Assignment 1

NAME: Nicholas Tolley
DUE DATE: February 14th, 6pm

Please post your completed homework (including the R code).

Problem I (100 pts)

This exercise focuses on the data sets country.rda, which can be found in Canvas in the Files item in the folder Data (file name: country.rda). It contains data called gapminder from the Gapminder non-profit organization. The data are 10,545 units and 9 variables. The variables are:

country

continent

year

infant_mortality: infant mortality rate (to interpret, if you divide by 10 you will obtain the percentage)

life_expectancy

fertility

population

gdp

continent

region

Read and load the data in R.

```r
library(tidyverse)
```

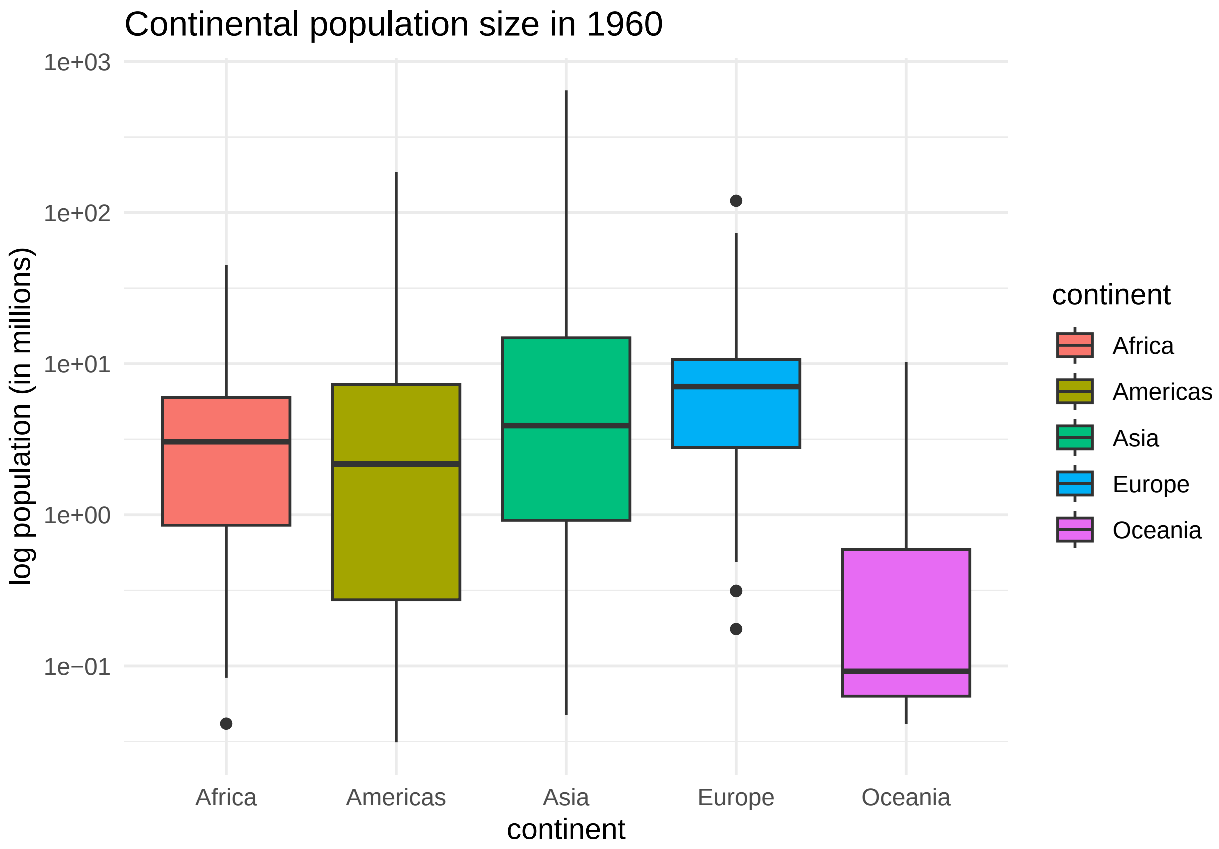
```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0     v purrr   1.0.1
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.3     v forcats 1.0.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(dslabs)
library(dplyr)

load(file='country.rda')
```

1) Perform a boxplot with ggplot in R where you can see in the x axis each continent and y axis the population just for year 1960, hint: use population/10^6 and use the log transformation. Be careful on writing the axes labels. What continent has the largest median population size? Check in R. Explain what you have found with theoretical details (outliers, symmetry of the distribution, median and interquartile range).

```
filtered_data_1960 <- filter(gapminder, year==1960)
p <- ggplot(filtered_data_1960, aes(x=continent, y=population/10^6, fill=continent)) + geom_boxplot() +
p
```



Based on the plot above, the continent with the largest median population size is Europe, meaning that countries in Europe had generally larger populations than other continents in 1960. The code below calculates the exact median population size:

```
filter(filtered_data_1960, continent=="Europe") %>%  pull(population) %>% median()
```

