# Receipt identification

## Project Background and Description

The goal of this project is to distinguish receipts from a particular STORE NAME vs others with 55%+ accuracy. Example: “Is this receipt a “Walmart” receipt – Yes/No?” The project files include 6,000+ receipt images taken from various camera angels, resolutions, etc, and are named according to the record id within the training/test dataset. Note: extracting information from the receipt such as UPC, dates, etc are optional and can be included as part of your results.csv submission.

## Project Scope (Python, R, Java, NodeJS, Tensorflow, OpenCV, Tesseract or similar)

The project can be completed using your language of choice, Python (preferred), R, Java, NodeJS, etc, but must use Tesseract, OpenCV or similar imaging engine/codebase.

## High-Level Requirements

The final project can include any additional features/functionality that is not listed below, however it must include the below items.

Your project must include the following:

* Correctly identify “Walmart” receipts vs “Non-Walmart” using Tesseract, OpenCV or similar library/engine. Can use any of the following for training your ML (Logo, Text, pattern matching, etc)
* When identifying “Walmart” receipts an accuracy of 0.55 or greater should be achieved when using Training vs Test data and images. For example: Every 100 receipts in the Test Data Set should have 55 correctly identified.
* Output should be saved as CSV (**results\_sample.csv –** [**download sample**](https://drive.google.com/file/d/0B4Ea5ipf_nEnUVc5bWFCTmx1LVE/view?usp=sharing)) in the following format

**EXT\_ID WalmartReceipt PredictionScore**

58baea8be4b0d23e37735ea9 TRUE 0.82

58baea8be4b0d23e37735ea9 FALSE 0.02

* **Optional - For those who wish to attempt, identify and extract any of the following information:**
  + UPC (12 digit number)
  + Date of Purchase
  + Store Address
  + Number of Items
  + Subtotal
  + Tax

## Resources

**CSV’s:** [**Training Data**](https://drive.google.com/file/d/0B4Ea5ipf_nEnc1laSzVrZ2VORWc/view?usp=sharing)  **/** [**Test Data**](https://drive.google.com/file/d/0B4Ea5ipf_nEnaXl6c2k5dFltS1U/view?usp=sharing)

**Note: If issues unzipping all 6,000 images, you can** [**Download Folder Here**](https://drive.google.com/drive/folders/0B4Ea5ipf_nEnMndLYzFWRHgyeU0?usp=sharing)

**Receipt Images 8GB** (17GB uncompressed): [Download Here](https://drive.google.com/file/d/0B4Ea5ipf_nEnRFVfcVFqTzZQd1E/view?usp=sharing)

Additional Links:

* <https://groups.google.com/forum/#!topic/tesseract-ocr/WHSFASNNEoQ>
* <https://github.com/mre/receipt-parser>
* <https://github.com/tesseract-ocr/tesseract/wiki/ImproveQuality>
* <http://stackoverflow.com/questions/31633403/tesseract-receipt-scanning-advice-needed>
* <https://github.com/pantacuzino/receipts/issues/2>

## Submitting your Project

**Deadline for project submission is May 1, 2017**

Upload your project (excluding images) to GitHub (or other file sharing site) and send an email to **internship@aeonpredictive.com** with **Subject: ”Data Science Challenge”** with a **URL** to your project.