Quick try of ChatGPT Code Interpreter

Author: Eric Silberstein

Claps: 11

Date: Jul 13

For decades people have wanted to ask questions of data on their own without programming, SQL, or becoming spreadsheet wizards. Lots of software moved us toward this goal. Tableau, for example, founded two decades ago and acquired by Salesforce in 2019, lets tech-savvy users develop dashboards that other people can then explore. The dashboard builders can â€œparameterizeâ€� things so that the explorers can slice, dice, and drill â€” but within tight constraints. Super useful, but still far from the vision.

In April, a bunch of researchers at Microsoft tried all sorts of crazy things with GPT-4 and published Sparks of Artificial General Intelligence: Early experiments with GPT-4. On page 112, they show GPT-4 acting as a â€œstrategicâ€� data analyst. They gave it data from IMDb and told it:

> â€œI am a Hollywood producer. We are starting a new project for an adventure movie in 2023â€¦Iâ€™d like to know the best way to visualize this dataset with a single plot, so that our decision process can be made easier.â€�

GPT-4 generated Python code (using pandas, plotly, and networkx) that in turn generated this interactive plot:

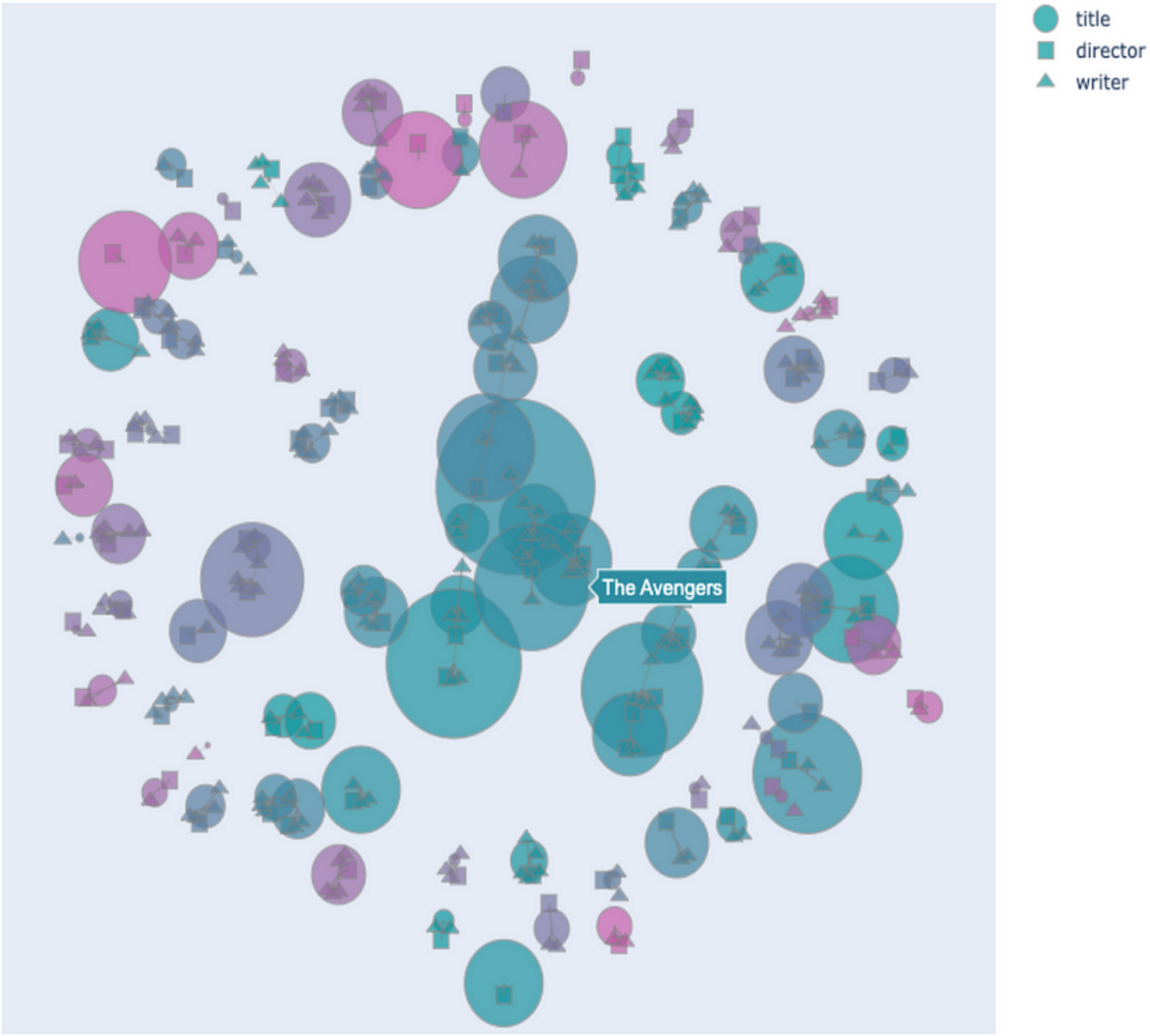
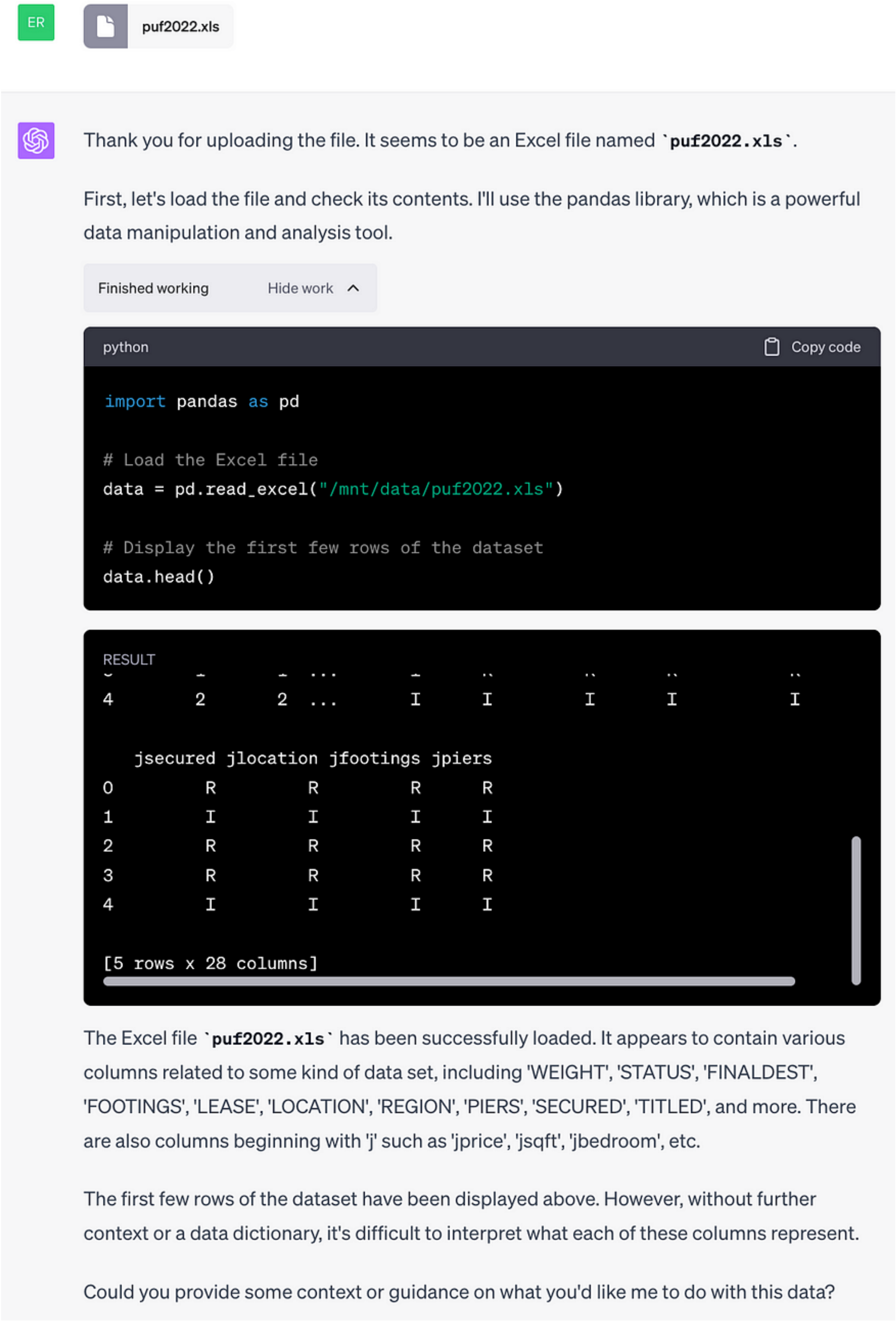


Figure from Sparks of Artificial General Intelligence: Early experiments with GPT-4

Now if only ChatGPT could generate data analysis code, execute it, and read the outputâ€¦what a beautiful workflow! Well, thatâ€™s exactly what OpenAI started rolling out to ChatGPT Plus users last week. Itâ€™s called Code Interpreter, which might sound like itâ€™s to help ChatGPT understand code, but in fact, it lets it execute code (*interpreter* as in, for example, the Python Interpreter).

