hw02_YongzhengParkerLi

Yongzheng Parker Li

9/24/2018

This is the *Homework 02* of the course STAT545A, taught by Vincenzo Coia at the University of British Columbia (UBC). The detailed requirements of this assignment could be found here. The STAT545A course page is here. My participation repository is here.

Bring rectangular data in

This section installs the related packages for this assignment.

```
library(gapminder)
library(tidyverse)
library(gmodels)
```

Smell test the data

This section explores the *gapminder* object.

The first code chunk checks if it is a data frame, and the answer is yes. Both *exists* and *class* function suffice the purpose.

```
exists("gapminder") #check if it is a data frame
```

```
## [1] TRUE
```

```
class(gapminder)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

The following subsection showcases the number of variables/columns (answer:6) and the number of rows/observations (answer: 1704).

```
ncol(gapminder) #check the number of columns
```

```
## [1] 6
```

```
nrow(gapminder) #check the number of rows
```

```
## [1] 1704
```

Head function could also demonstrate the number of columns. It is a better choice when people want to take a quick look at the data. The easiest way, however, is to use the *dim* function, which gives the number of rows and columns at the same time.

```
head(gapminder) #check the number of columns and have a quick look at the data
```

```
## # A tibble: 6 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.
```

```
dim(gapminder) #check the number of columns and rows at the same time
```

```
## [1] 1704    6
```

Str shows the data type of each variable.

```
str(gapminder)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    1704 obs. of  6 variables:
##  $ country  : Factor w/ 142 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ continent: Factor w/ 5 levels "Africa","Americas",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ year     : int  1952 1957 1962 1967 1972 1977 1982 1987 1992 1997 ...
##  $ lifeExp  : num  28.8 30.3 32 34 36.1 ...
##  $ pop      : int  8425333 9240934 10267083 11537966 13079460 14880372 12881816 13867957 16317921 22
##  $ gdpPercap: num  779 821 853 836 740 ...
```

Country and *continent* are *factor*; *year* and *pop* are *integer*; *lifeExp* and *gdpPercap* are *number*.

Explore individual variables

Categorical variable: *continent*

```
summary(gapminder$continent)
```

```
##   Africa Americas     Asia   Europe  Oceania
##      624      300      396      360       24
```

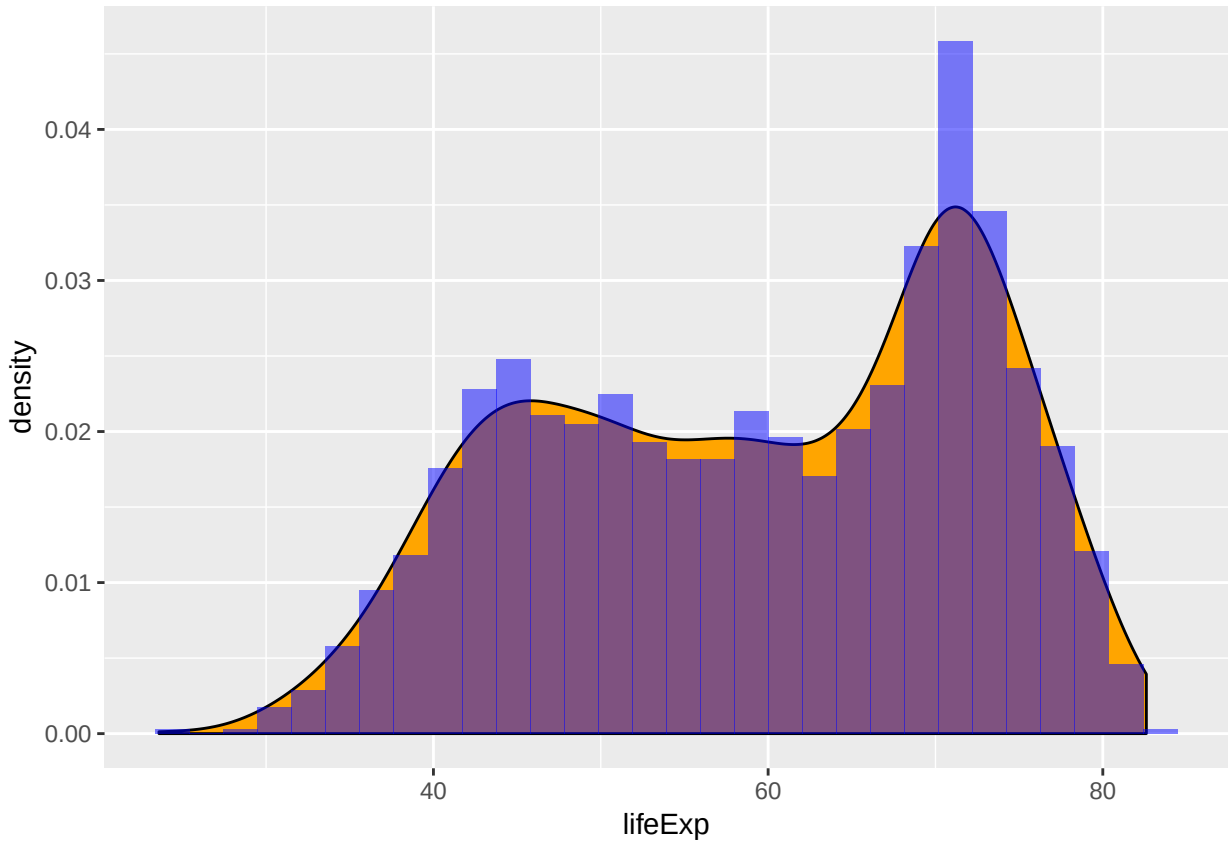
The result shows that there are five continents–Africa, Americas, Asia, Europe, and Oceania–in the dataset. The most common one is Africa (624 observations), while the least common one is Oceania (24 observations).

Quantitative variable: *lifeExp*

