AI Assignment 2

RWANDA REVENUE AUTHORITY
Bid Proposal
Group 1 Cohort 2

Agenda

- Introduction
- Prototype
- Budget
- Timeline
- Risk and Mitigation
- Recommendation
- Conclusion

Introduction

PURPOSE

To bid to RRA(Rwanda Revenue Authority) by demonstrating our technical capabilities through analyzing a financial and sentiment dataset and drawing useful inferences

MAIN OBJECTIVE

The main objective is to show RRA that we have the skills necessary to help them gain better insights from the vast amount of data which they collect on daily basis as well as sentiments from the media and other platforms.

Source: https://www.rra.gov.rw/

Stakeholders

#	CLIENT	TECH COMPANY
Name	RWANDA REVENUE AUTHORITY(RRA)	COHORT 2 GROUP 1
Phone	(250) 788 185 500	(250) 784 650 219
Email	info@rra.gov.rw	c1g1@alustudent.com

Problem statement

The Rwanda Revenue Authority (RRA); is a financial services institution mandated to collect tax revenue on behalf of the Republic of Rwanda.

As a result of the prevailing COVID- 19 pandemic, its tax collection figures for 2020-2021 fiscal year deteriorated by over 10% to a specific value of US1,3 Billion Dollars.

Invest in digital transformation

An effective revenue collection platform to influence their decision making process(Web, App etc)

Source: https://www.rra.gov.rw/

The Prototype

01	Data Collection	• Researching and downloading data • Importing Data to AI Environment • Understanding the Data
02	Data Preparation	• Missing values • Slicing
03	Data Transformation	• Renaming column titles
04	Data Integration	• Concatenation
05	Data Visualization	• Histograms • Bar charts • Scatter plot

Data Collection

Tax Revenue by Month

- Tax Collections by the Department of Revenue Services
- Contains 114 rows and 51 columns
- Every row contains tax revenues for every month of the year.
- While some of the values are null, most of the entries contain numerical variables.
- No duplicates

Airline Tweets

- Twitter data where contributors reviewed problems of each major U.S. airline.
- Contains 14640 rows and 15 columns
- "airline_sentiment" column that describes positive, negative, and neutral reviews.
- “Text” column that describes reviews from different users
- No duplicates

Source: https://catalog.data.gov/dataset/tax-revenue-by-month, https://www.kaggle.com/crowdflower/twitter-airline-sentiment

Data Preparation, Transformation and Integration

Missing Values: There are different ways to handle missing values. From dropping them, to filling them with zero, the mean, mode or the median.

- In our case, we filled all missing values in the tax dataframe with the mean.
- We also replaced the '-' filled value cells with 0

Slicing: To slice out a set of rows, we used the following syntax: dataframe[start:stop].

Renaming Column Names: Some of our column names were too long.

- We renamed our column names by shortening and simplifying their meaning.

Source: https://pandas.pydata.org/

Tax Revenue by Month Preview

	Month	Calendar Year	Fiscal Year	Withholding	Income Tax	Sales	Business Use	Room	Electricity	Business Entity	...	Nursing Home	Hospitals
1	August	2011	FY 2011-12	2.771524e+08	2.320268e+07	2.666841e+08	927436.95	8942131.10	0.0	484492.78	...	1358440.78	0.0
2	September	2011	FY 2011-12	3.980483e+08	2.748813e+08	2.896617e+08	2769766.79	10132507.23	0.0	824783.27	...	664491.99	0.0
3	October	2011	FY 2011-12	4.086132e+08	6.944227e+07	3.335668e+08	3798695.47	9770714.10	17767106.0	461952.67	...	33023090.06	84247905.0
4	November	2011	FY 2011-12	4.657074e+08	8.803835e+07	3.082123e+08	1837456.18	10088593.69	0.0	484246.81	...	2630223.01	3032664.0
5	December	2011	FY 2011-12	5.409526e+08	2.284256e+08	2.384391e+08	8393847.86	8472731.83	0.0	3841928.66	...	915106.47	0.0

Source: https://catalog.data.gov/dataset/tax-revenue-by-month, https://www.kaggle.com/crowdflower/twitter-airline-sentiment