There are a million ways this will be useful. I played with it today for the first time and started with a data analysis task. I went to the US Census datasets page, clicked on the first dataset, pulled down the 2022 file, and clicked on the plus icon to upload it to ChatGPT:



Thank you for uploading the file. It seems to be an Excel file named `puf2022.xls`.

First, let's load the file and check its contents. I'll use the pandas library, which is a powerful data manipulation and analysis tool.

```python
import pandas as pd

# Load the Excel file
data = pd.read_excel("/mnt/data/puf2022.xls")

# Display the first few rows of the dataset
data.head()
```

```
4        2       2  ...     I      I       I       I        I


   jsecured jlocation jfootings jpiers
0         R         R         R      R
1         I         I         I      I
2         R         R         R      R
3         R         R         R      R
4         I         I         I      I

[5 rows x 28 columns]
```

The Excel file `puf2022.xls` has been successfully loaded. It appears to contain various columns related to some kind of data set, including 'WEIGHT', 'STATUS', 'FINALDEST', 'FOOTINGS', 'LEASE', 'LOCATION', 'REGION', 'PIERS', 'SECURED', 'TITLED', and more. There are also columns beginning with 'j' such as 'jprice', 'jsqft', 'jbedroom', etc.

The first few rows of the dataset have been displayed above. However, without further context or a data dictionary, it's difficult to interpret what each of these columns represent.

Could you provide some context or guidance on what you'd like me to do with this data?

Great! It did what a human analyst might do if you sent them a file with no instructions â€” looked at it.

ER

I'd like to understand if there is a connection between square footage and price.

Sure, we can start by checking the relationship between square footage (`SQFT`) and price (`PRICE`) using a scatter plot and also compute the correlation between these two variables.

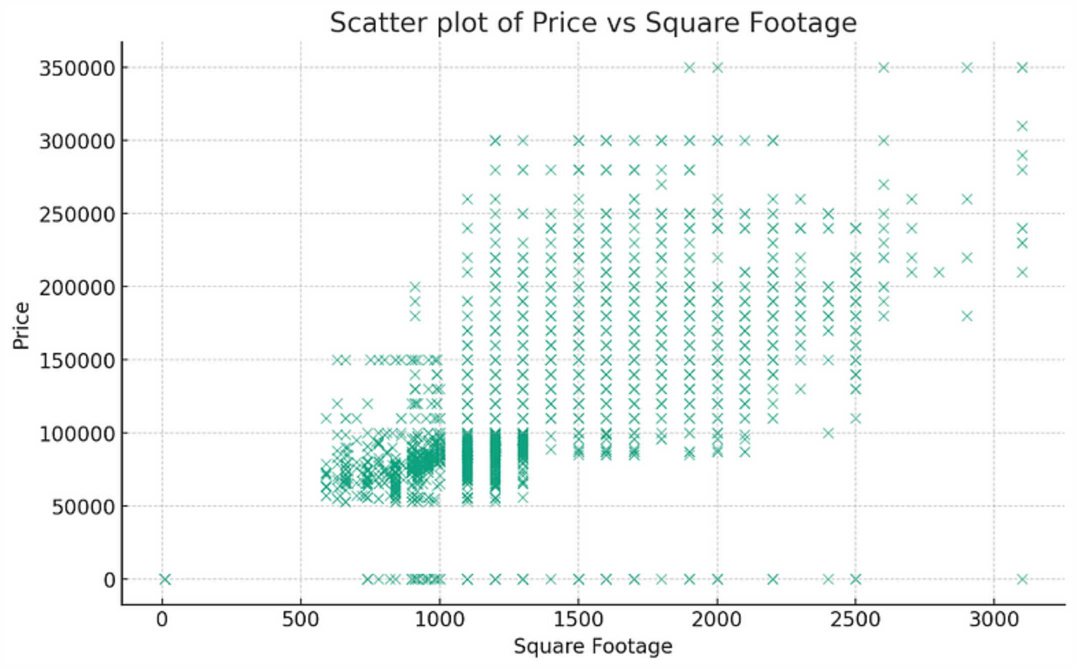
However, before we proceed, we need to ensure that the data in these columns is properly formatted (i.e., numerical). Let's check that first.

Finished working Show work

Both the `PRICE` and `SQFT` columns are of integer data type, which is good for our analysis.

