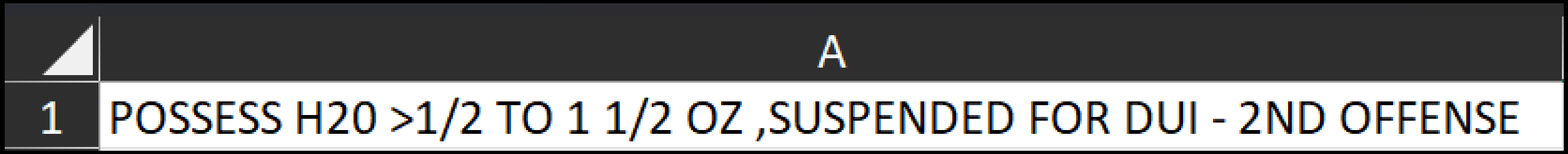
Substring lookups

- Excel
- Power Query
- Python
- R

Owen Price
linkedin.com/in/owenhprice

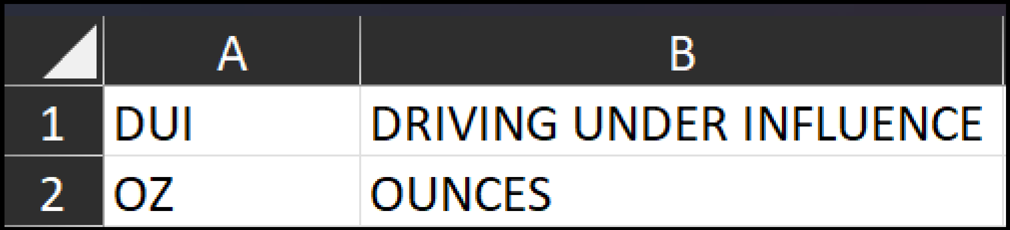
Word-based lookup of categories

Given this
(on Sheet1)

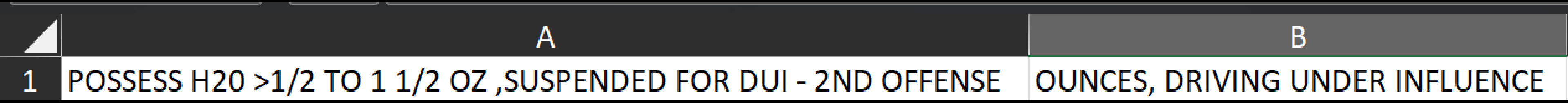


	A
1	POSSESS H20 >1/2 TO 1 1/2 OZ ,SUSPENDED FOR DUI - 2ND OFFENSE

And this
(on Sheet2)



	A	B
1	DUI	DRIVING UNDER INFLUENCE
2	OZ	OUNCES



	A	B
1	POSSESS H20 >1/2 TO 1 1/2 OZ ,SUSPENDED FOR DUI - 2ND OFFENSE	OUNCES, DRIVING UNDER INFLUENCE

Populate this

Owen Price
linkedin.com/in/owenhprice

This is a common task and each tool is competent but slightly different

The pattern is:

1. Split the input into words
2. Lookup each word in the category list
3. Join the found categories

Owen Price
linkedin.com/in/owenhprice

Excel

Split the input string into words

Lookup each word in column A of Sheet2. If found, return column B of Sheet2, otherwise “”

```
=LET(
    words, TEXTSPLIT(A1, " "),
    results, MAP(
        words,
        LAMBDA(x, XLOOKUP(x, Sheet2!A:A, Sheet2!B:B, ""))
    ),
    TEXTJOIN(", ", TRUE, results)
)
```

Join the list of found words with “, “

Power Query (M)

```
let
    Source = Excel.CurrentWorkbook(){[Name="Table1"]}[Content],

    AddColumn = Table.AddColumn(
        Source,
        "Categories",
        (ac) =>
            let
                words = Text.Split(ac[Column1]," "),
                results = List.Transform(
                    words,
                    (i)=>Table.SelectRows(
                            Table2,
                            (s)=>s[Column1]=i
                        )[Column2]{0}?
                )
            in
                Text.Combine(results,", ")
    )
in
    AddColumn
```

