School's Out: A Console-Based Application for Analysing Pupil Absence Data

Ivor Walker

Word Count: 1,437

Introduction

Understanding drivers behind pupil attendance enables governments to reward the highest performing schools, support those that are struggling, and ultimately prioritise resources towards initiatives that affect the most change in pupil attendance. This report details the development of a console-based application in Python and Apache Spark to analyse a large dataset of pupil absence in schools in England from 2006-2018. The application meets all requirements listed in Part 1, 2, and 3. 2

Design

main.py and Menu

main.py is the entry point of the application, which creates an instance of the Menu class. The Menu class starts the main menu loop, asking users for which analysis they would like to perform, getting the required data from the Absence class, and displaying the results using the View class. All expected exceptions (e.g View raising an invalid user input or SparkData raising a nonexistent local authority) are caught here.

View

The View class contains methods to display tables returned from the Absences class in a readable table format or chart, and methods to request input from the user. When the Menu class asks for user input, it also passes asses the expected format of the user input to the View class, which validates the input and returns it to the Menu class.

SparkData

Constructor

The constructor of SparkData starts the Spark session and loads a csv at the provided location into a Spark Dataframe. The user can input strings (e.g for analyses for a specific local authority) which should be case insensitive, but the data itself is in a specific case format. After loading data, I infer the case of each column based on a sample of each column (n = 50 by default) and create a map of all string columns with their case. I tranform any user input to the correct case of the requested column before searching. Certain supposedly numeric values (e.g ”sess_auth_ext_holiday”) sometimes contained symbols, which I infer to be missing data and replace with 0. I round all numeric values to five decimal places to improve output readability.

Getting subframes

SparkData contains methods to return subsets of any DataFrame in certain formats and aggregations. _get_frame returns all requested columns of the DataFrame, filtered by where the supplied filter_cols is in the corresponding list of values in filter_passes. _get_agg_frame returns a __get_grouped_frame of a _get_frame subframe. __get_grouped_frame is a transformation of any DataFrame that pivots on one specific column and aggregated (counts, sums, or mean) by another column. _get_multi_col_agg_frame stacks the columns of a _get_frame subframe, unpivots it, and returns a __get_grouped_frame of the unpivoted DataFrame - effectively aggregating by all columns selected in the original _get_frame query. Certain functionality cannot use Spark Dataframes as input (e.g matplotlib for charting), so _collect_to_dict takes an input DataFrame, collects it to the driver, computes means and confidence intervals, and returns a usable dictionary of the results.

Absences

Absences inherits from SparkData and passes the location of the absences file to SparkData's construction. It contains methods called by Menu to return data required specifically for analyses of absence data by wrapping around the subframe methods in SparkData.

It sets default values of filters which are always passed to _get_frame, e.g always excluding school type and geographical aggregation rows in the returned frames, and fetches some values for specific columns in the dataset which can be used instead of user input to simplify testing.

Comparing local authorities

Comparing local authorities by overall absence rate adjusts for the size of the local authority which could otherwise skew our results. The goal of governments is to reduce overall absence rates, but authorised absence rates are less of a problem because they represent unavoidable absences whereas unauthorised absences represent some disconnect between the school and the pupil. I compare all three figures to assess the performance of local authorities.

Students commit absences at different rates, and modelling showed that the percentage of enrolments that are persistent absentees has the largest effect on overall absence rates outside of time period and region. 3 I show this figure to partially explain the overall absence rate, and compare percentages of unauthorised absences by persistent absentees to show the proportion of pupils that are totally disconnected from the school and need the most support.

If the user chooses to compare more than five local authorities, I add a confidence interval around the mean of each category to enable conclusions on statistical significance, e.g a council has a higher percentage of unauthorised absences than the average council chosen.

Methodology for modelling absence rates

Alongside visualisations, I fit a model to precisely quantify how changes in regions and school types affect absences, and capture variance from other variables. I also included interaction terms between region and school type to explroe the effect of the relationship between region and school type on absences.

Absences contains methods for producing data required for modelling, a modelling framework, and a fitted model explaining overall absences. A Generalised Linear Model (GLM) retains explainable coefficients which is useful for explaining our response. A Poisson distribution fits our response variable best because it is a count, but I applied an offset of the number of possible sessions to account for different school sizes so we are effectively modelling the absence rate.

