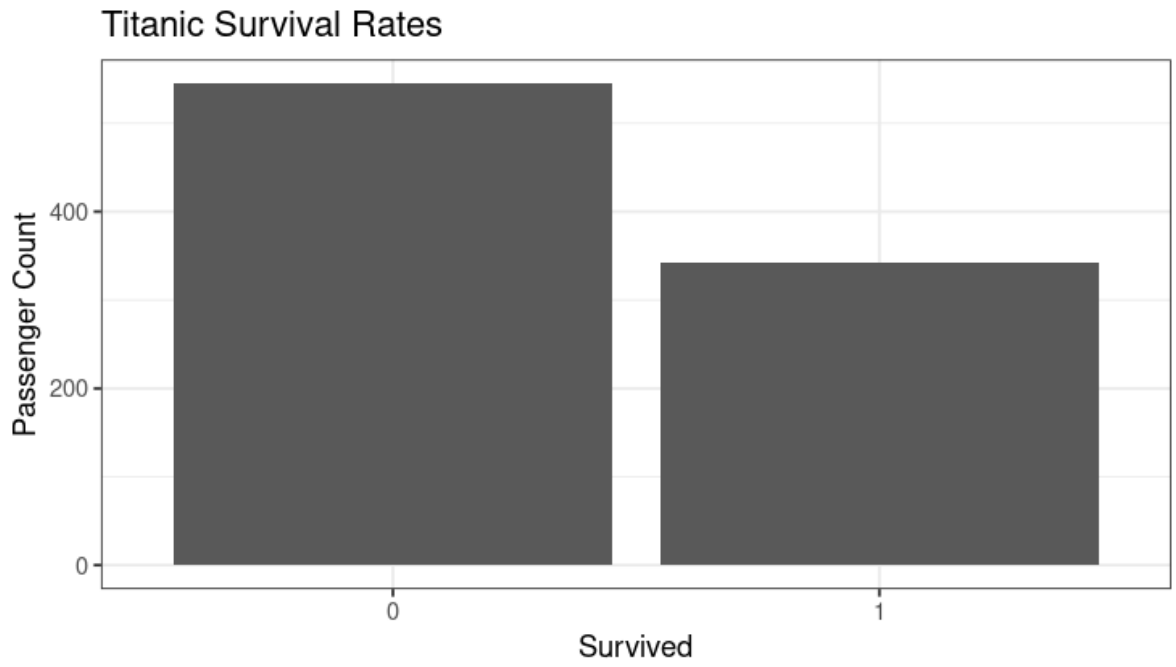
Report - Data Visualization

Name : Vineeth N Reddy

Titanic Data Set

- Survival Rate Graph

```
> ggplot(titanic, aes(x = Survived)) + theme_bw() + geom_bar() + labs(y =
"Passenger Count", title = "Titanic Survival Rates")
```



