2020-0701 IST 707 Data Analytics

Homework #2

Sathish Kumar Rajendiran

2020-07-17

Table of Contents

[Task Summary 2](#_Toc45928369)

[Data Exploration 2](#_Toc45928370)

[Summarization: 9](#_Toc45928371)

[Conclusion: 13](#_Toc45928372)

## **Task Summary**

Answer the data questions from the dataset given using data exploration, transformation and visualization techniques appropriately.

1. Dataset:

Small dataset that contains information from 5 schools (A, B, C, D and E) implementing same math course along 35 lessons in summer. The semester is about 3/4th through completion having spread of number of student’s record as below

* very ahead (more than 5 lessons ahead)
* middling (5 lessons ahead to 0 lessons ahead)
* behind (1 to 5 lessons behind)
* more behind (6 to 10 lessons behind)
* very behind (more than 10 lessons behind)
* completed (finished with the course)

1. Data Question(s):
   1. What’s the story(stories) in this data?

## **Data Exploration**

This step involves, various steps including pre-processing, data cleaning, aggregation and necessary transformation.

1. Import Libraries

A screenshot of a cell phone

Description automatically generated

1. Import Storyteller.csv into a data frame

Download and import the file into a R – data frame for further processing.

A screenshot of a cell phone

Description automatically generated

1. Pre-Processing

This step involves, analyzing the dataset for its structure, datatype, missing values, duplicate check and removal of unwanted columns.

3.1 Analyze Structure

This dataset contains 30 observations(rows) and 8 variables (columns). “School” being the only factor variable, while the remaining variables are numeric. In addition, it also suggests that column names need to be renamed properly without any special characters or numeric values.

A close up of a person

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

3.2 Analyze Data Spread

This step summarizes the data spread/distribution of the variables. It’s obvious that “Very.Ahead..5” has all zero values and can be ignored from this dataset as it doesn’t add or simply states that “No Students from any school are ahead in their math course”.

A picture containing ball

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Histogram, a visual interpretation of the data spread confirms the same.

A screenshot of a cell phone

Description automatically generated

A screenshot of a social media post

Description automatically generated

3.3 Data Wrangling

We may still need to perform this step, though the size of the dataset is very small. Following steps were performed,

* Re-Name columns – Special character and Numeric values are removed.

School 🡪 School (no change)

Section🡪 Section (no change)

Very.Ahead..5 🡪 VeryAhead

Middling..0 🡪 Middling

More.Behind..6.10 🡪 MoreBehind

Very.Behind..11🡪 VeryBehind

Completed 🡪 Completed (no change)

* Check for missing/incomplete values – No missing or incomplete values [Nulls, Blank values]
* Find NAs – No NAs found
* Check for duplicate records – 0 duplicate records were found
* Removal of unwanted or columns with all 0s – [VeryAhead] column has been ignored from the dataset as it has all 0s.

A screenshot of a social media post

Description automatically generated

Let’s, preview the dataset with proper column names below.

A screenshot of a social media post

Description automatically generated

A picture containing implement, stationary, pencil, game

Description automatically generated

1. Data Aggregation

Based on the current dataset, it is observed that

Schools: There are 5 schools in total; Namely, A, B, C, D & E

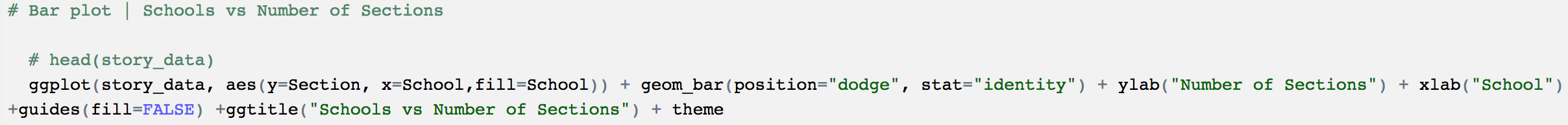
Section: Each school has varying number of sections

A close up of a clock

Description automatically generated

Grades/Categories: 5 meaningful (non-zero) categories; Remember, one column has been excluded as it had all 0s.

