Individual Assignment

Jack

2022/3/12

```
library(lubridate)
library(dplyr)
library(ggplot2)
```

1. Load projects.csv data.

```
data = read.csv("./projects.csv")
head(data)
```

```
##   id                       project_id teacher_id school_id school_name
## 1 45 d6fab27a79bb2ecf24aa4b915ba8be6f         43        42        HIST
## 2 50 3a19e83d3aa3f7872a7b4f1adbdbf383         48        47        HIST
## 3 53 b48e360a45375639e9c9b2bcdd8cedca         43        42        HIST
## 4 79 ee7a5e6cf495f2a4148889ff13d5bed0         74        70        HIST
## 5 82 3589ed60c7377545ed045f4236ea6b09         77        45        HIST
## 6 90 d8b7edd7234b9f228e553cdcdbc72e61         85        38        HIST
##   primary_focus_subject primary_focus_area secondary_focus_subject
## 1                     4                 1                      NA
## 2                     2                 2                      10
## 3                     4                 1                      20
## 4                     2                 2                       4
## 5                     1                 1                      20
## 6                     2                 2                      NA
##   secondary_focus_area resource_usage resource_type poverty_level   grade_level
## 1                   NA      essential      Supplies          high   Grades 9-12
## 2                    2     enrichment         Other          high    Grades 3-5
## 3                    4     enrichment      Supplies          high   Grades 9-12
## 4                    1      essential      Supplies          high Grades PreK-2
## 5                    4     enrichment      Supplies           low Grades PreK-2
## 6                   NA      essential         Books          high Grades PreK-2
##   vendor_shipping_charges sales_tax payment_processing_charges
## 1                     0.0      0.00                       5.21
## 2                    24.7     16.92                       3.70
## 3                     0.0      0.00                       5.35
## 4                     0.0     26.37                       4.32
## 5                     0.0     34.81                       5.71
## 6                    12.0     33.99                       5.57
##   fulfillment_labor_materials total_price_excluding_optional_support
## 1                          35                                387.57
```

```
## 2                          35                                327.30
## 3                          35                                396.79
## 4                          35                                353.92
## 5                          35                                455.95
## 6                          35                                457.99
##   total_price_including_optional_support students_reached total_donations
## 1                                 455.96               90              NA
## 2                                 385.06               22              NA
## 3                                 466.81               90              NA
## 4                                 416.38               24              NA
## 5                                 536.41               32              NA
## 6                                 538.81               24              NA
##   num_donors eligible_double_your_impact_match eligible_almost_home_match
## 1         NA                                 0                          0
## 2         NA                                 0                          0
## 3         NA                                 0                          0
## 4         NA                                 0                          0
## 5         NA                                 0                          0
## 6         NA                                 0                          0
##   funding_status date_posted date_completed date_thank_you_packet_mailed
## 1           live  02/16/2012          <NA>                         <NA>
## 2           live  02/16/2012          <NA>                         <NA>
## 3           live  02/16/2012          <NA>                         <NA>
## 4           live  02/16/2012          <NA>                         <NA>
## 5           live  02/16/2012          <NA>                         <NA>
## 6           live  02/16/2012          <NA>                         <NA>
##   date_expiration margin margin_percentage
## 1      07/14/2012     NA                NA
## 2      07/14/2012     NA                NA
## 3      07/14/2012     NA                NA
## 4      07/14/2012     NA                NA
## 5      07/13/2012     NA                NA
## 6      07/13/2012     NA                NA
##   summed_donations_excluding_optional_support
## 1                                          NA
## 2                                          NA
## 3                                          NA
## 4                                          NA
## 5                                          NA
## 6                                          NA
##   summed_donations_including_optional_support total_primary total_limited
## 1                                          NA            NA            NA
## 2                                          NA            NA            NA
## 3                                          NA            NA            NA
## 4                                          NA            NA            NA
## 5                                          NA            NA            NA
## 6                                          NA            NA            NA
##   total_matched total_primary_base total_limited_base total_matched_base
## 1            NA                NA                NA                NA
## 2            NA                NA                NA                NA
## 3            NA                NA                NA                NA
## 4            NA                NA                NA                NA
## 5            NA                NA                NA                NA
## 6            NA                NA                NA                NA
```

```
##   percent_primary percent_limited percent_matched percent_total date_ended
## 1              NA              NA              NA            NA 07/14/2012
## 2              NA              NA              NA            NA 07/14/2012
## 3              NA              NA              NA            NA 07/14/2012
## 4              NA              NA              NA            NA 07/14/2012
## 5              NA              NA              NA            NA 07/13/2012
## 6              NA              NA              NA            NA 07/13/2012
```

