Kateryna

Kononova

for humans on cats

**MACHINE LEARNING**

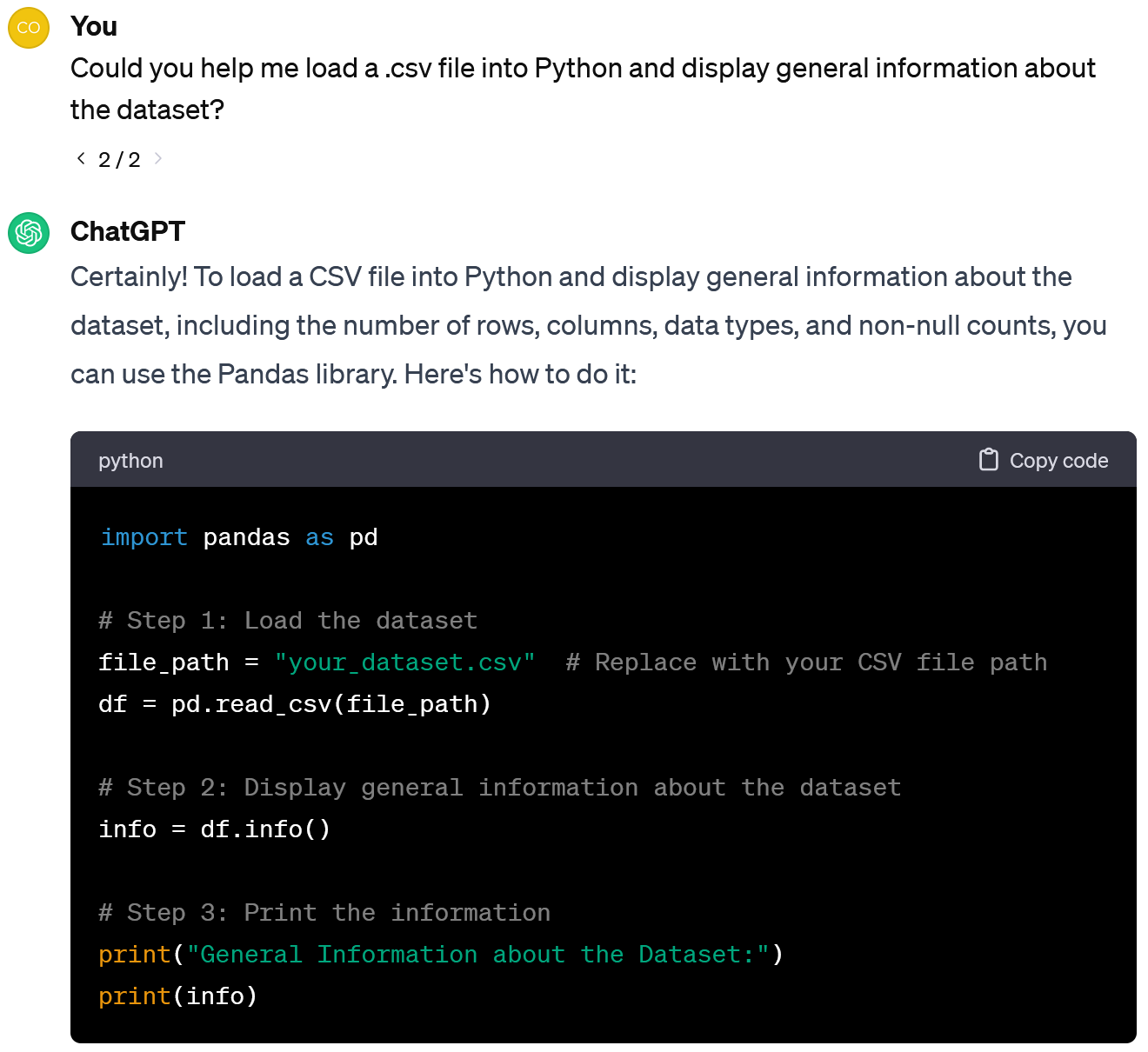


|  |  |
| --- | --- |
| C:\Users\kater\Documents\Comics\DALLE\icons\s31.png | **C:\Users\kater\Documents\Comics\DALLE\icons\s4_1.png**C:\Users\kater\Documents\Comics\DALLE\DALL·E 2023-12-21 05.21.56 - A black and white drawing of a cat looking at a mirror, and in the reflection, it sees itself as a cyber-cat. The scene depicts the cat from behind, g.png **Іntro** |
| C:\Users\kater\Documents\Comics\DALLE\Part 1\19_1.png  Once upon a time, there was a cat called Shylo. She was so smart that any cognitive scientist would call her a **"predictive machine"**. | But suddenly, the fridge creaked. The smell of food wafted in...  She was not interested in everything. Only a few things bothered her. |
| C:\Users\kater\Documents\Comics\DALLE\icons\s1.png  The sun won't disappear, but the food might. I'd better take both.  Shylo thought for a while. Should she run or keep lying down in the sun, which was warming her side so pleasantly?  She got up and ran to the kitchen with a loud "MEOW". | C:\Users\kater\Documents\Comics\DALLE\0.png |
| C:\Users\kater\Documents\Comics\DALLE\14.png  At the same time, like most felines, Shylo was a lazy one. "Who could do all this for me?" she wondered.  To describe this situation in scientific terms, this cat, like any predictive machine:   * collected **data**, * selected the most useful **features**, * made a **prediction** using her past experiences, * made a **decision** about how to proceed. | C:\Users\kater\Documents\Comics\DALLE\icons\m1.png |
| C:\Users\kater\Documents\Comics\DALLE\icons\s4.png  Who are you?  The strange creature had no smell. It couldn't be touched. All in all, it looked a bit artificial, but Shylo was not frightened…  The cat had a look at the people around her, who had their eyes on their computer screens... And suddenly she saw something. | C:\Users\kater\Documents\Comics\DALLE\15.png |
| C:\Users\kater\Documents\Comics\DALLE\icons\s1.pngC:\Users\kater\Documents\Comics\DALLE\Part 1\20.png  I am a virtual cat from the **Machine Learning** (ML) family.  Don't worry, I'm a predictive machine just like you. You can call me MyLo.  I've never heard of it... You must have a long pedigree. Perhaps, they paid a lot of money for you. | C:\Users\kater\Downloads\5 (1).png |
| C:\Users\kater\Documents\Comics\DALLE\icons\s1.png  Well, that's a bit strong. Imagine the Tree of Life, I would say that the ML family is included in the AI class. AI mimics animal behaviour. Some of them have physical bodies, some of them follow rules, some of them use algorithms, etc. AI is a big, big zoo.    Listen, Mylo, aren't you one of those creatures they call **Artificial Intelligence** (AI)?  C:\Users\kater\Documents\Comics\DALLE\icons\m1.pngC:\Users\kater\Documents\Comics\DALLE\icons\m1.png  Being from the ML family, I am just a complex algorithm trained to make predictions. It is like your brain that learns and predicts. | |
| C:\Users\kater\Documents\Comics\DALLE\icons\s31.pngC:\Users\kater\Documents\Comics\DALLE\icons\m1.png  The neural network is only one of my algorithms. I can do a lot of very complex things with it. You've probably heard of computer vision, Facebook recommendations, or Google Translate? But it is lots of learning and tons of data.  You may have just called me a house cat. This is wrong. If we continue with the biological analogy, I'm a creature of the famous Felidae family. I can easily become a lion or a tiger.  Oh, you mean you’re a sort of **Neural Network** (NN)? | |
| C:\Users\kater\Documents\Comics\DALLE\icons\s4.png  Well, yes, we have a lot in common. But they are interested in super-big data, and they have more tools, not just trained algorithms like me.  Data, data... You sound like a **Data Scientist** (DS). | |
| C:\Users\kater\Downloads\DALL·E 2023-12-28 21.59.16 - A black and white drawing of a tall and slim fractal taxonomy tree with sixteen distinct levels. The tree is elongated and slender, with each fractal .png  Species (Felis catus)  If I were NN I’d be here  Genus (Felis)  As ML, I’m here  Family (Felidae)  Order (Carnivora)  With DS we diverge here, our branches cross then  If I were AI  I’d be here  Phylum (Chordata)  Class (Mammalia)  Kingdom (Animalia) (Eukaryote)  Domain (Eukaryota)  You are on the top  Where am I? | |
| C:\Users\kater\Documents\Comics\DALLE\7.png  **Supervised Learning**  If we're talking about eating, that's like being told by your mum which foods are good and which are bad.  I have several options. Listen…  Yeah... And how do your algorithms learn? | C:\Users\kater\Downloads\4.pngC:\Users\kater\Documents\Comics\DALLE\DALL·E 2023-12-20 10.53.40 - A linear black and white drawing depicting a small cyber cat removing the head of a larger cyber cat, inspired by the provided sketch. The cyber cats .png |
|  | **Unsupervised Learning**  This is when you are trying to figure out yourself which food makes you jump, which makes you sleep, which makes you feel nauseous. |
| **Ensembles**  Well, it's like when you can share the experience with your friends. | C:\Users\kater\Downloads\DALL·E 2024-01-22 10.32.46 - A black and white drawing of multiple cyber cats sharing food. The cyber cats, each with unique mechanical and robotic features, are gathered around a.png |
| C:\Users\kater\Downloads\DALL·E 2024-01-22 10.27.14 - A black and white drawing of a cyber-cat carefully removing the head from a cyber-lion as part of a mechanical process or repair. Both the cyber-cat a.png  **Reinforcement Learning**  Here you seem to be walking through a labyrinth: if you turn into the right corridor, you get a sweet; if you make a mistake, no sweets, just the feeling that the food is moving away, but that's more how robots learn. | **Transfer Learning**  It's like taking the brain out of a lion that has been trained to hunt antelopes and using it to hunt mice. |
| Okay, so what does it give you? What can your trained algorithms do? | C:\Users\kater\Documents\Comics\DALLE\DALL·E 2023-12-20 10.59.06 - A linear black and white drawing of a cyber cat navigating a labyrinth. The cat is depicted with mechanical and robotic features, indicative of its cy.png |
| C:\Users\kater\Documents\Comics\DALLE\12.png  Mrrrrrr… | C:\Users\kater\Documents\Comics\DALLE\13.png  I can:   * predict * abstract * generalise * memorise * reproduce * choose the best |
| C:\Users\kater\Documents\Comics\DALLE\11.png  I've got nothing new to create.  So you won't hurt me? I've heard that people are afraid of being killed by AI. It seems I was misinformed.  Mrr, what can you not do? | C:\Users\kater\Documents\Comics\DALLE\9.png |
| C:\Users\kater\Documents\Comics\DALLE\DALL·E 2023-12-20 11.06.14 - A black and white drawing depicting two cats wearing hats, looking at an old plane taking off in the rain. The first cat is a normal cat, with fur and.png  By the way, have you seen Casablanca?  Mylo, I think this is the beginning of a beautiful friendship. |  |
|  |  |

