**TEAM ECD\_contest\_scriots\_5**

**1.0 Introduction**

Every child has a right to pass through a carefully crafted educational foundation. In fact, to attain the kinds of educational outcomes we are hoping for, it is important to help more children thrive before entering school by creating the correct educational foundation. Data Drive 2030 found, through a study conducted on more than 5000 young children aged between 50 and 59 months, that there are many concerning trends in the early child development data, one of which is socioeconomic. This is because the cost of education causes children from low-income families to perform worse than average and have lower Early Learning Outcomes Measures (ELOM score). However quite a few children coming from the poorest households (positive outliers) were doing quite well, but it is not enough to be satisfied with such number as only about 44.7% of learners are on track, 27% are fallen behind and 28.3% are fallen far behind. This analysis intends to indentify behaviors, environmental characteristics and combination thereof associated with ‘positive deviants’ with a view to understand why some children are outperforming their peers in spite facing the same challenging circumstances. The goal is to use this information to design aid programs to help children who are significantly behind to catch up.

The variables that will be taken into considerations include: Age, gender, concentration, diligence, grant, attention, interest, district, province, enumerator, language of the child, water, closing time, opening time, number of school days, district and facility. This variable will be discussed in details in the result section.

**2. 0 Methods and Methodology**

To achieve the objectives of this analysis, there is an attempt to provide answers to the following research questions; what features/factors or both are associated with outliers and how can these outliers be increased. Hence, the analysis is broken down into three sub stages which includes;

1. Data Cleaning to handle empty entries
2. The definition and tagging of each child observation as either positive deviant or non-deviant.
3. the analysis of the positive deviants to figure out features and factors which contributes to the increase of positive deviants

**I - Data Cleaning and Engineering**

This project uses Pandas (python library for data science) in Jupiter notebook to read data, clean and NumPy to perform analysis. Most columns in the dataset have nulls with some columns totally empty. Two operations were conducted to deal with nulls:

a). 60% thresh-hold was used to remove nulls in the side of the columns. Since this is an analysis task, 60% of the data can be used to represent the remaining 40%

b). 60% thresh-hold was also used to remove nulls on the side of the rows. This involved transposing the Data Frame and deleting entries with 60% nulls.

Since no machine learning model was required in our analysis, we opt to use the data with the nulls as it is without any form of imputation. This was also supported with the fact that we expect the outliers to be small in the dataset hence using measures of central tendency such as mean, mode and median may not favor the minority in the dataset hence:

I). We may miss the insights we are craving for seeing

ii). We may get a wrong insight altogether.

**II - Tagging of Each Child:**

This stage of the work leverages heavily on the outlier definition provided by data Drive 2030 and the grouping of children into socio-Economic groups based on the amount paid for their Early learning program per month. Positive Deviants, as defined, are learners who pay same amount of fees but perform two standard deviations above the expected score of learners in that group. The learners are also categorized into the below groups.

* Category 1 Rands 0 – 110
* Category 2 Rands 111 – 290
* Category 3 Rands 291 – 750
* Category 4 Rands 751 – 1750
* Category 5 Rands 1751 – upwards.

For ease of analysis, the dataset was split into first three focus categories and others left out. Then the outlier metric for each of the group was calculated using the above stated calculation. Using results from this calculation, each child observation was then labeled as positive deviant or non-deviant.

**III Analysis of the positive deviants**

Finally, Univariate and Multivariate analysis was performed on the features of interest and the observations recorded under the results section. On the top was the Age variable which played a significant role in determining the outliers. These features were looked at in several domains.

i). The child: These are features that are controlled by the child for example child age, gender, interest, concentration

ii). The learning environment: These include the presence of water, toilet etc

iii). Government aspect: These include the district, province etc

iv) Enumerator

**3.0 Results**

Age being one of the variables of interest had great impact on the outliers, most outliers in all three groups had an age above 56 months. In group one 67.5%, in group two 76.2 % and in group three 75.0%. Age below 56 months also existed but in smaller number.

Gender being another variable was dominated by females: 63.5% female in group one, 61.9% female in group two, 61.9.0% female in group three. This means that the female gender dominated

When it comes to grant, most outliers in group one received grants and the number reduced while moving to group three where the number who received grants had almost equal amount to those who dint.

