Practice set 1

```
testdf <-data.frame(row.names=c("Jack", "Rosa", "Dawn", "Vicki", "Blake", "Guillermo"),
        age=c(24, 23, NA, 25, 32, 19),
        city=c("Harlem", NA, "Queens", "Brooklyn", "Brooklyn", NA),
        gpa=c(3.5, 3.6, 4.0, NA, 3.8, NA))
testdf
```

Data Initialization

```
##           age     city gpa
## Jack       24   Harlem 3.5
## Rosa       23     <NA> 3.6
## Dawn       NA   Queens 4.0
## Vicki      25 Brooklyn  NA
## Blake      32 Brooklyn 3.8
## Guillermo  19     <NA>  NA
```

Part-A

Problem 1: Write a function of the following form:countNA(data, byrow = FALSE) • data: A data.frame for which to count the number of missing values • byrow: Should missing values be counted by row (TRUE) or by column (FALSE)?The function should return a named numeric vector giving the count of missing values (NAs) for each row oreach column of data(depending on the value of byrow). The names of the result should be the rownames() or colnames() of data, whichever is appropriate.

```
countNA<-function(data,byrow=FALSE) {
  if (byrow==TRUE){
    rowSums(is.na (data))
  }  else {
    colSums(is.na(data))
  }
}
```

Solution 1 of Problem 1

Output test for Solution 1

Output1-> Rowwise NAs count

```
countNA(testdf)
```

Output2-> Coloumnwise NAs count

```
##  age city  gpa
##    1    2    2
```

```
countNA(testdf,TRUE)
```

```
##      Jack      Rosa      Dawn     Vicki     Blake Guillermo
##         0         1         1         1         0         2
```

```
imputeNA <- function(data, use.mean = FALSE) {
    for (col in 1:length(data)) {
      if(sapply(data[col],class)=="numeric"){
        if(use.mean==TRUE){
          mean_data=colMeans(data[col],na.rm=TRUE)
          data[col]<-replace(data[col],is.na(data[col]),mean_data)
        } else {
          column<-data[,col]
          median_data<-median(column,na.rm=TRUE)
          data[col]=replace(data[col],is.na(data[col]),median_data)
        }
      }else {
        char<- table(data[,col])
        mode<-names(char)[char==max(char)]
        data[,col]<-replace(data[,col],is.na(data[,col]),mode)
      }
    }
  return (data)
  }
```

Problem 2: Write a function of the following form:imputeNA(data, use.mean = FALSE) • data: A data.frame for which to impute the missing values • use.mean: Use the mean instead of the median for imputing continuous values The function should return a modified copy of data with missing values (NAs) imputed. Continuous variables(numeric types) should be imputed using the median or mean (according to use.mean) of the non-missing values. Categorical variables (character or factor types) should be imputed using the mode. (You may find it useful to first create a function for calculating the mode.)

Output test for Solution 2

Output1-> Imputed Mean Values for NAs

Output2-> Imputed Median Values NAs

```
imputeNA(testdf,use.mean=TRUE)
```

In both the outputs NAs for cities were filled with Mode

```
##            age     city   gpa
## Jack      24.0   Harlem 3.500
## Rosa      23.0 Brooklyn 3.600
## Dawn      24.6   Queens 4.000
## Vicki     25.0 Brooklyn 3.725
## Blake     32.0 Brooklyn 3.800
## Guillermo 19.0 Brooklyn 3.725
```

```
imputeNA(testdf)
```

```
##           age     city gpa
## Jack       24   Harlem 3.5
## Rosa       23 Brooklyn 3.6
## Dawn       24   Queens 4.0
## Vicki      25 Brooklyn 3.7
```

```
## Blake      32 Brooklyn 3.8
## Guillermo  19 Brooklyn 3.7
```

