Code ▾

# Introduction to Data Exploration (4and5 chapter)

Bibek Sapkota

# Data Wrangling

Task 1: Load the library

Hide

```
library(tidyverse)
```

Task 2:Storing all data in a tidy format.

Hide

```
SeasonalTemps <- data.frame(Year = c(2015, 2016, 2017, 2018),
           Winter = c(40, 38, 42, 44),
           Spring = c(46, 40, 50, 48),
           Summer = c(70, 62, 81, 76),
           Fall = c(52, 46, 54, 56))
SeasonalTemps
```

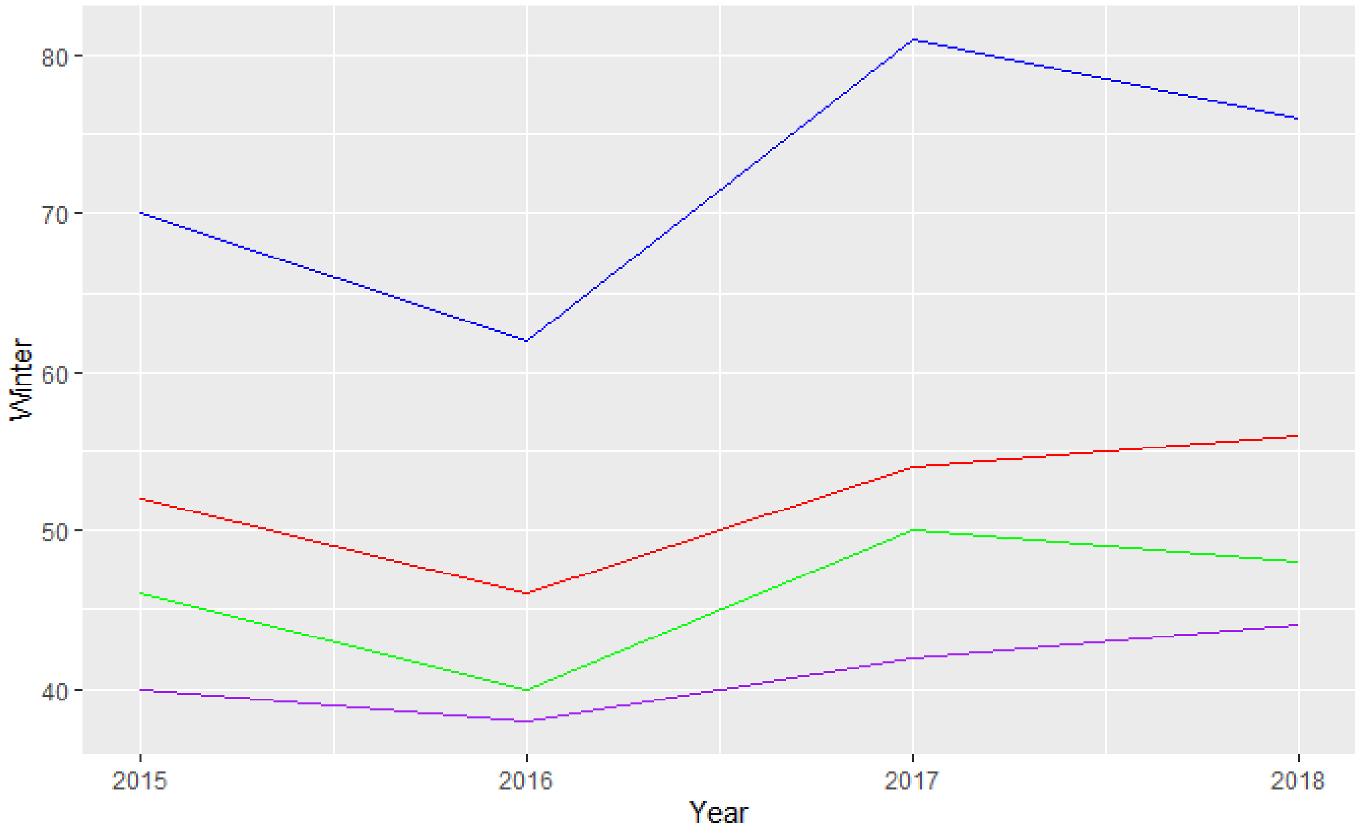
| Year | Winter | Spring | Summer | Fall |
|---:|---:|---:|---:|---:|
| <dbl> | <dbl> | <dbl> | <dbl> | <dbl> |
| 2015 | 40 | 46 | 70 | 52 |
| 2016 | 38 | 40 | 62 | 46 |
| 2017 | 42 | 50 | 81 | 54 |
| 2018 | 44 | 48 | 76 | 56 |

4 rows

Task 3:Ploting the data using ffplot.

Hide

```
ggplot(SeasonalTemps, aes(x = Year)) +
  geom_line(aes(y = Winter), color = "purple") +
  geom_line(aes(y = Spring), color = "green") +
  geom_line(aes(y = Summer), color = "blue") +
  geom_line(aes(y = Fall), color = "red")
```



Task 4:Adding a new columns.

Hide

```
LongTemps <- gather(data = SeasonalTemps, key = Season, value = AvgTemp, -Year)
LongTemps
```

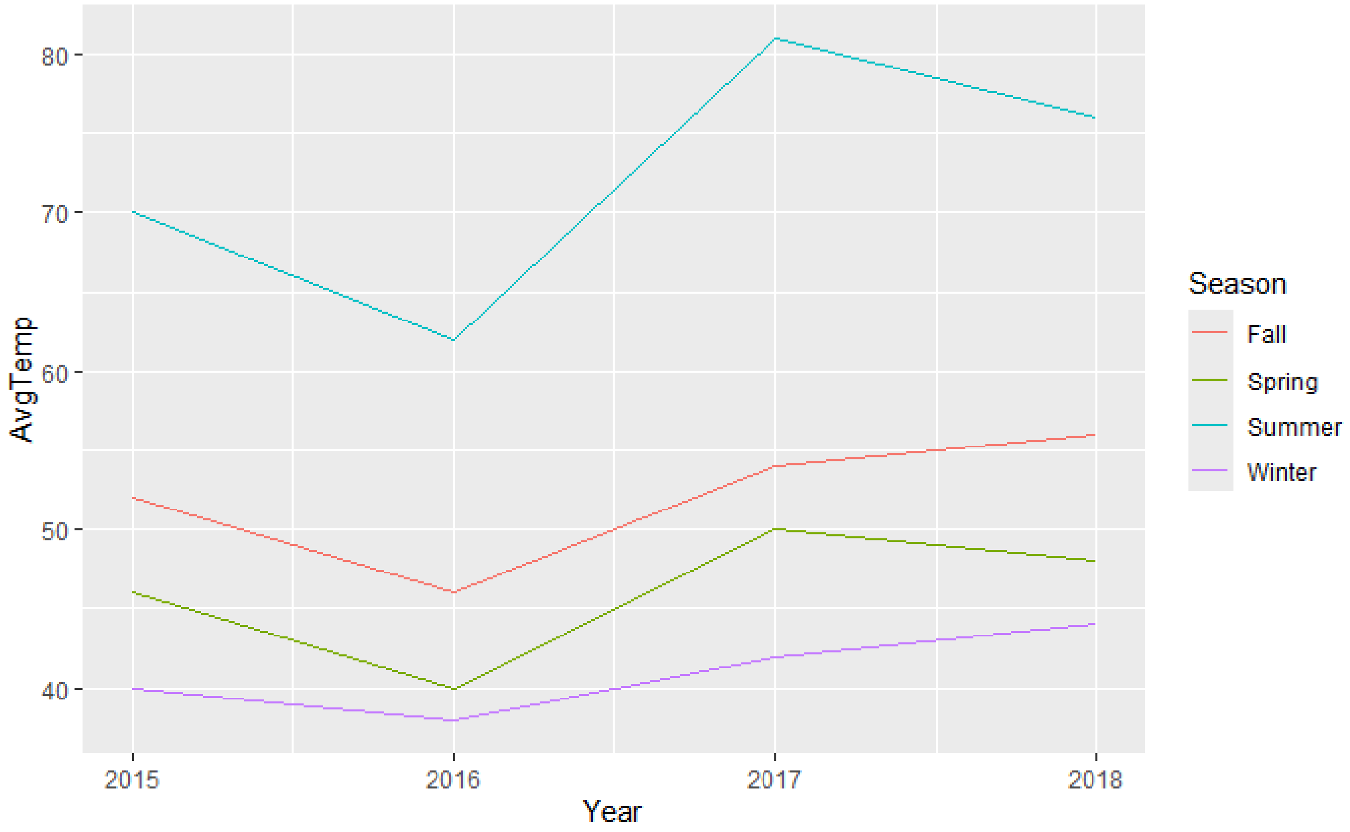
| Year | Season | AvgTemp |
|---:|:---|---:|
| <dbl> | <chr> | <dbl> |
| 2015 | Winter | 40 |
| 2016 | Winter | 38 |
| 2017 | Winter | 42 |
| 2018 | Winter | 44 |
| 2015 | Spring | 46 |
| 2016 | Spring | 40 |
| 2017 | Spring | 50 |
| 2018 | Spring | 48 |
| 2015 | Summer | 70 |
| 2016 | Summer | 62 |

1-10 of 16 rows Previous **1** 2 Next

Task 5:Ploting the LongTemps using ggplot.

Hide

```
ggplot(LongTemps, aes(x = Year, y = AvgTemp, color = Season)) +
  geom_line()
```



Task 6:This code reshapes LongTemps from long to wide format, spreading average temperatures across season columns into WideTemps.

Hide

```
WideTemps <- spread(LongTemps, Season, AvgTemp)
WideTemps
```

| Year<br><dbl> | Fall<br><dbl> | Spring<br><dbl> | Summer<br><dbl> | Winter<br><dbl> |
|---:|---:|---:|---:|---:|
| 2015 | 52 | 46 | 70 | 40 |
| 2016 | 46 | 40 | 62 | 38 |
| 2017 | 54 | 50 | 81 | 42 |
| 2018 | 56 | 48 | 76 | 44 |

4 rows

Task 7:It selects and orders the columns ‘Year’, ‘Winter’, ‘Spring’, ‘Summer’, and ‘Fall’ from the WideTemps data frame into a new data frame OrderWideTemps.

Hide

```
OrderWideTemps <- select(WideTemps, c(Year, Winter, Spring, Summer, Fall))
OrderWideTemps
```

| Year<br><dbl> | Winter<br><dbl> | Spring<br><dbl> | Summer<br><dbl> | Fall<br><dbl> |
|---|---|---|---|---|
| 2015 | 40 | 46 | 70 | 52 |
| 2016 | 38 | 40 | 62 | 46 |
| 2017 | 42 | 50 | 81 | 54 |
| 2018 | 44 | 48 | 76 | 56 |

4 rows

Task 8:It selects and displays the columns “Year,” “Winter,” “Spring,” and “Fall” from the data frame “WideTemps”.

Hide

```
select(WideTemps, c(Year, Winter, Spring, Fall))
```

| Year<br><dbl> | Winter<br><dbl> | Spring<br><dbl> | Fall<br><dbl> |
|---|---|---|---|
| 2015 | 40 | 46 | 52 |
| 2016 | 38 | 40 | 46 |
| 2017 | 42 | 50 | 54 |
| 2018 | 44 | 48 | 56 |

4 rows

Task 9:It selects summer and display oppositely.

Hide

```
select(WideTemps, -Summer)
```

| Year<br><dbl> | Fall<br><dbl> | Spring<br><dbl> | Winter<br><dbl> |
|---|---|---|---|
| 2015 | 52 | 46 | 40 |
| 2016 | 46 | 40 | 38 |
| 2017 | 54 | 50 | 42 |
| 2018 | 56 | 48 | 44 |

4 rows

Task 10:Loading the data in df and Displaying it.

