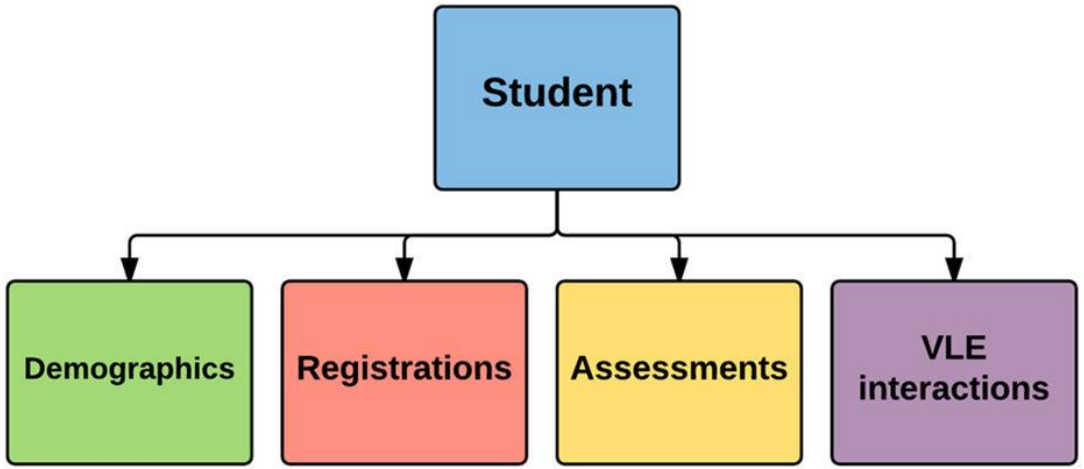
SUBMITTED ON
MARCH 5 2018

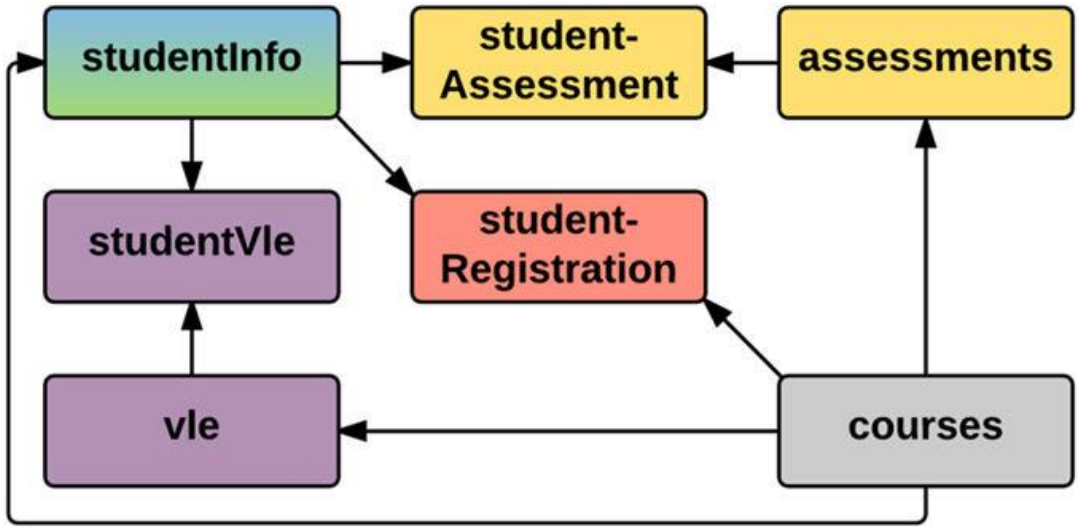
BY AASHRAY YADAV

Adobe Data Challenge

About the Data set

This data set is an open university database consisting of rich content on student performance based on various criteria including module assessment, demographic and interaction in the Virtual Learning Environment (VLE). It is aimed at improving the overall learning experience for students.





Database Hierarchy and Relationship

Metadata

The dataset contains the information about **22 courses**, **32,593 students**, their assessment results, and logs of their interactions with the VLE represented by daily summaries of student clicks (**10,655,280 entries**).

