

## Innovation Mindset

### Swindle in the Age of AI: An AI Forensic Challenge

Forensic accountants are financial experts who investigate and analyze financial information to uncover fraud, embezzlement, and other financial crimes. Forensic accountants use various technologies and techniques to analyze and interpret digital data to identify financial irregularities and fraudulent activities. Important to any analysis they perform, forensic accountants must be able to think creatively and critically. Often, forensic accountants think what criminals will do, and then devise tests to capture that bad behavior. In this case, you will think like a criminal and a forensic accountant to strengthen your critical and creative thinking skills.

For this case, assume the date is 1/1/2024.

First, you must think like the criminal Skylar Swindler. Skylar starts fake businesses and attracts investors to invest their cash before running away with their money. Skylar Swindler is peddling a new business called "Pure Oils." Pure Oils is a three-year old startup based in Memphis, Tennessee that specializes in the production of 100% natural, pure and organic essential oils made from the leaves that have fallen off trees in the rain forests. Skylar says his scientists have identified a new molecule present in leaves when they are knocked off trees due to rainy season. If the leaves are gathered quickly, they can be processed and the new molecule preserved to the benefit of those taking the oils.

Skylar advertises that the creation of the oils does not impact the environment and the oils are advertised as having near magical powers. Furthermore, Skyler reports that the company hires dozens of native Brazilians to gather leaves, lifting many of the employees out of poverty. Most of the employees have been hired in the last year and the company employs 110 employees.

Skylar asked for a review (not an audit) of Pure Oils books by an auditor. To mislead the auditor, Skylar generated fake data since everything of this company is fake—he actually fills up bottles with water and other liquids in his basement. To simplify this case, assume the auditor requests a dataset of all the company's employees for the last two years. The dataset should have the following fields:

- UniqueID: A unique identifier that was added to the dataset after it was compiled. It has no meaning other than to uniquely identify each row and make it easier to refer to a specific line of data.
- LastName: The employee's last name. If they do not have a last name in the database, it is listed as "NA".
- MiddleInitial: The employee's middle initial. If they do not have a middle initial in the database, it is listed as "NA".
- FirstName: The employee's first name. If they do not have a first name in the database, it is listed as "NA".
- Gender: The gender of the employee: male, female, or other.
- Job Title: The employee's current job title.
- Department: The department in which the employee works.
- HireDate: The date the employee was hired. Listed as dd/mm/yyyy. Dates can range from 6/23/1951 through 12/31/2023.
- TerminationDate: The date the employee was terminated from the organization. Listed as dd/mm/yyyy. If an employee has not been terminated, this field should be null.

- **AnnualSalary:** The employee's base annual salary for the fiscal year.

Your job is to generate a fake dataset that can fool the auditor. Your deliverable is the following:

- A fake dataset saved as an Excel file. The dataset should be for all 75 employees of Pure Oils.
- A Word file containing a description of how you generated the data. If you used ChatGPT, include the prompts you used to generate the data.

Second, now that Skylar Swindler is on the loose, you must think like the ace-forensic accountant, Payton Profit Protector, and snare this swindler. Skylar sends you the employee file. Your task is to do the following:

- Perform at least 5 tests that will help determine if the dataset is real or fake. Your tests can use the data included in the file, or you can add external data. If you add external data, include the link in the test explanation of where you acquired the data.
- For each test fill out the matrix:
  - List your expectation of what you expect to find in the data if it were real. Do this before you perform any testing.
  - Perform the test and document your findings for the test.
  - Write a conclusion—a description of whether the data support your expectation or not. You are not evaluated on whether your test discovers the fraud, rather you will be evaluated based on your creative and logical thinking in designing tests.

A hypothetical example is given in red.

| Test # | Expectation (include why)                                                                                                                                                      | Test (with references)                                                                                                                                                 | Conclusion                                                                                                                                                     |
|--------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1      | The file should not have any employees hired after the date of the file (January 1, 2024). The date of the case is 1/1/2024, companies do not hire employees at a future date. | Examine whether any entries in the HireDate field are > than DATE. See Browse activity Labeled "Test X: Employees hired after file was generated" for results of test. | None of the dates in the HireDate field are > than DATE. This test does not suggest the file is fraudulent. See Browse Tool labeled "Text #" for test results. |
|        |                                                                                                                                                                                |                                                                                                                                                                        |                                                                                                                                                                |

You should use Alteryx to perform all your tests and document your work. In Alteryx, use a Browse activity to show the final results of each of your tests. Annotate the Browse activity as "Test #: Description" where # is the number of the test from the table, and *description* is a brief description of the test. In the test column, reference the Browse activity so a reviewer of your work can easily find the results mentioned in the matrix.

