BANK LOAN ANALYSIS

Application data hyperlink

Previous_ data hyperlink

Project Description:

- Imagine you're a data analyst at a finance company that specializes in lending various types of loans to urban customers. Your company faces a challenge: some customers who don't have a sufficient credit history take advantage of this and default on their loans.
- The main aim of this project is to identify patterns that indicate if a customer will have difficulty paying their installments. This information can be used to make decisions such as denying the loan, reducing the amount of loan, or lending at a higher interest rate to risky applicants.
- The company wants to understand the key factors behind loan default so it can make better decisions about loan approval.

TECH STACK USED:

- I've used Microsoft Power point version 2309 to create this presentation.
- I've used Microsoft Excel version 2309 to implement the task assigned.
- I chose Microsoft Excel because it is the most convenient spreadsheet and can be used efficiently to view statistics and analyze the data set given very quickly.

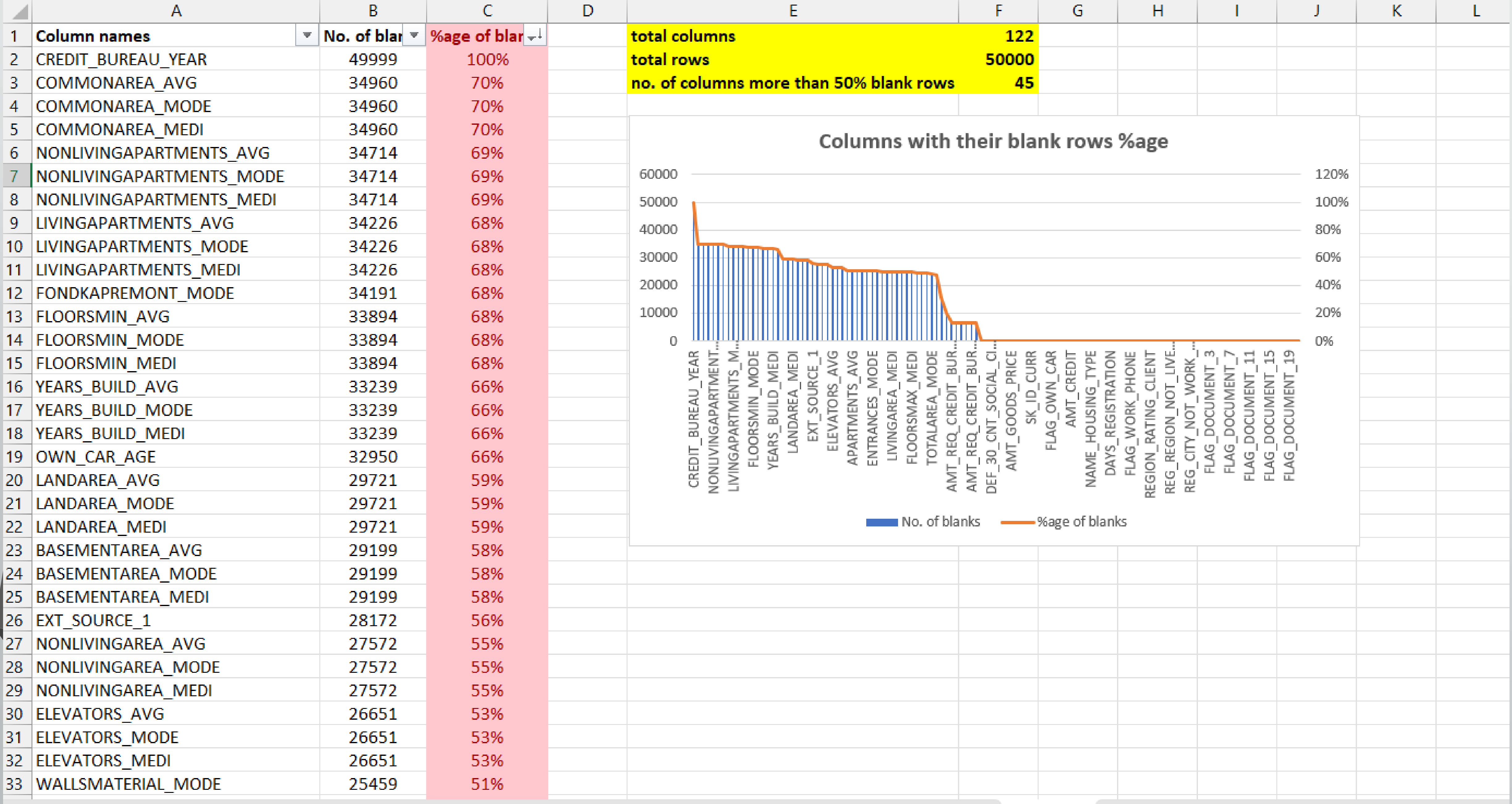
INSIGHTS

Files Description:

- **previous_application.csv:** Contains information about previous loan applications.
- **application_data.csv:** Provides details about the current loan applications.
- **columns_description.csv:** Describes the columns present in the other datasets, explaining what each column represents.

Task A:Identify Missing Data and Deal with it Appropriately

- **Description:** Identify the missing data in the dataset and decide on an appropriate method to deal with it using Excel built-in functions and features.
- To implement this task I've used excel functions to calculate blanks in the data in both the files which are application_data and previous_application .
- In application_data file there are total 122 columns and 50000 rows.
- From which 45 columns had blanks more than 50% , so it get deleted.
- I've used bar graph to visualize the blanks %age in columns.
- The details and the screenshot of the excel sheet is in the next slide



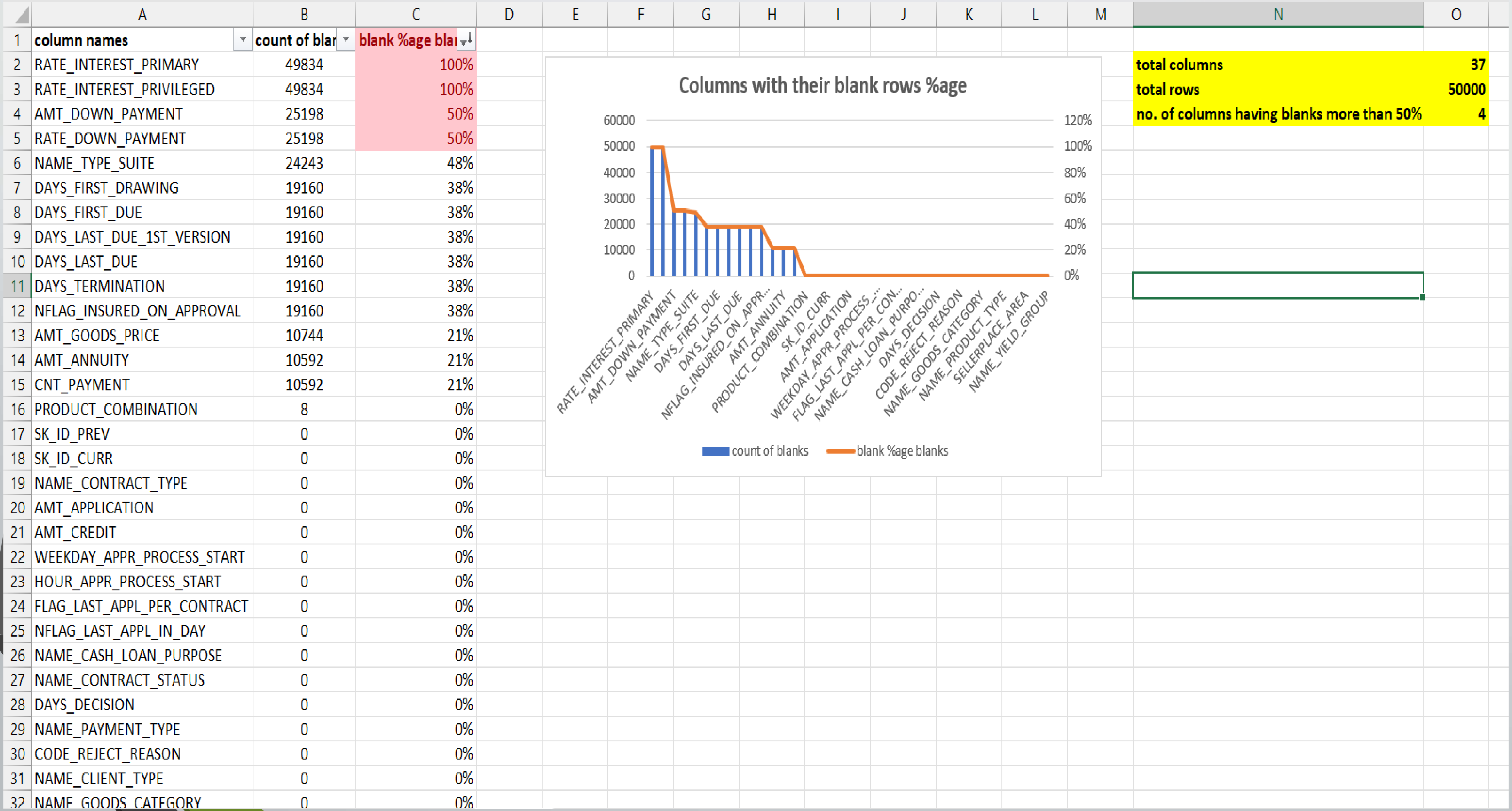
	A	B	C
1	**Column names**	**No. of blar**	**%age of blar**
2	CREDIT_BUREAU_YEAR	49999	100%
3	COMMONAREA_AVG	34960	70%
4	COMMONAREA_MODE	34960	70%
5	COMMONAREA_MEDI	34960	70%
6	NONLIVINGAPARTMENTS_AVG	34714	69%
7	NONLIVINGAPARTMENTS_MODE	34714	69%
8	NONLIVINGAPARTMENTS_MEDI	34714	69%
9	LIVINGAPARTMENTS_AVG	34226	68%
10	LIVINGAPARTMENTS_MODE	34226	68%
11	LIVINGAPARTMENTS_MEDI	34226	68%
12	FONDKAPREMONT_MODE	34191	68%
13	FLOORSMIN_AVG	33894	68%
14	FLOORSMIN_MODE	33894	68%
15	FLOORSMIN_MEDI	33894	68%
16	YEARS_BUILD_AVG	33239	66%
17	YEARS_BUILD_MODE	33239	66%
18	YEARS_BUILD_MEDI	33239	66%
19	OWN_CAR_AGE	32950	66%
20	LANDAREA_AVG	29721	59%
21	LANDAREA_MODE	29721	59%
22	LANDAREA_MEDI	29721	59%
23	BASEMENTAREA_AVG	29199	58%
24	BASEMENTAREA_MODE	29199	58%
25	BASEMENTAREA_MEDI	29199	58%
26	EXT_SOURCE_1	28172	56%
27	NONLIVINGAREA_AVG	27572	55%
28	NONLIVINGAREA_MODE	27572	55%
29	NONLIVINGAREA_MEDI	27572	55%
30	ELEVATORS_AVG	26651	53%
31	ELEVATORS_MODE	26651	53%
32	ELEVATORS_MEDI	26651	53%
33	WALLSMATERIAL_MODE	25459	51%

