

**TASK**

**Exploratory Data Analysis on the 2016 Matric Schools Report Data Set**

[](http://www.hyperiondev.com/portal/)

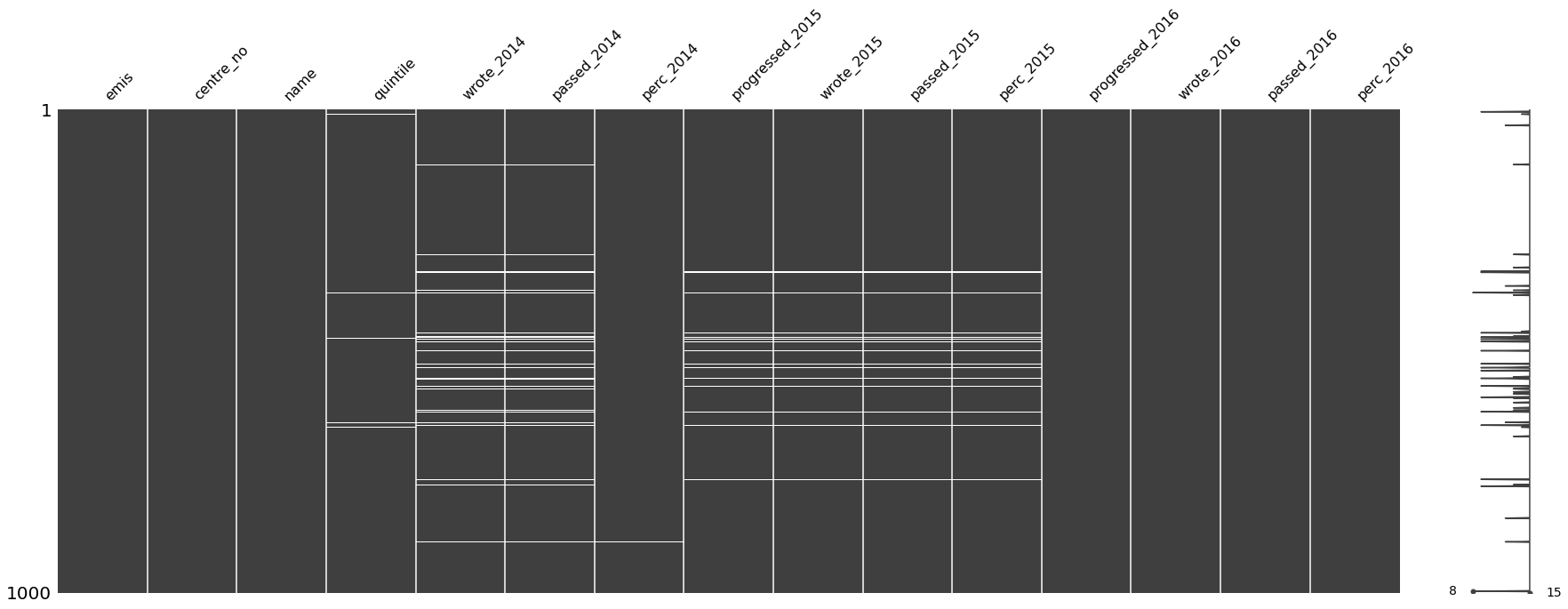
**Introduction**

The data set used in this report is the 2016 Matric Schools Report that shows the amount of learners that wrote, passed and was progressed, as well as the percentage for years 2014, 2015 and 2016. The data set consists of 1000 rows and 15 columns. Also included in the data set is the allocation of the schools in the various quintiles. Below is as sample of said data set:

