Dissertations EDA

Rui Almeida

This is a study of the master dissertations completed between 2013-2018 at ISEP. My goal is to consolidate some knowledge about R for exploratory data analysis that I obtained from reading the book R for Data Science. I also have to think about my dissertation topic this year and was curious about previous works from my university, so I will analyze a free dataset to answer some questions.

Let's start by loading the data:

```
data <- read_csv('./data/data.csv')
```

Data Exploration

```
glimpse(data)
```

```
## Observations: 292
## Variables: 13
## $ ID             <dbl> 201812975, 201813238, 201813297, 201813360, 201...
## $ Sex            <chr> "Feminino", "Masculino", "Masculino", "Feminino...
## $ Nationality    <chr> "Portugal", "Portugal", "Portugal", "Portugal",...
## $ Specialization <chr> "Engenharia de Software", "Engenharia de Softwa...
## $ Title          <chr> "Evaluating the Combination of Relaxation and A...
## $ Keywords       <chr> "Negociação de mapeamentos de ontologias; Argum...
## $ Date           <dttm> 2013-11-15, 2013-11-14, 2013-11-13, 2013-11-13...
## $ Grade          <dbl> 16, 16, 15, 15, 14, 14, 15, 14, 15, 17, 13, 14,...
## $ Advisors       <chr> "Nuno Alexandre Pinto Da Silva", "Lino Manuel B...
## $ Pages          <dbl> 93, 139, 145, 179, 87, 136, 124, 123, 108, 173,...
## $ Words          <dbl> 22292, 39826, 36143, 40198, 19187, 22224, 29549...
## $ Size           <dbl> 1.835146, 4.263048, 3.703902, 2.768955, 2.92813...
## $ WordsPerPage   <dbl> 239.6989, 286.5180, 249.2621, 224.5698, 220.540...
```

The table contains 292 rows and 13 columns. There is categorical data such as the Nationality and Specialization, discrete numerical data like the Grade and Pages and as continuous numerical data we have the WordsPerPage and Size.

What are the most common nationalities?

```
data %>%
  count(Nationality) %>%
  filter(n > 1) %>%
  arrange(desc(n))
```

```
## # A tibble: 5 x 2
##   Nationality               n
##   <chr>                 <int>
## 1 Portugal                276
## 2 Suíça                     4
## 3 Brasil                    2
## 4 Cabo Verde                2
## 5 República Árabe Síria     2
```

Out of 292 dissertations, 276 are from portuguese students. Not many students from other countries here.

Is there a significant difference between grades in specializations?

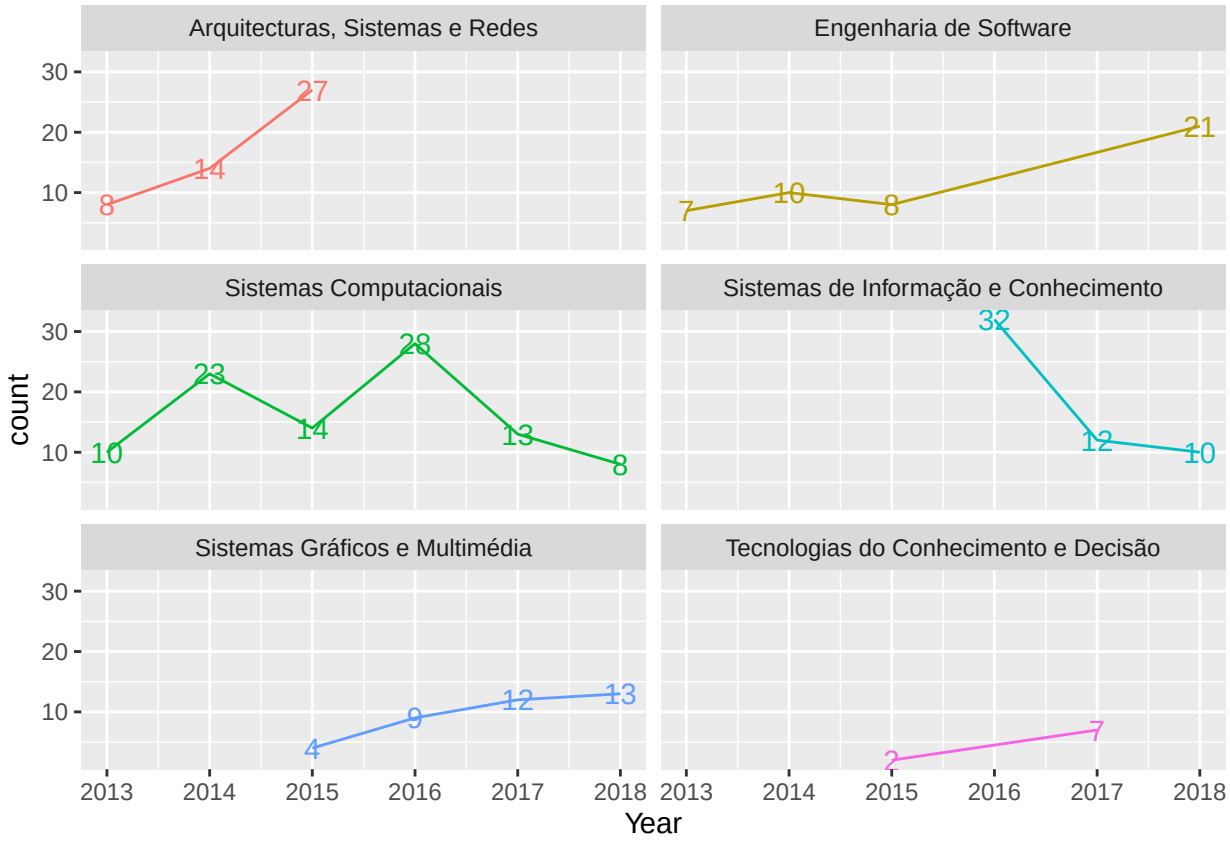
There have been a few changes to the specializations over the years. Here's the full list:

```
unique(data$Specialization)
```

```
## [1] "Engenharia de Software"
## [2] "Arquitecturas, Sistemas e Redes"
## [3] "Sistemas Computacionais"
## [4] "Tecnologias do Conhecimento e Decisão"
## [5] "Sistemas Gráficos e Multimédia"
## [6] "Sistemas de Informação e Conhecimento"
```

We can see how many students have delivered dissertations for each specialization.

```
data %>%
  group_by('Year' = year(Date), Specialization) %>%
  summarise(
    mean = mean(Grade),
    count = n()
  ) %>%
  ggplot(aes(Year, count, color=Specialization)) +
    geom_line(show.legend = FALSE) +
    geom_text(aes(label=count), show.legend = FALSE) +
    facet_wrap(~Specialization, nrow=3)
```