2. Create a new column called "project_order" that shows how many projects a teacher has created including the given project.

(1) Change the data type of *date_posted* column into date

```
# install.packages("lubridate")
library("lubridate")
data$date_posted<-mdy(data$date_posted)
class(data$date_posted)
```

```
## [1] "Date"
```

(2) Create a new column called "project_order" that shows how many projects a teacher has created including the given project.

```
data$project_order<-1
data %>%
  group_by(teacher_id) %>%
  arrange(teacher_id) %>%
  mutate(project_order = cumsum(project_order)) -> data
head(data[c("teacher_id", "date_posted", "project_order")], 20)
```

```
## # A tibble: 20 x 3
## # Groups:   teacher_id [4]
##    teacher_id date_posted project_order
##         <int> <date>              <dbl>
##  1         43 2012-02-16              1
##  2         43 2012-02-16              2
##  3         43 2010-12-13              3
##  4         43 2010-12-11              4
##  5         43 2009-09-10              5
##  6         43 2009-09-03              6
##  7         48 2012-02-16              1
##  8         74 2012-02-16              1
##  9         74 2012-02-11              2
## 10         74 2011-02-25              3
## 11         74 2010-11-22              4
## 12         74 2010-04-17              5
## 13         74 2009-10-08              6
## 14         77 2012-02-16              1
## 15         77 2011-12-06              2
## 16         77 2011-12-06              3
## 17         77 2011-11-21              4
## 18         77 2011-09-26              5
```

```
## 19         77 2011-07-20              6
## 20         77 2011-07-07              7
```

