Assessment 2: Data quality profiling and standardising

MA5831 – Advanced Data Processing and Analysis using SAS

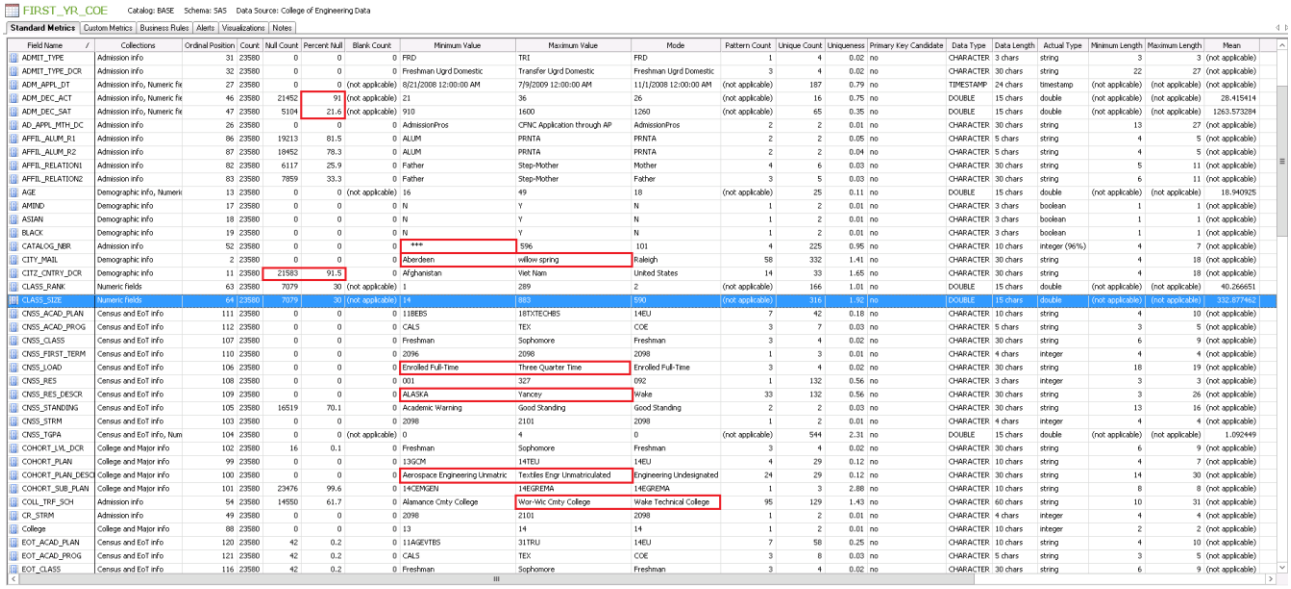
13848336 Nikki Fitzherbert

Question 1

There is no definitive definition of data quality; however, one definition indicates that high quality data is accurate, complete, consistent, current and each record refers to a unique entity. This definition, data profiling in DataFlux Data Management Studio and additional contextual information led the author to conclude that the available data for first year College of Engineering students was of average quality:

- The data was accurate insofar as being largely consistent with the information within the accompanying data dictionary.
- The data was incomplete as there were 70 fields in total with missing values (and thirteen fields with a percent null value of over 60 per cent).
- There were issues with data representation consistency for sixteen of the 129 fields.
- The data was relatively old as it related to first-year students who began in the fall semester of 2009.
- There were multiple records (rows of data) for each unique student id number.

The data profile extract shows some of the issues with the data.



FIRST_YR_COE Catalog: BASE Schema: SAS Data Source: College of Engineering Data

Standard Metrics | Custom Metrics | Business Rules | Alerts | Visualizations | Notes