Attention, concentration, diligence, interest was highly corelated. All the outliers in both groups were positive in the above three field. We expect children who observe these three things to do well in class.

All province in all groups had almost equal mean elom score. Western province had a higher percentage of learners (in comparison to their total number of learners in that province) as being positive deviance in group one followed by Mpumalanga and Gauteng. Eastern cape, Mpumalanga and Gauteng also topped in group two while northern cape, free state and Limpopo toping in group three.

In relation to the total number of outliers in each group: Eastern cape topped group one, Mpumalanga topped group two, and western cape topped group three.

For District, DC32 – 25 outliers, DC12 – 23 outliers, DC45 – 19 outliers topped group one. DC19 – 6 OUTLIERS, DC30 – 4 OUTLIERS, DC31 – 4 OUTLIERS topped group two. In group three, DC2 topped with 5 outliers followed by DC31, DC1, CPT, DC38, DC42, DC12 with all 1 outlier.

Some enumerators were associated with high number of positive deviant than others. In group one [2683.0] topped with 24 – outliers, followed by [1394.0] with 14 outliers. In group two, [563.0] topped with 8 – outliers followed by [2762.0] and [2983.0] both with 4 – outliers. In group Three, [20009.0] had 4 outliers which constitute to 27% of the total outliers in group three.

There was a significant impact, the language of the child caused on the outliers: isiXhosa, isisZulu and Setswana were in the top in group one. isisZulu, Sesotho, and Setswana were in the top of group two and in group three, Afrikaans and isiXhosa topped.

In terms of water, learners who have tap water in the building have a greater chance of becoming positive deviant in all three groups. Over 1100 in group one, overs 20 learners in group two and 9 learners in group 3(this is 60% of the total learners in group three).

Opening time and closing times were also important in determining positive deviance. In group one, opening time of [0700 hrs.] and [0800 hrs.] were dominant. In group two and group three, [0700 hrs. and 0600 hrs.] were dominant.

On closing time, [1400 hrs., 1500 hrs. and 1300 hrs.] dominated in group one, [1600 hrs., 1500 hrs.] dominated in group two and [1700 hrs. and 1600 hrs.] dominated in group three.

Number of school days is also a better concern for determining positive deviance. In all the three groups. Positive device thrives where there are 5 schooling days only.

Districts with the following id [366.0, 163.0,441.0] had high number of outliers in group one, while in group two, [207.0,90.0,494.0] dominated. The district with id [629.0] dominated in group three accounting for 33.3% of the total positive deviance in group three.

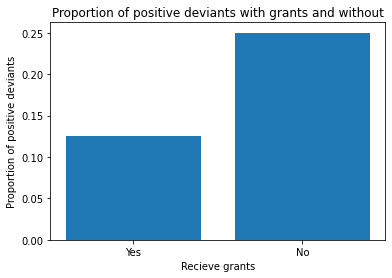
Our final variable of interest was the facility where these learners are. Some facilities had large numbers of positive deviance than others. In group one [1511.0] – 6 outliers and [1613.0, 285.0,615.0] – 4 outliers were dominant. In group two, [1943.0,1000.0115.0,515.0,1270.0] dominated with 2 outliers each. In group three, [463.0] dominated with 3 outliers accounting for 20% of the total outliers in group three.

The detailed output of the above results is present and well outlined in the notebook attached. Further discussions of these variables are in the discussion section.

**3.0 Discussion of Results**

1. **Grants:**

As against the hypothesis that the number of positive outliers should increase as the number of grants increase, our analysis indicates that the ratio of those with grants who becomes positive outliers are smaller compared to the ratio of those who have no grants who became positive outliers.