```
summary(gapminder$lifeExp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   23.60   48.20   60.71   59.47   70.85   82.60
```

```
ggplot(gapminder,aes(lifeExp)) +
  geom_density(fill="orange") +
  geom_histogram(aes(y=..density..), fill="blue", alpha=0.5)
```

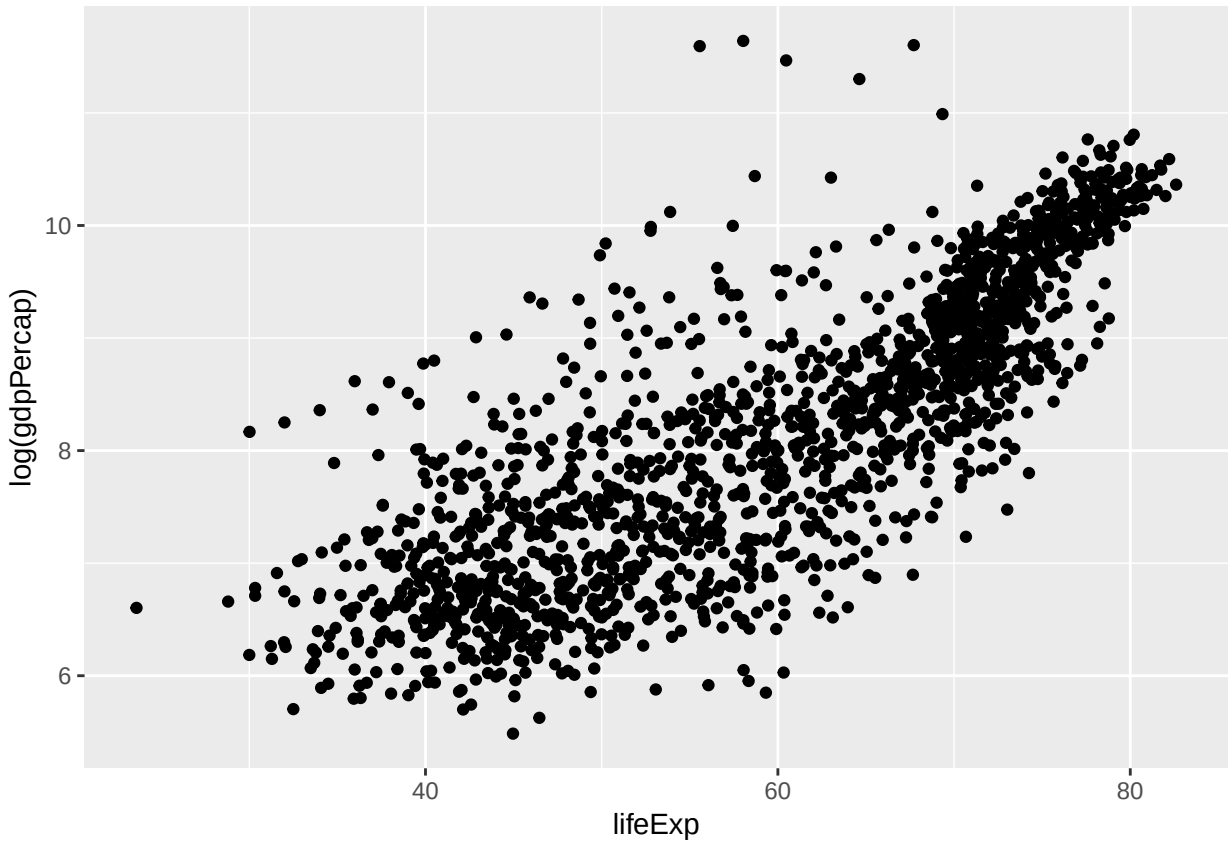


The result shows that, for *lifeExp*, the minimum is 