding and survival are
independent for number of deaths
# H1: Year banding and survival are
dependent for number of deaths
#----------------------------------------
----------------------------------------
---------#

#Create subset
Objective2_chisq <- subset(full_data_x4,
select = c(year_band, Survived_flag,
Deaths_Per_Year)) #keep column

Objective2_chisq <- Objective2_chisq %>%

group_by(year_band, Survived_flag) %>%

summarise(Count = n())
view(Objective2_chisq)

#Require a matrix of number of deaths
for chisq test
Not_Survived <- c(30,27,27, 20)
Survived <- c(7,11,16,12)

chisq_table = cbind(Not_Survived,
Survived)
rownames(chisq_table) = c("1990 - 1994",
"1995 - 1999", "2000 - 2004", "2005 +")
chisq_table

overall_chisq <- chisq.test(chisq_table)
overall_chisq

overall_chisq$expected

# RESULT: With p-value of 0.259, we
conclude that null hypothesis is not
rejected and conclude year and HDI
banding are dependent for number of
deaths.
#--------------------------------------

#========================================
========================================#

#------- Objective 3: --------#
# H0:
```

```
# H1: Mean of plane crash deaths for HDI
Banding 1, 2 and 3 are NOT equal
#------------------------------#

#------------------------------#
# Logistic regression
#------------------------------#
all(complete.cases(full_data_x4))

ggplot(full_data_x4, aes(x=HDI)) +
  geom_histogram(aes(y = ..density..),
fill = 'red', alpha = 0.5) +
  geom_density(colour = 'blue')+
  ggtitle("Histogram of HDI Score")+
  xlab("HDI score") +
  ylab("Density") +
  theme_minimal() +
  theme(plot.title = element_text(size =
16),
        panel.grid.major =
element_blank(),
        axis.title = element_text(size =
12, face = "bold"))

#------------------------------#
# Estimate the parameters
#------------------------------#
## Model Creation
full_data_x4$Survived_flag_Binary <-
ifelse(full_data_x4$Survived_flag
=="Survived", 1,0)

model <-
glm(Survived_flag_Binary~HDI+year_band,d
ata = full_data_x4, family = "binomial")
#REJECT
summary(model)
```