Data Analysis and Lab_Lab 3

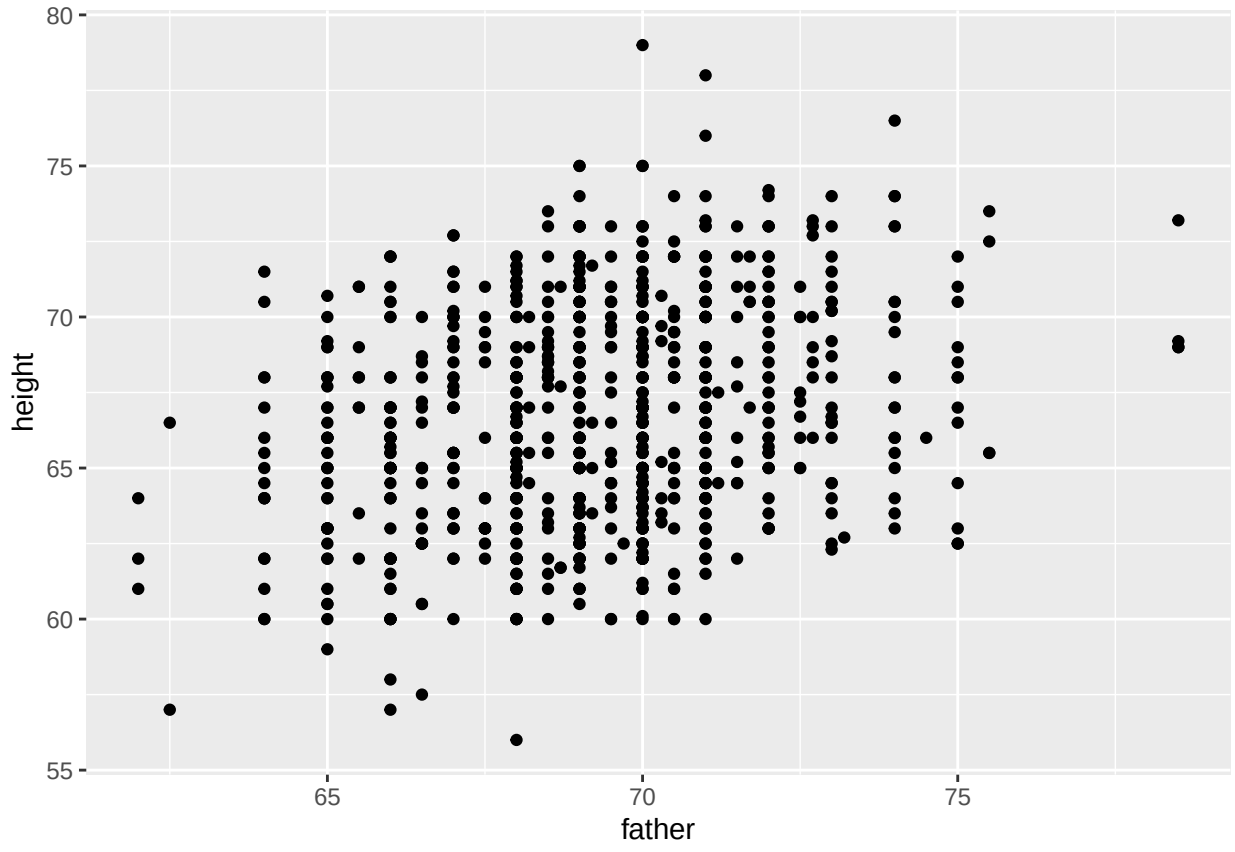
Na SeungChan

2023-09-24

Exercise 1

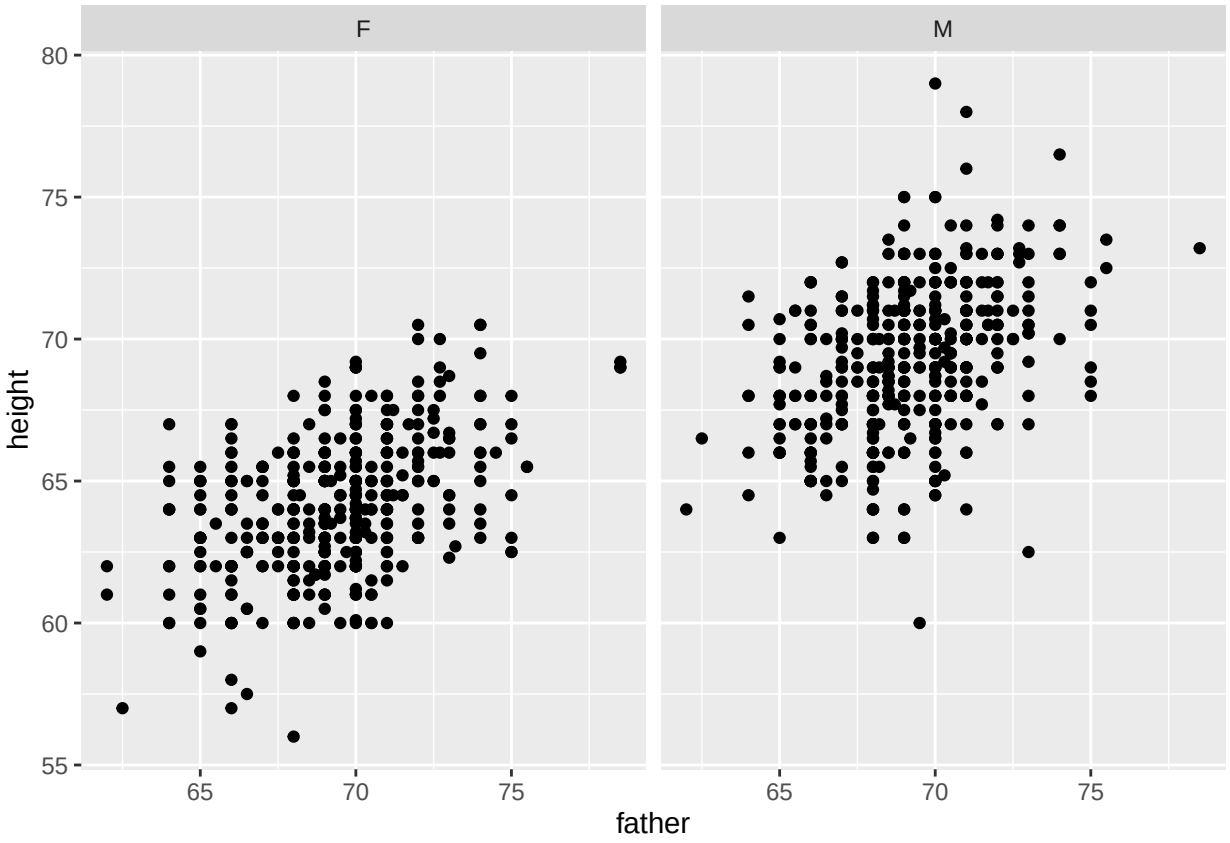
(a) Create a scatter plot of each person's height against their father's height

```
ggplot(data = Galton) +
  geom_point(mapping = aes(x = father, y = height))
```