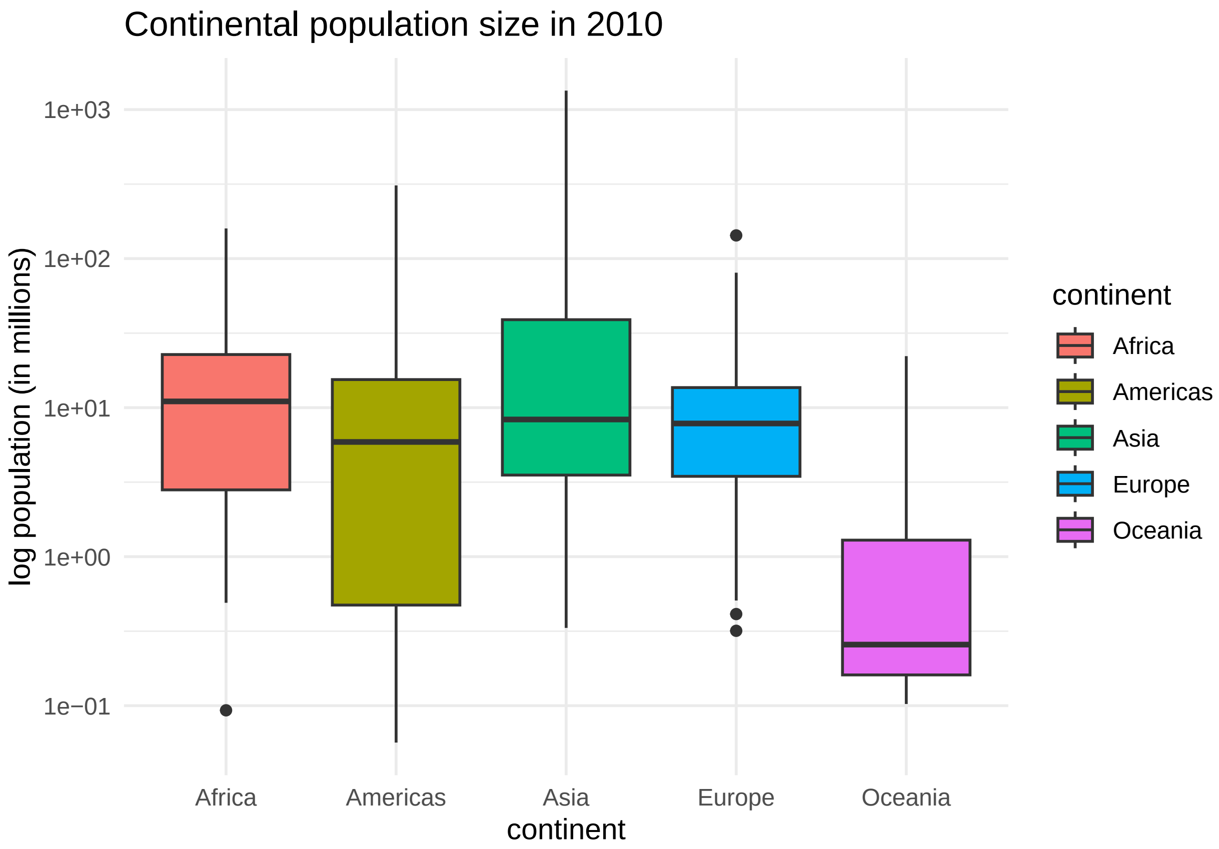
```
## [1] 7065525
```

As we can see, the median population size of Europe in 1960 was 7,065,525. In general we see that populations for each continent are symmetric around the median, that is the 1st and 3rd quartiles are roughly equidistant from the median. The notable exception being Oceania with a clear positive skew. However, noting that this data is a plotted on a log transformed y-axis, we can conclude that the country specific populations for all continents exhibit a rightward skew.

Only two continents exhibit outliers, with Africa having a country with a country population much smaller than the others. In contrast we see that Europe has outlier countries that are both larger and smaller than 1.5x the IQR (the threshold for being an outlier).

2) Perform a boxplot with ggplot where you can see in the x axis each continent and y axis the population just for year 2010, hint: use population/10ˆ6 and use the log transformation. What continent has the largest median population size in this year? Check in R. Explain what you have found with theoretical details (outliers, symmetry of the distribution, median and interquartile range).

```
filtered_data_2010 <- filter(gapminder, year==2010)
p <- ggplot(filtered_data_2010, aes(x=continent, y=population/10^6, fill=continent)) + geom_boxplot() +
p
```



3) What is the median population size for Africa to the nearest million? What proportion of countries in Europe have populations below 14 million? Perform in R.

The code below calculates the median population size for Africa

```
filter(gapminder, continent=="Africa") %>%  pull(population) %>% median(na.rm=TRUE) / 10^6
```

```
## [1] 5.570982
```

The code below calculates the proportion of countries in Europe with populations below 14 million.

