SEVERITY OF AVIATION DISASTERS: A CASE STUDY OF HISTORIC PLANE CRASHES

Executive Summary:

Aviation accidents are rare but significant. Advancements in respective safety, regulatory, and technological fields have helped curb the frequency of plane crashes in recent decades but there still exists a risk. Understanding the contributing factors that lead to plane crashes helps in preventing expensive loss, mitigating the chance of injuries, and saving lives.

These worthwhile outcomes have fueled the investigation of potential relationships between the mortality rates of plane crashes and a combination of:

- Human Development Index (HDI) score of the country it occurred it, as well as
- the year of the accident

by way of three distinct objectives that seek to address whether there is:

1) a difference in the mean of deaths resulting from plane countries by HDI score
2) a relationship between the mean of deaths resulting from plane countries by HDI score and the year of accident
3) a sufficient relationship between recorded deaths following a plane crash and HDI score to predict the approximate HDI score of the country

Plane crash data sourced from The International Disaster Database (*https://www.emdat.be/*) along with HDI Data provided by the United Nations website (*http://www.hdr.undp.org/*) provide an overview over the volume of deaths and injuries resulting from plane crashes, as well as the year, country and country's HDI score it occurred in. This information is then applied through several appropriate statistical tests to determine which of the accompanying hypotheses should be accepted.

The dominant message told by these outcomes indicate that although HDI score may not serve as a potentially predictive variable with applications to be used in anticipating crashes, there does appear to be at least a tenuous relationship between the volume of deaths resulting from plane crashes each year and a country's HDI score. This relationship is strengthened when taking into consideration the year the accident occurred; with the mid 1990's showing the highest proportion of deaths from the dataset.

Retrospective reporting of the total sum and mean of the total deaths by plane crash for HDI banding also indicate a couple of notable trends - although the median deaths for each HDI segment is similar, HDI 1 countries have a significantly higher upper quartile range of deaths compared to HD2 countries, and again to HD3 countries. This suggests that there have been several crashes in countries with lower HDI scores that were significantly more severe than higher-HDI countries have faced. This information could help spur funding from more developed nations to assist with plane-crashes in poorer HDI ratings countries as these countries are generally less-equipped to handle these incidents, and they have a higher likelihood of being more deadly than crashes faced by HD2 & HDI3 countries.

Introduction:

We are in an age of unparalleled aviation safety. Continuing technological and regulatory improvements have resulted in fewer and fewer plane crashes in the past few years. Understanding the causes of plane crashes allows regulators and aircraft developers to make changes so they are less likely to occur. Consequently, understanding the driving elements that influence the severity of plane crashes that *do* happen will provide opportunity to reduce the mortality and injury rate arising from aviation incidents.

The Human Development Index (HDI) is "*a summary measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable and have a decent standard of living*" (*Human Development Reports*, http://hdr.undp.org/en/content/human-development-index-hdi). The underlying motive of this paper is to explore if any relationships exist between plane crashes and HDI, with the main objectives of this report being to investigate whether there:

4) Is a difference in the mean of deaths resulting from plane countries by HDI score
5) Is a relationship between the mean of deaths resulting from plane countries by HDI score and the year of accident
6) Is there a relationship between the decade of a plane crash and the number of people who survived

DATA

For this analysis, three datasets were sourced from gapminder.org. Two of these contained the volume of Deaths and Injuries between 1970 to 2008 (39 years) across a span of 120 countries resulting from plane crashes, while the other dataset contained 26 years' worth of Human Development Index (HDI) data of 188 countries between the years of 1990-2015.

These datasets were then imported via use of the native RStudio function read.csv on default settings, and then transposed through use of the pivot_longer function available from the dplyr package. Following the transpose the datasets were able to be joined together on both Country_Name and observation year – through default use of the sqldf function and package of the same name. This package was also leveraged to create 3 ordinal levels of a new variable **HDI_Banding** – "1" being countries that had HDI <= 0.333 on that year, "3" being countries that had a HDI score >= 0.666 and "2" being anything in between. An additional column of **Year_Banding** was also created, the values of "1990 – 1994"," 1995 – 1999", "2000 - 2004" or "2005 +" depending on which year was listed for each observation. Finally, the year was then converted into numeric through use of the *as.numeric* function, employing the optional argument of "replace = T" to overwrite the original column. For the purposes of reproducibility the *set.seed* function was used with an input of **1234**. This was then applied to create a dataframe containing a random sample of 100 observations by using the *sample* function on standard settings.

An overview of the variables used in the analysis can be found in figure 3.1, as well as a brief summary table in item 3.2.

