Laser Lab 4 - Final Workflow

Jeanne McClure

2022-07-08

Contents

0.1 0. Introduction

Provide a brief overview or case study.

- Include your R1: *research questions!*

0.2 1. Prepare

Load Packages

```
#Load necessary packages
library(tidyverse)
library(here)
```

```
# install Latex - this may take a few minutes. After it is installed you do not need to keep it on your

install.packages("tinytex")
tinytex::install_tinytex()
```

0.3 2. Wrangle

0.3.1 a. *Import Data*

0.3.1.1 Data Source #1: Log Data Log-trace data is data generated from our interactions with digital technologies, such as archived data from social media postings. In education, an increasingly common source of log-trace data is that generated from interactions with LMS and other digital tools.

The data we will use has already been "wrangled" quite a bit and is a summary type of log-trace data: the number of minutes students spent on the course. While this data type is fairly straightforward, there are even more complex sources of log-trace data out there (e.g., time stamps associated with when students started and stopped accessing the course).

Let's use the `read_csv()` function from {readr} to import our `log-data.csv` file directly from our data folder and name this data set `time_spent`, to help us to quickly recollect what function it serves in this analysis:

```
#load with read_csv package
time_spent <- read_csv("~/RProj22/foundation_labs_2022/foundation_lab_2/data/log-data.csv")
```

```
## Rows: 716 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (4): course_id, gender, enrollment_reason, enrollment_status
## dbl (2): student_id, time_spent
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#read in data_to_explore
data_to_explore <- read_csv(here("data", "data_to_explore.csv"))
```

```
## Rows: 943 Columns: 34
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (8): student_id, subject, semester, section, gender, enrollment_reason...
## dbl  (23): total_points_possible, total_points_earned, proportion_earned, ti...
## dttm  (3): date_x, date_y, date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

0.4 3. Explore

```


#install package if this is first time using skimr


#load library
library(skimr)

#skim data
skim(data_to_explore)
```

0.4.0.1 A. TABLE SUMMARY

Table 1: Data summary

Name	data_to_explore
Number of rows	943
Number of columns	34
Column type frequency:	
character	8
numeric	23
POSIXct	3
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
student_id	0	1.00	2	6	0	879	0
subject	0	1.00	4	5	0	5	0
semester	0	1.00	4	4	0	4	0
section	0	1.00	2	2	0	4	0
gender	227	0.76	1	1	0	2	0
enrollment_reason	227	0.76	5	34	0	5	0
enrollment_status	227	0.76	7	17	0	3	0
course_id	281	0.70	12	13	0	36	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
total_points_possible	226	0.76	1619.55	387.12	1212.00	1217.00	1676.00	1791.00	2425.00	
total_points_earned	226	0.76	1229.98	510.64	0.00	1002.50	1177.13	1572.45	2413.50	
proportion_earned	226	0.76	0.76	0.25	0.00	0.72	0.86	0.92	1.01	
time_spent	232	0.75	1828.80	1363.13	0.45	895.57	1559.97	2423.94	8870.88	
time_spent_hours	232	0.75	30.48	22.72	0.01	14.93	26.00	40.40	147.85	
int	293	0.69	4.30	0.60	1.80	4.00	4.40	4.80	5.00	
val	287	0.70	3.75	0.75	1.00	3.33	3.67	4.33	5.00	
percomp	288	0.69	3.64	0.69	1.50	3.00	3.50	4.00	5.00	
tv	292	0.69	4.07	0.59	1.00	3.71	4.12	4.46	5.00	
q1	285	0.70	4.34	0.66	1.00	4.00	4.00	5.00	5.00	
q2	285	0.70	3.66	0.93	1.00	3.00	4.00	4.00	5.00	
q3	286	0.70	3.31	0.85	1.00	3.00	3.00	4.00	5.00	
q4	289	0.69	4.35	0.80	1.00	4.00	5.00	5.00	5.00	
q5	286	0.70	4.28	0.69	1.00	4.00	4.00	5.00	5.00	
q6	285	0.70	4.05	0.80	1.00	4.00	4.00	5.00	5.00	
q7	286	0.70	3.96	0.85	1.00	3.00	4.00	5.00	5.00	
q8	286	0.70	4.35	0.65	1.00	4.00	4.00	5.00	5.00	
q9	286	0.70	3.55	0.92	1.00	3.00	4.00	4.00	5.00	
q10	285	0.70	4.17	0.87	1.00	4.00	4.00	5.00	5.00	
post_int	848	0.10	3.88	0.94	1.00	3.50	4.00	4.50	5.00	
post_uv	848	0.10	3.48	0.99	1.00	3.00	3.67	4.00	5.00	
post_tv	848	0.10	3.71	0.90	1.00	3.29	3.86	4.29	5.00	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
post_percomp	848	0.10	3.47	0.88	1.00	3.00	3.50	4.00	5.00	

Variable type: POSIXct

skim_variable	n_missing	complete_rate	min	max	median	n_unique
date_x	393	0.58	2015-09-02 15:40:00	2016-05-24 15:53:00	2015-10-01 15:57:30	536
date_y	848	0.10	2015-09-02 15:31:00	2016-01-22 15:43:00	2016-01-04 13:25:00	95
date	834	0.12	2017-01-23 13:14:00	2017-02-13 13:00:00	2017-01-25 18:43:00	107

0.5 B. TIDY to EXPLORE