I tried to use all covariates initially, but the model failed to fit because some variables were perfectly "multi-colinear" (i.e a covariate was related to another covariate in the data). I created a "_print_high_collinearity" method to identify these variables and found that any school that had an associated academy type was a state-funded school. I removed all academy related covariates and the model fitted successfully. My final model is:

$$
\begin{aligned}
\text{overall absence sessions} &\sim \text{Poisson}(\lambda) \\
\log(\lambda) - \log(\text{total possible sessions}) &= \eta \\
\eta &= \beta_0 + \beta_1 \text{time period} + \beta_2 \text{region} \\
+\beta_4 \text{enrolments} &+ \beta_5 \% \text{ unauthorised sessions} \\
+\beta_6 \% \text{ overall absences committed by persistent absentees} &+ \beta_7 \% \text{ unauthorized absences committed by persistent absentees} \\
&+ \beta_8 \% \text{ enrolments that are persistent absentees}
\end{aligned} \tag{1}
$$

Results

Analysing pupil attendance by region over time

Attendance for all regions over time

I chose overall absence rate as the metric to represent pupil attendance to account for the different sizes of regions. Similarly, I compared the absence rate for each region against the mean absence rate for all regions instead of against the national absence rate to avoid larger regions skewing the results.

All regions has a significantly lower absence rate in 2018/19 than the mean council in 2006/07. The absence rate for the mean council falls drastically between 2006/07 and 2013/14, almost always being significantly lower than the previous year, but increases from 2014/15 until 2018/19. This increase is more intermittent and smaller in scale, with the new mean level remaining significantly lower than any year before 2013/14. 3

Improving and declining regions

Outer and Inner London has absence rates from 2018/19 that are within the confidence intervals for 2013/14, indicating these regions have significantly avoided the increased absence rates since 2014/15. Conversly, North East and Yorkshire and the Humber have absence rates from 2018/19 that are within the confidence intervals for 2011/12 and 2012/13, indicating these regions have not significantly benefitted from the overall decline in absence rates until 2013/14. 3
Overall, Outer London has the lowest absence rate across all years, whereas North East has the highest. 1

Exploring the relationship between school type, pupil absences and location

Overall absence rates by school type and region

Across all years, Inner London and Outer London have significantly lower absence rates than the mean region, whereas North East and Yorkshire and the Humber have significantly higher absence rates. 1 Special schools appear to have significantly higher absence rates than the mean school type 2.

Relationship between school type and region

outh West and East Midlands have significantly lower state funded secondary schools and special schools, and higher primary schools than the mean region. Outer London and Inner London have significantly higher state-funded secondary schools and special schools, but significantly fewer state-funded primary schools. East of England, South West, East Midlands and Yorkshire and the Humber have significantly lower special schools. **??**

Model results

All variables appear significant (see Discussion) but some have larger effect sizes than others. State funded secondary schools have a 2% lower absence rate and state funded primary schools have a 6% lower absence rate compared to special schools. Compared to East Midlands, some regions have higher absence rates (East of England: 3%, Yorkshire and the Humber: 2%, West Midlands: 2%, North West: 2%, North East: 1%, South East: 1%) and others have lower absence rates (Inner London: 4%, South West: 2%, Outer London: 2%). 3
These results differ from the previous analysis, e.g South West has an equal effect size as Outer London but did not have significantly lower absence rates than the mean region. This model is a more complete picture as it accounts for the effect of other variables on absence rates and the interaction components accounts for the relationship between school type and region.

Discussion

Case sensitivity

The standard approach to guaranteeing case sensitivity is to set the user input to lowercase and search on a lowercase version of the data. The size of the dataset means producing lowercase versions of the data is not feasible.
I transform the user input to the correct case before searching to bypass this issue, but this approach is sensitive to the case being applied consistently throughout the column being searched, e.g. local authority names are inferred to be proper case and nearly all have a lowercase "of" (e.g City of London), but in the dataset 'Isles of Scilly' has a title case "Of" so searching for "Isles Of Scilly" would not return any results as it would be transformed to "Isles of Scilly".