Variable	Format	Description
Country	Char	Country Name
Year	Num	Year of Observation
HDI_Banding	Char	Ordinal grouping of HDI scores into 3 Segments
DeathsInjuriesPerYear	Num	Volume of people affected from plane crashes
DeathsPerYear	Num	Volume of deaths resulting from plane crashes
InjuriesPerYear	Num	Volume of injuries resulting from plane crashes
RatioOfDeath	Float	The volume of people who died / by the total #

3.1 overview of variables used in the analysis

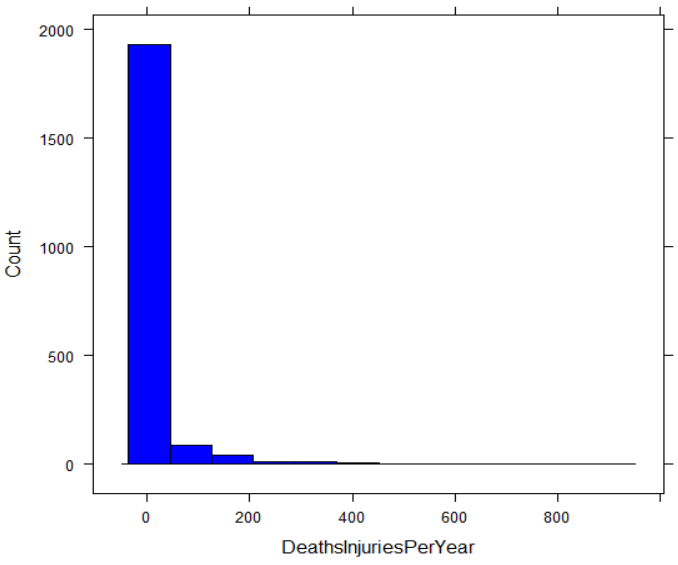
country	Year	HDI_Banding	DeathsInjuriesPerYear	DeathsPerYear	InjuriesPerYear	RatioOfDeath
Length:2082	Min. :1990	Length:2082	Min. : 0.00	Min. : 0.00	Min. : 0.000	Min. :0.0270
Class :character	1st Qu.:1995	Class :character	1st Qu.: 0.00	1st Qu.: 0.00	1st Qu.: 0.000	1st Qu.:0.9200
Mode :character	Median :2000	Mode :character	Median : 0.00	Median : 0.00	Median : 0.000	Median :1.0000
	Mean :1999		Mean : 12.42	Mean : 10.48	Mean : 1.938	Mean :0.9081
	3rd Qu.:2004		3rd Qu.: 0.00	3rd Qu.: 0.00	3rd Qu.: 0.000	3rd Qu.:1.0000
	Max. :2008		Max. :902.00	Max. :432.00	Max. :470.000	Max. :1.0000
						NA's :1741

Methods:

Note, use of any and all mentioned functions hereafter were run in R-Studio and are run under default settings unless otherwise specified.

Fortunately, most countries in a given year do not have a plane-crash occur. For the 4,680 observations recorded (120 countries*39 years), there were only 492 observations where people were injured or killed due to plane crash – i.e. a 10.51% chance of randomly selecting a year and country from this list that had a plane crash that year.

This resulted in a dataset that was severely positively skewed – revealed by figure 4.1



4.1 - 90% of observations had zero plane crashes recorded

This was skewness was further verified by employing the use of the skewness function in R on variable DeathsPerYear, which returned the value of **2.3952** – further substantiating that this variable was decidedly skewed.

Due to the abundance of zeroes the task of normalizing would require a slightly non-standard approach – including log10(+1) or a box-cox transformation however were ultimately unnecessary as the hypotheses did not require the observations with missing values to be passed through into the final analyses dataset. As a result observations where there no people were affected by plane crashes for that year were omitted.

The sqldf package and function of the same name in R-Studio was used to join the plane crash datasets

and HDI information together on both country name and observation year.
Preliminary investigations revealed a prevalence of missing values (i.e. not 0) for the period of 1970-1989 – particularly for low HDI countries. Coupled with the fact that HDI data was not available prior to 1990, the plane crash data was subset to start from 1990 onwards to ensure consistent observations for each country and year – and also to avoid any potential bias in the random sampling for HDI distribution.

Objective 1

With the data in a post-processed state some checks were necessary in order to determine which test would be most appropriate. The shapiro.test function in R-Studio on default settings returned a p-value of **3.691e-13** for the numeric variable DeathsPerYear, indicating that the values were far from normally distributed. Thus a non-parametric test was required to test 3 levels of a categorical variable, which satisfies the application of a Kruskal-Wallis test.