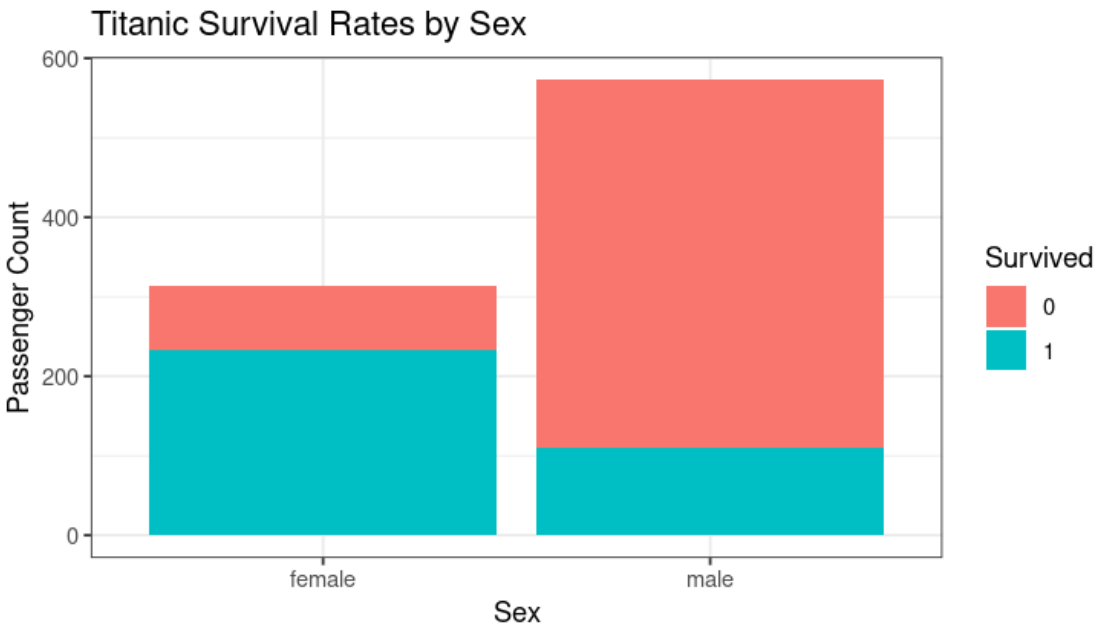
Note - 0 denotes not survived and 1 denotes survived

- Survival Percentage

```
0         1
0.6144307 0.3855693
```

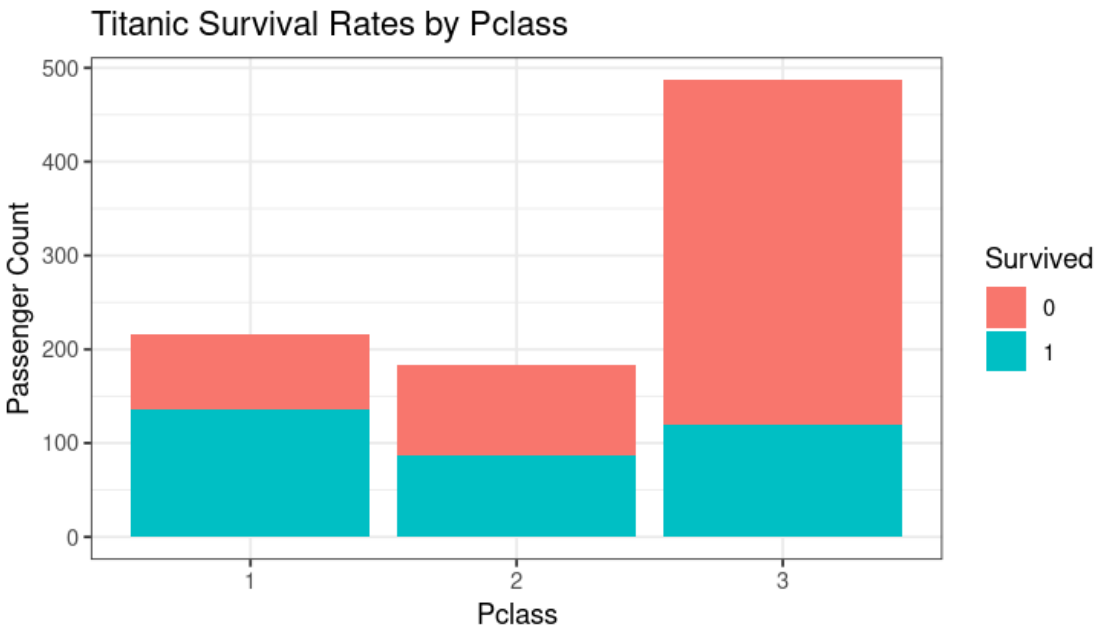
- Survival rate based on gender

```
> ggplot(titanic, aes(x = Sex, fill = Survived)) + theme_bw() + geom_bar() +
labs(y = "Passenger Count", title = "Titanic Survival Rates by Sex")
```