	E	F
1	**total columns**	**122**
2	**total rows**	**50000**
3	**no. of columns more than 50% blank rows**	**45**

Columns with their blank rows %age

Task A: Identify Missing Data and Deal with it Appropriately

- In previous_application file there are total 37 columns and 50000 rows.
- From which 4 columns had blanks more than 50% , so it get deleted.
- I've used bar graph to visualize the blanks %age in columns.
- The details and the screenshot of the excel sheet is in the next slide
- There is a sheet called CLEANED DATA which has only useful data in it.



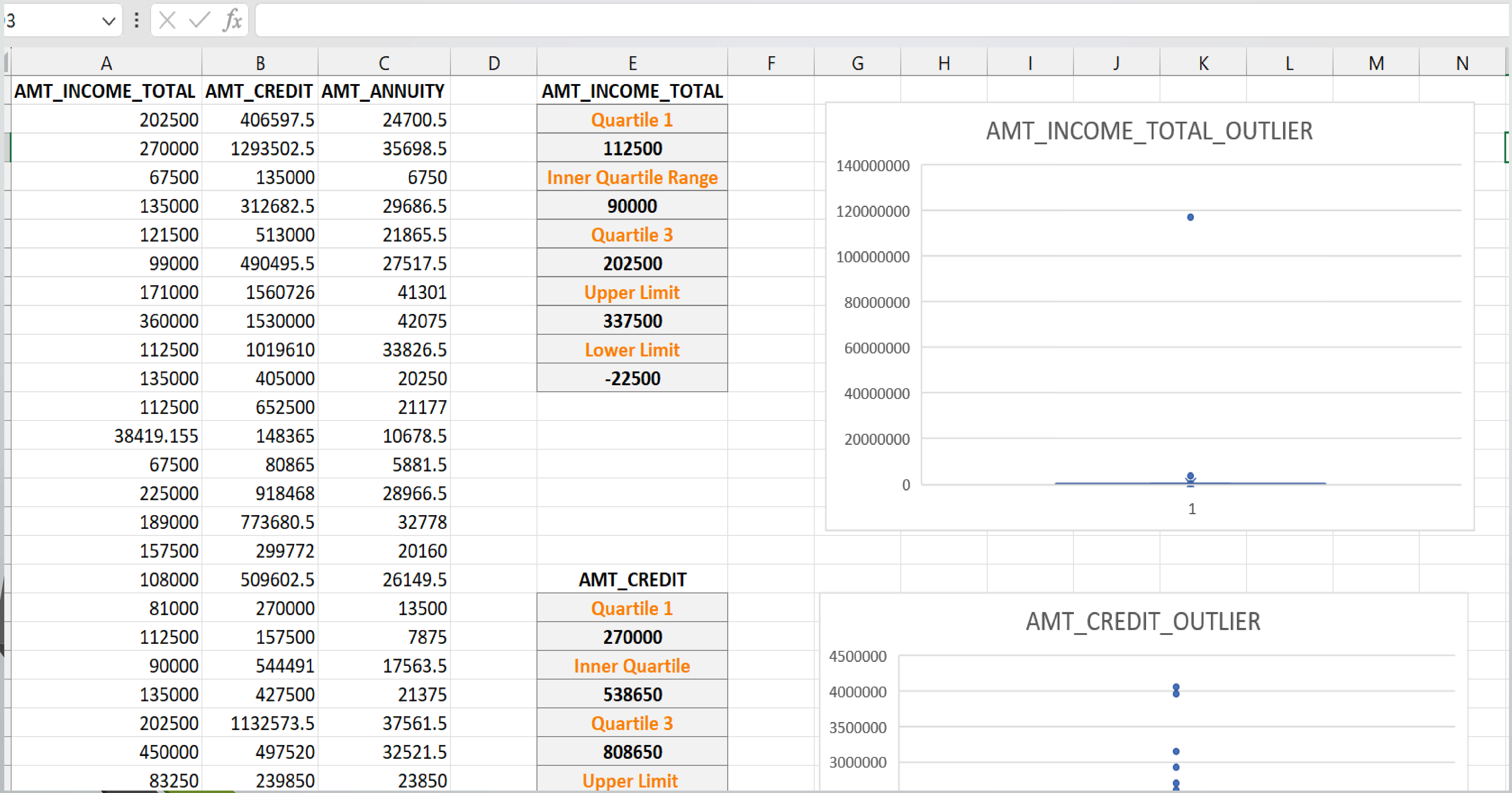
	A	B	C
1	column names	count of blar	blank %age blar
2	RATE_INTEREST_PRIMARY	49834	100%
3	RATE_INTEREST_PRIVILEGED	49834	100%
4	AMT_DOWN_PAYMENT	25198	50%
5	RATE_DOWN_PAYMENT	25198	50%
6	NAME_TYPE_SUITE	24243	48%
7	DAYS_FIRST_DRAWING	19160	38%
8	DAYS_FIRST_DUE	19160	38%
9	DAYS_LAST_DUE_1ST_VERSION	19160	38%
10	DAYS_LAST_DUE	19160	38%
11	DAYS_TERMINATION	19160	38%
12	NFLAG_INSURED_ON_APPROVAL	19160	38%
13	AMT_GOODS_PRICE	10744	21%
14	AMT_ANNUITY	10592	21%
15	CNT_PAYMENT	10592	21%
16	PRODUCT_COMBINATION	8	0%
17	SK_ID_PREV	0	0%
18	SK_ID_CURR	0	0%
19	NAME_CONTRACT_TYPE	0	0%
20	AMT_APPLICATION	0	0%
21	AMT_CREDIT	0	0%
22	WEEKDAY_APPR_PROCESS_START	0	0%
23	HOUR_APPR_PROCESS_START	0	0%
24	FLAG_LAST_APPL_PER_CONTRACT	0	0%
25	NFLAG_LAST_APPL_IN_DAY	0	0%
26	NAME_CASH_LOAN_PURPOSE	0	0%
27	NAME_CONTRACT_STATUS	0	0%
28	DAYS_DECISION	0	0%
29	NAME_PAYMENT_TYPE	0	0%
30	CODE_REJECT_REASON	0	0%
31	NAME_CLIENT_TYPE	0	0%
32	NAME_GOODS_CATEGORY	0	0%