Field Name	Collections	Ordinal Position	Count	Null Count	Percent Null	Blank Count	Minimum Value	Maximum Value	Mode	Pattern Count	Unique Count	Uniqueness	Primary Key Candidate	Data Type	Data Length	Actual Type	Minimum Length	Maximum Length	Mean
ADMIT_TYPE	Admission info	31	23580	0	0	0	FRD	TRI	FRD	1	4	0.02	no	CHARACTER	3 chars	string	3	3	(not applicable)
ADMIT_TYPE_DCR	Admission info	32	23580	0	0	0	Freshman Ugrd Domestic	Transfer Ugrd Domestic	Freshman Ugrd Domestic	3	4	0.02	no	CHARACTER	30 chars	string	22	27	(not applicable)
ADM_APPL_DT	Admission info, Numeric fo	27	23580	0	0	(not applicable)	8/21/2008 12:00:00 AM	7/9/2009 12:00:00 AM	11/1/2008 12:00:00 AM	(not applicable)	187	0.79	no	TIMESTAMP	24 chars	timestamp	(not applicable)	(not applicable)	(not applicable)
ADM_DEC_ACT	Admission info, Numeric fo	46	23580	21452	91	(not applicable)	21	36	26	(not applicable)	16	0.75	no	DOUBLE	15 chars	double	(not applicable)	(not applicable)	28.415414
ADM_DEC_SAT	Admission info, Numeric fo	47	23580	5104	21.6	(not applicable)	910	1600	1260	(not applicable)	65	0.35	no	DOUBLE	15 chars	double	(not applicable)	(not applicable)	1263.573284
AD_APPL_MTH_DC	Admission info	26	23580	0	0	0	AdmissionPros	CFNC Application through AP	AdmissionPros	2	2	0.01	no	CHARACTER	30 chars	string	13	27	(not applicable)
AFFIL_ALUM_R1	Admission info	86	23580	19213	81.5	0	ALUM	PRNTA	PRNTA	2	2	0.05	no	CHARACTER	5 chars	string	4	5	(not applicable)
AFFIL_ALUM_R2	Admission info	87	23580	18452	78.3	0	ALUM	PRNTA	PRNTA	2	2	0.04	no	CHARACTER	5 chars	string	4	5	(not applicable)
AFFIL_RELATION1	Admission info	82	23580	6117	25.9	0	Father	Step-Mother	Mother	4	6	0.03	no	CHARACTER	30 chars	string	5	11	(not applicable)
AFFIL_RELATION2	Admission info	83	23580	7859	33.3	0	Father	Step-Mother	Father	3	5	0.03	no	CHARACTER	30 chars	string	6	11	(not applicable)
AGE	Demographic info, Numeri	13	23580	0	0	(not applicable)	16	49	18	(not applicable)	25	0.11	no	DOUBLE	15 chars	double	(not applicable)	(not applicable)	18.940925
AMIND	Demographic info	17	23580	0	0	0	N	Y	N	1	2	0.01	no	CHARACTER	3 chars	boolean	1	1	(not applicable)
ASIAN	Demographic info	18	23580	0	0	0	N	Y	N	1	2	0.01	no	CHARACTER	3 chars	boolean	1	1	(not applicable)
BLACK	Demographic info	19	23580	0	0	0	N	Y	N	1	2	0.01	no	CHARACTER	3 chars	boolean	1	1	(not applicable)
CATALOG_NBR	Admission info	52	23580	0	0	0	***	596	101	4	225	0.95	no	CHARACTER	10 chars	integer (96%)	4	7	(not applicable)
CITY_MAIL	Demographic info	2	23580	0	0	0	Aberdeen	willow spring	Raleigh	58	332	1.41	no	CHARACTER	30 chars	string	4	18	(not applicable)
CITZ_CNTRY_DCR	Demographic info	11	23580	21583	91.5	0	Afghanistan	Viet Nam	United States	14	33	1.65	no	CHARACTER	30 chars	string	4	18	(not applicable)
CLASS_RANK	Numeric fields	63	23580	7079	30	(not applicable)	1	289	2	(not applicable)	166	1.01	no	DOUBLE	15 chars	double	(not applicable)	(not applicable)	40.266651
CLASS_SIZE	Numeric fields	64	23580	7079	30	(not applicable)	14	887	590	(not applicable)	316	1.32	no	DOUBLE	15 chars	double	(not applicable)	(not applicable)	332.877462
CNSS_ACAD_PLAN	Census and EoT info	111	23580	0	0	0	11BEBS	18TXTECHBS	14EU	7	42	0.18	no	CHARACTER	10 chars	string	4	10	(not applicable)
CNSS_ACAD_PROG	Census and EoT info	112	23580	0	0	0	CALS	TEX	COE	3	7	0.03	no	CHARACTER	5 chars	string	3	5	(not applicable)
CNSS_CLASS	Census and EoT info	107	23580	0	0	0	Freshman	Sophomore	Freshman	3	4	0.02	no	CHARACTER	30 chars	string	6	9	(not applicable)
CNSS_FIRST_TERM	Census and EoT info	110	23580	0	0	0	2096	2098	2098	1	3	0.01	no	CHARACTER	4 chars	integer	4	4	(not applicable)
CNSS_LOAD	Census and EoT info	106	23580	0	0	0	Enrolled Full-Time	Three Quarter Time	Enrolled Full-Time	3	4	0.02	no	CHARACTER	30 chars	string	18	19	(not applicable)
CNSS_RES	Census and EoT info	108	23580	0	0	0	001	327	092	1	132	0.56	no	CHARACTER	3 chars	integer	3	3	(not applicable)
CNSS_RES_DESCR	Census and EoT info	109	23580	0	0	0	ALASKA	Yancey	Wake	33	132	0.56	no	CHARACTER	30 chars	string	3	26	(not applicable)
CNSS_STANDING	Census and EoT info	105	23580	16519	70.1	0	Academic Warning	Good Standing	Good Standing	2	2	0.03	no	CHARACTER	30 chars	string	13	16	(not applicable)
CNSS_STRM	Census and EoT info	103	23580	0	0	0	2098	2101	2098	1	2	0.01	no	CHARACTER	4 chars	integer	4	4	(not applicable)
CNSS_TGPA	Census and EoT info, Num	104	23580	0	0	(not applicable)	0	4	0	(not applicable)	544	2.31	no	DOUBLE	15 chars	double	(not applicable)	(not applicable)	1.092449
COHORT_LVL_DCR	College and Major info	102	23580	16	0.1	0	Freshman	Sophomore	Freshman	3	4	0.02	no	CHARACTER	30 chars	string	6	9	(not applicable)
COHORT_PLAN	College and Major info	99	23580	0	0	0	13GCM	14TEU	14EU	4	29	0.12	no	CHARACTER	10 chars	string	4	7	(not applicable)
COHORT_PLAN_DESC	College and Major info	100	23580	0	0	0	Aerospace Engineering Unmatric	Textiles Engr Unmatriculated	Engineering Undesignated	24	29	0.12	no	CHARACTER	30 chars	string	14	30	(not applicable)
COHORT_SUB_PLAN	College and Major info	101	23580	23476	99.6	0	14CEMGEN	14EGREMA	14EGREMA	1	3	2.88	no	CHARACTER	10 chars	string	8	8	(not applicable)
COLL_TRF_SCH	Admission info	54	23580	14550	61.7	0	Alamance Cmty College	Wor-Wic Cmty College	Wake Technical College	95	129	1.43	no	CHARACTER	60 chars	string	10	31	(not applicable)
CR_STRM	Admission info	49	23580	0	0	0	2098	2101	2098	1	2	0.01	no	CHARACTER	4 chars	integer	4	4	(not applicable)
College	College and Major info	88	23580	0	0	0	13	14	14	1	2	0.01	no	CHARACTER	10 chars	integer	2	2	(not applicable)
EOT_ACAD_PLAN	Census and EoT info	120	23580	42	0.2	0	11AGEVTBS	31TRU	14EU	7	58	0.25	no	CHARACTER	10 chars	string	4	10	(not applicable)
EOT_ACAD_PROG	Census and EoT info	121	23580	42	0.2	0	CALS	TEX	COE	3	8	0.03	no	CHARACTER	5 chars	string	3	5	(not applicable)
EOT_CLASS	Census and EoT info	116	23580	42	0.2	0	Freshman	Sophomore	Freshman	3	4	0.02	no	CHARACTER	30 chars	string	6	9	(not applicable)

Question 2

Data profiling indicated that there were a number of fields that required cleansing before any sort of analysis or reporting could be performed. The table lists these fields along with their identified data quality issue(s):

Field Name(s)	Field Issue(s)	Example(s)
ADMIT_TYPE_DCR CITY_MAIL	Inconsistent capitalisation	Ugrd vs UGRD Aberdeen vs willow spring
EOT_LOAD CNSS_LOAD	Inconsistent hyphenation	Full-Time vs Quarter Time
GRADE	Categories with minor differences	A vs A- vs A+
NEW_ID	Multiple records per unique id number	-
CATALOG_NBR STATE_MAIL	Unexpected values	*** 51
STATE	Inconsistent capitalisation Unexpected values	North Carolina vs OHIO (states) vs BRAZIL (country)
CNSS_RES_DESCR EOT_RES_DESCR TUI_RES_CD_DCR	Inconsistent capitalisation Intermixed geographic information	Scotland (country) vs ALASKA (state) vs Yancey (county)
COHORT_PLAN_DESCR PLAN_FIRST_DCR PLAN_SECOND_DCR COLL_TRF_SCH	Inconsistent use of word abbreviations Categories with minor differences	Engr vs Engineering Cmty vs Technical
LST_HIGH_SCHOOL SUB_SECOND_DCR	Inconsistent use of word abbreviations Inconsistent capitalisation	High School vs hs Hs vs hs

Question 3

Outliers are data points that differ significantly from other observations and univariate outliers are usually defined as being some distance away from a field's mean or median. As such, the concept of outliers only applies to the numeric fields in the dataset.

Rather than using the outlier tab, which only showed a pre-specified number of minimum and maximum values for each field (the default is five), the author defined an outlier as being any value more than three standard deviations away from the mean and used visual inspection of the minimum and maximum metrics to determine the presence of possible outliers.