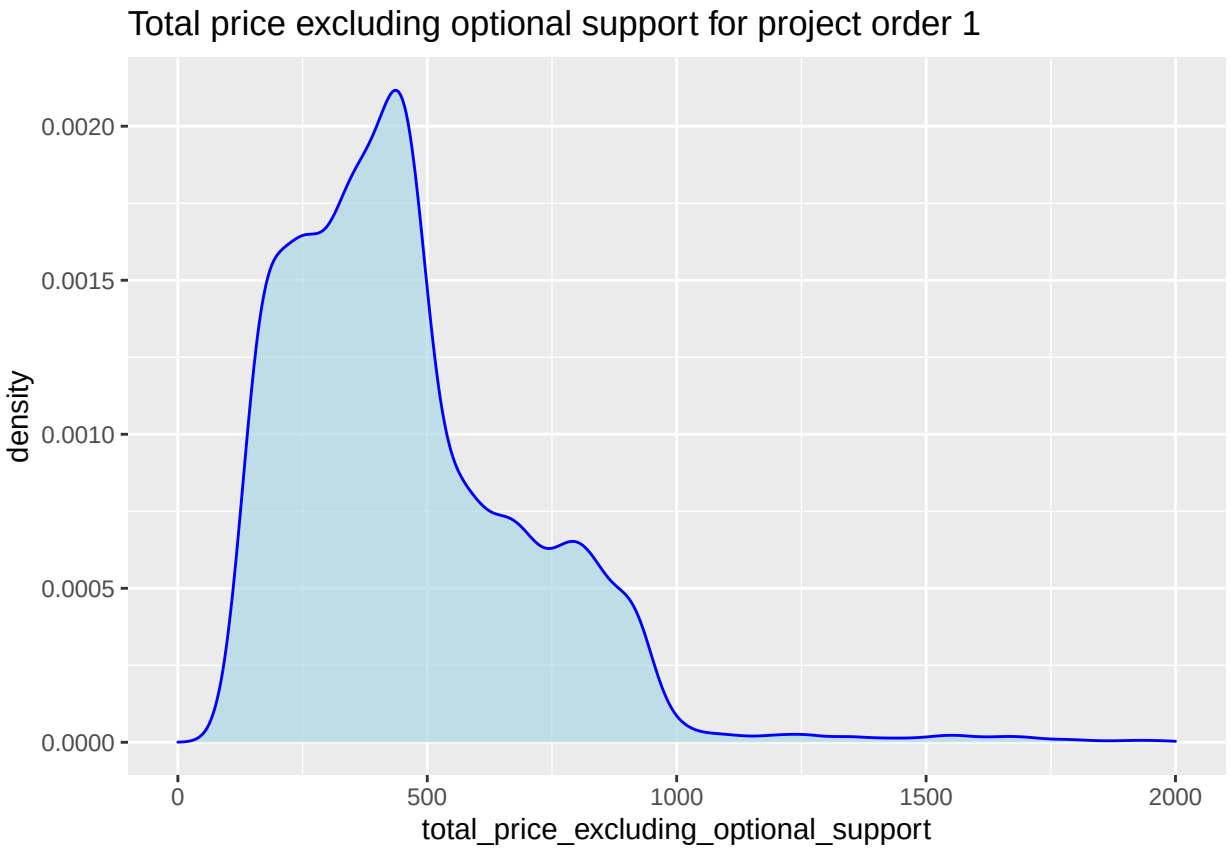
(3) Recode all values of *project_order* larger than 5 to 6. In other words, after the manipulation, the value of 6 in the *project_order* column should mean 6 or larger numbers.

```
data[which(data$project_order>=6), 'project_order'] <- 6
head(data[c("teacher_id", "date_posted", "project_order")], 20)
```