|  | **emis** | **centre\_no** | **name** | **quintile** | **wrote\_2014** | **passed\_2014** | **perc\_2014** | **progressed\_2015** | **wrote\_2015** | **passed\_2015** | **perc\_2015** | **progressed\_2016** | **wrote\_2016** | **passed\_2016** | **perc\_2016** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 200300003 | 4161001 | AMABELE SENIOR SECONDARY SCHOOL | 1.0 | 37.0 | 14.0 | 37.8 | 0.0 | 38.0 | 29.0 | 76.3 | 12 | 51 | 36 | 70.6 |
| **1** | 200300869 | 4161002 | BETHEL COLLEGE HIGH SCHOOL | 99.0 | 59.0 | 42.0 | 71.2 | 3.0 | 46.0 | 31.0 | 67.4 | 0 | 34 | 24 | 70.6 |
| **2** | 200300024 | 4161003 | BLYTHSWOOD INSTITUTION | 3.0 | 80.0 | 50.0 | 62.5 | 12.0 | 81.0 | 34.0 | 42.0 | 7 | 101 | 59 | 58.4 |
| **3** | 200300033 | 4161048 | BONGOLETHU SENIOR SECONDARY SCHOOL | 2.0 | 26.0 | 13.0 | 50.0 | 2.0 | 35.0 | 11.0 | 31.4 | 0 | 24 | 9 | 37.5 |
| **4** | 200300741 | 4161037 | BUTTERWORTH HIGH SCHOOL | 4.0 | 81.0 | 61.0 | 75.3 | 4.0 | 85.0 | 78.0 | 91.8 | 0 | 84 | 77 | 91.7 |

**DATA CLEANING**

I examined the data distribution of the data set using .descibe(). I used .isnull().sum() to calculate the number of missing value counts for each column. Additionally, I calculated the percentage of the missing data points and obtained a value of 1.19%. Consequently I concluded that the data is sufficient to work with. Below is a matix I pulled to better visualize the missing data points:



MISSING DATA

From my evaluation I determined that I have eight columns with missing data points:

Quintile – 9 missing data points

Wrote\_2014 – 44 missing data points

Passed\_2014 – 44 missing data points

Perc\_2014 – 2 missing data points

Progressed\_2015 – 20 missing data points

Wrote\_2015 – 20 missing data points

Passed\_2015 – 20 missing data points

Perc\_2015 – 20 missing data points

Quintile missing data points can be ascertained by looking at the area where the schools are located as the quintile number is depended on the socio-economic area of the school.

The other seven columns can only be completed by the schools themselves, as they would have the data on record.

DATA STORIES AND VISUALIZATIONS

I grouped the data by name and then obtained the average using .mean() and pulled the following table:

| **name** | **emis** | **centre\_no** | **quintile** | **wrote\_2014** | **passed\_2014** | **perc\_2014** | **progressed\_2015** | **wrote\_2015** | **passed\_2015** | **perc\_ 2015** | **progressed\_2016** | **wrote\_2016** | **passed\_2016** | **perc\_2016** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **a d tshayingca secondary school** | 200500039.0 | 4051033.0 | 3.0 | NaN | NaN | 0.0 | NaN | NaN | NaN | NaN | 0.0 | 209.0 | 126.0 | 60.0 |
| **a m sityana high school** | 200200001.0 | 4191001.0 | 2.0 | 42.0 | 25.0 | 60.0 | 3.0 | 41.0 | 16.0 | 39.0 | 19.0 | 45.0 | 18.0 | 40.0 |
| **a.m.zantsi senior secondary school** | 200600003.0 | 4141001.0 | 1.0 | 55.0 | 54.0 | 98.0 | 4.0 | 64.0 | 53.0 | 83.0 | 26.0 | 58.0 | 49.0 | 84.0 |
| **a.v platjie senior secondary school** | 200404017.0 | 4101001.0 | 3.0 | 40.0 | 28.0 | 70.0 | 13.0 | 52.0 | 26.0 | 50.0 | 0.0 | 21.0 | 14.0 | 67.0 |
| **abambo high school** | 200600004.0 | 4071001.0 | 2.0 | 104.0 | 39.0 | 38.0 | 1.0 | 119.0 | 51.0 | 43.0 | 0.0 | 85.0 | 44.0 | 52.0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **zwelivumile senior secondary school** | 200350966.0 | 4161048.0 | 2.0 | 76.0 | 49.0 | 62.0 | 28.0 | 124.0 | 72.0 | 55.0 | 42.0 | 120.0 | 64.0 | 53.0 |
| **zweliwelile senior secondary school** | 200300709.0 | 4161036.0 | 3.0 | 104.0 | 42.0 | 40.0 | 1.0 | 118.0 | 28.0 | 24.0 | 0.0 | 95.0 | 51.0 | 54.0 |
| **zwelixolile senior secondary school** | 200300711.0 | 4141026.0 | 2.0 | 18.0 | 9.0 | 50.0 | 0.0 | 12.0 | 9.0 | 75.0 | 0.0 | 20.0 | 17.0 | 85.0 |
| **zweliyandila high school** | 200200963.0 | 4181087.0 | 2.0 | 79.0 | 44.0 | 56.0 | 3.0 | 78.0 | 10.0 | 13.0 | 7.0 | 94.0 | 12.0 | 13.0 |
| **zwelodumo senior secondary school** | 200401225.0 | 4101056.0 | 3.0 | 112.0 | 58.0 | 52.0 | 58.0 | 165.0 | 56.0 | 34.0 | 0.0 | 76.0 | 47.0 | 62.0 |