* Middling
* Behind
* MoreBehind
* VeryBehind
* Completed



A screenshot of a cell phone

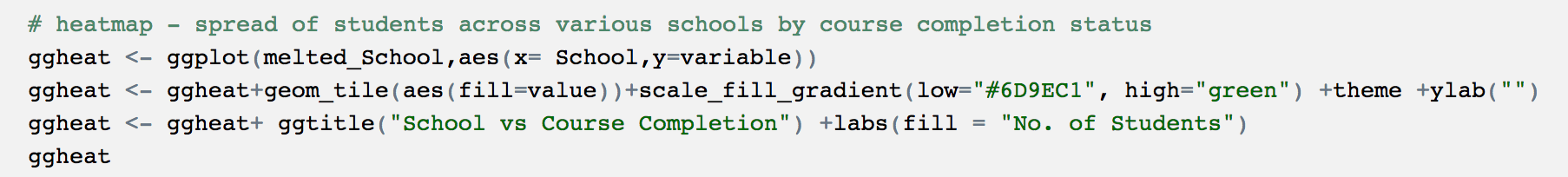
Description automatically generated

However, it doesn’t give lot of information as the number of students and their category are spread across and aggregated at School 🡪 Section level. In order to gain more insights, data has to be aggregated at School 🡪 Students and other numerical columns need to be melted or converted to rows.

A screenshot of a cell phone

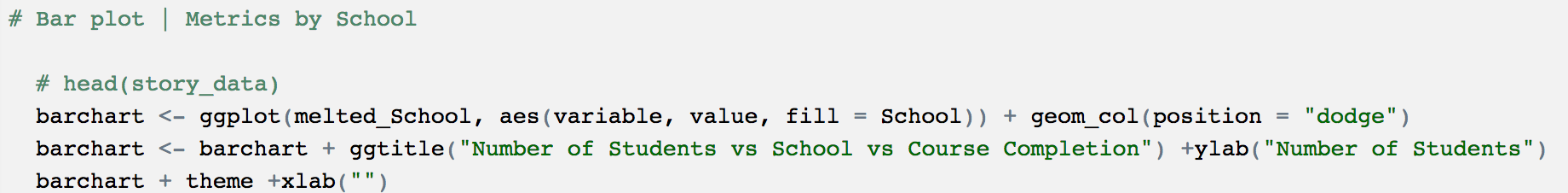
Description automatically generated

Now, let’s visualize the above aggregated data using a heatmap and bar chart.



A picture containing screenshot

Description automatically generated



A screenshot of a cell phone

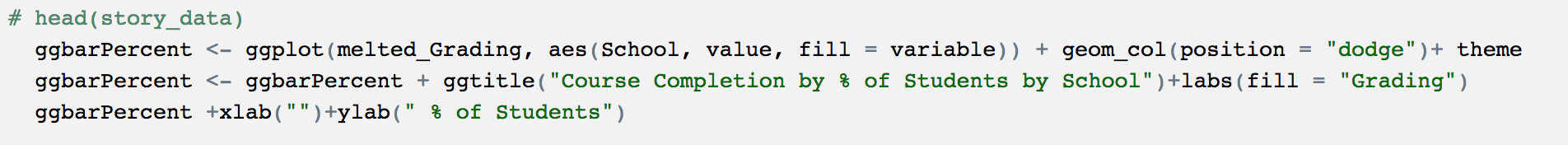
Description automatically generated

Number of students across schools are still not comparable because of varying number of sections and completion status across each school. So, let’s take % distribution of students across schools.

A screenshot of a social media post

Description automatically generated

Let’s visualize the aggregated value now.



