Final Stat Inference Project by Aryan Patel , Gurnit Chauhan, Gavin D'Cruz

Step 1 : Loading in the libraries

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.2     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.2     ✔ tibble    3.2.1
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force all
conflicts to become errors
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
##
## Loading required package: carData
##
##
## Attaching package: 'car'
##
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
##
## The following object is masked from 'package:purrr':
##
##     some
```

```
pop <- read.csv('population_total.csv')[, c('country', 'X2009')]
names(pop) <- c('country', 'population')
life <- read.csv('life_expectancy_years.csv')[, c('country', 'X2009')]
names(life) <- c('country', 'lifeexp')
income <- read.csv('income_per_person_gdppercapita_ppp_inflation_adjusted.csv')[, c('
country', 'X2009')]
names(income) <- c('country', 'income')
children <- read.csv('children_per_woman_total_fertility.csv')[, c('country', 'X2009'
)]
names(children) <- c('country', 'chdperwoman')
childmort <- read.csv('child_mortality_0_5_year_olds_dying_per_1000_born.csv')[, c('c
ountry', 'X2009')]
```

```
names(childmort) <- c('country', 'childmort')
co2 <- read.csv('co2_emissions_tonnes_per_person.csv')[, c('country', 'X2009')]
names(co2) <- c('country', 'co2')
gdp <- read.csv('gdppercapita_us_inflation_adjusted.csv')[, c('country', 'X2009')]
names(gdp) <- c('country', 'gdpcapita')
health <- read.csv('total_health_spending_per_person_us.csv')[, c('country', 'X2009')
]
names(health) <- c('country', 'healthspend')
popden <- read.csv('population_density_per_square_km.csv')[, c('country', 'X2009')]
names(popden) <- c('country', 'popdensity')
water <- read.csv('at_least_basic_water_source_overall_access_percent.csv')[, c('coun
try', 'X2009')]
names(water) <- c('country', 'water')
murders <- read.csv('murder_total_deaths.csv')[, c('country', 'X2009')]
names(murders) <- c('country', 'murders')
gap1 <- gapminder[c('country', 'continent')]
gap1 <- unique(gap1)
dflist <- list(pop, life, income, children, childmort, co2, gdp, health, popden, wate
r, murders, gap1)
mergedf <- Reduce(function(x,y) merge(x=x, y=y, by='country', all=TRUE), dflist)
mergedf <- na.omit(mergedf)
text.to.num <- function(x){
xx = as.character(x)
res=rep(0, length(x))
for(i in 1:length(x)){

1

if(is.na(xx[i])){
res[i]=NA
next
}
if (grepl("M", xx[i], ignore.case = TRUE)) {
res[i]=as.numeric(gsub("M", "", xx[i], ignore.case = TRUE)) * 1e6
} else if (grepl("k", xx[i], ignore.case = TRUE)) {
res[i]=as.numeric(gsub("k", "", xx[i], ignore.case = TRUE)) * 1e3
} else if (grepl("B", xx[i], ignore.case = TRUE)) {
res[i]=as.numeric(gsub("B", "", xx[i], ignore.case = TRUE)) * 1e9
} else {
res[i]=as.numeric(xx[i])
}
}
res
}
mergedf[,c(2:12)] <- lapply(mergedf[,c(2:12)] , text.to.num)
```

Checking dataframe

```
mergedf |>
head(5)
```

country <chr>	population <dbl>	lifeexp <dbl>	inco… <dbl>	chdperwo… <dbl>	childmort <dbl>	co2 <dbl>	gdpcapita <dbl>	healthsp <dbl>
1 Afghanistan	29200000	60.5	1960	5.82	88.0	0.29	526	
2 Albania	2950000	78.1	10700	1.65	13.3	1.56	3580	2
3 Algeria	36000000	74.5	11000	2.89	27.4	3.28	3920	1
6 Angola	23400000	60.2	7690	6.16	120.0	1.24	3990	1
9 Argentina	40900000	75.9	23500	2.37	14.4	4.57	13600	7

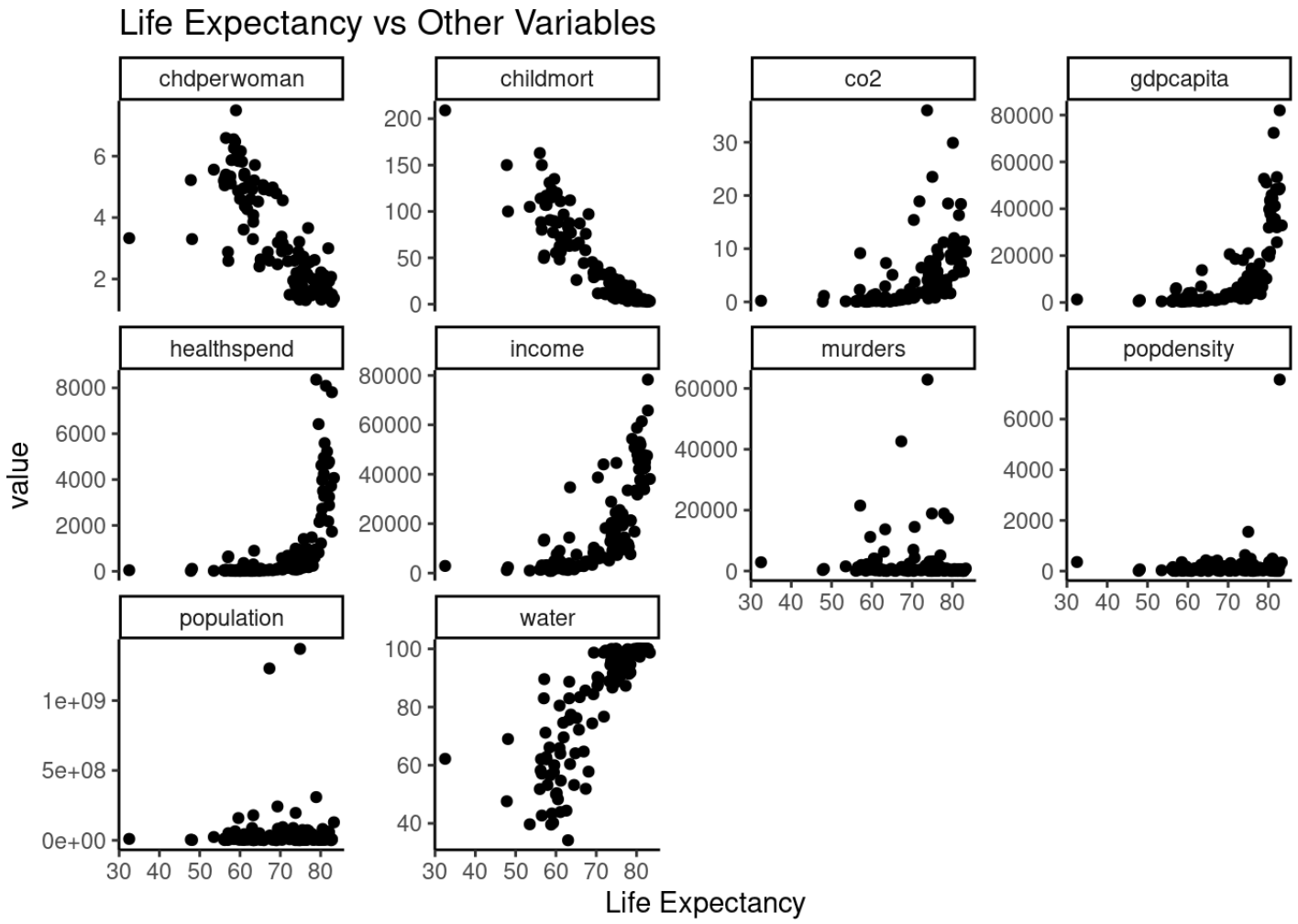
5 rows | 1-10 of 14 columns

Multiple scatterplot to view the relationship between numerical variables and life expectancy

