Week 6 report

Diana Hilleshein

2/22/2022

Preparations

```
setwd("C:/R scripts/DWA2022")
library(haven)

# import data
data <- read_sav("C:/R scripts/DWA2022/ESS9e03_1.sav")


save(data,file="data.csv")
head(data)
```

```
## # A tibble: 6 x 572
##   name  essround edition proddate  idno cntry    nwspol netusoft netustm ppltrst
##   <chr>    <dbl> <chr>   <chr>    <dbl> <chr+lb> <dbl+> <dbl+lb> <dbl+l> <dbl+l>
## 1 ESS9~        9 3.1     17.02.2~    27 AT [Aus~     60 5 [Ever~     180   2 [2]
## 2 ESS9~        9 3.1     17.02.2~   137 AT [Aus~     10 5 [Ever~      20   7 [7]
## 3 ESS9~        9 3.1     17.02.2~   194 AT [Aus~     60 4 [Most~     180   5 [5]
## 4 ESS9~        9 3.1     17.02.2~   208 AT [Aus~     45 5 [Ever~     120   3 [3]
## 5 ESS9~        9 3.1     17.02.2~   220 AT [Aus~     30 1 [Neve~      NA   5 [5]
## 6 ESS9~        9 3.1     17.02.2~   254 AT [Aus~     45 2 [Only~      NA   8 [8]
## # ... with 562 more variables: pplfair <dbl+lbl>, pplhlp <dbl+lbl>,
## #   polintr <dbl+lbl>, psppsgva <dbl+lbl>, actrolga <dbl+lbl>,
## #   psppipla <dbl+lbl>, cptppola <dbl+lbl>, trstprl <dbl+lbl>,
## #   trstlgl <dbl+lbl>, trstplc <dbl+lbl>, trstplt <dbl+lbl>, trstprt <dbl+lbl>,
## #   trstep <dbl+lbl>, trstun <dbl+lbl>, vote <dbl+lbl>, prtvtcat <dbl+lbl>,
## #   prtvtdbe <dbl+lbl>, prtvtdbg <dbl+lbl>, prtvtgch <dbl+lbl>,
## #   prtvtbcy <dbl+lbl>, prtvtecz <dbl+lbl>, prtvede1 <dbl+lbl>, ...
```

```
dim(data)#everything is right
```

```
## [1] 49519   572
```

As well as in a demo file, I will create the restricted data. I beleive it will will significantly increase speed of processing. We soulc remember, that img_avg will be a depended variable.
Used for *img_avg* variables: - *imsmetn* -Allow many/few immigrants of same race/ethnic group as majority (scale from 1 to 4, 1 = Allow many to come and live here, 4 = Allow none)
- *imdfetn* -Allow many/few immigrants of different race/ethnic group from majority (scale from 1 to 4, 1 =

Allow many to come and live here, 4 = Allow none)
- *impcntr* -Allow many/few immigrants from poorer countries outside Europe (scale from 1 to 4, 1 = Allow many to come and live here, 4 = Allow none)
General information on a participants
1) *stflife* -How satisfied with life as a whole (from 0 to 10, 0 = extremely dissatisfied, 10 = extremely satisfied)
2) *gndr*-Gender (1 = Male, 2 = Female)
3) *agea* -Age of respondent, calculated (integer)
4) *cntry* - Country of respondent (AT = Austria, BE = Belgium, BG = Bulgaria, CH = Switzerland, CY = Cyprus, CZ = Czechia, DE = Germany, DK = Denmark, EE = Estonia, ES = Spain, FI = Finland, FR = France, GB = United Kingdom, HR = Croatia, HU = Hungary, IE = Ireland, IS = Iceland, IT = Italy, LT = Lithuania, LV = Latvia, ME = Montenegro, NL = Netherlands, NO = Norway, PL = Poland, PT = Portugal, RS = Serbia, SE = Sweden, SI = Slovenia, SK = Slovakia)
5) *idno* - id number

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(ggplot2)

#create mean var
data$img_avg <- rowMeans(select(data, imwbcnt, imbgeco, imueclt), na.rm = TRUE)
head(data$img_avg)
```

```
## [1] 3.666667 4.666667 5.666667 5.000000 4.333333 6.000000
```

```
data6 <- data %>% filter(cntry == "FI") %>%
  select(., idno, cntry, agea, gndr, img_avg, stflife, ppltrst, polintr)
head(data6)
```

```
## # A tibble: 6 x 8
##    idno cntry             agea       gndr img_avg      stflife ppltrst   polintr
##   <dbl> <chr+lbl>    <dbl+lbl>  <dbl+lbl>   <dbl>    <dbl+lbl> <dbl+l> <dbl+lbl>
## 1    19 FI [Finland]        71 1 [Male]      5.67  9 [9]         7 [7] 2 [Quite~
## 2    57 FI [Finland]        29 1 [Male]      6     9 [9]         8 [8] 2 [Quite~
## 3    86 FI [Finland]        77 1 [Male]      6     8 [8]         7 [7] 2 [Quite~
## 4   120 FI [Finland]        46 1 [Male]      8    10 [Extreme~   7 [7] 2 [Quite~
## 5   164 FI [Finland]        57 2 [Female]    7.33  9 [9]         9 [9] 1 [Very ~
## 6   238 FI [Finland]        39 2 [Female]    4.67 10 [Extreme~   9 [9] 2 [Quite~
```

```
dim(data6)
```

```
## [1] 1755    8
```

For ANOVA analysis we need to have more than 3 categories in the variable.