Columns with their blank rows %age

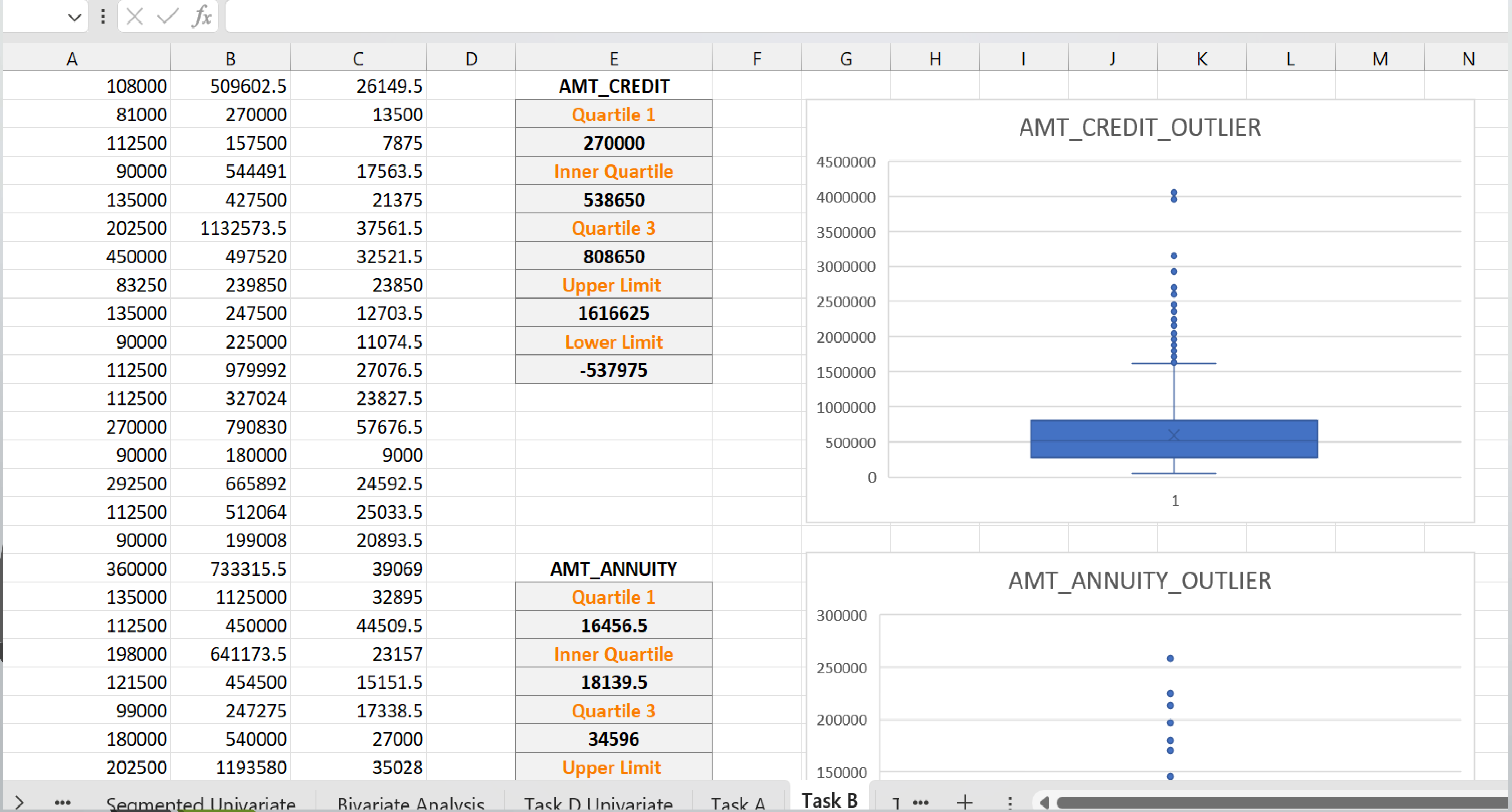
total columns	37
total rows	50000
no. of columns having blanks more than 50%	4

Task B: Identify Outliers in the Dataset

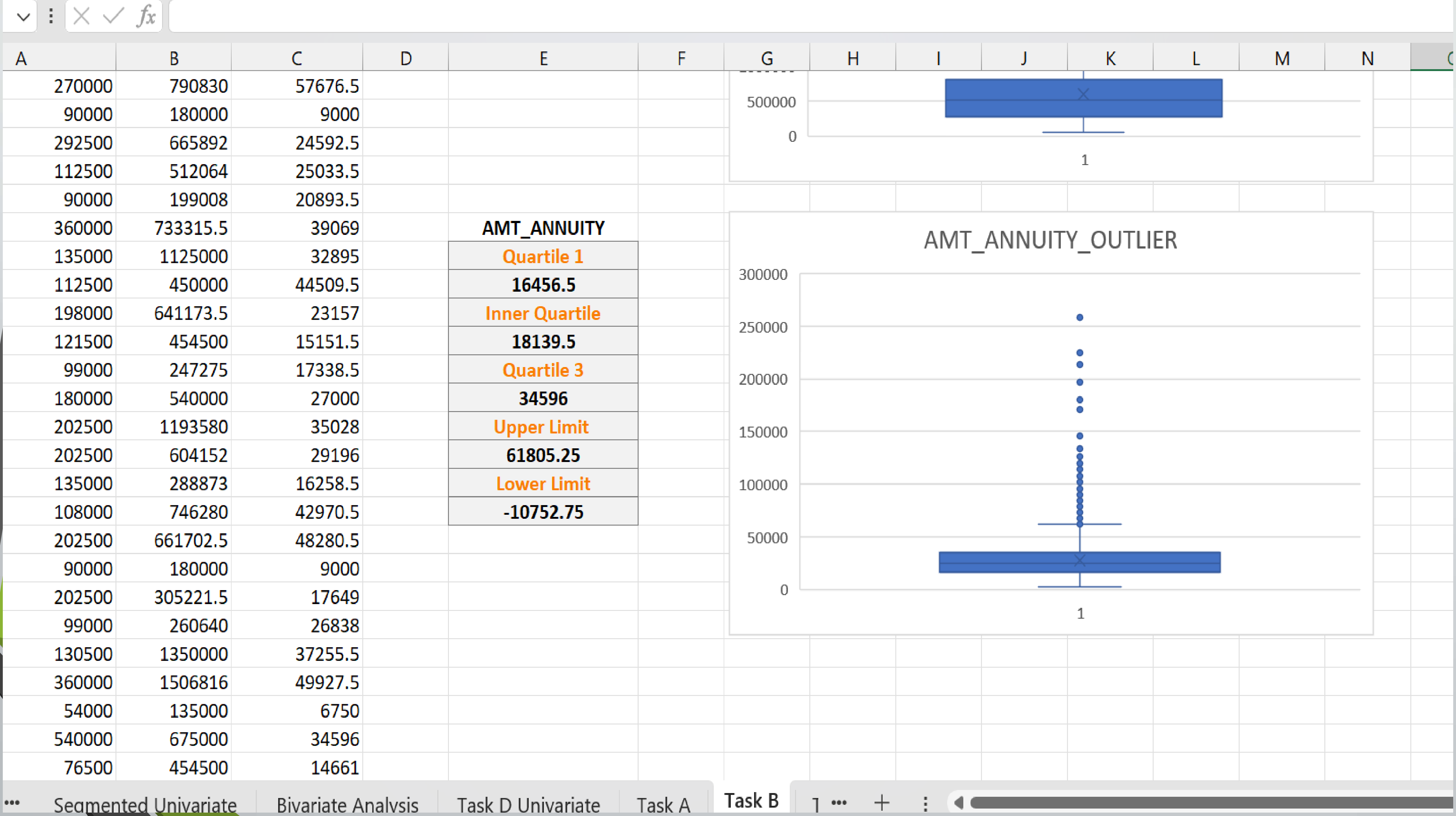
- **Description :** Detect and identify outliers in the dataset using Excel statistical functions and features, focusing on numerical variables.
- To identify outliers I’ve used Quartile function of excel and found IQR by subtracting Q3 with Q1 .
- I’ve plotted a Box and Whisker graph to visually identify the outliers easily.



A	B	C	D	E
AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY		AMT_INCOME_TOTAL
202500	406597.5	24700.5		Quartile 1
270000	1293502.5	35698.5		112500
67500	135000	6750		Inner Quartile Range
135000	312682.5	29686.5		90000
121500	513000	21865.5		Quartile 3
99000	490495.5	27517.5		202500
171000	1560726	41301		Upper Limit
360000	1530000	42075		337500
112500	1019610	33826.5		Lower Limit
135000	405000	20250		-22500
112500	652500	21177		
38419.155	148365	10678.5		
67500	80865	5881.5		
225000	918468	28966.5		
189000	773680.5	32778		
157500	299772	20160		
108000	509602.5	26149.5		AMT_CREDIT
81000	270000	13500		Quartile 1
112500	157500	7875		270000
90000	544491	17563.5		Inner Quartile
135000	427500	21375		538650
202500	1132573.5	37561.5		Quartile 3
450000	497520	32521.5		808650
83250	239850	23850		Upper Limit



A	B	C	D	E
108000	509602.5	26149.5		AMT_CREDIT
81000	270000	13500		Quartile 1
112500	157500	7875		270000
90000	544491	17563.5		Inner Quartile
135000	427500	21375		538650
202500	1132573.5	37561.5		Quartile 3
450000	497520	32521.5		808650
83250	239850	23850		Upper Limit
135000	247500	12703.5		1616625
90000	225000	11074.5		Lower Limit
112500	979992	27076.5		-537975
112500	327024	23827.5		
270000	790830	57676.5		
90000	180000	9000		
292500	665892	24592.5		
112500	512064	25033.5		
90000	199008	20893.5		
360000	733315.5	39069		AMT_ANNUITY
135000	1125000	32895		Quartile 1
112500	450000	44509.5		16456.5
198000	641173.5	23157		Inner Quartile
121500	454500	15151.5		18139.5
99000	247275	17338.5		Quartile 3
180000	540000	27000		34596
202500	1193580	35028		Upper Limit

