Shanti Greene

DATASCI 203

Lab 1

**Part 1: Multiple Choice**

1. e

2. a

3. c

4. b

5. d

6. b

7. f

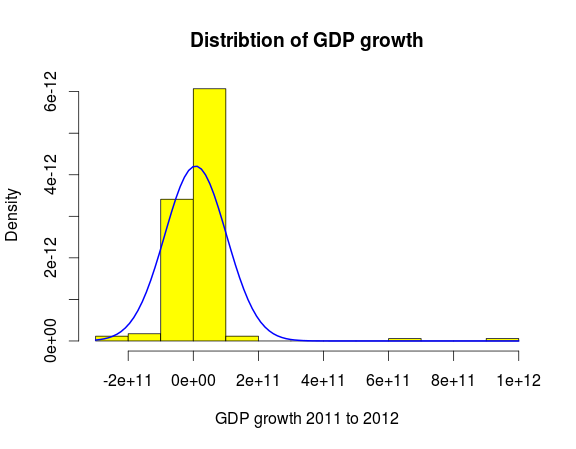
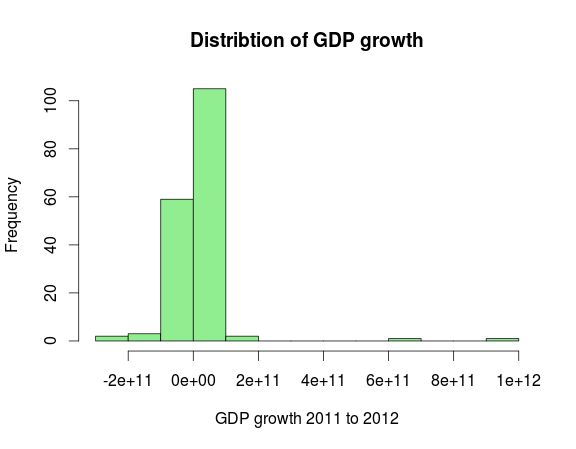
8. d

9. b

**Part 2: Data Analysis and Short Answer**

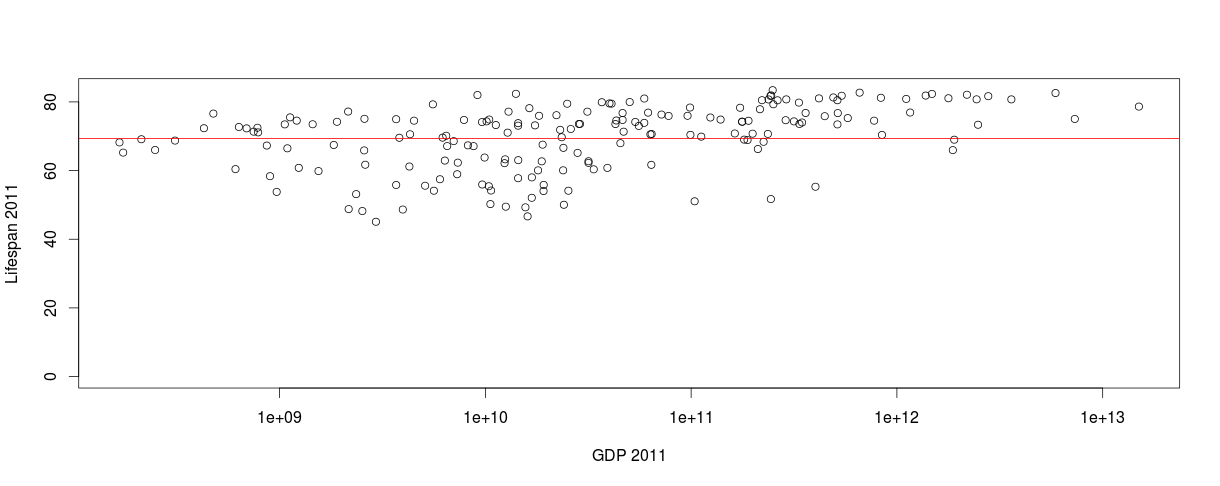
1. a. 7,172,376,796

b.



Slight positive skew, leptokurtic

c. 31 above average growth, 142 below average growth. Because of the positive skew we would expect the arithmetic mean to be skewed high, and therefore the majority of the distribution would be below the mean, which is what we see.