```
# using the `select()` and `filter()` functions. In the code chunk below,look at descriptive for just `

data_to_explore %>%
  select(c('subject', 'gender', 'proportion_earned', 'time_spent')) %>%
  filter(subject == "OcnA" | subject == "PhysA") %>%
  skim()
```

Table 5: Data summary

Name	Piped data
Number of rows	249
Number of columns	4
Column type frequency:	
character	2
numeric	2
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
subject	0	1.00	4	5	0	2	0
gender	48	0.81	1	1	0	2	0

Variable type: numeric

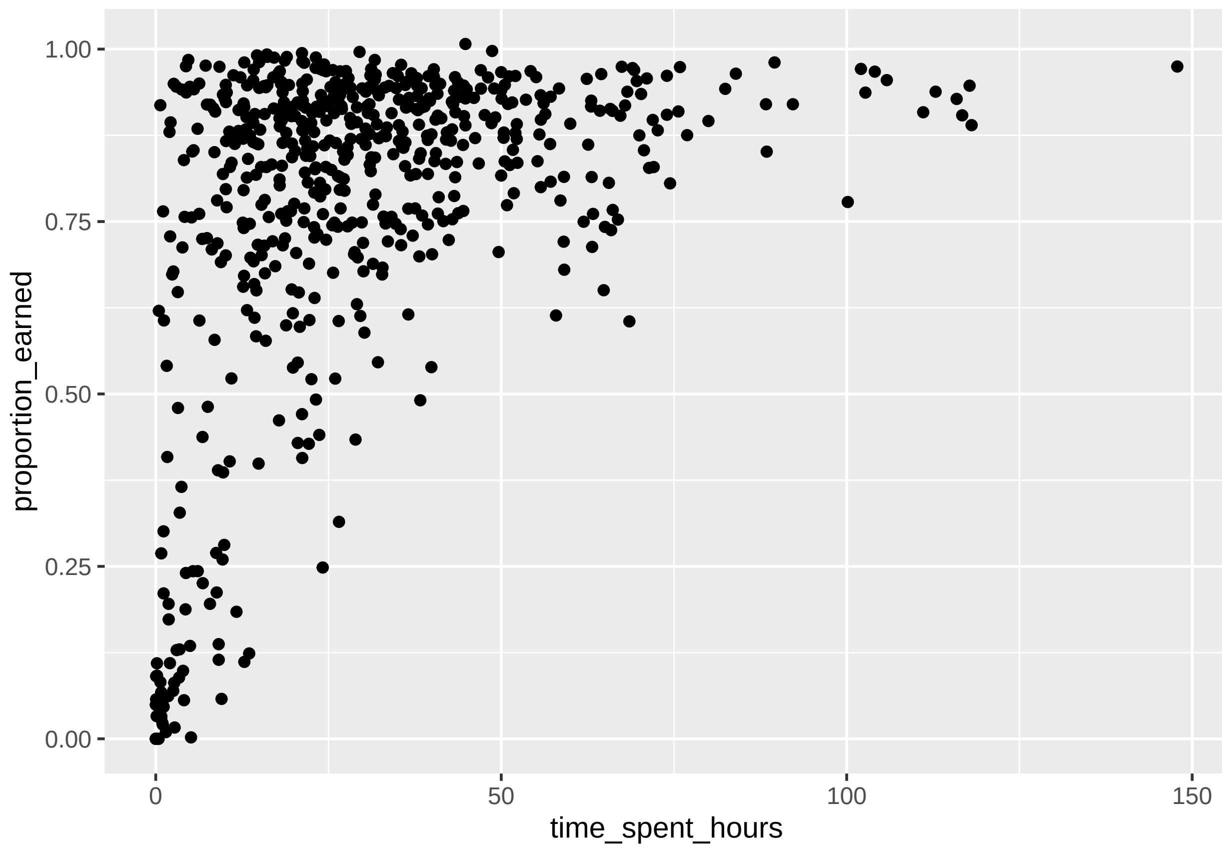
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
proportion_earned	48	0.81	0.78	0.24	0.00	0.73	0.86	0.94	1.00	
time_spent	48	0.81	1828.56	1374.13	0.58	943.07	1601.13	2356.88	8870.88	

0.5.0.1 B. DATA VIZ ggplot grammar - with layers

0.5.0.1.1 layers - Scatter Plot Basic graph 1. data 2. aes 3. geom

