# 1-setting-up-demo-doc

## Loading, setting up

Let's first load the {dataedu} package and the {tidyverse} suite of packages.

Click the green arrow to the first right of what is a "code chunk" below:

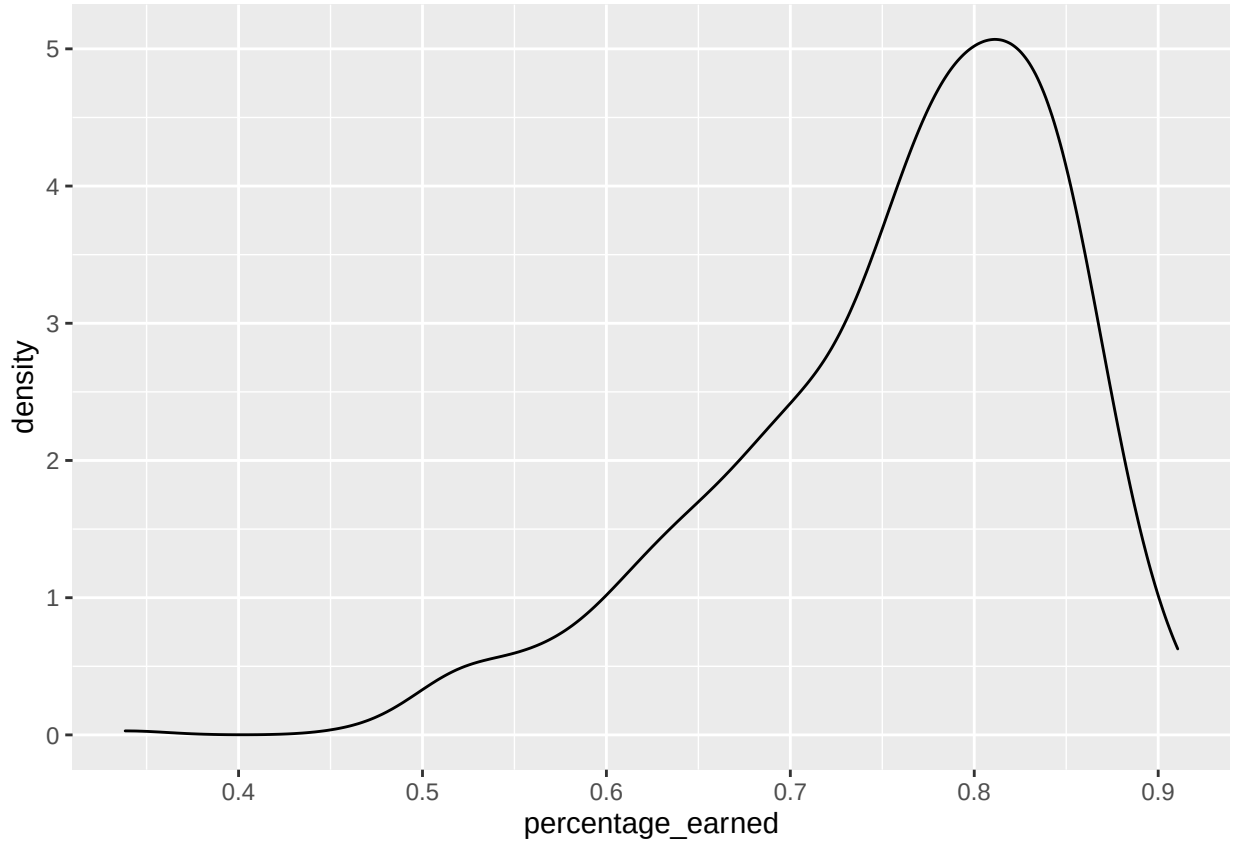
```
library(dataedu)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.1     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   0.8.5
## v tidyr   1.0.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

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

```
sci_mo_processed %>%
  ggplot(aes(x = percentage_earned)) +
  geom_density()
```




If you saw a plot, congratulations! Things are working just fine.

What do you notice about this plot? What do you wonder?

Here is the data set we will use; click the green arrow again:

```
sci_mo_processed
```

```
## # A tibble: 603 x 30
##    student_id course_id total_points_po~ total_points_ea~ percentage_earn~
##         <dbl> <chr>                <dbl>            <dbl>            <dbl>
##  1      43146 FrScA-S2~             3280             2220            0.677
##  2      44638 OcnA-S11~             3531             2672            0.757
##  3      47448 FrScA-S2~             2870             1897            0.661
##  4      47979 OcnA-S21~             4562             3090            0.677
##  5      48797 PhysA-S1~             2207             1910            0.865
##  6      51943 FrScA-S2~             4208             3596            0.855
##  7      52326 AnPhA-S2~             4325             2255            0.521
##  8      52446 PhysA-S1~             2086             1719            0.824
##  9      53447 FrScA-S1~             4655             3149            0.676
## 10      53475 FrScA-S1~             1710             1402            0.820
## # ... with 593 more rows, and 25 more variables: subject <chr>, semester <chr>,
## #   section <chr>, Gradebook_Item <chr>, Grade_Category <lgl>,
## #   FinalGradeCEMS <dbl>, Points_Possible <dbl>, Points_Earned <dbl>,
## #   Gender <chr>, q1 <dbl>, q2 <dbl>, q3 <dbl>, q4 <dbl>, q5 <dbl>, q6 <dbl>,
## #   q7 <dbl>, q8 <dbl>, q9 <dbl>, q10 <dbl>, TimeSpent <dbl>,
## #   TimeSpent_hours <dbl>, TimeSpent_std <dbl>, int <dbl>, pc <dbl>, uv <dbl>
```

## Selecting variables

Let's select only a few variables.

```
sci_mo_processed %>%
  select(student_id, course_id, percentage_earned)
```

```
## # A tibble: 603 x 3
##    student_id course_id     percentage_earned
##         <dbl> <chr>                     <dbl>
##  1      43146 FrScA-S216-02             0.677
##  2      44638 OcnA-S116-01              0.757
##  3      47448 FrScA-S216-01             0.661
##  4      47979 OcnA-S216-01              0.677
##  5      48797 PhysA-S116-01             0.865
##  6      51943 FrScA-S216-03             0.855
##  7      52326 AnPhA-S216-01             0.521
##  8      52446 PhysA-S116-01             0.824
##  9      53447 FrScA-S116-01             0.676
## 10      53475 FrScA-S116-02             0.820
## # ... with 593 more rows
```

**Try it out!**

Let's try to *include one additional variable* in your select function,

First, type the name of the data to view a summary of your data (including what variables are included in it):

```
sci_mo_processed
```

```
## # A tibble: 603 x 30
##    student_id course_id total_points_po~ total_points_ea~ percentage_earn~
##         <dbl> <chr>                <dbl>            <dbl>            <dbl>
##  1      43146 FrScA-S2~             3280             2220            0.677
##  2      44638 OcnA-S11~             3531             2672            0.757
##  3      47448 FrScA-S2~             2870             1897            0.661
##  4      47979 OcnA-S21~             4562             3090            0.677
##  5      48797 PhysA-S1~             2207             1910            0.865
##  6      51943 FrScA-S2~             4208             3596            0.855
##  7      52326 AnPhA-S2~             4325             2255            0.521
##  8      52446 PhysA-S1~             2086             1719            0.824
##  9      53447 FrScA-S1~             4655             3149            0.676
## 10      53475 FrScA-S1~             1710             1402            0.820
## # ... with 593 more rows, and 25 more variables: subject <chr>, semester <chr>,
## #   section <chr>, Gradebook_Item <chr>, Grade_Category <lgl>,
## #   FinalGradeCEMS <dbl>, Points_Possible <dbl>, Points_Earned <dbl>,
## #   Gender <chr>, q1 <dbl>, q2 <dbl>, q3 <dbl>, q4 <dbl>, q5 <dbl>, q6 <dbl>,
## #   q7 <dbl>, q8 <dbl>, q9 <dbl>, q10 <dbl>, TimeSpent <dbl>,
## #   TimeSpent_hours <dbl>, TimeSpent_std <dbl>, int <dbl>, pc <dbl>, uv <dbl>
```

Then, add a new variable to the code below after percentage_earned, being careful to type the new variable name as it appears in the data. When you're ready, click the green arrow to view the result.