|  |  |
| --- | --- |
| **C:\Users\kater\Documents\Comics\DALLE\Part 1\DALL·E 2023-12-21 07.11.58 - A black and white drawing of a cyber-cat dressed as a cook. The cyber-cat, with its mechanical and robotic features, is wearing a chef's hat and apron.png**  Listen, Mylo, if we're going to be friends, let me take you out to dinner. Fish, meat, canned food - what do you like best? | C:\Users\kater\Downloads\DALL·E 2024-01-22 10.45.35 - A black and white drawing of a cat being offered a choice of fish, meat, and canned food. The cat is depicted with an interested and curious expressio.png  Part 1. **Data and features** |
| C:\Users\kater\Documents\Comics\DALLE\Part 1\4 копия.png  I'm a virtual being. We only eat data. However not all data is edible. And there is still the matter of cooking.  Where do you get the data from? I take almost everything from the fridge, for example. | Preparing data? Wow... Mylo, give me the details. I love cooking! |
| Once I even caught a fish from the pool. What about you?  Although sometimes I go hunting for sparrows in the garden. Or mice. I can get them out of the ground. | C:\Users\kater\Documents\Comics\DALLE\DALL·E 2023-12-20 12.53.05 - A black and white drawing of a cat in the dynamic act of catching both a fish and a bird at the same time. The cat is in the foreground, its body stre1.png |
| C:\Users\kater\Documents\Comics\DALLE\DALL·E 2023-12-22 12.00.19 - A black and white drawing of a cyber-cat using advanced technology to predict how a billion cats are going to feel after eating. The cyber-cat is surr.png  For example, I can ask the cats in my neighborhood what they ate, when they ate it, how much they ate and how they felt afterwards. | C:\Users\kater\Documents\Comics\DALLE\icons\m1.png  To predict how a cat is going to feel after eating, I need to know what a billion cats ate and what happened afterwards.  It's like walking in their shoes.  For that, I can collect data **myself**. |
| You can also do it in an **automated way**.  It's boring, time-consuming and you won't collect much. But you'll definitely find something edible. | C:\Users\kater\Documents\Comics\DALLE\DALL·E 2023-12-22 12.03.46 - A black and white drawing of a cyber-cat interviewing the cats in a neighborhood about their eating habits. The cyber-cat is depicted with a notepad o.png |
| C:\Users\kater\Documents\Comics\DALLE\icons\m2.pngC:\Users\kater\Documents\Comics\DALLE\DALL·E 2023-12-20 17.43.27 - A black and white drawing depicting several webcams observing cats as they eat different types of food. Each webcam is focused on a different cat_ one.png  It’s like installing surveillance cameras and recording everything that happens around you. | Sure, it is faster and it gives you more data but there might be nothing edible. |
| C:\Users\kater\Documents\Comics\DALLE\icons\m2.png  Mylo, isn't it possible that there is a lot of data, but not much to eat?  This is the cool route, but the most expensive.  Take a chain of cat restaurants, for example. Record all the orders - who ate, what, when, and what feedback they left. Google does it, for example.  You can also **create the environment**. | C:\Users\kater\Documents\Comics\DALLE\DALL·E 2023-12-20 17.45.47 - A black and white drawing of a cats' restaurant. The scene shows a cozy and elegant dining area filled with cat-sized tables and chairs. There are cat.png |
| C:\Users\kater\Documents\Comics\DALLE\DALL·E 2023-12-20 18.51.00 - A black and white drawing of a cat with a curved back belching, with a tin of grass placed in front of it. The cat's posture suggests it has just fini1.png | C:\Users\kater\Documents\Comics\DALLE\icons\s2.pngC:\Users\kater\Documents\Comics\DALLE\icons\m1.png  For example, if you need to cleanse your stomach, then you eat grass. If you need to run, then you eat meat.  Of course it is. You should always choose only what is useful for a specific case. |
| C:\Users\kater\Documents\Comics\DALLE\icons\m2.pngC:\Users\kater\Documents\Comics\DALLE\icons\s1.png  Yes, that's right, Shylo.  Those useful things are called **features**, right? | C:\Users\kater\Documents\Comics\DALLE\16.png  But why prepare them if they are edible and useful? |
| C:\Users\kater\Documents\Comics\DALLE\DALL·E 2023-12-20 18.57.43 - A black and white drawing of a pig oinking while emerging from a tin labeled 'CATS FOOD'. The pig is depicted in a humorous and playful manner, squeez.png  Imagine your tin of food. Have you read the ingredients? There's meat, cereals, minerals, vegetable protein extracts, sugar, etc.  Don't get distracted. You better tell me, when you open it, have you ever seen a pig in there?  Sugar, really? | C:\Users\kater\Documents\Comics\DALLE\icons\m1.png |
| Never. All bones are removed, everything is cleaned, cut into small pieces so you can swallow it without choking.  At the same time, the jar is completely filled. There must be no air bubbles, otherwise it will spoil quickly. | C:\Users\kater\Documents\Comics\DALLE\DALL·E 2023-12-20 19.08.28 - A black and white drawing of a cat dressed as a chef ('cat-cooker'), surrounded by various animal parts such as horns, hoofs, and bones. The cat is we.png |
| C:\Users\kater\Documents\Comics\DALLE\DALL·E 2023-12-20 19.12.11 - A black and white drawing of four animals, each holding a numbered sign, depicted in a more realistic style. The pig, less cartoon-like, is holding a .png  That's right. In the virtual world, we call this handling **outliers, errors and missing data**, as well as scaling.  But first you have to put everything into numbers. That's the only way it works here.  So you can't create a record "rabbit" in your dataset. You have to code it somehow, or at least give it a number. | C:\Users\kater\Documents\Comics\DALLE\icons\m1.png |
| C:\Users\kater\Documents\Comics\DALLE\icons\s1.png  I look at the statistics and do the visualisations. Shall I show you?  Sure! How, then? You're a virtual cat and I'm a real one. How can we do anything together?  Mylo, how do you know that the features are ready? | C:\Users\kater\Documents\Comics\DALLE\DALL·E 2023-12-20 19.17.01 - A black and white drawing of a cyber-cat illustrating the process of cooking fish. The cyber-cat, depicted with mechanical and robotic features, is si.png |
| C:\Users\kater\Downloads\DALL·E 2024-01-22 10.58.13 - A black and white drawing of a cat stomping on a keyboard. The cat is depicted in an animated and playful posture, with its paws pressing down on the .png  Of course I do! Stomping on the keyboard is my favorite game! | Let's use one of my algorithms. There is a big linguistic model among the neural networks you mentioned. It's called [**ChatGPT**](https://openai.com/gpt-4).  C:\Users\kater\Documents\Comics\DALLE\icons\m1.png  But it takes a bit of typing. Shylo, do you know how to use a keyboard? |
| C:\Users\kater\Documents\Comics\DALLE\icons\m2.png  Please no pythons, I'm afraid of snakes!  Great! Then challenge me in Chat GPT. I'll put his answers into [**Python**](https://www.python.org/downloads). | C:\Users\kater\Documents\Comics\DALLE\DALL·E 2023-12-20 19.20.59 - A black and white drawing of a cat frightened by a snake. The cat is in an alarmed posture, with its back arched and fur standing on end. Its eyes are.png |
| C:\Users\kater\Documents\Comics\DALLE\Part 1\100.jpg  He loved the British comedy show Monty Python's Flying Circus.  No worries, Shylo. It's just the name of a programming language invented by Guido van Rossum almost fifty years ago. | C:\Users\kater\Documents\Comics\DALLE\icons\m1.png  You should watch it! There are some parallels with my virtual world. |
| C:\Users\kater\Documents\Comics\DALLE\icons\m2.png  Since we're talking about general concepts and all that, let's agree that you'll figure out how to install Python yourself. It's easy, but certain settings depend on your computer. | C:\Users\kater\Documents\Comics\DALLE\icons\s2.png  Ok, then I'm with you, let's go! |



Okay.

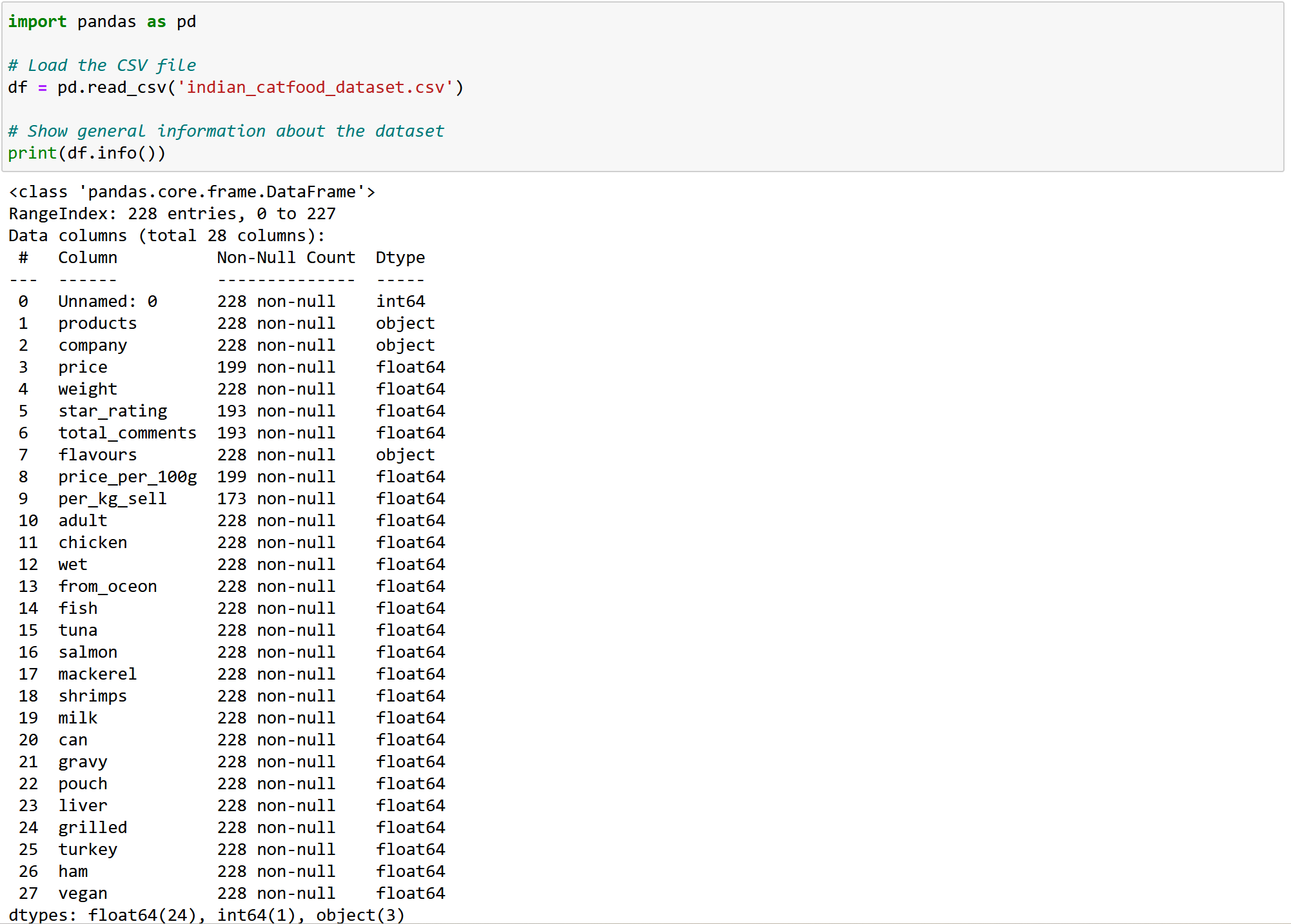
 Case 1. **Data pre-processing**

In fact, data collection is one of the most difficult and time-consuming stages, but it is more related to Data Science. Machine Learning is more about analysing data that already exists.

Oh, interesting what other cats eat!

Listen Shylo, let's put into practice everything we've talked about. I found a good [dataset about cat food](https://www.kaggle.com/datasets/herambsawant/amazon-india-cat-food-products-dataset) on the Kaggle. Instead of collecting data ourselves, let's look at what someone else has already collected.

Let’s download the data to our computer. And then we load it into Python. Look, I'll ask ChatGPT how to do that.



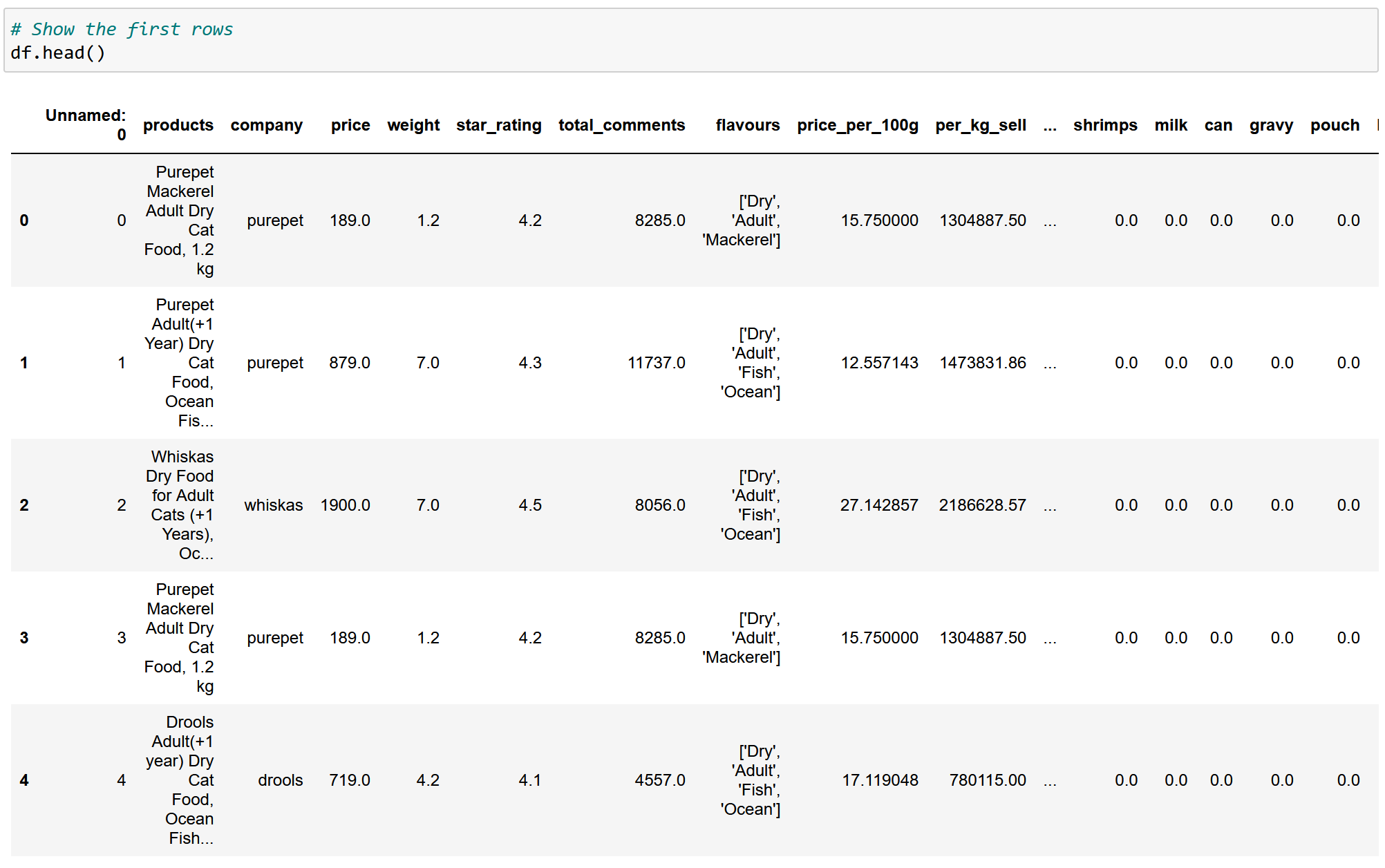
So we work with two windows in the browser? Pretty handy!

