705 HW1

2023-08-25

1.

```
library(magrittr)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.3.1
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(plotly)
```

```
## Warning: package 'plotly' was built under R version 4.3.1
```

```
##
## Attaching package: 'plotly'
```

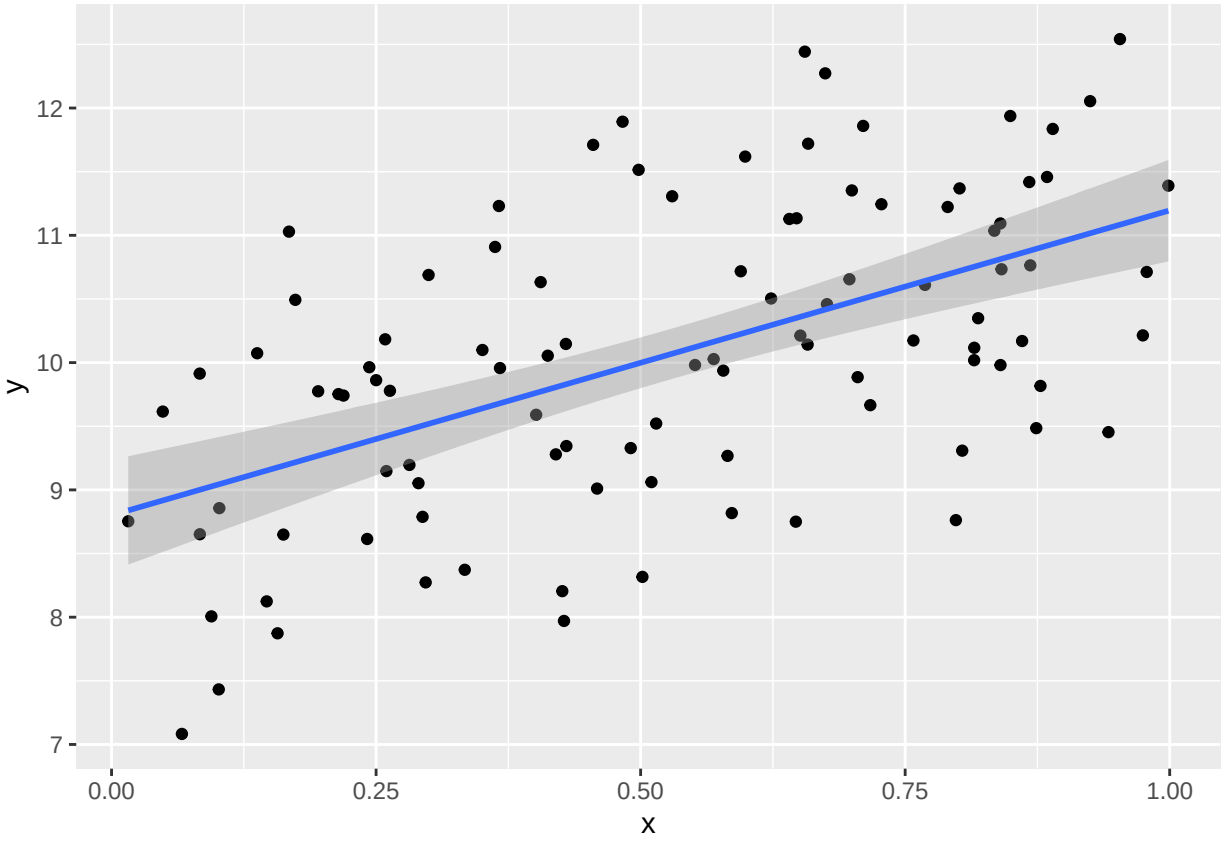
```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
set.seed(12)
xDF1 <- runif(100)  #how to simulate data before $>%
runif(100) %>% data.frame(y = 2*. + rnorm(100) + 9 , x=.)%>% ggplot(aes(x, y)) + geom_point() + geom_sm
```

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



2.

(1)

```
NOBEL   <- read.table(url("https://raw.githubusercontent.com/dsy109/Supplemental/main/Courses/705/nobel
```

The output of the structure of the Nobel Prize:

```
typeof(NOBEL)
```

```
## [1] "list"
```

Here the four variables has the following:

```
class(NOBEL$prize_year)
```

```
## [1] "integer"
```

```
class(NOBEL$category)
```

```
## [1] "character"
```

```r
typeof(NOBEL$gender)
```

```
## [1] "character"
```

```r
typeof(NOBEL$age)
```

```
## [1] "double"
```

(2) For the missing age:

```r
NOBEL$category <- as.factor(NOBEL$category)
```

```r
table(NOBEL$category[which(is.na(NOBEL$age))])
```

```
##
##  Chemistry  Economics Literature   Medicine      Peace    Physics
##          1          1          0          0         27          1
```

For the missing gender:

```r
table(NOBEL$category[which(is.na(NOBEL$gender))])
```

```
##
##  Chemistry  Economics Literature   Medicine      Peace    Physics
##          0          0          0          0         26          0
```

The Peace has the most missing values for each of these variables.