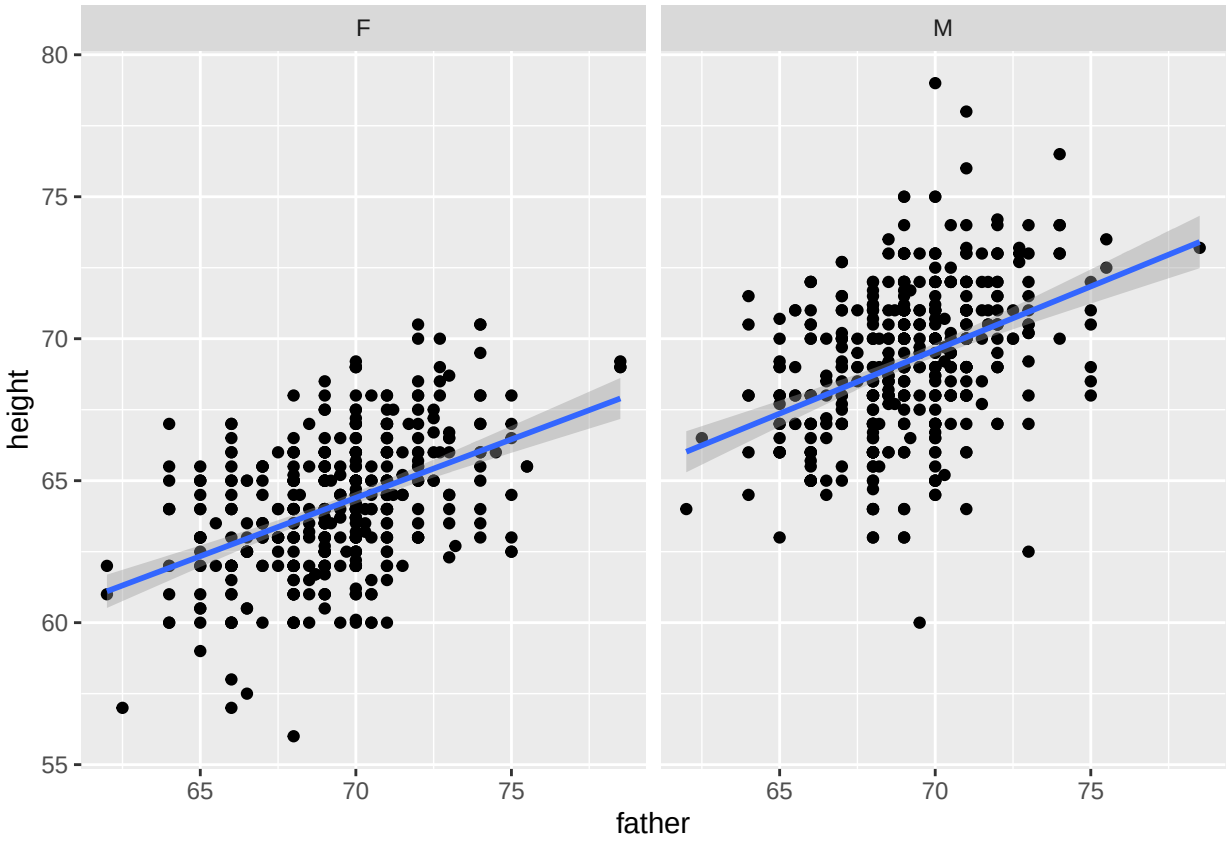
(b) Separate your plot into facets by sex

```
ggplot(data = Galton) +
  geom_point(mapping = aes(x = father, y = height)) +
  facet_wrap(~sex)
```



(c) Add regression lines to all of your facets; use **`geom_smooth(method = "lm")`**.

```
ggplot(data = Galton) +
  geom_point(mapping = aes(x = father, y = height)) +
  facet_wrap(~sex) +
  geom_smooth(mapping = aes(x = father, y = height), method = "lm", formula = y ~ x)
```



(d) Analyze the results of (a), (b) and (c) in 2 ~ 3 sentences.

산점도상 자녀의 키와 아버지의 키는 그래프로 봤을 때 양의 상관관계를 가지는 것으로 보인다. 이와 같은 결과는 '남성과 여성의 키 차이를 고려'하여 자녀의 성별에 따라 서로 다른 facet에 각 그래프를 그려도 동일하게 나타난다. 실제 이에 따라 회귀직선을 mapping해 본 경우에도 비슷한 결과가 나온다.