You see, now our dataset is loaded into a table called *df*.

There are 228 entries - one for each case of selling cat food. The features of each case are described in 28 string, numerical and binary columns: *product, company, price* (some entries are missing)*, weight, star\_rating* (some entries are missing)*, price\_per\_100g* (some entries are missing)*,* etc.

Copy the reply from ChatGPT and paste it into the window of the Jupiter notebook, a web-oriented environment that supports Python.

The [Jupiter notebook](https://www.anaconda.com/products/individual) must be added when you create a new project. If you have any problems with this, you can also ask ChatGPT.



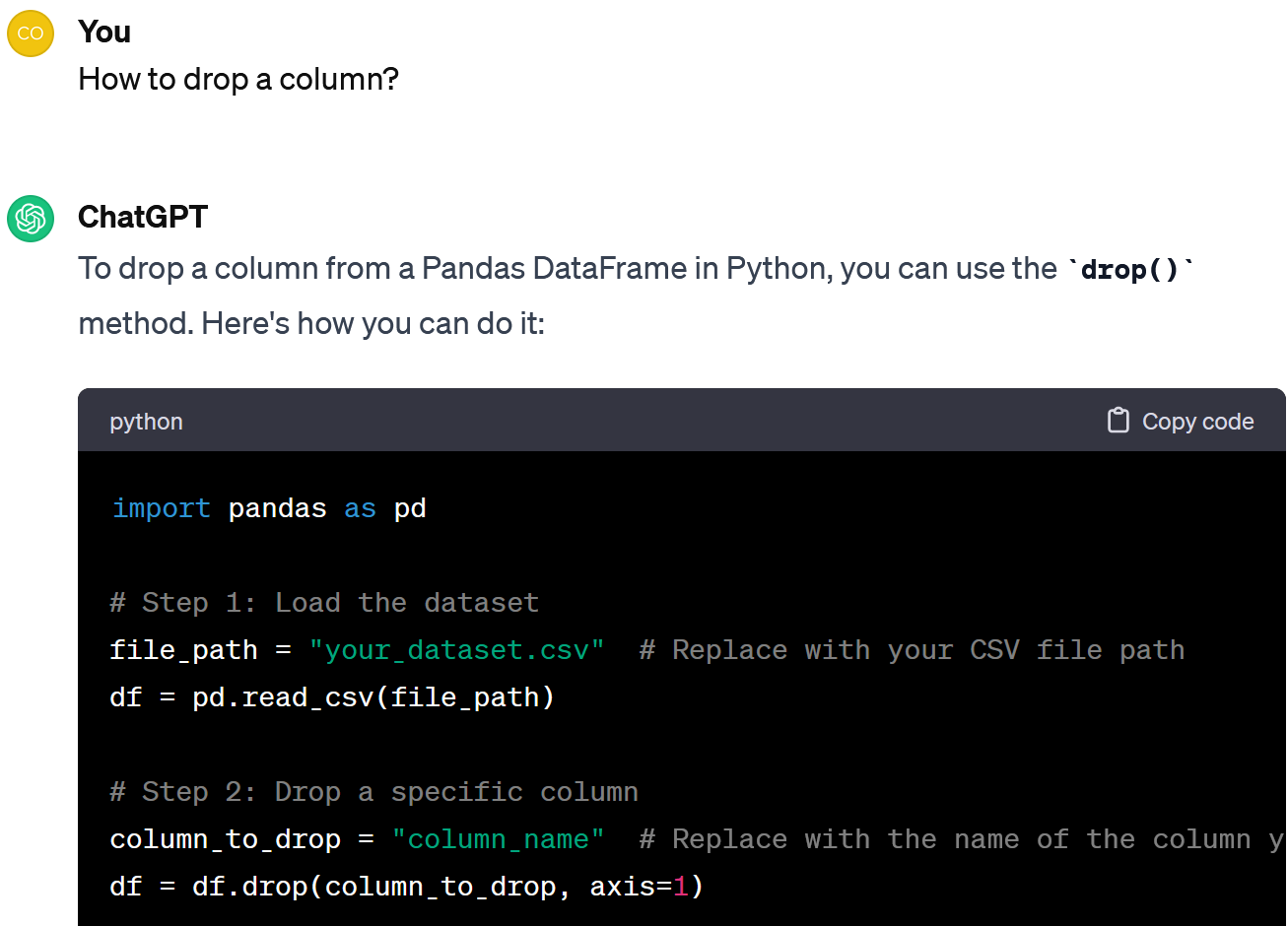
*flavours*: Flavors or ingredients of the cat food, stored as a list in string format.

Features from *chicken* to *vegan* are binary columns (0/1) indicating the presence of specific ingredients like whether the food is wet, contains fish or is vegan.

Let's look at these columns closer:

* *Unnamed: 0* is an index or identifier column,
* *products*: Generalized descriptions of the products,
* *company*: The company or brand of the cat food,
* *price*: The price of the product ,
* *weight*: Weight of the product in presumably kilograms,
* *star\_rating*: Average rating of the product,
* *total\_comments*: Reviews the product received,
* *price\_per\_100g*: Numerical data calculated based on *price* and *weight* ,
* *per\_kg\_sell*: Numerical data calculated based on *price*, *weight* and *total\_comments*,
* *adult*: A binary or categorical feature indicating if the product is suitable for adult cats.

Note that Python numbering starts with 0.



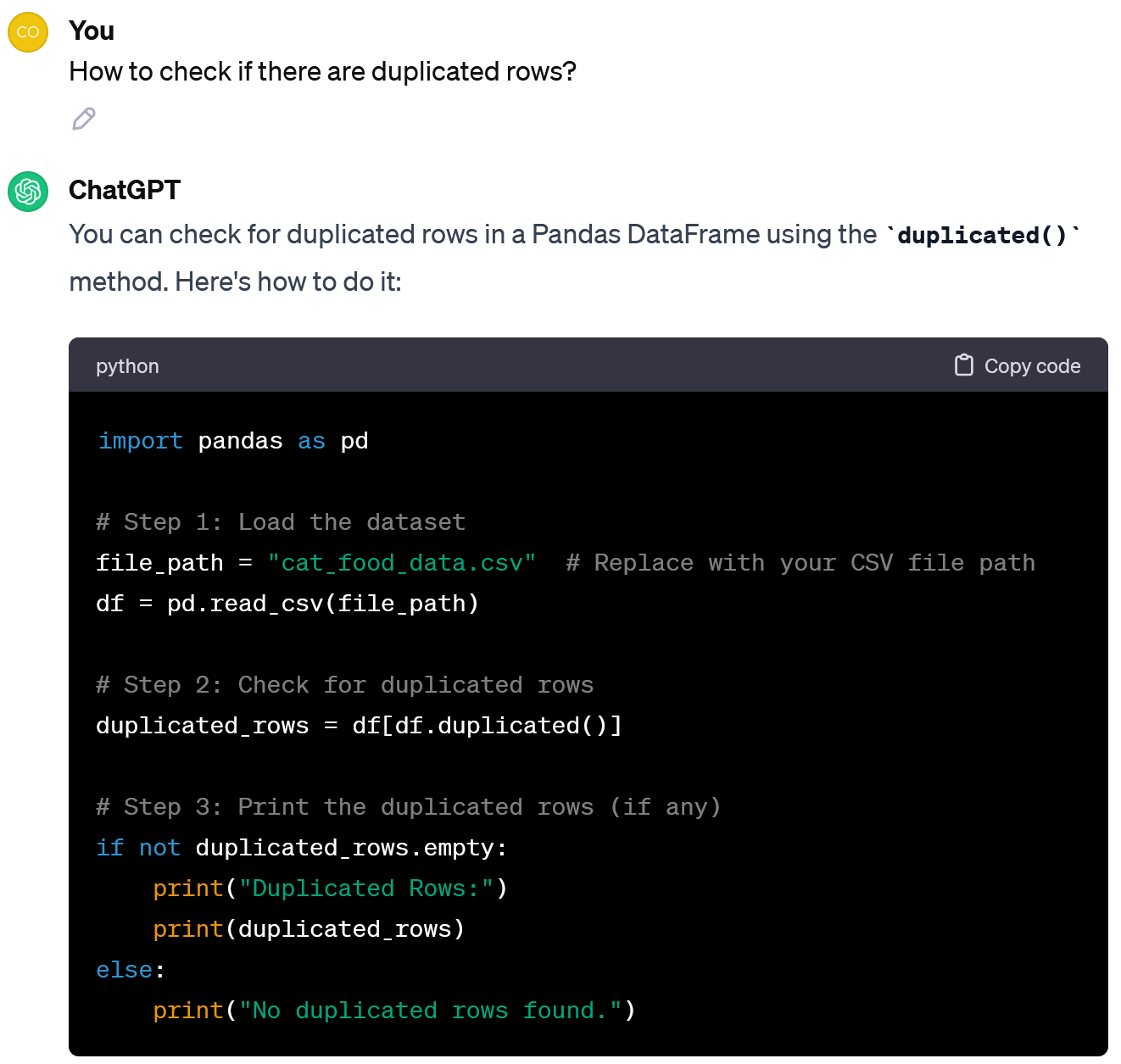
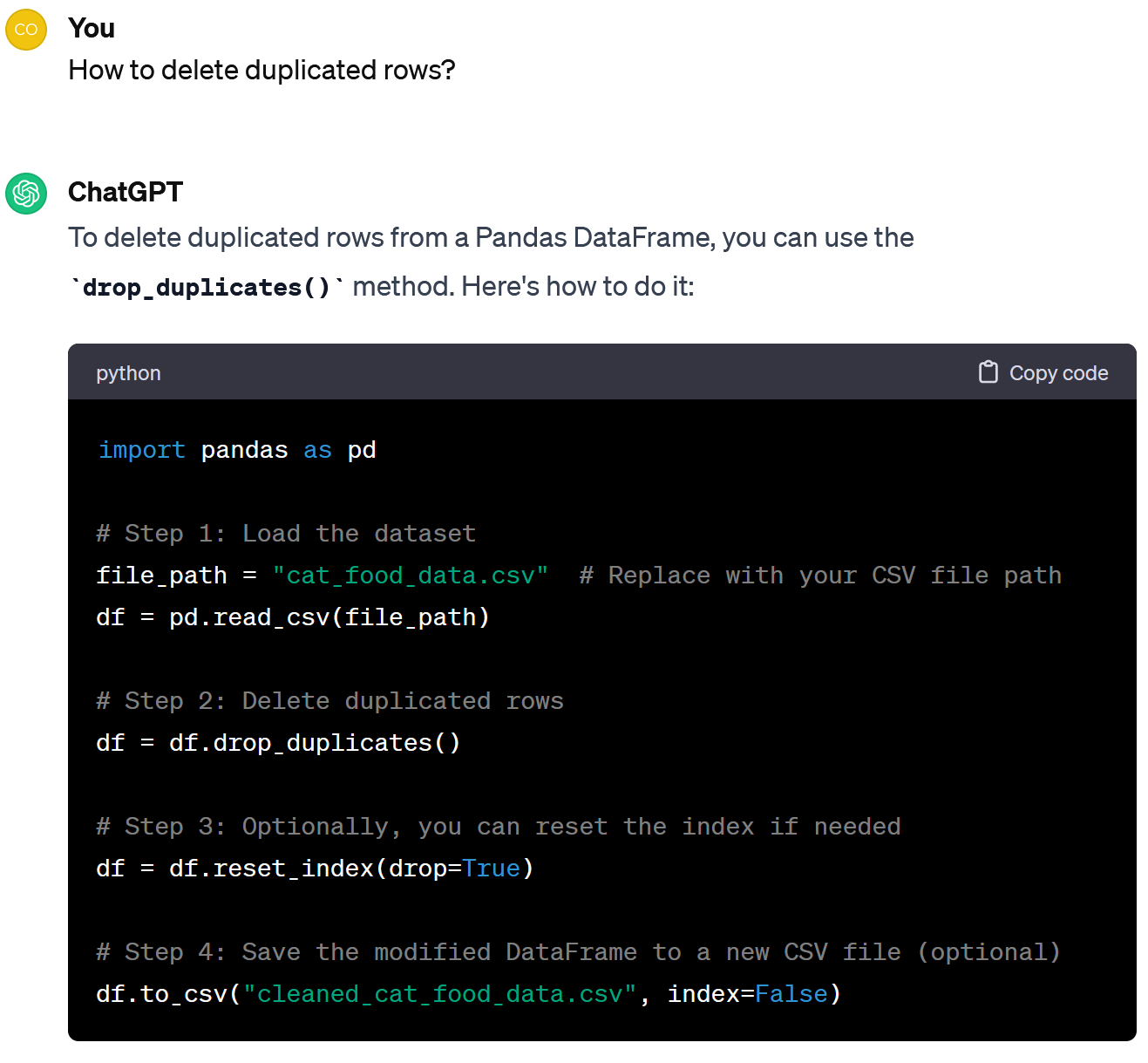
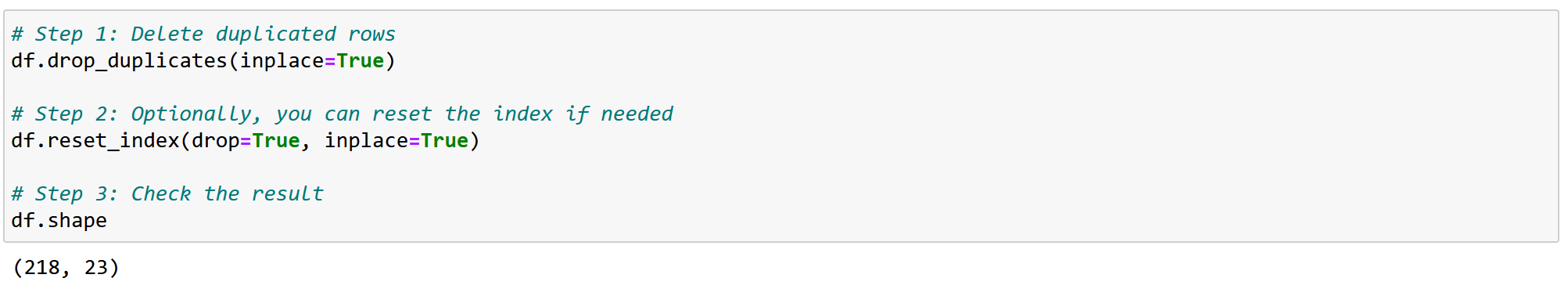
Good idea! In fact, I would also take *products* and *flavours*, because they duplicate the information from other columns; as well as *price\_per\_100g* and *per\_kg\_sell* because they are calculated based on other columns.