Objective 2

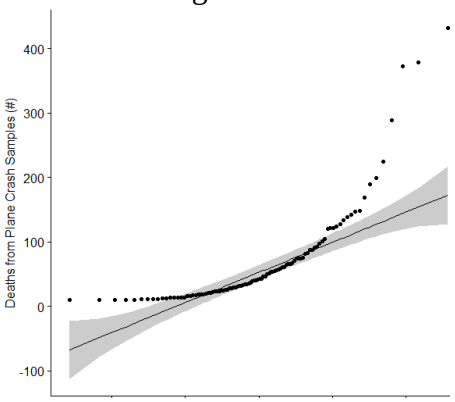
Following data exploration it was found that there were significantly fewer recorded deaths and injuries declared under HDI 1 countries. This resulted in a violation of one of the required assumptions of a Chi-Square Test of Independence which is that there are at least 5 counts of frequency for each group. This is potentially due to there being fewer HDI countries overall, or fact-recording faculties being underdeveloped for these nations at the time.
As a result the hypothesis was structured by testing whether there were observed relationships between the categorical variables of HDI_Banding and Year_Banding – for countries that fell into either HDI_Banding 2 or 3.
The sqldf package and function of the same name in R-Studio was used to sum the number of deaths recorded, grouped by HDI_Banding and Year_Banding – while filtering out observations that had HDI_Banding = 1.
The native R-Studio MATRIX function on default settings was then used to pass the results of the prior dataset into a Pearson's Chi-squared test of Independence– employing the use of the chisq.test function with the specification of explicitly disabling the use of Yates' Continuity Correction by specifying "correct=FALSE" as a secondary argument. This was required because this test does not use a numeric variable in the calculation, but instead counts the frequencies of the observations.
The formula for this test is computed as:

$$\chi^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$$

As the hypotheses for this test centered on the categorical variables of country and year-banding, it was necessary to remove the countries that were HDI Banding of 1 in order to satisfy the assumptions of the test.

Objective 3

Steering away from the hypotheses of Objective 2, the objective here was orientated towards exploring whether a significant relationship existed specifically between the decade a plane crash occurred in and the number of survivors from these crashes. The *ggqqplot* function on default settings was used to produce the below plot indicating the data did not appear to fall across a normal distribution – illustrated in figure 4.2



4.2 – Unnormalized data

This was further verified by the use of two functions - *shapiro.test* and *skewness* functions using DeathsPerYear variable.
The results of the Shapiro-Wilks test for normality indicated a p-value of **3.691e-13** – meaning we reject the null hypothesis of normalized data. Additionally, the test of skewness returned a value of **2.6387** – falling well outside the upper threshold of 1 as a limit, indicating that the data was highly right-skewed (the mean of the values was significantly larger than the median).
In an attempt to reduce these extremes to allow use of this variable in parametric tests, the log() function was applied to variable DeathsPerYear – in addition to another iteration of the shapiro-wilks and skewness functions. Logarithm was selected as a suitable transformation method as the numeric variable was right-skewed. The results of the shapiro-wilks returned a p-value of 0.0342 meaning the null hypothesis of normally-distributed data

could be accepted, while the output of the skewness test gave a figure of 0.2843 indicating the data was approximately symmetrical.
A before and after snapshot of this transformation can be seen below in figure 4.3

4.3 – *Before and after log transformation*

Using this dataset, two additional variables were then created – **Decade_Banding** with the values of "1990-1999" or "*2000-2008"* depending on when the accident occurred, and **Survived** which was a Boolean flag indicating a 1 if there were survivors from crashes in that year, or 0 if there were none.
Using the function glm and passing through this **Survived** column and the additional argument specifying the test should be conducted as "binomial", an iterative approach was taken where multiple subsets of the base data were run through this function for each distinct value of **Survived** and **Decade_Banding.**
Two assumptions were made for this test: the data displayed characteristics of both Homogeneity of Variance as well as variance-covariance homogeneity.

INVESTIGATION

Objectives and hypotheses of this analyses were as follows:
Objective 1: The first objective sought out to determine whether the mean number of deaths that occur due to plane crashes vary between more developed countries and lower developed countries. In effect this may be a result of greater safety controls in high HDI-scored countries being offset by a larger quantity of flights. While the opposite may be true for low scored-HDI countries – whereby fewer flights but higher potential of an incident do not wholly cancel each other out. Consequently, the hypotheses for this objective are:
H0 $\mu_1 = \mu_2$*: the mean number of deaths resulting from plane crashes by HDI are equal*
H1 $\mu_1 \neq \mu_2$*: the mean number of deaths resulting from plane crashes by HDI are NOT equal*

No assumptions were required for this test as the requirements were accounted for, and a 95% confidence interval was maintained.
Samples are random samples, or allocation to treatment group is random.
The two samples are mutually independent.
The measurement scale is at least ordinal, and the variable is continuous.
If the test is used as a test of dominance, it has no distributional assumptions. If it used to compare medians, the distributions must be similar apart from their locations.
The test is generally considered to be robust to ties. However, if ties are present they should not be concentrated together in one part of the distribution (they should have either a normal or uniform distribution)

Objective 2: Extending from the 1st objective, the second objective sought out to determine whether there was a relationship between the categorical variables of *HDI_Banding* and *Year_Banding*. This would further evaluate whether there were particular periods of time where a specific *HDI_Banding* showed significantly more deaths recorded than others. Thus the formed hypotheses for this were:

H0: there is no relationship between HDI Banding and Year Banding $H_0 : \mu_{New} - \mu_{Std} \leq 0$

H1: there is a relationship between HDI Banding and Year Banding $H_A : \mu_{New} - \mu_{Std} > 0$

Requirements of the ChiSquared Test of Independence include the requirement of one categorical variable, mutually exclusive groups, at least 5 frequencies for each group, and independence of observations, no assumptions were required as these conditions were tested.
A 95% confidence interval was maintained for the results of this test.