```
mergedf_num <- select_if(mergedf, is.numeric)


mergedf_num |>
  pivot_longer(cols = -lifeexp) |>
  ggplot(aes(x = lifeexp, y = value)) +
  geom_point() +
  facet_wrap(~name, scales = "free_y") +
  labs(x = "Life Expectancy") +
  ggtitle("Life Expectancy vs Other Variables") +
  theme_classic()
```

```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```



From analysis of these scatterplots we can see that gdp per capita, water, income, healthspend, all have a strong positive linear relationship to Life Expectancy

Print Correlation Coefficient of each variable vs life expectancy

```
for (col in names(mergedf_num)) {
  corr <- cor.test(mergedf_num$lifeexp, mergedf_num[[col]], method = "pearson")
  print(paste0("Correlation coefficient between lifeexp and ", col, ": ", corr$estima
te))
}
```

```
## [1] "Correlation coefficient between lifeexp and population: 0.0314539874200825"
## [1] "Correlation coefficient between lifeexp and lifeexp: 1"
## [1] "Correlation coefficient between lifeexp and income: 0.712285940262406"
## [1] "Correlation coefficient between lifeexp and chdperwoman: -0.786356169874379"
## [1] "Correlation coefficient between lifeexp and childmort: -0.914440060050424"
## [1] "Correlation coefficient between lifeexp and co2: 0.515708105751058"
## [1] "Correlation coefficient between lifeexp and gdpcapita: 0.640010295500812"
## [1] "Correlation coefficient between lifeexp and healthspend: 0.583595855145315"
## [1] "Correlation coefficient between lifeexp and popdensity: 0.143249094568269"
## [1] "Correlation coefficient between lifeexp and water: 0.826835056514479"
## [1] "Correlation coefficient between lifeexp and murders: -0.0346536651600101"
```

Checking Summary Stats of Numerical Variables

```
lapply(mergedf, summary)
```

```
## $country
##    Length     Class      Mode
##       124 character character
##
## $population
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 1.800e+05 4.572e+06 1.090e+07 4.935e+07 3.282e+07 1.370e+09
##
## $lifeexp
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   32.50   61.88   73.75   70.25   77.95   83.30
##
## $income
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     846    3198   10550   17960   29625   78300
##
## $chdperwoman
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.260   1.855   2.595   3.161   4.810   7.490
##
## $childmort
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.620   7.293  20.200  43.605  76.025 209.000
##
## $co2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0304  0.4140  2.3950  4.6335  7.2200 36.0000
##
## $gdpcapita
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     313    1245    4890   12501   15500   82000       1
##
## $healthspend
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.90   57.75  258.00 1126.04  943.25 8360.00
##
## $popdensity
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.74   25.38   76.85  186.28  144.00 7560.00
##
## $water
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   34.20   65.60   90.10   82.35   98.78  100.00
##
## $murders
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     2.51   144.50   486.50  2597.21  1317.50 62900.00
##
## $continent
##   Africa Americas     Asia   Europe  Oceania
##       48       20       25       29        2
```

Based on these summary statistics, I will make a scaled variable of each numerical variable in the data frame which may help with the regression models, also creating income levels making a new categorical variable

```
mergedf$population_scaled <- scale(mergedf$population)
mergedf$lifeexp_scaled <- scale(mergedf$lifeexp)
mergedf$income_scaled <- scale(mergedf$income)
mergedf$chdperwoman_scaled <- scale(mergedf$chdperwoman)
mergedf$childmort_scaled <- scale(mergedf$childmort)
mergedf$co2_scaled <- scale(mergedf$co2)
mergedf$gdpcapita_scaled <- scale(mergedf$gdpcapita)
mergedf$healthspend_scaled <- scale(mergedf$healthspend)
mergedf$popdensity_scaled <- scale(mergedf$popdensity)
mergedf$water_scaled <- scale(mergedf$water)
mergedf$murders_scaled <- scale(mergedf$murders)

mergedf$incomelevels <- cut(mergedf$income, breaks=c(0, 3198, 10550, 29625, 78300, In
f), labels=c("level1", "level2", "level3", "level4", "level5"))
```

```
for (col in names(mergedf_num)) {
  skewness <- skewness(mergedf_num[[col]], na.rm = TRUE)
  cat(sprintf("Skewness of %s: %f\n", col, skewness))
}
```

```
## Skewness of population: 6.797669
## Skewness of lifeexp: -0.772795
## Skewness of income: 1.100297
## Skewness of chdperwoman: 0.681625
## Skewness of childmort: 1.129061
## Skewness of co2: 2.385358
## Skewness of gdpcapita: 1.758686
## Skewness of healthspend: 2.090870
## Skewness of popdensity: 9.871978
## Skewness of water: -0.856945
## Skewness of murders: 5.478702
```

