PROJ3

Joon

April 16, 2018

```
library(gapminder)
```

```
## Warning: package 'gapminder' was built under R version 3.4.4
```

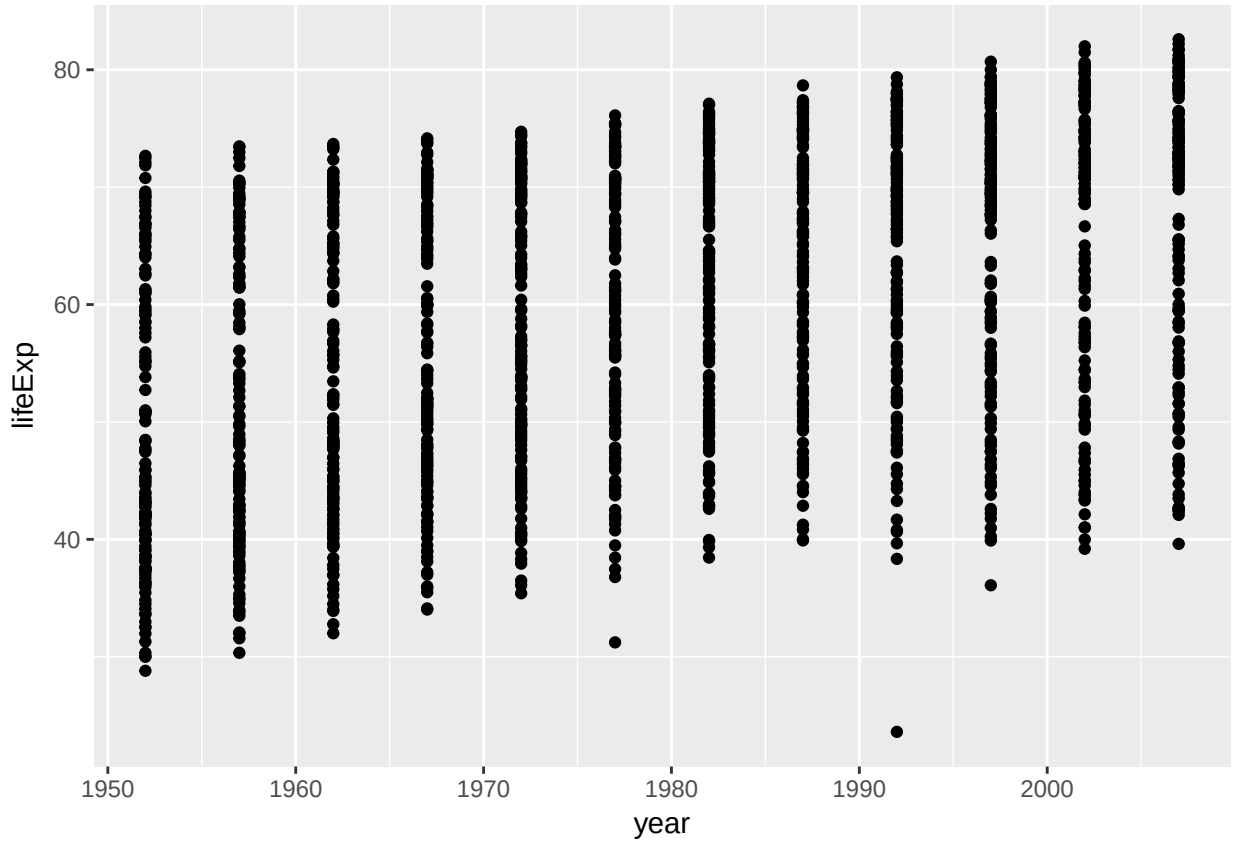
```
data(gapminder)

gapminder
```

```
## # A tibble: 1,704 x 6
##    country     continent  year lifeExp      pop gdpPercap
##    <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
##  1 Afghanistan Asia       1952    28.8  8425333       779
##  2 Afghanistan Asia       1957    30.3  9240934       821
##  3 Afghanistan Asia       1962    32.0 10267083       853
##  4 Afghanistan Asia       1967    34.0 11537966       836
##  5 Afghanistan Asia       1972    36.1 13079460       740
##  6 Afghanistan Asia       1977    38.4 14880372       786
##  7 Afghanistan Asia       1982    39.9 12881816       978
##  8 Afghanistan Asia       1987    40.8 13867957       852
##  9 Afghanistan Asia       1992    41.7 16317921       649
## 10 Afghanistan Asia       1997    41.8 22227415       635
## # ... with 1,694 more rows
```

