Bloomreach Engagement Support Team

Interview Assignment for the position of Product Support Specialist - Engagement

To find the answer for question “what could be a root cause?” I need to answer for several questions firstly. Initially I recognize following crucial questions:

- what are the most common SMTP error messages?
- which domains are affected with this error?
- can we relate number of mail messages sent with bounce rate for particular domains?

To complete it I need to analyse provided data set. As it contains a lot of data, some adjustment activities would be helpful to make analyse easier.

After that I need to familiarize with adjusted set and try to establish conclusions. It will allow me to provide hypothesis about the root cause and maybe to propose a solution.

All activities described above can be grouped into following steps:

1. Initial data review and SMTP responses recognition
2. Data set adjustment
3. Extraction of necessary data and transformation
4. Import into analytical environment and data analysis
5. Final conclusions

Each point is described in sections below. My hypothesis and propositions are underlined in data analysis section and summarized in final conclusions.

Step 1 - Initial data review and SMTP responses recognition

Data contains information about occurrence of various SMTP responses on pointed date. Date is provided in UNIX format. Both SMTP codes and UNIX date require further activities. Let's start with SMTP codes recognition.

We face several types of responses in input file. I have checked meaning of each of them using document available here. There is also website focused on SMTP codes which was helpful.

Following table presents codes and respective description:

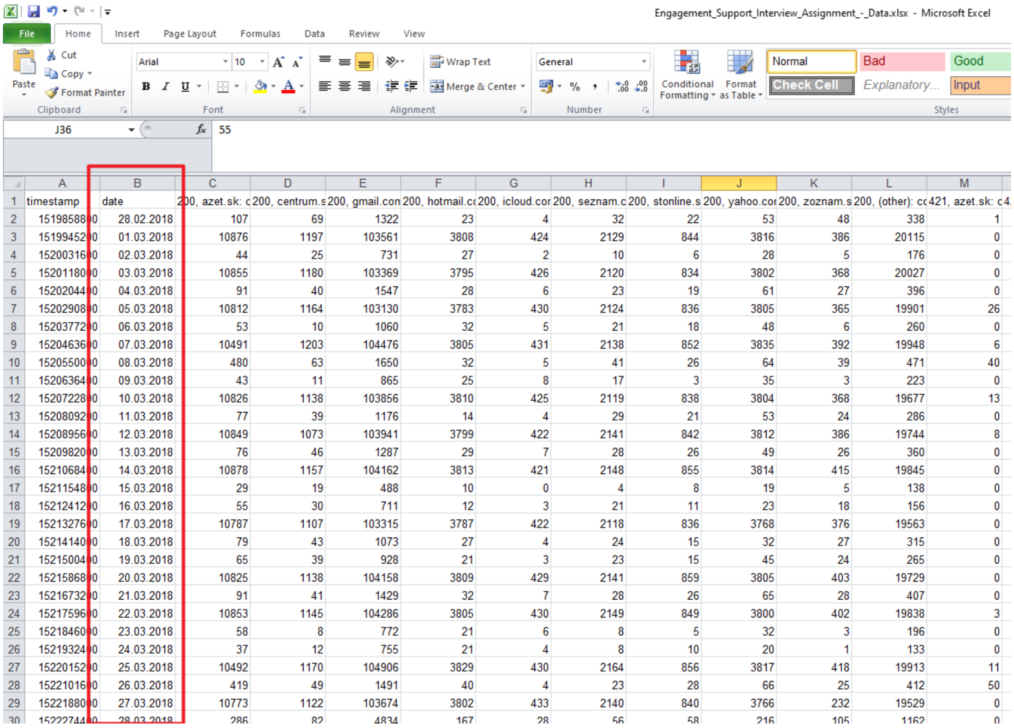
Code	Description
200 – Positive completion reply	The requested action has been successfully completed.
421 - Service not available	The Mail transfer service is unavailable. This can be caused by many things such as a server administrator stopping the mail service, or rebooting the mail server.
451 - Requested action aborted – Local error in processing	The action has been aborted by the ISP's server.
452 - Requested action not taken – Insufficient storage	This is usually caused by overloading mail server when attempting to send too many messages at once.
499 - Client closed request	It indicates that the client has closed the connection while the server is still processing the request. (not confirmed)
550 - Requested actions not taken mailbox unavailable	Recipient email address simply does not exist on the remote side.
552 - Requested mail actions aborted – Exceeded storage allocation	Mailbox has reached its maximum allowed size.
554 – Transaction failed	Recipient email address does not exist or there is anti-spam firewall.
605	Email address is currently suppressed by our system from further delivery attempts. (not confirmed)

Step 2 - Data set adjustment

As the task is focused on May campaign (increased bounce rate is also present in other months, but I treat it as out of scope) first thing to do is to filter data from May. To complete it I need to convert UNIX timestamp to date in 'user friendly' form. I do it directly in excel sheet using formula:

$$= ((('cell\ with\ UNIX\ timestamp'/60)/60)/24) + DATE(1970;1;1)$$

New column with DD.MM.YYYY format is now displayed next to timestamp:



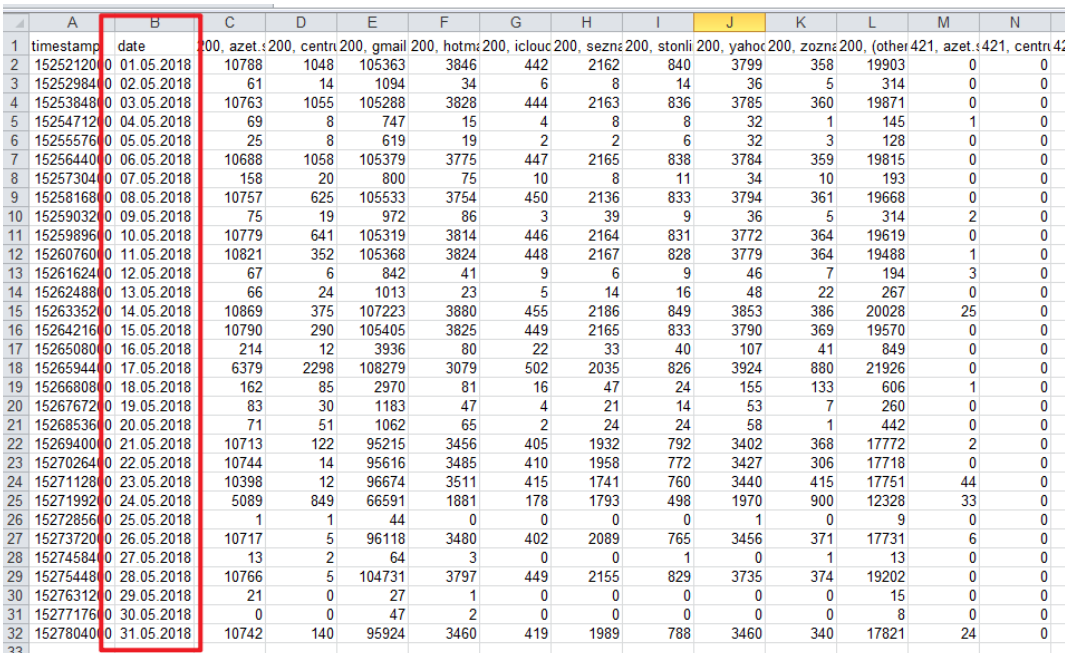
Engagement_Support_Interview_Assignment_-_Data.xlsx - Microsoft Excel

