**Customer Segmentation Project**

Week 8

July 18, 2022

Contents

[**1.** **Group Information** 2](#_Toc108941859)

[**2.** **Problem understanding** 2](#_Toc108941859)

[**3.** **Data understanding** 2](#_Toc108941859)

[**4.** **What type of data you have got for analysis?**............................................................................ 5](#_Toc108941859)

1. [**5.** **What are the problems in the data ( number of NA values, outliers , skewed etc)**…….6](#_Toc108941859)
2. **Group Information**

Group Name**:** M.A.S

Specialization: Data Science

Submitted to: Data Glacier canvas platform

Internship Batch: [LISUM10: 30](https://canvas.instructure.com/courses/4851447)

|  |  |  |  |
| --- | --- | --- | --- |
| **Group Members** | **Three members** | | |
| **Name** | **Email** | **Country** | **Collage/Company** |
| Moath Mohammed Bin Musallam | [moathmusallam@gmail.com](mailto:moathmusallam@gmail.com) | Saudi Arabia | University of East Anglia |
| Andrew Kojo Mensah-Onumah | [kojoakmo@gmail.com](mailto:kojoakmo@gmail.com) | Ghana | Data Glacier |
| Shaimaa Saleh Obad Al-khawlani | [s.khawlany@gmail.com](mailto:s.khawlany@gmail.com) | Yemen | Data Glacier |

1. **Problem description**

Most banks around the world have variant large customer base with different income levels, ages, characteristics, values and lifestyles.

XYZ bank wants to increase the production and the satisfactions of all customers categories by roll out Christmas offers to their customers.

But Bank does not want to roll out same offer to all customers instead they want to roll out personalized offer to particular set of customers. If they manually start understanding the category of customer then this will be not efficient and also, they will not be able to uncover the hidden pattern in the data (pattern which group certain kind of customer in one category).

1. **Data Understanding**

The existing data, which was provided by the bank, is the bank's customers data. However, the data contains many columns that will help the analytics team analyze the data and build a customer segmentation approach for the bank.

Since the data does not contain a dependent variable or (Target), We believe that machine learning (clustering) techniques would be appropriate to use for this type of data.

**Size:** 1000000 records, 48 columns.

## **Columns Description:**

|  |  |
| --- | --- |
| Column Name | Description |
| fecha\_dato | The table is partitioned for this column |
| ncodpers | Customer code |
| ind\_empleado | Employee index: A active, B ex employed, F filial, N not employee, P pasive |
| pais\_residencia | Customer's Country residence |
| sexo | Customer's sex |
| age | Age |
| fecha\_alta | The date in which the customer became as the first holder of a contract in the bank |
| ind\_nuevo | New customer Index. 1 if the customer registered in the last 6 months. |
| antiguedad | Customer seniority (in months) |
| indrel | 1 (First/Primary), 99 (Primary customer during the month but not at the end of the month) |
| ult\_fec\_cli\_1t | Last date as primary customer (if he isn't at the end of the month) |
| indrel\_1mes | Customer type at the beginning of the month ,1 (First/Primary customer), 2 (co-owner ),P (Potential),3 (former primary), 4(former co-owner) |
| tiprel\_1mes | Customer relation type at the beginning of the month, A (active), I (inactive), P (former customer),R (Potential) |
| indresi | Residence index (S (Yes) or N (No) if the residence country is the same than the bank country) |
| indext | Foreigner index (S (Yes) or N (No) if the customer's birth country is different than the bank country) |
| conyuemp | Spouse index. 1 if the customer is spouse of an employee |
| canal\_entrada | channel used by the customer to join |
| indfall | Deceased index. N/S |
| tipodom | Addres type. 1, primary address |
| cod\_prov | Province code (customer's address) |
| nomprov | Province name |
| ind\_actividad\_cliente | Activity index (1, active customer; 0, inactive customer) |
| renta | Gross income of the household |
| ind\_ahor\_fin\_ult1 | Saving Account |
| ind\_aval\_fin\_ult1 | Guarantees |
| ind\_cco\_fin\_ult1 | Current Accounts |
| ind\_cder\_fin\_ult1 | Derivada Account |
| ind\_cno\_fin\_ult1 | Payroll Account |
| ind\_ctju\_fin\_ult1 | Junior Account |
| ind\_ctma\_fin\_ult1 | Más particular Account |
| ind\_ctop\_fin\_ult1 | particular Account |
| ind\_ctpp\_fin\_ult1 | particular Plus Account |
| ind\_deco\_fin\_ult1 | Short-term deposits |
| ind\_deme\_fin\_ult1 | Medium-term deposits |
| ind\_dela\_fin\_ult1 | Long-term deposits |
| ind\_ecue\_fin\_ult1 | e-account |
| ind\_fond\_fin\_ult1 | Funds |
| ind\_hip\_fin\_ult1 | Mortgage |
| ind\_plan\_fin\_ult1 | Pensions |
| ind\_pres\_fin\_ult1 | Loans |
| ind\_reca\_fin\_ult1 | Taxes |
| ind\_tjcr\_fin\_ult1 | Credit Card |
| ind\_valo\_fin\_ult1 | Securities |
| ind\_viv\_fin\_ult1 | Home Account |
| ind\_nomina\_ult1 | Payroll |
| ind\_nom\_pens\_ult1 | Pensions |
| ind\_recibo\_ult1 | Direct Debit |

