Analysis (Lihong)

Lihong Wang

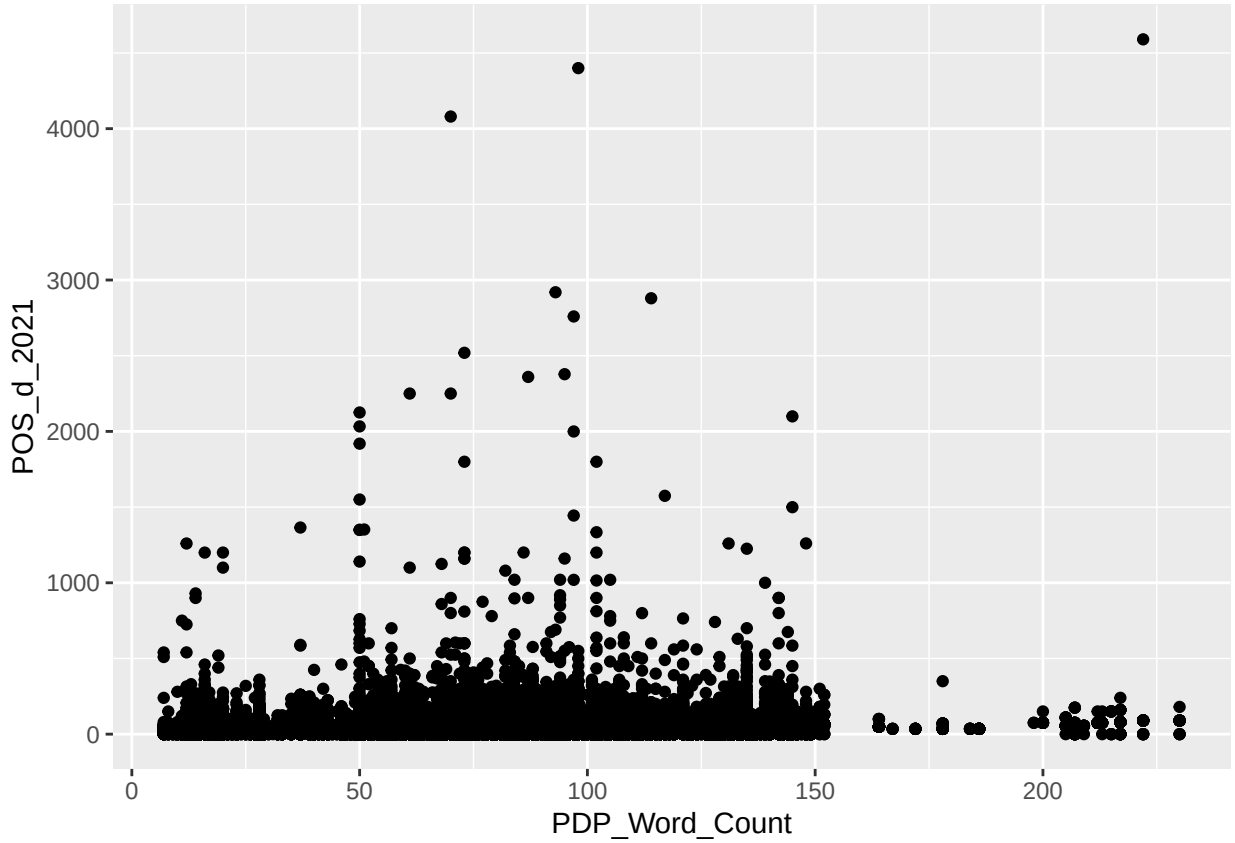
10/5/2021

Load packages & data

```
library(tidyverse)
library(knitr)
library(broom)
library(ggfortify)
library(pairwiseCI)
library(patchwork)
library(ggplot2)
library(readr)
library(reshape2)
library(readxl)
```

```
tdata1 <- read_excel("transaction_data_1-12.xlsx")
tdata2 <- read_excel("transaction_data_13-24.xlsx")
tdata3 <- read_excel("transaction_data_25-37.xlsx")
tdata4 <- read_excel("transaction_data_append.xlsx")
tdata <- data.frame(tdata1, tdata2, tdata3, tdata4)
cdata <- read_excel("convfunnel_data.xlsx")
tdata <- na.omit(tdata)
cdata <- na.omit(cdata)
```

```
overview <- ggplot(data = tdata, aes(x = PDP_Word_Count, y = POS_d_2021)) +
  geom_point()
overview
```



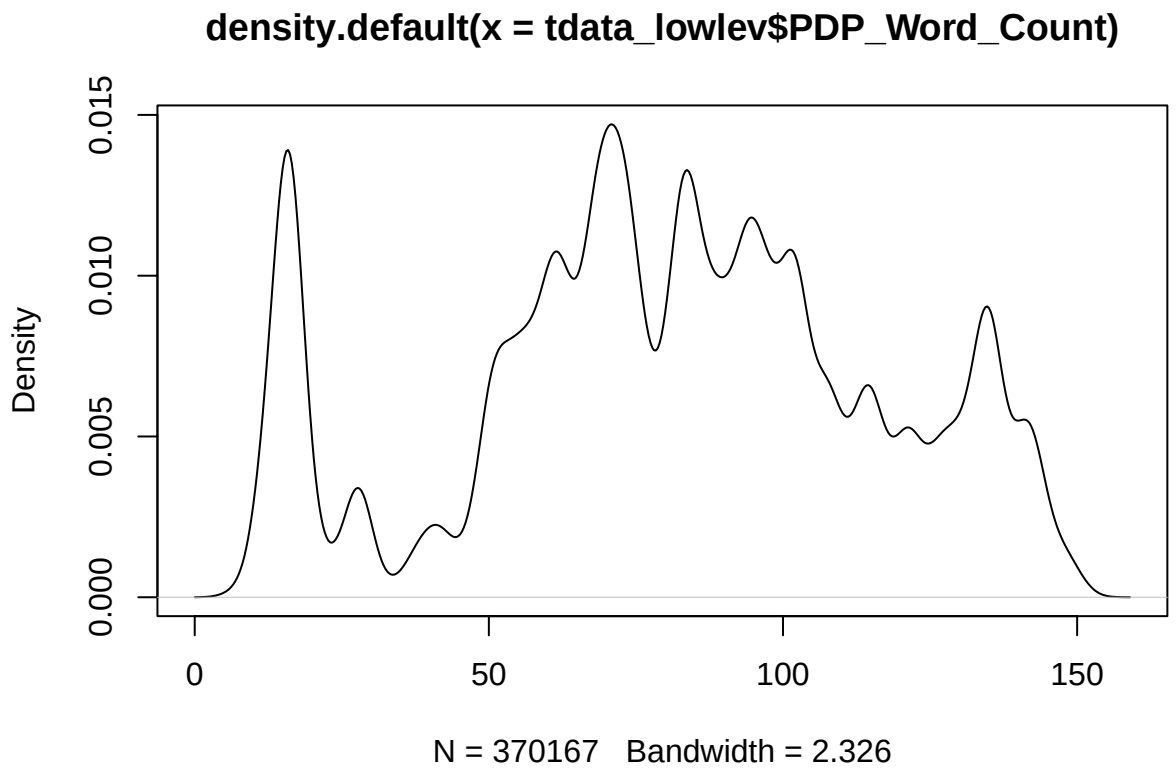

```