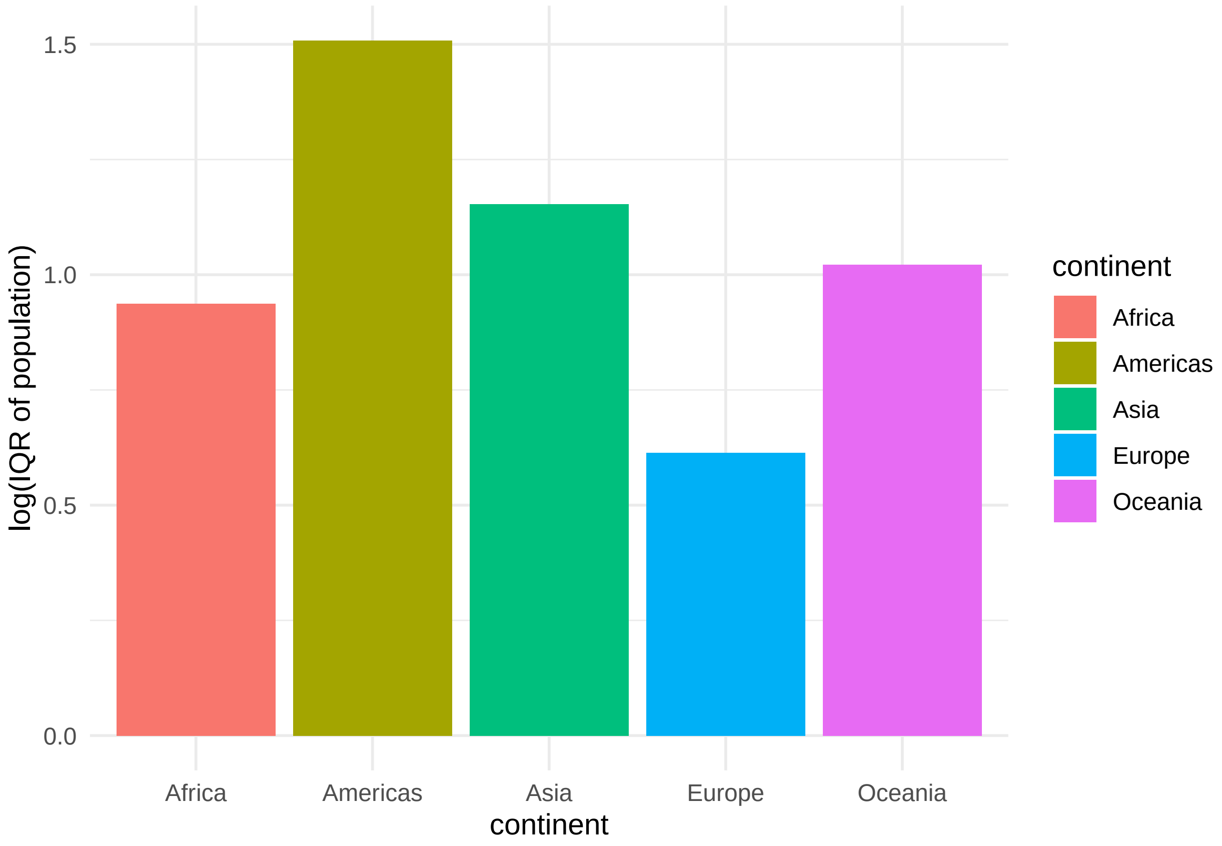
```
filtered_data_europe <- filter(gapminder, continent=="Europe")
num_countries <- nrow(filtered_data_europe)
nrow(filter(filtered_data_europe, population < (14 * 10^6))) / num_countries
```

```
## [1] 0.7390913
```

4) If we use a log transformation, which continent has the largest interquartile range? Perform and plot in R. Explain what you have found, give also the theoretical implications.

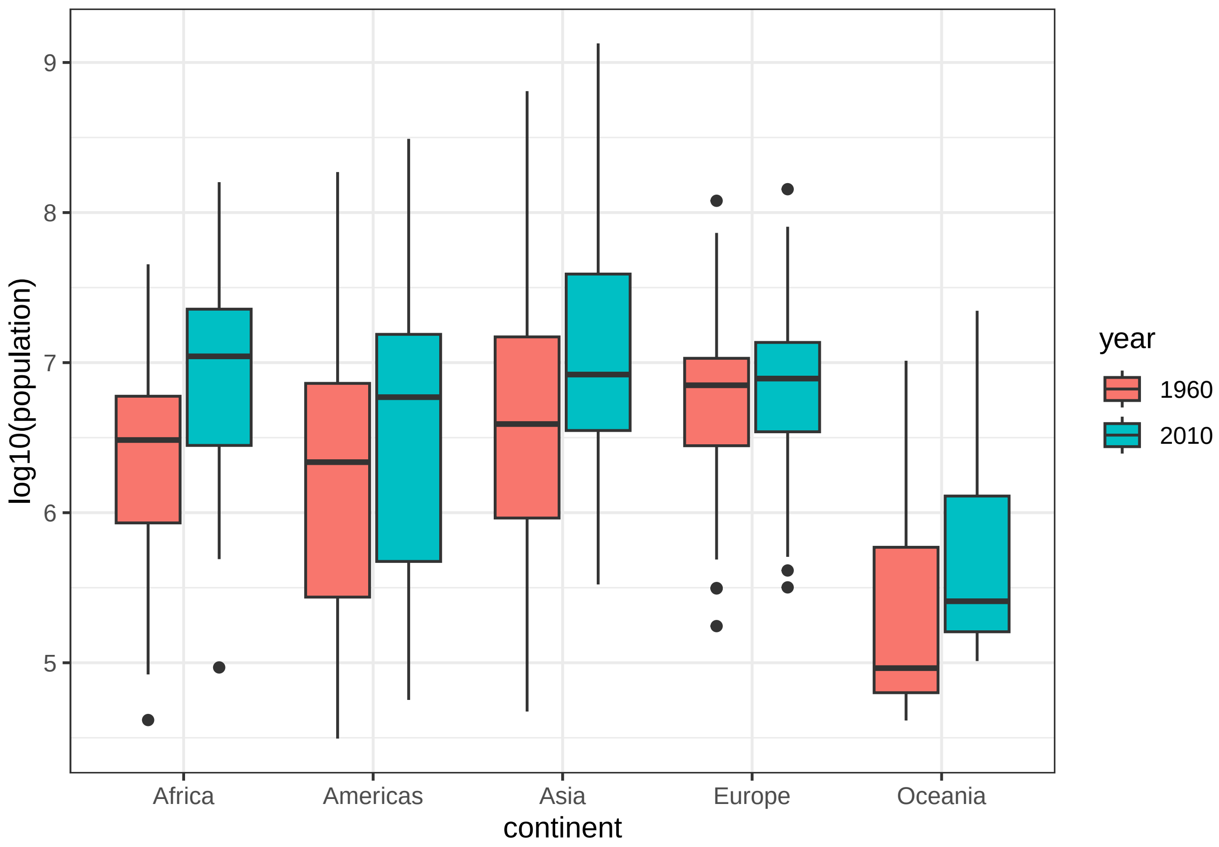
As the plot below shows, across all years in the data set the Americas show the largest interquartile range on log transformed data.

```
iqr_table <- group_by(gapminder, continent) %>% summarize(IQR=IQR(log10(population), na.rm=TRUE))
p <- ggplot(iqr_table, aes(x=continent, y=IQR, fill=continent)) + geom_col() + theme_minimal() + labs(y=
p
```



