directories and files:
Code and commands for data ingestion in a /data_ingest directory
ETL/cleaning code in a /etl_code directory
Profiling code in a /profiling_code directory
Screen shots that show my analytic running in a /screenshots directory - for every step

Build my code:
After cleaning I get 19 attributes about the url. In the mapreduce jobs, I calculated the average values of each 19 attributes when the url is a phishing website and when url is not a phishing website. This new data can be further compared to see which attributes may affect the result of if the url is a phishing website.

Run:
My code is simply two mapreduce jobs.
It can be run with the code:
hadoop jar Project.jar Project ProjectInput/Input.csv ProjectOutput
hadoop jar Project_1.jar Project_1 ProjectInput/Input.csv ProjectOutput_1

Result:
Result for analytics can be found in ProjectOutput and ProjectOutput_1

Input:
Input data for analytics and profiling can be found in user/sl7228/ProjectInput/Input.csv
Input data for profiling before cleaning can be found in user/sl7228/hw7/input/index.csv