These issues are extremely rare and does not remove these mixed case names from aggregates which makes this tradeoff worth it. Searching for all violations of the case mapping and adding them to some "edge case" mapping would incur the same cost as producing a lowercase version of the data. As a compromise, I manually include encountered edge cases in Absences and inform SparkData to transform these edge cases to the exception before searching. This approach is appropriate here because of the vanishingly small number of edge cases, but manually listing edge cases is not scalable for datasets with larger number of edge cases.

School types and absence rates

The chart proporting to show significant differences in absence rates between school types 2 includes the overall rate of absence, which biases the mean and associated confidence intervals downwards and creates the illusion that special schools have significantly higher absence rates than the mean school type. Comparing the differences between only three school types produces very large confidence intervals, so I included the total as a reference point. The proper statistical tool to compare the absence rates between school types is a t-test, but this is not implemented in Spark.

Significance in the model

The fitted model is overdispersed (estimated dispersion parameter: 152017105 / 276816 = 549.2, when it should be 1). Spark does not support quasi-poisson or negative binomial models that are needed to address this, so fixing this would require an external data analysis package which is forbidden by the spec. Consequently, the reported standard errors are underestimated leading to misleadingly small p-values. The model is also overpowered - the extremely large sample size is much higher than the minimum required to detect significance. Spark does not support power analysis required to find the minimum number of samples needed to detect a significant effect, but if I could perform this analysis I could drastically reduce the amount of data being handled to the minimum required. Because all covariates appear significant at this sample size and with this extreme overdispersion 3, I focus on interpreting the coefficients as effect sizes which remain unbiased in the face of dispersion.

Reflective summary

Data extraction and transformation

Although the SparkData class contains fewer methods and interactions with other classes, which are the typical sources of complexity, I spent the most time developing it because of the sheer level of 'precision' required when transforming data. Putting these transformation functions together to produce the desired output required relatively few lines of code, but how these transformation functions (e.g pivot) affect data was unintuitive because of how specific the transformations are to the data being transformed, and I spent a lot of time outside of the development environment thinking about how to transform the data to produce the desired output. However, I wish I spent more time thinking about exactly what transformations were required as I wasted time developing some methods (e.g the batch methods returning multiple DataFrames) that I later realised were not required.

Program design

Creating generic methods to return the types of subframes required proved to be a good decision as it abstracted the somewhat confusing transformations away from myself and instead put data extraction in terms of the "rows" and "columns" of the DataFrames that I needed which simplified the framework sitting ontop of SparkData.

Statistical analysis

I found the statistical tools that Spark offered to be frustratingly limited. Because the use of data analysis packages was forbidden in the assignment, I had to rely on simpler statistical tools e.g confidence intervals, and present flaws in my modelling approaches that I could not address because the appropriate procedure was not in Spark. Spark does not provide model diagnostics, so I had to implement methods to check for multicolinearity and overdispersion and could not check for other issues e.g outliers.

The lack of interpretable non-linear modelling frameworks (e.g generalised additive models) restricted the temporal and spacial modelling that I could perform. For example, the same page that contains this dataset also contains a dataset listing school postcodes. If I had more sophisticated statistical tools, I could join this data onto the existing data, use an API (e.g postcodes.io) to convert these postcodes to latitude and longitude, and use this to more completely model the effect of location on absence rates instead of in a non-contiguous way as I have done here.

Given that Spark is designed for large datasets, I found the lack of power analysis particularly strange as using the results from a power analysis could drastically reduce the amount of data that needs to be processed to discover relationships between data.

