2023 01 05 VB-STA5 Exam in Statistics

Thursday 5th of January.

The exam set consists of 3 main exercises with 9 sub exercises in total.

Each sub exercise is weighted equally when grading the hand-ins.

1. Gymnastics and figure skating.

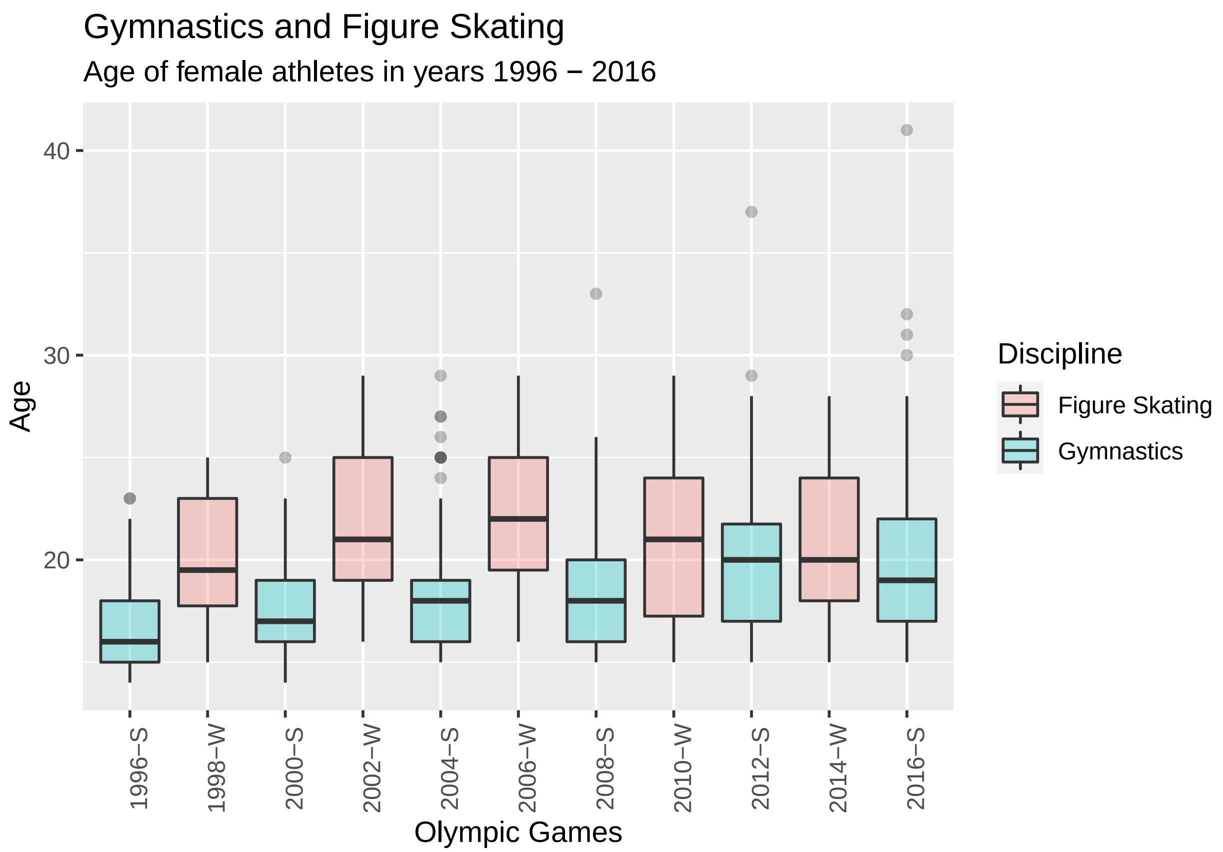
Dataset *data/gym_figskate.csv* contains information about Olympic athletes in Figure Skating and Gymnastics in years 1964 to 2016.

a) Recreate the plot.

```
sports <- readr::read_csv("data/gym_figskate.csv")
```

```
## Rows: 3143 Columns: 11
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (7): Name, Sex, Team, Games, Season, City, Sport
## dbl (4): Age, Height, Weight, Year
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
sports %>% filter(Year > 1994, Sex == 'F') %>%
  ggplot() +
  geom_boxplot(aes(Games, Age, fill = Sport), alpha = 0.3) +
  labs(subtitle = 'Age of female athletes in years 1996 - 2016',
       title = 'Gymnastics and Figure Skating',
       x = 'Olympic Games',
       y = 'Age',
       fill = 'Discipline') +
  theme(axis.text.x = element_text(angle = 90))
```



b) Describe the plot.

The plot shows Female Olympians in Gymnastics and Figure Skating in years 1996 to 2016. In the years before 2010 the age of Gymnast was lower than the figure skaters. There seems to be one athlete competing in subsequent Olympics in gymnastics up to age of 42. Only in a couple of cases athletes careers extend over age of 30. Athletes younger than 18 are quite prolific.

c) Is there significant difference in between average height of male figure skaters competing in 1972 Sapporo Olympics and 2002 in Salt Lake City. Conduct a suitable statistical test.

Difference of means t-test.

$H_0 : \mu_{m\_sapporo} - \mu_{m\_saltlakecity} = 0$

$H_A : \mu_{m\_sapporo} - \mu_{m\_saltlakecity} = 0 \neq 0$

H0: There is no difference between mean height for male figure scaters in Sapporo and Salt Lake City Olympics.

HA: There is a difference between mean height for male figure scaters in Sapporo and Salt Lake City Olympics.

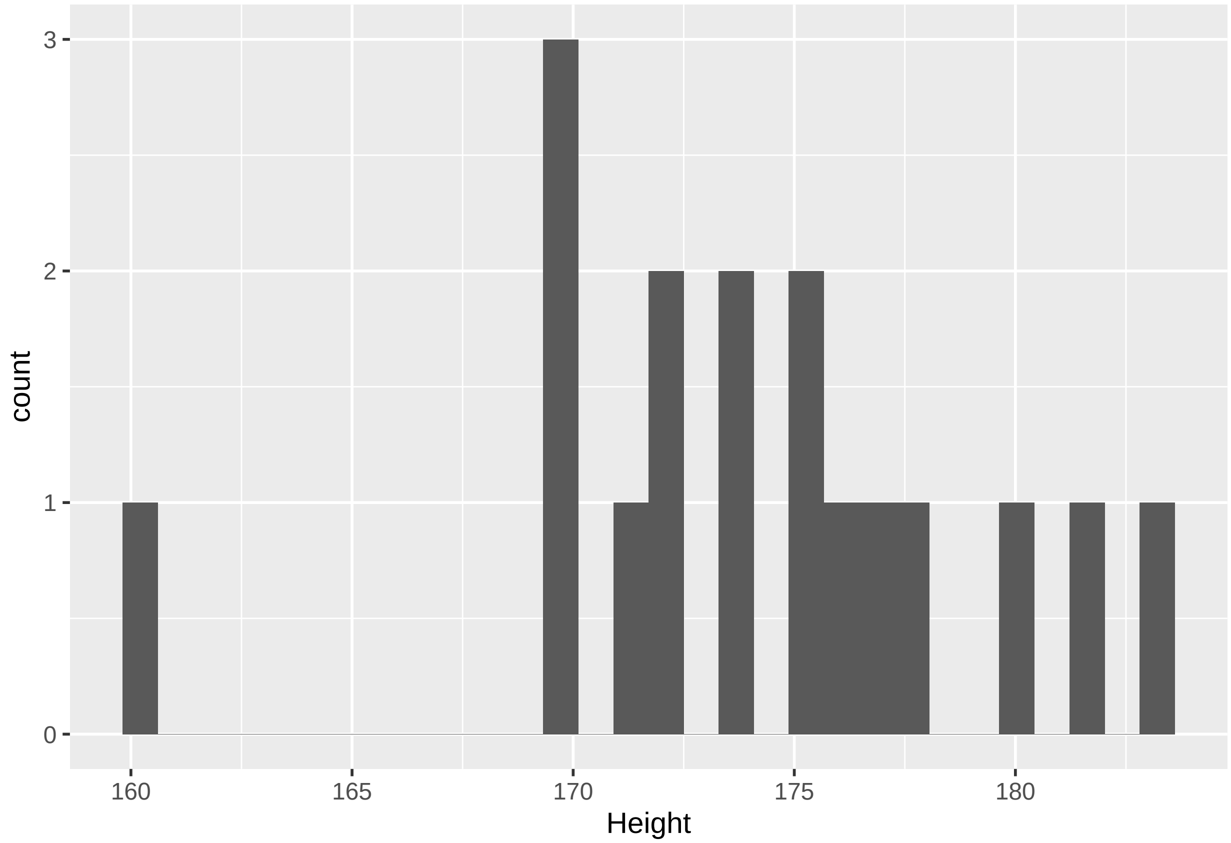
alpha significance level - 0.05

Conditions check:

Normality:

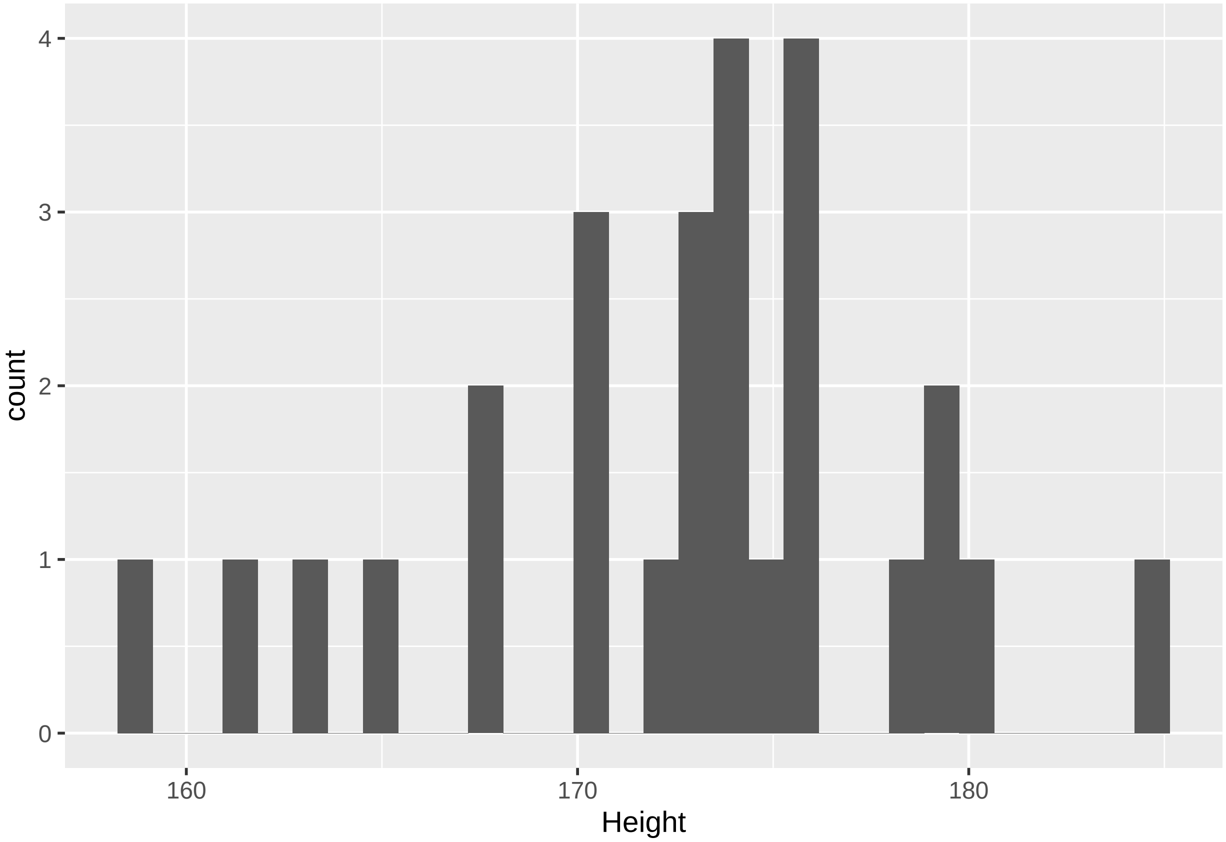
```
sports %>% filter(Sex == 'M') %>%
  filter(City %in% c('Sapporo')) %>%
  ggplot() +
  geom_histogram(aes(x = Height))
```

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



```
sports %>% filter(Sex == 'M') %>%
  filter(City %in% c('Salt Lake City')) %>%
  ggplot() +
  geom_histogram(aes(x = Height))
```

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



Hard to say with samples so small, but it seems that most variables follow normal distribution.

