Lab 04 – `tidyr`

Problem Sets

Question 1

Part A: Convert `bluechips` to long format where each stock closing price is in single column

Part B: Cast the dataset from Part A back to wide format

Solution

```
bluechips <- read.csv("https://remiller1450.github.io/data/bluechips.csv")

## Part A
dat <- pivot_longer(bluechips, cols = !Year,
             names_to = "Company",
             values_to = "Price")
head(dat, n = 3)
```

```
## # A tibble: 3 x 3
##    Year Company Price
##   <int> <chr>   <dbl>
## 1  2010 AAPL     7.64
## 2  2010 KO      28.5
## 3  2010 JNJ     64.7
```

```
## Part B
dat <- pivot_wider(dat, id_cols = Year,
                   names_from = Company,
                   values_from = Price)
head(dat, n = 3)
```

```
## # A tibble: 3 x 5
##    Year  AAPL    KO   JNJ   AXP
##   <int> <dbl> <dbl> <dbl> <dbl>
## 1  2010  7.64  28.5  64.7  40.9
## 2  2011 11.8   32.6  62.8  43.4
## 3  2012 14.7   35.1  65.9  48.4
```

Question 2

```
polls <- read.csv("https://remiller1450.github.io/data/polls2016.csv")
```

Part A: Using `polls` data, pivot wide or long so that each row associated with single poll

Part B: Use separate to split into two columns, one for name, the other for party affiliation

Solution

```
## Part A
dat <- pivot_longer(polls, cols = contains(c("Clinton", "Trump", "Johnson", "Stein")))
dim(dat)
```

```
## [1] 28  6
```

```
## Part B
separate(dat, col = name, sep = "\\.\\.", into = c("Name", "Party")) %>% head()
```

```
## # A tibble: 6 x 7
##   Poll     Date        Sample   MoE Name    Party value
##   <chr>    <chr>       <chr>  <dbl> <chr>   <chr> <int>
## 1 Monmouth 7/14 - 7/16 688 LV   3.7 Clinton D.       45
## 2 Monmouth 7/14 - 7/16 688 LV   3.7 Trump   R.       43
## 3 Monmouth 7/14 - 7/16 688 LV   3.7 Johnson L.        5
## 4 Monmouth 7/14 - 7/16 688 LV   3.7 Stein   G.        1
## 5 CNN/ORC  7/13 - 7/16 872 RV   3.5 Clinton D.       42
## 6 CNN/ORC  7/13 - 7/16 872 RV   3.5 Trump   R.       37
```

Question 3

The "airlines" data set (loaded below) contains data used in the article Should Travelers Avoid Flying Airlines That Have Had Crashes in the Past? that appeared on fivethirtyeight.com.

```
airlines <- read.csv("https://raw.githubusercontent.com/ds4stats/r-tutorials/master/tidying-data/data/a
```

Do following:

1. Pivot so last 6 columns in new var called accident
2. Separate into type and year
3. Widen so that each type its own column
4. Create plot showing avail_seat_km, fatal accidents, and year