Hide

```
df <- tibble(Horses = c("A", "B", "C"),
             Results = c("1-2-3", "3-1-2", "2-3-1"),
             TotalMinutes = c(3, 3, 3),
             TotalSeconds = c(12, 44, 15))
df
```

| Horses<br><chr> | Results<br><chr> | TotalMinutes<br><dbl> | TotalSeconds<br><dbl> |
|---|---|---:|---:|
| A | 1-2-3 | 3 | 12 |
| B | 3-1-2 | 3 | 44 |
| C | 2-3-1 | 3 | 15 |

3 rows

Task 11:Loading the data in df2 into Displaying it.

Hide

```
df2 <- separate(df, Results, c("FirstRace", "SecondRace", "ThirdRace"), sep = "-")
df2
```

| Horses<br><chr> | FirstRace<br><chr> | SecondRace<br><chr> | ThirdRace<br><chr> | TotalMinutes<br><dbl> | TotalSeconds<br><dbl> |
|---|---|---|---|---:|---:|
| A | 1 | 2 | 3 | 3 | 12 |
| B | 3 | 1 | 2 | 3 | 44 |
| C | 2 | 3 | 1 | 3 | 15 |

3 rows

Task 12:Seperating minutes and seconds by :

Hide

```
unite(df2, TotalTime, TotalMinutes, TotalSeconds, sep = ":")
```

| Horses<br><chr> | FirstRace<br><chr> | SecondRace<br><chr> | ThirdRace<br><chr> | TotalTime<br><chr> |
|---|---|---|---|---|
| A | 1 | 2 | 3 | 3:12 |
| B | 3 | 1 | 2 | 3:44 |
| C | 2 | 3 | 1 | 3:15 |

3 rows

Task 13:Taking the squareroot of a and Displaying it.

Hide

```
a <- 10
a <- a*2
a <- sqrt(a)
a
```

```
[1] 4.472136
```

Hide

```
a <- sqrt(10*2)
a
```

```
[1] 4.472136
```

Task 14:It splits the Results column of df into three columns (FirstRace, SecondRace, ThirdRace) using a hyphen (-) separator, and then combines TotalMinutes and TotalSeconds into a TotalTime column using a colon (:) separator.

Hide

```
unite(
  (separate
   (df,
    Results,
    c("FirstRace", "SecondRace", "ThirdRace"),
    sep = "-")),
  TotalTime,
  TotalMinutes,
  TotalSeconds,
  sep = ":")
```

| **Horses**<br>&lt;chr&gt; | **FirstRace**<br>&lt;chr&gt; | **SecondRace**<br>&lt;chr&gt; | **ThirdRace**<br>&lt;chr&gt; | **TotalTime**<br>&lt;chr&gt; |
|---|---|---|---|---|
| A | 1 | 2 | 3 | 3:12 |
| B | 3 | 1 | 2 | 3:44 |
| C | 2 | 3 | 1 | 3:15 |

3 rows

Task 15:calculating the mean of the numbers in the vector

Hide

```
Numbers <- c(5,10,15,20,25)

Numbers %>%
  mean()
```

```
[1] 15
```

Task 16:It find AvgTemp values across columns based on Season and Selects all columns except Summer.

Hide

```
LongTemps %>%
  spread(Season, AvgTemp) %>%
  select(-Summer)
```

| Year | Fall | Spring | Winter |
|---:|---:|---:|---:|
| <dbl> | <dbl> | <dbl> | <dbl> |
| 2015 | 52 | 46 | 40 |
| 2016 | 46 | 40 | 38 |
| 2017 | 54 | 50 | 42 |
| 2018 | 56 | 48 | 44 |

4 rows

Task 17:Reshaping the LongTemps data frame from long to wide format by spreading AvgTemp values across columns

Hide

```
LongTemps %>%
  spread(data = ., Season, AvgTemp) %>%
  select(-Summer)
```

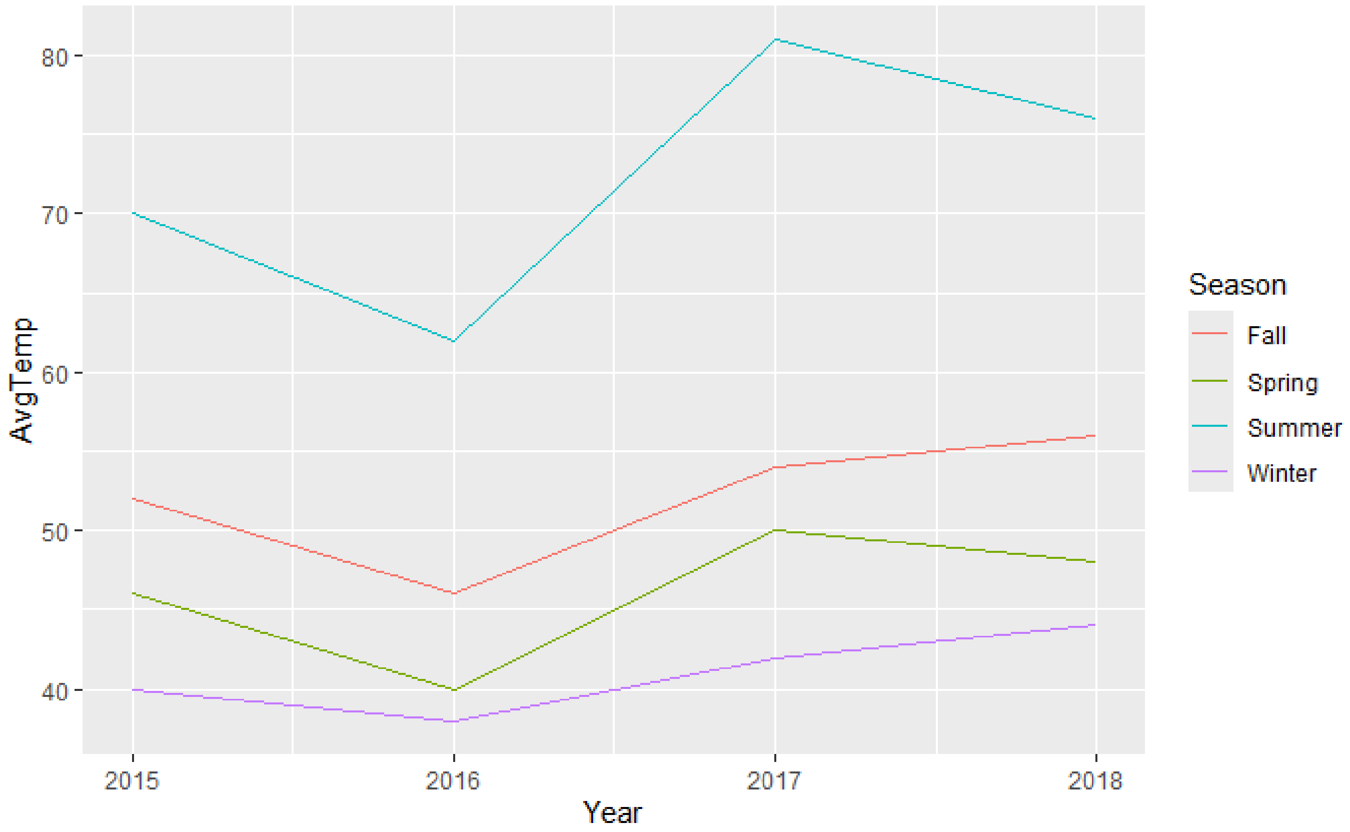
| Year | Fall | Spring | Winter |
|---:|---:|---:|---:|
| <dbl> | <dbl> | <dbl> | <dbl> |
| 2015 | 52 | 46 | 40 |
| 2016 | 46 | 40 | 38 |
| 2017 | 54 | 50 | 42 |
| 2018 | 56 | 48 | 44 |

4 rows

Task 18: create a line plot of average temperature over years), with each season represented by a different color.

Hide

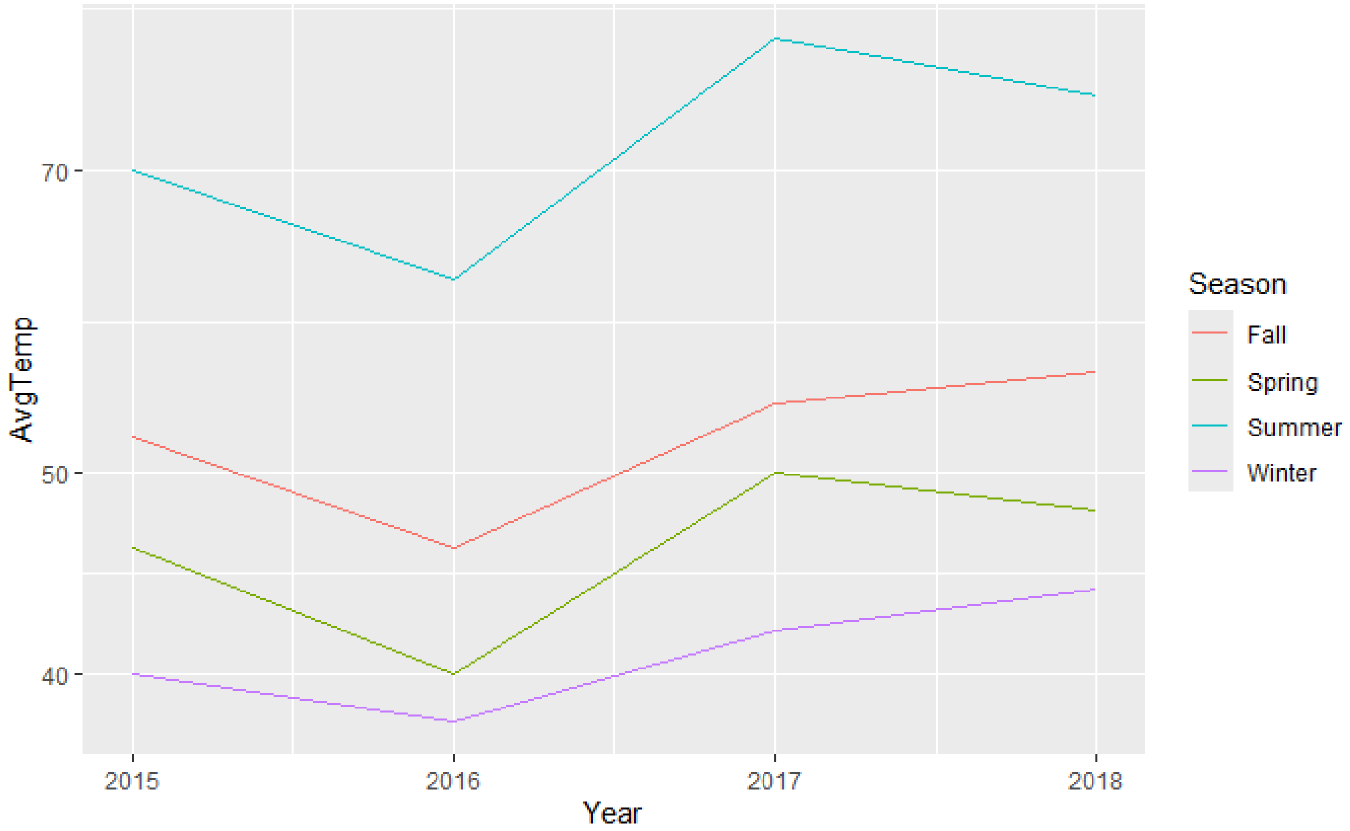
```
LongTemps %>%
  ggplot(aes(x = Year, y = AvgTemp, color = Season)) +
  geom_line()
```



Task 19:Ploting average temperature.