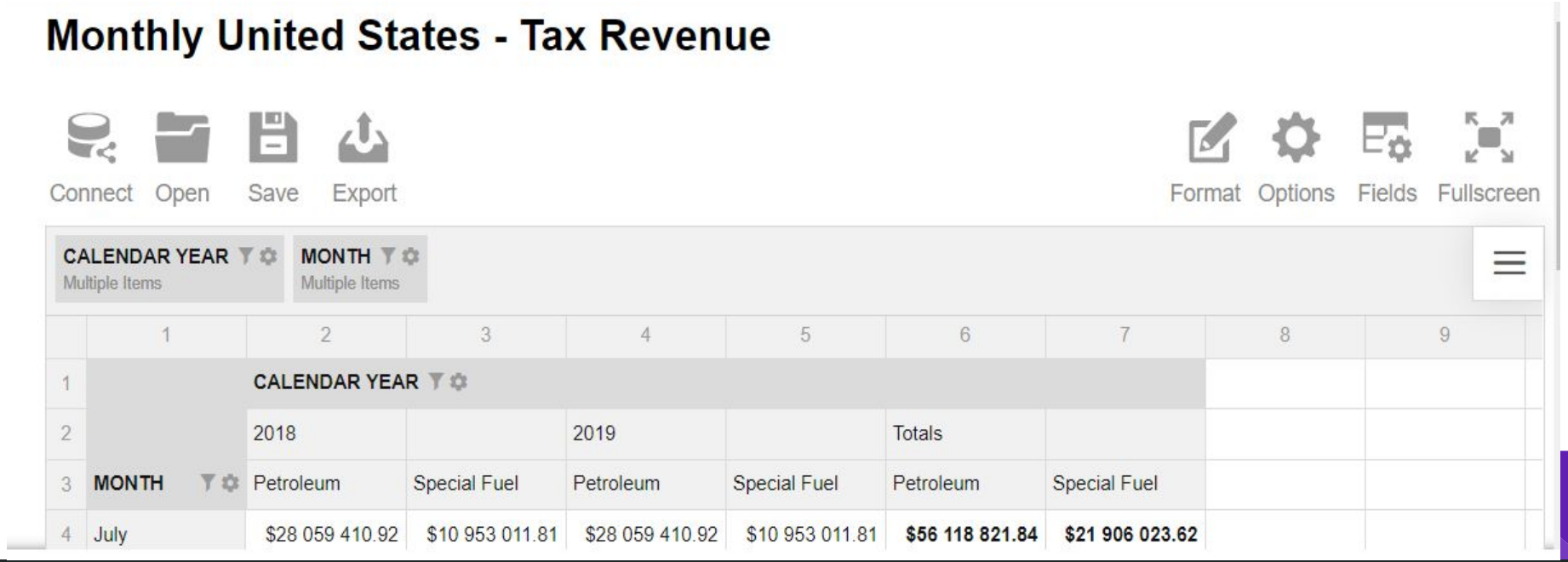
Airline Tweets Preview

	tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	airline_sentiment_gold	na
0	570306133677760513	neutral	1.0000	NaN	NaN	Virgin America	NaN	ca
1	570301130888122368	positive	0.3486	NaN	0.0000	Virgin America	NaN	jnar
2	570301083672813571	neutral	0.6837	NaN	NaN	Virgin America	NaN	yvonna
3	570301031407624196	negative	1.0000	Bad Flight	0.7033	Virgin America	NaN	jnar
4	570300817074462722	negative	1.0000	Can't Tell	1.0000	Virgin America	NaN	jnar

Source: https://catalog.data.gov/dataset/tax-revenue-by-month, https://www.kaggle.com/crowdflower/twitter-airline-sentiment

Pivot Table

We also integrated a pivot table of statistics that summarizes the data of the more extensive tax dataset. This summary includes sums, or averages which the pivot table groups together in a meaningful way.



	1	2	3	4	5	6	7	8	9
1		CALENDAR YEAR							
2		2018		2019		Totals			
3	MONTH	Petroleum	Special Fuel	Petroleum	Special Fuel	Petroleum	Special Fuel		
4	July	$28 059 410.92	$10 953 011.81	$28 059 410.92	$10 953 011.81	**$56 118 821.84**	**$21 906 023.62**		

Data Visualization

We used Matplotlib and Seaborn libraries in Python to visualize the data, and to find a trend in the tax revenues collected by the United States government.