You see: There are 23 columns left.

Note that I have not copied the whole answer, just the *drop* command. We have already fulfilled the first part, so there is no need to repeat it.

I also want to check how many columns are left, so I'll add the *len* command.

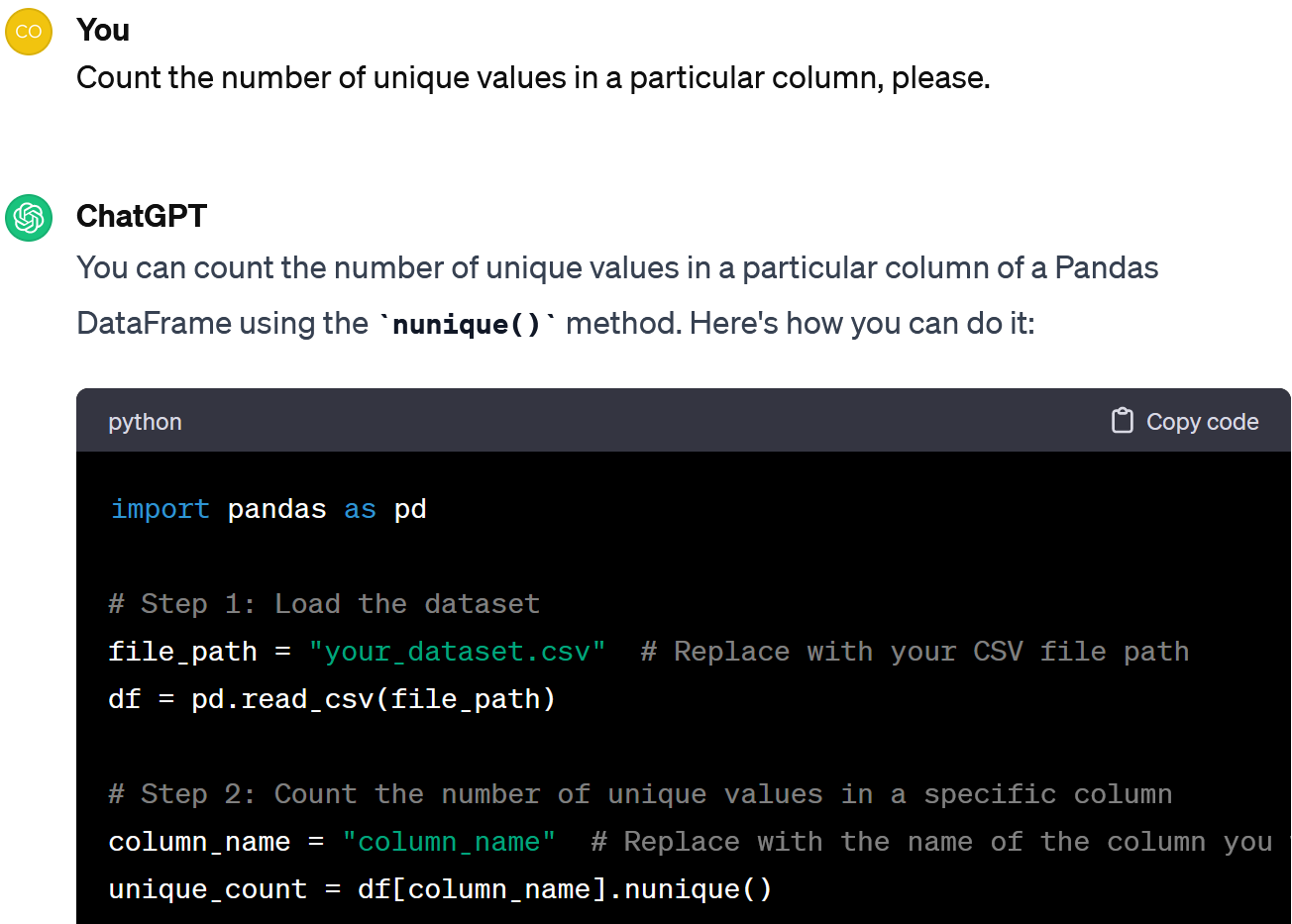
I don’t think that all columns are useful. Don’t you mind if we remove the *Unnamed: 0*, as it duplicates the index? Also it does not allow us to draw any conclusions about the food.



Let’s drop them and check how many columns and rows we will have then.

Processing the fragment from the ChatGPT response, we found 10 duplicate lines.

After deleting duplicated columns, let’s check if there are duplicated rows.

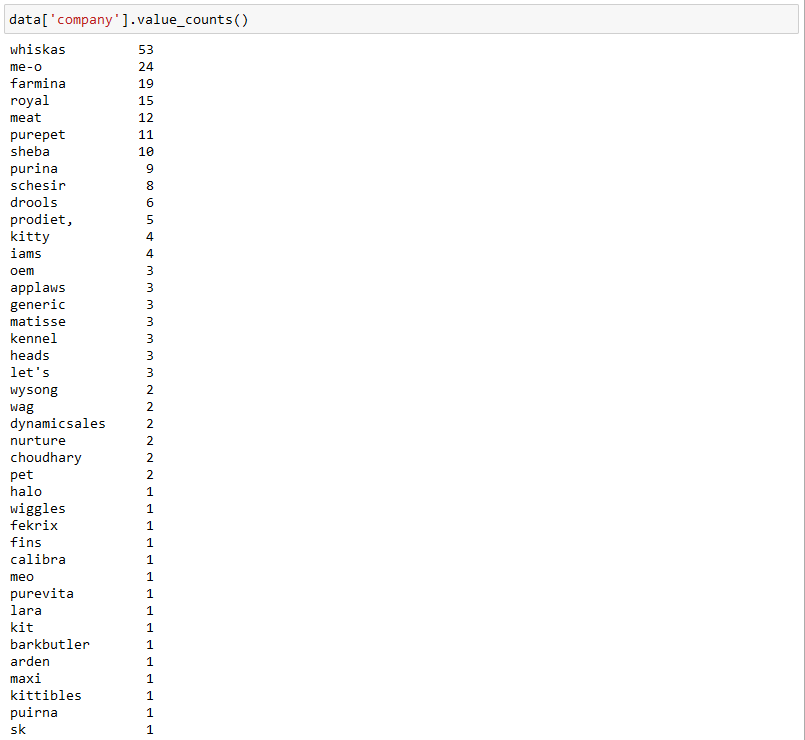
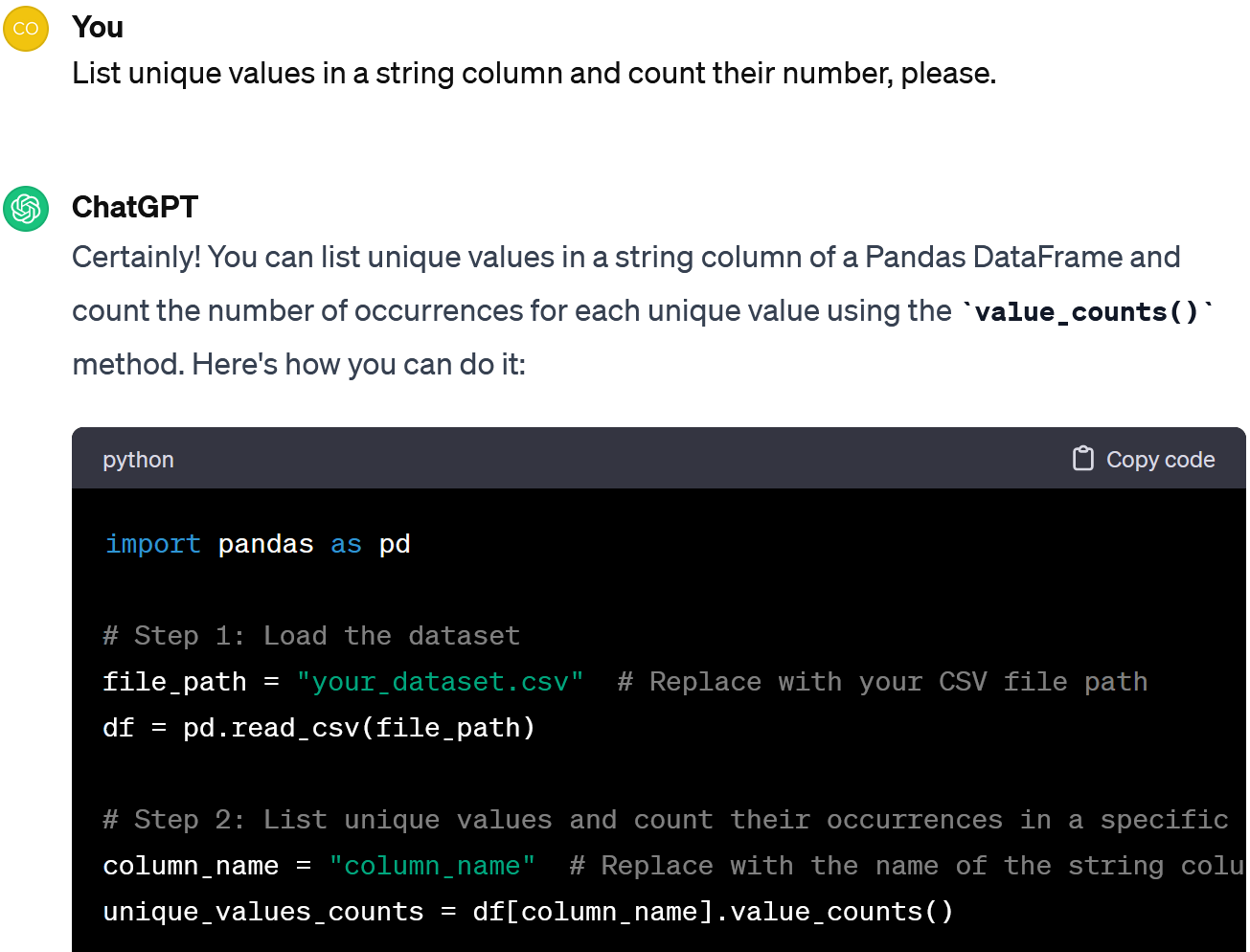
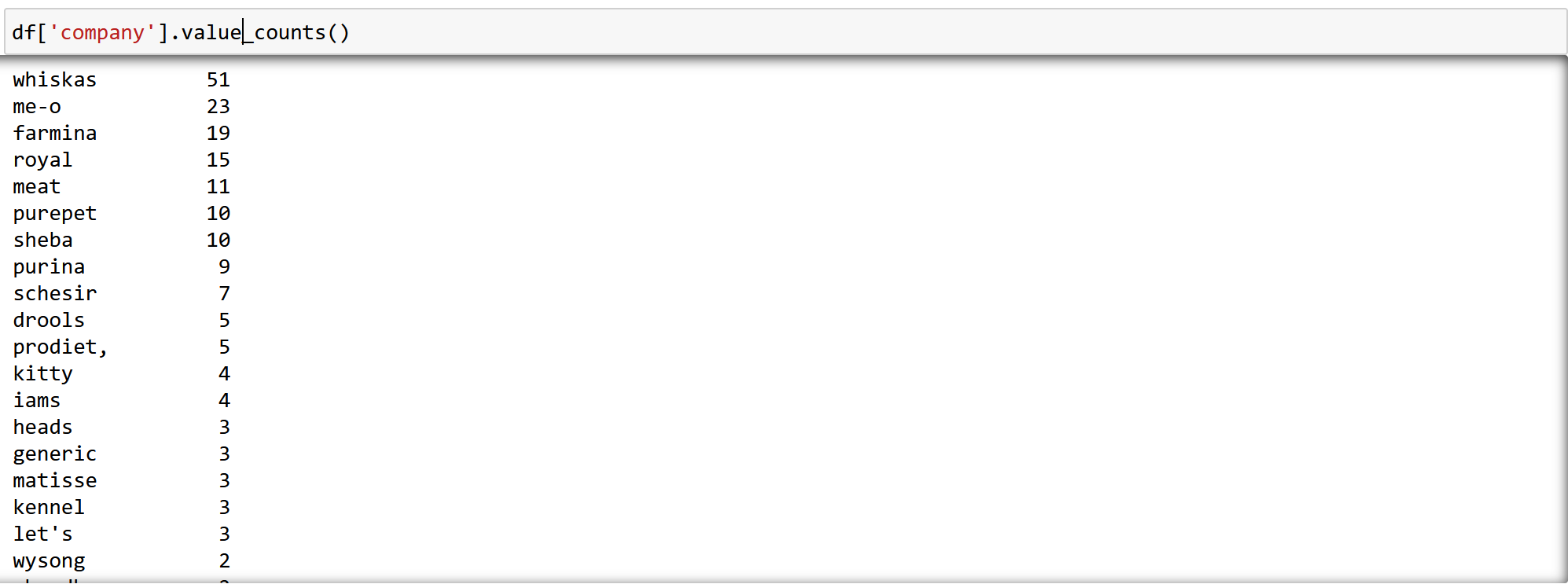
Wow, our database includes products from 41 companies.

