CS5052 Practical 1

Student ID: 210017213
Word count: 1494

Overview

From 2006-2018 the UK government recorded a variety of data regarding pupil absences in schools within England. This coursework task required utilising Python and Apache Spark to perform data preparation processes to the dataset, execute multiple queries followed by performing data analysis on the results generated.

Features Implemented

Feature	Section	Implemented
Read in dataset using Apache Spark	Part 1	Yes
Store the dataset using methods supported by Apache Spark		Yes
Query to search the dataset by local authority, showing number of pupil enrolments in each local authority by time period (Part 1 A)		Yes
Query to search the dataset by school type, showing the total number of pupils who were given authorised absences in a specific time period (Part 1 B)		Yes
Query to search for all unauthorised absences in a certain year, broken down by either region name or local authority name (Part 1 C)		Yes
Query to compare two local authorities of their choosing in a given year. Justify how you will compare and present the data (Part 2 A)	Part 2	Yes
Chart/explore the performance of regions in England from 2006-2018. Your charts and subsequent analysis in your report should answer the following questions (see Analysis, Part 2) (Part 2 B)		Yes
Explore whether there is a link between school type, pupil absences and the location of the school (see Analysis, Part 3)	Part 3	Yes

Code Lines

Citation	Functionality	Code lines
(a)	Writing to the parquet file	12-15
(b)	Replacing missing values with “None”	27-36
(c)	Ensuring time_period attribute was of type int	27-36
	Part 1 A	38 - 49
	Part 1 B	52 - 68
	Part 1 C	70 - 77
	Part 2 A	80 - 90
	Part 2 B	92 - 143
	Part 3	146 - 255

Design

Data flow diagram:

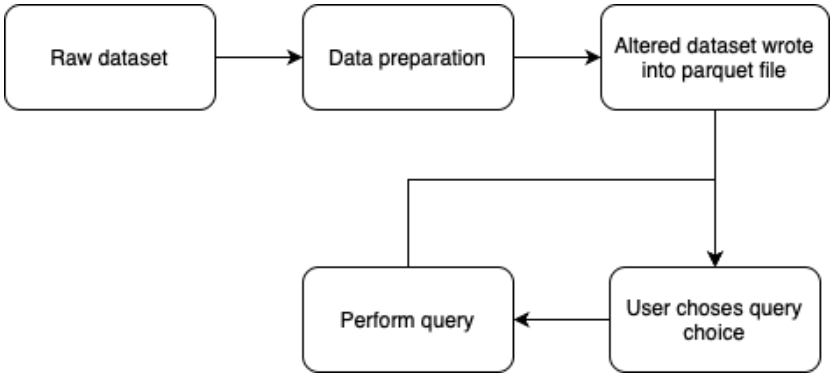

Figure 1 Data Flow Diagram

Figure 1 shows the process data follows from the raw dataset to the user being able to select from a range of queries to execute.

Implementation

Reading/Writing Dataset:

The raw dataset is read utilising spark functions available as part of the SparkSession, SQL PySpark package. After completing the data preparations steps (detailed below) the altered dataset was wrote into a parquet file (a) due to its higher performance and optimisation for performing queries due to its efficient data storage leading to minimised latency when accessing data [1].

Data Preparation:

Multiple data preparation processing steps were required to ensure the data integrity within the dataset. My strategy to deal with any missing numerical data was to replace with “None” (b), I chose this approach instead of replacing with zero because it could have an impact on any averaged percentage values if zero values are utilised in the calculation.

Further, to ensure data computing and integrity when utilising retrieved values from the dataset it was necessary to enforce type casting. For example, this was important for the *time_period* attribute to ensure it was of type integer (c).

Performing Queries

After analysing the data, I noticed that included in the data were rows with *school_type* of Total and this meant than when considering a whole region or local authority the Total row could be utilised rather than needing to perform sum aggregations on the different school types resulting in a simple query needing to be performed.

Analysis

Part 2:

Q1: “Are there any regions that have improved in pupil attendance over the years?”

