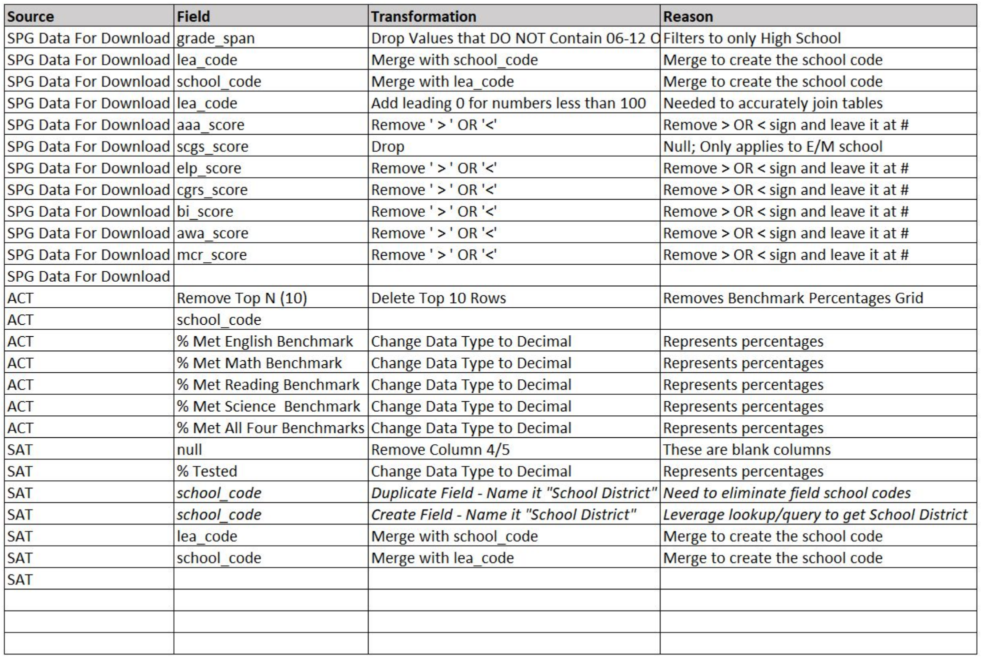
DS4A Data Cleaning Methodology Deliverable

From TAs: In general, this is a great opportunity to set a deadline for yourself to finish up your data pre-processing! A great format for this would be to prepare a Jupyter notebook where you show us your data and talk about how you did the following:

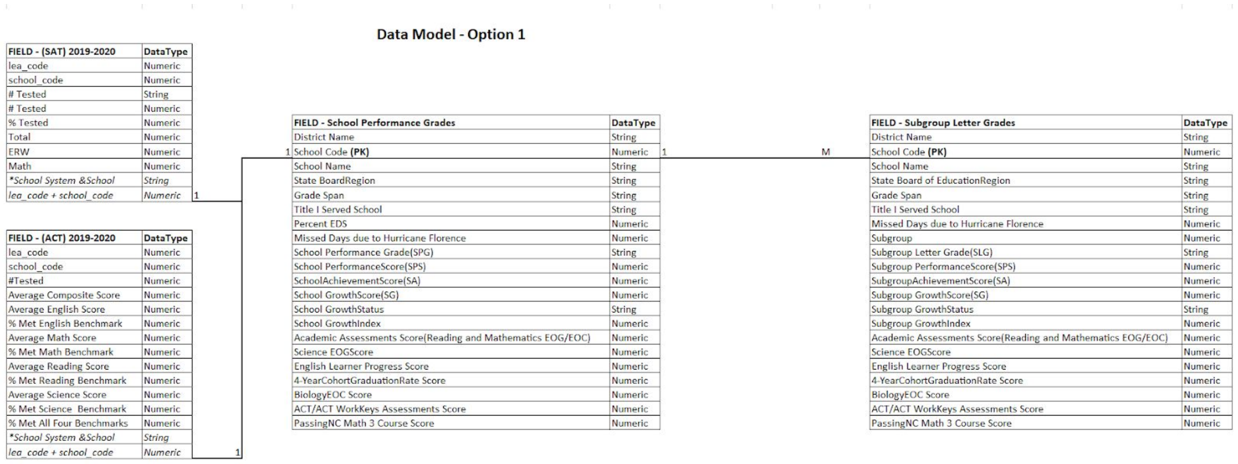
- What did you do with your null/NaN values? (ie. impute or drop)