```
data6 <- data6 %>%
  mutate(age3cat = cut(agea, breaks=c(0,39,65,Inf), labels=c("young", "middle-aged",
  ↪  "old")))
# check result!
table(data6$age3cat, useNA = "ifany")
```

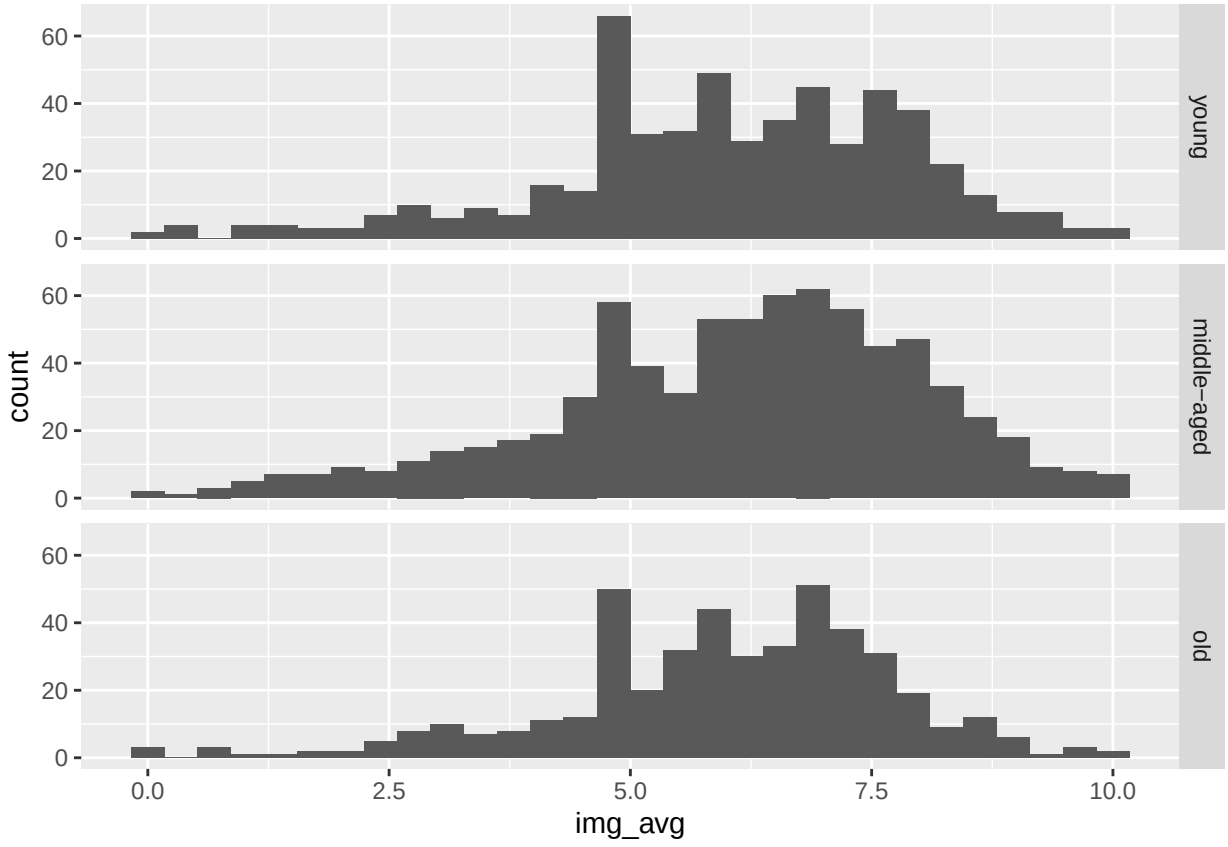
```
##
##       young middle-aged         old
##         544         752         459
```

We got number of cases in each age category.

```
# histograms by groups for img_avg
ggplot(data6, aes(x = img_avg)) +
  geom_histogram() + facet_grid(age3cat ~ .)
```

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

```
## Warning: Removed 7 rows containing non-finite values (stat_bin).
```



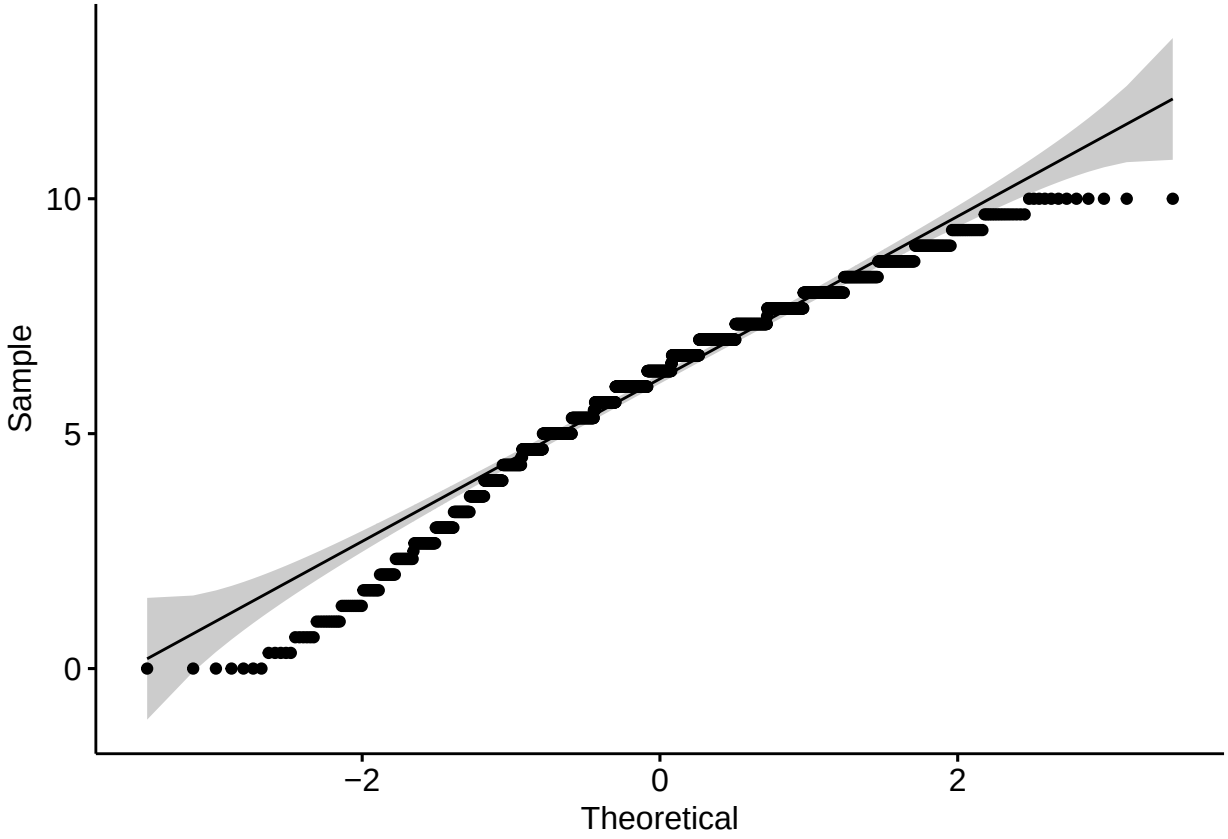
Vissual I can see that the histogram has a tail that goes to left. I will test if the data normally distributed.

