



AI as a Data Discovery Tool: Lessons Learned

Steve Cramer

Business & Entrepreneurship Librarian/Professor
smcramer@uncg.edu
www.linkedin.com/in/stevebizlib

Can the popular AI tools...

1. Access numeric datasets?
2. Utilize static tabular data?
3. Access paywalled content?

NAICS 518200 - Computing Infrastructure Providers, Data Processing, Web Hosting, and Related Services

Display records

Text search table:

Occupation code	Occupation title (click on the occupation title to view an occupational profile)	Level	Employment	Employment RSE	Percent of total employment	Median hourly wage	Mean hourly wage	Annual mean wage	Me wa RS
00-0000	All Occupations	total	482,080	0.0%	100.00%	\$49.20	\$55.88	\$116,230	
15-0000	Computer and Mathematical	major	189,600	1.1%	39.33%	\$56.93	\$58.91	\$122,530	
	ons	minor	181,380	1.1%	37.62%	\$56.93	\$58.88	\$122,480	
	velopers, Testers	broad	97,940	1.6%	20.32%	\$64.11	\$66.11	\$137,510	
	s	detail	73,990	1.6%	15.35%	\$65.59	\$68.72	\$142,940	
	Operations	major	73,930	1.1%	15.34%	\$82.71	\$90.28	\$187,790	
	Administrative Support	major	73,710	3.5%	15.29%	\$20.75	\$24.08	\$50,090	
	Social Operations	major	72,250	1.5%	14.99%	\$48.27	\$51.97	\$108,100	
	s Specialists	minor	60,670						
	Occupations	major	48,200						
	ies Managers	minor	34,260						
	ce Services	minor	31,730						



Explore Census Data

Learn about America's People, Places, and Economy

Find Tables, Maps, and more...



[Help](#) [Feedback](#) [Advanced Search](#)



SimplyAnalytics®

My assumptions:

1. Access numeric datasets? • ***No***
2. Utilize static tabular data? • ***Maybe***
3. Access paywalled content? • ***No way***

Agenda

1. My research class
2. AI assignments & lesson plans
3. How did AI perform?
4. Conclusion & best practices
5. Appendix (more thoughts from students; good prompting examples; interesting AI responses)

ENT/GEO/LIS/MKT 430/530/630: Researching Entrepreneurship & Economic Development Opportunities

<https://uncg.libguides.com/ent430-630>

Some of the SLOs:

4. Employ a variety of business information sources and strategies.
5. Accurately evaluate business research.
6. Demonstrate concepts of statistical literacy through efficient searching of data sets and meaningful application of data to market and industry analyses.
- 7. Learn effective use of AI for business research (new for 2024)**

New in the syllabus:

- Throughout the semester, we will be exploring the strengths, weaknesses, and best practices of using AI as a research tool for secondary data.
- You will not need to pay for any AI access.
- Note that in this class, you **don't** do writing projects and you are required to **attach** your data sources (typically spreadsheets, PDFs, or maps).

2. AI assignments & lesson plans

From the Demographics assignment

1. Use an AI tool of your choice to get a **housing profile** of San Francisco city, CA (not the county). Ask the AI to include **housing value** and **rent costs** in the profile. Of course, work with the AI and get specific.
2. Which AI did you use?
3. What were your prompts and the replies?
4. What sources did the AI use?
5. Finally, find the DP04 "Selected Housing Characteristics" table from the American Community Survey, 2022 1-year estimates, for San Francisco city. How does the AI's data **compare** to the data in this table?

“AI week”

Day 1 Agenda:

1. Welcome, hellos, check-ins
2. Your experience with AI at this point?
3. Working together on a real research question
4. Discussion of our best practices
5. Discussion of **ethical** aspects and concerns

“AI week”

Day 2 Agenda:

1. Welcome, hellos, check-ins
2. Working together on another real research question
3. Using the **same prompts** in three AI tools
4. Comparing to our core research tools
5. Final discussion
 - When is AI a good tool to use first?
 - For research like this, what is AI really good for?