The table indicates the outliers identified using this approach:

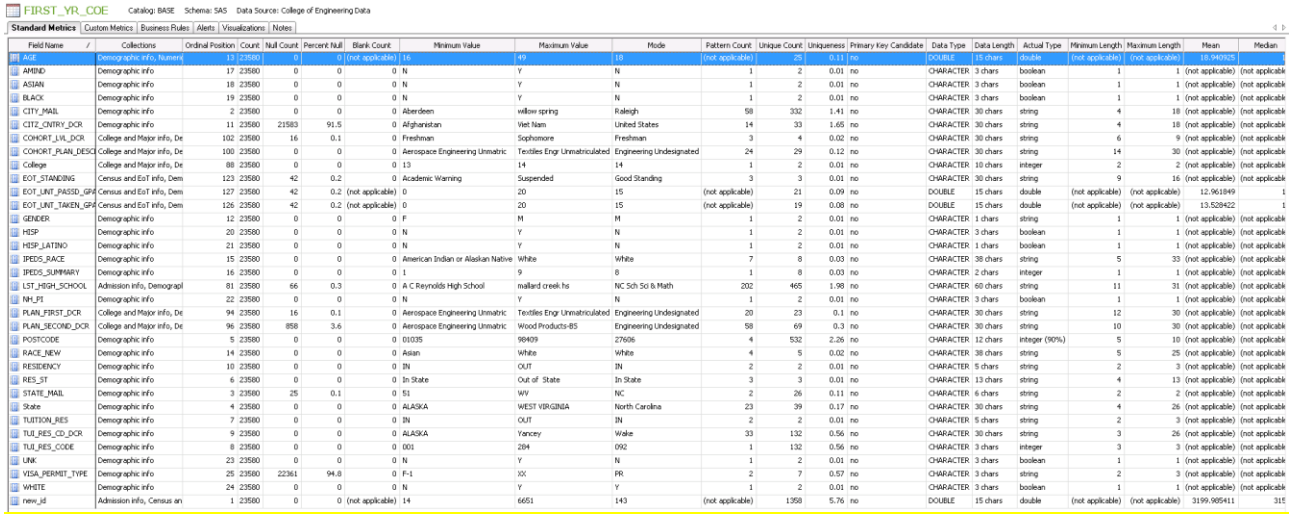
Field	Lower bound	Upper bound
ACT_ENGL	35	36
ACT_WRITING	32	35
AGE	29	49
CLASS_RANK	176	289
CLASS_SIZE	757	883
EOT_TOT_PASSD_PRGRSS	133	247
EOT_TOT_TAKEN_PRGRSS	134	247
EOT_UNT_PASSD_GPA	0	3
EOT_UNT_PASSD_PRGRSS	0	4
EOT_UNT_TAKEN_GPA	0	5
EOT_UNT_TAKEN_PRGRSS	1	8 (also 21)
HS_GPA	2.83	2.99
PERCENTILE	37	95
SAT_VERBAL	320	357
SAT_WRITING	220	341
TOTAL_SAT	860	871
TRF_SUMM_ATTEMP	118	194
TRF_SUMM_COMP	113	168

Question 4

For this case study, 33 fields (plus the student id field) were classified as demographic data. It was concluded that seven fields contained unexpected values, of which some have already been mentioned in the response to question two:

- Student ages ranging from as young as 16 to as old as 49.
- The minimum value for EOT_UNT_PASSD_GPA being 0, which indicated that at least one student failed to pass a single subject that semester.
- The pattern count for the STATE_MAIL and POSTCODE fields being more than one (the presence of a two-digit number pattern for the former and non-five-digit number patterns for the latter).
- Multiple records with the same student id number.
- A large number of missing values for CITZ_CNTRY_DCR (91.5 per cent).
- Categories for EOT_STANDING inconsistent with what was suggested by the data dictionary.

Below is an extract of the data profile for just the demographic fields.



FIRST_YR_COE Catalog: BASE Schema: SAS Data Source: College of Engineering Data

Standard Metrics | Custom Metrics | Business Rules | Alerts | Visualizations | Notes

Field Name	Collections	Ordinal Position	Count	Null Count	Percent Null	Blank Count	Minimum Value	Maximum Value	Mode	Pattern Count	Unique Count	Uniqueness	Primary Key Candidate	Data Type	Data Length	Actual Type	Minimum Length	Maximum Length	Mean	Median
AGE	Demographic info, Numeri	13	23580	0	0	(not applicable)	16	49	18	(not applicable)	25	0.11	no	DOUBLE	15 chars	double	(not applicable)	(not applicable)	18.940925	
AMIND	Demographic info	17	23580	0	0	0	N	Y	N	1	2	0.01	no	CHARACTER	3 chars	boolean	1	1	(not applicable)	(not applicable
ASIAN	Demographic info	18	23580	0	0	0	N	Y	N	1	2	0.01	no	CHARACTER	3 chars	boolean	1	1	(not applicable)	(not applicable
BLACK	Demographic info	19	23580	0	0	0	N	Y	N	1	2	0.01	no	CHARACTER	3 chars	boolean	1	1	(not applicable)	(not applicable
CITY_MAIL	Demographic info	2	23580	0	0	0	Aberdeen	willow spring	Raleigh	58	332	1.41	no	CHARACTER	30 chars	string	4	18	(not applicable)	(not applicable
CITZ_CNTRY_DCR	Demographic info	11	23580	21583	91.5	0	Afghanistan	Viet Nam	United States	14	33	1.65	no	CHARACTER	30 chars	string	4	18	(not applicable)	(not applicable
COHORT_LVL_DCR	College and Major info, De	102	23580	16	0.1	0	Freshman	Sophomore	Freshman	3	4	0.02	no	CHARACTER	30 chars	string	6	9	(not applicable)	(not applicable
COHORT_PLAN_DESC	College and Major info, De	100	23580	0	0	0	Aerospace Engineering Unmatric	Textiles Engr Unmatriculated	Engineering Undesignated	24	29	0.12	no	CHARACTER	30 chars	string	14	30	(not applicable)	(not applicable
College	College and Major info, De	88	23580	0	0	0	13	14	14	1	2	0.01	no	CHARACTER	10 chars	integer	2	2	(not applicable)	(not applicable
EOT_STANDING	Census and EoT info, Dem	123	23580	42	0.2	0	Academic Warning	Suspended	Good Standing	3	3	0.01	no	CHARACTER	30 chars	string	9	16	(not applicable)	(not applicable
EOT_UNT_PASSD_GPA	Census and EoT info, Dem	127	23580	42	0.2	(not applicable)	0	20	15	(not applicable)	21	0.09	no	DOUBLE	15 chars	double	(not applicable)	(not applicable)	12.961849	1
EOT_UNT_TAKEN_GPA	Census and EoT info, Dem	126	23580	42	0.2	(not applicable)	0	20	15	(not applicable)	19	0.08	no	DOUBLE	15 chars	double	(not applicable)	(not applicable)	13.528422	1
GENDER	Demographic info	12	23580	0	0	0	F	M	M	1	2	0.01	no	CHARACTER	1 chars	string	1	1	(not applicable)	(not applicable
HISP	Demographic info	20	23580	0	0	0	N	Y	N	1	2	0.01	no	CHARACTER	3 chars	boolean	1	1	(not applicable)	(not applicable
HISP_LATINO	Demographic info	21	23580	0	0	0	N	Y	N	1	2	0.01	no	CHARACTER	1 chars	boolean	1	1	(not applicable)	(not applicable
IPEDS_RACE	Demographic info	15	23580	0	0	0	American Indian or Alaskan Native	White	White	7	8	0.03	no	CHARACTER	38 chars	string	5	33	(not applicable)	(not applicable
IPEDS_SUMMARY	Demographic info	16	23580	0	0	0	1	9	8	1	8	0.03	no	CHARACTER	2 chars	integer	1	1	(not applicable)	(not applicable
LST_HIGH_SCHOOL	Admission info, Demograpl	81	23580	66	0.3	0	A C Reynolds High School	mallard creek hs	NC Sch Sci & Math	202	465	1.98	no	CHARACTER	60 chars	string	11	31	(not applicable)	(not applicable
NH_PI	Demographic info	22	23580	0	0	0	N	Y	N	1	2	0.01	no	CHARACTER	3 chars	boolean	1	1	(not applicable)	(not applicable
PLAN_FIRST_DCR	College and Major info, De	94	23580	16	0.1	0	Aerospace Engineering Unmatric	Textiles Engr Unmatriculated	Engineering Undesignated	20	23	0.1	no	CHARACTER	30 chars	string	12	30	(not applicable)	(not applicable
PLAN_SECOND_DCR	College and Major info, De	96	23580	858	3.6	0	Aerospace Engineering Unmatric	Wood Products-BS	Engineering Undesignated	58	69	0.3	no	CHARACTER	30 chars	string	10	30	(not applicable)	(not applicable
POSTCODE	Demographic info	5	23580	0	0	0	01035	98409	27606	4	532	2.26	no	CHARACTER	12 chars	integer (90%)	5	10	(not applicable)	(not applicable
RACE_NEW	Demographic info	14	23580	0	0	0	Asian	White	White	4	5	0.02	no	CHARACTER	38 chars	string	5	25	(not applicable)	(not applicable
RESIDENCY	Demographic info	10	23580	0	0	0	IN	OUT	IN	2	2	0.01	no	CHARACTER	5 chars	string	2	3	(not applicable)	(not applicable
RES_ST	Demographic info	6	23580	0	0	0	In State	Out of State	In State	3	3	0.01	no	CHARACTER	13 chars	string	4	13	(not applicable)	(not applicable
STATE_MAIL	Demographic info	3	23580	25	0.1	0	51	WV	NC	2	26	0.11	no	CHARACTER	6 chars	string	2	2	(not applicable)	(not applicable
State	Demographic info	4	23580	0	0	0	ALASKA	WEST VIRGINIA	North Carolina	23	39	0.17	no	CHARACTER	30 chars	string	4	26	(not applicable)	(not applicable
TUITION_RES	Demographic info	7	23580	0	0	0	IN	OUT	IN	2	2	0.01	no	CHARACTER	5 chars	string	2	3	(not applicable)	(not applicable
TUI_RES_CD_DCR	Demographic info	9	23580	0	0	0	ALASKA	Yancey	Wake	33	132	0.56	no	CHARACTER	30 chars	string	3	26	(not applicable)	(not applicable
TUI_RES_CODE	Demographic info	8	23580	0	0	0	001	284	092	1	132	0.56	no	CHARACTER	3 chars	integer	3	3	(not applicable)	(not applicable
UNK	Demographic info	23	23580	0	0	0	N	Y	N	1	2	0.01	no	CHARACTER	3 chars	boolean	1	1	(not applicable)	(not applicable
VISA_PERMIT_TYPE	Demographic info	25	23580	22361	94.8	0	F-1	XX	PR	2	7	0.57	no	CHARACTER	3 chars	string	2	3	(not applicable)	(not applicable
WHITE	Demographic info	24	23580	0	0	0	N	Y	Y	1	2	0.01	no	CHARACTER	3 chars	boolean	1	1	(not applicable)	(not applicable
new_id	Admission info, Census an	1	23580	0	0	(not applicable)	14	6651	143	(not applicable)	1358	5.76	no	DOUBLE	15 chars	double	(not applicable)	(not applicable)	3199.985411	315