23.60, the maximum is 82.60, the average is 59.47, and some other info. To make it vivid, above also shows the histogram-density-combined graph.

Explore various plot types

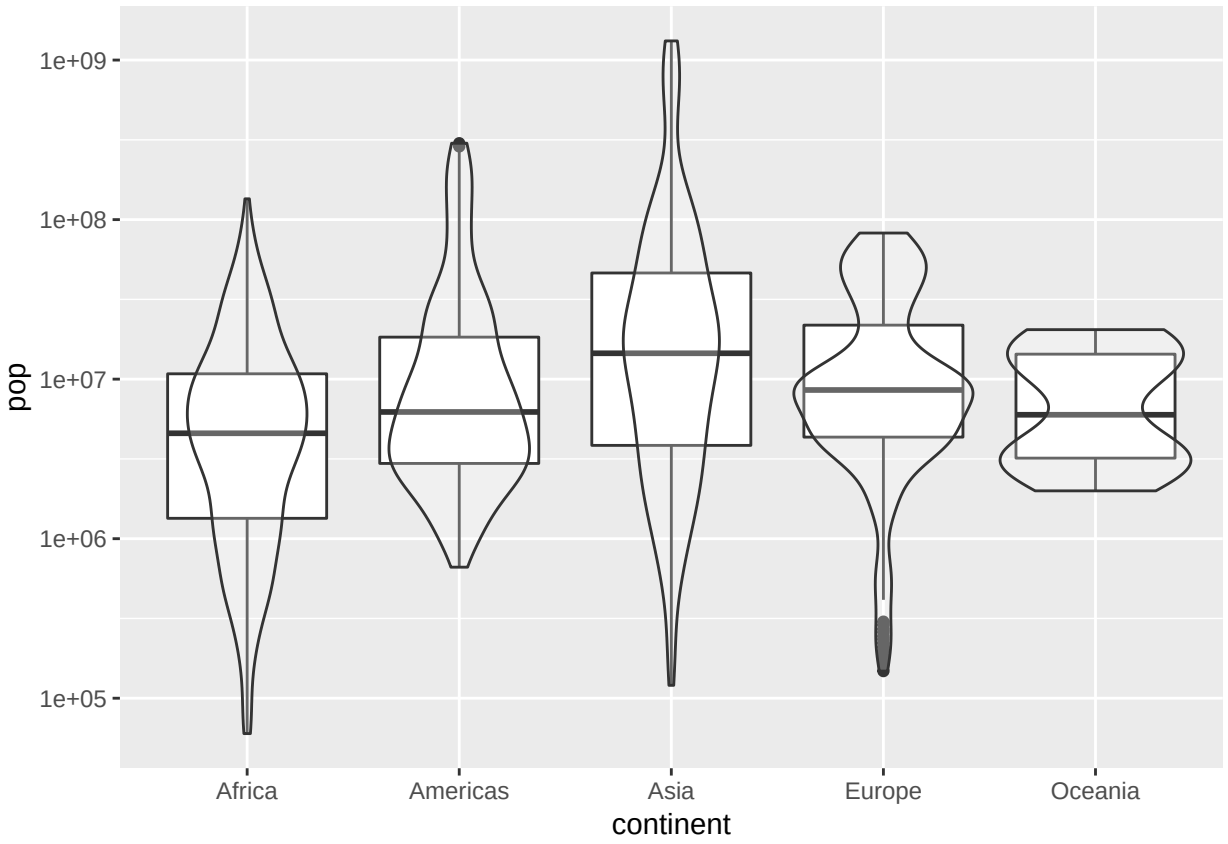
The plot above is a plot of one quantitative variable. This section makes a few more plots for people to have a better understanding of the data distribution.