Solution

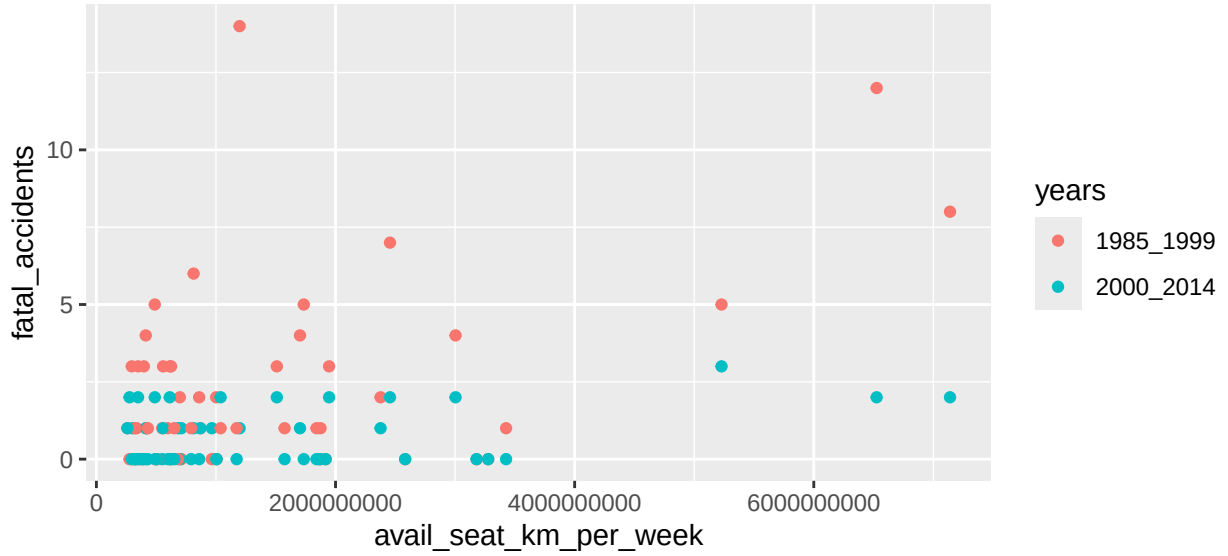
```
## Part A
tidy_airlines <- pivot_longer(airlines,
                              cols = !c(airline, avail_seat_km_per_week),
                              names_to = "accidents", values_to = "count")

## Part B
tidy_airlines <- separate(tidy_airlines, col = "accidents",
                          into = c("var", "years"), sep = "[.]")

## Part C
tidy_airlines <- pivot_wider(tidy_airlines,
                             id_cols = c(airline, avail_seat_km_per_week, years),
```

```
                            names_from = var, values_from = count)

ggplot(tidy_airlines, aes(x = avail_seat_km_per_week,
                          y = fatal_accidents, col = years)) +
  geom_point()
```



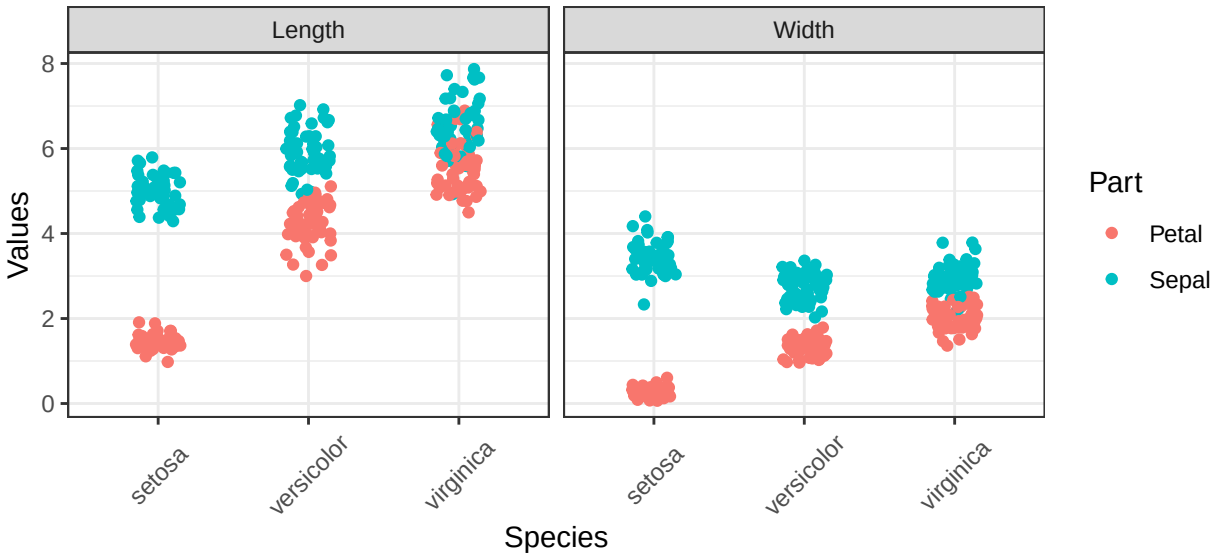
Question 4

Your goal is to recreate the following graphic using the `tidyr` functions covered in this lab from the `iris` dataset (code for creating plot shown below):

Solution

```
tt1 <- pivot_longer(iris, cols = !Species,
                    names_to = "Part", values_to = "Values")
tt2 <- separate(tt1, col = "Part", into = c("Part", "Measurement"), sep = "\\.")

library(ggplot2)
ggplot(tt2, aes(Species, Values, color = Part)) +
  geom_jitter(width = 0.15) +
  facet_wrap(~Measurement) + theme_bw() +
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5))
```


Length
Width
Values
8
6
4
2
0
setosa
versicolor
virginica
Species
Part
Petal
Sepal