Objective 3: As the prior hypotheses had focused on severity and mortality of plane crashes by HDI bandings, it was advantageous for this test to engage alternative metrics. The number of people who survived plane crashes was examined against the year bandings to explore the whether there were significant correlations. This was then plotted against the HDI bandings of 2 and 3 as HDI 1 countries had insufficient observations to meet the entry requirements for this test. A simple logistic regression was selected as the test of choice with the hypotheses:

H0: there is no relationship between the decade that a plane crash occurs in and the number of survivors

$H_0: \beta_k = 0$

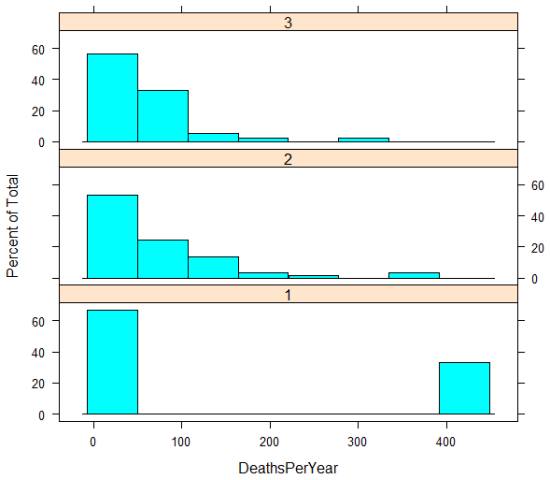
$H_1: \beta_k \neq 0$

H1: there is a significant relationship between the decade that a plane crash occurs in and the number of survivors

Results and Discussion

The overall theme of the results is varied with some instances of limited success. Relationships between HDI Banding to severity of plane crashes in terms of people affected or lives lost were not significant but there did appear to be some weak correlations, indicating that HDI 1 (low) countries had historic precedent of much more deadly aviation incidents compared to HDI 2 and 3 countries. Success was found in finding a strong relationship between decade of accident and survival rate – with plane crashes in the 21st century showing a much higher chance of survival compared to plane accidents that were during the 1990's.

Results of **Objective 1** Kruskal Wallis test – testing whether the mean volume of plane crash deaths is between HDI Bandings are equal – indicate that the mean for each group is not significantly different. (chi-squared Value = 0.75642, df = 2, p-value = 0.6851)
Therefore, there was insufficient evidence to reject the null hypotheses with a p-value of 0.6851 and significance level of 95% and conclude that the mean number of deaths resulting from plane crashes by HDI banding, are roughly the same.
This outcome is supported by histogram in figure 6.1, created via use of the histogram function in R-Studio sourced from the LATTICE package. As can be seen HDI Bandings 3 (high) and 2 (med) are approximately similar in terms of distribution of deaths, while HDI 1 initially a similar trend – with the majority of deaths falling <= 50. One notable difference is that HDI 1 has a significantly higher number of deaths resulting from one year – with 30% of the observations falling into the last segment. Had the hypothesis of this test centered solely on HDI 1 against either HDI 2 or HDI 3 countries, it is reasonable to expect a more significant p-value due to the overall dissimilar characteristics of HDI 1.



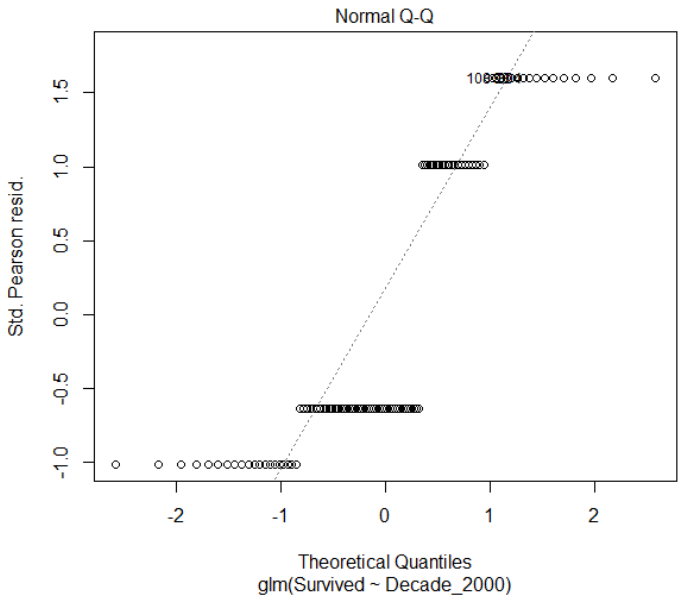
6.1 – roughly even means for each HDI banding

Results of **Objective 2**, consisted of a Chi-squared Test of Independence as means to determine whether there is a strong relationship between YearBanding and HDI_Banding in terms of the volume of deaths resulting from plane crashes.
indicate the null hypothesis of "there is no significant relationship between HDI Banding and Year Banding" (X-squared = 238.32, df = 3, p-value < 2.2e-16)
This output indicates that with a p-value that is < 2.2e-16 and significance level of 95% there is sufficient evidence to reject the null hypothesis and conclude that there is a relationship between the categorical variables of Year_Banding and HDI_Banding in terms of volume of deaths resulting from plane crashes.

