R Practice 2: `ggplot2` and Simple Regression

Ryan Safner

October 3, 2018

1. Install and load the package `gapminder`. Type `?gapminder` and hit enter to see a description of the data.

```
# install.packages("gapminder") #uncomment for initial installation
library("gapminder") # load gapminder
?gapminder
```

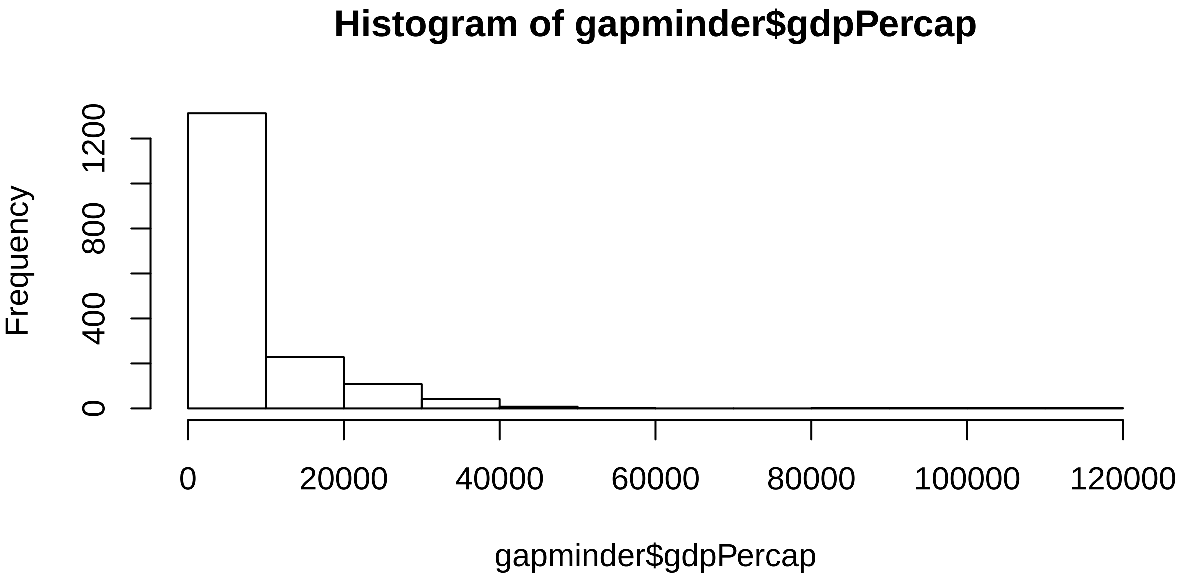
2. Get summary statistics of `gpdPercap`.

```
summary(gapminder$gdpPercap)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    241.2   1202.1   3531.8   7215.3   9325.5 113523.1
```

3. Use base `R`'s `hist()` function to plot a histogram of `gdpPercap`

```
hist(gapminder$gdpPercap)
```

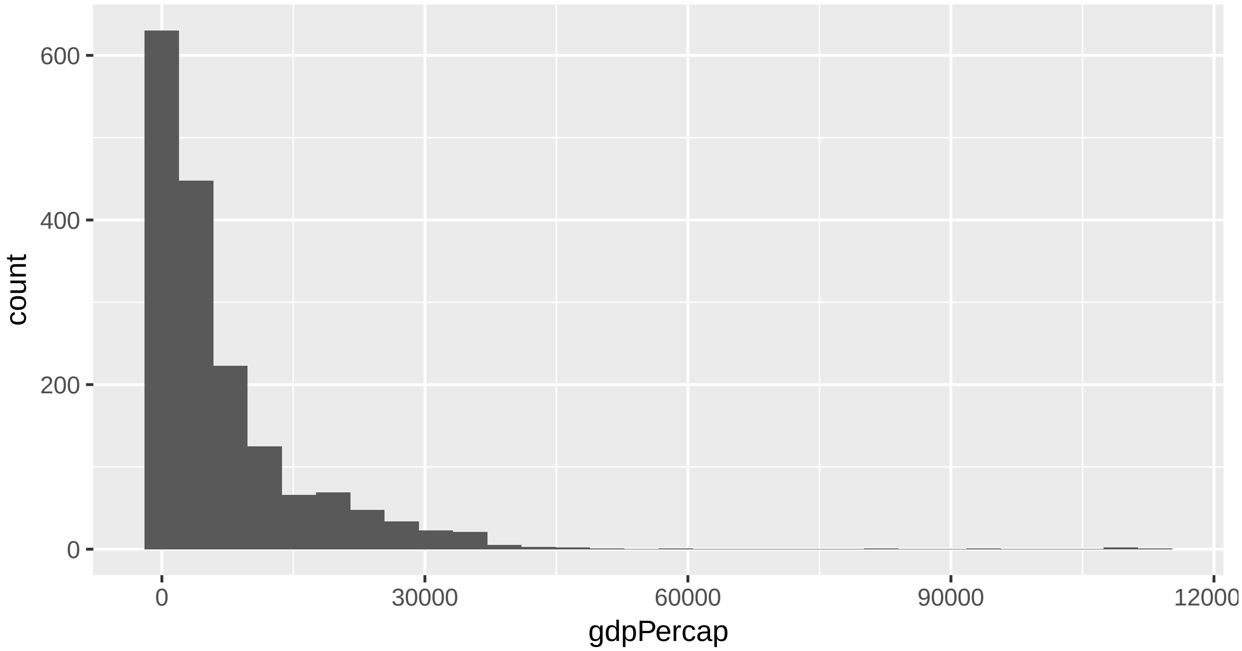


4. Now load and use `ggplot2` to create a histogram of `gdpPercap`. Remember your base layer must establish which `data frame` you are using (`gapminder`) and the base aesthetics `aes()` to define what variable is `x`. Your second layer is a `geom_histogram()`

```
library("ggplot2") # load ggplot2
```

```r
ggplot(gapminder,aes(x=gdpPercap))+
  geom_histogram()
```

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



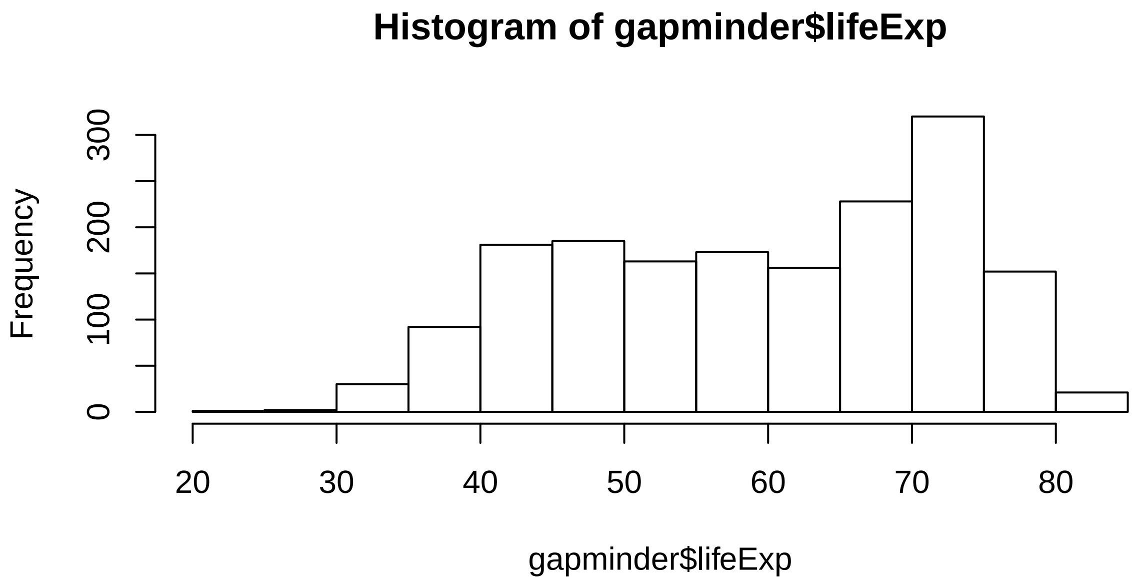
5. Get summary statistics of `lifeExp`.

