EXECUTIVE SUMMARY

This paper aimed to look at how life expectancy has changed between 1997 and 2016 and look at the relationship human development has on it. The purpose of this report was to test the following:

- Hypothesised that life expectancy has increased between 1997 and 2016.
- Hypothesized that there is a difference in life expectancy between human development in the year 2016.
- Hypothesised that larger percentage changes in human development led to larger percentage changes in life expectancies.

The data was sourced using gapminder and various tests were conducted, such as a t-test, Kruskal wallis test and a linear regression. Overall this study found that as human development increased, so too does life expectancy, though this increase occurs with diminished returns, that is countries which rank lower in development are able to have a greater percentage increase in life expectancy. As the world becomes more globalised and countries have better access to resources it suggests that overpopulation may become worse as people may live longer, stretching resources even further and the need for government intervention to ensure it is managed successfully.

INTRODUCTION

The human development index (HDI) is a statistic composite index of life expectancy, education, and per capita income indicators, which are used to rank countries into four tiers of human development. The HDI was developed by the United Nations in 1990 as a statistical measurement of human development. Prior to 1990 human development was measured by economic magnitude, though this method was considered too narrow to define human development.

Life expectancy is considered one of the most precise measurements among the components of the HDI, as other components of HDI such as GNI per capita comprise of measurement error. This paper aims to look at how life expectancy has changed between 1997 and 2016 due to the improvements in global development, attempting to demonstrate the relationship between life expectancy and the HDI.

DATA

Two data sets were downloaded from the Gapminder website (https://gapminder.org/data/) and later imported into R studio for pre-processing.

Human Development Data

The HDI data was downloaded from the 'society' folder and imported into R studio. The initial HDI data set comprised of 188 observations and 30 variables prior to data processing, with observations from 1990 until 2018. The HDI data set was processed in the following way:

- A second HDI data set was created by removing all na values utilising the na.omit function in R studio. Removal of na values resulted in 142 observations and 30 variables.
- A third HDI data set was created by creating a data frame of HDI data from 1997 to 2016.
- A new variable was created by taking the mean of HDI for each observation across the rows for the years between 1997 and 2016.
- Mean HDI was categorised into one of four groups (Dasic et al., 2020):
  - 0.00<HDI<=0.55: 1
  - 0.55<HDI<=0.70: 2
  - 0.70<HDI<=0.80: 3
  - 0.80<HDI<1.0: 4
- A final data set (hdi_3) was produced of HDI data which comprised of 3 variables and 142 observations. This data set

contained HDI as a categorical variable and HDI as a quantitative variable.

Life Expectancy Data

The life expectancy data was sourced from the 'health' folder and imported into R studio. The initial life expectancy data set compromised of 302 variables and 187 observations, from 1800 to 2100. The original life expectancy data set comprised of observed data between 1800 to 2016 and projected data from 2017 to 2099. The life expectancy data set was processed in the following way:

- Data loaded into R studio.
- Missing observations removed by na.omit function.
- Final data frame constructed (life_expectacny_2) with the following variables, country, 1997 life expectancy and 2016 life expectancy.

Complete data set

After the HDI and life expectancy data were processed as above, we then created our final data frame in R studio. An inner join utilising the *merge* function was conducted linking HDI_3 to life expectancy by country. Variables were renamed to make sense and the final data frame consisted of 142 observations and 5 variables.

Quantitative variables:

- Life expectancy 1997
- Life expectancy 2016
- Mean HDI between 1997 and 2016

Categorical variable

- HDI categorical variable

It should be noted that spelling variations in the country column between the HDI data set and life expectancy data set resulted in the loss of a small fraction of observations, considered negligible with such a large sample.

Methods

Three analytical tests were conducted on the completed data set. The statistical analysis of data included a, (1) t test, (2) kryskal-wallis test and (3) a non/linear regression analysis which was done with R and Rstudio(version 1.4.1106). All tests were conducted on the complete data set originally produced, with some additional data processing done.

The single tailed t test:

This test was used to test to determine if there had been an increase in life expectancy between 1997 and 2016.