Hide

```
LongTemps %>%
  ggplot(aes(x = Year, y = AvgTemp, color = Season)) +
  geom_line() +
  scale_y_log10()
```



Task 20:Multiplying avg temp by 2

Hide

```
LongTemps %>%
  mutate(TwiceTemp = AvgTemp * 2)
```

| Year<br>&lt;dbl&gt; | Season<br>&lt;chr&gt; | AvgTemp<br>&lt;dbl&gt; | TwiceTemp<br>&lt;dbl&gt; |
|---:|:---|---:|---:|
| 2015 | Winter | 40 | 80 |
| 2016 | Winter | 38 | 76 |
| 2017 | Winter | 42 | 84 |
| 2018 | Winter | 44 | 88 |
| 2015 | Spring | 46 | 92 |
| 2016 | Spring | 40 | 80 |
| 2017 | Spring | 50 | 100 |
| 2018 | Spring | 48 | 96 |
| 2015 | Summer | 70 | 140 |
| 2016 | Summer | 62 | 124 |

1-10 of 16 rows Previous **1** 2 Next

Task 21:This code adds new columns to the LongTemps data frame where TwiceTemp is twice the AvgTemp, TwiceSquaredTemp is the square of TwiceTemp, and YearSeason combines Year and Season values into a single string.

Hide

```
LongTemps %>%
  mutate(TwiceTemp = AvgTemp * 2,
         TwiceSquaredTemp = TwiceTemp^2,
         YearSeason = paste(Year, Season))
```

| Year<br>&lt;dbl&gt; | Season<br>&lt;chr&gt; | AvgTemp<br>&lt;dbl&gt; | TwiceTemp<br>&lt;dbl&gt; | TwiceSquaredTemp<br>&lt;dbl&gt; | YearSeason<br>&lt;chr&gt; |
|---:|:---|---:|---:|---:|:---|
| 2015 | Winter | 40 | 80 | 6400 | 2015 Winter |
| 2016 | Winter | 38 | 76 | 5776 | 2016 Winter |
| 2017 | Winter | 42 | 84 | 7056 | 2017 Winter |
| 2018 | Winter | 44 | 88 | 7744 | 2018 Winter |
| 2015 | Spring | 46 | 92 | 8464 | 2015 Spring |
| 2016 | Spring | 40 | 80 | 6400 | 2016 Spring |
| 2017 | Spring | 50 | 100 | 10000 | 2017 Spring |
| 2018 | Spring | 48 | 96 | 9216 | 2018 Spring |

| Year | Season | AvgTemp | TwiceTemp | TwiceSquaredTemp | YearSeason |
|---:|---|---:|---:|---:|---|
| <dbl> | <chr> | <dbl> | <dbl> | <dbl> | <chr> |
| 2015 | Summer | 70 | 140 | 19600 | 2015 Summer |
| 2016 | Summer | 62 | 124 | 15376 | 2016 Summer |

1-10 of 16 rows Previous **1** 2 Next

Task 22:This code creates a new column YearSeason in the data frame by combining the Season and Year values with additional text (“The”, “of”) using the paste() function.

Hide

```
LongTemps %>%
  mutate(YearSeason = paste("The", Season, "of", Year))
```

| Year | Season | AvgTemp | YearSeason |
|---:|---|---:|---|
| <dbl> | <chr> | <dbl> | <chr> |
| 2015 | Winter | 40 | The Winter of 2015 |
| 2016 | Winter | 38 | The Winter of 2016 |
| 2017 | Winter | 42 | The Winter of 2017 |
| 2018 | Winter | 44 | The Winter of 2018 |
| 2015 | Spring | 46 | The Spring of 2015 |
| 2016 | Spring | 40 | The Spring of 2016 |
| 2017 | Spring | 50 | The Spring of 2017 |
| 2018 | Spring | 48 | The Spring of 2018 |
| 2015 | Summer | 70 | The Summer of 2015 |
| 2016 | Summer | 62 | The Summer of 2016 |

1-10 of 16 rows Previous **1** 2 Next

Task 23:This code first doubles the AvgTemp values in the LongTemps data frame, and then creates a new data frame.

Hide

```
LongTemps %>%
  mutate(TwiceTemp = AvgTemp * 2) %>%
  transmute(AvgTemp = AvgTemp,
            TwiceSquaredTemp = TwiceTemp^2,
            YearSeason = paste(Year, Season))
```

| AvgTemp | TwiceSquaredTemp | YearSeason |
|---:|---:|---|
| <dbl> | <dbl> | <chr> |
| 40 | 6400 | 2015 Winter |
| 38 | 5776 | 2016 Winter |
| 42 | 7056 | 2017 Winter |

| AvgTemp<br><dbl> | TwiceSquaredTemp<br><dbl> | YearSeason<br><chr> |
|---:|---:|:---|
| 44 | 7744 | 2018 Winter |
| 46 | 8464 | 2015 Spring |
| 40 | 6400 | 2016 Spring |
| 50 | 10000 | 2017 Spring |
| 48 | 9216 | 2018 Spring |
| 70 | 19600 | 2015 Summer |
| 62 | 15376 | 2016 Summer |

1-10 of 16 rows Previous **1** 2 Next

Task 24:Turining iris dataset into tibble

Hide

```
iris <- as.tibble(iris)
```

Task 25:Displaying the data

Hide

```
iris
```

| Sepal.Length<br><dbl> | Sepal.Width<br><dbl> | Petal.Length<br><dbl> | Petal.Width<br><dbl> | Species<br><fctr> |
|---:|---:|---:|---:|:---|
| 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 4.9 | 3.0 | 1.4 | 0.2 | setosa |
| 4.7 | 3.2 | 1.3 | 0.2 | setosa |
| 4.6 | 3.1 | 1.5 | 0.2 | setosa |
| 5.0 | 3.6 | 1.4 | 0.2 | setosa |
| 5.4 | 3.9 | 1.7 | 0.4 | setosa |
| 4.6 | 3.4 | 1.4 | 0.3 | setosa |
| 5.0 | 3.4 | 1.5 | 0.2 | setosa |
| 4.4 | 2.9 | 1.4 | 0.2 | setosa |
| 4.9 | 3.1 | 1.5 | 0.1 | setosa |

1-10 of 150 rows Previous **1** 2 3 4 5 6 ... 15 Next

Task 26:Displaying only 1 columns.

Hide

```
iris[, 1]
```

| Sepal.Length<br>&lt;dbl&gt; |
|---:|
| 5.1 |
| 4.9 |
| 4.7 |
| 4.6 |
| 5.0 |
| 5.4 |
| 4.6 |
| 5.0 |
| 4.4 |
| 4.9 |

1-10 of 150 rows Previous **1** 2 3 4 5 6 ... 15 Next

Task 27:Displaying only 1st row.

Hide

```
iris[1, ]
```

| Sepal.Length<br>&lt;dbl&gt; | Sepal.Width<br>&lt;dbl&gt; | Petal.Length<br>&lt;dbl&gt; | Petal.Width<br>&lt;dbl&gt; | Species<br>&lt;fctr&gt; |
|---:|---:|---:|---:|:---|
| 5.1 | 3.5 | 1.4 | 0.2 | setosa |

1 row

Task 28:Displaying 1st 4 rows of dataset

Hide

```
iris[c(1,2,3,4), ]
```

| Sepal.Length<br>&lt;dbl&gt; | Sepal.Width<br>&lt;dbl&gt; | Petal.Length<br>&lt;dbl&gt; | Petal.Width<br>&lt;dbl&gt; | Species<br>&lt;fctr&gt; |
|---:|---:|---:|---:|:---|
| 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 4.9 | 3.0 | 1.4 | 0.2 | setosa |
| 4.7 | 3.2 | 1.3 | 0.2 | setosa |
| 4.6 | 3.1 | 1.5 | 0.2 | setosa |

4 rows

Task 29:Displaying 1st 4 rows of dataset

Hide

```
iris[1:4, ]
```

| Sepal.Length<br><dbl> | Sepal.Width<br><dbl> | Petal.Length<br><dbl> | Petal.Width<br><dbl> | Species<br><fctr> |
|---:|---:|---:|---:|:---|
| 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 4.9 | 3.0 | 1.4 | 0.2 | setosa |
| 4.7 | 3.2 | 1.3 | 0.2 | setosa |
| 4.6 | 3.1 | 1.5 | 0.2 | setosa |

4 rows

Task 30:Displaying intersection value of the 1row and 1column only

Hide

```
iris[1,1]
```

| Sepal.Length<br><dbl> |
|---:|
| 5.1 |

1 row

Task 31:Displayingintersection value in vector

Hide

```
iris[[1, 1]]
```

```
[1] 5.1
```

Task 32:Dispalying in vector format.

Hide

```
iris$Sepal.Length
```

```
  [1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4 4.9 5.4 4.8 4.8 4.3 5.8 5.7 5.4 5.1 5.7 5.1 5.4 5.1
4.6 5.1 4.8 5.0
 [27] 5.0 5.2 5.2 4.7 4.8 5.4 5.2 5.5 4.9 5.0 5.5 4.9 4.4 5.1 5.0 4.5 4.4 5.0 5.1 4.8 5.1 4.6
5.3 5.0 7.0 6.4
 [53] 6.9 5.5 6.5 5.7 6.3 4.9 6.6 5.2 5.0 5.9 6.0 6.1 5.6 6.7 5.6 5.8 6.2 5.6 5.9 6.1 6.3 6.1
6.4 6.6 6.8 6.7
 [79] 6.0 5.7 5.5 5.5 5.8 6.0 5.4 6.0 6.7 6.3 5.6 5.5 5.5 6.1 5.8 5.0 5.6 5.7 5.7 6.2 5.1 5.7
6.3 5.8 7.1 6.3
[105] 6.5 7.6 4.9 7.3 6.7 7.2 6.5 6.4 6.8 5.7 5.8 6.4 6.5 7.7 7.7 6.0 6.9 5.6 7.7 6.3 6.7 7.2
6.2 6.1 6.4 7.2
[131] 7.4 7.9 6.4 6.3 6.1 7.7 6.3 6.4 6.0 6.9 6.7 6.9 5.8 6.8 6.7 6.7 6.3 6.5 6.2 5.9
```

Task 33: calculating the mean

Hide

```
mean(iris$Sepal.Length)
```

```
[1] 5.843333
```

Hide

```
sd(iris$Sepal.Length)
```

```
[1] 0.8280661
```

Hide

```
cor.test(iris$Sepal.Length, iris$Sepal.Width)
```

```


    Pearson's product-moment correlation

data:  iris$Sepal.Length and iris$Sepal.Width
t = -1.4403, df = 148, p-value = 0.1519
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.27269325  0.04351158
sample estimates:
       cor
-0.1175698
```

Task 34:Displaying species=setosa only

Hide

```
iris[iris$Species == "setosa", ]
```

| Sepal.Length | Sepal.Width | Petal.Length | Petal.Width | Species |
|---:|---:|---:|---:|:---|
| <dbl> | <dbl> | <dbl> | <dbl> | <fctr> |
| 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 4.9 | 3.0 | 1.4 | 0.2 | setosa |
| 4.7 | 3.2 | 1.3 | 0.2 | setosa |
| 4.6 | 3.1 | 1.5 | 0.2 | setosa |
| 5.0 | 3.6 | 1.4 | 0.2 | setosa |
| 5.4 | 3.9 | 1.7 | 0.4 | setosa |
| 4.6 | 3.4 | 1.4 | 0.3 | setosa |
| 5.0 | 3.4 | 1.5 | 0.2 | setosa |
| 4.4 | 2.9 | 1.4 | 0.2 | setosa |
| 4.9 | 3.1 | 1.5 | 0.1 | setosa |

1-10 of 50 rows Previous **1** 2 3 4 5 Next

Task 35:Displaying whose speal.length is more than 7.5

Hide

```
iris[iris$Sepal.Length > 7.5, ]
```

| Sepal.Length | Sepal.Width | Petal.Length | Petal.Width | Species |
|---:|---:|---:|---:|:---|
| <dbl> | <dbl> | <dbl> | <dbl> | <fctr> |
| 7.6 | 3.0 | 6.6 | 2.1 | virginica |
| 7.7 | 3.8 | 6.7 | 2.2 | virginica |
| 7.7 | 2.6 | 6.9 | 2.3 | virginica |
| 7.7 | 2.8 | 6.7 | 2.0 | virginica |
| 7.9 | 3.8 | 6.4 | 2.0 | virginica |
| 7.7 | 3.0 | 6.1 | 2.3 | virginica |

