

**TASK**

**Exploratory Data Analysis on the 2016 Matric Year Results of each School Data Set**

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

**Introduction**

Summary of the data set

The data set inspects performances of both the quintiles and the high schools in regards to the tests taken within the periods of 2014 up to 2016

**DATA CLEANING**

# SUMMARY OF THE METHODS AND VISUALIZATIONS DONE DURING DATA CLEANING

Methods:

* Read the dataset as pandas csv
* Identified the columns
* Identified the missing values
* Visualized the missing values
* Imputated the missing values with bfill.

Visualizations:

* Bar graphs
* Dual bar graphs
* Pie charts

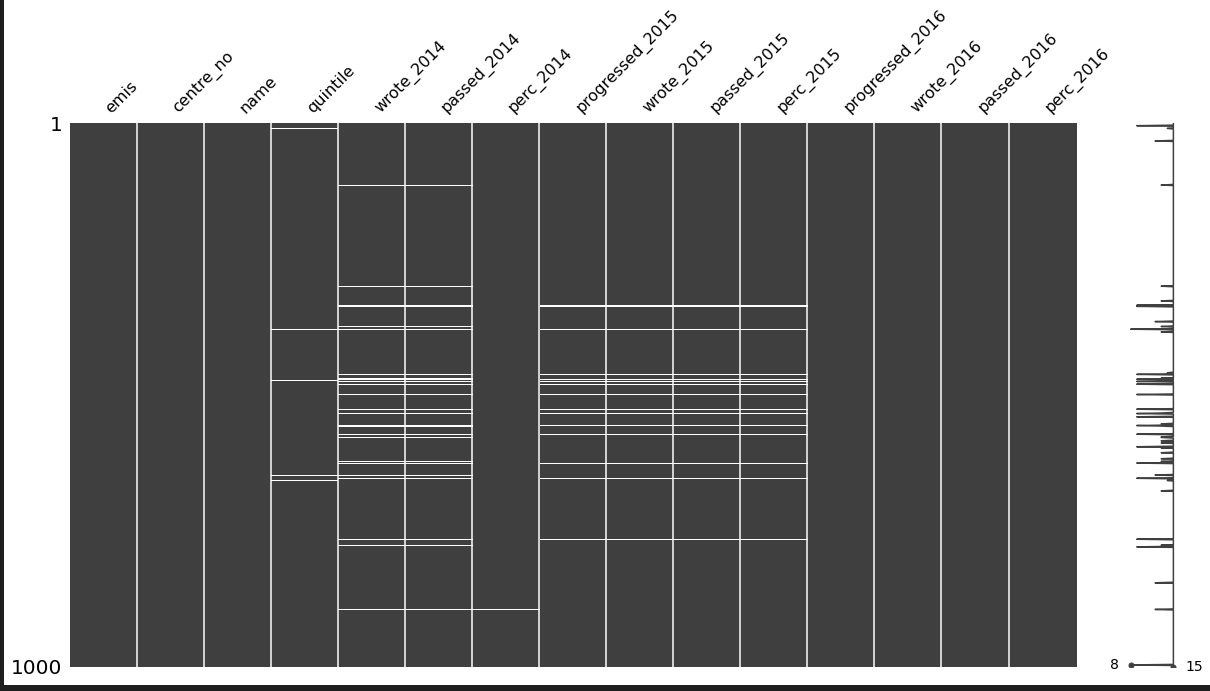
The EDA that was performed on the performance of tests for each school in the 2016 matric year were the following:

* Inspecting the data
  + How many schools belong to each quintile.
  + Quintiles for each year that have a 100% pass rate.
  + Top 5 Schools who wrote the most for each year.
  + Top 5 Schools who passed the most for each year.
  + top 5 schools with the highest percentage.
  + How many schools had a 100% pass rate.
* Visualizing the data
  + How many schools belong to each quintile.
  + Performance of schools for each year
  + Top ten Percentage of pass rate for each schools in each year in order from 2016 results compared to previous years
  + Most performing quintile for each year.

MISSING DATA

# ANY MISSING DATA? HOW DID YOU HANDLE IT?

Yes, there was missing data for the columns shown in the display below



I handled it by performing a dropna() bmethod on columns that had missing data. I decided it would be much simpler to replace empty values using cells before them and it did resolve with have no missing rows in each columns.