J36 55

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	timestamp	date	200, azet.sk: c	200, centrum.s	200, gmail.con	200, hotmail.c	200, icloud.cor	200, seznam.c	200, stonline.s	200, yahoo.coı	200, zoznam.s	200, (other): cc	421, azet.sk: c
2	1519858800	28.02.2018	107	69	1322	23	4	32	22	53	48	338	1
3	1519945200	01.03.2018	10876	1197	103561	3808	424	2129	844	3816	386	20115	0
4	1520031600	02.03.2018	44	25	731	27	2	10	6	28	5	176	0
5	1520118000	03.03.2018	10855	1180	103369	3795	426	2120	834	3802	368	20027	0
6	1520204400	04.03.2018	91	40	1547	28	6	23	19	61	27	396	0
7	1520290800	05.03.2018	10812	1164	103130	3783	430	2124	836	3805	365	19901	26
8	1520377200	06.03.2018	53	10	1060	32	5	21	18	48	6	260	0
9	1520463600	07.03.2018	10491	1203	104476	3805	431	2138	852	3835	392	19948	6
10	1520550000	08.03.2018	480	63	1650	32	5	41	26	64	39	471	40
11	1520636400	09.03.2018	43	11	865	25	8	17	3	35	3	223	0
12	1520722800	10.03.2018	10826	1138	103856	3810	425	2119	838	3804	368	19677	13
13	1520809200	11.03.2018	77	39	1176	14	4	29	21	53	24	286	0
14	1520895600	12.03.2018	10849	1073	103941	3799	422	2141	842	3812	386	19744	8
15	1520982000	13.03.2018	76	46	1287	29	7	28	26	49	26	360	0
16	1521068400	14.03.2018	10878	1157	104162	3813	421	2148	855	3814	415	19845	0
17	1521154800	15.03.2018	29	19	488	10	0	4	8	19	5	138	0
18	1521241200	16.03.2018	55	30	711	12	3	21	11	23	18	156	0
19	1521327600	17.03.2018	10787	1107	103315	3787	422	2118	836	3768	376	19563	0
20	1521414000	18.03.2018	79	43	1073	27	4	24	15	32	27	315	0
21	1521500400	19.03.2018	65	39	928	21	3	23	15	45	24	265	0
22	1521586800	20.03.2018	10825	1138	104158	3809	429	2141	859	3805	403	19729	0
23	1521673200	21.03.2018	91	41	1429	32	7	28	26	65	28	407	0
24	1521759600	22.03.2018	10853	1145	104286	3805	430	2149	849	3800	402	19838	3
25	1521846000	23.03.2018	58	8	772	21	6	8	5	32	3	196	0
26	1521932400	24.03.2018	37	12	755	21	4	8	10	20	1	133	0
27	1522015200	25.03.2018	10492	1170	104906	3829	430	2164	856	3817	418	19913	11
28	1522101600	26.03.2018	419	49	1491	40	4	23	28	66	25	412	50
29	1522188000	27.03.2018	10773	1122	103674	3802	433	2140	840	3766	232	19529	0
30	1522274400	28.03.2018	286	82	4834	167	28	56	58	216	105	1162	0

Step 3 - Extraction of necessary data and transformation

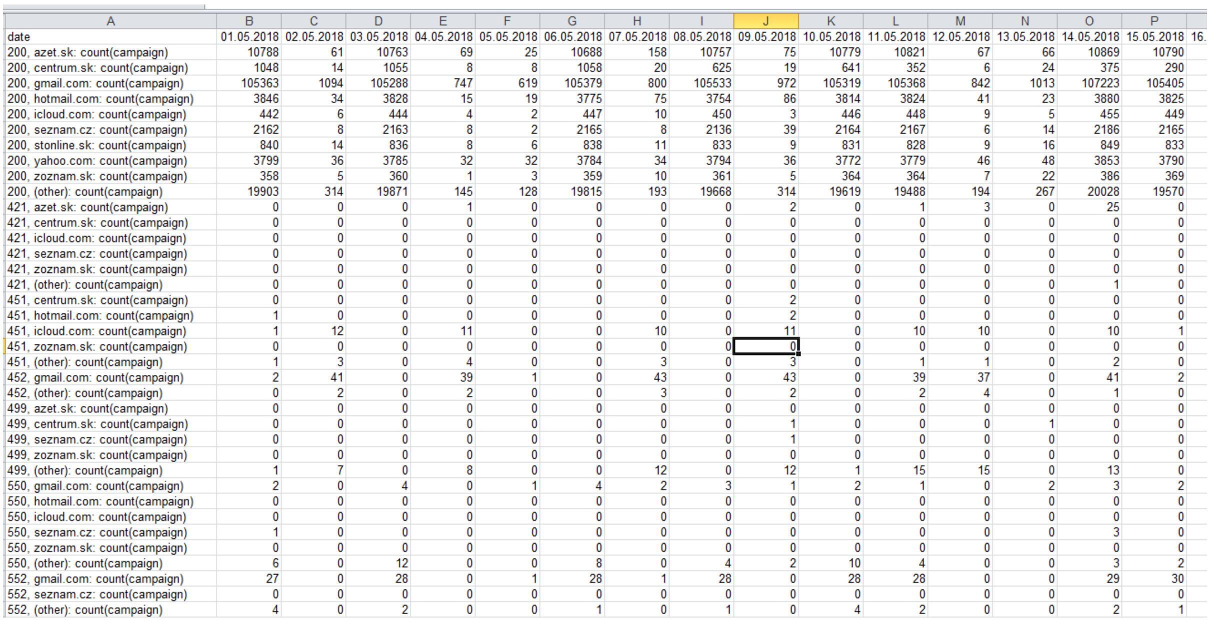
Now I clip data related to May into another sheet:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	timestamp	date	200, azet.s	200, centru	200, gmail	200, hotma	200, iclouc	200, sezna	200, stonli	200, yahoc	200, zozna	200, (other	421, azet.s	421, centru
2	1525212000	01.05.2018	10788	1048	105363	3846	442	2162	840	3799	358	19903	0	0
3	1525298400	02.05.2018	61	14	1094	34	6	8	14	36	5	314	0	0
4	1525384800	03.05.2018	10763	1055	105288	3828	444	2163	836	3785	360	19871	0	0
5	1525471200	04.05.2018	69	8	747	15	4	8	8	32	1	145	1	0
6	1525557600	05.05.2018	25	8	619	19	2	2	6	32	3	128	0	0
7	1525644000	06.05.2018	10688	1058	105379	3775	447	2165	838	3784	359	19815	0	0
8	1525730400	07.05.2018	158	20	800	75	10	8	11	34	10	193	0	0
9	1525816800	08.05.2018	10757	625	105533	3754	450	2136	833	3794	361	19668	0	0
10	1525903200	09.05.2018	75	19	972	86	3	39	9	36	5	314	2	0
11	1525989600	10.05.2018	10779	641	105319	3814	446	2164	831	3772	364	19619	0	0
12	1526076000	11.05.2018	10821	352	105368	3824	448	2167	828	3779	364	19488	1	0
13	1526162400	12.05.2018	67	6	842	41	9	6	9	46	7	194	3	0
14	1526248800	13.05.2018	66	24	1013	23	5	14	16	48	22	267	0	0
15	1526335200	14.05.2018	10869	375	107223	3880	455	2186	849	3853	386	20028	25	0
16	1526421600	15.05.2018	10790	290	105405	3825	449	2165	833	3790	369	19570	0	0
17	1526508000	16.05.2018	214	12	3936	80	22	33	40	107	41	849	0	0
18	1526594400	17.05.2018	6379	2298	108279	3079	502	2035	826	3924	880	21926	0	0
19	1526680800	18.05.2018	162	85	2970	81	16	47	24	155	133	606	1	0
20	1526767200	19.05.2018	83	30	1183	47	4	21	14	53	7	260	0	0
21	1526853600	20.05.2018	71	51	1062	65	2	24	24	58	1	442	0	0
22	1526940000	21.05.2018	10713	122	95215	3456	405	1932	792	3402	368	17772	2	0
23	1527026400	22.05.2018	10744	14	95616	3485	410	1958	772	3427	306	17718	0	0
24	1527112800	23.05.2018	10398	12	96674	3511	415	1741	760	3440	415	17751	44	0
25	1527199200	24.05.2018	5089	849	66591	1881	178	1793	498	1970	900	12328	33	0
26	1527285600	25.05.2018	1	1	44	0	0	0	0	1	0	9	0	0
27	1527372000	26.05.2018	10717	5	96118	3480	402	2089	765	3456	371	17731	6	0
28	1527458400	27.05.2018	13	2	64	3	0	0	1	0	1	13	0	0
29	1527544800	28.05.2018	10766	5	104731	3797	449	2155	829	3735	374	19202	0	0
30	1527631200	29.05.2018	21	0	27	1	0	0	0	0	0	15	0	0
31	1527717600	30.05.2018	0	0	47	2	0	0	0	0	0	8	0	0
32	1527804000	31.05.2018	10742	140	95924	3460	419	1989	788	3460	340	17821	24	0

