

## **PROJECT DESCRIPTION:**

To analyse and give suitable solutions for the given queries by using Microsoft Excel.

## **APPROACH:**

Understanding the given question and thinking logically and analytically to solve it.

## **TECH-STACK USED:**

The software used in these projects is Microsoft Excel required and perform cleaning using power query editor solutions and to represent the data in graphical and chart format.

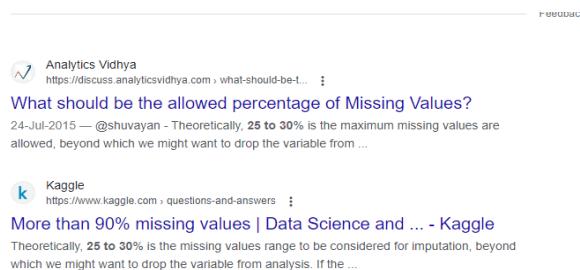
## **INSIGHTS:**

By using Microsoft Excel for performing the given task it is necessary to analyse the given table and provide the analysis solutions briefly.

## **RESULT:**

By solving the given problems, I have learned to perform statistical analysis using charts.

- **Identify** the missing data and use appropriate method to deal with it.  
(Remove columns/or replace it with an appropriate value)  
*Hint: Note that in EDA, since it is not necessary to replace the missing value, but if you have to replace the missing value, what should be the approach. Clearly mention the approach.*



The screenshot shows a search results page with two articles. The first article is from Analytics Vidhya, titled "What should be the allowed percentage of Missing Values?", dated 24-Jul-2015. It discusses the theoretical limit of 25 to 30% missing values. The second article is from Kaggle, titled "More than 90% missing values | Data Science and ... - Kaggle", also dated 24-Jul-2015. It suggests dropping variables with more than 90% missing values. Both articles include links to their respective websites.

Taking only values with null value less than 30% into consideration

## **APPLICATION\_DATA**

→ Counting the blank rows using

=COUNTBLANK(B2:B307512)

File	Home	Insert	Page Layout	Formulas	Data	Review	View	Help			
Clipboard	Font	Font	Font	Font	Font	Font	Font	Font			
Font	Font	Font	Font	Font	Font	Font	Font	Font			
Font	Font	Font	Font	Font	Font	Font	Font	Font			
Calibri	11	A <sup>1</sup>	A <sup>2</sup>	Wrap Text	General	%	0.00	AutSum			
Paste	B	I	U	Merge & Center	Conditional	Format as Table	Insert	Clear			
Clipboard	F1	Font	Font	Font	Format Styles	Cell Styles	Delete	Find & Select			
Font	Font	Font	Font	Font	Format Styles	Cell Styles	Format	Editing			
Font	Font	Font	Font	Font	Cells						
SUM											
=COUNTBLANK(B2:B307512)											
1	SK_ID_CURR	0	FLAG_PHONE	0	FLOORSMIN_AVG	208642	ENTRANCES_MEDI	154288	FLAG_DOCUMENT_10	0	EXT_SOL
2	TARGET	=COUNTBLANK(B2:B307512)	FLAG_EMAIL	0	LANDAREA_AVG	162590	FLOORSMAX_MEDI	153020	FLAG_DOCUMENT_11	0	OWNLCA
3	NAME_CONTRACT_TYPE	=COUNTBLANK(range)	OCCUPATION_TYPE	963931	LIVINGAPARTMENTS_AVG	210199	FLOORSMIN_MEDI	206643	FLAG_DOCUMENT_12	0	DECIPLA
4	CODE_GENDER	CNT_FAM_MEMBERS	2	LANDAREA_AVG	154350	LANDAPEA_MEDI	162590	FLAG_DOCUMENT_13	0	EXT_SOL	
5	FLAG_OWN_CAR	0	REGION_RATING_CLIENT	0	NONLIVINGAPARTMENTS_AVG	213514	LIVINGAPARTMENTS_MEDI	210199	FLAG_DOCUMENT_14	0	APARTIM
6	FLAG_OWN_REALTY	0	REGION_RATING_CLIENT_W_CIT	0	NONLIVINGARE_A_AVG	163682	LIVINGAREA_MEDI	154350	FLAG_DOCUMENT_15	0	BASEMEN
7	FLAG_DOCUMENT_16	0	REGION_RATING_CLIENT_W_CIT	0	NONLIVINGAREAS_AVG	163682	NONLIVINGAREAS_MEDI	213514	FLAG_DOCUMENT_16	0	FLAG_DOCUMENT_16
8	AMT_ANNUITY_TOTAL	0	HOUR_APPR_PROCESS_START	0	BASEMENTAREA_MODE	178943	NONLIVINGAREAS_MEDI	213514	FLAG_DOCUMENT_17	0	YEARS_E
9	AMT_CREDIT	0	REG_REGION_NOT_LIVE_REG	0	YEARS_BEGINEXPLUATATIO	150007	FONDKAPREMONT_MODE	210235	FLAG_DOCUMENT_18	0	COMMON
10	AMT_ANNUITY	12	REG_REGION_NOT_WORK_REG	0	YEARS_BUILD_MODE	204498	HOUSETYPE_MODE	154237	FLAG_DOCUMENT_19	0	ELEVATI
11	AMT_GOODS_PRICE	278	LIVE_REGION_NOT_WORK_REG	0	COMMONAREA_MODE	214865	TOTALAREA_MODE	148431	FLAG_DOCUMENT_20	0	ENTRAN
12	NAME_TYPE_SUITE	1232	REG_CITY_NOT_LIVE_CITY	0	ELEVATORS_MODE	163891	WALLSMATERIAL_MODE	156341	FLAG_DOCUMENT_21	41519	FLOORSI
13	NAME_TYPE_SUITE	0	REG_CITY_NOT_WORK_CITY	0	ENTRANCES_MODE	154828	EMERGENCYSTATE_MODE	147557	AMT_REQ_CREDIT_BUREAU	41519	FLORSSI
14	NAME_TYPE_SUITE	0	LIVE_CITY_NOT_WORK_CITY	0	FLOORSLIVE_MODE	163891	AMT_REQ_CREDIT_BUREAU	41519	AMT_REQ_CREDIT_BUREAU	41519	LANDARE
15	NAME_FAMILY_STATUS	0	ORGANIZATION_TYPE	0	FLOORSMIN_MODE	208642	DEF_60_CNT_SOCIAL_CIRC	1021	AMT_REQ_CREDIT_BUREAU	41519	LIVINGCAF
16	NAME_HOUSING_TYPE	0	EXT_SOURCE_1	173378	LANDAPEA_MODE	162590	DSB_60_CNT_SOCIAL_CIRC	1021	AMT_REQ_CREDIT_BUREAU	41519	LIVINGCAF
17	REGION_POPULATION_RELATION	0	EXT_SOURCE_2	660	LIVINGAPARTMENTS_MODE	210199	DEF_60_CNT_SOCIAL_CIRC	1021	AMT_REQ_CREDIT_BUREAU	41519	NONLIVI
18	DAYS_BIRTH	0	EXT_SOURCE_3	60965	LIVINGAREA_MODE	154350	DAYS_LAST_PHONE_CHANGE	1	AMT_REQ_CREDIT_BUREAU	41519	NONLIVI
19	DAYS_EMPLOYED	0	APARTMENTS_AVG	156091	NONLIVINGAPARTMENTS_MEDI	213514	FLAG_DOCUMENT_2	0	APARTIM		
20	DAYS_EMPLOYED	0	BASEMENTAREA_AVG	162590	NONLIVINGAREAS_MEDI	163682	FLAG_DOCUMENT_3	0	BASEMEN		
21	FLAG_DOCUMENT_4	PUBLISH	YEARS_BEGINEXPLUATATIO	150007	APARTMENTS_MEDI	156091	FLAG_DOCUMENT_4	0	YEARSE		
22	OWN_CAR_AGE	2023293	YEARS_BUILD_AVG	204498	BASEMENTAREAMEDI	173943	FLAG_DOCUMENT_5	0	YEARS_E		
23	FLAG_MOBILE	0	COMMONAREA_AVG	214865	YEARS_BEGINEXPLUATATIO	150007	FLAG_DOCUMENT_6	0	COMMON		
24	FLAG_EMP_PHONE	0	ELEVATORS_AVG	163891	YEARS_BUILD_MEDI	204498	FLAG_DOCUMENT_7	0	ELEVATI		
25	FLAG_WORK_PHONE	0	ENTRANCES_AVG	154828	COMMONAREA_MEDI	214865	FLAG_DOCUMENT_8	0	ENTRANC		
26	FLAG_CONT_MOBILE	0	FLOORSMAX_AVG	153021	ELEVATORS_MEDI	163891	FLAG_DOCUMENT_9	0	FLOORSI		
27	FLAG_DOCUMENT_10	0	FLOORSMIN_AVG	153020	FLAG_DOCUMENT_10	0	FLAG_DOCUMENT_10	0	FLOORSI		
28	FLAG_DOCUMENT_11	0	FLAG_DOCUMENT_11	0	FLAG_DOCUMENT_11	0	FLAG_DOCUMENT_11	0	FLAG_DOCUMENT_11		
29	FLAG_DOCUMENT_12	0	FLAG_DOCUMENT_12	0	FLAG_DOCUMENT_12	0	FLAG_DOCUMENT_12	0	FLAG_DOCUMENT_12		
30	FLAG_DOCUMENT_13	0	FLAG_DOCUMENT_13	0	FLAG_DOCUMENT_13	0	FLAG_DOCUMENT_13	0	FLAG_DOCUMENT_13		
31	FLAG_DOCUMENT_14	0	FLAG_DOCUMENT_14	0	FLAG_DOCUMENT_14	0	FLAG_DOCUMENT_14	0	FLAG_DOCUMENT_14		
32	FLAG_DOCUMENT_15	0	FLAG_DOCUMENT_15	0	FLAG_DOCUMENT_15	0	FLAG_DOCUMENT_15	0	FLAG_DOCUMENT_15		
33	FLAG_DOCUMENT_16	0	FLAG_DOCUMENT_16	0	FLAG_DOCUMENT_16	0	FLAG_DOCUMENT_16	0	FLAG_DOCUMENT_16		
34	FLAG_DOCUMENT_17	0	FLAG_DOCUMENT_17	0	FLAG_DOCUMENT_17	0	FLAG_DOCUMENT_17	0	FLAG_DOCUMENT_17		
35	FLAG_DOCUMENT_18	0	FLAG_DOCUMENT_18	0	FLAG_DOCUMENT_18	0	FLAG_DOCUMENT_18	0	FLAG_DOCUMENT_18		
36	FLAG_DOCUMENT_19	0	FLAG_DOCUMENT_19	0	FLAG_DOCUMENT_19	0	FLAG_DOCUMENT_19	0	FLAG_DOCUMENT_19		

- Calculating the blank rows percentage by dividing the number of blank rows by total number of rows
- And multiplying by 100

The screenshot shows a Microsoft Excel spreadsheet with a large dataset. The columns are labeled from EA to E. The rows contain various numerical values and categorical labels such as 'ENTRANCES\_MEDI', 'FLOORSMAX\_MEDI', etc. A formula  $(DX11/307511)*100$  is present in cell EK11. The Excel ribbon is visible at the top, and the status bar at the bottom indicates 'Ready'.

- Deleting the rows which have a more null percentage
- Taking the column OCCUPATION\_TYPE into consideration because it is having only 31% of null value and may contain some useful information for performing analysis

The screenshot shows a Microsoft Excel spreadsheet with a context menu open over a row of data. The columns are labeled R, S, T, U, V, W, X, Y, and Z. The rows contain various numerical values and categorical labels such as 'OWN\_CAR\_AGE', 'FLAG\_EMP\_PHONE', etc. The context menu includes options like Cut, Copy, Paste Options, Paste Special, Delete, Clear Contents, Format Cells, Column Width, Hide, and Unhide.

- For analysing purposes replacing y by 1 and n by 0 on FLAG\_OWN\_REALITY and FLAG\_OWN\_CAR

Screenshot of Power Query Editor showing the 'Replace Values' dialog box. The dialog is set to replace the value 'y' with '0' in the selected columns. The 'Advanced options' checkbox is checked. The main table view shows gender codes (M/F) and other columns like FLAG OWN CAR, FLAG OWN REALTY, CNT CHILDREN, and AMT INCOME TOTAL.

Screenshot of Power Query Editor showing the 'Replace Values' dialog box. The dialog is set to replace the value 'N' with '0' in the selected columns. The 'Advanced options' checkbox is checked. The main table view shows gender codes (M/F) and other columns like FLAG OWN CAR, FLAG OWN REALTY, CNT CHILDREN, and AMT INCOME TOTAL.

→ Replacing the negative column values by selecting the columns scientific>absolute value

Sheet1 - Power Query Editor

**Home** Transform Add Column View

Queries [3] application\_data application\_data (2) Sheet1

Table Any Column Text Column

Any Column

1.2 REGION\_POPULATION\_RELAT... DAYS\_BIRTH DAYS\_EMPLOYED DAYS

	A <sub>C</sub> _NAME_HOUSING_TYPE	A <sub>B</sub> _REGION_POPULATION_RELAT...	DAYS_BIRTH	DAYS_EMPLOYED	DAYS
1	House / apartment	0.018801	9461	-637	
2	House / apartment	0.003541	16765	-1188	
3	House / apartment	0.010032	19046	-225	
4	House / apartment	0.008019	19005	-309	
5	House / apartment	0.028663	19832	-308	
6	House / apartment	0.035792	16941	-1588	
7	House / apartment	0.035792	13778	-3130	
8	House / apartment	0.003122	18850	-449	
9	House / apartment	0.018634	20099	365243	-7427
10	House / apartment	0.019689	14469	-2019	-14437
11	House / apartment	0.0228	10197	-679	-4247
12	House / apartment	0.015221	20417	365243	-5246
13	House / apartment	0.031329	13439	-2717	-311
14	House / apartment	0.016612	14086	-3028	-643
15	House / apartment	0.010006	14583	-203	-615
16	Rented apartment	0.020713	8728	-1157	-3494
17	House / apartment	0.018634	12931	-1317	-6392
18	House / apartment	0.010966	9776	-191	-4143
19	House / apartment	0.04622	17718	-7804	-8751
20	House / apartment	0.015221	11348	-2038	-1021
21	House / apartment	0.015221	18252	-4286	-298
22	House / apartment	0.025164	14815	-1652	-2299
23	Rented apartment	0.020713	11146	-4906	-114

95 COLUMNS, 999+ ROWS Column profiling based on top 1000 rows PREVIEW DOWNLOADED AT 14:38

→ Count the value with high frequency for replacing the null rows presented in OCCUPATION\_TYPE

Sheet1 (2) - Power Query Editor

**Home** Transform Add Column View

Queries [1] Sheet1 (2)

Table

1.2 OCCUPATION\_TYPE Count

	A <sub>C</sub> _NAME_INCOME_TYPE	A <sub>B</sub> _OCCUPATION_TYPE	Count
1	Pensioner	null	55357
2	State servant	null	3787
3	Commercial associate	null	12297
4	Working	null	24920
5	Unemployed	null	22
6	Student	null	5
7	Maternity leave	null	1
8	Businessman	null	2

3 COLUMNS, 8 ROWS Column profiling based on top 1000 rows PREVIEW DOWNLOADED AT 16:54

→ Replacing the null rows using replace value option with a more frequency value pensioner

Screenshot of Power Query Editor showing the 'Replace Values' dialog box. The dialog shows 'Value To Find' as 'null' and 'Replace With' as 'Pensioner'. The main table view shows rows 1 through 23, with rows 15 through 23 highlighted in green. The 'APPLIED STEPS' pane shows the step 'Replaced Value'.

**Replace Values**

Replace one value with another in the selected columns.

Value To Find  
null

Replace With  
Pensioner

OK Cancel

	A <sub>C</sub> NAME_INCOME_TYPE	A <sub>B</sub> ORGANIZATION_TYPE	Count
1	Working		
2	State servant		
3	Working		
4	Working		
5	Working		
6	State servant		
7	Commercial associate		
8	State servant		
9	Pensioner		
10	Working		
11	Working		
12	Pensioner		
13	Working		
14	Working		
15	Working	Labors	
16	Working	Labors	
17	Working	Drivers	
18	Working	Labors	
19	Working	Labors	
20	State servant	Core staff	
21	Working	Labors	
22	Commercial associate	Sales staff	
23	Working	Sales staff	

- Count the value with high frequency for replacing the null rows presented in ORGANIZATION\_TYPE
- XNA are null values

Screenshot of Power Query Editor showing the 'Select Rows' dialog box. The dialog shows the condition '(ORGANIZATION\_TYPE) = "XNA"'. The main table view shows 2 rows: 'Pensioner' and 'Unemployed'. The 'APPLIED STEPS' pane shows the step 'Filtered Rows'.

**Select Rows**

Select rows based on the following conditions:

(ORGANIZATION\_TYPE) = "XNA"

	A <sub>C</sub> NAME_INCOME_TYPE	A <sub>B</sub> ORGANIZATION_TYPE	Count
1	Pensioner	XNA	55352
2	Unemployed	XNA	22

- Replacing the null rows using replace value option with a more frequency value pensioner

Screenshot of the Power Query Editor showing the 'Replace Values' dialog. The dialog is set to replace the value 'XNA' with 'Pensioner' in the selected column. The 'Advanced options' checkbox is checked. The background shows a table with columns: REG\_CITY\_NOT\_WORK\_CITY, LIVE\_CITY\_NOT\_WORK\_CITY, ORGANIZATION\_TYPE, and Column1.

→ Finding the age by the column DAYE\_BIRTH using =INT(DAYA\_BIRTH/365)

Screenshot of Microsoft Excel showing a formula being entered into cell R2: =INT(R2/365). The formula calculates the age from the 'DAYE\_BIRTH' column. The table has columns: NAME\_FAMILY\_STATUS, NAME\_HOUSING\_TYPE, REGION\_POPULATION\_RELATIVE, DAYS\_BIRTH, age, DAYS\_EMPLOYED, DAYS\_REGISTRATION, DAYS\_ID\_PUBLISH, FLAG\_MOBIL, and others.

→ Separating the values in a range to give a unique name for different ranges in the column AGE , AMT\_INCOME\_TOTAL and AMT\_CREDIT

Sheet1 (3) - Power Query Editor

**Add Conditional Column**

New column name: ATM\_INCOME\_TOTAL\_TABLE

Column Name	Operator	Value	Output
If	AMT_INCOME_TO...	is less than or equal to 50000	Then ABC 123 VERY LOW
Else If	AMT_INCOME_TO...	is less than or equal to 160000	Then ABC 123 LOW
Else If	AMT_INCOME_TO...	is less than or equal to 1300000	Then ABC 123 MEDIUM
Else If	AMT_INCOME_TO...	is less than or equal to 2000000	Then ABC 123 HIGH
Else If	AMT_INCOME_TO...	is less than or equal to 117000000	Then ABC 123 VERY HIGH
<b>Add Clause</b>			
Else	ABC 123		

**Query Settings**

**Properties**

**LIED STEPS**

Source: Sheet1 (3)

Navigation, Promoted Headers, Changed Type, Removed Columns, Added Conditional Column, Added Conditional Column1

PREVIEW DOWNLOADED AT 21:25

Sheet1 (2) - Power Query Editor

**Add Conditional Column**

New column name: ATM\_CREDIT\_TABLES

Column Name	Operator	Value	Output
If	AMT_CREDIT	is less than or equal to 100000	Then ABC 123 VERY LOW
Else If	AMT_CREDIT	is less than or equal to 500000	Then ABC 123 LOW
Else If	AMT_CREDIT	is less than or equal to 1500000	Then ABC 123 MEDIUM
Else If	AMT_CREDIT	is less than or equal to 2000000	Then ABC 123 HIGH
Else If	AMT_CREDIT	is greater than or equal to 2000000	Then ABC 123 VERY HIGH
<b>Add Clause</b>			
Else	ABC 123	null	

**Query Settings**

**Properties**

**LIED STEPS**

Source: Sheet1 (2)

Navigation, Promoted Headers, Changed Type, Removed Columns, Added Conditional Column, Replaced Value, Filtered Rows, Replaced Value1, Replaced Value2, Replaced Value3, Replaced Value4, Added Conditional Column1

PREVIEW DOWNLOADED AT 16:54

Screenshot of Microsoft Power Query Editor showing the "Add Conditional Column" dialog. The dialog is titled "Add Conditional Column" and contains a table for defining rules. The table has four columns: "Column Name", "Operator", "Value", and "Output". There are four rows defined:

- If age is less than or equal to 20, Then young adult
- Else If age is greater than or equal to 60, Then senior citizen
- Else If age is less than or equal to 25, Then adult
- Else If age is greater than 25, Then middle age

The "Else" row is currently empty. The "OK" button is visible at the bottom right of the dialog.