3. How did AI perform?

- As expected, AI tends to use statistics found in narrative sources:
 - Press releases (including from Census Bureau and BLS)
 - Newspaper articles
 - Reports ([BLS OOH](#); free content from the market research report vendors; trade associations)
- Can use [Census Quick Facts](#)
- Can't seem to handle long static tables
 - ex. BLS [Consumer Expenditure Survey](#) (PDF, XLS)
 - However...

CoPilot can understand big tables opened in Edge

The screenshot shows a Microsoft Edge browser window with a PDF document titled "cu-income-before-taxes-2022.pdf" open. The document contains a large table with 11 columns representing different income brackets and one column for "All consumer units". The table is titled "Table 1203. Income before taxes: Annual expenditure means, shares, standard errors, and relative standard errors, Consumer Expenditure Surveys, 2022".

On the right side of the browser, the Copilot chat interface is visible. It shows a message from Copilot stating that the conversation may have information that is not public and will not be saved. Below this, Copilot provides a summary of how spending on alcoholic beverages varies by income according to the table. The summary includes three bullet points:

- Lower Income (<\$15,000):**
 - Mean expenditure: **\$233**
 - Share of total expenditure: **0.7%**
- Middle Income (\$50,000 to \$69,999):**
 - Mean expenditure: **\$457**
 - Share of total expenditure: **0.8%**
- Higher Income (\$150,000 to \$199,999):**
 - Mean expenditure: **\$1,049**
 - Share of total expenditure: **0.9%**

Below the summary, Copilot notes that the data indicates that as income increases, the mean expenditure on alcoholic beverages also increases, but the share of total expenditure remains relatively stable.

The table in the background has the following structure:

Item	All consumer units	Less than \$15,000	\$15,000 to \$29,999	\$30,000 to \$39,999	\$40,000 to \$49,999	\$50,000 to \$69,999	\$70,000 to \$99,999	\$100,000 to \$149,999	\$150,000 to \$199,999	\$200,000 and more
Number of consumer units (in thousands) ¹	134,090	13,222	18,689	11,328	10,142	17,314	18,934	20,770	10,239	13,451
Consumer unit characteristics:										
Income before taxes										
Mean	\$94,003	\$7,624	\$22,485	\$34,988	\$44,518	\$59,200	\$83,894	\$121,806	\$171,270	\$322,568
SE	1,231.08	372.15	218.08	314.25	409.92	434.87	425.40	969.80	1,076.76	5,903.54
RSE	1.31	4.88	.97	.90	.92	.73	.51	.81	.63	1.83
Income after taxes										
Mean	83,195	9,810	24,622	37,370	45,658	58,126	79,460	110,099	147,787	250,412
SE	960.32	376.34	267.37	377.36	455.26	478.37	479.05	1,023.00	1,074.78	4,381.24
RSE	1.15	3.84	1.09	1.01	1.00	.82	.60	.93	.73	1.75
Age of reference person	52.1	54.4	60.5	55.6	53.6	50.7	48.7	48.2	48.6	49.4
Average number in consumer unit:										
People	2.4	1.6	1.7	2.1	2.2	2.4	2.7	2.9	3.2	3.2
Children under 18	.6	.3	.3	.5	.5	.5	.7	.7	.8	.8
Adults 65 and older	.4	.4	.6	.6	.6	.5	.4	.3	.3	.2
Earners	1.3	.4	.5	.8	1.0	1.3	1.6	1.9	2.1	2.1
Vehicles	1.9	.9	1.2	1.5	1.5	1.8	2.1	2.4	2.6	2.7
Percent distribution:										
Reference person:										
Men	49	42	39	42	46	49	52	56	56	57
Women	51	58	61	58	54	51	48	44	44	43
Housing tenure:										
Homeowner	65	38	53	57	57	62	68	77	85	90
With mortgage	38	11	15	21	26	34	42	57	66	70
Without mortgage	27	27	38	35	31	28	26	20	19	19
Renter	35	62	47	43	43	38	32	23	15	10
Race of reference person:										
Black or African-American	13	19	16	14	17	13	13	11	9	7
White, Asian, and all other races, not including Black or African-American	87	81	84	86	83	87	87	89	91	93
Hispanic or Latino origin of reference person:										
Hispanic or Latino	15	15	17	16	17	16	17	13	13	7
Not Hispanic or Latino	85	85	83	84	83	84	83	87	87	93
Education of reference person:										
Elementary (1-8)	3	5	4	4	3	2	2	(²)	(²)	(²)
High school (9-12)	28	45	43	40	33	30	24	18	12	6
College	69	49	53	55	63	67	74	81	86	93
Never attended and other	(³)	(²)	(²)	(²)	(²)	(²)	(²)	(²)	(²)	(²)
At least one vehicle owned or leased	89	63	78	90	89	92	95	97	97	97
Average annual expenditures										
Mean	\$72,967	\$31,066	\$37,534	\$45,424	\$49,951	\$57,942	\$70,871	\$89,354	\$116,773	\$167,088
SE	902.42	1,323.38	1,969.26	1,270.56	853.80	1,066.11	1,378.34	1,121.88	2,251.31	4,014.99
RSE	1.24	4.26	5.25	2.80	1.71	1.84	1.94	1.26	1.93	2.40
Food										
Mean	9,343	5,337	5,040	6,038	7,154	8,170	9,336	11,732	14,021	17,678
SE	17.2	17.2	13.4	12.2	14.3	14.4	13.2	12.4	12.6	10.6
RSE	1.8	3.2	2.7	2.0	2.0	1.7	1.4	1.1	0.9	0.6