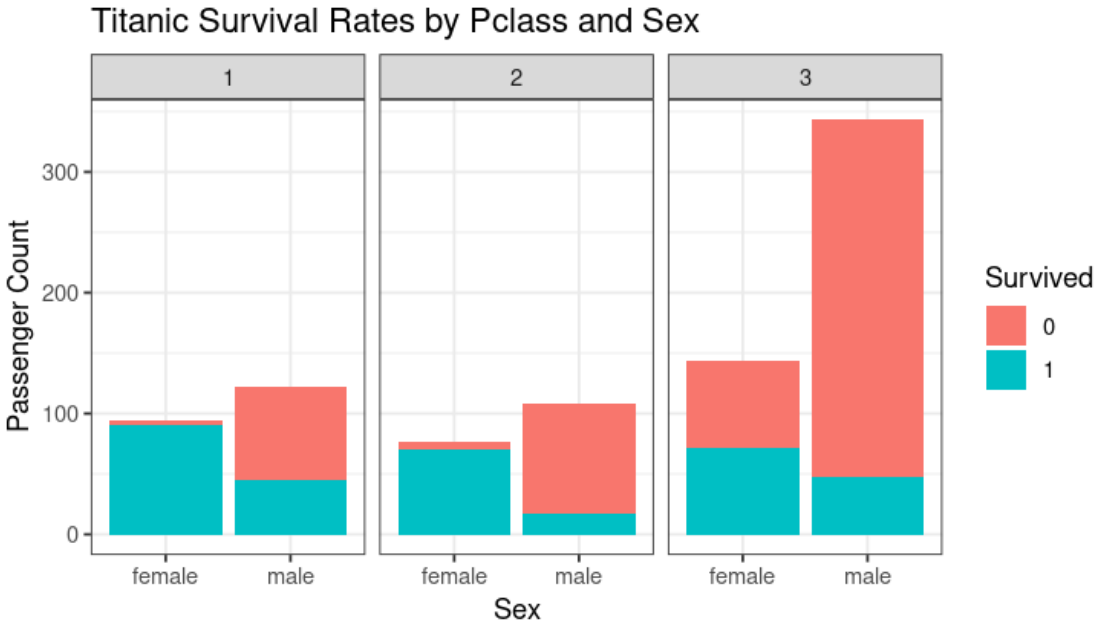
- Survival rate by class of ticket

```
> ggplot(titanic, aes(x = Pclass, fill = Survived)) + theme_bw() +
geom_bar() + labs(y = "Passenger Count", title = "Titanic Survival Rates by
Pclass")
```



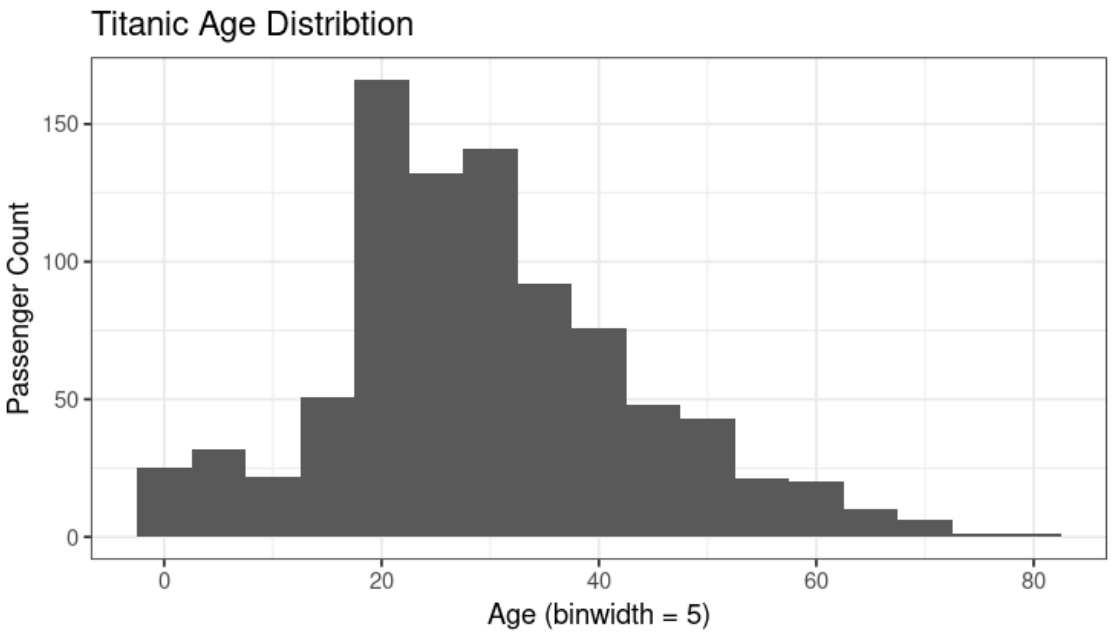
- Now let's combine the survival rate by class of ticket and gender together

```
> ggplot(titanic, aes(x = Sex, fill = Survived)) + theme_bw() + facet_wrap(~
Pclass) + geom_bar() + labs(y = "Passenger Count", title = "Titanic Survival
Rates by Pclass and Sex")
```



