AFRICAN LEADERSHIP UNIVERSITY RWANDA

CHAPTER 2: DATA ANALYTICS CHALLENGE

01/03/2021

[Department: Computer Science]

[Facilitator: Kudakwashe Dandajena]

PEER MEMBERS

- JOYCE NJERI jmiiri17@alustudent.com
- FATMATA ALICE KOROMA f.koroma@alustudent.com
- GILBERT SIBOMANA g.sibomana@alustudent.com
- AYO-FISHER OLUWAPAMILERIN o.ayofishe@alustudent.com
- CHRISTINE WASIKE c.wasike@alustudent.com

Abstract

This paper provides a review of the work that 1C2 (Cohort 2 Group 1) has done in line with data analysis on a financial data set, a sentiment data set and, a combination of both. The paper explains our prototype's implementation, the environment we need to scale the project to production, our team's budget, timelines, risk, and recommendations for its successful implementation.

Key Words

Data analysis, software development, sentiment analysis.

1. Introduction

1.1 Purpose

This document is our company's (Group 1 Cohort 2) bid to RRA(Rwanda Revenue Authority), which explains the implementation of our prototype, the environment we need to scale the project to production, our team's budget, timelines, risk, and recommendations for the successful implementation of the project.

1.2 Document Convention

This document follows the Latex format.

2. Stakeholders

2.1 Client Information

Name:

Rwanda Revenue Authority (RRA)

Address:

Kimihurura, Kigali, Rwanda

Phone Number:

(250) 788 185 500

Email Address:

info@rra.gov.rw

2.1 Contractor Information

Name:

1C2(Cohort 2 Group 1)

Address:

KG 126 St, Kigali, Rwanda

Phone Number:

(250) 784 650 219

Email Address:

info@alueducation.com

3. Our Prototype

3.1 Background Information

The aviation industry is substantial for every country's revenue system as it's the gateway to getting people in and out of the country. One of the biggest challenges in the aviation industry is that it is costly to maintain a good standard due to financial hurdles and legal complications.

In 2015 Virgin America(a US-based airline) experienced customer loss and profit loss when the government issued a license tax on special fuel(airplane fuel).

This made special fuel expensive for Virgin America, who decided not to increase ticket fares for customers but to cut costs. Unknowing to them, their cost-cutting affected their services, and they ended up losing customers.

Not only did they lose customers, but they also had their image tarnished as customers took it to Twitter(social media platform) to air their displeasure.

3.2 Prototype Implementation

3.2.1 Data Collection

3.2.1.1 Researching And Downloading Data

We researched using the prompt to know the kind of data that is needed to solve the problem. After our research, we downloaded data that we found relevant to solving the problem.

3.2.1.2 Importing Data To AI Environment

We used Pandas' to import the data set to Jupyter Notebook(AI environment)

3.2.1.3 Understand The Data

After importing it to Jupyter Notebook we used exploratory data analysis (EDA) to analyze, investigate and summarize the main characteristics of the data set.[1]

3.2.2 Data Preparation

3.2.2.1 Missing Values

There are different ways to handle missing values. From dropping them, to filling them with zero, the mean, mode, or the median. [2]

3.2.2.2 Slicing

Slicing is the process of selecting specific rows and columns of data based on some criteria. We used it in data preparation to highlight the unique rows and columns we wanted to work with.

3.2.3 Data Transformation

3.2.3.1 Renaming Column titles

Due to long column names, we used the Pandas' rename method to get short titles.

```
# Transforming Data: Renaming column titles
dict = {'Income Tax Estimates & Finals': 'Income Tax',
        'Sales and Use': 'Sales',
        'Room Occupancy': 'Room',
        'Electric Generation': 'Electricity',
        'Electronic Cigarette Products': 'Electronic Cigarettes',
        'Tobacco Products': 'Tobacco',
        'Real Estate Conveyance': 'Real Estate',
        'Petroleum Gross Earnings  ': 'Petroleum',
        'Prepaid Wireless E-9-1-1 Fee': 'Prepaid Wireless',
        'Nursing Home User Fee': 'Nursing Home',
        'Intermediate Care Facility': 'Intermediate Care',
        'Beverage Container Deposits': 'Beverage Containers'
       }

data.rename(columns=dict, inplace=True)
```

3.2.4 Data Integration

We integrated the sentiment data set with the financial data set in order to draw correlations and inferences.