```
#layer 1: add data and aesthetics mapping
ggplot(data_to_explore, #<<
       aes(x = time_spent_hours,
           y = proportion_earned)) +
#layer 2: +  geom function type
  geom_point() #<<
```

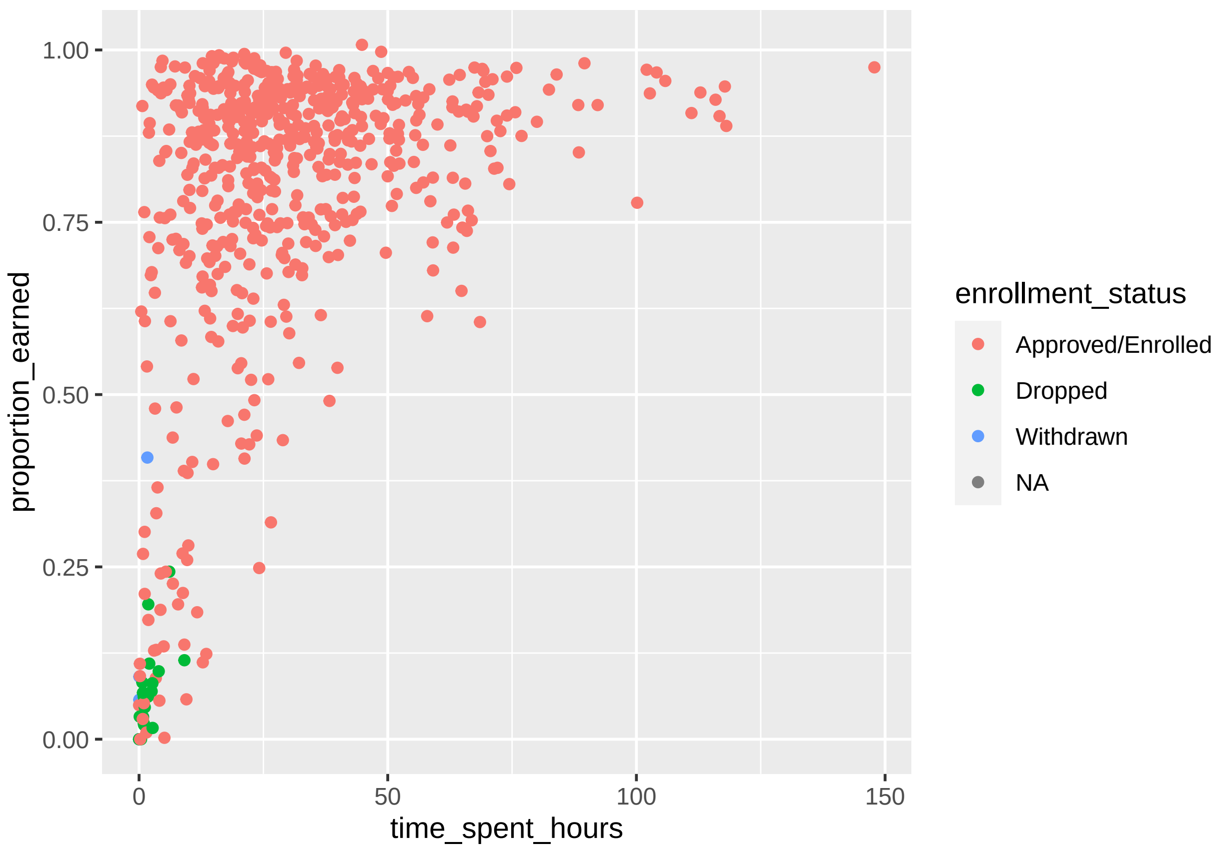
```
## Warning: Removed 345 rows containing missing values (geom_point).
```



Add **Scale** with different color for enrollment status.

```
#layer 1: add data and aesthetics mapping
#layer 3: add color scale by type
ggplot(data_to_explore,
       aes(x = time_spent_hours,
           y = proportion_earned,
           color = enrollment_status)) + #<<
#layer 2: +  geom function type
  geom_point()
```

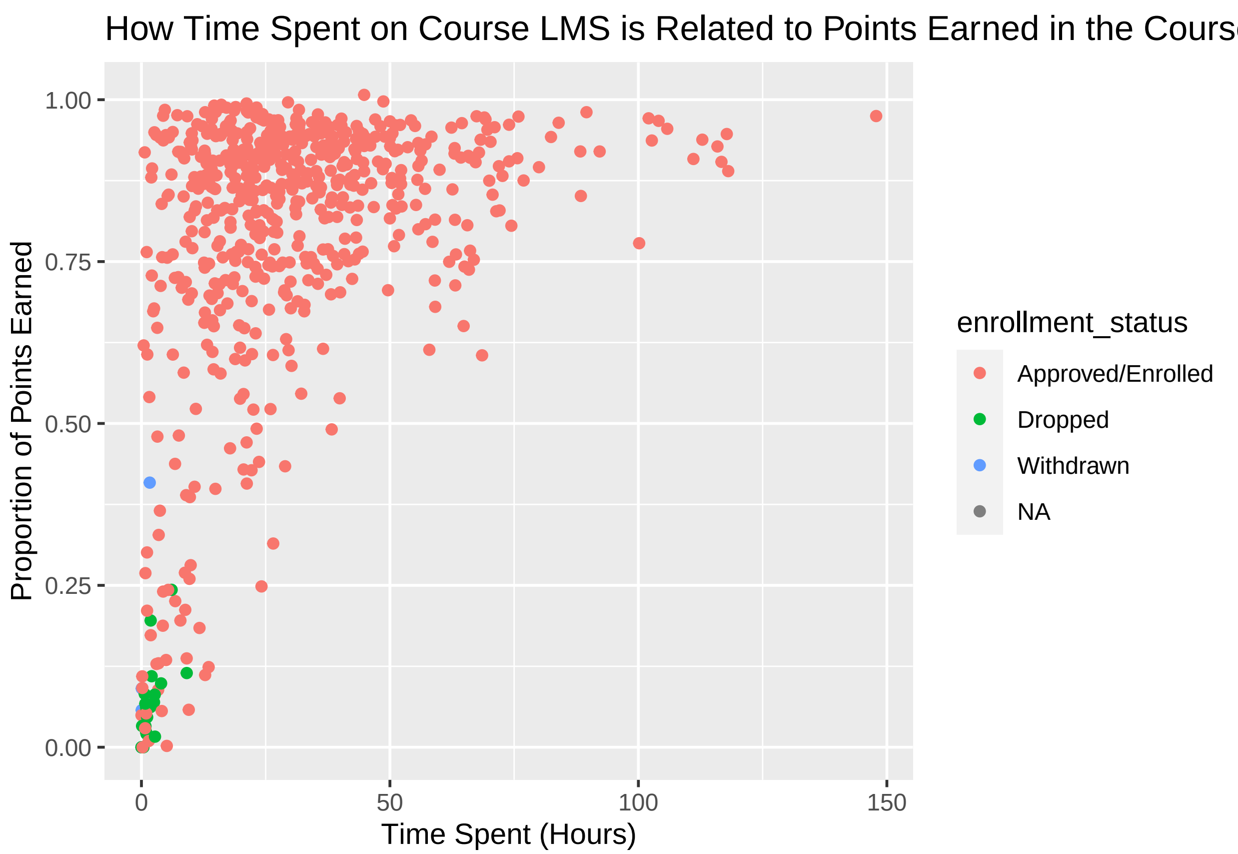
```
## Warning: Removed 345 rows containing missing values (geom_point).
```



Add another layer with **labs* labeling the title