Types of Graphs we used are:

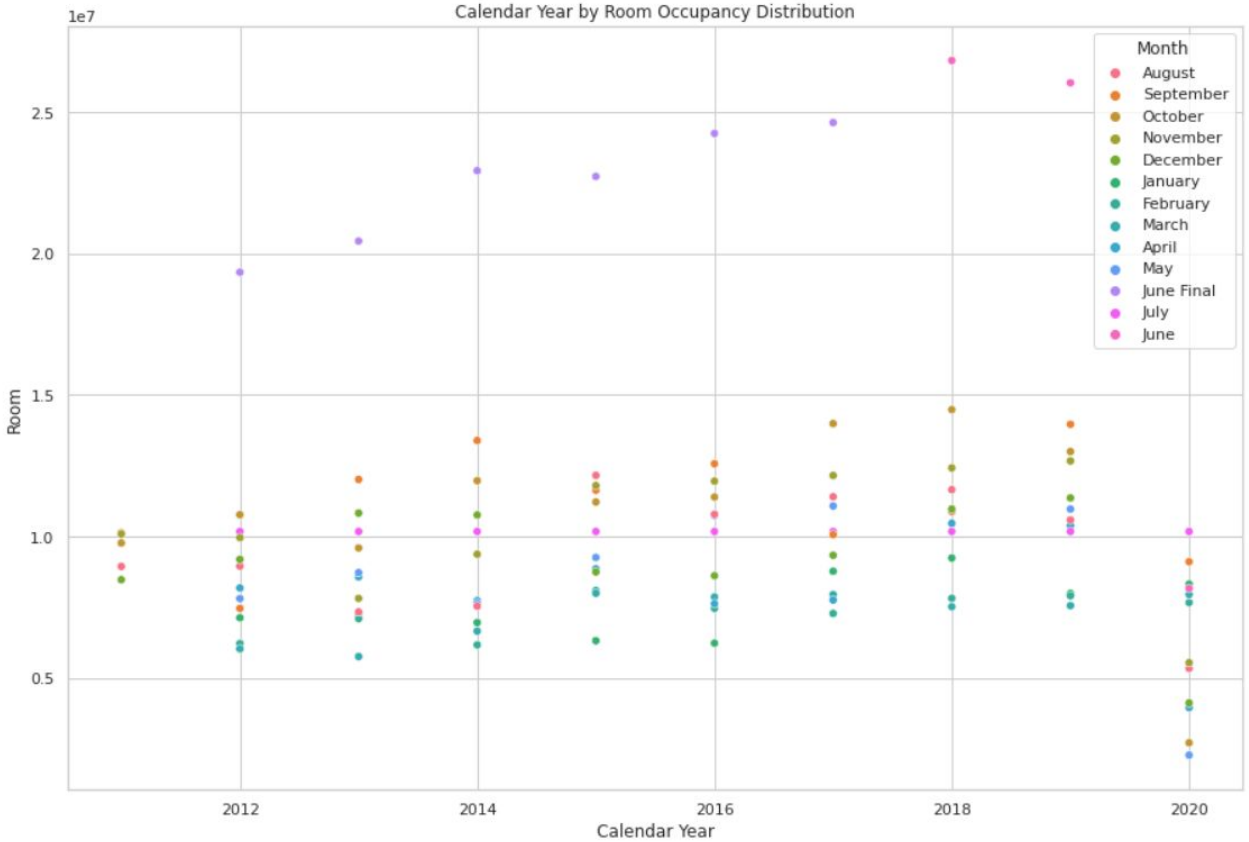
- Line Graph
- Scatterplot
- Histogram
- Bar Graph
- Density plot