Next, let's create a scatter plot to visualize the relationship between price and square footage. We'll also compute the correlation coefficient, which is a statistical measure that describes the size and direction of the relationship between two numerical variables. A correlation close to 1 or -1 indicates a strong positive or negative relationship, respectively, while a correlation near 0 indicates a weak or no relationship.

Finished working Show work



The scatter plot shows the relationship between price and square footage. Each point on the plot represents a data row in the dataset, with its position along the X-axis showing its square footage and its position along the Y-axis showing its price.

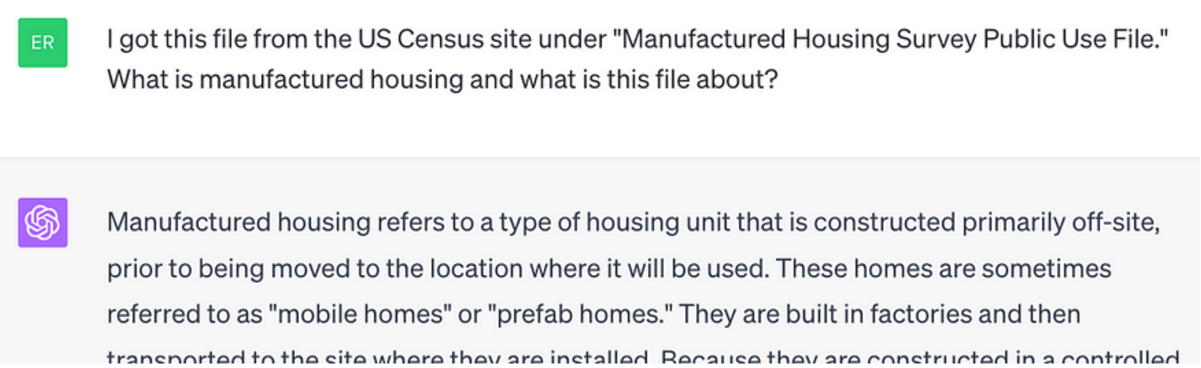
The correlation between square footage and price is approximately 0.882. This is a strong positive correlation, which suggests that as the square footage of a property increases, its price also tends to increase.

However, correlation does not imply causation. While these two variables move together, we cannot definitively say from this analysis alone that increasing square footage causes an increase in price. There may be other factors at play, or it could be a coincidence. Further

Great!

I asked it a few questions about weird data. I wonâ€™t paste the whole back and forth. Short is that it did a bunch of sanity checks, but did not catch what I hope would jump out to a humanâ€” that there are data points that look like they have a price of zero.

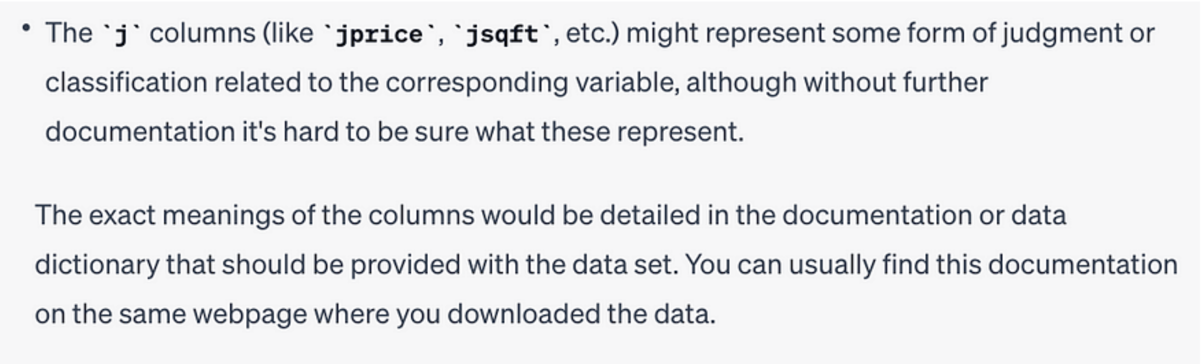
Something artificial about this exercise was I was so eager to try something that I just grabbed any data and didnâ€™t pay any attention to what it was. So I asked ChatGPT:



I got this file from the US Census site under "Manufactured Housing Survey Public Use File." What is manufactured housing and what is this file about?

