Quiz 01: Data wrangling

Aicha Sidiya

3/15/2023

Packages

```
library(tidyverse)
```

```
## Warning: package 'tibble' was built under R version 4.2.3
```

Data

```
lego <- read_csv("data/lego.csv")
```

Exercise 1

```
lego <- filter(lego, !is.na(retail_price), !is.na(pieces), pieces != 0)
```

Removing these sets shortned the dataframe but made it more expressive # Exercise 2

```
lego2 <- arrange(lego, desc(retail_price))
lego2[1:3, ] #print first 3 rows
```

```
## # A tibble: 3 x 10
##   id        name      theme~1 theme subth~2  year pieces minif~3 package retai~4
##   <chr>     <chr>     <chr>   <chr> <chr>   <dbl>  <dbl>   <dbl> <chr>     <dbl>
## 1 75192-1   Millenni~ Licens~ Star~ Ultima~  2017   7541       8 Box        800.
## 2 2000431-1 Connecti~ Educat~ Seri~ <NA>     2013   2455       0 <NA>       755.
## 3 75159-1   Death St~ Licens~ Star~ Ultima~  2016   4016      27 Box        500.
## # ... with abbreviated variable names 1: themegroup, 2: subtheme, 3: minifigs,
## #   4: retail_price
```

The three most expensive sets are Millennium Falcon which costs 800$ and has 7541 pieces, the second most expensive set is Connections Kit which costs 755$ and has 2455 pices, the third most expensive set is Death Star whcih costs 500$ and contains 4016 pieces.

Exercise 3

```
lego <- mutate(lego, price_per_piece=retail_price/pieces)
```

Exercise 4

```
lego4 <- arrange(lego, desc(price_per_piece))
select(lego4[1:5, ], name, themegroup, theme, pieces, price_per_piece)
```

```
## # A tibble: 5 x 5
##   name                          themegroup  theme      pieces price_per_piece
##   <chr>                         <chr>       <chr>       <dbl>           <dbl>
## 1 EV3 Intelligent Brick         Technical   Mindstorms      1           205.
## 2 Intelligent NXT Brick (Black) Technical   Mindstorms      1           170.
## 3 NXT Intelligent Brick         Technical   Mindstorms      1           170.
## 4 RCX Programmable LEGO Brick   Technical   Mindstorms      1           110
## 5 NXT DC Rechargeable Battery   Educational Education       1            80.0
```

The sets with the highest prices are the sets consisting of 1 piece only.

Exercise 5

```
lego5 <- filter(lego, theme == "The Lord of the Rings")
lego6 <- group_by(lego5, subtheme)
summarize(lego6, cheap_lego = min(retail_price),
          expensive_lego = max(retail_price))
```

```
## # A tibble: 3 x 3
##   subtheme                   cheap_lego expensive_lego
##   <chr>                           <dbl>          <dbl>
## 1 The Fellowship of the Ring       0              80.0
## 2 The Return of the King          20.0           100.
## 3 The Two Towers                   4             200.
```

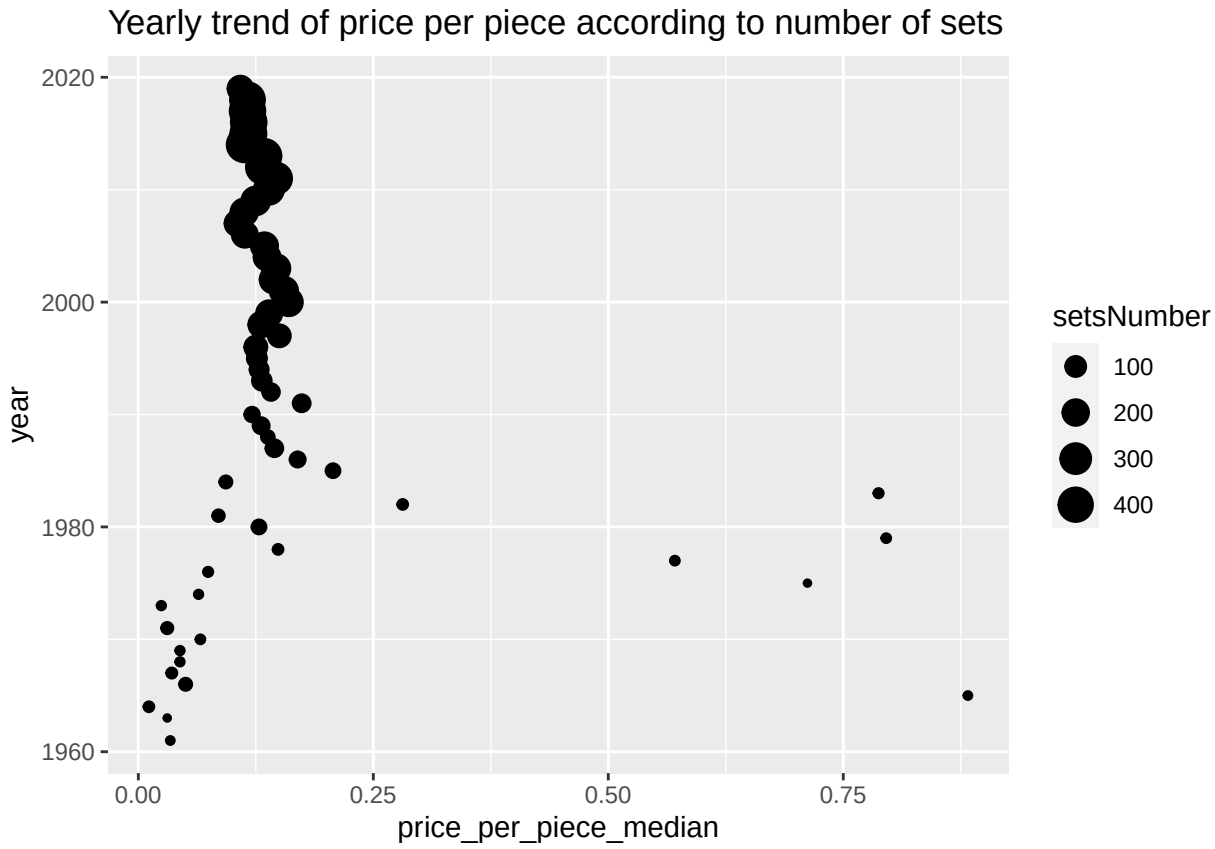
Exercise 6

```
lego7 <- group_by(lego, year)
yearly_trends <- summarize(lego7, setsNumber = n(), price_per_piece_median = median(price_per_piece))
```

Exercise 7

```
library(ggplot2)
ggplot(data=yearly_trends, mapping=aes(x=price_per_piece_median, y=year, size = setsNumber)) + geom_poi
```

```
## Warning: Removed 1 rows containing missing values (‘geom_point()‘).
```



The sisze of the sets increse over the years.