Question 5

Given that data quality issues had been identified for approximately 21 per cent of fields classified as demographic data (or 27 per cent if the eight Boolean race fields are considered a single field), it was concluded that the fields containing demographic data were not suitable for analysis and reporting. This was because the identified data quality issues would likely obscure any trends in the data and lead to misleading conclusions about the demographics of first-year students.

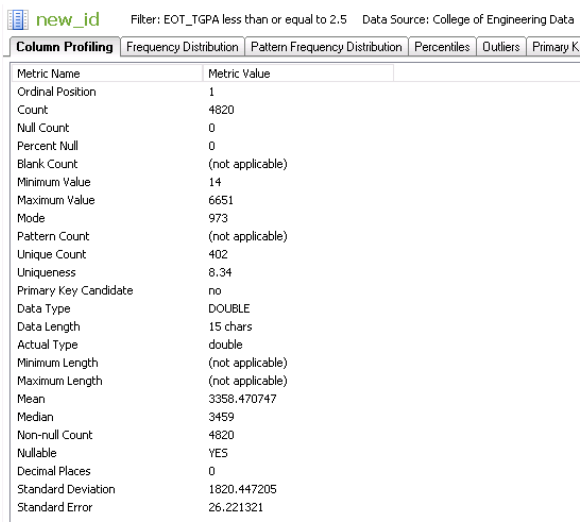
In order to increase the suitability of the data for the analysis and reporting of at-risk first-year students, the dataset needed to undergo a data-cleaning process. This involved making a number of changes to the data, including:

- Separating or otherwise identifying the different types of geographic information in fields such as STATE
- Correcting for inconsistent capitalisation and word abbreviations
- Consolidation of categories in fields such as PLAN_SECOND_DCR
- Data correction such as for POSTCODE
- Record consolidation so each student id appeared only once, and
- Data enrichment (if possible) to reduce the number of missing values.

Question 6

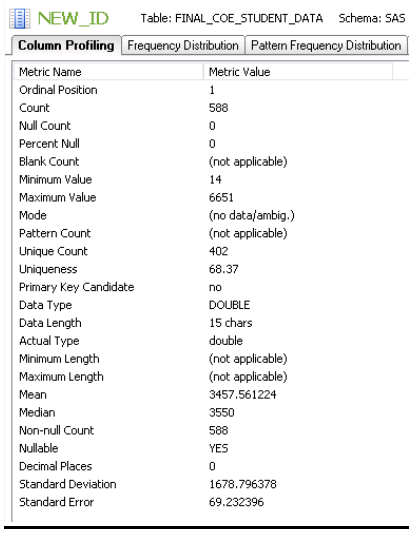
At-risk students were defined as first-year students with an end-of-term grade point average (EOT_GPA) value of 2.5 or less. Initial data profiling suggested that there were 4,820 at-risk students; however, a closer examination indicated that only 402 records had unique student ids. A profile of the data after it had undergone data cleansing and entity consolidation reduced the total number of records down to 588 but still showed a unique count of 402 (as per the images below). Therefore, the author concluded that the true number of at-risk students in the data was somewhere between 402 and 588.

