Repository and Mining of Temporal Data

RepoMining

User Manual

Team Members:
Jessica Nguy: jnguy2014@my.fit.edu
Siomara Nieves: snieves2014@my.fit.edu

Faculty Sponsor:
Dr. Philip Chan: pkc@fit.edu

Project Website:
https://jlnguy.github.io/RepoMining/

Table of Contents

Introduction

Repository and Mining of Temporal Data is an application designed to help users understand more about their data. It focuses on answering three main questions:

1.) Is there a significant change in the target variable?
2.) Why was there a significant change, and what are the top-k variables that have affected the target variable?
3.) What is the value in the next timestamp?

RepoMining aims to simplify the visualization and analysis of information (in the form of a .csv file with timestamps and data) by answering the previous questions, and by creating a system that is simple, easy to navigate, and easy to use.

Eight (8) visualizations will be returned to the user: Four (4) in Question 1, three (3) in Question 2, and one (1) in Question 3. The program analyzes data selected against all the data in the program provided that it has the same granularity and timestamp thresholds. The application is hosted on the web using Localhost.

The application has two types of users, Data Providers and Data Consumers. The first type of user contributes to the system by uploading their files, and the second type of user uploads the files (target variables) they want the program to analyze and obtain results on it.

The first question is answered by 4 graphs: time vs values, time vs change in values, time vs z-score, and time vs change in z-score. The visualizations include a zoom option for viewing the points and color coding in order to detect the most change. For the purposes of the program, Green indicates Low/No Change, Yellow indicates Some change, and Red indicates areas of High change and areas of interest.

The second question shows the top-k variables which could be affecting the target variable (of the Data Consumer) which is set to a default of three top variables. There top-k variables are picked depending if they have the same granularity (daily/monthly/yearly timestamps); if they fall within the same period of time; if they have data before the starting timestamp of the target; if they have data after the ending timestamp (for answering the third question using Linear Regression); and if they have correlation at all.

The first graph shows time vs the values of all 4 files, the target and other three variables in order to detect some type of pattern between them (ex. As the target variable goes up, the other goes down). The second graph shows the target vs the correlation values generated of the

three variables, with the scale being from 0.0 to 1.0. The closer to 1.0 a score is, the more correlated to the target variable it is. In only one case should the correlation be exactly 1.0: the target variable calculating correlation to itself. The correlation values are presented in absolute values. The last graph shows the top-1 correlation values calculated for each lag time.

Currently the program assumes a lag time of 5, with whatever the granularity indicates. For example, a granularity type of Yearly would assume that the lag time would go as far back as 5 Years. With a granularity type of Daily the lag time would go as far back as 5 Days.

The last question consists on only 1 graph which aims to 'predict' or forecast the next possible value for the target (whether it will go up or down in the next time period). It shows the time vs values and adds the newly generated point in red.

Glossary

Data Consumers - Users that are looking to use RepoMining to analyse their data. May or may not consist of uploading their own data to analyze, or searching through the repository for data to analyze.
Data Provider - Users looking to help populate the repository by uploading their own files.
Granularity - Refers to the timescale of a dataset (daily, monthly, yearly).
Q1 - Shorthand for Question 1.
Q2 - Shorthand for Question 2.
Q3 - Shorthand for Question 3.