The first step of the t test was to determine a hypothesis. The proposed hypothesis of the test was that the null hypothesis H0: The mean of differences between life expectancy between the two years 2016 and 1997 is 0 and the alternative hypothesis HA: The mean difference between life expectancy between the two years 2016 and 1997 is not 0 and has instead increased. This hypothesis assumes that over time there has been increases in many factors leading to favourable health outcomes. On rejection of the null hypothesis, the alternative is adopted and that there is evidence to suggest life expectancy has increased between 1997 and 2016.

The next thing to do was test for outliers calling on the boxplot(), followed by the qnorm() and qqline() function to test the data for the normality assumption.

After confirmation of normality and the removal of any possible outliers a new variable was created by taking the difference of 2016 life expectancy and 1997 life expectancy. Finally, the t.test() function was called for to test if the difference if life expectancy was greater than zero.

Kruskal Wallis test:

This test was conducted to determine if HDI influences life expectancy. This test looked at life expectancy in the 21st century using the 2016 life expectancy observations against the categorical HDI value.

The first step of the Kruskal wallis test was to determine a hypothesis. The proposed null hypothesis H0: That all groups come from the same population and HDI does not influence life expectancy in the 21st century, and the alternative hypothesis HD: That at least two of the populations are difference.

The next step was to ensure the assumptions of the Kruskal wallis test were not violated.

Utilising the nrow() function in R studio we were able to obtain individual counts of observations for when HDI categorical were set to 1,2,3 and 4, testing for the assumption that each category has 5 or more individual observations.

Finally, the Kruskal.test function was called for to test for the null hypothesis utilising the generic function.

Regression and correlation

The next question which was asked was if there was a relationship between a change in HDI and a change in life expectancy between 1997 and 2016. For this analysis, the percentage change for both variables were used to ensure equal scale.

The first step in this analysis was to view the data set and create the new variables. The new variables created are the percentage change for life expectancy, and percentage change for HDI between 1997 and 2016 using equation 1.

$$\% Change = 100 * \left(x_2 - x_1\right) / x_1$$

Equation 1: Percentage change equation

The next step was to naturally run a correlation, calling on the generic cor() function to determine the possible correlation and strength between percentage change in HDI (% HDI) and percentage change in life expectancy (% Life).

With evidence of a possible relationship exploratory analysis was conducted by visualisation. The plot() function was called for to visualise any possible relationship, and the boxplot() function was called for the detection of outliers.

Both % HDI and % Life had some evidence of extreme outliers and the Interquartile range (IQR) method rule was implemented to detect for possible outliers. The IQR method utilised the equation below to detect and remove values below the lower limit, and above the upper limit.

$$Q \leftarrow IQR\left(data\$varible, probs=c\left(.25,.75\right), na.rm=FALS\right.$$
$$iqr \leftarrow IQR\left(data\$varible\right)$$
$$uper\, r_{limit} \leftarrow Q\left[2\right] + 1.5 * iqr$$
$$Lowe\, r_{limit} \leftarrow Q\left[1\right] - 1.5 * iqr$$

A new data frame was then created, which resulted in the sample reducing in number of observations from 142 observations to 125 observations. All further analytics was done on this new data frame which had removed the outliers. The lm() function was called for, modelling % HDI as the predictor variable and % Life as the repose variable. The summary() function was called for the lm model produced. We then called for the fitted() function to test the fitted values of the linear model by looking at the discrepancy in % HDI.

The next step in the process was to test the assumptions of the linear model produced. The first step conducted was by calling the resid() function and plotting it utilising qqnorm() and qqline() to test for normality. We then tested for constant variance by plotting the fitted values against the residuals, calling for the plot() function and inserted an abline(). Finally, we check the independence of the residuals by plotting them in the sequence in which they appear in the data via the plot() function of the residuals.

We then called on the summary() function of the linear model to test for the significance of the model, the R-squared value and the significance of the slope.

