PSTAT 126 Project 1

Wilber Delgado & Mason Delan

2023-10-23

Data Description

The World Happiness Report is an annual publication that provides insights into the subjective well-being and happiness levels of countries around the world.

1. Country name: This column contains the names of different countries included in the report. Each row corresponds to a specific country.
2. Year: The year column represents the specific year for which the happiness data is recorded. It indicates the time period during which the survey or data collection took place.
3. Life Ladder: The Life Ladder column measures the overall subjective well-being or life satisfaction of individuals in a country. It is typically represented on a scale from 0 to 10, where higher values indicate greater life satisfaction.
4. Log GDP per capita: This column represents the logarithm of the Gross Domestic Product (GDP) per capita of a country, which is calculated by how much each country produces, divided by the number of people in the country. It serves as a measure of economic prosperity or standard of living, with higher values indicating higher GDP per person.
5. Social Support: The Social Support column assesses the availability and strength of social networks and support systems within a country. It measures the degree to which individuals have assistance from family, friends, and other social connections. The question associated with this column was “If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?”.
6. Health life expectancy: This column represents the average life expectancy or the number of years a person can expect to live in good health. It serves as an indicator of the physical health, mental health, and well-being of the population.
7. Freedom to make Life Choices: This column measures the extent to which individuals perceive having freedom and autonomy in making life choices, such as career, relationships, and personal decisions. Higher values indicate greater freedom. The question associated with this column is “Are you satisfied or dissatisfied with your freedom to choose what you do with your life?”.
8. Generosity: The Generosity column reflects the level of generosity or willingness to help others within a country’s population. It measures the frequency of charitable donations and voluntary work. The questions associated with this question is “Have you donated money to a charity in the past month?”.
9. Perception of Corruption: This column captures the perceived level of corruption within a country. It assesses the extent to which corruption is perceived to be prevalent in public institutions and the overall trust in the government and public sector. The questions associated with this column are “Is corruption widespread throughout the government or not” and “Is corruption widespread within businesses or not?”.
10. Positive affect: The Positive affect column measures the frequency and intensity of positive emotions experienced by individuals in a country. The result of this column is given by the average of yes or no answers based on emotions of laughter, enjoyment, and interest.

11. Negative affect: The Negative affect column assesses the frequency and intensity of negative emotions experienced by individuals in a country. The result of this column is given by the average of yes or no answers based on emotions of worry, sadness, and anger.

```
library("readxl")
WHR = read_excel("D:/WHR2023.xls")
```

```
str(WHR)
```

```
## tibble [2,199 × 11] (S3: tbl_df/tbl/data.frame)
##  $ Country name                    : chr [1:2199] "Afghanistan" "Afghanistan" "Afghanistan"
"Afghanistan" ...
##  $ year                            : num [1:2199] 2008 2009 2010 2011 2012 ...
##  $ Life Ladder                     : num [1:2199] 3.72 4.4 4.76 3.83 3.78 ...
##  $ Log GDP per capita              : num [1:2199] 7.35 7.51 7.61 7.58 7.66 ...
##  $ Social support                  : num [1:2199] 0.451 0.552 0.539 0.521 0.521 ...
##  $ Healthy life expectancy at birth: num [1:2199] 50.5 50.8 51.1 51.4 51.7 ...
##  $ Freedom to make life choices    : num [1:2199] 0.718 0.679 0.6 0.496 0.531 ...
##  $ Generosity                      : num [1:2199] 0.168 0.191 0.121 0.164 0.238 ...
##  $ Perceptions of corruption       : num [1:2199] 0.882 0.85 0.707 0.731 0.776 ...
##  $ Positive affect                 : num [1:2199] 0.414 0.481 0.517 0.48 0.614 ...
##  $ Negative affect                 : num [1:2199] 0.258 0.237 0.275 0.267 0.268 ...
```

```
summary(WHR)
```

```
##  Country name            year       Life Ladder    Log GDP per capita
##  Length:2199        Min.   :2005   Min.   :1.281   Min.   : 5.527
##  Class :character   1st Qu.:2010   1st Qu.:4.647   1st Qu.: 8.500
##  Mode  :character   Median :2014   Median :5.432   Median : 9.499
##                     Mean   :2014   Mean   :5.479   Mean   : 9.390
##                     3rd Qu.:2018   3rd Qu.:6.309   3rd Qu.:10.373
##                     Max.   :2022   Max.   :8.019   Max.   :11.664
##                                                    NA's   :20
##  Social support   Healthy life expectancy at birth Freedom to make life choices
##  Min.   :0.2282   Min.   : 6.72                    Min.   :0.2575
##  1st Qu.:0.7466   1st Qu.:59.12                    1st Qu.:0.6565
##  Median :0.8355   Median :65.05                    Median :0.7698
##  Mean   :0.8107   Mean   :63.29                    Mean   :0.7479
##  3rd Qu.:0.9048   3rd Qu.:68.50                    3rd Qu.:0.8594
##  Max.   :0.9873   Max.   :74.47                    Max.   :0.9852
##  NA's   :13       NA's   :54                       NA's   :33
##    Generosity       Perceptions of corruption Positive affect
##  Min.   :-0.33753   Min.   :0.0352            Min.   :0.1789
##  1st Qu.:-0.11212   1st Qu.:0.6881            1st Qu.:0.5717
##  Median :-0.02267   Median :0.7996            Median :0.6631
##  Mean   : 0.00010   Mean   :0.7452            Mean   :0.6521
##  3rd Qu.: 0.09207   3rd Qu.:0.8688            3rd Qu.:0.7379
##  Max.   : 0.70271   Max.   :0.9833            Max.   :0.8836
##  NA's   :73         NA's   :116               NA's   :24
##  Negative affect
##  Min.   :0.08274
##  1st Qu.:0.20766
##  Median :0.26067
##  Mean   :0.27150
##  3rd Qu.:0.32289
##  Max.   :0.70459
##  NA's   :16
```

```
library(skimr)

skim_summary <- skim(WHR[c('Life Ladder', 'Log GDP per capita', 'Social support', 'Healthy life expectancy at birth', 'Freedom to make life choices', 'Generosity', 'Perceptions of corruption', 'Positive affect', 'Negative affect')])

skim_summary
```