```
#testing for normal distribution
library(ggpubr)
ggqqplot(data6$img_avg)
```

```
## Warning: Removed 7 rows containing non-finite values (stat_qq).

## Warning: Removed 7 rows containing non-finite values (stat_qq_line).

## Warning: Removed 7 rows containing non-finite values (stat_qq_line).
```



Dots are not following like entirely, going out of confidence intervals on both ends.

```
#testing for normal distribution
ks.test(data6$img_avg, "pnorm")
```

```
## Warning in ks.test(data6$img_avg, "pnorm"): ties should not be present for the
## Kolmogorov-Smirnov test

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  data6$img_avg
## D = 0.95243, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

As the p-value is smaller than 0.05, I reject the normality hypothesis. Therefore, I need to use the nonparametric version, the kruskal.test function: https://rc2e.com/linearregressionandanova#recipe-id231

```
# H0: The mean is the same for all groups
# H1: At least the mean of one group is different
kruskal.test(data6$img_avg ~ data6$age3cat)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  data6$img_avg by data6$age3cat
## Kruskal-Wallis chi-squared = 2.1251, df = 2, p-value = 0.3456
```

Conventionally, p = 0.346 indicates that there is no a significant difference between the medians of three age groups. I don't need to perform multiple pairwise-comparison, but I will do it as I am interested

(alternative: p = 0.05 indicates that there is a significant difference between the medians of three age groups. But we don't know which pairs of groups are different. Next we will perform multiple pairwise-comparison to determine if the mean difference between specific pairs of group are statistically significant: https://www.spcforexcel.com/knowledge/comparing-processes/bonferronis-method)

```
#multiple pairwise-comparison with bonferroni adj
# pairwise t test with Bonferroni adjustment
#H0: There is no difference  in satisafction between youn, middle-age and old groups
#H1: There is change in satisafction between young, middle-age and old groups

## for normal distribution
#pairwise.t.test(data6$sat_avg, data6$age3cat, p.adj = "bonf")
#data6 %>% pairwise_t_test(sat_avg ~ age3cat, paired = TRUE, p.adjust.method =
↪  "bonferroni") #didnt work
# anotehr way
library(tidyverse)
library(ggpubr)
library(rstatix)
```

```
##
## Attaching package: 'rstatix'

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

```
pairwise.wilcox.test(data6$img_avg, data6$age3cat, p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  data6$img_avg and data6$age3cat
##
##             young middle-aged
## middle-aged 1.00  -
```

```
## old         1.00  0.45
##
## P value adjustment method: bonferroni
```

There rae no statistically significnat results.

```
library(summarytools)
```

```
##
## Attaching package: 'summarytools'

## The following object is masked from 'package:tibble':
##
##     view
```

```
with(data6,
stby(data=data6$img_avg,
INDICES=data6$age3cat, #form froups by gender => male/female ==> find summaries of social
↪  trust for each group
FUN=descr,
stats=c("mean","sd","min","med","max","skewness")))
```

```
## Descriptive Statistics
## img_avg by age3cat
## Data Frame: data6
## N: 544
##
##                   young   middle-aged     old
## --------------- ------- ------------- -------
##            Mean    6.09          6.15    6.05
##         Std.Dev    1.87          1.96    1.74
##             Min    0.00          0.00    0.00
##          Median    6.33          6.33    6.33
##             Max   10.00         10.00   10.00
##        Skewness   -0.68         -0.63   -0.76
```

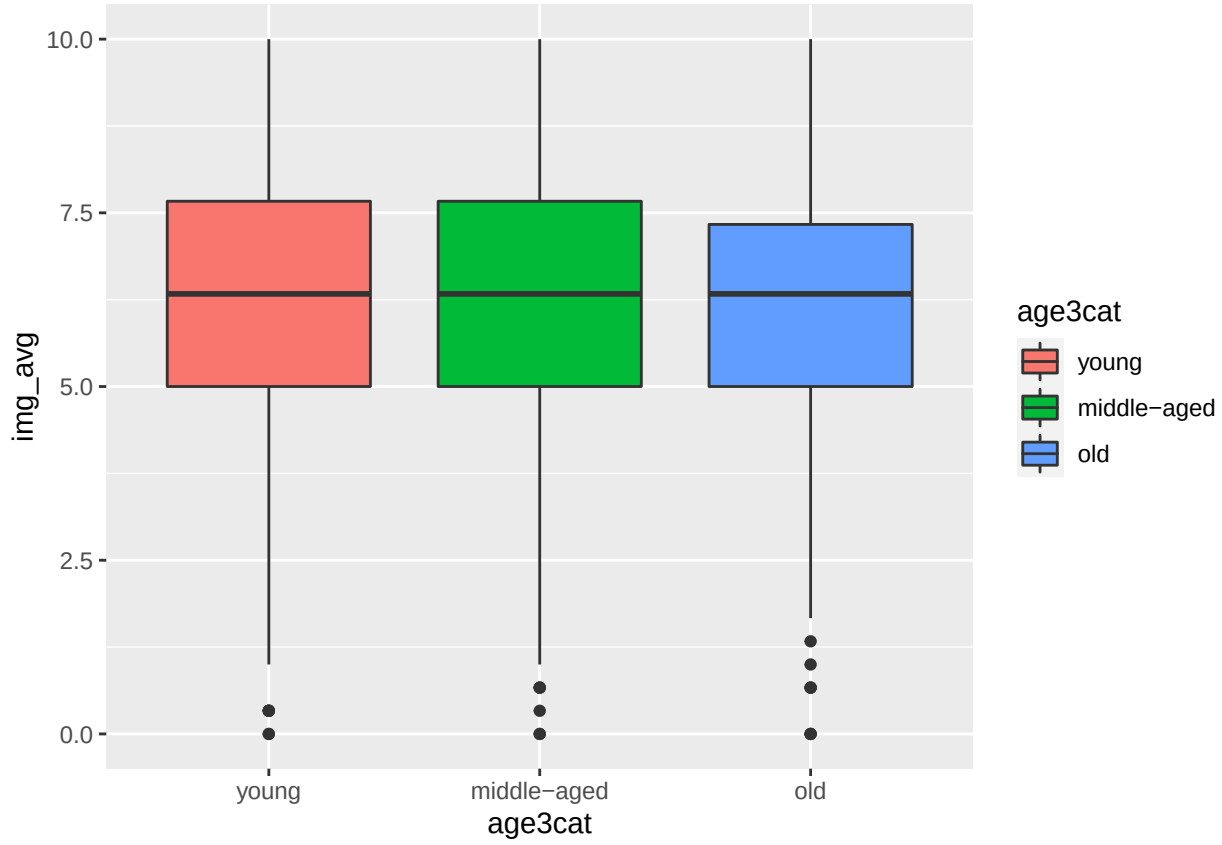
Multiple pairwise-comparison test indicates that there are no significnat differences between medians.

