*Genevieve Gonzales, Eddie Galarce Jr,Kristian Glorioso, Neil Angelo Martinez, Onofre Vidad*

*Date of Submission: December 5,2022*

**PROJECT ASSIGNMENT #1**

**INTRODUCTION**

For the past few years, Yellevate has been struggling with client disputes. Yellevate defines disputes as clients expressing dissatisfaction with the company’s services and refusing to pay for them. The team will be following Google’s five phases of data analysis: ask, prepare, process, analyze and share.

**ASK**

Statistically, approximately 20% of the Yellevate conflicts resulted in a payment opt-out. This has resulted in a 5% annual revenue loss (in USD). Management has now asked the data analyst team to assist in resolving the problem. The team must identify the reasons of these issues using data gathered by the company and devise practical methods to resolve them.

Executives at the company decided that the following information should be obtained to identify the circumstances around the dispute problem:

1. The processing time in which invoices are settled (average # of days rounded to a whole number).
2. The processing time for the company to settle disputes (average # of days rounded to a whole number).
3. Percentage of disputes received by the company that were lost (within two decimal places).
4. Percentage of revenue lost from disputes (within two decimal places)
5. The country where the company reached the highest losses from lost disputes (in USD).

**PREPARE** *(ROCCC Analysis)*

Reliable – The data has been made available by Refocus specifically for this project

Original - Subject to the terms and conditions of the Data License Agreement by Yellevate.

Comprehensive – Data dictionary has been provided through this [link](https://docs.google.com/spreadsheets/d/1bm4UvkcBfIS-blftRSafsg1TWTyVZ3SD/edit?rtpof=true&sd=true#gid=503846598).

Current – The latest data is 2022.

Cited - All the data needed for this project can be accessed through this [link](https://drive.google.com/file/d/1DXOaFAFF1Rqic1PWXvABxgCS4o1gYgxG/view).

**PROCESS**

The first step is to get the data from the provided link and import it into pgAdmin 4 for data cleaning. The code for importing data is shown below:

CREATE TABLE yellevate\_invoices (

country varchar,

customer\_id varchar,

invoice\_number numeric,

invoice\_date date,

due\_date date,

invoice\_amount numeric,

disputed numeric,

dispute\_lost numeric,

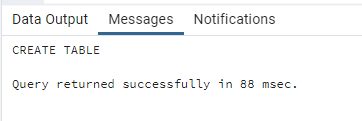
settled\_date date,

days\_settled integer,

days\_late integer

);

After creating the table, go to the side panel of pgAdmin4 and right click on the created table and click *Import/Export Data*. Click on the folder icon and locate the CSV file that you want to import. The format should be CSV and the encoding is UTF8.



**CLEAN DATA**

According to the data dictionary, the *disputed* field has 0 or 1 values. "1" signifies the customer disputed the invoice; "0" means they did not. When this table is provided to clients, they must understand what the numbers 0 and 1 signify. The team has added a new column that will display Disputed or Not Disputed.

ALTER TABLE yellevate\_invoices

ADD COLUMN category varchar,

UPDATE yellevate\_invoices

SET disputed\_ornot = (

CASE

WHEN disputed = 1 THEN 'Disputed'

ELSE 'Not Disputed'

END

);

Same goes with the *dispute\_lost* column, "1" means Yellevate lost the dispute and the dispute was resolved in favor of the customer, and the customer does not have to pay the invoice; "0" means the customer did not win the dispute, and they are legally required to pay the full invoice amount, it either means that Yellevate won the dispute, or there was no dispute in the first place.

ALTER TABLE yellevate\_invoices

ADD COLUMN win\_lost VARCHAR;

UPDATE yellevate\_invoices

SET win\_lost = (

CASE

WHEN dispute\_lost = 0 THEN 'Won'

ELSE 'Lost'

END

);

Lastly, a new column named *settlement* has been added to distinguish disputed that have been settled on time or settled late.

ALTER TABLE yellevate\_invoices

ADD COLUMN settlement varchar,

UPDATE yellevate\_invoices

SET settlement = (

CASE

WHEN days\_late > 0 THEN 'Late Settlement'

ELSE 'Settled on Time'

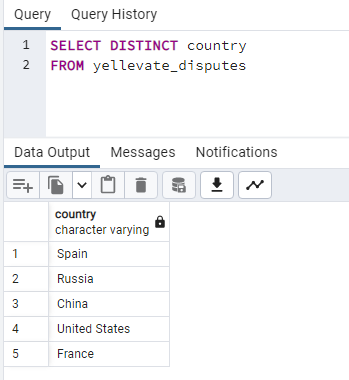
END

);

The country column is the one that can be checked for inconsistencies. We can see all of the countries listed by using the DISTINCT function. There are just five distinct countries after running the code.

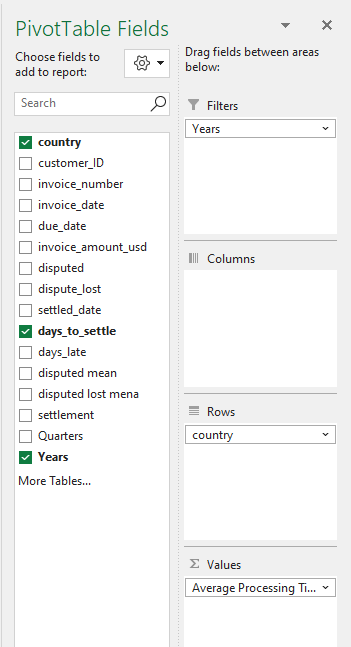
SELECT DISTINCT country

FROM yellevate\_invoices



**ANALYZE**

After cleaning the data, the CSV file was exported to Excel for further analysis. Creating pivot tables is the simplest technique to obtain averages. To calculate the average processing time of disputes, place the *country* column under rows and the *days\_to\_settle* column under values.



**SHARE**