We assume that observations are independent.

- short version

```
sports %>% filter(Sex == 'M') %>%
  filter(City %in% c('Sapporo', 'Salt Lake City')) %>%
  t.test(Height~City, data = .)
```

```
##
##  Welch Two Sample t-test
##
## data:  Height by City
## t = -0.81993, df = 36.357, p-value = 0.4176
## alternative hypothesis: true difference in means between group Salt Lake City and group Sapporo is no
## 95 percent confidence interval:
##  -4.963116  2.104728
## sample estimates:
## mean in group Salt Lake City        mean in group Sapporo
##                     172.6296                     174.0588
```

p-value is bigger than alpha significance level, thus we accept null hypothesis and reject the alternative. There is no statistically significant difference between height of malle figure skaters in Sapporo and Salt Lake City olympics.

- long version

```r
m_s <- sports %>% filter(Sex == 'M') %>%
  filter(City %in% c('Sapporo'))
m_slc <- sports %>% filter(Sex == 'M') %>%
  filter(City %in% c('Salt Lake City'))

(point_estimate <- mean(m_s$Height) - mean(m_slc$Height))
```

```
## [1] 1.429194
```

```r
(nrow(m_s))
```

```
## [1] 17
```

```r
(nrow(m_slc))
```

```
## [1] 27
```

```r
dof <- 16
```

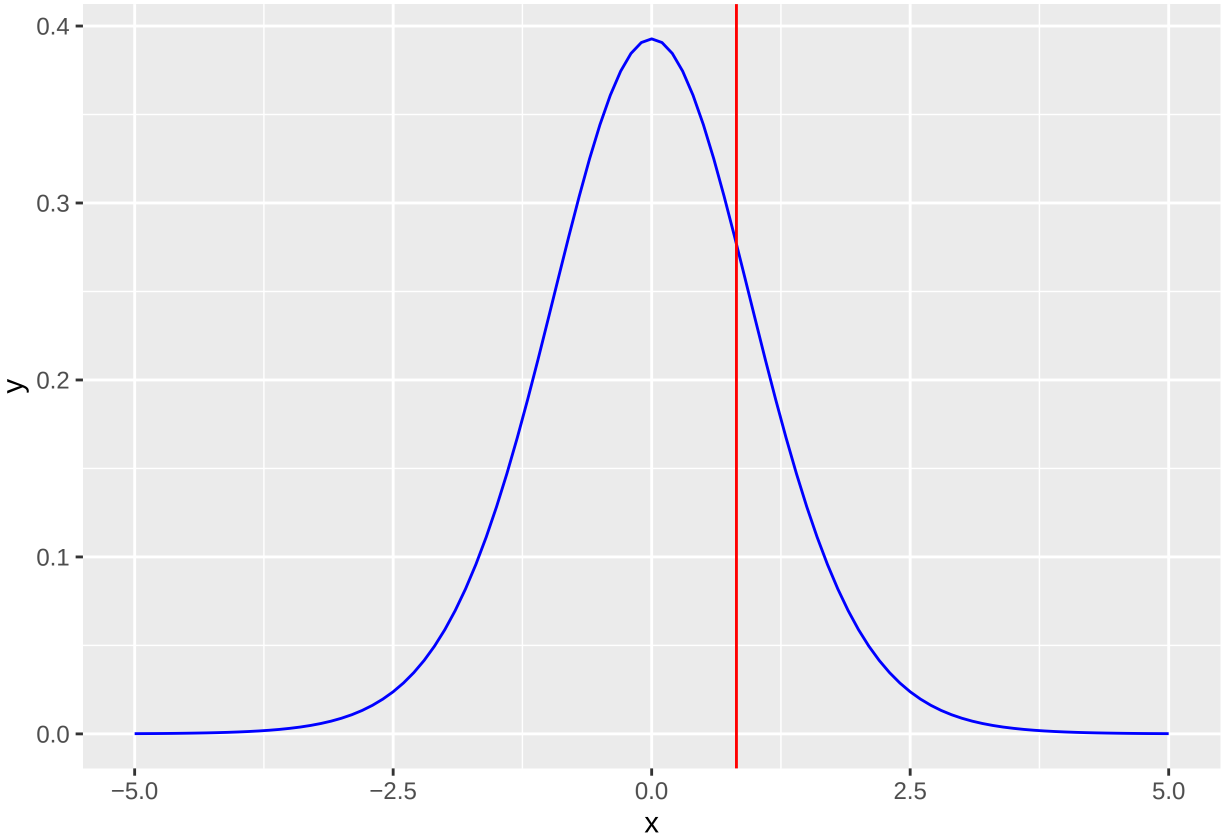
```r
(SE <- sqrt((sd(m_s$Height)^2/nrow(m_s)) + (sd(m_slc$Height)^2/nrow(m_slc))))
```

```
## [1] 1.743078
```

```r
(t_score <- (point_estimate - 0)/SE)
```

```
## [1] 0.8199253
```

```r
ggplot(data.frame(x = seq(-5, 5, length=100)), aes(x = x)) +
  stat_function(fun = dt, args = list(df = dof), color = 'blue') +
  geom_vline(aes(xintercept = t_score),  color = 'red')
```



```
(p_value <- 2 * (1- pt(t_score, df = dof)))
```

```
## [1] 0.4243054
```

p-value is bigger than alpha significance level, thus we accept null hypothesis and reject the alternative. There is no statistically significant difference between height of malle figure skaters in Sapporo and Salt Lake City olympics.

2. Candles Market

Dataset *candles_revenue.csv* contains information about candle market revenue around the world in Euros. Dataset *population_2020.csv* contains information about world population in 2020.

a) Join the two datasets.

```
candles <- readr::read_csv('data/candles_revenue.csv')
```

```
## Rows: 149 Columns: 15
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): Country
## dbl (14): 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, ...
```

```
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
population <- readr::read_csv('data/population_2020.csv')
```

```
## Rows: 188 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): Country Name, Country Code, Continent
## dbl (1): Pop. 2020
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
candles <- candles %>% left_join(population, by = c('Country' = 'Country Name'))
```

b) Calculate revenue per capita in year 2020. Present in descending order five countries with highest revenue per capita in Europe in format presented below.

```
candles %>%
  select(Country, `Pop. 2020`, Continent, `2020`) %>%
  mutate(`Revenue per capita in EUR` = `2020`/`Pop. 2020`) %>%
  filter(Continent == 'Europe') %>%
  select(1,5) %>%
  top_n(5, `Revenue per capita in EUR`) %>%
  arrange(desc(`Revenue per capita in EUR`)) %>% knitr::kable()
```

Country	Revenue per capita in EUR
Luxembourg	10.231291
Norway	8.218274
Switzerland	7.655825
Ireland	7.507511
Denmark	5.367489

c) In 2015 Yankee Candle company conducted a survey of random candle users in selected Nordics (Denmark, Finland, Iceland, Norway, and Sweden). The purpose of the survey was to evaluate customer needs. They survey encompassed 1000 people.

Country	n
Denmark	256
Finland	179
Iceland	98
Norway	193
Sweden	274

According to the Revenue in those countries in 2015, how many people should have been surveyed in each country? Is the survey distribution following the revenue distribution for those countries.

Chi square test for goodness of fit.

H0: Distribution of surveyed people within different nordic countries follows distribution of distribution of revenue of nordic countries.

HA: Distribution of surveyed people within different nordic countries doesn't follow distribution of distribution of revenue of nordic countries.

alpha significance level - 0.05

Conditions check:

- we assume that the dataset is independent
- expected cases should be more than 5

```
nordics_revenue <- candles %>%
  filter(Country %in% c('Denmark', 'Finland', 'Iceland', 'Norway', 'Sweden')) %>%
  select('Country', '2015') %>% right_join(survey)
```

```
## Joining, by = "Country"
```

```
(nordics_all <- sum(nordics_revenue$`2015`))
```

```
## [1] 145810000
```

```
(nordics_revenue <- nordics_revenue %>%
    mutate(prc = `2015`/nordics_all) %>%
    mutate(expected = prc * 1000))
```

```
## # A tibble: 5 x 5
##   Country   `2015`     n    prc expected
##   <chr>      <dbl> <dbl>  <dbl>    <dbl>
## 1 Denmark 30290000   256 0.208     208.
## 2 Finland 23470000   179 0.161     161.
## 3 Iceland  1750000    98 0.0120     12.0
## 4 Norway  39750000   193 0.273     273.
## 5 Sweden  50550000   274 0.347     347.
```

All expected values are above 5.

- short version

```
chisq.test(nordics_revenue$n, p=nordics_revenue$prc)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  nordics_revenue$n
## X-squared = 667.93, df = 4, p-value < 2.2e-16
```

We reject null hypothesis in favour of alternative. Distribution of surveyed customers within nordic countries is not the same as distribution of revenue.