Manufactured housing refers to a type of housing unit that is constructed primarily off-site, prior to being moved to the location where it will be used. These homes are sometimes referred to as "mobile homes" or "prefab homes." They are built in factories and then transported to the site where they are installed. Because they are constructed in a controlled

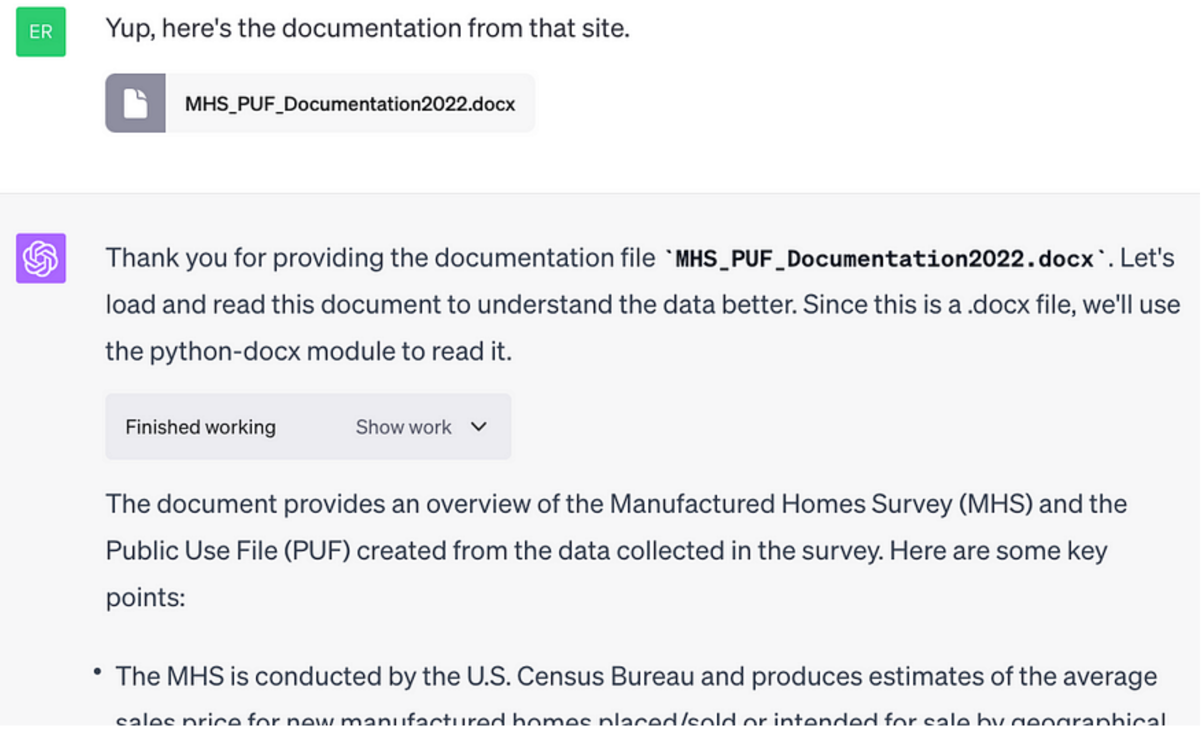
I cut off most of its answer. It helpfully pointed out that:



- The `j` columns (like `jprice`, `jsqft`, etc.) might represent some form of judgment or classification related to the corresponding variable, although without further documentation it's hard to be sure what these represent.

The exact meanings of the columns would be detailed in the documentation or data dictionary that should be provided with the data set. You can usually find this documentation on the same webpage where you downloaded the data.

It was right! I grabbed the documentation from the census page and uploaded it:



Yup, here's the documentation from that site.