```
sci_mo_processed %>%
  select(student_id, course_id, percentage_earned)
```

```
## # A tibble: 603 x 3
##    student_id course_id     percentage_earned
##         <dbl> <chr>                     <dbl>
##  1      43146 FrScA-S216-02             0.677
##  2      44638 OcnA-S116-01              0.757
##  3      47448 FrScA-S216-01             0.661
##  4      47979 OcnA-S216-01              0.677
##  5      48797 PhysA-S116-01             0.865
##  6      51943 FrScA-S216-03             0.855
##  7      52326 AnPhA-S216-01             0.521
##  8      52446 PhysA-S116-01             0.824
##  9      53447 FrScA-S116-01             0.676
## 10      53475 FrScA-S116-02             0.820
## # ... with 593 more rows
```

## Filtering variables

Next, let's explore filtering variables

```
sci_mo_processed %>%
  filter(percentage_earned >= .60)
```

```
## # A tibble: 563 x 30
##    student_id course_id total_points_po~ total_points_ea~ percentage_earn~
##         <dbl> <chr>                <dbl>            <dbl>            <dbl>
##  1      43146 FrScA-S2~             3280             2220            0.677
##  2      44638 OcnA-S11~             3531             2672            0.757
##  3      47448 FrScA-S2~             2870             1897            0.661
##  4      47979 OcnA-S21~             4562             3090            0.677
##  5      48797 PhysA-S1~             2207             1910            0.865
##  6      51943 FrScA-S2~             4208             3596            0.855
##  7      52446 PhysA-S1~             2086             1719            0.824
##  8      53447 FrScA-S1~             4655             3149            0.676
##  9      53475 FrScA-S1~             1710             1402            0.820
## 10      53475 FrScA-S2~             1209              977            0.808
## # ... with 553 more rows, and 25 more variables: subject <chr>, semester <chr>,
## #   section <chr>, Gradebook_Item <chr>, Grade_Category <lgl>,
## #   FinalGradeCEMS <dbl>, Points_Possible <dbl>, Points_Earned <dbl>,
## #   Gender <chr>, q1 <dbl>, q2 <dbl>, q3 <dbl>, q4 <dbl>, q5 <dbl>, q6 <dbl>,
## #   q7 <dbl>, q8 <dbl>, q9 <dbl>, q10 <dbl>, TimeSpent <dbl>,
## #   TimeSpent_hours <dbl>, TimeSpent_std <dbl>, int <dbl>, pc <dbl>, uv <dbl>
```

What do you think will happen if we add another condition to the filter statement?

**Try it out!**

Let's try to filter on TimeSpent, which is a variable for how many minutes students were on the course Learning Management System (LMS).

Be sure to run the result to see whether it did what you think it should do!

```
sci_mo_processed %>%
  filter(percentage_earned >= .60,
         TimeSpent > 10 )
```