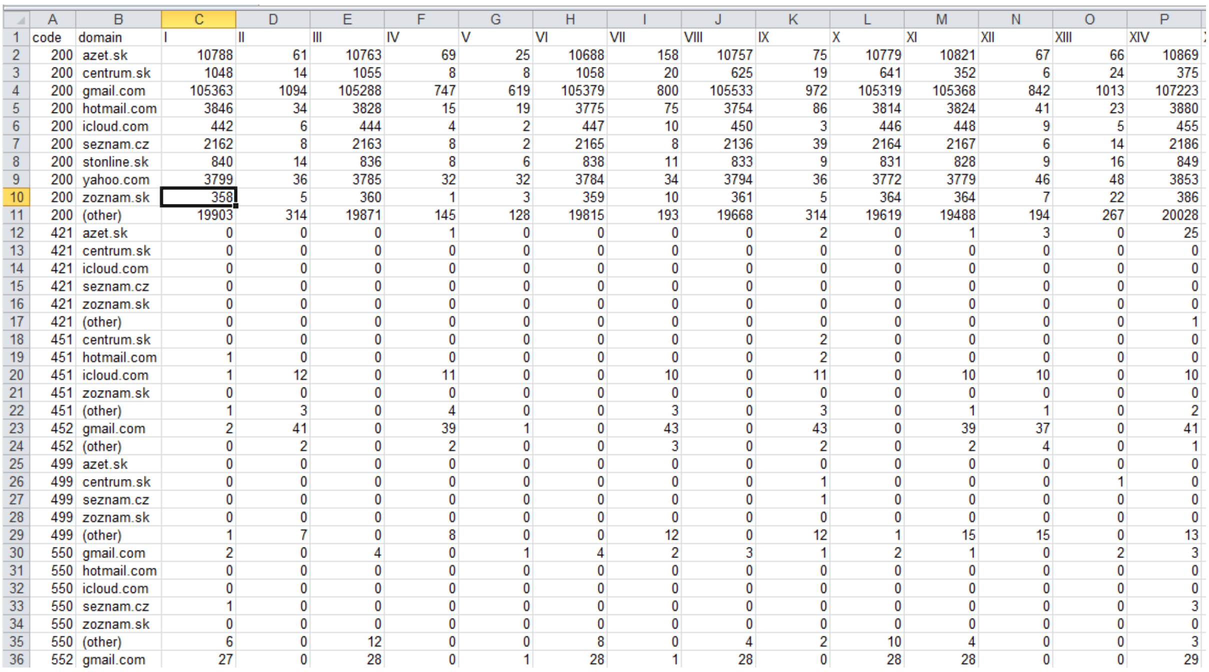
I have reduced the overall number of records but there is still a lot of columns. What is more each column contains 2 information – SMTP code and domain.

Analysis of data in this shape is inefficient and as a result it is difficult to find out answers for question listed at the beginning. My idea is to analyse particular codes and domains instead of dates so I will do transposition, separate codes and domains, add new column with total number of responses and load it as a database table to do SQL queries. SQL will allow me to do various checks helpful to answer posted questions. Let's begin with transposition. I will copy necessary data and paste it with transposition in new sheet:



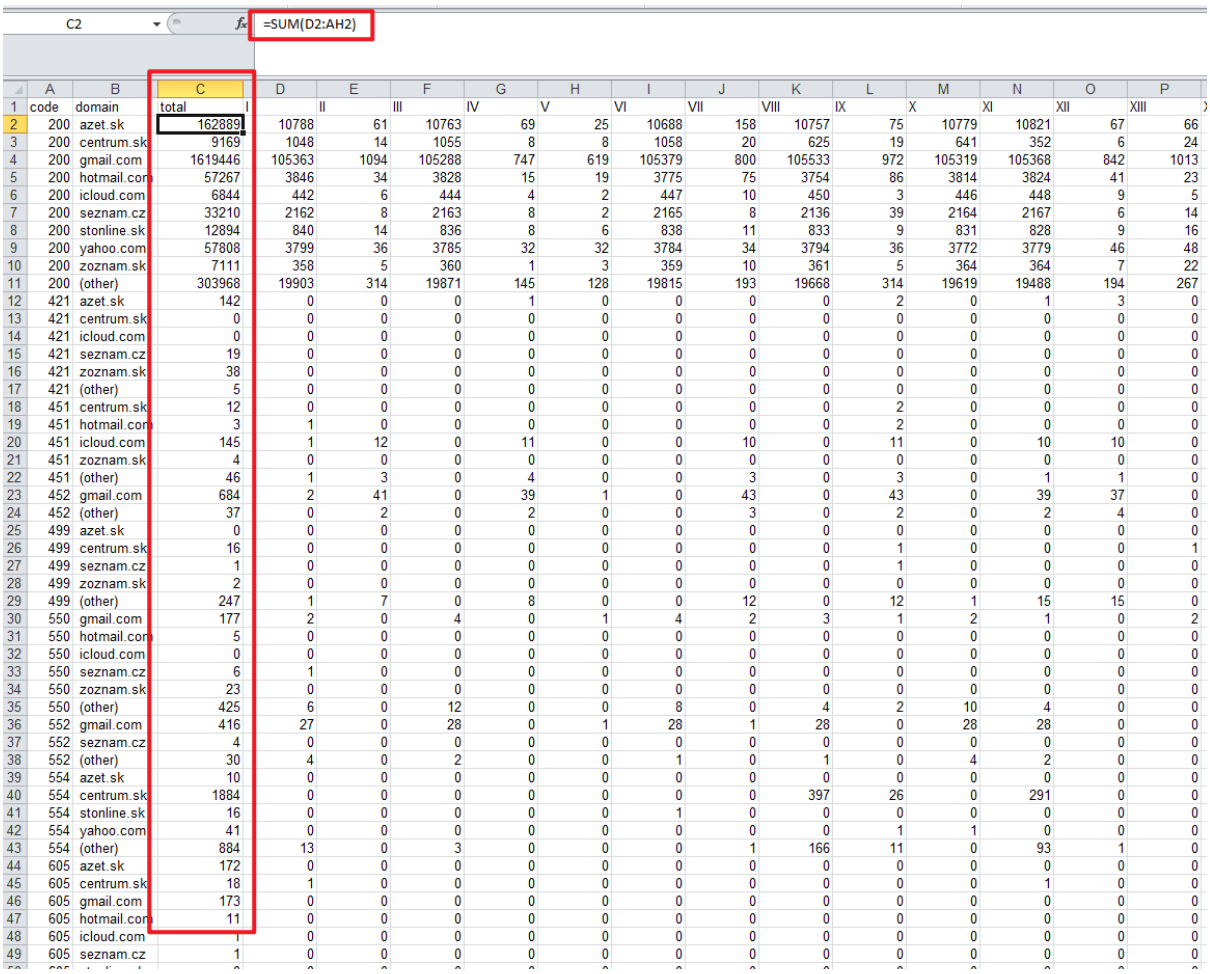
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	
date	01.05.2018	02.05.2018	03.05.2018	04.05.2018	05.05.2018	06.05.2018	07.05.2018	08.05.2018	09.05.2018	10.05.2018	11.05.2018	12.05.2018	13.05.2018	14.05.2018	15.05.2018	16.
200, azet.sk: count(campaign)	10788	61	10763	69	25	10688	158	10757	75	10779	10821	67	66	10869	10790	
200, centrum.sk: count(campaign)	1048	14	1055	8	8	1058	20	625	19	641	352	6	24	375	290	
200, gmail.com: count(campaign)	105363	1094	105288	747	619	105379	800	105533	972	105319	105368	842	1013	107223	105405	
200, hotmail.com: count(campaign)	3846	34	3828	15	19	3775	75	3754	86	3814	3824	41	23	3880	3825	
200, icloud.com: count(campaign)	442	6	444	4	2	447	10	450	3	446	448	9	5	455	449	
200, seznam.cz: count(campaign)	2162	8	2163	8	2	2165	8	2136	39	2164	2167	6	14	2186	2165	
200, stonline.sk: count(campaign)	840	14	836	8	6	838	11	833	9	831	828	9	16	849	833	
200, yahoo.com: count(campaign)	3799	36	3785	32	32	3784	34	3794	36	3772	3779	46	48	3853	3790	
200, zoznam.sk: count(campaign)	358	5	360	1	3	359	10	361	5	364	364	7	22	386	369	
200, (other): count(campaign)	19903	314	19871	145	128	19815	193	19668	314	19619	19488	194	267	20028	19570	
421, azet.sk: count(campaign)	0	0	0	1	0	0	0	0	2	0	1	3	0	25	0	
421, centrum.sk: count(campaign)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
421, icloud.com: count(campaign)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
421, seznam.cz: count(campaign)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
421, zoznam.sk: count(campaign)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
421, (other): count(campaign)	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	
451, centrum.sk: count(campaign)	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	
451, hotmail.com: count(campaign)	1	0	0	0	0	0	0	0	2	0	0	0	0	0	0	
451, icloud.com: count(campaign)	1	12	0	11	0	0	10	0	11	0	10	10	0	10	1	
451, zoznam.sk: count(campaign)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
451, (other): count(campaign)	1	3	0	4	0	0	3	0	3	0	1	1	0	2	0	
452, gmail.com: count(campaign)	2	41	0	39	1	0	43	0	43	0	39	37	0	41	2	
452, (other): count(campaign)	0	2	0	2	0	0	3	0	2	0	2	4	0	1	0	
499, azet.sk: count(campaign)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
499, centrum.sk: count(campaign)	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	
499, seznam.cz: count(campaign)	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	
499, zoznam.sk: count(campaign)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
499, (other): count(campaign)	1	7	0	8	0	0	12	0	12	1	15	15	0	13	0	
550, gmail.com: count(campaign)	2	0	4	0	1	4	2	3	1	2	1	0	2	3	2	
550, hotmail.com: count(campaign)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
550, icloud.com: count(campaign)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
550, seznam.cz: count(campaign)	1	0	0	0	0	0	0	0	0	0	0	0	0	3	0	
550, zoznam.sk: count(campaign)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
550, (other): count(campaign)	6	0	12	0	0	8	0	4	2	10	4	0	0	3	2	
552, gmail.com: count(campaign)	27	0	28	0	1	28	1	28	0	28	28	0	0	29	30	
552, seznam.cz: count(campaign)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
552, (other): count(campaign)	4	0	2	0	0	1	0	1	0	4	2	0	0	2	1	