```
# some figures and plots
data6 %>% group_by(age3cat)  %>%
  summarise(n = n(), mean_img_avg = mean(img_avg, na.rm = TRUE),
            sd_img_avg = sd(img_avg, na.rm = TRUE))
```

```
## # A tibble: 3 x 4
##   age3cat         n mean_img_avg sd_img_avg
##   <fct>       <int>        <dbl>      <dbl>
## 1 young         544         6.09       1.87
## 2 middle-aged   752         6.15       1.96
## 3 old           459         6.05       1.74
```

```
ggplot(data6, aes(x = age3cat, y = img_avg, fill = age3cat)) +
  geom_boxplot()
```

```
## Warning: Removed 7 rows containing non-finite values (stat_boxplot).
```



Linear regression model
Checking correlations

Variables:
- *stflife* -How satisfied with life as a whole (from 0 to 10, 0 = extremely dissatisfied, 10 = extremely satisfied)
- *agea* -Age of respondent, calculated (integer)
- *gndr*-Gender (1 = Male, 2 = Female)
- *cntry* - Country of respondent (AT = Austria, BE = Belgium, BG = Bulgaria, CH = Switzerland, CY = Cyprus, CZ = Czechia, DE = Germany, DK = Denmark, EE = Estonia, ES = Spain, FI = Finland, FR = France, GB = United Kingdom, HR = Croatia, HU = Hungary, IE = Ireland, IS = Iceland, IT = Italy, LT = Lithuania, LV = Latvia, ME = Montenegro, NL = Netherlands, NO = Norway, PL = Poland, PT = Portugal, RS = Serbia, SE = Sweden, SI = Slovenia, SK = Slovakia)
- *polintr* -How interested in politics (from 1 to 4, 1= Very interested ... 4 = Not at all)
- *imwbcnt* -Immigrants make country worse or better place to live (scale from 0 to 10, 0 = Worse place to live, 10 = Better place to live)
- *ppltrst* - Most people can be trusted or you can't be too careful (from 0 to 10, 0 = You can't be too careful ... 10 = people can be trusted)

Scales go from negative to positive for sat_avg and imbgeco, so we do not need to recode them. Scale for polintr does not have positive or negative meaning.

```
#corrmatrix <- cor(select(data6, img_avg, stflife, agea, gndr, cntry, polintr, ppltrst),
↪  use="complete.obs")
#corrmatrix
# gives me a mistake: "cor(data6, use = "complete.obs") : 'x' must be numeric"

sapply(data6, is.numeric) #checking wich columns are not appropriate//makes sence that
↪  cntry are not acceptable XD
```

```
##    idno   cntry    agea    gndr img_avg stflife ppltrst polintr age3cat
##    TRUE   FALSE    TRUE    TRUE    TRUE    TRUE    TRUE    TRUE   FALSE
```

```
corrmatrix <- cor(select(data6, img_avg, stflife, agea, gndr, polintr, ppltrst),
↪  use="complete.obs")
corrmatrix
```

```
##             img_avg     stflife        agea        gndr     polintr
## img_avg  1.00000000  0.21498599 -0.01525830 0.048953380 -0.20823574
## stflife  0.21498599  1.00000000  0.04749333 0.015555249 -0.06267876
## agea    -0.01525830  0.04749333  1.00000000 0.010983559 -0.06828374
## gndr     0.04895338  0.01555525  0.01098356 1.000000000  0.09937074
## polintr -0.20823574 -0.06267876 -0.06828374 0.099370735  1.00000000
## ppltrst  0.32626685  0.24339357  0.06112856 0.007792676 -0.06082827
##              ppltrst
## img_avg  0.326266853
## stflife  0.243393570
## agea     0.061128556
## gndr     0.007792676
## polintr -0.060828266
## ppltrst  1.000000000
```

```
#what happens if I use NAs
cor(select(data6, img_avg, agea, gndr, polintr)) #not good
```