→ Finding the column having the highest percentage of null values

## PREVIOUS\_APPLICATION

Screenshot of Microsoft Power Query Editor showing the "Transform" tab selected. A tooltip is displayed over the "Transform" ribbon icon, showing the status "Transforming data". The main area shows a table with several columns, including "DAYS\_FIRST\_DUE", "DAYS\_LAST\_DUE\_1ST\_VERSION", "DAYS\_LAST\_DUE", "DAYS\_TERMINATION", and "NFLAG\_INSURED\_ON\_APPR...". A context menu is open over the "DAYS\_LAST\_DUE" column, with the option "Remove Empty" highlighted. The "Query Settings" pane on the right shows the query name "previous\_application" and the applied step "Changed Type".

Queries [1] previous\_application

File Home Transform Add Column View

Transform ribbon: Close & Load, Refresh, Advanced Editor, Choose Columns, Remove Columns, Keep Rows, Remove Rows, Sort, Split Column, Group By, Replace Values, Data Type: Decimal Number, Merge Queries, Append Queries, Combine Files, Manage Parameters, Data source settings, New Source, Recent Sources, Enter Data, New Query.

Properties pane: Name (previous\_application), All Properties.

Applied Steps pane: Source, Promoted Headers, Changed Type, Removed Columns.

Query Settings pane: Name (previous\_application).

Table preview: RATE\_DOWN\_PAYM... RATE\_INTEREST\_PRIMARY RATE\_INTEREST\_PRIVILEGED NAME\_CASH\_LOAN\_PUR... NAME\_CONTRACT\_S...

Column profiling: 36 COLUMNS, 999+ ROWS. Column profiling based on top 1000 rows.

PREVIEW DOWNLOADED AT 11:54

→ And remove those columns

Queries [1] previous\_application

File Home Transform Add Column View

Transform ribbon: Close & Load, Refresh, Advanced Editor, Choose Columns, Remove Columns, Keep Rows, Remove Rows, Sort, Split Column, Group By, Replace Values, Data Type: Whole Number, Merge Queries, Append Queries, Combine Files, Manage Parameters, Data source settings, New Source, Recent Sources, Enter Data, New Query.

Properties pane: Name (previous\_application), All Properties.

Applied Steps pane: Source, Promoted Headers, Changed Type.

Query Settings pane: Name (previous\_application).

Table preview: DAYS\_FIRST\_DUE DAYS\_LAST\_DUE\_1ST\_VERSI... INFLAG\_INSURED\_ON\_APPR...

Column profiling: 37 COLUMNS, 999+ ROWS. Column profiling based on top 1000 rows.

PREVIEW DOWNLOADED AT 10:54

- Finding the mean value to replace the rows will null value only for the columns with a low null percentage or for the column which is important for analysis
- Finding the mean and replacing them with the null rows in the columns AMT\_ANNUITY, AMT\_GOODS\_PRICE, CNT\_PAYMENT

previous\_application - Power Query Editor

**Number Tools**

File Home Transform Add Column View Transform

Queries [1] previous\_application

= List.Average(#"Removed Columns"[AMT\_ANNUITY])  
15955.120659423315

**Query Settings**

- Properties Name previous\_application All Properties
- Applied Steps
  - Source
  - Promoted Headers
  - Changed Type
  - Removed Columns
  - Calculated Average

previous\_application - Power Query Editor

File Home Transform Add Column View

Close & Load Refresh Preview Advanced Editor

Choose Columns Remove Rows Keep Rows Remove Rows Sort Data Type: Decimal Number

Merge Queries Append Queries Combine Files Manage Parameters Data source settings Enter Data New Source Recent Sources

Queries [1] previous\_application

= Table.ReplaceValue(#"Removed Columns", null, 15955.12, Replacer.ReplaceValue, {"AMT\_ANNUITY"})

SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CRED
2030495	271877	Consumer loans	1730.43	17145	
2273188	270658	Consumer loans	9644.22	26550	
1232483	151612	Consumer loans	21307.455	126490.5	
2163253	154602	Consumer loans	4187.34	26955	
1285768	142748	Revolving loans	9000	180000	
2393109	396305	Cash loans	10181.7	180000	
1173070	199178	Cash loans	4666.5	45000	
1506815	166490	Cash loans	25454.025	450000	
1182516	267782	Cash loans	20361.6	405000	
1177447	903211	Cash loans	15444.12	0	

Replace Values

Replace one value with another in the selected columns.

Value To Find: null  
Replace With: 15955.12

OK Cancel

**Query Settings**

- Properties Name previous\_application All Properties
- Applied Steps
  - Source
  - Promoted Headers
  - Changed Type
  - Removed Columns
  - Replaced Value

Queries [1] previous\_application

File Home Transform Add Column View

Data Type: Decimal Number

Replace Values

Unpivot Columns

Move

Merge Columns

Statistics Standard Scientific

Trigonometry

Rounding

Format

Parse

Text Column

Number Column

Group By

Use First Row as Headers

Count Rows

Rename

Pivot Column

Convert to List

Split Column

Any Column

1 L2 AMT\_GOODS\_PRICE 17145 SATURDAY 15 Y

2 679671 607500 THURSDAY 11 Y

3 136444.5 112500 TUESDAY 11 Y

4 470790 450000 MONDAY 7 Y

5 404055 337500 THURSDAY 9 Y

6 340573.5

7 0

8 0

9 0

10 0

11 335754

12 246397.5

13 174361.5

14 57564

15 27252

16 119853

17 27297

18 180000 180000 FRIDAY 13 Y

19 180000 180000 THURSDAY 14 Y

20 49455 45000 SATURDAY 16 Y

21 491580 450000 MONDAY 6 Y

22 451777.5 405000 SATURDAY 4 Y

23 277487.2 THURSDAY 9 Y

36 COLUMNS, 999+ ROWS Column profiling based on top 1000 rows

PREVIEW DOWNLOADED AT 11:23

→ Replacing the negative column values by selecting the columns scientific>absolute value

Queries [1] previous\_application

File Home Transform Add Column View

Data Type: Whole Number

Replace Values

Unpivot Columns

Move

Merge Columns

Statistics Standard

Scientific

Absolute Value

Power

Square

Return the absolute value of numbers in the selected columns.

Exponent

Logarithm

Factorial

Text Column

Number Column

Group By

Use First Row as Headers

Count Rows

Rename

Pivot Column

Convert to List

Split Column

Any Column

1 L2 NFLAG\_LAST\_APPL\_IN\_D... 1 XAP Approved -73 Cash

2 1 XNA Approved -164 XNA

3 1 XNA Approved -301 Cash through the bank

4 1 XNA Approved -512 Cash through the bank

5 1 Repairs Refused -781 Cash through the bank

6 1 Everyday expenses Approved -684 Cash through the bank

7 1 XNA Canceled -14 XNA

8 1 XNA Canceled -21 XNA

9 1 XNA Canceled -386 XNA

10 1 XNA Canceled -57 XNA

11 1 XNA Approved -735 Cash through the bank

12 1 XNA Approved -815 Cash through the bank

13 1 XNA Approved -860 Cash through the bank

14 1 XAP Approved -408 Cash through the bank

15 1 XAP Approved -726 Cash through the bank

16 1 XAP Approved -699 Cash through the bank

17 1 XAP Approved -1473 Cash through the bank

18 1 XAP Approved -336 XNA

19 1 XNA Approved -700 Cash through the bank

20 1 Everyday expenses Refused -584 XNA

21 1 XNA Refused -401 Cash through the bank

22 1 XNA Approved -686 Cash through the bank

23 1 XNA Refused -799 XNA

26 COLUMNS, 999+ ROWS Column profiling based on top 1000 rows

PREVIEW DOWNLOADED AT 12:21

→ Changing the value refreshed and xna to repeater for considering NAME\_CLIENT\_TYPE as the target value to perform analysis as the target value not specified

previous\_application - Power Query Editor

**File** **Home** **Transform** **Add Column** **View**

**Table** **Transpose** **Reverse Rows** **Detect Data Type** **Pivot Column** **Convert to List** **Unpivot Columns** **Move** **Merge Columns** **Split Columns** **Format** **Text Column** **Number Column** **Statistics** **Standard** **Scientific** **Rounding** **Trigonometry** **Date** **Time** **Duration** **Extract Values** **Structured Column**

**Queries [1]** **fx** = Table.Group(#"Calculated Absolute Value", {"NAME\_CLIENT\_TYPE"}, {{"Count", each Table.RowCount(\_, Int64.Type)}})

**previous\_application**

A <sub>C</sub> NAME_CLIENT_TYPE	i <sub>2</sub> Count
1 Repeater	1231261
2 New	301363
3 Refreshed	135649
4 XNA	1941

**Query Settings**

**PROPERTIES** **Name**: previous\_application **All Properties**

**APPLIED STEPS**

- Source
- Promoted Headers
- Changed Type
- Removed Columns
- Replaced Value
- Replaced Value1
- Filtered Rows
- Calculated Absolute Value
- Grouped Rows

2 COLUMNS, 4 ROWS Column profiling based on top 1000 rows PREVIEW DOWNLOADED AT 12:26

previous\_application - Power Query Editor

**File** **Home** **Transform** **Add Column** **View**

**Table** **Transpose** **Reverse Rows** **Detect Data Type** **Pivot Column** **Convert to List** **Unpivot Columns** **Move** **Merge Columns** **Split Columns** **Format** **Text Column** **Number Column** **Statistics** **Standard** **Scientific** **Rounding** **Trigonometry** **Date** **Time** **Duration** **Extract Values** **Structured Column**

**Queries [1]** **fx** = Table.SelectRows(#"Calculated Absolute Value", each true)

**previous\_application**

A <sub>C</sub> NAME_CONTRACT_STA...	i <sub>2</sub> DAYS_DECISION	A <sub>C</sub> NAME_PAYMENT_TY...	A <sub>C</sub> CODE_REJECT_REASON	A <sub>C</sub> NAME_CLIENT_TYPE
1 Approved				
2 Approved				
3 Approved				
4 Approved				
5 Refused				
6 Approved				
7 Approved				
8 Canceled				
9 Canceled				
10 Canceled				
11 Approved				
12 Approved				
13 Approved				
14 Approved				
15 Approved				
16 Approved				
17 Approved				
18 Approved				
19 Approved				
20 Approved				
21 Refused				
22 Approved				
23 Approved				

**Replace Values**

Replace one value with another in the selected columns.

**Value To Find**: XNA **Replace With**: Repeated

**OK** **Cancel**

26 COLUMNS, 999+ ROWS Column profiling based on top 1000 rows PREVIEW DOWNLOADED AT 12:27

previous\_application (1) - Power Query Editor

**Transform**

File Home Transform Add Column View

Table Any Column Text Column Number Column Date & Time Column Structured Column

Queries [1] previous\_application...

Replace Values

Replace one value with another in the selected columns.

Value To Find: Refreshed

Replace With: Repeated

Advanced options

OK Cancel

Replaced Value

Properties Name: previous\_application (1)

Applied Steps Source: Navigation Promoted Headers Changed Type Replaced Value

## → Replacing the null values of NAME\_PAYME

previous\_application - Power Query Editor

File Home Transform Add Column View

Table Any Column Text Column Number Column Date & Time Column Structured Column

Queries [1] previous\_application

Count

NAME_PAYMENT_TYPE	Count
Cash through the bank	1033552
XNA	627384
Non-cash from your account	8193
Cashless from the account of...	1085

Properties Name: previous\_application

Applied Steps Source: Promoted Headers Changed Type Removed Columns Replaced Value Replaced Value1 Filtered Rows Calculated Absolute Value Filtered Rows1 Replaced Value2 Replaced Value3 Grouped Rows

2 COLUMNS, 4 ROWS Column profiling based on top 1000 rows

PREVIEW DOWNLOADED AT 12:34

## → NAME\_PAYMENT\_TYPE with the value with highest frequency

Replace Values

Replace one value with another in the selected columns.

Value To Find  
XNA

Replace With  
Cash through the bank

OK Cancel

	1 XAP	Approved	726 Cash through the bank
15	1 XAP	Approved	699 Cash through the bank
16	1 XAP	Approved	1473 Cash through the bank
17	1 XAP	Approved	336 Cash through the bank
18	1 XAP	Approved	700 Cash through the bank
19	1 XNA	Approved	584 Cash through the bank
20	1 Everyday expenses	Refused	401 Cash through the bank
21	1 XNA	Refused	686 Cash through the bank
22	1 XNA	Approved	720 Cash through the bank
23	f YNA	Refused	

PREVIEW DOWNLOADED AT 12:36

→ For analysing purposes replacing y by 1 and n by 0 on FLAG\_LAST\_APPL\_PER\_CONTACT

Replace Values

Replace one value with another in the selected columns.

Value To Find  
Y

Replace With  
1

OK Cancel

	10 Y	1 XAP
15	7 Y	1 XAP
16	12 Y	1 XAP
17	13 Y	1 XAP
18	14 Y	1 XNA
19	16 Y	1 Everyday expenses
20	6 Y	1 XNA
21	4 Y	1 XNA
22	9 Y	1 XNA

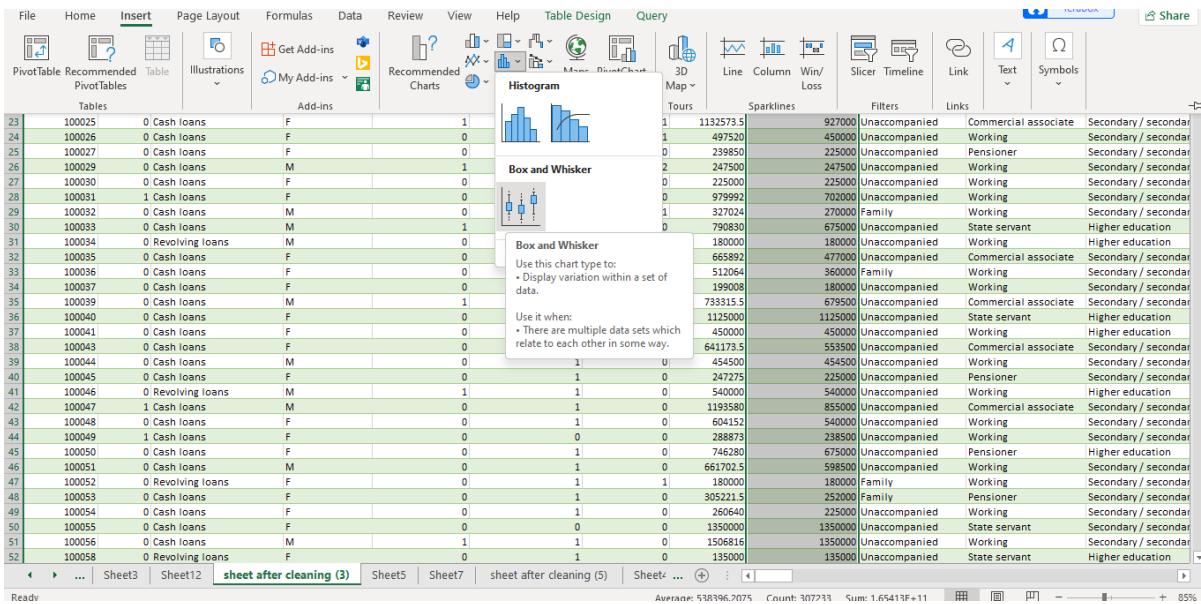
PREVIEW DOWNLOADED AT 12:36

Screenshot of Power Query Editor showing the 'Replace Values' dialog. The dialog is centered over a table with columns: FLAG\_LAST\_APPL\_PER\_CONTRACT, NFLAG\_LAST\_APPL\_IN\_D..., NAME\_CASH\_LOAN\_PUR..., NAME\_CONTRACT\_STA..., and DAYS\_DECISION. The 'Replace Values' dialog shows 'Value To Find' as 'N' and 'Replace With' as '0'. The 'Advanced options' section is collapsed. The main table preview shows rows 1 through 23. The 'Properties' pane on the right shows the query name as 'previous\_application'. The 'Applied Steps' pane lists steps like 'Replaced Value5'.

→ For finding the difference in the credit values before and at the time of applying subtracting AMT\_CREDIT and AMT\_APPLICATION

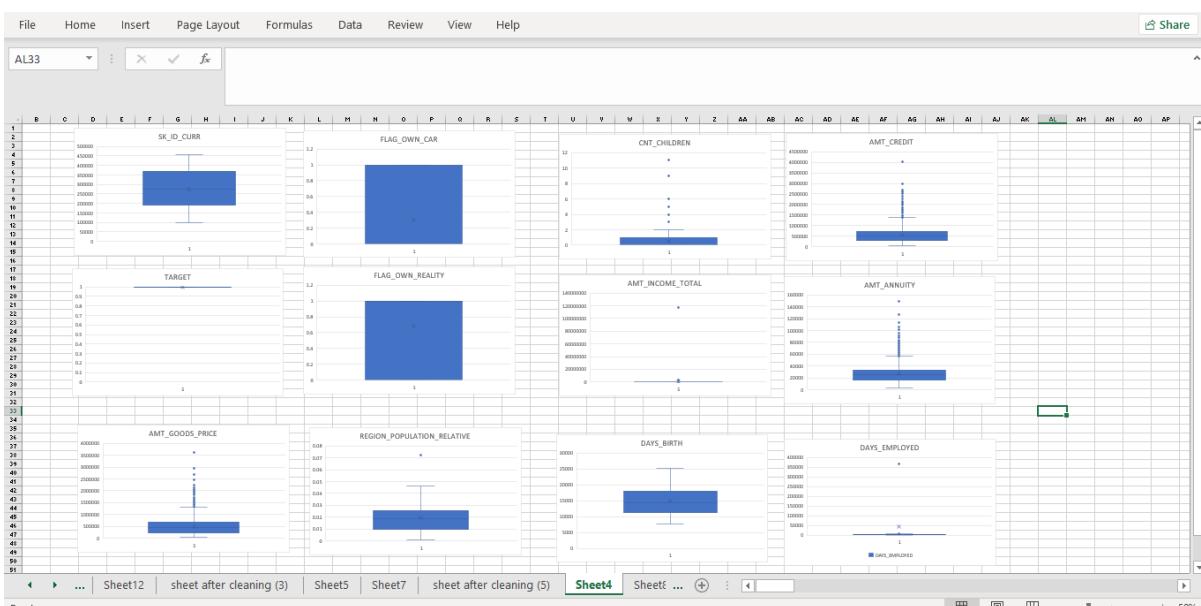
Screenshot of Power Query Editor showing the 'Subtract' dialog. The dialog is centered over a table with columns: NAME\_SELLER\_INDUST..., CNT\_PAYMENT, NAME\_YIELD\_GROUP, PRODUCT\_COMBINATION, and a calculated column 'Subtraction'. The 'Subtract' dialog shows 'Value' as 'AMT\_APPLICATION'. The main table preview shows rows 1 through 23. The 'Properties' pane on the right shows the query name as 'previous\_application'. The 'Applied Steps' pane lists steps like 'Inserted Subtraction'.

- Identify if there are outliers in the dataset. Also, mention why do you think it is an outlier. Again, remember that for this exercise, it is not necessary to remove any data points.