There are 3 types of data:

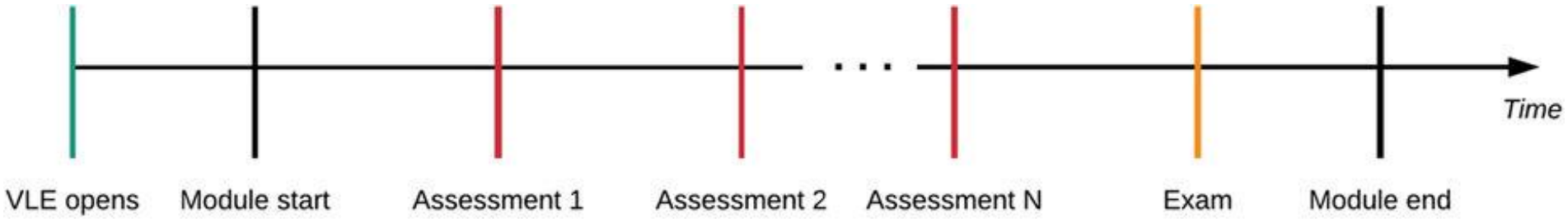
a) *Demographic*: basic information of students including age, gender, region, previous education, etc.

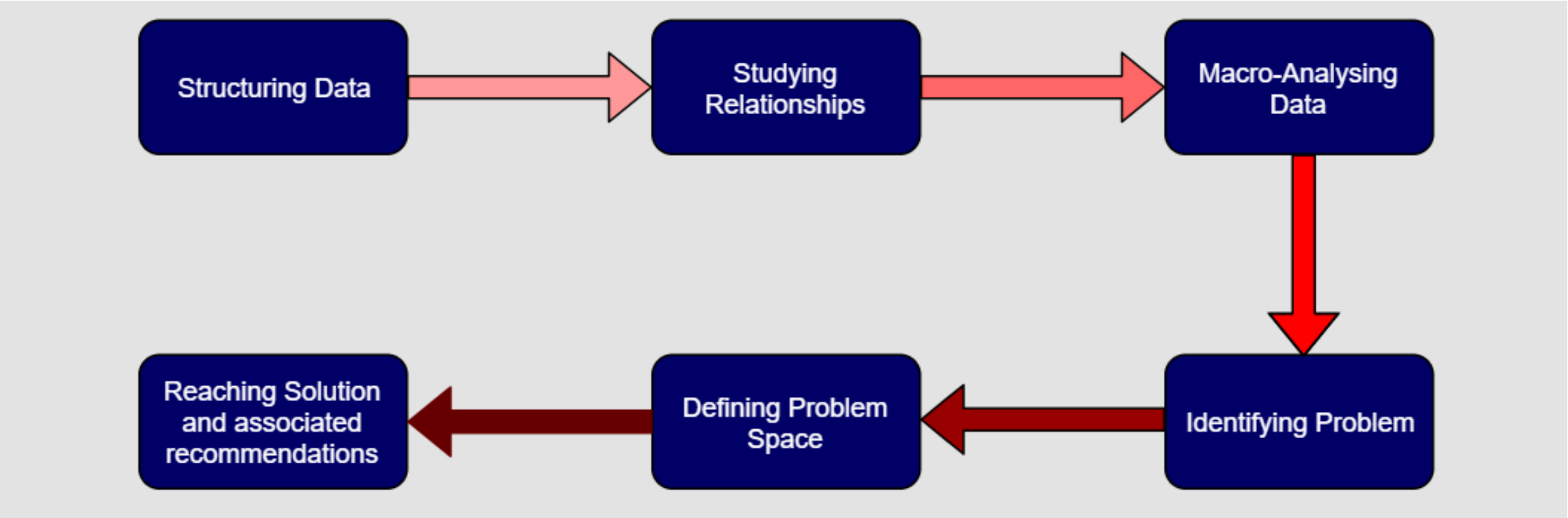
b) *Performance*: Student's results and achievements at the OU

c) *Learning Behaviour*: log of student activities and habits in the VLE.

Assessment Process Flow

Module-presentation content is usually available in VLE couple of weeks before official module start. During the presentation the students' their knowledge is evaluated in series of assessments, which defines the milestones in the module. At the end, there is usually the final exam.