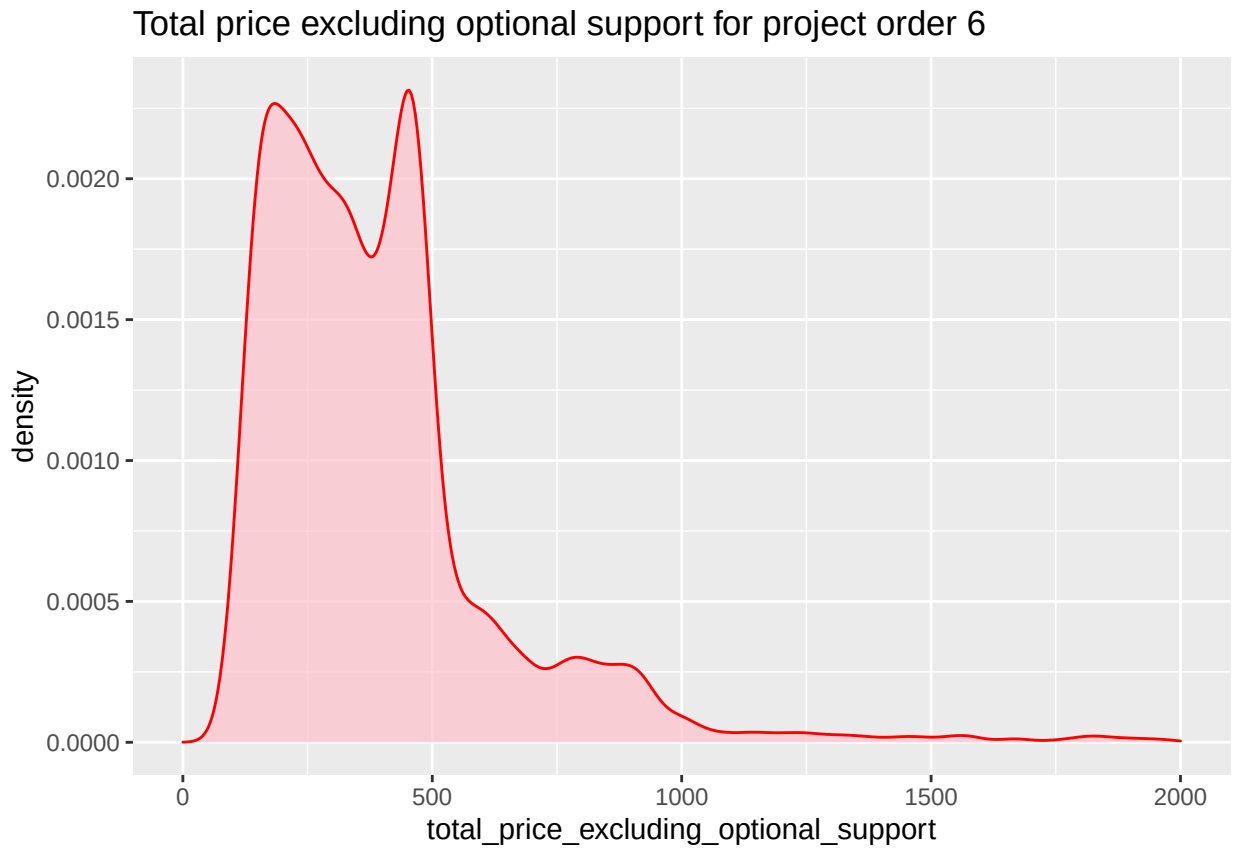
```
## # A tibble: 20 x 3
## # Groups:   teacher_id [4]
##    teacher_id date_posted project_order
##         <int> <date>              <dbl>
##  1         43 2012-02-16              1
##  2         43 2012-02-16              2
##  3         43 2010-12-13              3
##  4         43 2010-12-11              4
##  5         43 2009-09-10              5
##  6         43 2009-09-03              6
##  7         48 2012-02-16              1
##  8         74 2012-02-16              1
##  9         74 2012-02-11              2
## 10         74 2011-02-25              3
## 11         74 2010-11-22              4
## 12         74 2010-04-17              5
## 13         74 2009-10-08              6
## 14         77 2012-02-16              1
## 15         77 2011-12-06              2
## 16         77 2011-12-06              3
## 17         77 2011-11-21              4
## 18         77 2011-09-26              5
## 19         77 2011-07-20              6
## 20         77 2011-07-07              6
```

3. Graph the density of project sizes only for project_order==1 and project_order==6. As a variable for project sizes, use *total_price_excluding_optoinal_support*. If you use ggplot2 library, you can use the *geom_density* function.

```
ggplot() +
  geom_density(data=data[data$project_order==1,], mapping=aes(total_price_excluding_optional_support), a
  xlim(0,2000) +
  ggtitle( "Total price excluding optional support for project order 1")
```




```
ggplot() +
  geom_density(data=data[data$project_order==6,], mapping=aes(total_price_excluding_optional_support), a
  xlim(0,2000) +
  ggtitle("Total price excluding optional support for project order 6")
```




```
print("Test the distribution of project_order equal to 1")
```

```
## [1] "Test the distribution of project_order equal to 1"
```

```
data %>%
  filter(project_order == 1) %>%
  select(total_price_excluding_optional_support) %>%
  ks.test("pnorm")
```

```
## Adding missing grouping variables: `teacher_id`
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  .
## D = 1, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
print("Test the distribution of project_order equal to 6")
```

```
## [1] "Test the distribution of project_order equal to 6"
```

```
data %>%
  filter(project_order == 6) %>%
  select(total_price_excluding_optional_support) %>%
  ks.test("pnorm")
```

```
## Adding missing grouping variables: `teacher_id`
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  .
## D = 1, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Ans1. We can see that the p-value of both of two distribution is very small, which means that there is enough strong evidence to reject the hypothesis that the distribution is normal distribution. Therefore, I do not think these distribution are close to normal density.

Ans2. The distribution with project_order = 1 has only one peak, while the distribution with project_order = 6 has two peaks.

4. Demonstrate that Chebychev's inequality holds for the distribution of project sizes.