(3)

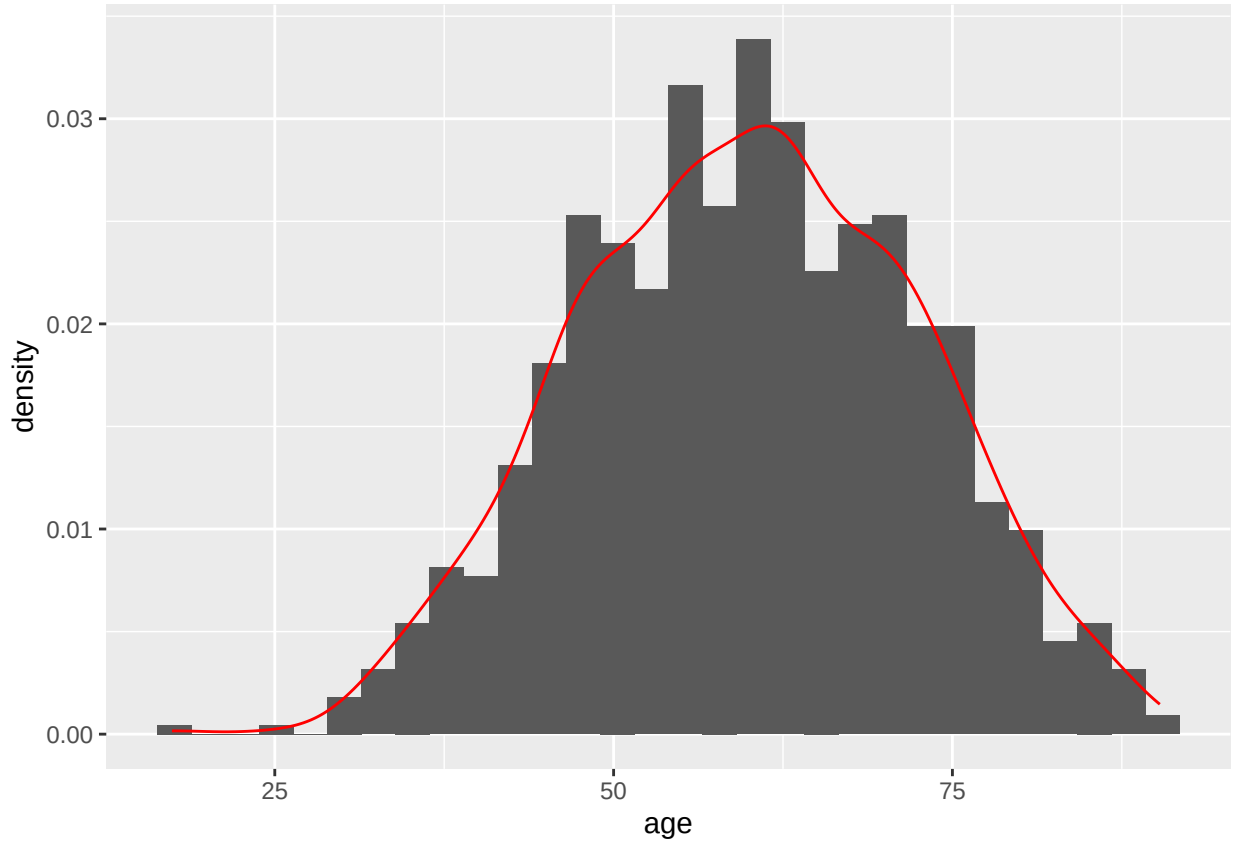
```r
ggplot(NOBEL, aes(x=age)) + geom_histogram(aes(y=..density..)) +   geom_density(color="red")
```

```
## Warning: The dot-dot notation (‘..density..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(density)‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

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