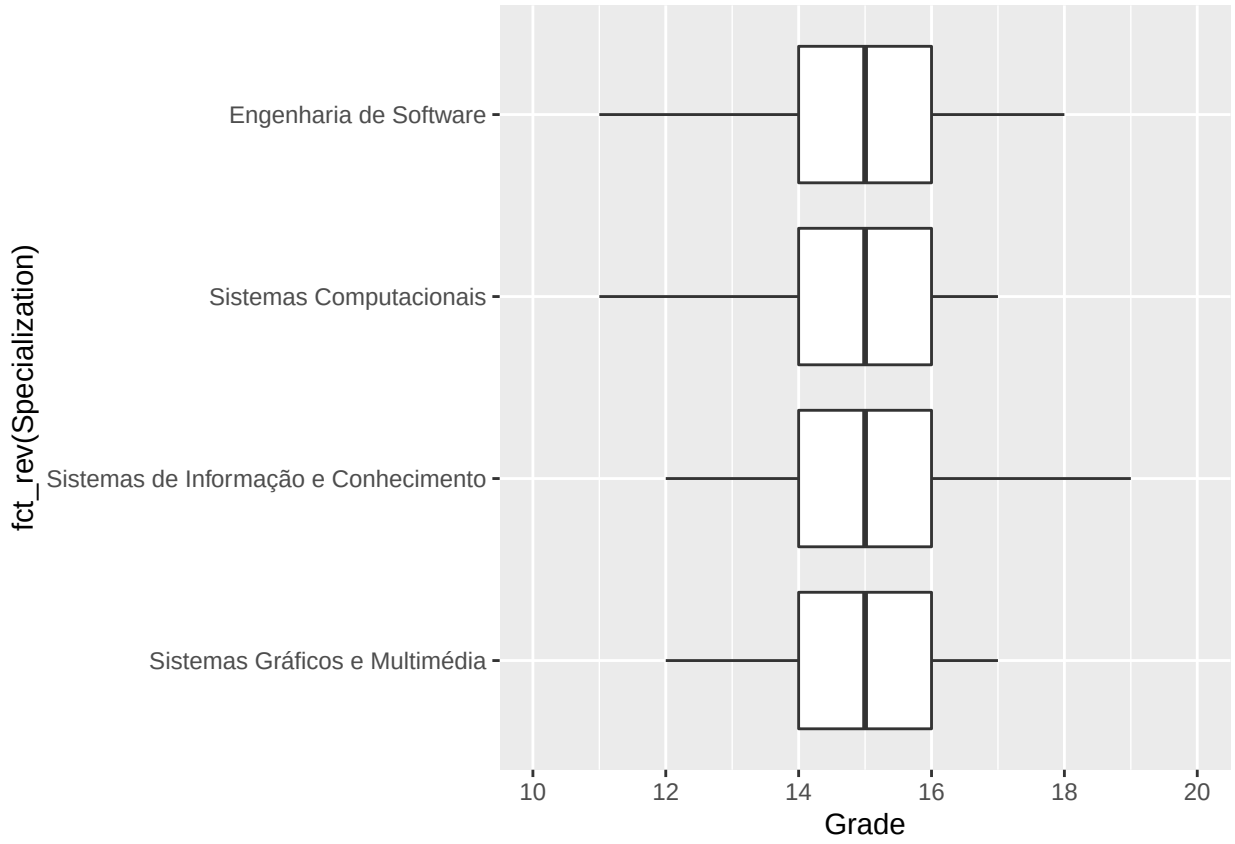


Weird that in 2016 and 2017 no dissertations were delivered in Software Engineering. I know that some specializations were renamed, so we'll apply a transformation.

```r
data <- data %>%
  mutate(
    Specialization = recode(
      Specialization,
      'Arquitecturas, Sistemas e Redes' = 'Sistemas Computacionais',
      'Tecnologias do Conhecimento e Decisão' = 'Sistemas de Informação e Conhecimento'
    )
  )
```

Now let's compare their distributions using boxplots.

```r
data %>%
  ggplot(aes(fct_rev(Specialization), Grade)) +
  geom_boxplot() +
  coord_flip() +
  scale_y_continuous(limits=c(10,20), breaks = seq(10, 20, 2))
```



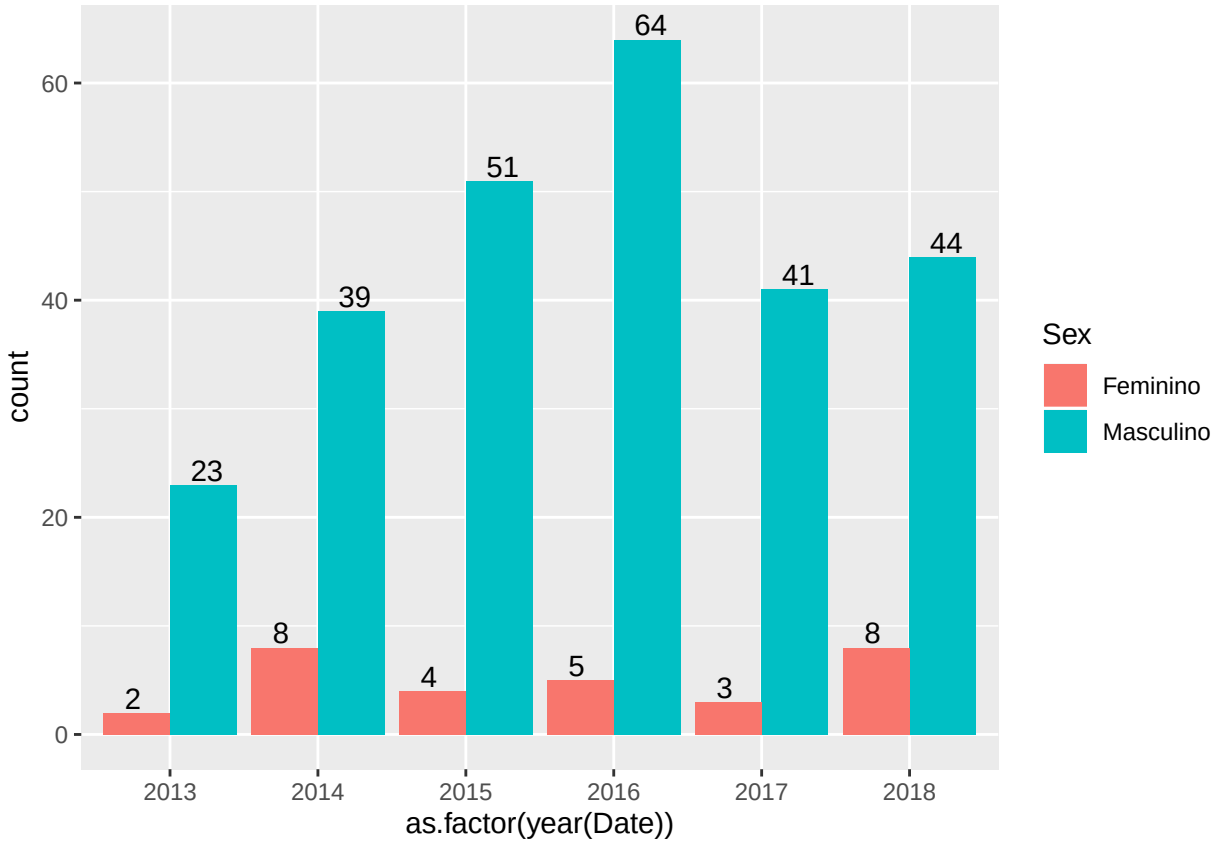

```r
# p-value > 0.05, cannot reject h0 (variance is the same)
anova(lm(Grade ~ Specialization, data = data))
```

```
## Analysis of Variance Table
##
## Response: Grade
##                 Df Sum Sq Mean Sq F value Pr(>F)
## Specialization   3   4.49  1.4961   0.811 0.4887
## Residuals      288 531.30  1.8448
```

