African Leadership University

Artificial Intelligence Course

AI Project 2: Data Analytics Challenge

Authors:
Neba Roland
Okoye-Nobert C.
Natnael Alemayehu
Christopher Kuzagbe

Facilitator:
Kudakwashe Dandajena

March 7, 2021

Contents

1 Problem Identification and Value

1.1 Problem Statement/Overview

Revenue collection is very crucial in every economy as this is the only way in which the government is able to raise funds for its daily operations. One of the major ways in which governments are able to generate revenue is through tax collection. In Rwanda, some of the tax payment plans for businesses operating in the region are value added tax, pay as you earn, withholding tax, and consumption tax [1]. Some of these taxes can be easily collected from payrolls whereby a percentage of an individual's salary is withheld by the paying institution and payment is made directly to RRA within 15 days of the payment and some taxes collected at customs before goods are released [1]. According to a report on the international growth center, with the recent covid-19 situation, The national government faces significant losses in expected VAT and income tax revenue, while local government also faces a loss in revenue from various fees [2].

As part of the plan to digitize the current Revenue collection in Rwanda, this paper seeks to analyze Revenue collected in South Africa to come up with a proof of concept that will be used by the Rwanda revenue authority to collect its revenue more efficiently

2 Data Selection

The selected data to develop a POC is the national revenue collection data from 2006/07 to 2017/18 provided by the National Treasury of the Republic of South Africa[3]. This dataset was selected as it resembles very closely to the type of data RRA would be collecting and can make it easier to drive analogous comparisons.

3 Environment Setup

To perform data analytics and gain insights, the analytics environment was set up and configured properly to allow the necessary analytics operations to be performed.

3.1 Libraries Used

The libraries used to develop this POC are Pandas and Matplotlib

- Pandas was used to read data from the dataset which was in excel format; It was also used to create dataframes.
- Matplotlib was used to perform various graph visualizations like line graphs and bar graphs

3.2 Programming Language

The primary programming language used is Python 3

3.3 Execution Environment

The Execution environment used is Jupyter Notebook.

4 Exploring Data

Before analysis was made the data was explored to get a general overview of what it has and what it lacked. During this analysis, the shape, description, the tail and head of the data were checked.

- Screenshot Of final TextData.Head

```
final_taxData.head()
```

	New Main Category	2006/07	2007/08	2008/09	2009/10	2010/11	2011/12	2012/13	2013/14	2014/15	2015/16	2016/17	2017/18
0	Taxes on income and profits	2.79991e+08	3.32058e+08	3.83483e+08	3.59045e+08	3.79941e+08	4.26584e+08	4.57314e+08	5.07759e+08	5.6179e+08	6.06821e+08	6.60586e+08	7.39153e+08
1	Personal income tax	1.40578e+08	1.68774e+08	1.95115e+08	2.05145e+08	2.26925e+08	2.504e+08	2.75822e+08	3.09834e+08	3.5295e+08	3.88102e+08	4.2581e+08	4.82086e+08
2	Corporate income tax	1.18999e+08	1.4012e+08	1.65378e+08	1.34883e+08	1.32902e+08	1.51627e+08	1.59259e+08	1.77324e+08	1.84925e+08	1.91152e+08	2.0509e+08	2.18692e+08
3	Secondary tax on companies/dividend withholdin...	1.52914e+07	2.05854e+07	2.00176e+07	1.54678e+07	1.71782e+07	2.19654e+07	1.97387e+07	1.73088e+07	2.12473e+07	2.39342e+07	2.571e+07	3.42369e+07
5	Tax on retirement funds	3.19053e+06	285357	143251	42699.2	2772.1	6665.25	159.437	0	0	0	0	0

- Screenshot Of TextData.Shape

```
[ ] taxData.shape

    (99, 15)
```

- Screenshot Of TextData.Describe

```
[ ] taxData.describe

    <bound method NDFrame.describe of                                                Main Category  ...       2017/18
    0                        Taxes on income and profits  ...  7.391526e+08
    1                                                NaN  ...  4.820859e+08
    2                                                NaN  ...           NaN
    3                                                NaN  ...  2.186918e+08
    4                                                NaN  ...  3.423692e+07
    ..                                               ...  ...           ...
    94                          Sales of capital assets  ...  8.374200e+04
    95  Financial transactions in assets and liabiliti...  ...  1.728272e+07
    96                          TOTAL NON-TAX REVENUE *13  ...  3.287996e+07
    97                         TOTAL MAIN BUDGET REVENUE   ...  1.242417e+09
    98                     National Revenue Fund receipts  ...  1.457800e+07

    [99 rows x 15 columns]>
```

Moreover, the data was discovered to have empty rows, rows containing NaN values.