Results

The single tailed t test: Has life expectancy increase between 2016 and 1997:

Exploratory analysis was conducted to test the validity of the assumptions. A Q-Q plot was produced (figure 1) testing for normality and as seen in figure 1 the data for life expectancy in both 1997 and 2016 appears approximately normal with skewness towards the higher values. However, when the sample size is large, t procedures are often robust against deviations from normality (The Basic Practise of Statistics, Third Edition, page 426). With the large number of data points (n=142) it allows us to invoke the theory of robustness for this procedure.

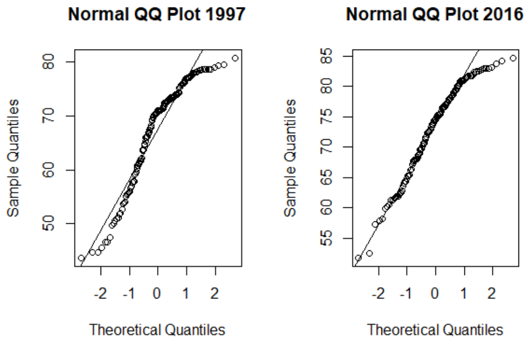

Figure 1: Normal Q-Q Plot for 1997 and 2016 life expectancy.

Boxplots are also produced to detect for outliers, and as seen in figure 2 no outliers are detected for life expectancy in 1997 and 2016. It also appears that the spread is somewhat identical as per the length of the whiskers on either side of the boxplot.

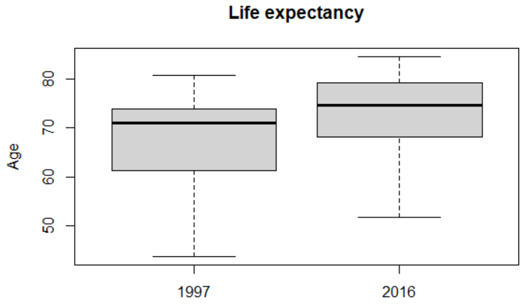

Figure 2: Box plot for 1997 and 2016 life expectancy.

Based on the assumption validity, and the single tailed t test it appears we reject the null hypothesis and adopt the alternative, that is there is evidence to suggest that life expectancy has increase between 2016 and 1997 (p=2.2x10^-16, 1=17.622, df=141).

Kruskal Wallis Chi square test

Exploratory analysis confirmed the validity of the assumptions. Table 1 demonstrates that in each HDI sample there were more than 35 individual observations.

Table 1: Frequency count of observations by HDI category.

HDI	Low	Medium	High	Very High
Count	38	33	34	37

Further visual analysis was conducted, and a boxplot was produced demonstrating visually that as HDI increases life expectancy also tends to increase with decreasing spread as indicated by the whisker length in figure 3.

Based on the assumption validity and the Kruskal-wallis rank sum test it appears we reject the null hypothesis and adopt the alternative, that there is evidence to suggest that life expectancy differs between at least two of the populations (p=2.2x10^-16).

Linear model

After removal of the outliers, figure 4 shows a possible linear relationship between % HDI and % Life.

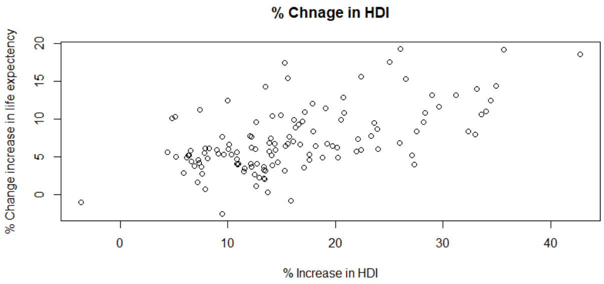

Figure 4: Scatterplot between % Increase in HDI and % Increase in life. Showing a possible postive relationship.

The proposed linear model produced in R studio gave the initial equation for the model produced.

$$y_{\%\,Life} = 2.24 + 0.304\,x_{\%HDI}$$

The next step was to test for the assumptions of the linear model by looking at the assumptions of the residuals. Figure 5 suggests that the residuals are approximately normally distributed.