6 rows

Task 36:Filtaring the species

Hide

```
iris %>%
  filter(Species == "setosa")
```

| Sepal.Length | Sepal.Width | Petal.Length | Petal.Width | Species |
|---:|---:|---:|---:|:---|
| <dbl> | <dbl> | <dbl> | <dbl> | <fctr> |
| 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 4.9 | 3.0 | 1.4 | 0.2 | setosa |
| 4.7 | 3.2 | 1.3 | 0.2 | setosa |
| 4.6 | 3.1 | 1.5 | 0.2 | setosa |
| 5.0 | 3.6 | 1.4 | 0.2 | setosa |
| 5.4 | 3.9 | 1.7 | 0.4 | setosa |
| 4.6 | 3.4 | 1.4 | 0.3 | setosa |
| 5.0 | 3.4 | 1.5 | 0.2 | setosa |
| 4.4 | 2.9 | 1.4 | 0.2 | setosa |
| 4.9 | 3.1 | 1.5 | 0.1 | setosa |

1-10 of 50 rows Previous **1** 2 3 4 5 Next

Task 37:Filtering the species, sepal.lenth and sepal.width

Hide

```
iris %>%
  filter(Species == "setosa" & Sepal.Length == 5.1 & Sepal.Width == 3.3)
```

| Sepal.Length<br><dbl> | Sepal.Width<br><dbl> | Petal.Length<br><dbl> | Petal.Width<br><dbl> | Species<br><fctr> |
|---:|---:|---:|---:|:---|
| 5.1 | 3.3 | 1.7 | 0.5 | setosa |

1 row

Task 38:Filteing the data

Hide

```
iris %>%
  filter(Species == "setosa" & Species == "versicolor")
```

0 rows

Hide

```
iris %>%
  filter(Species %in% c("setosa",
                        "versicolor"))
```

| Sepal.Length<br><dbl> | Sepal.Width<br><dbl> | Petal.Length<br><dbl> | Petal.Width<br><dbl> | Species<br><fctr> |
|---:|---:|---:|---:|:---|
| 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 4.9 | 3.0 | 1.4 | 0.2 | setosa |
| 4.7 | 3.2 | 1.3 | 0.2 | setosa |
| 4.6 | 3.1 | 1.5 | 0.2 | setosa |
| 5.0 | 3.6 | 1.4 | 0.2 | setosa |
| 5.4 | 3.9 | 1.7 | 0.4 | setosa |
| 4.6 | 3.4 | 1.4 | 0.3 | setosa |
| 5.0 | 3.4 | 1.5 | 0.2 | setosa |
| 4.4 | 2.9 | 1.4 | 0.2 | setosa |
| 4.9 | 3.1 | 1.5 | 0.1 | setosa |

1-10 of 100 rows Previous **1** 2 3 4 5 6 ... 10 Next

Task 39:Displaying the mean od sepal.length

Hide

```
mean(iris$Sepal.Length)
```

```
[1] 5.843333
```

Task 40:Filtering the data and then displaying the mean

Hide

```
setosa <- iris %>%
  filter(Species == "setosa")
virginica <- iris %>%
  filter(Species == "virginica")
versicolor <- iris %>%
  filter(Species == "versicolor")


mean(setosa$Sepal.Length)
```

```
[1] 5.006
```

Task 41:Displaying mean in different ways

Hide

```
mean(virginica$Sepal.Length)
```

```
[1] 6.588
```

Hide

```
mean(versicolor$Sepal.Length)
```

```
[1] 5.936
```

Task 42:Summarizing the species

Hide

```
iris %>%
  group_by(Species) %>%
  summarise(MeanSepalLength = mean(Sepal.Length))
```

| **Species** | **MeanSepalLength** |
|---|---:|
| <fctr> | <dbl> |
| setosa | 5.006 |
| versicolor | 5.936 |
| virginica | 6.588 |

3 rows

Task 43:This code calculates the deviation of each Sepal.Length value.

Hide

```
iris %>%
  group_by(Species) %>%
  mutate(SLDistanceFromMean = Sepal.Length - mean(Sepal.Length))
```

| Sepal.Length<br>&lt;dbl&gt; | Sepal.Width<br>&lt;dbl&gt; | Petal.Length<br>&lt;dbl&gt; | Petal.Width<br>&lt;dbl&gt; | Species<br>&lt;fctr&gt; | SLDistanceFromMean<br>&lt;dbl&gt; |
|---:|---:|---:|---:|:---|---:|
| 5.1 | 3.5 | 1.4 | 0.2 | setosa | 0.094 |
| 4.9 | 3.0 | 1.4 | 0.2 | setosa | -0.106 |
| 4.7 | 3.2 | 1.3 | 0.2 | setosa | -0.306 |
| 4.6 | 3.1 | 1.5 | 0.2 | setosa | -0.406 |
| 5.0 | 3.6 | 1.4 | 0.2 | setosa | -0.006 |
| 5.4 | 3.9 | 1.7 | 0.4 | setosa | 0.394 |
| 4.6 | 3.4 | 1.4 | 0.3 | setosa | -0.406 |
| 5.0 | 3.4 | 1.5 | 0.2 | setosa | -0.006 |
| 4.4 | 2.9 | 1.4 | 0.2 | setosa | -0.606 |
| 4.9 | 3.1 | 1.5 | 0.1 | setosa | -0.106 |

1-10 of 150 rows Previous **1** 2 3 4 5 6 ... 15 Next

Task 44:Calculating mean adn deviation.

Hide

```
iris %>%
  select(c(Sepal.Length, Species)) %>%
  group_by(Species) %>%
  mutate(SLDistanceFromGroupMean = Sepal.Length - mean(Sepal.Length)) %>%
  ungroup() %>%
  mutate(SLDistanceFromTotalMean = Sepal.Length - mean(Sepal.Length))
```

| Sepal.Length<br>&lt;dbl&gt; | Species<br>&lt;fctr&gt; | SLDistanceFromGroupMean<br>&lt;dbl&gt; | SLDistanceFromTotalMean<br>&lt;dbl&gt; |
|---:|:---|---:|---:|
| 5.1 | setosa | 0.094 | -0.74333333 |
| 4.9 | setosa | -0.106 | -0.94333333 |
| 4.7 | setosa | -0.306 | -1.14333333 |
| 4.6 | setosa | -0.406 | -1.24333333 |
| 5.0 | setosa | -0.006 | -0.84333333 |
| 5.4 | setosa | 0.394 | -0.44333333 |
| 4.6 | setosa | -0.406 | -1.24333333 |
| 5.0 | setosa | -0.006 | -0.84333333 |
| 4.4 | setosa | -0.606 | -1.44333333 |
| 4.9 | setosa | -0.106 | -0.94333333 |

1-10 of 150 rows Previous **1** 2 3 4 5 6 ... 15 Next

Task 45:Assigning the value and then displaying it.

Hide

```
MissingExample <- tibble(w = c(1, 2, 3),
                         x = c("A", "B", "C"),
                         y = c("do", "re", NA),
                         z = c(807, NA, 780))
MissingExample
```

| w<br>&lt;dbl&gt; | x<br>&lt;chr&gt; | y<br>&lt;chr&gt; | z<br>&lt;dbl&gt; |
|---:|:---|:---|---:|
| 1 | A | do | 807 |
| 2 | B | re | *NA* |
| 3 | C | *NA* | 780 |

3 rows

Task 46:Displaying the mean

Hide

```
mean(MissingExample$z)
```

```
[1] NA
```

Hide

```
mean(MissingExample$z, na.rm = TRUE)
```

```
[1] 793.5
```

Task 47:Summarizing the mean

Hide

```
MissingExample %>%
  filter(!is.na(z)) %>%
  summarise(Mean = mean(z))
```

| Mean<br>&lt;dbl&gt; |
|---:|
| 793.5 |

1 row

Task 48:It removes rows containing missing values

Hide

```
MissingExample %>%
  drop_na()
```

| w | x | y | z |
|---|---|---|---|
| <dbl> | <chr> | <chr> | <dbl> |
| 1 | A | do | 807 |

1 row

Task:

Hide

```
#MissingExample %>%
 # replace_na(list(y = "mi", z = "078"))
```

Task:

Hide

```
#TreeData
```

Task 49:Assigning tibble value and then displaying it

Hide

```
TreeData <- tibble(Site = c("A","A","A","B","B"),
                   Species = c("Red Maple", "Sugar Maple", "Black Cherry", "Red Maple", "Suga
r Maple"),
                   Count = c(10,5,15,8,19))
TreeData
```

| Site | Species | Count |
|---|---|---|
| <chr> | <chr> | <dbl> |
| A | Red Maple | 10 |
| A | Sugar Maple | 5 |
| A | Black Cherry | 15 |
| B | Red Maple | 8 |
| B | Sugar Maple | 19 |

5 rows

Task 50:Summarizing the species

Hide

```
TreeData %>%
  group_by(Species) %>%
  summarise(Mean = mean(Count), StandardDev = sd(Count))
```

| Species | Mean | StandardDev |
|---|---|---|
| <chr> | <dbl> | <dbl> |
| Black Cherry | 15 | *NA* |
| Red Maple | 9 | 1.414214 |

| Species<br>&lt;chr&gt; | Mean<br>&lt;dbl&gt; | StandardDev<br>&lt;dbl&gt; |
|---|---:|---:|
| Sugar Maple | 12 | 9.899495 |

3 rows

Hide

```
TreeData %>%
  complete(Site, Species, fill = list(Count = 0))
```

| Site<br>&lt;chr&gt; | Species<br>&lt;chr&gt; | Count<br>&lt;dbl&gt; |
|---|---|---:|
| A | Black Cherry | 15 |
| A | Red Maple | 10 |
| A | Sugar Maple | 5 |
| B | Black Cherry | 0 |
| B | Red Maple | 8 |
| B | Sugar Maple | 19 |

6 rows

Task 51:It fills missing combinations of Site and Species with Count = 0 in TreeData

Hide

```
TreeData %>%
  complete(Site, Species, fill = list(Count = 0)) %>%
  group_by(Species) %>%
  summarise(Mean = mean(Count), StandardDev = sd(Count))
```

| Species<br>&lt;chr&gt; | Mean<br>&lt;dbl&gt; | StandardDev<br>&lt;dbl&gt; |
|---|---:|---:|
| Black Cherry | 7.5 | 10.606602 |
| Red Maple | 9.0 | 1.414214 |
| Sugar Maple | 12.0 | 9.899495 |

3 rows

Task 52: Count the repreted value.