4.1 Cleaning Data

After the exploratory phase, the raw data had to be cleaned so that analysis and visualization takes place correctly. Cleaning the data entailed the following:

- Removing empty rows or rows with NaN value

Empty rows needed to be removed because they can be removed are not useful in any way to make analysis and can hinder/obscure the process as well.

- Combining two columns in two one

The tax data came with the various tax category and subcategory types and these categories were in separate columns which could have made analysis difficult. Hence, a merge was made between the rows of the respective columns to simplify the analysis process.

- Screenshot of cleaned_data

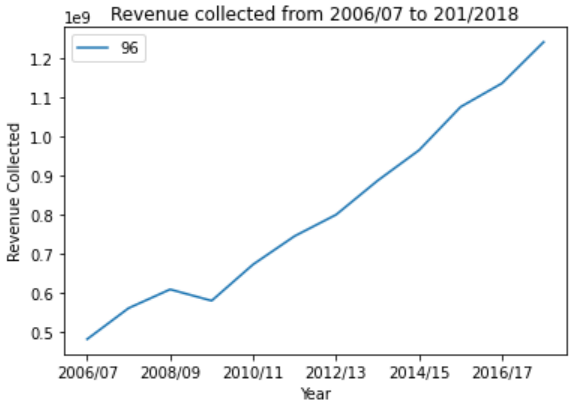
```
[ ] cleaned_taxData.head()
```

	2006/07	2007/08	2008/09	2009/10	2010/11	2011/12	2012/13	2013/14	2014/15	2015/16	2016/17	2017/18
0	279990516.0	332058296.0	383482732.0	3.590448e+08	3.799412e+08	4.265837e+08	4.573138e+08	5.077592e+08	5.617898e+08	6.068205e+08	6.605856e+08	7.391526e+08
1	140578347.0	168774352.0	195115008.0	2.051450e+08	2.269250e+08	2.503996e+08	2.758216e+08	3.098341e+08	3.529504e+08	3.881024e+08	4.258100e+08	4.820859e+08
3	118998582.0	140119631.0	165378278.0	1.348834e+08	1.329017e+08	1.516267e+08	1.592592e+08	1.773243e+08	1.849254e+08	1.911516e+08	2.050900e+08	2.186918e+08
4	15291351.0	20585421.0	20017580.0	1.546780e+07	1.717819e+07	2.196541e+07	1.973871e+07	1.730879e+07	2.124729e+07	2.393423e+07	2.571000e+07	3.423692e+07

5 Analysis and Insights

The first analysis that was performed entailed the understanding of the growth over the course of the year in the total revenue collection by the South African Revenue Service (SARS).