Source	Field	Transformation	Reason
SPG Data For Download	grade_span	Drop Values that DO NOT Contain 06-12 O	Filters to only High School
SPG Data For Download	lea_code	Merge with school_code	Merge to create the school code
SPG Data For Download	school_code	Merge with lea_code	Merge to create the school code
SPG Data For Download	lea_code	Add leading 0 for numbers less than 100	Needed to accurately join tables
SPG Data For Download	aaa_score	Remove ' > ' OR '<'	Remove > OR < sign and leave it at #
SPG Data For Download	scgs_score	Drop	Null; Only applies to E/M school
SPG Data For Download	elp_score	Remove ' > ' OR '<'	Remove > OR < sign and leave it at #
SPG Data For Download	cgrs_score	Remove ' > ' OR '<'	Remove > OR < sign and leave it at #
SPG Data For Download	bi_score	Remove ' > ' OR '<'	Remove > OR < sign and leave it at #
SPG Data For Download	awa_score	Remove ' > ' OR '<'	Remove > OR < sign and leave it at #
SPG Data For Download	mcr_score	Remove ' > ' OR '<'	Remove > OR < sign and leave it at #
SPG Data For Download			
ACT	Remove Top N (10)	Delete Top 10 Rows	Removes Benchmark Percentages Grid
ACT	school_code		
ACT	% Met English Benchmark	Change Data Type to Decimal	Represents percentages
ACT	% Met Math Benchmark	Change Data Type to Decimal	Represents percentages
ACT	% Met Reading Benchmark	Change Data Type to Decimal	Represents percentages
ACT	% Met Science Benchmark	Change Data Type to Decimal	Represents percentages
ACT	% Met All Four Benchmarks	Change Data Type to Decimal	Represents percentages
SAT	null	Remove Column 4/5	These are blank columns
SAT	% Tested	Change Data Type to Decimal	Represents percentages
SAT	*school_code*	*Duplicate Field - Name it "School District"*	*Need to eliminate field school codes*
SAT	*school_code*	*Create Field - Name it "School District"*	*Leverage lookup/query to get School District*
SAT	lea_code	Merge with school_code	Merge to create the school code
SAT	school_code	Merge with lea_code	Merge to create the school code
SAT			

- What feature engineering did you do?

- If you had multiple datasets how did you join all this info?



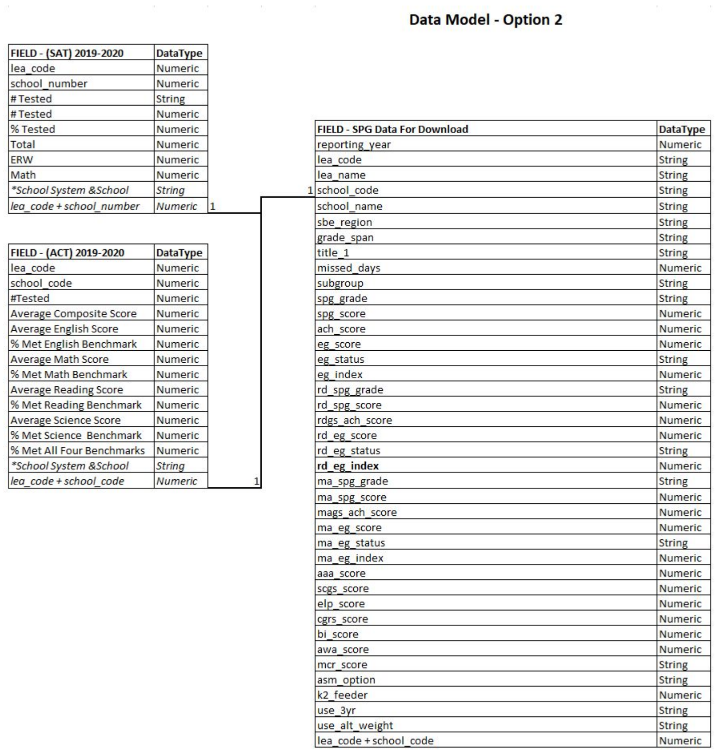
Data Model - Option 1

FIELD - (SAT) 2019-2020	DataType
lea_code	Numeric
school_code	Numeric
# Tested	String
# Tested	Numeric
% Tested	Numeric
Total	Numeric
ERW	Numeric
Math	Numeric
**School System &School*	*String*
lea_code + school_code	*Numeric*

FIELD - (ACT) 2019-2020	DataType
lea_code	Numeric
school_code	Numeric
#Tested	Numeric
Average Composite Score	Numeric
Average English Score	Numeric
% Met English Benchmark	Numeric
Average Math Score	Numeric
% Met Math Benchmark	Numeric
Average Reading Score	Numeric
% Met Reading Benchmark	Numeric
Average Science Score	Numeric
% Met Science Benchmark	Numeric
% Met All Four Benchmarks	Numeric
**School System &School*	*String*
lea_code + school_code	*Numeric*