Sure. Remember that we have 3 types of features: string, numeric and binary.

Let's start with strings and see how many brands we have in the dataframe.

Can we look at general statistics of these features?

Now our table has 218 rows - product descriptions - and 23 columns - product features.



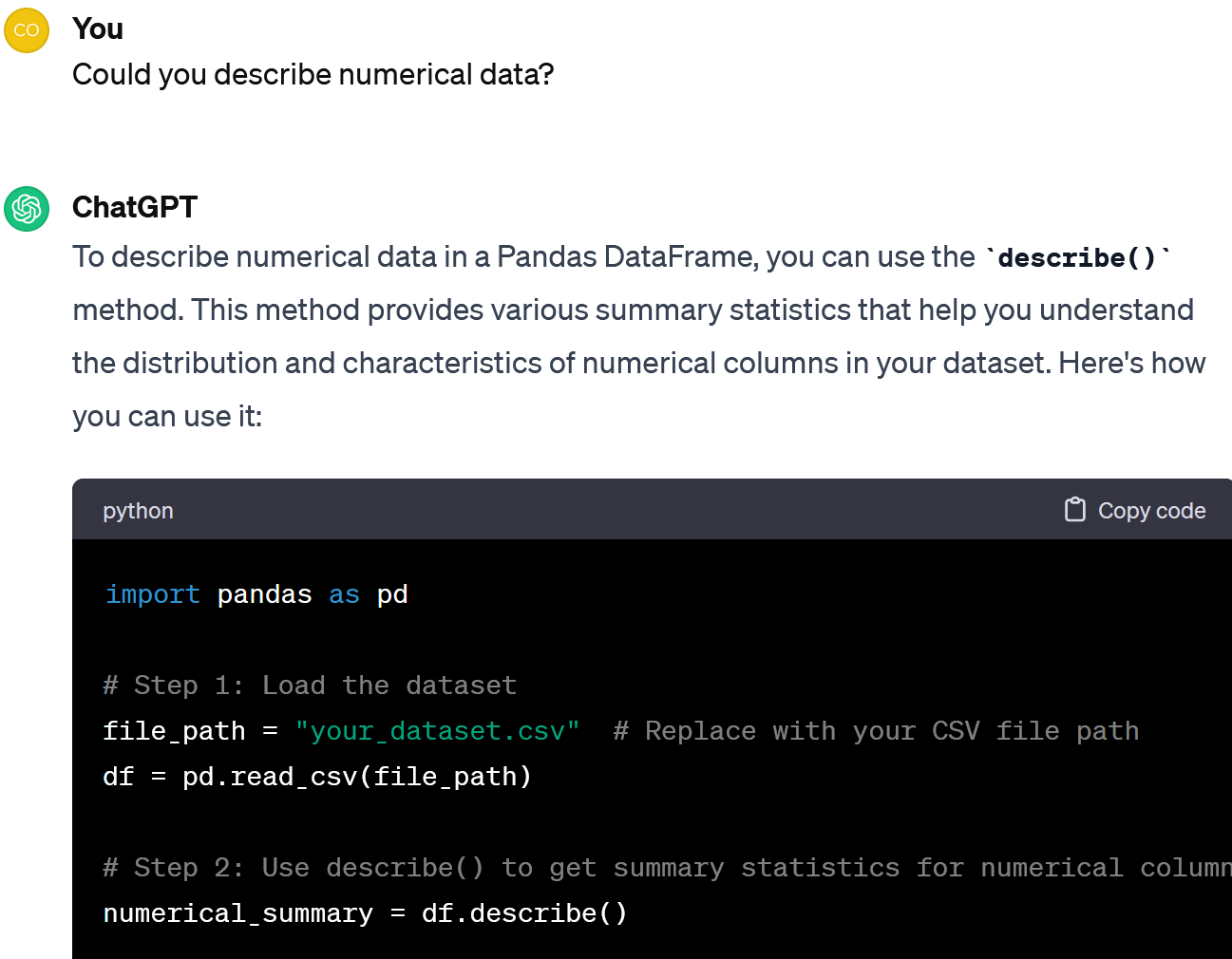
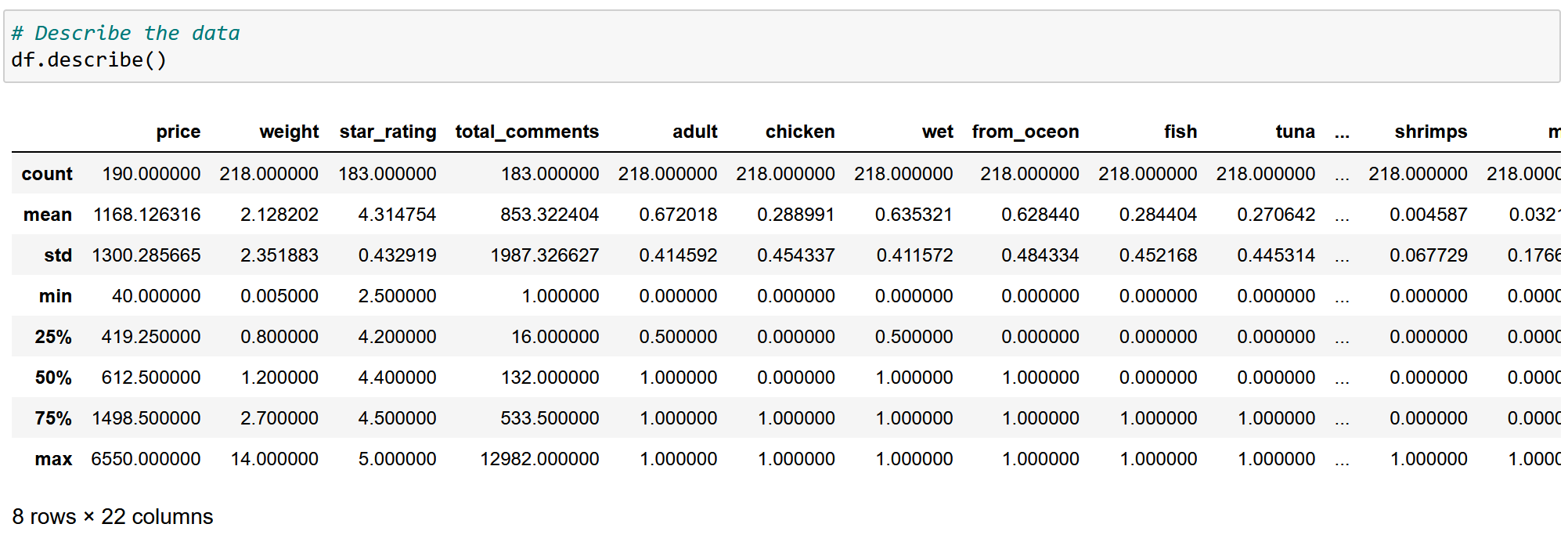
These brands (whiskas, me-o, farmina, royal, meat, purepet, sheba) sold 139 out of 218 products.

Look, only 7 of the 41 companies had products that were bought at least 10 times.

Do we have enough information on all the brands? How many products do we have for each company?

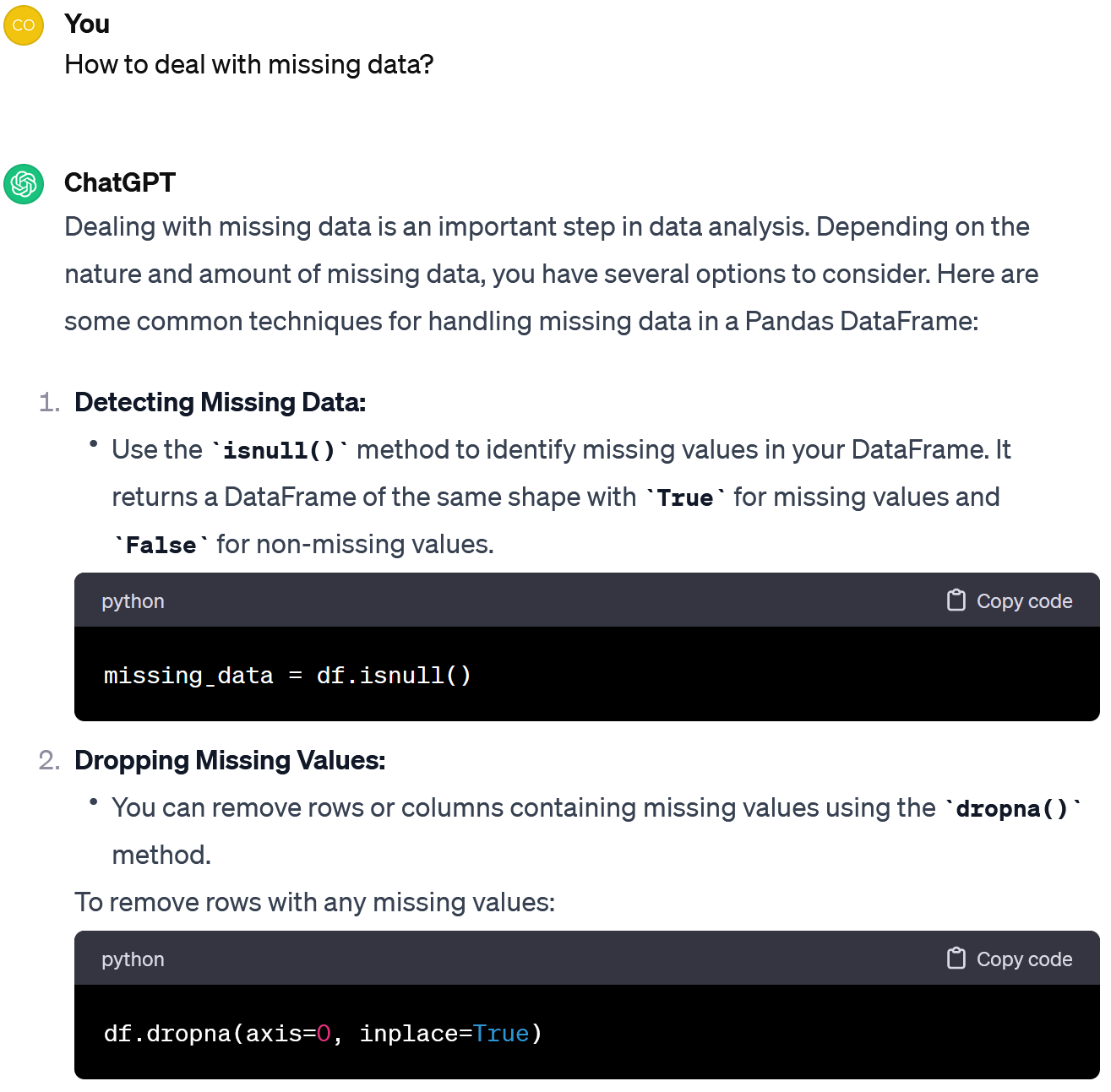
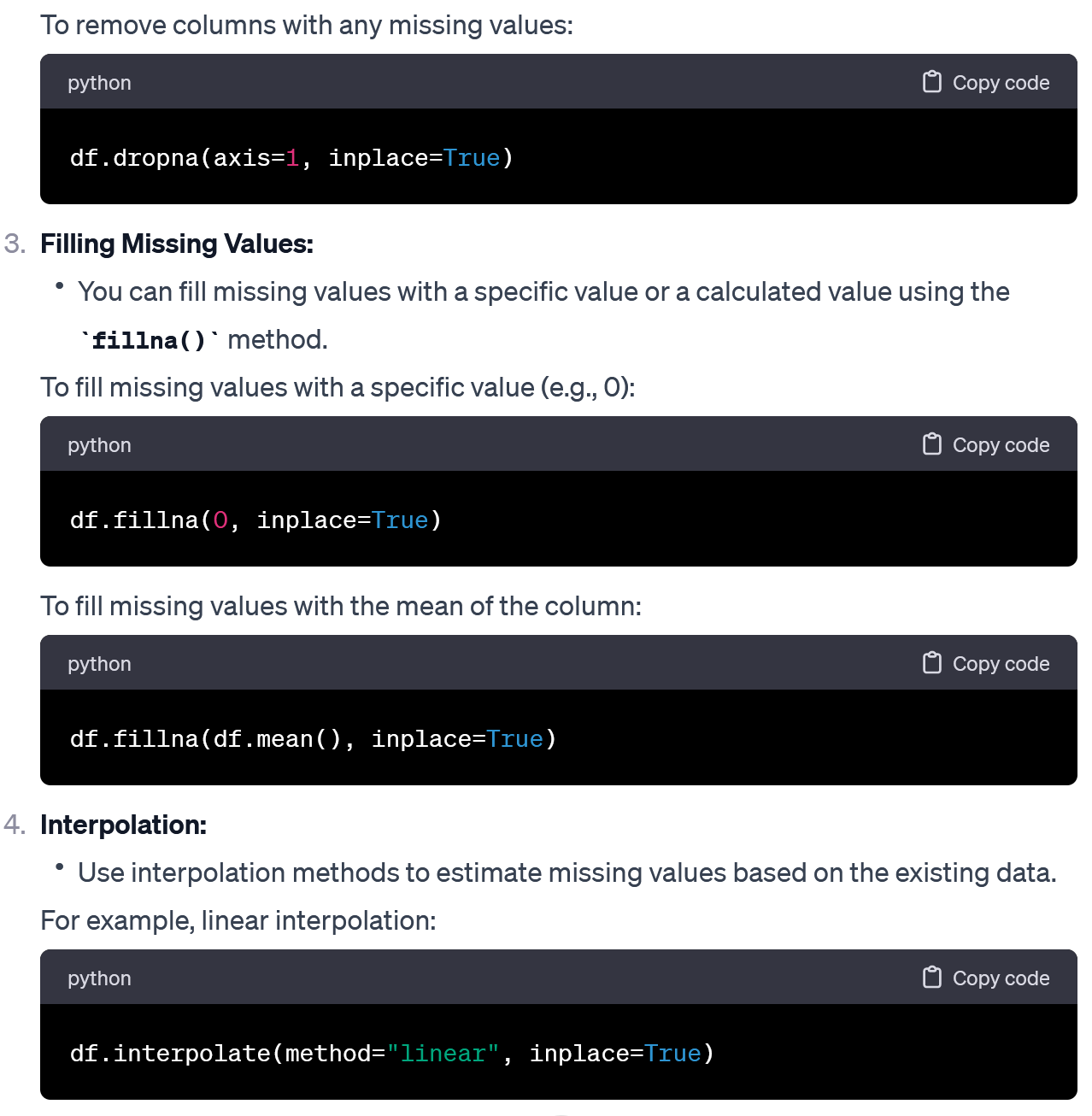
Look, only 7 of the 41 available manufacturers had products bought at least 10 times.