```
#layer 1: add data and aesthetics mapping
#layer 3: add color scale by type
ggplot(data_to_explore,
       aes(x = time_spent_hours,
           y = proportion_earned,
           color = enrollment_status)) +
#layer 2: +  geom function type
  geom_point() +
#layer 4: add lables
  labs(title="How Time Spent on Course LMS is Related to Points Earned in the Course", x="Time Spent (H
```

```
## Warning: Removed 345 rows containing missing values (geom_point).
```



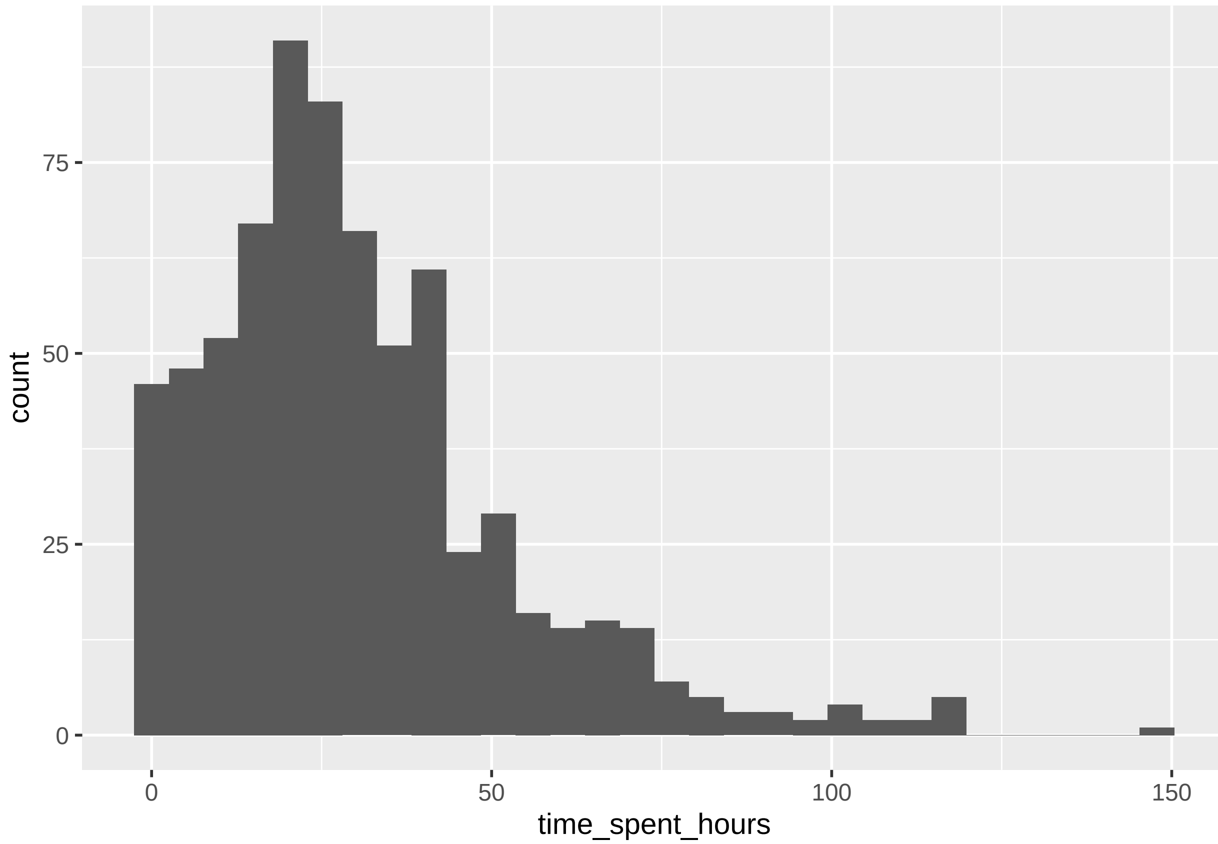
Add the **facet** layer

