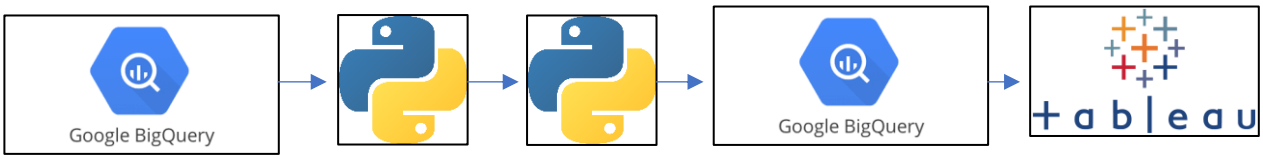
Leoranele Grace Q. Estrada
Brazil Real Estate Listing Analysis
Technical Documentation

I. Understanding the Dataset

The Brazil Real Estate listing is the dataset from Properati Platform. Understanding it can be quite complex as there is no data dictionary available. The following are the steps I took to understand:

- **Check the database table structure** – research about the dataset and Properati.
- **Check the database table structure** – there's a lot of tables in the database. The table is structured as listing per month having sales and rent on separate tables
- **Verify the uniqueness of the ID column** - I noticed that it is not unique because there are different properties with the same ID. My presumption is that is is the user ID.
- **Check unique values of category columns** – knowing the values a categorical column could take would be helpful in building an intuition on how it would be significant on the analysis.
- **Check range of numerical columns –** the numerical columns in the dataset has a lot of missing and invalid values, and it It is not practical to drop the rows containing the null values. Normally I would verify this with the process owner but since I can't, I just retained imputed the null values with 0 as an indicator.
- **Check outliers –** numerous outliers are also in the dataset specifically for the floors column which if used in the analysis, would yield incorrect results. For example, the maximum value for the floor column is 2,147,483,647.
- **Randomly checked property images –** checked some of the images to see if it aligns with the property description

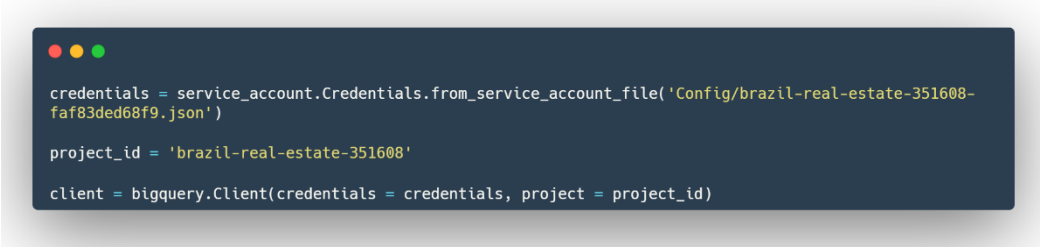
II. Analysis Workflow



III. Preparing Data for the Analysis

Since the data is in BigQuery, I had to learn about navigating in BigQuery first as I have no prior experience. After getting the gist of it, the following are my data preparation steps:

- Created a Python program, **"BigQuery Integration.py"**, that connects to Google BigQuery to extract data directly using python.

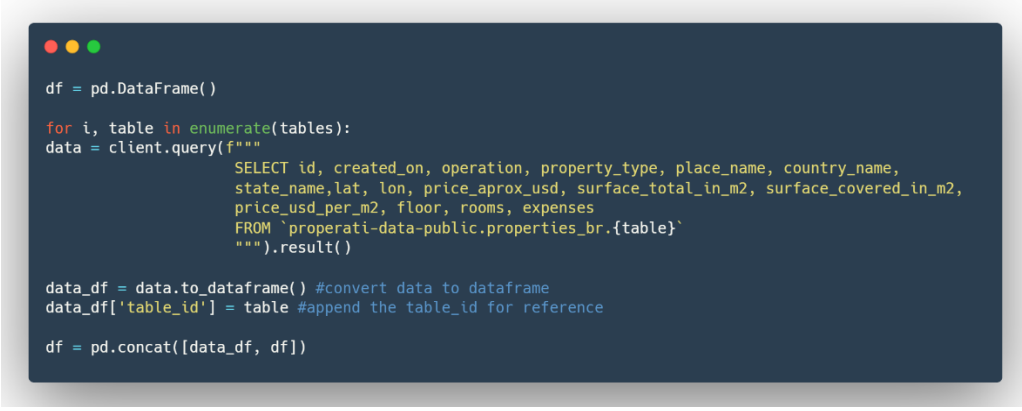


```
credentials = service_account.Credentials.from_service_account_file('Config/brazil-real-estate-351608-faf83ded68f9.json')

project_id = 'brazil-real-estate-351608'

client = bigquery.Client(credentials = credentials, project = project_id)
```