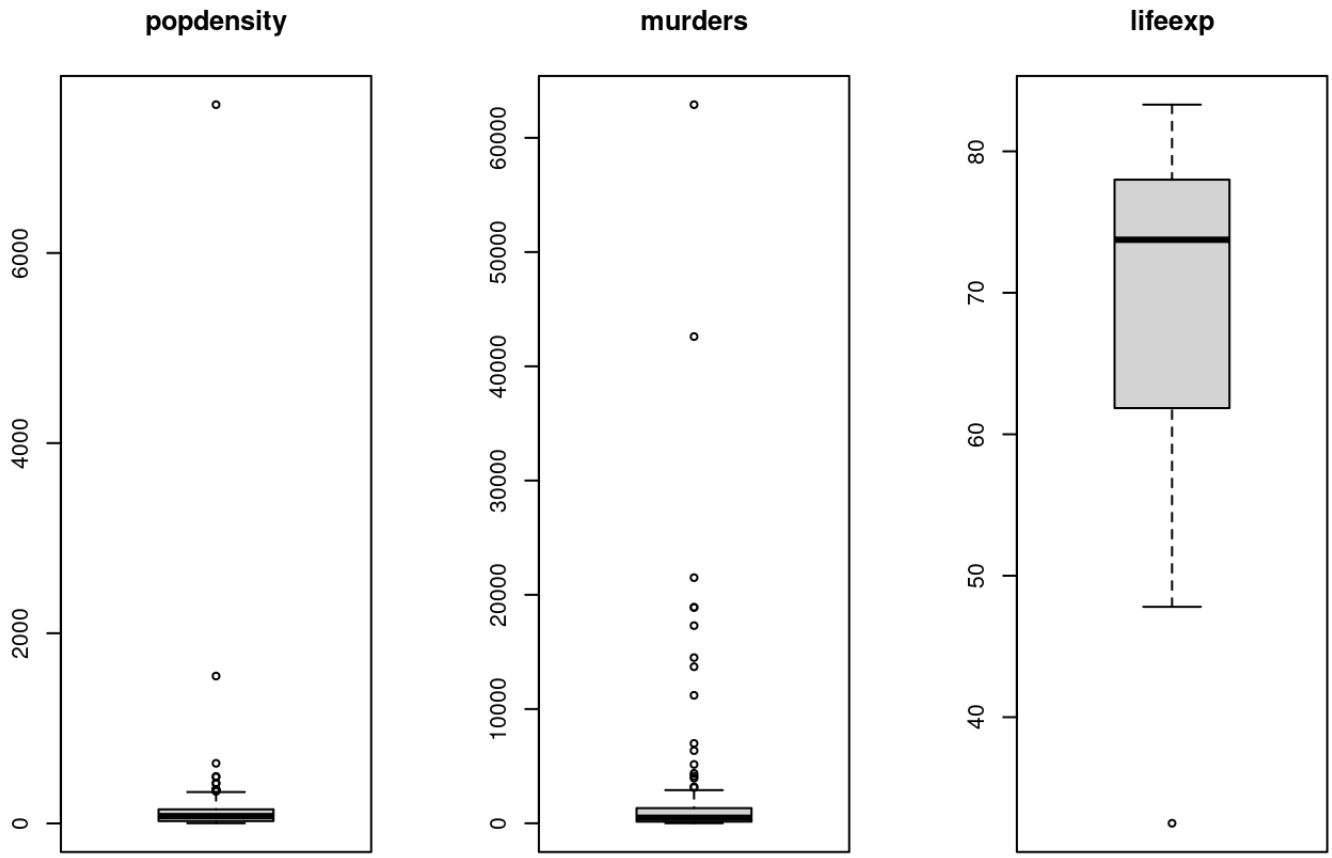
Due to high Skewness of Population, Pop Density, and Murders I will be making new variables for these taking the log function

```
mergedf$logpopulation <- log(mergedf$population)
mergedf$loggdpcapita <- log(mergedf$gdpcapita)
mergedf$loghealthspend <- log(mergedf$healthspend)
```

Outliers

```
cols <- c("popdensity", "murders", "lifeexp")


par(mfrow = c(1, length(cols))) # set up the panel
for (col in cols) {
  boxplot(mergedf_num[, col], main = col, xlab = "")
}
```




```
par(mfrow = c(1, 1))
```

*From viewing these multiple boxplots of each numeric column in the data frame we can see that the data clearly contains some outliers in some of the numeric columns for example outliers are evident in popdensity, murders, and lifeexp, as they have points that are outside of the whiskers making these points to be outliers through exploratory data analysis.**