```
project_size <- data$total_price_excluding_optional_support
size_mean <- mean(project_size)
size_sd <- sd(project_size)
poss_2 <- sum((size_mean - 2*size_sd) < project_size & project_size < (size_mean + 2*size_sd))/length(pr

poss_3 <- sum((size_mean - 3*size_sd) < project_size & project_size < (size_mean + 3*size_sd))/length(pr

poss_2
```

```
## [1] 0.9999583
```

```
poss_3
```

```
## [1] 0.9999583
```

We can find that when k = 2, the data in the two standard deviations of the mean is larger than 75%. When k = 3, the data in the three standard deviations of the mean is larger than 89%. So Chebychev's inequality holds for the distribution of project sizes.

5. Using the *projects* table, it is time to create a new data set that you will name "teachers." This table should have two columns for each teacher.

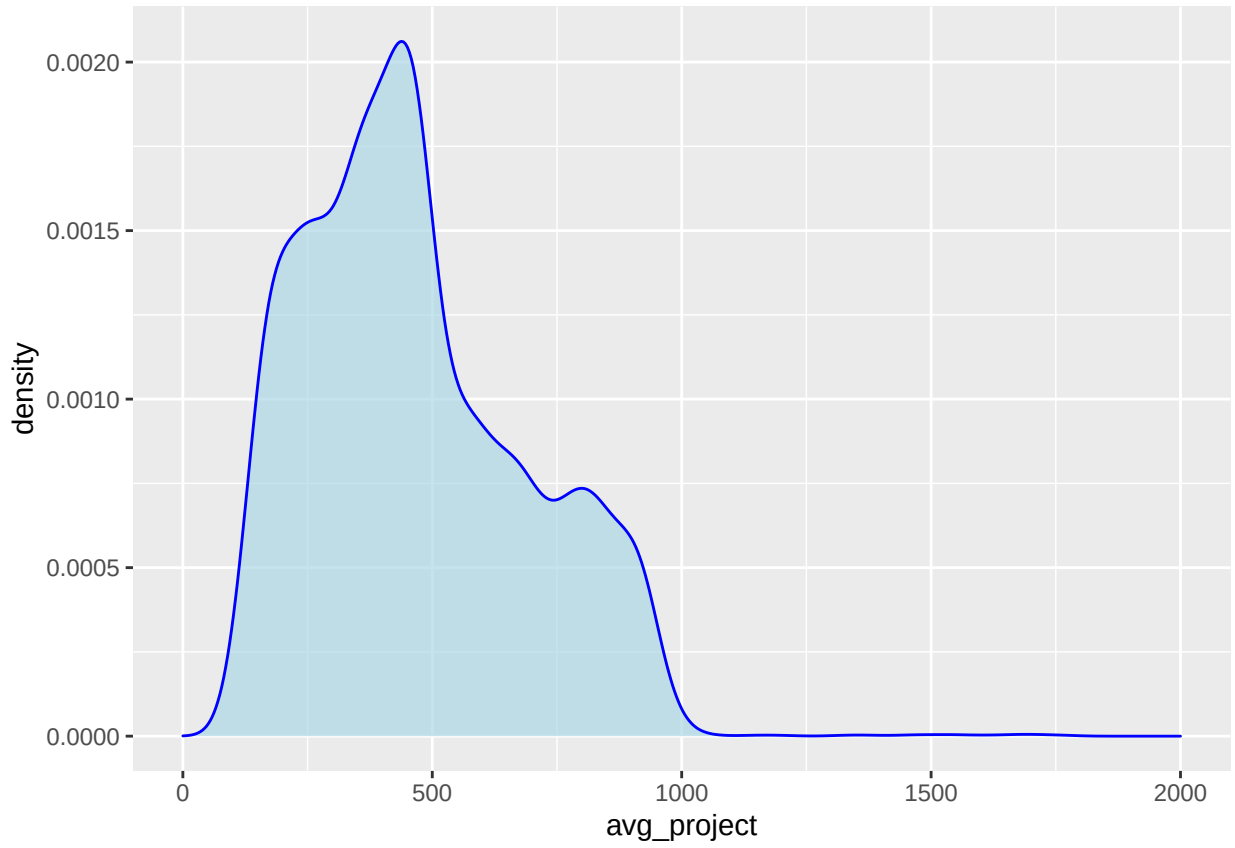
```
teachers <- data.frame(teacher_id = unique(data$teacher_id))
teachers$last<- (data %>%
                   group_by(teacher_id) %>%
                   arrange(teacher_id) %>%
                   summarise(maximum=max(project_order)))$maximum