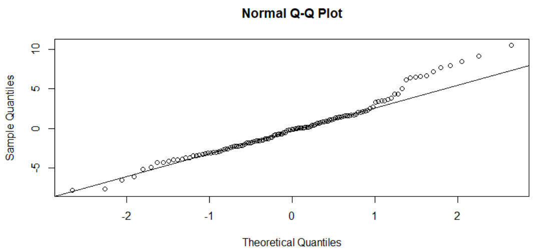

Figure 5: Test of normality for residuals of linear model. Suggesting the residuals are approximately normally distributed.

In figure 6 we observe that we are unable to see a pattern between the residuals against the fitted values, thus we can assume constant variation.

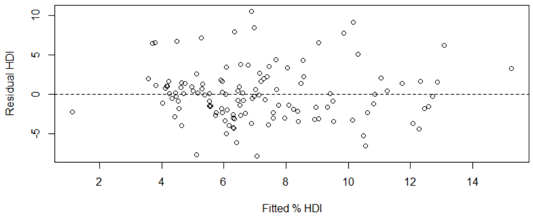

Figure 6: Scatter plot to test for constant variance of the residuals against fitted model.

Finally, figure 7 suggests that there is no discernible pattern in the residuals, thus the assumption of impendence for the residuals is valid.

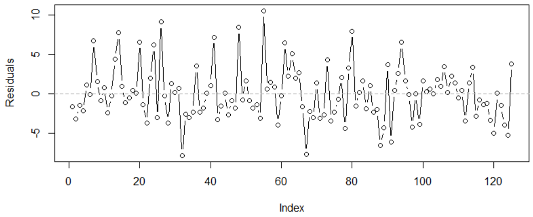

Figure 7: Plot of residuals in the order which they apper. No discernible pattern present.

From the summary output of the linear model, we observe the R Squared value at 0.3504 (F=66.3, p<0.05) and the population slope being significantly different to zero (t=8.145, p<0.05).

Overall, the linear model satisfies the assumptions as demonstrated, though weakly models % Life from % HDI, though suggests a relationship does exist. The null hypothesis is rejected, and the alternative is adopted, thus the slop is not equal to zero in the linear model.

Discussion

The aim of this paper was designed to look at how life expectancy has changed between 1997 and 2016. There was evidence to suggest that life expectancy has increased. It was also demonstrated that there is a difference in life expectancy between countries depending on their level of human development. Finally, it could be suggested based on the linear model that a country which undergoes greater development will lead to a greater percentage change in life expectancy, looking at the raw data we observed that countries with a higher percentage increase in HDI tend to come from countries with a low HDI ranking, suggesting diminishing returns to life expectancy change as development improves. This would suggest that as a country moves up in HDI ranking, whilst they may live longer, the improvements are not as great percentage wise.

R Code:

Clear working enviroment

~~~
```{r}
rm(list=ls())
```
~~~

# HDI Dataset possessing BEGIN

##Step 1: load data

##Step 2: Remove na values

##Step 3: Design new data set 15 years 1997 - 2011

##Step 4: Find mean of HDI between 1997 and 2011

##Step 5: Categorise mean HDI

##Step 7: Clean R enviroment and rename final data frame

###Step 1: Load HDI data into working environment and test successful by str(x)

~~~
```{r}
hdi <- read.csv("hdi_human_development_index.csv", header=TRUE)
```
~~~

###Step 2: Remove na values

~~~
```{r}
hdi_1 <- na.omit(hdi)
```
~~~

###Step 3:Built HDI data set for year 1997 - 2016

~~~
```{r}
hdi_2 <- data.frame(hdi_1$country,
            hdi_1$X1997,
            hdi_1$X1998,
            hdi_1$X1999,
            hdi_1$X2000,
            hdi_1$X2001,
            hdi_1$X2002,
            hdi_1$X2003,
            hdi_1$X2004,
            hdi_1$X2005,
            hdi_1$X2006,
            hdi_1$X2007,
            hdi_1$X2008,
            hdi_1$X2009,
            hdi_1$X2010,
            hdi_1$X2011,
            hdi_1$X2012,
            hdi_1$X2013,
            hdi_1$X2014,
            hdi_1$X2015,
            hdi_1$X2016)
```
~~~

