Hi, I'm Freddie, what is your name?
>: C. Ilin

Nice to meet you Cornelia. How can I help you today?
>: I would like to analyze some hospital discharge data (really big!), can you help me with this?

Ok, I can help you with this.
What is your computer's operating system?
>: Windows
Nice. Your drive letter is now set to N:\

What is your input path?
>: Classes, AAE875, DataAnalytics, FinalProgram, Input
Your input path is now set to Classes\AAE875\DataAnalytics\FinalProgram\Input

What is your output path?
>: Classes, AAE875, DataAnalytics, FinalProgram, Output
Your output path is now set to Classes\AAE875\DataAnalytics\FinalProgram\Output

What are the names of your data files?
>: SPARCS2014.csv, SPARCS2015.csv, SCARCS2016.csv

What is the data structure you would like to work with?
>: list(csv)

Ok, good choice. Tell me what you would like to do next?
>: How many columns and row are in my data?

Let me check... oh... this data is really big, C. Ilin!
You have [add here] inpatient discharges and [add here] variables that document these observations.
>: Ok, sounds a bit intimidating but let's move on.
>: My final objective is to run a linear regression but I would like to clean the data and run some statistics first.

No problem, I can help you with this. Let's clean the data first.
Would you like to drop observations with missing values?
>: Yes, please.

I have removed all the missing values in your data.
You now have **X million** inpatient discharges and **Y variables** that document these observations.
>: Still lots of data to play with, right?

Right... would you like to remove data outliers?
>: Yes, let's do this as well.

I have removed all outliers in your data.
You now have [add here] inpatient discharges and [add here] variables that document these observations.
>: Got it, thanks - you are really good at this!
>: Freddie, let's run some summary statistics now.

I am afraid this is not the best way to move forward. The variable names in your data are not consistent over time.
Let me put these in a table for you. Please see below:
[Print table here]
>: I see... variable names are not consistent over time, good catch. Could you please change the names for years 2014, 2015 to align with those in year 2016?

Sure thing! I will use a dictionary for this. Processing...
Variable names match those in year 2016 now. Please see below:
[Print table here]
>: Great! Are we ready to move to the descriptive statistics part?

Yes, unless you want to do more data cleaning?
>: Let's keep going, we may see more inconsistencies when we plot statistics

Sure thing! I can plot some graphs for you. What do you want to know more precisely?
>: I would like to see how many patients with asthma conditions were admitted in the hospital by year and type of admission

This is a very interesting question, C. Ilin. Let me prepare some graphs for you...
Here is what I found:
[Plot graphs here]
>: Hmm, that's very interesting. [Comment more on the results here]

Indeed, very interesting. Would you like me to plot some more graphs?
>: Yes, please. Let's see what is the major payor for patients with asthma conditions.

It seems that there are 10 types of payment sources. I think a pie chart is more appropriate here. Do you agree?
>: Yes, let's plot a pie chart

Interesting results again. Please see below what I found in the data:
[Plot graphs here]
>: This is really cool stuff, Freddie!

C.Ilin - I agree with you. In the meanwhile, I need a break. I will be back in 15 minutes.
>: Enjoy!

Ok, I am back. You mentioned your main objective is to fit a linear regression model. I am ready whenever you are ready.
>: Let's do it, Freddie.

What is your question?
>: I would like to see if there is any casual effect between hospital length of stay for patients with asthma conditions and health insurance status.

This sounds like a very interesting question. What would be the control variables then?
>: Let's try these: gender, race, type of admission, patient disposition, health service area, facility, year

Ok, I will put all these in an OLS model.
Here are the results:
[Print regression analysis results]
>: Freddie, this is fantastic! Thanks much for your help!

Of course, any time.
Goodbye C.Ilin
>: Goodbye Freddie