```
#layer 1: add data and aesthetics mapping
#layer 3: add color scale by type
ggplot(data_to_explore, aes(x = time_spent_hours, y = proportion_earned, color = enrollment_status)) +
#layer 2: +  geom function type
  geom_point() +
#layer 4: add lables
    x_lab(title="How Time Spent on Course LMS is Related to Points Earned in the Course",
       x="Time Spent (Hours)",
       y = "Proportion of Points Earned")
#layer 5: add facet wrap
  facet_wrap(~ subject)
```

0.5.0.1.2 layers- Histogram

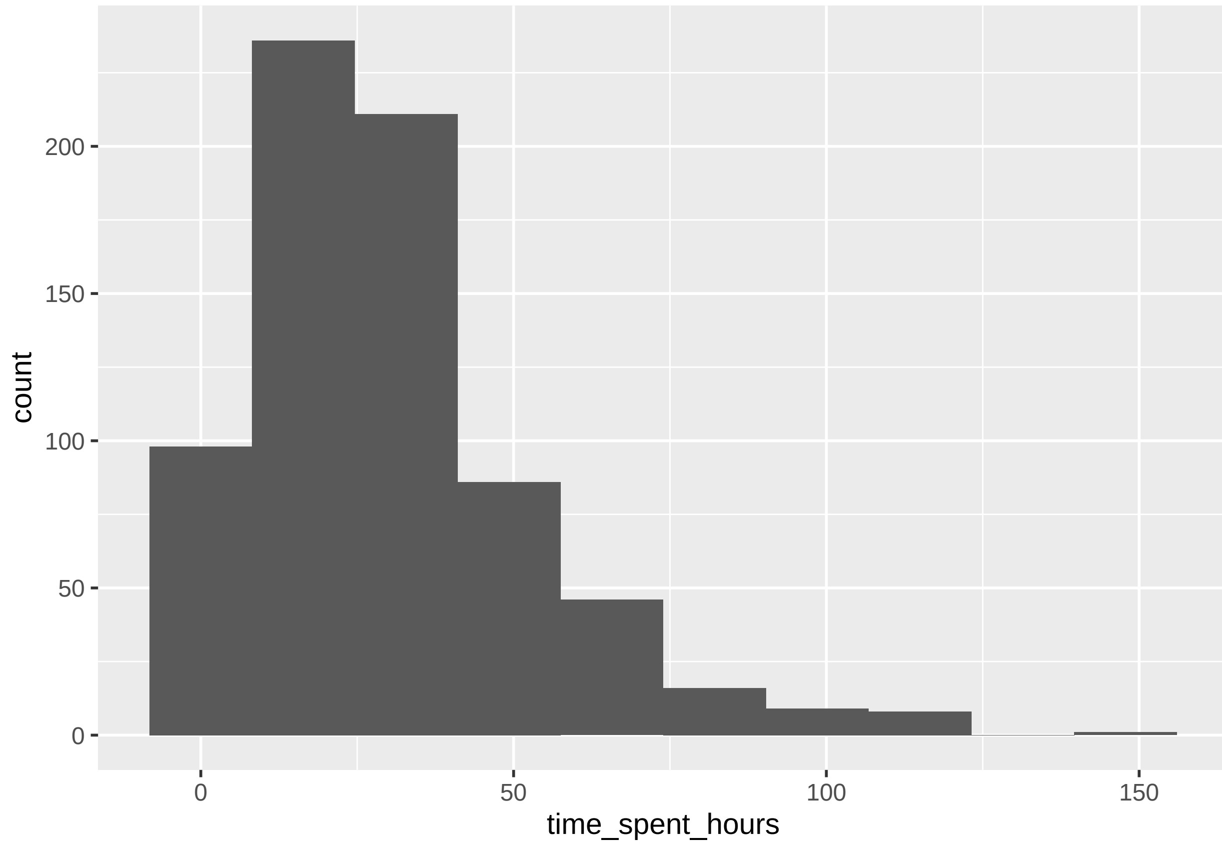
Create a basic histogram using the 'geom_hist()' function

```
# Layer 1: add data and aesthetic mapping
data_to_explore %>% #<<
  ggplot(aes(x = time_spent_hours)) +
# layer 2: add histogram geom
  geom_histogram()
```



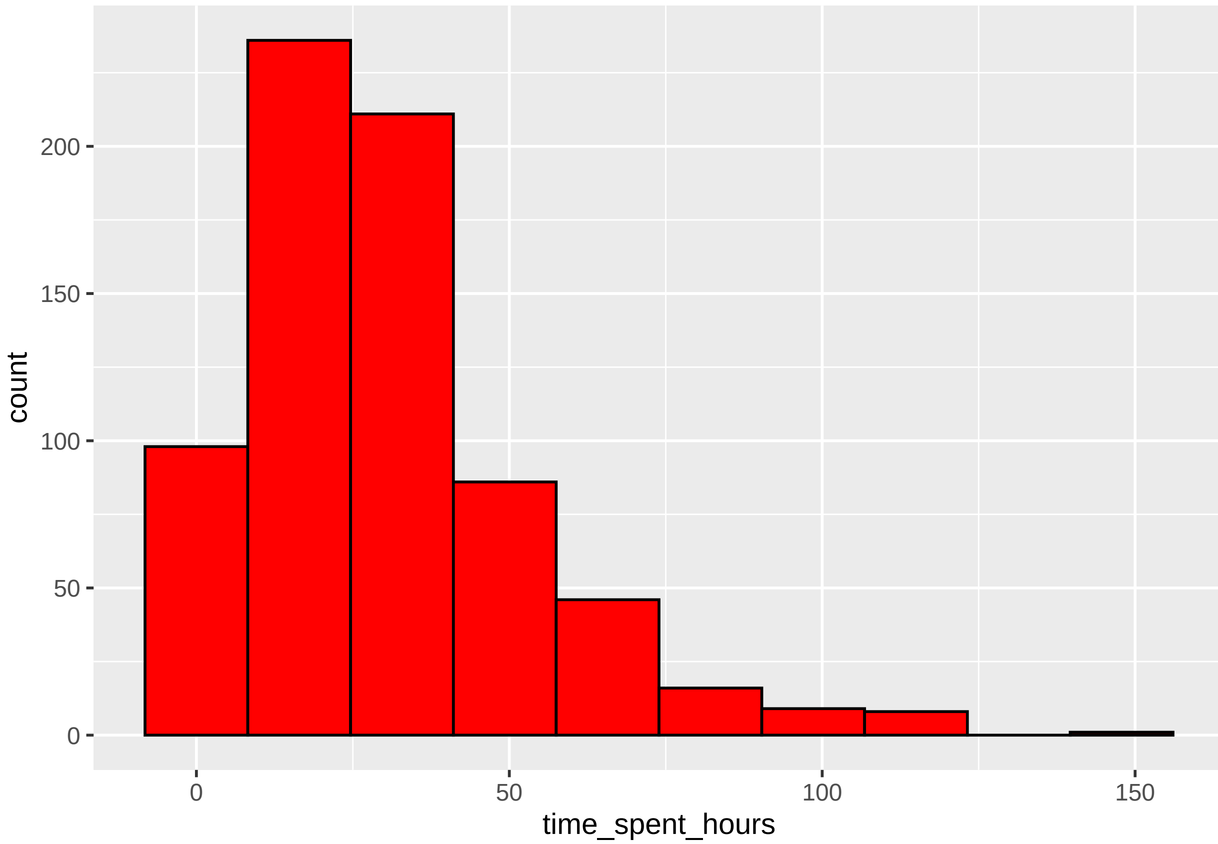
Change bin size