teachers$avg_project<- (data %>%
              group_by(teacher_id) %>%
                arrange(teacher_id) %>%
              select(total_price_excluding_optional_support) %>%
              summarise(avg_project=mean(total_price_excluding_optional_support)))$avg_project
head(teachers, 20)
```

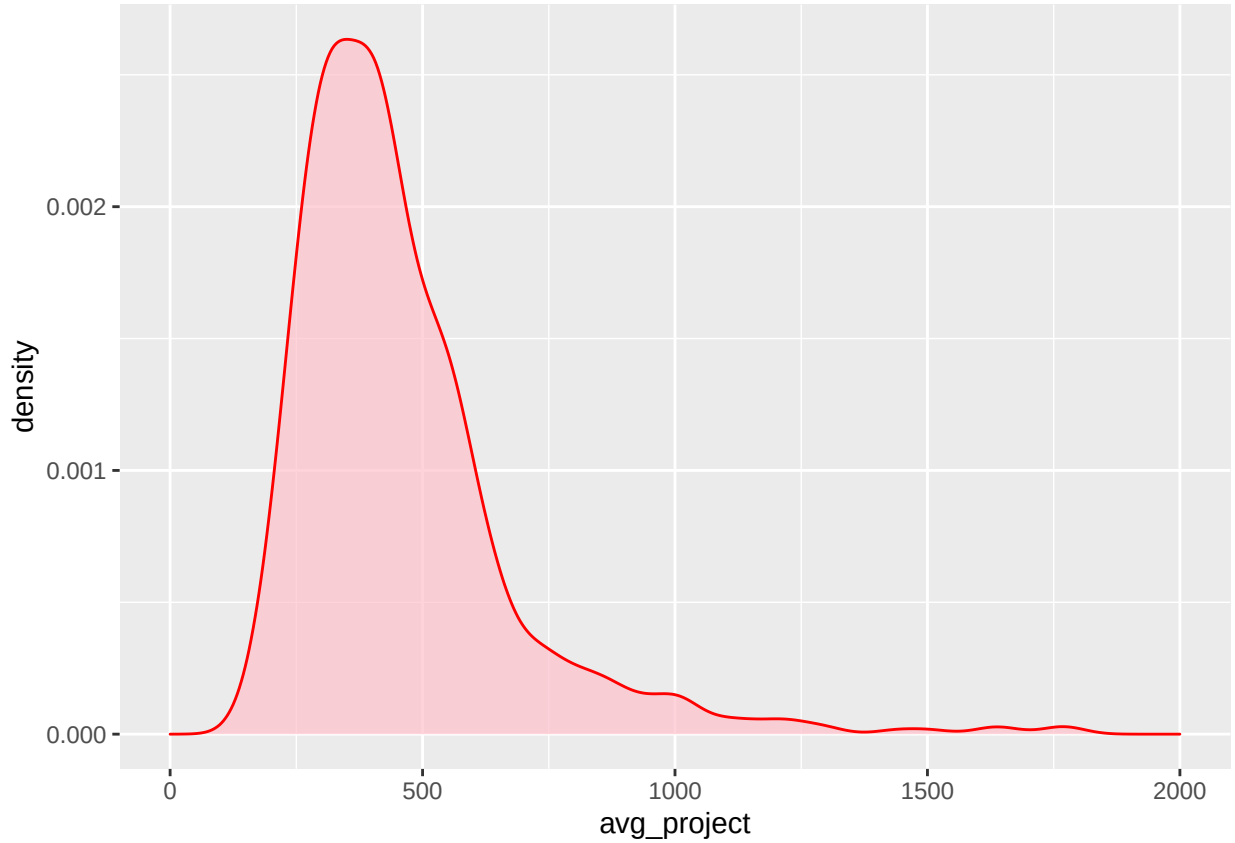
```
##    teacher_id last avg_project
## 1          43    6    560.2400
## 2          48    1    327.3000
## 3          74    6    295.3533
## 4          77    6    592.2843
## 5          85    6    487.3357
## 6          94    2    563.6800
## 7         114    2    288.6300
## 8         150    6    246.7897
## 9         152    1    270.0000
## 10        154    1    900.0800
## 11        155    1    751.6300
## 12        156    6    423.4108
## 13        166    6    818.7517
## 14        169    3    516.2967
## 15        177    3    569.1067
## 16        202    1    324.8100
## 17        223    2    581.1800
## 18        262    1    219.2100
## 19        263    1    507.5300
## 20        267    6    290.1000
```

6. Using the new "teachers" table, graph the densities of average project size only for last==1 and last==6.