```
##         img_avg       agea       gndr polintr
## img_avg       1         NA         NA      NA
## agea         NA 1.00000000 0.01212829      NA
## gndr         NA 0.01212829 1.00000000      NA
## polintr      NA         NA         NA       1
```

Correlations:

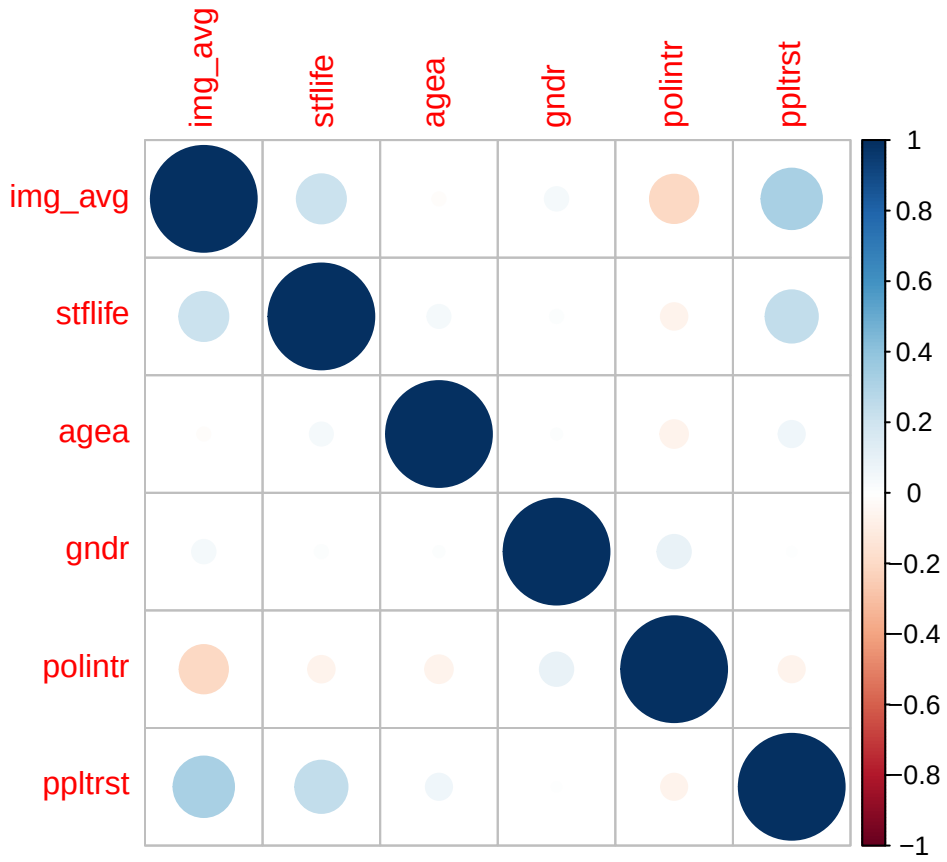
- stflife & **img_avg** (r= 0.215) - the more people satisfied with life and government, the more they accept immigration
- **img_avg** & ppltrst (r = 0.326) - the more people accept immigrants, the more people trust other people

- **img_avg** & polintr (r = -0.208) - the more people accept immigrants, the lest they interested in politics
- stflife & ppltrst (r = 0.243) - the more people satisfied with life and government, the more they trust other people

```
# a plot
library(corrplot)
```

```
## corrplot 0.92 loaded
```

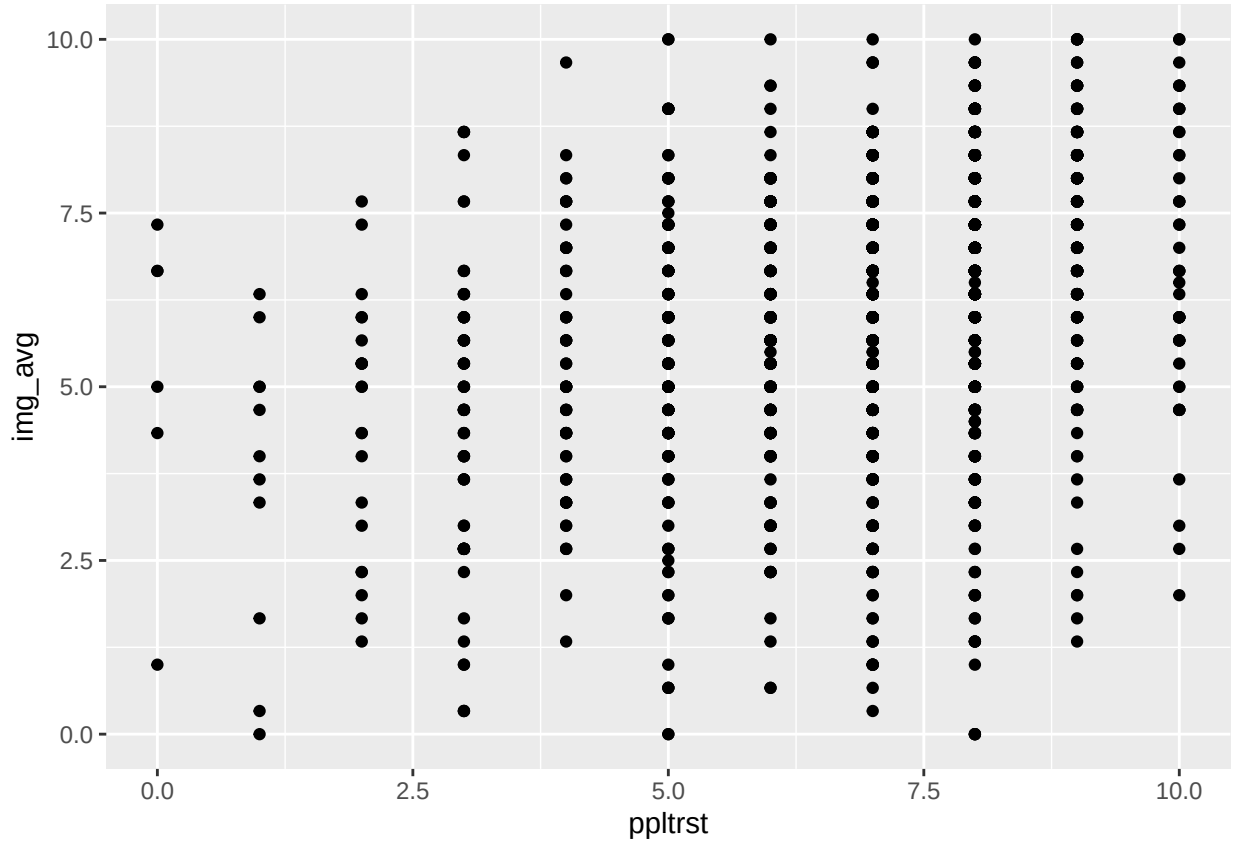
```
corrplot(corrmatrix)
```




```
# linear regression with one predictor
# Scatter plot
ggplot(data6, aes(x = ppltrst, y = img_avg)) +
  geom_point()
```

```
## Don't know how to automatically pick scale for object of type haven_labelled/vctrs_vctr/double. Defau
```