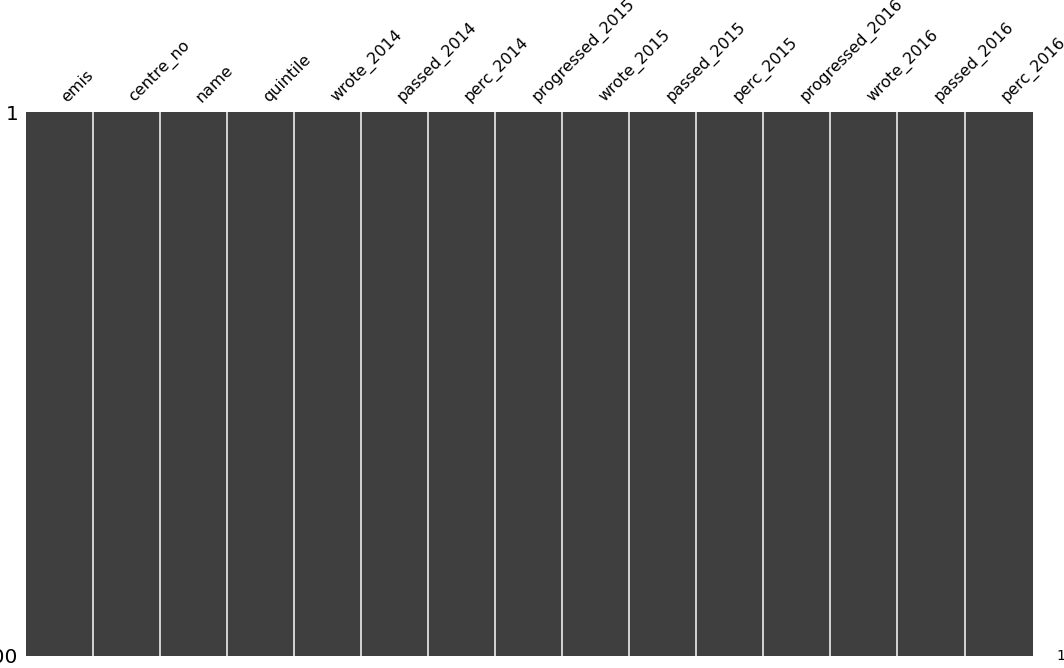
e.g below:

df = df.fillna(method='bfill',axis=0).fillna(0)

The shape is as follows:

(943, 15)

Display of the imputation below:



I then performed a filter to remove all the rows from a column value of quintile 99. AS this value seems to be odd one out because of how high the number is.

df = df[df['quintile'] != 99.0]

The shape is as follows:

(892, 15)

DATA STORIES AND VISUALIZATIONS

# THIS IS THE BULK OF THIS PROJECT. EXTRACT STORIES AND ASSUMPTIONS BASED ON VISUALIZATIONS OF THE DATA  
  
***Quintile EDA***

How many schools belong to each Quintile?

quintile

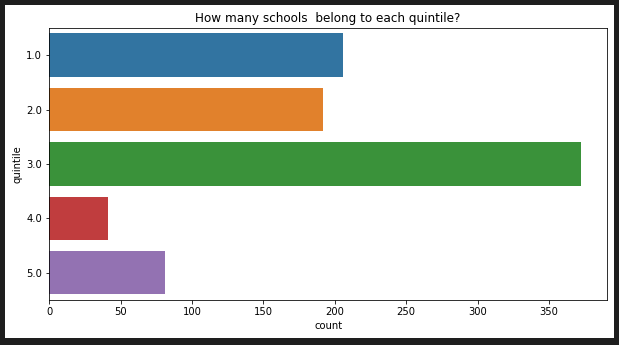
3.0 372

1.0 206

2.0 192

5.0 81

4.0 41



As you can see from both the data and the chart, quintile 3 seems to have the bulk of the schools as compared to other quintiles.

Also, I did more investigating to see which quintile seems to contain 100% pass rates for all 3 years of 2014,2015 and 2016.