But they sold 144 out of 173 products, so it makes sense to analyse them only (*whiskas, me-o, farmina, royal, meat, purepet* та *sheba*).

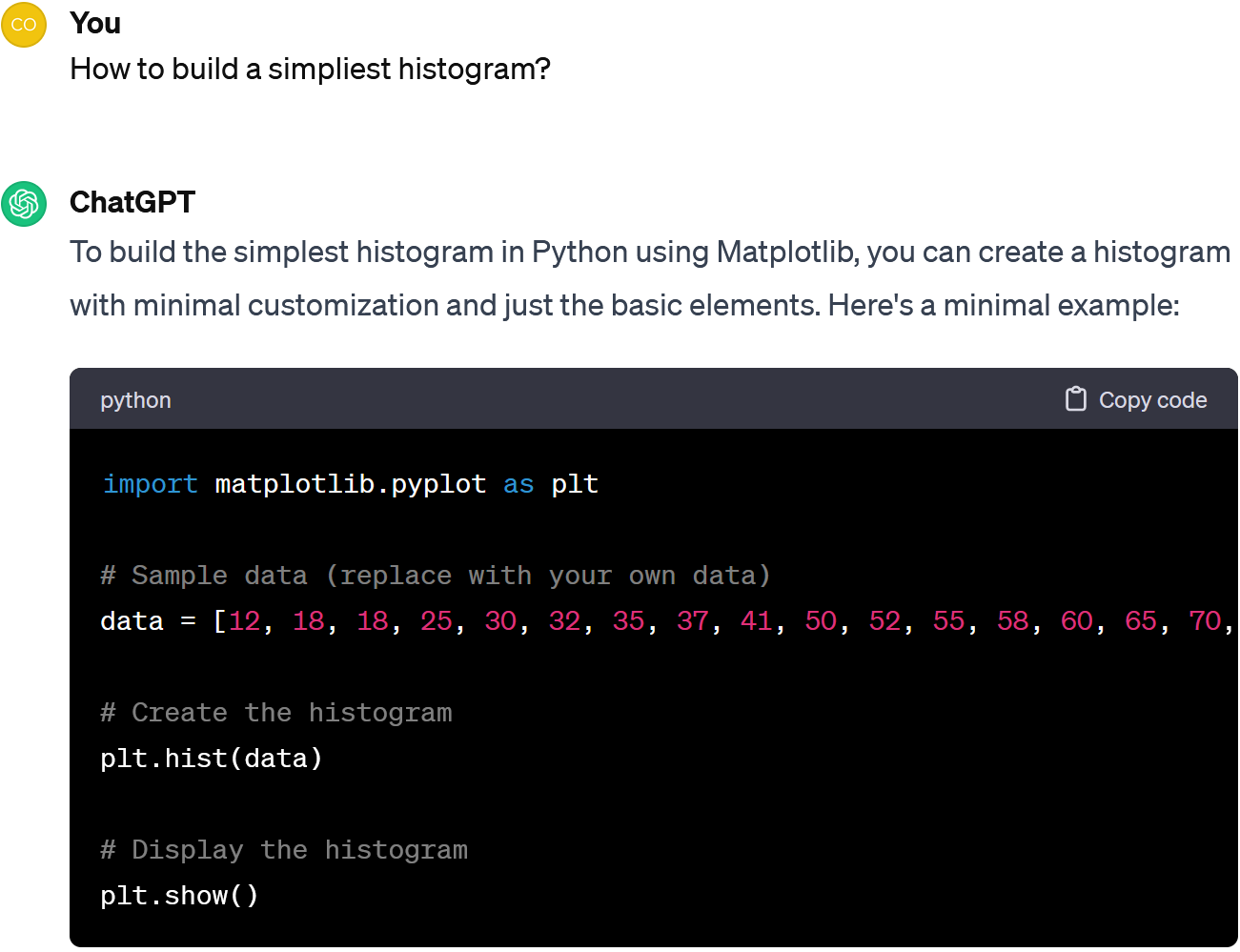


After executing the code, we can analyse the number of non-null values in each numerical column, the mean, the standard deviation, the minimum and maximum values, the quartiles.

Now let's now take a closer look at the numerical data.



As I see, we have only 190 non-null prices and 183 non-null ratings out of 218 entries in our dataframe. How to process the rest?



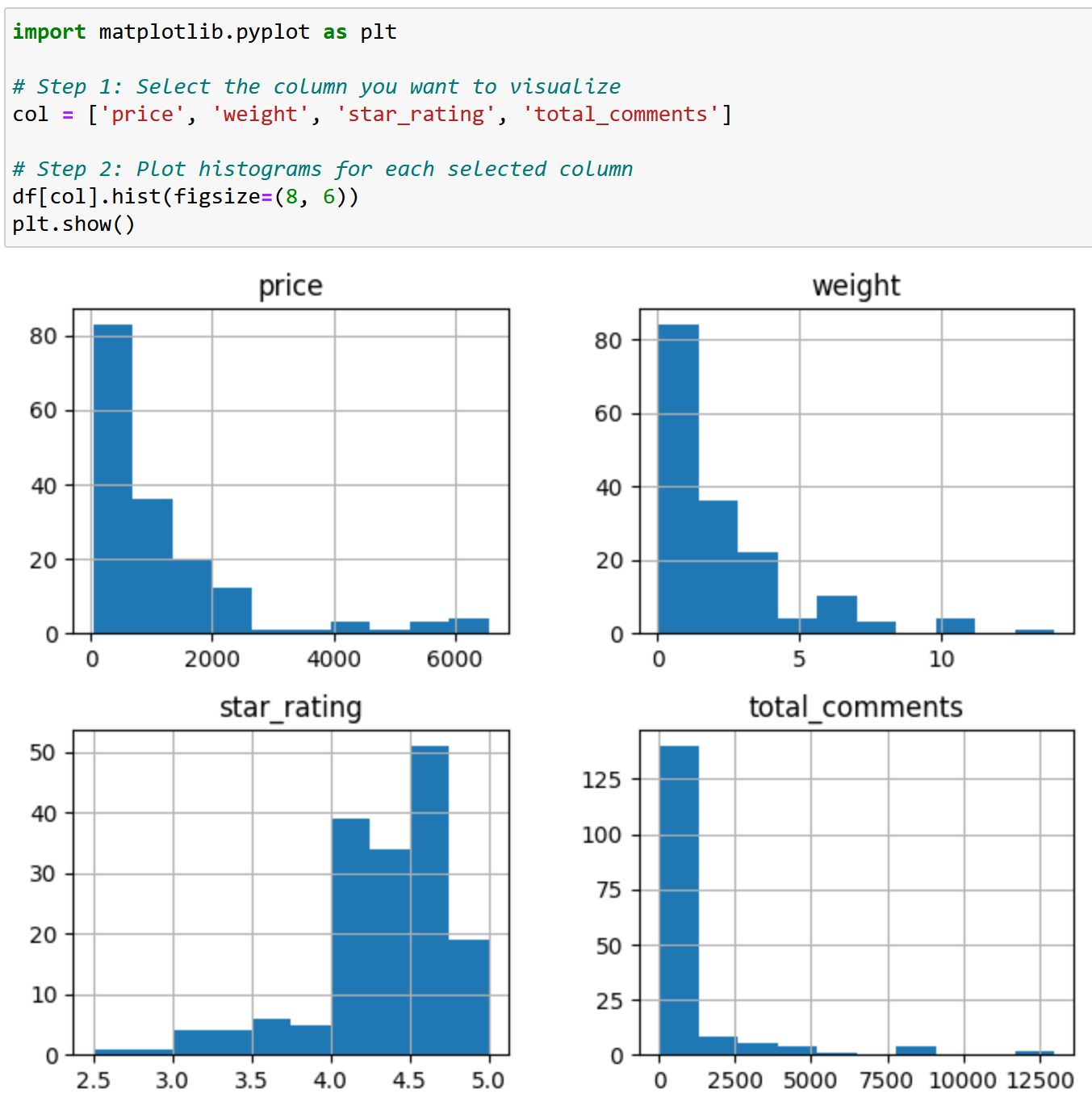
The ChatGPT also offered 5 other ways to deal with missing data. Let's take the easiest one - drop the missing values. But remember that this is not the best idea - you can lose a lot of valuable things by cleaning up the data in such a crude way.

So we need to check how many records are left in the table after dropping missing data.

Look, there are 164 rows. Not too bad.