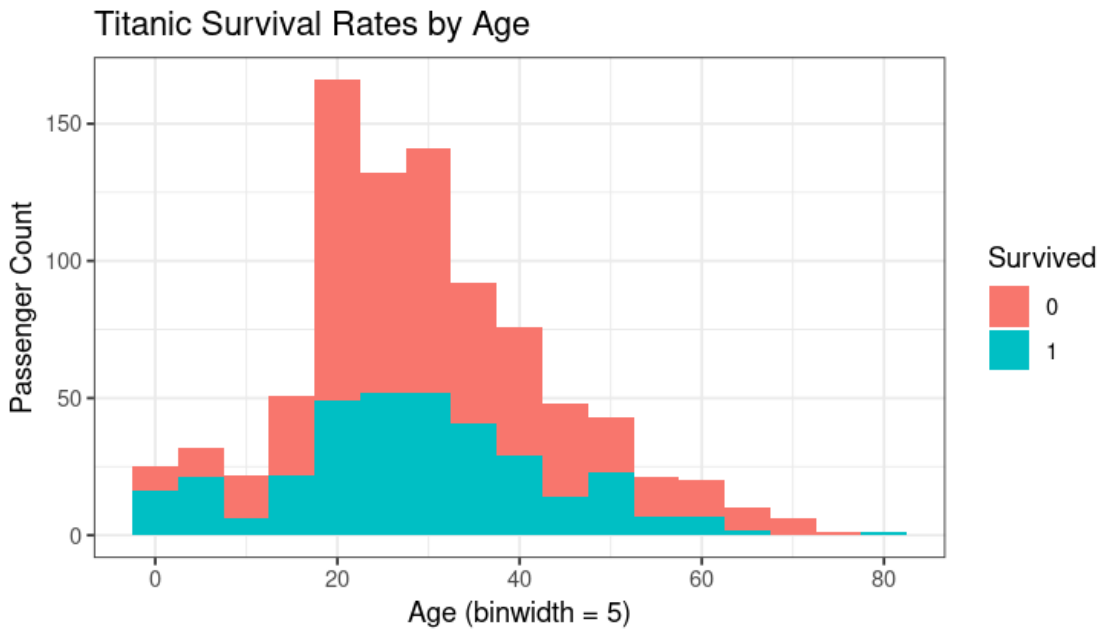
- Let's look at the distribution graph of various passenger's age

```
> ggplot(titanic, aes(x = Age)) + theme_bw() + geom_histogram(binwidth = 5)
+ labs(y = "Passenger Count", x = "Age (binwidth = 5)", title = "Titanic Age
Distribtion")
```