```
## Warning: Removed 30 rows containing non-finite values (‘stat_bin()‘).
```

```
## Warning: Removed 30 rows containing non-finite values (‘stat_density()‘).
```



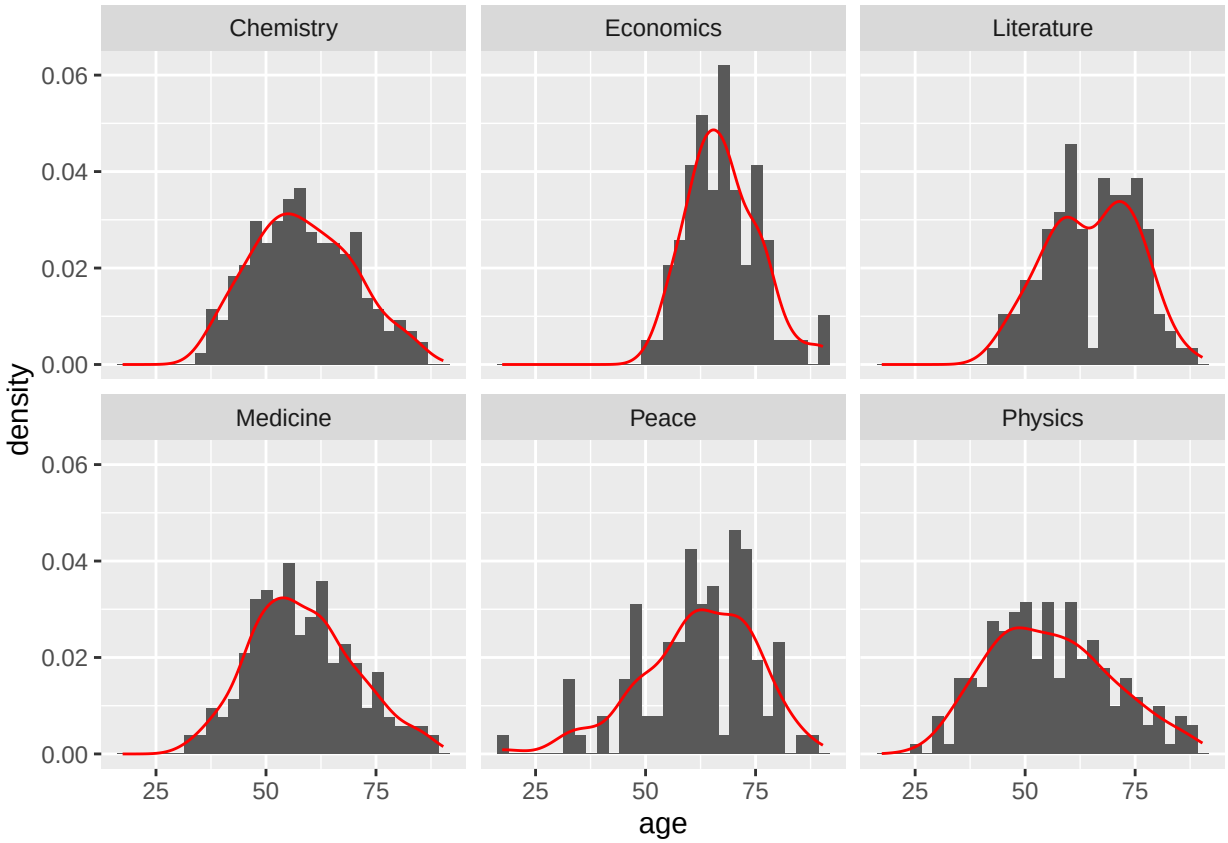
I think the distribution of ages is a like a normal distribution with mean around 58. Although has a little bit left skewed, but basically show a normal distribution.

```
ggplot(NOBEL, aes(x=age)) +
  geom_histogram(aes(y=..density..)) +
   facet_wrap(vars(category)) +
  geom_density(color="red")
```

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