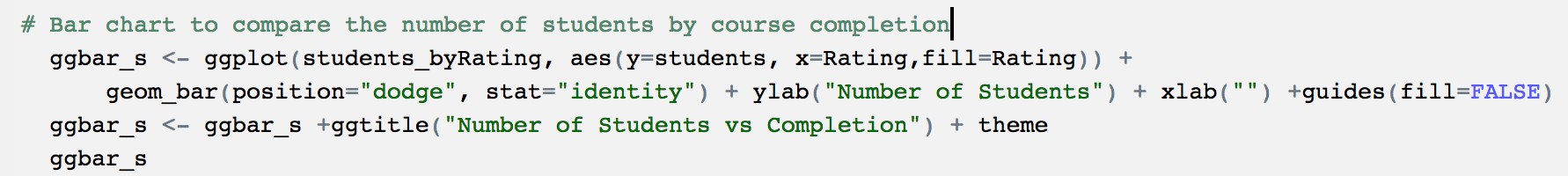
A screenshot of a cell phone

Description automatically generated

In addition, lets aggregate all students by school.

A screenshot of a cell phone

Description automatically generated



A screenshot of a cell phone

Description automatically generated

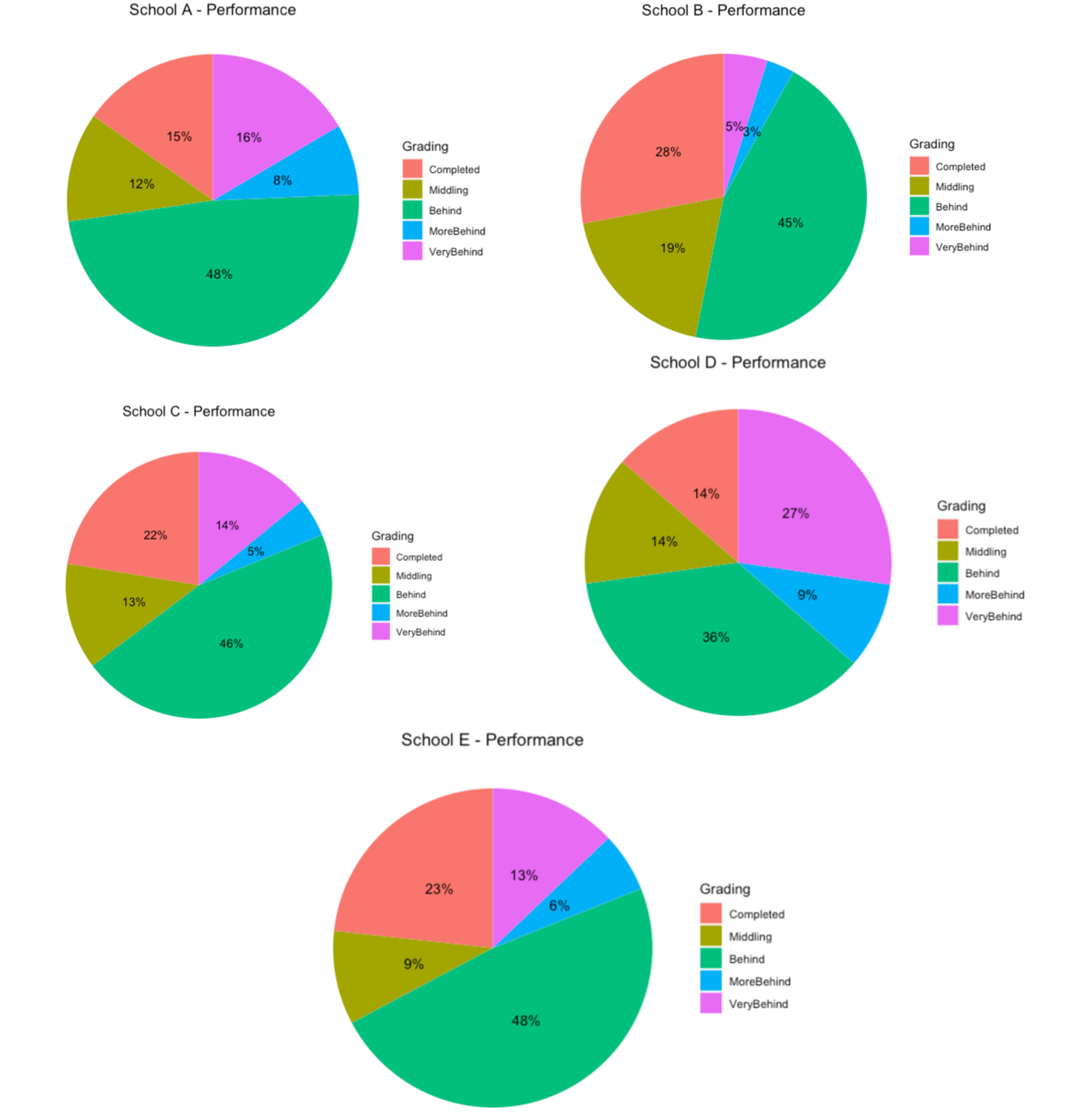
## **Summarization:**

Let’s, run through few scripts for the purpose of summarizing.

* Pic Chart to display the performance of each school by % of students enrolled into the math class across different sections and their completion status

A screenshot of a social media post

Description automatically generated



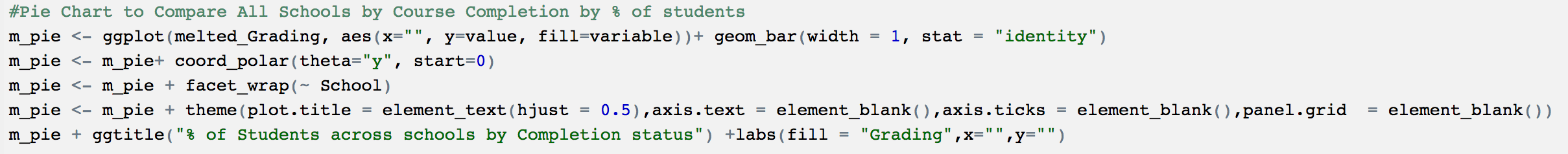
* Table to display total number of students at each school by number of sections offered

A screenshot of a cell phone

Description automatically generated

A picture containing beach

Description automatically generated



A close up of graphics

Description automatically generated

* Table to display Average number of students at each school by course completion