```
# Layer 1: add data and aesthetic mapping
data_to_explore %>%
  ggplot(aes(x = time_spent_hours)) +
# layer 2: add histogram geom
# layer 3a: add bin size
  geom_histogram(bins = 10)
```



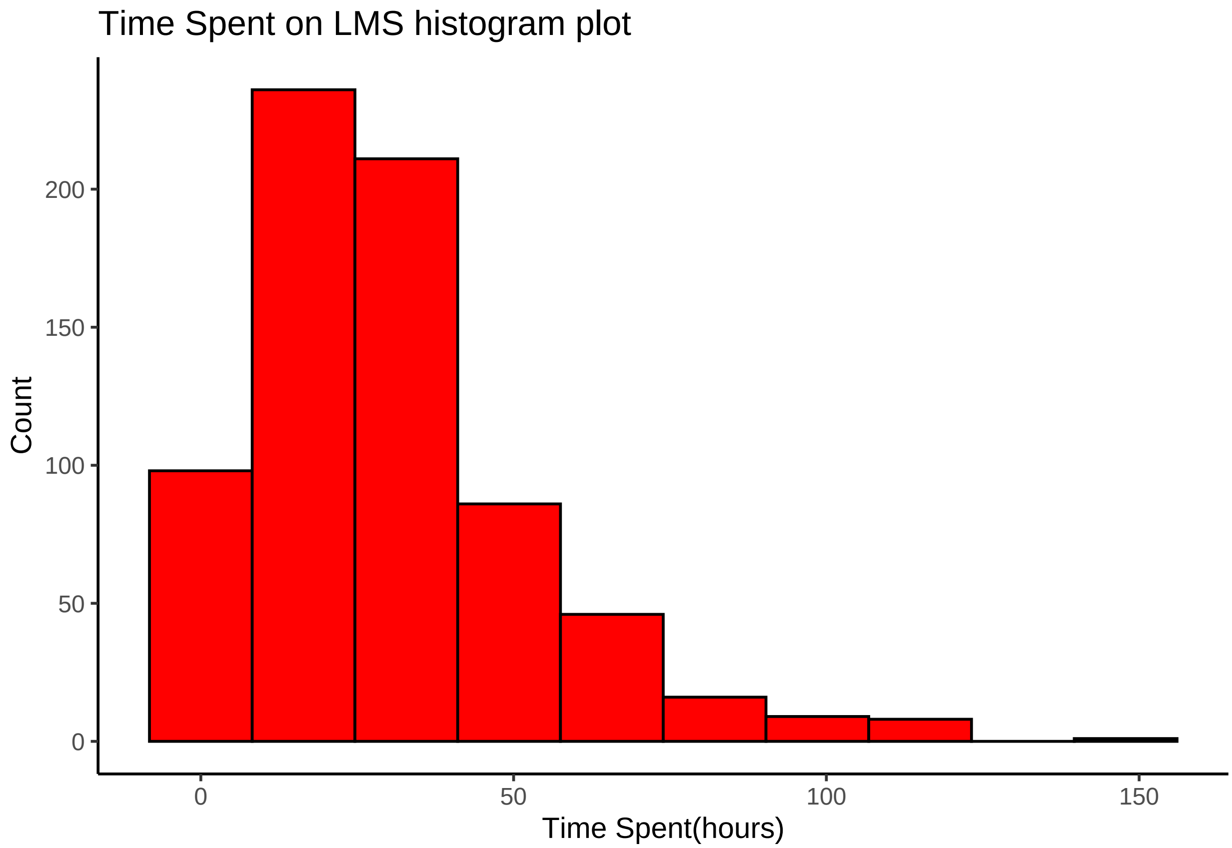
Add color label to make bins stand out