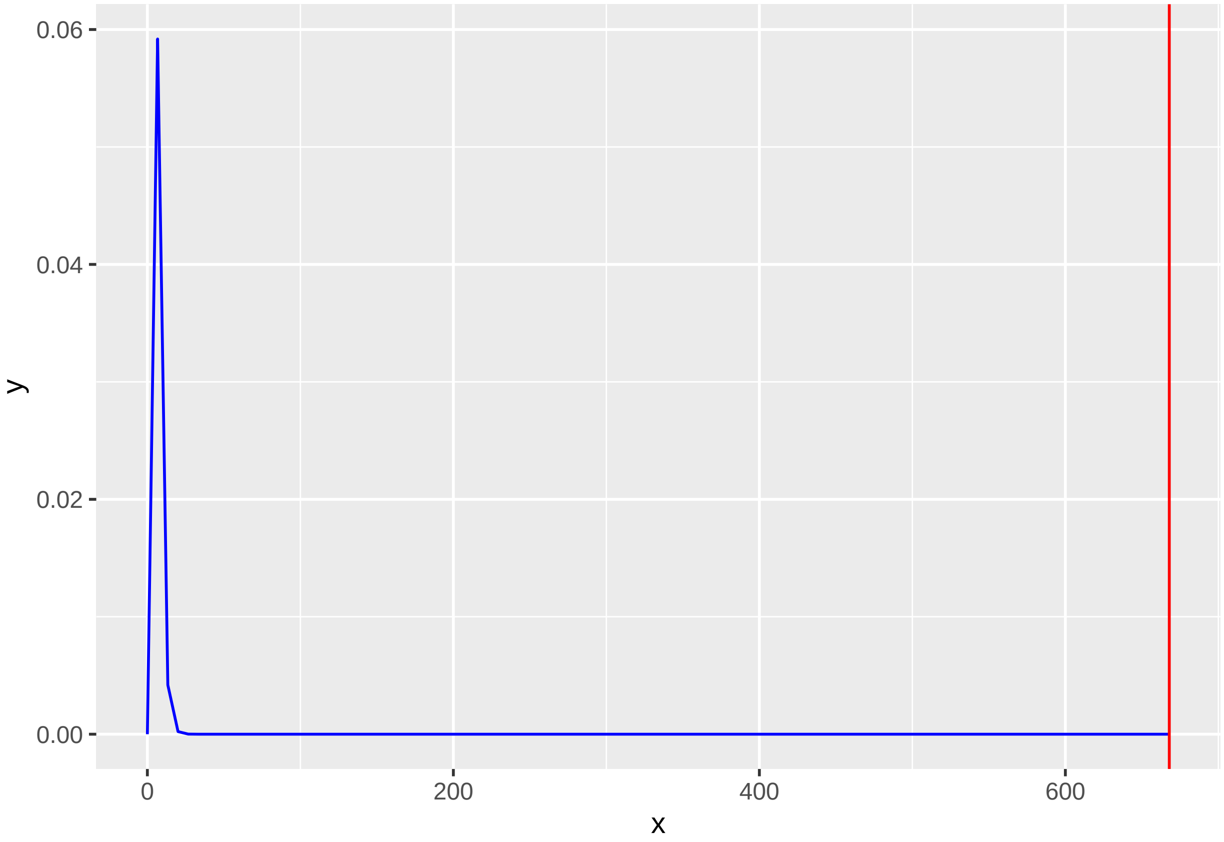
- long version

```
(chi2_stat <- sum(((nordics_revenue$n - nordics_revenue$expected)^2)/nordics_revenue$expected))
```

```
## [1] 667.9312
```

```
dof <- 4
```

```
ggplot(data.frame(x = seq(0, 100, length=100)), aes(x = x)) +
  stat_function(fun = dchisq, args = list(df = dof), color = 'blue') +
  geom_vline(aes(xintercept = chi2_stat),  color = 'red')
```



```
(p_value <- 1 - pchisq(chi2_stat, df = dof))
```

```
## [1] 0
```

We reject null hypothesis in favour of alternative. Distribution of surveyed customers within nordic countries is not the same as distribution of revenue.

3. Students expenses.

Dataset *UniversityStudentsMonthlyExpenses.csv* contains information about monthly expenses of randomly sampled students in U.S.A. in the 2000s.

a) Create a multiple regression model to predict students Monthly Expenses and tune it.

```
students <- readr::read_csv('data/UniversityStudentsMonthlyExpenses.csv')
```

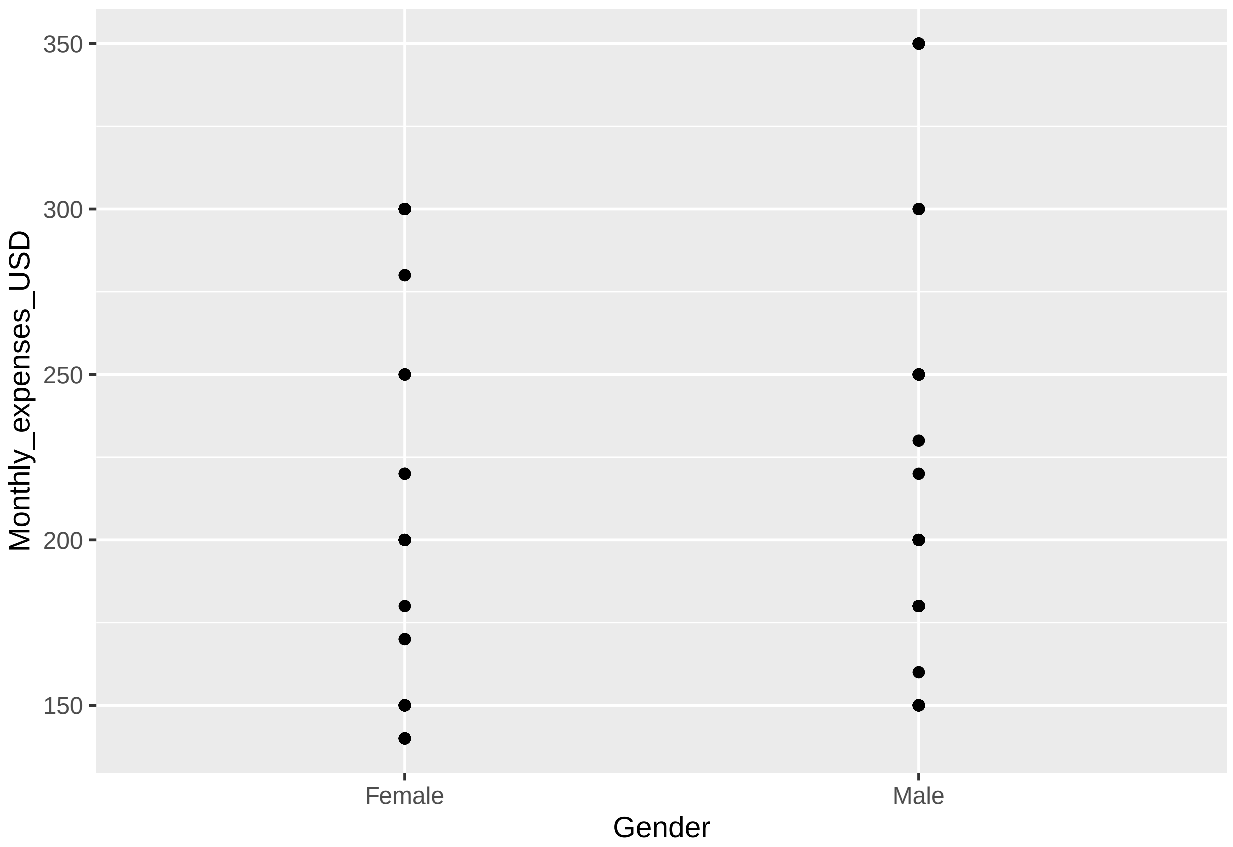
```
## Rows: 105 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (10): Gender, Living, Scholarship, Part_time_job, Transporting, Smoking,...
## dbl  (3): Age, Study_year, Monthly_expenses_USD
##
## i Use 'spec()' to retrieve the full column specification for this data.
## i Specify the column types or set 'show_col_types = FALSE' to quiet this message.
```

```
colnames(students)
```

```
##  [1] "Gender"                   "Age"
##  [3] "Study_year"               "Living"
##  [5] "Scholarship"              "Part_time_job"
##  [7] "Transporting"             "Smoking"
##  [9] "Drinks"                   "Games_and_Hobbies"
## [11] "Cosmetics_and_Self_care" "Monthly_Subscription"
## [13] "Monthly_expenses_USD"
```

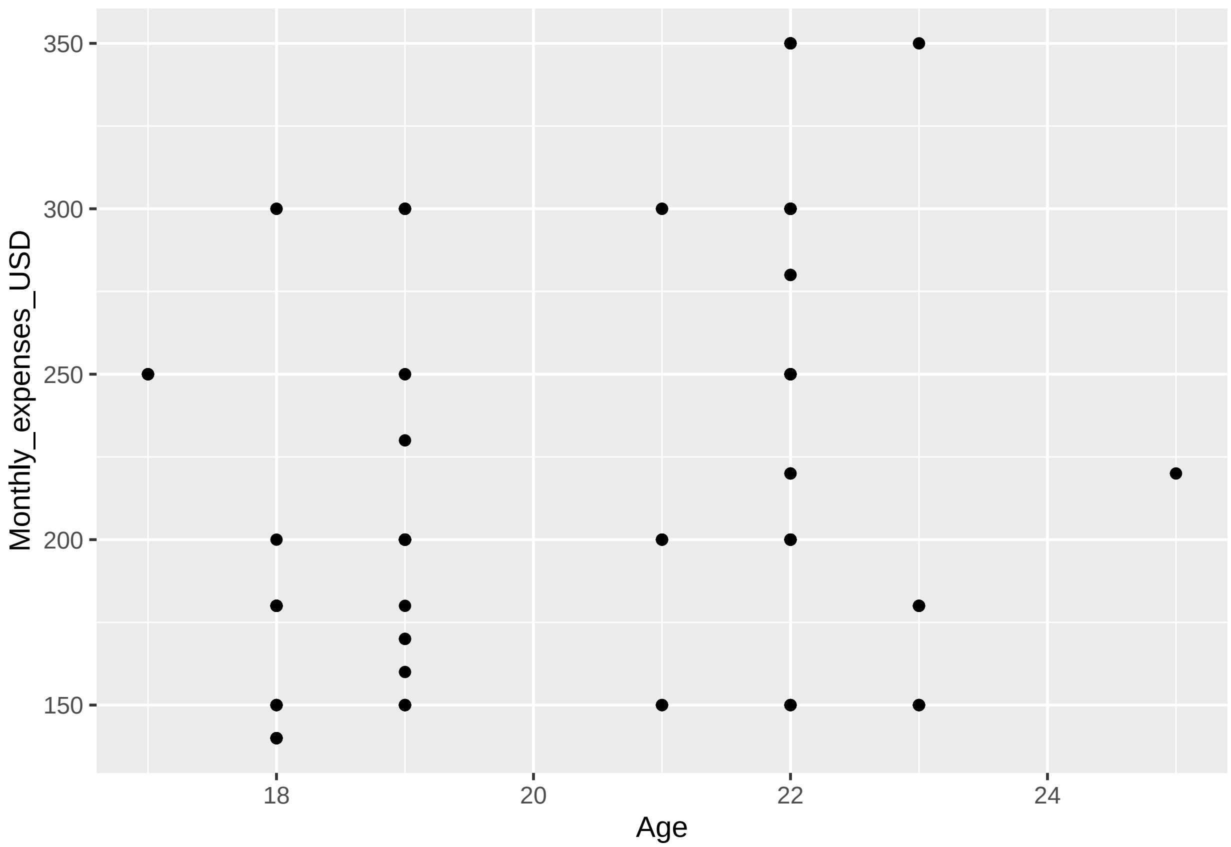
```
ggplot(students, aes(y = `Monthly_expenses_USD`, x =Gender)) +
  geom_point()
```

```
## Warning: Removed 6 rows containing missing values (geom_point).
```