Data summary

Name	…[]
Number of rows	2199
Number of columns	9
_______________________	

Column type frequency:	
numeric	9
________________________	
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Life Ladder	0	1.00	5.48	1.13	1.28	4.65	5.43	6.31	8.02	▁▂▇▇▃
Log GDP per capita	20	0.99	9.39	1.15	5.53	8.50	9.50	10.37	11.66	▁▃▆▇▅
Social support	13	0.99	0.81	0.12	0.23	0.75	0.84	0.90	0.99	▁▁▂▆▇
Healthy life expectancy at birth	54	0.98	63.29	6.90	6.72	59.12	65.05	68.50	74.47	▁▁▁▃▇
Freedom to make life choices	33	0.98	0.75	0.14	0.26	0.66	0.77	0.86	0.99	▁▂▅▇▆
Generosity	73	0.97	0.00	0.16	-0.34	-0.11	-0.02	0.09	0.70	▃▇▃▁▁
Perceptions of corruption	116	0.95	0.75	0.19	0.04	0.69	0.80	0.87	0.98	▁▁▁▃▇
Positive affect	24	0.99	0.65	0.11	0.18	0.57	0.66	0.74	0.88	▁▁▆▇▆
Negative affect	16	0.99	0.27	0.09	0.08	0.21	0.26	0.32	0.70	▃▇▃▁▁