we were told that we should do one point per country

Exercise 1

```
library(tidyr)
library(ggplot2)
gapminder %>%
    ggplot(aes(x=year,y=lifeExp))+ geom_point()
```



##Question 1: It appears that the general trend behind this plot is that the average life expectancy of the data is increasing

Question 2:

The life expectancy distribution per year is skewed for each year as the "violin" balloons out further on one end indicating a larger distribution on that end. For example years starting from 1950 to 1970 are larger at the bottom adn thus skewed towards that direction while years past that have a more top heavy distribution and thus are skewed more heavily towards the upper range.

I would describe the data as unimodal as the violin plots peak around one value for each of the years which are reasonably assumed to be the mode of the life expectancy for that year.

Question 3:

I would reject the null hypothesis as there does seem to be a definable trend between year and the increase in life expectancy

Question 4:

The violin plot would have a positive relationship with years and residuals similar to the trend observed in life expectancy and year

Question 5:

The violin plot Should be centered around 0 with variations between each violin plot.

Exercise 2

```
library(tidyr)
library(ggplot2)

d2<-lm(lifeExp ~ year,data = gapminder)
  broom::tidy(d2)
```

```
##          term     estimate   std.error statistic      p.value
## 1 (Intercept) -585.6521874 32.31396452 -18.12381 2.897807e-67
## 2        year    0.3259038  0.01632369  19.96509 7.546795e-80
```

QUestion 6:

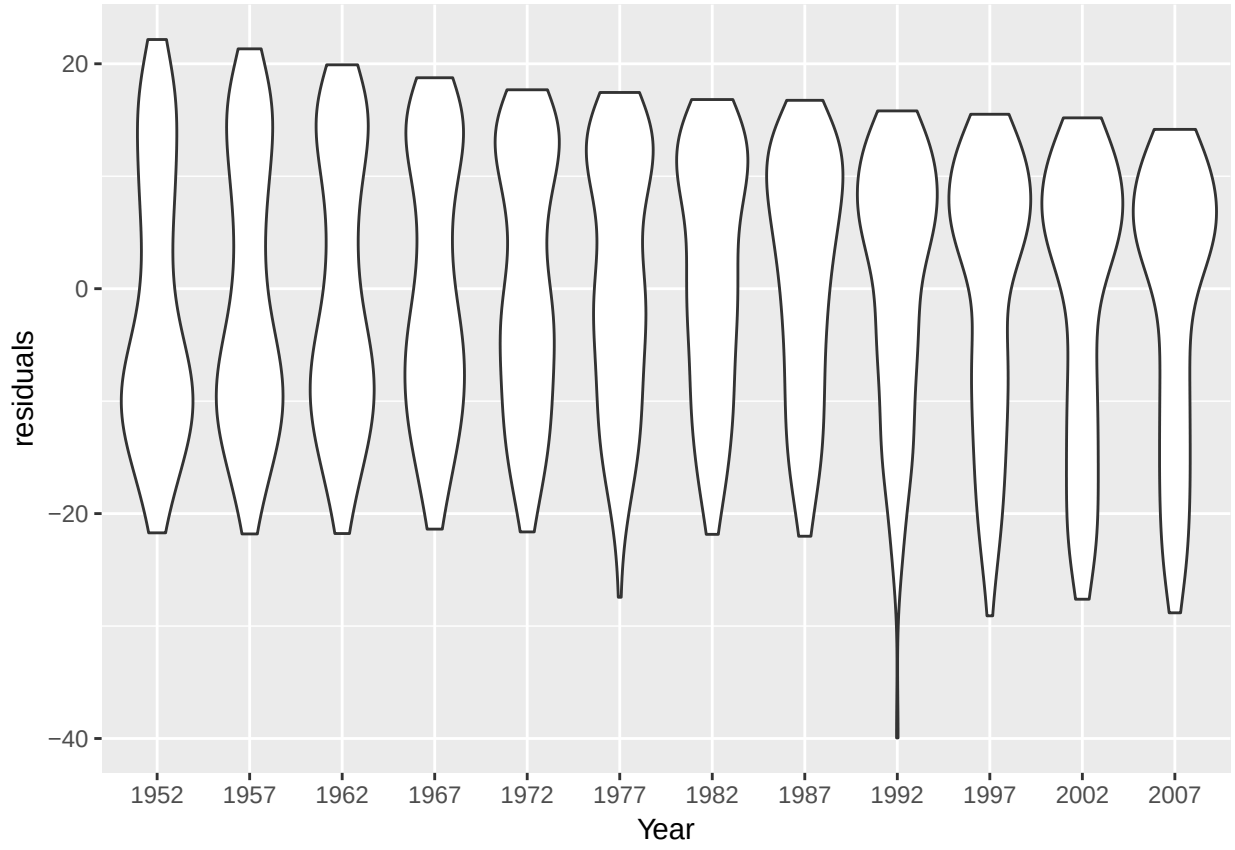
according to the linear model the life expectancy increases by .32 per year