Appendix

Feature	Coefficient	P-value
time_period_201819	0.95	0.00
time_period_201718	0.97	0.00
time_period_201617	0.95	0.00
time_period_201516	0.95	0.00
time_period_201415	0.95	0.00
time_period_201314	0.93	0.00
time_period_201213	1.00	0.00
time_period_201112	0.97	0.00
time_period_201011	1.01	0.00
time_period_200910	1.01	0.00
time_period_200809	1.01	0.00
region_name_Yorkshire and the Humber	1.02	0.00
region_name_West Midlands	1.02	0.00
region_name_South West	0.98	0.00
region_name_South East	1.01	0.00
region_name_Outer London	0.98	0.00
region_name_North West	0.98	0.00
region_name_North East	1.01	0.00
region_name_Inner London	0.96	0.00
region_name_East of England	1.03	0.00
school_type_State-funded secondary	0.98	0.00
school_type_State-funded primary	0.94	0.00
enrolments	1.00	0.00
unauthorised_percent	0.99	0.00
overall_percent_persistent_absentees	1.01	0.00
unauthorised_percent_persistent_absentees	1.01	0.00
enrolments_persistent_absentees_percent	1.03	0.00
Yorkshire and the Humber:State-funded secondary	0.98	0.00
West Midlands:State-funded secondary	0.98	0.00
South West:State-funded secondary	1.03	0.00
South East:State-funded secondary	1.00	0.00
Outer London:State-funded secondary	1.02	0.00
North West:State-funded secondary	1.01	0.00
North East:State-funded secondary	0.99	0.00
Inner London:State-funded secondary	1.02	0.00
East of England:State-funded secondary	0.98	0.00
Yorkshire and the Humber:State-funded primary	0.99	0.00
West Midlands:State-funded primary	0.98	0.00
South West:State-funded primary	1.05	0.00
South East:State-funded primary	0.99	0.00
Outer London:State-funded primary	1.03	0.00
North West:State-funded primary	1.01	0.00
North East:State-funded primary	1.01	0.00
Inner London:State-funded primary	1.04	0.00
East of England:State-funded primary	0.99	0.00

Table 3: Table of all coefficients and p-values for the model of absence rates. The reference level for region name is East Midlands, for school type is special school, and for year is 2007/08. Coefficients are the estimated effect size of the variable on the absence rate, e.g state funded primary schools in Yorkshire and the Humber have 1% lower absence rates than special schools in East Midlands. All variables are significant predictors of absence rates ($p < 0.001$).

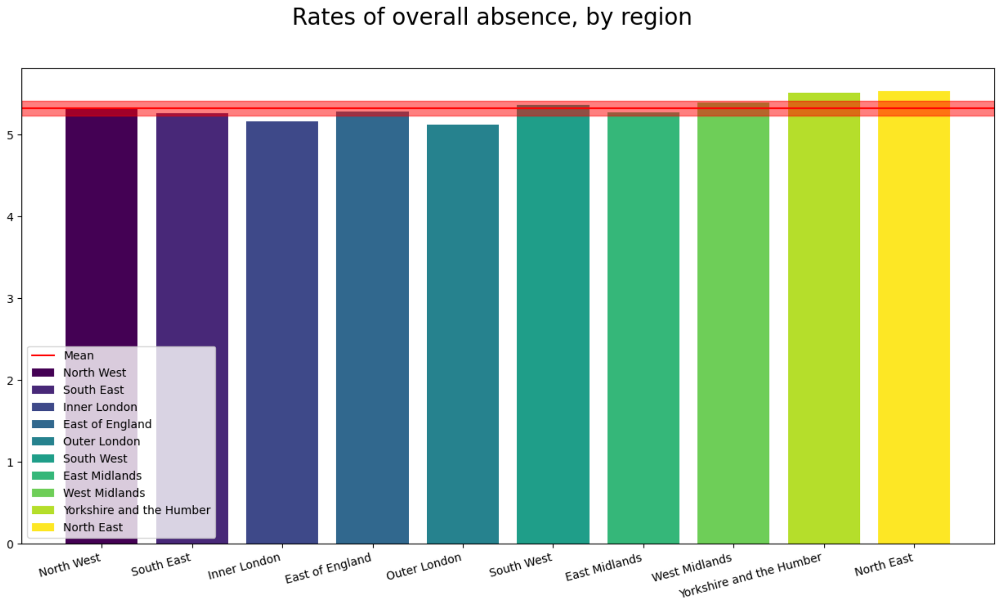

Figure 1: Bar chart of overall absence rates by region, with a line representing the mean absence rate and a shaded area representing the 95% confidence interval of the mean. Values below or above the shaded area are significantly different from the mean region's absence rate, e.g Inner London has significantly lower absence rates than the mean region.