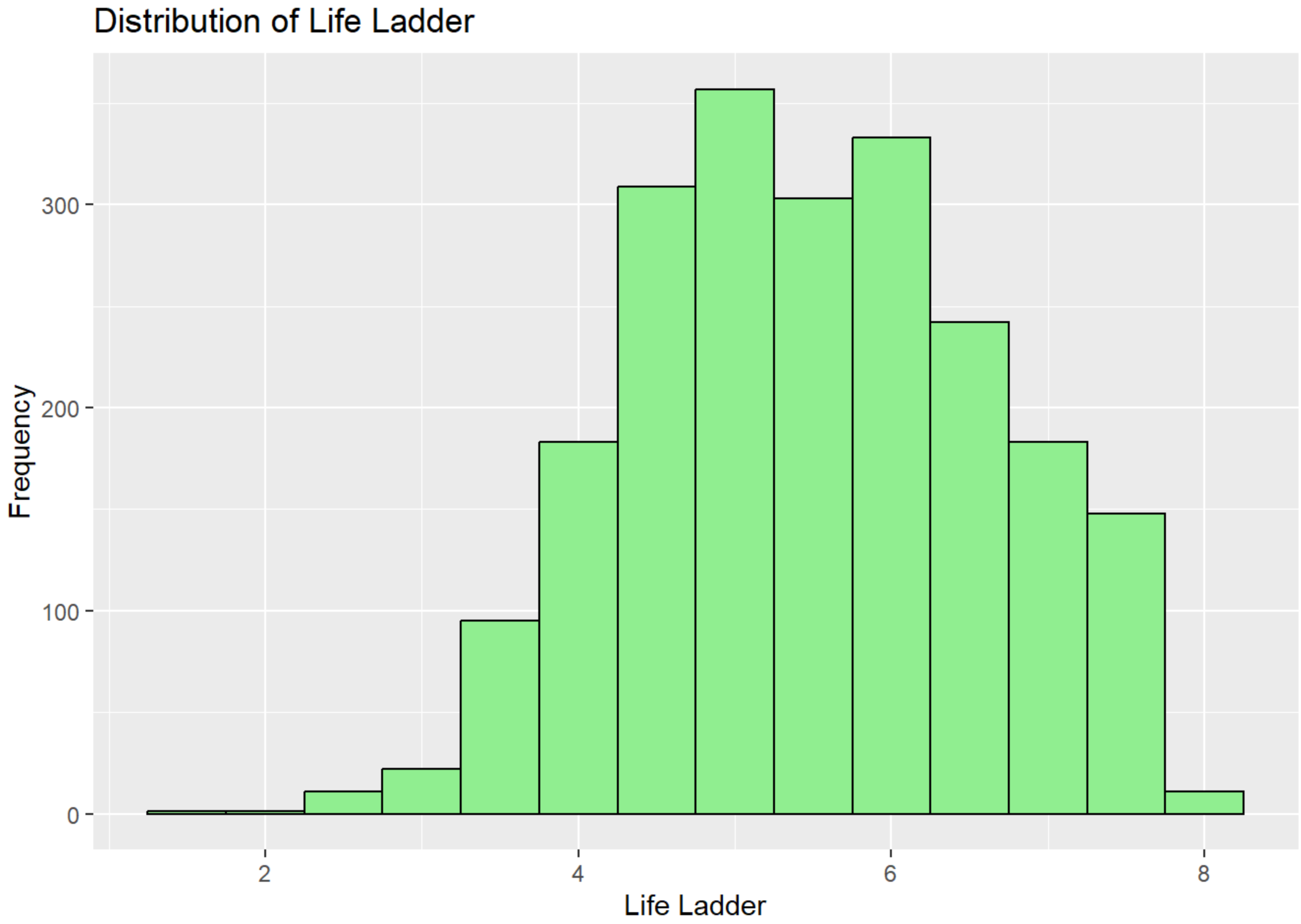
```
sum(is.na(WHR))
```

```
## [1] 349
```

The skim summary shows us the sum of missing data per variable. We can see that ‘Perceptions of corruption’ has the highest number of missing values, which may suggest that our analysis of this variable may be skewed in our project. Furthermore, we see that the ‘Life Ladder’ variable has no missing values which may suggest that the researchers primarily focus on the Life Ladder question when surveying people. By calculating the number of missing values, we see that there is a total of 349.

```
library(ggplot2)

ggplot(WHR, aes(x = `Life Ladder`)) +
  geom_histogram(binwidth = 0.5, fill = "lightgreen", color = "black") +
  labs(x = "Life Ladder", y = "Frequency", title = "Distribution of Life Ladder")
```