5) Compare (in ggplot) the countries in term of population using boxplots with two different colours for year 1960 and 2010. Try to make a plot exactly as Figure 1 (just the colour can change).

```
compare_df <- rbind(filtered_data_1960, filtered_data_2010)
compare_df$year <- as.character(compare_df$year)
p <- ggplot(compare_df, aes(x=continent, y=log10(population), fill=year)) + geom_boxplot() + theme_bw()
p
```



6) Compare (in ggplot) the countries in term of infant mortality using boxplots with two different colours for year 1960 and 2010. Try to make a plot exactly as Figure 2 (just the colour can change).

```
p <- ggplot(compare_df, aes(x=continent, y=infant_mortality, fill=year)) + geom_boxplot(na.rm = TRUE) +
p
```

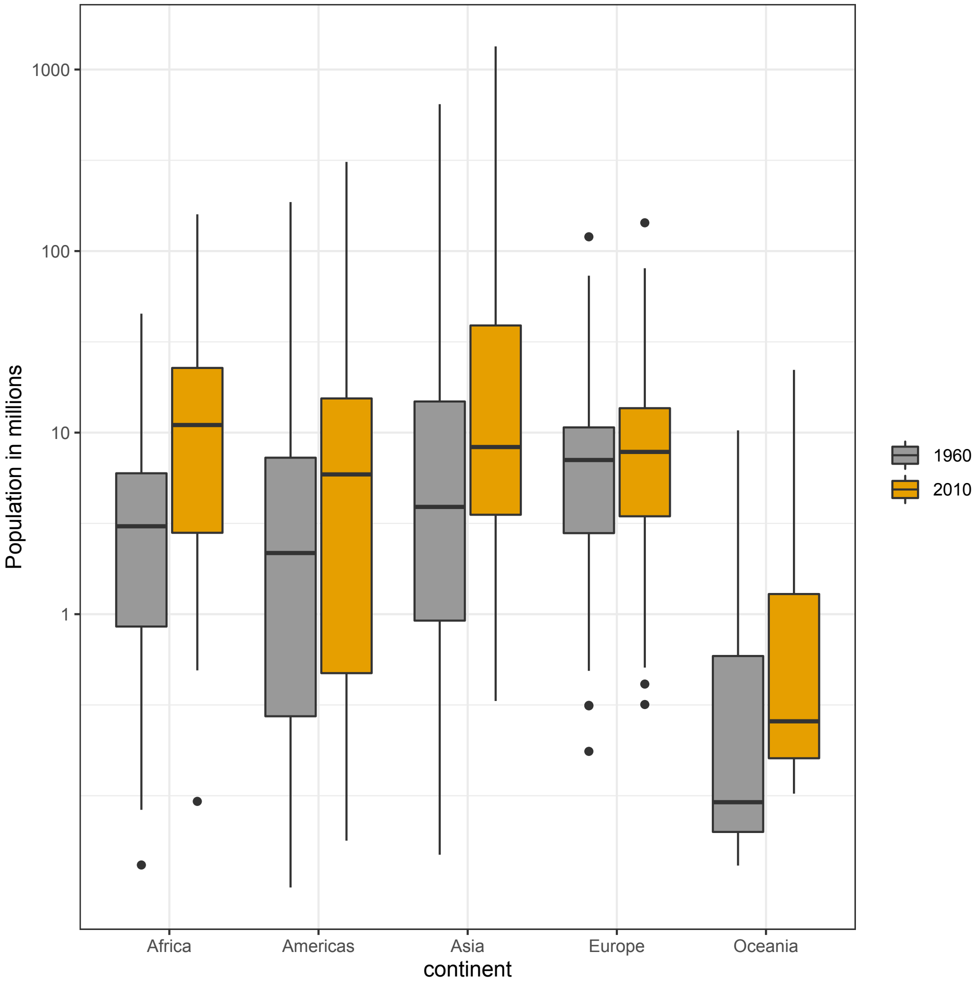
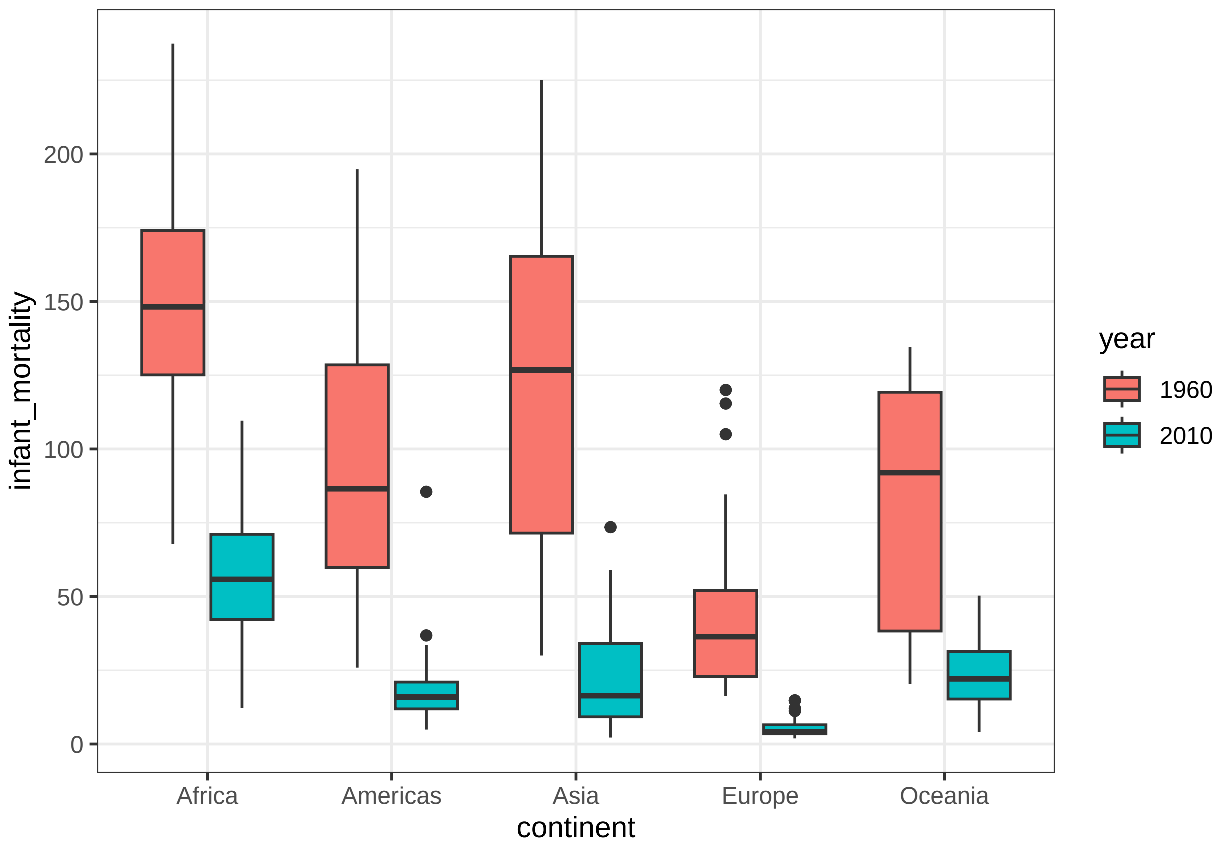