Data profile reports prior to (LHS) and after (RHS) entity consolidation


new_id Filter: EOT_TGPA less than or equal to 2.5 Data Source: College of Engineering Data

Column Profiling | Frequency Distribution | Pattern Frequency Distribution | Percentiles | Outliers | Primary K

Metric Name	Metric Value
Ordinal Position	1
Count	4820
Null Count	0
Percent Null	0
Blank Count	(not applicable)
Minimum Value	14
Maximum Value	6651
Mode	973
Pattern Count	(not applicable)
Unique Count	402
Uniqueness	8.34
Primary Key Candidate	no
Data Type	DOUBLE
Data Length	15 chars
Actual Type	double
Minimum Length	(not applicable)
Maximum Length	(not applicable)
Mean	3358.470747
Median	3459
Non-null Count	4820
Nullable	YES
Decimal Places	0
Standard Deviation	1820.447205
Standard Error	26.221321


NEW_ID Table: FINAL_COE_STUDENT_DATA Schema: SAS

Column Profiling | Frequency Distribution | Pattern Frequency Distribution

Metric Name	Metric Value
Ordinal Position	1
Count	588
Null Count	0
Percent Null	0
Blank Count	(not applicable)
Minimum Value	14
Maximum Value	6651
Mode	(no data/ambig.)
Pattern Count	(not applicable)
Unique Count	402
Uniqueness	68.37
Primary Key Candidate	no
Data Type	DOUBLE
Data Length	15 chars
Actual Type	double
Minimum Length	(not applicable)
Maximum Length	(not applicable)
Mean	3457.561224
Median	3550
Non-null Count	588
Nullable	YES
Decimal Places	0
Standard Deviation	1678.796378
Standard Error	69.232396

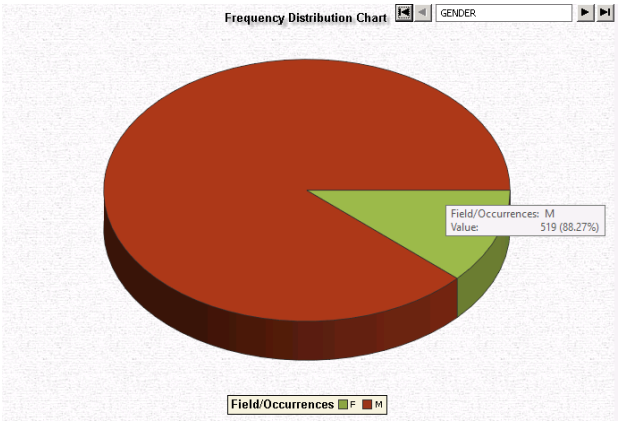
Question 7

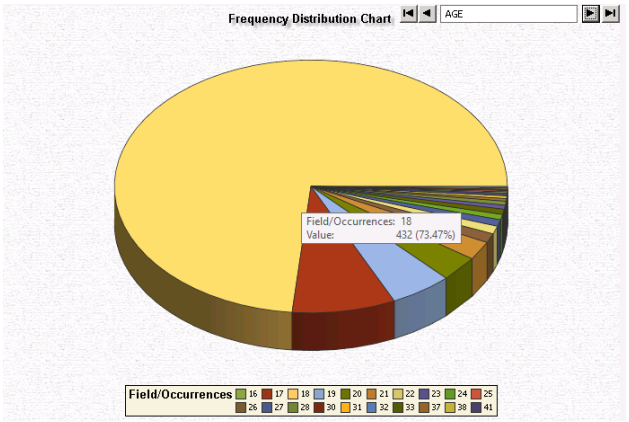
Analysis of the data indicated the presence of several notable demographic patterns for at-risk students:

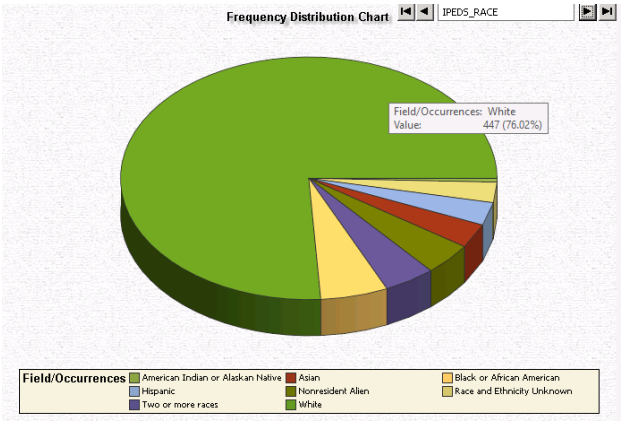
- Most were male (88.3 per cent).
- Close to 90 per cent were aged between 17 and 19 years of age, with most (73.5 per cent) being 18 years old.
- About three-quarters identified as white.
- Just less than half (44.4 per cent) were citizens of the United States.
- Most (over 90 per cent) were local to North Carolina.
- Approximately two-thirds resided in Wake, Guilford or Mecklenburg counties, or the state of New Jersey, Virginia, Pennsylvania or Maryland.
- Approximately two-thirds had listed the cities of Raleigh, Greensboro or Charlotte on their application.
- Most were freshmen (85.0 per cent).
- Just over half (56.6 per cent) were in good academic standing at the end of term.

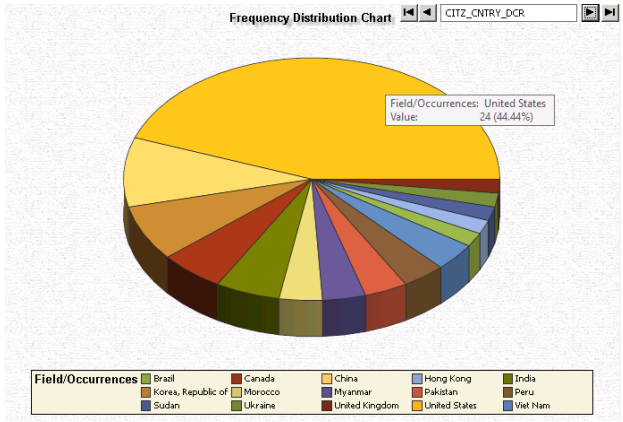
However, it must be noted that it was not investigated as to whether these demographic patterns were also shared with the overall first-year College of Engineering student cohort.

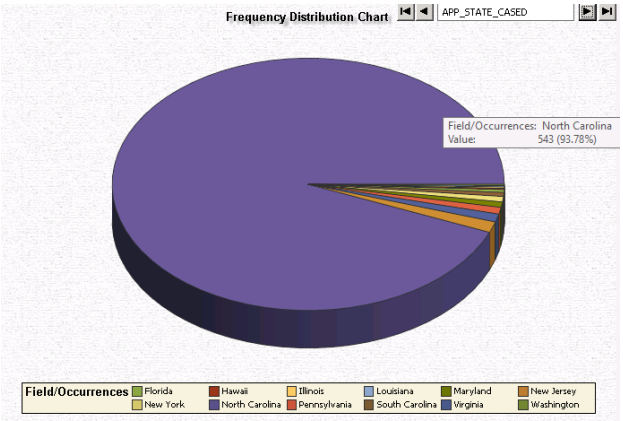
The frequency distribution charts for these insights are below:

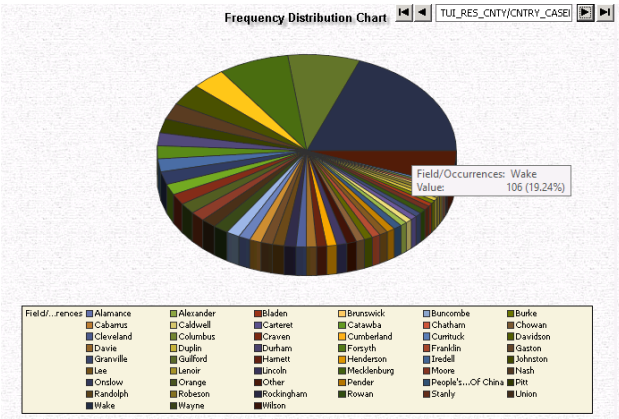


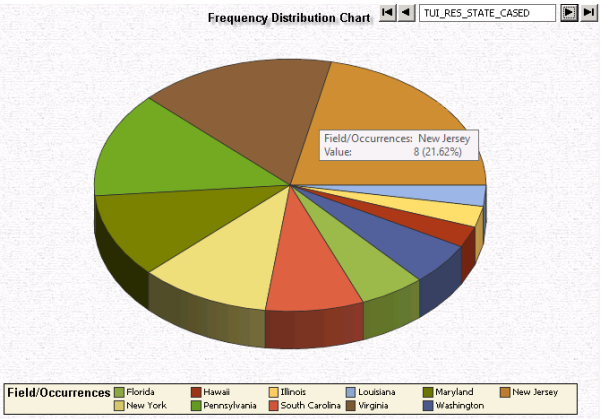


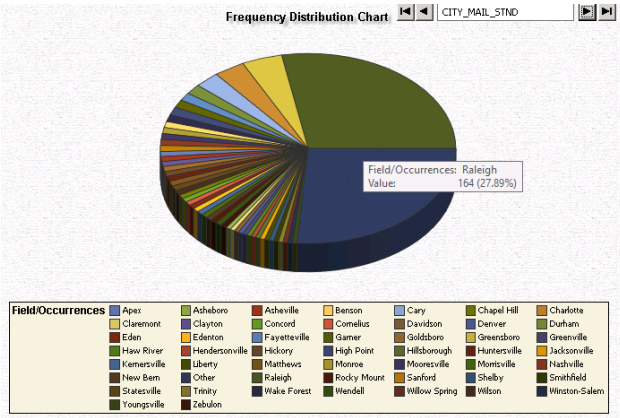


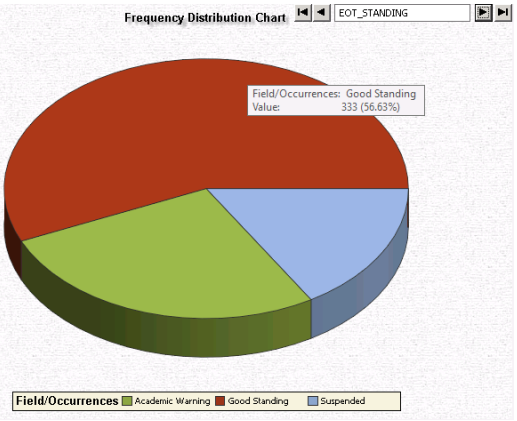


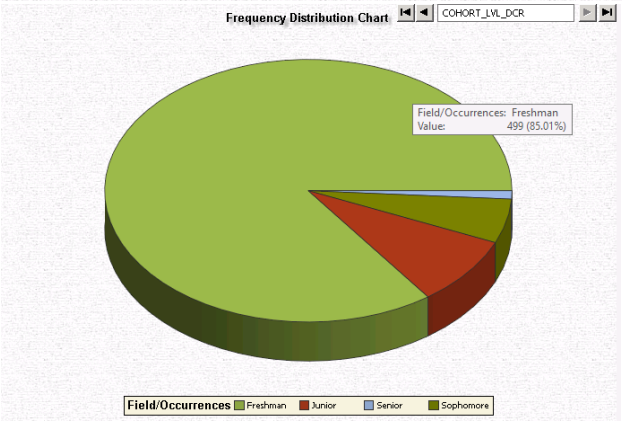

Frequency Distribution Chart
APP_STATE_CASED
Field/Occurrences: North Carolina
Value: 543 (93.78%)
Field/Occurrences
Florida
Hawaii
Illinois
Louisiana
Maryland
New Jersey
New York
North Carolina
Pennsylvania
South Carolina
Virginia
Washington


Frequency Distribution Chart
TUI_RES_CNTY/CNTRY_CASE
Field/Occurrences: Wake
Value: 106 (19.24%)
Field/...rences
Alamance
Alexander
Bladen
Brunswick
Buncombe
Burke
Cabarrus
Caldwell
Carteret
Catawba
Chatham
Chowan
Cleveland
Columbus
Craven
Cumberland
Currituck
Davidson
Davie
Duplin
Durham
Forsyth
Franklin
Gaston
Granville
Guilford
Harnett
Henderson
Iredell
Johnston
Lee
Lenoir
Lincoln
Mecklenburg
Moore
Nash
Onslow
Orange
Other
Pender
People's...Of China
Pitt
Randolph
Robeson
Rockingham
Rowan
Stanly
Union
Wake
Wayne
Wilson


Frequency Distribution Chart
TUI_RES_STATE_CASED
Field/Occurrences: New Jersey
Value: 8 (21.62%)
Field/Occurrences
Florida
Hawaii
Illinois
Louisiana
Maryland
New Jersey
New York
Pennsylvania
South Carolina
Virginia
Washington


Frequency Distribution Chart
CITY_MAIL_STND
Field/Occurrences: Raleigh
Value: 164 (27.89%)
Field/Occurrences
Apex
Asheboro
Asheville
Benson
Cary
Chapel Hill
Charlotte
Claremont
Clayton
Concord
Cornelius
Davidson
Denver
Durham
Eden
Edenton
Fayetteville
Garner
Goldsboro
Greensboro
Greenville
Haw River
Hendersonville
Hickory
High Point
Hillsborough
Huntersville
Jacksonville
Kernersville
Liberty
Matthews
Monroe
Mooresville
Morrisville
Nashville
New Bern
Other
Raleigh
Rocky Mount
Sanford
Shelby
Smithfield
Statesville
Trinity
Wake Forest
Wendell
Willow Spring
Wilson
Winston-Salem
Youngsville
Zebulon


Frequency Distribution Chart
EOT_STANDING
Field/Occurrences: Good Standing
Value: 333 (56.63%)
Field/Occurrences
Academic Warning
Good Standing
Suspended


Frequency Distribution Chart
COHORT_LVL_DCR
Field/Occurrences: Freshman
Value: 499 (85.01%)
Field/Occurrences
Freshman
Junior
Senior
Sophomore

Question 8

In order to improve the quality of data to be used in reports, it needs to undergo data-cleansing (including consolidation) and enrichment processes. This involves one or more sub-processes depending on the particular issues with a dataset and includes:

