Data Analytics Mini Project 2
M.Tech- Group 9

Name of the Student: Akansha Tyagi
Student ID: 201761001
Role: Question 1(b)

Name of the Student: Hemant Aggarwal
Student ID: 201761002
Role: Question 3(b)

Name of the Student: Vijay Deshpande
Student ID: 201761003
Role: Question 1(a)

Name of the Student: Jayant Mahawar
Student ID: 201761004
Role: Question 3(a)

Name of the Student: Madhuri Mahawar
Student ID: 201761005
Role: Question 2(a)

Name of the Student: Rashmi Maheshwari
Student ID: 201761006
Role: Question 2(b)

Indian Institute of Information Technology, Vadodara, Gujarat.

Exercise 1:

(a) Read about the Happy Planet Index on Wikipedia:

https://en.wikipedia.org/wiki/Happy_Planet_Index

Be sure to read the methodology behind the index and its criticism. Download the data for 2012 from http://www.happyplanetindex.org/data/.

Examine the distribution of the HPI variable graphically. What would be appropriate measures of center and spread of this distribution --- (mean, SD) or (median, IQR). Justify your answers.

Make scatterplots of HPI against each of the three variables on which the index is best. Comment on what you see. Will it be appropriate to use correlation to summarize the relationship of HPI with the other three variables? If yes, provide the correlations. Explain your answers. Sample correlation can be obtained using the 'cor' function in R.

Solution:

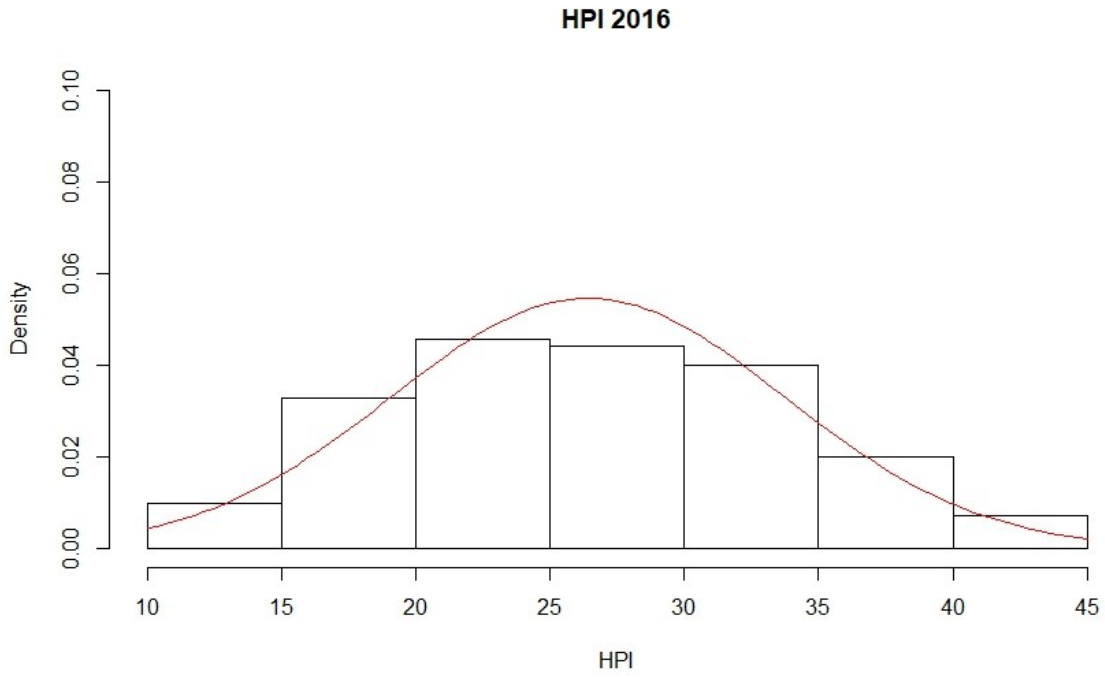

Fig. 1: Graphical view of HPI of given data

As we can see in Fig. 1, histogram plot of happy planet index, conincides with normal distribution graph.