```
# Layer 1: add data and aesthetic mapping
data_to_explore %>%
  ggplot(aes(x = time_spent_hours)) +
# layer 2: add histogram geom
# layer 3a: add bin size
#layer 3b: add color
  geom_histogram(bins = 10,
                 fill = "red",
                 colour = "black")
```



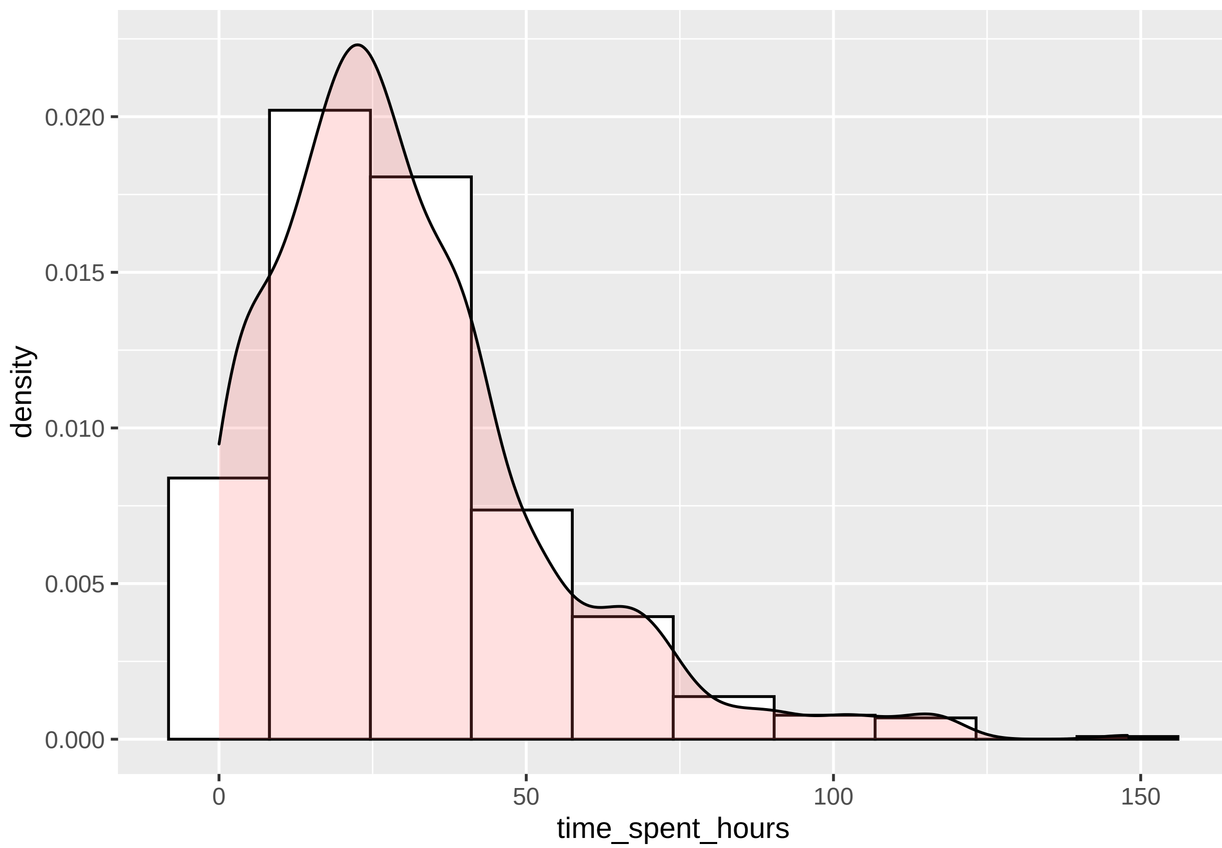
Add labels and add a theme for a clean aesthetic

```
# Layer 1: add data and aesthetic mapping
data_to_explore %>%
  ggplot(aes(x = time_spent_hours)) +
# layer 2: add histogram geom
# layer 3a: add bin size
# layer 3b: add color
  geom_histogram(bins = 10, fill = "red", colour = "black")+
#layer 4: add Labels
  labs(title="Time Spent on LMS histogram plot",x="Time Spent(hours)", y = "Count")+
  theme_classic()
```



Create a histogram with density plot

```
data_to_explore%>%
  ggplot(aes(x=time_spent_hours)) +
  geom_histogram(aes(y=..density..), colour="black", fill="white", bins = 10)+
 geom_density(alpha=.2, fill="#FF6666")
```




```
  labs(title="Time Spent on LMS histogram/density plot",x="Time Spent(hours)", y = "Density")+
  theme_classic()
```

```
## NULL
```

0.6 4. Model

Quantify the insights using mathematical models.

0.6.0.1 A. MATHMATICAL

Does time spent predict grade earned?

```
# Use linear regression model
lm(proportion_earned ~ time_spent_hours,
   data = data_to_explore)
```

```
##
## Call:
## lm(formula = proportion_earned ~ time_spent_hours, data = data_to_explore)
##
## Coefficients:
##      (Intercept)  time_spent_hours
##         0.624306          0.004792
```

```
# Add predictor variable for science
lm(proportion_earned ~ time_spent_hours + int,
   data = data_to_explore)
```

```
##
## Call:
## lm(formula = proportion_earned ~ time_spent_hours + int, data = data_to_explore)
##
## Coefficients:
##      (Intercept)  time_spent_hours               int
##         0.449657          0.004255          0.046283
```

```
# save the model
m1 <- lm(proportion_earned ~ time_spent_hours + int, data = data_to_explore)
```

Run a summary model for the model you just created called, `m1`.

```
#run the summary
summary(m1)
```

```
##
## Call:
## lm(formula = proportion_earned ~ time_spent_hours + int, data = data_to_explore)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66705 -0.07836  0.05049  0.14695  0.35766
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.449657   0.066488   6.763 3.54e-11 ***
## time_spent_hours 0.004255   0.000410  10.378  < 2e-16 ***
## int              0.046282   0.015364   3.012  0.00271 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2142 on 536 degrees of freedom
##   (404 observations deleted due to missingness)
## Multiple R-squared:  0.1859, Adjusted R-squared:  0.1828
## F-statistic: 61.18 on 2 and 536 DF,  p-value: < 2.2e-16
```

```
#install apaTables if this is your first time - do you remember how?