Problem Flowchart

Demographic

Assessment

Interaction &Content

Structuring the Data

	code_module	code_presentation	id_student	highest_education	imd_band	age_band	final_result
0	AAA	2013J	11391	HE Qualification	90-100%	55<=	Pass
1	AAA	2013J	28400	HE Qualification	20-30%	35-55	Pass
2	AAA	2013J	30268	A Level or Equivalent	30-40%	35-55	Withdrawn
3	AAA	2013J	31604	A Level or Equivalent	50-60%	35-55	Pass
4	AAA	2013J	32885	Lower Than A Level	50-60%	0-35	Pass

Demographic

	id_student	code_module	code_presentation	Mean_Score	CMA_Score	TMA_Score	Exam_Score
0	6516	AAA	2014J	61.800000	NaN	61.800000	NaN
1	8462	DDD	2013J	87.666667	NaN	87.666667	NaN
2	8462	DDD	2014J	86.500000	NaN	86.500000	NaN
3	11391	AAA	2013J	82.000000	NaN	82.000000	NaN
4	23629	BBB	2013B	82.500000	100.0	65.000000	NaN

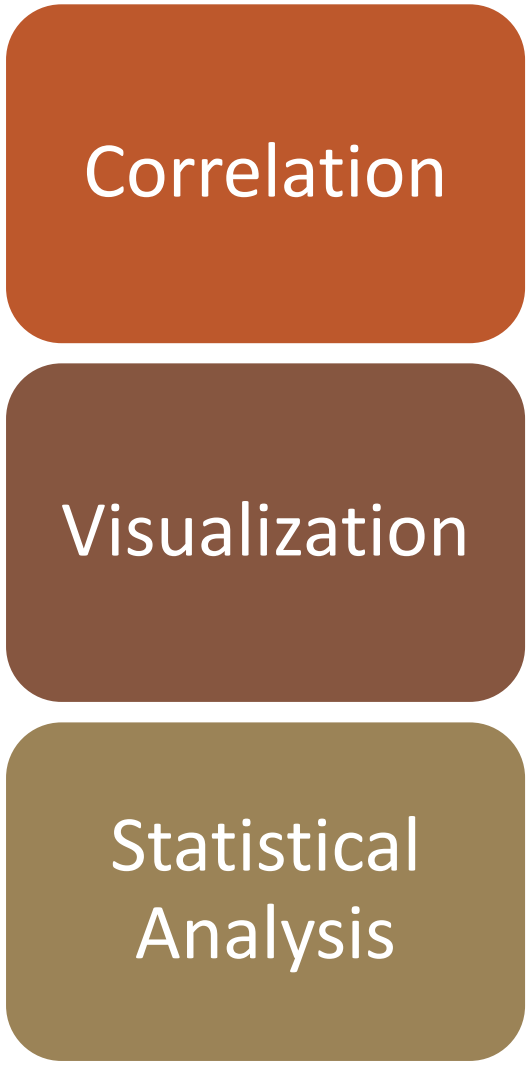
Assessment

	id_student	code_module	code_presentation	sum_click	days_interacted	daily_click	count dataplus	count dualpane	count externalquiz	count folder	...	count ouelluminate	count ouwiki
0	6516	AAA	2014J	2791	159	17.553459	4.0	0.0	0.0	0.0	...	0.0	0.0
1	24734	AAA	2014J	499	56	8.910714	4.0	0.0	0.0	0.0	...	0.0	0.0
2	26192	AAA	2014J	2223	118	18.838983	4.0	0.0	0.0	0.0	...	0.0	0.0
3	28061	AAA	2014J	1590	148	10.743243	4.0	0.0	0.0	0.0	...	0.0	0.0
4	31600	AAA	2014J	429	19	22.578947	4.0	0.0	0.0	0.0	...	0.0	0.0