All specializations are very similar in terms of grading.

How many men and women have graduated over the years?

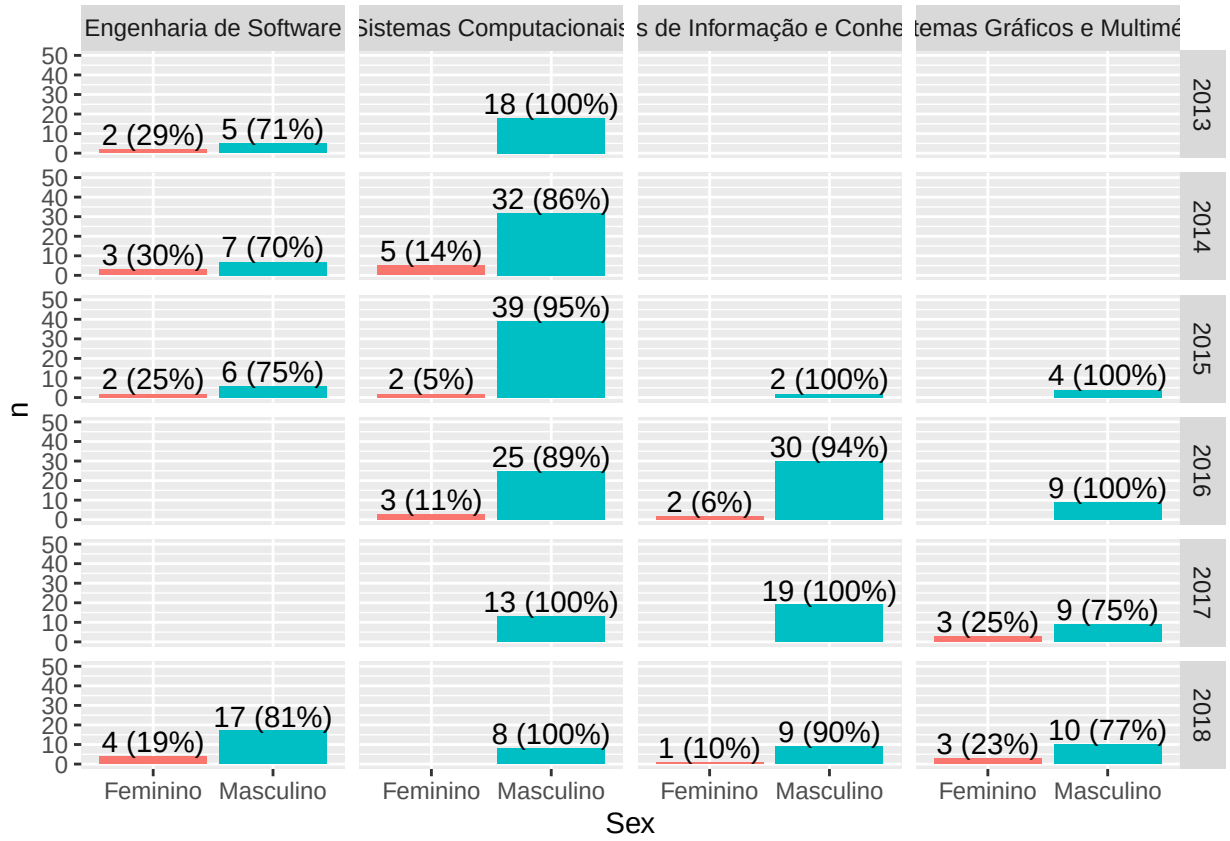
```
ggplot(data, aes(x = as.factor(year(Date)), fill = Sex)) +
  geom_bar(position = 'dodge') +
  geom_text(stat = 'count', aes(label=..count.., vjust = -0.2), position = position_dodge(width = 1))
```



There are 262 male students and 30 female students that finished their disserations between 2013 and 2018.