Now I will separate first column, remove unnecessary string ‘*count(campaign)*’, and rename columns to be more meaningful and database friendly:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	code	domain	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV
2	200	azet.sk	10788	61	10763	69	25	10688	158	10757	75	10779	10821	67	66	10869
3	200	centrum.sk	1048	14	1055	8	8	1058	20	625	19	641	352	6	24	375
4	200	gmail.com	105363	1094	105288	747	619	105379	800	105533	972	105319	105368	842	1013	107223
5	200	hotmail.com	3846	34	3828	15	19	3775	75	3754	86	3814	3824	41	23	3880
6	200	icloud.com	442	6	444	4	2	447	10	450	3	446	448	9	5	455
7	200	seznam.cz	2162	8	2163	8	2	2165	8	2136	39	2164	2167	6	14	2186
8	200	stonline.sk	840	14	836	8	6	838	11	833	9	831	828	9	16	849
9	200	yahoo.com	3799	36	3785	32	32	3784	34	3794	36	3772	3779	46	48	3853
10	200	zoznam.sk	358	5	360	1	3	359	10	361	5	364	364	7	22	386
11	200	(other)	19903	314	19871	145	128	19815	193	19668	314	19619	19488	194	267	20028
12	421	azet.sk	0	0	0	1	0	0	0	0	2	0	1	3	0	25
13	421	centrum.sk	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	421	icloud.com	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	421	seznam.cz	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	421	zoznam.sk	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	421	(other)	0	0	0	0	0	0	0	0	0	0	0	0	0	1
18	451	centrum.sk	0	0	0	0	0	0	0	0	2	0	0	0	0	0
19	451	hotmail.com	1	0	0	0	0	0	0	0	2	0	0	0	0	0
20	451	icloud.com	1	12	0	11	0	0	10	0	11	0	10	10	0	10
21	451	zoznam.sk	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	451	(other)	1	3	0	4	0	0	3	0	3	0	1	1	0	2
23	452	gmail.com	2	41	0	39	1	0	43	0	43	0	39	37	0	41
24	452	(other)	0	2	0	2	0	0	3	0	2	0	2	4	0	1
25	499	azet.sk	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	499	centrum.sk	0	0	0	0	0	0	0	0	1	0	0	0	1	0
27	499	seznam.cz	0	0	0	0	0	0	0	0	1	0	0	0	0	0
28	499	zoznam.sk	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	499	(other)	1	7	0	8	0	0	12	0	12	1	15	15	0	13
30	550	gmail.com	2	0	4	0	1	4	2	3	1	2	1	0	2	3
31	550	hotmail.com	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	550	icloud.com	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	550	seznam.cz	1	0	0	0	0	0	0	0	0	0	0	0	0	3
34	550	zoznam.sk	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	550	(other)	6	0	12	0	0	8	0	4	2	10	4	0	0	3
36	552	gmail.com	27	0	28	0	1	28	1	28	0	28	28	0	0	29

It's against relational databases rules but in this case it will be useful to add another column with sum of responses from entire month:



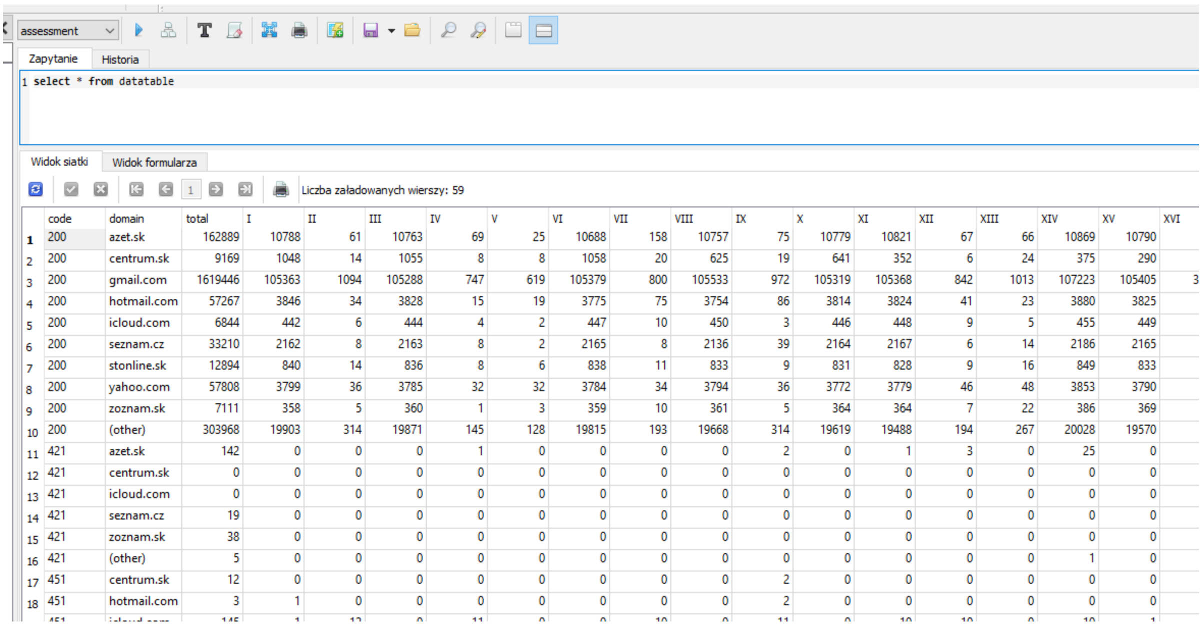
C2 =SUM(D2:AH2)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	code	domain	total	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII
2	200	azet.sk	162889	10788	61	10763	69	25	10688	158	10757	75	10779	10821	67	66
3	200	centrum.sk	9169	1048	14	1055	8	8	1058	20	625	19	641	352	6	24
4	200	gmail.com	1619446	105363	1094	105288	747	619	105379	800	105533	972	105319	105368	842	1013
5	200	hotmail.com	57267	3846	34	3828	15	19	3775	75	3754	86	3814	3824	41	23
6	200	icloud.com	6844	442	6	444	4	2	447	10	450	3	446	448	9	5
7	200	seznam.cz	33210	2162	8	2163	8	2	2165	8	2136	39	2164	2167	6	14
8	200	stonline.sk	12894	840	14	836	8	6	838	11	833	9	831	828	9	16
9	200	yahoo.com	57808	3799	36	3785	32	32	3784	34	3794	36	3772	3779	46	48
10	200	zoznam.sk	7111	358	5	360	1	3	359	10	361	5	364	364	7	22
11	200	(other)	303968	19903	314	19871	145	128	19815	193	19668	314	19619	19488	194	267
12	421	azet.sk	142	0	0	0	1	0	0	0	0	2	0	1	3	0
13	421	centrum.sk	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	421	icloud.com	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	421	seznam.cz	19	0	0	0	0	0	0	0	0	0	0	0	0	0
16	421	zoznam.sk	38	0	0	0	0	0	0	0	0	0	0	0	0	0
17	421	(other)	5	0	0	0	0	0	0	0	0	0	0	0	0	0
18	451	centrum.sk	12	0	0	0	0	0	0	0	0	2	0	0	0	0
19	451	hotmail.com	3	1	0	0	0	0	0	0	0	2	0	0	0	0
20	451	icloud.com	145	1	12	0	11	0	0	10	0	11	0	10	10	0
21	451	zoznam.sk	4	0	0	0	0	0	0	0	0	0	0	0	0	0
22	451	(other)	46	1	3	0	4	0	0	3	0	3	0	1	1	0
23	452	gmail.com	684	2	41	0	39	1	0	43	0	43	0	39	37	0
24	452	(other)	37	0	2	0	2	0	0	3	0	2	0	2	4	0
25	499	azet.sk	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	499	centrum.sk	16	0	0	0	0	0	0	0	0	1	0	0	0	1
27	499	seznam.cz	1	0	0	0	0	0	0	0	0	1	0	0	0	0
28	499	zoznam.sk	2	0	0	0	0	0	0	0	0	0	0	0	0	0
29	499	(other)	247	1	7	0	8	0	0	12	0	12	1	15	15	0
30	550	gmail.com	177	2	0	4	0	1	4	2	3	1	2	1	0	2
31	550	hotmail.com	5	0	0	0	0	0	0	0	0	0	0	0	0	0
32	550	icloud.com	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	550	seznam.cz	6	1	0	0	0	0	0	0	0	0	0	0	0	0
34	550	zoznam.sk	23	0	0	0	0	0	0	0	0	0	0	0	0	0
35	550	(other)	425	6	0	12	0	0	8	0	4	2	10	4	0	0
36	552	gmail.com	416	27	0	28	0	1	28	1	28	0	28	28	0	0
37	552	seznam.cz	4	0	0	0	0	0	0	0	0	0	0	0	0	0
38	552	(other)	30	4	0	2	0	0	1	0	1	0	4	2	0	0
39	554	azet.sk	10	0	0	0	0	0	0	0	0	0	0	0	0	0
40	554	centrum.sk	1884	0	0	0	0	0	0	0	397	26	0	291	0	0
41	554	stonline.sk	16	0	0	0	0	0	1	0	0	0	0	0	0	0
42	554	yahoo.com	41	0	0	0	0	0	0	0	0	1	1	0	0	0
43	554	(other)	884	13	0	3	0	0	0	1	166	11	0	93	1	0
44	605	azet.sk	172	0	0	0	0	0	0	0	0	0	0	0	0	0
45	605	centrum.sk	18	1	0	0	0	0	0	0	0	0	0	1	0	0
46	605	gmail.com	173	0	0	0	0	0	0	0	0	0	0	0	0	0
47	605	hotmail.com	11	0	0	0	0	0	0	0	0	0	0	0	0	0
48	605	icloud.com	1	0	0	0	0	0	0	0	0	0	0	0	0	0
49	605	seznam.cz	1	0	0	0	0	0	0	0	0	0	0	0	0	0

