



Overview of Learning Analytics

Learning Analytics Section

Agenda



- Attendance Check (**Remember to check your attendance on EdX**)
- Housekeeping
- R Homework Walkthrough
- Work with Your Group

Housekeeping



- Remember to check your attendance on EdX
- Add to Edx & Slack
- Take the survey on Edx Week 1
- Post questions in #help or go to office hours
- Form study group, create a private slack channel “DIS20X-[Group Name]” and add your section leader

Reading Discussion



Talk about the question in your group:

- What kind of data/system would help you answer your question?
What level of data would it be (micro-, meso, macro-)?
- If you can only use traditional methods (e.g., survey, interview, etc.), how would it be different from using big data methods?
- What do you think are the advantages of using big data methods?
What do you think are some challenges in using big data methods?

R Homework Walkthrough


Open your RStudio, download the hw file and load the dataset.

Understand the context and the dataset:

- Each column indicates a variable
- Each row indicates a record of event (Log data)

		studentID <int>	itemid <int>	correctonfirstattempt <int>	attempts <int>	hints <int>	seconds <int>	full_start_time <fctr>
→	1	136	90	1	1	0	58	01-OCT-04 07.44.43.000000 AM
→	2	136	91	0	1	3	91	10-DEC-04 09.27.20.000000 AM
→	3	136	92	1	1	0	11	10-DEC-04 09.28.51.000000 AM
→	4	136	93	1	1	0	10	10-DEC-04 09.29.02.000000 AM
→	5	136	94	0	2	0	43	10-DEC-04 09.29.12.000000 AM
→	6	136	95	1	1	0	13	15-OCT-04 07.44.14.000000 AM

R Homework Walkthrough




	studentID <int>	itemid <int>	correctonfirstattempt <int>	attempts <int>	hints <int>	seconds <int>	full_start_time <fctr>
1	136	90	1	1	0	58	01-OCT-04 07.44.43.000000 AM
2	136	91	0	1	3	91	10-DEC-04 09.27.20.000000 AM
3	136	92	1	1	0	11	10-DEC-04 09.28.51.000000 AM
4	136	93	1	1	0	10	10-DEC-04 09.29.02.000000 AM
5	136	94	0	2	0	43	10-DEC-04 09.29.12.000000 AM
6	136	95	1	1	0	13	15-OCT-04 07.44.14.000000 AM

6 rows | 1-8 of 12 columns

Whenever you look at a new dataset..

1. Check the number of columns and rows
2. Check the unique number of students, items → HW Question 3,4
3. Check the level of dataset → Is this at the student level? student-attempt level? Or student-item level?

R Homework Walkthrough



	studentID <int>	itemid <int>	correctonfirstattempt <int>	attempts <int>	hints <int>	seconds <int>	full_start_time <fctr>
1	136	90	1	1	0	58	01-OCT-04 07.44.43.000000 AM
2	136	91	0	1	3	91	10-DEC-04 09.27.20.000000 AM
3	136	92	1	1	0	11	10-DEC-04 09.28.51.000000 AM
4	136	93	1	1	0	10	10-DEC-04 09.29.02.000000 AM
5	136	94	0	2	0	43	10-DEC-04 09.29.12.000000 AM
6	136	95	1	1	0	13	15-OCT-04 07.44.14.000000 AM

6 rows | 1-8 of 12 columns

Whenever you look at a new dataset..

1. Check the number of columns and rows
2. Check the unique number of students, items → HW Question 3,4
3. Check the level of dataset → Is this at the student level? student-attempt level?
Or student-item level?

R Homework Walkthrough

3

	studentID <int>	itemid <int>	correctonfirstattempt <int>	attempts <int>	hints <int>	seconds <int>	full_start_time <fctr>
1	136	90	1	1	0	58	01-OCT-04 07.44.43.000000 AM
2	136	91	0	1	3	91	10-DEC-04 09.27.20.000000 AM
3	136	92	1	1	0	11	10-DEC-04 09.28.51.000000 AM
4	136	93	1	1	0	10	10-DEC-04 09.29.02.000000 AM
5	136	94	0	2	0	43	10-DEC-04 09.29.12.000000 AM
6	136	95	1	1	0	13	15-OCT-04 07.44.14.000000 AM

6 rows | 1-8 of 12 columns

Questions:

What variables can you think of if the dataset is on the student-attempt level?