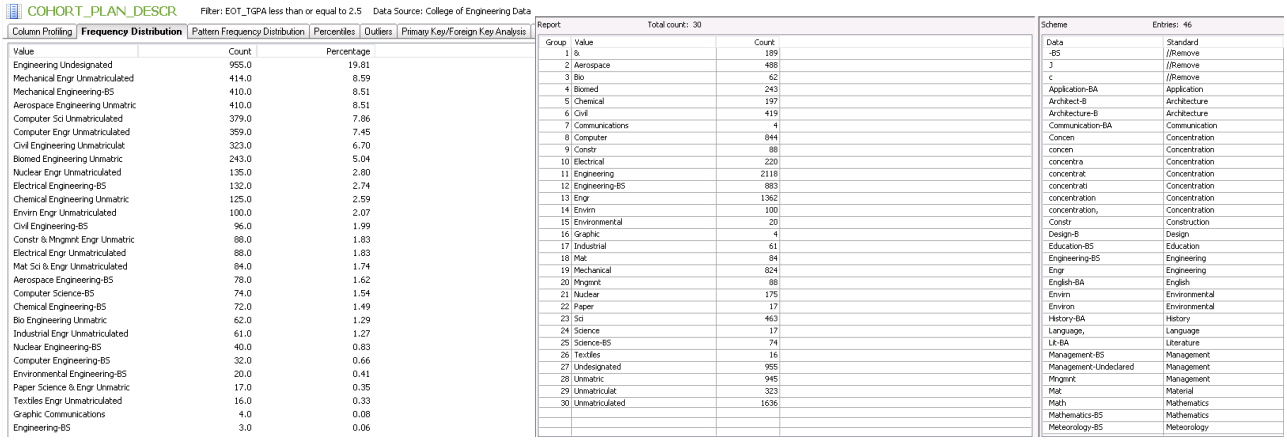
- Parsing to separate text strings into component tokens,
- Standardising data to present it in a consistent manner and/or preferred style,
- Correcting data values using other data within the dataset or external information,
- Re-ordering and/or selecting only relevant fields,
- Identifying the type of data within a field,
- Dealing appropriately with missing data, and
- Matching and consolidating data records so the same physical object is represented once in the dataset.

Question 9 – Standardisation

Data standardisation is a broad term describing any process generating a standard representation of data values through the use of a definition or scheme. The columns of demographic data that required data standardisation were:

- COHORT_PLAN_DESCR
- LST_HIGH_SCHOOL
- PLAN_FIRST_DCR
- PLAN_SECOND_DCR
- POSTCODE
- STATE_MAIL

The following is an example of a field that required standardisation and an extract of the schema used:



COHORT_PLAN_DESCR Filter: EOT_TGPA less than or equal to 2.5 Data Source: College of Engineering Data

Column Profiling | **Frequency Distribution** | Pattern Frequency Distribution | Percentiles | Outliers | Primary Key/Foreign Key Analysis

Value	Count	Percentage
Engineering Undesignated	955.0	19.81
Mechanical Engr Unmatriculated	414.0	8.59
Mechanical Engineering-BS	410.0	8.51
Aerospace Engineering Unmatric	410.0	8.51
Computer Sci Unmatriculated	379.0	7.86
Computer Engr Unmatriculated	359.0	7.45
Civil Engineering Unmatriculat	323.0	6.70
Biomed Engineering Unmatric	243.0	5.04
Nuclear Engr Unmatriculated	135.0	2.80
Electrical Engineering-BS	132.0	2.74
Chemical Engineering Unmatric	125.0	2.59
Environ Engr Unmatriculated	100.0	2.07
Civil Engineering-BS	96.0	1.99
Constr & Mngmnt Engr Unmatric	88.0	1.83
Electrical Engr Unmatriculated	88.0	1.83
Mat Sci & Engr Unmatriculated	84.0	1.74
Aerospace Engineering-BS	78.0	1.62
Computer Science-BS	74.0	1.54
Chemical Engineering-BS	72.0	1.49
Bio Engineering Unmatric	62.0	1.29
Industrial Engr Unmatriculated	61.0	1.27
Nuclear Engineering-BS	40.0	0.83
Computer Engineering-BS	32.0	0.66
Environmental Engineering-BS	20.0	0.41
Paper Science & Engr Unmatric	17.0	0.35
Textiles Engr Unmatriculated	16.0	0.33
Graphic Communications	4.0	0.08
Engineering-BS	3.0	0.06

Report Total count: 30

Group	Value	Count
1	&	189
2	Aerospace	488
3	Bio	62
4	Biomed	243
5	Chemical	197
6	Civil	419
7	Communications	4
8	Computer	844
9	Constr	88
10	Electrical	220
11	Engineering	2118
12	Engineering-BS	883
13	Engr	1362
14	Environ	100
15	Environmental	20
16	Graphic	4
17	Industrial	61
18	Mat	84
19	Mechanical	824
20	Mngmnt	88
21	Nuclear	175
22	Paper	17
23	Sci	463
24	Science	17
25	Science-BS	74
26	Textiles	16
27	Undesignated	955
28	Unmatric	945
29	Unmatriculat	323
30	Unmatriculated	1636

Scheme Entries: 46

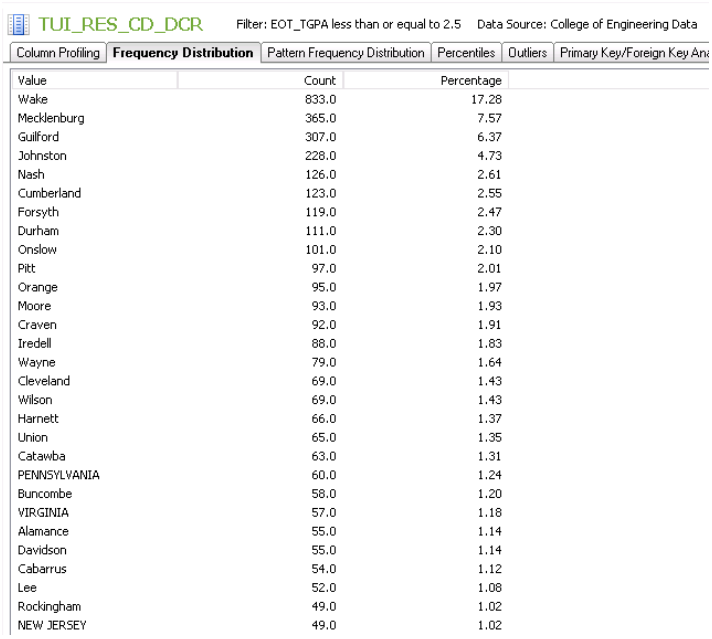
Data	Standard
-BS	//Remove
J	//Remove
c	//Remove
Application-BA	Application
Architect-B	Architecture
Architecture-B	Architecture
Communication-BA	Communication
Concen	Concentration
concen	Concentration
concentra	Concentration
concentrat	Concentration
concentrati	Concentration
concentration	Concentration
concentration,	Concentration
Constr	Construction
Design-B	Design
Education-BS	Education
Engineering-BS	Engineering
Engr	Engineering
English-BA	English
Environ	Environmental
Environ	Environmental
History-BA	History
Language,	Language
Lit-BA	Literature
Management-BS	Management
Management-Undeclared	Management
Mngmnt	Management
Mat	Material
Math	Mathematics
Mathematics-BS	Mathematics
Meteorology-BS	Meteorology

Question 10 – Casing

Casing involves the application of context-sensitive case rules to text strings. The columns of demographic data that required a case change for consistency were:

- CITY_MAIL
- LST_HIGH_SCHOOL
- STATE
- TUI_RES_CD_DCR

The following is an example of a field that required casing:



TUI_RES_CD_DCR Filter: EOT_TGPA less than or equal to 2.5 Data Source: College of Engineering Data