Part-B

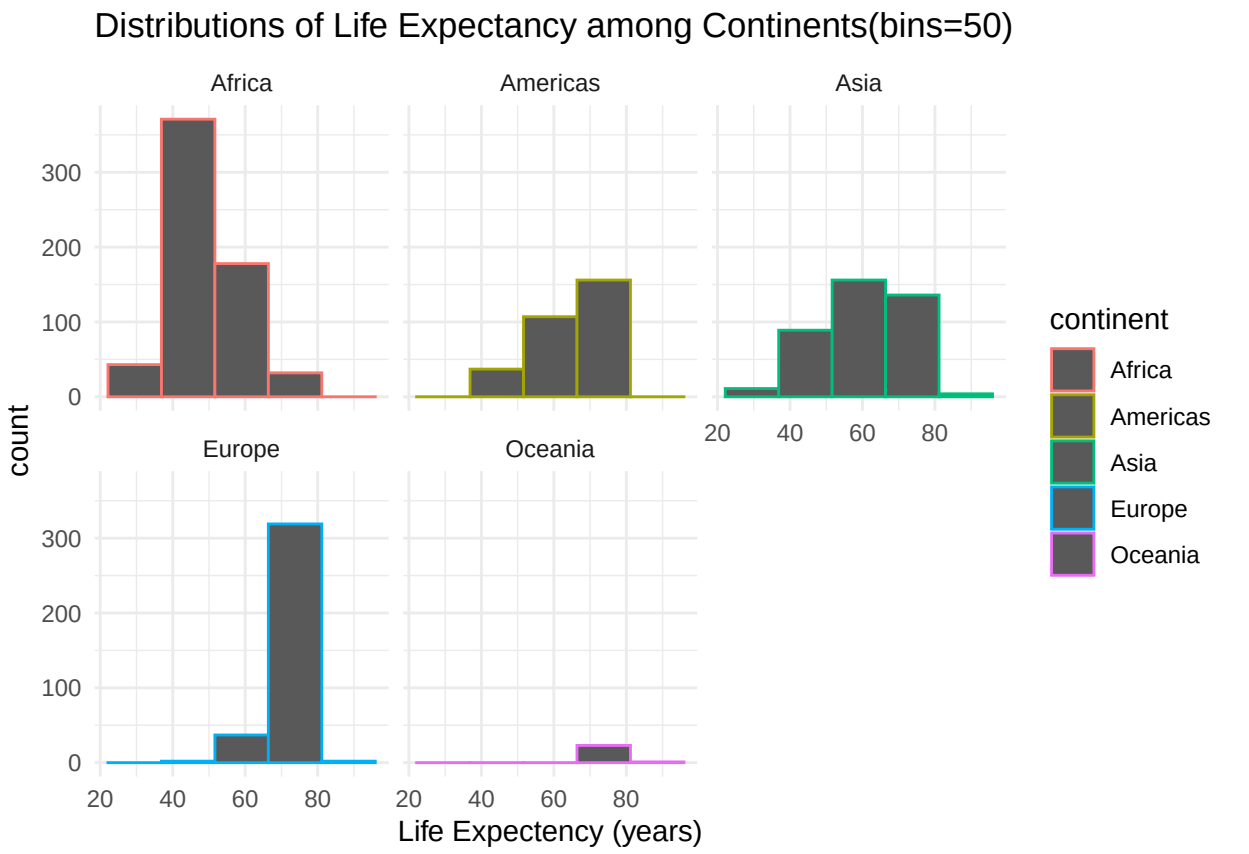
Problem 3: Visualize the distribution of life expectancy using histograms or density plots, faceted by continent. Describe the shape of the distributions of life expectancy, and any relationships you notice between life expectancy and continent.

Analysis on the Histogram "Distributions of Life Expectancy among Continents(bins=50)" plotted from the gapminder dataset. – Histogram depicts the distribution of Life Expectancy over 5 continents.

– From the plot, we can deduce that the highest life expectancy is from Europe continent.

– Asia and Americas show a similar inclining curve while Africa represents a declining curve while Ocenia has the least life expectency among all the continents.

```
library("ggplot2")
library("gapminder")
ggplot(gapminder,aes(x=lifeExp,color=continent))+
  geom_histogram(bins=5)+
  labs(title="Distributions of Life Expectancy among Continents(bins=50)",
       x="Life Expectency (years)")+
  theme_minimal()+
  facet_wrap(~continent)
```



Problem 4: For each continent, use box-and-whisker plots to visualize the distribution of life expectancy over time (i.e.,for each year in the dataset from 1952 to 2007). Comment on any trends you see in the evolution of life expectancy over the years.