3.2.5 Data Visualization

We have used Matplotlib[4] and Seaborn libraries in Python to visualize the data and to find a trend in the tax revenues collected by the United States government.[1]

```
In [32]: data[['Income Tax','Sales']].plot(figsize=(15,10), style=['-','--'], lw=2)
Out[32]: <AxesSubplot:>
```

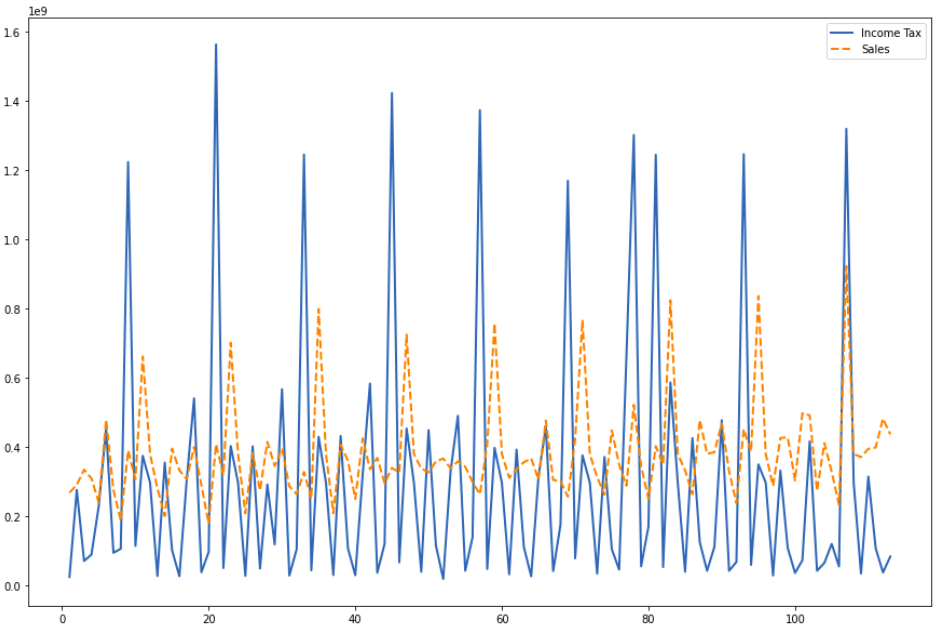

Figure 1: Line graphs comparing Income Tax to Sales[6]

```
sns.scatterplot(x='Calendar Year', y='Room',hue = 'Month',data=data,)
plt.title('Calendar Year by Room Occupancy Distribution')
Out[33]: Text(0.5, 1.0, 'Calendar Year by Room Occupancy Distribution')
```