#load packages
library(apaTables)
# use the {apaTables} package to create a nice regression table that could be used for later publicatio
apa.reg.table(m1, filename = "lm-table.doc")
```

```
##
##
## Regression results using proportion_earned as the criterion
```

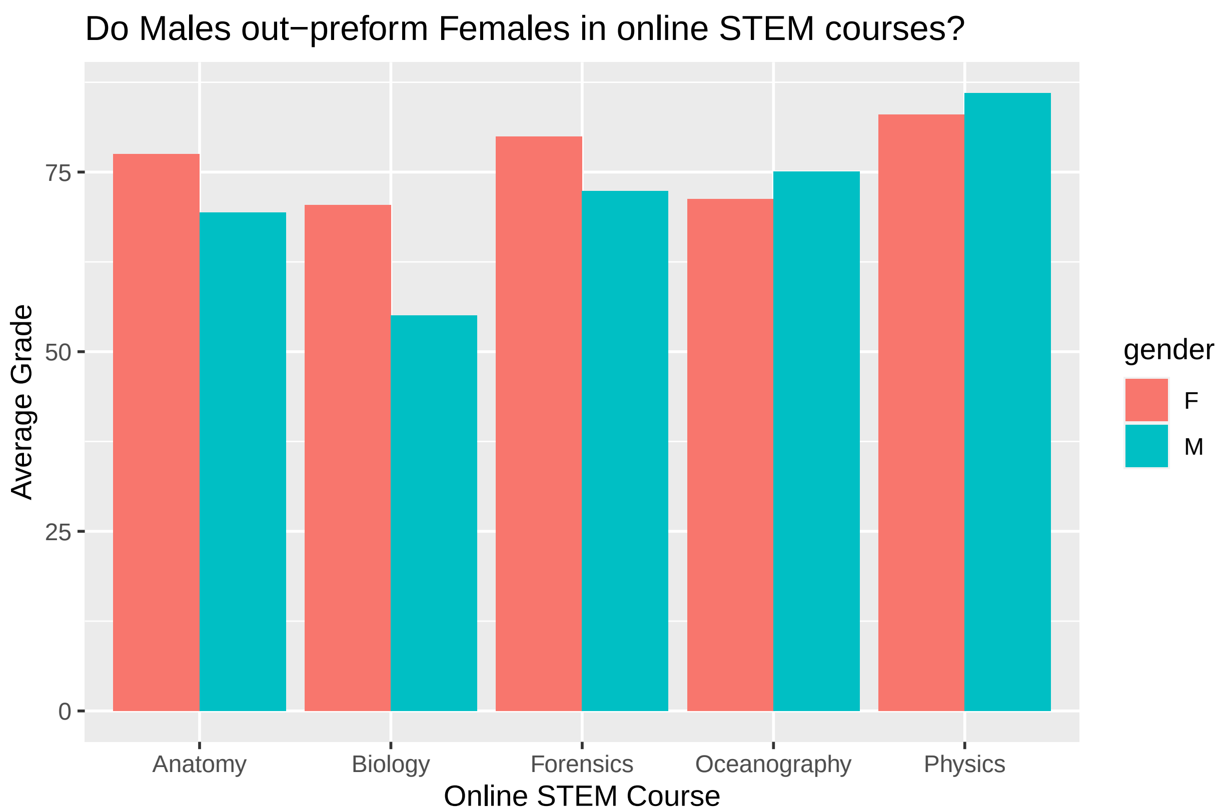
```
##
##
##         Predictor      b     b_95%_CI beta  beta_95%_CI sr2  sr2_95%_CI     r
##       (Intercept) 0.45** [0.32, 0.58]
##  time_spent_hours 0.00** [0.00, 0.01] 0.41 [0.33, 0.48] .16  [.11, .22] .41**
##               int 0.05** [0.02, 0.08] 0.12 [0.04, 0.19] .01 [-.00, .03] .15**
##
##
##
##              Fit
##
##
##
##      R2 = .186**
##  95% CI[.13,.24]
##
##
## Note. A significant b-weight indicates the beta-weight and semi-partial correlation are also signifi
## b represents unstandardized regression weights. beta indicates the standardized regression weights.
## sr2 represents the semi-partial correlation squared. r represents the zero-order correlation.
## Square brackets are used to enclose the lower and upper limits of a confidence interval.
## * indicates p < .05. ** indicates p < .01.
##
```

0.7 Communicate

RQ1: Do males outperform females in online STEM courses?

```r
data_viz <- data_to_explore %>%
  select(subject, gender, proportion_earned) %>%  # reduced
  mutate(subject = recode(subject,
                          "AnPhA" = "Anatomy",
                          "BioA" = "Biology",
                          "FrScA" = "Forensics",
                          "OcnA" =  "Oceanography",
                          "PhysA" = "Physics")) %>%
  mutate(grade = proportion_earned * 100) %>%
  # filter(!is.na(gender)) %>%
  na.omit() %>% # removed all NAs instead of just those for gender
  group_by(subject, gender) %>% # grouped by subject and gender
  summarise(grade = mean(grade),
            sd = sd(grade))# calculated mean and sd for grade and saved as grade again

  ggplot(data_viz, aes(x = subject, y = grade,
                       fill = gender)) +
  geom_bar(stat = "identity",
           position = position_dodge()) +
  labs(title = "Do Males out-preform Females in online STEM courses?",
       caption = "Online STEM course performance, why is there still a gender gap?",
       y = "Average Grade",
       x = "Online STEM Course")
```



Online STEM course performance, why is there still a gender gap?