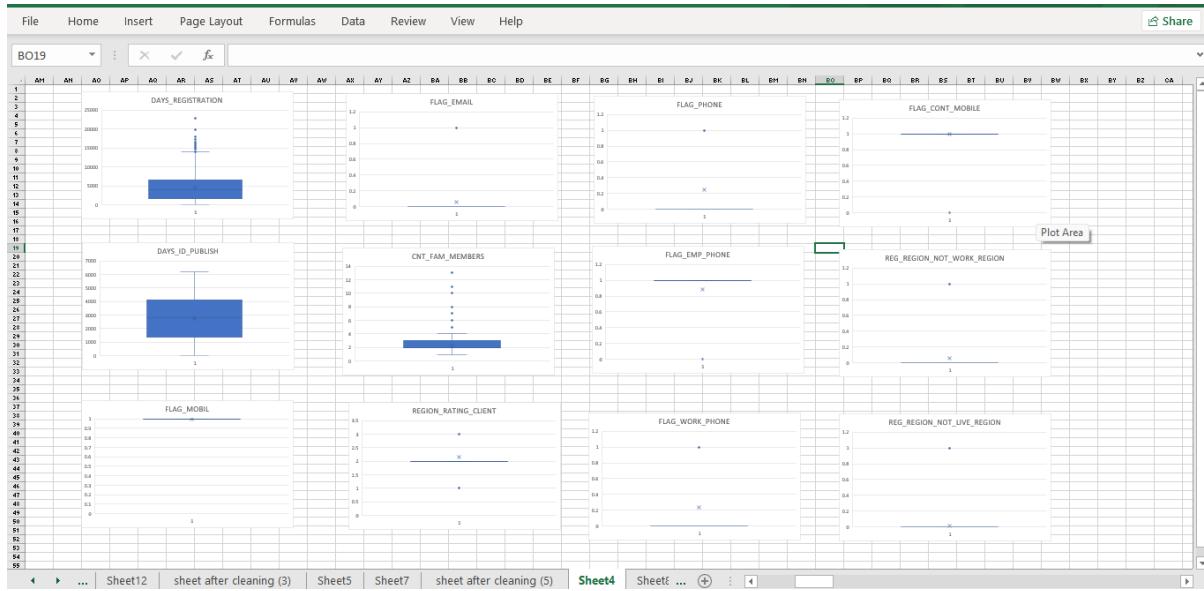
→ To find the outliers use the box and Whisker chart option in excel

## APPLICATION\_DATA

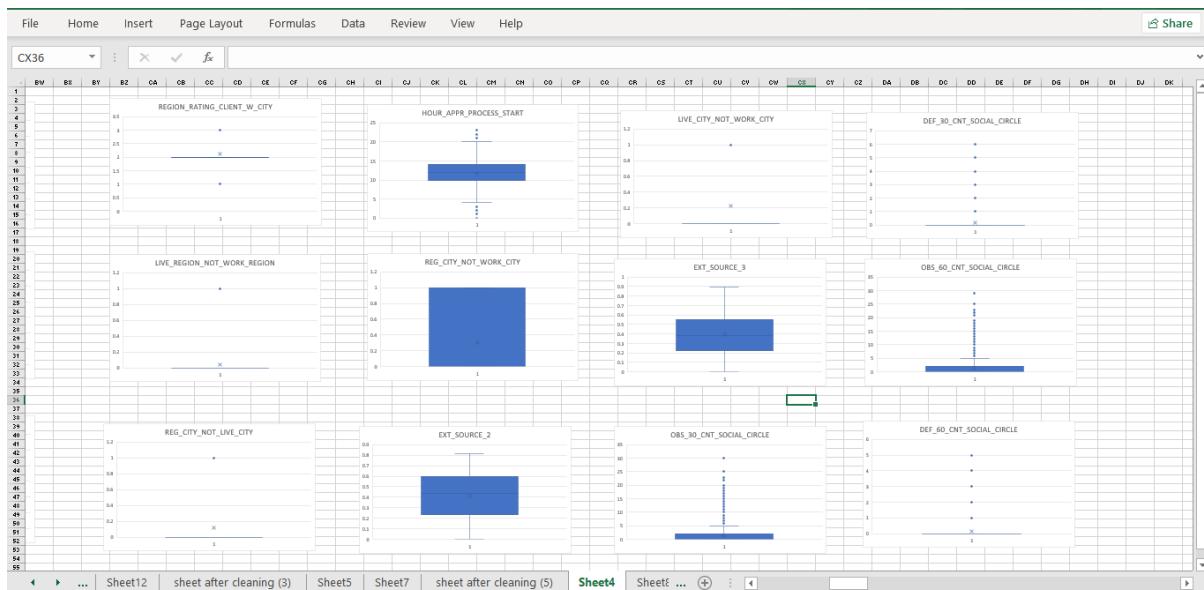


- CNT\_CHILDREN, AMT\_CREDIT, AMT\_ANNUITY, AMT\_GOODS\_PRICE have more number of outliers with thin box
- AMT\_INCOME\_TOTAL, DAYS\_EMPLOYED have very thin boxes and only one or two outliers
- SK\_ID\_CURR, DAYS\_BIRTH, REGION\_POPULATION\_RELATIVE have 1 or no outliers

→ FLAG\_OWN\_CAR,FLAG\_OWN\_REALITY have boxes with high length and no outliners

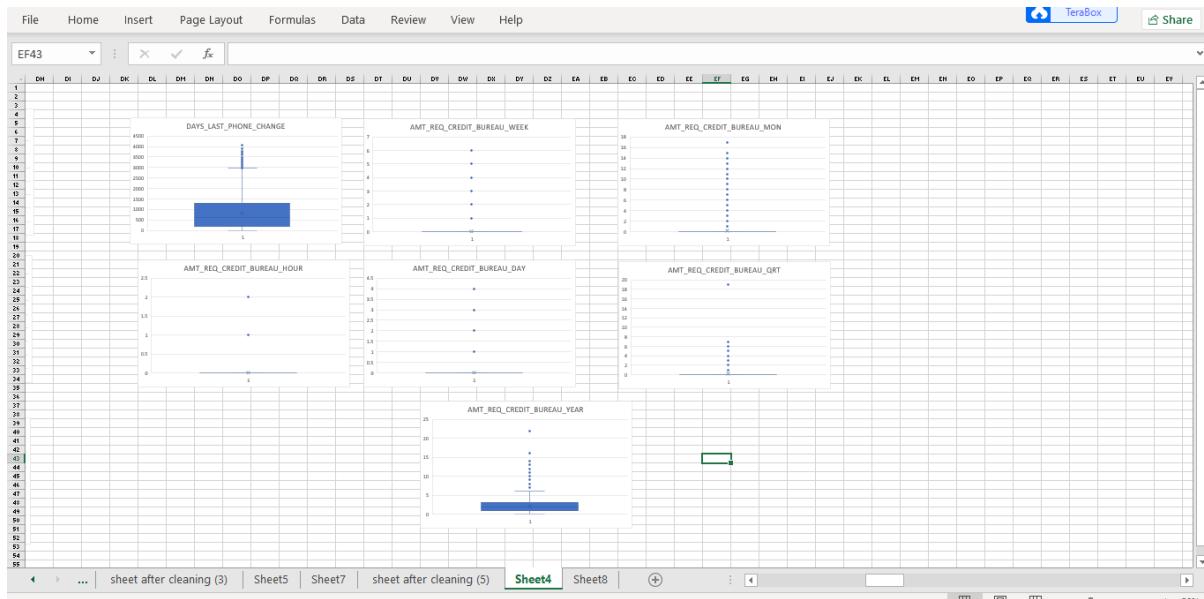


- DAYS\_REGISTRATION,CNT\_FAM\_MEMBERS have more outliners
- FLAG\_EMAIL,FLAG\_PHONE,FLAG\_CONTACT\_MOBILE,FLAG\_EMP\_PHONE,REG\_REGION\_NOT\_WORK\_REGION,FLAG\_MOBIL,REGION\_RATING\_CLIENT\_FLAG\_WORK\_PHONE,REG\_REGION\_NOT\_LIVE\_REGION have very thin box with one or no outliners
- DAYS\_ID\_PUBLISH have no outliners



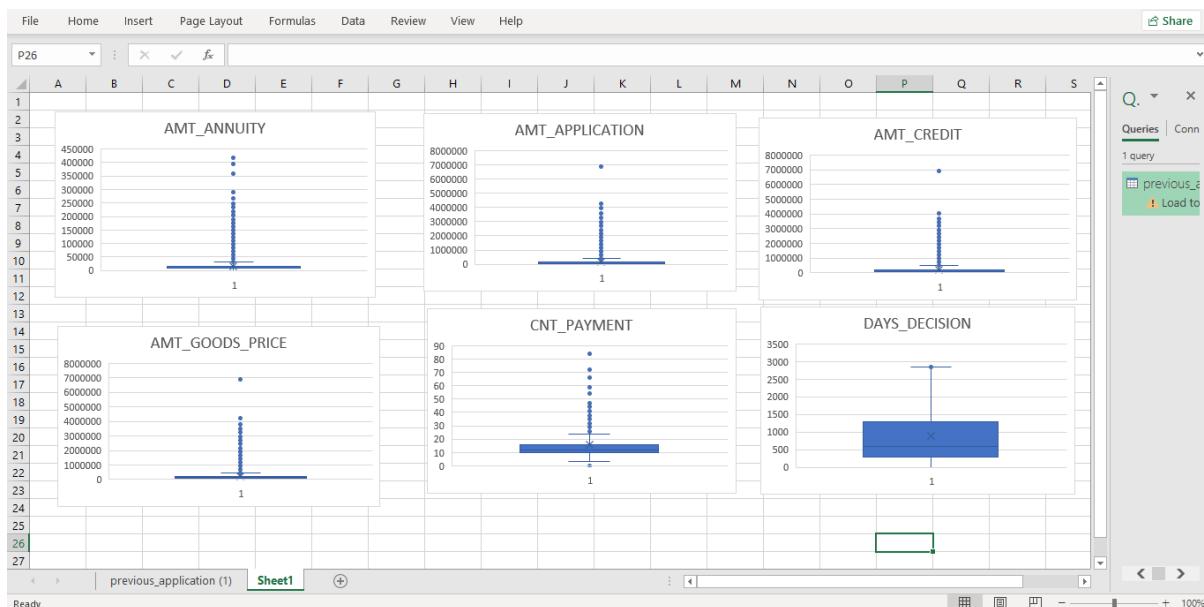
- LIVE\_CITY\_NOT\_WORK\_CITY,REGION\_RATING\_CLIENT\_W\_CITY,LIVE\_REGION\_NOT\_WORK\_REGION,REG\_CITY\_NOT\_LIVE\_CITY have a very thin box and one or two outliners

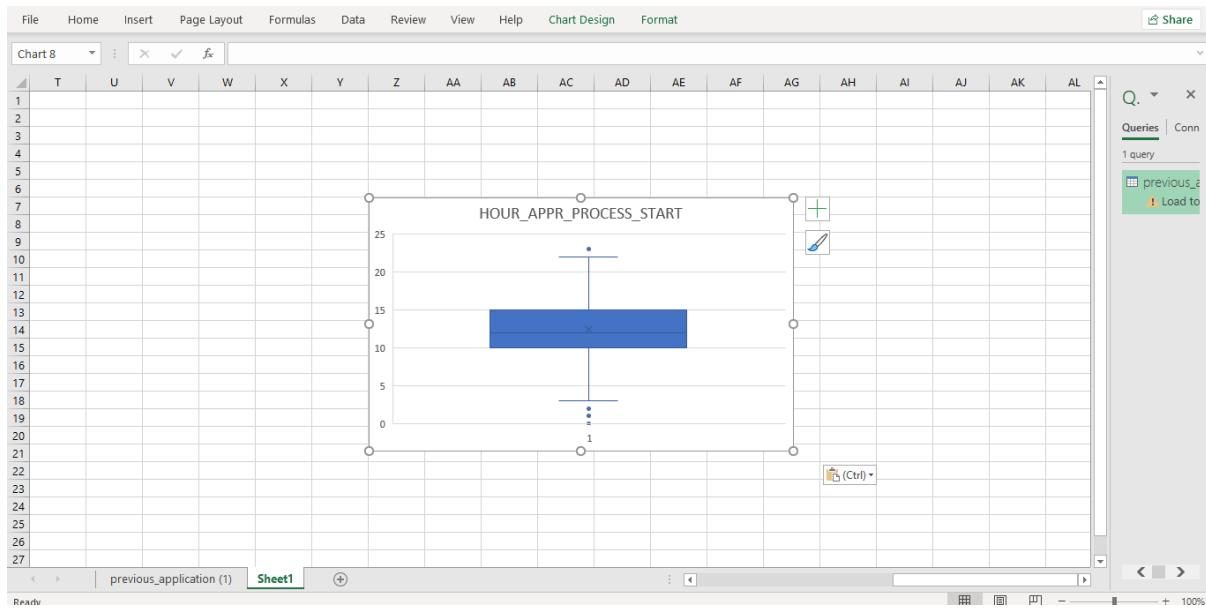
- OBS\_30\_CNT\_SOCIAL\_CIRCLE, DFF\_60\_SOCIAL\_CIRCLE, OBS\_60\_CNT\_SOCIAL\_CIRCLE, DFF\_30\_SOCIAL\_CIRCLE, HOUR\_APPR\_PROCESS\_START have thin box and more outliers
- EXT\_SOURCE\_2, EXT\_SOURCE\_3, REG\_CITY\_NOT\_WORK\_CITY have no outliers



- DAY\_LAST\_PHONE\_CHANGE, AMT\_REQ\_CREDIT\_BUREAU\_YEAR have more outliers
- AMT\_REQ\_CREDIT\_BUREAU\_WEEK, AMT\_REQ\_CREDIT\_BUREAU\_MON, AMT\_REQ\_CREDIT\_BUREAU\_DAY, AMT\_REQ\_CREDIT\_BUREAU\_QRT have very thin box and more outliers
- AMT\_REQ\_CREDIT\_BUREAU\_HOUR has only 2 outliers with thin box

## PREVIOUS\_APPLICATION





- AMT\_ANNUITY ,AMT\_APPLICATION , AMT\_GOODS\_PRICE ,AMT\_CREDIT,CNT\_PAYMENT have very thin box and also high outliers
- DAYS\_DECISION have only one outlier
- HOUR\_APPR\_PROCESSOR\_START have 2-4 outliers
- Identify if there is data imbalance in the data. Find the ratio of data imbalance.  
*Hint: Since there are a lot of columns, you can run your analysis in loops for the appropriate columns and find the insights.*

## APPLICATION\_DATA

	TARGET	Count
1	0	282686
2	1	24825

**APPLIED STEPS**

- Source
- Navigation
- Promoted Headers
- Changed Type
- Replaced Value
- Grouped Rows
- Sorted Rows**

- Finding the Ratio Between target 1 and 0

=target 0/target 1 = $282686/24825 = 11.3:0$

- This analysis is done using pivotchart

Screenshot of Microsoft Excel showing a PivotTable and PivotChart interface.

**PivotTable Fields:**

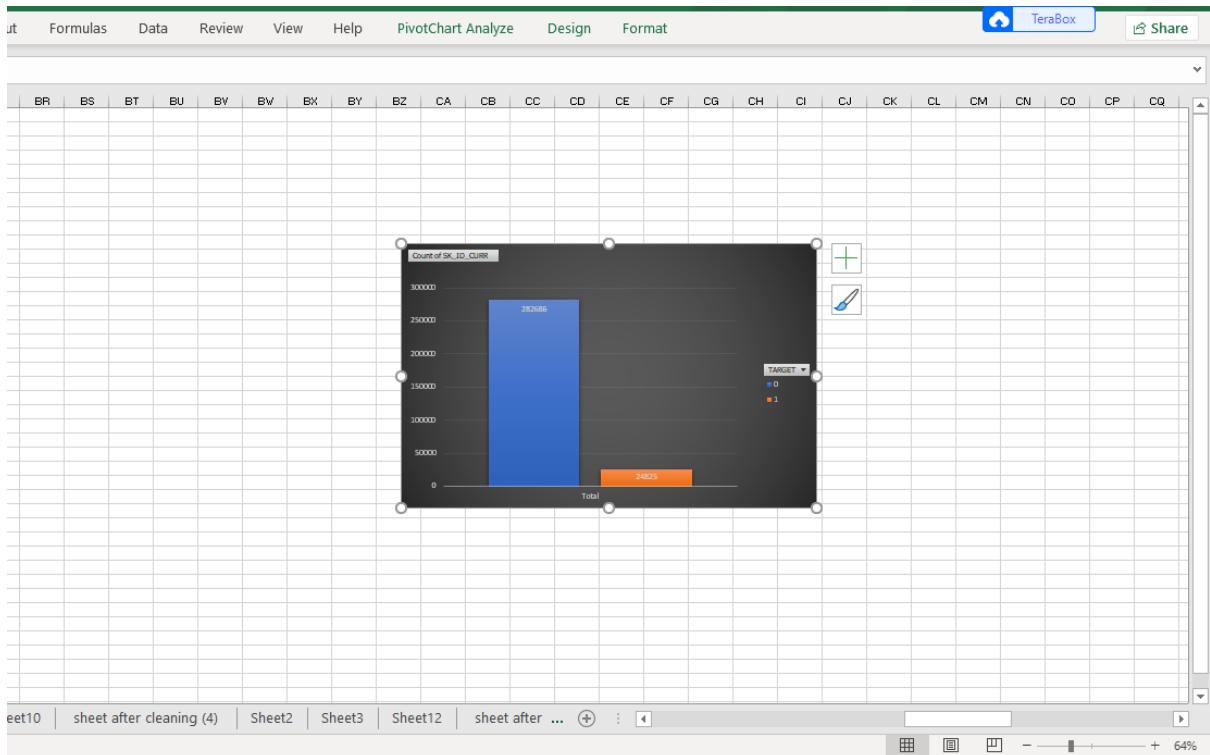
- Rows: SK\_ID\_CURR
- Columns: Sex (F/M), Is\_Infant (0/1), Is\_Married (0/1), Is\_Student (0/1), Is\_Working (0/1)
- Data: Count of SK\_ID\_CURR

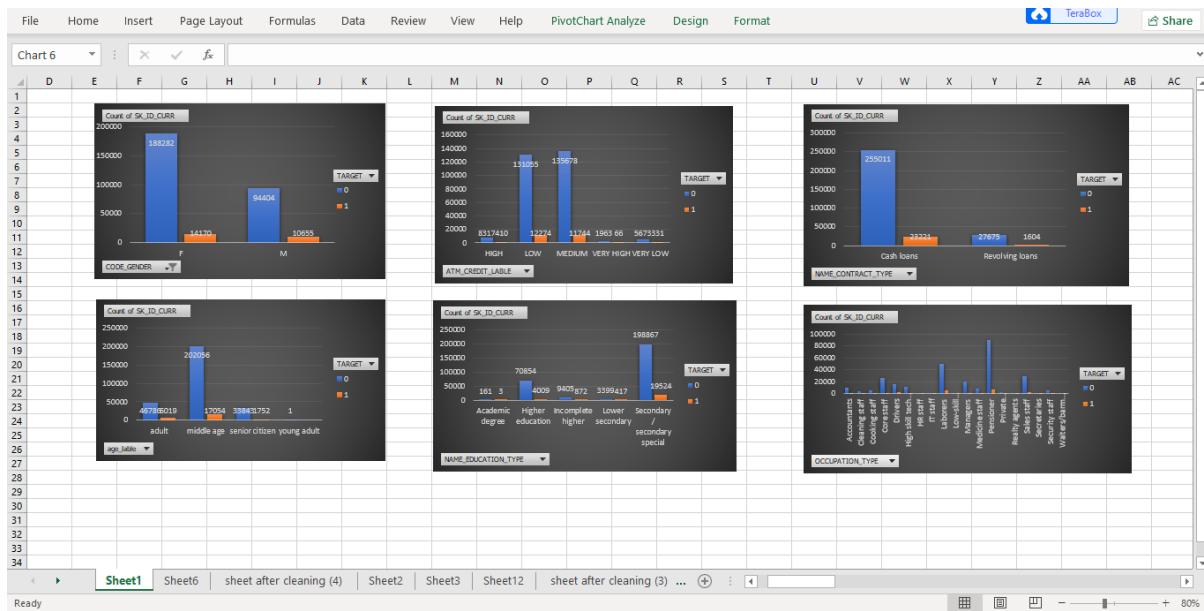
**PivotChart Fields:**

- Legend: TARGET (0, 1)
- Series: Count of SK\_ID\_CURR
- Target: 24825

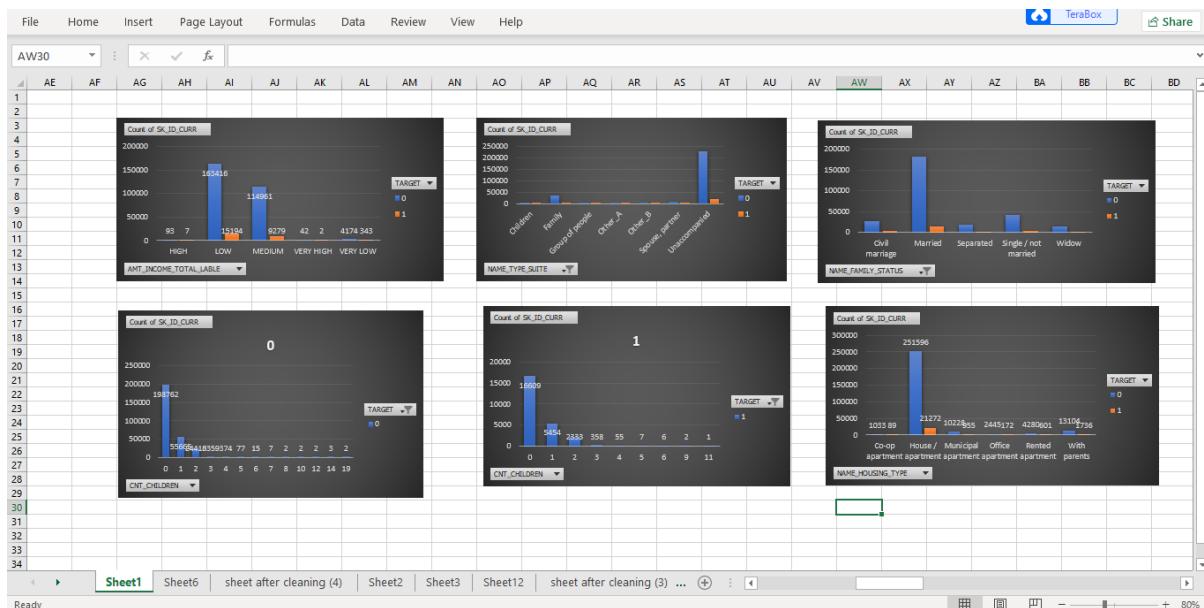
**Table Data (Partial):**

SK_ID_CURR	Sex	Is_Infant	Is_Married	Is_Student	Is_Working	Count of SK_ID_CURR	TARGET		
100025	F	0	1	0	1	1	0		
100026	F	0	0	0	1	0	0		
100027	F	0	1	0	0	0	0		
100029	M	1	0	2	247500	225000	Unaccompanied		
100030	F	0	1	0	0	0	702000	Unaccompanied	
100031	F	0	1	0	0	0	979992	Family	
100032	M	0	1	1	0	1	327024	270000	Unaccompanied
100033	M	1	1	0	0	0	790830	675000	Unaccompanied
100034	M	0	1	0	0	0	180000	180000	Unaccompanied
100035	F	0	1	0	0	0	665892	477000	Unaccompanied
100036	F	0	1	0	0	0	512064	360000	Family
100037	F	0	0	0	0	0	199008	180000	Unaccompanied
100039	M	1	0	1	0	1	733315.5	679500	Unaccompanied
100040	F	0	1	0	0	0	1125000	1125000	Unaccompanied
100041	F	0	0	0	0	0	450000	450000	Unaccompanied
100043	F	0	1	2	0	2	641173.5	553500	Unaccompanied
100044	M	0	1	0	0	0	454500	454500	Unaccompanied
100045	F	0	1	0	0	0	247275	225000	Unaccompanied
100046	M	1	1	0	0	0	540000	540000	Unaccompanied
100047	M	0	1	0	0	0	1193580	855000	Unaccompanied
100048	F	0	1	0	0	0	604152	540000	Unaccompanied
100049	F	0	0	0	0	0	288873	238500	Unaccompanied
100050	F	0	1	0	0	0	746280	675000	Unaccompanied
100051	M	0	1	0	0	0	661702.5	598500	Unaccompanied
100052	F	0	1	1	0	1	180000	180000	Family
100053	F	0	1	0	0	0	305221.5	252000	Family
100054	F	0	1	0	0	0	260640	225000	Unaccompanied
100055	F	0	0	0	0	0	1350000	1350000	Unaccompanied
100056	M	1	1	0	0	0	1506816	1350000	Unaccompanied
100058	F	0	1	0	0	0	135000	135000	Unaccompanied