Interaction & Content

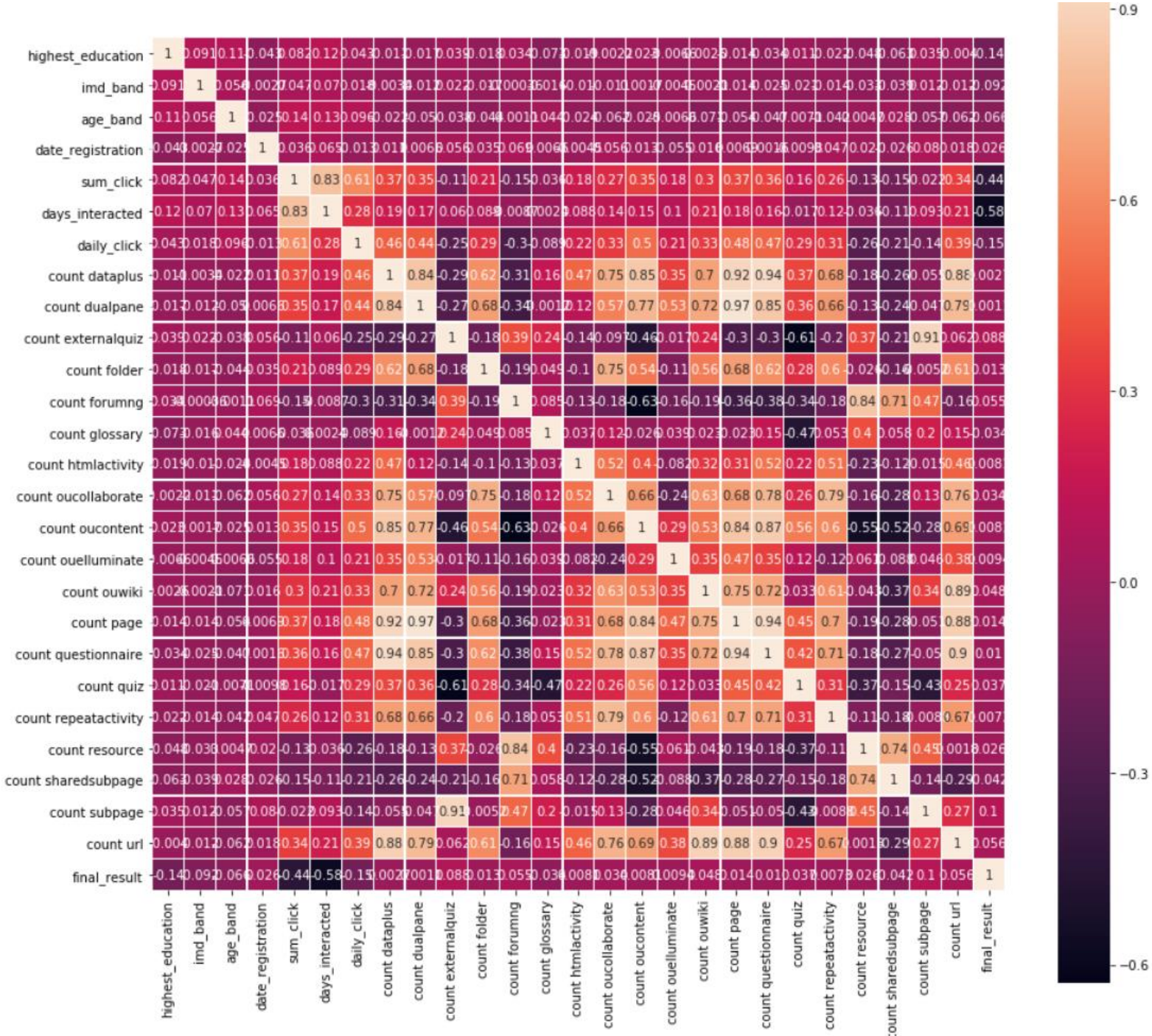
	highest_education	imd_band	age_band	date_registration	sum_click	days_interacted	daily_click	count dataplus	count dualpane	count externalquiz	...	count ouwiki	count page	c
0	3	8	2	-52.0	2791.0	159.0	17.553459	4.0	0.0	0.0	...	0.0	0.0	
1	3	3	2	-137.0	646.0	56.0	11.535714	0.0	0.0	7.0	...	4.0	0.0	
2	3	3	2	-38.0	10.0	1.0	10.000000	0.0	0.0	5.0	...	1.0	0.0	
3	3	9	2	-159.0	934.0	40.0	23.350000	4.0	0.0	0.0	...	0.0	0.0	
4	1	2	0	-47.0	161.0	16.0	10.062500	0.0	0.0	0.0	...	0.0	0.0	