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

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

6. Use base `R`'s `hist()` function to create a histogram of `lifeExp`.

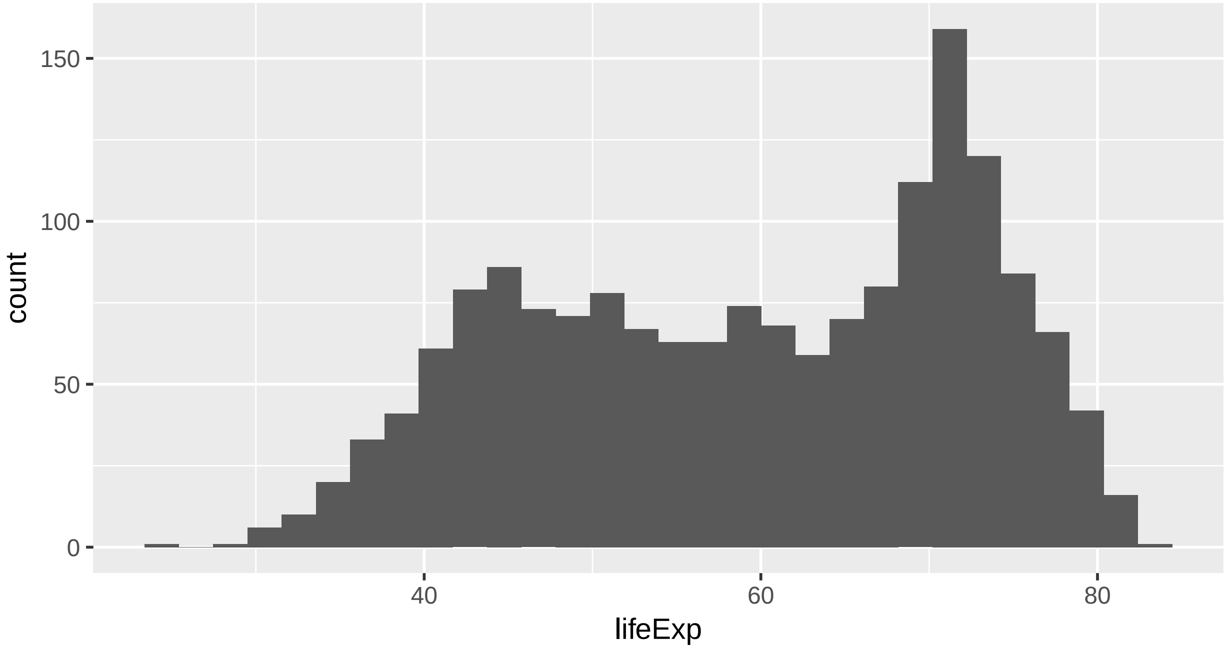
```r
hist(gapminder$lifeExp)
```



7. Use `ggplot2` to create a histogram of `lifeExp`.

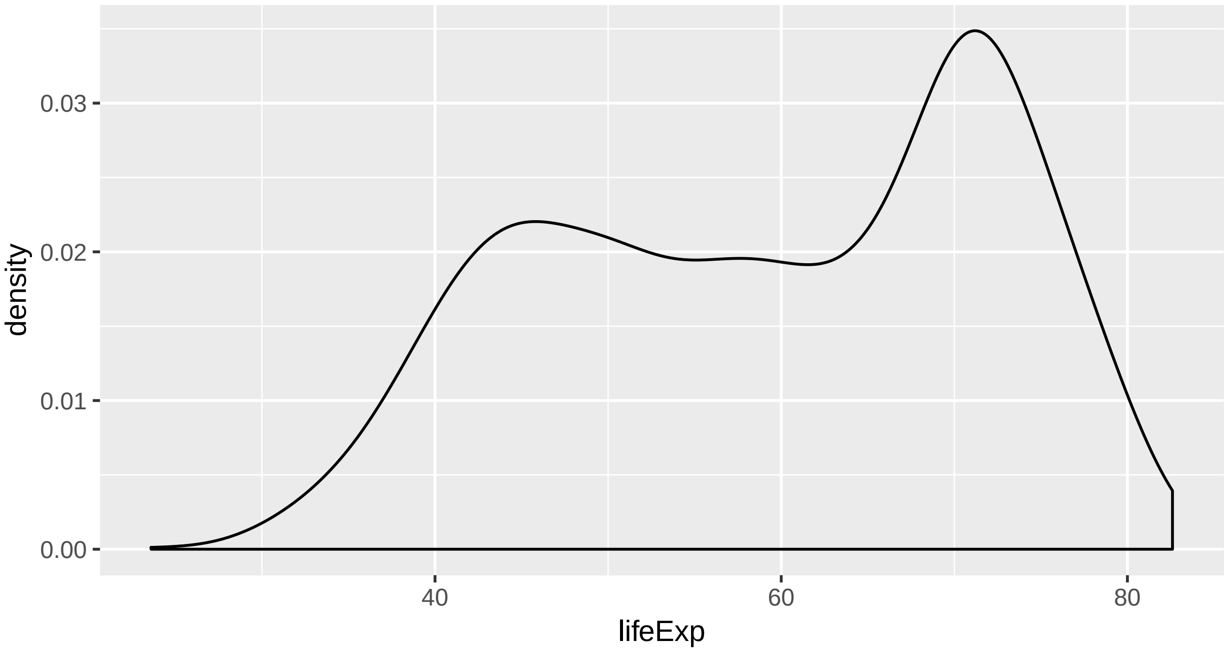
```
ggplot(gapminder,aes(x=lifeExp))+
    geom_histogram()
```

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



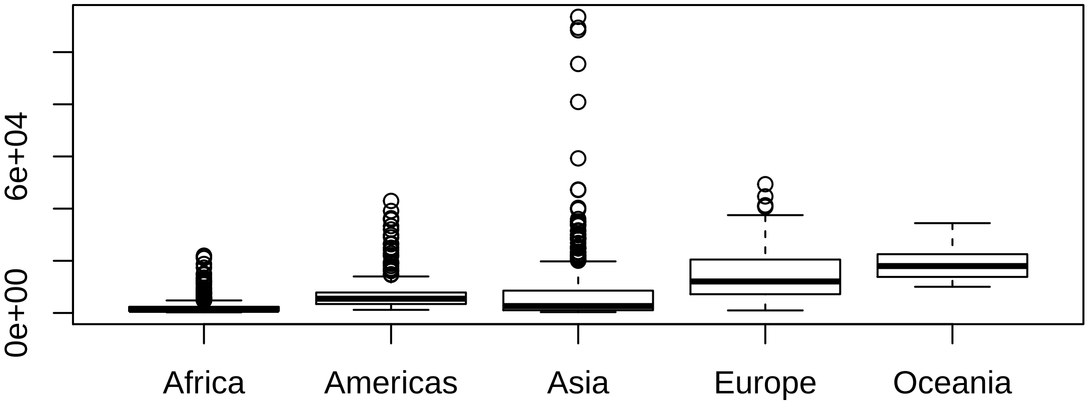
8. Instead of a histogram, make a density plot of `lifeExp` with `geom_density()`

```
ggplot(gapminder, aes(x=lifeExp))+
  geom_density()
```



