Python Data Analysis - Final project
Hanan SALAM

Context

This final project is based on a real data analyst life scenario:

You've just been hired as a Data Analyst at "Home Credit", a loan company, and a business owner comes to you.

Business Owner: "Hey you're the new kid? Doing nerdy stuff eh? Here is some data for ya".

He hands you a USB key

Business Owner: "You're welcome. Now do some magic with it".

And then he stays there looking fixedly at your laptop screen…
awkward silence
It's like he's waiting for you to do something

Alright the context is set, now you know what to do (or not). Let's look at the data!

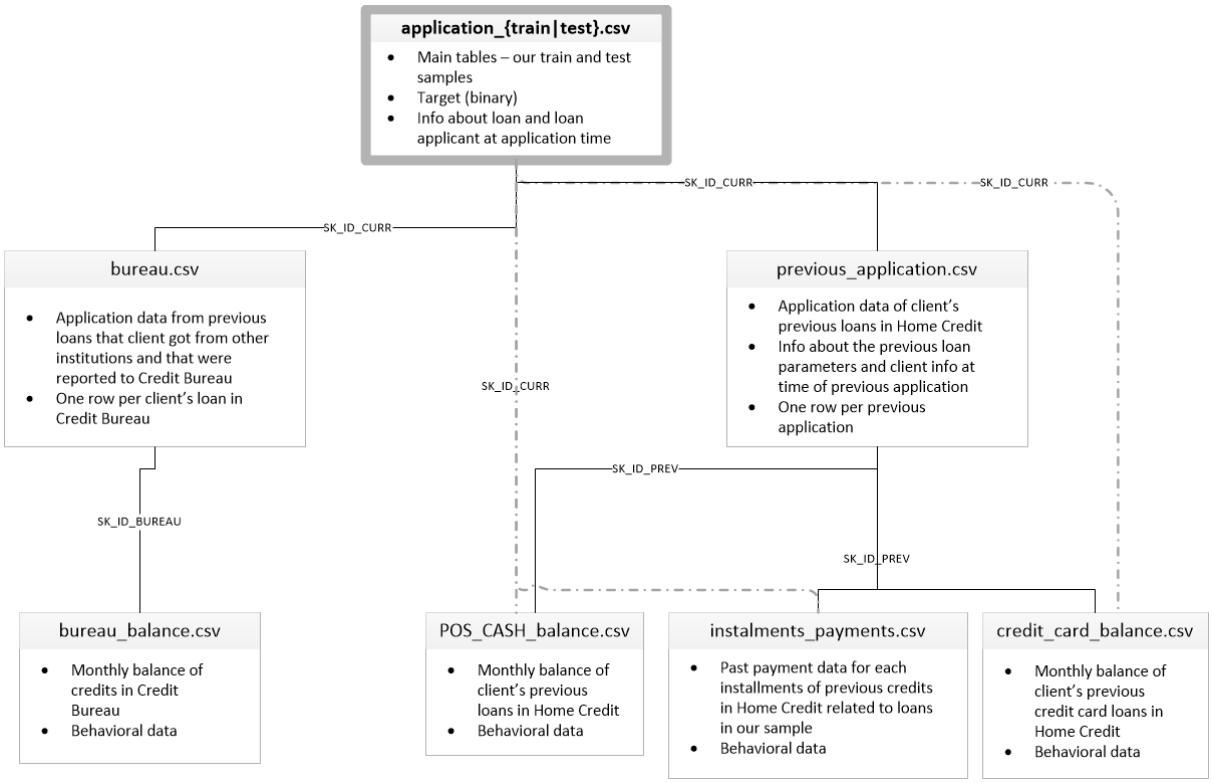
Data

You find 8 csv files inside his USB device, along with a readme file describing which file is what :

- **application_{train|test}.csv**

    - This is the main table, broken into two files for Train (with TARGET) and Test (without TARGET).
    - Static data for all applications. One row represents one loan in our data sample.
- **bureau.csv**
    - All client's previous credits provided by other financial institutions that were reported to Credit Bureau (for clients who have a loan in our sample).
    - For every loan in our sample, there are as many rows as number of credits the client had in Credit Bureau before the application date.
- **bureau_balance.csv**
    - Monthly balances of previous credits in Credit Bureau.
    - This table has one row for each month of history of every previous credit reported to Credit Bureau – i.e the table has (#loans in sample * # of relative previous credits * # of months where we have some history observable for the previous credits) rows.
- **POS_CASH_balance.csv**
    - Monthly balance snapshots of previous POS (point of sales) and cash loans that the applicant had with Home Credit.
    - This table has one row for each month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in our sample – i.e. the table has (#loans in sample * # of relative previous credits * # of months in which we have some history observable for the previous credits) rows.
- **credit_card_balance.csv**
    - Monthly balance snapshots of previous credit cards that the applicant has with Home Credit.
    - This table has one row for each month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in our sample – i.e. the table has (#loans in sample * # of relative previous credit cards * # of months where we have some history observable for the previous credit card) rows.
- **previous_application.csv**
    - All previous applications for Home Credit loans of clients who have loans in our sample.
    - There is one row for each previous application related to loans in our data sample.
- **installments_payments.csv**
    - Repayment history for the previously disbursed credits in Home Credit related to the loans in our sample.
    - There is a) one row for every payment that was made plus b) one row each for missed payment.
    - One row is equivalent to one payment of one installment OR one installment corresponding to one payment of one previous Home Credit credit related to loans in our sample.
- **HomeCredit_columns_description.csv**
    - This file contains descriptions for the columns in the various data files.

It looks like data about loan applications and applicants.



(This diagram gives you joining keys to merge datasets altogether)

Ok that dude has given us food for data scientists, you can tell because the application data is split into two files, one for training a predictive model and another one to test it. You're not data scientist or madam Irma so you're not going to predict anything. But you want this guy to know how good you are, he can't tell the difference between data analysts and data scientists anyway.. We're all data somethings to him, let's teach him about Data Analysts!

As a Data Analyst you're thing is to tell stories about data, stories no one ever told, and whether the story is true or not, you have to be convincing. So, let's forget about the predicting thing and make him a story he'll find interesting as a business owner.

The last file (**HomeCredit_columns_description.csv**) looks like a good start.

Ok, funny.. seriously what should I do ?

Make a notebook telling interesting things about this data, tell a story (or many) using everything you learned. There are many hidden stories in data. You have to submit at least a notebook and any resources you used (like images or any other files).

Here are the criteria we will use to assess your work:

Is it meaningful?

As a data analyst you have to produce something meaningful enough, just plotting random data is not going to work. Like a story your analysis should have some kind of logical progression.

How well did you use the technical knowledge you've been taught?
Obviously, the way you use everything you learned during the lectures is going to be assessed.

Cleanliness, aesthetics and clearness of your notebook
Is your analysis full of unused code? Is it difficult to read? Have you tried to make it easy and enjoyable to read?

Innovation
Creativity, surprising things or any good initiatives you take are potential bonus points.

Careful:
This work is individual, plagiarism is going to be measured by both machines and humans. Too many similarities between your work and any online or python buddy work will result in grade penalties.

Good Luck!