I looked at the amount of schools per quintile and obtained the following result:

3.0 390

1.0 220

2.0 196

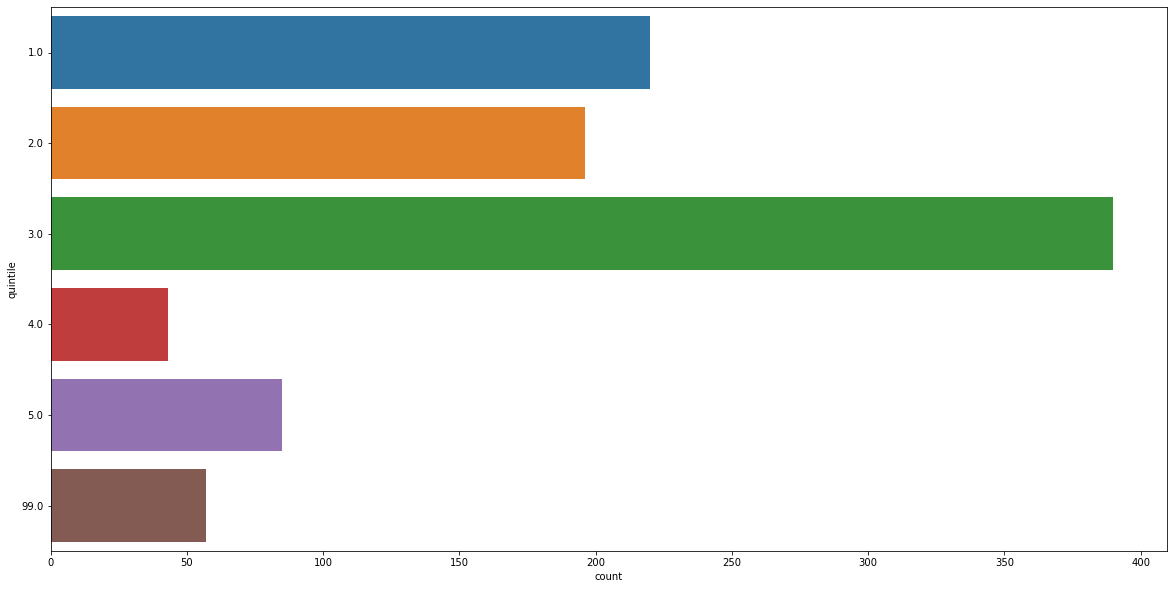
5.0 85

99.0 57

4.0 43

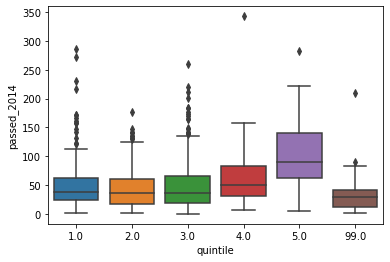
Name: quintile, dtype: int64

I used a count plot to present the above information visually:



After realizing that the CAPS curriculum was introduced to Grade 12 in 2014, I decided to compare the pass rate and progressed rate of 2014 to that of 2016 (two years down the line). Further, I examined whether or not the quintile rating of a school played a role in this result. As seen from above results, there are schools that fall into a quintile 99. Upon further examination of these schools, I observed that they were either Special Needs Schools, Christian Academies or Business Schools.