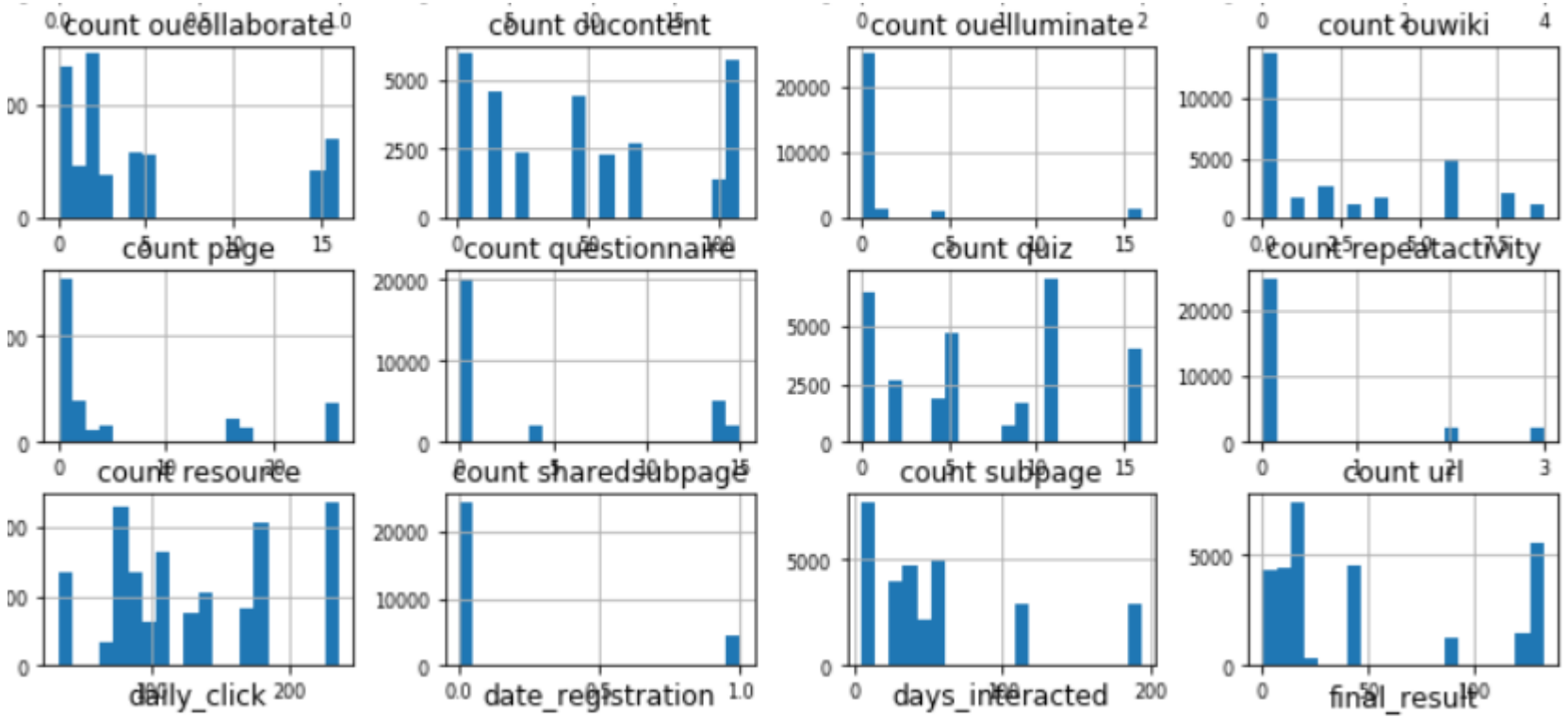
Aggregation: Cleaning and Binning



Studying Relationships

Pearson's Correlation Matrix





Visualization viz. Histograms

Statistical Analysis of Features

F-Score

```
[  2.78290473e+02   1.07793400e+02   6.82772597e+01   3.53120495e+01
   2.52443089e+03   5.40489970e+03   2.58279869e+02   6.87385148e+00
   8.31306443e+00   7.74752386e+01   8.11432955e+00   3.17289580e+01
   8.36717835e+01   1.69309161e+01   1.59482749e+01   1.03152943e+01
   1.65103392e+01   2.88438334e+01   3.89441508e+00   2.93300734e+00
   7.50047080e+01   2.10393111e+00   4.52483010e+01   4.50862861e+01
   1.07402939e+02   3.24747863e+01]
```

P-Value

```
[  4.06439709e-178   2.09123792e-069   5.42963246e-044   9.02799339e-023
   0.00000000e+000   0.00000000e+000   1.90343594e-165   1.26569656e-004
   1.59775179e-005   6.52106585e-050   2.12812211e-005   1.81574915e-020
   6.71718914e-054   5.53167086e-011   2.34124465e-010   8.79375768e-007
   1.02593773e-010   1.29575925e-018   8.55933677e-003   3.21029255e-002
   2.53689490e-048   9.74123576e-002   3.64841282e-029   4.63869699e-029
   3.72573516e-069   6.02164896e-021]
```

Note: Higher F-score and Lower P-value indicate importance of feature

Naïve Bayes

Logistic Regression

Random Forest

Macro-Analysis

Naïve Bayes

	precision	recall	f1-score	support
0	0.35	0.27	0.30	2119
1	0.69	0.50	0.58	8644
2	0.32	0.30	0.31	4680
3	0.44	0.71	0.55	5011
avg / total	0.51	0.48	0.48	20454

	precision	recall	f1-score	support
0	0.35	0.26	0.30	905
1	0.69	0.51	0.59	3714
2	0.34	0.32	0.33	1998
3	0.45	0.71	0.55	2150
avg / total	0.51	0.49	0.49	8767

```