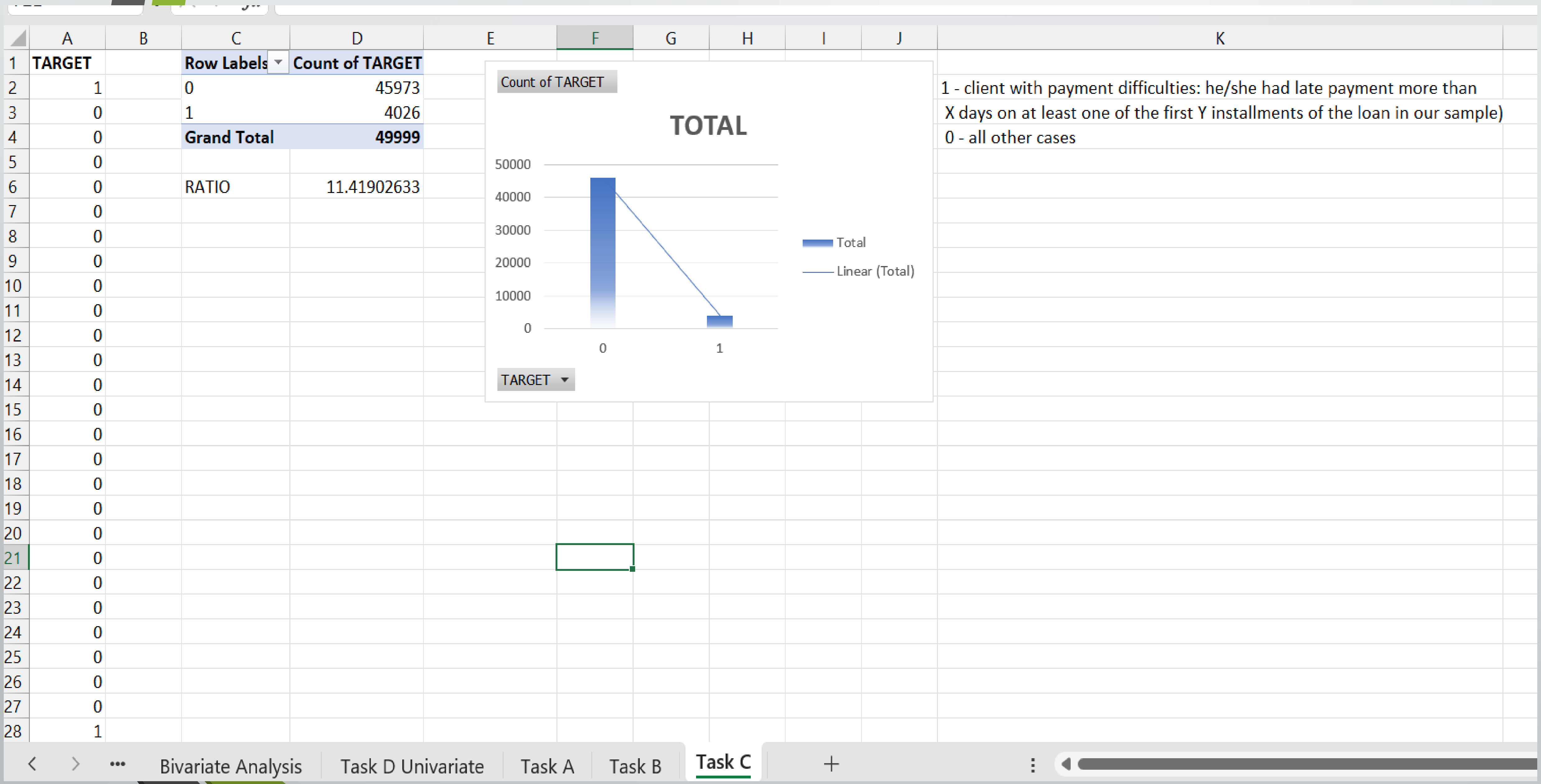


270000	790830	57676.5
90000	180000	9000
292500	665892	24592.5
112500	512064	25033.5
90000	199008	20893.5
360000	733315.5	39069
135000	1125000	32895
112500	450000	44509.5
198000	641173.5	23157
121500	454500	15151.5
99000	247275	17338.5
180000	540000	27000
202500	1193580	35028
202500	604152	29196
135000	288873	16258.5
108000	746280	42970.5
202500	661702.5	48280.5
90000	180000	9000
202500	305221.5	17649
99000	260640	26838
130500	1350000	37255.5
360000	1506816	49927.5
54000	135000	6750
540000	675000	34596
76500	454500	14661

AMT_ANNUITY
Quartile 1
16456.5
Inner Quartile
18139.5
Quartile 3
34596
Upper Limit
61805.25
Lower Limit
-10752.75

Task C: Analyze Data Imbalance:

- **Description** : Determine if there is data imbalance in the loan application dataset and calculate the ratio of data imbalance using Excel functions.
- Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models.
- To visually see the difference of data imbalance I've created a pivot bar chart that shows the data imbalance of the Target column.
- I've used pivot table to count the no. of clients having payment difficulties (1) and clients with all other cases(0).



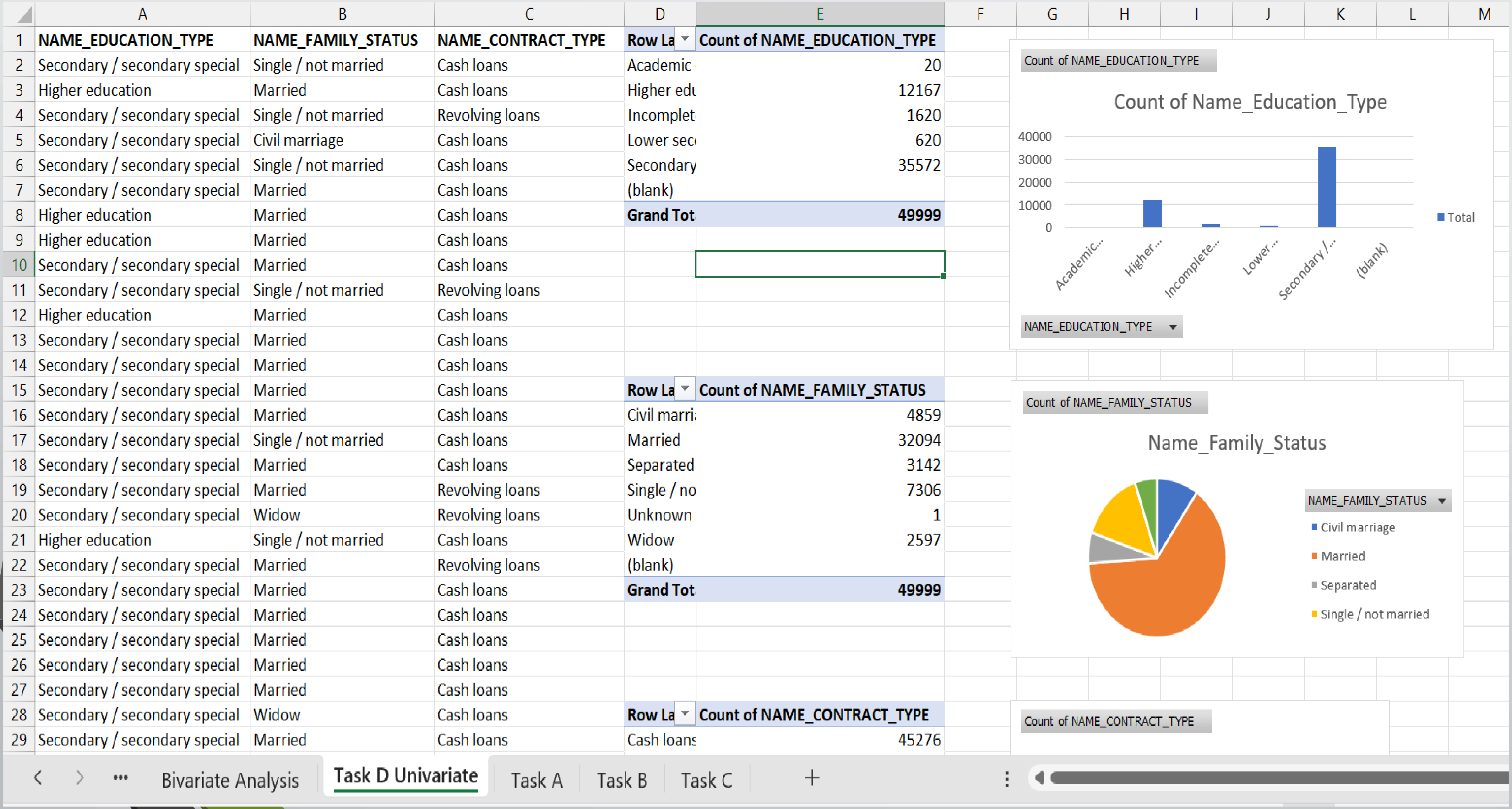
	A	B	C	D	E	F	G	H	I	J	K
1	TARGET		Row Labels	Count of TARGET							
2	1		0	45973							1 - client with payment difficulties: he/she had late payment more than
3	0		1	4026							X days on at least one of the first Y installments of the loan in our sample)
4	0		**Grand Total**	**49999**							0 - all other cases
5	0										
6	0		RATIO	11.41902633							
7	0										
8	0										
9	0										
10	0										
11	0										
12	0										
13	0										
14	0										
15	0										
16	0										
17	0										
18	0										
19	0										
20	0										
21	0										
22	0										
23	0										
24	0										
25	0										
26	0										
27	0										
28	1										