Finally, the results of **Objective 3** entailed the evaluation of the output following a logistic regression. The glm function with specification of "binomial" test-type was used for 5 iterations – one for each distinct value of HDI Banding and Survived. The input and output of each of these tests are listed as follows:

HDI – 1: (z-value=-0.010,df=99,p-value=0.992)
HDI – 2: (z-value=-1.185, df=98, p-value= 0.236)
HDI – 3:(z-value0.4183,df=98,p-value= 0.1809)
Decade1990: (z-value=-2.175, df=98 , p-value=0.0296)
Decade2000: (z-value=2.175, df=98 , p-value=0.0296)

The key observations from these results are that there is no significance for the first 3 iterations of the logistic regression for each of the three HDI bandings – indicating that the volume of people surviving a plane crash in a HDI1, HDI2, or HDI3 country does not differ by a statistically significant amount.
However, for the fourth and fifth iterations of the regression, using inputs of Survival outcome against each of the two decades of 1990 and 2000 – the p-values for both are statistically significant. The intercept values output are -9.280 and 0.9280 for the decades of 1990 and 2000 respectively, indicating that there is an inverse relationship and that passengers involved in air accidents during the 1990's were significantly less likely to survive compared to those involved in plane crashes in the 21st century.
As a result there was insufficient evidence to reject the null hypothesis using for the test outcomes focused on HDI bandings – *however* - there was sufficient evidence to reject the null hypothesis and state there is a strong relationship between the variable Decade_Banding and the frequency of Survived with significance level of 95%. It can thus be concluded that plane crashes over time have had statistically significant decline in mortality rate. This is captured in figure 6.2 below – produced using the plot function in RStudio on default settings.



Concluding Remarks

In recent decades the frequency and severity of plane crashes has declined thanks to advancements in technology and greater regulatory and safety controls. Despite this there is likely to be a non-zero chance of a plane crash occ
For this report the chosen hypotheses were dependent on the few available variables, which limited the breadth and depth of available research. For objective 3 as part of the exploratory analysis, a linear regression was created exploring whether there is a potential relationship between HDI banding and survival rate, however no significant relationship was found. It is acknowledged that there is opportunity for further investigation between plane crashes and a country's HDI by focusing more on the volume and frequency of plane crashes rather than the volume of people killed or injured. Future research would be useful in

References

APA style referencing has resulted in references linked inline through the above text where needed.