```
## # A tibble: 553 x 30
##    student_id course_id total_points_po~ total_points_ea~ percentage_earn~
##         <dbl> <chr>                <dbl>            <dbl>            <dbl>
##  1      43146 FrScA-S2~             3280             2220            0.677
##  2      44638 OcnA-S11~             3531             2672            0.757
##  3      47448 FrScA-S2~             2870             1897            0.661
##  4      47979 OcnA-S21~             4562             3090            0.677
##  5      48797 PhysA-S1~             2207             1910            0.865
##  6      52446 PhysA-S1~             2086             1719            0.824
##  7      53447 FrScA-S1~             4655             3149            0.676
##  8      53475 FrScA-S2~             1209              977            0.808
##  9      54066 OcnA-S11~             4641             3429            0.739
## 10      54282 OcnA-S11~             3581             2777            0.775
## # ... with 543 more rows, and 25 more variables: subject <chr>, semester <chr>,
## #   section <chr>, Gradebook_Item <chr>, Grade_Category <lgl>,
## #   FinalGradeCEMS <dbl>, Points_Possible <dbl>, Points_Earned <dbl>,
## #   Gender <chr>, q1 <dbl>, q2 <dbl>, q3 <dbl>, q4 <dbl>, q5 <dbl>, q6 <dbl>,
## #   q7 <dbl>, q8 <dbl>, q9 <dbl>, q10 <dbl>, TimeSpent <dbl>,
## #   TimeSpent_hours <dbl>, TimeSpent_std <dbl>, int <dbl>, pc <dbl>, uv <dbl>
```

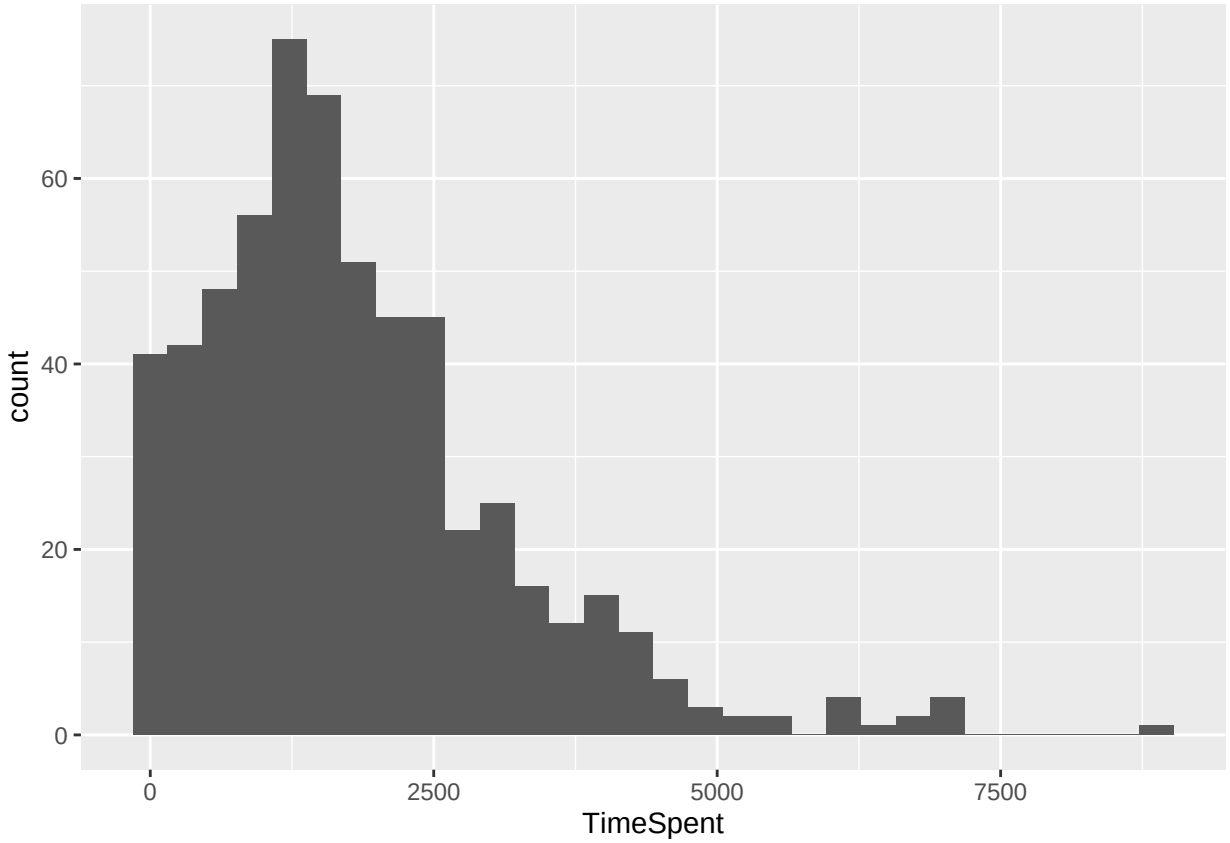
## Creating a plot (with ggplot2)

What do you think this code will do?