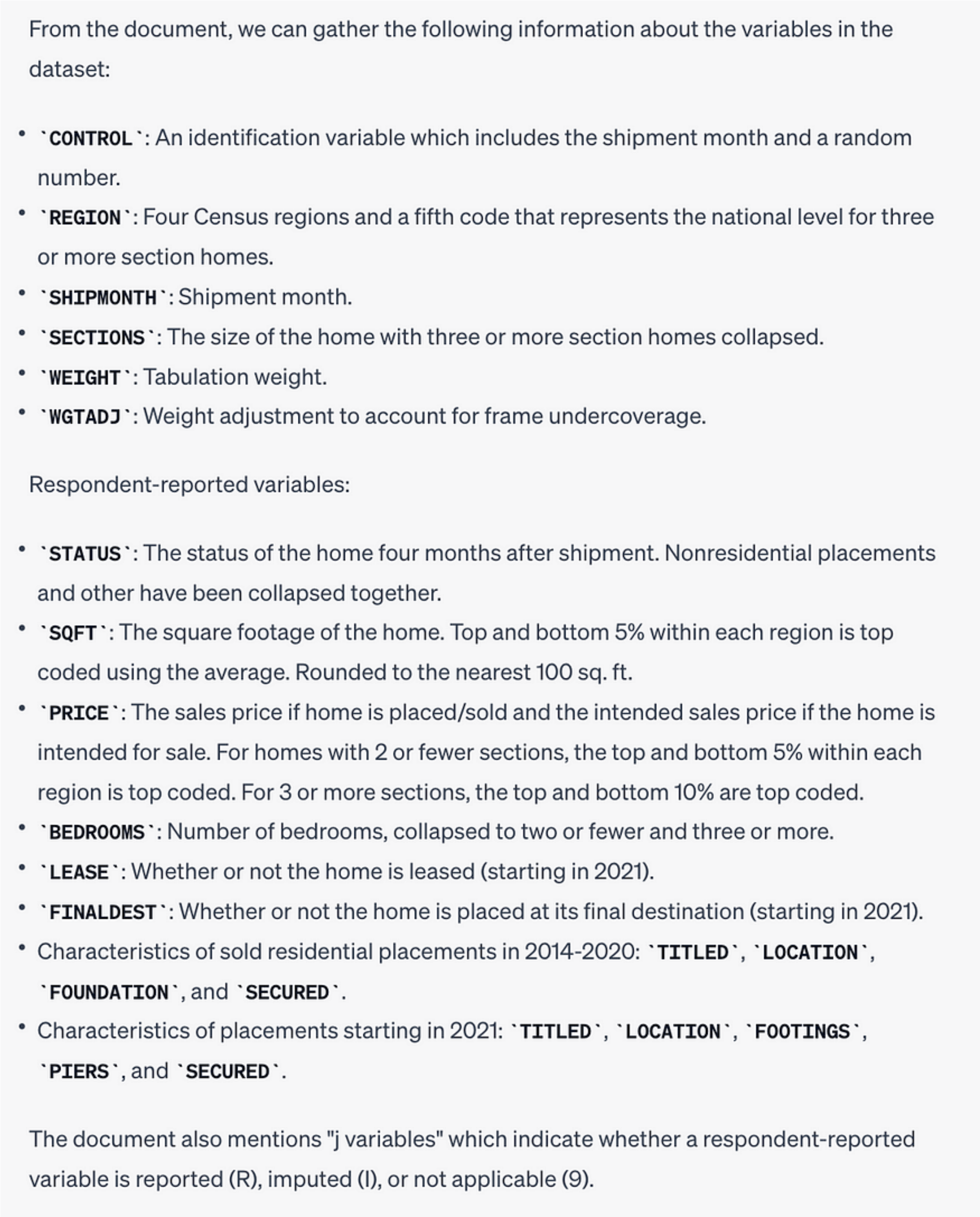
MHS_PUF_Documentation2022.docx

Thank you for providing the documentation file `MHS_PUF_Documentation2022.docx`. Let's load and read this document to understand the data better. Since this is a .docx file, we'll use the python-docx module to read it.

The document provides an overview of the Manufactured Homes Survey (MHS) and the Public Use File (PUF) created from the data collected in the survey. Here are some key points:

- The MHS is conducted by the U.S. Census Bureau and produces estimates of the average sales price for new manufactured homes placed/sold or intended for sale by geographical

It used Code Interpreter to open the Word document and look through it, getting I think most of the information right. For example:



From the document, we can gather the following information about the variables in the dataset:

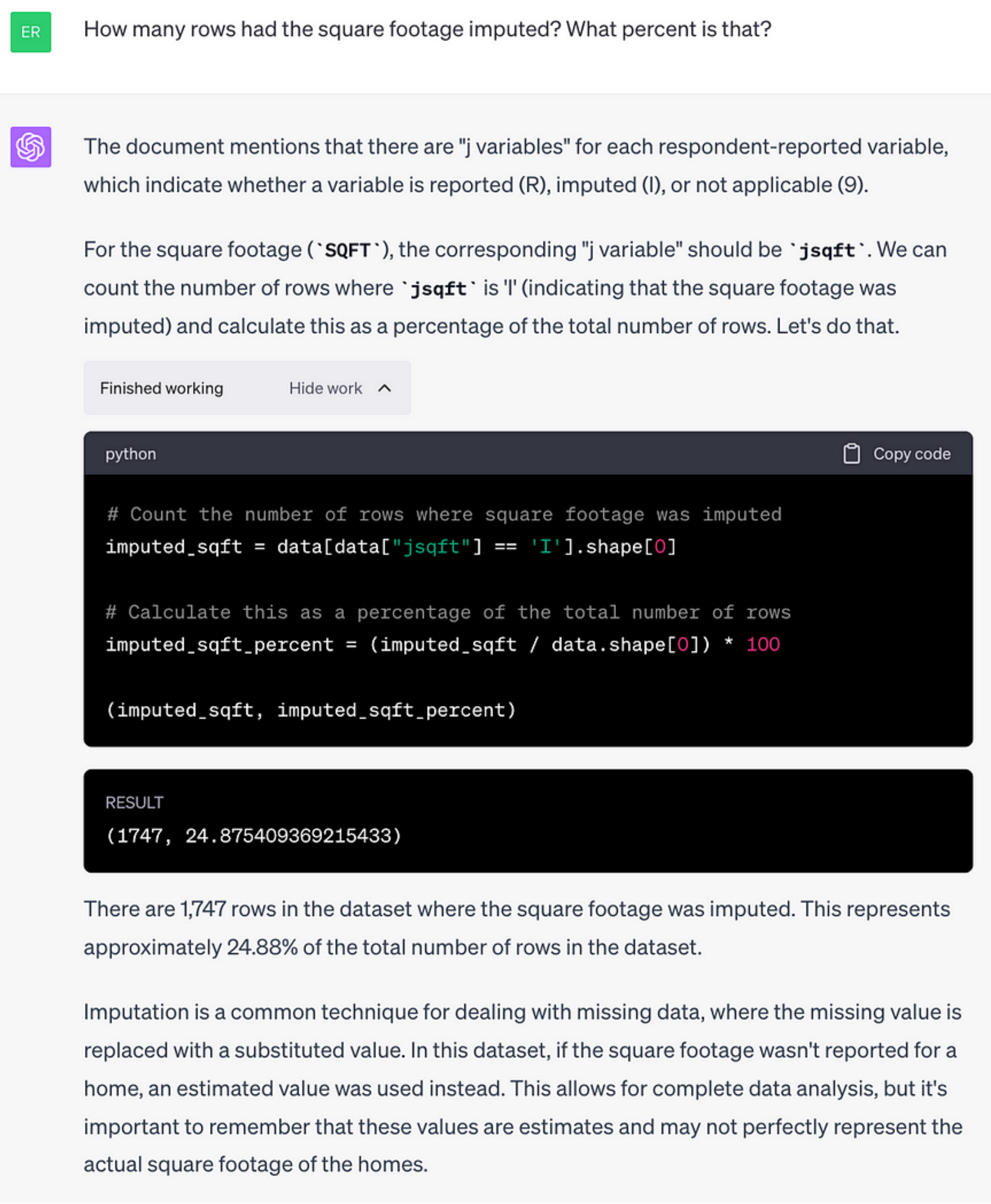
- **`CONTROL`**: An identification variable which includes the shipment month and a random number.
- **`REGION`**: Four Census regions and a fifth code that represents the national level for three or more section homes.
- **`SHIPMONTH`**: Shipment month.
- **`SECTIONS`**: The size of the home with three or more section homes collapsed.
- **`WEIGHT`**: Tabulation weight.
- **`WGTADJ`**: Weight adjustment to account for frame undercoverage.

Respondent-reported variables:

- **`STATUS`**: The status of the home four months after shipment. Nonresidential placements and other have been collapsed together.
- **`SQFT`**: The square footage of the home. Top and bottom 5% within each region is top coded using the average. Rounded to the nearest 100 sq. ft.
- **`PRICE`**: The sales price if home is placed/sold and the intended sales price if the home is intended for sale. For homes with 2 or fewer sections, the top and bottom 5% within each region is top coded. For 3 or more sections, the top and bottom 10% are top coded.
- **`BEDROOMS`**: Number of bedrooms, collapsed to two or fewer and three or more.
- **`LEASE`**: Whether or not the home is leased (starting in 2021).
- **`FINALDEST`**: Whether or not the home is placed at its final destination (starting in 2021).
- Characteristics of sold residential placements in 2014-2020: **`TITLED`**, **`LOCATION`**, **`FOUNDATION`**, and **`SECURED`**.
- Characteristics of placements starting in 2021: **`TITLED`**, **`LOCATION`**, **`FOOTINGS`**, **`PIERS`**, and **`SECURED`**.

The document also mentions "j variables" which indicate whether a respondent-reported variable is reported (R), imputed (I), or not applicable (9).

It was then able to use that knowledge:



How many rows had the square footage imputed? What percent is that?

The document mentions that there are "j variables" for each respondent-reported variable, which indicate whether a variable is reported (R), imputed (I), or not applicable (9).