tdata_lowlev <- tdata %>% filter(POS_p_2021 < 10 & POS_d_2021 < 155)
tdata_highlev <- tdata %>% filter(POS_p_2021 >= 10 | POS_d_2021 >= 155)
tdata_lowlev <- tdata_lowlev %>% filter(PDP_Word_Count < 155)
tdata_highlev <- tdata_highlev %>% filter(PDP_Word_Count < 155)
```

```
lowlev_sales <- sum(tdata_lowlev$POS_d_2021)
highlev_sales <- sum(tdata_highlev$POS_d_2021)
(highlev_sales/(lowlev_sales+highlev_sales))*100
```

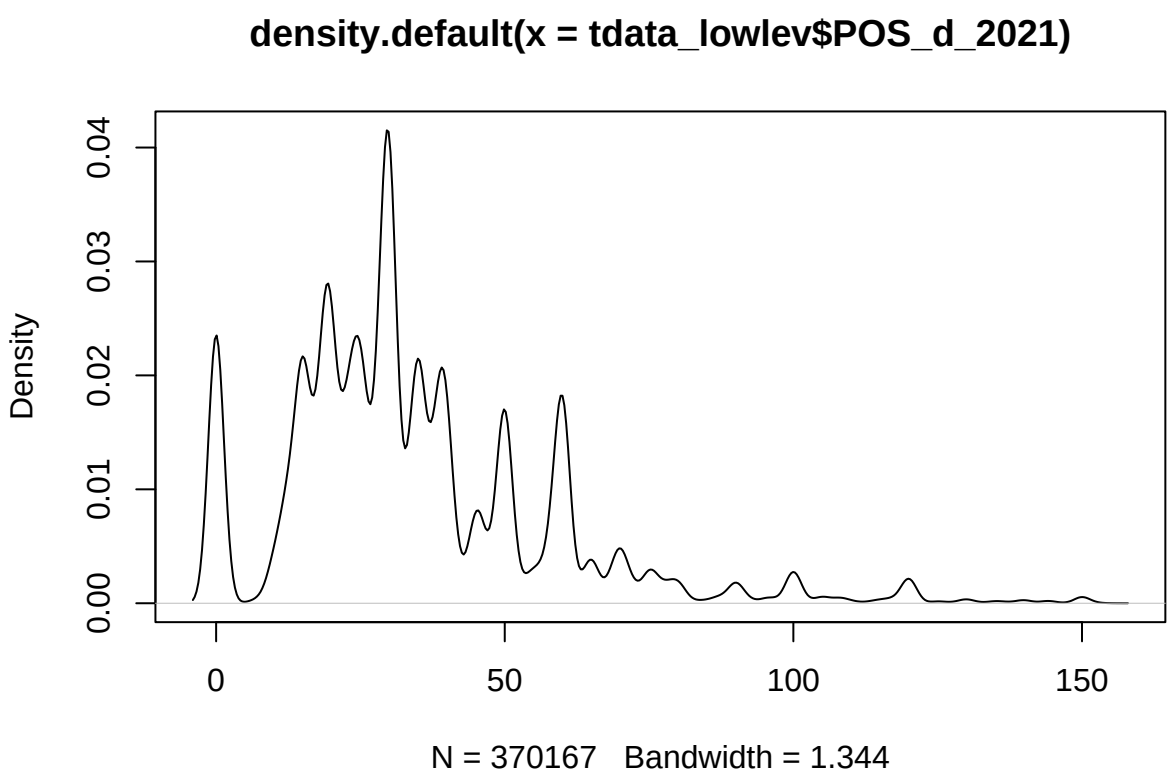
```
## [1] 4.539638
```

```
density_word <- density(tdata_lowlev$PDP_Word_Count)
density_dollars <- density(tdata_lowlev$POS_d_2021)

plot(density_word)
```

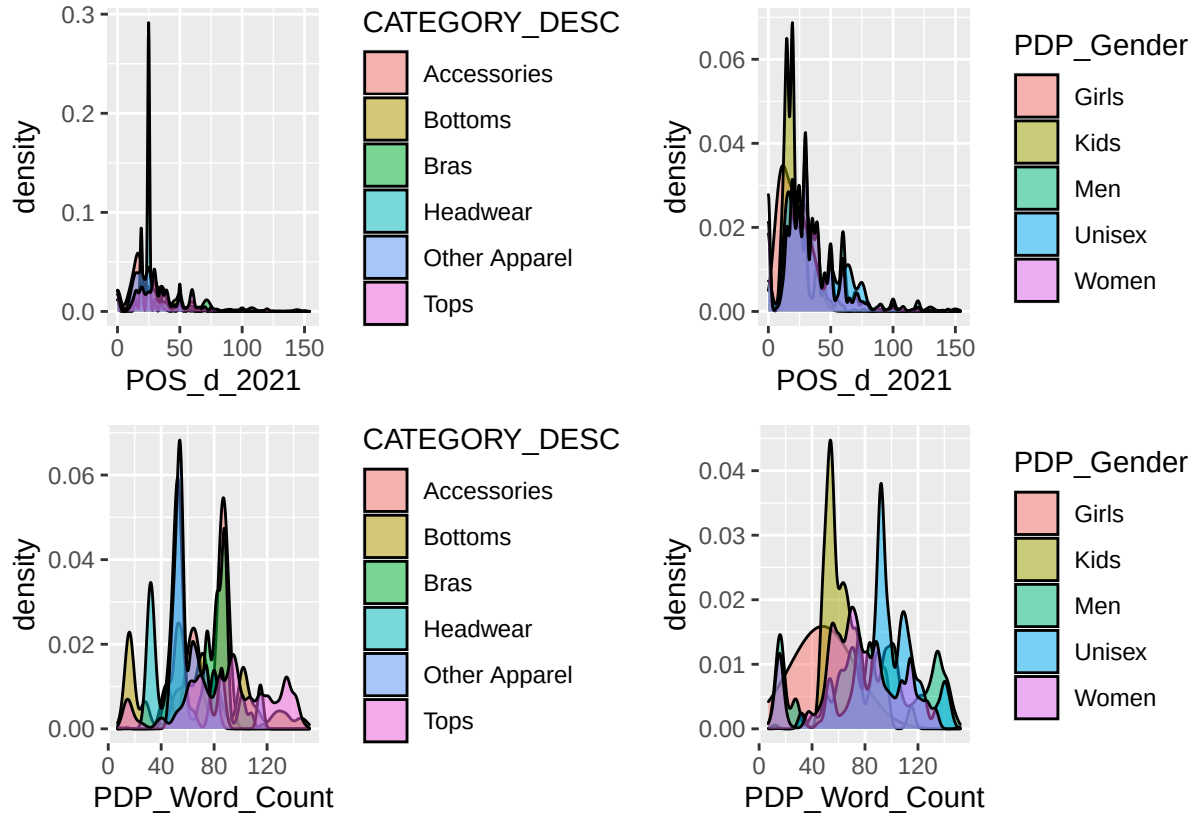



```
plot(density_dollars)
```




```
p3 <- ggplot(data = tdata_lowlev, aes(x = POS_d_2021, fill = CATEGORY_DESC)) +
  geom_density(alpha = 0.5)