Question 7:

Yes as there is a definable increase of life expectancy per year and thus a definable relationship between year and life expectancy

Exercise 3

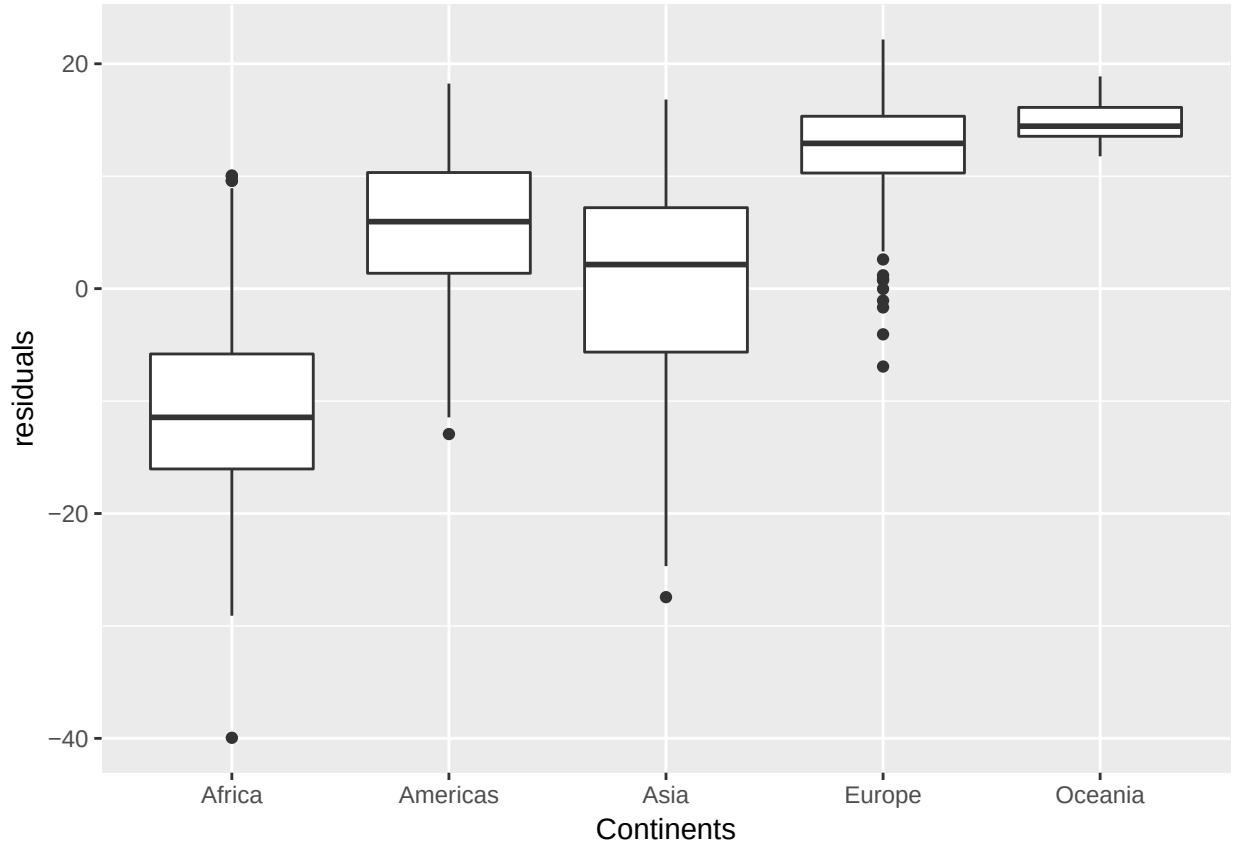
```
library(ggplot2)
library(broom)
library(tidyr)

augmented_gapminder <- d2 %>% augment()

augmented_gapminder %>%
  ggplot(aes(x=factor(year),y=.resid)) +
  geom_violin() +
  xlab("Year") +
  ylab("residuals")
```