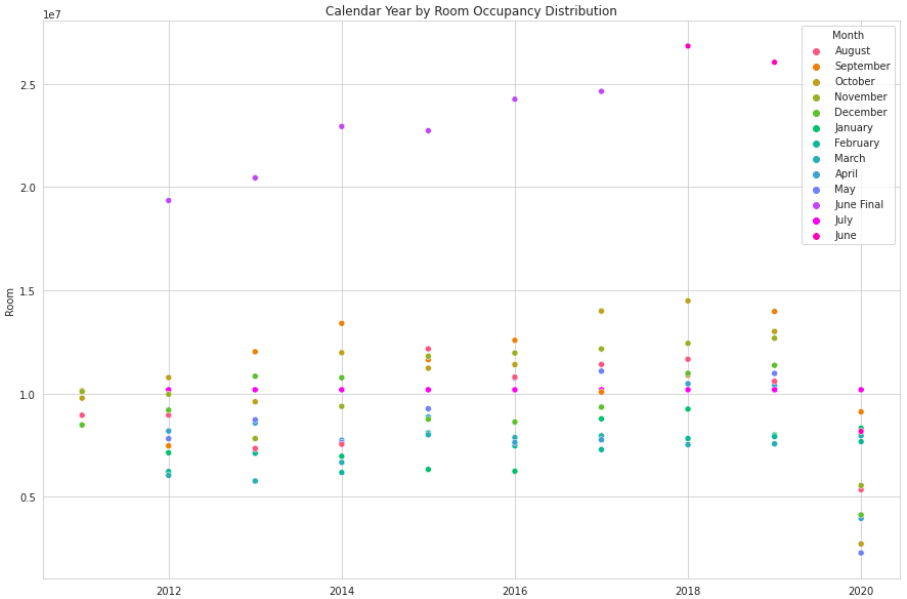

Figure 2: A scatter plot demonstrating the Occupancy Distribution across a year

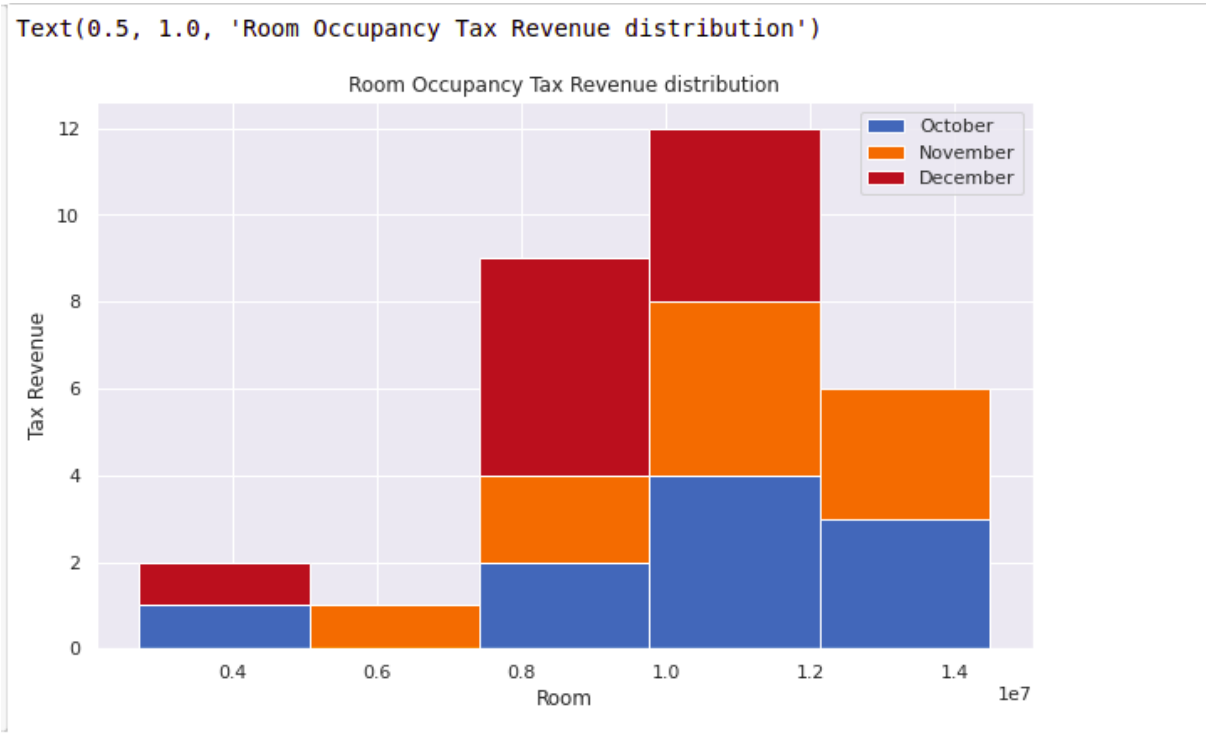

Figure 3: Tax Revenue based on Room Occupancy

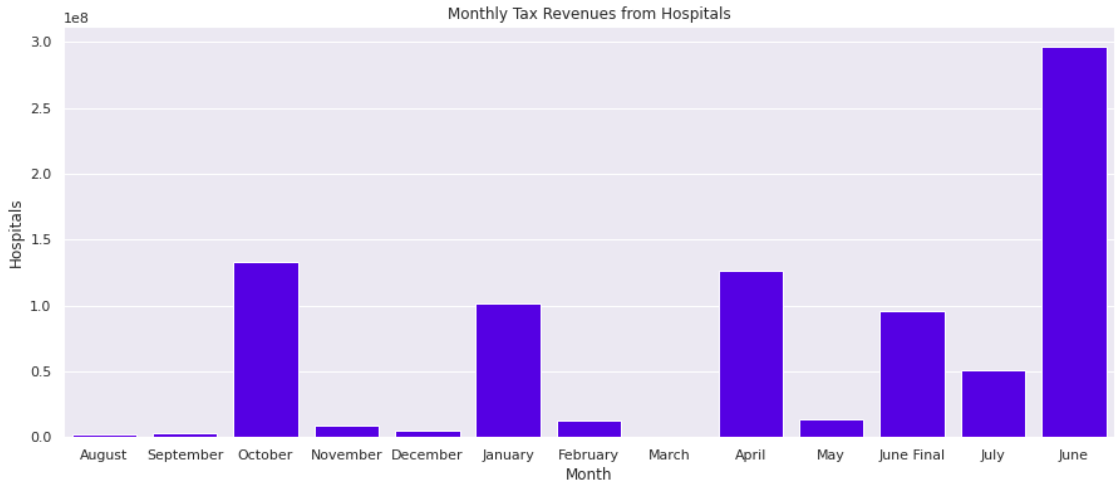

Figure 4: Monthly Tax Revenues from Hospitals

Figure 5: Word Cloud Part 1[6]