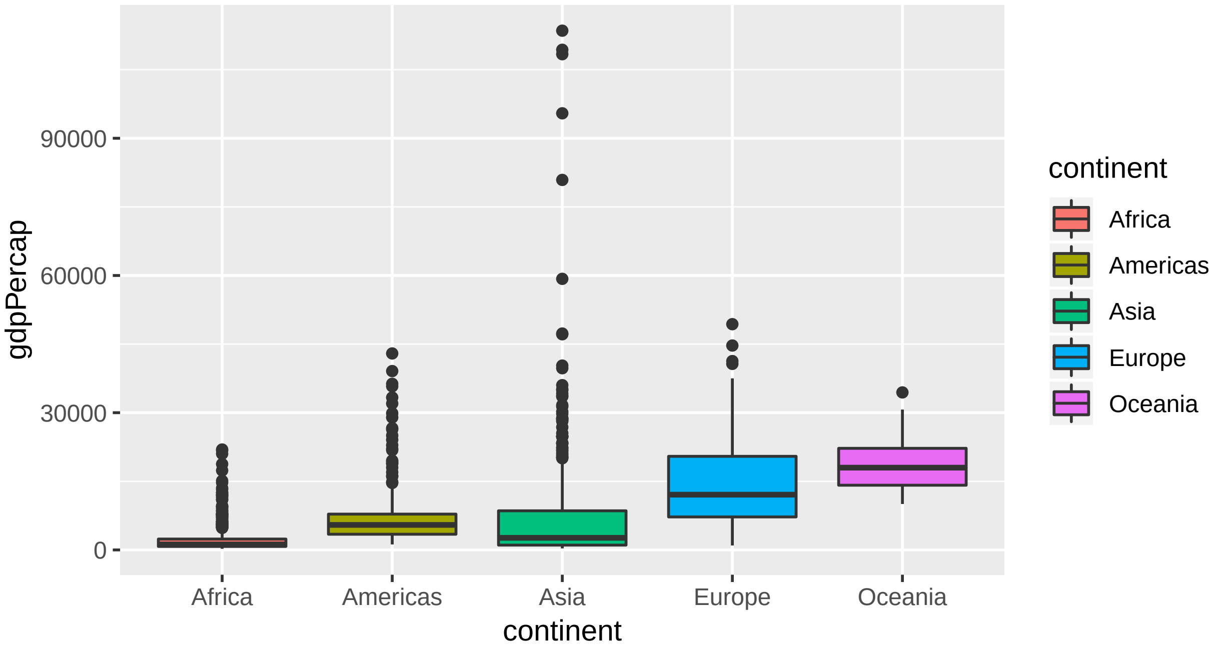
9. Using base R's `boxplot()` function, create a boxplot of `gpdPercap` by `continent`.

```
boxplot(gdpPercap~continent, data=gapminder)
```



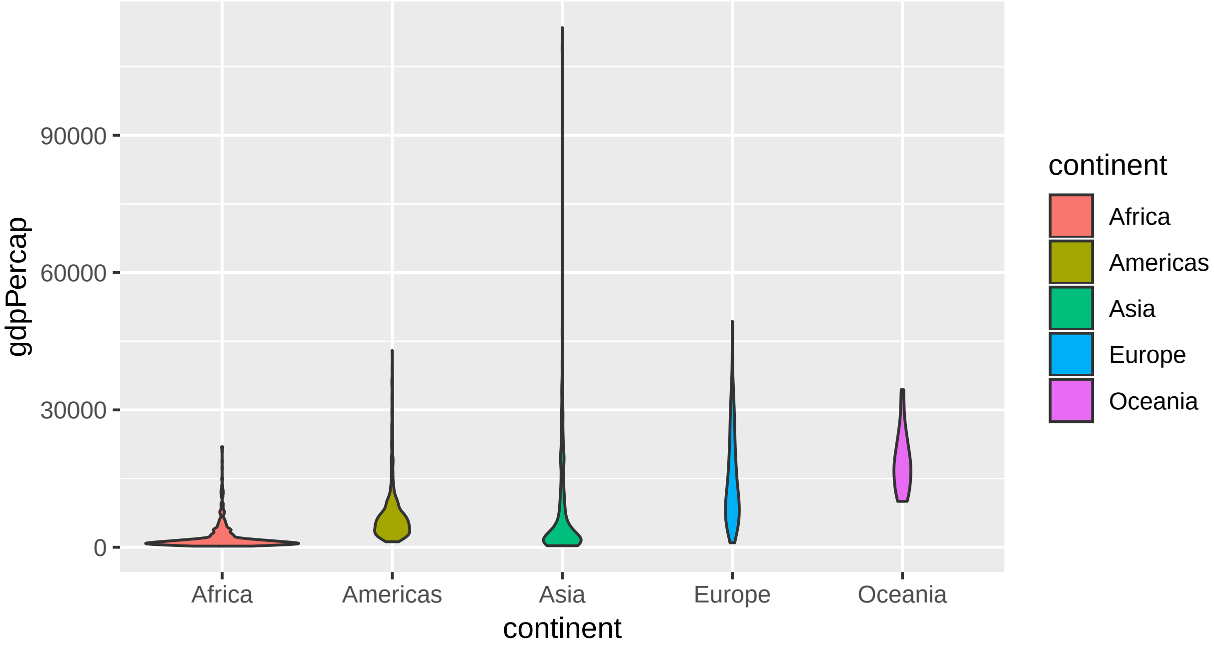
10. Now do the same with `ggplot2`. In your initial aesthetics, set `x` as `continent`, `y` as `gdpPercap` and `fill` (color) by `continent`. Your `geom` layer is `geom_boxplot()`.

```
ggplot(gapminder, aes(x=continent, y=gdpPercap, fill=continent))+
  geom_boxplot()
```