confusion matrix:
 [[ 239  467   92  107]
 [ 396 1895  691  732]
 [  34  256  643 1065]
 [  17  141  465 1527]]
Traning accuracy: 0.48381734624034417, Testing accuracy: 0.4909319037298962
```

Logistic Regression

	precision	recall	f1-score	support
0	0.55	0.07	0.13	2119
1	0.64	0.85	0.73	8644
2	0.45	0.26	0.33	4680
3	0.58	0.69	0.63	5011
avg / total	0.57	0.60	0.55	20454

	precision	recall	f1-score	support
0	0.50	0.06	0.11	905
1	0.64	0.86	0.73	3714
2	0.45	0.24	0.32	1998
3	0.58	0.69	0.63	2150
avg / total	0.56	0.59	0.55	8767

```
confusion matrix:
 [[  54  813   16   22]
 [  53 3198  269  194]
 [   2  653  489  854]
 [   0  358  319 1473]]
traning accuracy: 0.5966559108242886, testing accuracy: 0.5947302383939774
```

Random Forest

	precision	recall	f1-score	support
0	0.31	0.74	0.43	2119
1	0.69	0.49	0.58	8644
2	0.47	0.38	0.42	4680
3	0.62	0.67	0.64	5011
avg / total	0.58	0.54	0.54	20454

	precision	recall	f1-score	support
0	0.27	0.69	0.39	905
1	0.66	0.46	0.54	3714
2	0.44	0.35	0.39	1998
3	0.60	0.64	0.62	2150
avg / total	0.55	0.50	0.51	8767

confusion matrix:

```
[[ 622  240   24   19]
 [1424 1724  409  157]
 [ 158  412  691  737]
 [  74  244  451 1381]]
```

traning accuracy: 0.5356409504253447, testing accuracy: 0.503935211588913

Identifying Problem

```
In [190]:
1 wnp=13839/29221 * 100
2 wnp

Out[190]: 47.359775503918414
```

Problem: 47% students have either failed or withdrawn from the program

1. Focus on Final Result as decisive factor
2. Diving Deep into regressive cases of ***'Fail'*** and ***'Withdrawn'***

Process	Normalizing data
Spot the gold	Spotting pertinent features
Apply	Running models

