

## Sales List

**Notes:** This python scripts scrapes PDFs creates a CSV output file with the following items: **Item #, ID Number, County Access, Min Bid.** The filename is static and not unique.

PDF

**Sales List Advertisement for: SIERRA COUNTY**  
**SIERRA COUNTY TREASURER**  
**SIERRA COUNTY ASSESSOR**  
**Sale Location: 100 N. Date Ste 13 Truth or Consequences**  
**SALE DATE: 07-17-2019 at 10 AM**

Item #1

ID Number 21 494 2015 /County Access #:0001252  
Assessed Owner ARREY PIERRE U & ARREY MARK G  
Simple Description IN ARREY ON HWY 187  
Minimum Bid \$1,400.00  
Property Description SECT-23 TWP-17S RGE-05W 2.69 ACRES IN SE4NE4 13408 HWY 187

Item #2

ID Number 21 1337 2015 /County Access #:0017745  
Assessed Owner BLEA ISIDRO  
Simple Description IN T OR C OFF 3RD AVE  
Minimum Bid \$500.00  
Property Description YUCCA HEIGHTS BLK 1 667 SQ FT IN LOT 19 3775 SQ FT IN LOT 20

Item #3

ID Number 21 2416 2015 /County Access #:0015288  
Assessed Owner CAUGHRON INVESTMENTS LLC  
Simple Description IN T OR C ON WATERDOG RD  
Minimum Bid \$900.00  
Property Description MONTICELLO CANYON SUB PHASE II LOT 24 46 WATERDOG RD

Item #4

ID Number 21 2657 2015 /County Access #:0001049  
Assessed Owner DAHLIN TERESA  
Simple Description IN T OR C ON 3RD ST  
Minimum Bid \$3,200.00  
Property Description ORIGINAL TNS BLK 69 N2 OF LOTS 1 & 2 DBA: DEBBIES DANCE STUDIO  
117 3RD

Item #5

ID Number 21 3365 2015 /County Access #:0003792  
Assessed Owner DARM LYNDA L  
Simple Description IN TOR C ON BIRCH ST  
Minimum Bid \$900.00  
Property Description ORIGINAL TNS BLK 24 LOT 1, 2 620 BIRCH

Item #6

ID Number 21 3423 2015 /County Access #:0008255  
Assessed Owner DAVIS SUSAN IRENE  
Simple Description IN TOR C ON 7TH AVE  
Minimum Bid \$2,800.00

## Python Code (20190715\_SalesList.py)

```
# Sales List
# Accumatch
# by Michael Keller
# July 15, 2019
# requires pdfplumber, re, datetime

# libraries
import re
import pdfplumber
from datetime import datetime

#start timer
start = datetime.now()

# Messages
messageRun = 'Reading \'input.pdf\' file . . .'
messageSaveBeg = 'Saving \''
messageSaveEnd = '\' file . . .'
messageEOL = '. . . end of line'

# Boolean
countyFound = False
itemFound = False

# Regular expressions
regNumbers = '[^0-9]'

# Variables
outputType = '.csv'
outputFile = ''
filename = ''
count = 0
outString = []
itemNumberStart = 'Item'
idNumberStart = 'ID Number'
```

```

minimumBidStart = 'Minimum Bid'
tempString = ''
space = ' '
nospace = ''
comma = ','
_return = '\r'
colHeadings = 'Item #, ID Number, County Access, Min Bid'

# PDF File Open
pdf = pdfplumber.open("input.pdf")

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 begging of lines for data to be pulled
        itemNumber = item[0:4]
        idNumber = item[0:9]
        minBid = item[0:11]

        if(itemNumber == itemNumberStart):
            # Item #1
            carp = item.split('#',2)
            tempString = carp[1].strip()
        elif(idNumber == idNumberStart):
            # ID Number 33 8963 2015 /County Access# R08963
            carp = item.rsplit('#',2)
            carp = re.sub(regNumbers,'',carp[1])
            carp2 = item.split('/',2)
            carp3 = item[10:len(carp2[0])].strip()
            tempString = tempString + comma + carp3 + comma + carp

```

```
elif(minBid == minimumBidStart):
    # Minimum Bid $1,400.00
    carp = item.rsplit('$',2)
    carp2 = carp[1].replace(comma,nospace)
    tempString = tempString + comma + carp2
    outString.append(tempString)

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
    carp = text[0].rsplit(':',2)
    filename = carp[1].strip()
    # Create filename based on county name found in PDF
    outputFile = filename + outputType

#PDF File Close
pdf.close()

print(messageSaveBeg + outputFile + messageSaveEnd)

# Create and open output file
file = open(outputFile,'w')

# Write out lines from array
file.write(colHeadings + _return)

# Now dump string array into file
for taxLine in outString:
    file.write(taxLine + _return)

print(messageEOL)

# Close files
pdf.close()
file.close()
```

```
#end timer
```

```
print('Runtime: ' + str(datetime.now() - start))
```

```
# EOF
```

## CSV Output

	A	B	C	D
1	Item #	ID Number	County Access	Min Bid
2	1	21 494 2015	1252	1400
3	2	21 1337 2015	17745	500
4	3	21 2416 2015	15288	900
5	4	21 2657 2015	1049	3200
6	5	21 3365 2015	3792	900
7	6	21 3423 2015	8255	2800
8	7	21 3630 2015	14253	4500
9	8	21 4774 2015	4743	1600
10	9	21 5351 2015	13310	500
11	10	21 5572 2015	2677	1500
12	11	21 5903 2015	18237	400
13	12	21 6086 2015	16174	2900
14	13	21 6668 2015	286	1100
15	14	21 6825 2015	4764	2400
16	15	21 6998 2015	9310	700
17	16	21 6999 2015	6252	900
18	17	21 7479 2015	3637	2100
19	18	21 7638 2015	5676	600
20	19	21 7757 2015	4851	2500
21	20	21 8043 2015	2327	200
22	21	21 8092 2015	1924	2100
23	22	21 8114 2015	4368	2100
24	23	21 10568 2015	3100	1800
25	24	21 11033 2015	4652	1100
26	25	21 11141 2015	6189	200
27	26	21 11500 2015	5920	1400
28	27	21 11760 2015	11953	1500

Michael Keller  
Accumatch  
20190715\_SalesList  
July 17, 2019