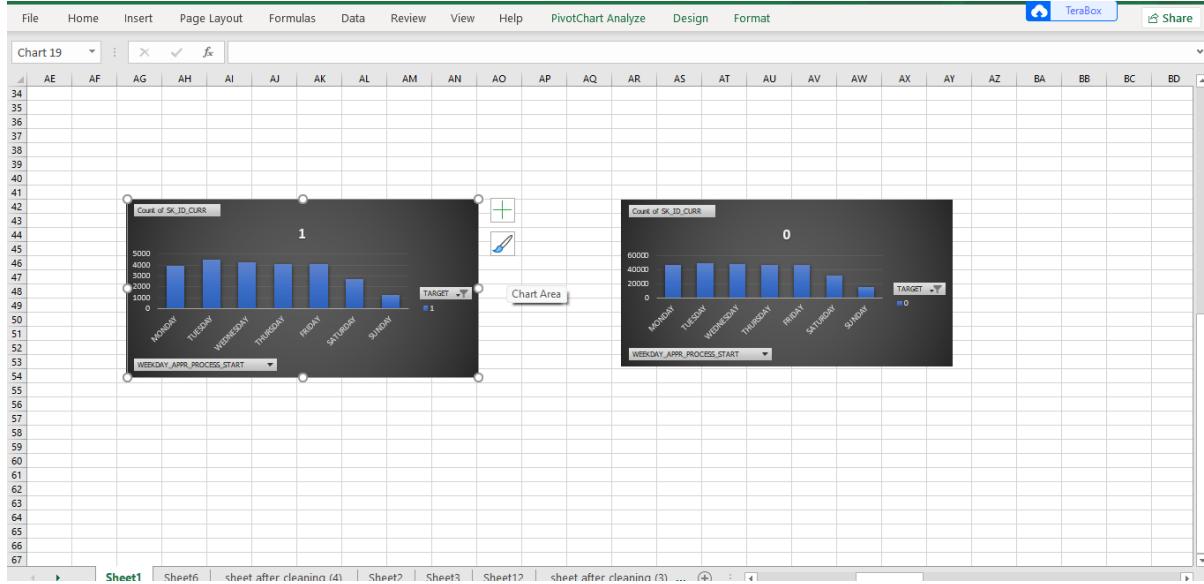


- Both in target 0 and target 1 most people applied for cash loans
  - The people who applied for loans most of them are female
  - By considering the people who applied for loans only one person have aged from 18-20
  - People in the age group 25-60 have been highest in the count
  - Students doing higher education or secondary /secondary special are those who have the highest count in loans accepted



- ➔ People having AMT\_INCOME\_TOTAL medium or low have the highest count in the loan application accepted
  - ➔ Clients with no children applied for the loan the most when compared to others

- In the NAME\_HOUSING\_TYPE column people having housing as a house/apartment have the highest count in the loan application accepted
- The married client are highest count both in target 1 and 2



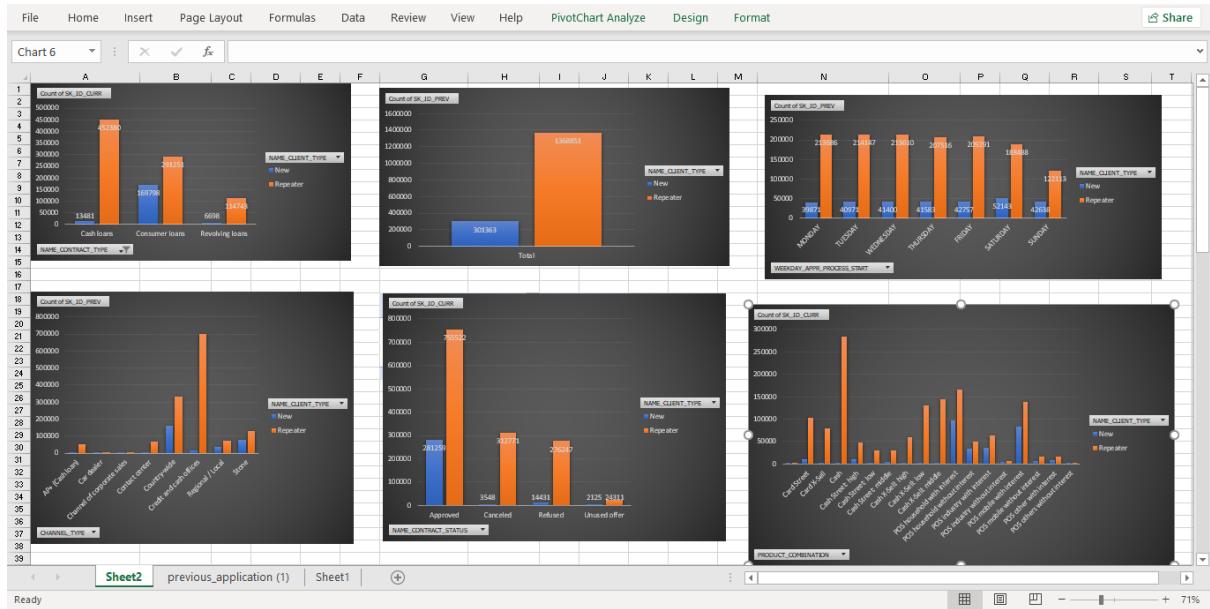
- By performing analysis column vice we can see that the maximum and minimum values are same in both the target 0 and 1 on all the columns

## PREVIOUS\_APPLICATION

NAME_CLIENT_TYPE	Count
Repeater	1368851
New	301363

- Finding the Ratio Between new and repeater client

=repeater/new=1368851/301363 =4.5:0



- Both in new and repeater most people applied for cash loans
- The loans that got approved are very high when compared to those cancelled or refused
- In the CHANNEL\_TYPE most are from credit and cash officers for repeater clients
- Most of the new clients are have countryside as their channel type

- Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms.

## UNIVARIATE

Using Histograms and descriptive analysis

## APPLICATION\_DATA

Tables	Add-ins	Histogram	Box	Tours	Sparklines	Filters	Links	TeraBox
100041	1	1	1	112500	1710000	U 0138029	18/24	b2/z
100047	1	0	1	202500	855000	0.025164	17482	1262
100049	1	0	0	135000	855000	0.007305	13384	3597
100051	1	0	1	315000	1710000	0.028663	24794	365243
100052	1	1	1	157500	1710000	0.035792	16667	200
100130	1	0	1	292500	1710000	0.007236	10526	267
100160	1	0	1	157500	1710000	0.035792	11667	200
100181	1	0	1	157500	1710000	0.022625	17538	774
100192	1	0	0	111915	1710000	0.016304	7989	150
100209	1	0	1	315000	1710000	0.034641	17326	1038
100214	1	0	1	202500	1710000	0.032561	10127	111
100246	1	0	1	135000	1710000	0.008019	12824	154
100273	1	0	1	157500	1710000	0.004849	23127	9274
100275	1	1	1	73541	1710000	0.011937	1471	2526
100286	1	1	1	121500	1710000	0.014464	17847	448
100295	1	0	1	225000	1019205	0.072508	11356	602
100300	1	0	0	63000	426645	0.018634	12209	1572
100301	1	0	1	571486.5	22468.5	0.022625	8292	166
100316	1	1	1	36000	284400	0.013949	20778	3633
100338	1	1	1	157500	497530	0.024561	13989	1550
100396	1	0	0	112500	417024	0.025164	13707	365243
100401	1	0	1	202500	343683	0.010966	20207	7809
100424	1	0	1	112500	117162	0.018851	18302	136
100439	1	0	1	81000	12433.5	0.018851	8039	8039
100452	1	0	1	171000	1009566	0.024705	8039	8039
100472	1	1	1	135000	545040	0.051164	18767	2046
100477	1	1	1	112500	284400	0.008474	13923	519
100485	1	1	0	270000	810000	0.015221	13715	151
100490	1	0	0	135000	750000	0.026392	11383	651
100540	1	0	1	157500	338832	0.018029	9289	1003
100541	1	1	0	180000	900000	0.018209	10616	1065
100547	1	1	0	211500	450000	0.026392	10037	148
100567	1	1	1	99000	180000	0.028663	8063	781
100615	1	1	0	225000	203760	0.072508	13481	234

Screenshot of Microsoft Excel showing the Data Analysis Tools ribbon tab selected. The Data ribbon tab is also visible. The Data Analysis Tools pane is open, displaying the message: "Data Analysis Tools Tools for financial and scientific data analysis." The main worksheet contains a large dataset with columns labeled SK\_ID\_CURR, TARGET, FLAG\_OWN\_CAR, FLAG\_OWN\_REALTY, CNT\_CHILDREN, AMT\_INCOME\_TOTAL, AMT\_CREDIT, AMT\_ANNUITY, AMT\_GOODS\_PRICE, and REGION\_POPULATION\_RELATIVE.

Screenshot of Microsoft Excel showing the Data Analysis Tools ribbon tab selected. The Data ribbon tab is also visible. A context menu is open over the dataset, showing options like Data Analysis, Descriptive Statistics, and other statistical tools. The Data Analysis Tools pane is open, displaying the message: "Data Analysis Tools Tools for financial and scientific data analysis." The main worksheet contains a large dataset with columns labeled SK\_ID\_CURR, TARGET, FLAG\_OWN\_CAR, FLAG\_OWN\_REALTY, CNT\_CHILDREN, AMT\_INCOME\_TOTAL, AMT\_CREDIT, AMT\_ANNUITY, AMT\_GOODS\_PRICE, and REGION\_POPULATION\_RELATIVE.

File Home Insert Page Layout Formulas Data Review View Help Table Design Query

**SECURITY WARNING** External Data Connections have been disabled [Enable Content](#)

H1 AMT\_ANNUITY

SK_ID_CURR	TARGET	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE
100003	0	0	0	0	270000	1293502.5	35698.5	1129500	0.003541
100004	0	1	1	0	67500	135000	6750	135000	0.010032
100006	0	0	1	0	135000	312682.5	29686.5	297000	0.008015
100007	0	0	1	0	121500	513000	21865.5	513000	0.028665
100008	0	0	1	0	99000	490495.5	27517.5	454500	0.035792
100009	0	1	1					1395000	0.035792
100010	0	1	1					1530000	0.003122
100011	0	0	1					913500	0.018634
100012	0	0	1					405000	0.019685
100014	0	0	1					652500	0.0228
100015	0	0	1					135000	0.015223
100016	0	0	1					67500	0.031325
100017	0	1	0					697500	0.016612
100018	0	0	1					679500	0.010006
100019	0	1	1					247500	0.020715
100020	0	0	0					387000	0.018634
100021	0	0	1					270000	0.010966
100022	0	0	1					157500	0.04622
100023	0	0	1					454500	0.015223
100024	0	1	1					427500	0.015223
100025	0	1	1					927000	0.025164
100026	0	0	0					450000	0.020715
100027	0	0	0					226000	0.006160

Descriptive Statistics

Input: Input Range: \$F:\$F OK Cancel Help  
Grouped By: Columns Labels in first row  
Output options: Output Range: New Worksheet Ply: New Workbook  
Summary statistics: Confidence Level for Mean: 95 %  
Kth Largest: Kth Smallest:

Average: 27163.62335 Count: 282675 Sum: 7678450067

File Home Insert Page Layout Formulas Data Review View Help

R22 AMT\_ANNUITY

Book123 - Excel

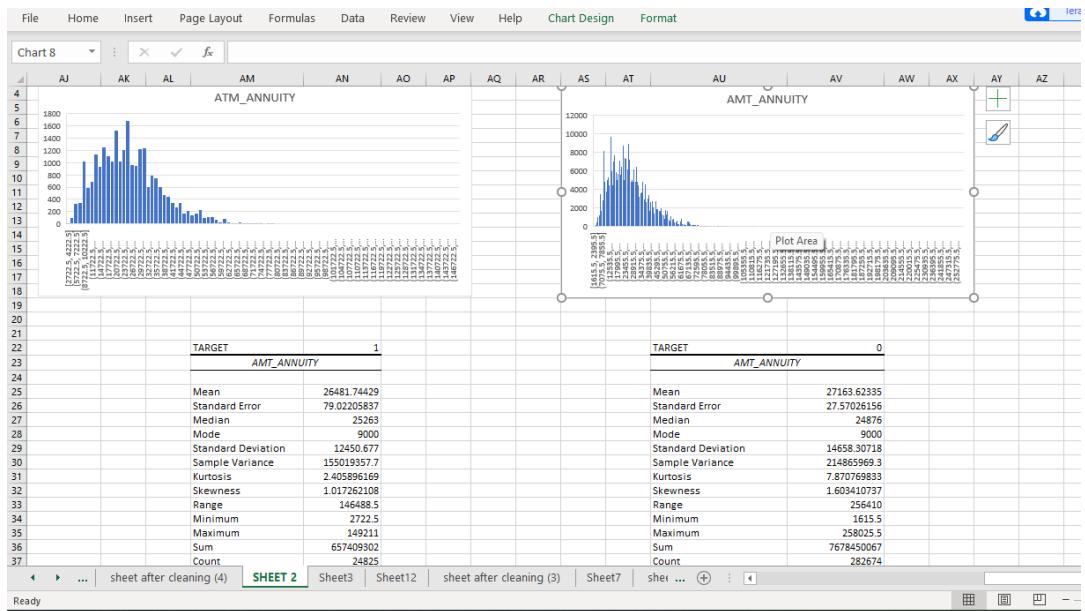
AMT_INCOME_TOTAL	AMT_INCOME_TOTAL
14000	35000
12000	30000
10000	25000
8000	20000
6000	15000
4000	10000
2000	5000
0	0

TARGET

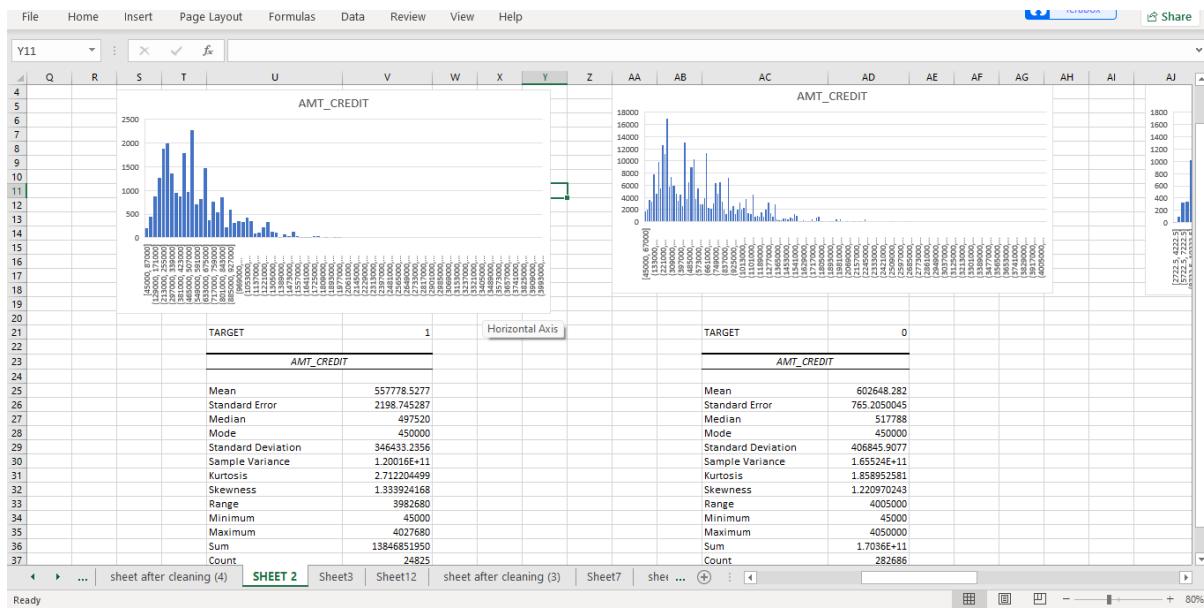
AMT_INCOME_TOTAL	
1	Mean: 165611.7603
1	Standard Error: 4739.015419
1	Median: 135000
1	Mode: 135000
1	Standard Deviation: 746676.9534
1	Sample Variance: 5.57526E+11
1	Kurtosis: 24150.78782
1	Skewness: 154.3468239
1	Range: 116374350
1	Minimum: 25650
1	Maximum: 117000000
1	Sum: 411131965
1	Count: 24825

AMT_INCOME_TOTAL	
0	Mean: 169077.7223
0	Standard Error: 207.7862696
0	Median: 148500
0	Mode: 135000
0	Standard Deviation: 110476.2685
0	Sample Variance: 12205005907
0	Kurtosis: 3419.565662
0	Skewness: 27.8256627
0	Range: 17974440
0	Minimum: 25650
0	Maximum: 18000090
0	Sum: 47795904396
0	Count: 282686