4. Conclusion & best practices

1. AI **can't** access datasets and paywalled data.
2. Student: “You often get the **best** data from the datasets and the paywalled research tools”
3. Yet AI can be useful as a **discovery tool** for alternative sources, usually pages with data extracted from datasets or paywalled content.
 - You can ask the AI to only use information from PDF and XLS files
4. Use **power words** (smart words) and be as detailed as possible in your prompting...

A student prompt:

“What are the start up costs for an animal hospital in the United States? Act as a **business analyst**. Consider the **different types** of animal hospitals, such as small clinics or large hospitals with in-house labs and surgeries, and the **geographical location**, such as urban or rural areas. Include expenses for **equipment, rent, labor** and **training** costs, and costs for obtaining **licenses** and **permits** necessary for running an animal hospital. Provide a detailed cost analysis with considerations of all of these variables. If possible, include results from the **Economic Census.**”

Student comments & suggestions (more in the appendix)

- “From the experience in this exercise, the key to using AI for data research is to be as **specific** as possible. Adding keywords that specify subtopics, geography, and even the specific source, yields better results.”
- “For quality research, you need to know and include the **best sources/datasets** in your prompts”
- “Ask for the resources to be provided in a concise **table** with **links**.”

- "AI can provide useful **brainstorming** support and can dramatically reduce the time it takes to create an **initial framework** for doing data research but it does **not** replace doing the **actual research** and understanding the **nuances** of how various data interconnects to create a full picture."
- "While AI is useful as a "search" engine, it is **not** a "**decision**" engine. AI is good at finding resources, but it **cannot critically evaluate** the scholastic values of the sources it finds. As such, AI can be a useful tool that complements research, but it cannot replace the role of **critical human evaluation**."

5. Appendix

If anyone finding this in Canvas wants to see more examples

Student comments & suggestions

- “My suggestion is to be as specific as possible and to be okay with digging deeper if necessary. Be prepared with several ways to ask for the same thing.”
- “For quality research, need to know and include the best sources/datasets in your prompts”
- “I also asked for the resources to be provided in a concise table with website links.”
- "I find that it is very helpful to have an existing knowledge base on the topic about which one is inquiring but it seems possible to use AI to educate yourself about an overview of the topic if necessary."
- “Overall, I really liked that Bing was able to deliver on more current market trend data, which only further supports the findings from the ACS tables.”

- "AI can provide useful brainstorming support and can dramatically reduce the time it takes to create an initial framework for doing data research but it does not replace doing the actual research and understanding the nuances of how various data interconnects to create a full picture."
- "Artificial Intelligence can be helpful when researching for information on different industries. While AI is useful as a "search" engine, it is not a "decision" engine. AI is good at finding resources, but it cannot critically evaluate the scholastic values of the sources it finds. As such, AI can be a useful tool that complements research, but it cannot replace the role of critical human evaluation."
- "Using AI for industry research is not very reliable as it pulls from sources that aren't verified. Also, it can incorrectly cite sources and/or information. I think it is good to just get a broad idea or figure out what questions you really want to know"

Example of students generating “smart words” based on this research question

I was searching for industry information of meal kit/delivery (e.g., Blue Apron, HelloFresh) in the U.S. But I was not able to locate anything. Could you kindly help me?