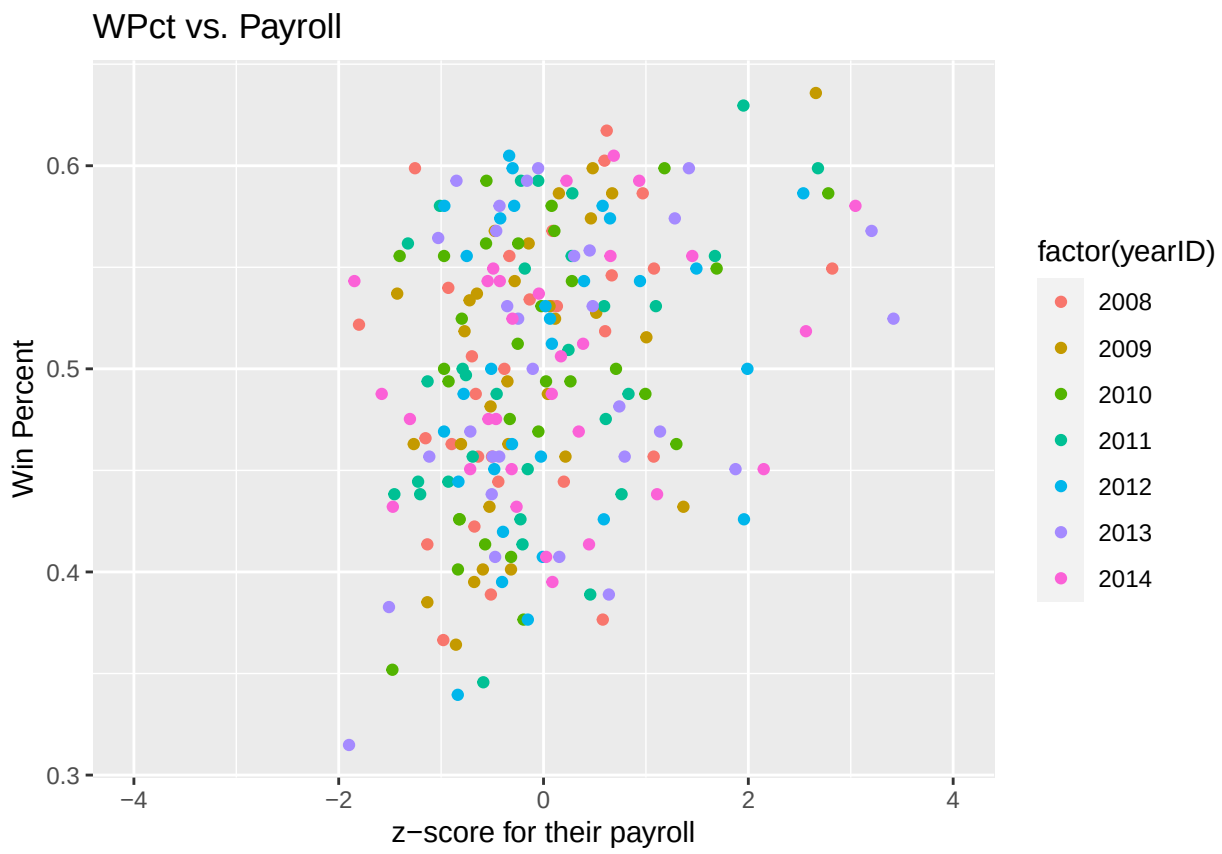
Exercise 2

```
MLB_teams
```

(a) See how many variables you can illustrate on a single plot in R. The current record is 7. [Note: this is not good graphical practice—it is merely an exercise to help you understand how to use visual cues and aesthetics!]

```
MLB_teams %>%
  mutate(z_pay = (payroll-mean(payroll))/sd(payroll)) %>%
  ggplot() +
  geom_point(mapping = aes(x = z_pay, y = WPct, color = factor(yearID))) +
  xlim(-4,4) +
```