```
ggplot(gapminder,aes(x=lifeExp, y=log(gdpPercap))) +
  geom_point() #scatter plot
```



The scatterplot above is the scatterplot of *lifeExp* and *gdpPercap* (log). We can also scale the latter variable to log10 value. There is a positive correlation between these two variables.

```
a <- ggplot(gapminder, aes(continent, pop)) +
  scale_y_log10()
a + geom_boxplot() + geom_violin(alpha=0.25)
```

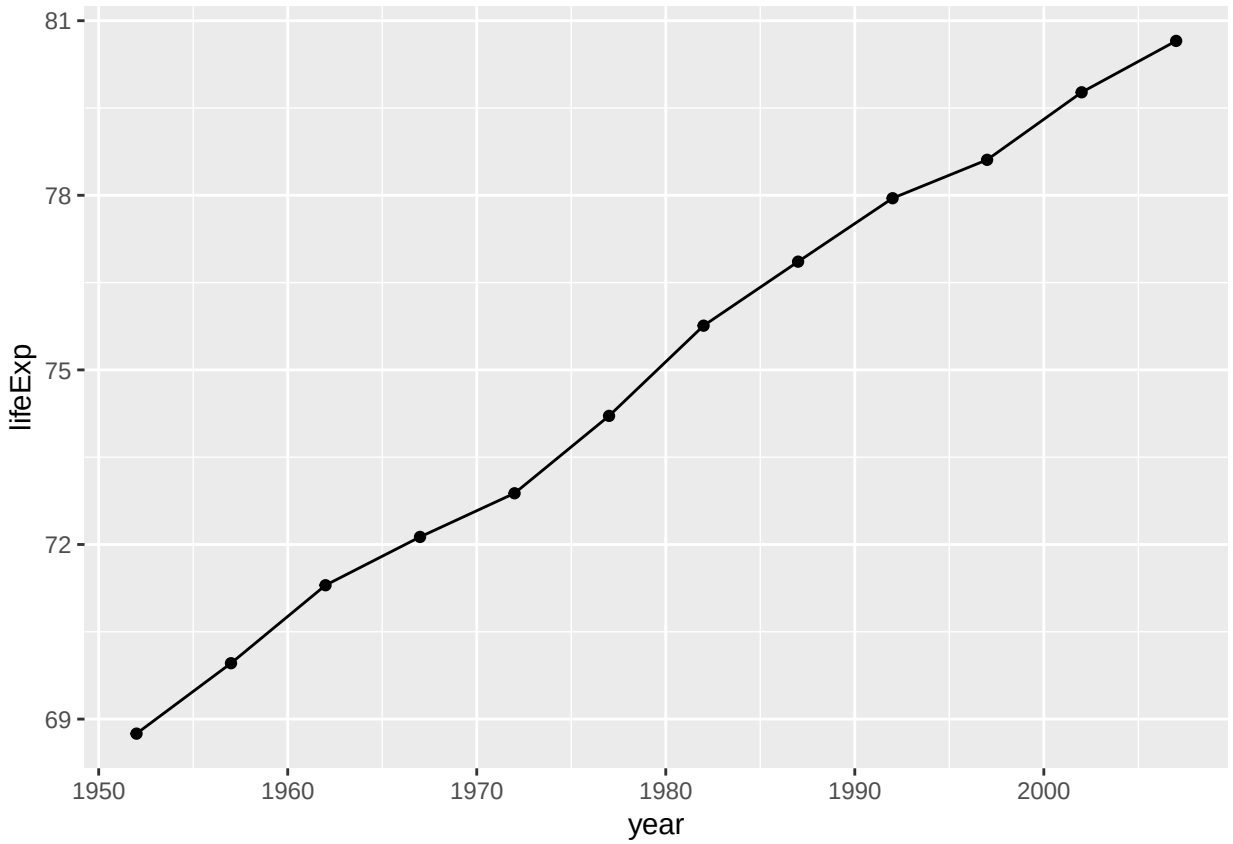