Task D: Perform Univariate, Segmented Univariate, and Bivariate Analysis

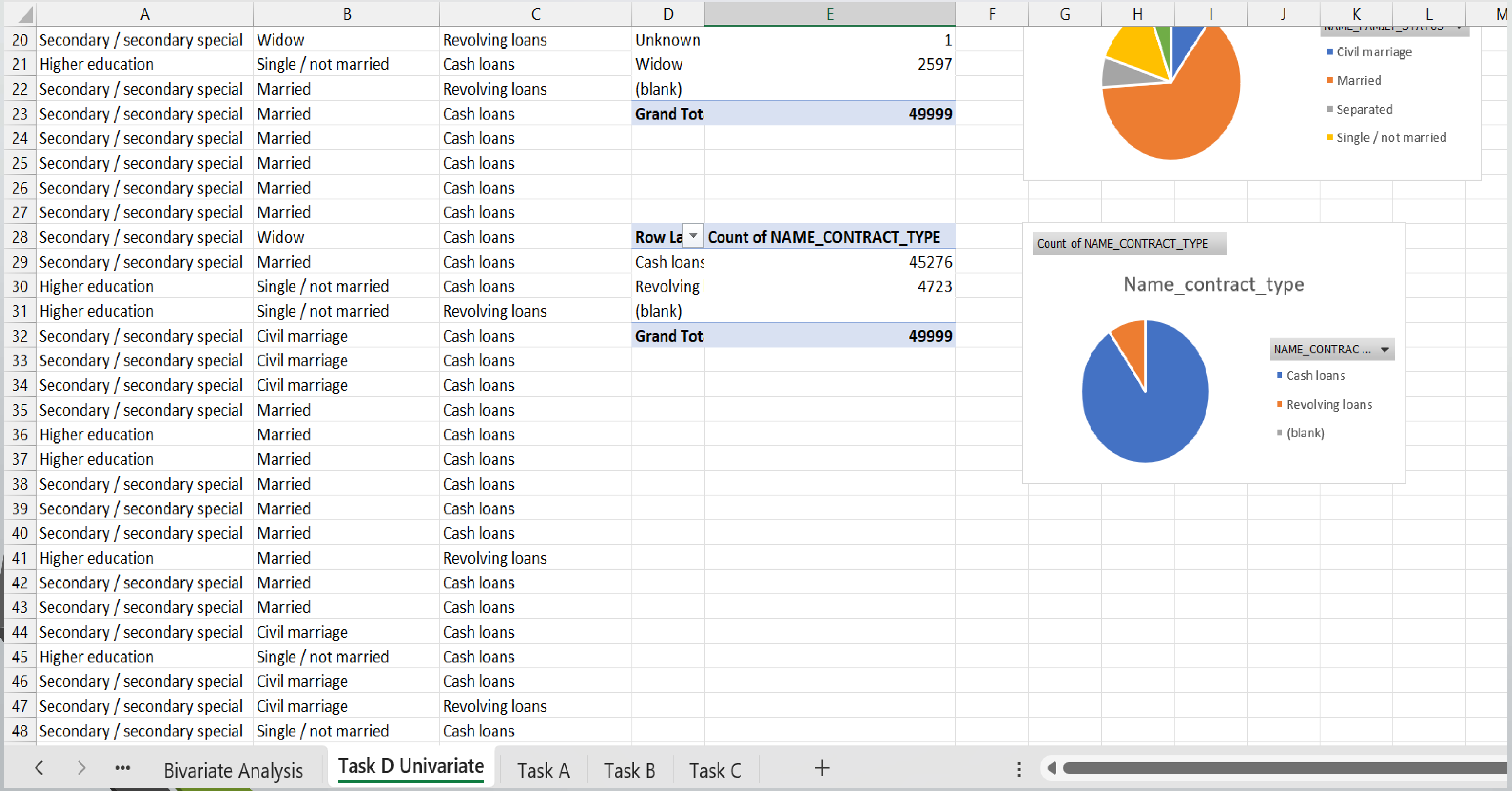
- **Description**: Perform univariate analysis to understand the distribution of individual variables, segmented univariate analysis to compare variable distributions for different scenarios, and bivariate analysis to explore relationships between variables and the target variable using Excel functions and features.
- I’ve performed all described analysis which are shown in the next slides.

Univariate Analysis:

- As the name suggests, Univariate analysis explores one variable in a data set, separately.
- Next are examples of three univariate analysis performed in the working file of our data set.
- I’ve performed univariate analysis on three columns using pivot table.
- I’ve also added pivot charts to understand the analysis better.



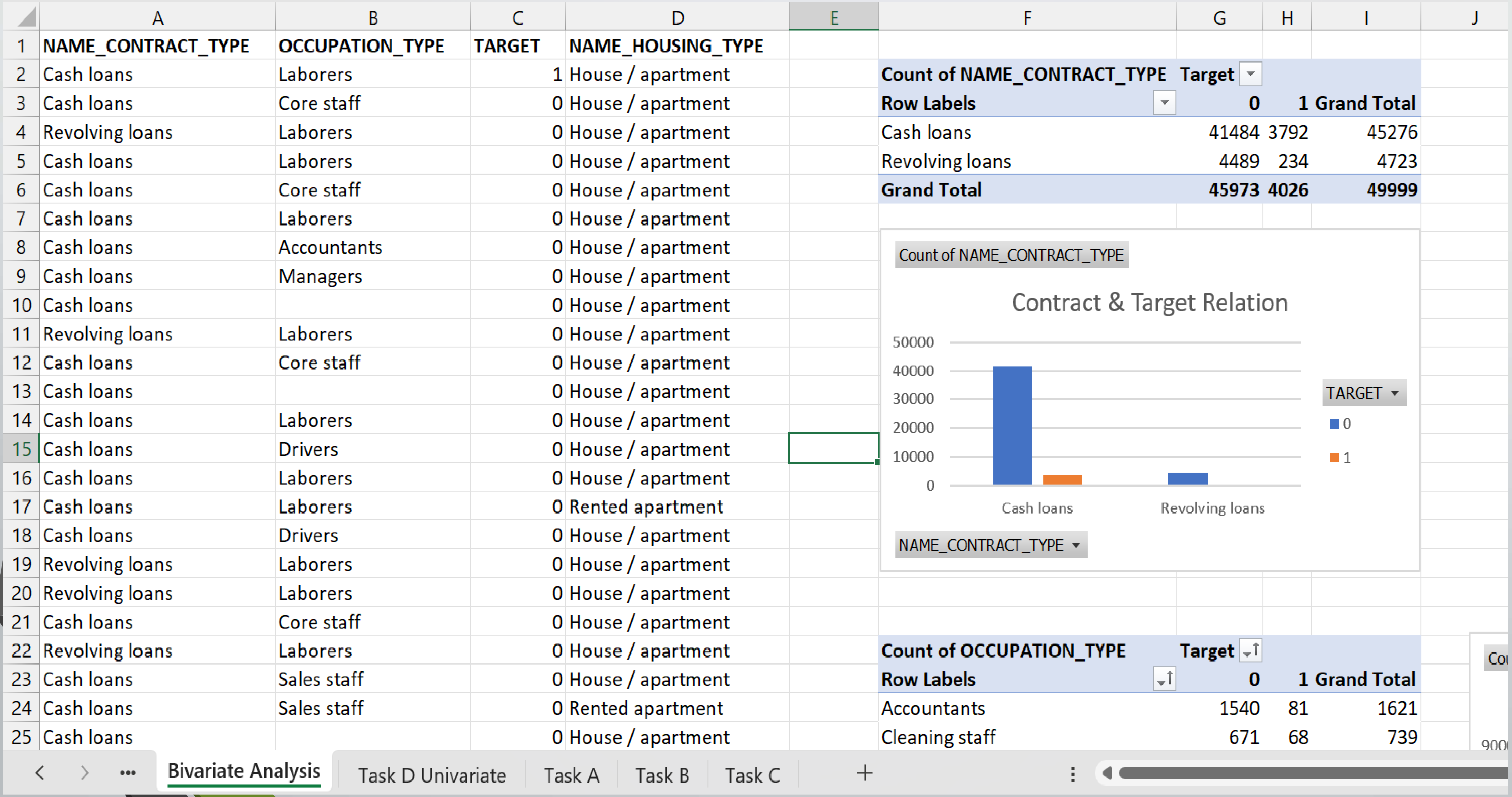
	A	B	C	D	E
1	NAME_EDUCATION_TYPE	NAME_FAMILY_STATUS	NAME_CONTRACT_TYPE	Row La	Count of NAME_EDUCATION_TYPE
2	Secondary / secondary special	Single / not married	Cash loans	Academic	20
3	Higher education	Married	Cash loans	Higher edu	12167
4	Secondary / secondary special	Single / not married	Revolving loans	Incomplet	1620
5	Secondary / secondary special	Civil marriage	Cash loans	Lower sec	620
6	Secondary / secondary special	Single / not married	Cash loans	Secondary	35572
7	Secondary / secondary special	Married	Cash loans	(blank)	
8	Higher education	Married	Cash loans	Grand Tot	49999
9	Higher education	Married	Cash loans		
10	Secondary / secondary special	Married	Cash loans		
11	Secondary / secondary special	Single / not married	Revolving loans		
12	Higher education	Married	Cash loans		
13	Secondary / secondary special	Married	Cash loans		
14	Secondary / secondary special	Married	Cash loans		
15	Secondary / secondary special	Married	Cash loans	Row La	Count of NAME_FAMILY_STATUS
16	Secondary / secondary special	Married	Cash loans	Civil marri	4859
17	Secondary / secondary special	Single / not married	Cash loans	Married	32094
18	Secondary / secondary special	Married	Cash loans	Separated	3142
19	Secondary / secondary special	Married	Revolving loans	Single / no	7306
20	Secondary / secondary special	Widow	Revolving loans	Unknown	1
21	Higher education	Single / not married	Cash loans	Widow	2597
22	Secondary / secondary special	Married	Revolving loans	(blank)	
23	Secondary / secondary special	Married	Cash loans	Grand Tot	49999
24	Secondary / secondary special	Married	Cash loans		
25	Secondary / secondary special	Married	Cash loans		
26	Secondary / secondary special	Married	Cash loans		
27	Secondary / secondary special	Married	Cash loans		
28	Secondary / secondary special	Widow	Cash loans	Row La	Count of NAME_CONTRACT_TYPE
29	Secondary / secondary special	Married	Cash loans	Cash loans	45276



