CIND 820 XJH Final Project by Sylvia Pereira

Supervisor: Dr. Tamer Abdou tamer.abdou@torontomu.ca

Data Preparation

The data preparation was performed in Alteryx Designer Desktop software.

Glossary
VLE: Virtual Learning Environment

Below you can follow the steps I used to prepare the dataset for the Exploratory Data Analysis (EDA) report.

Original datasets can be found at: https://analyse.kmi.open.ac.uk/open_dataset

<table>
<tr>
<td>

Step 1

StudentInfo.csv

- 32,593 records originally and 12 columns.
- Filter students that Pass or Fail the course and remove undesired columns.
- Other final results such as withdrawn, distinction, etc. are not part of this analysis.

</td>
<td>

Result: Table with 4 columns and 19,413 records.

Results - Select (44) - Output

4 of 4 Fields | Cell Viewer | 19,413 records displayed

Record	code_module	code_presentation	id_student	final_result
1	AAA	2013J	11391	Pass
2	AAA	2013J	28400	Pass
3	AAA	2013J	31604	Pass
4	AAA	2013J	32885	Pass
5	AAA	2013J	38053	Pass
6	AAA	2013J	45462	Pass
7	AAA	2013J	45642	Pass
8	AAA	2013J	52130	Pass
9	AAA	2013J	53025	Pass
10	AAA	2013J	57506	Pass
11	AAA	2013J	58873	Pass
12	AAA	2013J	59185	Pass

</td>
</tr>
<tr>
<td>

Step 2

StudentVle.csv

- 10,655,280 records originally and 6 columns.

vle.csv

- 6.364 records originally and 6 columns.

These two datasets were combined to summarize the VLE interactions per student. Duplicates and undesired fields were removed.

</td>
<td>

Result: Table with 6 columns and 10,655,280 records.

Results - Join (48) - Out - Join

6 of 6 Fields | Cell Viewer | * 26,761 of 10,655,280 records displayed(partial results)

Record	id_student	code_module	code_presentation	id_site	sum_click	activity_type
1	331358	AAA	2013J	546914	1	resource
2	617805	FFF	2014B	779811	1	url
3	402806	FFF	2014B	779811	1	url
4	631311	FFF	2014B	779811	1	url
5	631311	FFF	2014B	779811	1	url
6	627620	FFF	2014B	779811	1	url
7	589548	FFF	2014B	779811	1	url
8	600702	FFF	2014B	779811	1	url
9	628846	FFF	2014B	779811	1	url
10	629540	FFF	2014B	779811	1	url
11	625753	FFF	2014B	779811	1	url
12	607453	FFF	2014B	779811	3	url

</td>
</tr>
</table>

Step 3

Using the Join tool, the results of **StudentInfo.csv** and the combined dataset from **StudentVle.csv** + **vle.csv** were joined again to form the big dataset used in the association analysis.

Result: Table with 7 columns and 8,754,616 records.

Results - Join (41) - Out - Join

7 of 7 Fields | Cell Viewer | * 23,996 of 8,754,616 records displayed(partial results)

Record	id_student	course_name	code_presentation	id_site	activity_type	sum_click	final_result
1	349182	DDD	2013J	674010	subpage	1	Fail
2	349182	DDD	2013J	674449	url	1	Fail
3	349182	DDD	2013J	674100	subpage	2	Fail
4	349182	DDD	2013J	674100	subpage	2	Fail
5	349182	DDD	2013J	674100	subpage	1	Fail
6	349182	DDD	2013J	673740	oucontent	11	Fail
7	349182	DDD	2013J	673740	oucontent	4	Fail
8	349182	DDD	2013J	673740	oucontent	2	Fail
9	349182	DDD	2013J	673740	oucontent	5	Fail
10	349182	DDD	2013J	673740	oucontent	1	Fail
11	349182	DDD	2013J	673727	oucontent	2	Fail
12	349182	DDD	2013J	673727	oucontent	1	Fail

Step 4

Using the Find Replace tool, the fields course_name and code_presentation were replaced by one field called **course_year_month** to clarify the information.

The old fields course_name and code_presentation were dropped from the dataset as well as the id_site field, since it's not relevant for this analysis.

Result: Table with 5 columns and 8,754,616 records.