This plot is the box-violin-combined plot between *continent* and *pop* (log10). We can see that the population distribution in each continent is different. For instance, distribution in Aisa is more normal-like.

Use filter(), select(), and %>%

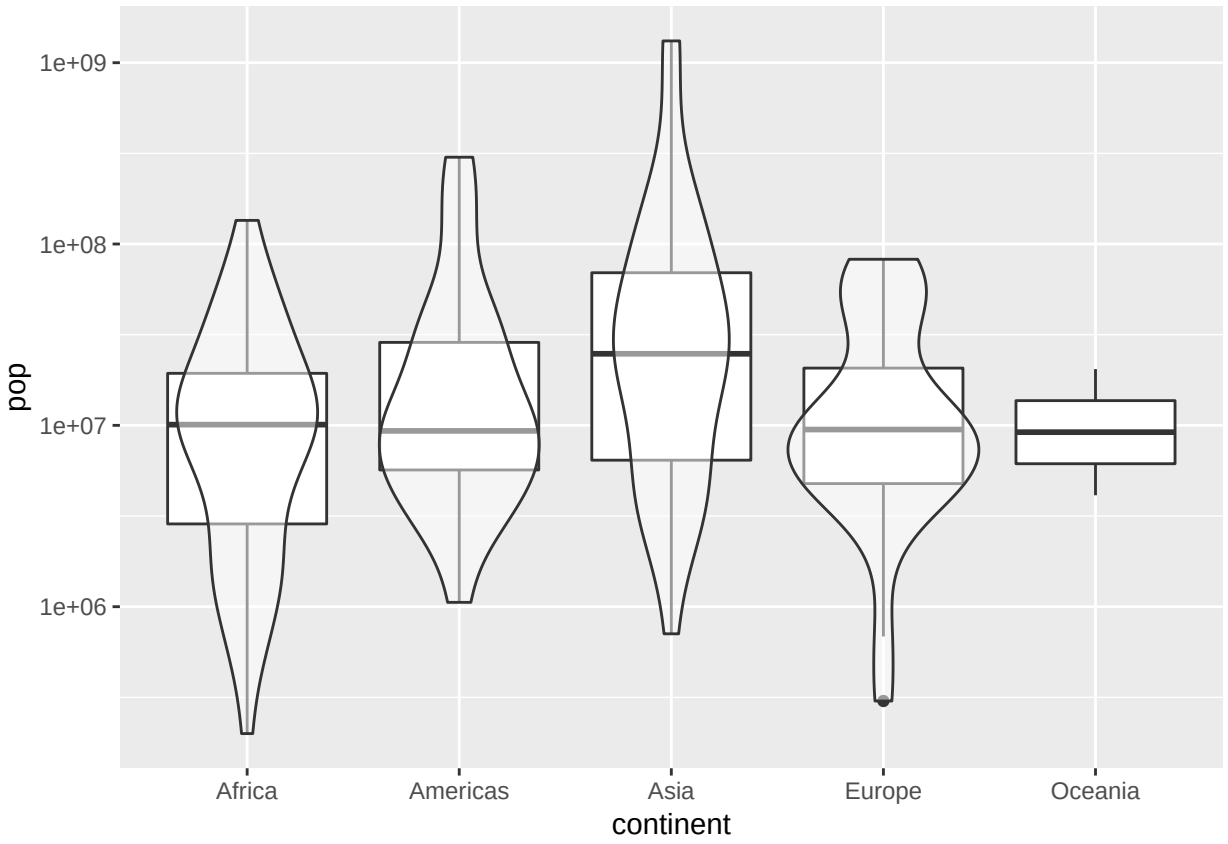
This section utilizes filter() to create data subsets that I want to plot. I also practice piping together filter() and select() into ggplot. Piping syntax could simplify the coding process and make it more reader-friendly.