	A	B	C	D	E
20	Secondary / secondary special	Widow	Revolving loans	Unknown	1
21	Higher education	Single / not married	Cash loans	Widow	2597
22	Secondary / secondary special	Married	Revolving loans	(blank)	
23	Secondary / secondary special	Married	Cash loans	**Grand Tot**	**49999**
24	Secondary / secondary special	Married	Cash loans		
25	Secondary / secondary special	Married	Cash loans		
26	Secondary / secondary special	Married	Cash loans		
27	Secondary / secondary special	Married	Cash loans		
28	Secondary / secondary special	Widow	Cash loans	**Row La**	**Count of NAME_CONTRACT_TYPE**
29	Secondary / secondary special	Married	Cash loans	Cash loans	45276
30	Higher education	Single / not married	Cash loans	Revolving	4723
31	Higher education	Single / not married	Revolving loans	(blank)	
32	Secondary / secondary special	Civil marriage	Cash loans	**Grand Tot**	**49999**
33	Secondary / secondary special	Civil marriage	Cash loans		
34	Secondary / secondary special	Civil marriage	Cash loans		
35	Secondary / secondary special	Married	Cash loans		
36	Higher education	Married	Cash loans		
37	Higher education	Married	Cash loans		
38	Secondary / secondary special	Married	Cash loans		
39	Secondary / secondary special	Married	Cash loans		
40	Secondary / secondary special	Married	Cash loans		
41	Higher education	Married	Revolving loans		
42	Secondary / secondary special	Married	Cash loans		
43	Secondary / secondary special	Married	Cash loans		
44	Secondary / secondary special	Civil marriage	Cash loans		
45	Higher education	Single / not married	Cash loans		
46	Secondary / secondary special	Civil marriage	Cash loans		
47	Secondary / secondary special	Civil marriage	Revolving loans		
48	Secondary / secondary special	Single / not married	Cash loans		

Bivariate Analysis:

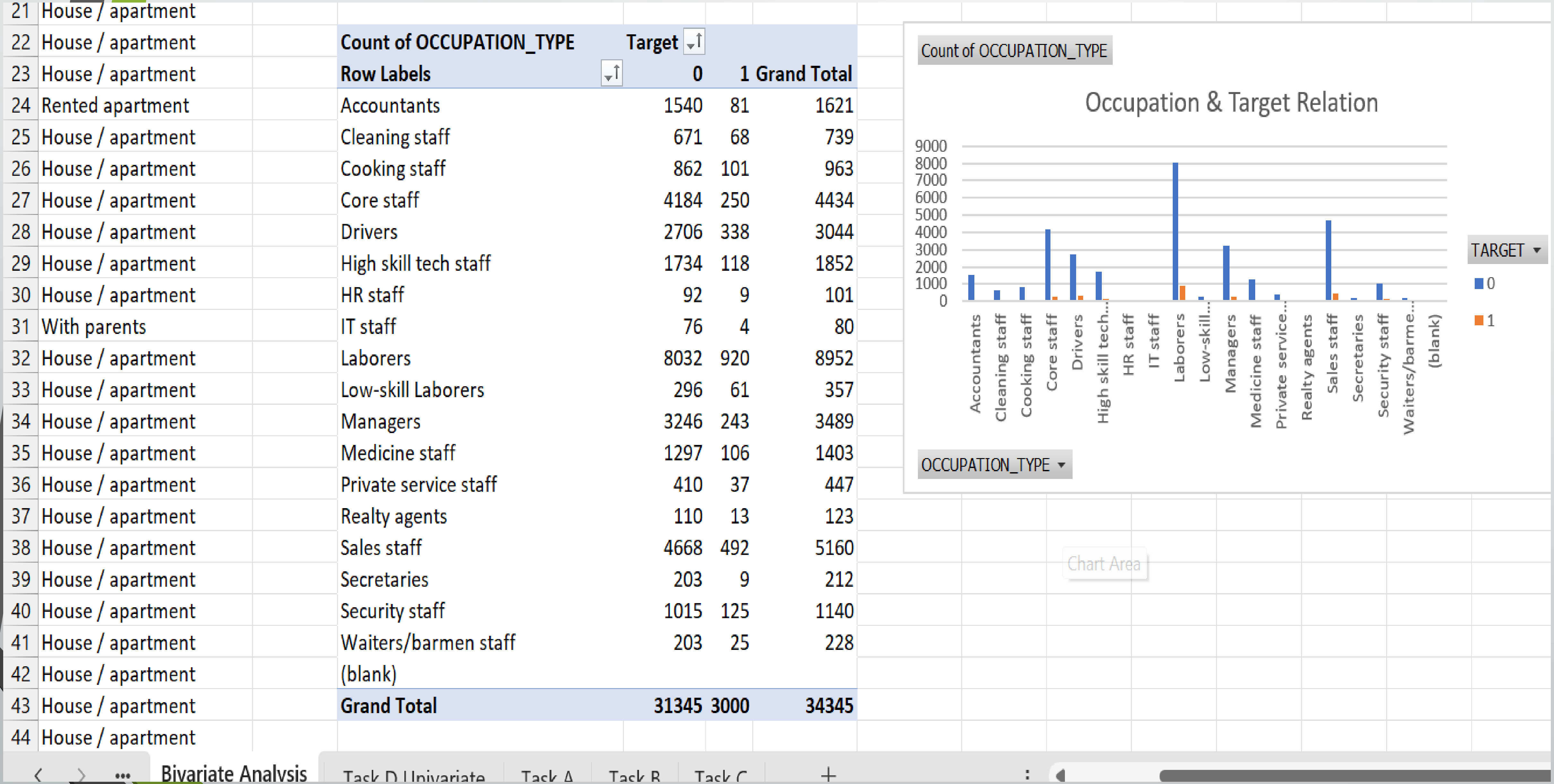
- Bivariate analysis is stated to be an analysis of any concurrent relation between two variables or attributes.
- Next are the examples of three bivariate analysis performed in the working file of our data set.
- I've performed bivariate analysis on three columns using pivot table.
- I've also added pivot charts to understand the analysis better.