What variables can you think of if the dataset is on the student level?

R Homework Walkthrough

3

	studentID <int>	itemid <int>	correctonfirstattempt <int>	attempts <int>	hints <int>	seconds <int>	full_start_time <fctr>
1	136	90	1	1	0	58	01-OCT-04 07.44.43.000000 AM
2	136	91	0	1	3	91	10-DEC-04 09.27.20.000000 AM
3	136	92	1	1	0	11	10-DEC-04 09.28.51.000000 AM
4	136	93	1	1	0	10	10-DEC-04 09.29.02.000000 AM
5	136	94	0	2	0	43	10-DEC-04 09.29.12.000000 AM
6	136	95	1	1	0	13	15-OCT-04 07.44.14.000000 AM

6 rows | 1-8 of 12 columns

Questions:

What variables can you think of if the dataset is on the student-attempt level?

→ 'hints' column can be 'hint_used' with binary values instead of numeric values

What variables can you think of if the dataset is on the student level?

→

R Homework Walkthrough

3

	studentID <int>	itemid <int>	correctonfirstattempt <int>	attempts <int>	hints <int>	seconds <int>	full_start_time <fctr>
1	136	90	1	1	0	58	01-OCT-04 07.44.43.000000 AM
2	136	91	0	1	3	91	10-DEC-04 09.27.20.000000 AM
3	136	92	1	1	0	11	10-DEC-04 09.28.51.000000 AM
4	136	93	1	1	0	10	10-DEC-04 09.29.02.000000 AM
5	136	94	0	2	0	43	10-DEC-04 09.29.12.000000 AM
6	136	95	1	1	0	13	15-OCT-04 07.44.14.000000 AM

6 rows | 1-8 of 12 columns

Questions:

What variables can you think of if the dataset is on the student-attempt level?

→ 'hints' column can be 'hint_used' with binary values instead of numeric values

What variables can you think of if the dataset is on the student level?

→ 'attempts' column can be 'total_number_attempts' that is sum of attempts that a student tried

R Homework Walkthrough



Understand the context and the dataset:

- Each column indicates a variable
- Each row indicates a record of event

R Homework Walkthrough



Understand the context and the dataset:

- Each column indicates a variable
- Each row indicates a record of event


Question: What variables can you think of if the dataset is on the attempt-student level? What variables can you think of if the dataset is on the student level?

R Homework Walkthrough



Useful base R functions (Ask how to use these functions in R with ? or ask our virtual TA!):

- `$`, `c()`
- `dim()`, `nrow()`
- `head()` - return the first part of vector, matrix, table, data frame or function - `head(x)`
- `length()` - get the length of **vector/lists and factors** - `length(x)`
- `unique()` - return a **vector, data frame or array** with duplicated elements/rows removed
- Calculation: `mean()`, `sum()`, ...
 - `mean()` & `sum()` turns the variable into a boolean when a condition is given in the parenthesis, e.g., `mean(x == 1)`, `(asm$attempts > 1)`
- `table()` - return two-way cross table or two-way frequency table along with proportion - `table(x)`, `table(x, y)`