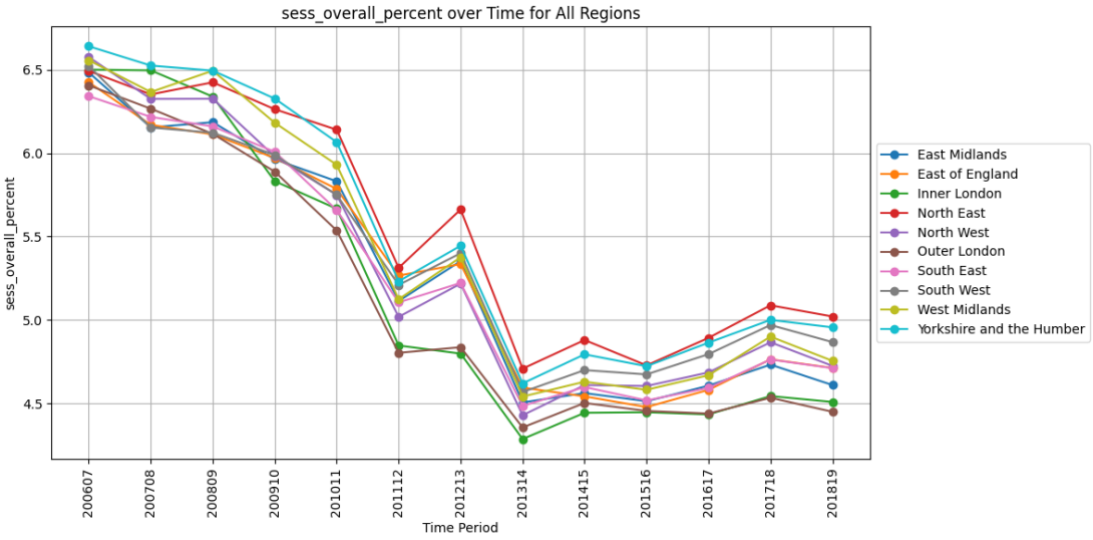

Figure 2 Overall Absence Percentage Per Region Over Time

Illustrated in Figure 2, it can be observed that the overall absence rate decreases in all regions from 2006/07 – 2011/12 until a spike in 2012/13 after which the absence rate drops overall from 2012/13 – 2018/19. From the beginning of the results recording in 2006/07 until 2018/19 the absence rate drops in every region, showing that all regions have less absent pupils by percentage of total pupils meaning that more pupils are attending school.

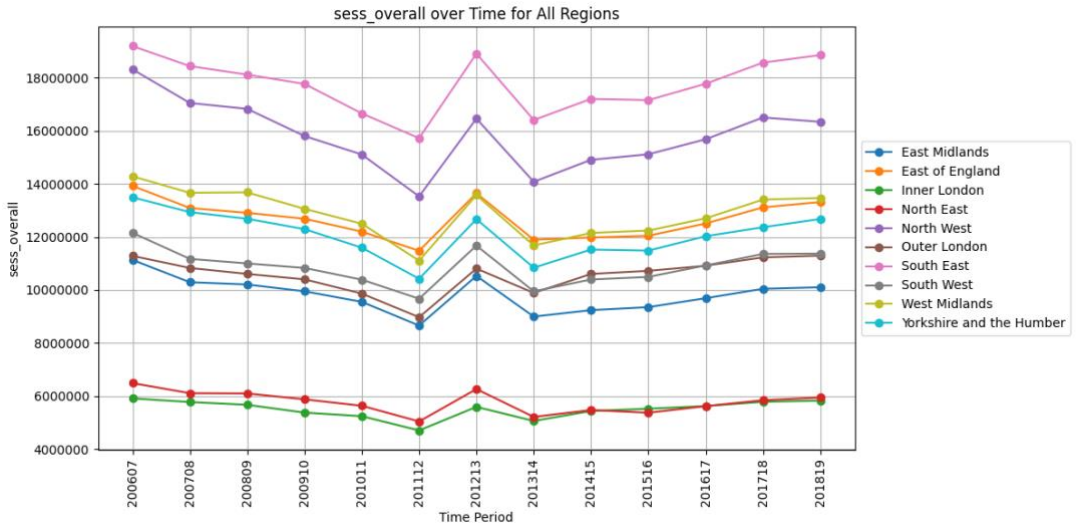

Figure 3 Overall Absences Per Region Over Time

Figure 3 shows the overall absences (*sess_overall*) per region for each time period in chronological order. Pupil attendance relates to when pupils are present within school, meaning this can be measured and evaluated using the overall amount of pupil absences. From Figure 3 it can be observed that in all regions from 2006/07 – 2011/12 decreased in overall absences, meaning an increase in pupil attendance. Further, from Figure 3 2006/07 – 2018/19 it can be observed all regions have decreased in total absences or remained relatively similar in the case of Inner or Outer London. However, the overall absences metric does not necessarily provide a clear metric to measure the pupil attendance as it does not consider the number of pupils who are not absent, the pupils that are attending school.