- Revenue collected from 2006/07 to 201/2018

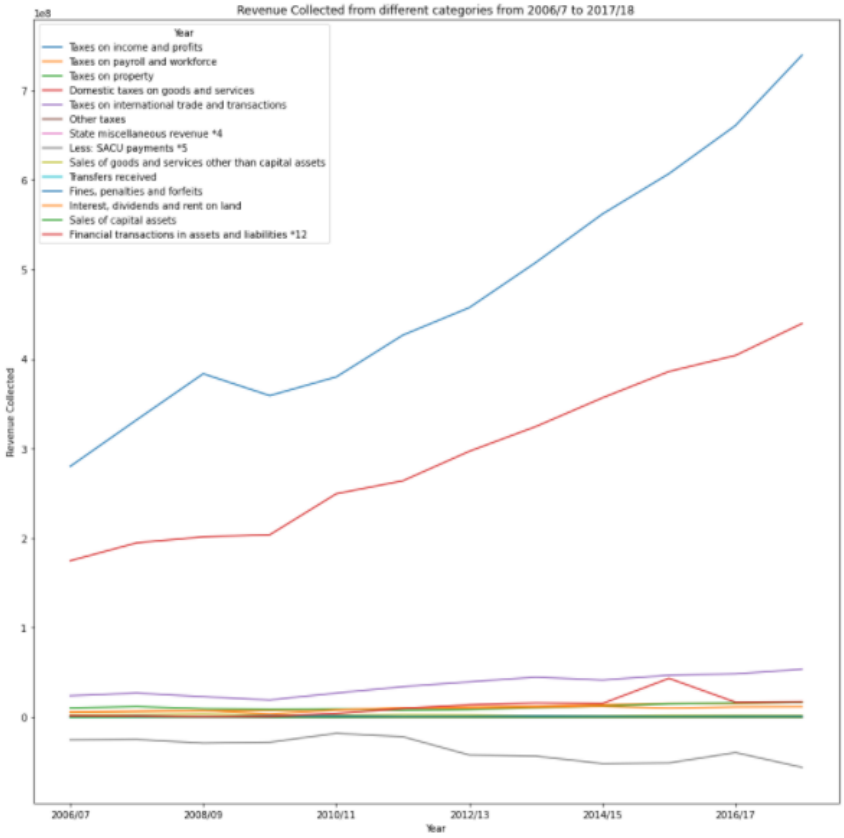


According to the graph, it is clear that the revenue collection has been increasing

at a linear rate at a high gradient year on year. This can lead us to deduce that SARS has been able to cope up with the increasing socio-economic activities in SA and more or less effectively collect taxes.

The second Analysis that was performed entailed the understanding of how the major tax types have been collected throughout the year and which major tax type contributed the most.

- Revenue Collected from different categories from 2006/7 to 2017/18

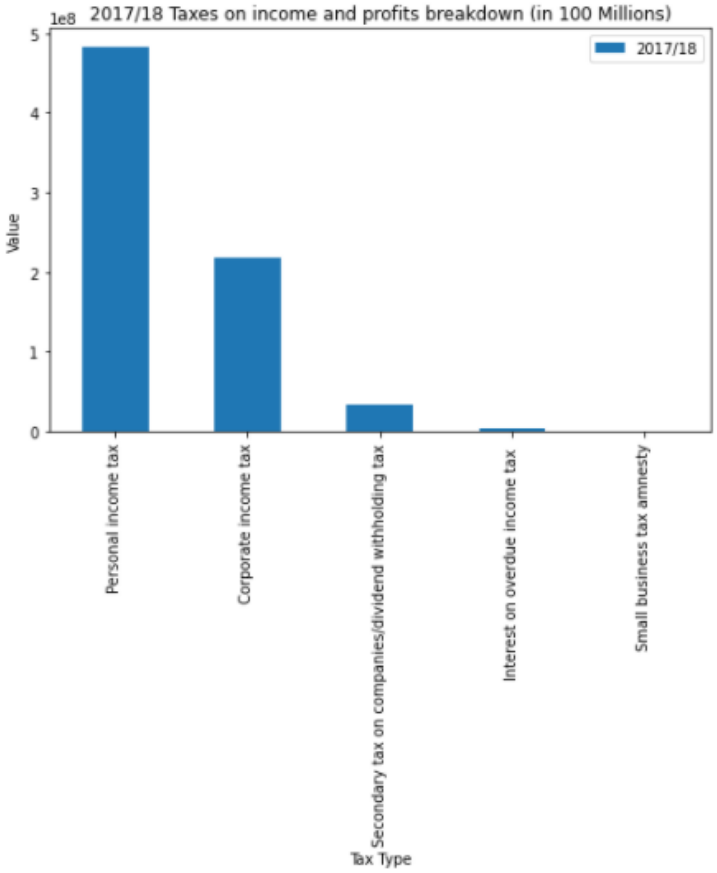


From the graph above it is seen that apart from the tax on income and profit and Domestic tax on goods and service, the other tax types have been constant in value. If SARS wanted to increase their revenue collection, this is an important aspect to look. This could mean two things: the taxes with minimal change have lower economic activity associated with them or the tax collection in these tax types is not effective and hence the actual amount that should be collected is not getting collected.

Moreover, it is seen that tax on income and profit contributed more than any other tax type.

The next analysis made was intended to find out how the sub categories within the top major tax category stand out and which sub categories are the highest and lowest contributors.