Data is ready to be loaded as a database table.

Step 4 - Import into analytical environment and data analysis

I use SQLite RDBMS as it is free software. Data is extracted from excel into csv format and then loaded into database table:



```
select * from datatable
```

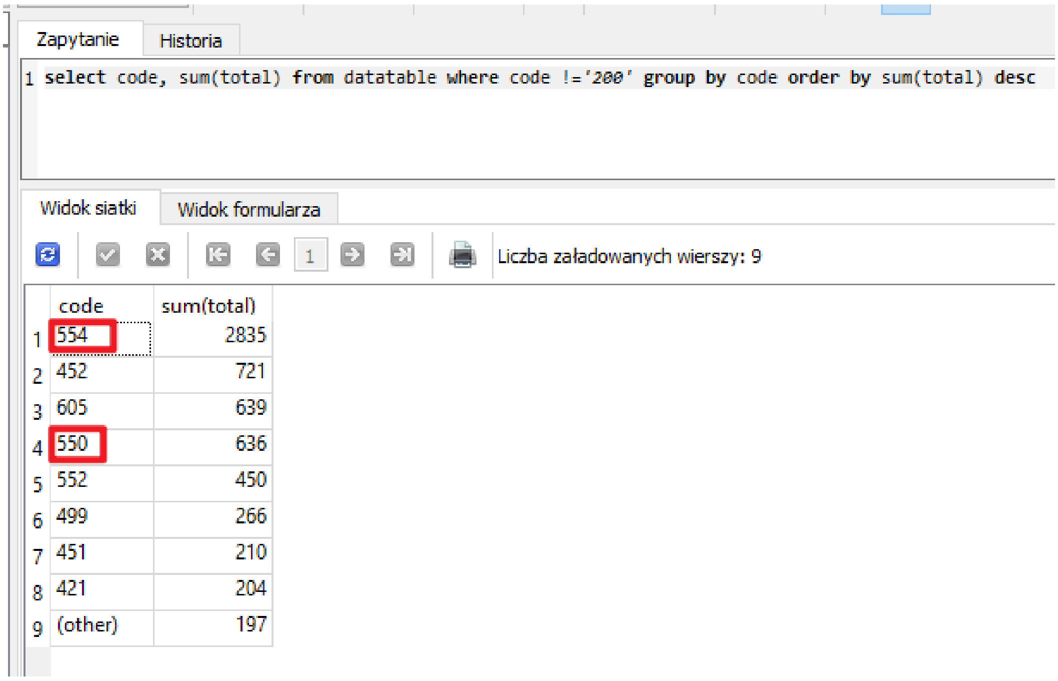
Liczba załadowanych wierszy: 59

	code	domain	total	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	XVI
1	200	azet.sk	162889	10788	61	10763	69	25	10688	158	10757	75	10779	10821	67	66	10869	10790	
2	200	centrum.sk	9169	1048	14	1055	8	8	1058	20	625	19	641	352	6	24	375	290	
3	200	gmail.com	1619446	105363	1094	105288	747	619	105379	800	105533	972	105319	105368	842	1013	107223	105405	3
4	200	hotmail.com	57267	3846	34	3828	15	19	3775	75	3754	86	3814	3824	41	23	3880	3825	
5	200	icloud.com	6844	442	6	444	4	2	447	10	450	3	446	448	9	5	455	449	
6	200	seznam.cz	33210	2162	8	2163	8	2	2165	8	2136	39	2164	2167	6	14	2186	2165	
7	200	stonline.sk	12894	840	14	836	8	6	838	11	833	9	831	828	9	16	849	833	
8	200	yahoo.com	57808	3799	36	3785	32	32	3784	34	3794	36	3772	3779	46	48	3853	3790	
9	200	zoznam.sk	7111	358	5	360	1	3	359	10	361	5	364	364	7	22	386	369	
10	200	(other)	303968	19903	314	19871	145	128	19815	193	19668	314	19619	19488	194	267	20028	19570	
11	421	azet.sk	142	0	0	0	1	0	0	0	0	2	0	1	3	0	25	0	
12	421	centrum.sk	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
13	421	icloud.com	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
14	421	seznam.cz	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
15	421	zoznam.sk	38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
16	421	(other)	5	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	
17	451	centrum.sk	12	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	
18	451	hotmail.com	3	1	0	0	0	0	0	0	0	2	0	0	0	0	0	0	

Code 554, 550

Now database table is ready to do analysis. Let’s start with check which code was returned most often:

select code, sum(total) from datatable where code !='200' group by code order by sum(total) desc



```
select code, sum(total) from datatable where code !='200' group by code order by sum(total) desc
```