Additional Data Transformation

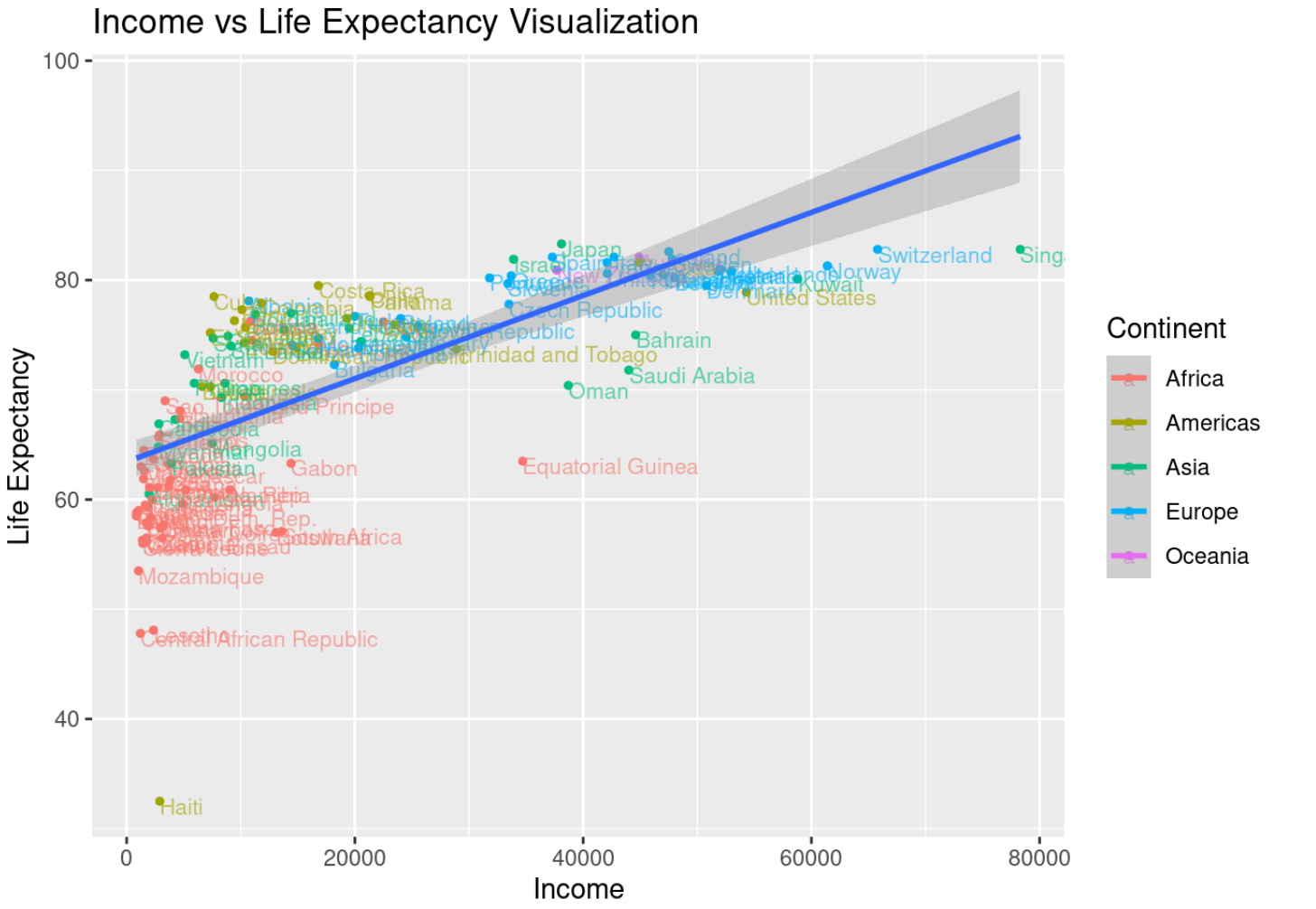
```
mergedf <- mergedf %>% mutate(water2=water^2)
mergedf <- mergedf %>% mutate(childmort2=childmort^2)
mergedf <- mergedf %>% mutate(chd2=chdperwoman^2)
mergedf <- mergedf %>% mutate(logwater=log(water))
mergedf <- mergedf %>% mutate(logchildmort=log(childmort))
mergedf <- mergedf %>% mutate(logchd=log(chdperwoman))
mergedf <- mergedf %>% mutate(logco2=log(co2))
```

Exploratory Data Analysis First ggplot to view Country and their respective 'lifeexp vs income'

```
mergedf |>
  ggplot(aes(x = income, y = lifeexp, color = continent)) +
  geom_point(size = 1) +
  geom_text(aes(label = country), hjust = 0, vjust = 0.8, size = 3,alpha = 0.6) +
  labs(x = "Income", y = "Life Expectancy", color = "Continent") +
  ggtitle("Income vs Life Expectancy Visualization") +
  stat_smooth(method = "lm", se = TRUE, level = 0.95, aes(group = 1))
```

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

```
## Warning: The following aesthetics were dropped during statistical transformation:
colour
## ℹ This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```

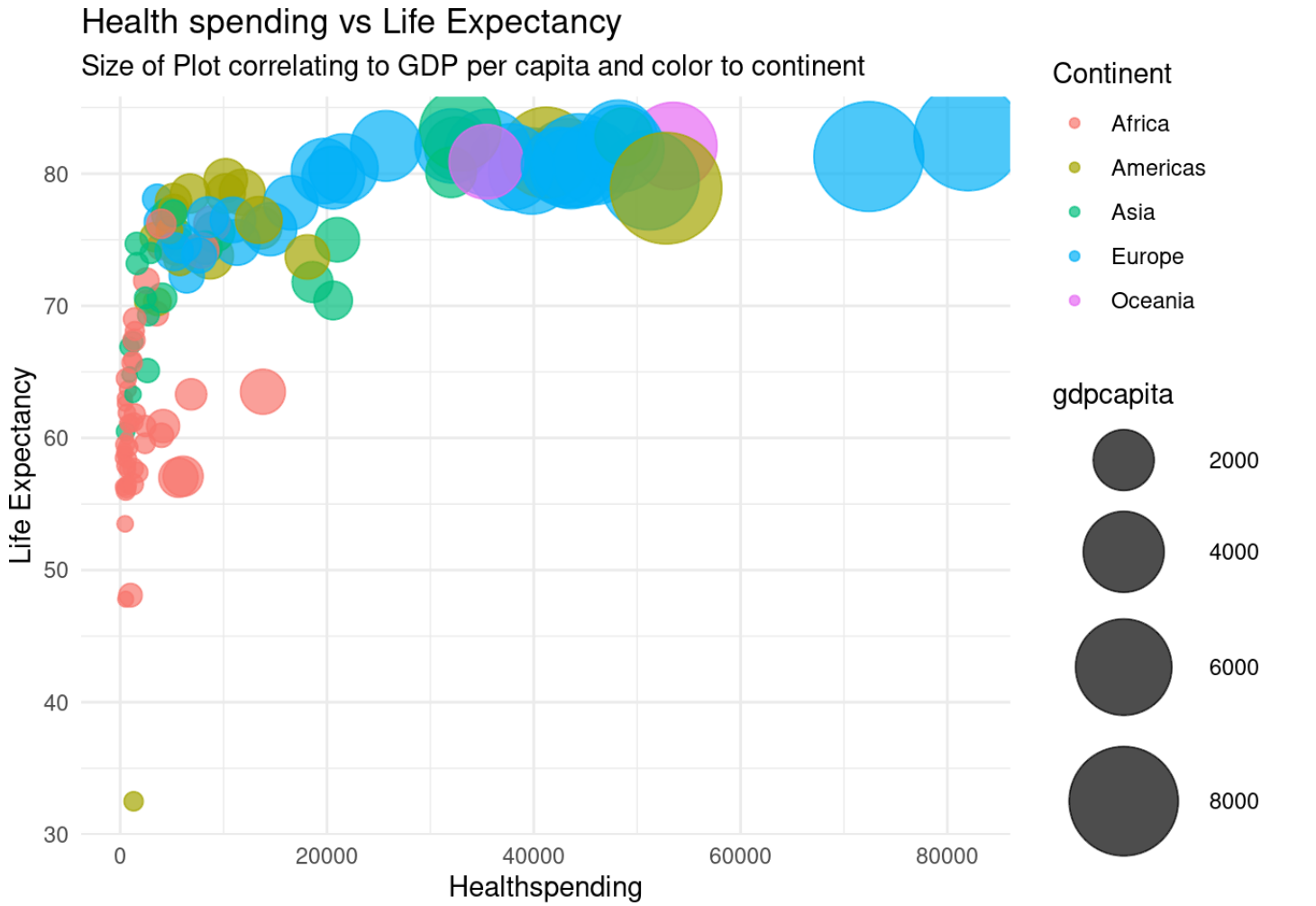


Graph indicates a positive linear relationship between the Life Expectancy and Income Variables.Nations in Africa seem to have the lowest income and life expectancy while nations in Europe seem to have the highest life expectancy and income.

Second ggplot to impact of Health Spending on Life expectancy based on gdp per capita and continent