p4 <- ggplot(data = tdata_lowlev, aes(x = POS_d_2021, fill = PDP_Gender)) +
  geom_density(alpha = 0.5)
p5 <- ggplot(data = tdata_lowlev, aes(x = PDP_Word_Count, fill = CATEGORY_DESC)) +
  geom_density(alpha = 0.5)
p6 <- ggplot(data = tdata_lowlev, aes(x = PDP_Word_Count, fill = PDP_Gender)) +
  geom_density(alpha = 0.5)

(p3 + p4)/(p5 + p6)
```




```
Word_anova <- aov(PDP_Word_Count ~ PDP_Gender, data = tdata_lowlev)
tidy(Word_anova)
```

```
## # A tibble: 2 x 6
##   term           df      sumsq   meansq statistic p.value
##   <chr>       <dbl>      <dbl>    <dbl>     <dbl>   <dbl>
## 1 PDP_Gender      4  12611328. 3152832.     2581.       0
## 2 Residuals  370162 452115022.    1221.       NA       NA
```

```
# 1-(0.05/5) = 0.99 Bonferroni correction
pairwiseCI(PDP_Word_Count ~ PDP_Gender, data = tdata_lowlev,
           conf.level = 0.99, var.equal = TRUE)
```

```
##
## 99 %-confidence intervals
##  Method:  Difference of means assuming Normal distribution and equal variances
##
##
##               estimate   lower   upper
## Kids-Girls      16.934   5.895  27.973
## Men-Girls       35.134   9.411  60.857
## Unisex-Girls    47.539  32.365  62.713
## Women-Girls     26.305   6.544  46.065
## Men-Kids        18.201  16.597  19.804
## Unisex-Kids     30.605  29.623  31.588
## Women-Kids       9.371   8.135  10.607
## Unisex-Men      12.405  11.762  13.047
```

```
## Women-Men      -8.829  -9.161  -8.498
## Women-Unisex  -21.234 -21.748 -20.721
##
##
```

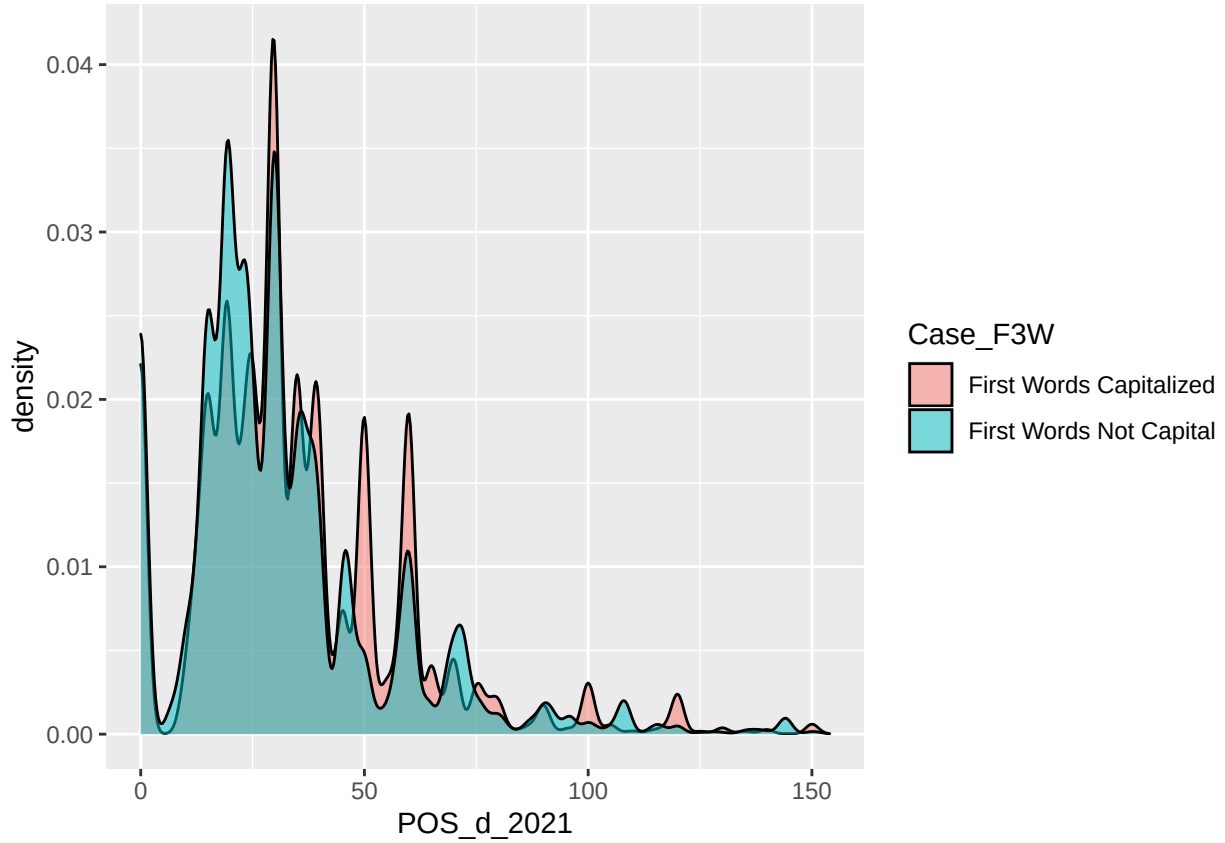
```
Sales_anova <- aov(POS_d_2021 ~ PDP_Gender, data = tdata_lowlev)
tidy(Sales_anova)
```

```