Hide

```
LongTreeData <- TreeData %>%
  uncount(Count)

LongTreeData
```

| Site | Species |
|---|---|
| <chr> | <chr> |
| A | Red Maple |
| A | Red Maple |
| A | Red Maple |
| A | Red Maple |
| A | Red Maple |
| A | Red Maple |
| A | Red Maple |
| A | Red Maple |
| A | Red Maple |
| A | Red Maple |

1-10 of 57 rows Previous **1** 2 3 4 5 6 Next

Hide

```
LongTreeData %>%
  count(Site, Species)
```

| Site | Species | n |
|---|---|---:|
| <chr> | <chr> | <int> |
| A | Black Cherry | 15 |
| A | Red Maple | 10 |
| A | Sugar Maple | 5 |
| B | Red Maple | 8 |
| B | Sugar Maple | 19 |

5 rows

Task: Changing the given name of coloumn to a more descriptive name

Hide

```
LongTreeData %>%
  count(Site, Species) %>%
  rename(Count = n)
```

| Site | Species | Count |
|---|---|---:|
| <chr> | <chr> | <int> |
| A | Black Cherry | 15 |
| A | Red Maple | 10 |
| A | Sugar Maple | 5 |

| Site<br>&lt;chr&gt; | Species<br>&lt;chr&gt; | Count<br>&lt;int&gt; |
|---|---|---:|
| B | Red Maple | 8 |
| B | Sugar Maple | 19 |

5 rows

# Introduction to Data Analysis

Task 1: Install Gapminder package

Hide

```
install.packages("gapminder")
```

```
Error in install.packages : Updating loaded packages
```

Task 2: Import Dataset

Hide

```
library(gapminder)
```

Task 3: Check the first six rows and coloumns of the dataset

Hide

```
head(gapminder_unfiltered)
```

| country<br>&lt;fctr&gt; | continent<br>&lt;fctr&gt; | year<br>&lt;int&gt; | lifeExp<br>&lt;dbl&gt; | pop<br>&lt;int&gt; | gdpPercap<br>&lt;dbl&gt; |
|---|---|---:|---:|---:|---:|
| Afghanistan | Asia | 1952 | 28.801 | 8425333 | 779.4453 |
| Afghanistan | Asia | 1957 | 30.332 | 9240934 | 820.8530 |
| Afghanistan | Asia | 1962 | 31.997 | 10267083 | 853.1007 |
| Afghanistan | Asia | 1967 | 34.020 | 11537966 | 836.1971 |
| Afghanistan | Asia | 1972 | 36.088 | 13079460 | 739.9811 |
| Afghanistan | Asia | 1977 | 38.438 | 14880372 | 786.1134 |

6 rows

Task 4: Summarize the data

Hide

```
summary(gapminder_unfiltered)
```

```
          country        continent          year         lifeExp           pop               g
dpPercap
 Czech Republic:  58   Africa  : 637   Min.   :1950   Min.   :23.60   Min.   :5.941e+04   Mi
n.   :   241.2
 Denmark       :  58   Americas: 470   1st Qu.:1967   1st Qu.:58.33   1st Qu.:2.680e+06   1st
Qu.:  2505.3
 Finland       :  58   Asia    : 578   Median :1982   Median :69.61   Median :7.560e+06   Med
ian :  7825.8
 Iceland       :  58   Europe  :1302   Mean   :1980   Mean   :65.24   Mean   :3.177e+07   Mea
n   : 11313.8
 Japan         :  58   FSU     : 139   3rd Qu.:1996   3rd Qu.:73.66   3rd Qu.:1.961e+07   3rd
Qu.: 17355.8
 Netherlands   :  58   Oceania : 187   Max.   :2007   Max.   :82.67   Max.   :1.319e+09   Ma
x.   :113523.1
 (Other)       :2965
```

Task 5: Import psych

Hide

```
library(psych)
```

Task 6: Describe

Hide

```
describe(gapminder_unfiltered)
```

| | v...<br>&lt;dbl&gt; | n<br>&lt;dbl&gt; | mean<br>&lt;dbl&gt; | sd<br>&lt;dbl&gt; | median<br>&lt;dbl&gt; | trimmed<br>&lt;dbl&gt; | mad<br>&lt;dbl&gt; | |
|---|---|---|---|---|---|---|---|---|
| country* | 1 | 3313 | 93.30 | 53.42 | 92.00 | 93.47 | 68.20 | |
| continent* | 2 | 3313 | 3.12 | 1.40 | 3.00 | 3.08 | 1.48 | |
| year | 3 | 3313 | 1980.29 | 16.93 | 1982.00 | 1980.57 | 22.24 | 19 |
| lifeExp | 4 | 3313 | 65.24 | 11.77 | 69.61 | 66.63 | 8.45 | |
| pop | 5 | 3313 | 31773251.41 | 104501904.44 | 7559776.00 | 12014676.53 | 8509318.95 | 594 |
| gdpPercap | 6 | 3313 | 11313.82 | 11369.01 | 7825.82 | 9629.08 | 9200.98 | 2 |

6 rows | 1-9 of 13 columns

Task 7: Check the number of unique countries

Hide

```
length(unique(gapminder_unfiltered$country))
```

```
[1] 187
```

Task 8: Check the number of unique countries using n_distinct

Hide

```
dplyr::n_distinct(gapminder_unfiltered$country)
```

```
[1] 187
```

Task 9: Attaching package dplyr

Hide

```
library(dplyr)
```

Task 9: print extreme numbers using filter

Hide

```
gapminder_unfiltered %>%
  filter(lifeExp == min(lifeExp))
```

| country<br><fctr> | continent<br><fctr> | year<br><int> | lifeExp<br><dbl> | pop<br><int> | gdpPercap<br><dbl> |
|---|---|---:|---:|---:|---:|
| Rwanda | Africa | 1992 | 23.599 | 7290203 | 737.0686 |

1 row

Task 10:

Hide

```
gapminder_unfiltered %>%
  filter(gdpPercap == min(gdpPercap))
```

| country<br><fctr> | continent<br><fctr> | year<br><int> | lifeExp<br><dbl> | pop<br><int> | gdpPercap<br><dbl> |
|---|---|---:|---:|---:|---:|
| Congo, Dem. Rep. | Africa | 2002 | 44.966 | 55379852 | 241.1659 |

1 row

Task 11: Checking FSU continent

Hide

```
gapminder_unfiltered %>%
  filter(continent == "FSU")
```

| country<br><fctr> | continent<br><fctr> | year<br><int> | lifeExp<br><dbl> | pop<br><int> | gdpPercap<br><dbl> |
|---|---|---:|---:|---:|---:|
| Armenia | FSU | 1992 | 68.663 | 3378331 | 1442.938 |
| Armenia | FSU | 1997 | 70.377 | 3059000 | 1791.347 |
| Armenia | FSU | 2002 | 71.403 | 3013818 | 2692.304 |
| Armenia | FSU | 2007 | 71.965 | 2971650 | 4942.544 |
| Belarus | FSU | 1973 | 72.710 | 9236465 | 4958.608 |

| country | continent | year | lifeExp | pop | gdpPercap |
|---|---|---|---|---|---|
| <fctr> | <fctr> | <int> | <dbl> | <int> | <dbl> |
| Belarus | FSU | 1990 | 71.150 | 10215208 | 6807.781 |
| Belarus | FSU | 1991 | 70.580 | 10244639 | 6693.188 |
| Belarus | FSU | 1992 | 70.170 | 10306362 | 6014.406 |
| Belarus | FSU | 1993 | 69.050 | 10360516 | 5528.263 |
| Belarus | FSU | 1994 | 68.820 | 10387841 | 4868.616 |

1-10 of 139 rows Previous **1** 2 3 4 5 6 ... 14 Next

Task 12: Import ggplot2 library

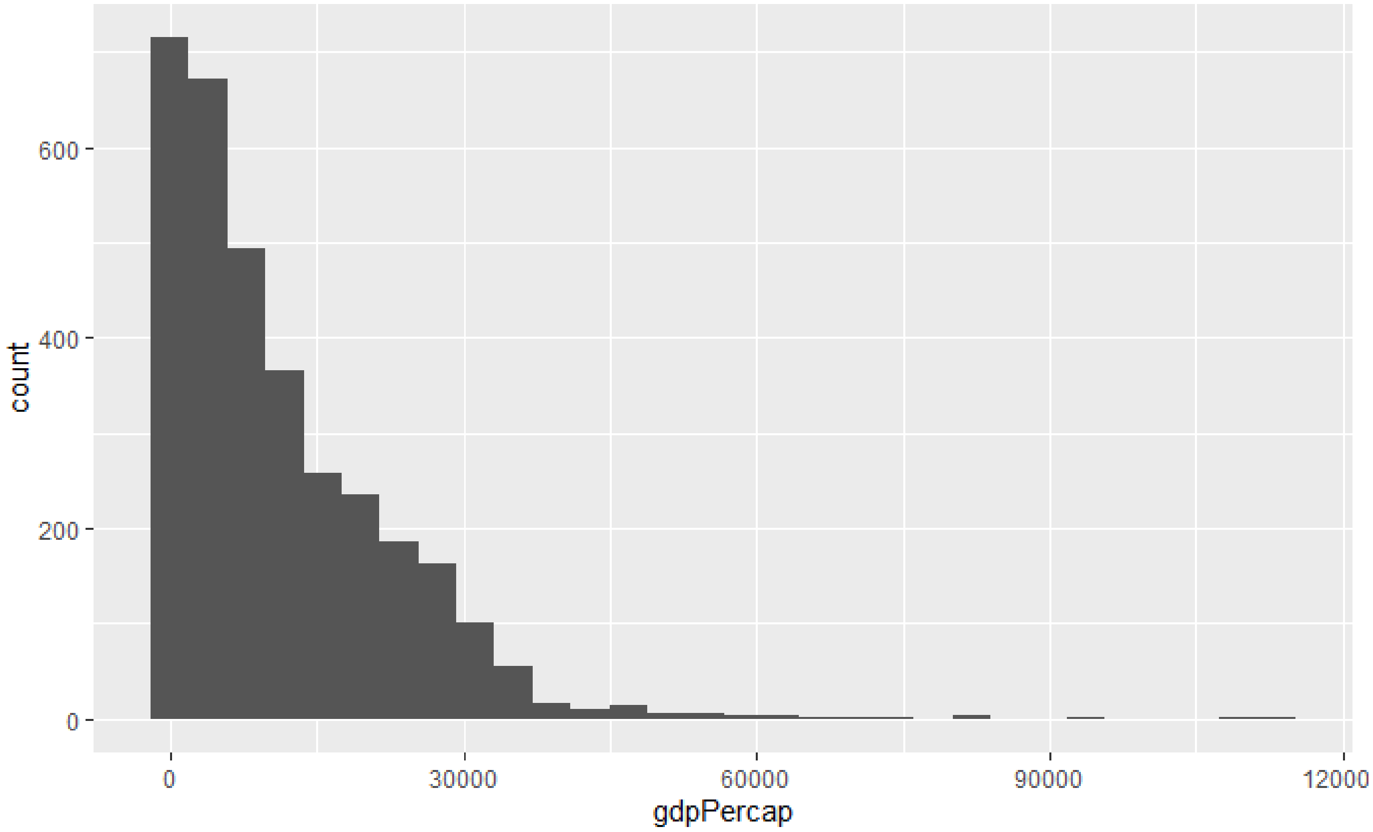
Hide

```
library(ggplot2)
```

Task 13: Visualizing unfiltered data

Hide

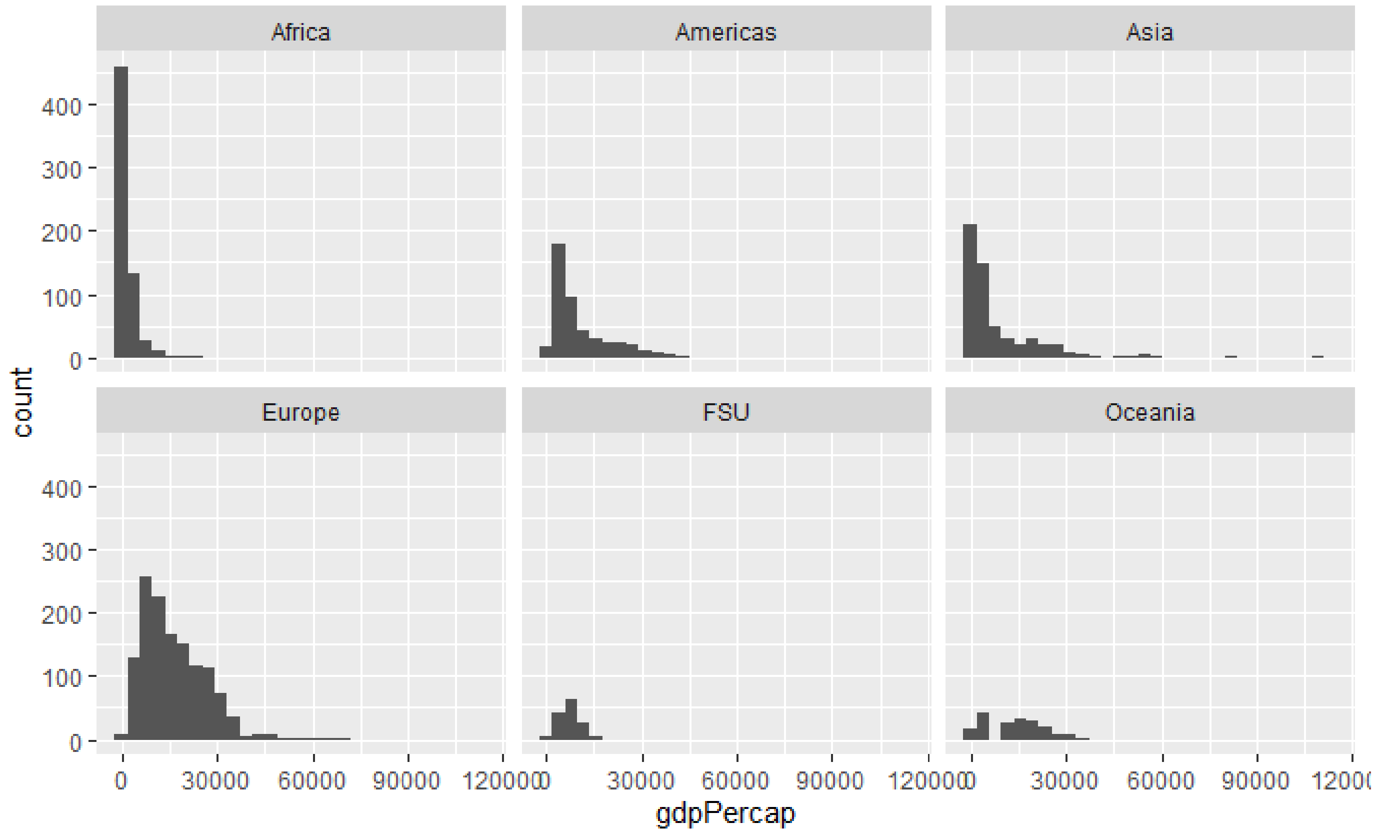
```
ggplot(gapminder_unfiltered, aes(gdpPercap)) +
  geom_histogram()
```