Results - Select (5) - Output

5 of 5 Fields | Cell Viewer | * 23,689 of 8,754,616 records displayed(partial results)

Record	id_student	course_year_month	activity_type	sum_click	final_result
1	349182	DDD 2013_October	subpage	1	Fail
2	349182	DDD 2013_October	url	1	Fail
3	349182	DDD 2013_October	subpage	2	Fail
4	349182	DDD 2013_October	subpage	2	Fail
5	349182	DDD 2013_October	subpage	1	Fail
6	349182	DDD 2013_October	oucontent	11	Fail
7	349182	DDD 2013_October	oucontent	4	Fail
8	349182	DDD 2013_October	oucontent	2	Fail
9	349182	DDD 2013_October	oucontent	5	Fail
10	349182	DDD 2013_October	oucontent	1	Fail
11	349182	DDD 2013_October	oucontent	2	Fail
12	349182	DDD 2013_October	oucontent	1	Fail

Step 5

To reduce the number of records the fields: id_student, course_year_month, activity_type were grouped while the function "sum" was used for sum_clicks.

Then, the table was sorted by ID_Student field.

Result: Table with 5 fields and 167,933 records

Results - Sort (19) - Output

5 of 5 Fields | Cell Viewer | * 23,947 of 167,933 records displayed(partial results)

Record	ID_Student	course_year_month	Activity_Type	Sum_click	final_result
1	6516	AAA 2014_October	homepage	497	Pass
2	6516	AAA 2014_October	forumng	451	Pass
3	6516	AAA 2014_October	subpage	143	Pass
4	6516	AAA 2014_October	dataplus	21	Pass
5	6516	AAA 2014_October	resource	31	Pass
6	6516	AAA 2014_October	oucontent	1505	Pass
7	6516	AAA 2014_October	url	143	Pass
8	11391	AAA 2013_October	oucontent	553	Pass
9	11391	AAA 2013_October	url	5	Pass
10	11391	AAA 2013_October	forumng	193	Pass
11	11391	AAA 2013_October	homepage	138	Pass
12	11391	AAA 2013_October	resource	13	Pass

Step 6

With the Formula tool, three new columns were added in the dataset: access, Pass and Fail.

I populated the **access** column with a binary attribute corresponding to the visits in the VLE. The interaction of a user with a resource will be represented by **1** if the user visited the resource at least once and **0** if the resource was not visited at all.

The Pass and Fail columns were also populated with binary attributes and the columns **final_result** and sum_click were dropped using the Select tool since they will not be used in the next steps.

Result:

Results - Select (67) - Output

6 of 6 Fields | Cell Viewer | * 25,204 of 167,933 records displayed(partial results)

Record	ID_Student	course_year_month	Activity_Type	access	Pass	Fail
1	6516	AAA 2014_October	homepage	1	1	0
2	6516	AAA 2014_October	forumng	1	1	0
3	6516	AAA 2014_October	subpage	1	1	0
4	6516	AAA 2014_October	dataplus	1	1	0
5	6516	AAA 2014_October	resource	1	1	0
6	6516	AAA 2014_October	oucontent	1	1	0
7	6516	AAA 2014_October	url	1	1	0
8	11391	AAA 2013_October	oucontent	1	1	0
9	11391	AAA 2013_October	url	1	1	0
10	11391	AAA 2013_October	forumng	1	1	0
11	11391	AAA 2013_October	homepage	1	1	0
12	11391	AAA 2013_October	resource	1	1	0

Step 7

In the final step, using the CrossTab tool, I pivoted the orientation of the table by moving vertical data onto the horizontal axis.

The data corresponding to learner's trajectories through the VLE are shown on which each row, and each column corresponds to a particular resource within the VLE.

Results - Data Cleansing (36) - Output26

24 of 24 Fields | Cell Viewer | * 17,356 of 19,077 records displayed(partial results)

Record	ID_Student	course_year_month	Pass	Fail	dataplus	dualpane	externalquiz	folder	forumng
1	186670	BBB 2013_February	0	1	0	0	0	0	1
2	590974	CCC 2014_October	0	1	0	0	0	0	0
3	466425	FFF 2013_October	0	1	1	0	0	1	1
4	628759	GGG 2014_February	0	1	0	0	0	0	1
5	341509	BBB 2013_February	1	0	0	0	0	0	1
6	645371	BBB 2014_October	0	1	0	0	0	0	0
7	40419	DDD 2013_October	1	0	0	0	1	0	1
8	680569	GGG 2014_October	1	0	0	0	0	0	1
9	591655	CCC 2014_October	0	1	0	1	0	0	1
10	585603	BBB 2013_October	1	0	0	0	0	0	1
11	650424	FFF 2014_October	1	0	1	1	0	0	1
12	2497624	BBB 2014_October	1	0	0	0	0	0	1

Note: Although we have different site_id representing each resource present in the VLE for this association analysis project I will be using the main categories represented by names.