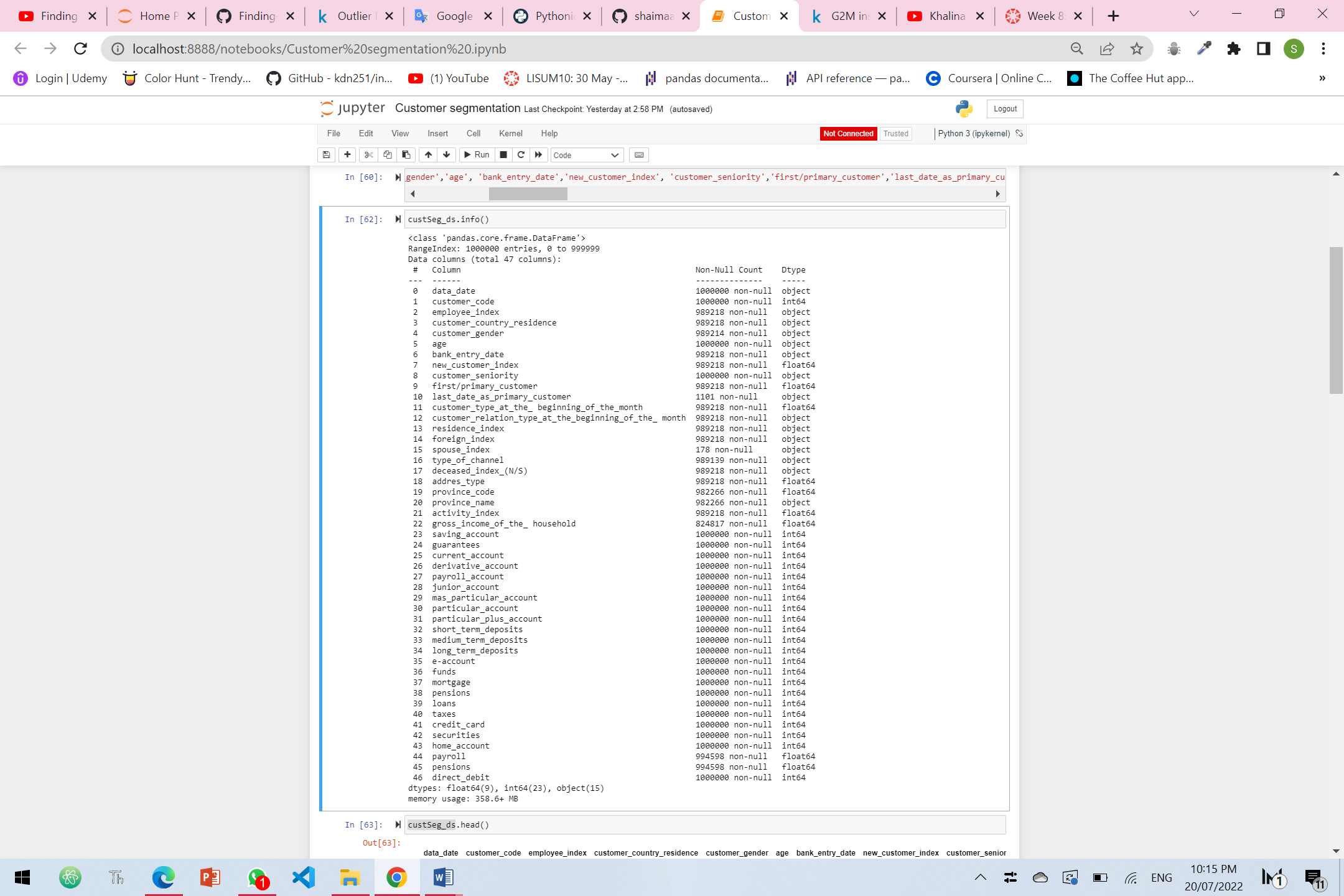
1. **What type of data you have got for analysis?**

The dataset provided was CSV format. The dataset contains 1000000 rows and 48 columns. The datasets mostly contain numerical and categorical data types.