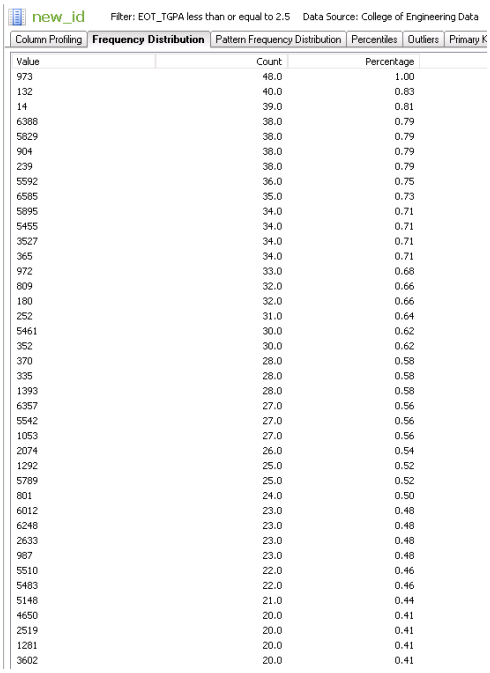
Column Profiling | **Frequency Distribution** | Pattern Frequency Distribution | Percentiles | Outliers | Primary Key/Foreign Key Ana

Value	Count	Percentage
Wake	833.0	17.28
Mecklenburg	365.0	7.57
Guilford	307.0	6.37
Johnston	228.0	4.73
Nash	126.0	2.61
Cumberland	123.0	2.55
Forsyth	119.0	2.47
Durham	111.0	2.30
Onslow	101.0	2.10
Pitt	97.0	2.01
Orange	95.0	1.97
Moore	93.0	1.93
Craven	92.0	1.91
Iredell	88.0	1.83
Wayne	79.0	1.64
Cleveland	69.0	1.43
Wilson	69.0	1.43
Harnett	66.0	1.37
Union	65.0	1.35
Catawba	63.0	1.31
PENNSYLVANIA	60.0	1.24
Buncombe	58.0	1.20
VIRGINIA	57.0	1.18
Alamance	55.0	1.14
Davidson	55.0	1.14
Cabarrus	54.0	1.12
Lee	52.0	1.08
Rockingham	49.0	1.02
NEW JERSEY	49.0	1.02

Question 11 – Entity resolution

Entity resolution involves using user-specified match rules to identify and merge records that belong to the same physical object. For the College of Engineering data, records were clustered on NEW_ID and then the record with the highest value for EOT_UNT_TAKEN_GPA and ADM_APPL_DATE was chosen at the surviving record for the cluster.

The image illustrates why entity resolution was required:



new_id Filter: EOT_TGPA less than or equal to 2.5 Data Source: College of Engineering Data

Column Profiling | **Frequency Distribution** | Pattern Frequency Distribution | Percentiles | Outliers | Primary K

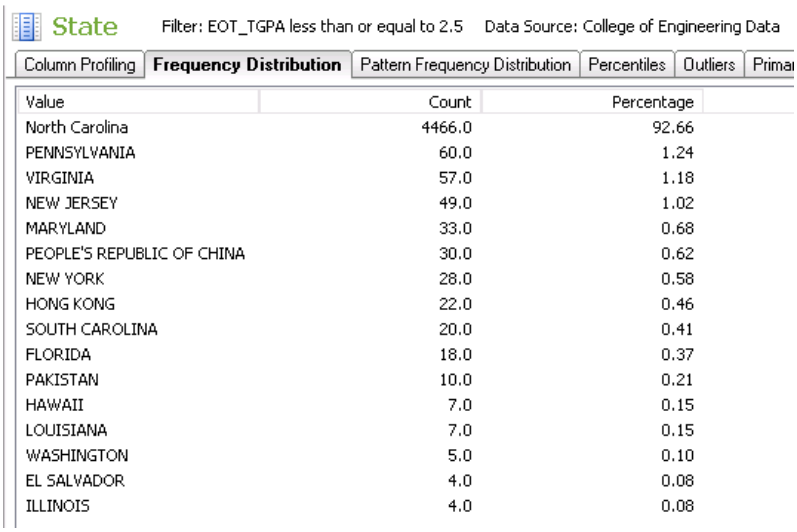
Value	Count	Percentage
973	48.0	1.00
132	40.0	0.83
14	39.0	0.81
6388	38.0	0.79
5829	38.0	0.79
904	38.0	0.79
239	38.0	0.79
5592	36.0	0.75
6585	35.0	0.73
5895	34.0	0.71
5455	34.0	0.71
3527	34.0	0.71
365	34.0	0.71
972	33.0	0.68
809	32.0	0.66
180	32.0	0.66
252	31.0	0.64
5461	30.0	0.62
352	30.0	0.62
370	28.0	0.58
335	28.0	0.58
1393	28.0	0.58
6357	27.0	0.56
5542	27.0	0.56
1053	27.0	0.56
2074	26.0	0.54
1292	25.0	0.52
5789	25.0	0.52
801	24.0	0.50
6012	23.0	0.48
6248	23.0	0.48
2633	23.0	0.48
987	23.0	0.48
5510	22.0	0.46
5483	22.0	0.46
5148	21.0	0.44
4650	20.0	0.41
2519	20.0	0.41
1281	20.0	0.41
3602	20.0	0.41

Question 12 – Parsing

Parsing separates text strings into semantically-atomic tokens. The four columns for which parsing could have potentially added to the value of the data were:

- COHORT_PLAN_DESCR
- PLAN_FIRST_DCR
- STATE
- TUI_RES_CD_DCR

The image illustrates one such example:


State Filter: EOT_TGPA less than or equal to 2.5 Data Source: College of Engineering Data

Column Profiling | **Frequency Distribution** | Pattern Frequency Distribution | Percentiles | Outliers | Primar

Value	Count	Percentage
North Carolina	4466.0	92.66
PENNSYLVANIA	60.0	1.24
VIRGINIA	57.0	1.18
NEW JERSEY	49.0	1.02
MARYLAND	33.0	0.68
PEOPLE'S REPUBLIC OF CHINA	30.0	0.62
NEW YORK	28.0	0.58
HONG KONG	22.0	0.46
SOUTH CAROLINA	20.0	0.41
FLORIDA	18.0	0.37
PAKISTAN	10.0	0.21
HAWAII	7.0	0.15
LOUISIANA	7.0	0.15
WASHINGTON	5.0	0.10
EL SALVADOR	4.0	0.08
ILLINOIS	4.0	0.08

Question 13

The process of standardising the data and in particular the right-fielding of the geographical information fields very much highlighted the incompleteness of the first-year College of Engineering demographic data. This was also supported by the discovery that the entity resolution process to produce a dataset containing only one record per unique student id was not fully successful despite multiple attempts using different field combinations. As a result, the author concluded that full data for the relevant (demographic) fields for analysis and reporting was not available.

This conclusion meant that the identification of notable demographic trends for at-risk students had to be treated with some caution, particularly decisions were made based on these results.

In order to assist with data analysis, data enrichment would be of significant assistance. This would involve using external data from other university datasets to attempt to complete the records containing missing information and finishing the entity resolution process. Examples of suitable data could include a student's full mailing address or contact phone number.

Word count: 1,568 words