A2Q2

Ian Murphy

2019-09-19

Question 2

Start by loading packages.

```
# load your packages here:
library(gapminder)
library(tidyverse)
```

Range of Values

Our goal is to investigate a categorical variable and a quantitative variable. For this exercise, the categorical variable will be *continent* and the quantitative variable will be

```
gapminder %>%
  select(continent) %>%
  distinct(continent)
```

```
## # A tibble: 5 x 1
##   continent
##   <fct>
## 1 Asia
## 2 Europe
## 3 Africa
## 4 Americas
## 5 Oceania
```

Looking at the output, we see that there are 5 distinct categories for continent: Asia, Europe, Africa, Americas, Oceania. This is the possible range.

For our continuous variable, we will choose lifeExp, only from 2007. This will make the analysis more clear.

```
lifeExp2007 <- gapminder %>%
                 select(country, lifeExp, year) %>%
                 filter(year == 2007)

lifeExp2007
```

```
## # A tibble: 142 x 3
##    country     lifeExp  year
##    <fct>         <dbl> <int>
##  1 Afghanistan    43.8  2007
##  2 Albania        76.4  2007
##  3 Algeria        72.3  2007
##  4 Angola         42.7  2007
##  5 Argentina      75.3  2007
##  6 Australia      81.2  2007
##  7 Austria        79.8  2007
##  8 Bahrain        75.6  2007
```

```
##  9 Bangladesh     64.1  2007
## 10 Belgium        79.4  2007
## # ... with 132 more rows
```

Now, we can look at the range of possible values.

```
lifeExp2007 %>%
  select(lifeExp) %>%
  range()
```

```
## [1] 39.613 82.603
```

So, our range of possible values is 39.613 to 82.603.

Typical Values

For the categorical variable, we can look at the distribution of the continents. To do so, it may just be nice to know the count for each continent, then present that in a graph:

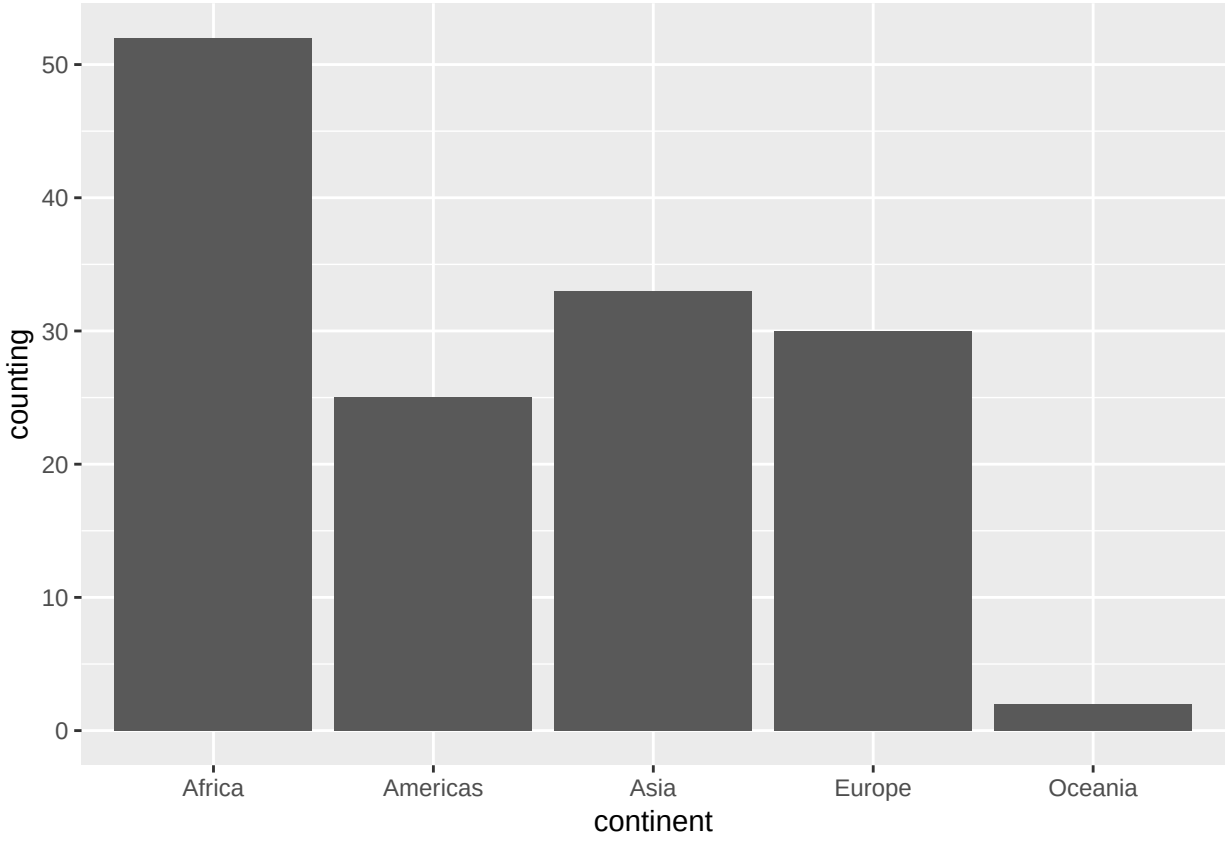
```
dist_cont <- gapminder %>%
  select(continent, country, year) %>%
  filter(year == 2007) %>%
  select(continent) %>%
  group_by(continent) %>%
  summarize(n()) %>%
  rename(counting = "n()")

dist_cont
```

```
## # A tibble: 5 x 2
##   continent counting
##   <fct>        <int>
## 1 Africa          52
## 2 Americas        25
## 3 Asia            33
## 4 Europe          30
## 5 Oceania          2
```