Since the data has no target value, the unsupervised learning (clustering) is the best algorithm to use for this kind of data.

Fewer categorical columns have higher cardinality, i.e, they have more than 10 categories. Most of the categorical columns are binary. Among the numerical features, only the `renta` variable is continuous. The rest are integers. It is important to note that some of the binary categorical columns are of float data type.

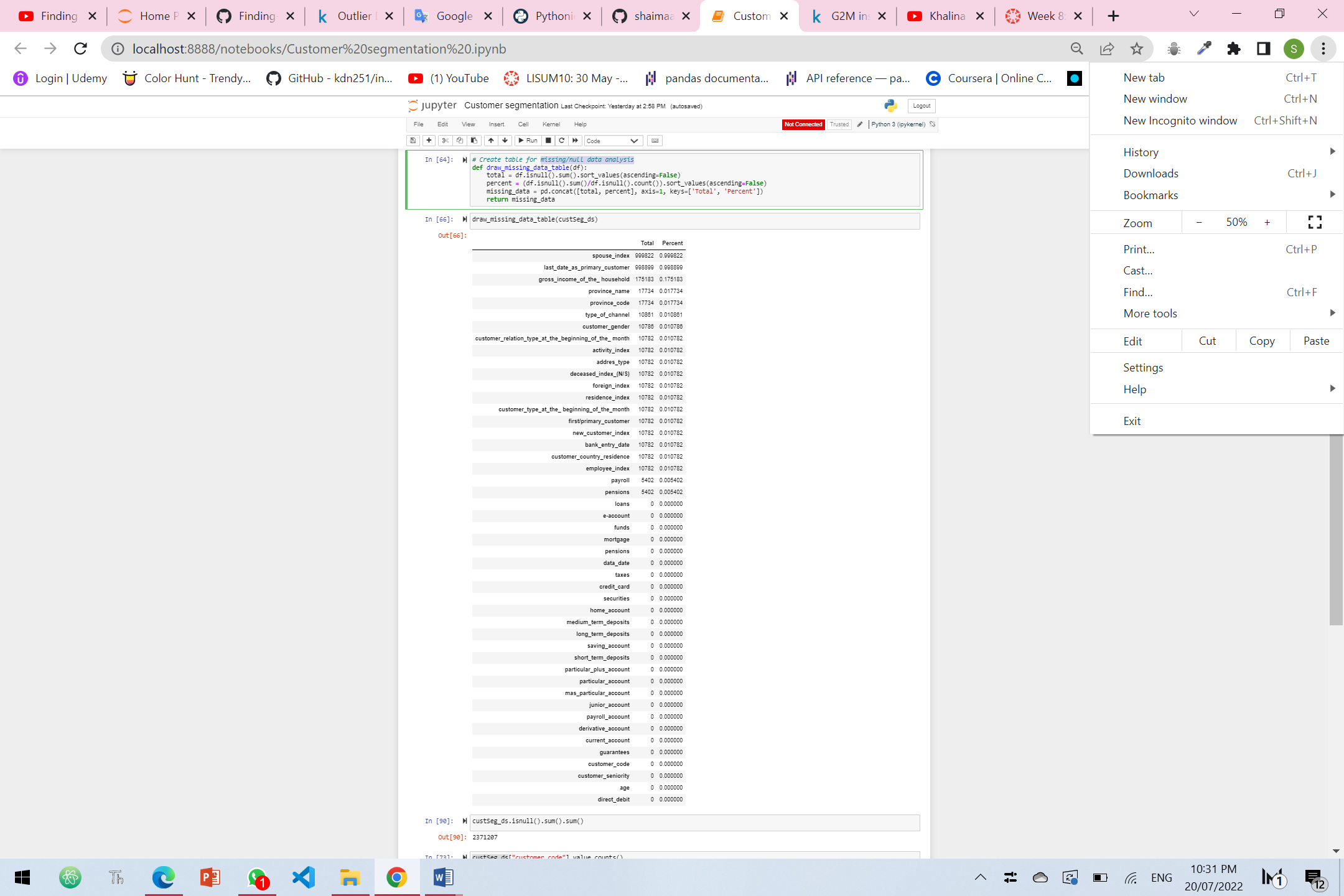
Below, we have attached snapshots of the datasets and its data types.