Appendix

Code used to produce the above results:

```
usRp <-"https://cran.csiro.au/"

if(!require(ggplot2))   install.packages("ggplot2"  ,repos = usRp)
if(!require(dplyr))     install.packages("dplyr"    ,repos = usRp)
if(!require(tidyverse)) install.packages("tidyverse",repos = usRp)
if(!require(backports)) install.packages("backports",repos = usRp)
if(!require(ggpubr))    install.packages("ggpubr"   ,repos = usRp)
if(!require(evaluate))  install.packages("evaluate" ,repos = usRp)
if(!require(magrittr))  install.packages("magrittr" ,repos = usRp)
if(!require(qqplotr))   install.packages("qqplotr"  ,repos = usRp)
if(!require(psych))     install.packages("psych"    ,repos = usRp) #install
PSYCH pack
if(!require(tidyr))     install.packages("tidyr"    ,repos = usRp)
if(!require(sqldf))     install.packages("sqldf"    ,repos = usRp)
if(!require(textir))    install.packages("textir"   ,repos = usRp)
if(!require(geoR))      install.packages("geoR"     ,repos = usRp)
if(!require(moments))   install.packages("moments"  ,repos = usRp)
if(!require(mblm))      install.packages("mblm"     ,repos = usRp)
if(!require(countrycode)) install.packages("countrycode"    ,repos = usRp)
if(!require(forcats))  install.packages("forcats"   ,repos = usRp)
if(!require(ggeasy))   install.packages("ggeasy"    ,repos = usRp)
if(!require(lattice))  install.packages("lattice"   ,repos = usRp)
if(!require(MASS))     install.packages("MASS"      ,repos = usRp)
if(!require(epitools)) install.packages("epitools"  ,repos = usRp)
if(!require(lmtest))   install.packages("lmtest"    ,repos = usRp)
if(!require(robust))   install.packages("robust"    ,repos = usRp)
if(!require(boot))     install.packages("boot"      ,repos = usRp)
if(!require(car))      install.packages("car"       ,repos = usRp)
if(!require(caret))    install.packages("caret"     ,repos = usRp)

library(ggplot2)
library(dplyr)
library(tidyverse)
library(backports)
library(ggpubr)
library(evaluate)
library(magrittr)
library(qqplotr)
library(psych)       #assign PSYCH pack
library(tidyr)
library(sqldf)
library(textir)
library(geoR)
library(moments)
library(mblm)
library(countrycode)
library(forcats)
library(ggeasy)
library(lattice)
library(MASS)
library(epitools)
library(lmtest)
library(robust)
library(boot)
library(car)
library(caret)

par(mfrow=c(1,1))

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

#-------------
# Import Data
#-------------

loc_DS_HDI                  <- "C:/Users/jarma/Desktop/Jarg Assignment 3 -
MA5820/Import Data/hdi_human_development_index.csv"
loc_DS_PlaneCrashes_Affected    <- "C:/Users/jarma/Desktop/Jarg
Assignment 3 - MA5820/Import
Data/plane_crash_affected_annual_number.csv"
loc_DS_PlaneCrashes_Deaths      <- "C:/Users/jarma/Desktop/Jarg
Assignment 3 - MA5820/Import
Data/plane_crash_deaths_annual_number.csv"
export_loc                  <- "C:/Users/jarma/Desktop/Jarg Assignment 3 -
MA5820/Import Data/SummaryExport.xlsx"

DS_HDI <- read.csv(loc_DS_HDI)   #import HDI
DS_CRASH_AFF <- read.csv(loc_DS_PlaneCrashes_Affected)  #import
Crash Affected
DS_CRASH_DTH <- read.csv(loc_DS_PlaneCrashes_Deaths)    #import
Crash Death

DS_HDI        #view HDI
DS_CRASH_AFF   #view Crash Affected
DS_CRASH_DTH   #view Crash Death

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

#---------------------------------------------
# Validate dataset formats, and brief summaries
#---------------------------------------------


str(DS_HDI)            #confirm the data has imported in correct variable format
(i.e. not char)
summary(DS_HDI)        #basic summary statistics
describe(DS_HDI)       #a few more summary statistics using PSYCH

str(DS_CRASH_AFF)      #confirm the data has imported in correct variable
format (i.e. not char)
summary(DS_CRASH_AFF)   #basic summary statistics
describe(DS_CRASH_AFF)  #a few more summary statistics using PSYCH

str(DS_CRASH_DTH)      #confirm the data has imported in correct variable
format (i.e. not char)
summary(DS_CRASH_DTH)   #basic summary statistics
describe(DS_CRASH_DTH)  #a few more summary statistics using PSYCH


TRANS_HDI      <- DS_HDI %>% pivot_longer(-country, names_to =
"year", values_to = "hdi")
TRANS_CRASH_AFF <- DS_CRASH_AFF %>% pivot_longer(-country,
names_to = "year", values_to = "DeathsInjuriesPerYear")
TRANS_CRASH_DTH <- DS_CRASH_DTH %>% pivot_longer(-country,
names_to = "year", values_to = "DeathsPerYear")

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

#------------------------------
# Final Clean Datasets - merged
#------------------------------

fullMerge <- sqldf(" SELECT  HDI.*
                    ,AFF.DeathsInjuriesPerYear
                    ,DTH.DeathsPerYear
                FROM      TRANS_HDI      as HDI
                INNER JOIN TRANS_CRASH_AFF as AFF  on    HDI.YEAR
= AFF.YEAR
                                    and HDI.COUNTRY = AFF.COUNTRY
                INNER JOIN TRANS_CRASH_DTH as DTH  on    HDI.YEAR
= DTH.YEAR
                                    and HDI.COUNTRY = DTH.COUNTRY
                ORDER BY YEAR")

fullMerge1      <- transform(fullMerge,CleanYear = substr(year,2,5))
CharYear_ToNum  <- as.numeric(fullMerge1$CleanYear,replace = T)