Figure 1: Population per country



7) What proportion of the infant mortality is between 100 and 150? And what is its approximation to a normal distribution of this proportion? What proportion of the infant mortality is greater than 150? And what is its approximation to a normal distribution of the proportions? Perform in R. Explain what you have found, give also the theoretical implications and what theorem or statistical assumptions you are using.

The code below calculates the proportion of infant mortality between 100 and 150 across all years/countries in the dataset. We can see that only 12.6% of the dataset fits this requirement.

```
infant_mortality_100_150 <- filter(gapminder, infant_mortality <= 150, infant_mortality > 100)
nrow(infant_mortality_100_150) / nrow(gapminder)
```

```
## [1] 0.1257468
```

To find the approximation to the normal distribution of this proportion, we simply need to calculate the mean and standard deviation of the data. This parameterizes a normal approximation, which can be visualized against a histogram of the real data.

```
mean_100_150 <- mean(infant_mortality_100_150$infant_mortality, na.rm = TRUE)
sd_100_150 <- sd(infant_mortality_100_150$infant_mortality, na.rm = TRUE)

p <- ggplot(infant_mortality_100_150, aes(x=infant_mortality)) + geom_histogram(aes(y = ..density..),fil
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
```

```
p
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
```

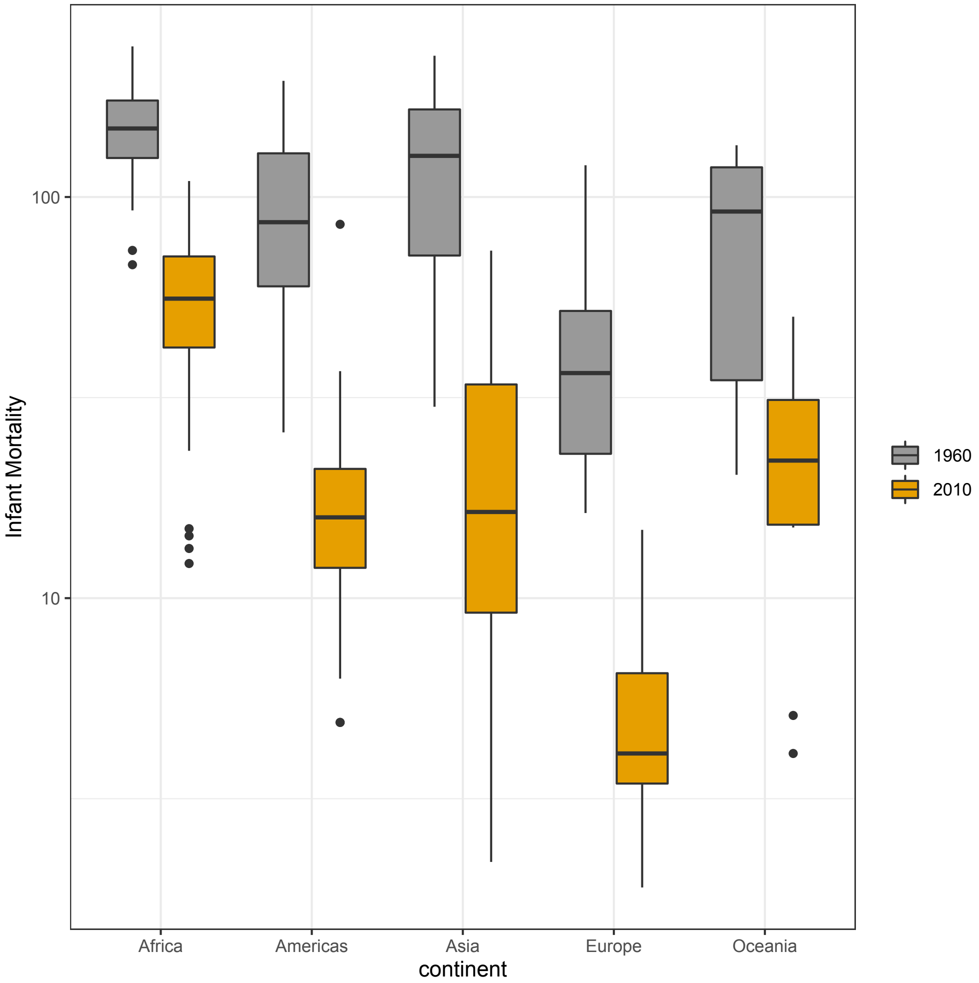
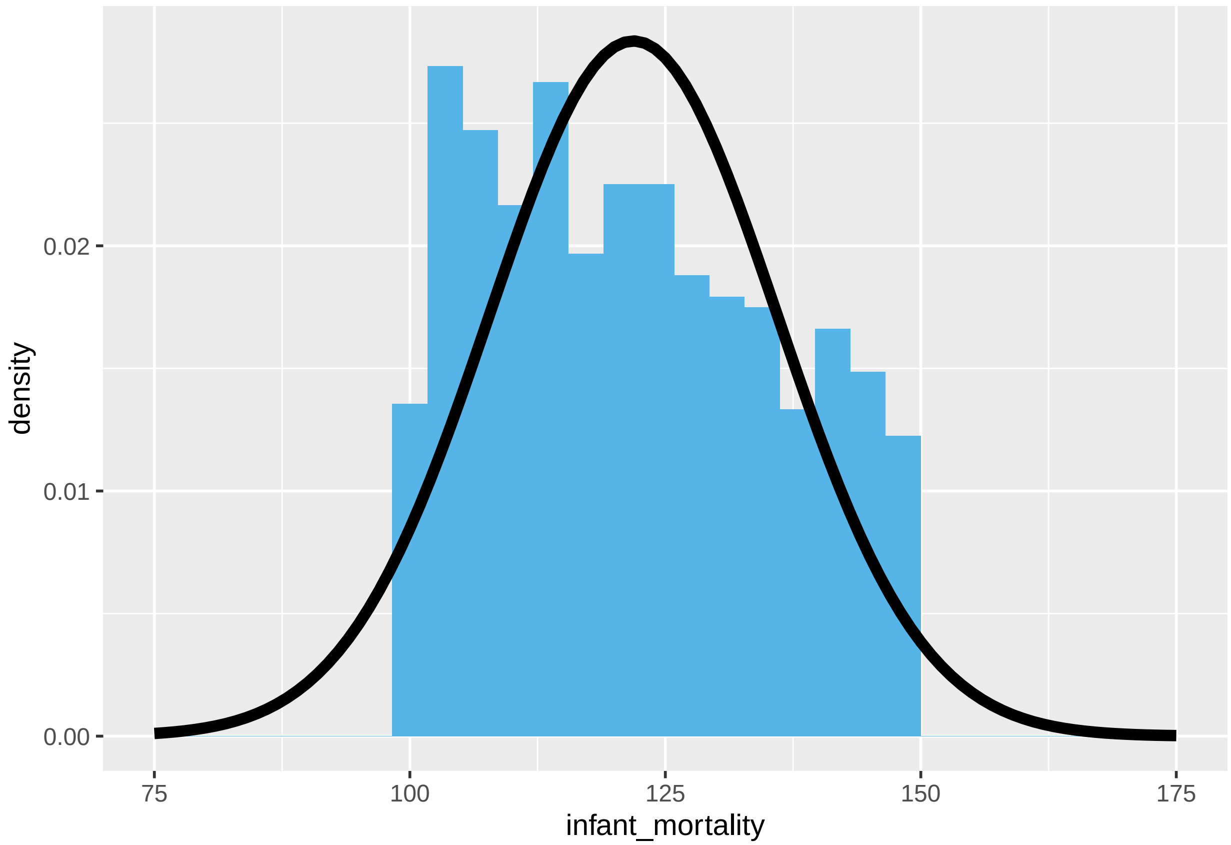