Getting Started

Setting Up the System

Download the repository

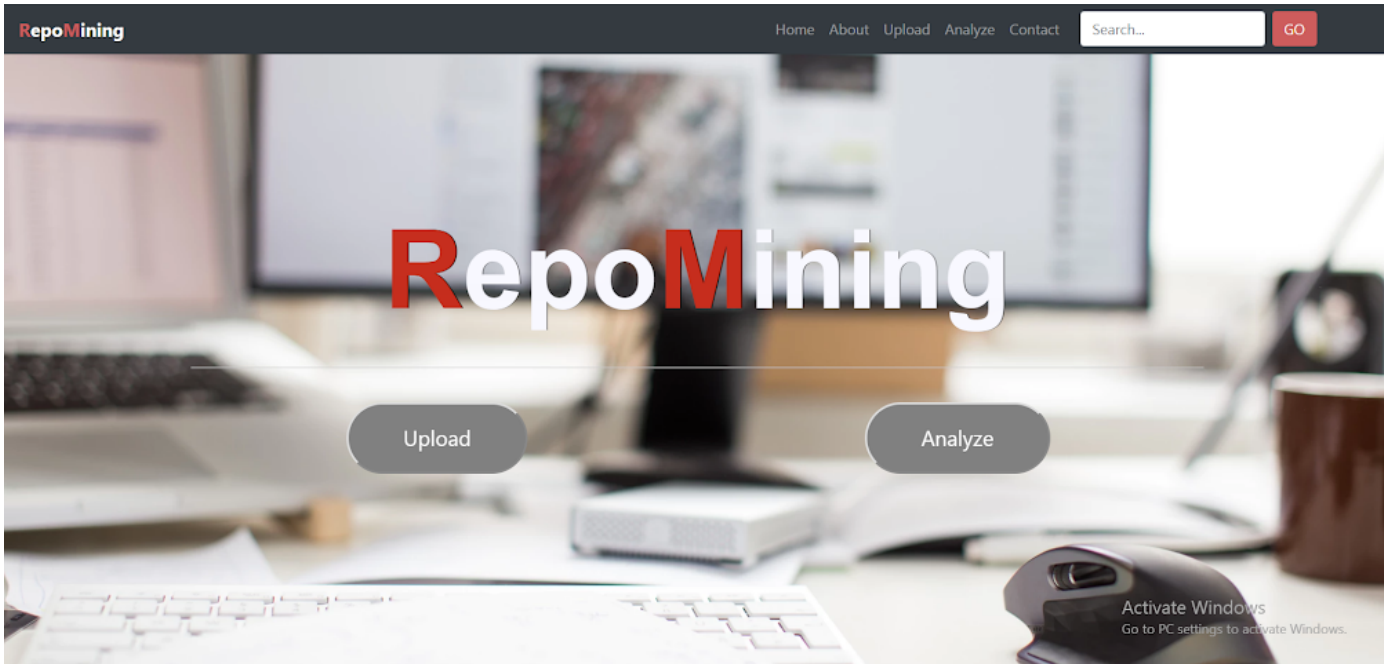
Open cmd (or terminal)