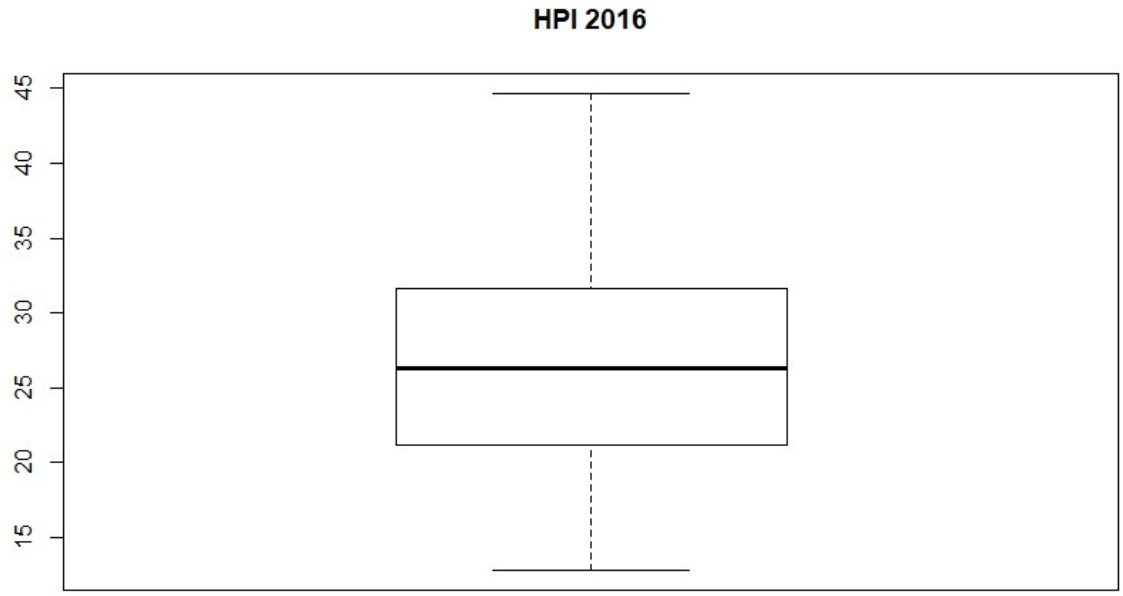

Fig 2: Boxplot of HPI

We considered mean and standard deviation as the appropriate measures of the distribution. The boxplot and histogram shows that the distribution is symmetic and approximately normal. The median coincides with the mean

b)