Question 8

The plot matches my expectations set in q4. We can see that the mode of each violin plot is slowly increasing to a positive mode and by the end of 2007 most of the data values are above 0.

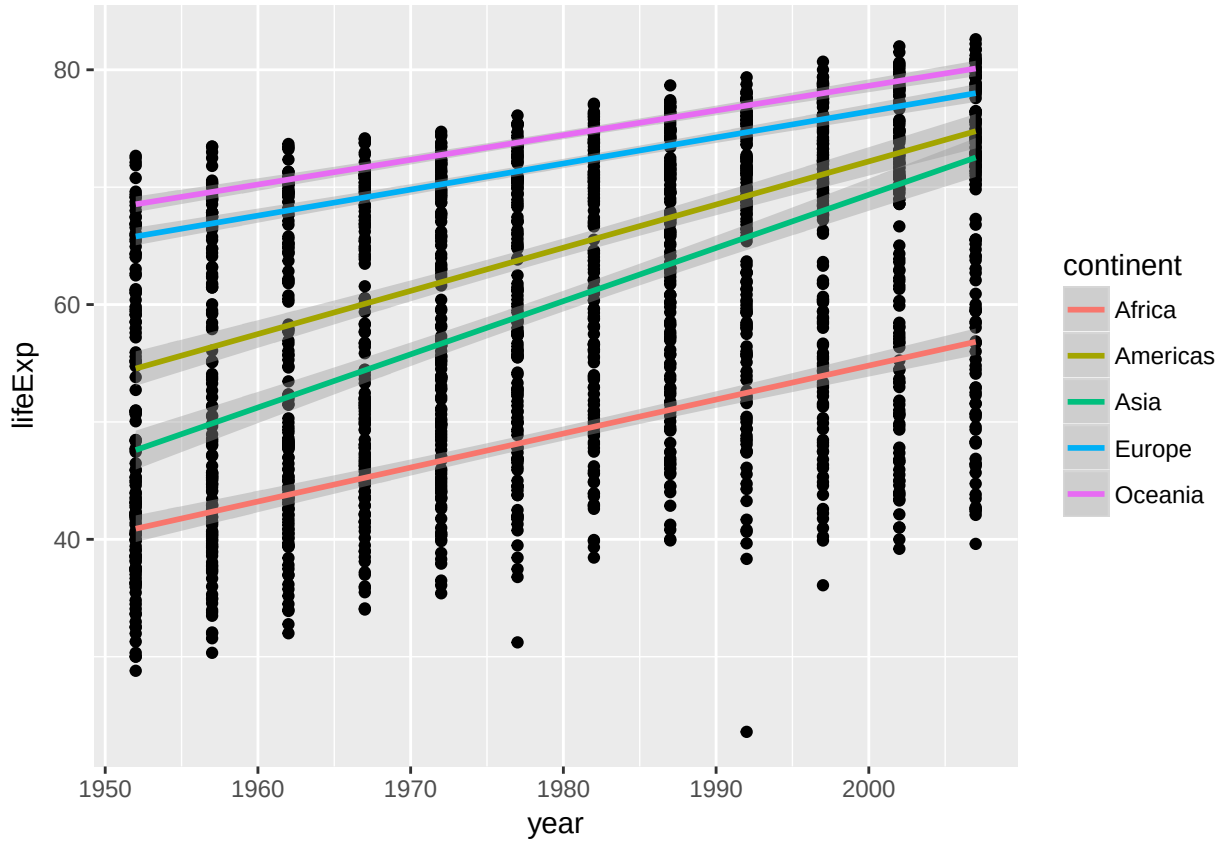
```
big_gapminder <- merge(gapminder,augmented_gapminder, by=c("lifeExp","year"))

big_gapminder %>%
  ggplot(aes(x=continent,y=.resid)) +
  geom_boxplot() +
  xlab("Continents") +
  ylab("residuals")
```