Part and bulletpoint in spec	**Implemented feature to achieve spec**	**Line in Menu class**	**Line in Absences class**	**Line in SparkData class**	**Line in View class**
1a, 1b	Read and store dataset	16	30	30	NA
1c	Pupil enrolments over time by local authority	88	146	542	134
1d	Authorised absences by school type in year	109	209	542	134
1d extension	Authorised absence reasons by school type in year	141	253	784	134
1e	Unauthorised absences by region name or local authority in year	172	359	542	134
2a	Compare local authorities in year	209	422	784	134, 155
2b	Absences by local authority over time	247	477	542	134, 260
3	Chart link between school type, pupil absences, and school location	262	521, 558, 592	542	134, 260, 155
3	Model pupil absences using school type and school location	262, 289	651, 636, 810, 793, 816	585	26

Table 2: Table of all implemented requirements and their locations in the code. For example, the third bulletpoint of the first part of the spec asks to select pupil enrolments over time by local authority. Choosing this option in the menu runs line 88 in the Menu class, which gets data by calling line 146 in the Absences class (in turn calling line 542 in the SparkData class) and displays it by calling line 134 in the View class.

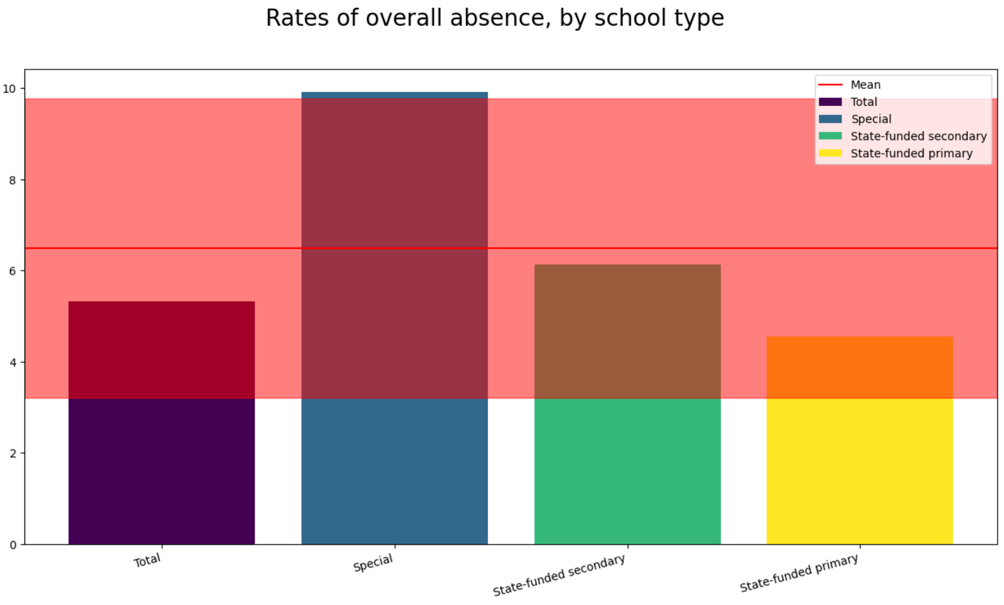

Figure 2: Bar chart of overall absence rates by school type, with a line representing the mean absence rate and a shaded area representing the 95% confidence interval of the mean. Values below or above the shaded area are significantly different from the mean absence rate, e.g special schools have significantly higher absence rates than the mean school type. However, this chart is a misleading representation (see Discussion).