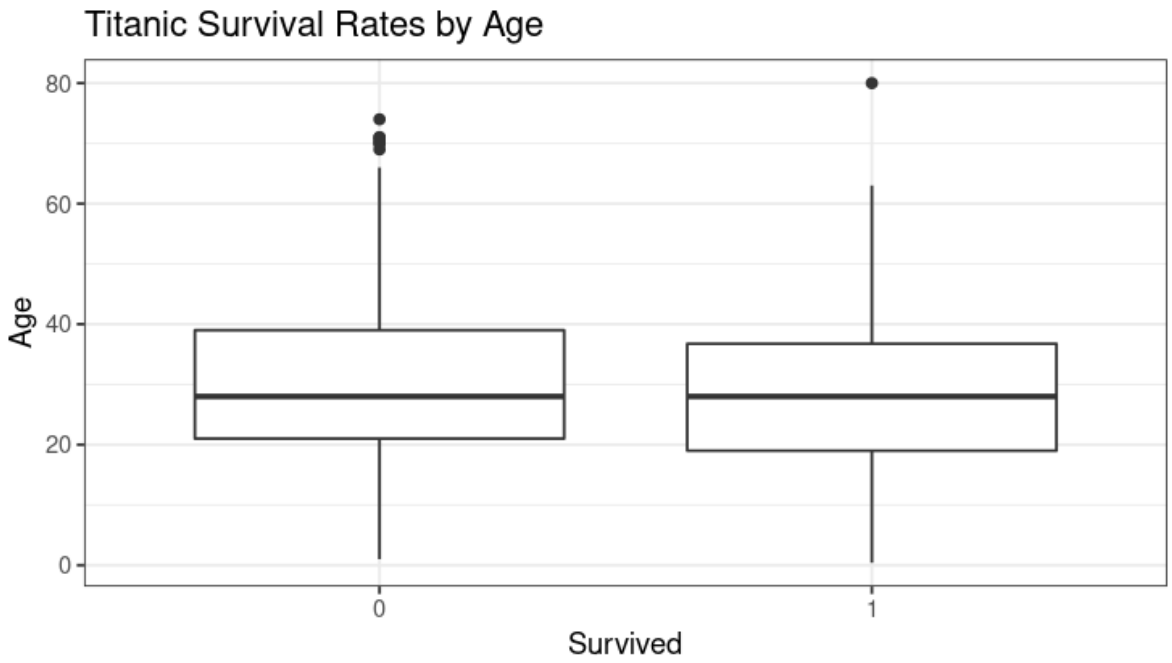


- With the help of the above graph one can derive the survival rate based on the age

```
> ggplot(titanic, aes(x = Age, fill = Survived)) + theme_bw() +
geom_histogram(binwidth = 5) + labs(y = "Passenger Count", x = "Age
(binwidth = 5)", title = "Titanic Survival Rates by Age")
```




```
> ggplot(titanic, aes(x = Survived, y = Age)) + theme_bw() + geom_boxplot() +
labs(y = "Age", x = "Survived", title = "Titanic Survival Rates by Age")
```



Box Plot allows us to find the median, maximum value, minimum value and outliers.

For not survived (0 - approximation from the graph)

1. Median - 28
2. Minimum - 1
3. Maximum - 65
4. Q1 - 21
5. Q3 - 39
6. Outliers present (suspected anomalies)

For survived (1 - approximation from the graph)

1. median - 28

2. Minimum - 0
3. Maximum - 62
4. Q1 - 19
5. Q3 - 36
6. One outlier present

- Survival rate by age when segmented by gender and class of tickers

```
> ggplot(titanic, aes(x = Age, fill = Survived)) + theme_bw() +
facet_wrap(Sex ~ Pclass) + geom_histogram(binwidth = 5) +labs(y = "Age", x =
"Survived", title = "Titanic Survival Rates by Age, Pclass and Sex")
```