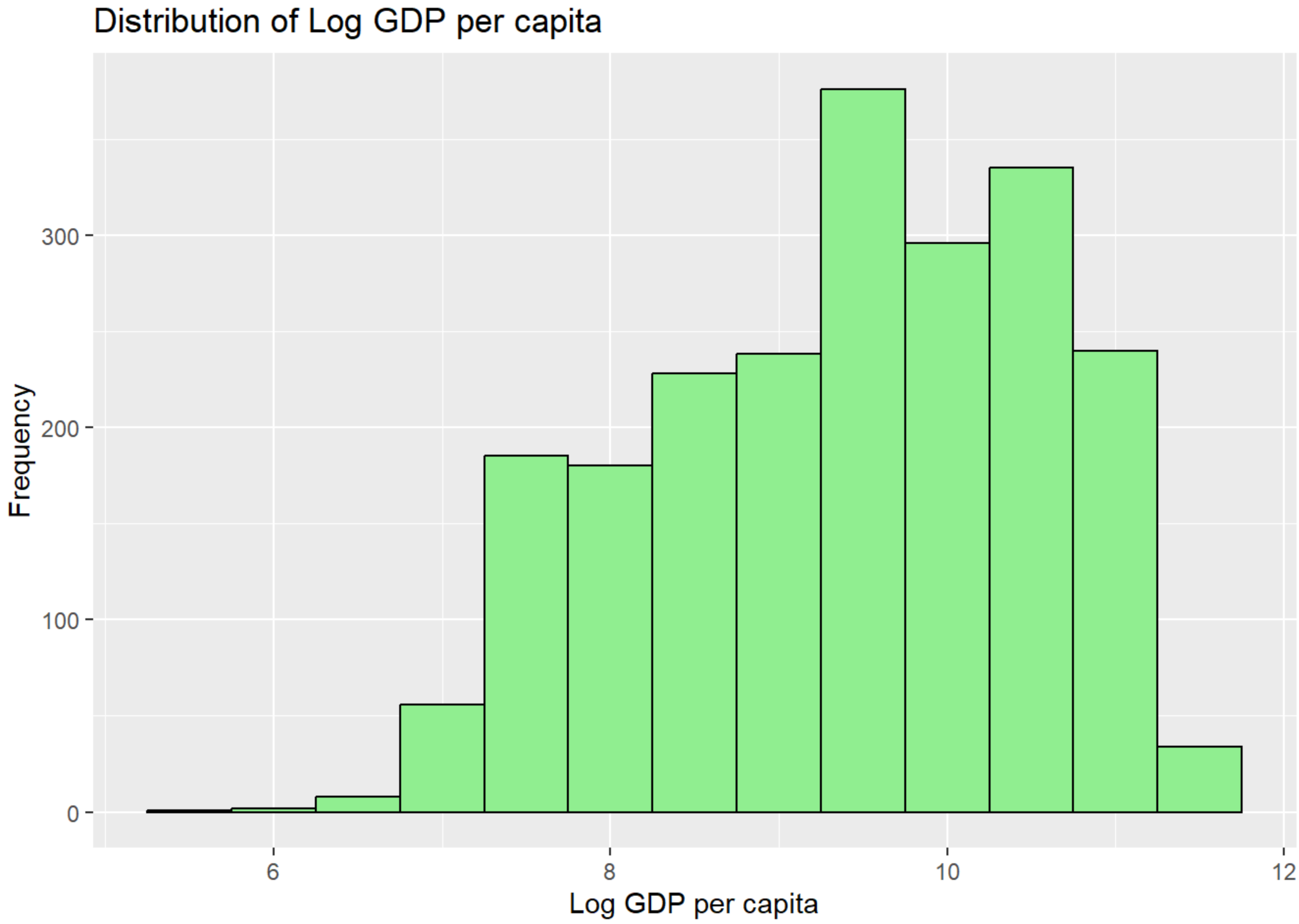


Worldwide from 2005 to 2023, we see that Life Ladder has higher frequencies for values approximately at 5 & 6, with 2 and 8 seeming to be the least frequent in the data. Based on this we can assume that a majority of people are content with their lives at a average (subjective) well being . The graphic of this model is bell-shaped.

```r
library(ggplot2)

ggplot(WHR, aes(x = `Log GDP per capita`)) +
  geom_histogram(binwidth = 0.5, fill = "lightgreen", color = "black") +
  labs(x = "Log GDP per capita", y = "Frequency", title = "Distribution of Log GDP per capita")
```

```
## Warning: Removed 20 rows containing non-finite values (`stat_bin()`).
```