###Rename column country

~~~
```{r}
names(hdi_2)[names(hdi_2)=="hdi_1.country"] <- "country"
~~~

~~~
```
~~~

###Step 4: Create mean of HDI, less columne 1 (country names) 1997-2011 (15 years)

~~~
```{r}

hdi_2$meanhdi <- rowMeans(hdi_2[,-1])

```
~~~

###Step 5: Categorize meanHDI value

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7507379/#:~:text=and%20now%20they%20are%20classified,level%20of%20human%20development%20and

The above provides a classification of the HDI grouping as low, med, high, very high

~~~
```{r}

hdi_2$hdicat[hdi_2$meanhdi>=0 & hdi_2$meanhdi<0.55] <- 1

hdi_2$hdicat[hdi_2$meanhdi>=0.55 & hdi_2$meanhdi<0.70] <- 2

hdi_2$hdicat[hdi_2$meanhdi>=0.70 & hdi_2$meanhdi<0.8] <- 3

hdi_2$hdicat[hdi_2$meanhdi>=0.80 & hdi_2$meanhdi<1] <- 4

```
~~~

###Create final HDI data frame with varibles of intrest for final data frame

~~~
```{r}

hdi_3 <- data.frame(hdi_2$country,

            hdi_2$hdicat,

            hdi_2$meanhdi,

            hdi_2$perchhdi)

names(hdi_3)[names(hdi_3)=="hdi_2.country"] <- "country"

```
~~~

#Life expecnty data processing: Begin

##Important information about data set: https://www.gapminder.org/data/documentation/gd004/

##Data set grouped into 1970-2016: This is observed data

##Data set grouped into 2017-2099: This is projection data from UN - we will not analyze this data

##This project will only be concenred with 1970 - 2016 data

### Step 1: load data

### Step 2: Remove na values

### Step 3: Design new data set 1997 and 2011

### Step 4: Clean R environment: remove data frames no longer required

###Step 1: load data: Load life expec data into working environment and test successful by str(x)

~~~
```{r}
~~~

~~~
life_expectancy <- read.csv("life_expectancy_years.csv", header=TRUE)

str(life_expectancy)


```



###Step 2: Remove na values

```{r}

life_expectancy_1 <- na.omit(life_expectancy)

```


###Step 3: Design new data set 1997 and 2016

```{r}

life_expectancy_2 <- data.frame(life_expectancy_1$country,

                         life_expectancy_1$X1997,

                         life_expectancy_1$X2016)

```


#Rename column country

```{r}

names(life_expectancy_2)[names(life_expectancy_2)=="life_expectancy_1.country"] <- "country"

```


#Create a new data frame

###Step 1: rename linking key in hdi_2 and life_expectency_2 data frame

###Step 2: Remove na values

###Step 3: Clean R environment: remove data frames no longer required

###Step 4: Rename varibles in final data set


###Step 1: link data frames to create one data frame

```{r}

hdi_life <- merge(hdi_3, life_expectancy_2, by="country")

```


###Step 2: Remove na values

```{r}

hdi_life_1 <- na.omit(hdi_life)

```


###Step 4: Rename varibles

```{r}

colnames(hdi_life_1)

names(hdi_life_1)[names(hdi_life_1)=="hdi_2.hdicat"] <- "hdi_categorical"

names(hdi_life_1)[names(hdi_life_1)=="life_expectancy_1.X1997"] <- "life_expectancy_1997"

names(hdi_life_1)[names(hdi_life_1)=="life_expectancy_1.X2016"] <- "life_expectancy_2016"

colnames(hdi_life_1)

```
~~~

#Test 1: Has there been a change in life expectancy from the year 1997 to 2016)

Important information about data set: https://www.gapminder.org/data/documentation/gd004/