```
sci_mo_processed %>%
  ggplot(aes(x = TimeSpent)) +
  geom_histogram()
```

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

```
## Warning: Removed 5 rows containing non-finite values (stat_bin).
```




**Try it out!**

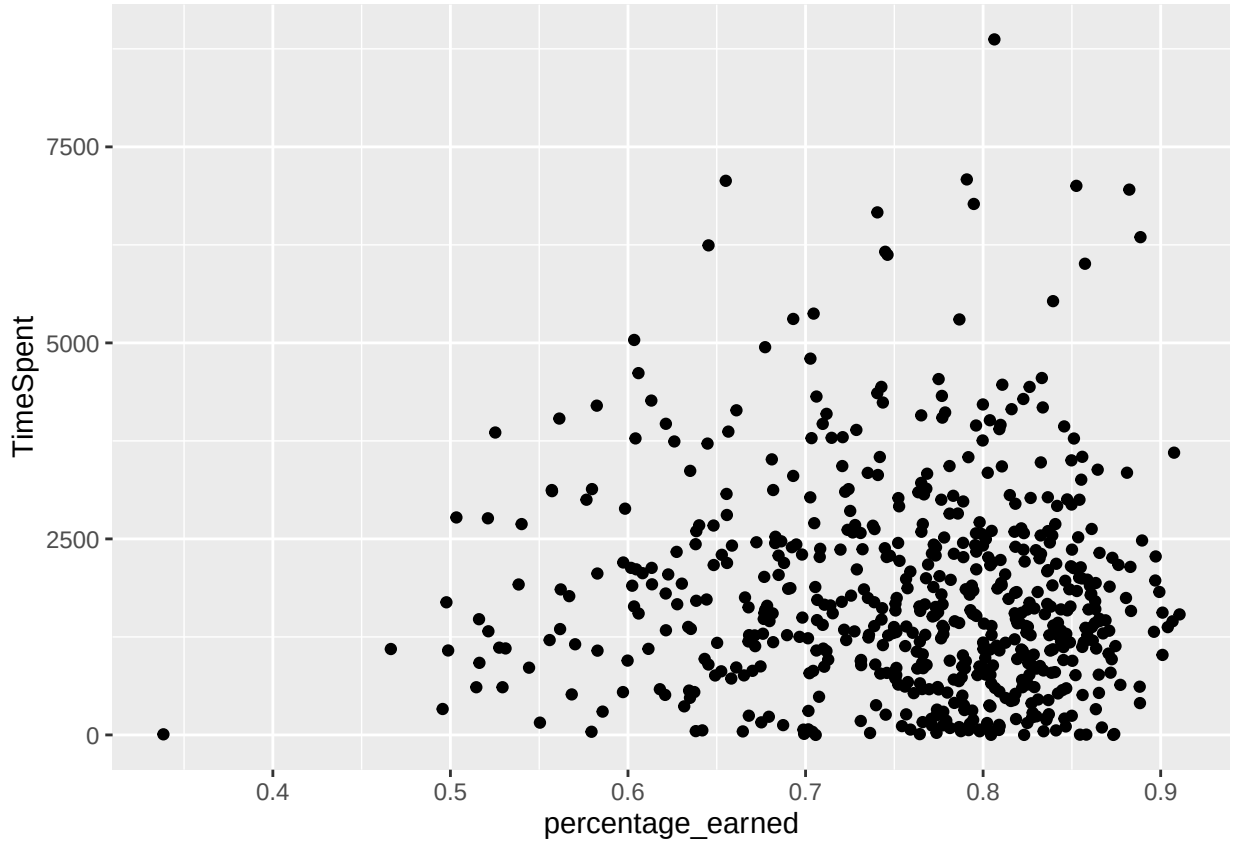
Now, add TimeSpent to the code below as the variable that will correspond to the $y$-axis.

Notice that instead of using geom_density, we're using geom_point. What do you think this will create?

Be sure to run the code chunk to see the result.