Then I compared the amount of students that passed in the year 2014 to the quintile rating and compiled the following graph using a boxplot:



As observed in the visualization above, a school with a higher quintile rating had a higher pass rate for the year 2014.

I decided to pull a table that was grouped by quintile, name and wrote\_2014, as can be seen below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **quintile** | **name** | **wrote\_2014** |
| **0** | 1.0 | a.m.zantsi senior secondary school | 55.0 |
| **1** | 1.0 | amabele senior secondary school | 37.0 |
| **2** | 1.0 | baleni senior secondary school | 121.0 |
| **3** | 1.0 | bambilanga senior secondary school | 31.0 |
| **4** | 1.0 | barnard molokoane ss | 95.0 |
| **...** | ... | ... | ... |
| **981** | 99.0 | vaal christian ci/s | 58.0 |
| **982** | 99.0 | vukuzenzele special school | 3.0 |
| **983** | 99.0 | zamokuhle special school | 6.0 |
| **984** | 99.0 | zanokhanyo baptist comprehensive sch | NaN |
| **985** | 99.0 | zenith ci/s | 40.0 |

986 rows × 3 columns

I did the same for passed\_2014 and obtained the following result:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **quintile** | **name** | **passed\_2014** |
| **0** | 1.0 | a.m.zantsi senior secondary school | 54.0 |
| **1** | 1.0 | amabele senior secondary school | 14.0 |
| **2** | 1.0 | baleni senior secondary school | 87.0 |
| **3** | 1.0 | bambilanga senior secondary school | 30.0 |
| **4** | 1.0 | barnard molokoane ss | 82.0 |
| **...** | ... | ... | ... |
| **981** | 99.0 | vaal christian ci/s | 41.0 |
| **982** | 99.0 | vukuzenzele special school | 3.0 |
| **983** | 99.0 | zamokuhle special school | 6.0 |
| **984** | 99.0 | zanokhanyo baptist comprehensive sch | NaN |
| **985** | 99.0 | zenith ci/s | 34.0 |

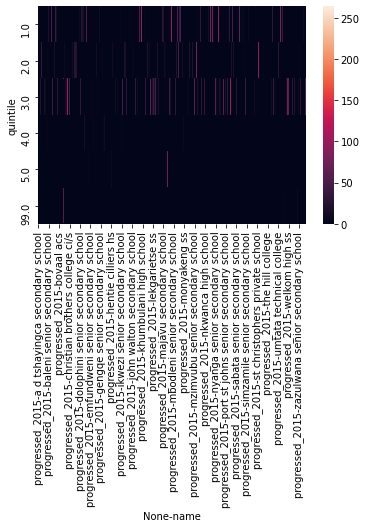
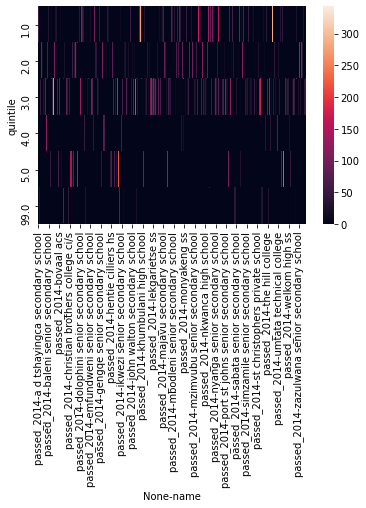
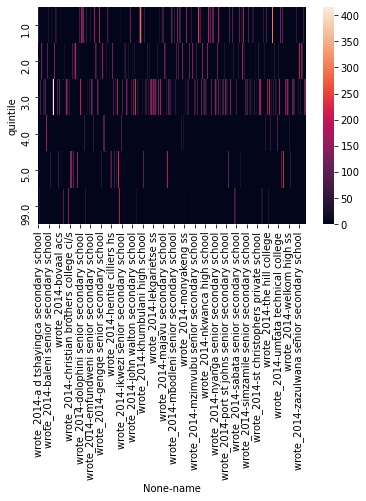
986 rows × 3 columns