2. a. 

The graph shows average lifespan by country for 2011 plotted against GDP 2011 by country. The GDP axis is shown on a logarithmic scale. There does not appear to be a particularly strong relationship, but when a linear model is build and the line is added we do see a slight positive slope. Inspecting the goodness of fit statistics the coefficient for GDP is statistically significant (p<0.05), however the model can only explain around 5% of the variance in Lifespan (R2 = 0.05).

**R Code**

#import data from csv file into frame GDP\_World\_Bank

GDP\_World\_Bank <- read.csv("~/Dropbox/Berkeley MIDS/W203/Lab1\_2\_2\_2/Lab1/GDP\_World\_Bank.csv")

#add new vector "gdp\_growth\_2011\_2012" to the frame, assign the values as the difference between the 2012 and 2011 vectors

GDP\_World\_Bank["gdp\_growth\_2011\_2012"] <- GDP\_World\_Bank$gdp2012 - GDP\_World\_Bank$gdp2011

#find mean of "gdp\_growth\_2011\_2012" vector, ignoring null values assign it to variable "mean\_gdp\_growth\_2011\_2012"

mean\_gdp\_growth\_2011\_2012 <- mean(GDP\_World\_Bank$gdp\_growth\_2011\_2012, na.rm=TRUE)

#display histogram of gdp\_growth\_2011\_2012 vector (recoded in a vector, removing NA values)

grow<-subset(GDP\_World\_Bank$gdp\_growth\_2011\_2012, !(is.na(GDP\_World\_Bank$gdp\_growth\_2011\_2012)))

hist(grow, xlab="GDP growth 2011 to 2012", main="Distribtion of GDP growth", col="light green", breaks=10)

hist(grow, xlab="GDP growth 2011 to 2012", main="Distribtion of GDP growth", col="yellow", breaks=10, freq=FALSE)

curve(dnorm(x, mean=mean(grow), sd=sd(grow)), add=TRUE, col="blue", lwd=2)

#add new vector "high\_growth" to the frame, assign the value TRUE when the value of "gdp\_growth\_2011\_2012" is greater than "mean\_gdp\_growth\_2011\_2012"

GDP\_World\_Bank["high\_growth"] <- GDP\_World\_Bank$gdp\_growth\_2011\_2012 > mean\_gdp\_growth\_2011\_2012

#display frequency table for "high\_growth" vector

table(GDP\_World\_Bank$high\_growth)

#add Lifespan by country data to a new frame

LifeSpan\_2011 <- read.csv("~/Dropbox/Berkeley MIDS/W203/Lab1\_2\_2\_2/Lab1/LifeSpan\_2011.csv")

#merge lifespan and gdp frames on Country store in a new frame GDPLife

GDPLife <- merge(GDP\_World\_Bank, LifeSpan\_2011, by.x="Country", by.y="Country", all = TRUE)

#check for data that did not merge correctly

subset(GDPLife, is.na(GDPLife$Lifespan\_2011))

#import edited lifespan data

LifeSpan\_2011\_edited <- read.csv("~/Dropbox/Berkeley MIDS/W203/Lab1\_2\_2\_2/Lab1/LifeSpan\_2011\_edited.csv")

#merge edited lifespan data with GDP data into a GDPLife frame

GDPLife <- merge(GDP\_World\_Bank, LifeSpan\_2011\_edited, by.x="Country", by.y="Country", all = TRUE)

#create new data frame where we remove countries that are missing data in either lifespan or gdp2011

GDPLifeFull <- subset(GDPLife, !(is.na(GDPLifeFull$gdp2011)))

GDPLifeFull <- subset(GDPLifeFull, !(is.na(GDPLifeFull$Lifespan\_2011)))

#plot gdp2011 versus Lifespan 2011 using a log scale for gdp2011

plot(xy.coords(GDPLifeFull$gdp2011,GDPLifeFull$Lifespan\_2011), log="x", ylim=c(0,max(GDPLifeFull$Lifespan\_2011)), xlab="GDP 2011", ylab="Lifespan 2011")

#create linear model of Lifespan from gdp2011

lmodel <- lm(GDPLifeFull$Lifespan\_2011 ~ GDPLifeFull$gdp2011)

#display model fit statistics

summary(lmodel)

#add a fitted line to the graph

abline(lmodel, col="red")