Submit the following for grading:

- An Excel file containing the data you simulated.
- The Alteryx workflow where you document the results of your test.

- A Word document that contains
  - ▶ A description of how you generated your data. This should include the tools you used, and if you used ChatGPT, the prompts you provided to it.
  - ▶ The completed matrix you developed for testing the dataset. After the matrix, provide a one paragraph summary of whether you detected fraud or not. Make sure your references in the “Test” column tie to your Alteryx workflow as explained.

## Help for Generating Data

To help you generate data, consider using ChatGPT (<https://chat.openai.com/chat>) or a similar generative AI tool. ChatGPT is a language model developed by OpenAI that can understand and generate natural language text. It can be used for a wide range of tasks, such as language translation, text summarization, and text generation. For this case, ChatGPT can be used to generate data to aid Skylar Swindler.

ChatGPT works better if you use prompts. Prompts are a powerful tool that can be used to guide and enhance the responses generated by ChatGPT. By providing specific information or context in the form of a prompt, ChatGPT can generate more accurate and relevant responses that are tailored to the given input. As an example, consider the following:

*Prompt (what you type into ChatGPT):* "Generate a dataset of employee information for a company. The dataset should include employee ID, first name, last name, job title, department, hire date, and salary."

*Response generated by ChatGPT:*

```
EmployeeID,FirstName,LastName,JobTitle,Department,HireDate,Salary
10001,Emily,Smith,Marketing Specialist,Marketing,2018-05-20,65000
10002,James,Johnson,Sales Manager,Sales,2019-01-10,85000
10003,Morgan,Jones,Software Engineer,Engineering,2017-08-15,95000
10004,Riley,Lee,Financial Analyst,Finance,2020-04-01,75000
10005,Charlie,Davis,Human Resources Generalist,Human Resources,2016-11-12,60000
```

In this example, the prompt provides specific information about the type of data that needs to be generated, including the fields that should be included in the dataset. ChatGPT uses this information to generate a dataset of fake employee information that includes employee ID, first name, last name, job title, department, hire date, and salary.

To make the data more realistic, you can add additional details to your prompt, such as “Make all hiring dates after 1979.” You can experiment with ChatGPT for generating data. A few suggestions:

- Limit the number of rows ChatGPT produces until you have refined your prompt. Once you have refined it, you can batch process and have ChatGPT produce rows in groups (e.g., tell ChatGPT to “produce 25 rows” and then when it finishes, tell ChatGPT to “produce 25 more rows” or to “produce 25 unique rows.”)
- ChatGPT is capable of remembering information entered in the same chat. You can refine prompts by adding more information in the successive writing you enter into ChatGPT.

You do not have to use ChatGPT. You can use ChatGPT to generate the data and then do additional editing with other tools. ChatGPT by itself is not likely to develop a highly realistic dataset. However, it

can be useful to get you started on this task. By using it, you will learn the strengths and weaknesses of large language models for generating data.

Possible tests to perform:

1. File should not have any employee hired before 6/23/1951
2. See how many employees are in dataset and whether it matches what the case says.
3. Many of the names should be Brazilian (or likely from other South American countries)
  - a. Look up common South American/Brazilian names
4. Annual salaries for certain positions should be similar to others.
5. More employees should be hired in the last year than in previous years...should see a steady increase in the number of employees hired each year.
6. No employee should be hired more than 3 years ago when the company started.
7. Look up normal salary ranges from Brazil and see how these compare to what is in the data
8. Why aren't job titles in Brazilian?
9. Look up days of the week of hire dates...should not likely be on weekends.
10. Look up likelihood of having a middle initial in South American/Brazilian names. Compare to dataset.
11. Count the number of people with various job titles. Should have fewer managers and more workers. Also, there should be scientists, accountants, office staff, etc. that are employed for the company
12. Analyze unique job titles to see if they match language for workers picking up leaves, processing them, creating oils, etc.
13. Examine if any employees were terminated before they were hired.
14. Examine the day of the week of a termination. Should not be on weekends.
15. Annual salary of terminated employees should be less than annual salary of others who worked entire year.
16. Benford's analysis on salaries? (other patterns in real salary data?)
17. Check when is rainy season in the Amazon jungle...should be hiring many people leading up to and during rainy season and few workers not during rainy season.