```
## Warning: Removed 7 rows containing missing values (geom_point).
```

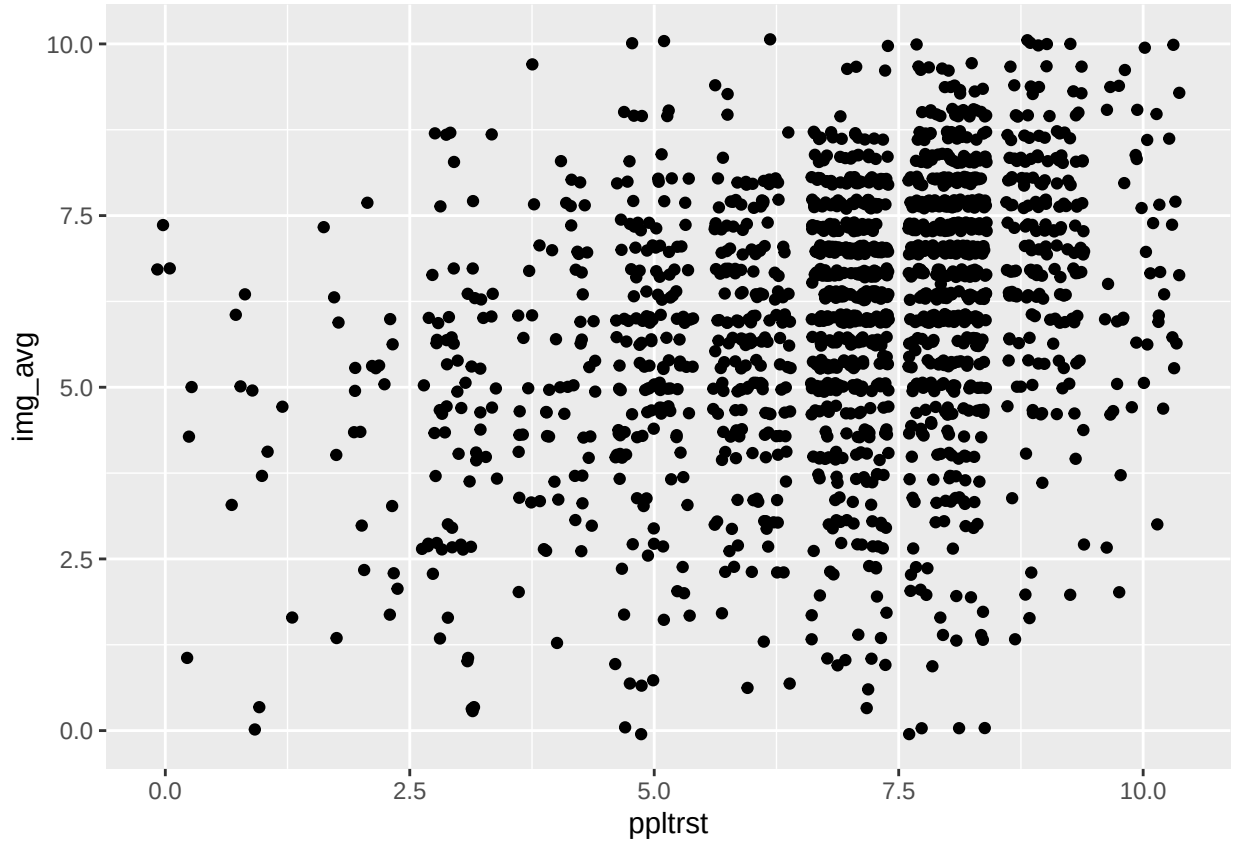


I can see that it follows a line/pattern.

```
# Sometimes some jittering (in Finnish: täristää) is needed.
ggplot(data6, aes(x = ppltrst, y = img_avg)) +
  geom_point(position = "jitter")
```

```
## Don't know how to automatically pick scale for object of type haven_labelled/vctrs_vctr/double. Defau
```

```
## Warning: Removed 7 rows containing missing values (geom_point).
```



"It is typically used to better visualize overlapping values, such as integer covariates. This helps grasp where the density of observations is high." <= https://stackoverflow.com/questions/17547699/what-does-the-jitter-function-do-in-r#:~:text=Jittering%20indeed%20means%20just%20adding,amount%2Dparameter%20is%20not%20provided
I guess we can see that the plot is skewed: the density ob observations is bigger from 5 to 8 on the ppltrst scale and from 5 to 8 on the img_avg scale. In other words the majority of answers are located in upper right corner that is above median.

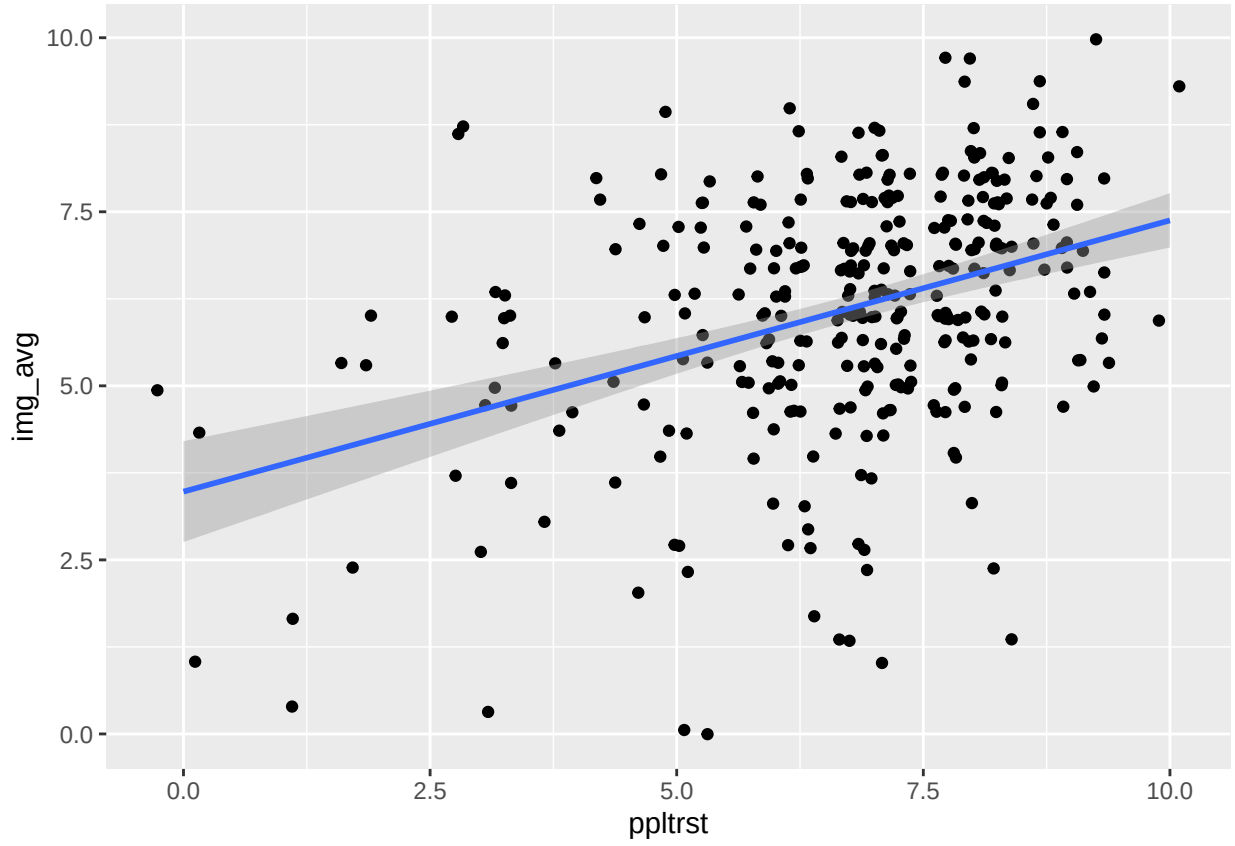
```
# add regression line and e.g. just part of the data (persons under 30 years old)
data6 %>% filter(agea <= 30) %>%
  ggplot(., aes(x = ppltrst, y = img_avg)) +
  geom_point(position = "jitter") +
  geom_smooth(method = lm)
```