```
sci_mo_processed %>%
  ggplot(aes(x = percentage_earned, y = TimeSpent)) +
  geom_point()
```

```
## Warning: Removed 5 rows containing missing values (geom_point).
```




---

## Rendering this document to a PDF report

- change the name at the very top of this document
- click "Knit" and check out the result!

## Visualizing a larger dataset (leap)

Can you filter or select variables from the dataset below?

Can you select a smaller number of variables from the dataset below?

```
tt_tweets
```

```
## # A tibble: 4,418 x 90
##    user_id status_id created_at          screen_name text  source
##    <chr>   <chr>     <dttm>              <chr>       <chr> <chr>
##  1 115921~ 11631542~ 2019-08-18 18:22:42 MKumarYYC   "Fir~ Twitt~
##  2 107332~ 11632475~ 2019-08-19 00:33:11 cizzart     "El ~ Twitt~
##  3 107332~ 11450435~ 2019-06-29 18:57:17 cizzart     "Pro~ Twitt~
##  4 107332~ 11168648~ 2019-04-13 00:45:15 cizzart     "#Ar~ Twitt~
```

```
##  5 107332~ 11228824~ 2019-04-29 15:17:02 cizzart     "Pes~ Twitt~
##  6 107332~ 11176387~ 2019-04-15 04:00:17 cizzart     "Dat~ Twitt~
##  7 107332~ 11245531~ 2019-05-04 05:55:32 cizzart     "El ~ Twitt~
##  8 107332~ 11407021~ 2019-06-17 19:25:50 cizzart     "#da~ Twitt~
##  9 107332~ 11325299~ 2019-05-26 06:12:46 cizzart     "El ~ Twitt~
## 10 107332~ 11233585~ 2019-04-30 22:48:43 cizzart     "Vis~ Twitt~
## # ... with 4,408 more rows, and 84 more variables: display_text_width <dbl>,
## #   reply_to_status_id <chr>, reply_to_user_id <chr>,
## #   reply_to_screen_name <chr>, is_quote <lgl>, is_retweet <lgl>,
## #   favorite_count <int>, retweet_count <int>, quote_count <int>,
## #   reply_count <int>, hashtags <list>, symbols <list>, urls_url <list>,
## #   urls_t.co <list>, urls_expanded_url <list>, media_url <list>,
## #   media_t.co <list>, media_expanded_url <list>, media_type <list>,
## #   ext_media_url <list>, ext_media_t.co <list>, ext_media_expanded_url <list>,
## #   ext_media_type <chr>, mentions_user_id <list>, mentions_screen_name <list>,
## #   lang <chr>, quoted_status_id <chr>, quoted_text <chr>,
## #   quoted_created_at <dttm>, quoted_source <chr>, quoted_favorite_count <int>,
## #   quoted_retweet_count <int>, quoted_user_id <chr>, quoted_screen_name <chr>,
## #   quoted_name <chr>, quoted_followers_count <int>,
## #   quoted_friends_count <int>, quoted_statuses_count <int>,
## #   quoted_location <chr>, quoted_description <chr>, quoted_verified <lgl>,
## #   retweet_status_id <chr>, retweet_text <chr>, retweet_created_at <dttm>,
## #   retweet_source <chr>, retweet_favorite_count <int>,
## #   retweet_retweet_count <int>, retweet_user_id <chr>,
## #   retweet_screen_name <chr>, retweet_name <chr>,
## #   retweet_followers_count <int>, retweet_friends_count <int>,
## #   retweet_statuses_count <int>, retweet_location <chr>,
## #   retweet_description <chr>, retweet_verified <lgl>, place_url <chr>,
## #   place_name <chr>, place_full_name <chr>, place_type <chr>, country <chr>,
## #   country_code <chr>, geo_coords <list>, coords_coords <list>,
## #   bbox_coords <list>, status_url <chr>, name <chr>, location <chr>,
## #   description <chr>, url <chr>, protected <lgl>, followers_count <int>,
## #   friends_count <int>, listed_count <int>, statuses_count <int>,
## #   favourites_count <int>, account_created_at <dttm>, verified <lgl>,
## #   profile_url <chr>, profile_expanded_url <chr>, account_lang <lgl>,
## #   profile_banner_url <chr>, profile_background_url <chr>,
## #   profile_image_url <chr>
```