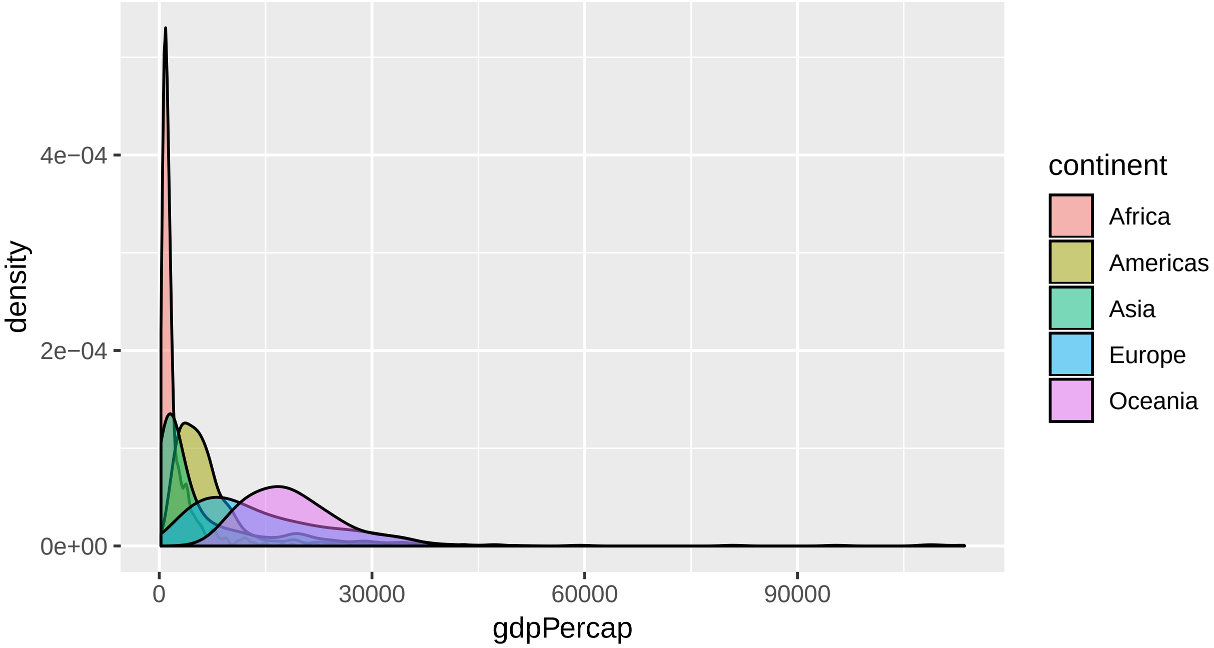
11. The nice thing about building plots one layer at a time is that we can use different `geoms` on the same base layer. Replicate your answer to #10 and instead of `geom_boxplot()`, try a "Violin plot" with `geom_violin().`

```
ggplot(gapminder, aes(x=continent, y=gdpPercap, fill=continent))+
  geom_violin()
```



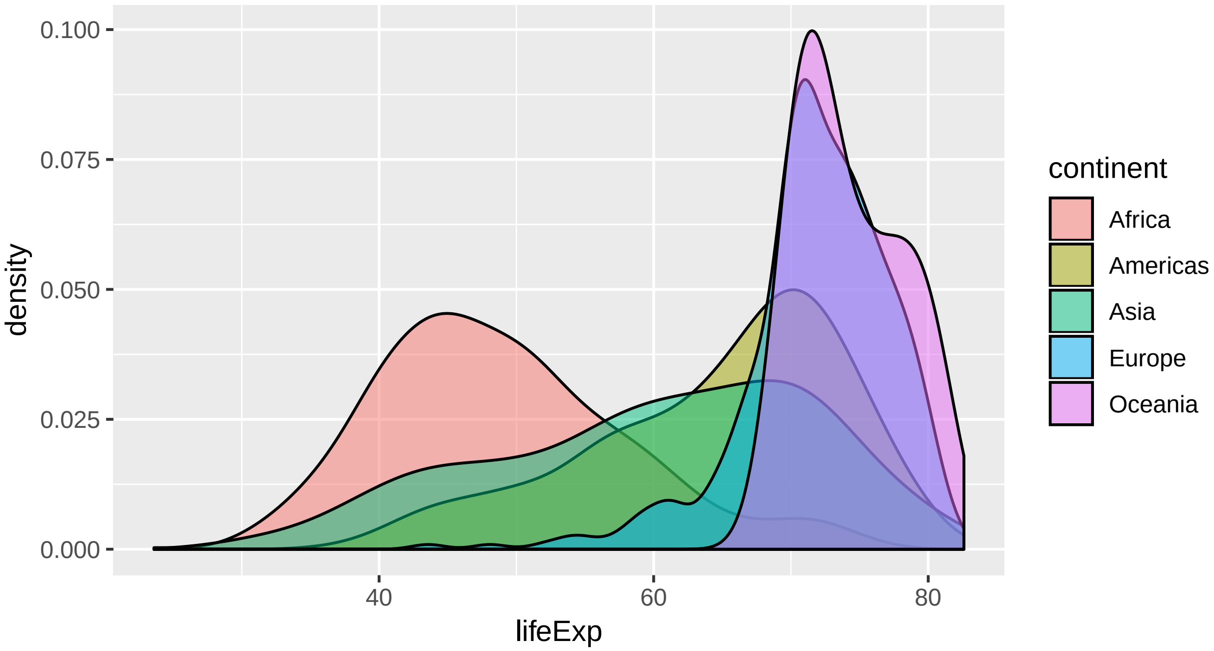
12. Use what you've learned so far to make a density plot of `gdpPercap` by `continent`. Note your only variable here is `x`. Add an option to your `geom_density` layer of setting `alpha=0.5` (to make plots more transparent).

```
ggplot(gapminder, aes(x=gdpPercap, fill=continent))+
  geom_density(alpha=0.5)
```



13. Do the same thing for `lifeExp`

```
ggplot(gapminder, aes(x=lifeExp, fill=continent))+
  geom_density(alpha=0.5)
```