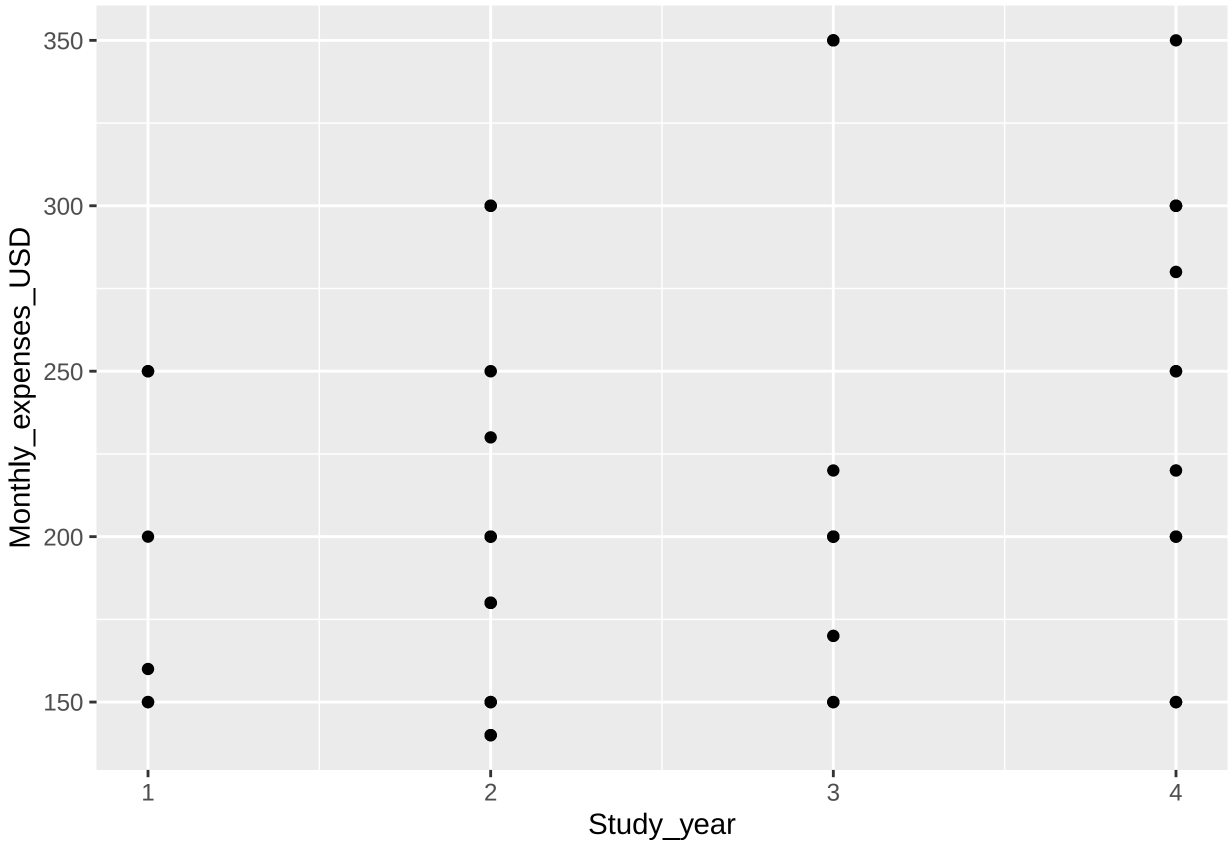
```
ggplot(students, aes(y = `Monthly_expenses_USD`, x =Age)) +
  geom_point()
```

```
## Warning: Removed 6 rows containing missing values (geom_point).
```



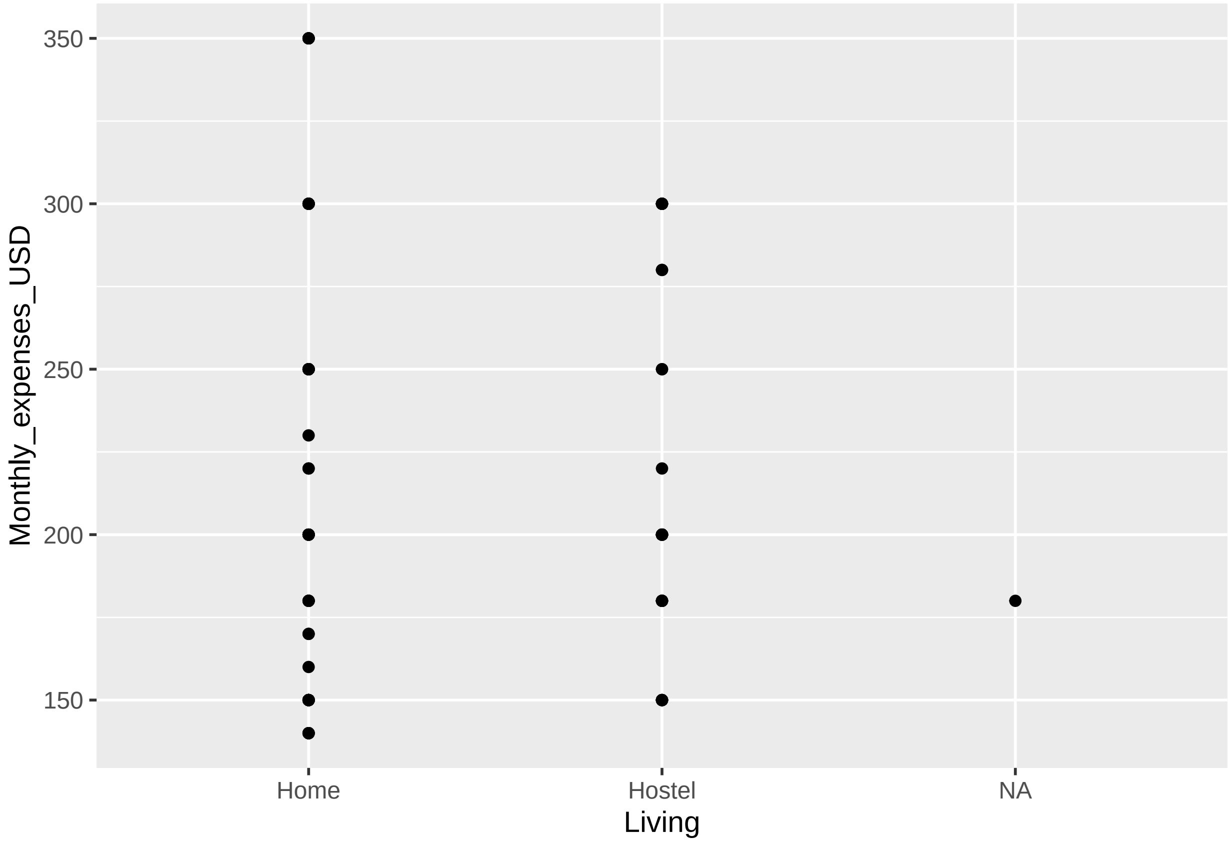
```
ggplot(students, aes(y = `Monthly_expenses_USD`, x =`Study_year`)) +
  geom_point()
```

```
## Warning: Removed 10 rows containing missing values (geom_point).
```



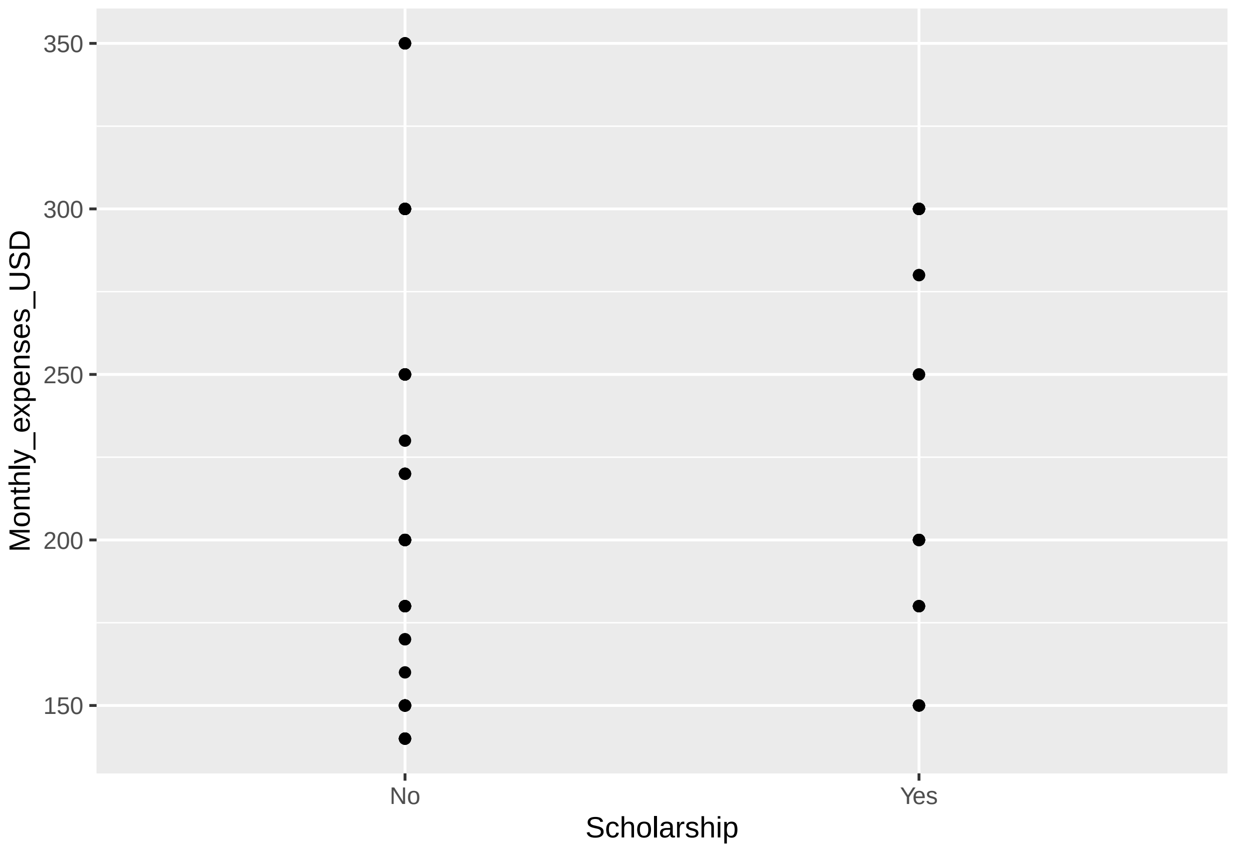
```
ggplot(students, aes(y = `Monthly_expenses_USD`, x =Living)) +
  geom_point()
```

```
## Warning: Removed 6 rows containing missing values (geom_point).
```