###Step 1: Perfrom somme exploratory data analysis procedures

###Step 2: State the null and alternative hypothesis

###Step 3: State the significance level of the test

###Step 4: Choose the statstical test and check assumptions

###Step 5: Calculate the test stastic and determine if it is extreme

###Step 6:: Interpret findings

###Step 1: Perform some exploratory data analysis procedures

###Detect for outliers

~~~
```{r}
par(mfrow=c(1,1))
boxplot(hdi_life_1$life_expectancy_1997, hdi_life_1$life_expectancy_2016,
    names=c("1997", "2016"),
    ylab="Age",
    main="Life expectancy")
```
~~~

No extreme outliers are detected for 1997 and 2016.

As all points in 2016 are greater then 1997 we are curious to see

if life expecency has increase from 1997 to 2016

###Detect for normality - as we will attempt a parametric test

~~~
```{r}
par(mfrow=c(1,2))
qqnorm(hdi_life_1$life_expectancy_1997,
    main="Normal QQ Plot 1997")
qqline(hdi_life_1$life_expectancy_1997,
    main="Normal QQ Plot 1997")

qqnorm(hdi_life_1$life_expectancy_2016,
    main="Normal QQ Plot 2016")
qqline(hdi_life_1$life_expectancy_2016,
    main="Normal QQ Plot 2016")
```
~~~

The data possesses some deviations from normality. However, we note that when

the sample size is large, t procedures are often robust against slight deviations

from normality (The Basic Practice of Statistics, Third Edition, page 426).

The large number of data points (n=142) allows us to invoke this theory of robustness.

###Step 2: State the null and alternative hypothesi

Ho: mu2016 = mu1997

HA: mu2016 > mu1997

###Step 3: State the significance level of the test

This is the probability of committing a type I error. We use the common

significance level of 5% here. Therefore ?? =0.05.

###Step 4: Choose the statstical test and check assumptions

We choose a paired-sample t-test here. As usual, we choose a "t-test" here because

we do not know the population standard deviations and must estimate it through

the sample standard deviation. We choose a "paired-sample" test as we are considering

samples of male and female youth alcohol consumption from the same country.

The paired-sample t-test assumes:

1- The data are a simple random sample. This was stated in the question.

2- Observations comes from a population which is normally distributed.

Although this does not appear to be true, the large sample size (n=142)

allows us to invoke robustness properties of t procedures.

###Step 5: Calculate the test stastic and determine if it is extreme

~~~~
```{r}

# Deterimne by hand

hdi_life_1$difi <- hdi_life_1$life_expectancy_2016 - hdi_life_1$life_expectancy_1997

difi <- hdi_life_1$life_expectancy_2016 - hdi_life_1$life_expectancy_1997

difbar <- mean(difi)

s <- sd(difi)

n <- length(difi)

std.err = s/sqrt(n)

tstatistic <- difbar / std.err

print(paste("The t-statistic is",signif(tstatistic,4)))

alpha = 0.05

tstar <- qt(1-alpha, df=n-1)

print(paste("tstar is",signif(tstar,4)))


# Call for R built in t.test

zt.test=t.test(difi,mu=0,alternative = "greater")

zt.test

```
~~~~

#TEST 2: Kryskal-Wallis test

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7507379/#:~:text=and%20now%20they%20are%20classified,level%20of%20human%20development%20and

###Step 1: Perfrom exploratory data analysis

~~~~
```{r}

par(mfrow=c(1,1))
~~~~

~~~
boxplot(hdi_life_1$life_expectancy_2016~hdi_life_1$hdi_categorical)

```
~~~

###Step 2: State the hypothesis

H0: All groups come from the same population and the alternative is that at least two of the population are diff

We are investigating 4 distributions without relying on the assumption of normality and will conduct

a kruskal-wallis test.