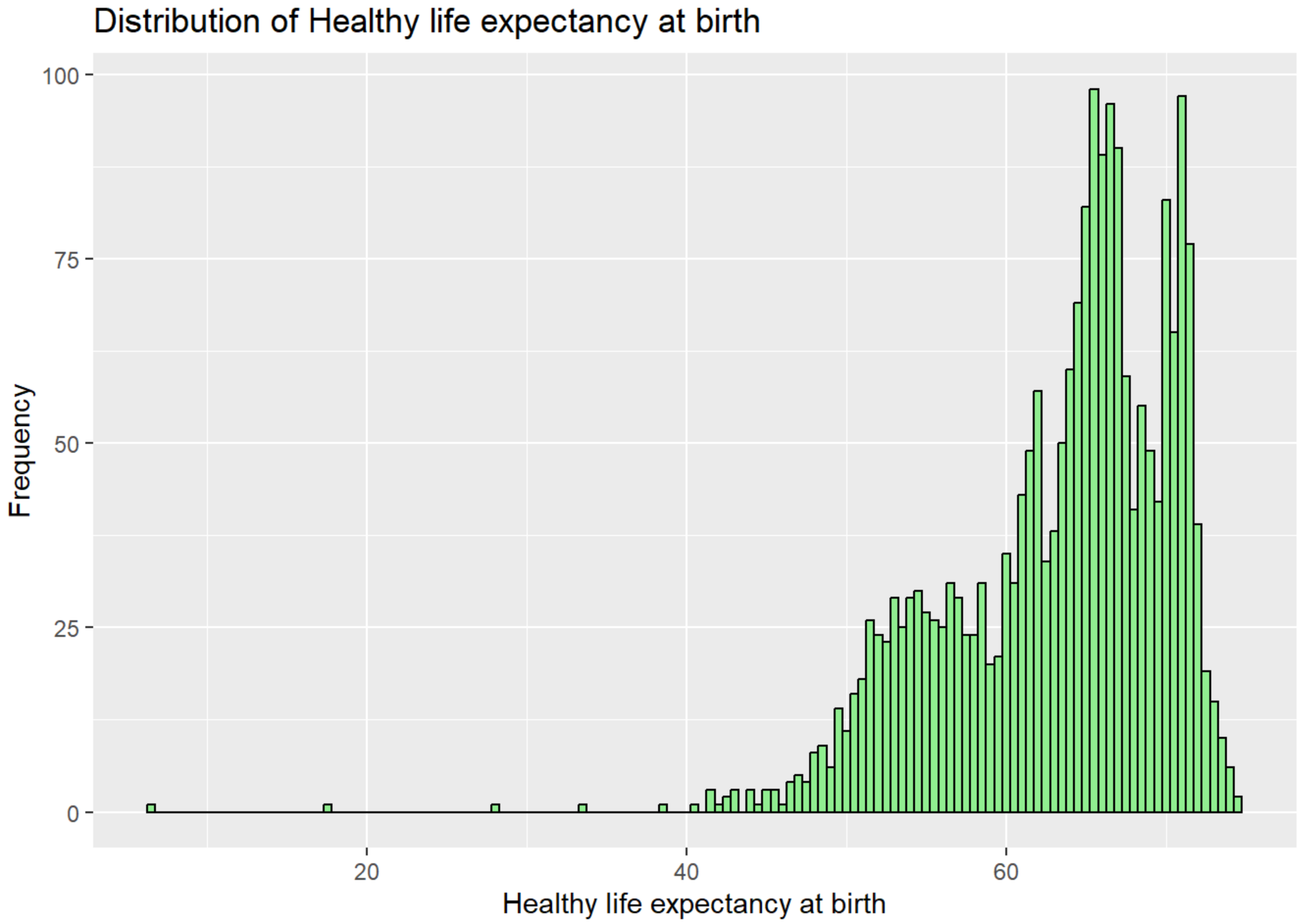


Here we see that the Log GDP per capita has high frequencies among all countries. We can interpret this due to the non-bell shaped like curve.

```
library(ggplot2)

ggplot(WHR, aes(x = `Healthy life expectancy at birth`)) +
  geom_histogram(binwidth = 0.5, fill = "lightgreen", color = "black") +
  labs(x = "Healthy life expectancy at birth", y = "Frequency", title = "Distribution of Healthy
life expectancy at birth")
```