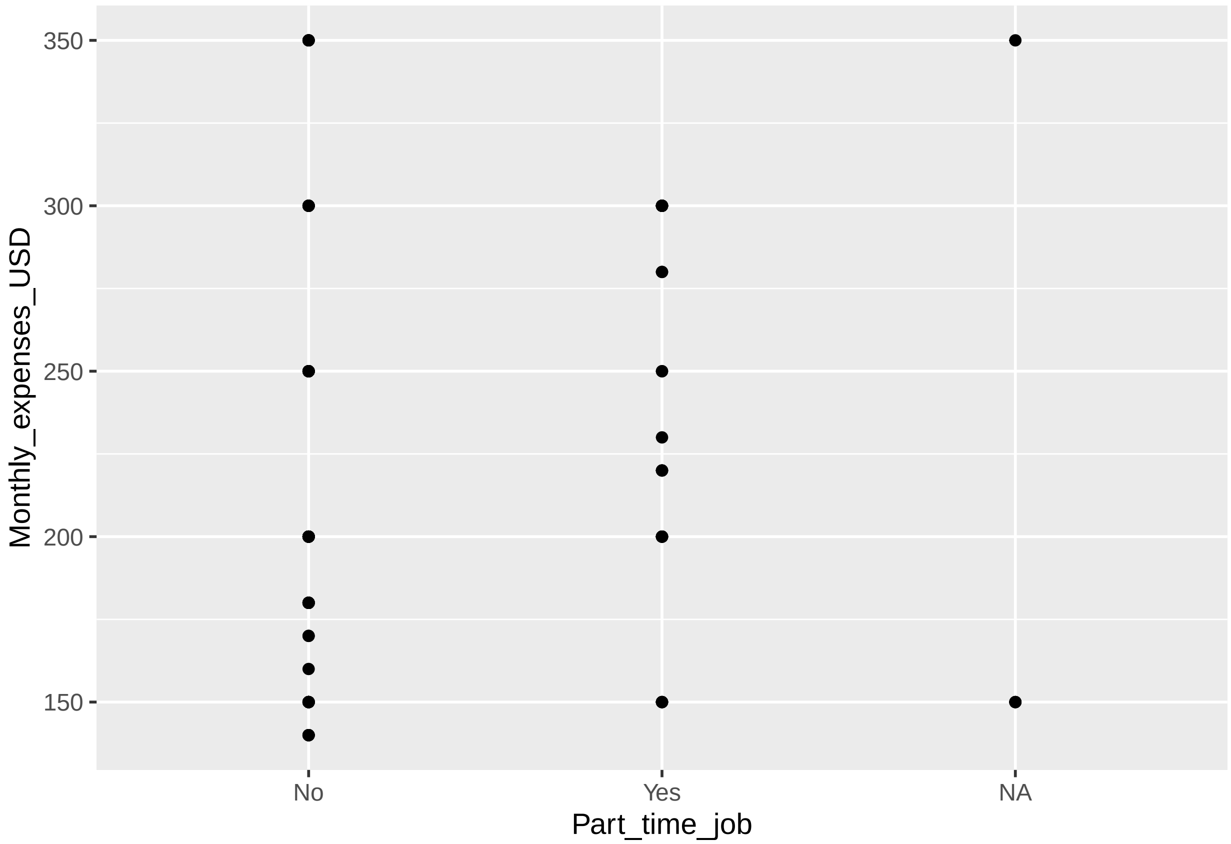
```
ggplot(students, aes(y = `Monthly_expenses_USD`, x =Scholarship)) +
  geom_point()
```

```
## Warning: Removed 6 rows containing missing values (geom_point).
```



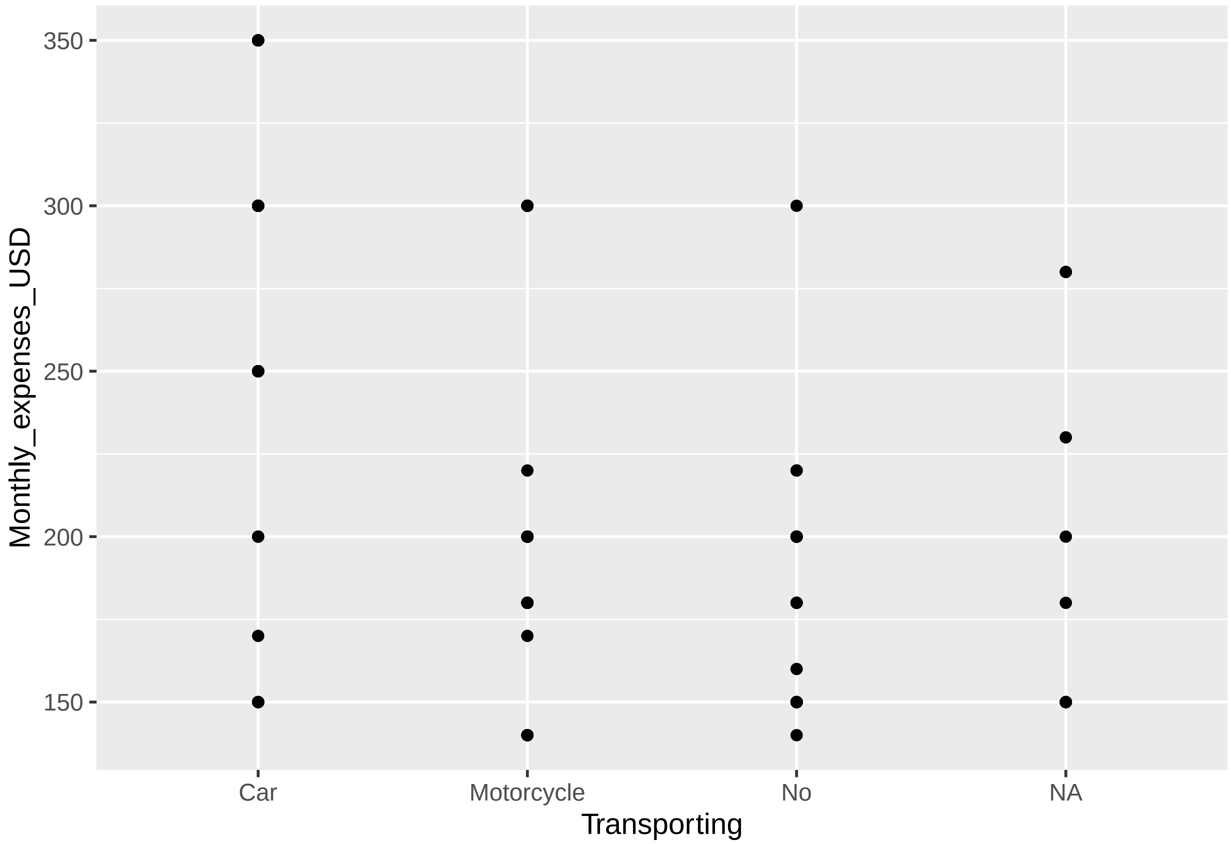
```
ggplot(students, aes(y = `Monthly_expenses_USD`, x =`Part_time_job`)) +
  geom_point()
```

```
## Warning: Removed 6 rows containing missing values (geom_point).
```



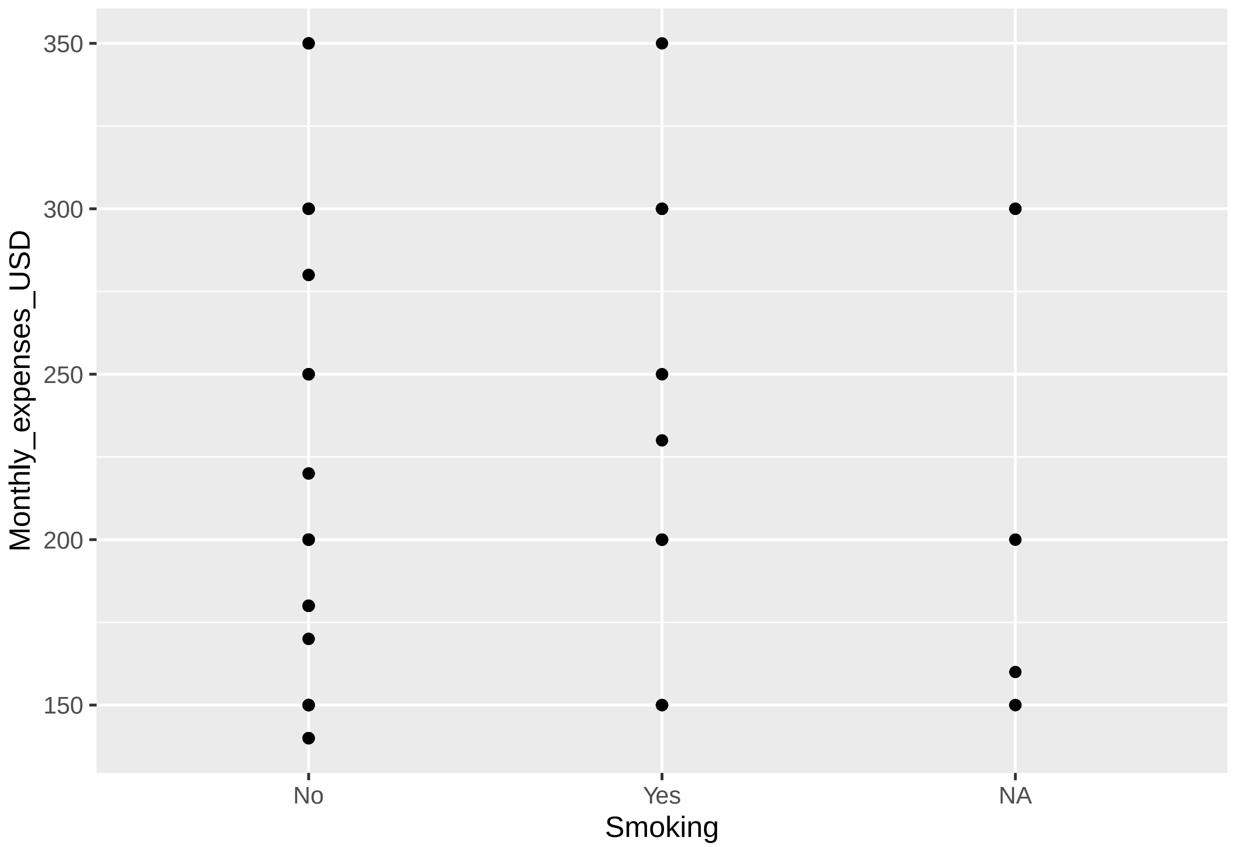
```
ggplot(students, aes(y = `Monthly_expenses_USD`, x =Transporting)) +
  geom_point()
```

```
## Warning: Removed 6 rows containing missing values (geom_point).
```