I also repeated it for progressed\_2015:

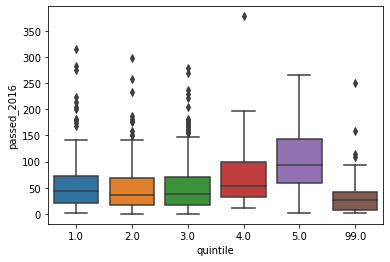
|  | **quintile** | **name** | **progressed\_2015** |
| --- | --- | --- | --- |
| **0** | 1.0 | a.m.zantsi senior secondary school | 4.0 |
| **1** | 1.0 | amabele senior secondary school | 0.0 |
| **2** | 1.0 | baleni senior secondary school | 16.0 |
| **3** | 1.0 | bambilanga senior secondary school | 19.0 |
| **4** | 1.0 | barnard molokoane ss | 57.0 |
| **...** | ... | ... | ... |
| **981** | 99.0 | vaal christian ci/s | 0.0 |
| **982** | 99.0 | vukuzenzele special school | 4.0 |
| **983** | 99.0 | zamokuhle special school | 0.0 |
| **984** | 99.0 | zanokhanyo baptist comprehensive sch | 8.0 |
| **985** | 99.0 | zenith ci/s | 3.0 |

986 rows × 3 columns

I visualized the information in the tables above using heat maps, as seen below:



I repeated the above assessments for the year 2016 as well. The first visualization is a box plot of the amount of students that passed per quintile for 2016:

****

As observed for the previous years, the schools with a higher quintile rating had a better pass rate for the year 2016.

I pulled a table that was grouped by quintile, name and wrote\_2016:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **quintile** | **name** | **wrote\_2016** |
| **0** | 1.0 | a.m.zantsi senior secondary school | 58.0 |
| **1** | 1.0 | amabele senior secondary school | 51.0 |
| **2** | 1.0 | baleni senior secondary school | 120.0 |
| **3** | 1.0 | bambilanga senior secondary school | 48.0 |
| **4** | 1.0 | barnard molokoane ss | 131.0 |
| **...** | ... | ... | ... |
| **981** | 99.0 | vaal christian ci/s | 47.0 |
| **982** | 99.0 | vukuzenzele special school | 9.0 |
| **983** | 99.0 | zamokuhle special school | 6.0 |
| **984** | 99.0 | zanokhanyo baptist comprehensive sch | 11.0 |
| **985** | 99.0 | zenith ci/s | 44.0 |

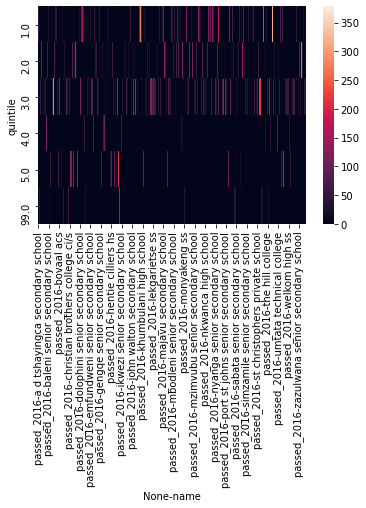
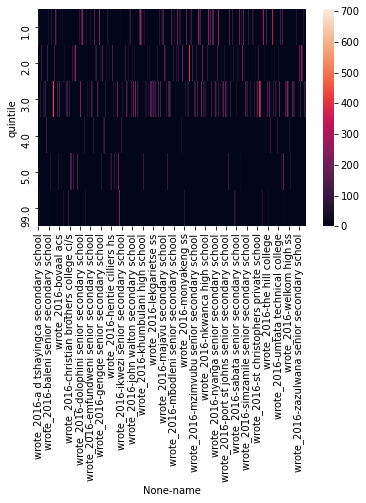
986 rows × 3 columns