Question 9

There definitely seems to be a relationship between continent and life expectancy which suggests that life expectancy over time and the rate at which it is increasing depends on which continent you are observing.

```
gapminder %>%
  ggplot(aes(x=year,y=lifeExp)) +
  geom_point() +
  geom_smooth(aes(color = continent),method=lm)
```



Question 10

```
d6<-lm(lifeExp ~ year*continent,data = gapminder)
  broom::tidy(d6)
```

```
##                       term      estimate    std.error   statistic
## 1              (Intercept) -524.25784607  32.96342596 -15.9042281
## 2                     year    0.28952926   0.01665177  17.3872996
## 3       continentAmericas -138.84844718  57.85057778  -2.4001220
## 4           continentAsia -312.63304922  52.90355242  -5.9094907
## 5         continentEurope  156.84685210  54.49775866   2.8780423
## 6        continentOceania  182.34988290 171.28298566   1.0646118
## 7  year:continentAmericas    0.07812167   0.02922373   2.6732271
## 8      year:continentAsia    0.16359314   0.02672470   6.1214213
## 9    year:continentEurope   -0.06759712   0.02753003  -2.4553961
## 10  year:continentOceania   -0.07925689   0.08652512  -0.9159986
##         p.value
## 1  3.436134e-53
## 2  1.953998e-62
## 3  1.649695e-02
## 4  4.139916e-09
## 5  4.051687e-03
## 6  2.872034e-01
## 7  7.584665e-03
## 8  1.149941e-09
```

```
## 9  1.417280e-02
## 10 3.597980e-01
```

QUestion 11

Most of the variables are not significantly different from 0 for example year has a value of 1.953998e-62. Some variables that might be are the continent of Oceania, Europe and the year. Others include the interaction such as year.continentOceania,year.continentEurope that are significantly different from 0 compared to the other variables.

Question 12

```
d6[[1]][2]
```

```
##      year
## 0.2895293
```

```
d6[[1]][7]
```

```
## year:continentAmericas
##             0.07812167
```

```
averages <- c(d6[[1]][2],d6[[1]][2]+d6[[1]][7],d6[[1]][2]+d6[[1]][8],d6[[1]][2]+d6[[1]][9],d6[[1]][2]+d
continents <- c('Africa','America','Asia','Europe','Oceania')

estimates <- data.frame(continents,averages)

estimates
```

```
##   continents  averages
## 1     Africa 0.2895293
## 2    America 0.3676509
## 3       Asia 0.4531224
## 4     Europe 0.2219321
## 5    Oceania 0.2102724
```

We see from our dataframe that year is our default and so to find the average of all the other continents by adding them together to access each variable i had to access the 2d array which is the reason for the odd syntax d6[[1]][n] where n is the position of the yearcontinent or year variable

```
nova2 <-anova(d2)
nova6 <-anova(d6)
nova2
```

```
## Analysis of Variance Table
##
## Response: lifeExp
##             Df Sum Sq Mean Sq F value    Pr(>F)
## year         1  53919   53919   398.6 < 2.2e-16 ***
## Residuals 1702 230229     135
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
nova6
```