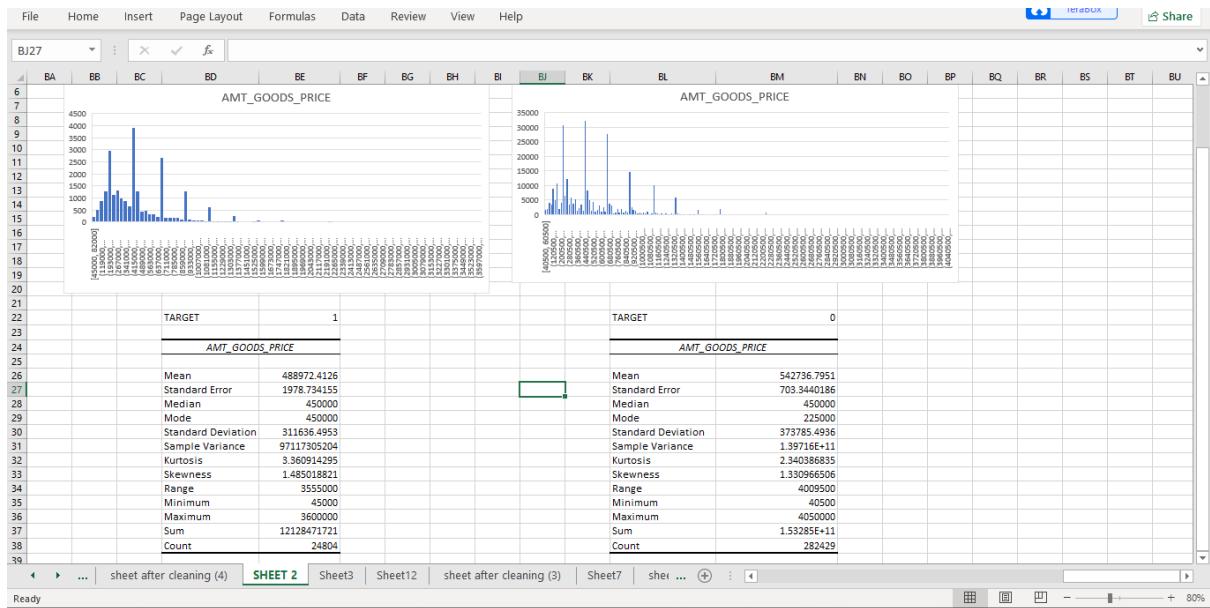
- The lower the variance lesser the values are scatter
- Target 1 has fewer scatter values when compared to Target 0
- And have positive skewness



- Target 0 has fewer scatter values when compared to target 1
- And have positive skewness
- Target 1 also has more outliers

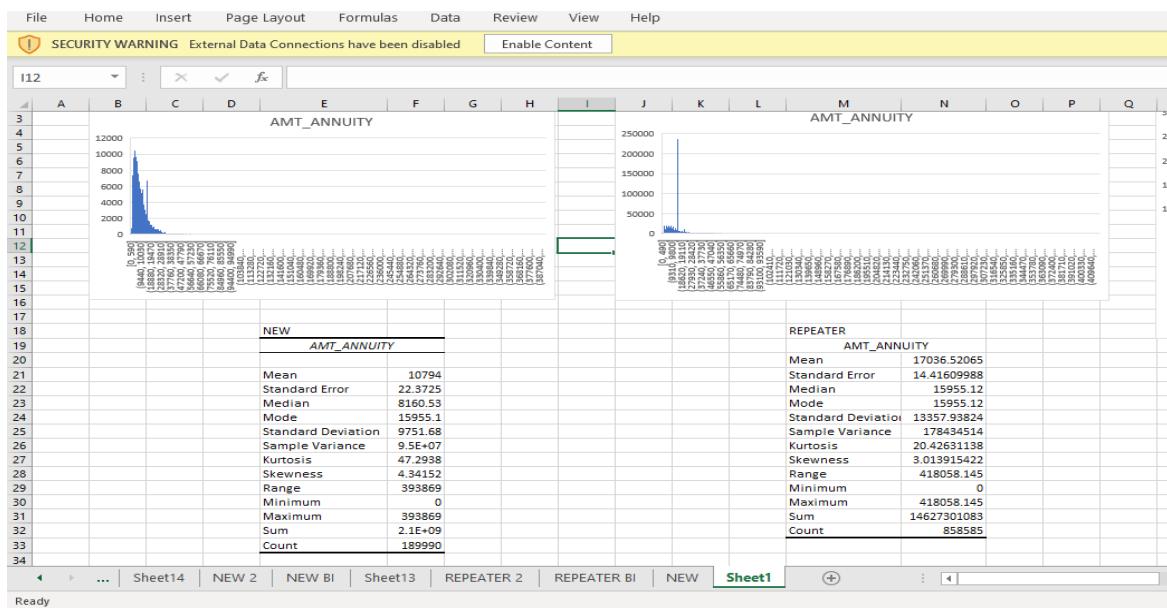


- Target 1 has fewer scatter values when compared to Target 0
- And have positive skewness
- Target 0 also has more outliers

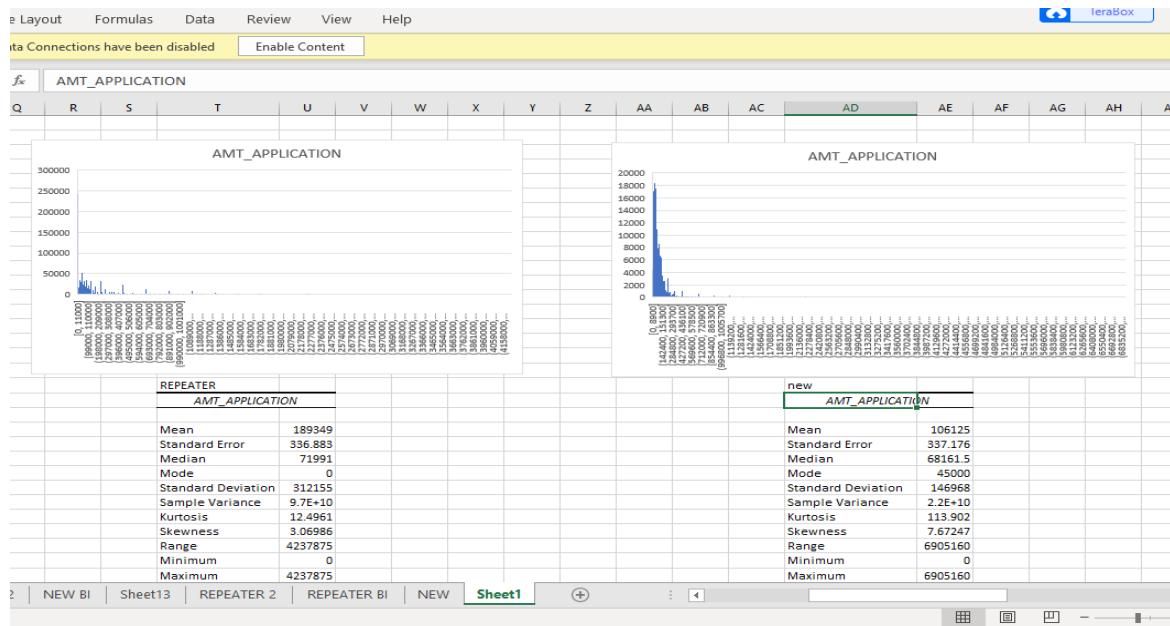


- Target 0 has fewer scatter values when compared to target 1
- And have positive skewness
- Target 0 also has more outliers

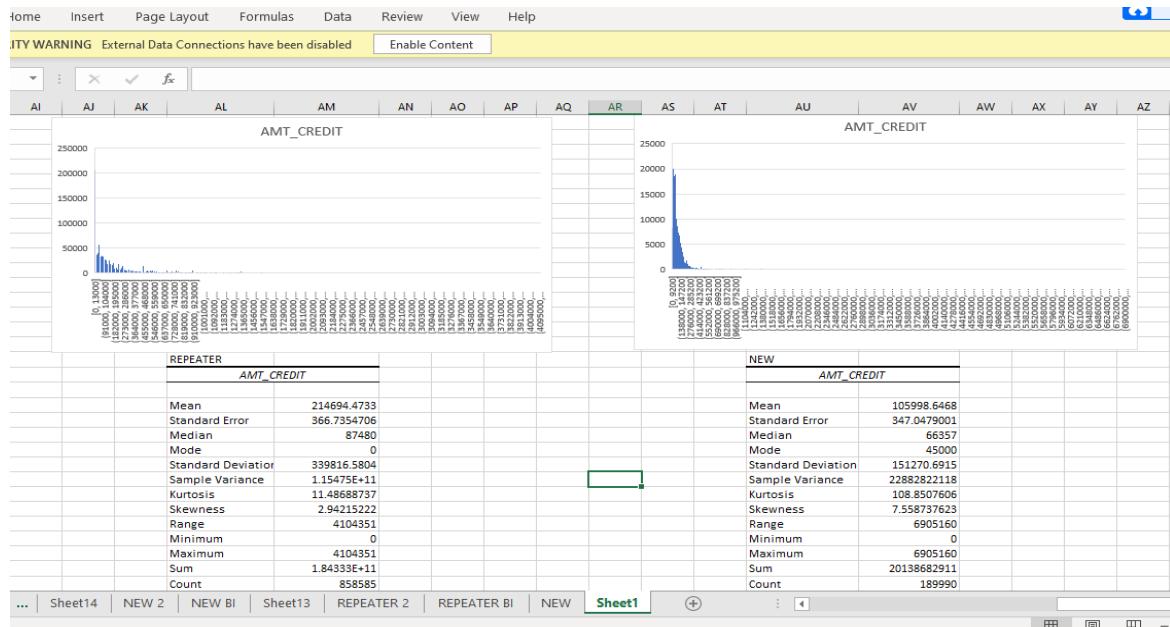
## PREVIOUS\_APPLICATION



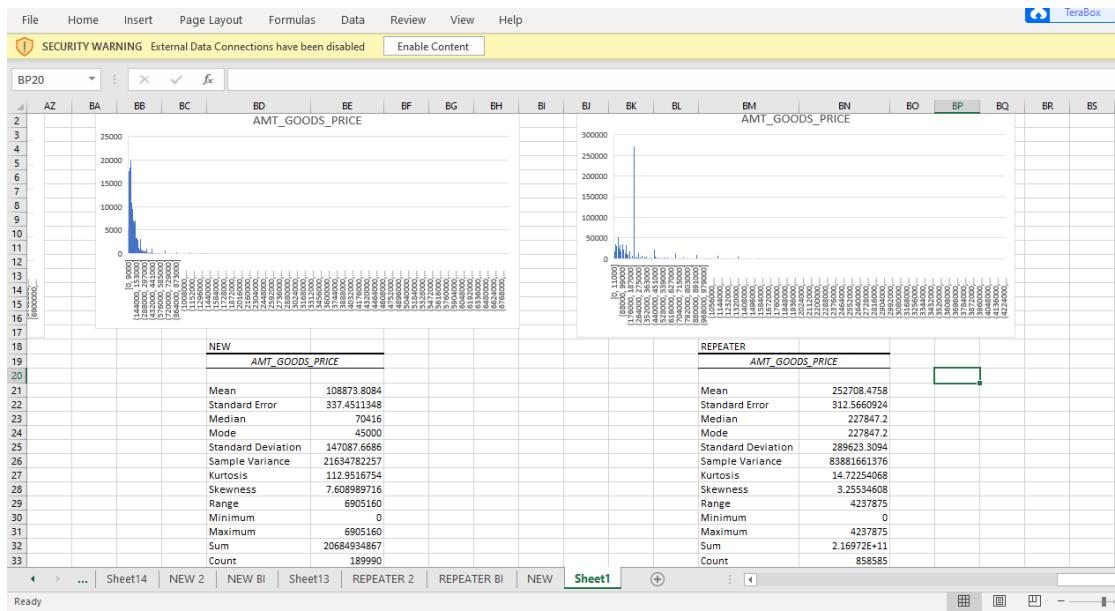
- Repeater clients has fewer scatter values when compared to new clients
- And have positive skewness



- Repeater clients has high variance when compared to new clients
- And have positive skewness
- Repeater have more outliers



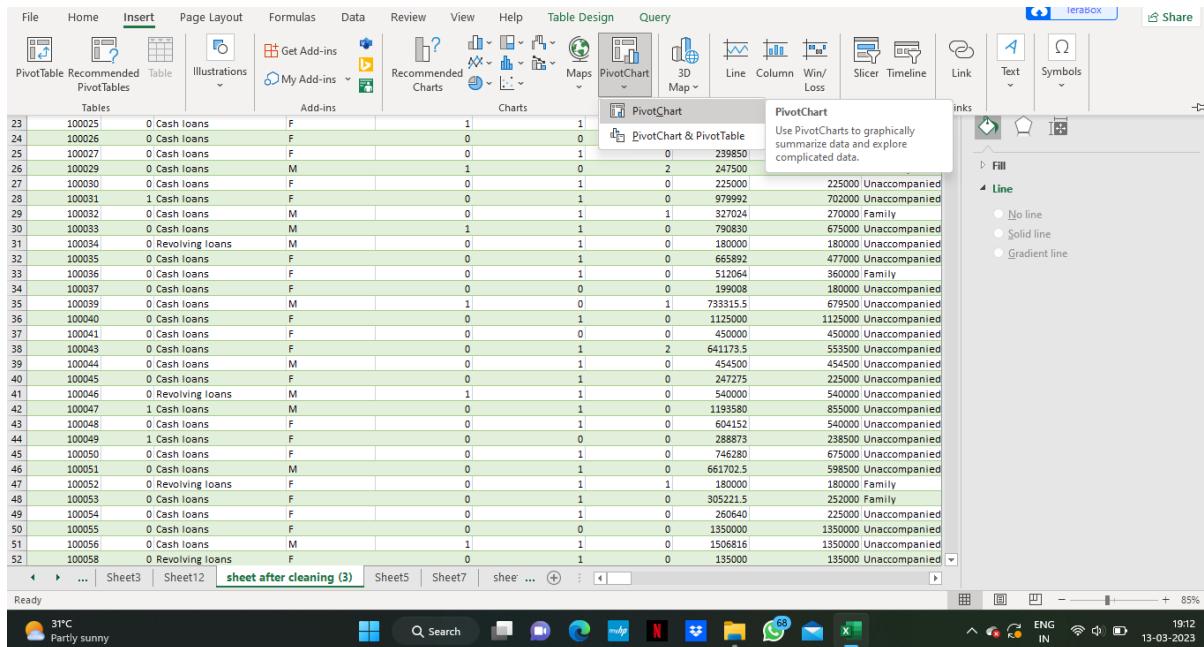
- Repeater clients has more scatter values when compared to new clients
- And have positive skewness
- Repeater have more outliers



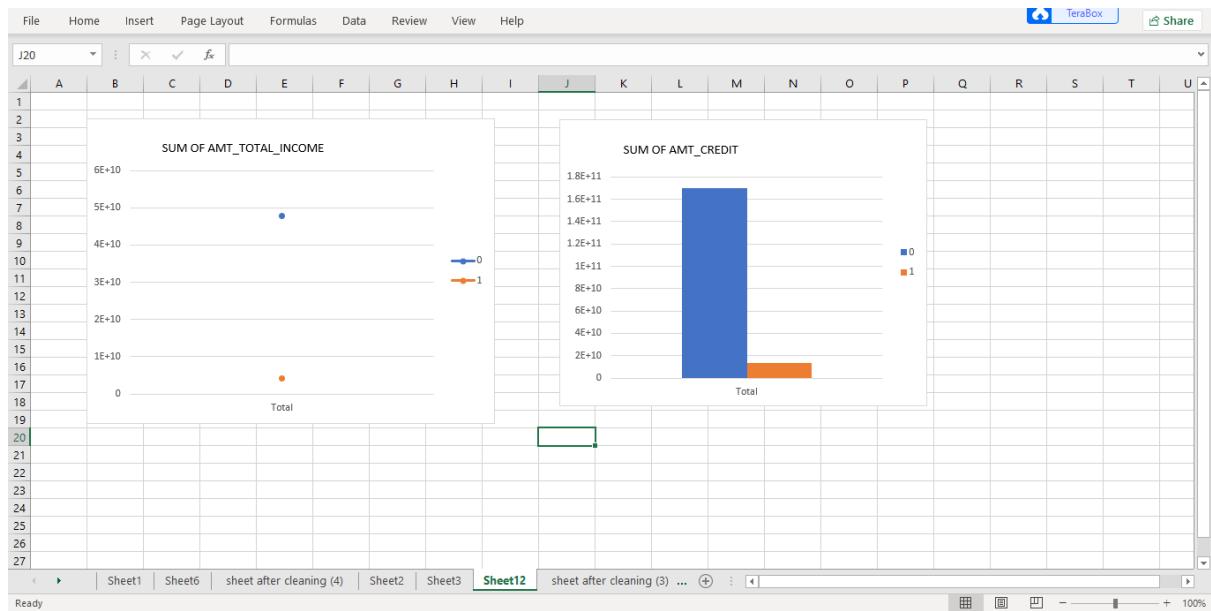
- Repeater clients has high variance when compared to new clients
- And have positive skewness
- Repeater have more outliers