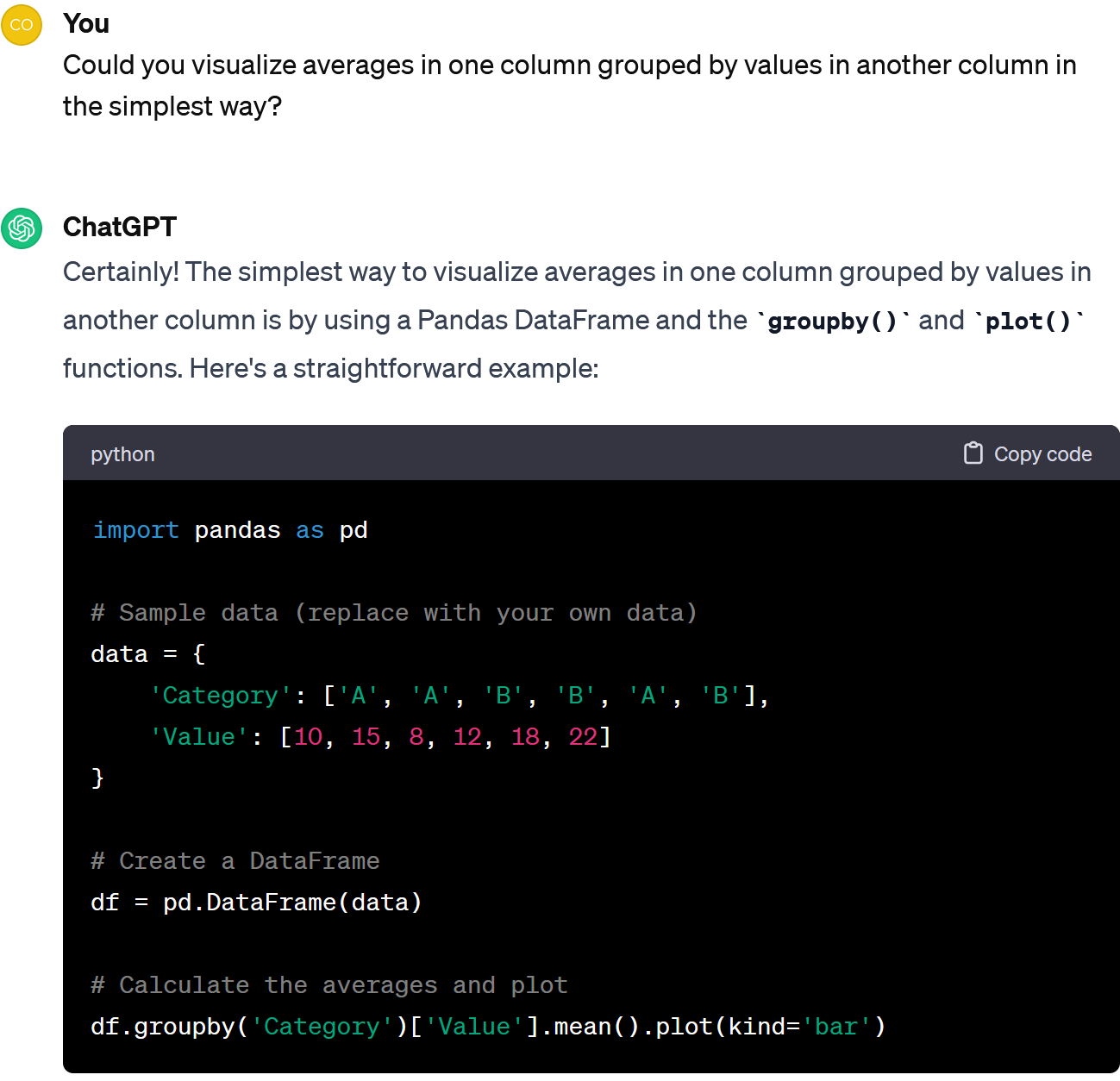
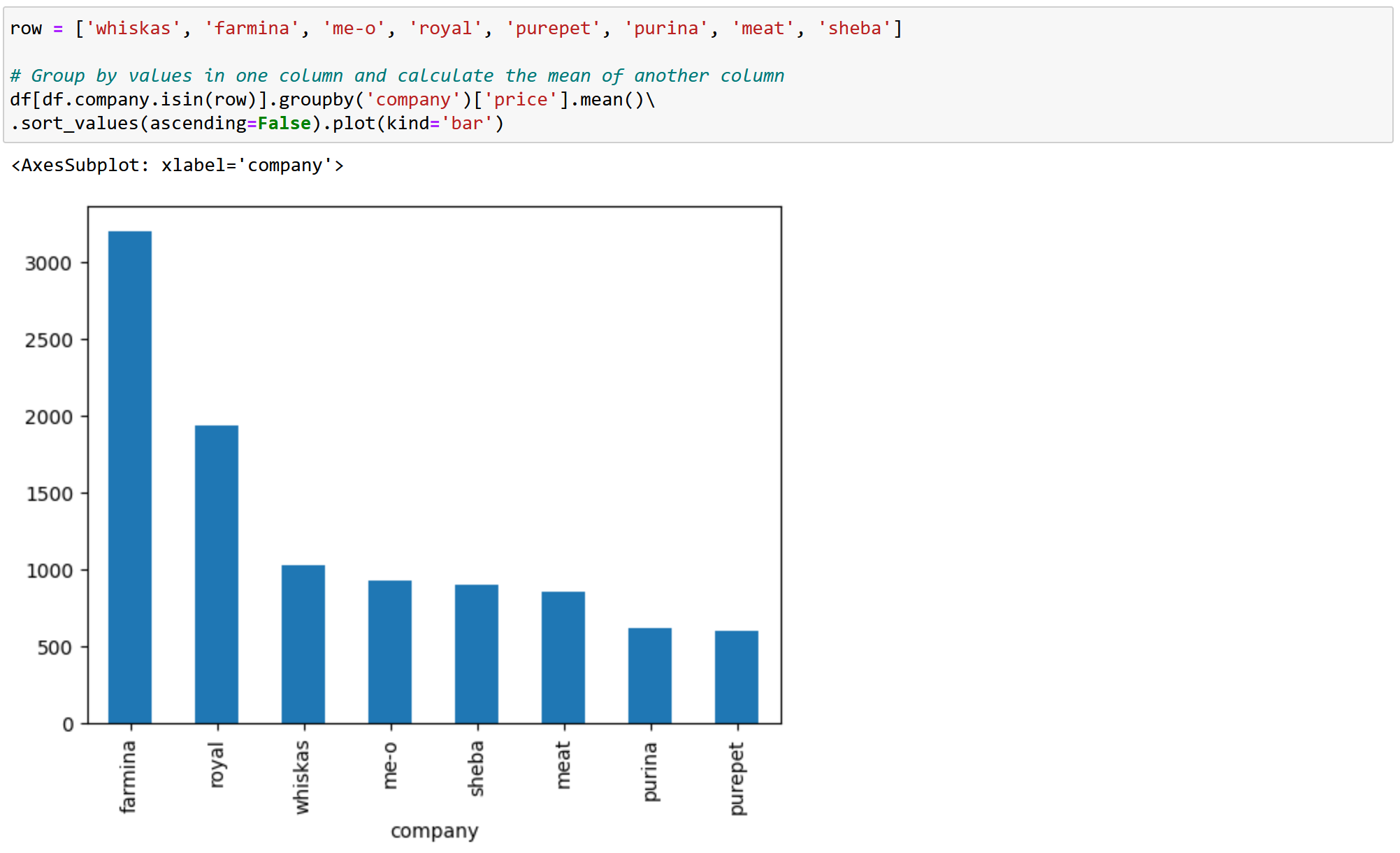
I was also surprised by the maximum price of 6.5K, while the minimum price is 40, given the descriptive statistics we saw above. Perhaps there is an error?

Let's visualize the *price*, *weight*, *star\_rating* and *total\_comments.* That should help.



As you can see, most of the products cost up to 1K, but there are also more expensive ones. At the same time, while most come in packages of up to 2kg, there are also 14kg packages. Perhaps the larger packs cost more?

Or is it just that consumers like them more? You see, although the average rating is more than four, some products received more than 12K comments.

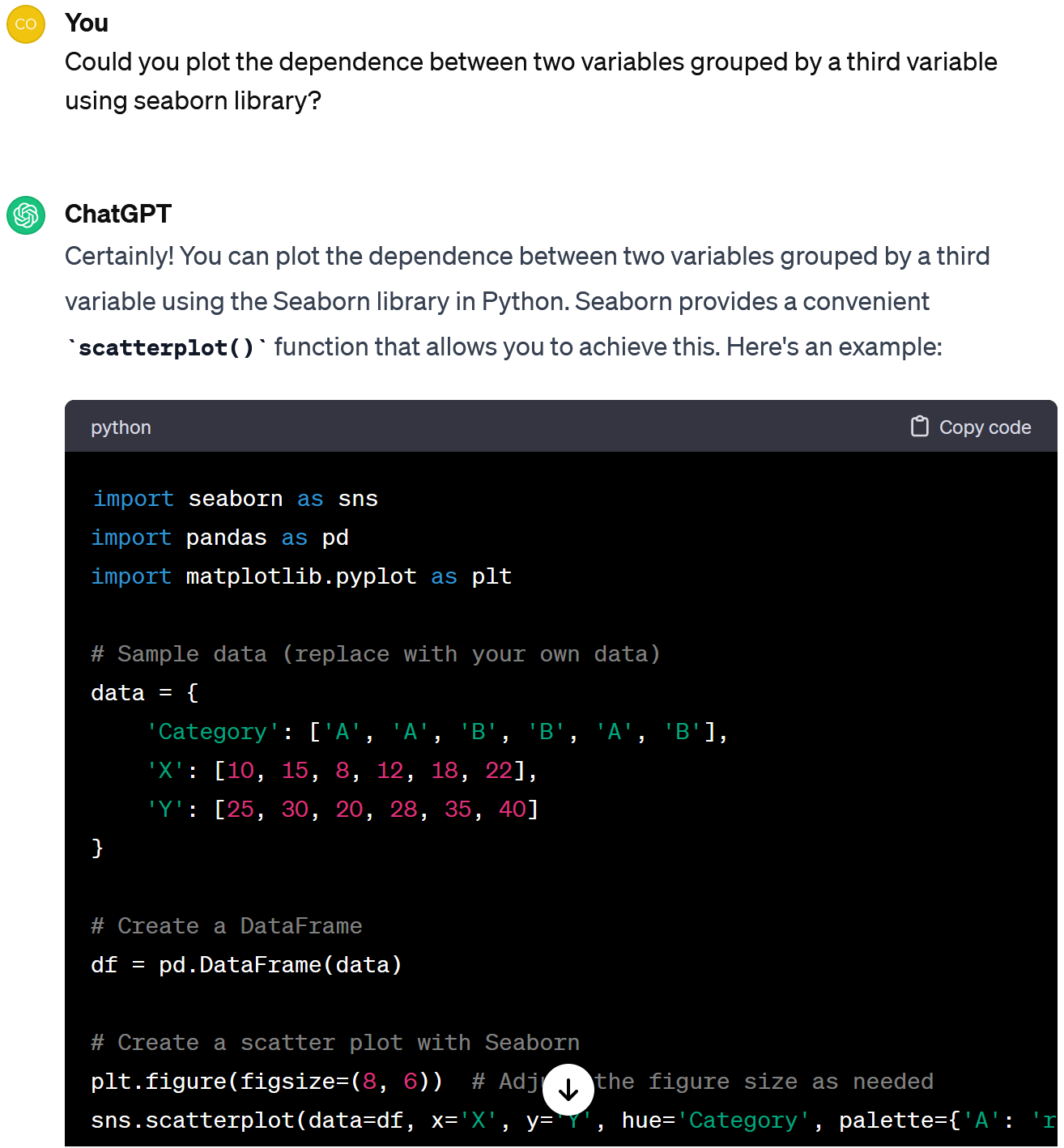
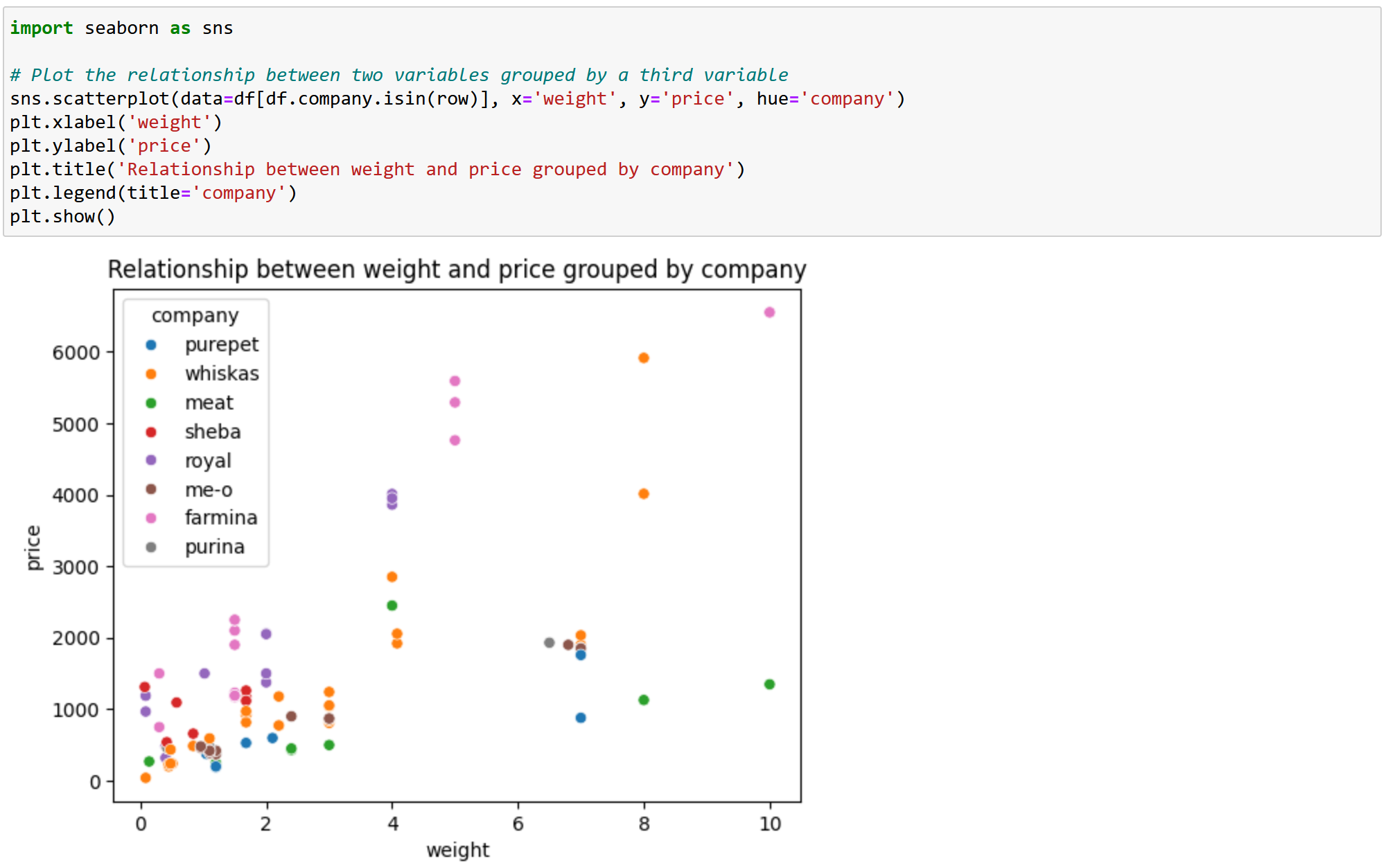


We can see that *Royal* and *Farmina* products are more expensive than other brands.

Agree. Let's compare the average prices of different manufacturers of cat food, focusing only on well represented brands (those with more than 10 products in our dataset).