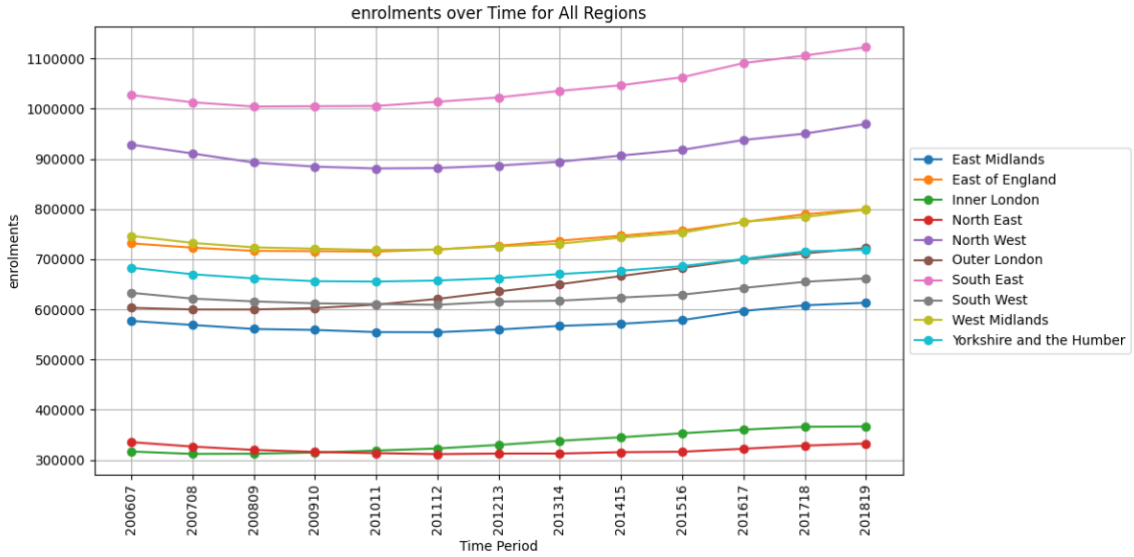

Figure 4 Enrolments by Region over Time

With the number of total pupils present in schools fluctuating year on year shown in Figure 4. The absences represented as a percentage of total pupils shown in Figure 2 is a more appropriate measure to gain an insight into the impact on pupil attendance over the time period 2006/07 – 2018/19.

Q2: "Are there any regions that have worsened?"

From analysing the overall absences (Figure 3) and overall absence by percentage (Figure 2) it can't be said that any region has worsened from 2006/07 – 2018/19 as all regions have fewer overall absences, or similar. However, it can be observed that from 2013/14 – 2018/19 there is a trend that all regions have experienced an increase in absences in total and by percentage of pupils.

Q3: "Which is the overall best/worst region for pupil attendance?"

To determine the best and worst region for pupil attendance I used the overall absence rate as this provides the most standardised metric across all regions as they have different total pupils. From analysing Figure 3 it can be observed that the North London region has on average the highest absence rate across time periods compared to the

other regions meaning it is the worst region for pupil attendance as it on average has the most proportion of pupils not attending school. The best region for pupil attendance is determined using a similar approach and it can be seen from Figure 3 that both the Inner and Outer London have the lowest absence rate throughout the time periods. This translates into the Outer London and Inner London regions have the best pupil attendance compared to the other regions.

Part 3:

To determine whether a link exists between the school type, pupil absences and location of school I decided to focus on the different region locations available in the dataset and compare the absences of the different school types. Further, I utilised the *sess_overall_percent* attribute, the overall absence rate and the *sess_overall* attribute, the overall absence count. These were chosen as they provide an overall picture of the absences in regions by the school type. The overall absence rate is important to include as it is a more standardised measurement for analysing the absences per school type. My approach to perform the analysis was to create multiple graphical representations relating to the overall absence rate and overall absence count measurements.

The first graphical representation Figure 5 shows the average overall absence rate across regions for all school types.