## SEGMENTED UNIVARIATE

This analysis is done using pivotchart

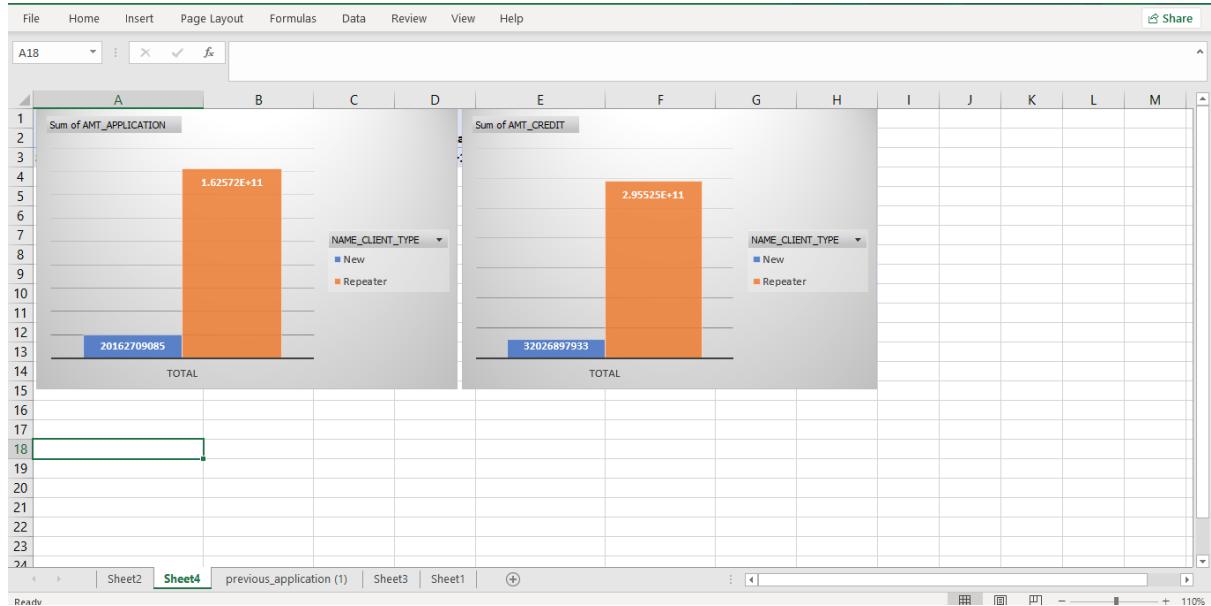


## APPLICATION\_DATA



→ Target 0 has the highest total in both AMT\_TOTAL\_INCOME and AMT\_CREDIT

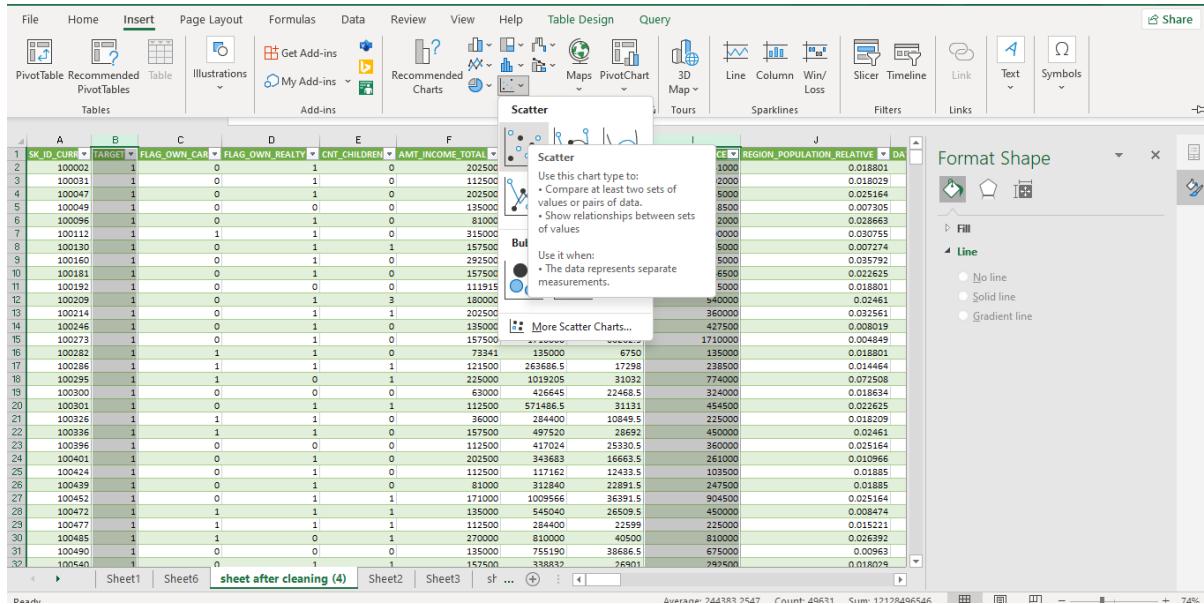
## PREVIOUS\_APPLICATION



→ There is an increase in credit for both new and repeater clients

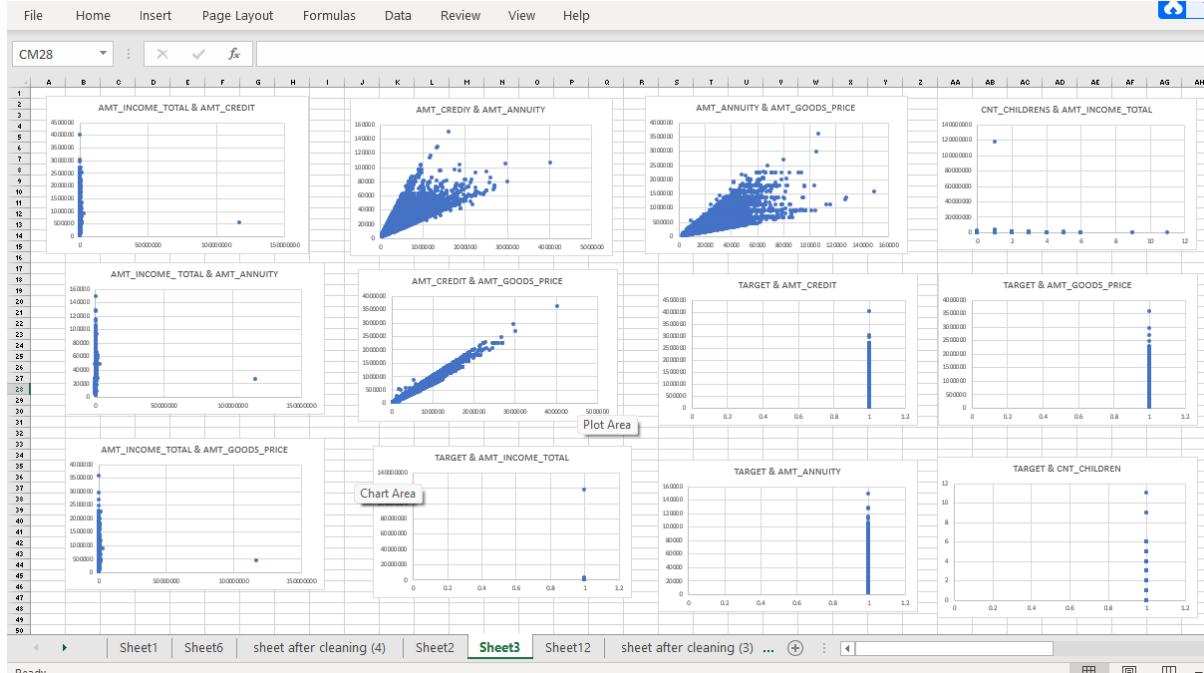
## BIVARIATE

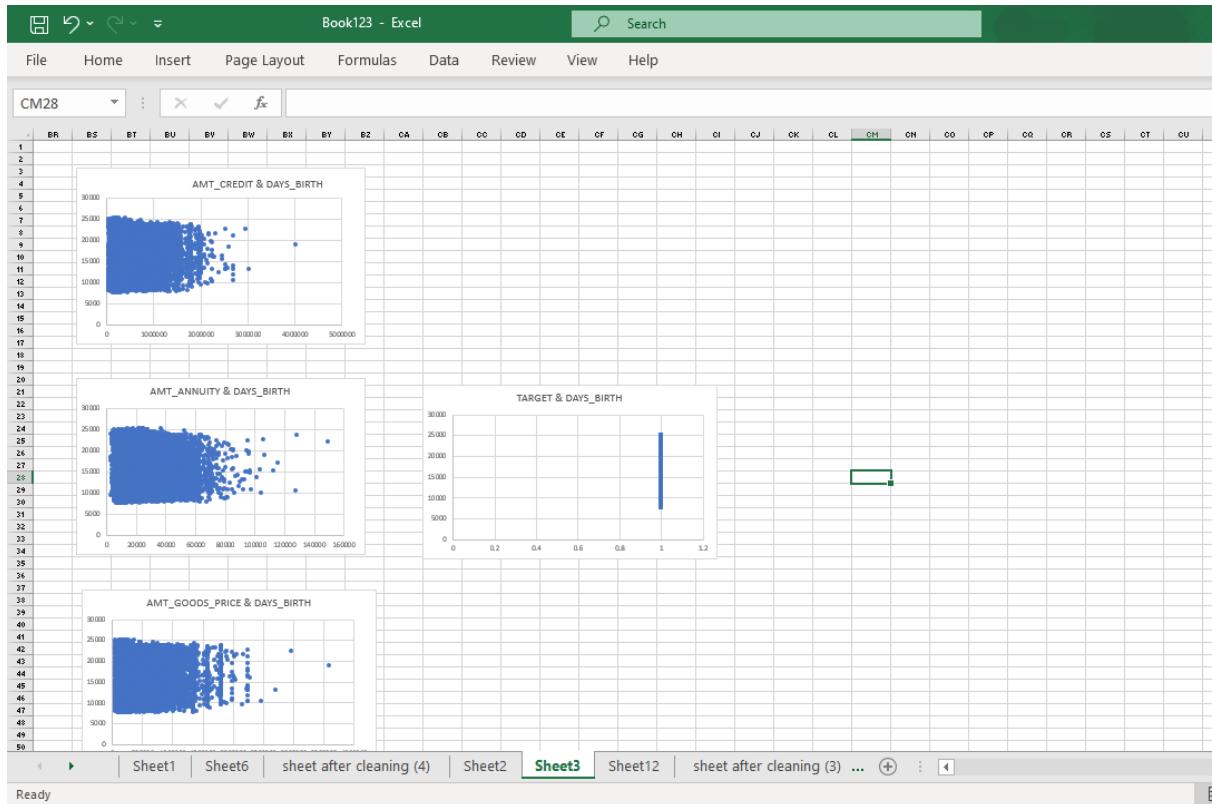
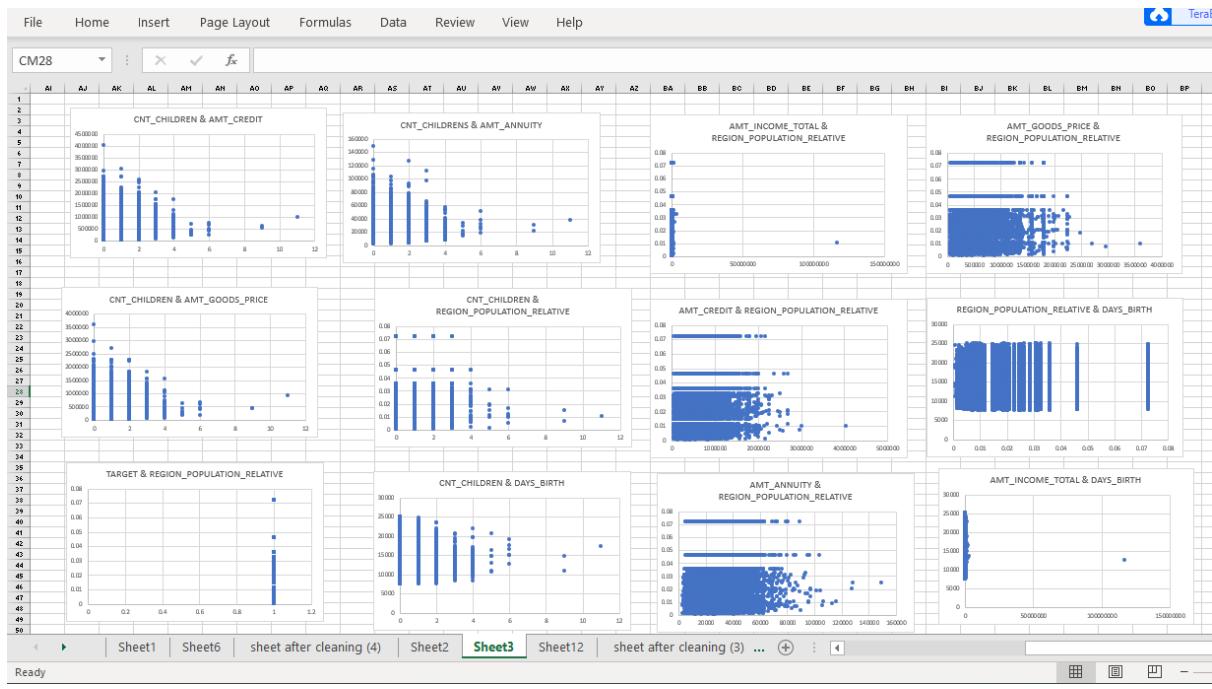
Scatter charts are used to perform bivariate analysis



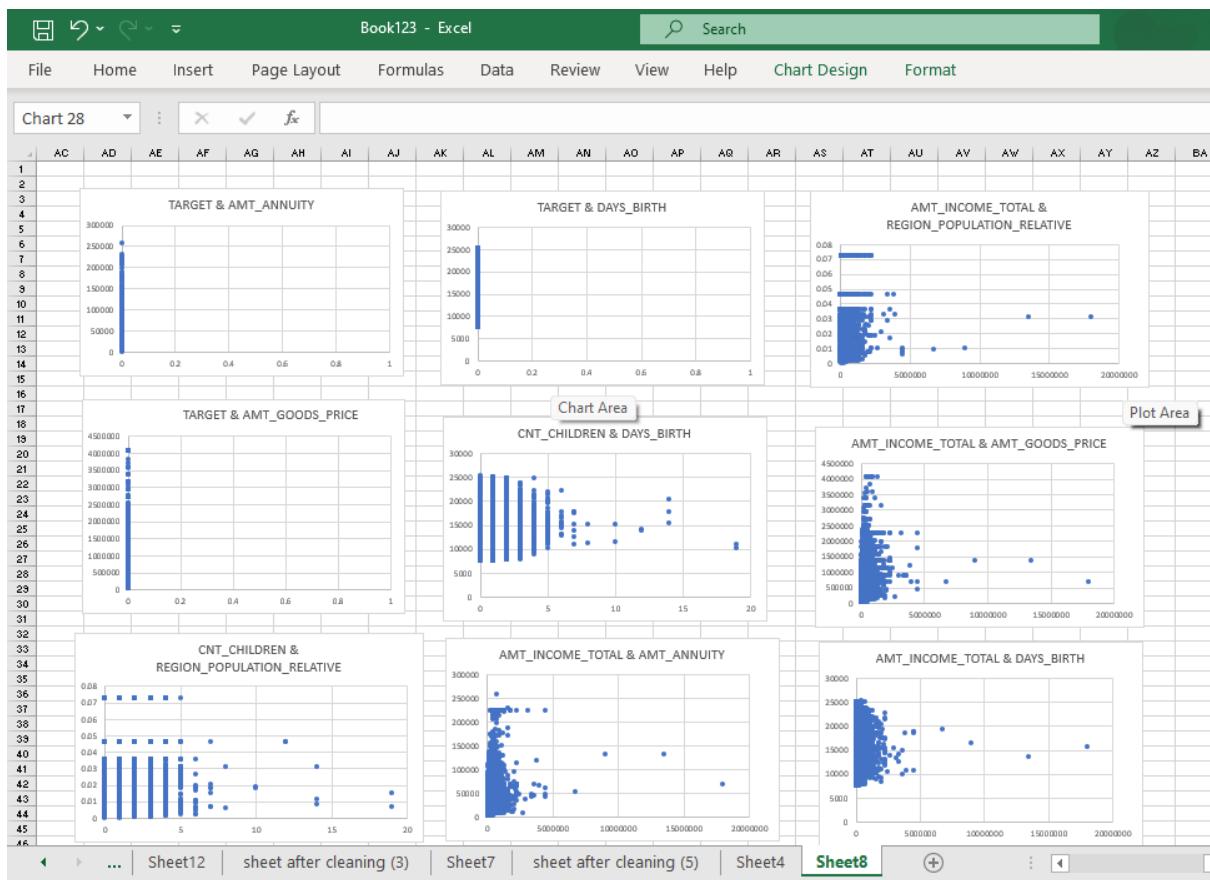
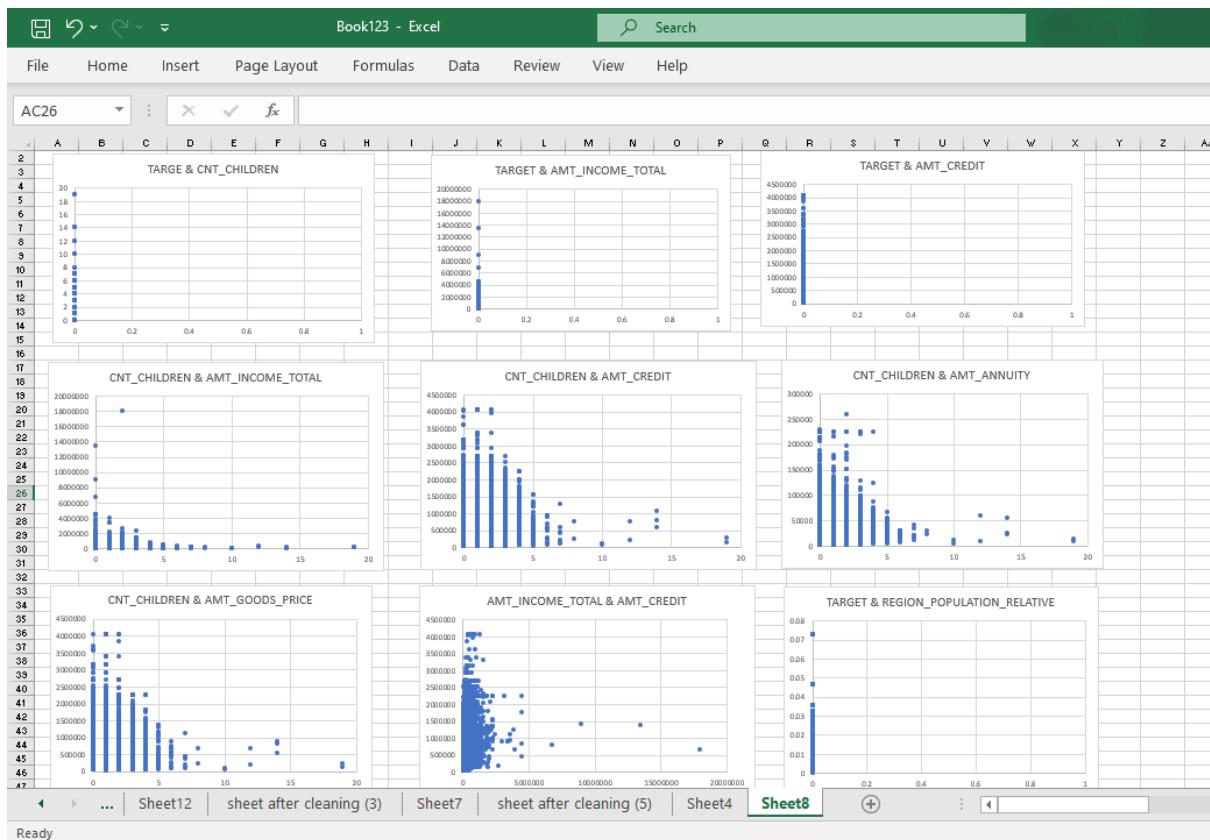
## APPLICATION\_DATA

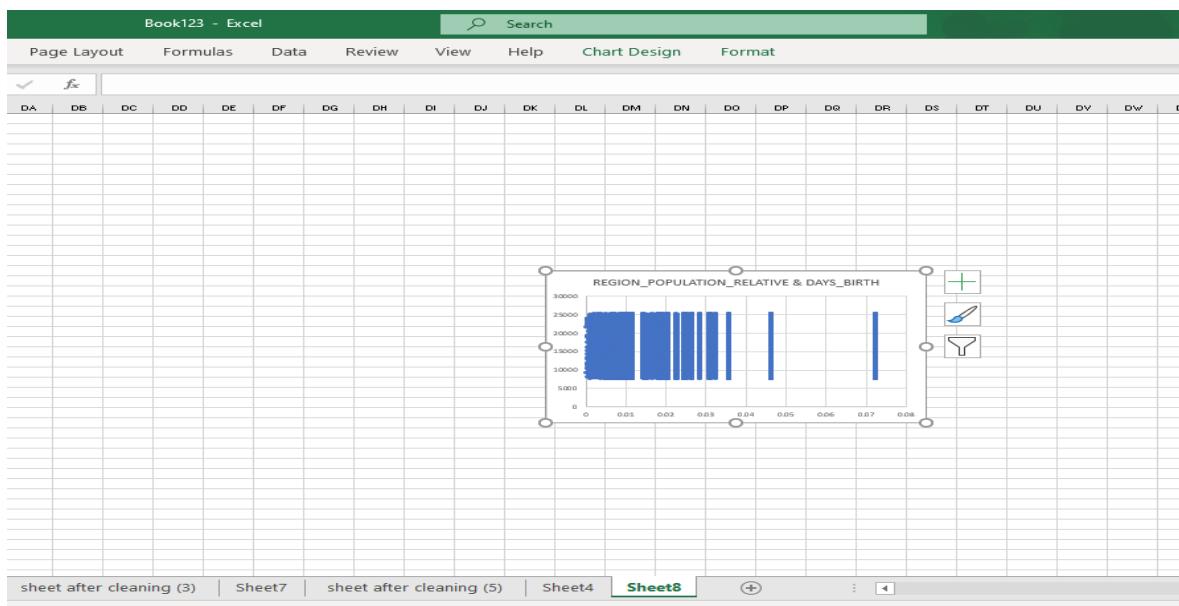
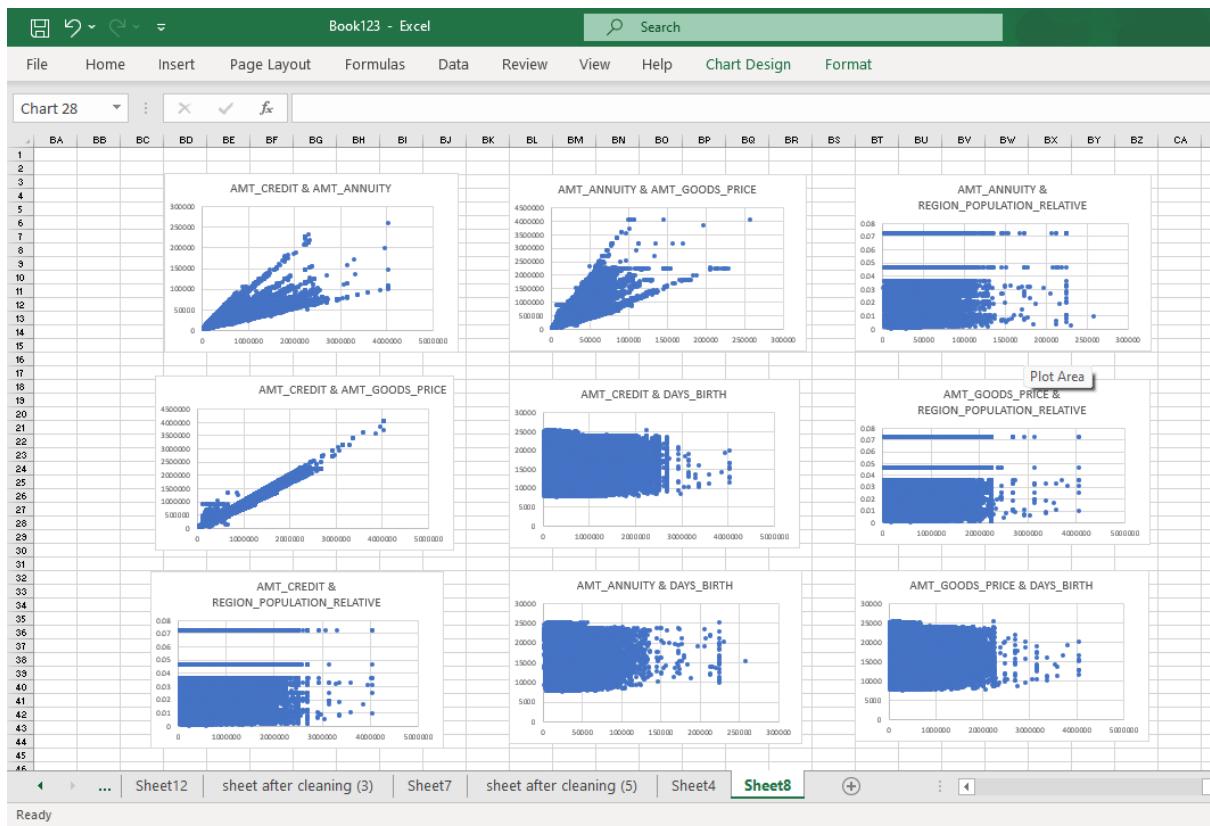
### TARGET 1