```
## Warning: Removed 30 rows containing non-finite values (`stat_bin()`).
```

```
## Warning: Removed 30 rows containing non-finite values (`stat_density()`).
```



Here I noticed that the winner's age of the Chemistry, Medicine and Physics shows a slightly right skewed normal distribution which centered around 50. For the winner's age of Economics, although it also shows a slightly right skewed normal distribution, its mean is larger than the previous three categories, with a larger variance. The distribution of winner's age of the peace, shows a left skewed distribution which centered around 55. The winner's age of Literature shows a double peak distribution which centered at 58.

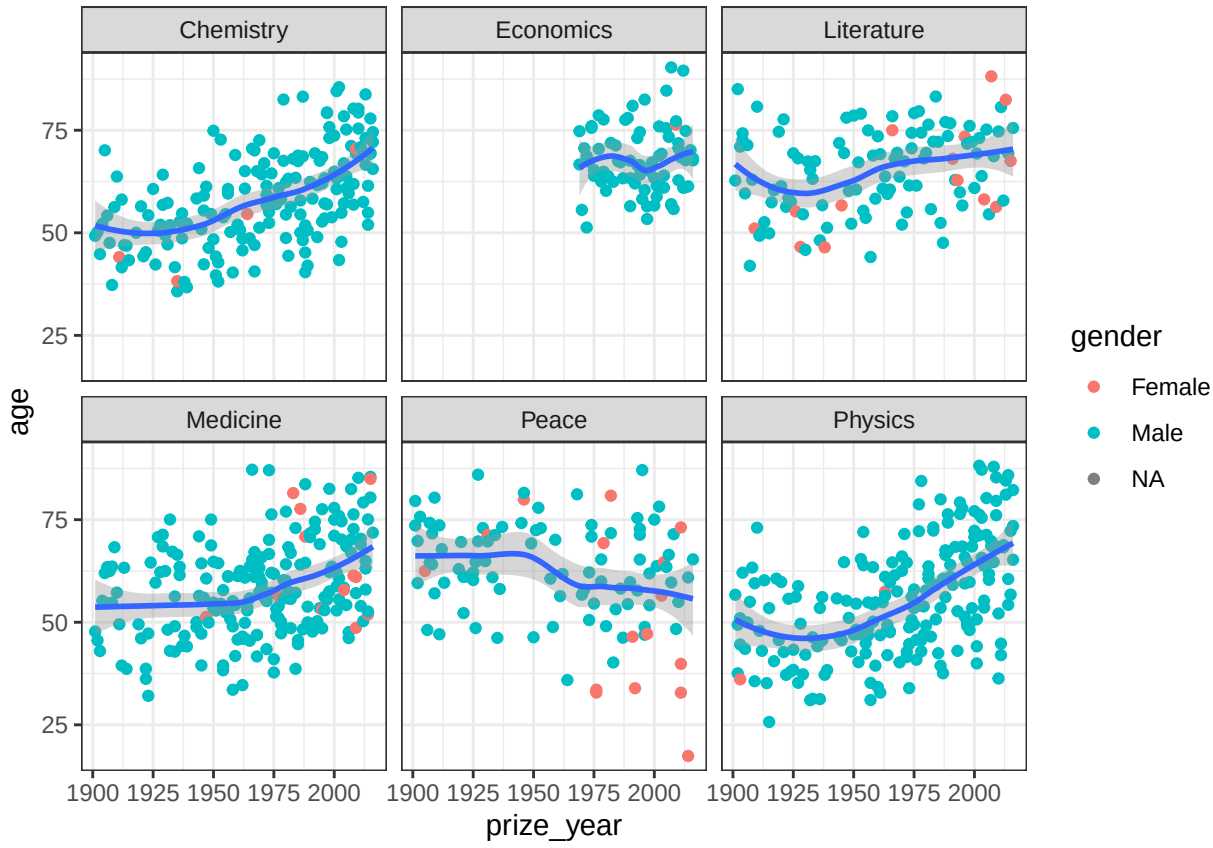
(4)

```
base_size <- 11
ggplot(NOBEL, aes(x=prize_year,y =age  )) +
  geom_point(aes(colour=gender)) +
   facet_wrap(vars(category)) +
   geom_smooth(method = loess)+
   theme_bw(base_size = 11,
  base_line_size = base_size/22,
  base_rect_size = base_size/22
)
```

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

```
## Warning: Removed 30 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 30 rows containing missing values (`geom_point()`).
```



From the gender part, I notice that the men were the majority of the winners.

For the Chemistry, Medicine, Physics and Literature nobel prize, with the passage of time, the winner's age is increasing. On the other hand, the winner of peace's age is decreasing. Economic prize winner's age shows fluctuate among different years.

3.