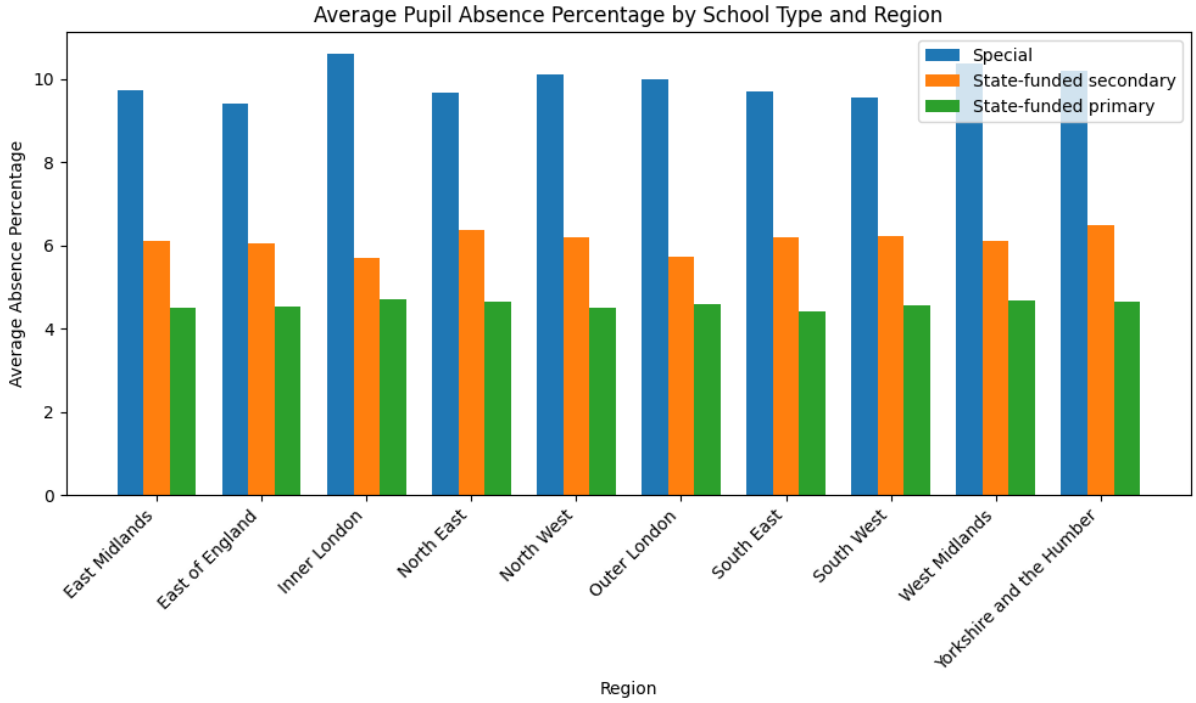

Figure 5 Average Absence Rate by School Type and Region

The second graphical representation Figure 6 shows the average pupil absence count from all time periods across the regions.