Figure 6: Word Cloud Part 2

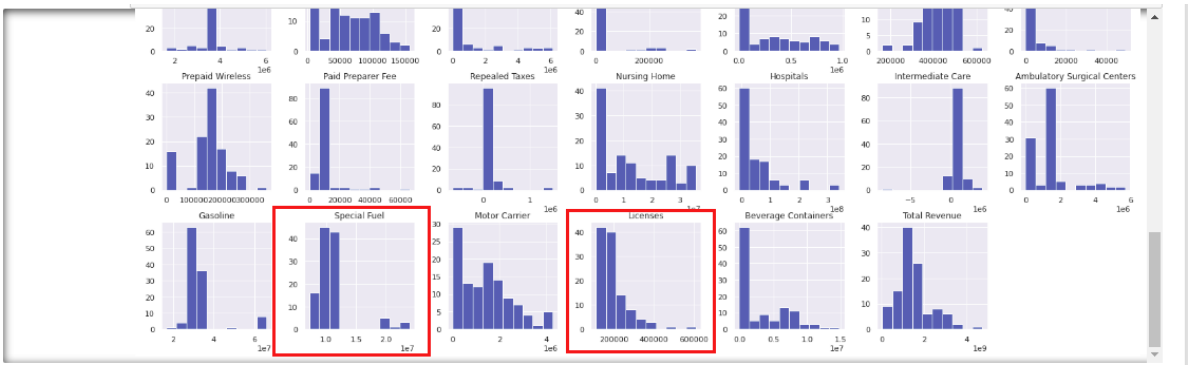

Figure 7:Histogram graph showing frequency distributions in a bar set-up

Figure 8: Density plot showing the probability density function of the variable in form of probability distribution.

3.3 Environment Needed

Hardware Environment

- 4GB Ram at least, 8GB is recommended.
- 500GB Drive
- Operating system(Win, Mac OS, Linux).

Software Environment

- Jupyter Notebook
- WebDataRocks (Pivot Table)
- Python libraries (Pandas, Matplotlib, Seaborn, WordCloud)

4. Administration Information

4.1 Budget

Table 1. Budget Breakdown

Expenses	Cost (Per month)	Role Breakdown	Description
Data Analysts' Quotation	$5,000	Head Analyst - $3,000 Data Scientist - $2,000	Money set aside to pay the Data Analysts working on the product prototype backend. This includes the models and visualization and exposes the most critical data points for a light and fast system.
Software Development Quotation	$9,000	Head Developer- $4,000 Project Manager- $2,000 Frontend Developer- $1,500 Backend Developer- $1,500	Money allocated towards developing the product, i.e., Mobile Application, and the backend system

4.2 Timelines

Table 2. Project Timelines

Months	Tasks	Status
February	Source data	Completed
	Prepared the data	Completed
	Visualized the data	Completed
March	Created prototype	Completed
	Created bid proposal	Completed
	Submitted bid	Completed
	Bid approval	In Progress
April	Exposing data endpoints to the development team	Not Started
	Implementation of the software prototype	Not Started

	Testing the software prototype	Not Started
May	Build the final product	Not Started
	Production release	Not Started

4.3 Risk

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.

4.4 Mitigating Risks

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

5. Recommendations

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.

6. Conclusion

- Exploratory Data Analysis (EDA) is the best way to understand and summarise the characteristics and behaviors of the data.
- We can also use plotting techniques to validate the hypothesis which is made about the data.
- EDA also helps us to understand which model will fit best for predictions about the data set.
- It also reduces our efforts at the time of machine learning model building

With the data set of our choice, we have been able to demonstrate our prowess by properly manipulating the data and performing operations. The group achieved the following results:

1. Manipulated the data, performed operations and procedures with the data.
2. Used the data set to come up with recommendations.

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

[4]"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].

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