Task 14: Visualizing unfiltered data for every continent

Hide

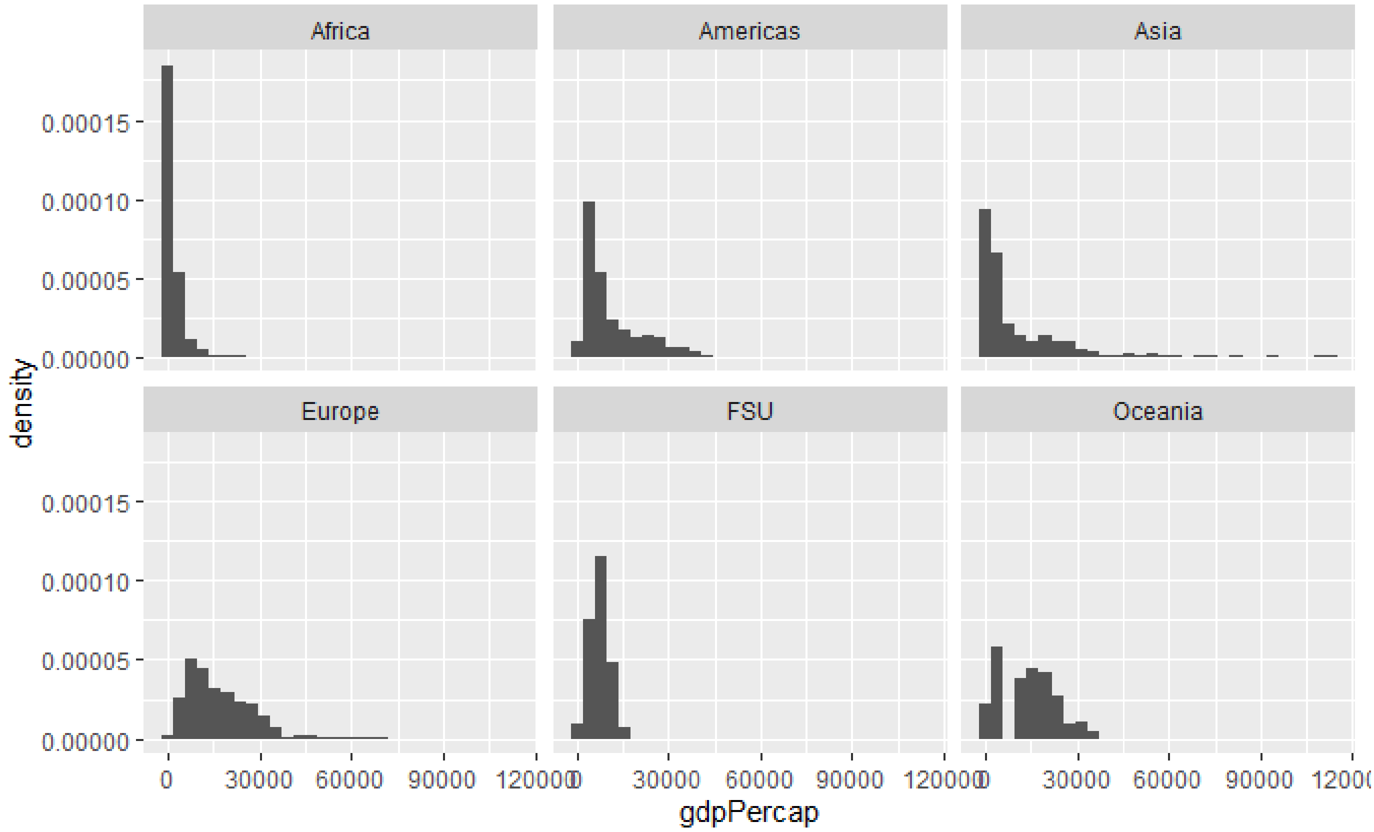
```
ggplot(gapminder_unfiltered, aes(gdpPercap)) +
  geom_histogram() +
  facet_wrap(~ continent)
```



Task 15: Checking outliers density in all continents

Hide

```
ggplot(gapminder_unfiltered, aes(gdpPercap, ..density..)) +
  geom_histogram() +
  facet_wrap(~ continent)
```



Task 16: Using frequency polygons

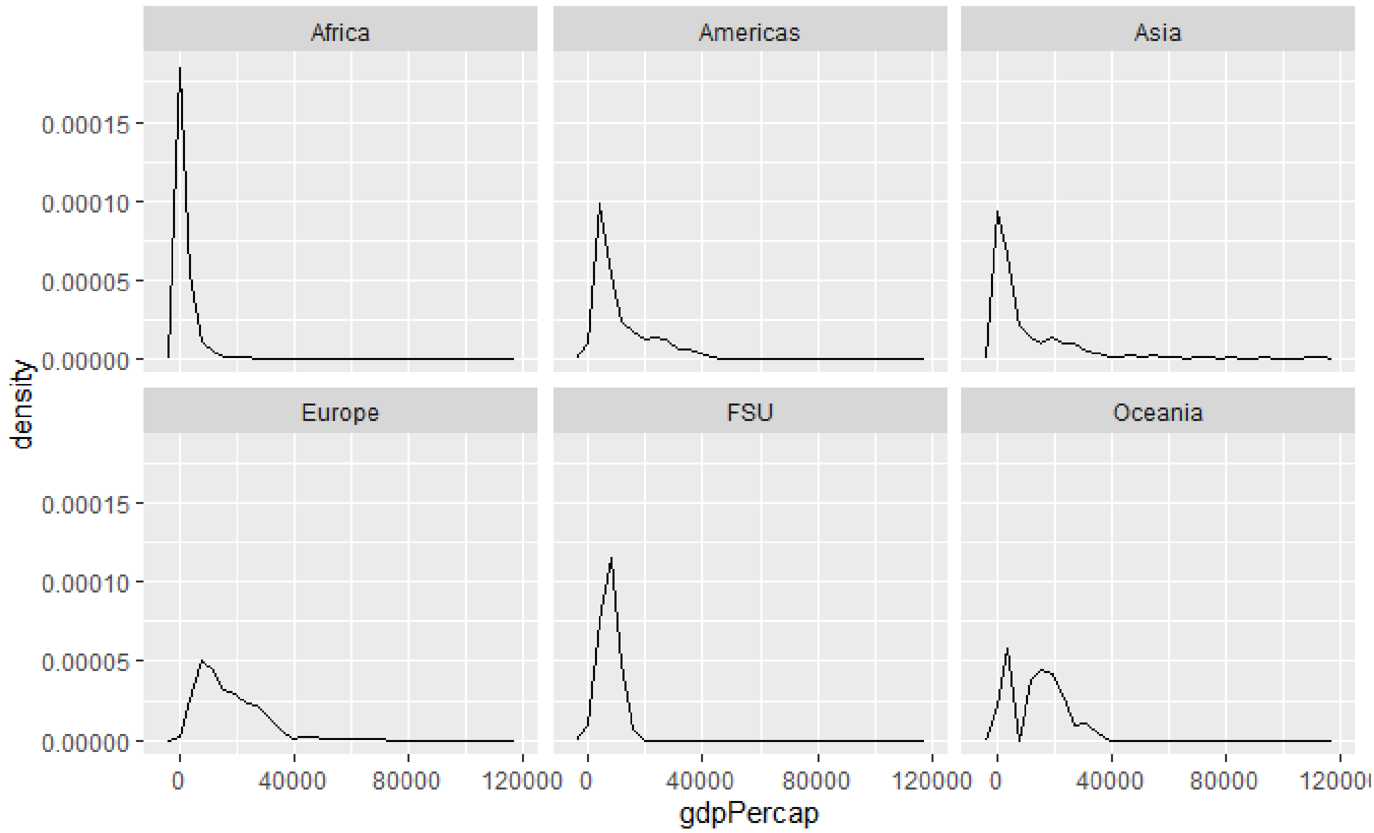
Hide

```
ggplot(gapminder_unfiltered, aes(gdpPercap, ..density..)) +
  geom_freqpoly() +
  facet_wrap(~ continent)
install.packages("gapminder")
```

```
WARNING: Rtools is required to build R packages but is not currently installed. Please downlo
ad and install the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
```

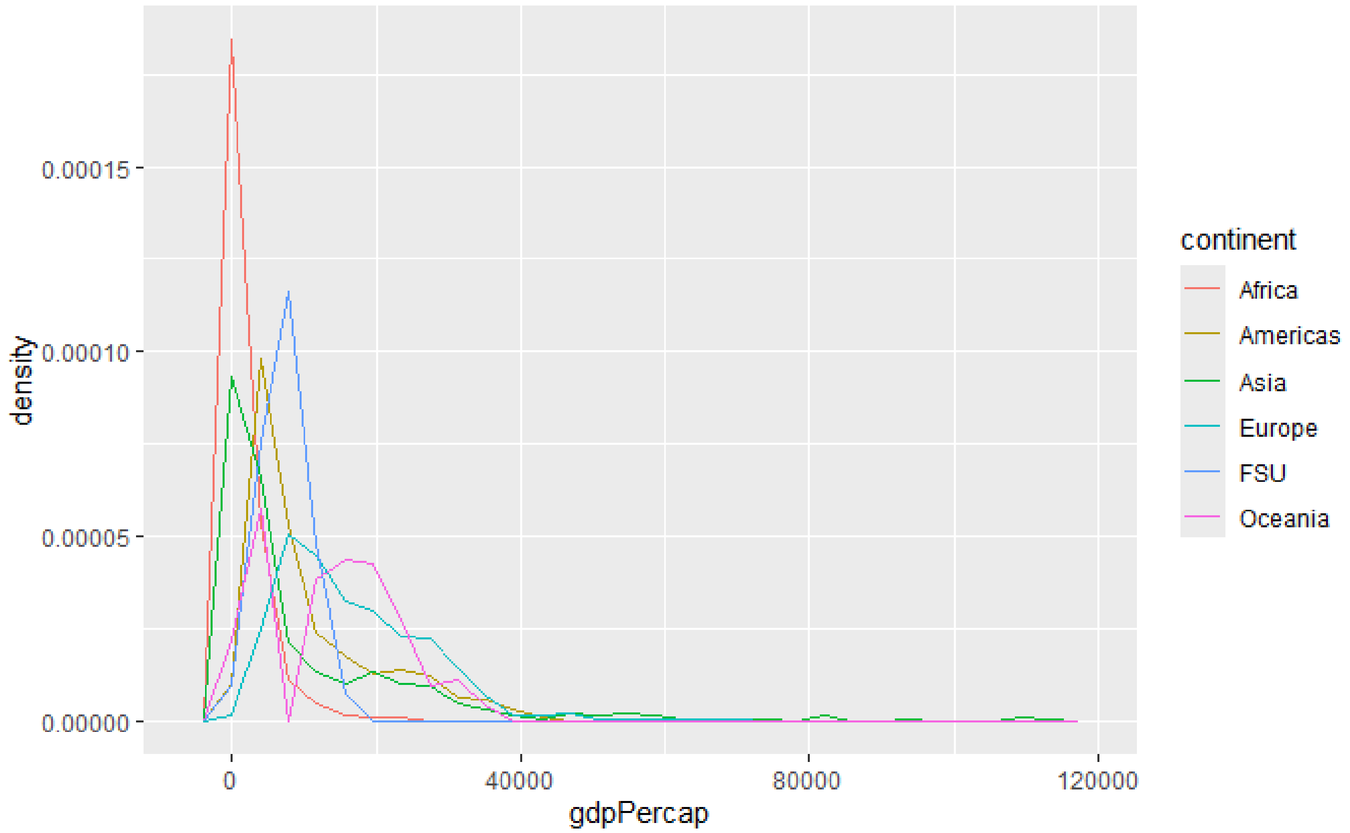
```
Warning in install.packages :
  package ‘gapminder’ is in use and will not be installed
```