I did the same for passed\_2016 and obtained the following result:

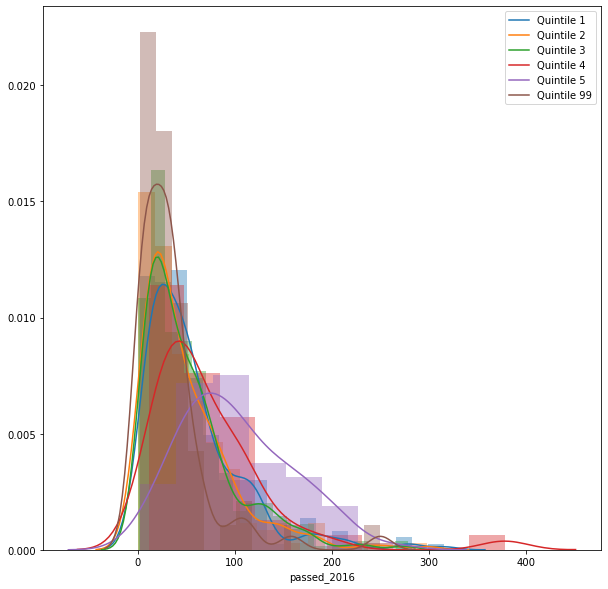
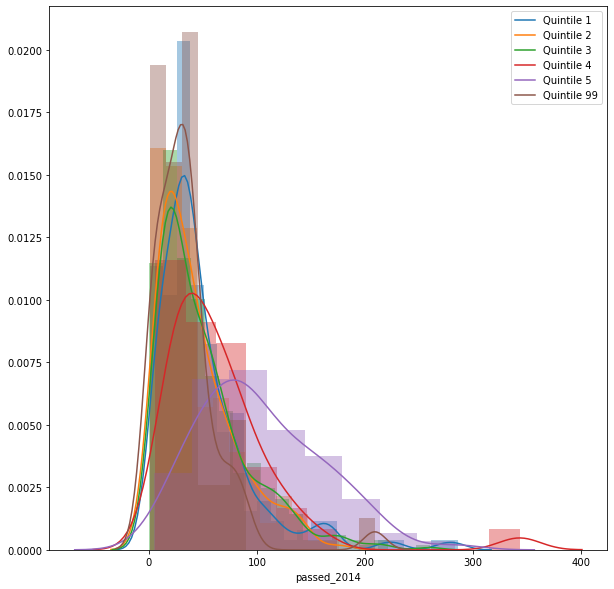
|  |  |  |  |
| --- | --- | --- | --- |
|  | **quintile** | **name** | **passed\_2016** |
| **0** | 1.0 | a.m.zantsi senior secondary school | 49.0 |
| **1** | 1.0 | amabele senior secondary school | 36.0 |
| **2** | 1.0 | baleni senior secondary school | 51.0 |
| **3** | 1.0 | bambilanga senior secondary school | 25.0 |
| **4** | 1.0 | barnard molokoane ss | 128.0 |
| **...** | ... | ... | ... |
| **981** | 99.0 | vaal christian ci/s | 46.0 |
| **982** | 99.0 | vukuzenzele special school | 4.0 |
| **983** | 99.0 | zamokuhle special school | 3.0 |
| **984** | 99.0 | zanokhanyo baptist comprehensive sch | 8.0 |
| **985** | 99.0 | zenith ci/s | 42.0 |

986 rows × 3 columns

I used the information above to compile the following heat maps:



I pulled two last visualizations to show the pass rate of 2014 and 2016 across the quintiles. One can observe that over the course of two years, the pass rate has increased, as the confidence with the new curriculum increased.



**THIS REPORT WAS WRITTEN BY : Ryno Viljoen**

