Washington County

Notes: This python scripts scrapes PDFs and creates a csv output file with the following items: **PIN, Total Tax, Total Paid, Total Due**. The first page of the PDF contains not tax data and is skipped. The output filename is not unique and is static.

PDF

Lender Property Association Report

Service Company	Group Code	Lender Code	PIN	Owner Name	Loan Number	Total Tax	Total Paid	Total Due
9710000		9716000	07.029.20.43.0008	NAUGLE TODD D & AMY L	1092	5,082.00	0.00	5,082.00
9710000		9716000	12.032.20.31.0005	FIRKUS JEFFREY A	1092	3.00	0.00	3.00
9710000		9716000	15.027.21.23.0012	HUGHES TIMOTHY A & JENNIFER	1092	3,516.00	0.00	3,516.00
9710000		9716000	29.031.21.13.0003	MARKS LISA L	1092	2,370.00	0.00	2,370.00
					9716000 Total:	10,971.00	0.00	10,971.00
9710000	0	900652	02.028.20.12.0038	CHUDYK ANN I	1090	2,084.00	0.00	2,084.00
9710000	0	900652	02.029.20.31.0003	KRENGEL KRIS	1090	934.00	0.00	934.00
9710000	0	900652	02.029.20.31.0005	KRENGEL KRIS	1090	290.00	0.00	290.00
9710000	0	900652	03.029.20.24.0014	PATRICK RUSSELL A & JACQUELYN	1090	3,034.00	0.00	3,034.00
9710000	0	900652	03.029.20.44.0024	ANDERSON CARY J	1090	1,466.00	0.00	1,466.00
9710000	0	900652	03.029.20.44.0030	ANDERSEN ARTHUR W II	1090	988.00	0.00	988.00
9710000	0	900652	03.029.20.44.0074	BORICH KEVIN A	1090	1,170.00	0.00	1,170.00
9710000	0	900652	04.029.20.11.0045	FISCHER ABBEY	1090	800.00	0.00	800.00
9710000	0	900652	04.029.20.11.0047	FISCHER JAMIE H & JILL A	1090	734.00	0.00	734.00
9710000	0	900652	04.029.20.11.0093	LIND JUNE E	1090	1,428.00	0.00	1,428.00
9710000	0	900652	04.029.20.11.0131	FLANDERS COURT LLC	1090	1,998.00	0.00	1,998.00
9710000	0	900652	04.029.20.11.0144	CLAUSON JUDITH M	1090	1,752.00	0.00	1,752.00
9710000	0	900652	04.029.20.13.0043	VERNON DAVID A & MONIQUE R	1090	2,948.00	0.00	2,948.00
9710000	0	900652	04.029.20.14.0069	OVADIA CLIFFORD G & KIMBERLEE M NELSC	1090	3,336.00	0.00	3,336.00
9710000	0	900652	04.029.20.21.0078	RANDALL LAURA A	1090	1,916.00	0.00	1,916.00
9710000	0	900652	04.029.20.23.0011	KNUTSON ROBERT A & ROSEMARY	1090	2,342.00	0.00	2,342.00
9710000	0	900652	04.029.20.32.0055	OTLEY ANTHONY J & KIMBERLY	1090	2,518.00	0.00	2,518.00
9710000	0	900652	07.029.20.11.0008	DIETERLE JASON P & NATALIE A	1090	14,112.00	0.00	14,112.00
9710000	0	900652	07.029.21.22.0021	BETTER THAN PAR DIVERSIFIED LLC	1090	3,108.00	0.00	3,108.00
9710000	0	900652	08.029.20.42.0005	WETTEROFF MARYLIN A	1090	1,830.00	0.00	1,830.00
9710000	0	900652	08.032.21.23.0021	MILDON BENJAMIN N	1090	2,972.00	0.00	2,972.00
9710000	0	900652	09.029.20.32.0006	ROSE JAMES M & SUSAN	1090	2,384.00	0.00	2,384.00
9710000	0	900652	10.029.20.11.0016	JOHNSON CHARLES D & ANN E	1090	1,718.00	0.00	1,718.00
9710000	0	900652	10.029.20.11.0063	CROWDER ERIN M	1090	1,604.00	0.00	1,604.00
9710000	0	900652	10.029.20.11.0079	NOREEN JOHN E	1090	1,398.00	0.00	1,398.00
9710000	0	900652	10.029.20.11.0101	GALLICHANT NICHOLAS A & YANDRA L RIAS	1090	2,122.00	0.00	2,122.00
9710000	0	900652	10.029.20.11.0102	GALLICHANT NICHOLAS A & YANDRA L RIAS	1090	112.00	0.00	112.00
9710000	0	900652	10.029.20.12.0033	GEDATUS MARIBETH	1090	1,410.00	0.00	1,410.00
9710000	0	900652	10.029.20.14.0005	ST CLAIRE BRIAN P & KARI A	1090	1,806.00	0.00	1,806.00
9710000	0	900652	10.029.20.14.0009	GOULETTE PAUL D & MARY C	1090	1,790.00	0.00	1,790.00
9710000	0	900652	10.029.20.14.0074	SWANGSTUE GARY A & VICKIE L	1090	1,830.00	0.00	1,830.00
9710000	0	900652	11.028.20.13.0101	THRON DENNIS W & SHEREE L	1090	8.00	0.00	8.00
9710000	0	900652	11.028.20.13.0102	THRON DENNIS W & SHEREE L	1090	2,766.00	0.00	2,766.00
9710000	0	900652	11.028.20.34.0058	SCHNEIDER JACK L	1090	1,898.00	0.00	1,898.00