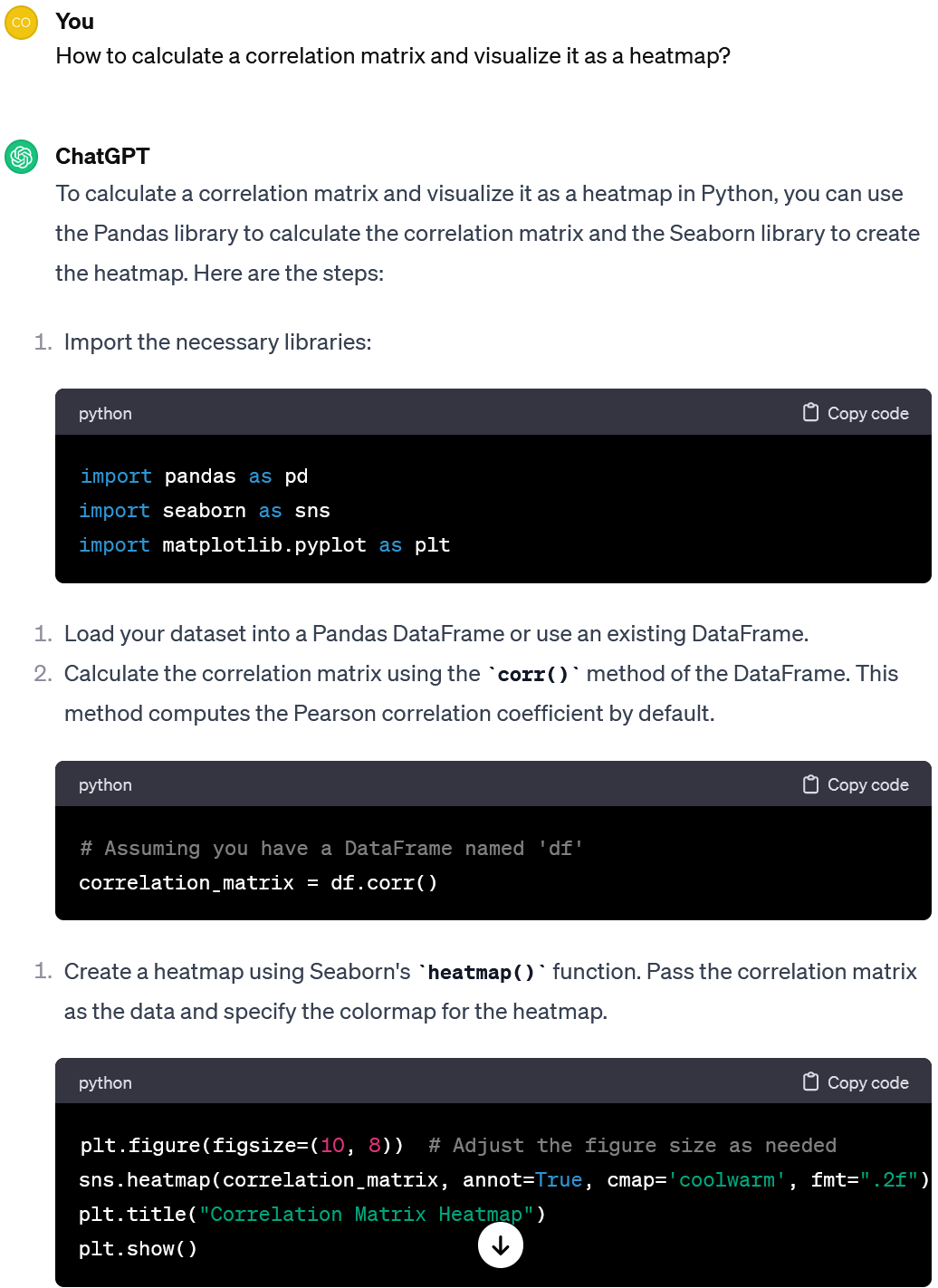
Hmm, cool. But that doesn't help us understand if there are errors in the price.

Could this difference be explained by the pricing policies of the brands? Let's have a look.



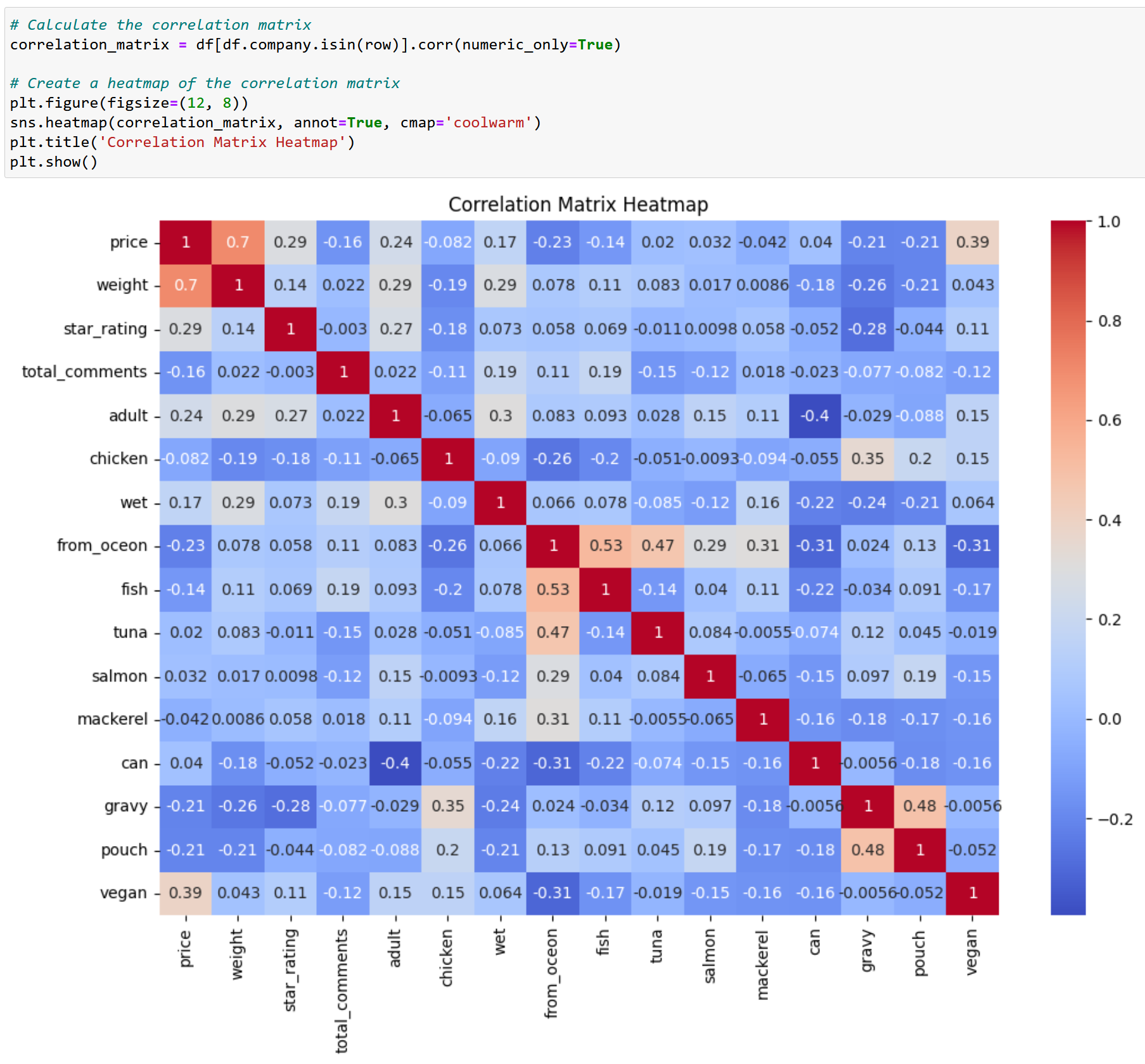
On the graph we can see that the prices by brand are in line with the packaging, so it seems that there are no significant errors here. Anyway, this factor (the company) needs to be taken into account for further analysis.

I agree. Let's see if prices are balanced in relation to weight by brand.



It is clear, but how do you see what has the strongest effect on the price, or on the product rating, or on the number of comments, etc.?

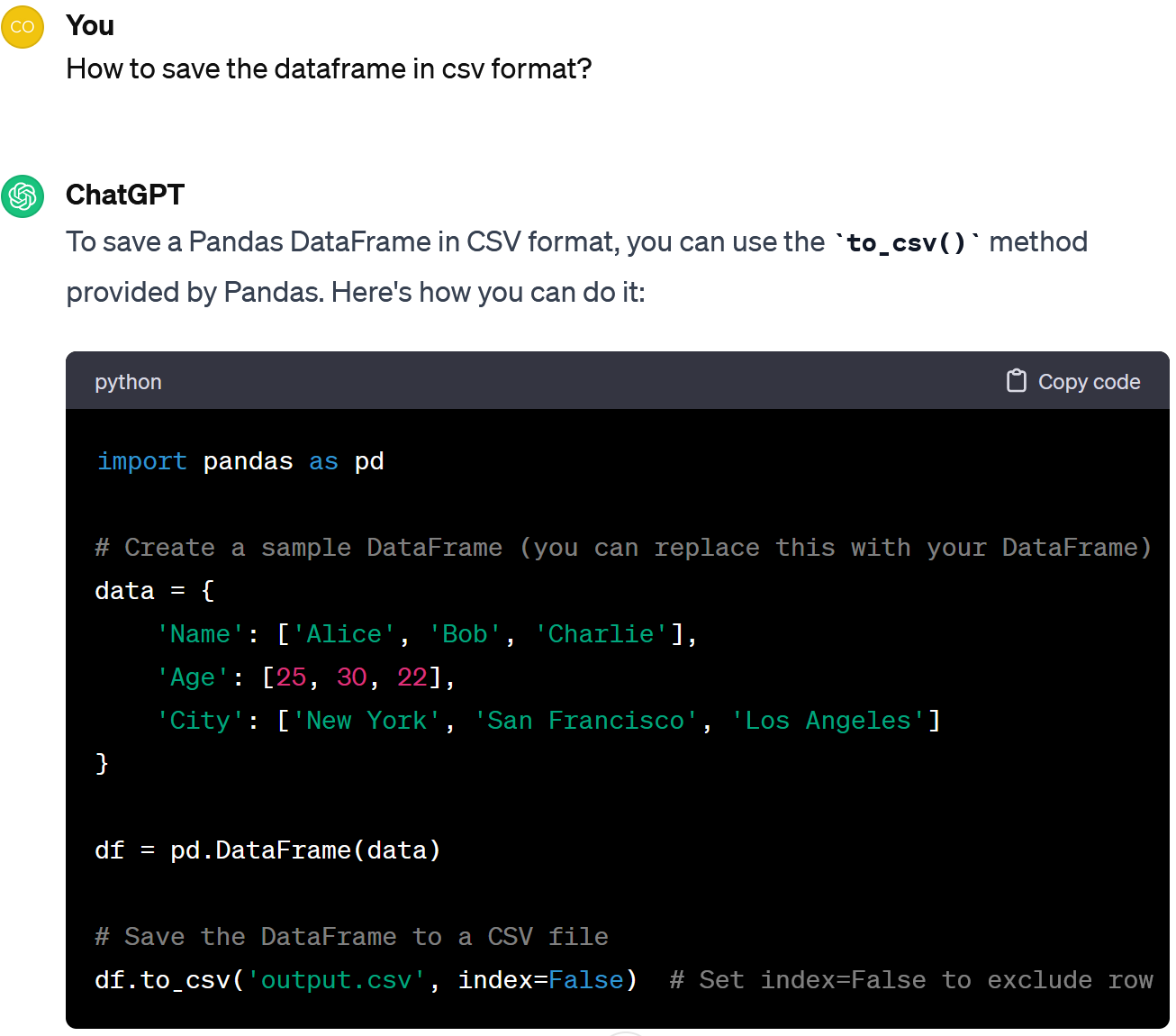
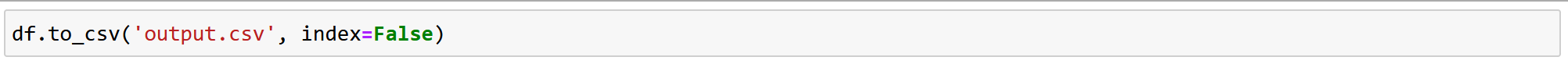
Yep, there is a correlation matrix! Let's ask how to build and visualise it.



Please note! There are 16 features in the matrix. I've removed some binary variables with low variance (indicating the presence of certain ingredients in the cat food). Ask ChatGPT and also do it yourself.

I see that among the quantitative variables, the *price* is most influenced by the *weight* - the correlation coefficient reaches 70%.

There is also a correlation between some ingredients (e.g. *from*\_*ocean*, *fish* and *tuna*; or *gravy* and *pouch*). When modelling, we should drop such features. Do it yourself!



Let's summarise. At the data pre-processing stage, we

1. Downloaded the data and understood what it was about.
2. Checked the number of entries, and the number of features.
3. Cleaned the table of the duplicates.
4. Cleaned the table of records where some data was missing.
5. Analysed the general statistics of the data.
6. Used visualisation to study the brand pricing policies.
7. Created a correlation matrix to analyse the co-influence of variables.

All done!



Now our data is ready for further analysis. Let's save it.