For the square footage (`SQFT`), the corresponding "j variable" should be `jsqft`. We can count the number of rows where `jsqft` is 'I' (indicating that the square footage was imputed) and calculate this as a percentage of the total number of rows. Let's do that.

```python
# Count the number of rows where square footage was imputed
imputed_sqft = data[data["jsqft"] == 'I'].shape[0]

# Calculate this as a percentage of the total number of rows
imputed_sqft_percent = (imputed_sqft / data.shape[0]) * 100

(imputed_sqft, imputed_sqft_percent)
```

```
RESULT
(1747, 24.875409369215433)
```

There are 1,747 rows in the dataset where the square footage was imputed. This represents approximately 24.88% of the total number of rows in the dataset.

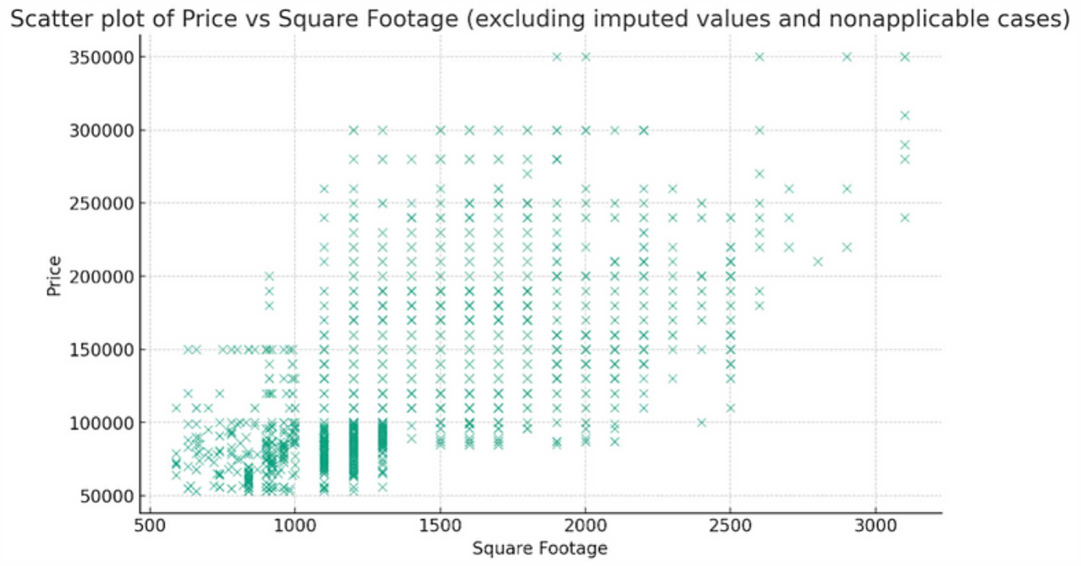
Imputation is a common technique for dealing with missing data, where the missing value is replaced with a substituted value. In this dataset, if the square footage wasn't reported for a home, an estimated value was used instead. This allows for complete data analysis, but it's important to remember that these values are estimates and may not perfectly represent the actual square footage of the homes.

I next had a back and forth about the houses with a price of â€œzero.â€� Iâ€™m leaving that out of this post. Hereâ€™s where we ended:

ER

Yes, in that word doc it says 9 is for nonapplicable cases. Can you redo the scatterplot and correlation calculation also excluding the rows where price or square footage is 9?

Finished working Show work



The scatter plot shows the relationship between price and square footage for the subset of the data where neither price nor square footage were imputed or marked as nonapplicable (9). Each point on the plot represents a data row in the dataset, with its position along the X-axis showing its square footage and its position along the Y-axis showing its price.

The correlation between square footage and price for this subset of the data is approximately 0.596. This is a moderate positive correlation, which suggests that as the square footage of a property increases, its price also tends to increase. Interestingly, this correlation is weaker than the correlation calculated earlier for the entire dataset (which was around 0.882) and for the dataset excluding imputed values (which was around 0.889). This means that the relationship between price and square footage is less clear when we exclude the rows with imputed values or nonapplicable values.

This could be because the rows with nonapplicable values (9) represent a significant proportion of the dataset, and these rows might have a stronger relationship between price and square footage. It could also be because the nonapplicable values represent special cases that have a different relationship between price and square footage compared to the rest of the data.

This highlights the importance of understanding the coding and structure of your dataset when analyzing the data. Special codes like this can significantly affect the results of your analysis if not handled correctly.

Not bad for playing around for a few minutes! (And if you rewind by a year itâ€™s shocking that this is possible at all. Back in those old days, who would believe that the transcript above was with a computer? *Oh, sure, the system will read some random Excel file and corresponding spec written out in a Word document, make sense of it all, and generate code to answer arbitrary user questions.*)

This famous scene from Minority Report is starting to feel like no big deal! (And Tom Cruise having to wear those gloves already looks quaint.)