World Trends

- Summary of the data set (StringsAsFactors = TRUE)

```
> head(data)
          Country.Name Country.Code
1                Aruba          ABW
2          Afghanistan          AFG
3               Angola          AGO
4              Albania          ALB
5 United Arab Emirates          ARE
6            Argentina          ARG
        Region Year Fertility.Rate
1 The Americas 1960          4.820
2         Asia 1960          7.450
3       Africa 1960          7.379
4       Europe 1960          6.186
5  Middle East 1960          6.928
6 The Americas 1960          3.109

> str(data)
'data.frame':   374 obs. of  5 variables:
 $ Country.Name  : Factor w/ 187 levels "Afghanistan",..: 8 1 4 2 176 6 7 5
9 10 ...
```

```
 $ Country.Code  : Factor w/ 187 levels "ABW","AFG","AGO",..: 1 2 3 4 5 6 7
8 9 10 ...
 $ Region        : Factor w/ 6 levels "Africa","Asia",..: 6 2 1 3 4 6 2 6 5
3 ...
 $ Year          : int  1960 1960 1960 1960 1960 1960 1960 1960 1960 1960
...
 $ Fertility.Rate: num  4.82 7.45 7.38 6.19 6.93 ...

> summary(data)
              Country.Name  Country.Code
 Afghanistan         :  2   ABW    :  2
 Albania             :  2   AFG    :  2
 Algeria             :  2   AGO    :  2
 Angola              :  2   ALB    :  2
 Antigua and Barbuda:  2   ARE    :  2
 Argentina           :  2   ARG    :  2
 (Other)             :362   (Other):362
          Region          Year
 Africa       :106   Min.   :1960
 Asia         : 66   1st Qu.:1960
 Europe       : 80   Median :1986
 Middle East  : 24   Mean   :1986
 Oceania      : 26   3rd Qu.:2013
 The Americas: 72   Max.   :2013

 Fertility.Rate
 Min.   :1.124
 1st Qu.:2.243
 Median :3.994
 Mean   :4.191
 3rd Qu.:6.252
 Max.   :8.187
```

- Turning the year into a factor

```
> temp <- factor(data$Year)
> levels(temp)
[1] "1960" "2013"
```

- Split the data frame into 2013 and 1960

```
> data1960 <- data[data$Year==1960,]
> data2013 <- data[data$Year==2013,]
```

- Add additional data frames

```
> add1960 <- data.frame(Code=Country_Code,
Life.Exp=Life_Expectancy_At_Birth_1960)
> add2013 <- data.frame(Code=Country_Code,
Life.Exp=Life_Expectancy_At_Birth_2013)
```

```
> summary(add1960)
     Code              Life.Exp
```