Now that we have the distribution of continents, perhaps we can graph this as a bar chart to make it more apparent how they are distributed:

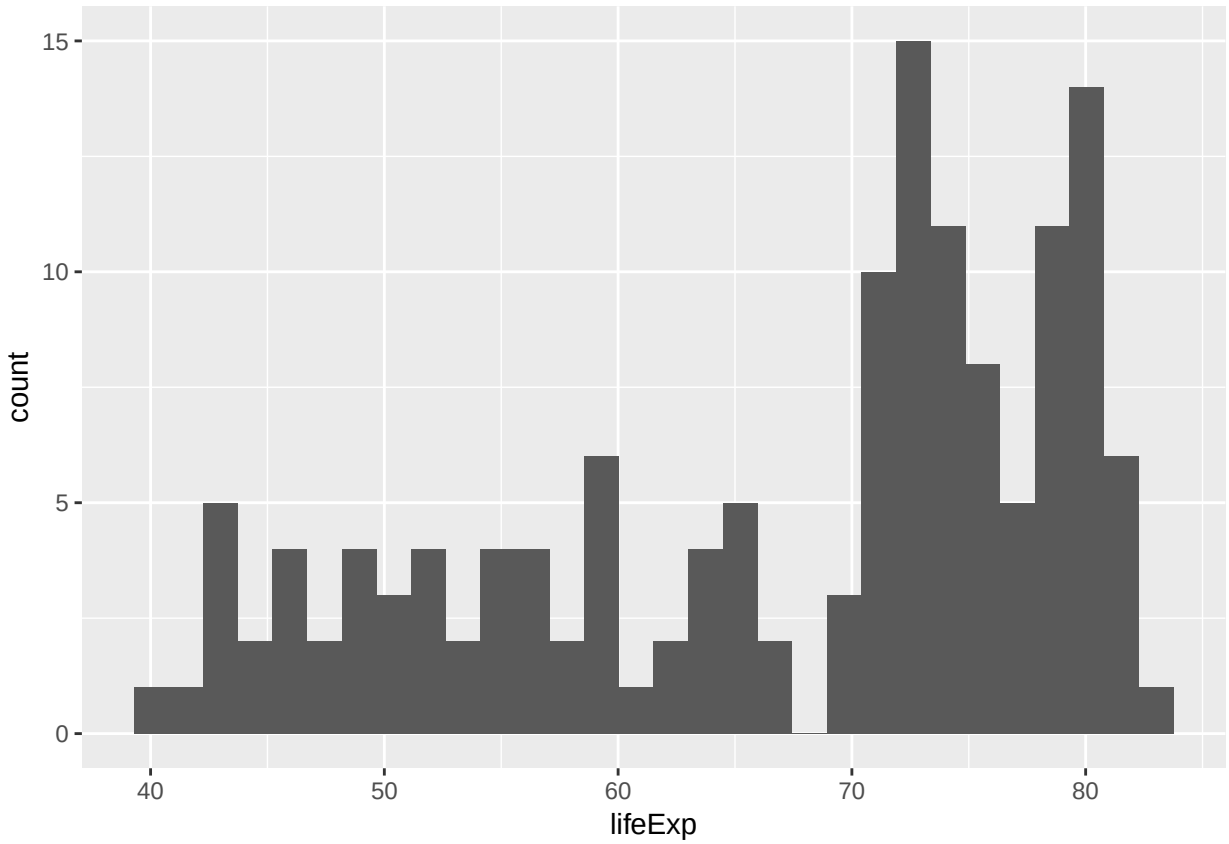
```
dist_cont %>%
  ggplot(aes(continent, counting)) +
  geom_bar(stat="identity")
```



For the distribution of life expectancy, we can use a typical histogram since it is a continuous variable.

```
lifeExp2007 %>%
  select(lifeExp) %>%
  ggplot(aes(lifeExp)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



This shows a "bimodel" type of data. It's fairly uniform for below 60, then it turns somewhat normal afterwards.