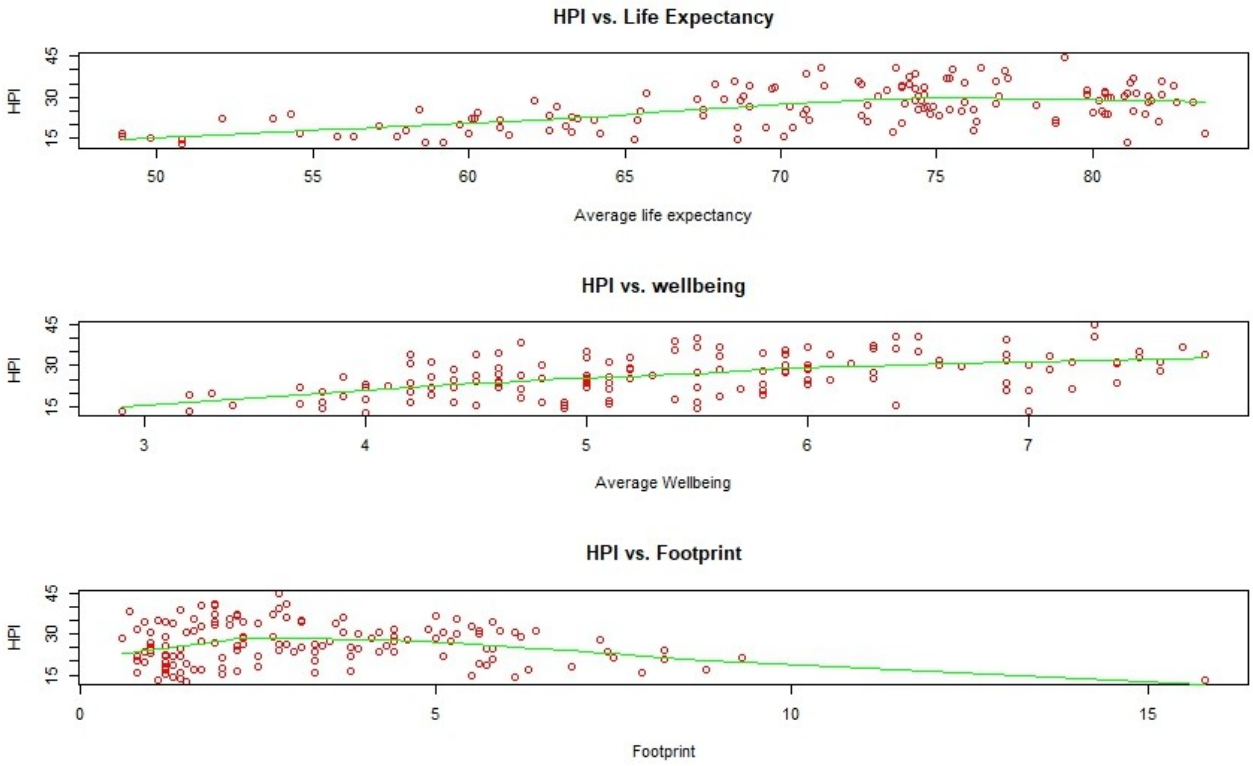

Fig. 3: HPI vs three variables

It will be appropriate to use correlation to summarize the relationship of HPI with the other three variables.
Correaltion b/w average life expectancy and HPI is 0.5405386
Correaltion b/w average wellbeing and HPI is 0.5096469
Correaltion b/w footprint and HPI is -0.1306051

Positive correlation value shows that 2 variables are directly propotional and negative value shows 2 variables are inversly propotional to each other. We can also see this realtion in Fig. 3 also.

Exercise 2:

A study shows that 61 of 414 adults who grew up in a single-parent household report that they suffered at least one incident of abuse during childhood. By contrast, 74 of 501 adults who grew up in two-parent households report abuse.

(a) Is there a difference in single-parent and two-parent households when it comes to reporting abuse? Answer this question by computing an appropriate 95% confidence interval. What assumptions, if any, did you make to compute the interval in (a)?

(b) Do the assumptions seem reasonable?

Solution:

Let A denote group of adults who grew up in a single-parent household report that they suffered at least one incident of abuse during childhood.
Let B denote group of adults who grew up in a two-parent household report that they suffered at least one incident of abuse during childhood.

Probabiltity of occurance of A (P_A) = 61/414 = 0.147
Probabiltity of occurance of A (P_B) = 74/501 = 0.148

Let us assume that the null hypothesis H0 : there is no difference in single-parent and two-parent households when it comes to reporting abuse that is
$H_0 : \pi_A = \pi_B$
Alternative hypothesis $H_A: \pi_A \neq \pi_B$

a) Using 95% confidence interval to test the null hypothesis

Confidence interval for the difference of proportions

$$= (Pa - Pb) \pm Z_{\alpha/2} \sqrt{\frac{Pa(1-Pa)}{na} + \frac{Pb(1-Pb)}{nb}}$$

Calculating the value of $Z_{\alpha/2}$ in R we get $Z_{\alpha/2} = 1.96$

$$= (0.147343 - 0.1477046) \pm 1.96 * \sqrt{\frac{0.147343*(1-0.147343)}{414} + \frac{0.1477046*(1-0.1477046)}{501}}$$

$$= -0.0003615956 \pm 0.04580191$$

Lower Limit = -0.0465251

Upper Limit = 0.04580191

Zero is between -0.0465251 and 0.04580191 so that the null hypothesis is accepted.

b) Critical Region Testing

Two sample Z-test of propartion

$$= \frac{Pa - Pb - D}{\sqrt{\frac{Pa*(1-Pa)}{na} + \frac{Pb*(1-Pb)}{nb}}}$$

$$= \frac{0.147343 - 0.1477046 - 0}{\sqrt{\frac{0.147343*(1-0.147343)}{414} + \frac{0.1477046*(1-0.1477046)}{501}}}$$

= -0.015

Acceptance and rejection regions. This is two sided test: thus we divide α by 2.
We find $Z_{\alpha/2}$ where α = 0.05 and - $Z_{\alpha/2}$ = - 1.96 and $Z_{\alpha/2}$ = 1.96
Therefore, Z lies between the range (-1.96,1.96).

Result : The evidence against H0 is insufficient because |Z| < 1.96. Although sample proportion of resposenses of abused adults are unequal, the difference between them appear too small to claim that population proportions are different.

P-value Testing

We obtained a test statistic :

Z_{obs} = -0.0.15

The P-value for this test equals

$$P = P\{|Z| \geq |-0.015|\} = 2 \times (1 - \phi(0.015)) = 2 \times (1 - 0.5059839) = 0.9880322$$

The P-value obtained is rather high (greater than 0.1) so the null hypothesis is accepted. Given H0, there is a 98% chance of observing what we really observed.