Defining the problem space

```
# Normalizing X

from sklearn.preprocessing import StandardScaler
X_norm = StandardScaler().fit_transform(X)

# Calculating F-score and P-value for each of the features

from sklearn.feature_selection import f_classif, SelectKBest
Fs, pval = f_classif(X_norm,Y)
print(Fs)
print(pval)

```

```
[  1.87993810e+01   5.31132202e+00   1.12599132e+01   1.03307726e+02
   2.89995089e+02   5.47320017e+02   5.06199690e-01   1.26287631e-01
   1.25574017e+01   3.10933469e+01   7.46416048e+00   4.03389080e+00
   5.67112699e+01   4.18315973e+01   3.06780237e+01   2.96894946e+01
   2.48197084e+01   2.01258763e-01   1.00456089e+00   7.15323893e-02
   4.64652505e+01   6.12031091e+00   5.56350508e+01   1.24379118e+02
   5.04933144e+01   5.41295596e+00]
[  1.46241022e-005   2.12020051e-002   7.94119177e-004   3.48911807e-024
   2.25384135e-064   9.58824589e-119   4.76800619e-001   7.22318054e-001
   3.95940610e-004   2.50516940e-008   6.30192151e-003   4.46143757e-002
   5.35934339e-014   1.02825184e-010   3.10150114e-008   5.15762742e-008
   6.37104325e-007   6.53713747e-001   3.16226943e-001   7.89122538e-001
   9.71201099e-012   1.33756360e-002   9.24425467e-014   9.22914594e-029
   1.25420530e-012   2.00022056e-002]
```

Normalization

```
# Finding 5 most significant features out of the 26 features for deeper analysis
selectK = SelectKBest(f_classif, k = 5).fit(X_norm,Y)
selectK.get_support()
```

```
array([False, False, False,  True,  True,  True, False, False, False,
       False, False, False,  True, False, False, False, False, False,
       False, False, False, False, False,  True, False, False], dtype=bool)
```

Spot *'right'* relevance

Naïve Bayes – 59%

(^49%)

Logistic Regression – 63%
(^59%)

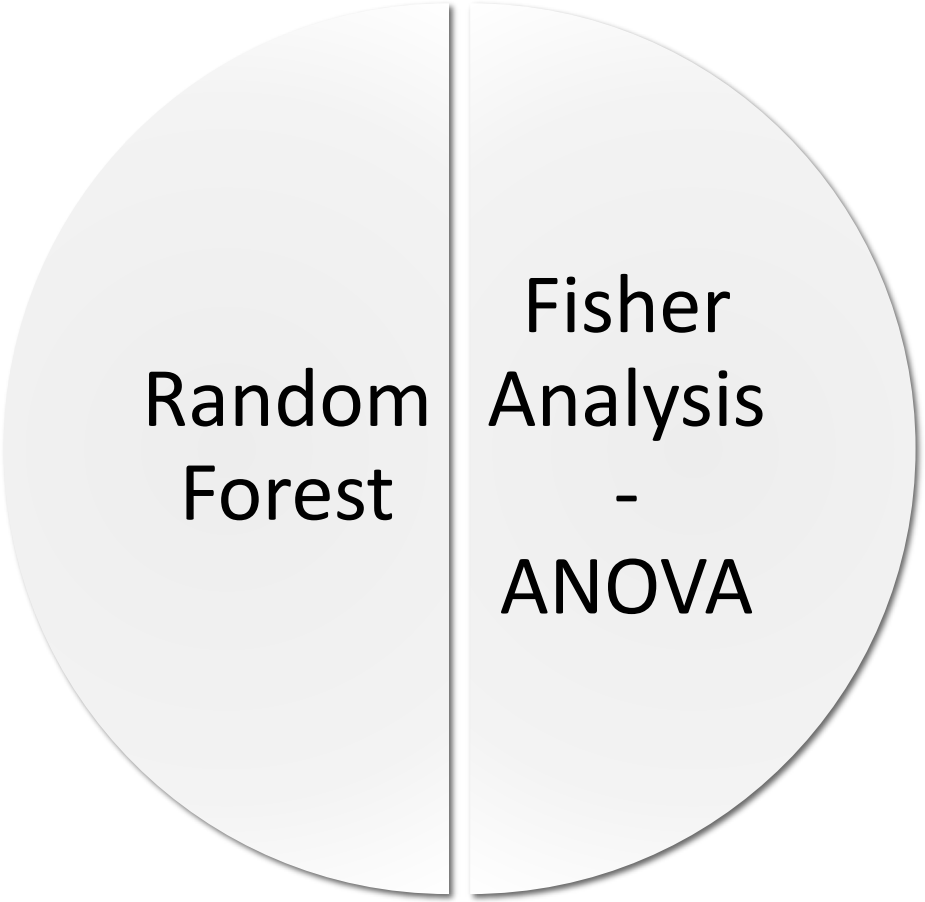
Random Forest – 63%
(^49%)