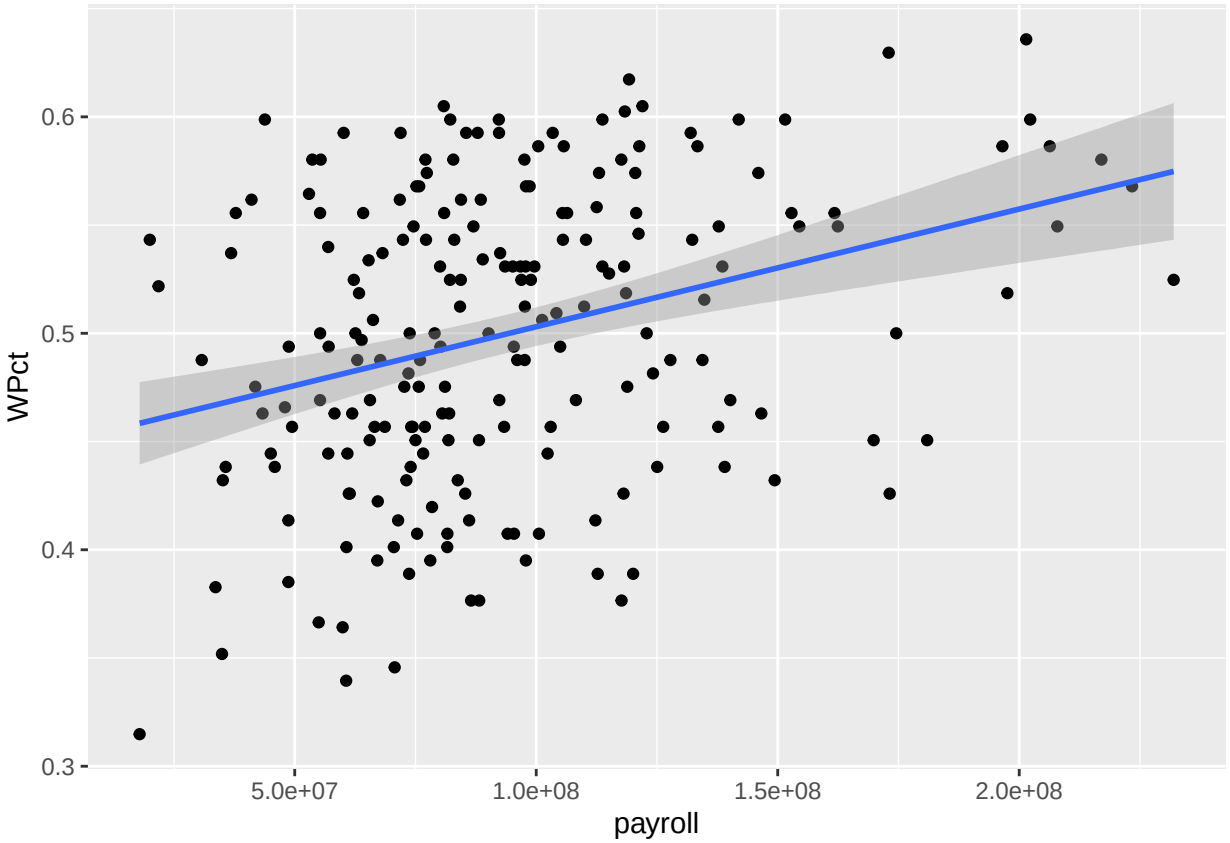
```
  xlab('z-score for their payroll') +
  ylab('Win Percent') +
  labs(title = 'WPct vs. Payroll')
```