Go in the ‘repo’ folder

cd repo

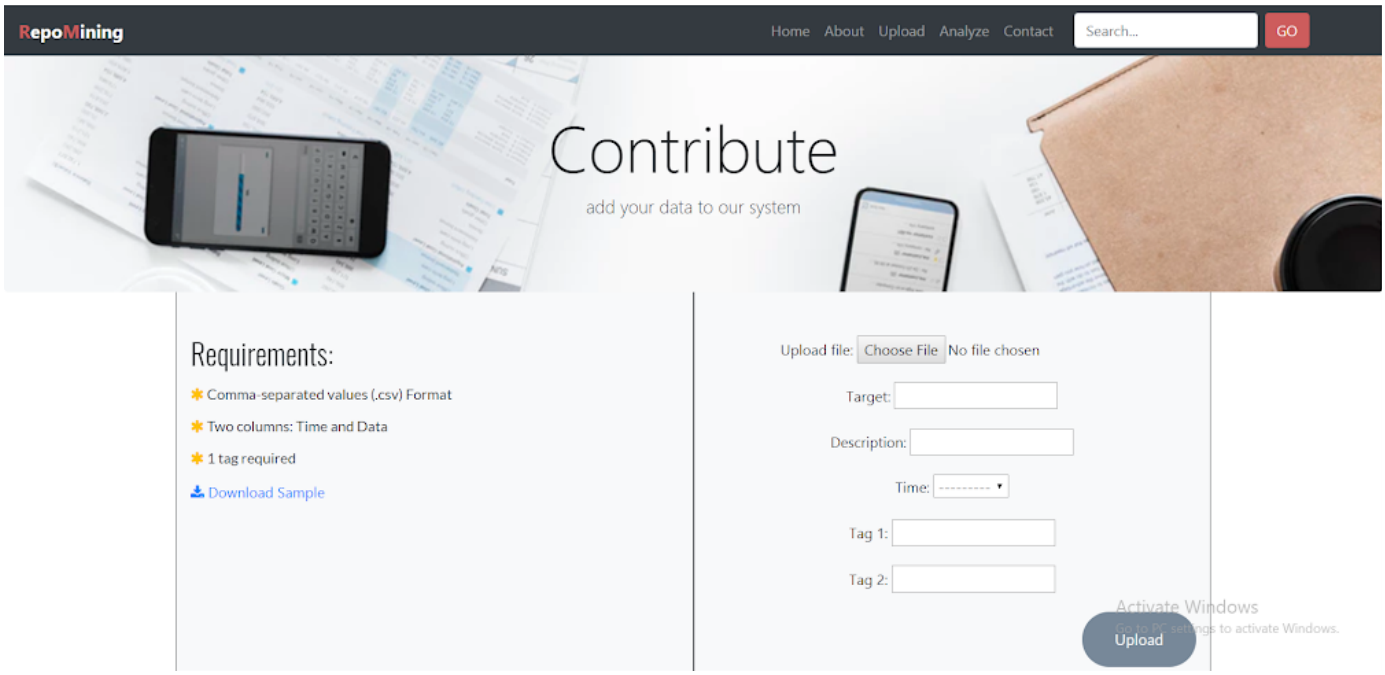
Type python manage.py runserver

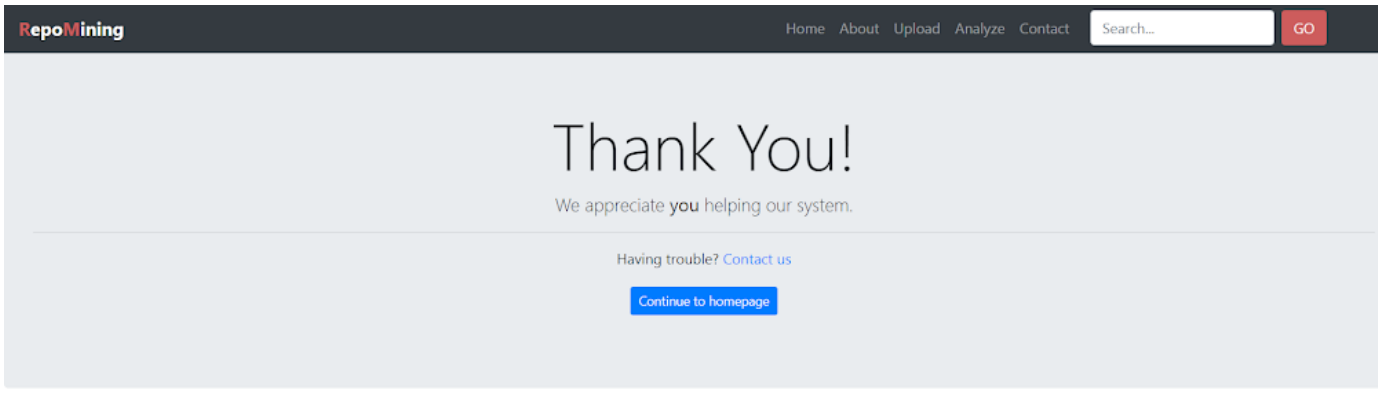
The server will be available at 127. .0.0.1:800 (LocalHost)

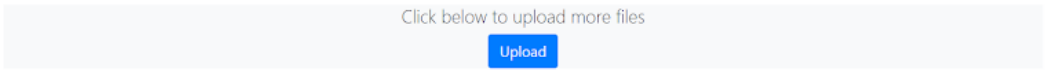


Uploading a file

Click the Upload button on the navigation bar or on the homepage
Select a file (.csv file)
Add metadata (target variable, description, tags, granularity)
Upload
If the file is accepted, it will be saved on the system's database and redirected to a thank-you page