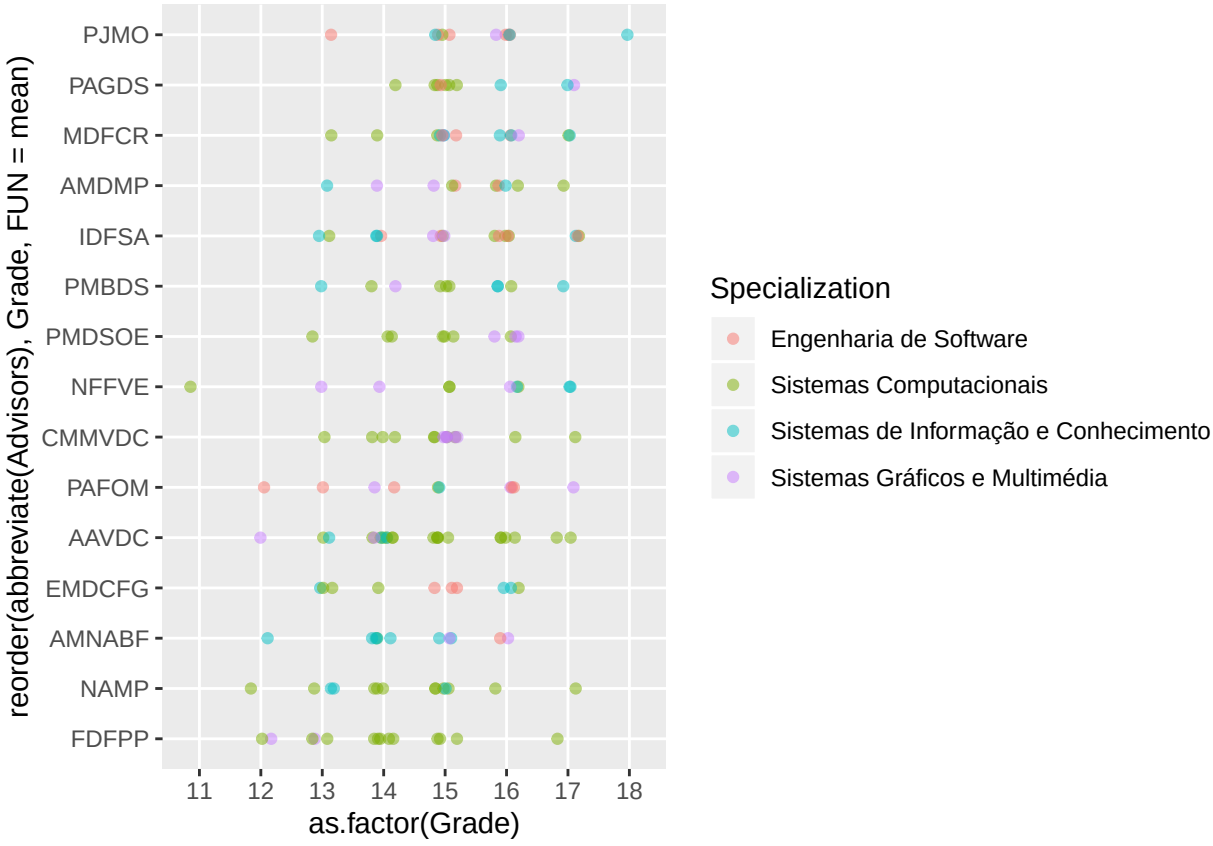
I'm curious about their distribution over the years and specializations.

```
data %>%
  group_by('Year' = year(Date), Specialization) %>%
  count(Sex) %>%
  mutate(freq = round(n / sum(n), 2) * 100) %>%
  ggplot(aes(Sex, n, fill = Sex)) +
    geom_bar(stat = 'identity') +
    geom_text(stat = 'identity', aes(label = paste0(n, ' (', freq, '%)'), vjust = -0.2)) +
    ylim(0, 50) +
    facet_grid(Year ~ Specialization) +
    theme(legend.position = 'none')
```



Do advisors accept students from multiple specializations?

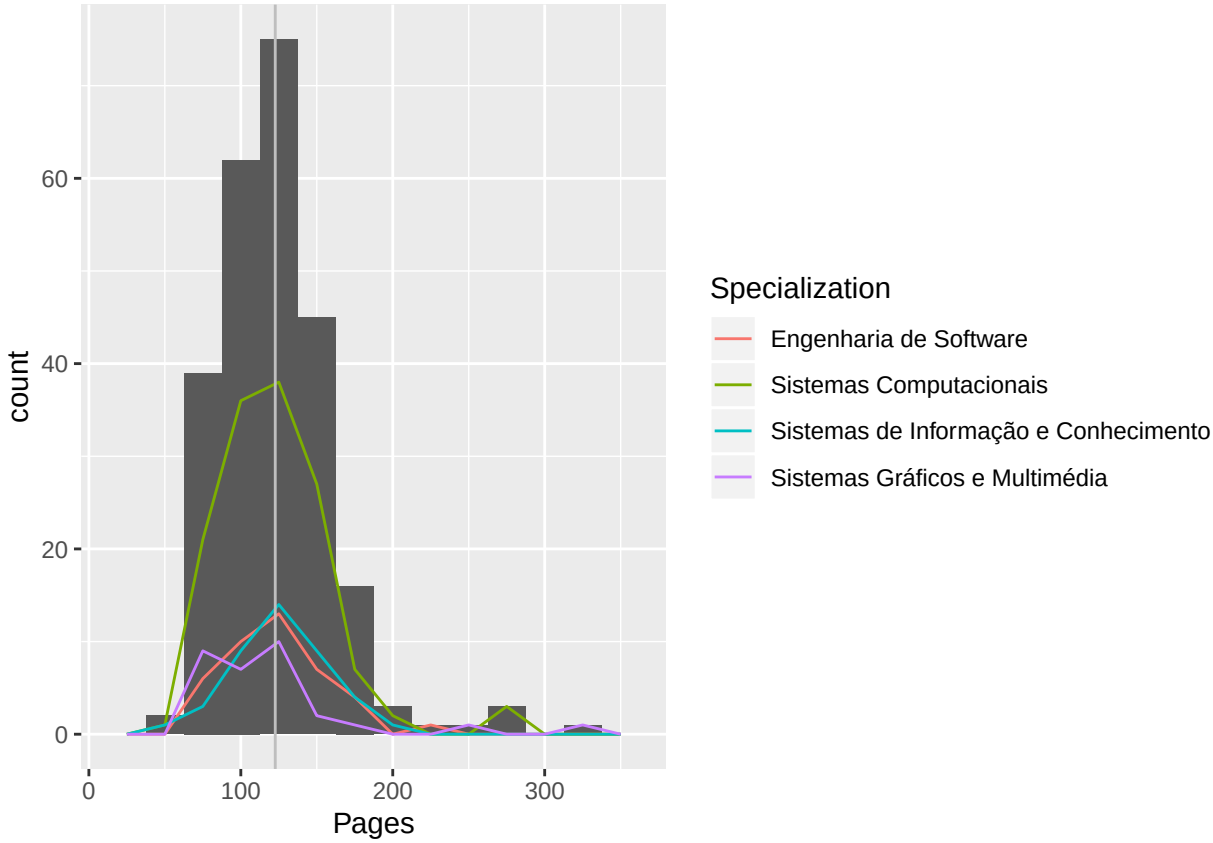
```
topAdvisors <- (data %>% count(Advisors) %>% filter(n > 9))$Advisors
data %>%
  filter(Advisors %in% topAdvisors) %>%
  ggplot(aes(reorder(abbreviate(Advisors), Grade, FUN = mean), as.factor(Grade))) +
  geom_point(alpha = 1/2, position = position_jitter(w = 0, h = 0.2), aes(color=Specialization)) +
  coord_flip()
```



The advisors with most students don't limit themselves to a single specialization.