- Market size, demand, demographics/ psychographics
- Top competitors and market share
- Financial information/benchmarking
- What geography to focus on?
- Sustainability considerations?

Prompt engineering example from a student

- What marketing strategies work best for the performance sock industry?
- What marketing strategies work best for the performance sock industry? Focus on the United States market. Only use information that focuses on running or performance socks. Include some industry or market statistics. Respond as if you are a market research analyst.
- I don't want information about socks that aren't performance socks. I only want information from running-specific or sock industry sources.

Series of prompts from a smart and organized student:

- 1: "Act as a venture capitalist and list what statistical industry data is relevant to deciding whether to invest in a national chain of animal hospitals"
- 2: "Create a list of what data sources to use for each item in the list above"
- 3: "List recent industry statistical data for each item. Pull statistics from the sources listed where possible"

Honest ChatGPT confession

- First prompt: “What is the profit margin for brewery startups in North Carolina?”
- [A long answer was provided with some statistics]
- Second Prompt: “What are your sources for the above answer?”
- “My apologies, but I don’t have direct access to external sources or the internet to provide specific citations. The information I shared is based on my existing knowledge up to a certain point in time (around 2021) and general industry trends. For more accurate and up-to-date data, I recommend consulting reliable sources such as industry reports, financial analyses, or expert advice. If you’re considering starting a brewery, conducting thorough research and seeking professional guidance would be prudent.”

Gemini on research strategies

“Remember, thorough market research is crucial. Visit these potential locations, talk to local residents and existing businesses, and assess the overall vibe to understand if it aligns with your vision and target audience. By combining information on general trends, local specifics, and your unique concept, you can make an informed decision about the best location for your used bookstore in North Carolina.”
[Gemini]

Again

- Using Google Gemini, I asked “What are the financial benchmarking for used bookstores in Greensboro, North Carolina”?
- “Unfortunately, obtaining specific financial benchmarks for independent used bookstores in Greensboro, NC, is challenging due to several factors:
 - Privacy of Financial Information: Most independent businesses, including bookstores, are not obligated to publicly disclose their financial information.
 - Limited Industry Data: Industry reports and data for used bookstores are often broad and may not provide specific details for a particular city or region.
- However, you can explore alternative resources to gain insights into the financial aspects of operating a used bookstore in Greensboro”

ChatGPT on research strategies

As of my last update in January 2022, I don't have access to real-time or region-specific data, including the profit margins for brewery startups in North Carolina. Profit margins for brewery startups can vary widely depending on various factors such as location, size of the brewery, operational efficiency, market demand, competition, and more.

To get accurate and up-to-date information on profit margins for brewery startups in North Carolina, you may need to consult industry reports, market research, or reach out to local brewery associations or industry experts who may have access to such data. Additionally, speaking directly with brewery owners in North Carolina could provide valuable insights into the financial aspects of running a brewery in the region."

Caution from ChatGPT

“As of my last update, the U.S. Census Bureau's American Community Survey (ACS) provides data on median gross rent for San Francisco. However, please note that this data might not reflect the most current rental prices, as the ACS is conducted over a period of time and the data is released periodically. According to the ACS data for 2019, the median gross rent in San Francisco was approximately \$1,994 per month. However, it's important to recognize that this figure represents the median for all rental units in San Francisco and may not capture the full range of rental prices across different neighborhoods and housing types. For the most up-to-date and detailed rental price information, I recommend consulting real estate websites, rental listing platforms, or contacting local property management companies in San Francisco. These sources often provide more current and specific data on rental prices in different neighborhoods and for various types of rental properties.”

"I can't do that for you"

- The ChatGPT response when a student in April 2023 gave it questions from my trade data assignment
- That student told me he had begun trying AI for my homework just to see what would happen