R Function	Python Equivalent	What It Does
\$ 	<code>df['column_name']</code> or <code>df.column_name</code>	Accesses elements of a list or columns of a dataframe by name.
<code>c()</code>	<code>list1 + list2</code> or <code>np.concatenate()</code>	Combines values into a vector or list.
<code>dim()</code> / <code>nrow()</code>	<code>df.shape</code> / <code>len(df)</code> or <code>df.shape[0]</code>	Retrieves the dimensions of an array, matrix, or dataframe/Retrieves the number of rows in a dataframe or matrix.
<code>head()</code>	<code>df.head()</code>	Returns the first part (rows) of a dataframe.
<code>length()</code>	<code>len(list)</code> or <code>len(df)</code>	Gets the length of a vector, list, or the number of rows in a dataframe.
<code>unique()</code>	<code>pd.unique(df['column'])</code> or <code>np.unique(array)</code>	Returns unique elements from a vector, dataframe column, or array.
<code>mean()</code> , <code>sum()</code>	<code>np.mean(array)</code> , <code>np.sum(array)</code> or <code>df['column'].mean()</code> , <code>df['column'].sum()</code>	Performs basic statistical calculations like mean and sum over arrays or dataframe columns.
<code>table()</code>	<code>pd.crosstab(index=df['column1'], columns=df['column2'])</code>	Returns a cross-tabulation of two (or more) factors, akin to a frequency table.

R Homework Walkthrough



The tidyverse package:

- Install and use a package: `install.packages()`, `library()`
- Manipulations by row: `filter()`, `arrange()`
- Manipulations by column: `select()`, `mutate()`
- Manipulations by group: `group_by()`, `%>%`, `summarise()`
 - Think of the pipe `%>%` as another way of saying “and then do the following:”
 - It feeds your dataset into new functions as you modify it

R Homework Walkthrough



The tidyverse package:

- Install and use a package: `install.packages()`, `library()`
- Get an overview of the dataset: `glimpse()`
- Manipulations by row: `filter()`, `arrange()`, ...
- Manipulations by column: `select()`, `mutate()`, ...
- Manipulations by group: `group_by()`, `%>%`, `summarise()`
 - Think of the pipe `%>%` as another way of saying “and then do the following:”
 - It feeds your dataset into new functions as you modify it

Question: Work in groups. Create a student level dataset – Compute the average minutes that every student spend in questions that they answered correctly in their first attempt (it's OK to ignore students who did not get any questions correct in their first attempt).

R Homework Walkthrough

The tidyverse package:

- Install and use a package: `install.packages("tidyverse")`
- Get an overview of the dataset: `library(tidyverse)`
- Manipulations by row: `filter()`, `select()`
- Manipulations by column: `select()`
- Manipulations by group: `group_by()`
 - Think of the pipe `%>%` as a "chain" of operations following:
 - It feeds your dataset into the next function

Question: Work in groups. Create a student dataset that every student spend in questions that they got correct. OK to ignore students who did not get any questions correct in their first attempt.

```
> student <- asm %>%
+   filter(correctonfirstattempt == 1) %>%
+   mutate(minutes = seconds/60) %>%
+   group_by(studentID) %>%
+   summarise(
+     mean_minutes = mean(minutes)
+   )
> head(student)
# A tibble: 6 × 2
  studentID mean_minutes
  <int>         <dbl>
1     136         0.442
2     137         0.534
3     139         0.505
4     140         0.552
5     141         1.04
6     143         0.514
```

R Homework Walkthrough



Visualization functions:

- Base R functions
 - **Histogram (one variable)**
`hist(x)` or `hist(x, breaks = ..., main = ..., xlim = ..., ylim = ..., ...)`
 - **Scatter Plot (two continuous variables)**
`plot(x,y)`
 - **Box Plot (two variables, one continuous variable and one categorical variable)**
`boxplot(dependent_variable ~ independent_variable, dataset_name)`

Continuous variable Categorical variable



Interpret the plots

R Homework Walkthrough



Visualization functions:

- `ggplot()`: it is included in the tidyverse package and follows tidyverse syntax
 - Histogram
`ggplot() + geom_histogram() + ...`
 - Scatter Plot
`ggplot() + geom_point() + ...`
 - Box Plot
`ggplot() + geom_boxplot() + ...`

Resource (R for Data Science): <https://r4ds.had.co.nz/data-visualisation.html>



Resources

- <https://swirlstats.com/>
- R for Data Science: <https://r4ds.had.co.nz/index.html>
- R Markdown Tutorial: <https://rmarkdown.rstudio.com/>
- On-campus R resource
<https://cscu.cornell.edu/workshops/current-schedule/>



Remaining Questions?

-