100% pass rate for each quintile in the year 2014:

**quintile perc\_2014**

**5.0 100.0 22**

**4.0 100.0 7**

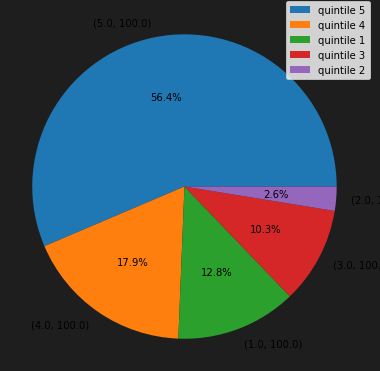
**1.0 100.0 5**

**3.0 100.0 4**

**2.0 100.0 1**

**dtype: int64**

Chart for 100% pass rate for each quintile in the year 2014:



100% pass rate for each quintile in the year 2015:

**quintile perc\_2015**

**5.0 100.0 22**

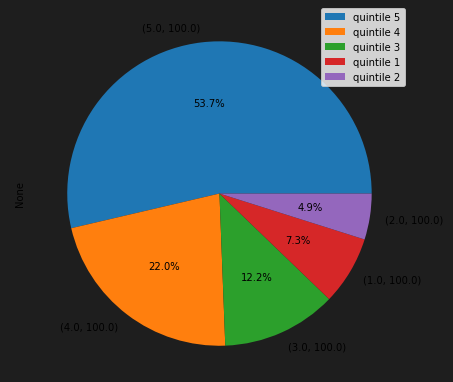
**4.0 100.0 9**

**3.0 100.0 5**

**1.0 100.0 3**

**2.0 100.0 2**

**dtype: int64**

Chart for 100% pass rate for each quintile in the year 2015: 

100% pass rate for each quintile in the year 2016:

**quintile perc\_2016**

**5.0 100.0 28**

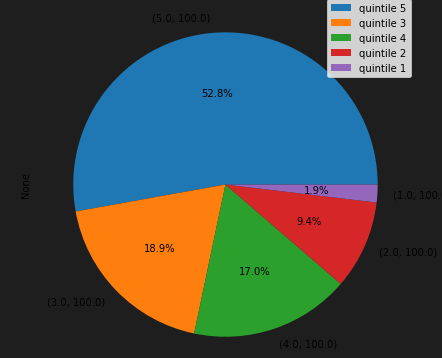
**3.0 100.0 10**

**4.0 100.0 9**

**2.0 100.0 5**

**1.0 100.0 1**

**dtype: int64**

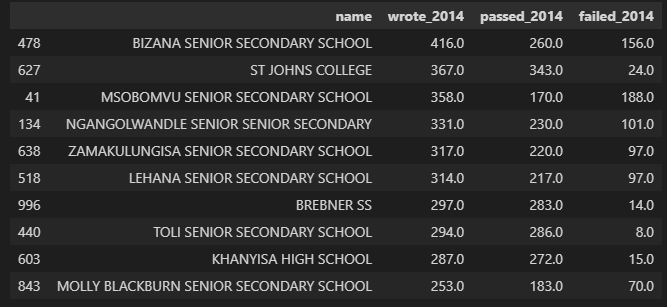
Chart for 100% pass rate for each quintile in the year 2016: 