```
ggplot() +
  geom_density(data=teachers[teachers$last==1,], mapping=aes(avg_project), alpha=0.7, colour="blue", fi
  xlim(0, 2000)
```




```
ggplot() +
  geom_density(data=teachers[teachers$last==6,], mapping=aes(avg_project), alpha=0.7, colour="red", fil
  xlim(0, 2000)
```




```
print("Test the distribution of last equal to 1")
```

```
## [1] "Test the distribution of last equal to 1"
```

```
teachers %>%
  filter(last == 1) %>%
  select(avg_project) %>%
  ks.test("pnorm")
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  .
## D = 1, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
print("Test the distribution of last equal to 6")
```

```
## [1] "Test the distribution of last equal to 6"
```

```
teachers %>%
  filter(last == 6) %>%
  select(avg_project) %>%
  ks.test("pnorm")
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  .
## D = 1, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Ans1. We can see that the p-value of both of two distribution is very small, which means that there is enough strong evidence to reject the hypothesis that the distribution is normal distribution. Therefore, I do not think these distribution are close to normal density.

Ans2. Compared to the distribution with last=1, the "avg_project" with "last" equal to 6 is more concentrated and smoother, with only 1 peak. The distribution of last=1 fluctuates greatly and does not tend to be flat.

7. Interpret the differences between outputs of step 3 and outputs of step 6. What would you conclude in relation to the given research interest?

Ans. In step 6, our avg_project is equivalent to averaging the teacher's Total price excluding optional support for project order. In general, we can find that the distribution in step6 will be slightly more concentrated and smoother than that in step1, which is more obvious in the distribution of the number of 6. The reason for this may be that teachers who have done multiple projects will be more proficient than those who have only once, which can help them pay more attention to the rationality of size each time they set out a project, and keep it at a stable level instead of huge changing.