```
COVID   <- read.csv(url("https://raw.githubusercontent.com/dsy109/Supplemental/main/Courses/705/COVID.t
```

(1)

```
first.digit  <- function(x) {
                a <- trunc(log10(x))+1

                return(trunc(x/10^(a-1)))

}
```

(2)

```
power_32 <- 2^seq(1:1000)
mytable <- table(sapply(power_32, function(x) {
  first.digit(x)}))
```

```
mytable/1000
```

```
##
##     1     2     3     4     5     6     7     8     9
## 0.301 0.176 0.125 0.097 0.079 0.069 0.056 0.052 0.045
```

(3)

```
power_33 <- 3^seq(1:500)
mytable3 <- table(sapply(power_33, function(x) {
  first.digit(x)}))
```

```
mytable3/500
```

```
##
##     1     2     3     4     5     6     7     8     9
## 0.300 0.176 0.124 0.098 0.080 0.066 0.058 0.052 0.046
```

(4)

```
my_covid <- COVID$Deaths
```

```
mytable4 <- table(sapply(my_covid, function(x) {
  first.digit(x)}))
```

```
mytable4/192
```

```
##
##          1          2          3          4          5          6          7
## 0.38541667 0.14583333 0.13541667 0.07291667 0.04687500 0.07291667 0.05208333
##          8          9
## 0.05208333 0.03645833
```

(5)

```
my_df5 <- data.frame(
  Leading_Digit = 1:9,
  two_power = mytable/1000,
  three_power = mytable3/500,
  Prop_COVID = mytable4/192
)
```

```
my_df5$two_power.Var1 <- NULL

my_df5$three_power.Var1 <- NULL
my_df5$Prop_COVID.Var1 <- NULL
```

```
my_df5$three_power.Freq
```

```
## [1] 0.300 0.176 0.124 0.098 0.080 0.066 0.058 0.052 0.046
```

```
summary(lm(Prop_COVID.Freq~two_power.Freq + three_power.Freq ,data = my_df5))
```

```
##
## Call:
## lm(formula = Prop_COVID.Freq ~ two_power.Freq + three_power.Freq,
##     data = my_df5)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.047534 -0.014814  0.001206  0.015904  0.028100
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.02974    0.01653  -1.799    0.122
## two_power.Freq    6.76054    6.76793   0.999    0.356
## three_power.Freq -5.49290    6.80445  -0.807    0.450
##
## Residual standard error: 0.02751 on 6 degrees of freedom
## Multiple R-squared:  0.953,  Adjusted R-squared:  0.9373
## F-statistic: 60.84 on 2 and 6 DF,  p-value: 0.0001038
```

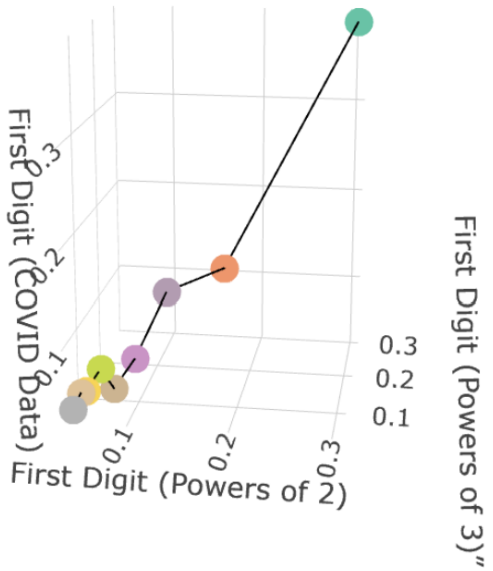
Here the R^2 is 0.953. The adjusted R^2 is 0.9373.

(5)

```
p <- plot_ly(
  x = my_df5$two_power.Freq,
  y = my_df5$three_power.Freq ,
  z = my_df5$Prop_COVID.Freq,
  type = 'scatter3d' ,
  mode='lines',
  line = list(color = 'black')
)  %>% add_markers( data = my_df5,  color =  ~factor( Leading_Digit), mode = 'markers' ) %>%
  layout( title ='Benford's Law Demonstration',scene = list(xaxis = list(title = 'First Digit (Powers o
     yaxis = list(title = 'First Digit (Powers of 3)"'),  zaxis = list(title = 'First Digit (COVID Data
                                                                  ))
```

```
knitr::include_graphics("Screenshot 2023-08-30 154216.png")
```

Benford's Law Demonstration



4.