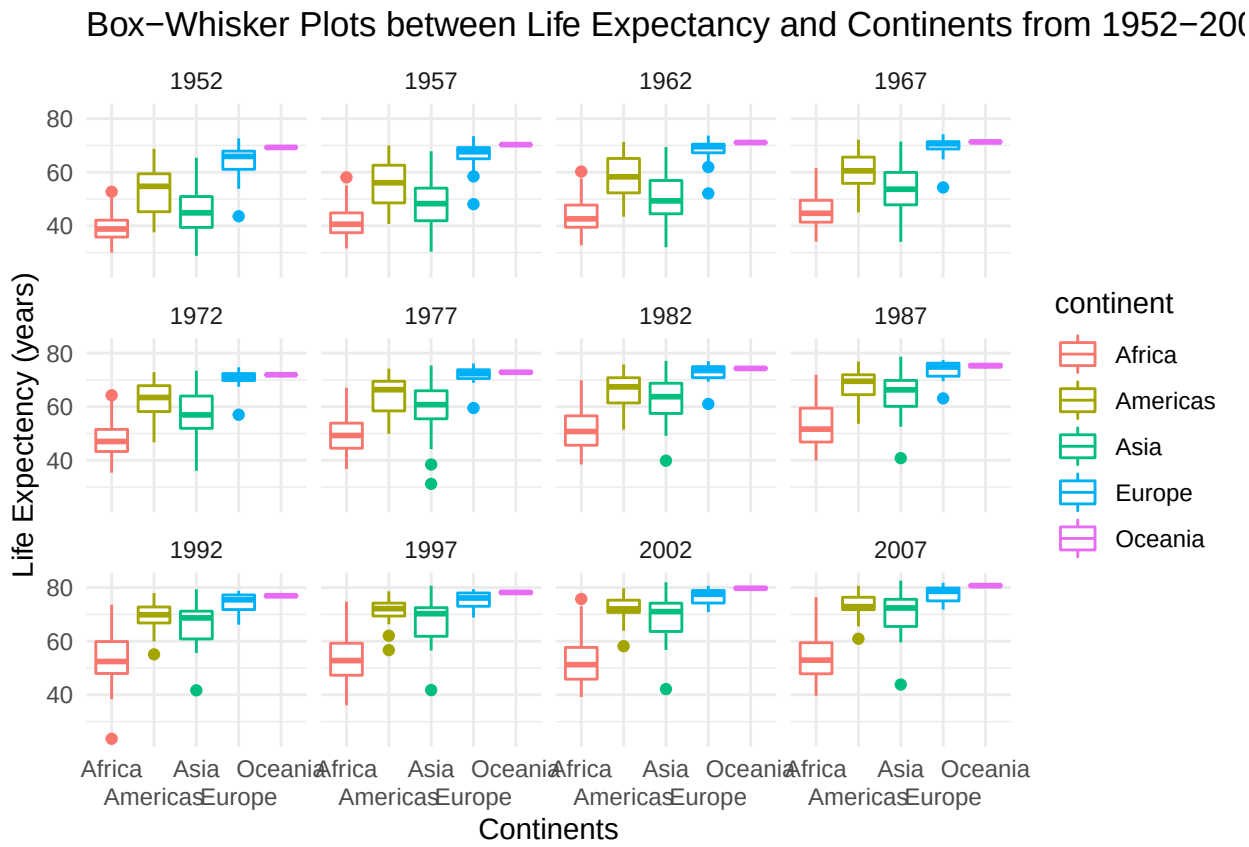
The below Box and Whisker plots represent the change in trends of life expectency over the years 1952 to 2007 among 5 continents (represented in a different colour each) – From 1952, the inter quartile range has increased for Africa and the plot has consistently maintained itself in a left skewed manner with few outliers more than the upper quartile in the first 2 decades [1950s-60s] and gradually reduced over the time till 2007.

– The Americas and the Asia interquartile ranges kept oscillating over the years with noticeable outliers in every alternative plot. However, the box plot moved from the lower expectancy towards the higher expectancy from 1952 to 2007.

– Europe's interquartile range fluctuated a bit over the initial decades and has been left skewed for the entirety of the plot.

– Ocenia's interquartile range has been consistent in being the least and also has zero outliers.

```
ggplot(gapminder,aes(x=continent,y=lifeExp,color=continent))+
  geom_boxplot()+
  labs(title="Box-Whisker Plots between Life Expectancy and Continents from 1952-2007",
       x="Continents",y="Life Expectency (years)")+
  theme_minimal()+
  facet_wrap(~year)+
  scale_x_discrete(guide=guide_axis(n.dodge=2.8))
```



Problem 5:Use a scatter plot to visualize the relationship between life expectancy and GDP per capita, including separate trend lines for each continent. Comment on the relationship between life expectancy and GDP per capita and any outliers you may notice.

This scatter plot shows the direct relation between Life Expectency and GDPperCapita Income
– We notice that there is an exponential increase in the life expectancy with the increase in the gdppercapita income.

– While the relation may be true for the most part, we can notice that there are two exceptions in this relation (Africa and Asia), they showcase a declining graph after a certain point of exponential growth.

– The exception is due to the increased number of outliers from the continents Africa and Asia.

```
ggplot(gapminder,aes(x=gdpPercap,y=lifeExp,color=continent,))+
  geom_point()+
  labs(title="Scatter plot between GDP per Capita Income and Life Expectency ",
       x="GDP per Capita Income",y="Life Expectency (years)")+
  geom_smooth(color="grey9")+
  theme_minimal()+
  facet_wrap(~continent)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```