How many pages do dissertations have on average?

```
data %>%
  ggplot(aes(Pages)) +
  geom_histogram(binwidth = 25) +
  geom_freqpoly(aes(color=Specialization), binwidth = 25) +
  geom_vline(xintercept=mean(data$Pages, na.rm = TRUE), col='gray')
```

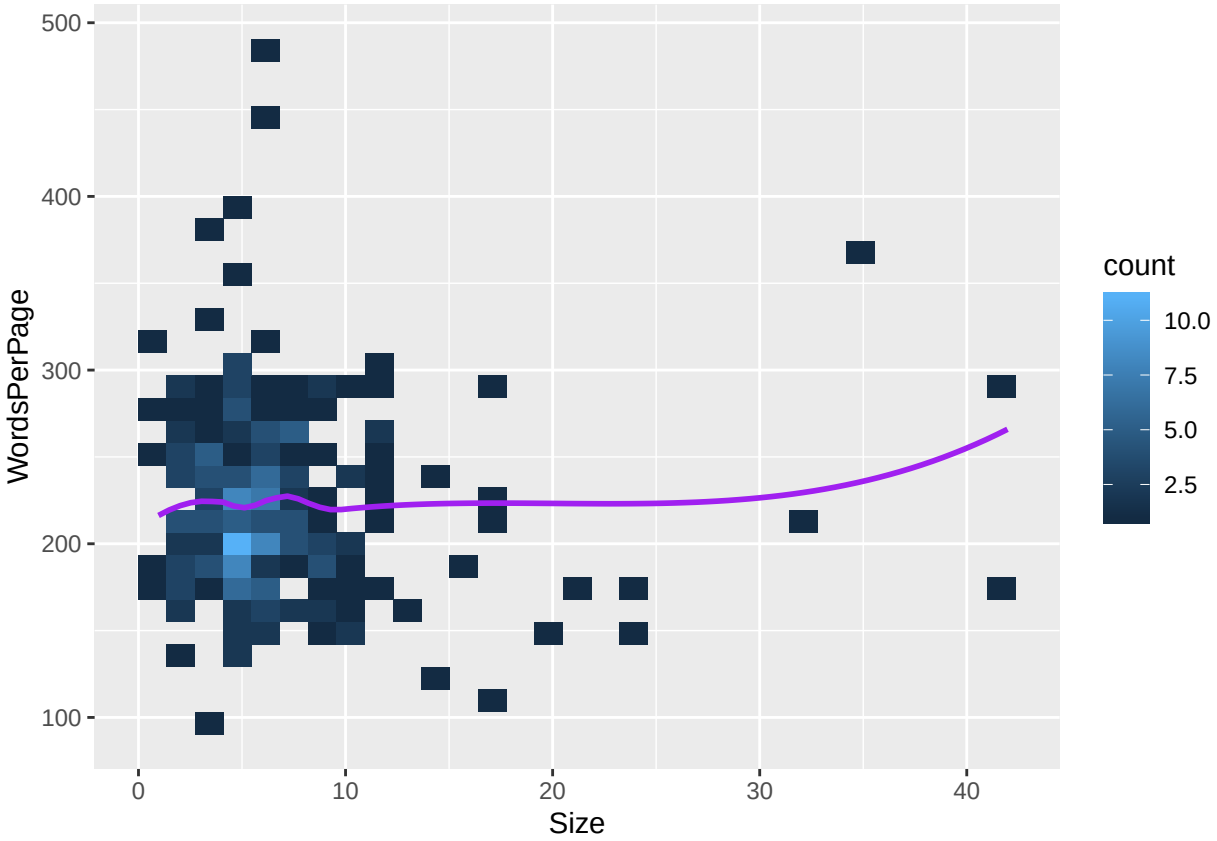



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   59.00   98.75  119.00  122.67  144.00  330.00      44
```

On average the dissertations analyzed have 123 pages.

Is there a correlation between the size and the number of words per page?

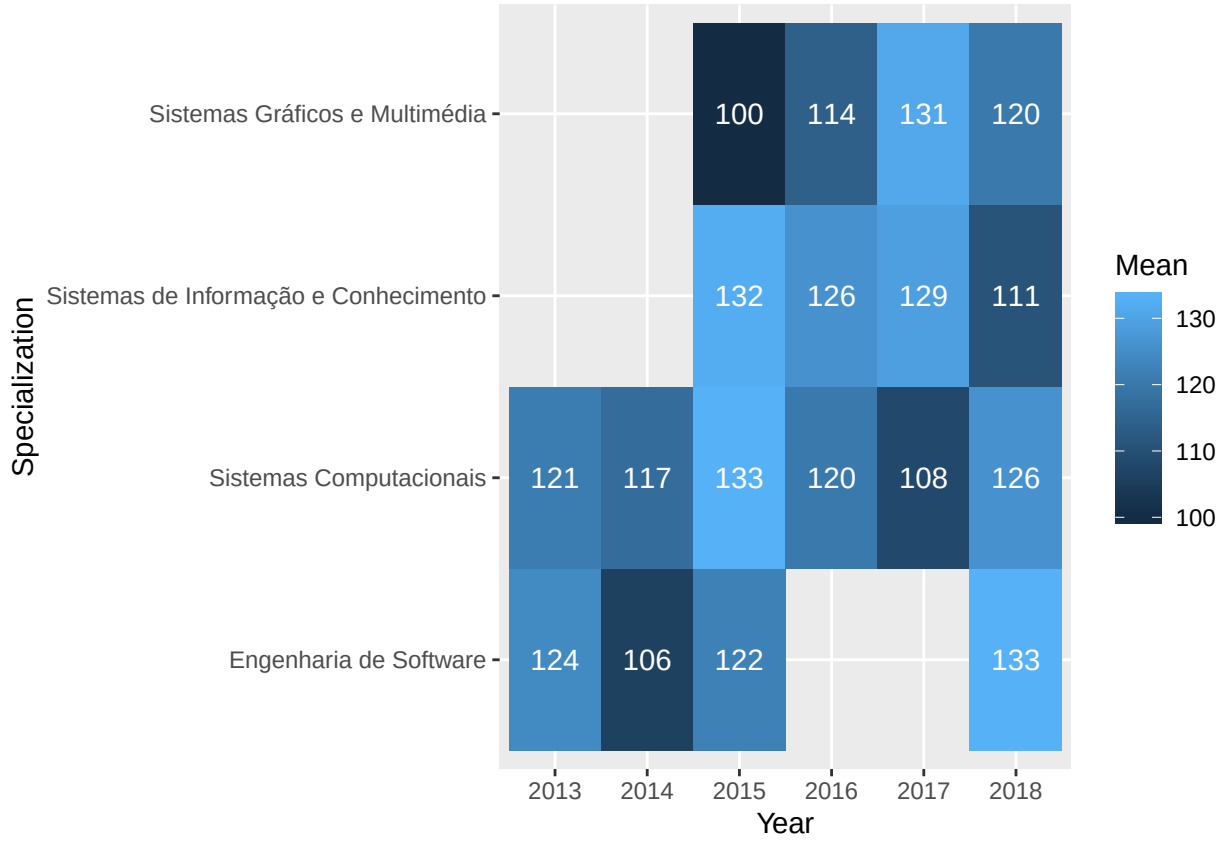
```
data %>%
  ggplot(aes(Size, WordsPerPage)) +
  geom_bin2d() + # to prevent overplotting: binning, transparency, jitter
  geom_smooth(color = 'purple', se = F)
```