```
library(HoRM)
```

```
## Warning: package 'HoRM' was built under R version 4.3.1
```

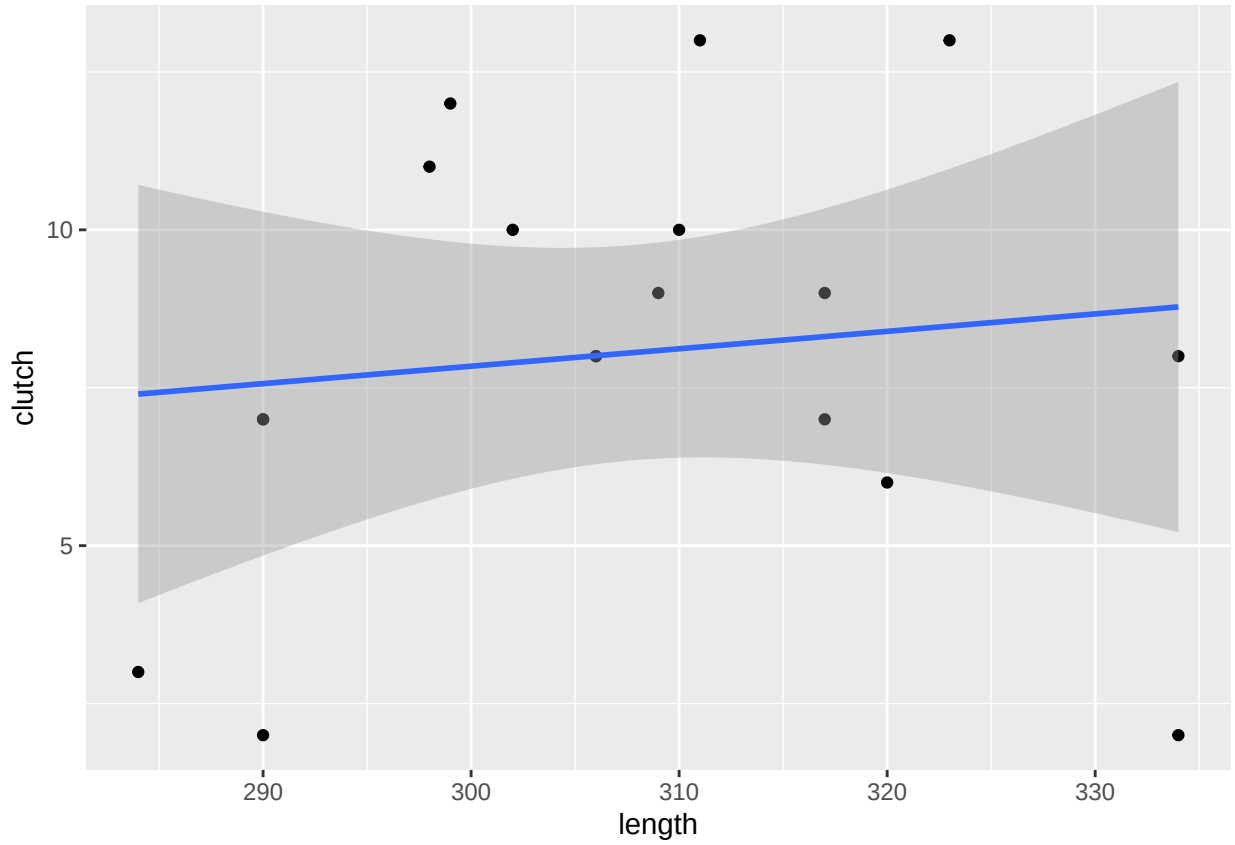
```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
data("tortoise")
```

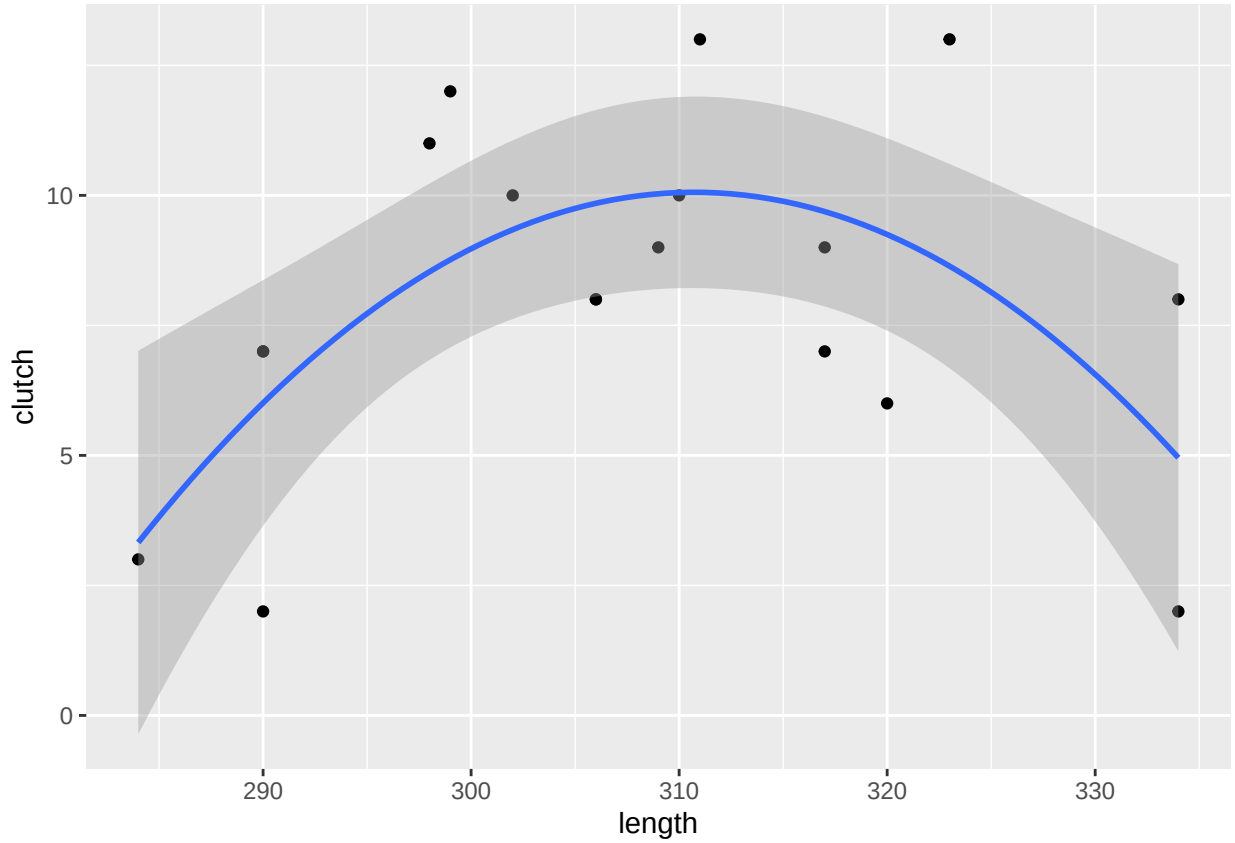
Linear regression line:

```
ggplot(tortoise,aes(length,clutch)) +
  geom_point() +
  geom_smooth(method='lm')
```

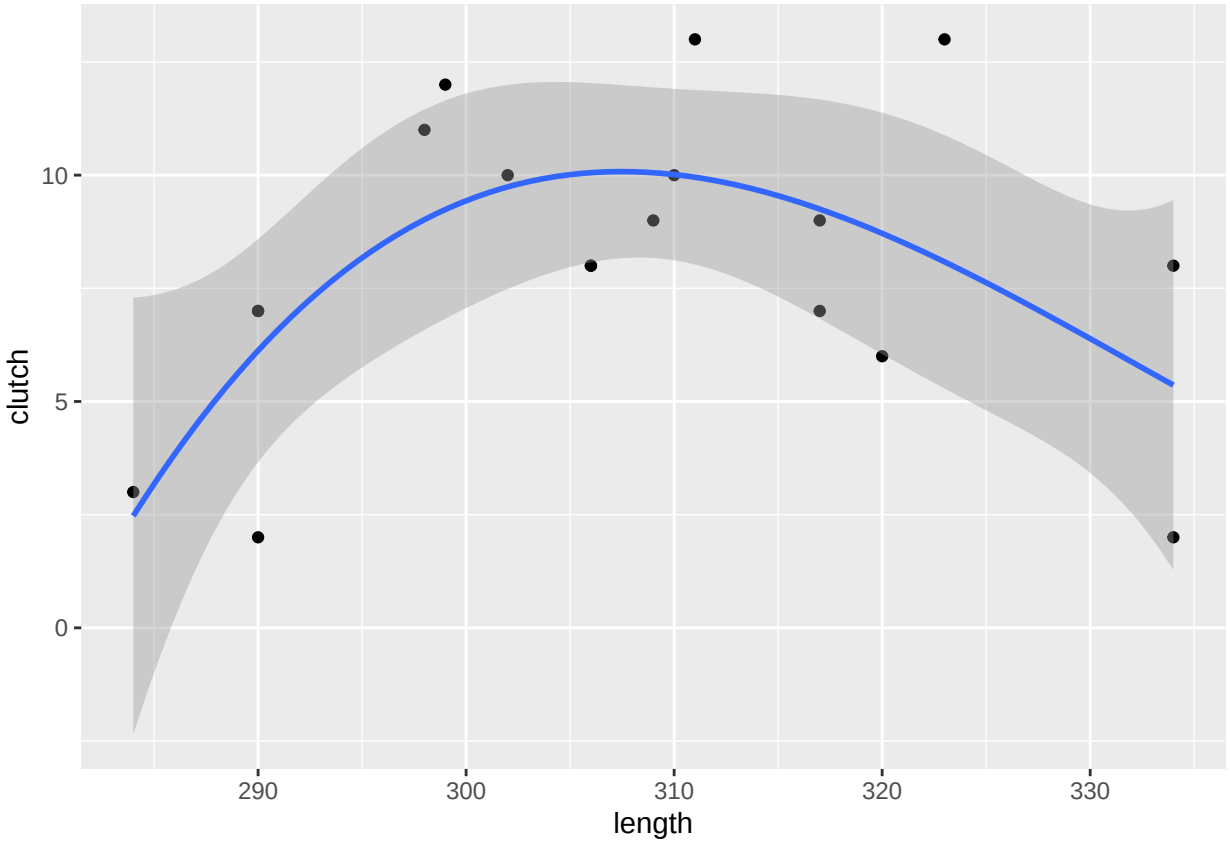
```
## `geom_smooth()` using formula = 'y ~ x'
```