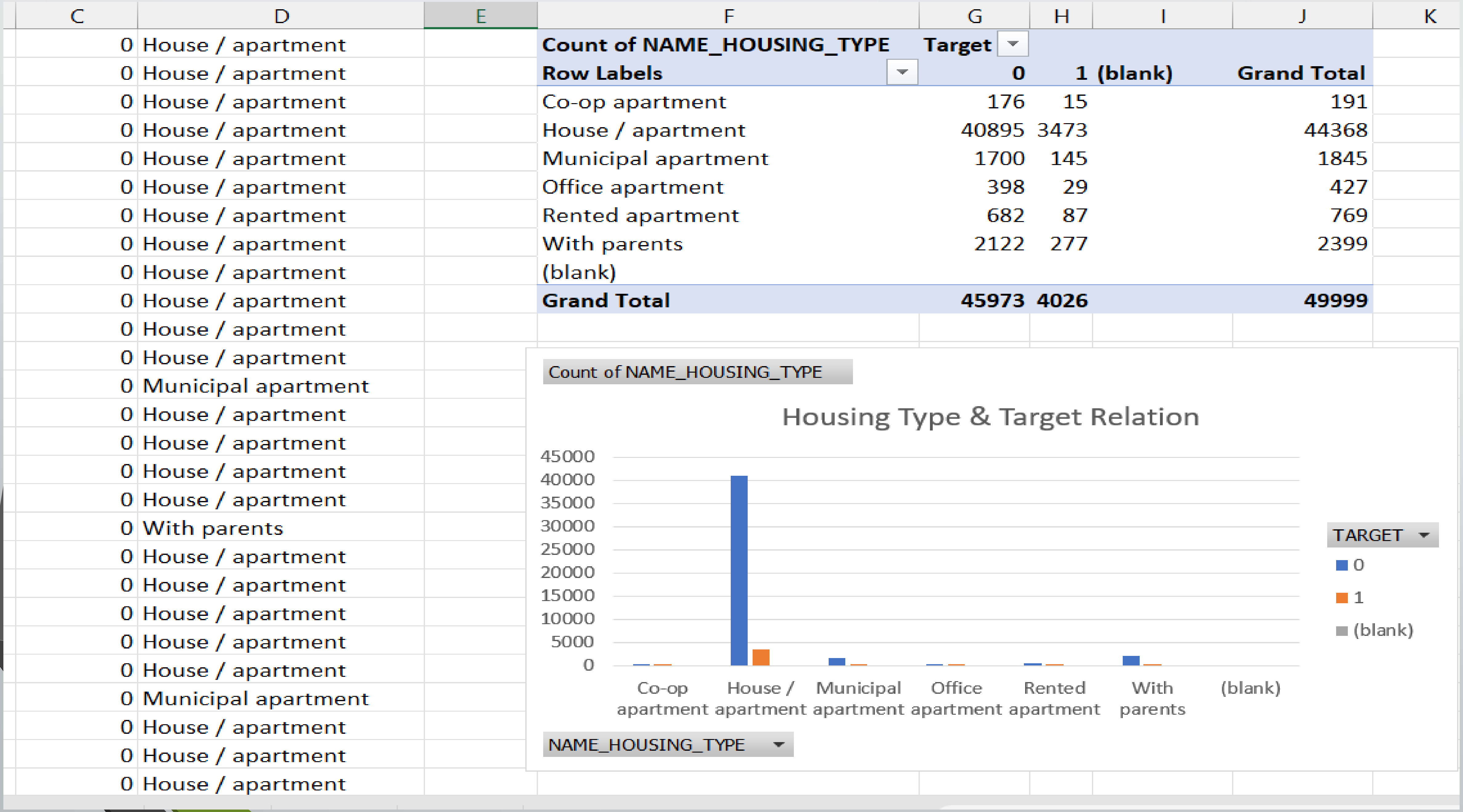
	A	B	C	D
1	NAME_CONTRACT_TYPE	OCCUPATION_TYPE	TARGET	NAME_HOUSING_TYPE
2	Cash loans	Laborers	1	House / apartment
3	Cash loans	Core staff	0	House / apartment
4	Revolving loans	Laborers	0	House / apartment
5	Cash loans	Laborers	0	House / apartment
6	Cash loans	Core staff	0	House / apartment
7	Cash loans	Laborers	0	House / apartment
8	Cash loans	Accountants	0	House / apartment
9	Cash loans	Managers	0	House / apartment
10	Cash loans		0	House / apartment
11	Revolving loans	Laborers	0	House / apartment
12	Cash loans	Core staff	0	House / apartment
13	Cash loans		0	House / apartment
14	Cash loans	Laborers	0	House / apartment
15	Cash loans	Drivers	0	House / apartment
16	Cash loans	Laborers	0	House / apartment
17	Cash loans	Laborers	0	Rented apartment
18	Cash loans	Drivers	0	House / apartment
19	Revolving loans	Laborers	0	House / apartment
20	Revolving loans	Laborers	0	House / apartment
21	Cash loans	Core staff	0	House / apartment
22	Revolving loans	Laborers	0	House / apartment
23	Cash loans	Sales staff	0	House / apartment
24	Cash loans	Sales staff	0	Rented apartment
25	Cash loans		0	House / apartment

Count of NAME_CONTRACT_TYPE	Target		
Row Labels	0	1	Grand Total
Cash loans	41484	3792	45276
Revolving loans	4489	234	4723
Grand Total	**45973**	**4026**	**49999**

Count of OCCUPATION_TYPE	Target		
Row Labels	0	1	Grand Total
Accountants	1540	81	1621
Cleaning staff	671	68	739



21	House / apartment				
22	House / apartment	**Count of OCCUPATION_TYPE**	**Target**		
23	House / apartment	**Row Labels**	**0**	**1**	**Grand Total**
24	Rented apartment	Accountants	1540	81	1621
25	House / apartment	Cleaning staff	671	68	739
26	House / apartment	Cooking staff	862	101	963
27	House / apartment	Core staff	4184	250	4434
28	House / apartment	Drivers	2706	338	3044
29	House / apartment	High skill tech staff	1734	118	1852
30	House / apartment	HR staff	92	9	101
31	With parents	IT staff	76	4	80
32	House / apartment	Laborers	8032	920	8952
33	House / apartment	Low-skill Laborers	296	61	357
34	House / apartment	Managers	3246	243	3489
35	House / apartment	Medicine staff	1297	106	1403
36	House / apartment	Private service staff	410	37	447
37	House / apartment	Realty agents	110	13	123
38	House / apartment	Sales staff	4668	492	5160
39	House / apartment	Secretaries	203	9	212
40	House / apartment	Security staff	1015	125	1140
41	House / apartment	Waiters/barmen staff	203	25	228
42	House / apartment	(blank)			
43	House / apartment	**Grand Total**	**31345**	**3000**	**34345**
44	House / apartment				

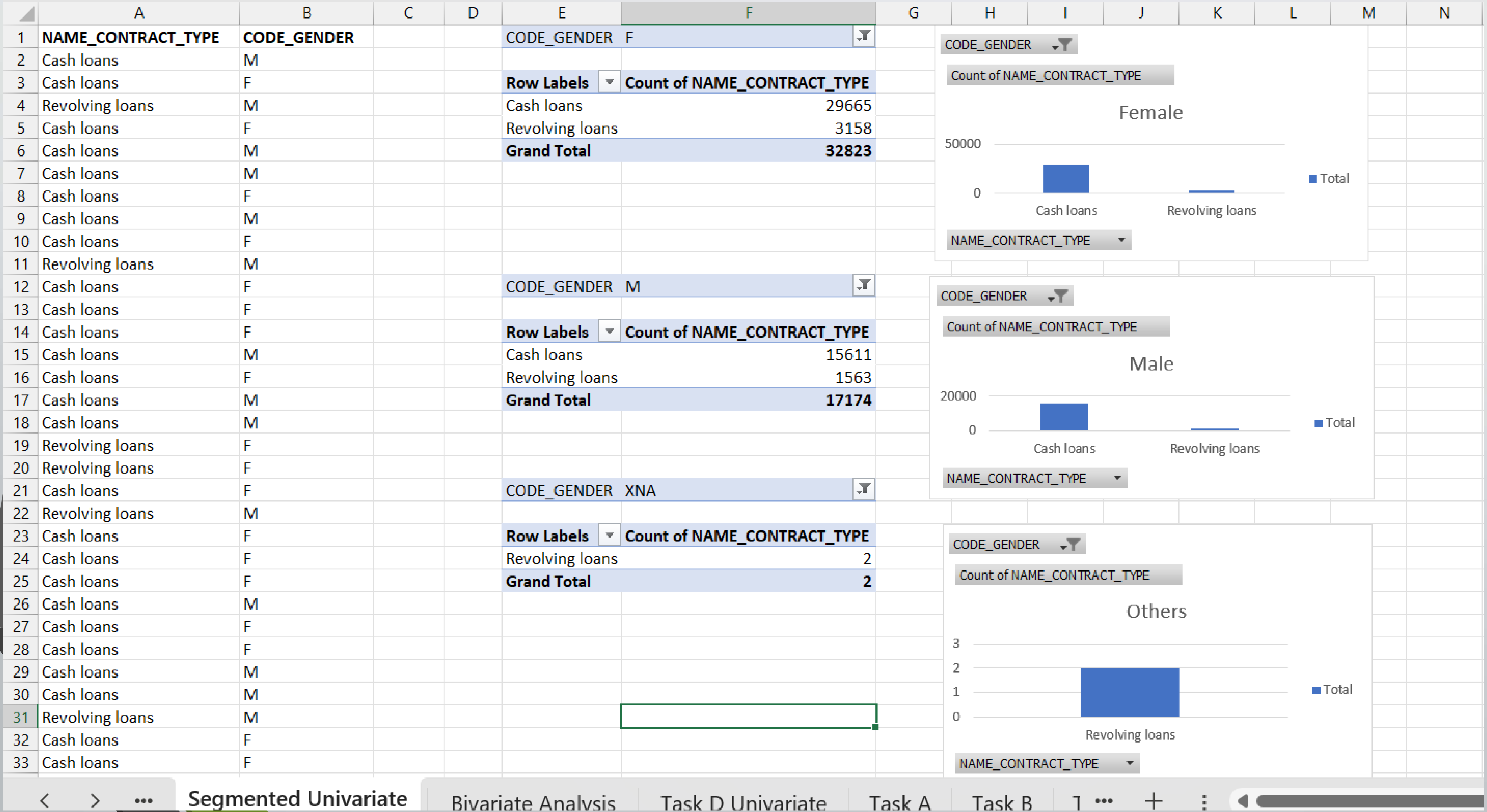


C	D
0	House / apartment
0	House / apartment
0	House / apartment
0	House / apartment
0	House / apartment
0	House / apartment
0	House / apartment
0	House / apartment
0	House / apartment
0	House / apartment
0	House / apartment
0	House / apartment
0	Municipal apartment
0	House / apartment
0	House / apartment
0	House / apartment
0	House / apartment
0	With parents
0	House / apartment
0	House / apartment
0	House / apartment
0	House / apartment
0	House / apartment
0	Municipal apartment
0	House / apartment
0	House / apartment
0	House / apartment