The graph shows no correlation between the 2 variables, so dissertations with more pages do not usually have more words per page.

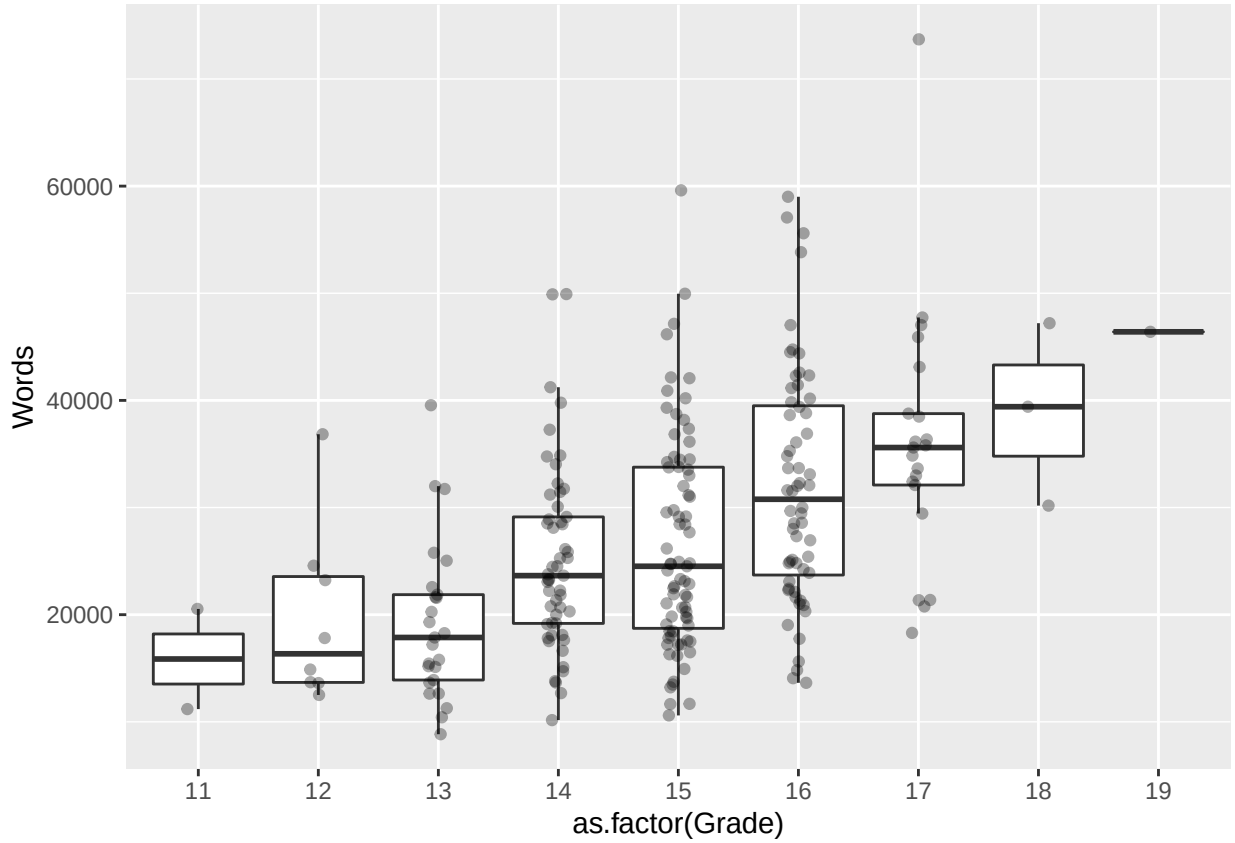
What is the mean number of pages by year and specialization?

```
data %>%
  group_by('Year' = year(Date), Specialization) %>%
  summarise(m = round(mean(Pages, na.rm = TRUE), 0)) %>%
  ggplot(aes(as.factor(Year), Specialization)) +
    geom_tile(mapping = aes(fill = m)) +
    geom_text(aes(label = m), color='white') +
    labs(x = 'Year', fill = 'Mean')
```



How is the distribution of the number of words by each grade?

```
data %>%
  ggplot(aes(as.factor(Grade), Words)) +
  geom_boxplot(outlier.shape = NA) +
  geom_point(alpha = 1/3, position = position_jitter(w = 0.1, h = 0))
```



What are common keywords?