MergeBack       <- data.frame(fullMerge1,CharYear_ToNum)

#------- Before Removing zeros from   -------#
#Identify observations that are not recorded properly
#Histogram of illogical observations
ggplot(fullMerge1, aes(x=DeathsPerYear)) +
  geom_histogram(colour="dodgerblue",fill=rgb(1,.54,0,.7), bins = 30) +
  scale_y_continuous(name="count") +
  labs(title="Histogram of number of people killed in air accidents from 1970
to 2008")
```

```
fullMerge1_final <- sqldf(" SELECT  a.country
                          ,CharYear_ToNum as Year
                          ,case   when a.hdi <= 0.333                    then '1'
                                  when a.hdi > 0.333 and a.hdi <= 0.666   then '2'
                                  when a.hdi > 0.666 and a.hdi <= 1       then '3'
                           else '4' end as HDI_Banding
                          ,case   when CharYear_ToNum between 1990 and 1994
then '1990 - 1994'
                                  when CharYear_ToNum between 1995 and 1999
then '1995 - 1999'
                                  when CharYear_ToNum between 2000 and 2004
then '2000 - 2004'
                           else '2005 +' end as Year_Banding
                          ,case   when (DeathsInjuriesPerYear - DeathsPerYear) > 0
then 1 else 0 end as Survived
                          ,case   when CharYear_ToNum between 1990 and 1999
then '1990 - 1999'
                                  when CharYear_ToNum between 2000 and 2008
then '2000 - 2008'
                           else '' end as Decade_Banding
                          ,count(DeathsInjuriesPerYear) as Count_Crashes
                          ,DeathsInjuriesPerYear
                          ,DeathsPerYear
                          ,(DeathsInjuriesPerYear - DeathsPerYear)           as
InjuriesPerYear
                          ,cast(DeathsPerYear as float)/DeathsInjuriesPerYear as
RatioOfDeath
                  FROM MergeBack as a
                  WHERE a.hdi is not null
                    and DeathsPerYear > 0
                      group by
country,Year,HDI_Banding,Year_Banding,Survived,Decade_Banding")

fullMerge1_final

fullMerge1_final$continent <- countrycode(sourcevar = fullMerge1_final[,
"country"],
                          origin = "country.name",
                          destination = "continent")

seed <- set.seed(1234)
Random_Sample <- sample(1:nrow(fullMerge1_final), 100)
fullMerge2 <- fullMerge1_final[Random_Sample, ]
fullMerge2


# Deaths per year by continent Bar chart:
fullMerge2 %>%
  mutate(name = fct_reorder(continent, desc(DeathsPerYear))) %>%
  ggplot( aes(x=continent, y=DeathsPerYear)) +
  geom_bar(stat="identity", fill="#f68060", alpha=.6, width=.4) +
  coord_flip() +
  ylab("Deaths Per Year") +
  xlab("Continents") +
  ggtitle("Plane Crash Deaths Per Year By Continents") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))

# Deaths per year by continent Bar chart:
fullMerge2 %>%
  mutate(name = fct_reorder(HDI_Banding, desc(DeathsPerYear))) %>%
  ggplot( aes(x=HDI_Banding, y=DeathsPerYear)) +
  geom_bar(stat="identity", fill="#f68060", alpha=.6, width=.4) +
  coord_flip() +
  ylab("Deaths Per Year") +
  xlab("HDI Banding") +
  ggtitle("Plane Crash Deaths Per Year By HDI Banding") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))

#=====================================================
=====================================================
====================#
# Hypothesis Testing
#=====================================================
=====================================================
====================#
#------- Objective 1: --------#
# H0: Mean of plane crash deaths for HDI Banding 1, 2 and 3 are equal
# H1: Mean of plane crash deaths for HDI Banding 1, 2 and 3 are NOT equal
#-----------------------------#

ggqqplot(fullMerge2$DeathsPerYear, ylab="Deaths from Plane Crash
Samples (#)") #distribution check
shapiro.test(fullMerge2$DeathsPerYear) #Normality test - Not normally
distributed

histogram(~ DeathsPerYear | HDI_Banding,
          data=fullMerge2,
          layout=c(1,3))

kruskal.test(DeathsPerYear ~ HDI_Banding, data = fullMerge2) # REJECT
NULL HYPOTHESIS

# RESULT: With significance level of 0.05, there is not enough evidence to
reject null hypothesis with p-value of 0.6851
#         and conclude that mean is same for each HDI banding which is
consistent with historgram we generated
#-------------------------------------

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

#------- Objective 2: --------#
# H0: Mean of plane crash deaths for HDI Banding 1, 2 and 3 are equal
# H1: Mean of plane crash deaths for HDI Banding 1, 2 and 3 are NOT equal
#-----------------------------#

# Summarise data by
fullMerge2_chisq <- sqldf("SELECT  Year_Banding
                          ,Survived
                          ,sum(DeathsPerYear) as DeathsPerYear
                  FROM fullMerge2
                  where HDI_Banding <> '1'
                  group by Year_Banding, Survived")

fullMerge2_chisq

Chisq_table <- matrix(c(555,786,527,770,960,1410,834,525),nrow=4,ncol=2)
Chisq_table
chisq.test(Chisq_table, correct=FALSE)