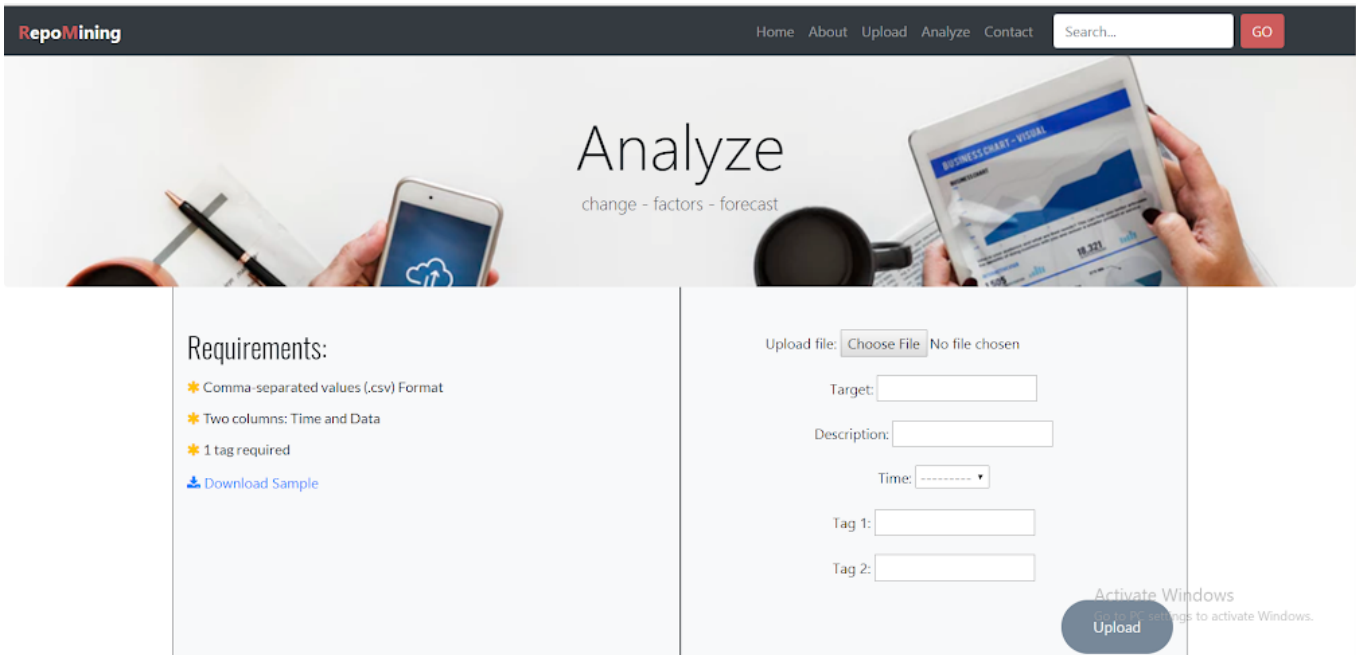






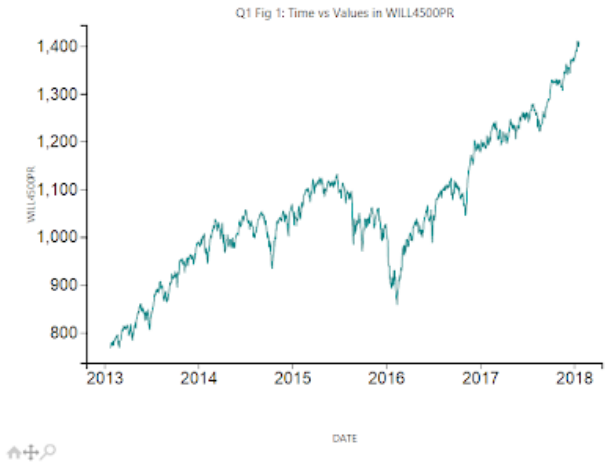
Analyzing a Target

Click the Analyze button on the navigation bar or on the homepage
Select the target file to analyze
Add metadata (target variable, description, tags, granularity)
Click Upload
Wait (less than 1 minute, depending on the file size)
Results will be displayed on the new page (Q1, Q2, Q3)