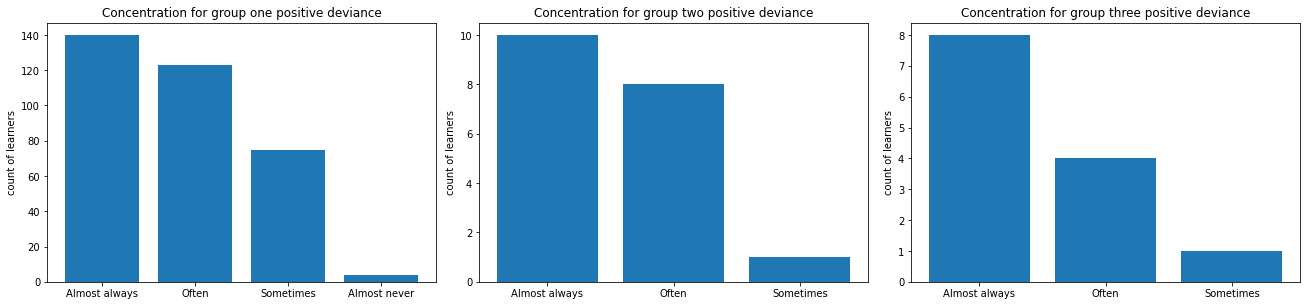
25 out of about every 100 children who did not receives grant becomes positive outlier against about 12 out of every 100 who receives.

There was also a significant trend in the number of those learners who receive grant to those who did not from one group to another. Most learners in group one thrives well as positive deviant in the presence of grants. Moving toward group three, these learns who need support start to reduce and in group three the ratio is almost 1:1.

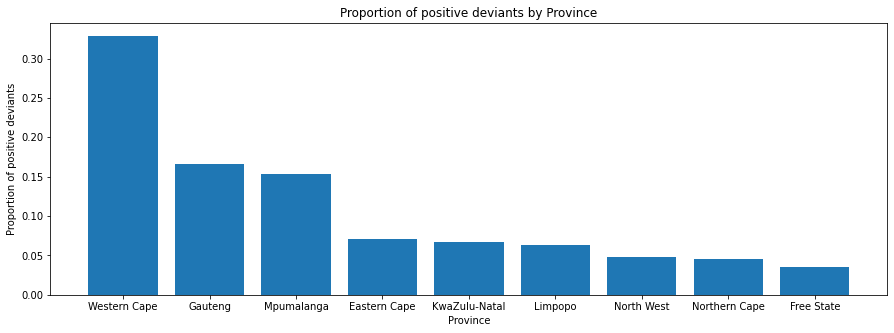
Good support inform of grants should be given to learners who are at poor financial background as this form part of their motivation to becoming positive deviant as seen in group one.

1. **Attentiveness concentration, diligence and interest**

Attentiveness concentration, diligence and interest plays huge role in the making of positive deviants. Those who almost always pay attention, concentration, diligence and interest have higher changes of being a positive outlier. These features are also linearly correlated, meaning, an attempt to boost one of these characteristics will automatically increase the others.

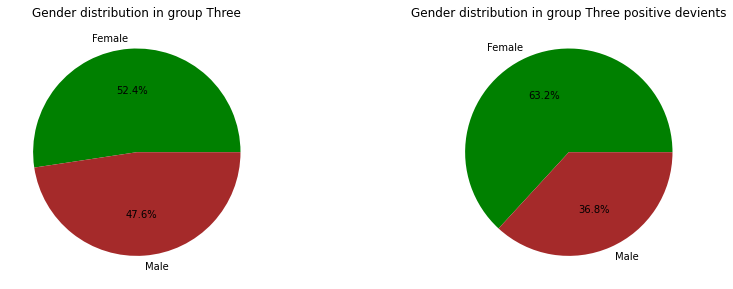


1. **Province**

Generally, Eastern cape, Mpumalanga, Western cape are outperforming learners from the other province. Although, to back the claim up, we need a look at the proportion of learners who made the positive outlier’s class from each of the province

**Iv. Gender:**

A female child is more likely to be a positive deviant than the male child in any of the three classes if gender alone is used as a metric. Considering the proportional distribution, the difference isn’t too large but still worthy of exploration.

****

**v). Enumerator:**

Some enumerators seam to have higher number of positive deviant learners and these may be a concern. Detailed information concerning these enumerators should be given to explain this phenomenon. Enumerator with Id 20009.0 can be traced as an outlier in this case. This person had 6 outliers in group one and led group three with 4 outliers.

**vi) Language of the child**

isiXhosa, isiZulu, Setswana, Sesotho, and Afrikaans can be considered as an outlier. Afrikaans for example had 6 outliers ingroup three which is 40% of the total learners in group three. Most learners having the above languages as their primary language thrive as outliers better than others.