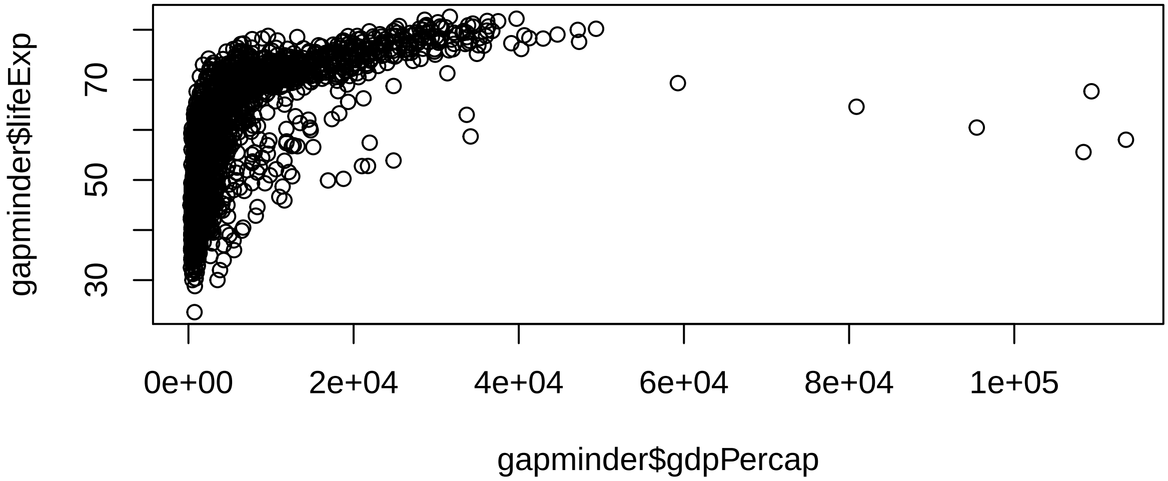


14. Now let's try to estimate the following relationship.

$$\widehat{\text{Life Expectancy}} = \beta_0 + \beta_1 \text{GDP Per Capita}$$

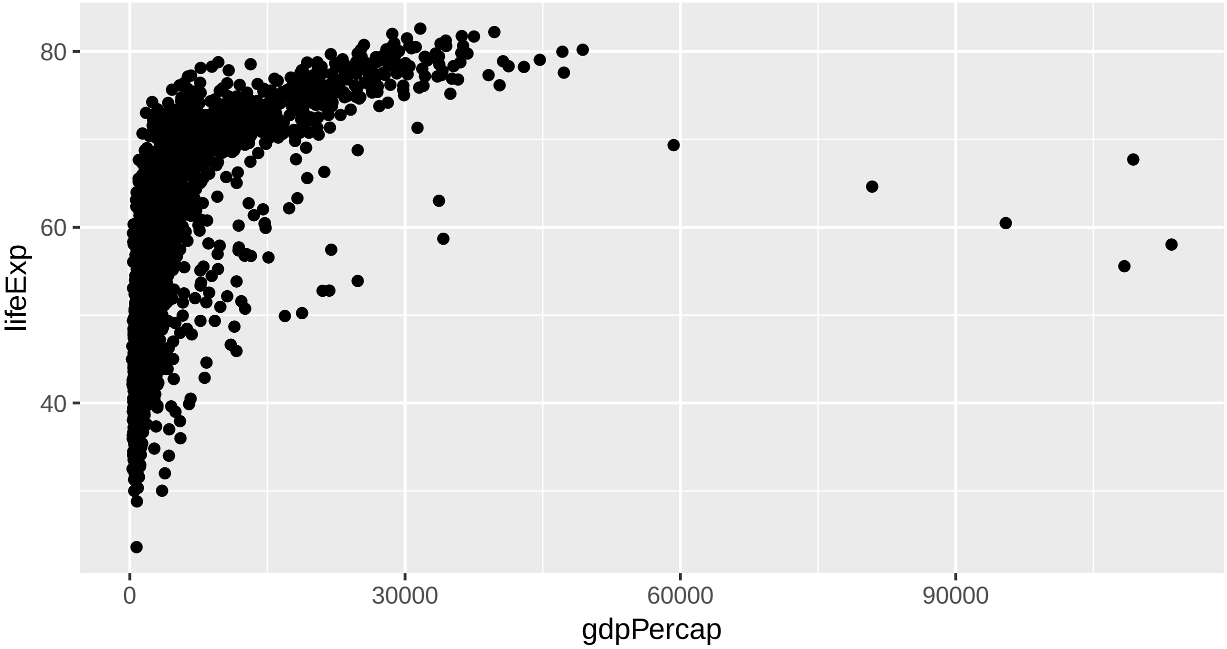
First, use base R to make a scatterplot of these two variables with `plot()`. Be sure to signify `x` and `y` using the `data.frame$variable` syntax.

```
plot(gapminder$gdpPercap,gapminder$lifeExp)
```



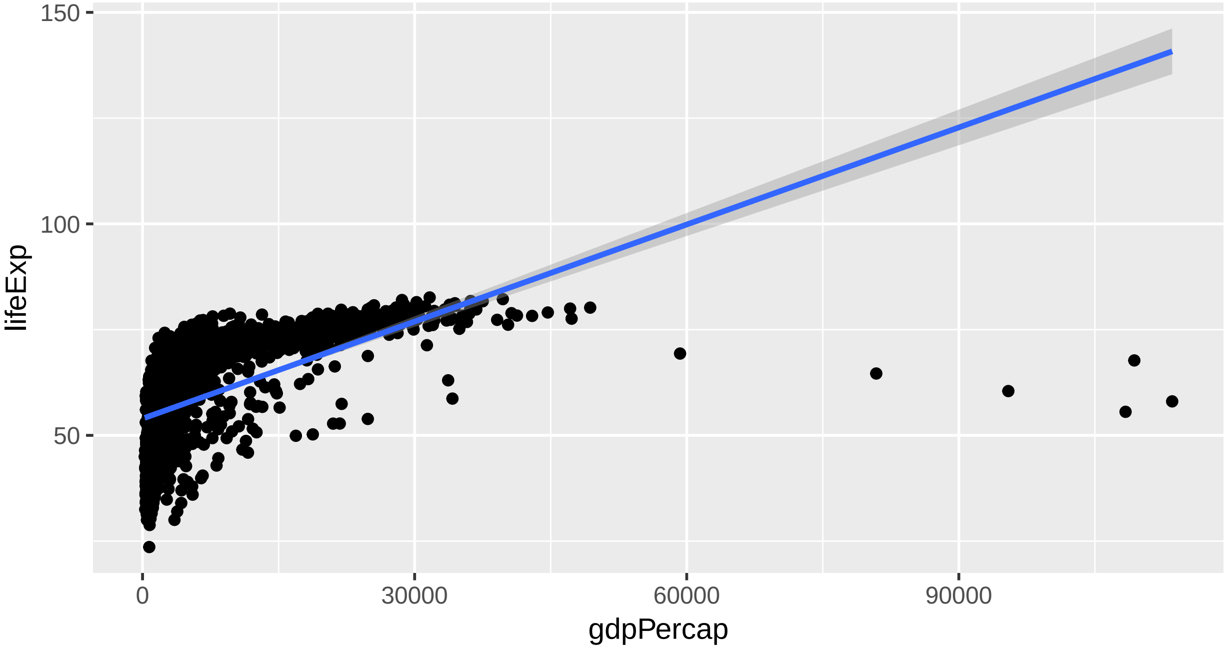
15. Now let's try with ggplot2. For your base layer, consider in your aesthetics what is x and what is y. We want our data to manifest as data points, so use `geom_point()` as your second layer. Be sure to save this as some object.

```
# my object is called p
p<-ggplot(gapminder,aes(x=gdpPercap,y=lifeExp))+
  geom_point()
p
```



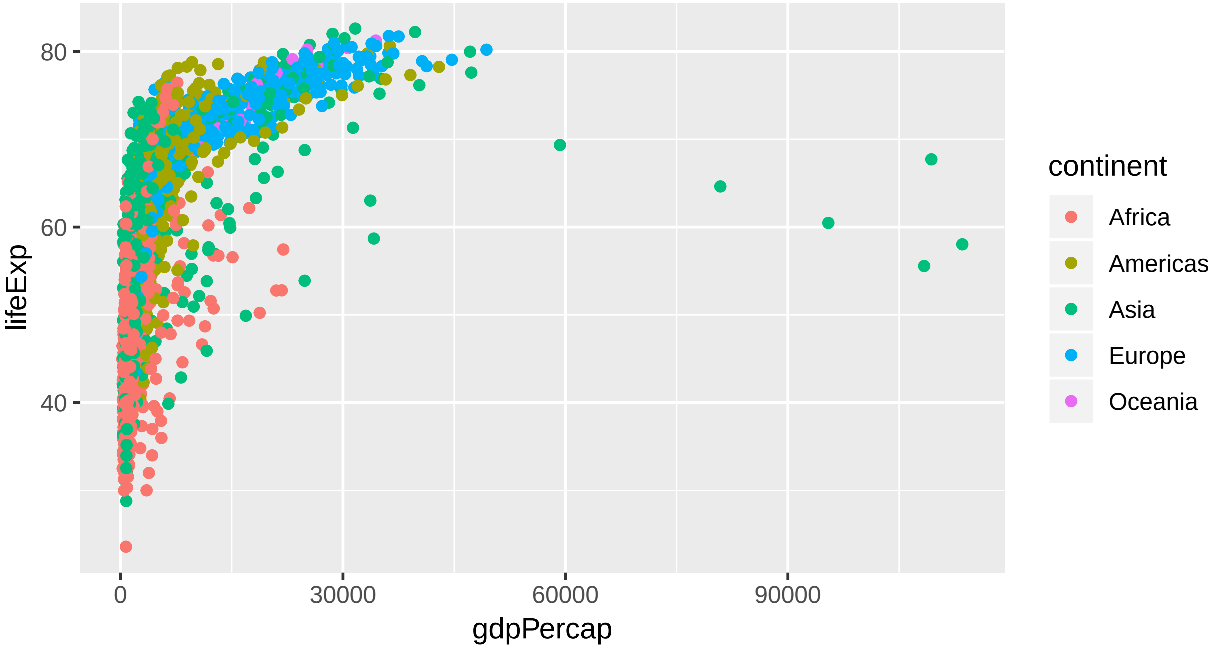
16. Now on top of the existing plot, let's add a regression line. Redefine your object to be itself `+geom_smooth(method="lm")` to add the regression line (`geom_smooth` creates a smooth line, and `lm` stands for linear model, i.e. OLS regression).

```
p<-p+geom_smooth(method="lm")
p
```



