chatgpt cheatsheet for coding in data analysis

JOE SQUIRE

starting & troubleshooting

Creating a Shell Code to Kick Start Your Analysis

Get a jump start on your project and by describing your dataset to ChatGPT and explaining what you want to do with it. This produces a code shell that you can tweak or iterate on to get to your final version.

Key items to include in prompt:
- Dataset name and type (i.e. Iris.csv)
- Granularity of dataset (i.e. what each row indicates)
- Column header names
- Which columns you want to transform with (i.e. group, sum, etc.)
- What you want to accomplish
  - create a new grouped data
  - perform statistical testing
  - generate EDA
- What language you want to do this in (i.e. R, Python, SQL, etc.)

Example:
I have a data set named {DATASET NAME}. Each row in the data set is {GRANULARITY}. The columns in this data set are delimited in triple quotes """{COLUMN NAMES}""". I am interested in the relationship between: {COLUMN NAMES}. {DESCRIBE WHAT THE COLUMNS INDICATE AND WHAT YOU WANT TO ACCOMPLISH IN YOUR ANALYSIS}. Could you suggest approaches to determining the {RELATIONSHIP OF THE COLUMNS}. Perform this in R.

Error Handling

Trouble shooting/debugging can be easily done using delimiters and a simple prompt using the code and error messages.

Key items to include in prompt:
- Original code that is throwing error
- Error message

Example:
I have this code in triple quotes """{ORIGINAL CODE}""" that is throwing this error in triple back ticks '''{ERROR MESSAGE}'''. What could be the issue and can you suggest a fix to the code?

BE AWARE OF DATA PRIVACY

Consider the rule of thumb "If you don't want it on the front page of the newspaper, don't put it in ChatGPT

joining tables

Key Items to consider:
- Table Keys
- Column names for both tables
- Other key elements from the 'Creating a Shell Code' section

Example:
I have two data sets named {DATASET NAME 1} and {DATASET NAME 2} . The columns in {DATASET NAME 1} are delimited in triple quotes """{COLUMN NAMES 1}""" and the columns in {DATASET NAME 2} are delimited in triple hashtags ###{COLUMN NAMES 1}###.

Example (cont.):
Each row in {DATASET NAME 1} is {GRANULARITY 1} and each row in {DATASET NAME 2} is {GRANULARITY 2}.

{JOIN TYPE} {DATASET NAME 1} to {DATASET NAME 2} using column {KEY COLUMN}.

******continue prompt in similar fashion to 'Creating a Shell Code' section***

improving your prompt

Use Delimiters for Clarity

Delimiters help ChatGPT better understand what you are asking by providing more context.

Delimiters to use:
- Triple Quotes: """
- Triple backticks: '''
- Triple dashes: ---
- Angle brackets: <>
- XML tags: <tag></tag>
- Triple Hashtags: ###

Example:
Explain the code delimited by the triple quotes """{YOUR CODE}"""

Provide a Successful Example

Have a piece of code you used before and want to get something similar but with a different dataset? Feed that code into the prompt. This is known as 'few shot prompting'.

Key items to include in prompt:
- Same as in 'Creating a Shell Code'
- Add in the example code you want to mimic

Example:
{SHELL CODE PROMPT}. Here is a successful example that was done on a different dataset enclosed in triple dashes: ---{EXAMPLE CODE YOU WANT TO MIMIC}---

Including a data dictionary in code shell prompt

Including certain elements of a dataset/table's data dictionary improves the cognitive ability of ChatGPT to work with your data.

Key items to include from your data dictionary:
- Data Element Name
- Data Type
- Description

Example:
Delimited in triple hashtags is the data dictionary for {DATASET NAME}, use this dictionary in your reasoning for the solution you come up with ###{DATA DICTIONARY}###. I am interested in the relationship between: {COLUMN NAMES}. {DESCRIBE WHAT YOU WANT TO ACCOMPLISH IN YOUR ANALYSIS}. Could you suggest approaches to determining the {RELATIONSHIP OF THE COLUMNS}. Perform this in Python.

<u>PRO TIP</u>: learn at least the basics of the coding language you will use ChatGPT to write in.
ChatGPT can code from scratch without anyone needing to know coding. But you will be at a huge loss in debugging and tweaking the code to fit your specific needs if you do not understand the basics of the language you are having ChatGPT write in

chatgpt cheatsheet for coding in data analysis

JOE SQUIRE

improving your prompt (cont)

Outline Analysis Steps for ChatGPT

Telling ChatGPT specific steps to work through reduces hallucinations and allows the model time to "think" and process the entire prompt systematically as it works.

Key Items to include
- Original prompt
- Steps you want ChatGPT to work through

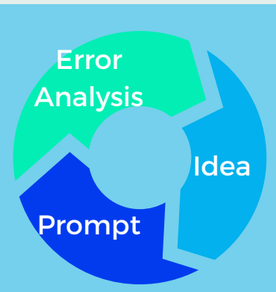
Example:
{SHELL CODE PROMPT}.
Execute in order of the following actions:
- Step 1: {ACTION, i.e. perform preliminary data inspection looking at first 5 rows, data types, number of non-null values in each column, etc.}
- Step 2: {ACTION, i.e. check the number of missing values in each column, replace missing values with mean, etc.}
- Step 3: {ACTION, i.e. identify outliers using boxplot visualization, perform univariant analysis with histogram or bar charts, etc}
- Step 4...

Separate each step answer with line breaks.

Iteration

Your first prompt rarely gets you the best answer. Iterating and trying out various tactics will improve your prompt immensely.

Key considerations:
- Can you clarify/be more specific?
- Try reducing the complexity of the prompt
  - Work in stages vs all at once



BE AWARE OF PASSWORDS, ETC.

If you have an API key or username/password in your code, best practice is to use packages or environmental variables to obscure them. Otherwise, make sure you remove them before pasting code into ChatGPT.

logic checking your outputs

Instruct the model work out it's own solution

ChatGPT can generate inaccurate answers. To help prevent this, you can check the output code with an additional prompt asking it to solve the problem again and compare it to the original code. This give the model time to reason and helps prevent hallucinations.

Key Items to consider:
- This will not catch inaccuracies in how your data interacts since you have not fed the data directly to ChatGPT
- This will look for logical & technical fallacies in the code produced based on your prompt

Example:
Your task is to produce approaches to determining the {RELATIONSHIP OF THE COLUMNS X & Y} in {DATASET NAME}.

First work out your own solution to the problem then compare it to the code in triple quotes and evaluate if the code in triple quotes is correct technically and in reasoning. Don't decide if the code in triple quotes is sound until you have done the problem yourself.

Problem: {ORIGINAL PROMPT THAT GENERATED THE ORIGINAL CODE}
Code: """{ORIGINAL CODE}"""

Respond if the original code is sound technically and in reasoning and why it is or is not sound.

rapid development in a new environment

Translating code into a another language

ChatGPT is good at being able to translate code from language into another, allowing you to rapidly develop a similar solution in new environments. This is especially useful for leveraging libraries/functions not in the current environment.

Key Items to consider:
- Not all languages can do the same thing or should.
- Be aware you will likely need to tweak the translated output

Example:
Here is a script in R delimited by triple hashtags ###{YOUR CODE }###.
Translate this into Python. Please call out any libraries or functions that may not work as expected in the translation and explain why. Make a recommendation on an substitutions for these libraries. If there is no good substitution please say so and explain why.