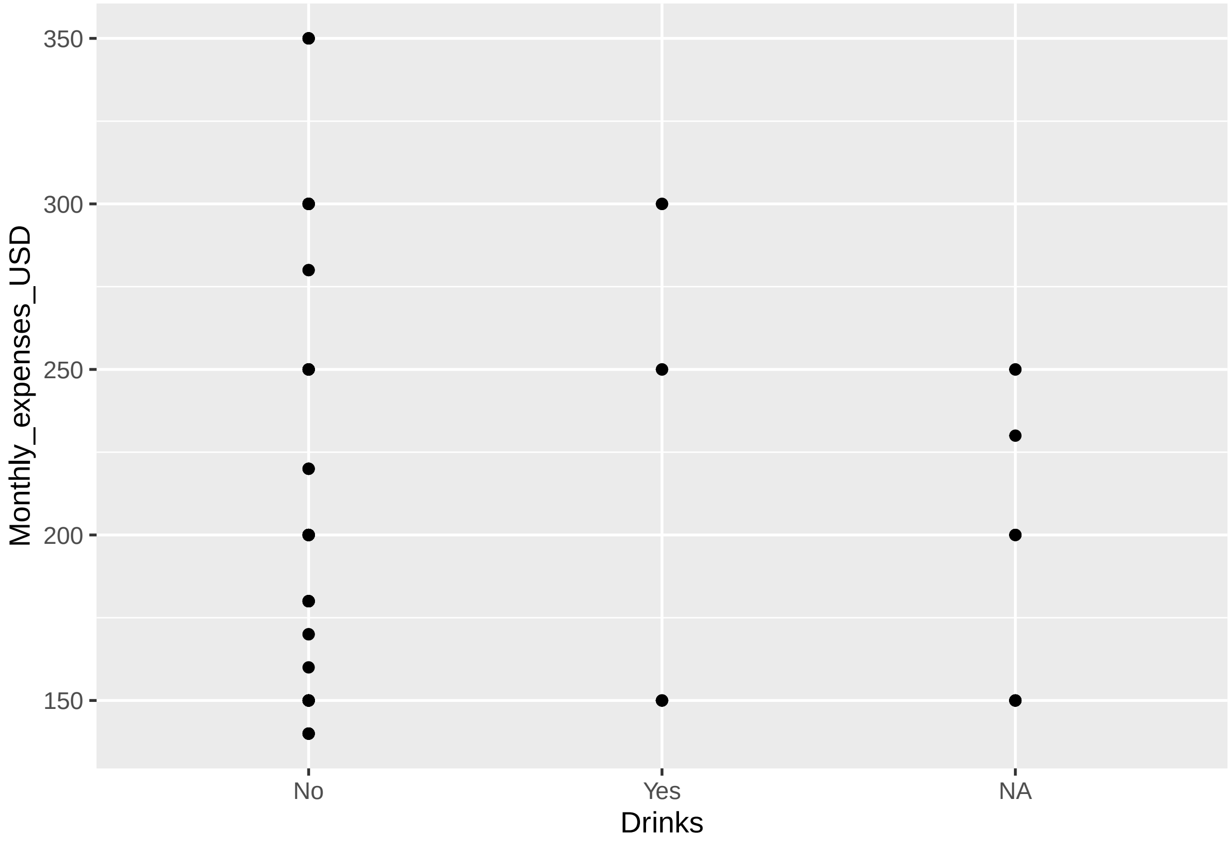
```
ggplot(students, aes(y = `Monthly_expenses_USD`, x =Smoking)) +
  geom_point()
```

```
## Warning: Removed 6 rows containing missing values (geom_point).
```



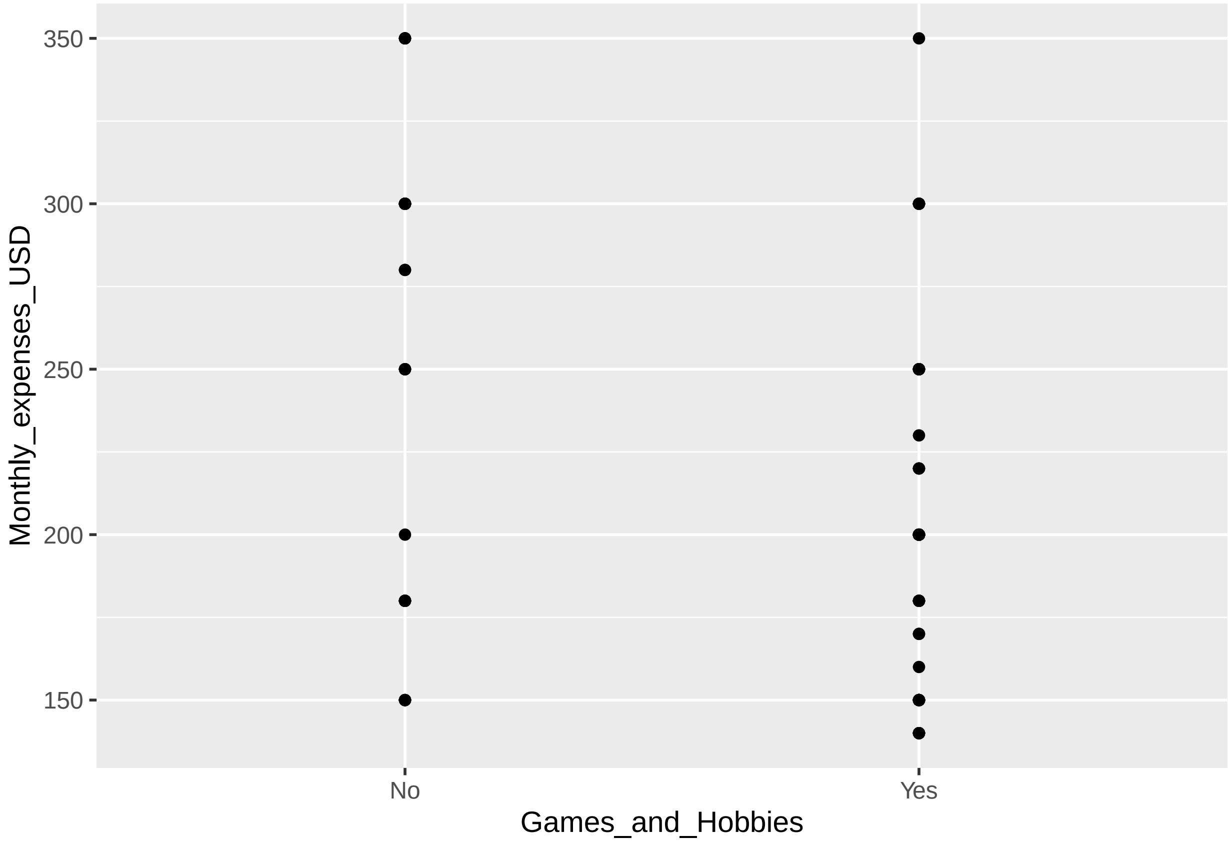
```
ggplot(students, aes(y = `Monthly_expenses_USD`, x =Drinks)) +
  geom_point()
```

```
## Warning: Removed 6 rows containing missing values (geom_point).
```



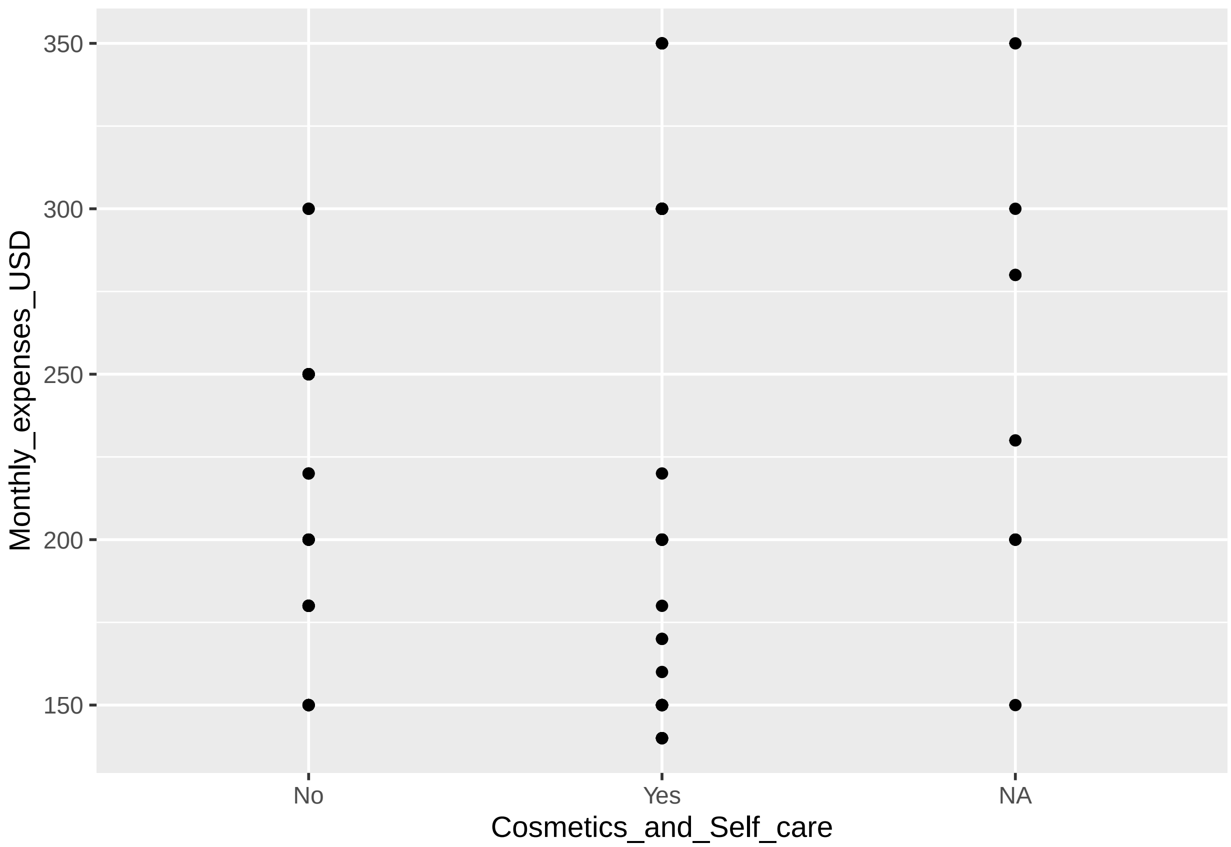
```
ggplot(students, aes(y = `Monthly_expenses_USD`, x =`Games_and_Hobbies`)) +
  geom_point()
```

```
## Warning: Removed 6 rows containing missing values (geom_point).
```