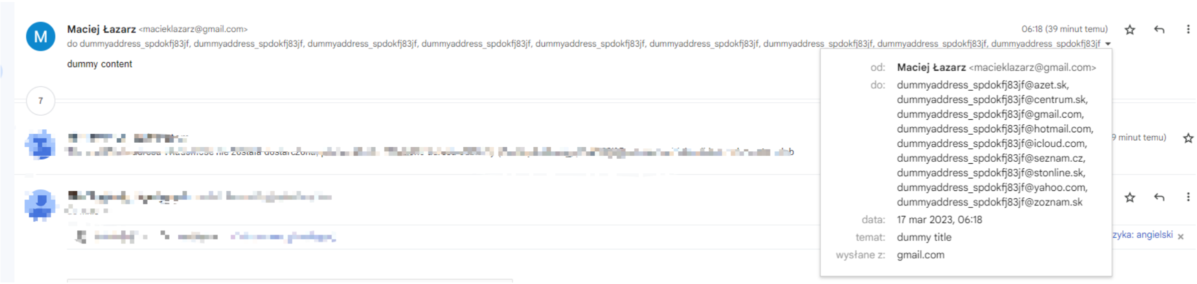
Liczba załadowanych wierszy: 9

	code	sum(total)
1	554	2835
2	452	721
3	605	639
4	550	636
5	552	450
6	499	266
7	451	210
8	421	204
9	(other)	197

We can see 554 code in position number 1 with large advantage before 452 in position number 2. There is also great number in case of code 550 and I intentionally consider it before I consider codes between 554 and 550 because both 550 and 554 are returned in similar situation – not valid mail address. This is one of supposed reasons of failure for mentioned codes and the second is anti-spam firewall. In this situation (if I was granted with access) I would check if there is some additional message returned with code which will clarify what is the reason of failure, because with error code there could be also message like *"550 Invalid recipient"* or "*550 User account is unavailable"* attached. It will allow me to list invalid mail address and inform client about it.

In case of failure caused by anti-spam firewall I would notify person responsible for the mailing tool about that fact pointing cases (email addresses) when it occur to allow him to do proper reconfiguration or changes in the tool to avoid firewall.

I will do small check. Let's send a mail to dummy addresses for each domain and analyze response:



Maciej Łazarz <macieklazarz@gmail.com>
do dummyaddress_spdokfj83jf, dummyaddress_spdokfj83jf, dummyaddress_spdokfj83jf, dummyaddress_spdokfj83jf, dummyaddress_spdokfj83jf, dummyaddress_spdokfj83jf, dummyaddress_spdokfj83jf, dummyaddress_spdokfj83jf, dummyaddress_spdokfj83jf
06:18 (39 minut temu)
dummy content

od: Maciej Łazarz <macieklazarz@gmail.com>
do: dummyaddress_spdokfj83jf@azet.sk, dummyaddress_spdokfj83jf@centrum.sk, dummyaddress_spdokfj83jf@gmail.com, dummyaddress_spdokfj83jf@hotmail.com, dummyaddress_spdokfj83jf@icloud.com, dummyaddress_spdokfj83jf@seznam.cz, dummyaddress_spdokfj83jf@stonline.sk, dummyaddress_spdokfj83jf@yahoo.com, dummyaddress_spdokfj83jf@zoznam.sk
data: 17 mar 2023, 06:18
temat: dummy title
wysłane z: gmail.com

[redacted]

[redacted]

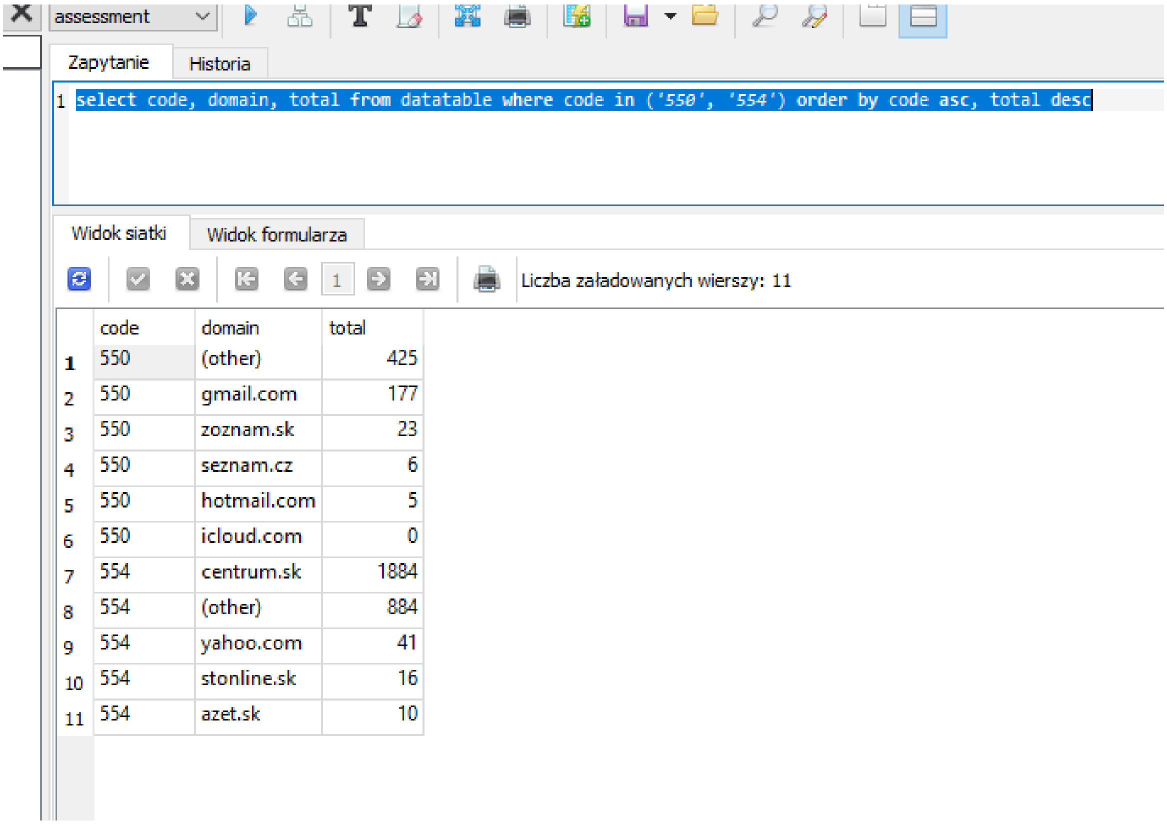
Answers are:

Domain	Response
azet.sk	550 5.1.1 Recipient address rejected
centrum.sk	550 #5.1.0 Address rejected.
gmail.com	550 5.1.1 The email account that you tried to reach does not exist.
hotmail.com	550 5.5.0 Requested action not taken: mailbox unavailable
icloud.com	550 5.1.1 user does not exist
seznam.cz	550 5.1.1 sorry, no such mailbox here
stonline.sk	550 5.1.1 Recipient address rejected: User unknown in local recipient table
yahoo.com	552 1 Requested mail action aborted, mailbox not found
zoznam.sk	550 5.1.1 Recipient address rejected: User unknown in virtual mailbox table

As we can see in almost all cases we have 550 response. It leads me to conclusion that in case of 550 code the reason is wrong address while in case of 554 the reason is probably firewall.

Now let's check if these errors are most common for some particular domains:

select code, domain, total from datatable where code in ('554', '550') order by code asc, total desc



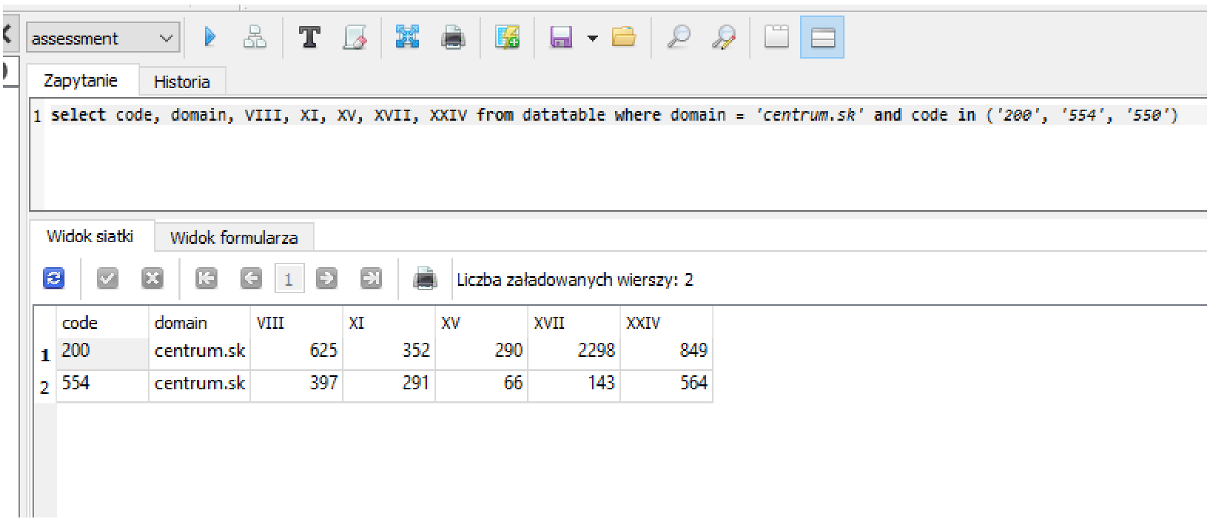
```
select code, domain, total from datatable where code in ('550', '554') order by code asc, total desc
```