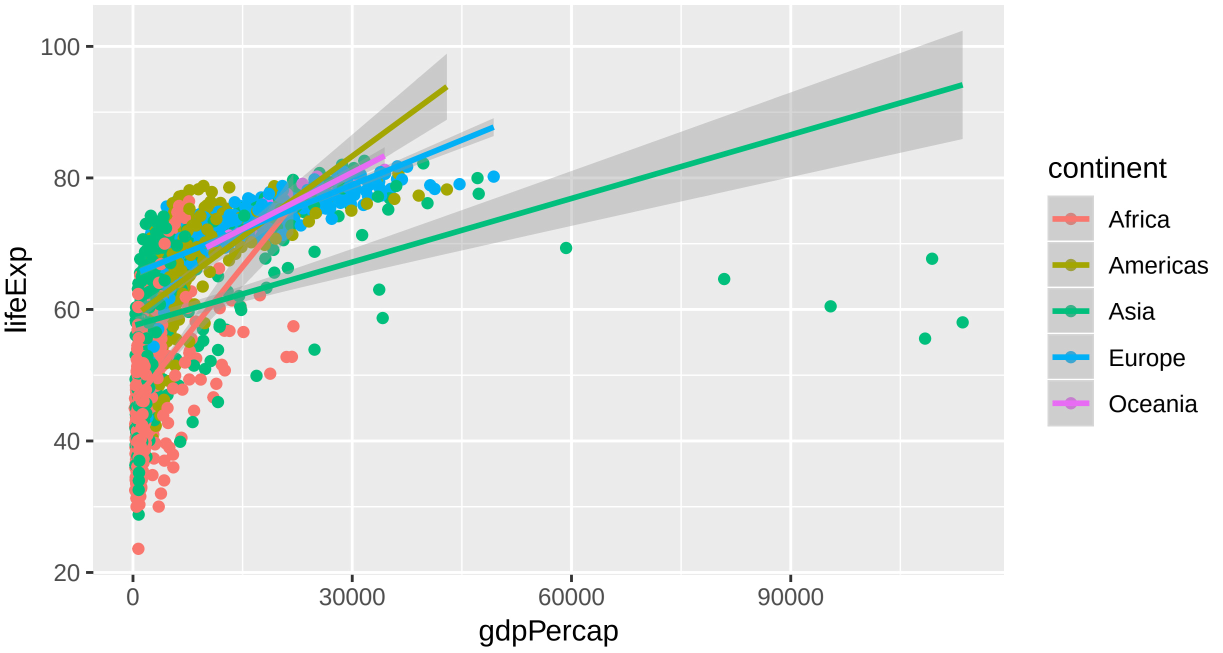
17. Now let's spice this up a bit. Recreate your plot but this time, include in your base layer's aesthetics (in addition to defining x **and** y**)** color=continent **to color by continent.**

```
p<-ggplot(gapminder,aes(x=gdpPercap,y=lifeExp,color=continent))+
  geom_point()
p
```



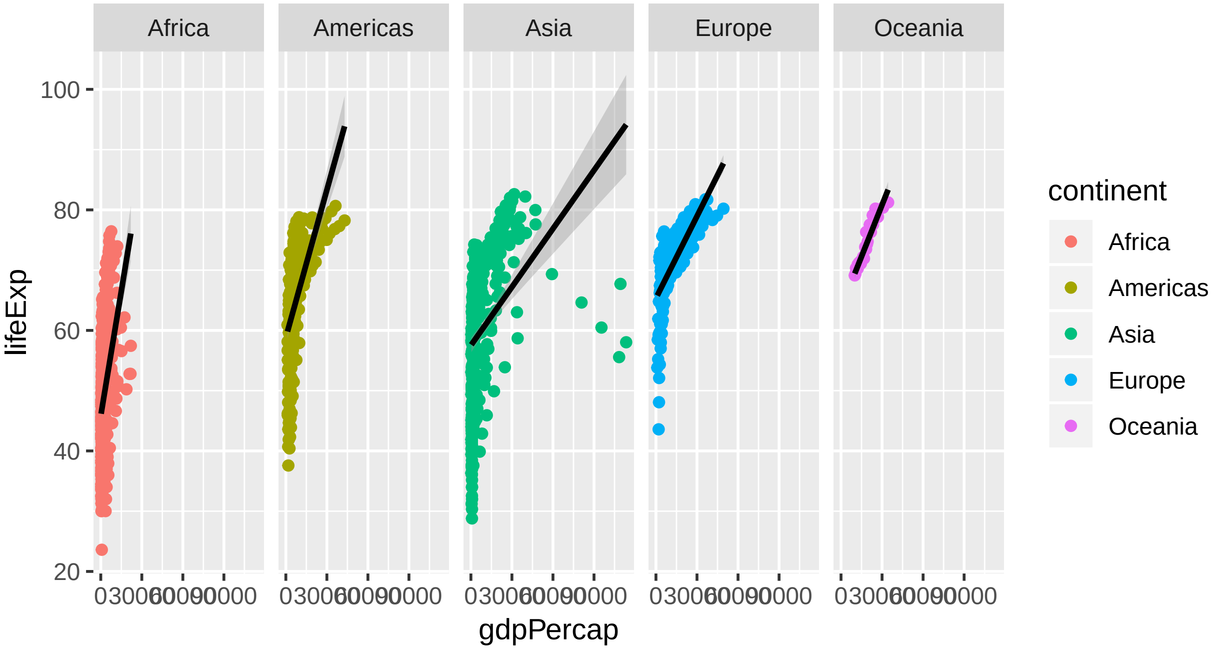
18. Now add a regression line. Notice that since we initially defined in the base layer to color by continent, it also creates different colored lines, one for each continent.

```
p<-ggplot(gapminder,aes(x=gdpPercap,y=lifeExp,color=continent))+
  geom_point()+geom_smooth(method="lm")
p
```



19. Let's try facetting. Add to your previous plot `+facet_grid(cols=vars(continent))`. This creates a grid of individual plots, one for each continent, and arranges them into columns (`cols`) by the variable `continent`.

```
p<-ggplot(gapminder,aes(x=gdpPercap,y=lifeExp,color=continent))+
  geom_point()+geom_smooth(method="lm",color="black")+facet_grid(cols=vars(continent))
p
```



20. Let's try only looking at the year 2002. We can use the `subset()` function to create another data frame for only the year 2000 like `gapminder.2002<-subset(gapminder, year==2002)`. Next, get summary statistics for the gdp per capita in 2002.