This test requires trhat there are at least three indepdent SRSs

There are at least 5 observations in each sample

###Assumption test (count observations - boxplot)

~~~
```{r}
nrow(hdi_life_1[hdi_life_1$hdi_categorical==1,])
nrow(hdi_life_1[hdi_life_1$hdi_categorical==2,])
nrow(hdi_life_1[hdi_life_1$hdi_categorical==3,])
nrow(hdi_life_1[hdi_life_1$hdi_categorical==4,])

boxplot(hdi_life_1$life_expectancy_2016~hdi_life_1$hdi_categorical,
         names=c("Low", "Medium", "High", "Very High"),
        xlab="HDI Ranking",
        ylab="Life Expectency",
        main="Life expectancy by HDI")
```
~~~

# Step 3: Perfom test

~~~
```{r}
zk <- kruskal.test(life_expectancy_2016~hdi_categorical, data=hdi_life_1)
zk
```
~~~

We find evidence to suggest that life expectency differs between at least two populations of HDI (p=2.2x10-16)

## DELETE??

~~~
###Step 1: Produce a boxplot for outlier detection

```{r}

boxplot(life_expectancy_2$life_expectancy_1.X1997,life_expectancy_2$life_expectancy_1.X2008, names=c('1997', '2008'),

       xlab="Year",

       ylab="Age",

         main="Life expectency for 1978 and 2008 for multiple countries")

```

The boxplot produced is indicative of a possible outlier. We will utilize

IQR method to determine if we remove this outlier




################################################################################################## DELETE


##Step 1.2: Utilise IQRR method to detect for outlier for data set ##

Q <- quantile(life_expectancy_2$life_expectancy_1.X1978)

iqr <- IQR(life_expectancy_2$life_expectancy_1.X1978)

up <- Q[2]+1.5*iqr

low <- Q[1]-1.5*iqr

life_expectancy_2[which(life_expectancy_2$life_expectancy_1.X1978<30),]

life_expectancy_3 <- life_expectancy_2[-28,]
~~~

~~~
################################################################################################## DELETE


## Based on IQR method the outlier will be retained in the data set ##

################## CONFRIM ABOVE CODE ###############################

qqnorm(life_expectancy_1$X1998)

qqline(life_expectancy_1$X1998)

qqnorm(life_expectancy_1$X2008)

qqline(life_expectancy_1$X2008)


## We have decided to remove the outlier for parametric test



## Step 2: Test normality with Q-Q plot


## DELETE
~~~

#Test 3: Regression analysis

###Step 1: Ensure we have all the variables needed in the data set

###Step 2: Test for a correlation

In this section the aim was to do a linear regresion by looking at % change in both HDI and life

expectency between 1997 and 2016, that is does a large percentage change in development lead to

a large percentage chnage in life expectancy

###Step 1: Ensure we have all the varibles needed in the data set

We need to create a new varbile known as life expetecnay percetnage chnage as this is not in the final data set and

the same for HDI

###% Change HDI 1997 to 2016

```{r}
hdi_2$perchhdi <- 100*(hdi_2$hdi_1.X2016 - hdi_2$hdi_1.X1997)/hdi_2$hdi_1.X1997

hdi_life_1$perch <- 100*(hdi_life_1$life_expectancy_2016 - hdi_life_1$life_expectancy_1997)/hdi_life_1$life_expectancy_1997
```

###Step 2: Test for a correlation

Produce visual aide to see if relationship exisits

```{r}
plot(x=hdi_life_1$hdi_2.perchhdi, y=hdi_life_1$perch,

    xlab="% Increase in HDI",

          ylab="% Change increase in life expectency",

    main="% Chnage in HDI")


```

Pearson correlation test: Indicates a possible positive correlation

```{r}
```

zpe1 <- cor(x=hdi_life_1$hdi_2.perchhdi, y=hdi_life_1$perch)

zpe1

#plot(x=hdi_life_1$hdi_2.meanhdi, y=hdi_life_1$perch,

#   xlab="% Increase in HDI",

#   ylab="% Change increase in life expectency",

#   main="% Chnage in HDI")

#zpe2 <- cor(hdi_life_1$hdi_2.meanhdi, hdi_life_1$hdi_2.perchhdi)