```
## Don't know how to automatically pick scale for object of type haven_labelled/vctrs_vctr/double. Defau
```

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

```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```

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



Regression model

```
# At first a simple regression model
lm1 <- lm(img_avg ~ ppltrst, data = data6)
summary(lm1)
```

```
##
## Call:
## lm(formula = img_avg ~ ppltrst, data = data6)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4752 -1.0860  0.2101  1.1961  4.5992
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.65969    0.17429   21.00   <2e-16 ***
## ppltrst      0.35194    0.02436   14.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.774 on 1746 degrees of freedom
```

```
##   (7 observations deleted due to missingness)
## Multiple R-squared:  0.1068, Adjusted R-squared:  0.1063
## F-statistic: 208.8 on 1 and 1746 DF,  p-value: < 2.2e-16
```

A nice hint from Maria: Hint. For model comparison use AIC and/or BIC. The model with the lowest AIC and BIC score is preferred.

```
AIC(lm1)
```

```
## [1] 6968.993
```

```
BIC(lm1)
```

```
## [1] 6985.392
```

```
# Add second predictor
lm2 <- lm(img_avg ~ ppltrst + stflife, data = data6)
summary(lm2)
```

```
##
## Call:
## lm(formula = img_avg ~ ppltrst + stflife, data = data6)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6037 -0.9727  0.2217  1.2019  5.0509
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.48877    0.25517   9.753  < 2e-16 ***
## ppltrst      0.31358    0.02486  12.616  < 2e-16 ***
## stflife      0.17848    0.02859   6.244 5.35e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.755 on 1744 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.126,  Adjusted R-squared:  0.125
## F-statistic: 125.7 on 2 and 1744 DF,  p-value: < 2.2e-16
```

```
AIC(lm2)
```

```
## [1] 6928.664
```

```
BIC(lm2)
```

```
## [1] 6950.527
```

```r
lm3 <- lm(img_avg ~ ppltrst + stflife + ppltrst, data = data6)
summary(lm3)
```

```
##
## Call:
## lm(formula = img_avg ~ ppltrst + stflife + ppltrst, data = data6)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6037 -0.9727  0.2217  1.2019  5.0509
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.48877    0.25517   9.753  < 2e-16 ***
## ppltrst      0.31358    0.02486  12.616  < 2e-16 ***
## stflife      0.17848    0.02859   6.244 5.35e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.755 on 1744 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.126,  Adjusted R-squared:  0.125
## F-statistic: 125.7 on 2 and 1744 DF,  p-value: < 2.2e-16
```

```r
AIC(lm3) #smaller then the previous values
```

```
## [1] 6928.664
```

```r
BIC(lm3)
```

```
## [1] 6950.527
```

```r
lm4 <- lm(img_avg ~ stflife, data = data6)
summary(lm4)
```

```
##
## Call:
## lm(formula = img_avg ~ stflife, data = data6)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3548 -1.0885  0.2448  1.3119  4.7103
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.95844    0.23709  16.696   <2e-16 ***
## stflife      0.26626    0.02896   9.196   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.833 on 1745 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.04622,    Adjusted R-squared:  0.04567
## F-statistic: 84.56 on 1 and 1745 DF,  p-value: < 2.2e-16
```

```
AIC(lm4)
```

```
## [1] 7079.247
```

```
BIC(lm4)
```

```
## [1] 7095.644
```

The model 2 and 3 gave the smallest AIC/BIC values. I would prefer the model 3 since all the predictors are statistically significnat in it.

R Squared is showing proportion of variance in the dependent variable that can be explained by the independent variables. Adjusted R squared is modified version of R**2 that was adjusted for the number of predictors in the model. Looking at the output of the model, I can conclude that the model explains 12.6% of the proportion of the variance or 12.5% for Adjusted R Squared.

The equation for the model is "img_avg = 2.489+ (0.314 *"ppltrst") + (0.179* "stflife"). The model is statistically significant ($p < 0.05$). The model shows that with the increase of ppltrst and stflife by 1, the img_avg is increasing by 0.493 and the img_avg value would be 2.982.

It means that with increase of trust in people and satisfaction with life and governement policy by one unit, the acceptance of immigrants is increased by 0.493 units.