## TARGET 0



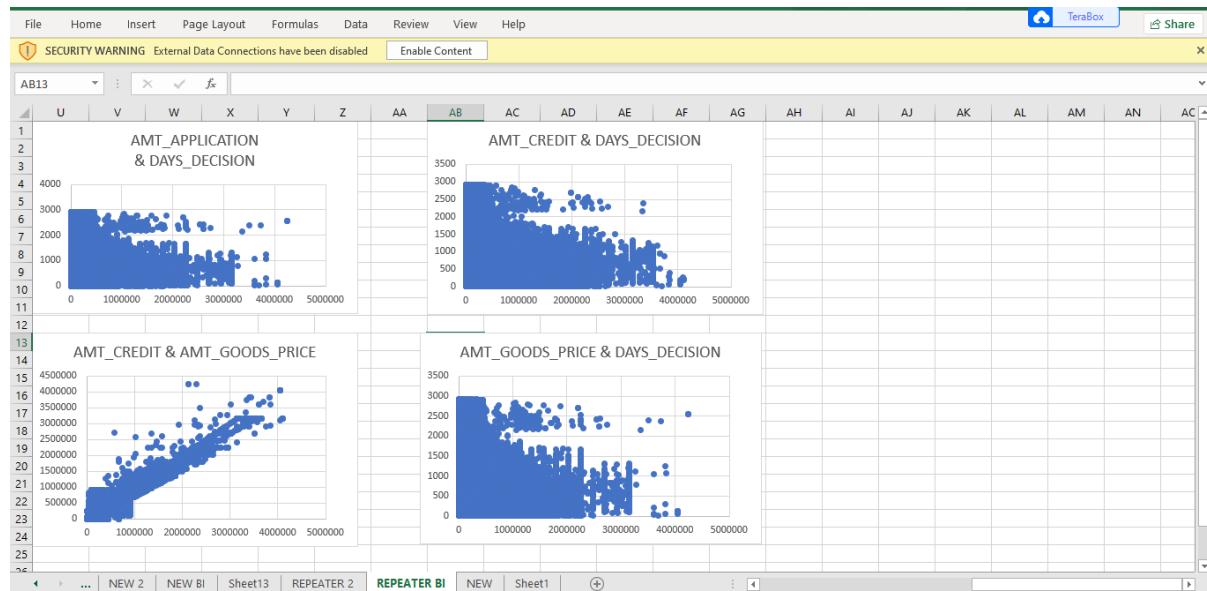
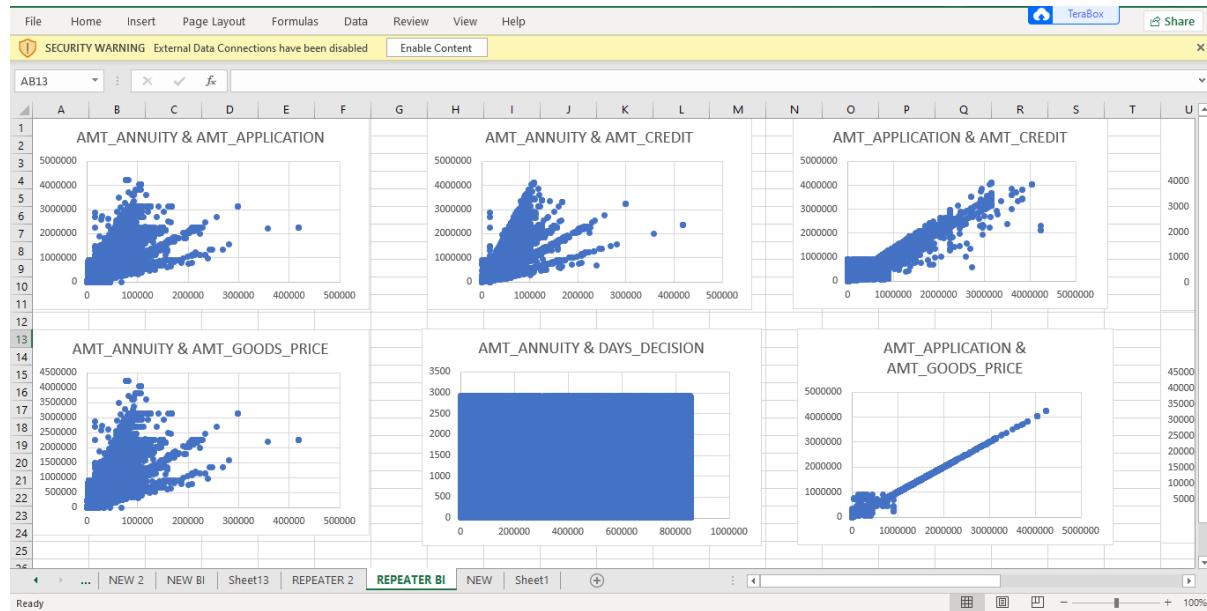


For Both Target 0 and 1

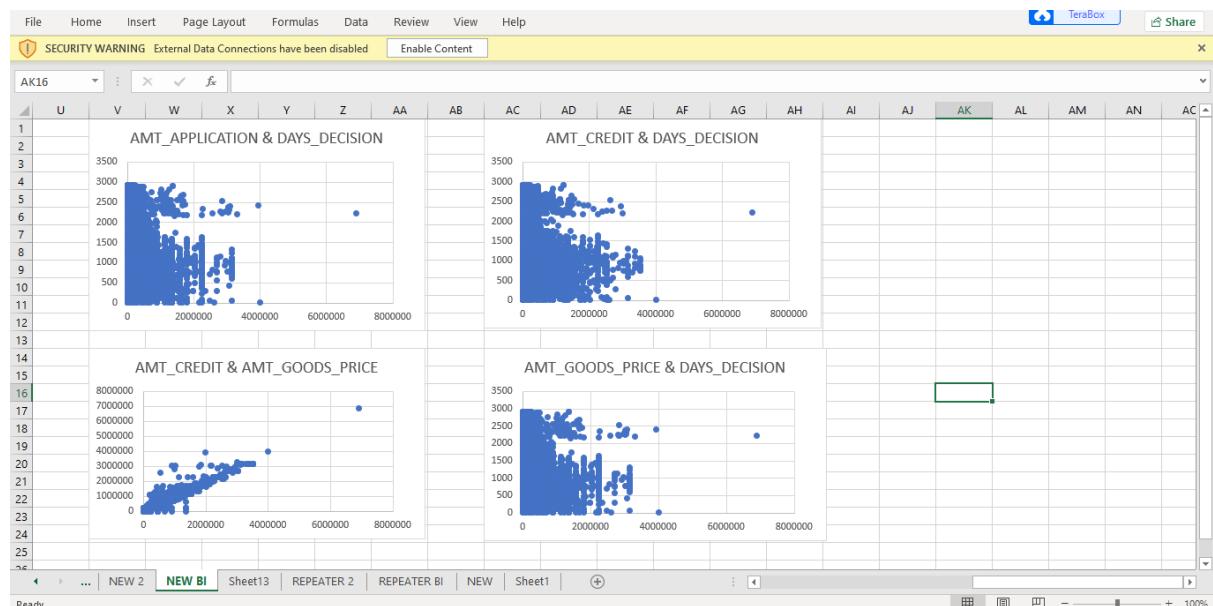
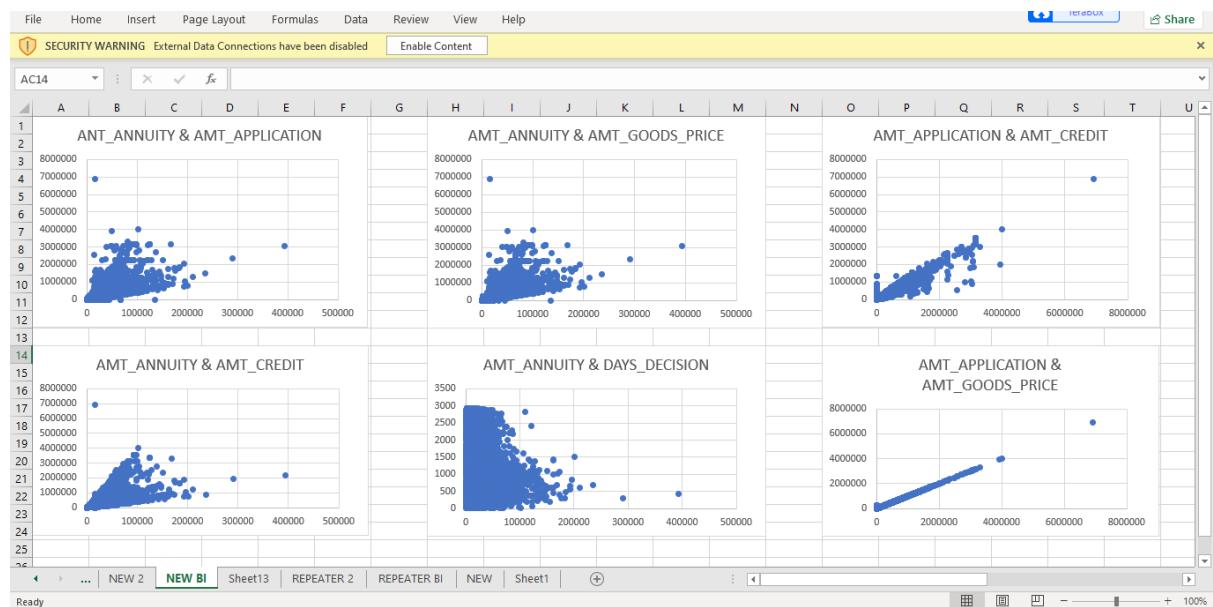
- These charts are used for finding the correlation between two variables
- Straight line represent higher correlation and acattered vales represent lesser correlation
- From the data given in the Scatter chat we can see that variables with higher correlation are AMT\_CREDIT & AMT\_GOODS\_PRICE

## **PREVIOUS\_APPLICATION**

### **REPEATER CLIENT**



## NEW CLIENT



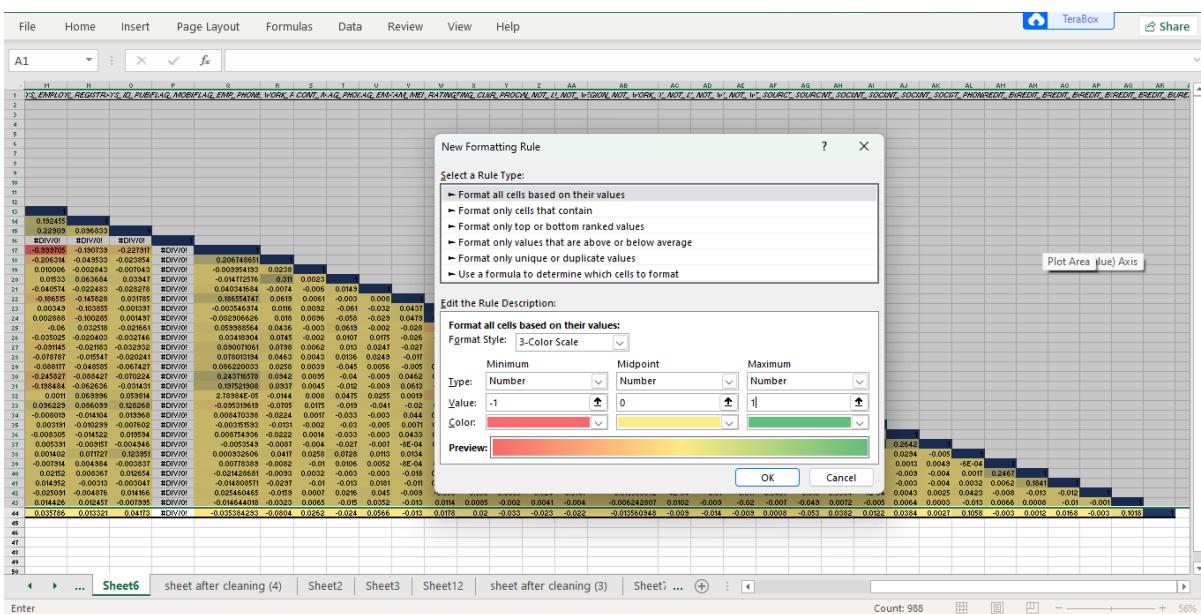
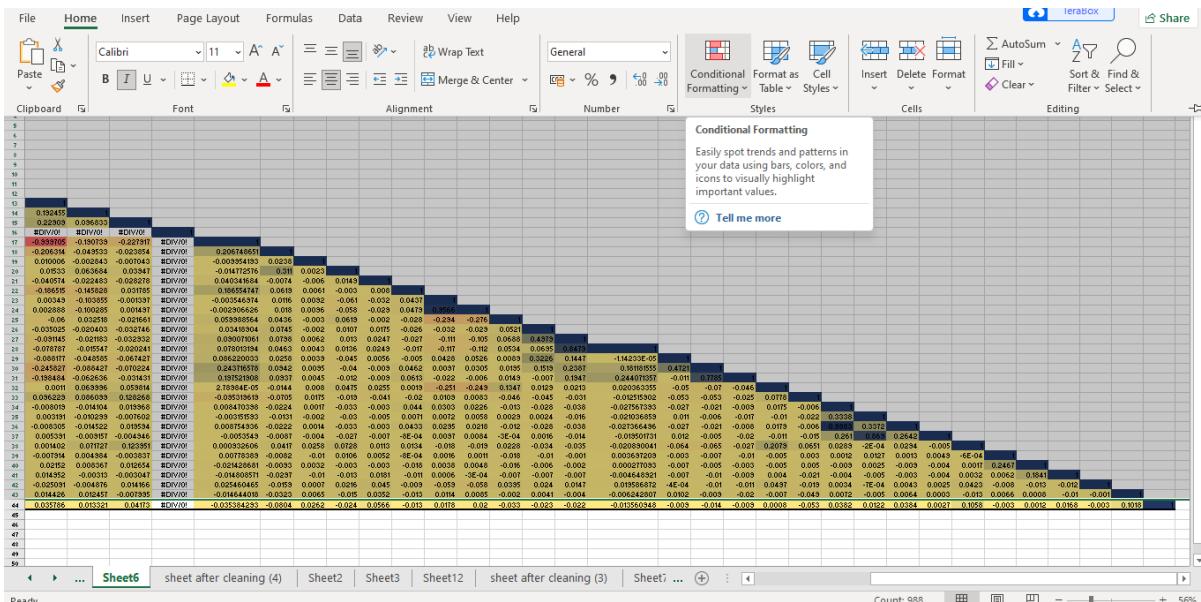
For Both New and Repeater client

- From the data given in the Scatter chart we can see that variables with higher correlation are AMT\_APPLICATION & AMT\_GOODS\_PRICE
- Find the top 10 **correlation** for the Client with payment difficulties and all other cases (Target variable). Note that you have to find the top correlation by segmenting the data frame w.r.t to the target variable and then find the top

correlation for each of the segmented data and find if any insight is there. Say, there are 5+1(target) variables in a dataset: Var1, Var2, Var3, Var4, Var5, Target. And if you have to find top 3 correlation, it can be: Var1 & Var2, Var2 & Var3, Var1 & Var3. Target variable will not feature in this correlation as it is a categorical variable and not a continuous variable which is increasing or decreasing.

- To find the correlation between two variables
- Data->Data Analysis->correlation

→ Conditional Formatting is used to give the required colours for specific values

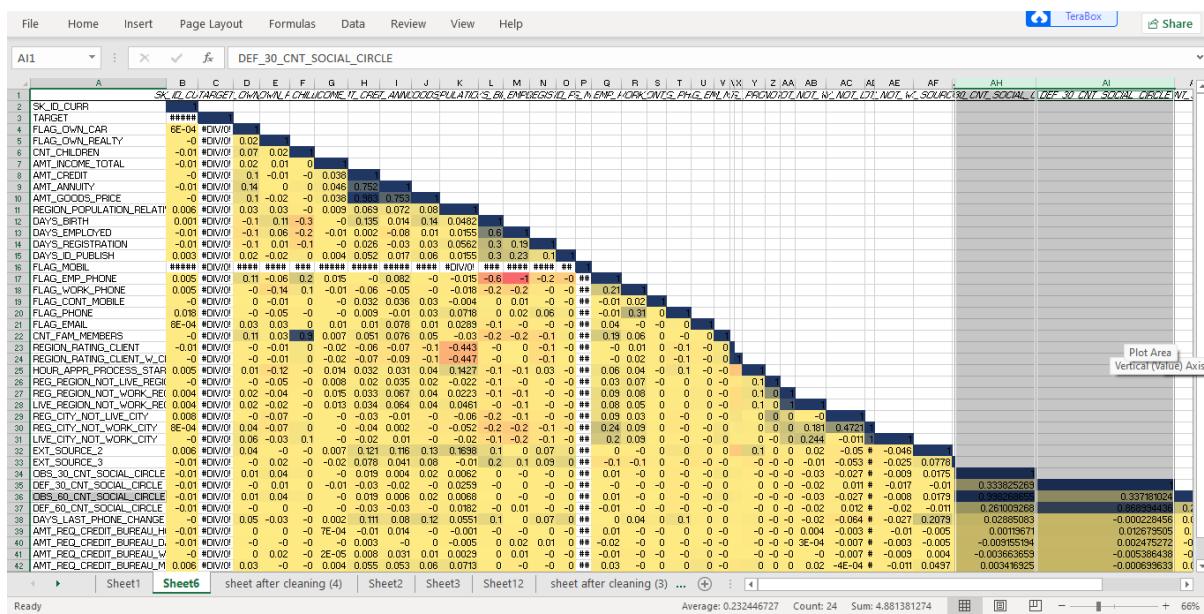


## APPLICATION\_DATA

### TARGET 1

- RED represents values with a correlation between -1 to 0
- Yellow represents value with a correlation between 0 to 1
- Blue represents values with a correlation between 0 to 1