(b) Use the **MLB_teams** data in the **mdsr** package to create an informative data graphic that illustrates the relationship between winning percentage **WPct** and payroll **payroll** in context.

```
ggplot(data = MLB_teams, mapping = aes(x = payroll, y = WPct)) +
  geom_point() +
  geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Overall there is a positive correlation between payroll and WPct. However, overall payroll tends to increase as time goes.

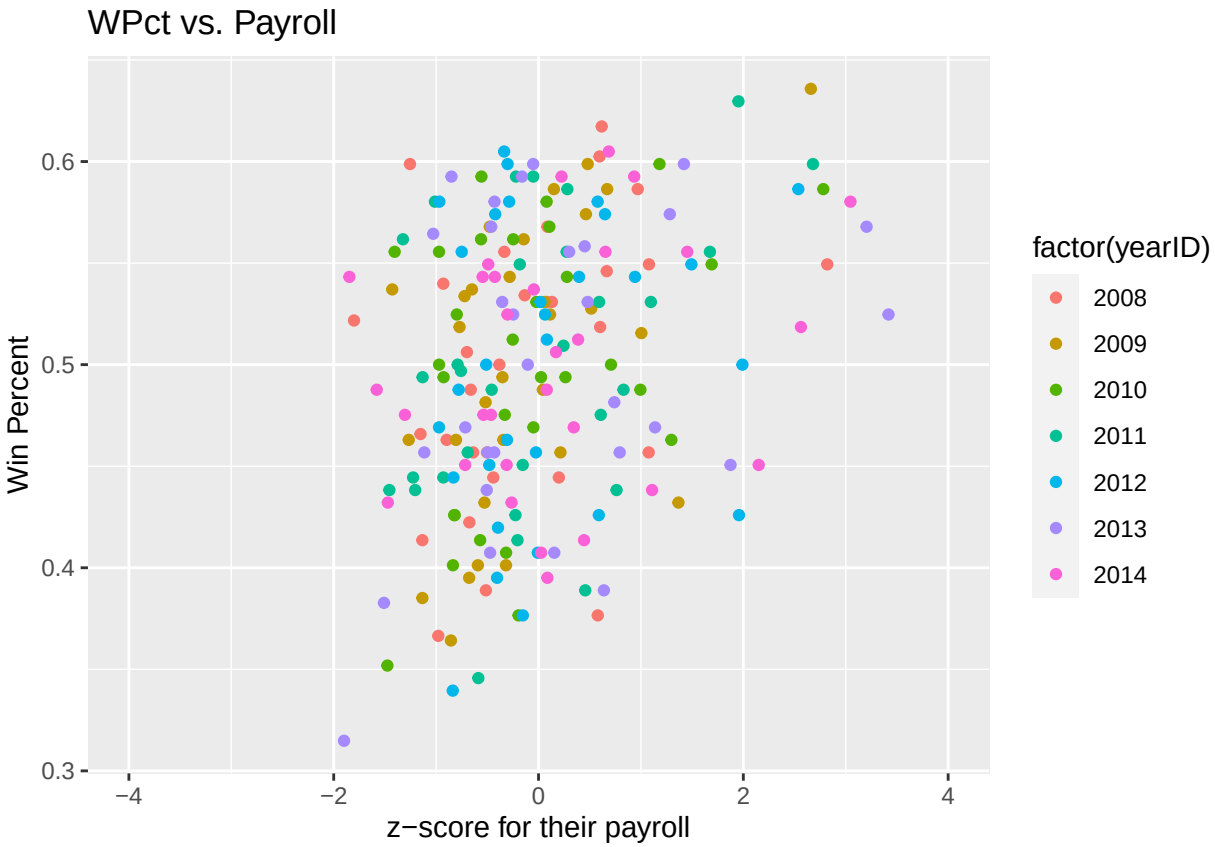
```
MLB_teams %>%
  group_by(yearID) %>%
  summarise(sum(payroll))
```

```
## # A tibble: 7 x 2
##   yearID `sum(payroll)`
##    <int>          <dbl>
## 1   2008     2684858670
## 2   2009     2664726994
## 3   2010     2721359865
## 4   2011     2784505291
## 5   2012     2932741192
## 6   2013     3034525648
## 7   2014     2994000466
```

So, we will use z-score. In z-scores, we can see teams below -2.0 did not exist, but teams with z-scores over 3.0 was 3 teams.

```
MLB_teams %>%
  mutate(z_pay = (payroll-mean(payroll))/sd(payroll)) %>%
  ggplot() +
  geom_point(mapping = aes(x = z_pay, y = WPct, color = factor(yearID))) +
```

```
  xlim(-4,4) +
  xlab('z-score for their payroll') +
  ylab('Win Percent') +
  labs(title = 'WPct vs. Payroll')
```



Exercises 3 : Write a code to create a data object named **`Binary_medv`** whose value is **`"rich"`** if the value of **`medv`** is greater than 25, **`"not so"`** if not. Use both **`with()`** and **`ifelse()`**.

```
my_Boston <- Boston %>%
  mutate(Binary_medv = with(Boston, ifelse(medv > 25, 'rich', 'not so'))) %>%
  relocate(Binary_medv)
head(my_Boston)
```

```
##   Binary_medv    crim zn indus chas   nox    rm  age    dis rad tax ptratio
## 1      not so 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3
## 2      not so 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8
## 3        rich 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8
## 4        rich 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7
## 5        rich 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7
## 6        rich 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7
```

```
##    black lstat medv
## 1 396.90  4.98 24.0
## 2 396.90  9.14 21.6
## 3 392.83  4.03 34.7
## 4 394.63  2.94 33.4
## 5 396.90  5.33 36.2
## 6 394.12  5.21 28.7
```