Applying models

Solution and Recommendations

Final Selection and Impact

Final Verdict



Feature Selection and Impact

ANOVA Outcomes

High F-Score indicates low P-Value and therefore, establishes *days_interacted* and *sum_click* as most important features

```
[('days_interacted', 9.5882458873469527e-119),
 ('sum_click', 2.2538413501212967e-64),
 ('date_registration', 3.4891180678744023e-24),
 ('count externalquiz', 2.5051693997621333e-08),
 ('highest_education', 1.4624102188195593e-05),
 ('count dualpane', 0.00039594060968423632),
 ('age_band', 0.00079411917697016013),
 ('imd_band', 0.021202005077540741),
 ('daily_click', 0.47680061894077441),
 ('count dataplus', 0.72231805431037888)]
```

Random Forest Results

These values depict correlation between *final_result* and each of these features. Higher the value, more profound is the impact of the feature on *final_result*.

Again, **days_interacted** and **sum_click come** out to be the most important features

```
[('days_interacted', 0.22166085169416039),
 ('sum_click', 0.16383936034879856),
 ('daily_click', 0.076950181595113326),
 ('date_registration', 0.073294938890613406),
 ('count forumng', 0.061326417390364908),
 ('count oucollaborate', 0.049702775783114161),
 ('count url', 0.047080880177469163),
 ('count subpage', 0.046651909011428697),
 ('count resource', 0.041994474337744878),
 ('count quiz', 0.041719530407678625),
 ('count oucontent', 0.032742978031556957),
 ('imd_band', 0.030899821599997157),
 ('highest_education', 0.017520745473669851),
 ('count sharedsubpage', 0.015727993111166395),
 ('count page', 0.014879944506678816),
 ('age_band', 0.013198726848108964),
 ('count ouwiki', 0.012708834380392763),
 ('count glossary', 0.010175828117641493),
 ('count dualpane', 0.0069707126533300554),
 ('count externalquiz', 0.0057450404971586362),
 ('count questionnaire', 0.004258112287083159),
 ('count ouelluminate', 0.003250984096252965),
 ('count htmlactivity', 0.0027621766949978989),
 ('count folder', 0.0021351275158227813),
 ('count dataplus', 0.0014578985198987419),
 ('count repeatactivity', 0.0013437560297572993)]
```

Key Takeaways

The 4 features on the right come out to be the most pertinent once. If the withdrawal/failure rate is to be decreased, then the University must pay emphasis on these features. *Date_registration*, although important, can be easily fixed by establishing most effective registration date.

Therefore, much of the effort must go into ensuring high output/input ratio for **days_interacted**, **sum_clicks**, and **daily_click**.*

*Targeting 3 features only

```
[('days_interacted', 3),
 ('sum_click', 3),
 ('date_registration', 3),
 ('daily_click', 3),
```

Final Recommendations

The main reason for students in a Virtual Environment can be identified as **LACK OF MOTIVATION**.

In order to positively drive this initiative, the university must take the following steps:

1) Award course credits only based on minimum *days_interacted*

2) Incentivize completing courses by offering lower merit awards for Distinction

3) Increase participation(*sum_clicks* and *daily_clicks*) by weighing final grade(~50%) on participation

4) Penalizing students for Failing/Withdrawing by blocking any scholarships/grants