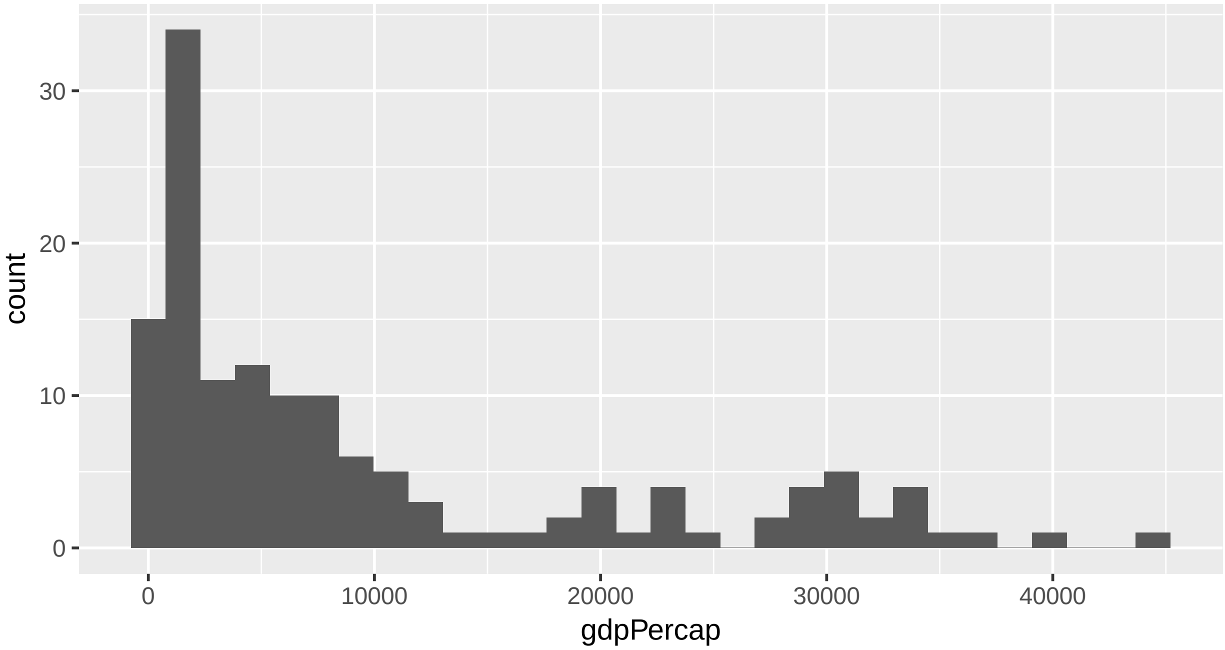
```
gapminder.2002<-subset(gapminder, year==2002)
summary(gapminder.2002$gdpPercap)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   241.2  1409.6  5319.8  9917.9 13359.5 44684.0
```

21. Plot a histogram of gdp per capita in `ggplot2` for 2002

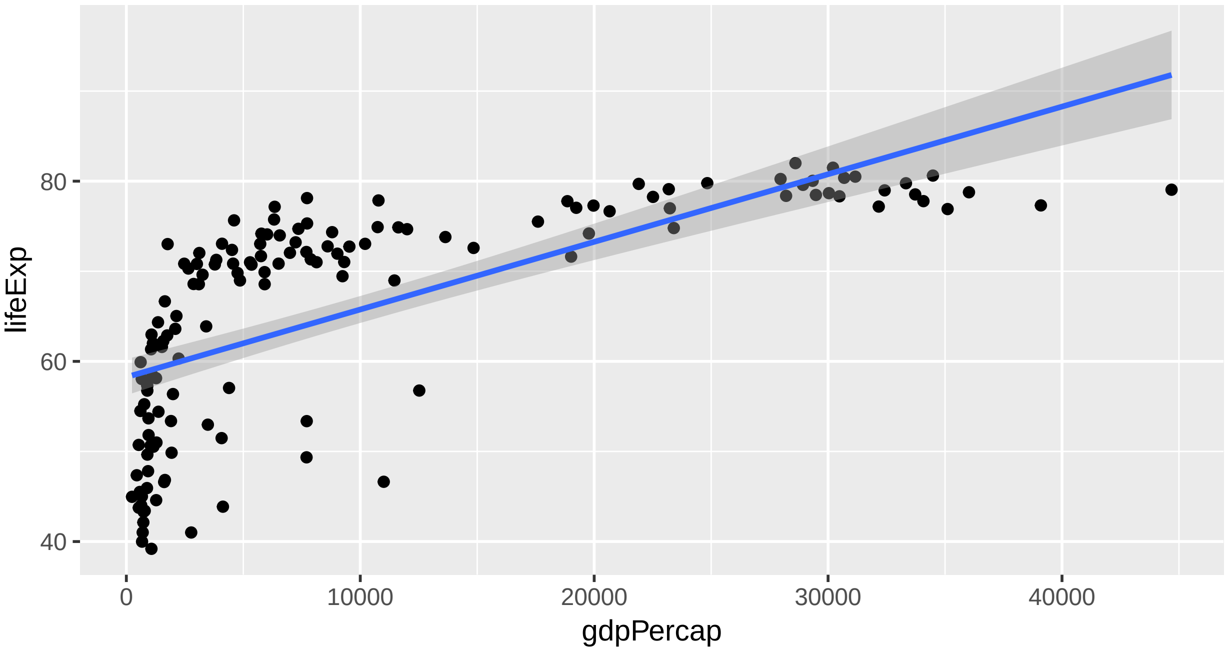
```
ggplot(gapminder.2002,aes(x=gdpPercap))+
  geom_histogram()
```

