Enhanced Data Analytics Test Work

Marie Hoie, MSBA

Agenda
- Keyword Analysis
- Fuzzy Matching
- Benford’s Law
- Data Visualization
Keyword Analysis

Why Keyword Analysis?

- Most commonly, keyword audits are done for search engine optimization
  - Increases understanding of target demographic

- In audit, the same principle applies: increasing understanding of each client and its employees
  - Most commonly occurring activity in the GL, AJEs, etc.
  - Activities you don’t expect to see
  - Keyword usage by employee
Keyword Analysis in IDEA

Columns C through G:

=IF(ISNUMBER(FIND($C$1,$B2)),"1","0")

Using keyword data from each column header, this formula searches for those keywords in column B.

Column B:

=LOWER( TRIM( SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(A2,"",""),"",""),"",""),"",""),"",""),"",""),"",""),"",""),"","")))

Using raw data from column A, this formula clears punctuation, excess spacing, and capitalization to enable keyword searches and totals in columns C through G.
Keyword Analysis in Excel

Columns C through G:

After keywords are totaled, these columns are not in the correct format to be summed; simply selecting the columns and changing to a number format won’t work.

Enter a “1” into an unused cell. Copy the cell.

Right-click on a cell you need to reformat (e.g. cell C2).

Click “Paste Special”, then select “Multiply” and hit OK.

The cell will then be appropriately formatted. Drag this formatting down to all subsequent cells in the column. Repeat this action for all keyword columns. The columns will now all sum correctly.

Fuzzy Matching
Why Fuzzy Matching?

- Identifies non-exact matches between two fields
- Results include all possible matches between a specified threshold percentage (e.g. 80% similarity) and 100% similarity
- In audit, this can be used in many capacities:
  - Ghost employees or vendors
  - Employees with undisclosed conflicts of interest
  - Year-on-year analysis when IDs or account codes don’t match

Fuzzy Matching in IDEA

- Join Employees, Vendors, and Invoice Processing files
- Calculate similarity index to assess strength of each match
- Isolate active employees who match vendors not currently listed as employees
- Target those results with current FY invoice activity
An add-in must first be installed to perform Excel fuzzy matching: “Fuzzy Lookup Add-in for Excel”

Set up both data sources to be compared
• Turn each data source into an Excel table
• The data source tables can either be in the same worksheet together or in different worksheets
• The fuzzy matching results can either be placed in the same worksheet as the tables or in a new worksheet

Select “Fuzzy Lookup” tab to start match
Designate Left Table and Right Table to populate field names below
Select field to compare between the two tables, then click the button between tables to add a fuzzy join binding
Customize the results table you want to see by removing output columns if desired
To allow only one match per line item, keep “Number of Matches” at 1; otherwise, more matches can be added
Move the similarity threshold slider to restrict the number of results
## Fuzzy Matching in Excel