Exercise 3:

According to the credit rating agency Equifax (in US), credit limits on newly issued credit cards increased between January 2011 and May 2011. Suppose that random samples of 400 credit cards issued in January 2011 and 500 credit cards issued in May 2011 had average credit limits of $2635 and $2887, respectively. Suppose that the sample standard deviations of these two samples were $365 and $412, respectively.

(a) Construct an appropriate 95% confidence interval for the difference in mean credit limits of all credit cards issued in January 2011 and in May 2011. Interpret your results. Be sure to justify your choice of the interval.

(b) In order to minimize the risk, similar to that during 2008, if such a situation comes, Equifax also wanted to find any change in variability so as to caution the credit card company. As a data analyst at Equifax, how can you help them?

Solution:

a)

n = 400
m = 500
$\bar{x}$ = \$2635
$\bar{y}$ = \$2887
σ_1 = \$365
σ_2 = \$412

As samples are large, we assume them as normal.

Using below formula we will find confidence interval:

$$\bar{X} - \bar{Y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}$$

So the confidence interval is -252 ± 50.82 = (-302.82, -201.18)
It means in May, 2011 resulted in a \$252 increase of the mean credit limit, with a 95% confidence margin of \$50.82.

b)

Null hypothesis H_0: $\sigma_1^2 = \sigma_2^2$
Alternate hypothesis H_A: $\sigma_1^2 \neq \sigma_2^2$

$F_{obv} = (412)^2/(365)^2 = 1.27$
Degree of freedom numerator: 499
Degree of freedom denominator: 399

P-value = $P(F \geq 1.27) = 0.0063$
which is less than significance value, $\alpha = 0.05$
$F_{crit} = 1.17$
$F_{crit} < F_{obv}$
Therefore null hypothesis is rejected and the variances of both are different.

R Code:

Q1)

```
> summary(hpi2016)
> sd(hpi2016$Happy.Planet.Index)
> hist(hpi2016$Happy.Planet.Index, freq = FALSE, ylim = c(0,0.1), xlab = "HPI", main = "HPI
2016")
> curve(dnorm(x, mean = 26.41, sd = 7.31),col= "RED", add = TRUE)
> boxplot(hpi2016$Happy.Planet.Index, main = "HPI 2016")
par(mfrow=c(3,1))
> plot(hpi2016$Average.Life..Expectancy, hpi2016$Happy.Planet.Index, col = "RED", xlab =
"Average life expectancy", ylab = "HPI", main= "HPI vs. Life Expectancy") +
lines(lowess(hpi2016$Average.Life..Expectancy, hpi2016$Happy.Planet.Index), col="GREEN")
> plot(hpi2016$Average.Wellbeing..0.10., hpi2016$Happy.Planet.Index, col = "RED", xlab =
"Average Wellbeing", ylab = "HPI", main= "HPI vs. wellbeing") +
lines(lowess(hpi2016$Average.Wellbeing..0.10., hpi2016$Happy.Planet.Index), col="GREEN")
> plot(hpi2016$Footprint..gha.capita., hpi2016$Happy.Planet.Index, col = "RED", xlab =
"Footprint", ylab = "HPI", main= "HPI vs. Footprint") +
lines(lowess(hpi2016$Footprint..gha.capita., hpi2016$Happy.Planet.Index), col="GREEN")
> cor(hpi2016$Average.Life..Expectancy, hpi2016$Happy.Planet.Index)
> cor(hpi2016$Average.Wellbeing..0.10., hpi2016$Happy.Planet.Index)
> cor(hpi2016$Footprint..gha.capita., hpi2016$Happy.Planet.Index)
```

Q2)

```
>Pa<-0.147343
>Pb<-0.1477046
>na<-414
>nb<-501
>L<- (Pa-Pb)-1.96 * sqrt((Pa*(1-Pa)/na+Pb*(1-Pb)/nb))
>R<- (Pa-Pb)+1.96 * sqrt((Pa*(1-Pa)/na+Pb*(1-Pb)/nb))
```

(Calculate Confidence Interval)

```
>alpha<-0.05
>cal_alpha<-qnorm(1-alpha/2)
```

(calculate $Z_{\alpha/2}$)

```
>Z<-(Pa-Pb-D)/sqrt((Pa*(1-Pa))/na+(Pb*(1-Pb))/nb)
```

(Calculate critical region)

```
>P<-2*(1-pnorm(0.015))
```

(Calculate P-value)

Q3)

```
> pf(1.27, 499,399, lower.tail = FALSE)
> qf(c(0.05,0.95),df1 = 499, df2 = 399)
```