Python Code (20190711_WashingtonPDF.py)

```python
# Read and dump PDF file data to text file

# Accumatch

# by Michael Keller

# June 28, 2019

# requires pdfplumber, re


# libraries

import pdfplumber

import re


# Booleans

countyFound = False


# Variables

outputType = '.csv'

outputFile = ''

comma = ','

outString = []

_return = '\r'

spaces = '\s{2,}'

space = ' '

nospace = ''

count = 0

regPIN = '\d{2}[.]\d{3}[.]\d{2}[.]\d{2}[.]\d{4}'

colHeadings = 'PIN,Total Tax,Total Paid,Total Due'


# Messages

messageRun = 'Reading input.pdf file . . .'

messageSaveBeg = 'Saving \''

messageSaveEnd = '\' file . . .'

messageEOL = '. . . end of line'


# PDF File Open

pdf = pdfplumber.open("input.pdf")
```

```python
print(messageRun)


for page in pdf.pages:

    p1 = pdf.pages[count]

    count = count + 1

    pltext = p1.extract_text()

    text = pltext.splitlines()


    for item in text:

        # Search for PIN in line read from PDF

        PIN_Found = re.search(regPIN, item)

        # Only run if PIN is found in line

        if(PIN_Found is not None):

            # Separate PIN from the rest of the line

            PIN = item[PIN_Found.start():PIN_Found.end()]

            # Grab the rest of the line using the PIN found variable

            restOfLine = item[PIN_Found.end():len(item)].strip()

            # Take out commas in amounts

            restOfLine = restOfLine.replace(comma,nospace)

            # Remove double spaces in rest of line

            restOfLine_mod = re.sub(spaces, space, restOfLine)

            # Split from the left four ways

            carp = restOfLine_mod.rsplit(' ',4)

            # Add only the last three items or amounts found from the split

            outString.append(str(PIN) + comma + str(carp[2]) + comma + str(carp[3]) + comma +
str(carp[4]))

        else:

            # Will run the first time if pin is not found in line

            if(countyFound == False):

                # Do not run this if statement again i.e. change boolean to True

                countyFound = True

                # Read first line in PDF and create filename based on the text found.

                filename = text[0].strip()

                # Create filename based on county name found in PDF

                outputFile = filename + outputType


print(messageSaveBeg + outputFile + messageSaveEnd)
```

```python
# Create and open output file

file = open(outputFile,'w')


# Add column headings to first line in file

file.write(colHeadings + _return)


# Now dump string array into file

for taxLine in outString:

    file.write(taxLine + _return)


# PDF File Close

pdf.close()

file.close()


print(messageEOL)


# EOF
```

CSV Output

	PIN	Total Tax	Total Paid	Total Due
1	PIN	Total Tax	Total Paid	Total Due
2	07.029.20.43.0008	$5,082.00	$0.00	$5,082.00
3	12.032.20.31.0005	$3.00	$0.00	$3.00
4	15.027.21.23.0012	$3,516.00	$0.00	$3,516.00
5	29.031.21.13.0003	$2,370.00	$0.00	$2,370.00
6	02.028.20.12.0038	$2,084.00	$0.00	$2,084.00
7	02.029.20.31.0003	$934.00	$0.00	$934.00
8	02.029.20.31.0005	$290.00	$0.00	$290.00
9	03.029.20.24.0014	$3,034.00	$0.00	$3,034.00
10	03.029.20.44.0024	$1,466.00	$0.00	$1,466.00
11	03.029.20.44.0030	$988.00	$0.00	$988.00
12	03.029.20.44.0074	$1,170.00	$0.00	$1,170.00
13	04.029.20.11.0045	$800.00	$0.00	$800.00
14	04.029.20.11.0047	$734.00	$0.00	$734.00
15	04.029.20.11.0093	$1,428.00	$0.00	$1,428.00
16	04.029.20.11.0131	$1,998.00	$0.00	$1,998.00
17	04.029.20.11.0144	$1,752.00	$0.00	$1,752.00
18	04.029.20.13.0043	$2,948.00	$0.00	$2,948.00
19	04.029.20.14.0069	$3,336.00	$0.00	$3,336.00
20	04.029.20.21.0078	$1,916.00	$0.00	$1,916.00
21	04.029.20.23.0011	$2,342.00	$0.00	$2,342.00
22	04.029.20.32.0055	$2,518.00	$0.00	$2,518.00
23	07.029.20.11.0008	$14,112.00	$0.00	$14,112.00
24	07.029.21.22.0021	$3,108.00	$0.00	$3,108.00
25	08.029.20.42.0005	$1,830.00	$0.00	$1,830.00
26	08.032.21.23.0021	$2,972.00	$0.00	$2,972.00
27	09.029.20.32.0006	$2,384.00	$0.00	$2,384.00
28	10.029.20.11.0016	$1,718.00	$0.00	$1,718.00

Michael Keller
Accumatch
20190711_WashingtonPDF
July 17, 2019