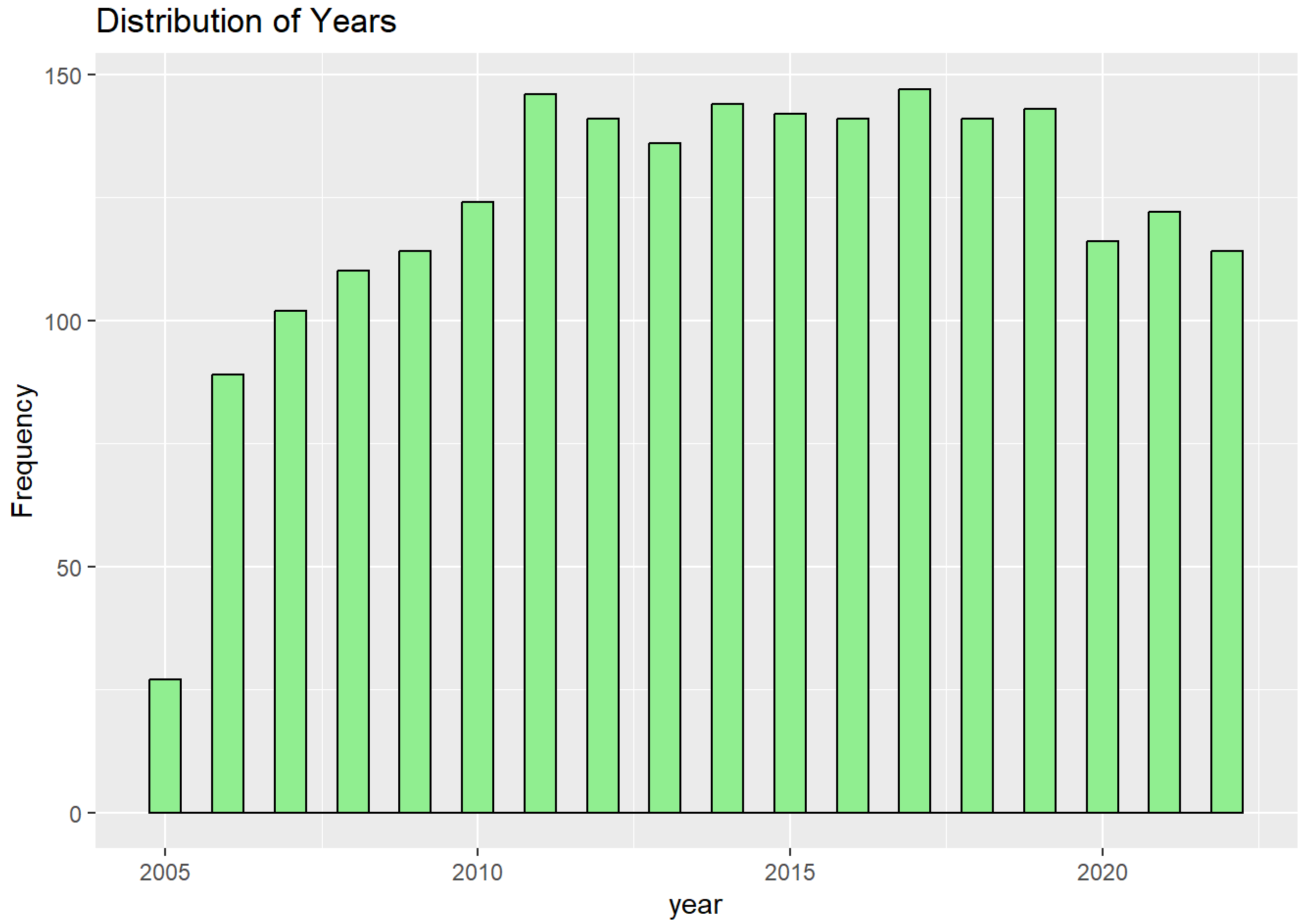
```
## Warning: Removed 54 rows containing non-finite values (`stat_bin()`).
```



Based on the plot, we can observe that there is an exponential shape. This makes sense considering that the worlds average life expectancy in 2023 is 73.16. We can see that 40 years old has a much smaller frequency than 60 years old showing that on average people are more likely to stay healthy up until their 60s where in the late 60s. There is a noticeable decline in frequencies indicating that people are less healthy at those respective ages.

```
library(ggplot2)

ggplot(WHR, aes(x = `year`)) +
  geom_histogram(binwidth = 0.5, fill = "lightgreen", color = "black") +
  labs(x = "year", y = "Frequency", title = "Distribution of Years")
```



Viewing this plot we can see that 2005 seems to be the year when the collection of data began. It exponentially increased until the year 2011, the following years show a significant plateau.