Split the input string into words

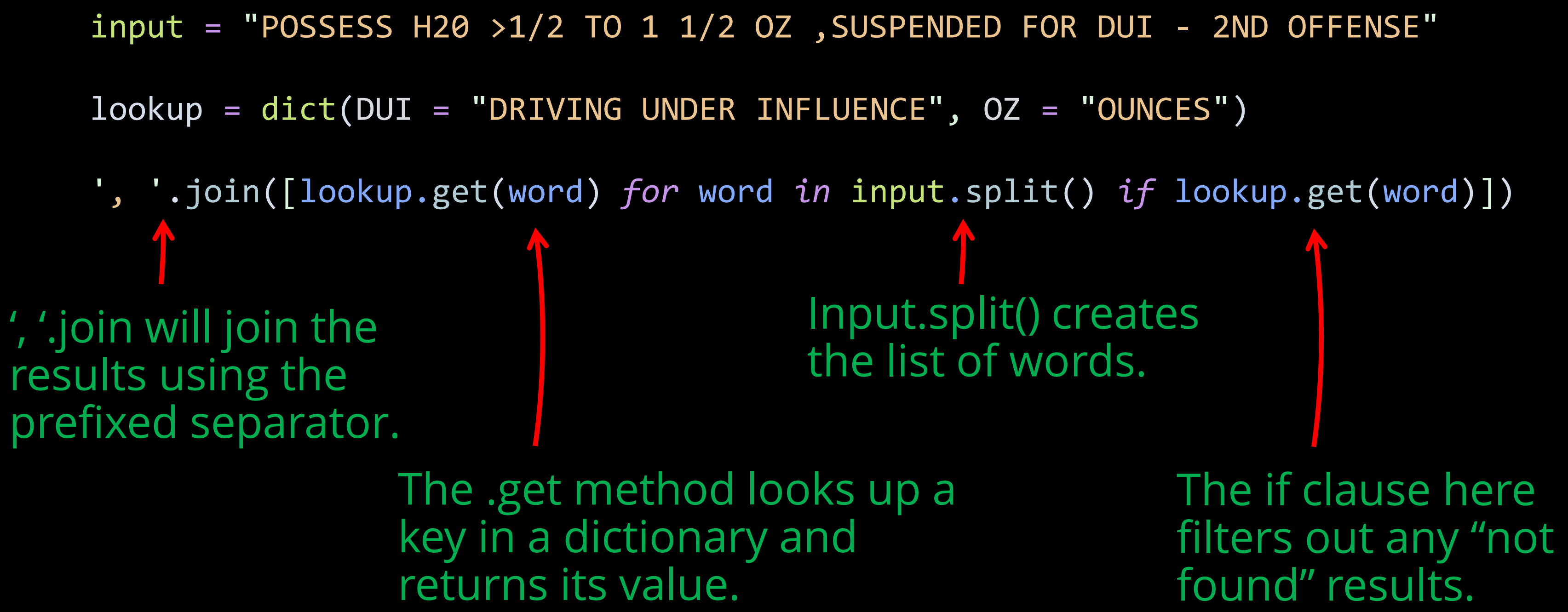
For each word, select rows from Table2 where Column1 is equal to that word. Return the value from Column2 in the first row of the result, or null if there were no rows (achieved with ?)

Join the list of found words with “, “

Owen Price
linkedin.com/in/owenhprice

Python

List comprehensions!



```
input = "POSSESS H20 >1/2 TO 1 1/2 OZ ,SUSPENDED FOR DUI - 2ND OFFENSE"

lookup = dict(DUI = "DRIVING UNDER INFLUENCE", OZ = "OUNCES")

', '.join([lookup.get(word) for word in input.split() if lookup.get(word)])
```

‘, ‘.join will join the results using the prefixed separator.