## # A tibble: 2 x 6
##   term           df      sumsq  meansq statistic p.value
##   <chr>       <dbl>      <dbl>   <dbl>     <dbl>   <dbl>
## 1 PDP_Gender      4   1192294. 298073.      539.       0
## 2 Residuals  370162 204816716.    553.       NA       NA
```

```
# 1-(0.05/5) = 0.99 Bonferroni correction
pairwiseCI(POS_d_2021 ~ PDP_Gender, data = tdata_lowlev,
           conf.level = 0.99, var.equal = TRUE)
```

```
##
## 99 %-confidence intervals
##  Method:  Difference of means assuming Normal distribution and equal variances
##
##
##               estimate   lower   upper
## Kids-Girls      0.9073 -7.7616  9.5761
## Men-Girls      15.3718 -0.4799 31.2235
## Unisex-Girls   15.7735 -0.5048 32.0518
## Women-Girls    13.0059 -2.2838 28.2957
## Men-Kids       14.4645 13.4752 15.4538
## Unisex-Kids    14.8662 13.8310 15.9014
## Women-Kids     12.0987 11.1419 13.0554
## Unisex-Men      0.4017 -0.0090  0.8124
## Women-Men      -2.3658 -2.5840 -2.1477
## Women-Unisex   -2.7675 -3.1846 -2.3504
##
##
```

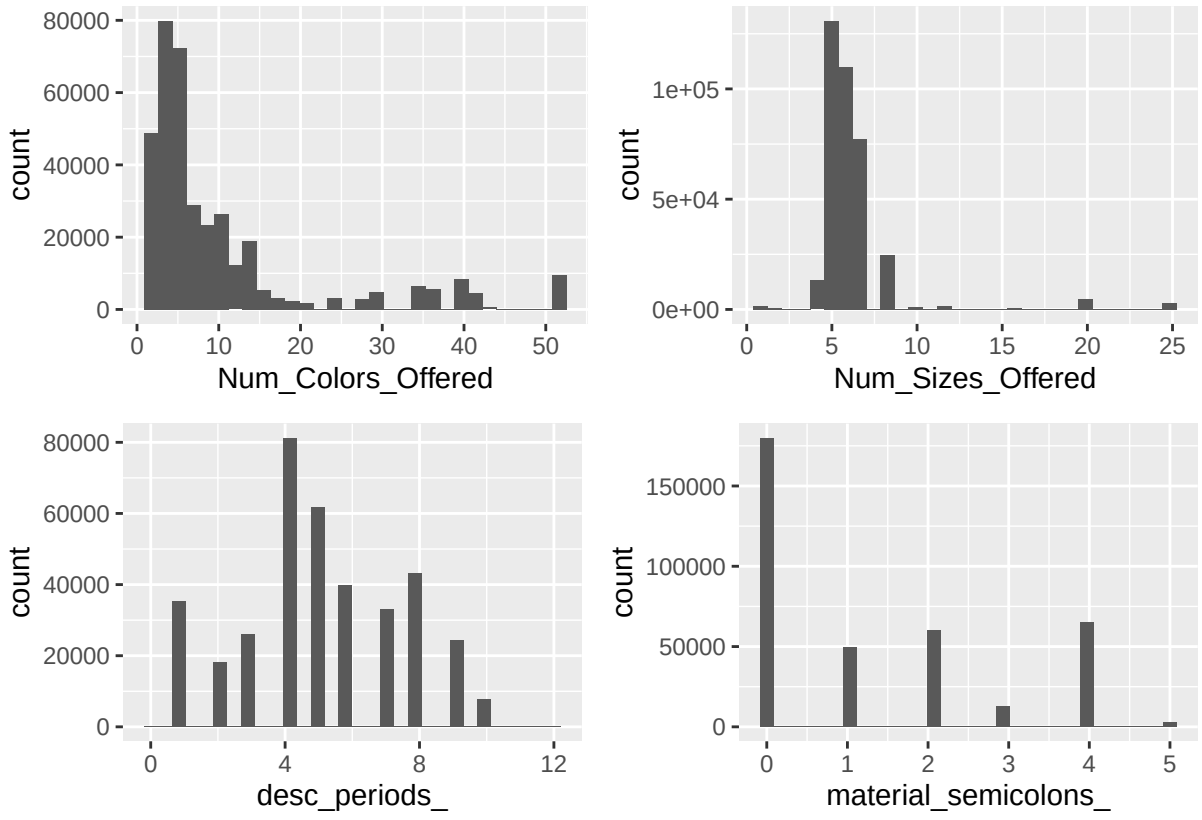
```
ggplot(data = tdata_lowlev, aes(x = POS_d_2021, fill = Case_F3W)) +
  geom_density(alpha = 0.5)
```




```
# tdata_Women <- tdata_lowlev %>% filter(PDP_Gender == "Women")
```

```
p7 <- ggplot(data = tdata_lowlev, aes(x = Num_Colors_Offered)) +
  geom_histogram()
p8 <- ggplot(data = tdata_lowlev, aes(x = Num_Sizes_Offered)) +
  geom_histogram()
p9 <- ggplot(data = tdata_lowlev, aes(x = desc_periods_)) +
  geom_histogram()
p10 <- ggplot(data = tdata_lowlev, aes(x = material_semicolons_)) +
  geom_histogram()

(p7+p8)/(p9+p10)
```




```
tdata_lowlev <- tdata_lowlev %>%
  mutate(Squre_PDP_Word_Count = (PDP_Word_Count)^2)
```