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



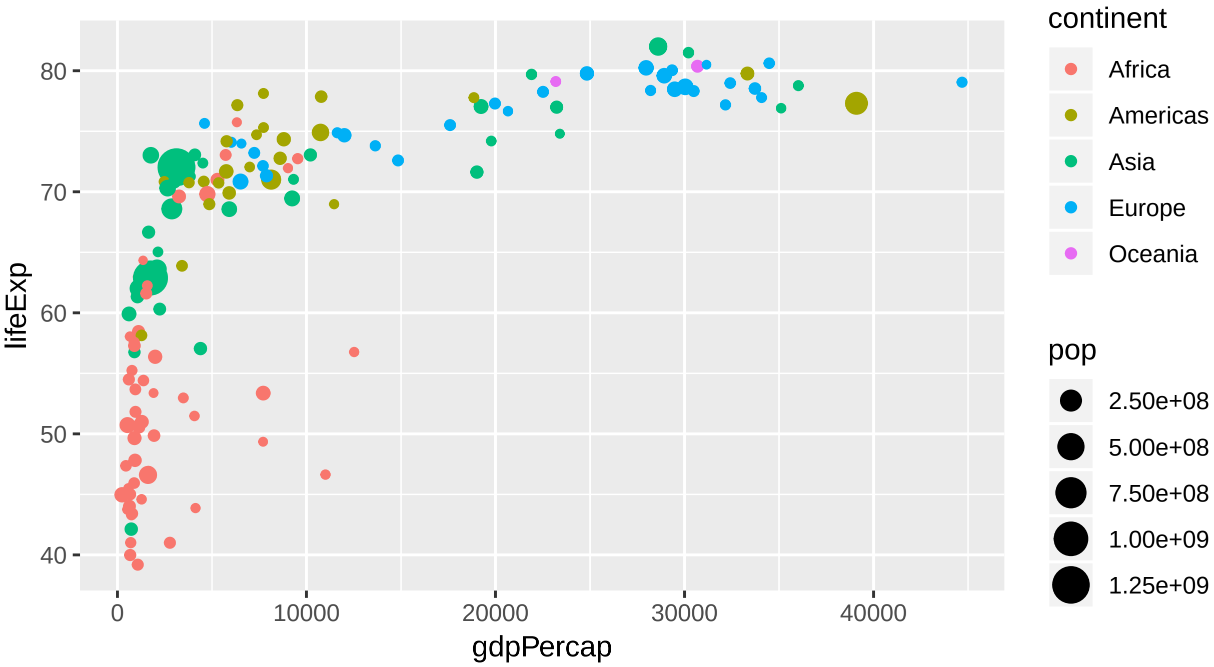
22. Plot a scatterplot with `ggplot2` for 2002 gdp per capita (x) vs. life expectancy (y)

```
ggplot(gapminder.2002,aes(x=gdpPercap,y=lifeExp))+
  geom_point()+geom_smooth(method="lm")
```



23. Now let's add more information to our scatterplot. Add an option to the `geom_point()` to plot `size=pop`.

```
ggplot(gapminder.2002,aes(x=gdpPercap,y=lifeExp, color=continent))+
  geom_point(aes(size=pop))
```



Regression Analysis

24. Now let's turn away from data visualization to more technical analysis with regression. For more information and examples, see lecture 7. Run a regression of life expectancy on gdp per capita. `summary()` your regression. What are:

- $\hat{\beta}_0$
- $\hat{\beta}_1$
- $SE(\hat{\beta}_0)$
- $SE(\hat{\beta}_1)$
- R^2
- SER

```
reg<-lm(lifeExp~gdpPercap,data=gapminder)
summary(reg)
```

```
##
## Call:
## lm(formula = lifeExp ~ gdpPercap, data = gapminder)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -82.754  -7.758   2.176   8.225  18.426
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.396e+01  3.150e-01  171.29   <2e-16 ***
## gdpPercap   7.649e-04  2.579e-05   29.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.49 on 1702 degrees of freedom
## Multiple R-squared:  0.3407, Adjusted R-squared:  0.3403
## F-statistic: 879.6 on 1 and 1702 DF,  p-value: < 2.2e-16
```

- $\hat{\beta}_0 = 53.96$
- $\hat{\beta}_1 = 0.0007649$ (for every 1 \$ increase in GDP, life expectancy increases by 0.0007649 years)
- $SE(\hat{\beta}_0) = 0.315$
- $SE(\hat{\beta}_1) = 0.00002579$
- $R^2 = 0.3407$ (our model explains 34% of the total variation in Life Expectancy)
- $SER = 10.49$ (the average prediction is off by 10.49 years of Life Expectancy)

25. Is $\hat{\beta}_1$ statistically significantly different from 0 (i.e. $H_0 : \beta_1 = 0, H_1 : \beta_1 \neq 0$)? How do you know? See lecture 8 for more help.

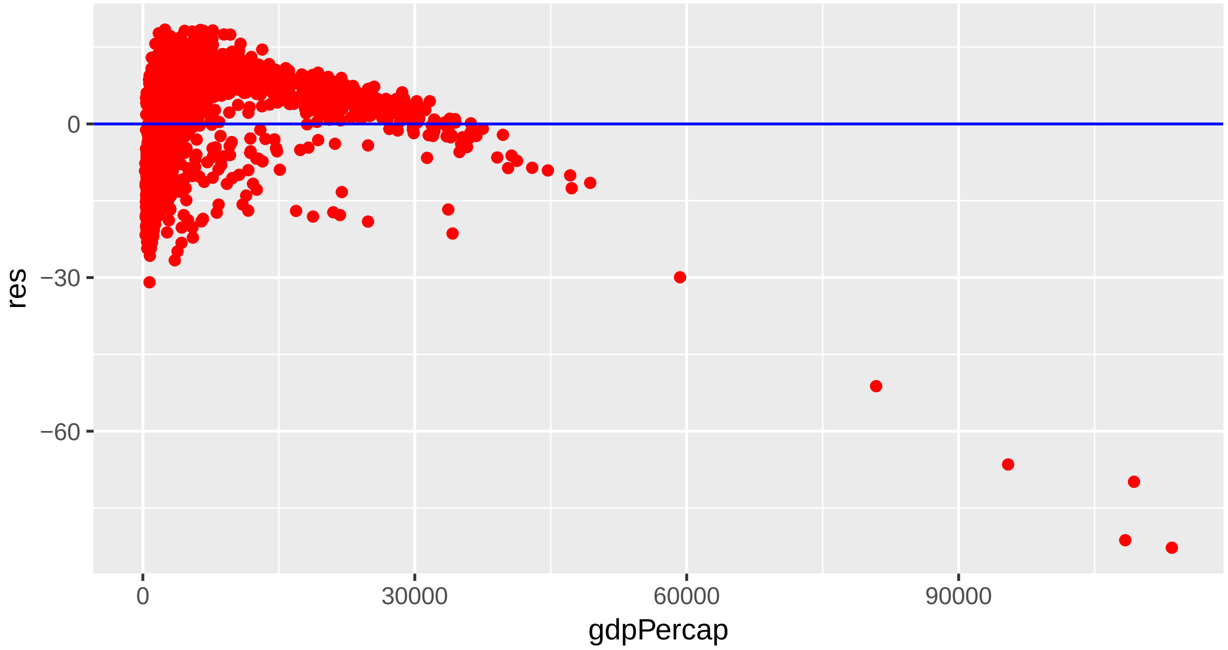
Yes. The t-statistic is 29.66, which is very large, and the p-value is 0.000000000000002, which is very small.

We also see that the estimated slope is more than twice its' standard error: $0.0007649 > 2(0.00002579)$

26. Save the residuals and plot them in a residual plot (using the residuals as y instead of `lifeExp`). Add a horizontal line at 0 with `geom_vline(yintercept=0)`

```
gapminder$res<-residuals(reg)

ggplot(gapminder,aes(x=gdpPercap,y=res))+
  geom_point(color="red")+
  geom_hline(yintercept=0,color="blue")
```



27. Install and then load stargazer to output your regression into a table. For simplicity, set type=text for now. Verify where everything is that you found for question #24.

```
#install.packages("stargazer") # install for first time
library("stargazer")
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
stargazer(reg,type="text")
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                               lifeExp
## -----------------------------------------------
## gdpPercap                    0.001***
##                             (0.00003)
##
## Constant                    53.956***
##                              (0.315)
##
## -----------------------------------------------
## Observations                  1,704
## R2                            0.341
## Adjusted R2                   0.340
## Residual Std. Error     10.491 (df = 1702)
## F Statistic          879.577*** (df = 1; 1702)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```