Count of NAME_HOUSING_TYPE	**Target**			
Row Labels	**0**	**1**	**(blank)**	**Grand Total**
Co-op apartment	176	15		191
House / apartment	40895	3473		44368
Municipal apartment	1700	145		1845
Office apartment	398	29		427
Rented apartment	682	87		769
With parents	2122	277		2399
(blank)				
Grand Total	**45973**	**4026**		**49999**

Segmented Univariate Analysis:

- Segmented Univariate analysis is one of the simplest form of visualization to analyze data.
- Next is an example of segmented univariate analysis performed in the working file of our data set.
- I’ve performed segmented univariate analysis on columns using pivot table.
- I’ve also added pivot charts to understand the analysis better.



	A	B
1	NAME_CONTRACT_TYPE	CODE_GENDER
2	Cash loans	M
3	Cash loans	F
4	Revolving loans	M
5	Cash loans	F
6	Cash loans	M
7	Cash loans	M
8	Cash loans	F
9	Cash loans	M
10	Cash loans	F
11	Revolving loans	M
12	Cash loans	F
13	Cash loans	F
14	Cash loans	F
15	Cash loans	M
16	Cash loans	F
17	Cash loans	M
18	Cash loans	M
19	Revolving loans	F
20	Revolving loans	F
21	Cash loans	F
22	Revolving loans	M
23	Cash loans	F
24	Cash loans	F
25	Cash loans	F
26	Cash loans	M
27	Cash loans	F
28	Cash loans	F
29	Cash loans	M
30	Cash loans	M
31	Revolving loans	M
32	Cash loans	F
33	Cash loans	F

CODE_GENDER F

Row Labels	Count of NAME_CONTRACT_TYPE
Cash loans	29665
Revolving loans	3158
Grand Total	**32823**

CODE_GENDER M

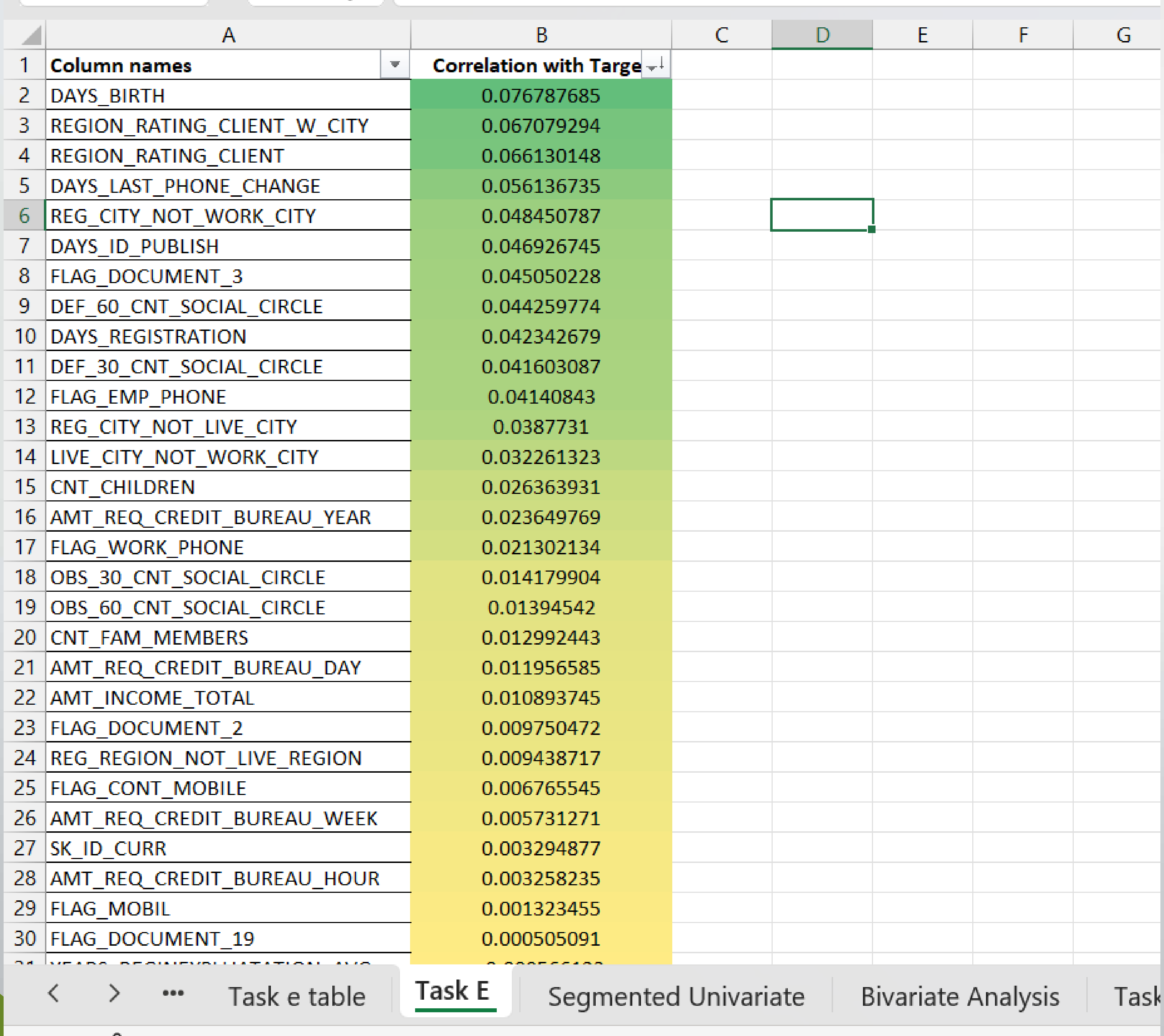
Row Labels	Count of NAME_CONTRACT_TYPE
Cash loans	15611
Revolving loans	1563
Grand Total	**17174**

CODE_GENDER XNA

Row Labels	Count of NAME_CONTRACT_TYPE
Revolving loans	2
Grand Total	**2**

Task E:Identify Top Correlations for Different Scenarios:

- **Description:** Segment the dataset based on different scenarios (e.g., clients with payment difficulties and all other cases) and identify the top correlations for each segmented data using Excel functions.
- I’ve find the correlation between target and all other columns respectively.
- Then using conditional formatting highlighted the correlations arranged in descending order.



	Column names	Correlation with Targe
1	Column names	Correlation with Targe
2	DAYS_BIRTH	0.076787685
3	REGION_RATING_CLIENT_W_CITY	0.067079294
4	REGION_RATING_CLIENT	0.066130148
5	DAYS_LAST_PHONE_CHANGE	0.056136735
6	REG_CITY_NOT_WORK_CITY	0.048450787
7	DAYS_ID_PUBLISH	0.046926745
8	FLAG_DOCUMENT_3	0.045050228
9	DEF_60_CNT_SOCIAL_CIRCLE	0.044259774
10	DAYS_REGISTRATION	0.042342679
11	DEF_30_CNT_SOCIAL_CIRCLE	0.041603087
12	FLAG_EMP_PHONE	0.04140843
13	REG_CITY_NOT_LIVE_CITY	0.0387731
14	LIVE_CITY_NOT_WORK_CITY	0.032261323
15	CNT_CHILDREN	0.026363931
16	AMT_REQ_CREDIT_BUREAU_YEAR	0.023649769
17	FLAG_WORK_PHONE	0.021302134
18	OBS_30_CNT_SOCIAL_CIRCLE	0.014179904
19	OBS_60_CNT_SOCIAL_CIRCLE	0.01394542
20	CNT_FAM_MEMBERS	0.012992443
21	AMT_REQ_CREDIT_BUREAU_DAY	0.011956585
22	AMT_INCOME_TOTAL	0.010893745
23	FLAG_DOCUMENT_2	0.009750472
24	REG_REGION_NOT_LIVE_REGION	0.009438717
25	FLAG_CONT_MOBILE	0.006765545
26	AMT_REQ_CREDIT_BUREAU_WEEK	0.005731271
27	SK_ID_CURR	0.003294877
28	AMT_REQ_CREDIT_BUREAU_HOUR	0.003258235
29	FLAG_MOBIL	0.001323455
30	FLAG_DOCUMENT_19	0.000505091

THANKYOU