Dataset Setup Guide

This guide will help you download and set up the Brazilian E-Commerce Public Dataset for the BrazilRetail-BI project.

Prerequisites

- Python 3.8+
- Kaggle account
- Kaggle API credentials

Step 1: Install Kaggle CLI

Install the Kaggle command-line tool:

```
pip install kaggle
```

Step 2: Get Your Kaggle API Credentials

1. Go to Kaggle.com and sign in to your account
2. Click on your profile picture (top right) ·"Account"
3. Scroll down to the "API" section
4. Click "Create New API Token"
5. This will download a `kaggle.json` file containing your API credentials

Step 3: Place API Credentials

1. Move the downloaded `kaggle.json` file to your project root directory:

```
BrazilRetail-BI/
··· kaggle.json
··· ...
```

2. **Security Note:** The `kaggle.json` file contains sensitive information. Make sure it's added to your `.gitignore` file (which it should be with the current configuration).

Step 4: Download the Dataset

From your project root directory, run:

```
kaggle datasets download -d olistbr/brazilian-ecommerce
```

This will download the `brazilian-ecommerce.zip` file to your current directory.

Step 5: Extract the Dataset

Unzip the downloaded file:

```
unzip brazilian-ecommerce.zip
```

This will create several CSV files in your current directory.

Step 6: Organize Data Files

1. Create the `data` directory if it doesn't exist:

   ```
   mkdir -p data
   ```

2. Move all the CSV files to the `data` directory:

   ```
   mv *.csv data/
   ```

3. Clean up the zip file:

   ```
   rm brazilian-ecommerce.zip
   ```

Final Directory Structure

After setup, your project should look like this:

```
BrazilRetail-BI/
··· data/
·   ··· olist_customers_dataset.csv
·   ··· olist_orders_dataset.csv
·   ··· olist_order_items_dataset.csv
·   ··· olist_products_dataset.csv
·   ··· olist_sellers_dataset.csv
·   ··· olist_order_payments_dataset.csv
·   ··· olist_order_reviews_dataset.csv
·   ··· product_category_name_translation.csv
·   ··· ...
··· docs/
·   ··· dataset_setup.md
··· etl/
··· db_schema/
··· .env
··· requirements.txt
··· ...
```

Verification

You can verify the setup by checking that all CSV files are present:

```
ls -la data/
```

You should see 9 CSV files from the Brazilian E-Commerce dataset.

Next Steps

Once the data is set up, you can proceed with:

1. Running the ETL pipeline (`python etl/main.py`)
2. Creating the database schema (`python db_schema/create_schema.py`)
3. Building your dashboards and analytics

Troubleshooting

- **Permission denied**: Make sure `kaggle.json` has the correct permissions: `chmod 600 kaggle.json`
- **API errors**: Verify your Kaggle account has API access enabled
- **Download fails**: Check your internet connection and Kaggle API limits
- **Files not found**: Ensure you're running commands from the project root directory