Source: https://riptutorial.com/ebook/matplotlib

Line Graph

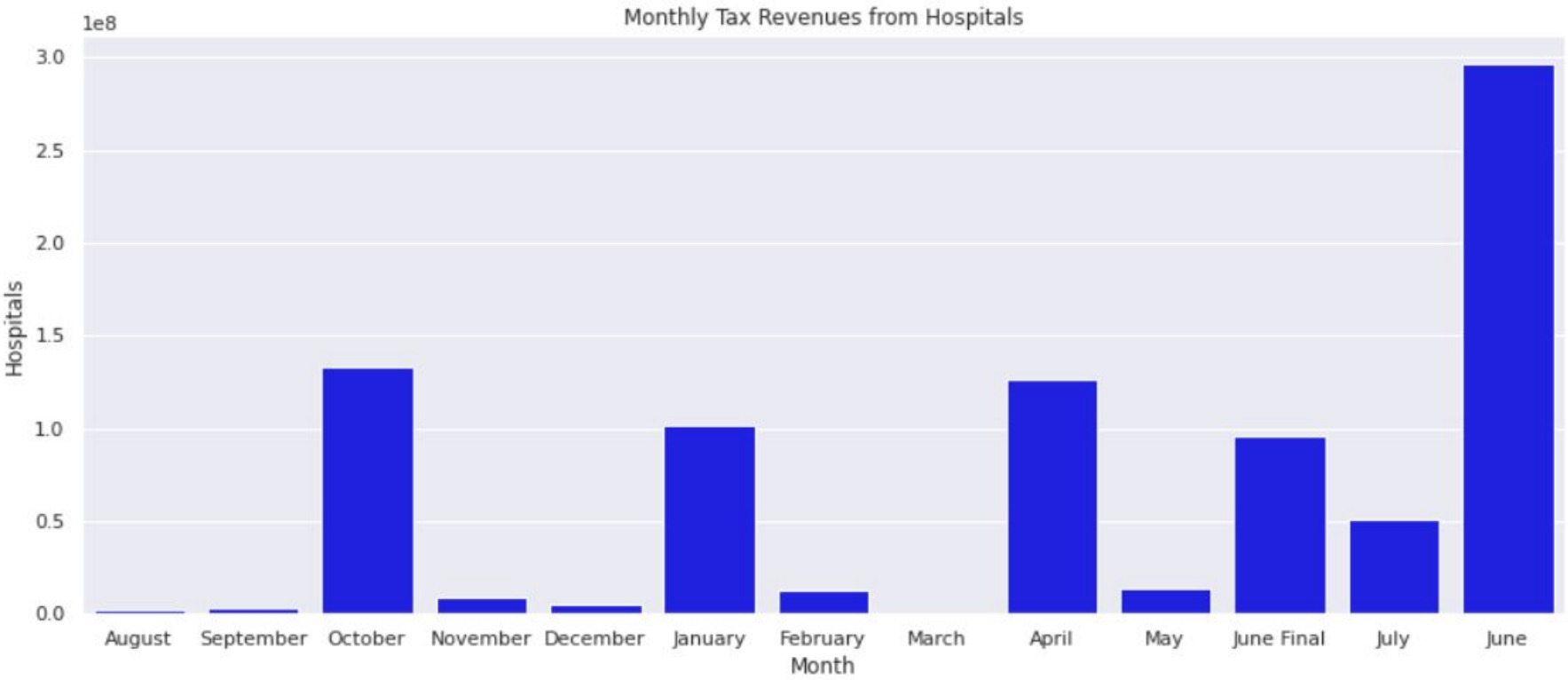
Source: https://riptutorial.com/ebook/matplotlib

Scatterplot



Source: https://riptutorial.com/ebook/matplotlib

Bar Graph



Source: https://riptutorial.com/ebook/matplotlib

Word Cloud

Source: https://kavita-ganesan.com/word-cloud-for-data-scientists/

Environment

Hardware Environment

- Computer or a Laptop

Software Environment

- Operating system (Windows, Mac OS, Linux).
- Jupyter Notebook
- Python libraries (Pandas, Matplotlib, Seaborn, WordCloud)