prop.test(Chisq_table) # Proportion for groups 1 to 4

# RESULT: With p-value of 2.2e-16, we conclude that null hypothesis is
rejected and conclude year and Survival in plane crash are dependent
#-------------------------------------

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

#------- Objective 3: --------#
# H0: Mean of plane crash deaths for HDI Banding 1, 2 and 3 are equal
# H1: Mean of plane crash deaths for HDI Banding 1, 2 and 3 are NOT equal
#-----------------------------#

#-----------------------------#
# Logistic regression
#-----------------------------#
Decade_1990=fullMerge2$Decade_Banding=="1990 - 1999"
Decade_2000=fullMerge2$Decade_Banding=="2000 - 2008"

Decade_1990=as.numeric(Decade_1990)
Decade_2000=as.numeric(Decade_2000)

HDI_1=fullMerge2$HDI_Banding=="1"
HDI_2=fullMerge2$HDI_Banding=="2"
HDI_3=fullMerge2$HDI_Banding=="3"

HDI_1=as.numeric(HDI_1)
HDI_2=as.numeric(HDI_2)
HDI_3=as.numeric(HDI_3)
```

```
glm_model1 <- glm(Survived~HDI_1,data = fullMerge2, family =
"binomial")  #REJECT
summary(glm_model1)

glm_model2 <- glm(Survived~HDI_2,data = fullMerge2, family =
"binomial")  #REJECT
summary(glm_model2)

glm_model3 <- glm(Survived~HDI_3,data = fullMerge2, family =
"binomial")  #REJECT
summary(glm_model3)

glm_model4 <- glm(Survived~Decade_1990,data = fullMerge2, family =
"binomial")  #ACCEPT
summary(glm_model4)

glm_model5 <- glm(Survived~Decade_2000,data = fullMerge2, family =
"binomial")  #ACCEPT
summary(glm_model5)


confint.default(glm_model4) # Confidence Intervals - Model 4
confint.default(glm_model5) # Confidence Intervals - Model 5

ggqqplot(fullMerge2$Count_Crashes, ylab="Deaths from Plane Crash
Samples (#)") #distribution check

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

################################                THE END
###################################

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

# Summarise data by
test <- sqldf("SELECT  count (country) as country
                FROM fullmerge1_final")
test


#------- Attempt to Linear regression -------#

ggqqplot(fullMerge2$Count_Crashes, ylab="Deaths from Plane Crash
Samples (#)") #distribution check


plot(fullMerge2$DeathsPerYear)
boxplot(fullMerge2$DeathsPerYear ~ fullMerge2$HDI_Banding, xlab =
"HDI Banding", ylab="Deaths from Plane Crash (#)")

shapiro.test(fullMerge2$DeathsPerYear) #Normality test
skewness(fullMerge2$DeathsPerYear)     # Skewness before transformation


hist(fullMerge2$DeathsPerYear, xlab="Deaths Per Year", main="Before Log
Transformation")        # Shows if the data is normally distributed - it is right
skewed
hist(log(fullMerge2$DeathsPerYear), xlab="Deaths Per Year - Log
Transformation", main="After Log Transformation")  # Checking to see if
log10 transformation is normally distributed

summary(fullMerge2$DeathsPerYear)
summary(log(fullMerge2$DeathsPerYear))

#****** Log Transformation *****

fullMerge2$Log_DeathsPerYear        <- log(fullMerge2$DeathsPerYear)
fullMerge2$Log_InjuriesPerYear <- log(fullMerge2$InjuriesPerYear)

#------- After Transformation -------#

shapiro.test(fullMerge2$Log_DeathsPerYear) #Normality test
skewness(fullMerge2$Log_DeathsPerYear) # Skewness after transformation

par(mfrow=c(1,1))
hist(log(fullMerge2$DeathsPerYear), xlab="Deaths Per Year - Log
Transformation", main="After Log Transformation")  # Checking to see if
log10 transformation is normally distributed

plot(log(fullMerge2$DeathsPerYear), ylab="Deaths Per Year - log
Transformation")

#------------------------------#
# Estimate the parameters
#------------------------------#

ggplot(fullMerge2, aes(Year_Banding, Survived)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)

## Model Creation
model <- glm(Survived ~ Count_Crashes,family=binomial,data=fullMerge2)
summary(model)

anova(model, test="Chisq")

ci = confint.default(model); ci

# Summarise data by
model1 <- lm(DeathsPerYear ~ Decade_Banding+HDI_Banding,
data=fullMerge2)
summary(model1)

model2 <- lm(InjuriesPerYear ~ HDI_Banding, data=fullMerge2)
summary(model2)

summary(lm(InjuriesPerYear ~ I(HDI_Banding == 1) + I(HDI_Banding == 2)
+ I(HDI_Banding == 3), data = fullMerge2))

predict(model2, interval = "confidence", level = 0.95)

Deaths.fitted <- fitted(model2)

plot(Deaths.fitted, fullMerge2$InjuriesPerYear, xlab="Fitted Deaths from
plane crash (#)", ylab="Actual Deaths from plane crash (#)")
abline(0,1, lty=2)

Deaths.resid <- resid(model2)

qqnorm(Deaths.resid)
qqline(Deaths.resid)

plot(Deaths.resid, type="b", ylab="Residuals")
abline(h=0, lty=2, col="grey")

dwtest(model1)

plot(Deaths.fitted, Deaths.resid, xlab="Fitted Deaths from plane crash (#)",
ylab="Residual")
abline(h=0, lty=2)

Anova(model1, model2)

#------- Attempt to try Robust regression -------#

robust_model = lmRob(formula = Log_DeathsPerYear ~ HDI_Banding,
data=fullMerge2)
summary(robust_model)

summary(rr.huber <- rlm(formula = Log_DeathsPerYear ~ HDI_Banding,
data=fullMerge2))

fit.mod = dynlm(Log_DeathsPerYear ~ HDI_Banding)

fit.compare = fit.models(list(Robust = "lmRob",
                              "LS" = "lm"), formula = Log_DeathsPerYear ~
HDI_Banding, data = fullMerge2)
summary(fit.compare)
```