- Created a SQL Query to query relevant data to be loaded in the DataFrame.
  Also added the table_id column as a reference for the table name



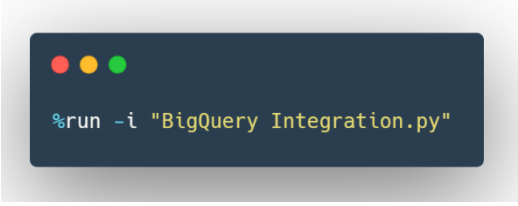
```
df = pd.DataFrame()

for i, table in enumerate(tables):
data = client.query(f"""
                    SELECT id, created_on, operation, property_type, place_name, country_name,
                    state_name,lat, lon, price_aprox_usd, surface_total_in_m2, surface_covered_in_m2,
                    price_usd_per_m2, floor, rooms, expenses
                    FROM `properati-data-public.properties_br.{table}`
                    """).result()

data_df = data.to_dataframe() #convert data to dataframe
data_df['table_id'] = table #append the table_id for reference

df = pd.concat([data_df, df])
```

- The Python program is then executed within the Jupyter Notebook to facilitate reproducibility of the data extraction up to the data cleaning.



```
%run -i "BigQuery Integration.py"
```

Data Cleaning

The codes and detailed documentation for my data cleaning is in the **"Brazil Real Estate Listings – Data Preprocessing.ipynb"** file

To summarize, the steps I didare the following:

- Merging the **rent and sell** dataset
- Changing the column **data types**
- Add list_date columns based on the table_id
- Create new DataFrame for unique property listings
- Add new columns:
    - list_date – based on the table_id, indicates the first month the property was listes
    - operation_ind – indicator for the operation if it is for rent, sell, initially rent then changed to sell (rent to sell), and vice-versa (sell to rent)
    - list_duration – calculated column for the duration of listing:
      maximum list_date – minimum list_date
    - status – whether the listing is active or inactive based on the list date
    - owner_id – renamed the id to owner_id
    - property_id – unique id created to identify unique listing
      owner_id + state_name + place_name

- Load the cleaned dataset to Google BigQuery and used It as the data source for the Tableau Visualization.

```
credentials = service_account.Credentials.from_service_account_file('Config/brazil-real-estate-351608-faf83ded68f9.json')

project_id = 'brazil-real-estate-351608'

pandas_gbq.to_gbq(data_unique, table_id, project_id=project_id, if_exists = 'replace')
```

IV. **Thought process for the Insights**

Since this is the first time for me to analyze this data, I decided to get the profile of the property listing first like the most expensive house, least expensive house, average prices, distribution of property types, etc. It was more on the discovery or exploratory phase that just describes the current state of the properties in Properati for me to get a general understanding of what properties are in the platform. This serves as the foundation of the analysis.

The next one would be to identify some patterns and feature values that deviate from the patterns like the price trends, property type distribution per state, etc.

The last one was about the platform metrics. I think this is helpful to know as this gives us a general idea of the value that the platform gives to the users or in other words, how well this platform serves as a tool to connect property sellers to potential buyers. Below are some of the questions I include for this section:

- How many of the total listings are already sold through the platform?
- How many months does it take for a property to be sold? These are one of the questions

Statistical analysis like checking for relationships between the features might also be included but this is out of the scope of my current analysis.

V. **Thought process for the data visualization**

With the data visualization, the idea is to have high-level charts on the top, and as I move along, more specific charts are created that is filtered based on the selections on the high-level charts. This format also helps to easily drill-down on the data to discover unique patterns, and general insights.

Link to the Data Visualization:
https://prod-apnortheast-a.online.tableau.com/t/graceestrada/views/TM-BrazilRealEstateListing_v2/Dashboard?:origin=card_share_link&:embed=n