Moving points in graph

Hover over graph (Q1, Q2, or Q3)
Click the pan button
Move around the graph

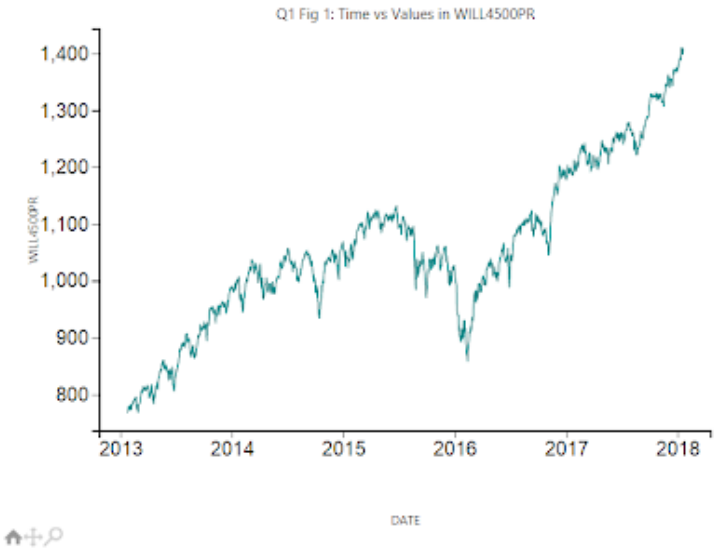


Zooming in points in graph

Hover over graph (Q1, Q2, or Q3)
Click the magnifier button on the lower left
Select the points

Reset graph

Hover over graph (Q1, Q2, or Q3)
Click the house button on the lower left



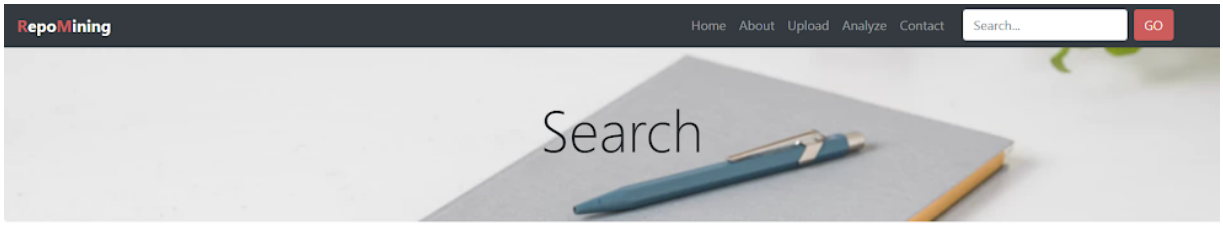
Interpreting Results

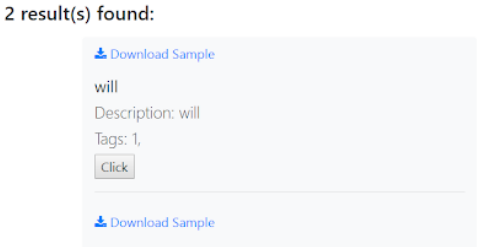
At the moment of writing this document, the website does not handle showing the predicted vs actual results of the analysis, but can be viewed on the Command-Prompt/Terminal.

```
Analysis:
R^2 prediction:  0.8582406760014895
---
Mean Absolute Error: 41.57615061070526
Mean Squared Error: 2775.5948483795796
Root Mean Squared Error: 52.68391451268195
Percent Error: 4.051213614638982
Percent Difference: 8.102427229277964
---

   Coefficient
0     0.110758
1     0.034280
2     0.034280

Last timestamp on graph:  2018-01-18 00:00:00
Last value on graph:  1404.46
X-Value:  2018-01-19 00:00:00
Y-Value:  1322.0403898712543
```