Liczba załadowanych wierszy: 11

	code	domain	total
1	550	(other)	425
2	550	gmail.com	177
3	550	zoznam.sk	23
4	550	seznam.cz	6
5	550	hotmail.com	5
6	550	icloud.com	0
7	554	centrum.sk	1884
8	554	(other)	884
9	554	yahoo.com	41
10	554	stonline.sk	16
11	554	azet.sk	10

As we can see centrum.sk domain is most affected in case of 554 code. In this case let's take a closer look into emails sent to centrum.sk domain:

select code, domain, VIII, XI, XV, XVII, XXIV from datatable where domain = 'centrum.sk' and code in ('200', '554', '550')



```
select code, domain, VIII, XI, XV, XVII, XXIV from datatable where domain = 'centrum.sk' and code in ('200', '554', '550')
```

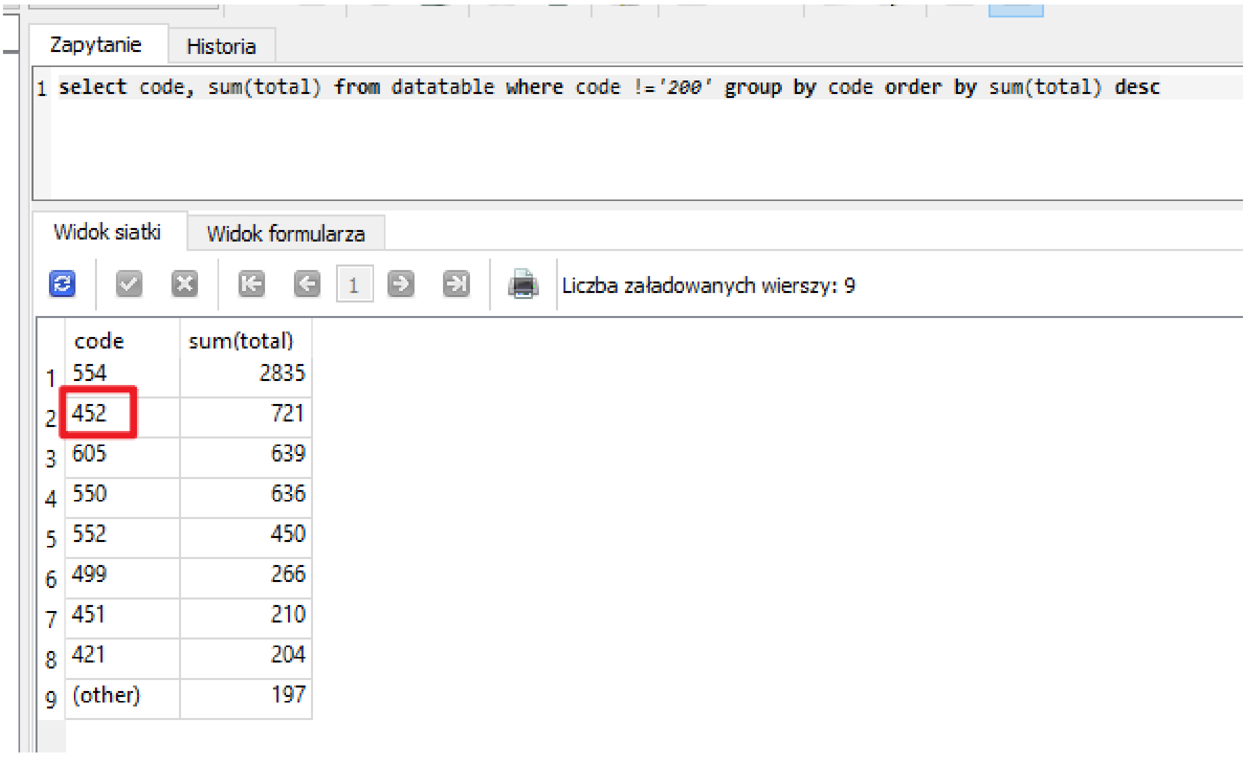
Liczba załadowanych wierszy: 2

	code	domain	VIII	XI	XV	XVII	XXIV
1	200	centrum.sk	625	352	290	2298	849
2	554	centrum.sk	397	291	66	143	564

554 code occurred among huge number of successfully delivered messages. It allows me to conclude that email recipients added sender to SPAM list.

Nevertheless, code comes with description so I would check the description in production database to make sure if my hypothesis is correct.

Code 452



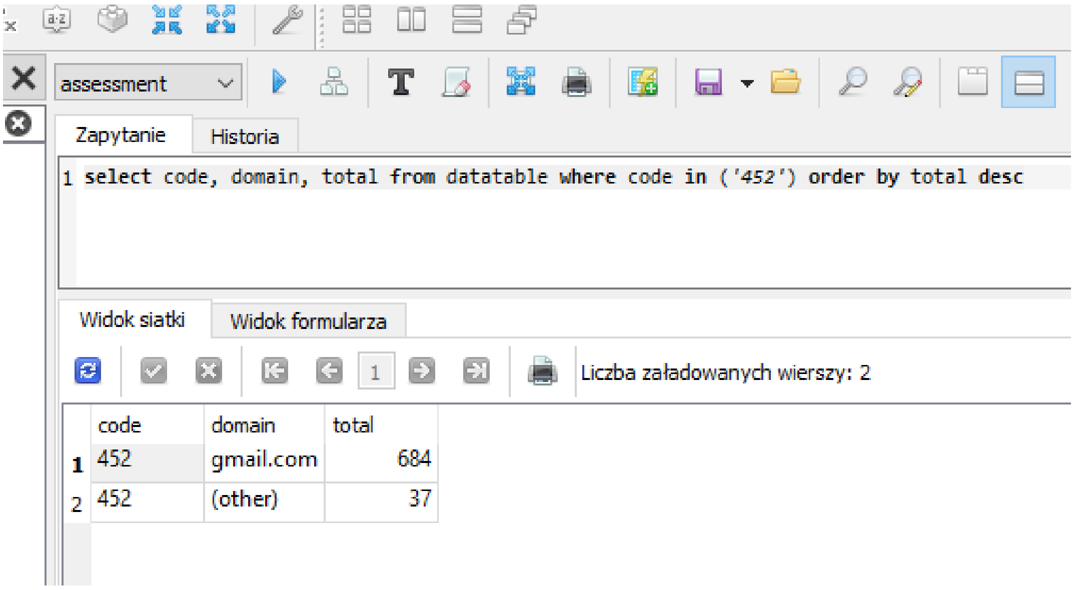
```
select code, sum(total) from datatable where code !='200' group by code order by sum(total) desc
```

Liczba załadowanych wierszy: 9

	code	sum(total)
1	554	2835
2	452	721
3	605	639
4	550	636
5	552	450
6	499	266
7	451	210
8	421	204
9	(other)	197

Code 452 is second on our list. It indicates that server is overloaded with too many messages. I suppose it occurs when huge number of mails is tried to be sent at the same time. Let's check it in our table:

select code, domain, total from datatable where code in ('452') order by total desc