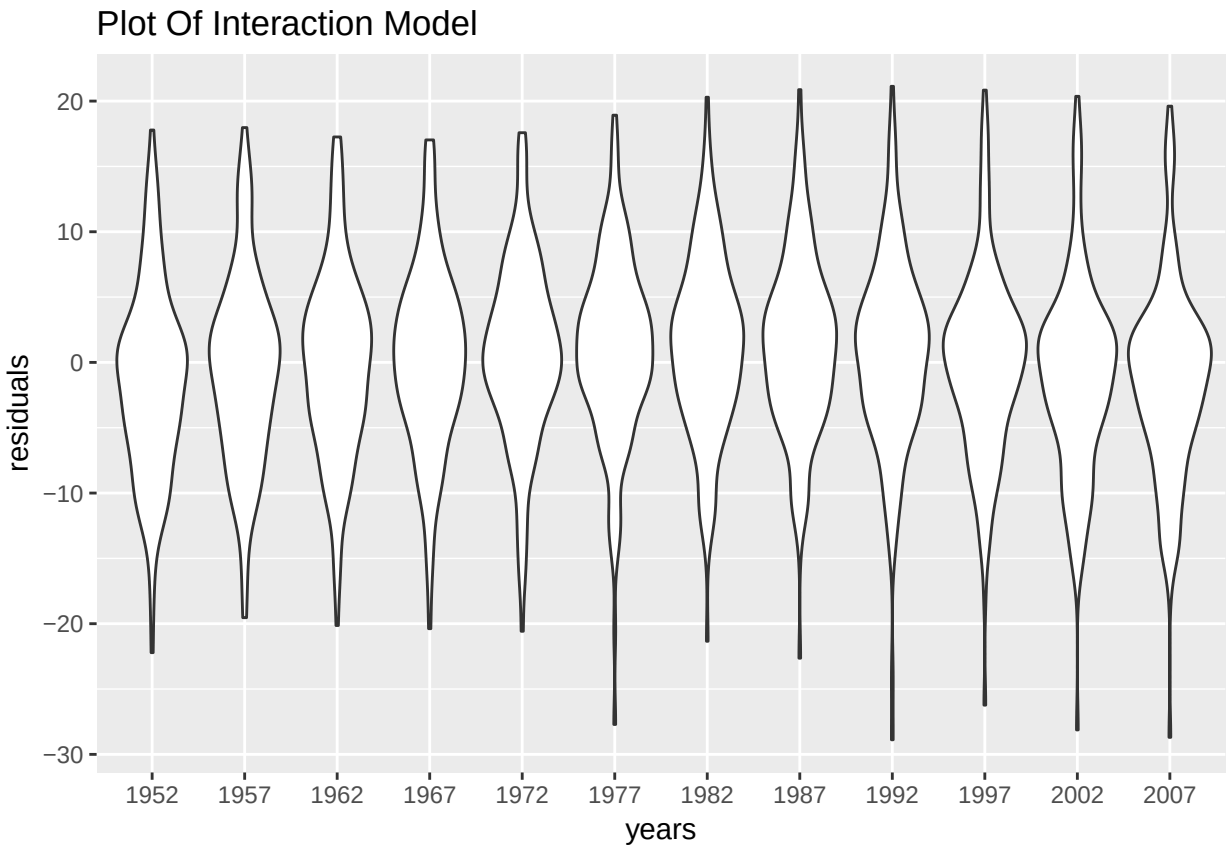
```
## Analysis of Variance Table
```

```
##
## Response: lifeExp
##                   Df Sum Sq Mean Sq  F value    Pr(>F)
## year               1  53919   53919 1046.028 < 2.2e-16 ***
## continent          4 139343   34836  675.812 < 2.2e-16 ***
## year:continent     4   3566     892   17.296 6.463e-14 ***
## Residuals       1694  87320      52
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

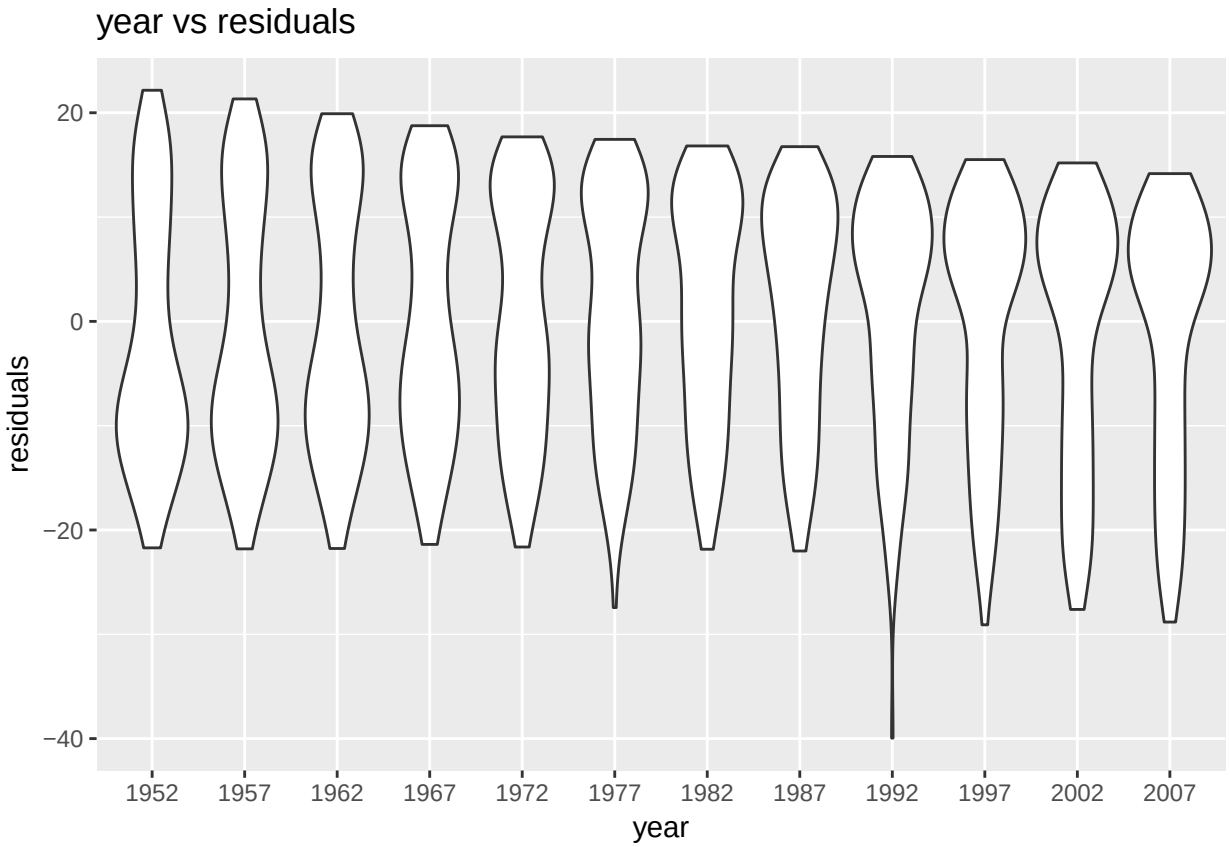
Question 13

Comparing the two linear regression models we see that the model in exercise 6 which defines an interaction between continent and year is a better model than just the year only model. We can infer this from the F values and the probability columns. The F stat of the continent-year model are all over 1 which indicates a relationship between the two variables. This is also supported by the probabilities associated with the f statistic using the f distribution. All the probability values for the f stats are close to 0 which indicates extremely strong evidence that there is an association with the specified interaction. This gives us a clearer picture than the year only model does as we have determined that coninent does in fact affect the life expectancy.