Target Variable Search

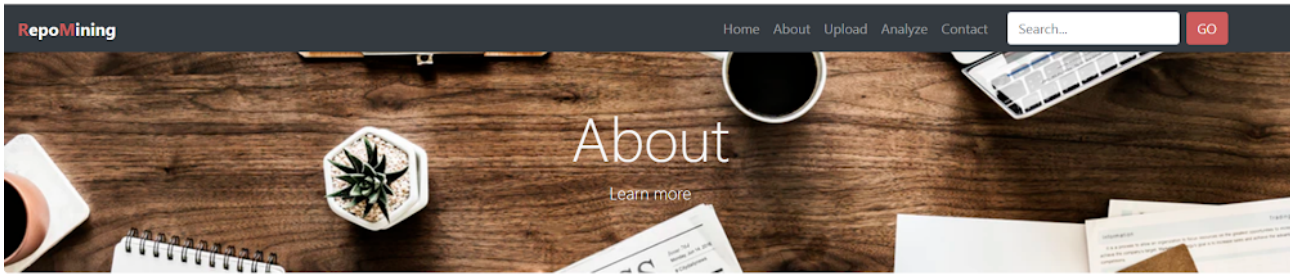
Go on the navigation bar (while on any page of the website)
Type on the search bar
Click enter
Visualize the search result





Navigating through the site

Go around the site as any other webpage
All buttons work as desired redirecting to the correct pages



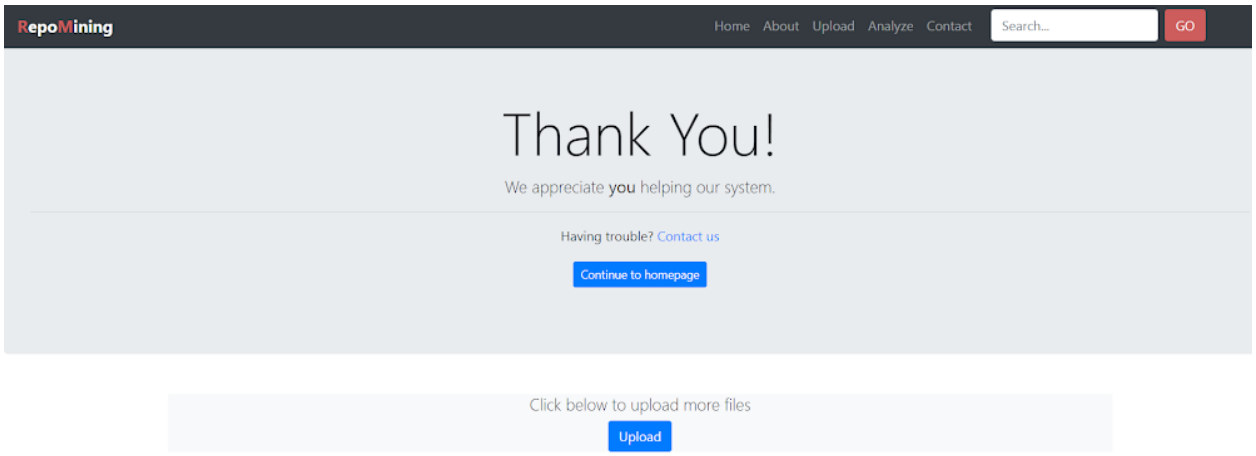
RepoMining is a web-application designed to simplify the appearance of **data mining**, **collection** and **analysis** as well as be able to predict future trends based on a variety of types of information.

The system saves .csv files uploaded by Data Providers and allows users to calculate and visualize the change in their target data, the correlation between variables that are affecting it, and the next possible date and value in the future.

Example for Data Providers

Contributing with their data to the system:

1. Click the ‘Upload’ button on the home page or navigation bar
2. Choose a file (.csv format, two columns)
3. Fill out metadata (target, description, granularity, tags)
4. Click ‘Upload’
5. If performed correctly, a thank you page should load.



Example for Data Consumers

Analyzing their data:

1. Click the ‘Analyze’ button on the homepage or navigation bar
2. Choose the file to be analyzed (.csv format, two columns)
3. Fill out the metadata (target, description, granularity, tags)
4. Click ‘Upload’
5. Next page will show the results.