```
wordcloud(data$Keywords, min.freq = 5, random.order = F)
```

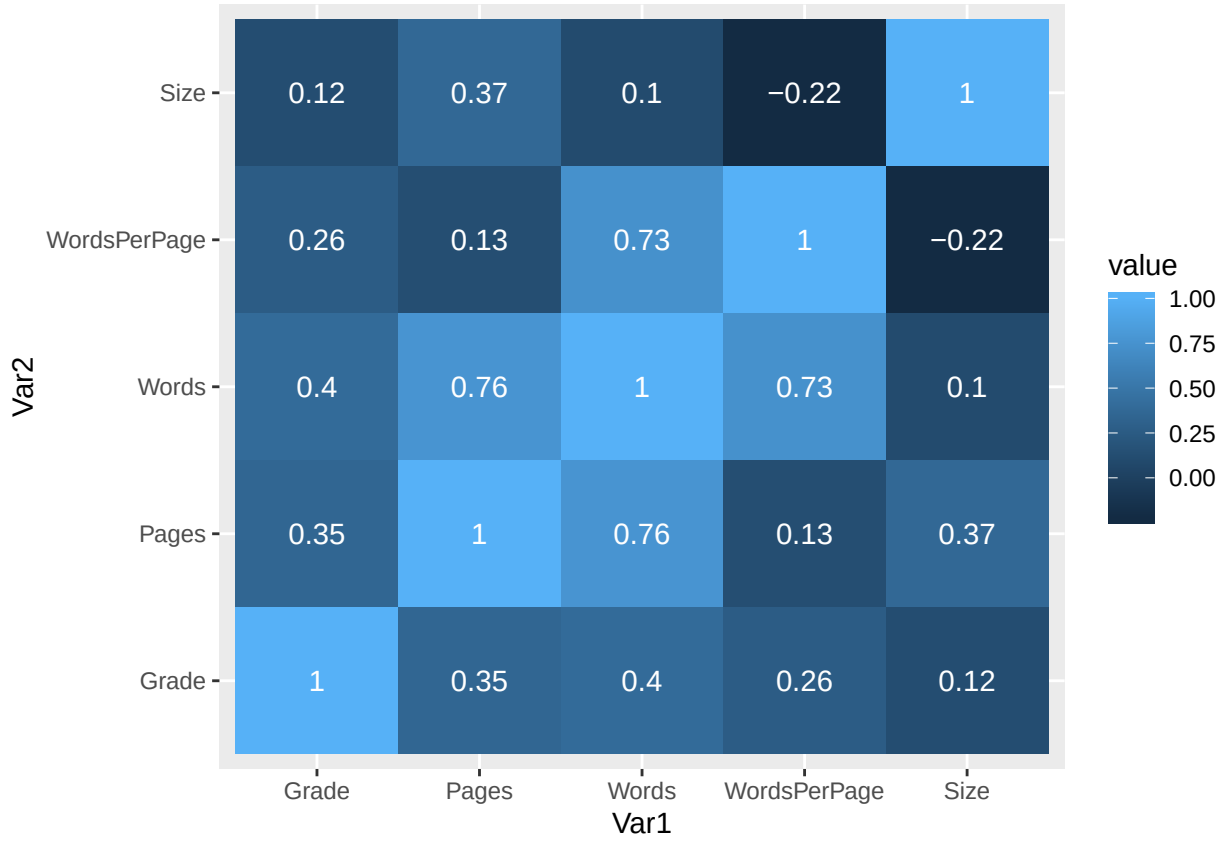
Modelling

Just as an example, I will try to predict something. We need to start by dealing with missing values. I have opted to remove these rows, along with some outliers.

```
data2 <- data %>% filter(!is.na(Words) & Words < 40000)
```

A correlation matrix is helpful to see how variables are related to each other.

```
data2 %>%
  select(Grade, Pages, Words, WordsPerPage, Size) %>%
  cor() %>%
  round(2) %>%
  melt() %>%
  ggplot(mapping = aes(Var1, Var2, fill = value)) +
  geom_tile() +
  geom_text(aes(label = value), color = 'white')
```



As expected there is a strong correlation between the number of pages and the number of words. The calculated metric WordsPerPage also has a strong correlation with the number of words.

Linear Regression

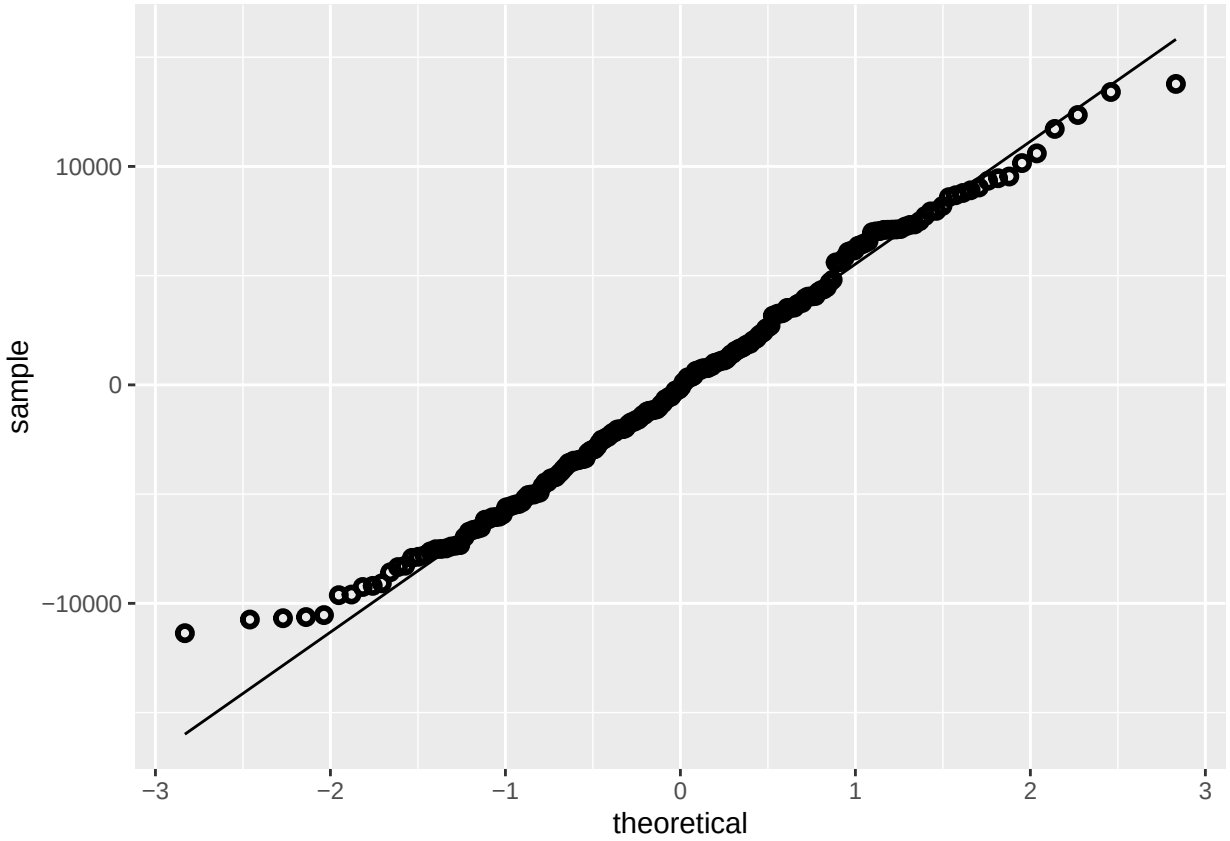
```
mod1 <- lm(Words ~ WordsPerPage + Pages, data = data2) # with + the variables are independent
mod2 <- lm(Words ~ WordsPerPage * Pages, data = data2) # * denotes interactions between vars
mod3 <- lm(Words ~ WordsPerPage, data = data2)

data2 %>%
  gather_predictions(mod1, mod2, mod3) %>%
  ggplot(aes(WordsPerPage, Words, color=Pages)) +
    geom_point(alpha = 0.5) +
    geom_line(aes(y = pred), colour = 'red', size = 1) +
    facet_wrap(~model)
```

The first two models are overfitting, for obvious reasons. I will pick the third model and verify the linear regression model assumptions.

Residuals should be normally distributed with a mean of zero:

```
ggplot(NULL, aes(sample = residuals(mod3))) +
  stat_qq(shape=21, stroke=1.5) +
  stat_qq_line()
```




```
# p-value > 0.05, cannot reject h0 (normal distribution)
shapiro.test(residuals(mod3))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(mod3)
## W = 0.99007, p-value = 0.1441
```

```
# p-value > 0.05, cannot reject h0 (mean = 0)
t.test(residuals(mod3))
```

```
##
##  One Sample t-test
##
## data:  residuals(mod3)
## t = -1.387e-15, df = 215, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -723.8942  723.8942
## sample estimates:
##     mean of x
## -5.094063e-13
```

Residuals should be independent:

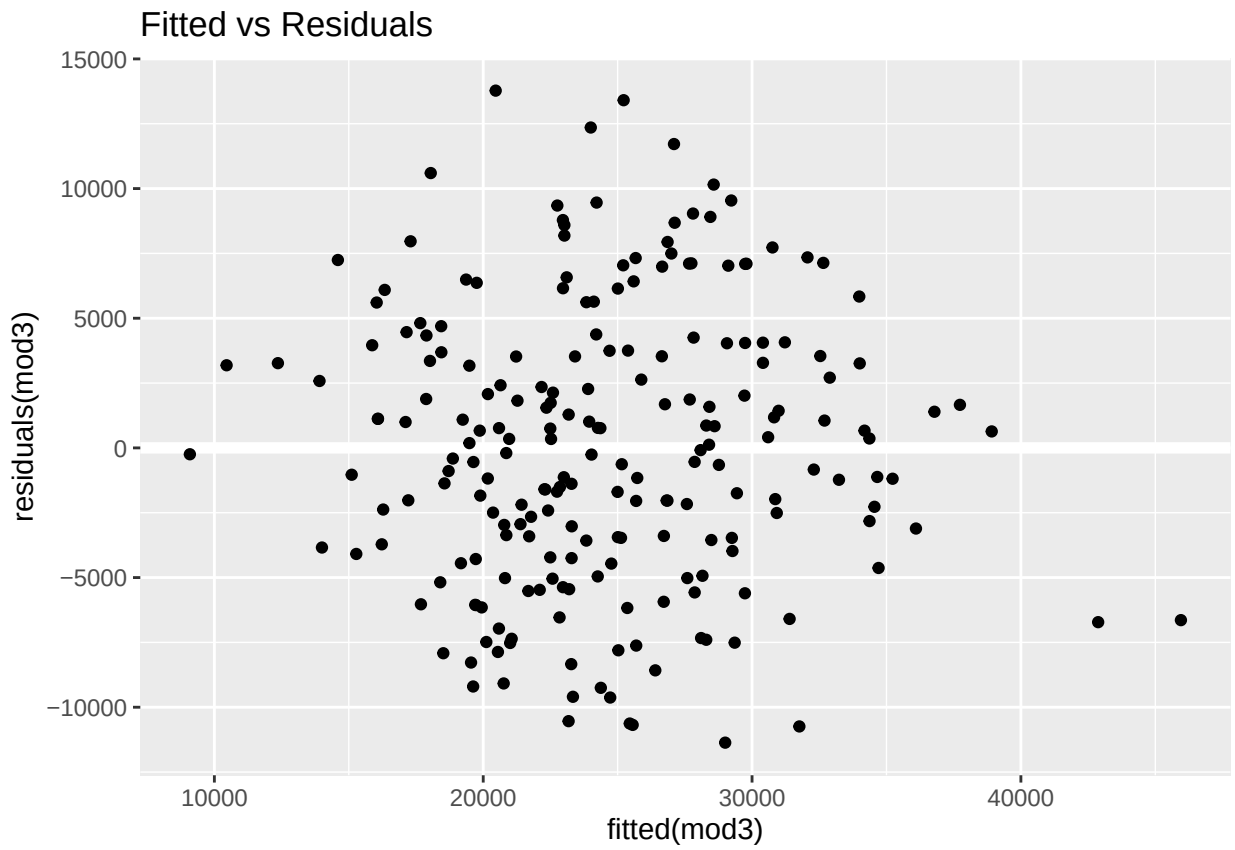
```
library(car)
# p-value > 0.05, cannot reject h0 (residuals are independent)
```

```
durbinWatsonTest(mod3)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.07219216        2.1387   0.306
##  Alternative hypothesis: rho != 0
```

Residuals should have equal variance:

```
ggplot(NULL, aes(fitted(mod3), residuals(mod3))) +
  geom_ref_line(h = 0) +
  geom_point() +
  labs(title = 'Fitted vs Residuals')
```




```
library(lmtest)
# p-value > 0.05, cannot reject h0 (variances are equal)
bptest(mod3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod3
## BP = 0.11644, df = 1, p-value = 0.7329
```

```
library(gvlma)
gvlma(mod3)
```

```
##
```

```
## Call:
## lm(formula = Words ~ WordsPerPage, data = data2)
##
## Coefficients:
##  (Intercept)  WordsPerPage
##      -3483.6         130.8
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = mod3)
##
##                     Value p-value                Decision
## Global Stat        3.8854 0.42174 Assumptions acceptable.
## Skewness           0.6941 0.40478 Assumptions acceptable.
## Kurtosis           2.7670 0.09623 Assumptions acceptable.
## Link Function      0.2367 0.62662 Assumptions acceptable.
## Heteroscedasticity 0.1876 0.66488 Assumptions acceptable.
```

```
summary(mod3)
```

```
##
## Call:
## lm(formula = Words ~ WordsPerPage, data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11366.3  -3875.1   -142.5   3701.5  13776.3
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3483.55    1848.84  -1.884   0.0609 .
## WordsPerPage   130.79       8.45  15.479   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5410 on 214 degrees of freedom
## Multiple R-squared:  0.5282, Adjusted R-squared:  0.526
## F-statistic: 239.6 on 1 and 214 DF,  p-value: < 2.2e-16
```

```
names(summary(mod3))
```

```
##  [1] "call"          "terms"         "residuals"     "coefficients"
##  [5] "aliased"       "sigma"         "df"            "r.squared"
##  [9] "adj.r.squared" "fstatistic"    "cov.unscaled"
```

```
summary(mod3)$adj.r.squared
```

```
## [1] 0.525998
```

All the assumptions have been verified. The Ajusted R-squared shows that 52.6% of the Words variability is explained by the WordsPerPage.

Conclusion

I have recently learned a lot about exploratory data analysis and R. It is helpful to know how to present information to other people using appropriate data visualization techniques. I also want to improve my knowledge about the machine learning domain because it is increasingly being used to solve hard problems in the real world.