#zpe2

###Step 3: Begin a regresion analysis

```{r}

plot(x=hdi_life_1$hdi_2.perchhdi, y=hdi_life_1$perch,

    xlab="% Increase in HDI",

    ylab="% increase in life expectency",

       main="% Change between 1997 and 2016")

```

Detection of outliers/extreme values

```{r}

par(mfrow=c(1,1))

boxplot(hdi_life_1$hdi_2.perchhdi,hdi_life_1$perch, names=c("% Change HDI", "% Change Life"))

```

Possible evidence of outliers in the data set based on prelimerly results

###Detection of outlines utilizing IQR method

```{r}

#Find Q1 and Q2 Value of % change HDI

Qhdi <- quantile(hdi_life_1$hdi_2.perchhdi, probs=c(0.25,0.75), na.rm=FALSE)

# Solve IQR % change HDI

iqrhdi <- IQR(hdi_life_1$hdi_2.perchhdi)

# Select the upper limit for % change HDI

uphdi <- Qhdi[2]+1.5*iqrhdi

# Select the lower limit for % change HDI

lowhdi <- Qhdi[1]-1.5*iqrhdi

# Find Q1 and Q2 Value of % change life

Qlife <- quantile(hdi_life_1$perch, probs=c(0.25,0.75), na.rm=FALSE)

# Solve IQR % change life

iqrlife <- IQR(hdi_life_1$perch)

# Select the upper limit for % change life

uplife <- Qlife[2]+1.5*iqrlife
```

# Select the lower limit for % change life

lowlife <- Qlife[1]-1.5*iqrlife

# Produce a new adjusted data set which removes outlines based on IQR method for % HDI

eliminated_data_1 <- subset(hdi_life_1,

hdi_life_1$hdi_2.perchhdi > lowhdi

& hdi_life_1$hdi_2.perchhdi < uphdi

& hdi_life_1$perch > lowlife

& hdi_life_1$perch < uplife)

# Produce new boxplot

```{r}

boxplot(eliminated_data_1$hdi_2.perchhdi, eliminated_data_1$perch, names=c("% Change HDI", "% Change Life"))

```

# We have removed the outleirs and will now plot the data set

```{r}

plot(x=eliminated_data_1$hdi_2.perchhdi, y=eliminated_data_1$perch,

xlab="% Increase in HDI",

ylab="% Change increase in life expectency",

main="% Chnage in HDI")

```

# Estimate parameters

# We use the lm() to find the linear regression output all the relevent info using summary()

```{r}

zmodel <- lm(eliminated_data_1$perch~eliminated_data_1$hdi_2.perchhd)

```

Call to find our equation

y(% life)=Intercept + HDI%chnage*value

```{r}

lm(formula = eliminated_data_1$perch~eliminated_data_1$hdi_2.perchhd)

summary(zmodel)

```

Based on the output of the zmodel, it would appear the regression indicates

the model does not suite lineraity as R2 is resonablly low (0.3504).

the population slope is signif diff to zero (t123=8.145, p<3.61X10^-13)

The pop R2 is sig diff to zero (F1,123=66.34, p<3.606x10^-13) residue

Compute fitted value of linear model

```{r}
hdi.fitted <- fitted(zmodel)
plot(hdi.fitted, eliminated_data_1$perch, xlab="Fited model chnageHDI", ylab="actual chnageHID")
abline(0,1, lty=2)
```

The scatter plot shows there is discrepancy between fitted values vs actual

# Compute residuals of model

```{r}
hdi.resid <- resid(zmodel)
```

# Normality of residuals

```{r}
qqnorm(hdi.resid)
qqline(hdi.resid)
```

The QQ plot shows there is slight devation from normality

# Test constant variance

```{r}
plot(hdi.fitted, hdi.resid, xlab="Fitted % HDI", ylab="Residual HDI")
abline(h=0, lty=2)
```

# Independence of the residuals

```{r}
plot(hdi.resid, type="b", ylab="Residuals")
abline(h=0, lty=2, col="grey")
```

```{r}
summary(zmodel)
```