```
mergedf |>
  ggplot(aes(x = gdpcapita, y = lifeexp, size = healthspend, color = continent)) +
  geom_point(alpha = 0.7) +
  scale_size(range = c(2, 20)) +
  labs(x = "Healthspending", y = "Life Expectancy", color = "Continent", size = "gdpc
apita") +
  ggtitle("Health spending vs Life Expectancy",subtitle = "Size of Plot correlating t
o GDP per capita and color to continent") +
  theme_minimal()
```

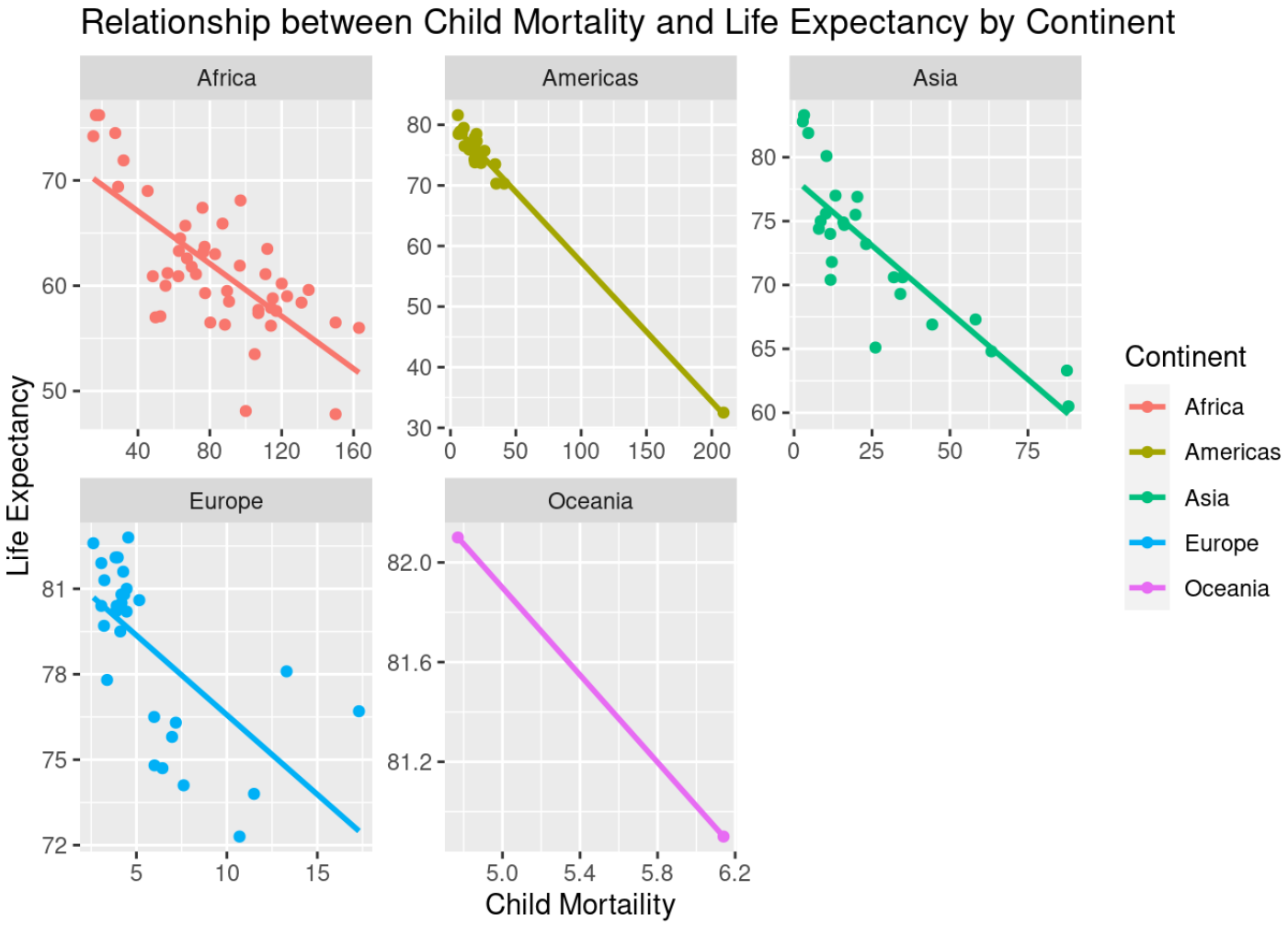
```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```



From viewing this ggplot, we can see there is a positive relationship between HealthSpending and GDP per capita as the size of the bubbles are increasing as Health Spending increases. We can also see that the lower the health spending and gdp per capita the lower life expectancy as well. We can also take note of the European nations which are the highest in life expectancy who are also having the highest gdp per capita as well as Health Spending. On the other hand, we can see that African nations have the lowest life expectancy while also having the lowest GDP per capita and Health Spending.

```
mergedf |>
  ggplot(aes(x = childmort, y = lifeexp, color = continent)) +
  geom_point() +
  stat_smooth(method = "lm", se = FALSE, aes(group = 1)) +
  facet_wrap(~ continent, nrow = 2, scales = "free") +
  labs(x = "Child Mortaility", y = "Life Expectancy", color = "Continent") +
  ggtitle("Relationship between Child Mortality and Life Expectancy by Continent")
```

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



From viewing this plot using, stratified scatter plots with paneling, we can see the strong negative linear relationship between Child Mortality and Life Expectancy, regardless of the continent the nation is in.