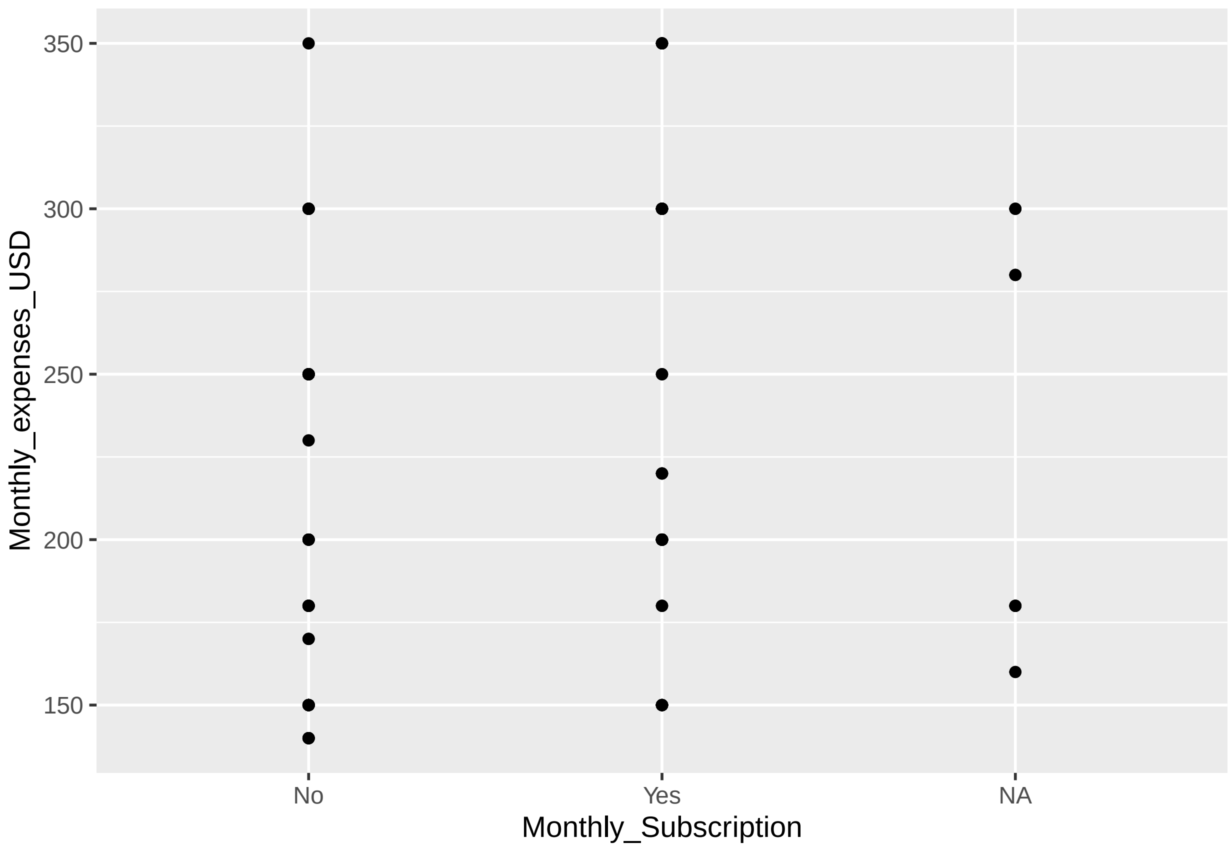
```
ggplot(students, aes(y = `Monthly_expenses_USD`, x =`Cosmetics_and_Self_care`)) +
  geom_point()
```

```
## Warning: Removed 6 rows containing missing values (geom_point).
```



```
ggplot(students, aes(y = `Monthly_expenses_USD`, x =`Monthly_Subscription`)) +
  geom_point()
```

```
## Warning: Removed 6 rows containing missing values (geom_point).
```



A lot of categorical variables, so hard to estimate if linear or not.

```
fit <- lm(`Monthly_expenses_USD`~ Gender +
                                Age +
                                Study_year +
                                Living +
                                Scholarship +
                                Part_time_job +
                                Transporting +
                                Smoking +
                                Drinks +
                                Games_and_Hobbies +
                                Cosmetics_and_Self_care +
                                Monthly_Subscription,
                                data=students)
summary(fit)
```

```
##
## Call:
## lm(formula = Monthly_expenses_USD ~ Gender + Age + Study_year +
##     Living + Scholarship + Part_time_job + Transporting + Smoking +
##     Drinks + Games_and_Hobbies + Cosmetics_and_Self_care + Monthly_Subscription,
##     data = students)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -81.928 -16.416  -0.787  16.686  82.341
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -75.359    104.848  -0.719 0.475643
## GenderMale                  -38.491     19.644  -1.959 0.055642 .
## Age                          20.345      6.454   3.152 0.002737 **
## Study_year                  -17.055     11.020  -1.548 0.128017
## LivingHostel                 26.599     20.411   1.303 0.198486
## ScholarshipYes              -24.055     17.743  -1.356 0.181257
## Part_time_jobYes            -40.462     28.144  -1.438 0.156764
## TransportingMotorcycle      -54.179     16.156  -3.353 0.001528 **
## TransportingNo              -91.265     21.593  -4.227 0.000101 ***
## SmokingYes                  -13.245     23.947  -0.553 0.582680
## DrinksYes                    90.506     30.753   2.943 0.004918 **
## Games_and_HobbiesYes        -24.129     13.582  -1.776 0.081737 .
## Cosmetics_and_Self_careYes   -1.706     15.004  -0.114 0.909907
## Monthly_SubscriptionYes      36.401     15.733   2.314 0.024833 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.34 on 50 degrees of freedom
##   (41 observations deleted due to missingness)
## Multiple R-squared:  0.6276, Adjusted R-squared:  0.5308
## F-statistic: 6.482 on 13 and 50 DF,  p-value: 5.447e-07
```

R2 backwards approach of model tunning.

```
fit <- lm(`Monthly_expenses_USD`~ Gender +
                                Age +
                                Study_year +
                                Living +
                                Scholarship +
                                Part_time_job +
                                Transporting +
                                Smoking +
                                Drinks +
                                Games_and_Hobbies +
                                #Cosmetics_and_Self_care +
                                Monthly_Subscription,
                                data=students)
summary(fit)
```

```
##
## Call:
## lm(formula = Monthly_expenses_USD ~ Gender + Age + Study_year +
##     Living + Scholarship + Part_time_job + Transporting + Smoking +
##     Drinks + Games_and_Hobbies + Monthly_Subscription, data = students)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -79.902 -16.183  -1.925  15.569  84.620
##
```

```
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -66.184     96.079  -0.689 0.493868
## GenderMale               -37.547     15.617  -2.404 0.019667 *
## Age                       19.721      5.801   3.400 0.001275 **
## Study_year               -15.710      9.639  -1.630 0.108966
## LivingHostel              25.408     19.120   1.329 0.189493
## ScholarshipYes           -21.989     15.710  -1.400 0.167323
## Part_time_jobYes         -41.779     25.552  -1.635 0.107853
## TransportingMotorcycle   -55.499     14.929  -3.718 0.000479 ***
## TransportingNo           -91.121     19.904  -4.578  2.8e-05 ***
## SmokingYes               -17.666     18.034  -0.980 0.331673
## DrinksYes                 93.911     27.365   3.432 0.001157 **
## Games_and_HobbiesYes     -24.326     13.041  -1.865 0.067573 .
## Monthly_SubscriptionYes   34.625     13.524   2.560 0.013287 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 40.77 on 54 degrees of freedom
##   (38 observations deleted due to missingness)
## Multiple R-squared:  0.6294, Adjusted R-squared:  0.5471
## F-statistic: 7.643 on 12 and 54 DF,  p-value: 5.074e-08
```

```
summary(fit)
```

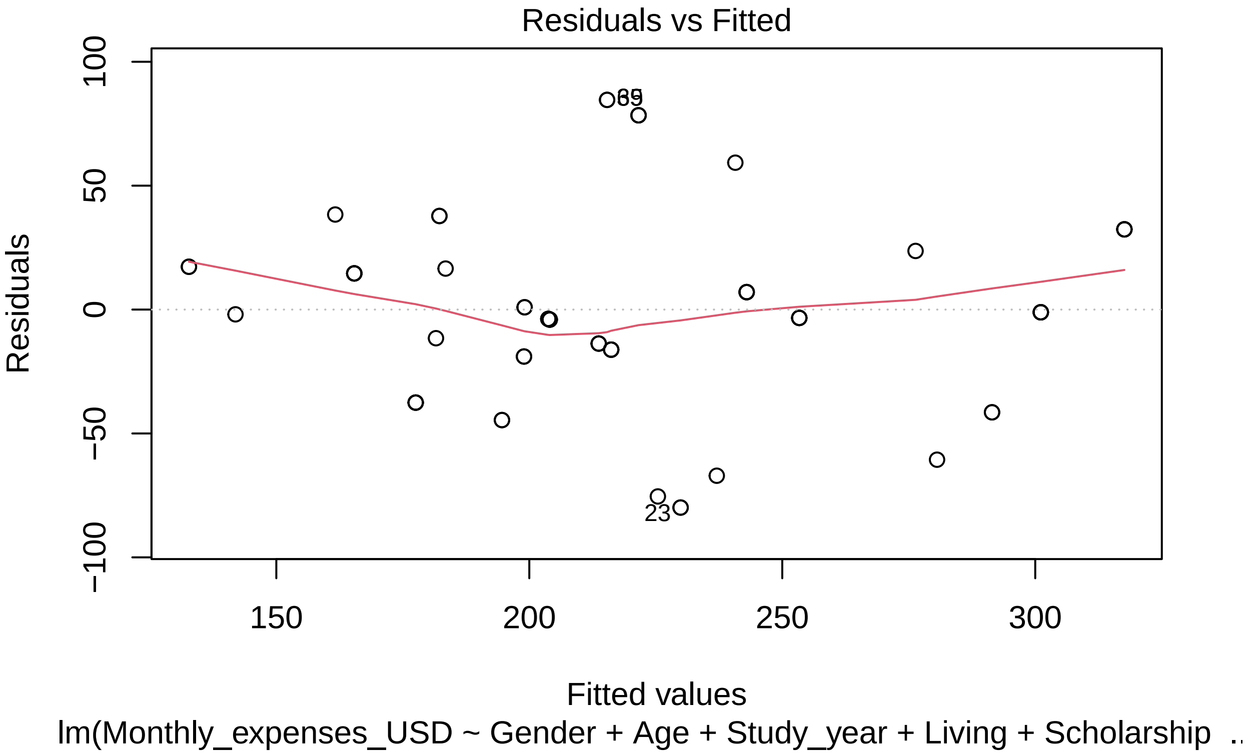
```
##
## Call:
## lm(formula = Monthly_expenses_USD ~ Gender + Age + Study_year +
##     Living + Scholarship + Part_time_job + Transporting + Smoking +
##     Drinks + Games_and_Hobbies + Monthly_Subscription, data = students)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -79.902 -16.183  -1.925  15.569  84.620
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -66.184     96.079  -0.689 0.493868
## GenderMale               -37.547     15.617  -2.404 0.019667 *
## Age                       19.721      5.801   3.400 0.001275 **
## Study_year               -15.710      9.639  -1.630 0.108966
## LivingHostel              25.408     19.120   1.329 0.189493
## ScholarshipYes           -21.989     15.710  -1.400 0.167323
## Part_time_jobYes         -41.779     25.552  -1.635 0.107853
## TransportingMotorcycle   -55.499     14.929  -3.718 0.000479 ***
## TransportingNo           -91.121     19.904  -4.578  2.8e-05 ***
## SmokingYes               -17.666     18.034  -0.980 0.331673
## DrinksYes                 93.911     27.365   3.432 0.001157 **
## Games_and_HobbiesYes     -24.326     13.041  -1.865 0.067573 .
## Monthly_SubscriptionYes   34.625     13.524   2.560 0.013287 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 40.77 on 54 degrees of freedom
##   (38 observations deleted due to missingness)
## Multiple R-squared:  0.6294, Adjusted R-squared:  0.5471
## F-statistic: 7.643 on 12 and 54 DF,  p-value: 5.074e-08
```