Task 17: Combining different histograms as one

Hide

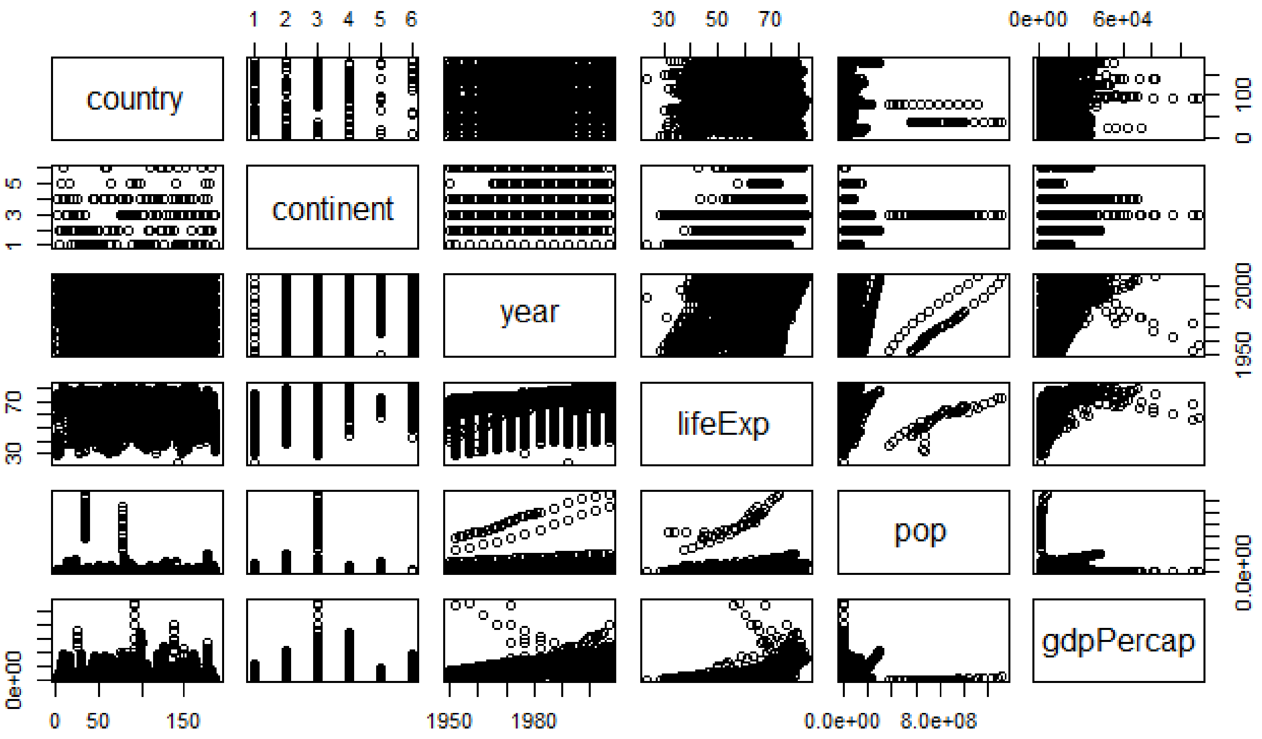
```
ggplot(gapminder_unfiltered, aes(gdpPercap, ..density.., color = continent)) +
  geom_freqpoly()
```



Task 17: Using pairs to check scatterplots between each of variables

Hide

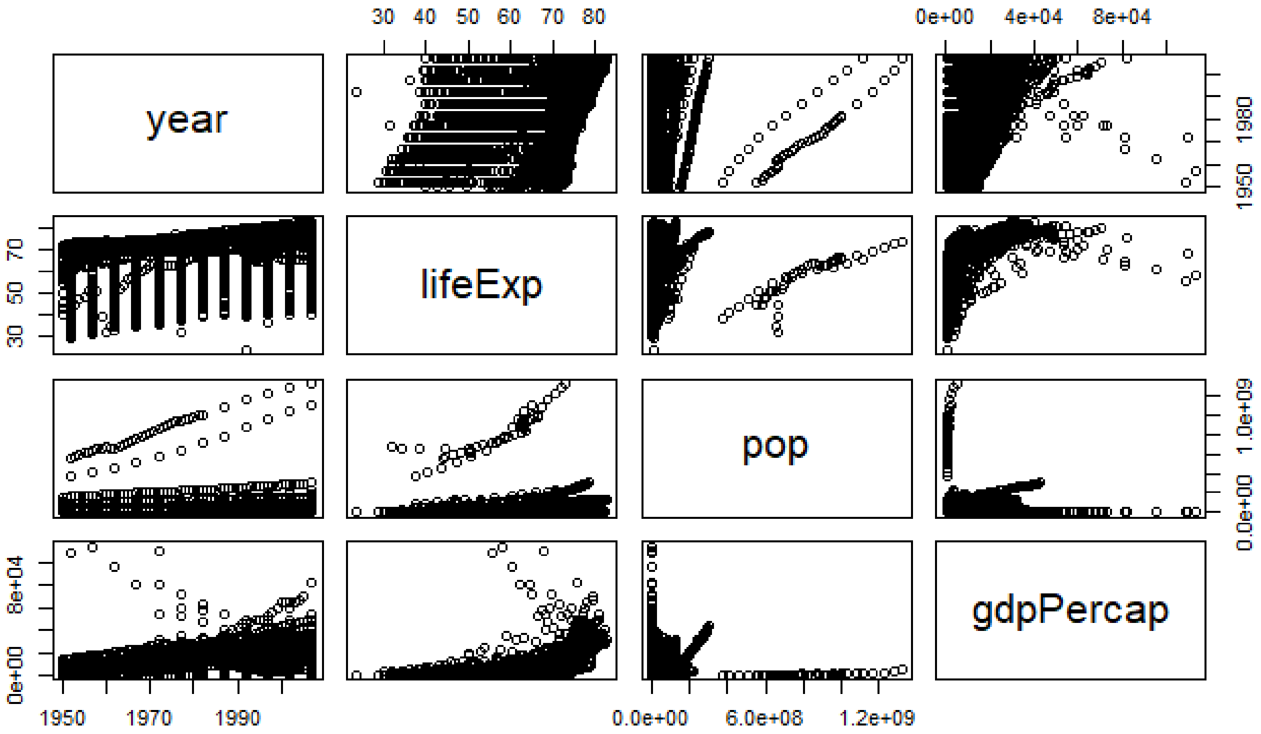
```
pairs(gapminder_unfiltered)
```



Task 18: Create new dataset using select

Hide

```
gp <- select(gapminder_unfiltered, 3:6)
pairs(gp)
```



Task 19:

Hide

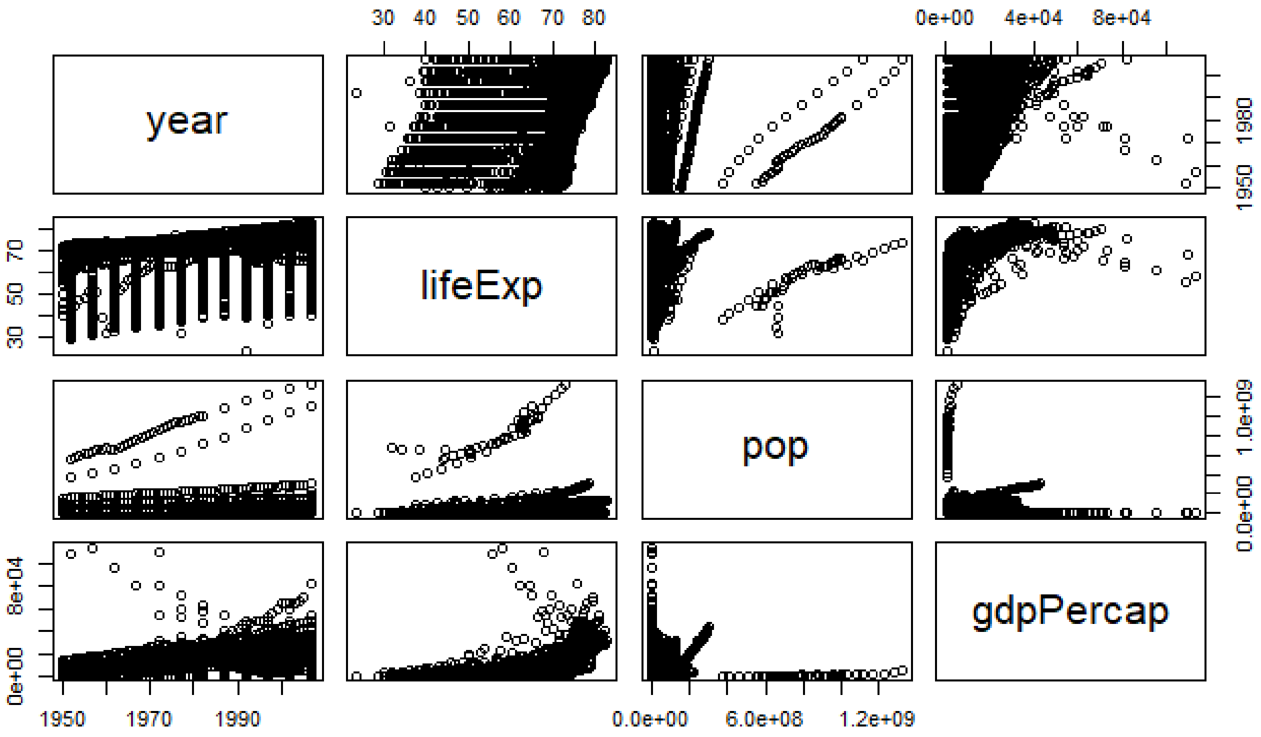
```
head(select_if(gapminder_unfiltered, is.numeric))
```

| year | lifeExp | pop | gdpPercap |
|---|---|---|---|
| <int> | <dbl> | <int> | <dbl> |
| 1952 | 28.801 | 8425333 | 779.4453 |
| 1957 | 30.332 | 9240934 | 820.8530 |
| 1962 | 31.997 | 10267083 | 853.1007 |
| 1967 | 34.020 | 11537966 | 836.1971 |
| 1972 | 36.088 | 13079460 | 739.9811 |
| 1977 | 38.438 | 14880372 | 786.1134 |