The .get method looks up a key in a dictionary and returns its value.

Input.split() creates the list of words.

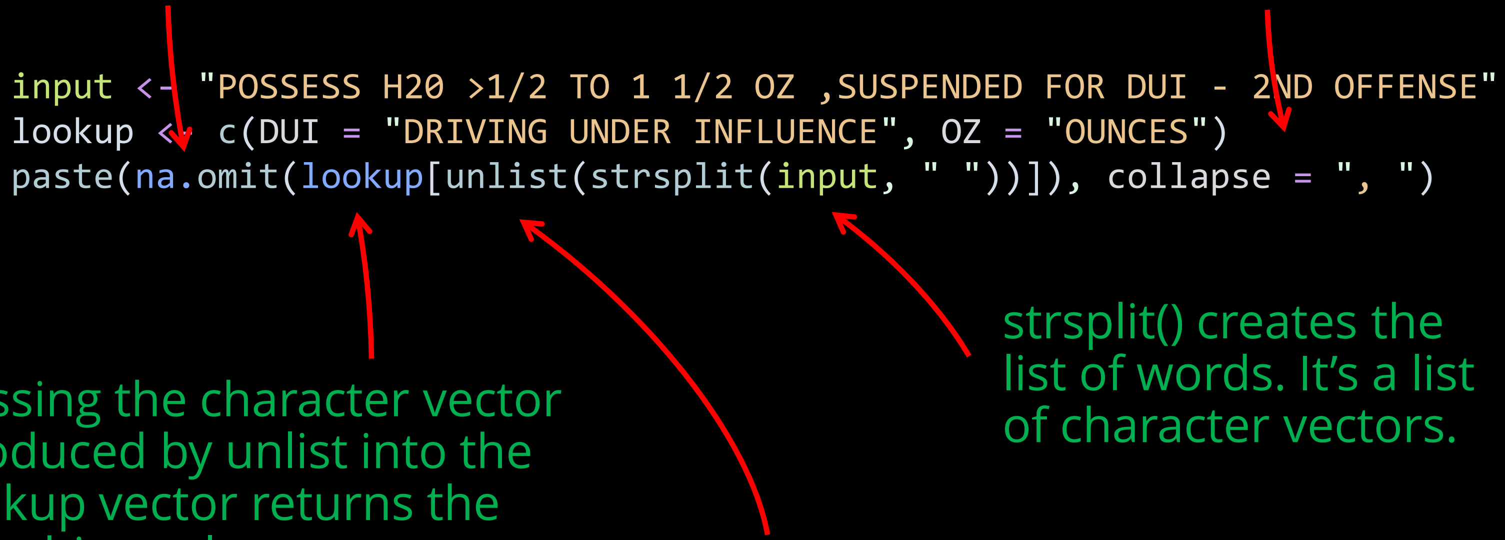
The if clause here filters out any “not found” results.

Owen Price
linkedin.com/in/owenhprice

R

na.omit removes na values (i.e. words not found)

paste with collapse creates the comma separated output



```
input <- "POSSESS H20 >1/2 TO 1 1/2 OZ ,SUSPENDED FOR DUI - 2ND OFFENSE"
lookup <- c(DUI = "DRIVING UNDER INFLUENCE", OZ = "OUNCES")
paste(na.omit(lookup[unlist(strsplit(input, " "))]), collapse = ", ")
```

Passing the character vector produced by unlist into the lookup vector returns the matching values or na

strsplit() creates the list of words. It’s a list of character vectors.

unlist converts the list of character vectors into a single character vector containing the words in the input string

Owen Price
linkedin.com/in/owenhprice

Takeaways:

1. Most tools can do what you need.
2. Some make it easier than others.
3. Understanding the steps needed is most important.

Owen Price
linkedin.com/in/owenhprice