Figure 2: Infant Mortality per country

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

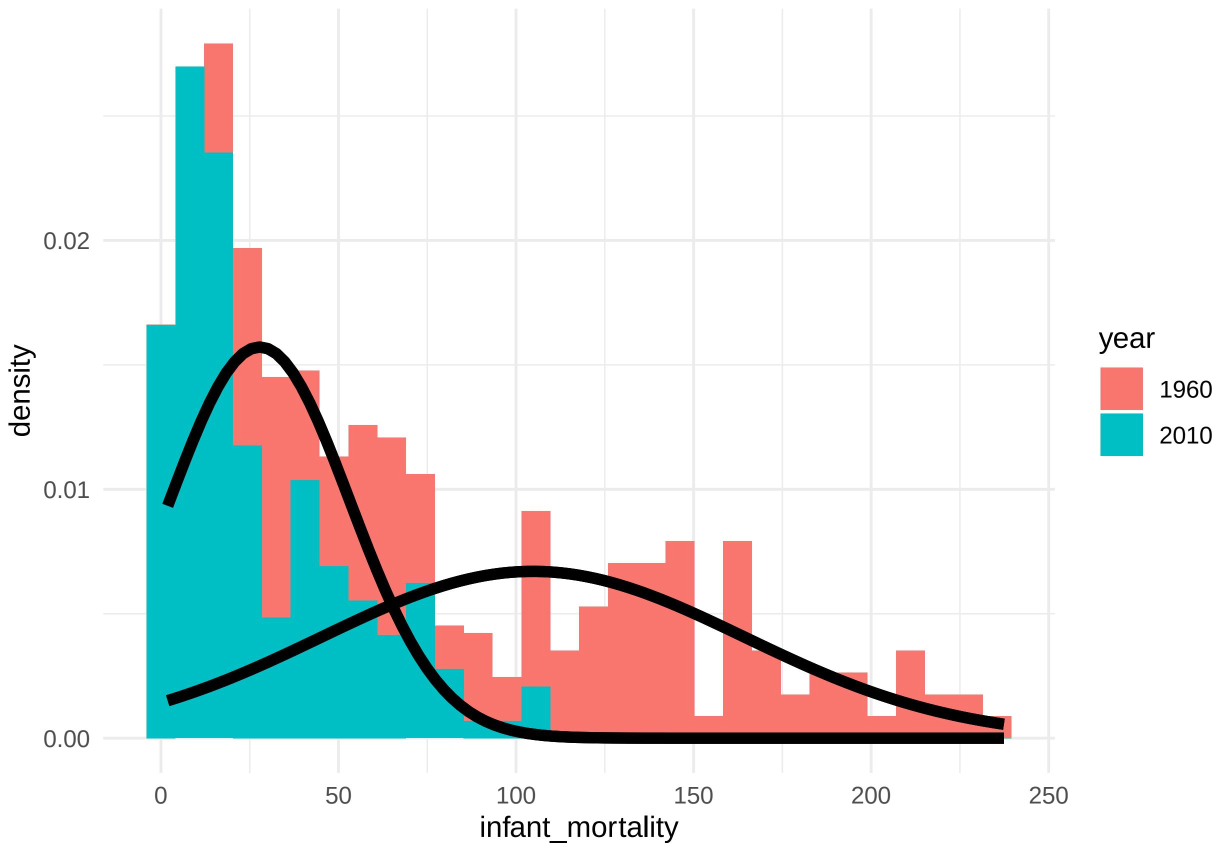


8) In ggplot perform both a histogram and a curve for infant mortality, use two different colours for year 1960 and 2010. Does the infertility follow a normal distribution? Are the two distribution symmetric or skewed? Be careful to choose the right x axis for the histogram.

```
p <- ggplot(compare_df, aes(x=infant_mortality, fill=year)) + geom_histogram(aes(y = ..density..), na.r
p
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Multiple drawing groups in `geom_function()`
## i Did you use the correct group, colour, or fill aesthetics?
## Multiple drawing groups in `geom_function()`
## i Did you use the correct group, colour, or fill aesthetics?
```