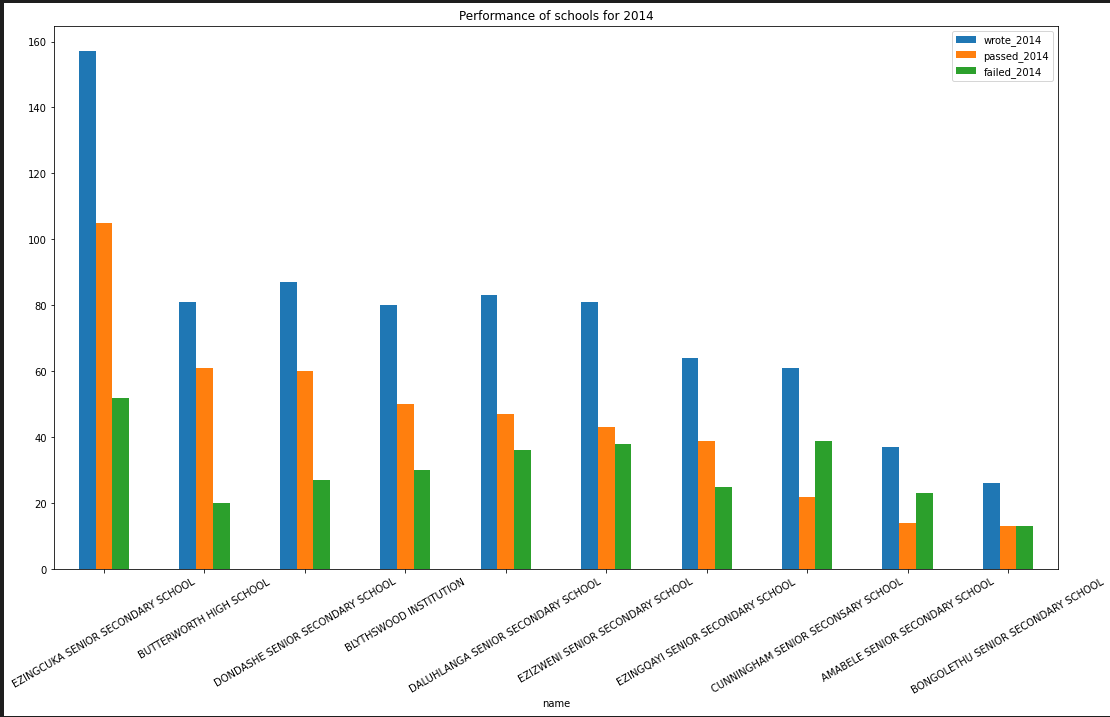
As you can see from the data above, quintile 5 seems to be the one with the most 100% pass rates for all three years. It seems to have had the most. Also there was an increase for quintile 5 in regards to having more 100% pass rates.

***Performance of Schools of each Year.***

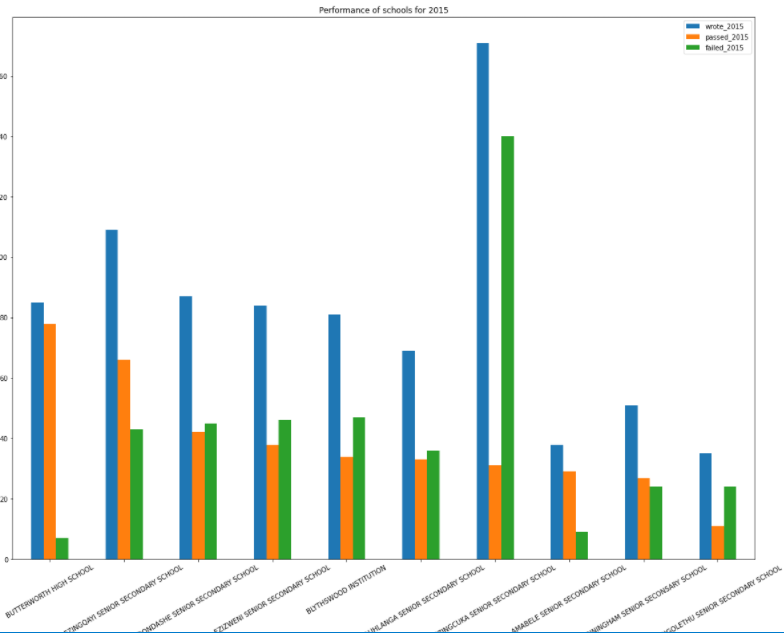
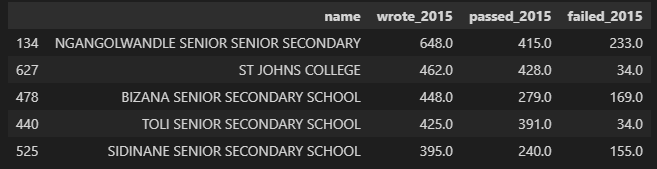
For this, I was analysing the data to see the top ten schools that performed that passed as compared to the number of written and failed for each year…

Top ten schools that passed based on the year 2014:

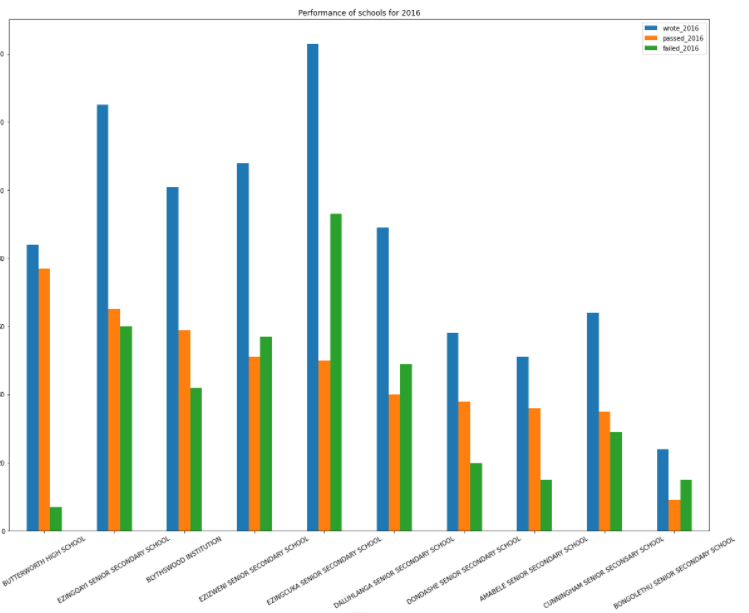
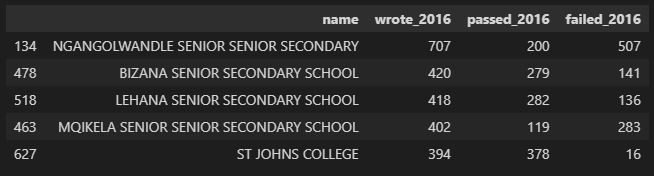




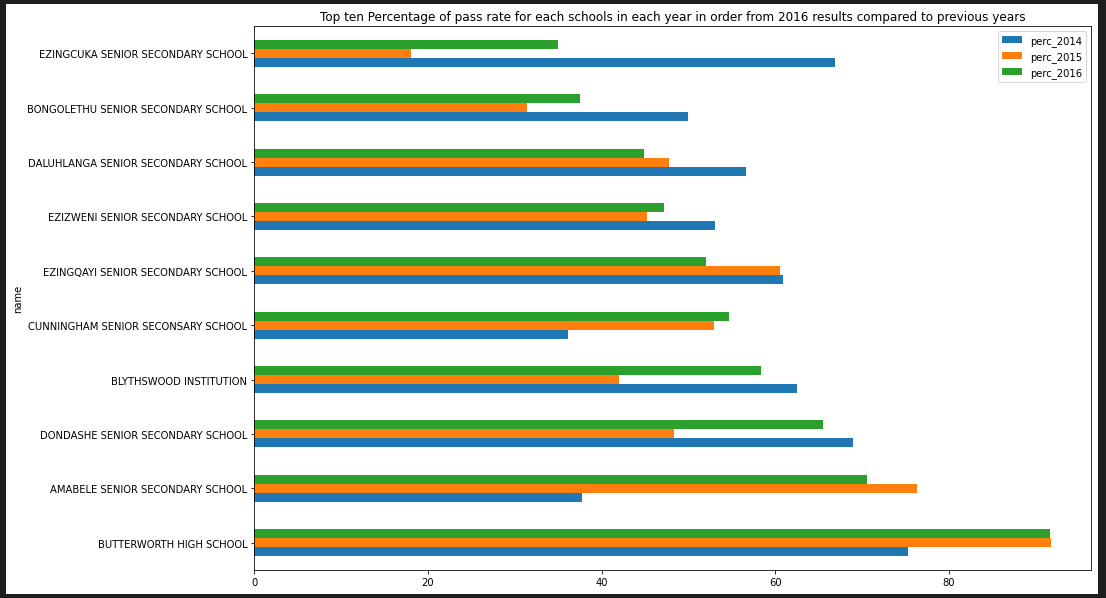
Top 5 schools that passed based on the year 2015:



Top 5 schools that passed based on the year 2016:



Las thing I wanted to find out was the top ten percentage of pass rate for each schools in each year in order from 2016 results compared to previous years. Only Created a chart and I see that Butterworth Highschool seems to have had some of the best performances in regards to exams and all the previous data that was presented at the top.



**THIS REPORT WAS WRITTEN BY : Naledi Motau**