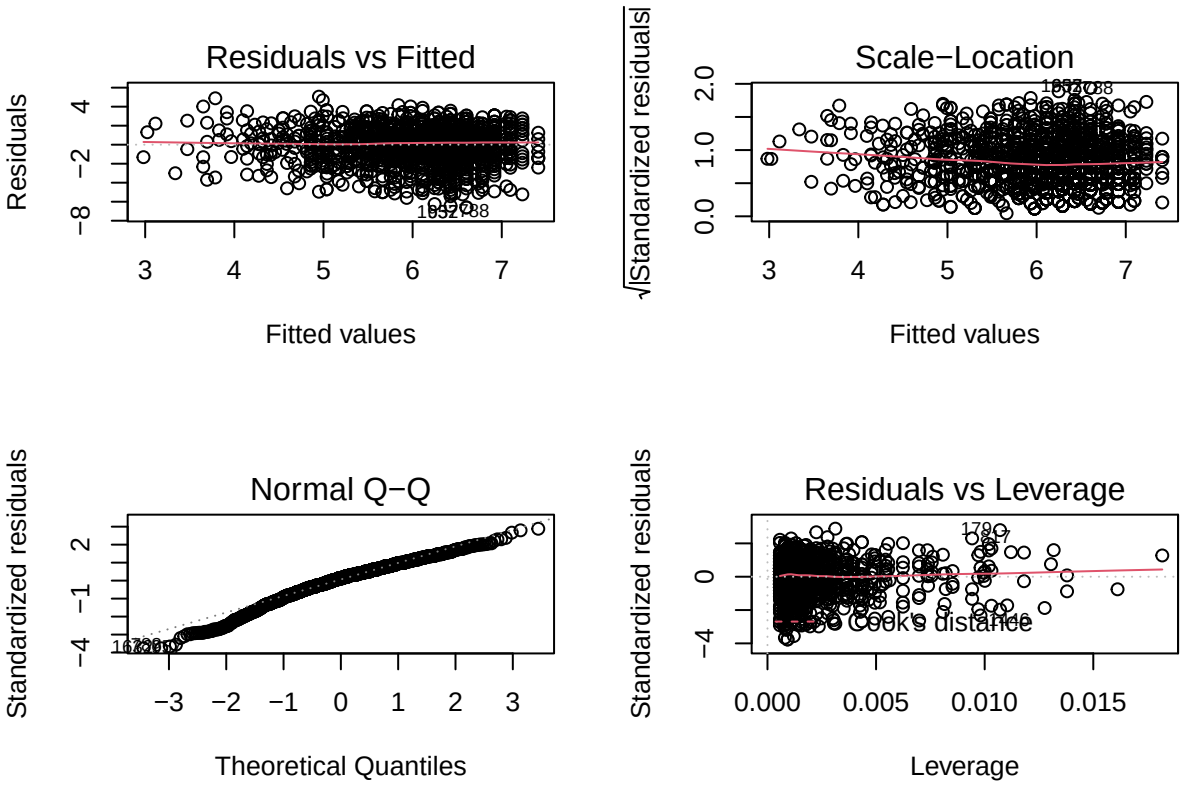
```
cor(select(data6, stflife, ppltrst), use="complete.obs")
```

```
##           stflife   ppltrst
## stflife 1.0000000 0.2436024
## ppltrst 0.2436024 1.0000000
```

Note, that with increase in satisfaction with life and governement policy, there si increase in people trust. The predictors are correlated.

Diagnostic

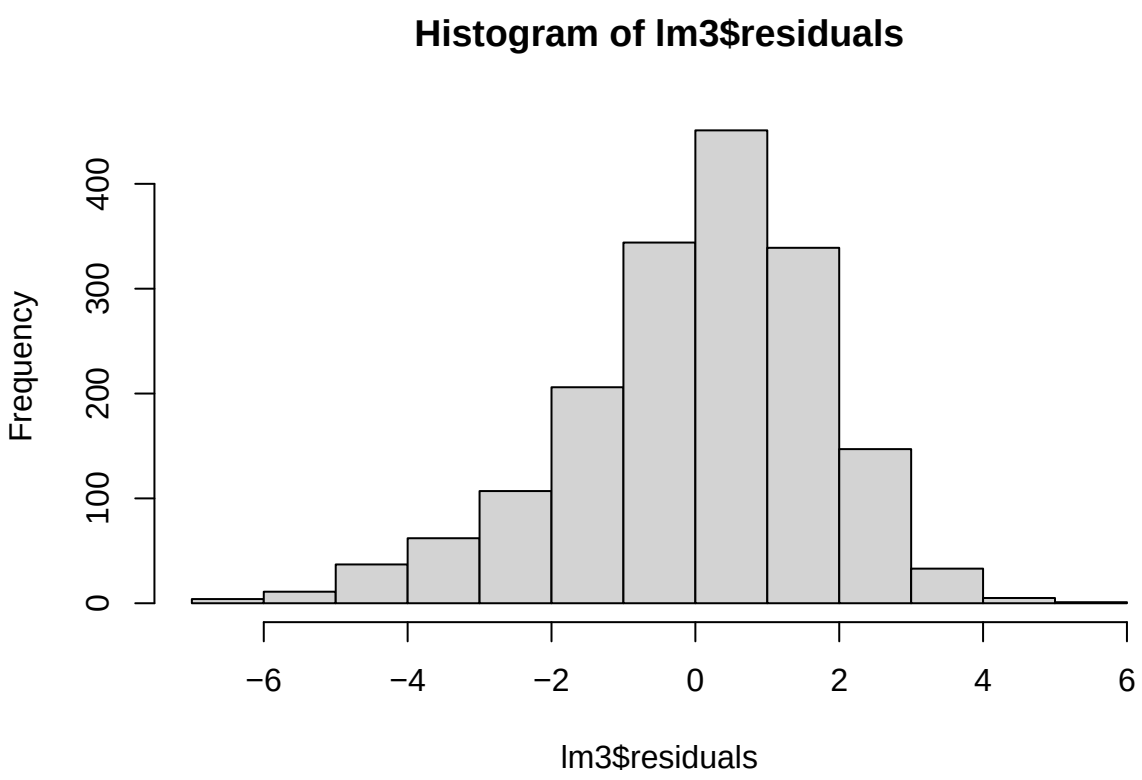
```
# diagnostic plots
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(lm3)
```



Looking at the Residual vs Fitted plot, I can say that there is some pattern in how the residuals are spreaded, they are more frequent between 6 and 7 on fitted residuals axis. Also, points on theleft seem to be too lonely. Normal Q-Q plot shows that the fit to the normal distribution is not perfect, point follow the line, but violate it on the ends. Looking at the Residuals vs Leverage plot, we conclude that there can be few observation that stands out significantly (on the right side of the plot), but they doent seem to affect the fit line strongly.

```
# Test Normality of Residuals
# H0: residuals are normally distributed

hist(lm3$residuals)
```



As I said, the residuals seem to be skewed to right.

```
ks.test(x = lm3$residuals, y = pnorm)
```

```
## Warning in ks.test(x = lm3$residuals, y = pnorm): ties should not be present for
## the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  lm3$residuals
## D = 0.14865, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

As the p-value is smaller than 0.05, I reject the normality hypothesis.