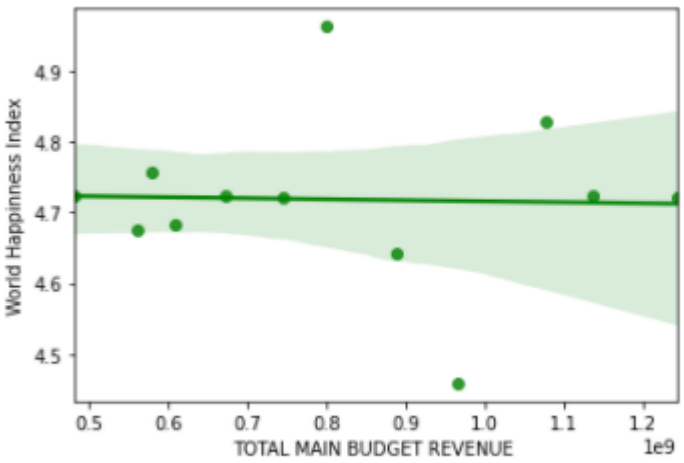
- 2017/18 Taxes on income and profits breakdown (in 100 Millions)



As seen in the bar graph personal income tax is the biggest contributor followed by corporate income tax. Small business tax amnesty and interest on overdue income tax are the lowest ones. This finding led to the following suggestions in terms of SARS ability to increase tax collection.

- Increasing the small business tax (which is out of SARS mandate, however can influence decision to be made in the parliament [4])
- Ensuring the systems in place to collect taxes from small businesses are functioning effectively.
- Increasing overdue income tax interest rate

The last analysis made was finding out the relationship between South Africa happiness index and total revenue collection.



We can deduce from the correlation graph that there is no relationship between the happiness index and the main budget revenue since the line passes straight through.

6 Recommendation for RRA

This POC was a demonstration of how data could be used to drive better decision making and increase tax collection. However, the data used here is not detailed and big enough as it would have been if we were working on developing the actual analysis systems at RRA. Hence, it is imperative for a robust and thorough analysis RRA needs to collect and provide various data points if we are to build an analysis and recommendation system.

Data required:

- Revenue collected by tax types
- Revenue collected by districts
- The internet connectivity level of the districts
- Projected revenue (preferably on a quarterly or bi-yearly basis so that comparisons can be made between projected and actual revenues and identify areas of improvement quickly)
- User satisfactions on RRA services (this would allow RRA to understand how its customer service impacts the interest in people to pay taxes in time and with honesty)

- Internal employee performance monitoring data (ideally one that has check-in and check-out times, employee role and average number of customers/cases handled by the employee if any)

Human Resource Required

- Our team has the right skill sets required to develop an analysis and recommendation system; we have data engineers (to structure data pipelines), data scientists (to create analytical models and frameworks) and software engineers (to develop the analytics application which would be accessible by senior management and other authorized personnel)

7 Conclusion

In a nutshell, even though the data-set used for this analysis isn't for Rwanda, we can clearly see the correlation between the state or Rwanda's Revenue collection and that of South Africa which makes it the best fit for us to deduce some insight to help curtail the problem at hand. From the data exploration through to the cleaning of the data to analysing and getting insights, following the recommendation provided above will aid in getting to the button of the problem and eventually make heads way.

8 Links

- A Github link that contains the Jupyter notebook file, along with a copy of this latext file can be found at
  https://github.com/NatnaelAlemayehu/SA_TaxRevenue_Analysis.git
- The presentation link shows us as the tech companies requesting for the biding proposal for the RRA project
  https://youtu.be/Wmj7TGOb7est
- This is the Jupyter Notebook link where all the data manipulation and analysis were made.
  https://github.com/NatnaelAlemayehu/SA_TaxRevenue_Analysis/blob/main/C1Group1_AI_Data_Analysis.ipynb

9 References

[1] Business procedures.rdb.rw. 2021. Business Procedures in Rwanda. [online] Available at: https://businessprocedures.rdb.rw/menu/12?l=en [Accessed 6 March 2021].

[2] Bower, J., Apell, D., Twum, A. and Umulisa, A., 2020. Rwanda's response to COVID-19 and future challenges - IGC. [online] IGC. Available at: https://www.theigc.org/blog/rwandas-response-to-covid-19-and-future-challenges/ [Accessed 6 March 2021].

[3] Treasury.gov.za. 2021. National Treasury. [online] Available at: http://www.treasury.gov.za/documents/national20budget/2017/excelFormat.aspx [Accessed 7 March 2021]

[4] Sars.gov.za. 2021. Mandate. [online] Available at: https://www.sars.gov.za/About/HowTax/Pages/Mandate.aspx [Accessed 7 March 2021].