**Zoetis Capstone – Transition Document**

Ayushee Nigam | Chengdao Yang

Dingran Lu | Emilia Dwyer

Spring 2019

The following are descriptions of the files handed off in May 2019 including summaries of logic flow, package requirements, and use going forward.

**pdf\_csv.ipynb**

* Purpose: convert third party incoming PDFs into CSV files
* Package Requirements: tabula-py, PyPDF2
* Use:
  1. Create a folder in your own directory called *Data*
  2. In the first variable, *path\_master*, assign it to the folder directory from previous step
  3. Drop incoming files into new directory
  4. Run

**Antibiotics\_by\_Risk.ipynb**

* Purpose: merge Antibiotics by Risk CSVs
* Package Requirements: regex
* Use:
  1. In the first variable, *path\_master*, on line 4, assign it to the same directory as pdf\_csv.ipynb’s variable with the addition of */Csv/*
  2. Run *Setup* block
  3. Run *General Functions* block
  4. *Format Functions* block:
     + Represents functions to read in different types of CSV patterns that could appear when converted from PDFs. It is not guaranteed that this list is exhaustive
     + *type3* function is the ‘base case’ that all first pages are read in as but outside of this, all other formats were discovered by inspecting the processed CSVs manually in Python.
     + It is recommended to inspect all CSVs and write each its own function at this point. Can use previous cases in notebook if they appear and add new format functions as they appear
  5. *Clean* block:
     + All *file\_num*==1 will be read in using *type3*
* Read in each new file with the appropriate *type#* formats defined from *Format Functions* block
  1. Run *Additives* block
  2. View fully merged dataframe in *Final df* block and saved out file under title *df\_ABR.csv* in directory

**Statistical\_learnig\_ABR.ipynb**

* Purpose: setup ML analyses for future that will become useful as more data is acquired
* Package Requirements: pandas, numpy, sklearn, and plotly
* Use:
  1. Simply change the path in the 4th cell of the notebook,
  2. Click and run all cell.
  3. All graphs are interactive for users to do EDA
  4. This ipython notebook’s main purpose is EDA and some ML potientially