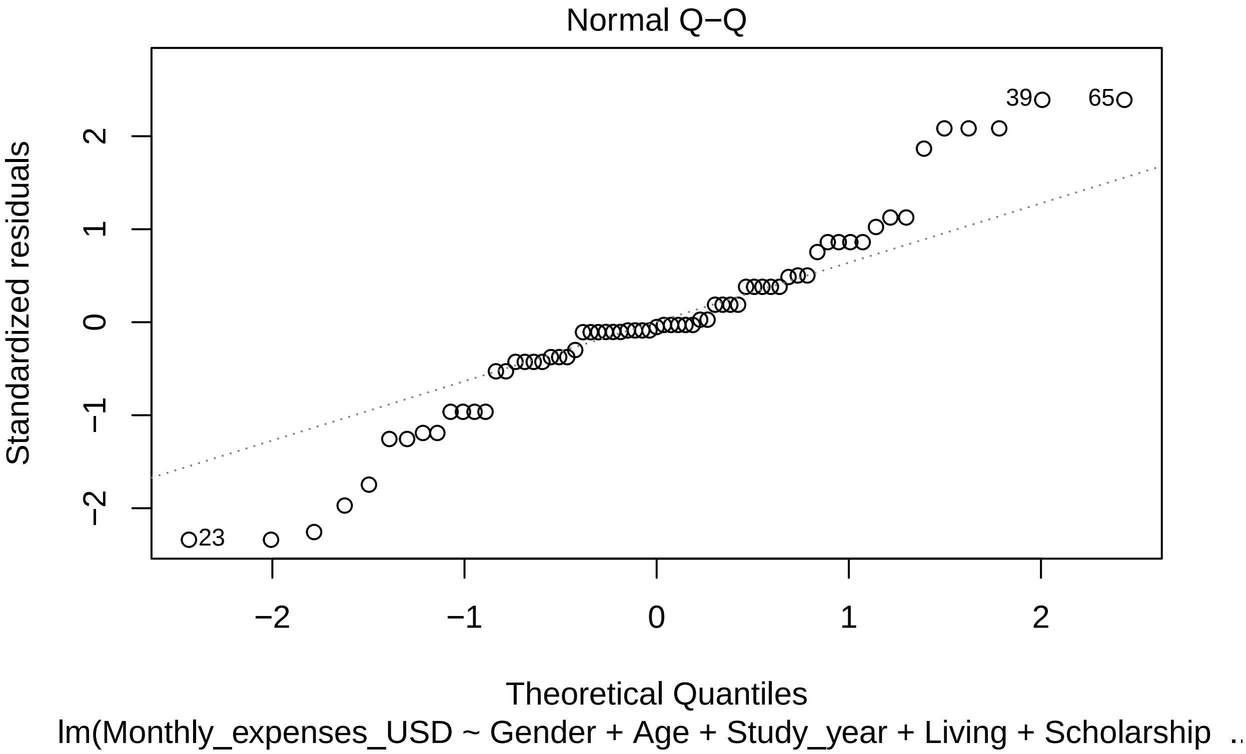
Monthly expenses in USD = - 66.184 - 37.547 (if male) + 19.721 * Age - 15.710 * study year + 25.408 (if living in a Hostel) - 21.989 (if having a scholarship) - 41.779(If have part time job) - 55.499 (if transporting on a motorcycle) - 91.121 (if walking) - 17.666 (if smoking) + 93.911 (if drinking) -24.326 (if spending on games and hobbies) + 34.625 (if have monthly subscriptions)

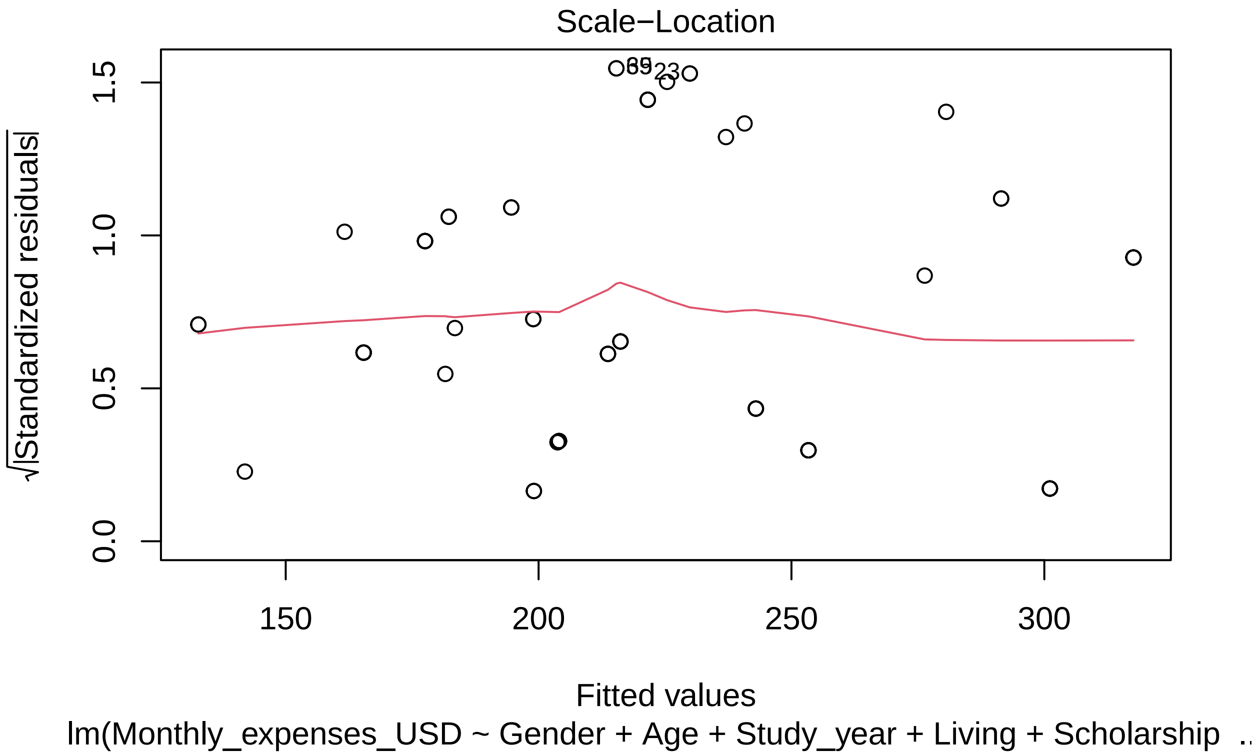
b) Is the model valid.

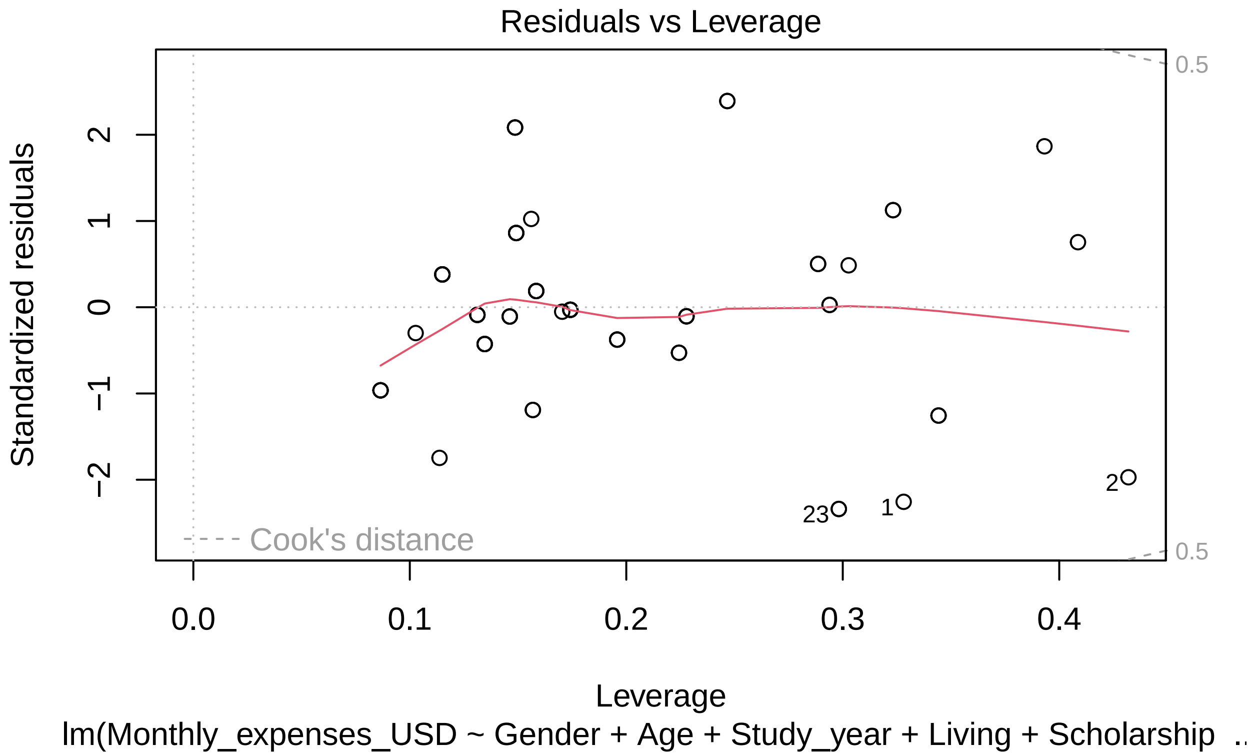
- Linearity - already checked
- independence - we assume it's independently sampled
- constant distrubution of residuals
- Normal distribution of residuals.

```
plot(fit)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Monthly_expenses_USD ~ Gender + Age + Study_year + Living + Scholarship ..
−2
−1
0
1
2
23
39
65


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
35
69
23
150
200
250
300
Fitted values
lm(Monthly_expenses_USD ~ Gender + Age + Study_year + Living + Scholarship ..



There is no identifiable trend in the distribution of residuals along the model. The normality could be questioned, Normal qq plot suggests big deviations from normal distribution.

c) Predict monthly expenses for student 17721.

Student	17721
Gender	Female
Age	22
Study Year	2
Living	Home
Scholarship	No
Part time job	Yes
Transport	Car
Smoking	Yes
Drinks	Yes
Games and Hobbies	No
Cosmetics, Self-care	Yes
Monthly Subscriptions	Yes

```
fit$coefficients['(Intercept)'] +
  fit$coefficients['GenderMale'] * 0 +
  fit$coefficients['Age'] * 22 +
  fit$coefficients['Study_year'] * 2 +
  fit$coefficients['LivingHostel'] * 0 +
  fit$coefficients['ScholarshipYes'] * 0 +
  fit$coefficients['Part_time_jobYes'] * 1 +
  fit$coefficients['TransportingMotorcycle'] * 0 +
  fit$coefficients['TransportingNo'] * 0 +
  fit$coefficients['SmokingYes'] * 1 +
  fit$coefficients['DrinksYes'] * 1 +
  fit$coefficients['Games_and_HobbiesYes'] * 0 +
  fit$coefficients['Monthly_SubscriptionYes'] * 1
```

```
## (Intercept)
##    405.3476
```

Student 17721 will spend about 405$ a month.