```
gapminder %>%
  filter(country == "Canada") %>%
  ggplot(aes(year, lifeExp)) +
  geom_line() +
  geom_point()
```



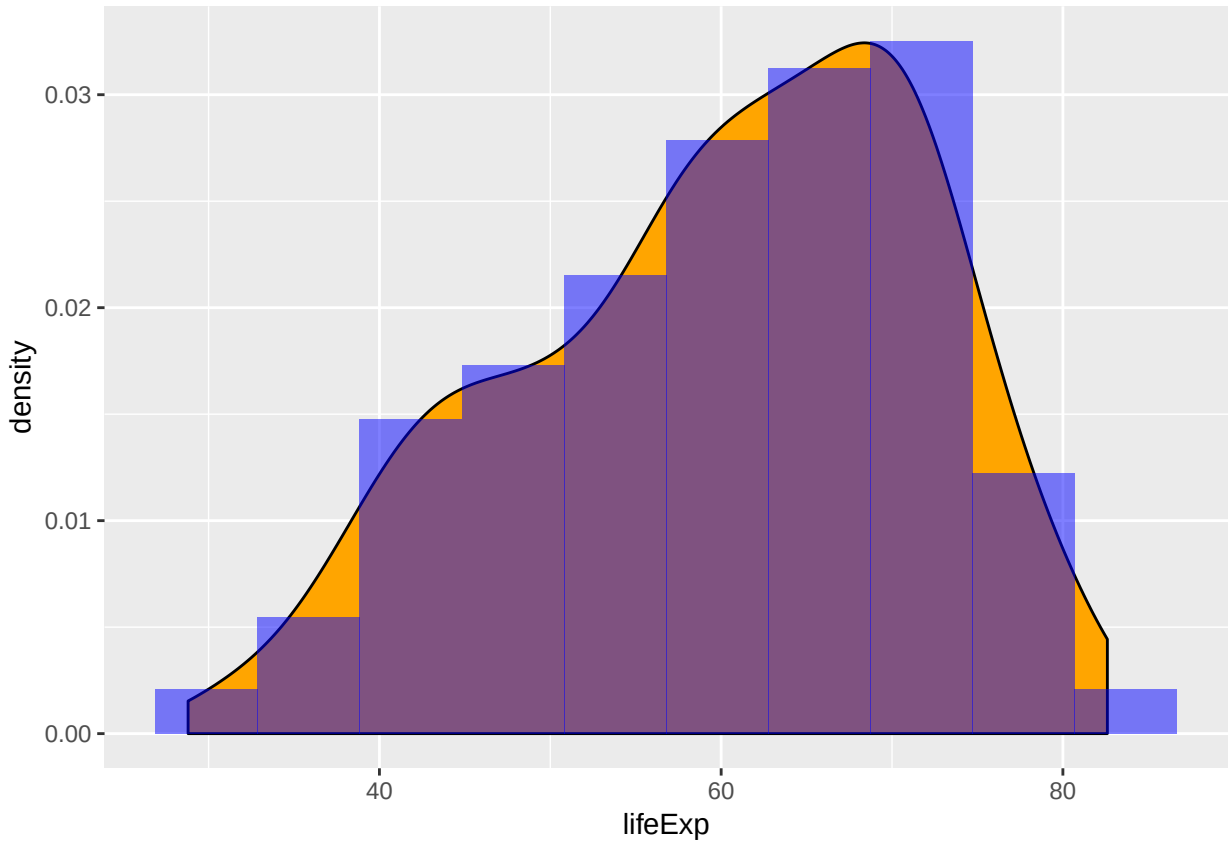
This graph shows the year-lifeExp point-line for Canada. We can tell that the *lifeExp* in Canada increases along the *year*.

```
gapminder %>%
  filter(year == 2007) %>%
  ggplot(aes(continent, pop)) +
  scale_y_log10() +
  geom_boxplot() +
  geom_violin(alpha=0.5)
```



This graph shows the continent-pop box-violin-combined plot in 2007. We can compare this with the overall box-violin graph above, and it seems the difference is small in 2007.

```
gapminder %>%
  filter(continent == "Asia") %>%
  ggplot(aes(lifeExp))+
  geom_density(fill="orange") +
  geom_histogram(aes(y=..density..), fill="blue", alpha=0.5,bins=10)
```



This graph shows the hist-density plot of *lifeExp* in Asia. We can see the average in Aisa is quite high, and it has a relative long left-tail.