<table>
<thead>
<tr>
<th>EMP ID</th>
<th>EMP NAME</th>
<th>EMP ADDRESS</th>
<th>VEND ID</th>
<th>VEND NAME</th>
<th>VEND ADDRESS</th>
<th>EMPLOYEE/Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Robin Paulden</td>
<td>727 W Center Ave, Show Low, AZ 85091</td>
<td>V0458</td>
<td>Paul Robin</td>
<td>727 W Center Ave, Show Low, AZ 85091</td>
<td>1 1.0000</td>
</tr>
<tr>
<td>2</td>
<td>Tiffany Place</td>
<td>PO Box 172 Phoenix AZ 85004</td>
<td>V0131</td>
<td>Vector Textbooks</td>
<td>PO Box 32, Phoenix, AZ 85004</td>
<td>1 0.9541</td>
</tr>
<tr>
<td>3</td>
<td>Juliana Prescott</td>
<td>10781 E Saguaroon, Scottsdale AZ 85260</td>
<td>V1577</td>
<td>Richard Gray Consulting</td>
<td>1209 W Gables Phoenix 85021</td>
<td>0 0.0000</td>
</tr>
<tr>
<td>4</td>
<td>Melody Gray</td>
<td>1209 W Gables Drive, Phoenix, AZ 85021</td>
<td>V0111</td>
<td>Glendale School District</td>
<td>PO Box 464, Glendale AZ 85307</td>
<td>0 0.7717</td>
</tr>
<tr>
<td>5</td>
<td>John Bloomfield</td>
<td>PO Box 464, Glendale AZ 85307</td>
<td>V0221</td>
<td>Shinnock Foods</td>
<td>2518 N Pheasant Pl, Mesa AZ 85012</td>
<td>0 0.6799</td>
</tr>
<tr>
<td>6</td>
<td>David Holmes</td>
<td>2518 N Pheasant Place Apt 311, Mesa AZ 85202</td>
<td>V2219</td>
<td>Shinnock Foods</td>
<td>2518 N Pheasant Pl, Mesa AZ 85012</td>
<td>0 0.6799</td>
</tr>
</tbody>
</table>

1. Vendor listed as Employee; exact similarity match, transposed first and last names
2. Vendor listed as Employee; near-identical similarity match, different PO boxes
3. No vendor match
4. Vendor not listed as Employee; 77% similarity match, same address, different names
5. Vendor listed as Employee; exact similarity match, different vendor names
6. Vendor not listed as Employee; 68% similarity match, different addresses and names

---

## Benford’s Law

[Benford’s Law Image]
Why Benford’s Law?

- Identifies abnormal distribution of natural sets of numbers to detect various data integrity issues:
  - Avoided threshold values
  - Rounding off operations
  - Number invention
  - Number padding
  - Entry mistakes

- Benford’s Law tests are only appropriate for large unrestricted and continuous number sets (e.g. JE’s, PO’s, p-cards, etc.)

- Files with less than 1,000 records should not typically be considered for these tests

Benford’s Law in IDEA

First Digit –
A high-level test for obvious reasonableness on the proportions of the first digit of a set of numbers, either positive or negative values but not both at once; the first digit can never be zero.

If the dataset fails this test, the First Two Digits and First Three Digits tests should be conducted.

First Digit test is not to be used for sample selections.
Benford’s Law in IDEA

**Second Digit**

A high-level test for conformity on the proportions of the second digit of a set of numbers, either positive or negative values but not both at once.

If the dataset fails this test, the First Two Digits and First Three Digits tests should be conducted.

Second Digit test is not to be used for sample selections.

---

**First Two Digits**

Tighter focus on the proportions of the counts of the first two digits of a set of numbers.

Particularly used for analysis of avoided threshold values.

Useful for selecting audit targets for review.
Benford’s Law in IDEA

**Summation**
Calculates the sum of the occurrences for each combination of first two digits, rather than the counts

Used in conjunction with results of the First Two Digits test.

Useful for selecting audit targets for review.

---

**First Three Digits**
Tighter focus on the proportions of the first three digits of a set of numbers; particularly used for revealing rounding off operations or number duplication

Not typically meaningful for number sets with under 10,000 records.

Very useful for sample selections.
Benford’s Law in IDEA

Last Two Digits –
Analysis of frequency of either the last two integers or first two decimals of a set of numbers.

Particularly used for situations in which number padding, number invention, or rounding is suspected.

Benford’s Law in Excel

- Many online guides discuss application of Benford’s Law in Excel
  - Association of Certified Fraud Examiners
    - PDF guide discussing statistical significance and fraud detection
  - Journal of Accountancy
    - How-To guide for manual creation of First Digit test in Excel

- Convenient template on Audit Monk set up for multiple Benford’s tests
  - First Digit
  - Second Digit
  - First Two Digits
  - First Three Digits
Benford’s Law in Excel

- Identify numeric data you wish to test
  - If largely positive data (e.g. debits), omit negative values
  - If largely negative data (e.g. credits), omit positive values

- Paste numeric data into column B of the template and all Benford’s data will be auto-generated in subsequent tabs

Apply conditional formatting to help ID values in excess of expected First and Second digit distributions
Benford’s Law in Excel

- Apply conditional formatting to “Top X” number of greatest differences for First 2 and First 3 Digit distributions
- Use these results for sampling

Data Visualization
Why Data Visualization?