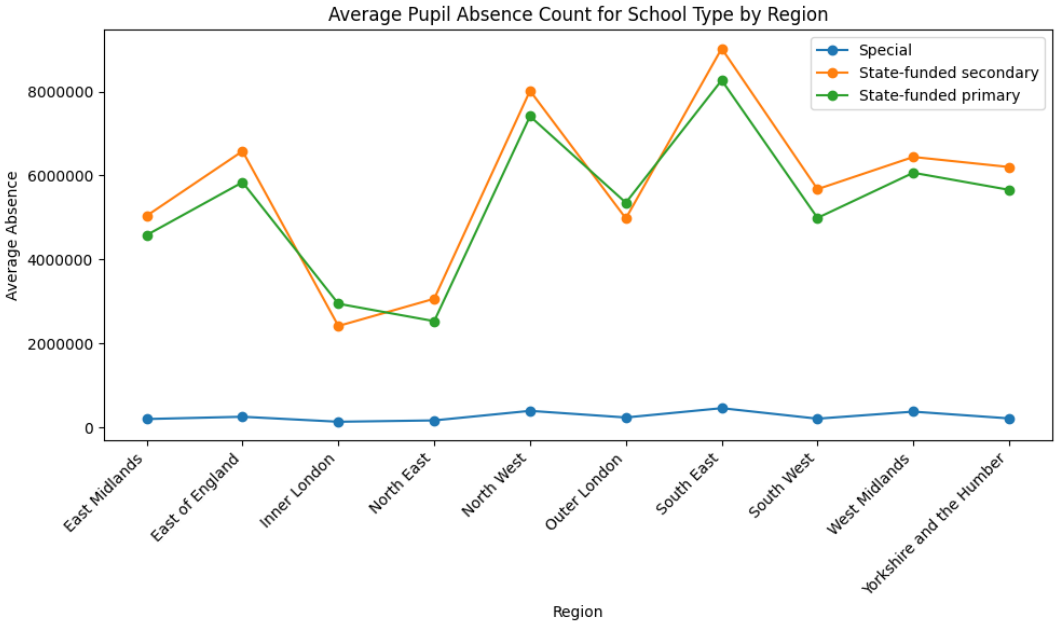

Figure 6 Average Absence Count for School Type by Region

It can be seen in Figure 5 that schools of the type Special consistently have the highest average overall absence rate. This highlights that the school type is a strong factor in the pupil absence rates. School type is further shown to be a factor as the other school types both produce consistent results across all regions regarding the average overall absence rate.

From Figure 6 it can be seen more clearly that the regional location does influence the number of absences in State-funded primary and secondary school types. However, for schools of type special the regional location appears to have little impact, but this is due to the smaller number of pupils in schools of the Special type.

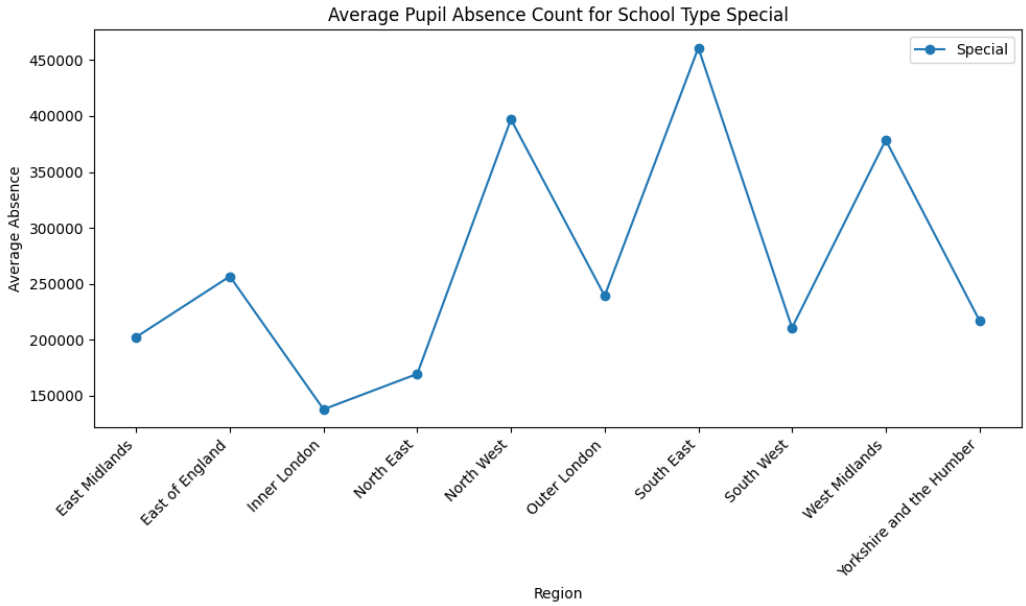

Figure 7 Overall Absence Count Across Regions for Special School Type

When isolating the overall absence count across regions of the school type Special it can be noticed that all of the school types follow a similar trend in terms of the regions with the highest amounts of overall absences.

From this, my analysis had led me to conclude that the region does play a factor in the absence rate as it can be seen to impact all school types separately. Further the type of school, not regarding the region does also affect the absence rates. However, from the evidence I have gathered I do not believe together there is a link between all three aspects of location school type and number of absences.

Reflective Summary

This practical provided a helpful insight and use of a new tool that I will now be able to confidentially use in the future. A challenge I faced was deciding the best approach of what graphs to make to provide enough information to complete thorough analysis. However after much thought and research into different techniques I was able to make more informed choices reagarding my approach.

References

[1] “Data Bricks,” [Online]. Available: https://www.databricks.com/glossary/what-is-parquet. [Accessed 05 03 2025].