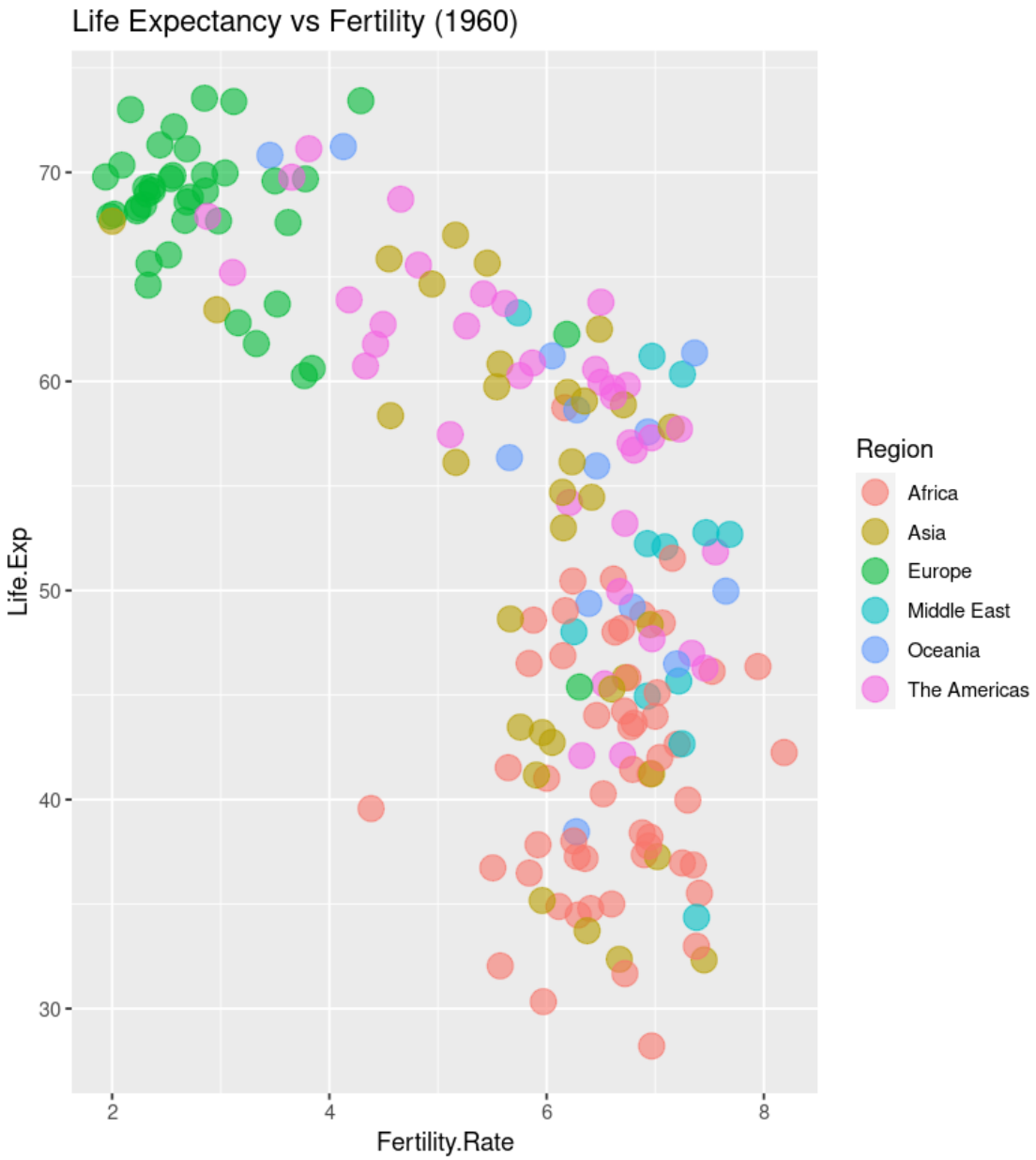
```
 Length:187         Min.   :28.21
 Class :character   1st Qu.:43.47
 Mode  :character   Median :54.70
                    Mean   :53.73
                    3rd Qu.:64.05
                    Max.   :73.55

> summary(add2013)
     Code              Life.Exp
 Length:187         Min.   :48.94
 Class :character   1st Qu.:64.52
 Mode  :character   Median :73.25
                    Mean   :70.76
                    3rd Qu.:76.84
                    Max.   :83.83
```