- Graphical representation of data that is easier to digest than the data source tables or spreadsheets from which they’re made
  - Tell a story about their source data
  - Allow for quicker/easier detection of outliers and trends
  - Increase overall understanding of “normal” client activity
- Many tools exist to make imaginative viz’s easier to create
  - Tableau - https://usergroups.tableau.com/phoenix
  - Power BI

### Tableau Visualizations

AJE Keywords by User & Reference

<table>
<thead>
<tr>
<th>User</th>
<th>Keyword</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>USER 1</td>
<td>CORRES</td>
<td>JE InTouch</td>
</tr>
<tr>
<td>USER 2</td>
<td>CORRES</td>
<td>JE InTouch</td>
</tr>
<tr>
<td>USER 3</td>
<td>CHANGE</td>
<td>JE InTouch Transfers</td>
</tr>
<tr>
<td>USER 3</td>
<td>CHANGE</td>
<td>JE Other</td>
</tr>
<tr>
<td>USER 3</td>
<td>CHANGE</td>
<td>JE Payroll Code Correction</td>
</tr>
<tr>
<td>USER 3</td>
<td>CHANGE</td>
<td>JE Code Correction</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>USER</th>
<th>TRANSF</th>
<th>JE InTouch Transfers</th>
</tr>
</thead>
<tbody>
<tr>
<td>USER</td>
<td>TRANSF</td>
<td>JE Other</td>
</tr>
<tr>
<td>USER</td>
<td>TRANSF</td>
<td>JE Payroll Code Correction</td>
</tr>
<tr>
<td>USER</td>
<td>TRANSF</td>
<td>JE Code Correction</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>USER</th>
<th>USER 3</th>
<th>JE InTouch Transfers</th>
</tr>
</thead>
<tbody>
<tr>
<td>USER</td>
<td>USER 3</td>
<td>JE Other</td>
</tr>
<tr>
<td>USER</td>
<td>USER 3</td>
<td>JE Payroll Code Correction</td>
</tr>
<tr>
<td>USER</td>
<td>USER 3</td>
<td>JE Code Correction</td>
</tr>
</tbody>
</table>

**Multi-Axis Tornado Chart**

- **Number of Records**: 1,094
- **Debt**: $9,157,495

- **User Contributions**:
  - USER 1: $325
  - USER 2: $322
  - USER 3: $279,642
  - USER 4: $9,000
  - USER 5: $2,256
  - USER 6: $2,850

- **Keywords**:
  - JE InTouch: $325
  - JE InTouch Transfers: $279,642
  - JE Other: $9,000
  - JE Payroll Code Correction: $2,256
  - JE Code Correction: $2,850

- **Reference**:
  - USER 1: JE InTouch
  - USER 2: JE InTouch
  - USER 3: JE InTouch Transfers
  - USER 3: JE Other
  - USER 3: JE Payroll Code Correction
  - USER 3: JE Code Correction

- **Notes**:
  - USER 4: JE InTouch Transfers
  - USER 5: JE Other
  - USER 6: JE Payroll Code Correction
  - USER 6: JE Code Correction

- **Data Source**:
  - https://usergroups.tableau.com/phoenix
Tableau Visualizations

Multi-Category Stacked Bar Chart

Top 10 Paid Employees in Top 4 Paycodes

Questions?

Marie Hoie, MSBA
602.277.9449 x 348
marie.hoie@heinfeldmeech.com