FIELD - School Performance Grades	DataType
District Name	String
School Code **(PK)**	Numeric
School Name	String
State BoardRegion	String
Grade Span	String
Title I Served School	String
Percent EDS	Numeric
Missed Days due to Hurricane Florence	Numeric
School Performance Grade(SPG)	String
School PerformanceScore(SPS)	Numeric
SchoolAchievementScore(SA)	Numeric
School GrowthScore(SG)	Numeric
School GrowthStatus	String
School GrowthIndex	Numeric
Academic Assessments Score(Reading and Mathematics EOG/EOC)	Numeric
Science EOGScore	Numeric
English Learner Progress Score	Numeric
4-YearCohortGraduationRate Score	Numeric
BiologyEOC Score	Numeric
ACT/ACT WorkKeys Assessments Score	Numeric
PassingNC Math 3 Course Score	Numeric

FIELD - Subgroup Letter Grades	DataType
District Name	String
School Code **(PK)**	Numeric
School Name	String
State Board of EducationRegion	String
Grade Span	String
Title I Served School	String
Missed Days due to Hurricane Florence	Numeric
Subgroup	Numeric
Subgroup Letter Grade(SLG)	String
Subgroup PerformanceScore(SPS)	Numeric
SubgroupAchievementScore(SA)	Numeric
Subgroup GrowthScore(SG)	Numeric
Subgroup GrowthStatus	String
Subgroup GrowthIndex	Numeric
Academic Assessments Score(Reading and Mathematics EOG/EOC)	Numeric
Science EOGScore	Numeric
English Learner Progress Score	Numeric
4-YearCohortGraduationRate Score	Numeric
BiologyEOC Score	Numeric
ACT/ACT WorkKeys Assessments Score	Numeric
PassingNC Math 3 Course Score	Numeric



Data Model - Option 2

FIELD - (SAT) 2019-2020	DataType
lea_code	Numeric
school_number	Numeric
# Tested	String
# Tested	Numeric
% Tested	Numeric
Total	Numeric
ERW	Numeric
Math	Numeric
**School System &School*	*String*
lea_code + school_number	*Numeric*

FIELD - (ACT) 2019-2020	DataType
lea_code	Numeric
school_code	Numeric
#Tested	Numeric
Average Composite Score	Numeric
Average English Score	Numeric
% Met English Benchmark	Numeric
Average Math Score	Numeric
% Met Math Benchmark	Numeric
Average Reading Score	Numeric
% Met Reading Benchmark	Numeric
Average Science Score	Numeric
% Met Science Benchmark	Numeric
% Met All Four Benchmarks	Numeric
**School System &School*	*String*
lea_code + school_code	*Numeric*

FIELD - SPG Data For Download	DataType
reporting_year	Numeric
lea_code	String
lea_name	String
school_code	String
school_name	String
sbe_region	String
grade_span	String
title_1	String
missed_days	Numeric
subgroup	String
spg_grade	String
spg_score	Numeric
ach_score	Numeric
eg_score	Numeric
eg_status	String
eg_index	Numeric
rd_spg_grade	String
rd_spg_score	Numeric
rdgs_ach_score	Numeric
rd_eg_score	Numeric
rd_eg_status	String
rd_eg_index	Numeric
ma_spg_grade	String
ma_spg_score	Numeric
mags_ach_score	Numeric
ma_eg_score	Numeric
ma_eg_status	String
ma_eg_index	Numeric
aaa_score	Numeric
scgs_score	Numeric
elp_score	Numeric
cgrs_score	Numeric
bi_score	Numeric
awa_score	Numeric
mcr_score	String
asm_option	String
k2_feeder	Numeric
use_3yr	String
use_alt_weight	String
lea_code + school_code	Numeric

- Did you employ any kind of feature selection methods? (ie. lasso or ridge regression)

No

- Did you look at the distribution of your variables and if so did you perform any kind of transformation and if so, why? (ie. normalization or log transformation)

Yes to consider if there was any unimodal or bimodal observations

- How did you address outliers?

- Talk about any technique-specific cleaning you had to do (ie. the steps of text pre-processing if you're doing NLP)

- If you didn't have a chance to finish up your pre-processing, what are your future steps?

Data cleaning notes

- **Dropping "reporting year column" there is only one uniques value.**
- **It is necessary to start with grade_span column, we drop the elementary and middle schools, we keep only high schools.**

The minimum number of scores needed for an indicator to be used is 30. Students must be enrolled in the same school for at least half the instructional period to be included.

title_1 Schools 'nan' values will be 'N'

spg_grade 'nan' values will be considered 'I'

eg_status 'nan' values will be considered 'I'

spg_score 'nan' values will be 999 (with the understanding that any score higher than 100 means that there is insufficient data)

eg_score 'nan' values will be 999 (with the understanding that any score higher than 100 means that there is insufficient data)