```
ggplot(tortoise,aes(length,clutch)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x + I(x^2))
```




```
ggplot(tortoise,aes(length,clutch)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x + I(x^2)+I(x^3))
```



Here I get that the above 3 regression. Here I noticed that the curve model is better than the linear model. However, I did not see a lot of difference between the quadratic and cubic model.

(b) ANOVA:

```
my_1 <-  lm(clutch ~  length,data = tortoise)
my_2 <-  lm(clutch ~ length + I(length^2),data = tortoise)
my_3 <-  lm(clutch ~ length + I(length^2)+I(length^3),data = tortoise)
```

```
anova(my_1,my_2)
```

```
## Analysis of Variance Table
##
## Model 1: clutch ~ length
## Model 2: clutch ~ length + I(length^2)
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     16 186.15
## 2     15 106.97  1    79.178 11.102 0.00455 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(my_3,my_2)
```

```
## Analysis of Variance Table
##
```

```
## Model 1: clutch ~ length + I(length^2) + I(length^3)
## Model 2: clutch ~ length + I(length^2)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     14 104.18
## 2     15 106.97 -1    -2.797 0.3759 0.5496
```

```
anova(my_1,my_3)
```

```
## Analysis of Variance Table
##
## Model 1: clutch ~ length
## Model 2: clutch ~ length + I(length^2) + I(length^3)
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     16 186.15
## 2     14 104.18  2    81.975 5.5082 0.01719 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Stepwise variable selection using AIC:

```
step(my_3)
```

```
## Start:  AIC=39.6
## clutch ~ length + I(length^2) + I(length^3)
##
##               Df Sum of Sq    RSS    AIC
## - I(length^3)  1    2.7970 106.97 38.080
## - I(length^2)  1    3.2099 107.39 38.149
## - length       1    3.6537 107.83 38.223
## <none>                     104.18 39.603
##
## Step:  AIC=38.08
## clutch ~ length + I(length^2)
##
##               Df Sum of Sq    RSS    AIC
## <none>                     106.97 38.080
## - I(length^2)  1    79.178 186.15 46.051
## - length       1    79.879 186.85 46.119

##
## Call:
## lm(formula = clutch ~ length + I(length^2), data = tortoise)
##
## Coefficients:
## (Intercept)       length  I(length^2)
##  -8.999e+02    5.857e+00   -9.425e-03
```

Calculating the BIC

```
step(my_3,k=log(length(tortoise$length)))
```

```
## Start:  AIC=43.16
## clutch ~ length + I(length^2) + I(length^3)
##
##                Df Sum of Sq    RSS    AIC
## - I(length^3)   1    2.7970 106.97 40.751
## - I(length^2)   1    3.2099 107.39 40.820
## - length        1    3.6537 107.83 40.894
## <none>                      104.18 43.164
##
## Step:  AIC=40.75
## clutch ~ length + I(length^2)
##
##                Df Sum of Sq    RSS    AIC
## <none>                      106.97 40.751
## - I(length^2)   1    79.178 186.15 47.832
## - length        1    79.879 186.85 47.900

##
## Call:
## lm(formula = clutch ~ length + I(length^2), data = tortoise)
##
## Coefficients:
## (Intercept)       length  I(length^2)
##  -8.999e+02    5.857e+00   -9.425e-03
```

Here I choose 3 different ways to select the best model. Anova ,AIC and BIC. For the Anova, it agree with your assessment in the previous part. However, the AIC and BIC model agree with curve model is better than the linear model, but the AIC and BIC procudure said that the quadratic model is enough and cubic term is required.