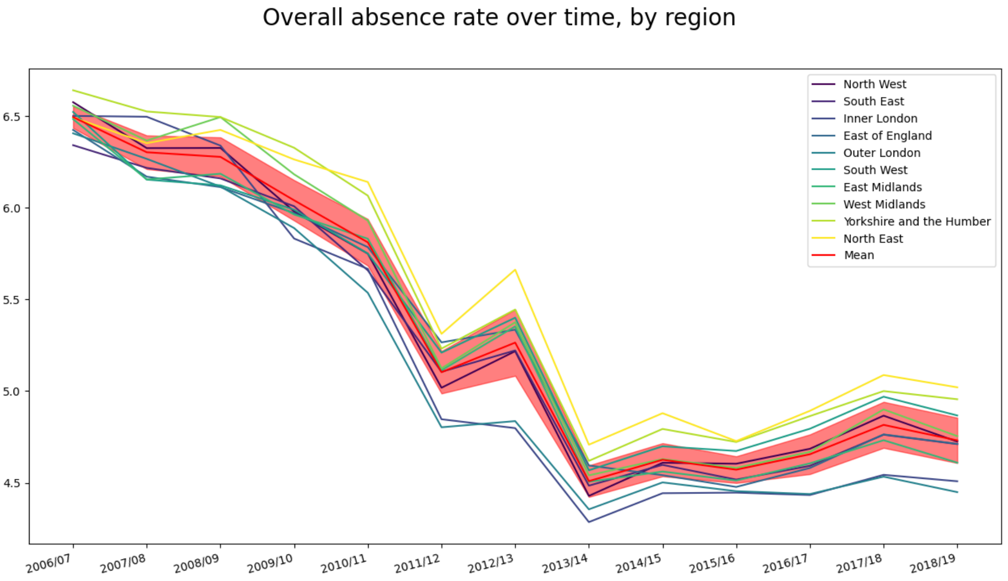

Figure 3: Line chart of overall absence rates over time by region, with a line representing the mean absence rate and a shaded area representing the 95% confidence interval of the mean. Values below or above the shaded area are significantly different from the mean region's absence rate for that year, e.g Yorkshire and the Humber has significantly higher absence rates than the mean region in 2006/07. A mean absence rate for one year is significantly different from another year if the new mean is outside the shaded area of the old mean, e.g the mean absence rate in 2006/07 is significantly higher than the mean absence rate in 2007/08.

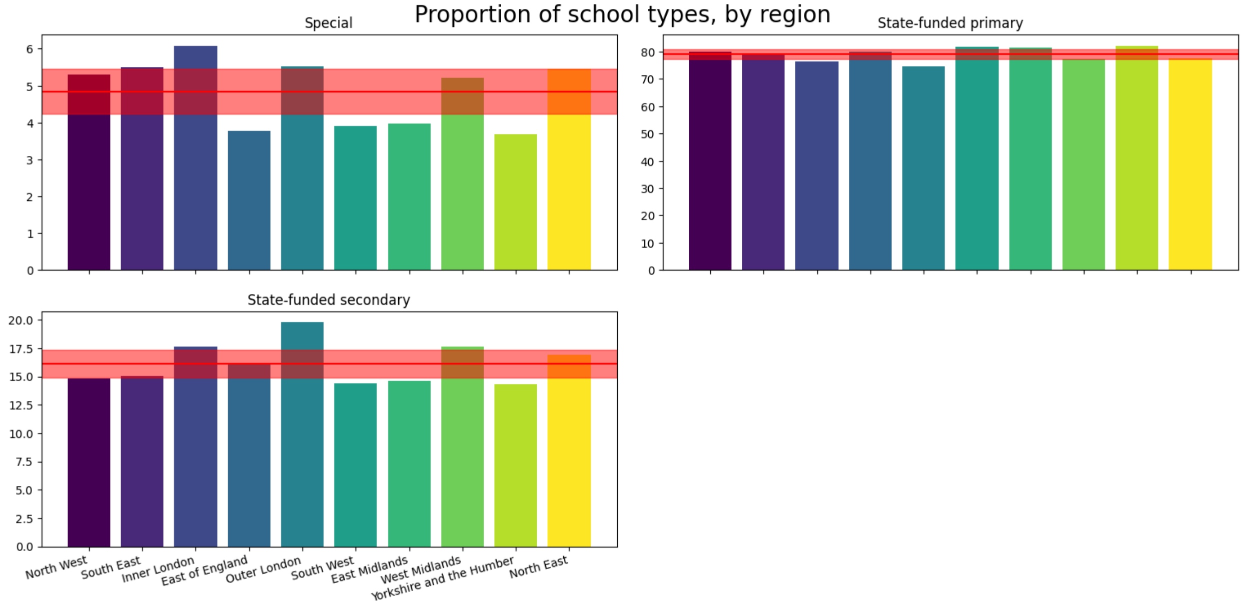

Figure 4: Bar chart of the proportion of each school type by region, with a line representing the mean proportion of each school type and a shaded area representing the 95% confidence interval of the mean. Values below or above the shaded area are significantly different from the mean proportion of that school type, e.g Inner London has significantly more special schools than the mean region.