Regression Models

Initial Model 1

```
reg1 <- lm(data=mergedf, lifeexp ~ water + water2 + childmort + continent)
get_regression_summaries(reg1)
```

r_squared <dbl>	adj_r_squared <dbl>	mse <dbl>	rmse <dbl>	sigma <dbl>	statistic <dbl>	p_value <dbl>	df <dbl>	n... <dbl>
0.869	0.861	12.00856	3.465337	3.583	109.802	0	7	124

1 row

Initial Model 2

```
reg2 <- lm(data=mergedf, lifeexp ~ childmort*water + continent + chdperwoman)
get_regression_summaries(reg2)
```

r_squared <dbl>	adj_r_squared <dbl>	mse <dbl>	rmse <dbl>	sigma <dbl>	statistic <dbl>	p_value <dbl>	df <dbl>	n... <dbl>
0.889	0.882	10.13844	3.184091	3.306	115.47	0	8	124

1 row

Initial Model 3

```
reg3 <- lm(data=mergedf, lifeexp ~ logchildmort*logwater + continent)
get_regression_summaries(reg3)
```

r_squared <dbl>	adj_r_squared <dbl>	mse <dbl>	rmse <dbl>	sigma <dbl>	statistic <dbl>	p_value <dbl>	df <dbl>	n... <dbl>
0.854	0.846	13.32473	3.650306	3.774	97.319	0	7	124

1 row

Reasons for rejecting the above models

#The above code displays some of our preliminary models, which were not selected as our final model for several reasons. One model showed a non-significant p-value for the childmort variable, while the other two models had lower-than-expected R-squared values and higher RMSE values. Our final model, which we will explain below, outperformed all other tested models in terms of metrics.

Final Regression Model

```
finalmodel <- lm(lifeexp~(income_scaled*gdpcapita_scaled)+ childmort + chd2 + water2
+(chdperwoman_scaled*childmort)+ (healthspend*gdpcapita) ,data = mergedf)
summary(finalmodel)
```

```
##
## Call:
## lm(formula = lifeexp ~ (income_scaled * gdpcapita_scaled) + childmort +
##     chd2 + water2 + (chdperwoman_scaled * childmort) + (healthspend *
##     gdpcapita), data = mergedf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9741 -1.3631  0.0497  1.8665  9.1432
##
## Coefficients: (1 not defined because of singularities)
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   7.990e+01  3.932e+00  20.321  < 2e-16 ***
## income_scaled                -2.259e+00  1.193e+00  -1.894  0.06076 .
## gdpcapita_scaled              2.878e+00  1.962e+00   1.467  0.14527
## childmort                    -2.189e-01  1.683e-02 -13.007  < 2e-16 ***
## chd2                         -7.209e-01  2.639e-01  -2.732  0.00732 **
## water2                        6.680e-04  2.486e-04   2.687  0.00831 **
## chdperwoman_scaled            4.610e+00  2.442e+00   1.888  0.06166 .
## healthspend                   8.928e-04  8.614e-04   1.036  0.30227
## gdpcapita                            NA         NA      NA       NA
## income_scaled:gdpcapita_scaled  8.240e-01  6.723e-01   1.226  0.22287
## childmort:chdperwoman_scaled    9.689e-02  1.842e-02   5.260 7.01e-07 ***
## healthspend:gdpcapita         -2.871e-08  1.620e-08  -1.772  0.07910 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.942 on 112 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.9142, Adjusted R-squared:  0.9066
## F-statistic: 119.4 on 10 and 112 DF,  p-value: < 2.2e-16
```

Steps to Reach Best Model

Our final model to predict life expectancy included interactions between income_scales and gdpcapita_scales, child mortality, child per women squared, water squared, interaction between child per woman squared and child mort, and finally the interaction between health spending and gdpcapita. Log terms were created for some variables but not included in this model, as these variables were either slightly skewed or skewed, taking the log function would allow the distribution to be normal and help with prediction but including these log variables did not result in a statistically significant p value for that predictor. Some of the variables also included

in this prediction are scaled as they have a lot of variability and will allow variables to be on the same scale as others and have been included due to some of there significance. From the beginning exploratory analysis conducted, it was clear some variables such as income, gdp, water, health spending all had a positive linear relationship with life expectancy while other variables such as childmort had a clear negative linear relationship. Using this information we were able to different combinations of interactions terms, second order terms, building upon on our initial models and slightly adjusting the model each time to gain a higher r^2 led to this final model which has an adjusted r^2 of 90.66%. While our main method of obtaining this final model was starting off with a simple model, changing predictors, interaction terms, and order of terms, to obtain a higher r^2 each time, the exploratory data analysis is truly what guided this project and allowed us to pick the proper variables for predicting life expectancy to create a model with high adjusted r^2. Our group spent a lot of time trying each predictor with the scaled version, 2nd order, and including each predictor in a reasonable interaction term, for example water^2 proved to be a very important predictor that is statistically significant as well as child per women^2. Scaling also did not seem to have a major impact on the p value of most predictors for most predicotrs made little to no difference. Interaction terms such as child mortaility and child per women proved to be very significant at childmort:chdperwoman_scaled = p value= 7.01e-07. Ultimately, we decided on this final model due to the adjusted r^2 above 90%.Some of predictor/interaction terms although not signicant are very close to being signicant and increase the adjusted r^2 of the model so we have chosen to keep them

Multi-Collinearity in Model

```
cor(mergedf$chd2, mergedf$childmort)
```

```
## [1] 0.8407783
```

```
cor(mergedf$childmort, mergedf$water2)
```

```
## [1] -0.8637717
```

```
cor(mergedf$chd2, mergedf$water2)
```

```
## [1] -0.8644546
```

```
cor(mergedf$healthspend, mergedf$income)
```