```
  aug_d6 <-d6 %>% augment()

aug_d6 %>%
  ggplot(aes(x=factor(year),y=.resid)) +
  geom_violin() +
  ggtitle("Plot Of Interaction Model ") +
  xlab("years") +
  ylab("residuals")
```

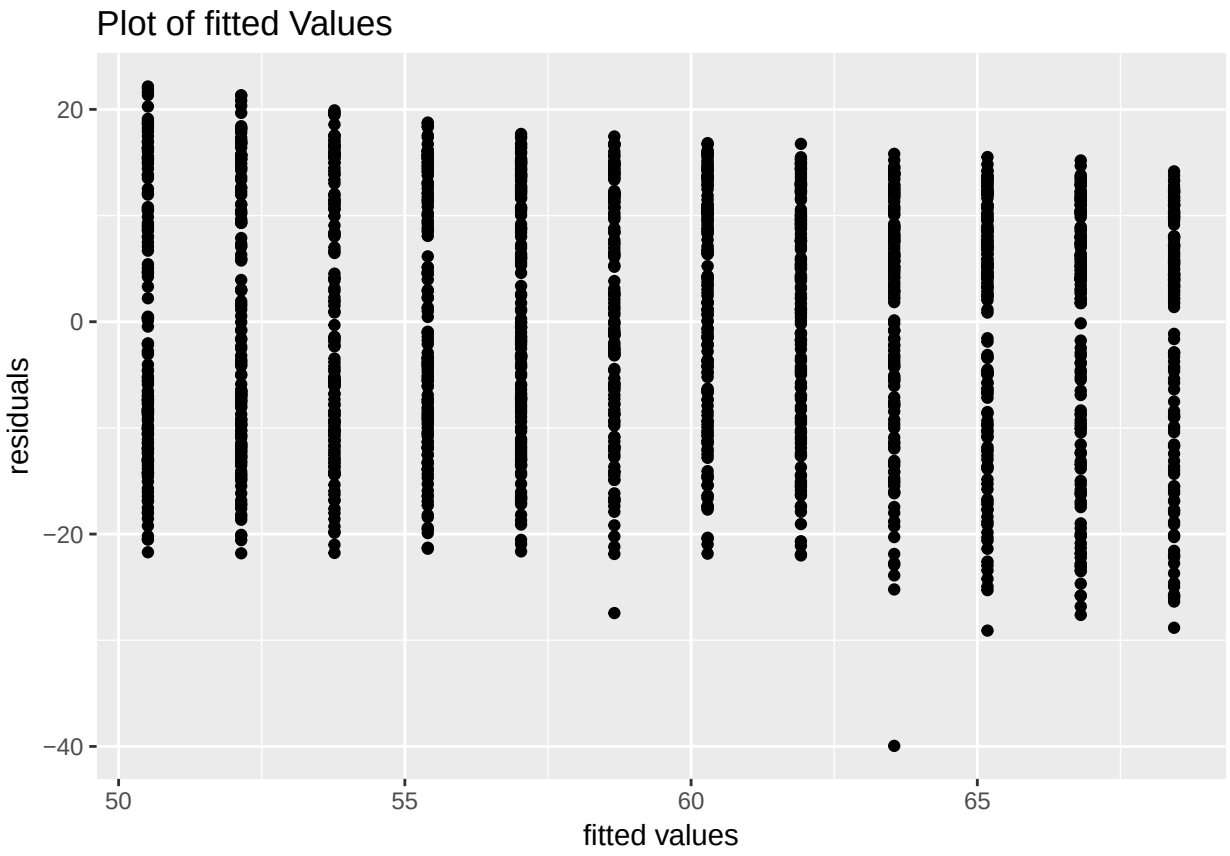



```
d2%>%ggplot(aes(x=factor(year),y=.resid)) +
 geom_violin()+
 ggtitle("year vs residuals") +
 xlab("year") +
 ylab("residuals")
```



Hector described the model for the fitted vs residuals to be "It's the value predicted by the regression model for each observation in the dataset.". SO I used the augmented gapminder df which was pretty much the df in exercise 2 just augmented.

```
 aug_d2 <- d2%>% augment()

aug_d2 %>%
  ggplot(aes(x=.fitted,y=.resid)) +
  geom_point() +
  ggtitle("Plot of fitted Values") +
  xlab("fitted values") +
  ylab("residuals")
```


Plot of fitted Values
residuals
fitted values
20
0
−20
−40
50
55
60
65