A picture containing table

Description automatically generated

A screenshot of a cell phone

Description automatically generated

* Correlation between variables

A screenshot of a cell phone

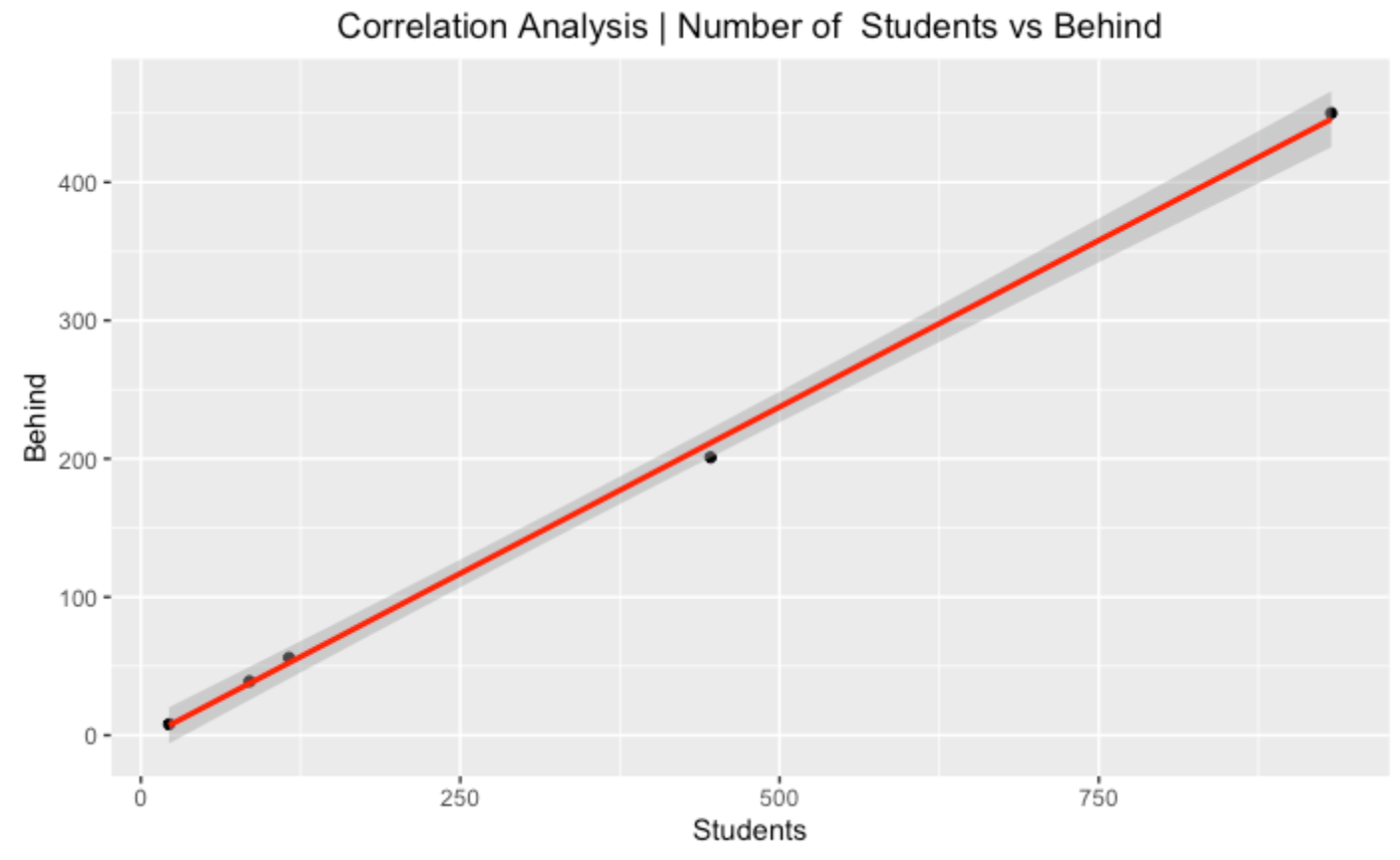
Description automatically generated

A picture containing game

Description automatically generated

A picture containing game

Description automatically generated



A picture containing game

Description automatically generated

## **Conclusion:**

After analyzing the dataset for the 5 schools, below are observations that could be inferred from the dataset having only 30 rows and 8 columns. So, may be a larger dataset might have helped further on the analysis.

* 1601 Students enrolled on the math course during summer term across 5 schools on 30 sections.
* Schools A & B had to majority of students (~86%) across 25 sections; However, ~50% still behind on their course.
* School B has the highest completion rate (~28%) and least Very behind rate (~5%). Having more than 400 students, making it the strong contender for best school.
* Section D has the least number of students enrolled (22); However, performance wise School D rates the lowest with 27% students under Very behind category. With only 13% of its students completed the course. Which is lowest in term of % completion by schools.
* On an average all the schools having almost half if its students behind their schedule is alarming. May be the subject itself is harder for most or season (summer) could be another reason. Making it harder for students to balance schoolwork.
* None of the students were on the very Ahead list, confirming the above point.
* Correlation between various factors like Number of Students vs Completion/Behind shows a linear correlation; Suggesting as the number of students increase, the % of Completion/Behind also increases. However, Correlation between Number of Sections vs Completion/Behind doesn’t necessarily comply. There are few exceptions and possible outliers. May be more data could help justify these.