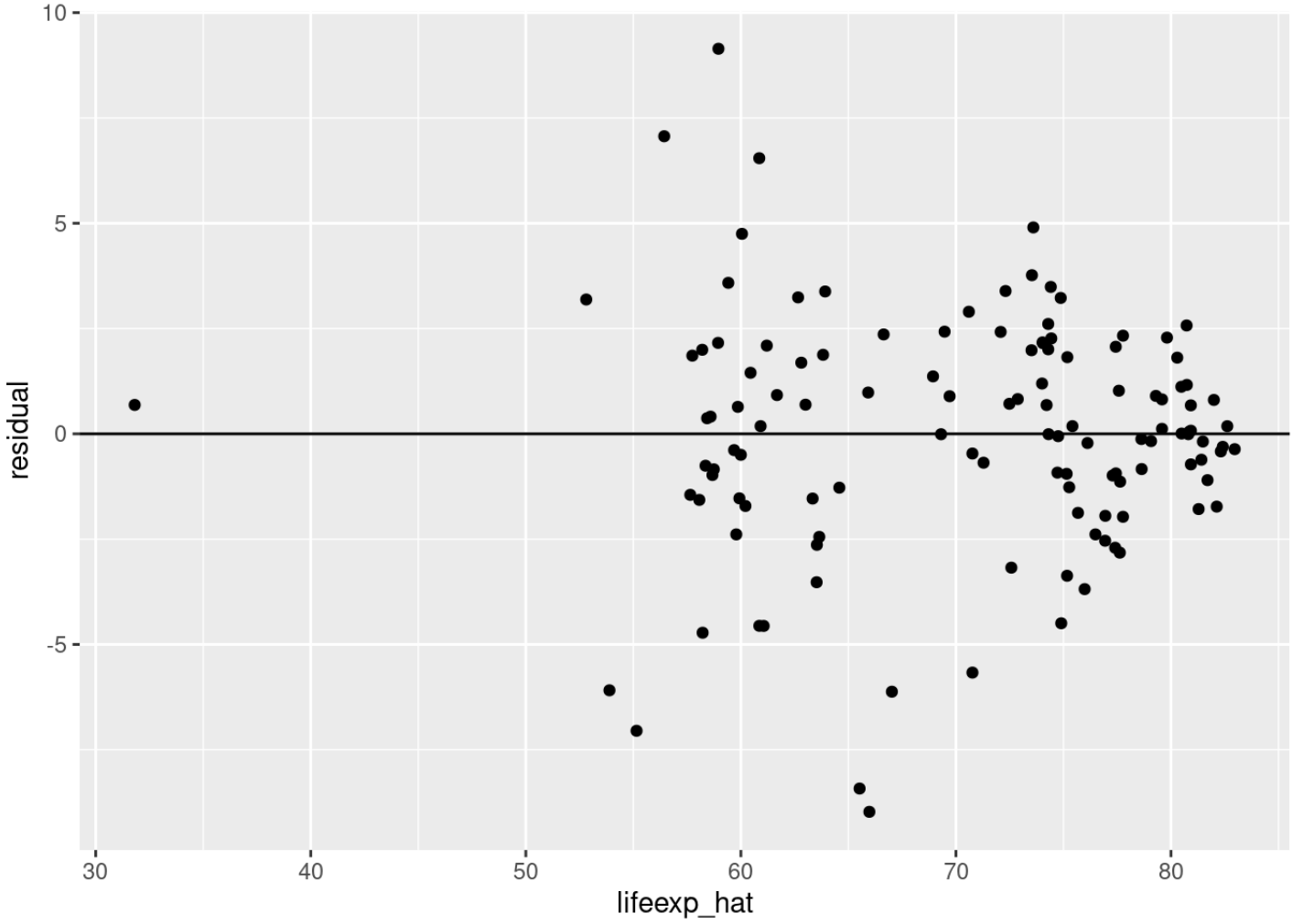
```
## [1] 0.8401676
```

Multi-Collinearity exists between childmort and chd2, childmort and water, chd2 and water, and healthspend and income. Due to the high correlation coefficients, it is apparent that child mortality is likely to increase as the number of children per woman rises. Similarly, it is reasonable that child mortality would increase with low

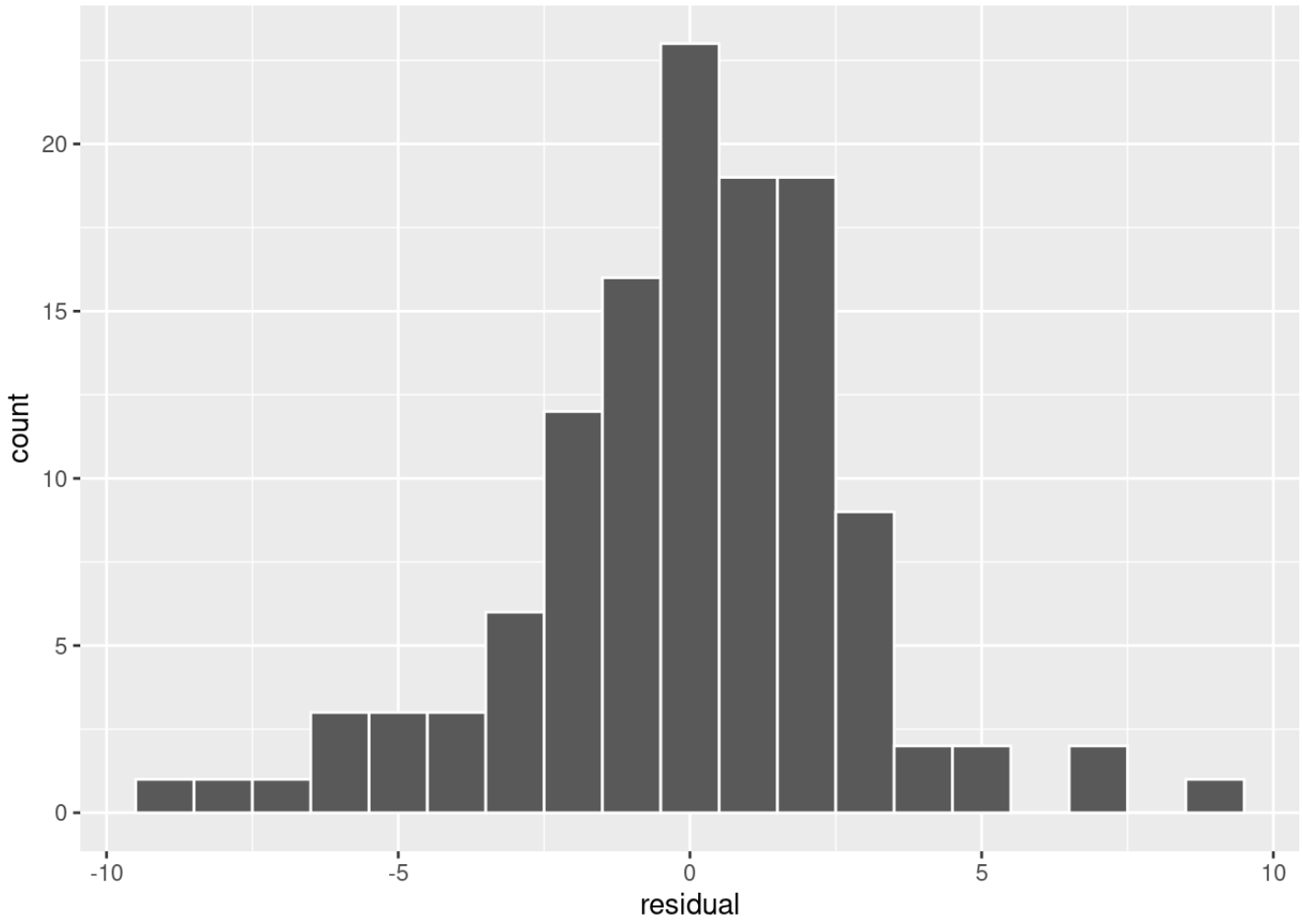
water levels, and children per women increase if there is lower water levels since the child mortality is higher when there is higher children per women. In addition, the strong correlation coefficients of healthspend and income also proves Multi-Collinearity. Although these variables exhibit Multi-Collinearity, we chose to include them in our model since this phenomenon was already present in the dataset. Without them, we would not have been able to use any variables to explain life expectancy.

