Email Bot Detection Analysis

This repository contains a comprehensive analysis pipeline for detecting potential bot behavior in email engagement data.

📁 File Structure

```
email-bot-detection/
├── README.md
├── main_email_bot_analysis.py      # Main script to run complete analysis
├── 01_setup_and_imports.py         # Package installation and imports
├── 02_database_connection.py       # Database connection and data extraction
├── 03_data_preprocessing.py        # Data cleaning and feature engineering
├── 04_exploratory_data_analysis.py # EDA and statistical analysis
├── 05_visualization_functions.py   # Visualization and plotting functions
├── 06_bot_detection_analysis.py    # Bot detection algorithms and analysis
├── 07_summary_and_export.py        # Summary reporting and data export
└── output/                          # Directory for exported results
    ├── email_engagement_sample.csv
    ├── suspicious_accounts.csv
    ├── browser_bot_statistics.csv
    └── summary_statistics.csv
```

🚀 Quick Start

Option 1: Run Complete Analysis

```
bash

python main_email_bot_analysis.py
```

Option 2: Run Individual Components

Execute files in order:

```
bash

python 01_setup_and_imports.py
python 02_database_connection.py
python 03_data_preprocessing.py
# ... continue with other files
```

📊 Analysis Components

1. Data Extraction

- Connects to PostgreSQL/Redshift database
- Extracts email engagement data (sends, opens, clicks)
- Joins multiple tables for comprehensive view

2. Data Preprocessing

- Converts timestamp fields to datetime
- Creates time difference features
- Handles categorical variables
- Creates bot detection flags

3. Exploratory Data Analysis

- Descriptive statistics
- Email frequency analysis
- Time-based pattern analysis
- Correlation analysis

4. Visualizations

- Time distribution plots
- Correlation heatmaps
- Browser analysis
- Pairplot analysis

5. Bot Detection

- Fast open detection (< 10 seconds)
- Immediate click analysis
- Suspicious account identification
- Browser/device pattern analysis

6. Results Export

- Sample dataset export
- Suspicious accounts list

- Browser statistics
- Summary report

🔧 Requirements

```python
pandas>=1.3.0
numpy>=1.21.0
matplotlib>=3.4.0
seaborn>=0.11.0
sqlalchemy>=1.4.0
psycopg2>=2.8.0
evidently>=0.1.0
missingno>=0.5.0
```

📈 Key Features

- **Bot Detection**: Identifies potential automated behavior
- **Performance Optimized**: Handles large datasets efficiently
- **Comprehensive Visualizations**: Multiple chart types for insights
- **Export Ready**: Results saved in CSV format
- **Modular Design**: Each component can be run independently

🤖 Bot Detection Criteria

The analysis identifies potential bots based on:

- Opens occurring within 10 seconds of email send
- Simultaneous open and click events
- Consistent timing patterns across multiple emails
- Unusual browser/device combinations

📊 Output Files

- **email_engagement_sample.csv**: Sample of processed data
- **suspicious_accounts.csv**: List of potentially suspicious email accounts
- **browser_bot_statistics.csv**: Bot behavior statistics by browser
- **summary_statistics.csv**: High-level summary metrics

🔒 Security Notes

- Database credentials should be stored in environment variables
- Consider using configuration files for production deployments
- Implement proper access controls for sensitive data

📝 License

This project is for internal use and analysis purposes.

👥 Contributors

Data Science Team - Email Analytics Division

For questions or issues, please contact the Data Science team.