Budget

Estimated Total

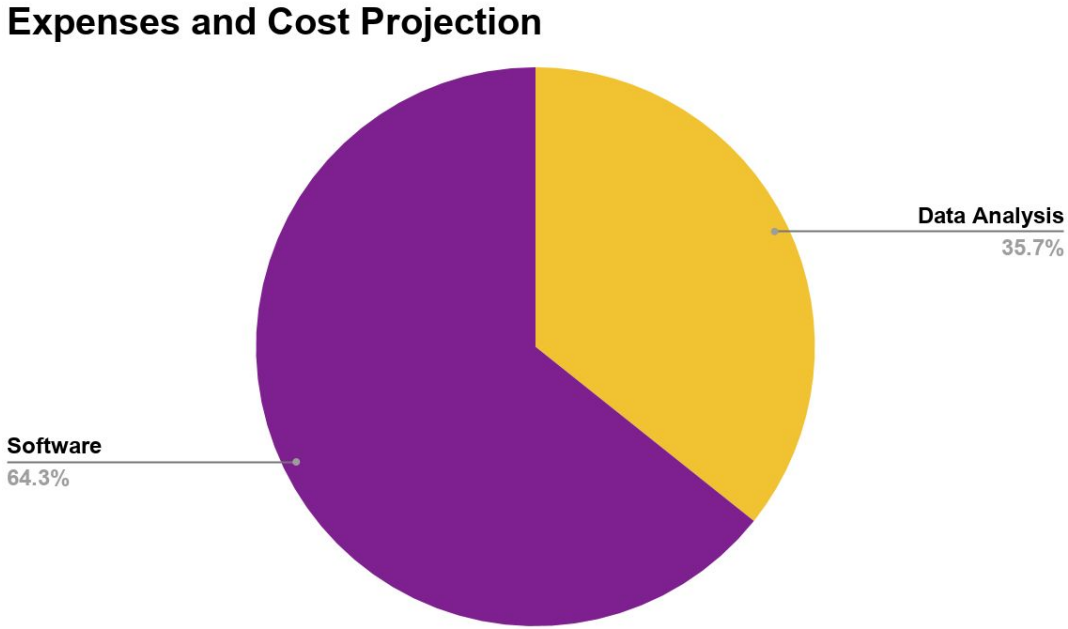
$ 14,000

Data Analysis **$9000**

- Head Analyst - $3,000
- Data Scientist - $2,000

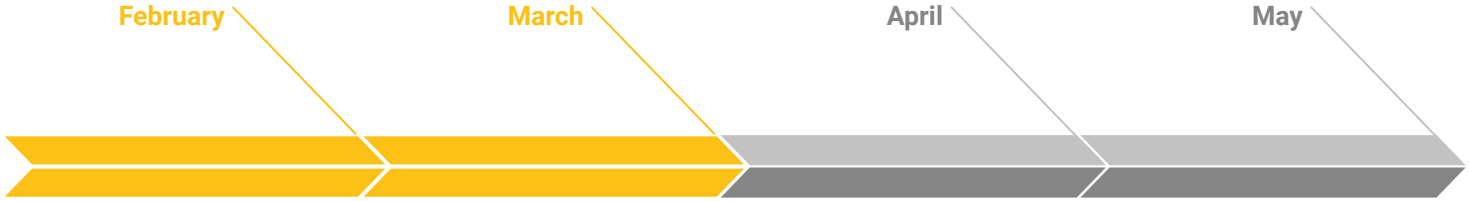
Software Development **$5000**

- Head Developer- $4,000
- Project Manager- $2,000
- Frontend Developer- $1,500
- Backend Developer- $1,500



Source: https://docs.google.com/spreadsheets/d/1mVfUTTIXQHOiyNWajdVUJMBZYjX00TAWXke3bc4KNeg/edit?usp=sharing

Timeline



- Source data
- Prepared the data

- Prototype
- Bid proposal

- Implementation
- Testing

- Product Build
- Product release

Risks and Mitigation

- Lack of adequate data may result in inaccurate findings. Having more data provides a more substantial basis for making fact-based decisions that can better inform company strategies.

- To mitigate the risk of inadequate data, the institution must maintain a steady stream of data supply through customer review collection.

Recommendation

We recommend Virgin America to increase their ticket prices and also restructure their target market to a customer demographic that can afford it.

As an alternative to increasing prices, we recommend that they should ensure proper allocation of funds to departments that interact directly with the customer base.

In conclusion:

- We often make assumptions about a business and figure out decisions without a firm base.
- Exploratory Data Analysis is a great methodology to visualise the data using different charts and graphs and they, in turn, provide an affirmation to our hypothesis
- Plotting techniques also help to validate the hypothesis which is made about the data.

References

[1] "What is Exploratory Data Analysis?", Ibm.com, 2020. [Online]. Available: https://www.ibm.com/cloud/learn/exploratory-data-analysis. [Accessed: 04- Mar- 2021].

[2] Part 4: Data Management and Analysis, Reporting and Disseminating Results. WHO STEPS Surveillance, 2017, pp. 4-1-1 to 4-4-1.

[3] A. Miller, "Introduction to Using Excel® Pivot Tables and Pivot Charts to Increase Efficiency in Library Data Analysis and Illustration", Missouri State University Library Administration, vol. 54, no. 2, pp. 94-106, 2014. Available: https://bearworks.missouristate.edu/cgi/viewcontent.cgi?article=1001&context=articles-lib. [Accessed 5 March 2021].

[4] M. Wood, Python and Matplotlib Essentials for Scientists and Engineers, 3rd ed. Quebec: Morgan & Claypool, 2015, p. 150pp.

[5]"What's new in each version — seaborn 0.11.1 documentation", Seaborn.pydata.org, 2021. [Online]. Available: https://seaborn.pydata.org/whatsnew.html. [Accessed: 05- Mar- 2021].

[6] Z. Luvsandorj, "Simple word cloud in Python", Towards Data Science, 2020. [Online]. Available: https://towardsdatascience.com/simple-wordcloud-in-python-2ae54a9f58e5. [Accesses: 05- Mar -2021].

[7]J. Vanschoren, "An Exploration of Techniques Used in Data Analytics to Produce Analysed Data in Graphical Format", Masters, Eindhoven University of Technology, 2018.