File	Home	Insert	Page Layout	Formulas	Data	Review	View	Help	TeraBox	Share											
A33	EXT_SOURCE_3																				
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	
1	SK_ID_CURR	SK_ID_CURR	TARGETING_OWN_LOAN_AC	CHILD_INCOME_I	AMT_CREDIT	AMT_ANNUITY	GODOS_PULATIUS	DAYS_BIRTH	DAYS_EMPLOYED	REGISTR_S_ID	PUBLAG_NOTE	EMP_WORKF	CONT_HAG	PHOLAG_EMPL							
2	SK_ID_CURR	#D0V0	0.0060	0.0060	0.0238	0.07	0.752194736	0.07169025	0.0078	0.192455151	0.0638	0.02	0.01	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		
3	TARGET	#D0V0	0.0006	0.0006	0.0238	0.07	0.0052	0.0002	-0.002	0.003813	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		
4	FLAG_OWN_CAR	#D0V0	-0.001	0.0001	0.0238	0.07	0.0052	0.0002	-0.002	0.003813	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		
5	FLAG_OWN_REALTY	#D0V0	-0.001	0.0001	0.0238	0.07	0.0052	0.0002	-0.002	0.003813	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		
6	AMT_INCOME_TOTAL	#D0V0	-0.001	0.0011	0.0073	0.005	0.75263935	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		
7	AMT_CREDIT	#D0V0	-0.0001	0.0001	0.005595	-0.0012	-0.0008	0.003786	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		
8	AMT_ANNUITY	#D0V0	-0.0001	0.0001	0.005595	-0.0012	-0.0008	0.003786	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		
9	AMT_GOODS_PRICE	#D0V0	-0.0002	0.0001	0.005693	-0.0011	-0.0008	0.003786	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		
10	REGION_POPULATION_RELATIV	#D0V0	0.0003	0.0023	0.0033	-0.032	0.0005	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		
11	REGION_TYPE_SUITE	#D0V0	0.0003	0.0023	0.0033	-0.032	0.0005	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		
12	DAYS_EMPLOYED	#D0V0	-0.0005	0.0125	0.0058	-0.183	-0.015	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		
13	DAYS_REGISTRATION	#D0V0	-0.0005	0.0068	0.0135	-0.143	-0.002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		
14	DAYS_PUBLISH	#D0V0	0.0025	0.0029	-0.0197	0.032	0.0041	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		
15	FLAG_MOBIL	#D0V0	#D0V0	#D0V0	#D0V0	#D0V0	#D0V0	#D0V0	#D0V0	#D0V0	#D0V0	#D0V0	#D0V0	#D0V0	#D0V0	#D0V0	#D0V0	#D0V0	#D0V0		
16	FLAG_EMP_PHONE	#D0V0	0.0059	0.0057	0.0044	0.025	0.002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	
17	FLAG_CONT_MOBILE	#D0V0	-0.0003	0.0001	0.0051	-0.0065	0.0004	0.0009	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	
18	FLAG_PHONE	#D0V0	0.0182	0.0182	-0.0001	-0.0438	-0.024	-0.0045	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		
19	FLAG_EMAIL	#D0V0	0.0009	0.0009	0.0289	0.0273	0.014	0.00989	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		
20	CNT_FAM_MEMBERS	#D0V0	-0.0004	0.0001	0.0269	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		
21	REGION_RATING_CLIENT	#D0V0	-0.0012	0.0001	0.0269	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		
22	REGION_RATING_CLIENT_W_CITY	#D0V0	-0.0004	0.0001	0.0269	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		
23	HOUR_APPR_PROCESS_START	#D0V0	0.0183	-0.1174	-0.024	0.0377	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		
24	REG_REGION_NOT_LIVE_REGION	#D0V0	-0.0004	-0.0041	-0.0538	0.0274	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		
25	REG_REGION_NOT_WORK_REGION	#D0V0	0.01843	-0.0424	-0.021	0.01453	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		
26	LIVE_REGION_NOT_WORK_REGION	#D0V0	0.0183	-0.0884	-0.021	0.01341	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		
27	REG_CITY_NOT_WORK_CITY	#D0V0	0.01843	-0.0424	-0.021	0.01453	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		
28	LIVE_CITY_NOT_WORK_CITY	#D0V0	0.01843	-0.0424	-0.021	0.01453	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		
29	LIVE_REGION_NOT_LIVE_CITY	#D0V0	0.01843	-0.0424	-0.021	0.01453	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		
30	LIVE_CITY_NOT_LIVE_CITY	#D0V0	0.01843	-0.0424	-0.021	0.01453	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		
31	LIVE_REGION_NOT_WORK_REGION	#D0V0	0.01843	-0.0424	-0.021	0.01453	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		
32	LIVE_CITY_NOT_WORK_CITY	#D0V0	0.01843	-0.0424	-0.021	0.01453	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001		
33	EXT_SOURCE_3	#D0V0	-0.0009	0.0009	0.0209	-0.02	-0.051	0.077683467	0.04145205	0.05789	-0.001	0.017620323	0.006226868	0.0060933	0.01282618	#D0V0	-0.01	-0.07	0.075	-0.01	-0.041
34	OBS_30_CNT_SOCIAL_CIRCLE	#D0V0	0.0183	0.0394	0.028	-0.0047	0.044852328	0.0198	0.005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	
35	OBS_60_CNT_SOCIAL_CIRCLE	#D0V0	0.0183	0.0394	0.028	-0.0047	0.044852328	0.0198	0.005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005		
36	DEF_30_CNT_SOCIAL_CIRCLE	#D0V0	0.0183	0.0394	0.028	-0.0047	0.044852328	0.0198	0.005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005		
37	DEF_60_CNT_SOCIAL_CIRCLE	#D0V0	0.0183	0.0394	0.028	-0.0047	0.044852328	0.0198	0.005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005		
38	DAYS_LST_LST_PHONE_CHANGE	#D0V0	0.0002	0.04902	-0.030	0.012	0.00243	0.070851482	0.070869392	0.188	0.0551	0.011958914	0.000402407	0.0771277	0.123*	#D0V0	0.047	0.0259	0.0728	0.0193	
39	AMT_REQ_CREDIT_BUREAU_HI	#D0V0	0.0044	0.0001	-0.04	-0.04	0.00066	0.001043333	0.0045	-0.0005	-0.014654464	0.000719434	0.00484	-0.0003	0.0004	0.0004	-0.0005	-0.0004	-0.0005	0.0004	
40	AMT_REQ_CREDIT_BUREAU_MID	#D0V0	-0.0087	-0.0006	0.013	-0.0003	0.00030802	-0.00024	0.0005	-0.0005	0.00742621	0.02503039	0.008837	0.002654	-0.0002	-0.0002	-0.0002	-0.0002	-0.0002	-0.0002	
41	AMT_REQ_CREDIT_BUREAU_W	#D0V0	0.0001	0.0001	-0.02	0.018	0.00451	0.01204561	0.008	0.0028	0.00537057	-0.004958	0.0003	-0.0036	0.0001	-0.01	-0.03	-0.03	-0.03	0.0045	
42	AMT_REQ_CREDIT_BUREAU_M	#D0V0	0.0261	-0.0002	0.013	-0.004	0.00537068	0.005437398	0.058	0.01703	0.00571308	-0.0029354	0.00484	-0.0016	0.0001	-0.03	-0.0007	-0.026	0.0045		



TARGET 0

File Home Insert Page Layout Formulas Data Review View Help

REG\_CITY\_NOT\_WORK\_CITY

A	B	C	D	E	F	G	H	I	J	K	L	M	N	
SK_ID_CURR	TARGET	G_OWN_C_G_OWN_REL	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	GOODS_FPOPULATION	DAYS_BIRTH	DAYS_EMPLOYED	AYS_REGISTRATION				
3	TARGET	#DIV/0!												
4	FLAG_OWN_CAR	0.00121603	#DIV/0!	1										
5	FLAG_OWN_REALTY	0.00084137	#DIV/0!	-0.0052										
6	CNT_CHILDREN	-0.0007158	#DIV/0!	0.10565	-0.00404									
7	AMT_INCOME_TOTAL	0.00173896	#DIV/0!	0.18422	0.00206	0.027397188	1							
8	AMT_CREDIT	-0.000342	#DIV/0!	0.11717	-0.04165	0.003081225	0.34279945	1						
9	AMT_ANNUITY	6.823E-05	#DIV/0!	0.14178	-0.00601	0.020905318	0.418952886	0.771308946						
10	AMT_GOODS_PRICE	-0.0002051	#DIV/0!	0.12114	-0.0477	-0.000524874	0.349461894	0.987250457	0.776685775					
11	REGION_POPULATION_RELATIVE	0.00035999	#DIV/0!	0.04138	0.1354	-0.024362658	0.167850636	0.100603799	0.120988482	0.10383	1			
12	DAYS_BIRTH	0.00134617	#DIV/0!	-0.13706	0.11998	0.336966484	0.062609158	0.047370881	-0.01262771	0.04456	0.25254	1		
13	DAYS_EMPLOYED	0.00174396	#DIV/0!	-0.15956	0.07141	-0.245170465	-0.140392466	-0.070104514	-0.104978408	-0.06861	-0.0072	0	Plot Area (blue) Axis	
14	DAYS_REGISTRATION	0.00147511	#DIV/0!	-0.09038	0.02594	0.185791545	0.064937112	-0.013477233	-0.039435609	-0.01591	0.05208	0.383151	1	
15	DAYS_ID_PUBLISH	7.7019E-05	#DIV/0!	-0.01793	-0.00585	0.028705653	-0.022869393	0.00146417	0.014112898	0.00365	0.00107	0.271314	0.27666316	0.100236041
16	FLAG_MOBIL	0.00292507	#DIV/0!	-0.0026	-0.00125	0.00108046	0.007033387	0.001498	0.000160729	0.00146	0.0014	0.00326	0.000925528	0.000127538
17	FLAG_EMP_PHONE	-0.0017175	#DIV/0!	0.15909	0.07089	0.244274063	0.140743926	0.071256808	0.105644511	0.06977	0.00709	-0.62207	-0.999756229	-0.212296555
18	FLAG_WORK_PHONE	-0.000854	#DIV/0!	0.01462	-0.11155	0.055452828	-0.032404083	-0.017118982	-0.022467404	0.00504	-0.01454	-0.17229	-0.235247172	-0.0578474782
19	FLAG_CONT_MOBILE	0.00331947	#DIV/0!	-0.00732	0.0984	-0.001201475	-0.018813912	0.023096227	0.021396507	0.02010	-0.0131	0.15528	0.013137215	0.004408002
20	FLAG_PHONE	0.00140815	#DIV/0!	-0.00869	-0.04099	-0.029929193	0.00268842	0.026742619	0.012231373	0.0422	0.09148	0.041835	0.015782355	0.075103827
21	FLAG_EMAIL	0.00022925	#DIV/0!	0.03234	0.02941	0.023459914	0.08370512	0.017073526	0.071207216	0.01723	0.04082	-0.09079	-0.064849668	-0.035468204
22	CNT_FAM_MEMBERS	-0.0027897	#DIV/0!	0.15377	0.0056	0.078571369	0.034255958	0.064553958	0.075787127	0.06281	-0.02342	-0.28582	-0.238801269	-0.175630327
23	REGION_RATING_CLIENT	-0.0005192	#DIV/0!	-0.02228	0.00265	0.022842107	-0.186573418	-0.103336744	-0.132128435	-0.10438	-0.539	-0.02333	0.038327694	-0.07584587
24	REGION_RATING_CLIENT_W_CITY	-0.000739	#DIV/0!	-0.02112	0.00316	0.021866354	-0.020466109	-0.112237751	-0.145153055	-0.1126	-0.5373	-0.00078	0.040461022	-0.069356888
25	HOUR_APPR_PROCESS_START	-0.0001146	#DIV/0!	0.01385	-0.10158	-0.005243998	0.076742921	0.053618783	0.053588877	0.06277	0.17281	-0.09592	-0.095454898	0.008043656
26	REG_REGION_NOT_LIVE_REGION	0.00010418	#DIV/0!	-0.00022	-0.0346	-0.012342406	0.0683052	0.024616818	0.041991896	0.02673	0.00451	-0.06625	-0.037502465	-0.028699505
27	REG_REGION_NOT_WORK_REGION	0.00083585	#DIV/0!	0.04096	-0.03133	0.010856632	0.137173505	0.053735255	0.080665424	0.05492	0.06009	-0.09791	-0.111011014	-0.037870295
28	LIVE_REGION_NOT_WORK_REGION	0.00278434	#DIV/0!	0.04875	-0.1828	0.017326477	0.127701208	0.054250263	0.075486712	0.05441	0.08396	-0.07165	-0.089820206	-0.028474855
29	REG_CITY_NOT_LIVE_CITY	-0.0028847	#DIV/0!	0.00561	-0.06088	0.021586823	0.010566688	-0.025035944	-0.005674568	-0.02488	-0.04822	-0.17977	-0.092841765	-0.06421592
30	REG_CITY_NOT_WORK_CITY	-0.0016955	#DIV/0!	0.08071	-0.06117	0.072192663	-0.015702549	0.001519501	-0.01674	-0.04152	-0.2419	-0.258121321	-0.098919288	

File Home Insert Page Layout Formulas Data Review View Help

LIVE\_REGION\_NOT\_WORK\_REGION

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	S	T	V	W	X	Y	Z	AA	AB	AC			
SK_ID_CURR	TARGET	G_OWN_C_G_OWN_REL	CNT_CHILDREN	INCOME_TL_CRF_ANN_GOODS_FPOPULATION	REGISTRYS_ID_PUEAG_MOBMP_WORK_ONT_IQ_PHM_RATING_C_PRC_NOT_REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION																							
3	TARGET	#DIV/0!																											
4	FLAG_OWN_CAR	0.00121603	#DIV/0!	1																									
5	FLAG_OWN_REALTY	0.00084137	#DIV/0!	-0	1																								
6	CNT_CHILDREN	-0.0007158	#DIV/0!	0.1	-0																								
7	AMT_INCOME_TOTAL	0.00173896	#DIV/0!	0.2	0.0	0.027397	1																						
8	AMT_CREDIT	-0.000342	#DIV/0!	0.1	-0.00058	0.0342799	1	1																					
9	AMT_ANNUITY	6.823E-05	#DIV/0!	0.1	-0.00052	0.0349462	1	0.78																					
10	AMT_GOODS_PRICE	-0.0002051	#DIV/0!	0.00205	-0.00052	0.0349462	1	0.78																					
11	REGION_POPULATION_RELATIVE	0.00035999	#DIV/0!	0.001	-0.02436	0.0167851	0	0.12	0.10383																				
12	DAYS_BIRTH	0.00134617	#DIV/0!	-0.12	-0.03	0.06456	0.025	1																					
13	DAYS_EMPLOYED	0.00174396	#DIV/0!	-0.02	-0.06	0.04456	0.025	1																					
14	DAYS_REGISTRATION	0.00147511	#DIV/0!	-0.01	-0.05	0.06565	0.001	0.008	0	0.15	0	0.004	0.0015	-0.054	-0.054	-0.054	-0.054	-0.054	-0.054	-0.054	-0.054	-0.054	-0.054	-0.054	-0.054	-0.054			
15	DAYS_ID_PUBLISH	7.7019E-05	#DIV/0!	-0.002	-0.04	0.03313	0	0.008	0	0.004	0.001	0.0002	0	0.0002	0	0.0002	0	0.0002	0	0.0002	0	0.0002	0	0.0002	0	0.0002	0		
16	FLAG_MOBIL	0.00292507	#DIV/0!	-0	-0	0.00108	0.000733	0	0	0	0.00146	0.00014	0.000326	0	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004		
17	FLAG_EMP_PHONE	-0.0017175	#DIV/0!	0.2	-0.1	0.044274	0.1407427	0	0.11	0.06977	0.00709	-0.62207	-0.999756229	-0.212296555	-0.275468	-0.000897													
18	FLAG_WORK_PHONE	-0.000854	#DIV/0!	0	-0.1	0.055453	-0.0324	-0	-0	0.00504	-0.01434	-0.17229	-0.235247172	-0.0578474782	-0.046546	0.0009286	0.23541												
19	FLAG_CONT_MOBILE	0.000331947	#DIV/0!	-0	-0.01	-0.0012	-0.01881	0	0.02	0.0019	-0.0131	0.15528	0.013137215	0.004408002	0.001507	-8.14E-05	-0.01307	0.021378											
20	FLAG_PHONE	0.00140815	#DIV/0!	-0	-0	-0.02993	-0.002685	0	0.01	0.04225	0.01048	0.041835	0.015782355	0.075103827	0.036247	0.0011852	-0.01508	0.292623	-0.00665										
21	FLAG_EMAIL	0.00022925	#DIV/0!	0	0.03	0.02346	0.083705	0	0.07	0.01725	0.04082	-0.0679	-0.064849668	-0.035468204	-0.027576	0.0004617	-0.011862	-0.0052											
22	CNT_FAM_MEMBERS	-0.0027897	#DIV/0!	0.2	0.01	0.078571	0.034256	0	0.08	0.06281	-0.02342	-0.28582	-0.238301269	-0.175630327	0.02044	0.0003114	0.23785	0.067452	-0.0003										
23	REGION_RATING_CLIENT	-0.0005192	#DIV/0!	-0	-0.02	0.022842	-0.18657	-0	-0.1	-0.10438	-0.539	-0.00233	0.038327694	-0.07584587	0.008994	0.0001612	-0.0380	0.005128	0.0133										
24	REGION_RATING_CLIENT_W_CITY	-0.000739	#DIV/0!	-0	-0.02	0.021865	-0.20047	-0	-0.1	-0.1126	-0.5373	-0.00078	0.040461022	-0.069356888	0.011734	8.411E-05	-0.0402	0.01118	0.0140										
25	HOUR_APPR_PROCESS_START	-0.0001146	#DIV/0!	0	-0.1	-0.00524	0.076743	0	0.05	0.06277	0.17281	-0.09592	-0.095454898	0.008043656	0.0030398	-0.000526	0.09511	0.036978	-0.0014										
26	REG_REGION_NOT_LIVE_REGION	0.00010418	#DIV/0!	-0	-0.01234	0.06851	0	0.04	0.02673	0.00431	-0.06625	-0.037502465	-0.028699505																

The white box are present because target has only one value 0 or 1 so the correlation can't be found  
If two variables have higher correlation they are inversely proportional to each other

## POSITIVE CORRELATION

- AMT\_CREDIT and AMT\_GOODS\_PRICE
  - CNT\_CHILDRENS and CNT\_FAM\_MEMBERS
  - REG\_REGION\_NOT\_WORK\_REGION and LIVE\_REGION\_NOT\_WORK\_REGION\
  - REG\_CITY\_NOT\_WORK\_CITY and LIVE\_CITY\_NOT\_WORK\_CITY
  - AMT\_CREDIT and AMT\_ANNUITY
  - AMT\_ANNUITY and AMT\_GOODS\_PRICE
  - BIRTH\_DAYS and REGION\_POPULATION\_RELATIVE
  - OBS\_30\_CNT\_SOCIAL\_CIRCLE and OBS\_60\_CNT\_SOCIAL\_CIRCLE
  - DEF\_30\_CNT\_SOCIAL\_CIRCLE and DEF\_60\_CNT\_SOCIAL\_CIRCLE
  - AMT\_INCOME\_TOTAL and AMT\_ANNUITY

## NEGATIVE CORRELATION

- DAYS\_EMPLOYED and FLAG\_EMP\_PHONE
  - DAYS\_BIRTH and FLAG\_EMP\_PHONE
  - REGION\_POPULATION\_RELATIVE and REGION\_RATING\_CLIENT
  - REGION\_POPULATION\_RELATIVE and REGION\_RATING\_CLIENT W CITY

## **PREVIOUS APPLICATION**

- RED represents values with a correlation between -1 to 0
  - Yellow represents value with a correlation between 0 to 1
  - Green represents values with a correlation between 0 to 1

REPEATER CLIENT

## NEW CLIENT

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	SK_ID_PREV	SK_ID_CURR	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	GOODS_PRICE	PRICAPR	PROCESS_ST	APPL_PER_COS	LAST_APPL_IN_DAYS	DECISION	SELLERPLACE_AREA	CNT_PAYMENT
2	SK_ID_PREV	1											
3	SK_ID_CURR	0.000230452	1										
4	AMT_ANNUITY	0.006750727	0.000642928	1									
5	AMT_APPLICATION	0.000238201	0.002243683	0.771100716	1								
6	AMT_CREDIT	0.000305898	0.002367897	0.775470794	0.975473034	1							
7	AMT_GOODS_PRICE	0.010051614	0.002094662	0.807865559	0.945082587	0.936944	1						
8	HOUR_APPR_PROCESS_S	-0.003037857	0.002920558	-0.027745929	-0.0120328	-0.0176	-0.035291964	1					
9	FLAG_LAST_APPL_PER_C	-0.0060086399	-0.00060176	0.028750312	0.011231988	-0.02465	-0.008529997	0.007526587	1				
10	NFLAG_LAST_APPL_IN_D	-0.003703419	0.000101603	0.027923031	0.007495774	-0.02294	-0.006935398	0.004922584	0.732724698	1			
11	DAYS_DECISION	-0.015687062	0.000235815	-0.203515026	-0.09806959	-0.0895	-0.213969034	0.037634749	-0.041249191	-0.031132219	1		
12	SELLERPLACE_AREA	-0.00033393	-0.02224521	-0.01121002	-0.01355	-0.022434743	0.019365553	0.001172569	0.000562104	0.026844005	1		
13	CNT_PAYMENT	0.010522604	0.001188788	0.391005352	0.647002377	0.634267	0.647095105	-0.047924331	0.099303247	0.071927305	-0.194395058	-0.016297614	1
14													
15													
16													
17													
18													
19													
20													
21													
22													
23													
24													
25													
26													
27													

## POSITIVE CORRELATION

- AMT\_APPLICATION and AMT\_GOODS\_PRICE
- AMT\_APPLICATION and AMT\_CREDIT
- AMT\_CREDIT and AMT\_GOODS\_PRICE
- AMT\_ANNUITY and AMT\_GOODS\_PRICE
- AMT\_ANNUITY and AMT\_APPLICATION
- AMT\_ANNUITY and AMT\_CREDIT
- AMT\_GOODS\_PRICE and CNT\_PAYMENT
- CNT\_PAYMENT and AMT\_CREDIT
- CNT\_PAYMENT and AMT\_APPLICATION

## NEGATIVE CORRELATION

- CNT\_PAYMENT and AMT\_APPLICATION
- AMT\_GOODS\_PRICE and DAYS\_DECISION
- DAYS\_DECISION and CNT\_PAYMENT
- DAYS\_DECISION and AMT\_ANNUITY