6 rows

Hide

```
pairs(select_if(gapminder_unfiltered, is.numeric))
```



Hide

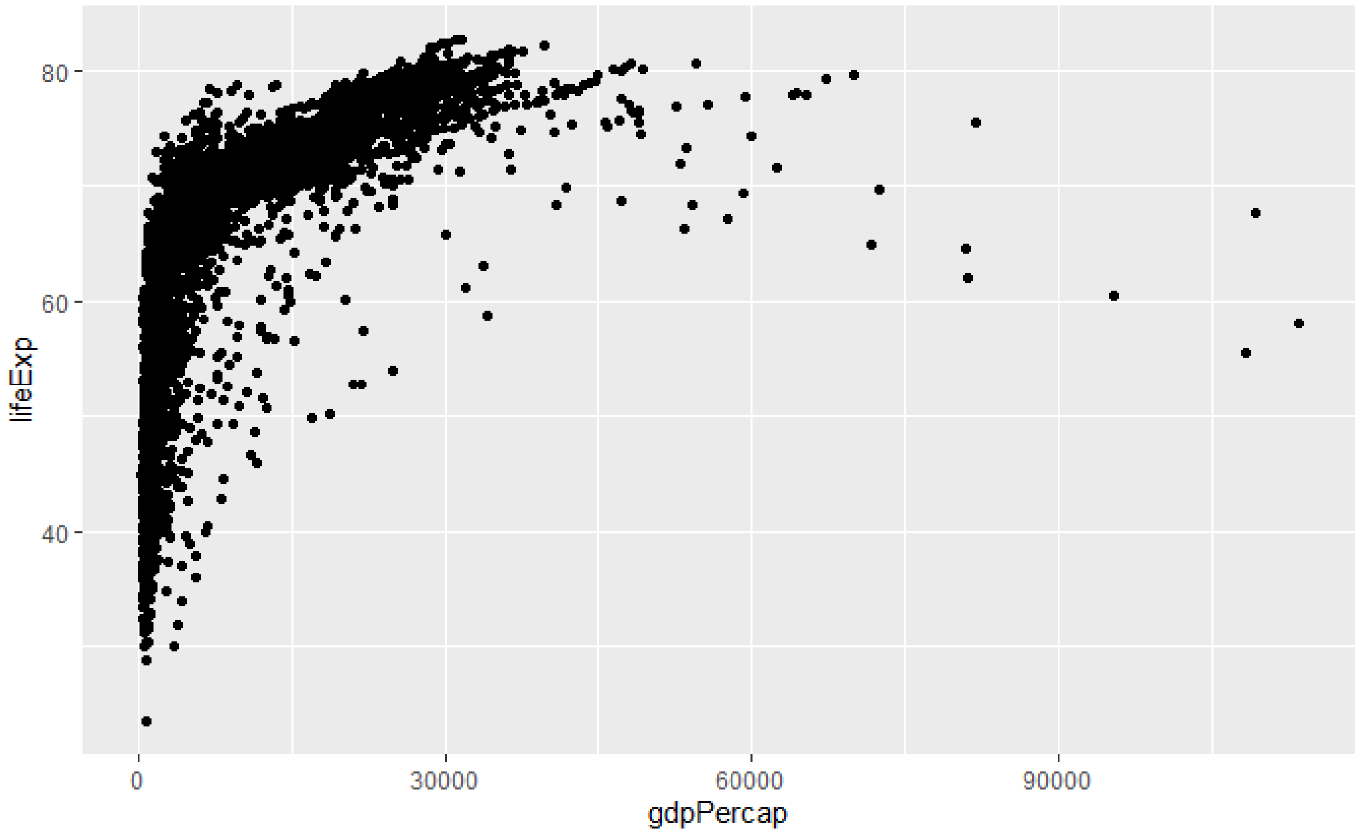
```
cor(select_if(gapminder_unfiltered, is.numeric))
```

```
                year      lifeExp          pop   gdpPercap
year      1.00000000  0.383616006  0.013368315  0.31440915
lifeExp   0.38361601  1.000000000 -0.006116394  0.63376069
pop       0.01336832 -0.006116394  1.000000000 -0.04595259
gdpPercap 0.31440915  0.633760687 -0.045952593  1.00000000
```

###Analysing Patterns Task 1: Creating a scatterplot for two variables

Hide

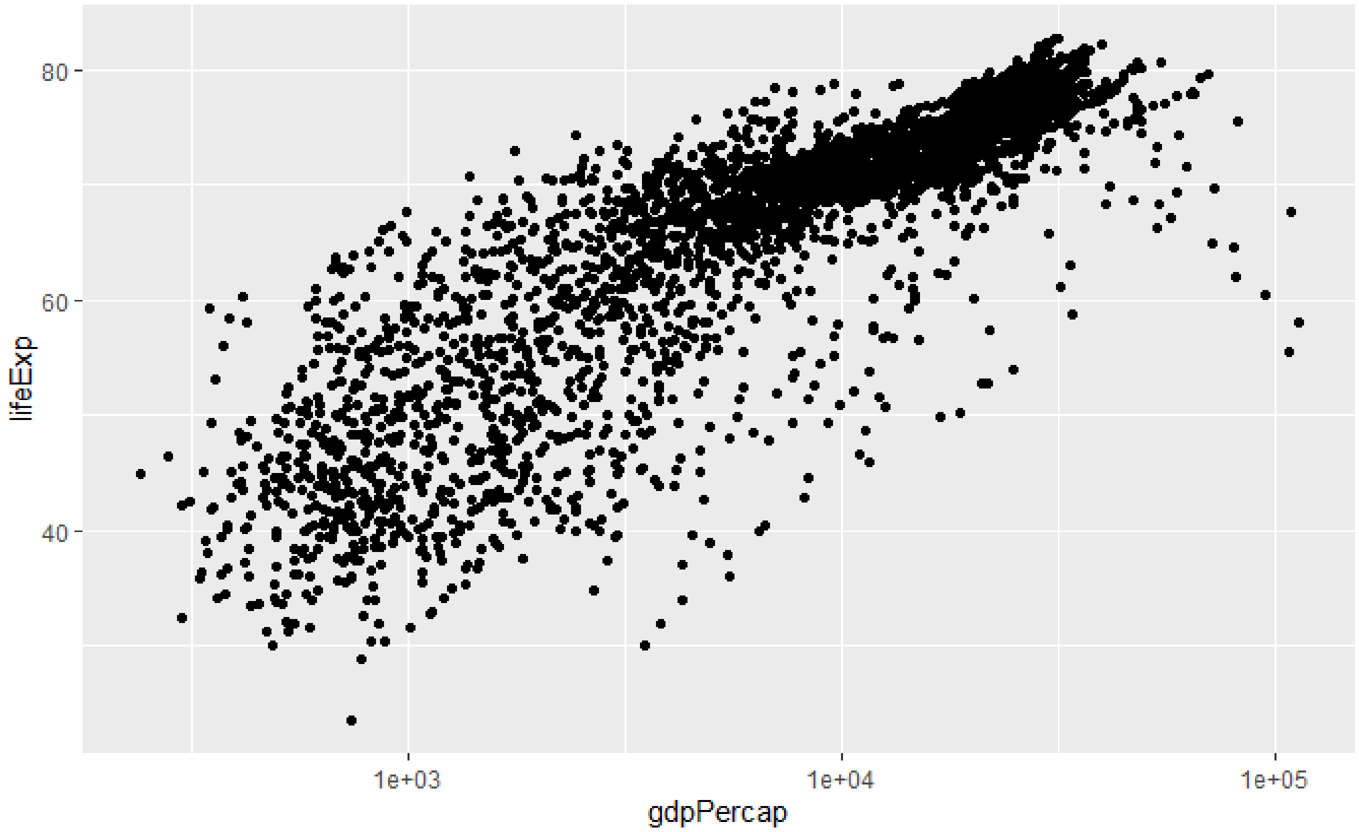
```
ggplot(gapminder_unfiltered, aes(gdpPercap, lifeExp)) +
  geom_point()
```



Task 2:

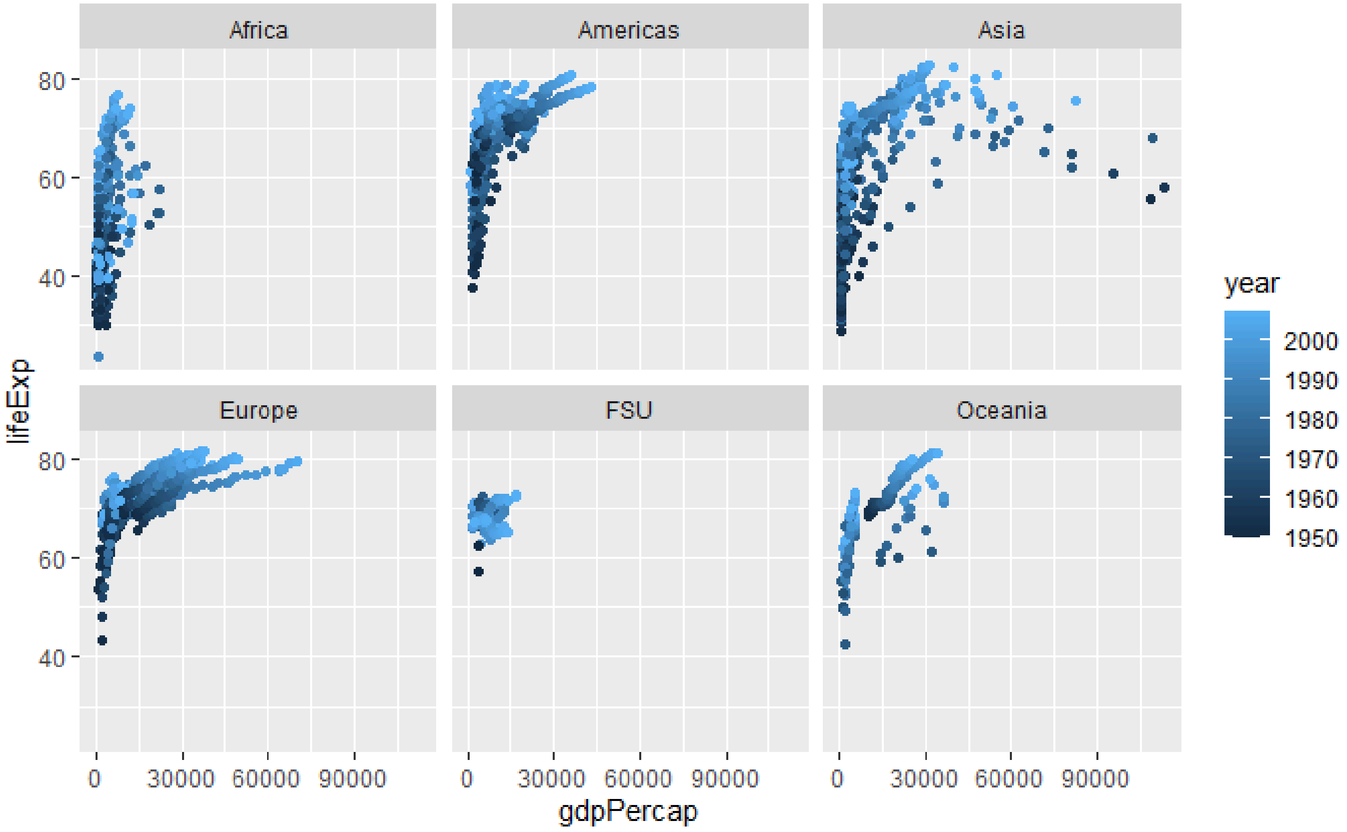
Hide

```
ggplot(gapminder_unfiltered, aes(gdpPercap, lifeExp)) +
  geom_point() +
  scale_x_log10()
```



Hide

```
ggplot(gapminder_unfiltered, aes(gdpPercap, lifeExp, color = year)) +
  geom_point() +
  facet_wrap(~ continent)
```


Africa
Americas
Asia
Europe
FSU
Oceania
lifeExp
gdpPercap
year