Residual Analysis

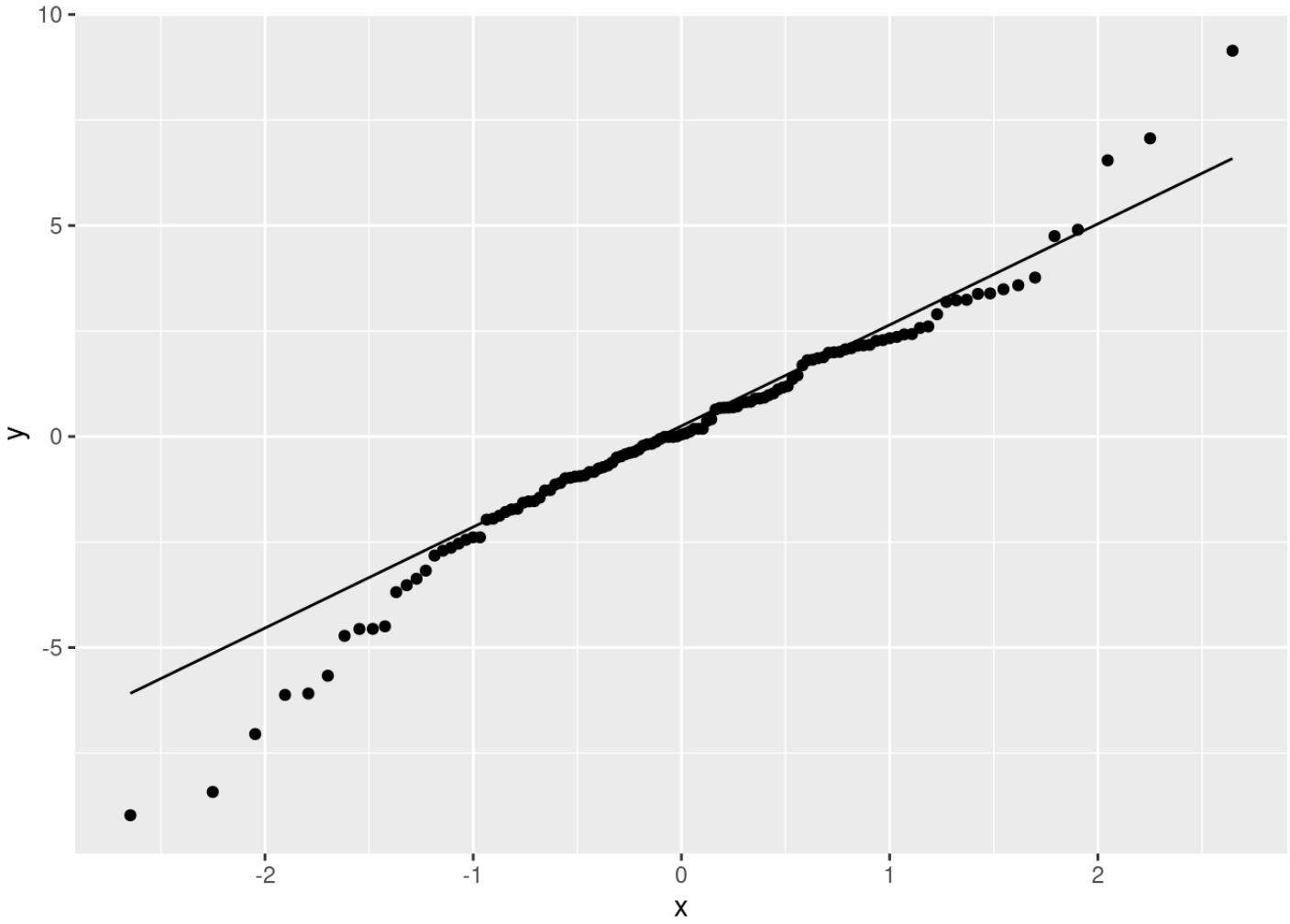
```
points1 <- get_regression_points(finalmodel)
ggplot(data=points1, aes(x=lifeexp_hat, y=residual)) + geom_point() + geom_hline(yint
ercept=
0)
```




```
ggplot(data=points1, aes(x=residual)) + geom_histogram(binwidth=1, color="white")
```




```
ggplot(data=points1, aes(sample=residual)) + stat_qq() + stat_qq_line()
```



#Mean 0: Assumption not violated

#Constant variance: Violated due to the clear cone-shaped pattern evident in the scatter plot

#Normality: Even though the histogram is normally distributed, but there are irregularities in the QQ plot.

#Independence: Assumption violated as there is a clear straight line and correlation in the scatterplot.

Based on the report we presented, we can conclude that child mortality, number of children per woman, water levels, and continent can be used to predict and explain life expectancy using the given dataset. Also, based on the summary statistics, the p-values for childmort, chd2, water2, and the interaction between childmort and chdperwoman_scaled in the final model appear to be statistically significant. We arrived at this conclusion using a model that demonstrated a high R-squared value, low RMSE, and a significant p-value for the overall model. However, we recognize that our model was not perfect as there was Multi-Collinearity among the explanatory variables and some of the assumptions for multiple linear regression were violated for our model. Despite these limitations, our model was the best out of all the models we tested both intuitively and statistically.