- Now we merge the data sets which we split

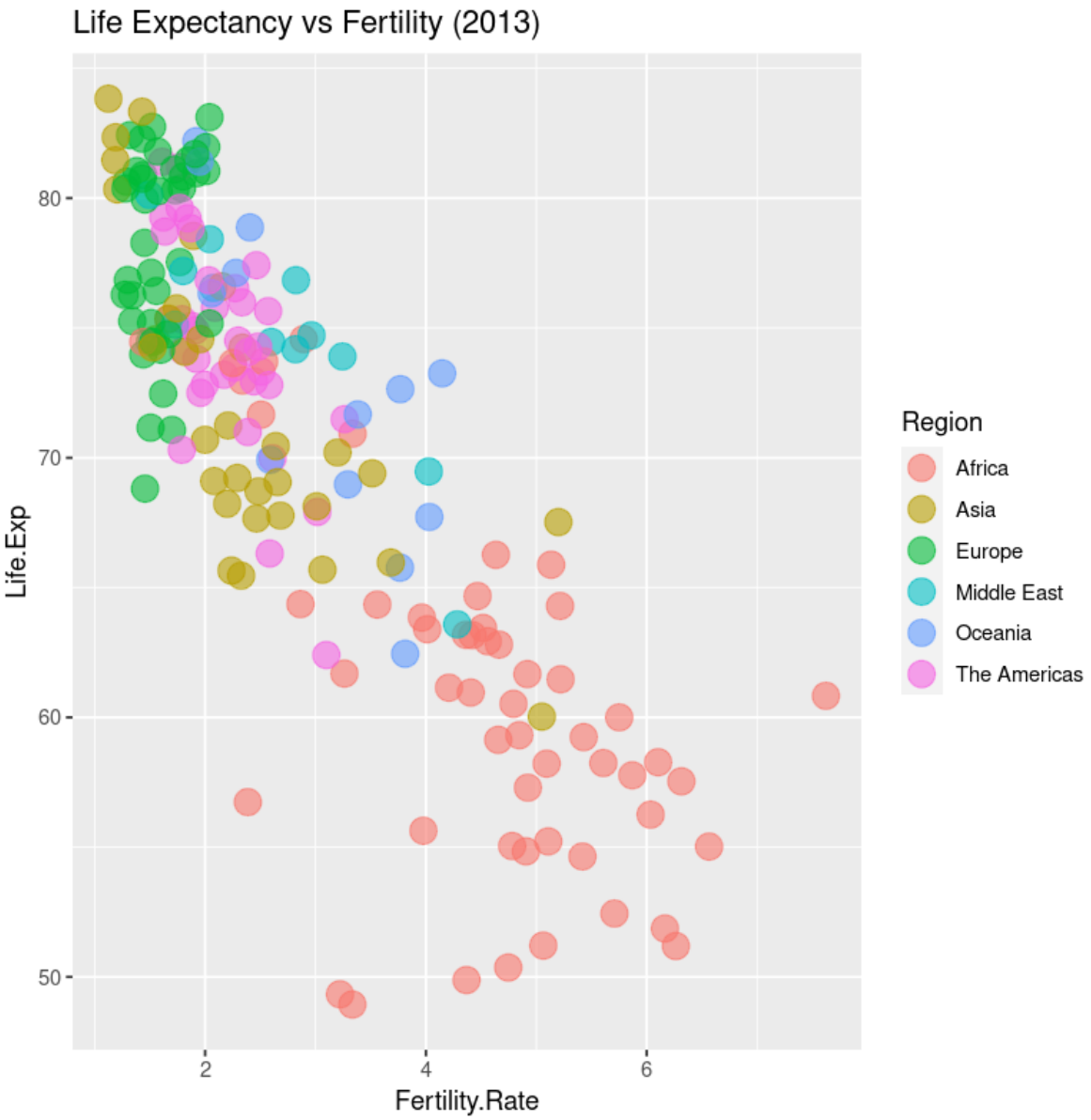
```
> merged1960 <- merge(data1960, add1960, by.x="Country.Code", by.y="Code")
> merged2013 <- merge(data2013, add2013, by.x="Country.Code", by,y="Code")
```

- Visualize the data

```
> qplot(data=merged1960, x=Fertility.Rate, y=Life.Exp, colour=Region,
size=I(5), alpha = I(0.6), main="Life Expectancy vs Fertility (1960)")
```




```
> qplot(data=merged2013, x=Fertility.Rate, y=Life.Exp, colour=Region, size=I(5),
alpha = I(0.6), main="Life Expectancy vs Fertility (2013)")
```



Analysis based on the graph

1. African countries have high fertility rate (more children per woman) and very low life expectancy
2. European countries have low fertility rate (less children per woman) and very high life expectancy.
3. In 53 years, the fertility rate in African countries have dropped but life expectancy has increased.
4. In European countries these green ones show the number of children between two and four and in 2013 it's dropped between one and two children. So less than two children on average along European country but the life expectancy has increased.