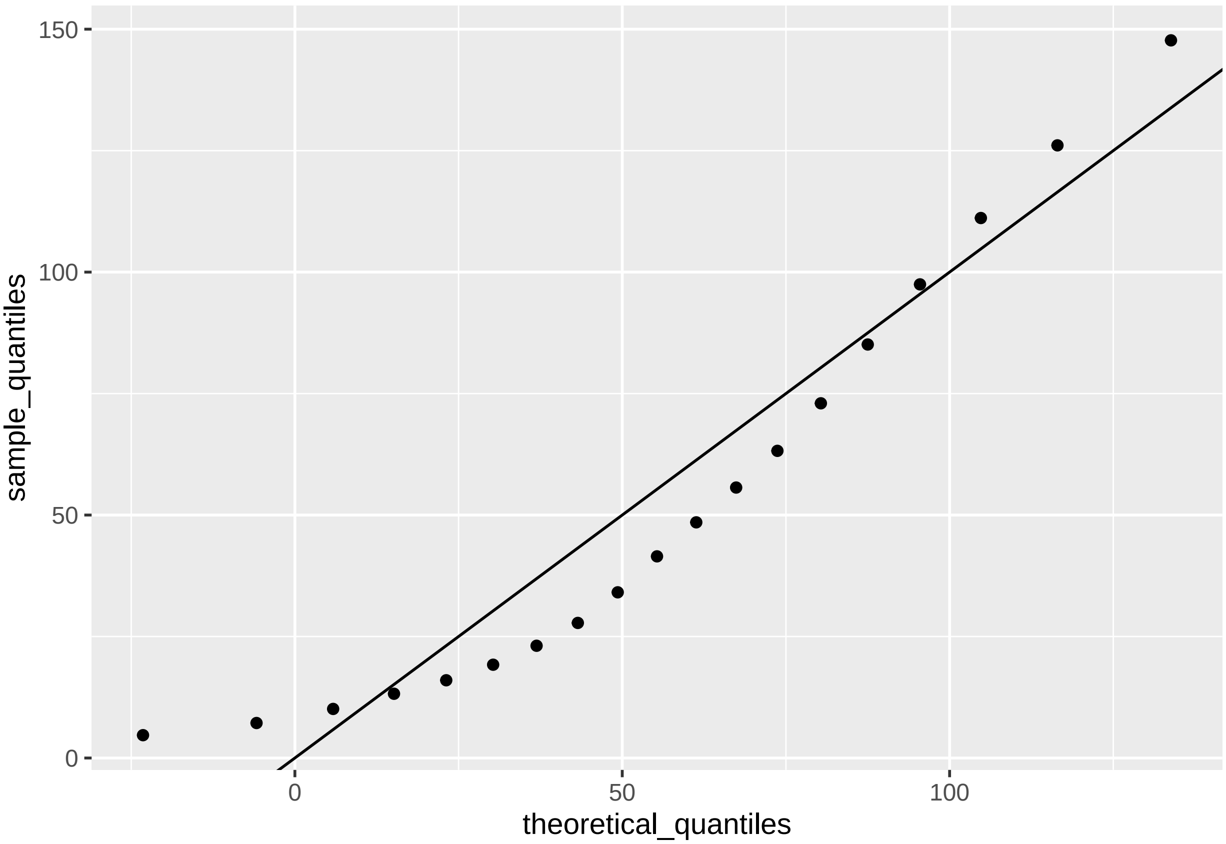
9) Perform a q-q plot for the infant mortality to check if this variables is normal or not. Write the label of the axes in the appropriate way. Explain what you have found including the theoretical implications.

```
p <- seq(0.05, 0.95, 0.05)

sample_quantiles <- quantile(gapminder$infant_mortality, p, na.rm=TRUE)
theoretical_quantiles <- qnorm(p, mean = mean(gapminder$infant_mortality,na.rm=TRUE), sd = sd(gapminder$

qplot(theoretical_quantiles, sample_quantiles) + geom_abline()
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
```



10) Compare the regions in term of infant mortality for the two years 1960 and 2010 using the barplot. Try to make a plot exactly as Figure 3 (just the colour can change). Create a binary variable, with value 1 if life-expectancy is greater than 65 (group A), 0 otherwise (group B). Then perform the appropriate two sample test to know if the mean of the infant mortality is the same in group A and B (remember to check the variance of the two groups). Explain your results.

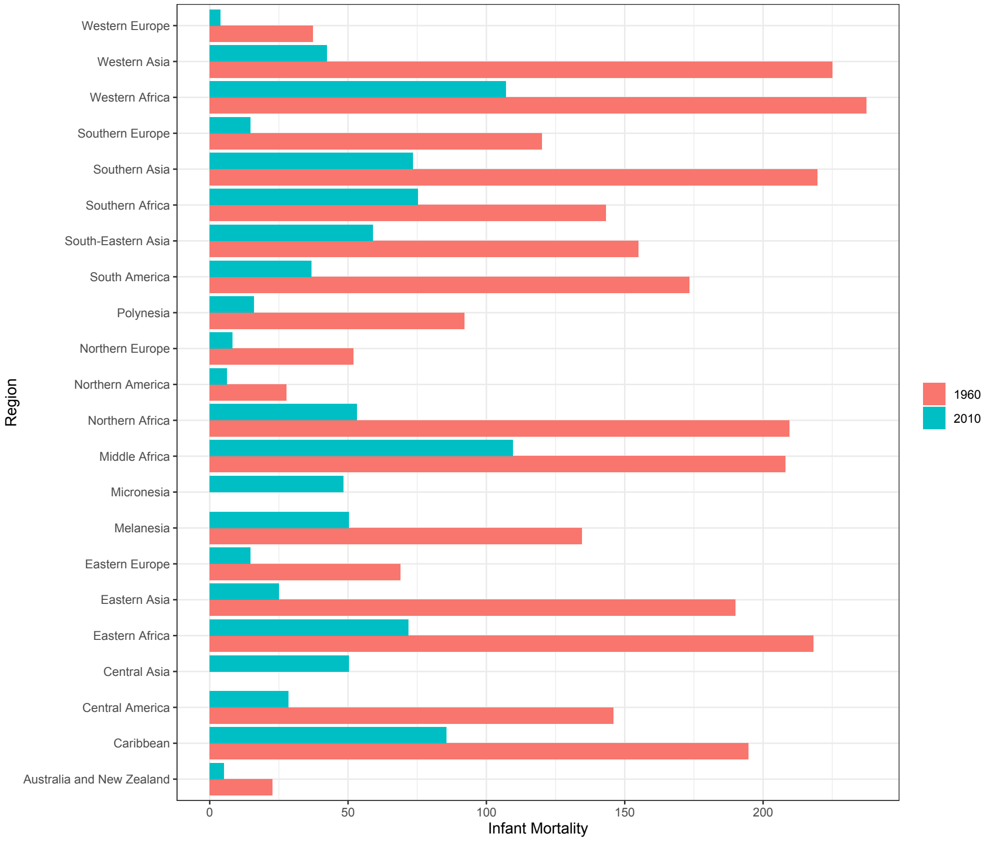


Figure 3: Infant Mortality per region