1. **What are the problems in the data ( number of NA values, outliers , skewed etc)**

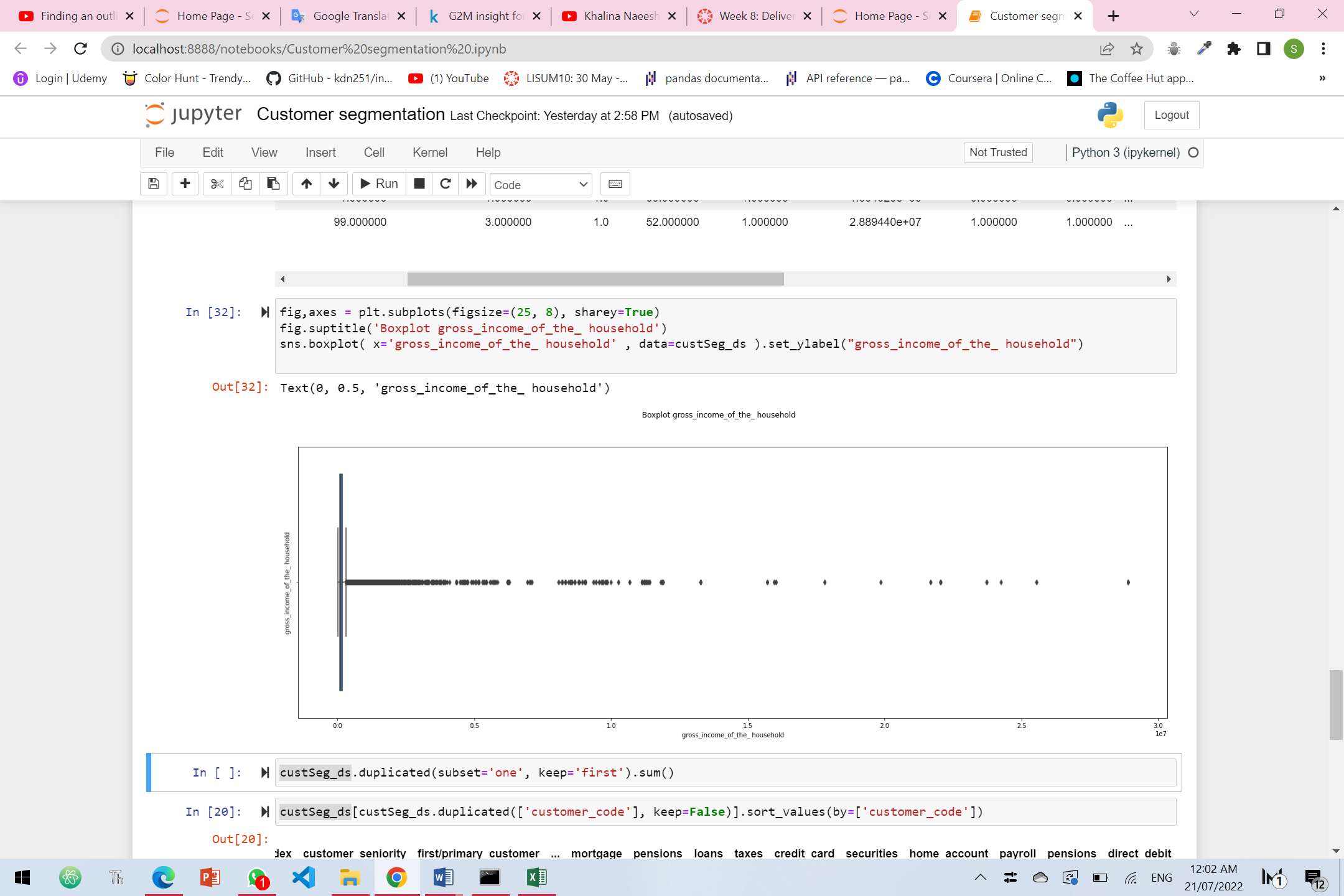
We started initial analysis and we can say that the dataset has some following problems:

1. **Missing/null data:**



From the above analysis we found that the columns have 2371207 missing data, and the columns are listed above.

1. **Outliers**



As per the above there are significant outliers in “gross\_income\_of\_the\_ household”

, Upper and Lower Limits of gross\_income\_of\_the\_ household is:

[-66219.10500000001, 301223.41500000004]