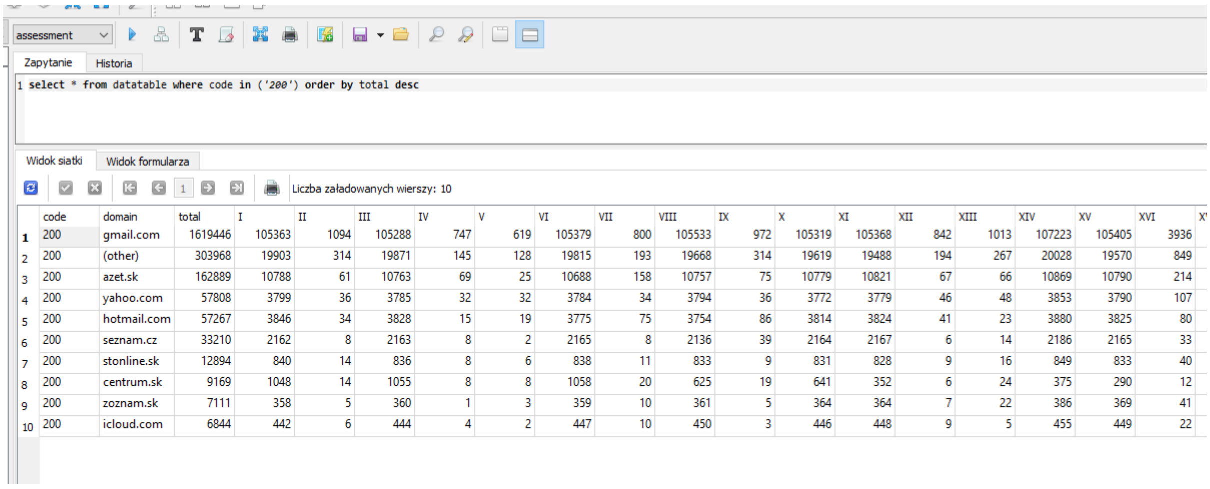
```
select code, domain, total from datatable where code in ('452') order by total desc
```

Liczba załadowanych wierszy: 2

	code	domain	total
1	452	gmail.com	684
2	452	(other)	37

As we can see in this case gmail.com is mostly affected. I assume that it may be caused by the fact that gmail is really popular mail domain (so it needs to serve huge number of requests at the same time) and additionally as we can see below number daily messages delivered there is higher than in case of rest domains summed up:

*select * from datatable where code in ('200') order by total desc*



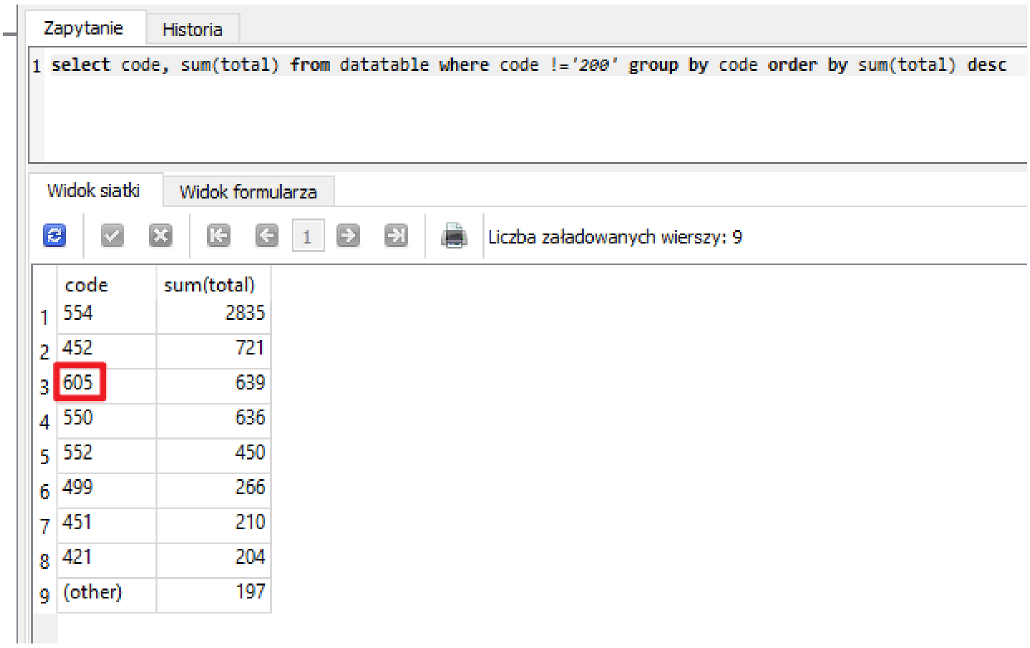
```
select * from datatable where code in ('200') order by total desc
```

Liczba załadowanych wierszy: 10

	code	domain	total	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	XVI
1	200	gmail.com	1619446	105363	1094	105288	747	619	105379	800	105533	972	105319	105368	842	1013	107223	105405	3936
2	200	(other)	303968	19903	314	19871	145	128	19815	193	19668	314	19619	19488	194	267	20028	19570	849
3	200	azet.sk	162889	10788	61	10763	69	25	10688	158	10757	75	10779	10821	67	66	10869	10790	214
4	200	yahoo.com	57808	3799	36	3785	32	32	3784	34	3794	36	3772	3779	46	48	3853	3790	107
5	200	hotmail.com	57267	3846	34	3828	15	19	3775	75	3754	86	3814	3824	41	23	3880	3825	80
6	200	seznam.cz	33210	2162	8	2163	8	2	2165	8	2136	39	2164	2167	6	14	2186	2165	33
7	200	stonline.sk	12894	840	14	836	8	6	838	11	833	9	831	828	9	16	849	833	40
8	200	centrum.sk	9169	1048	14	1055	8	8	1058	20	625	19	641	352	6	24	375	290	12
9	200	zoznam.sk	7111	358	5	360	1	3	359	10	361	5	364	364	7	22	386	369	41
10	200	icloud.com	6844	442	6	444	4	2	447	10	450	3	446	448	9	5	455	449	22

In my opinion there is a possibility to deal with this problem by changing ratio of emails sent to gmail domain. There are various mechanisms responsible for mailing queue. From my experience I know that in case of Python there is Celery service responsible for queue management. We don't know what kind of mechanism is utilized in this case but I suppose that there is a possibility to decrease ratio of message sending.

Code 605



```
select code, sum(total) from datatable where code !='200' group by code order by sum(total) desc
```

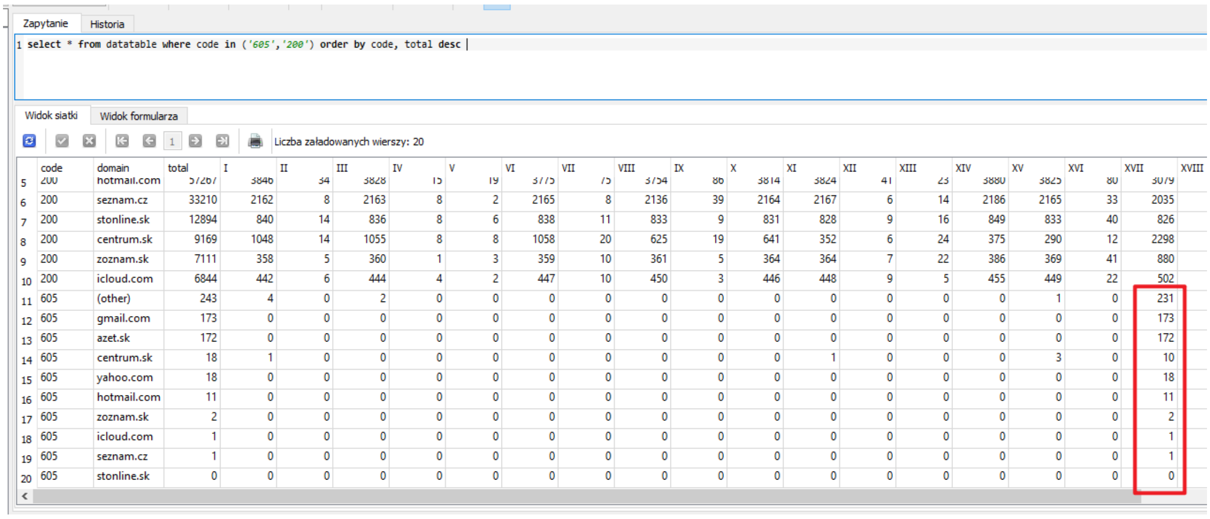
Liczba załadowanych wierszy: 9

	code	sum(total)
1	554	2835
2	452	721
3	605	639
4	550	636
5	552	450
6	499	266
7	451	210
8	421	204
9	(other)	197

It was difficult to find the meaning of code 605 and I'm still not 100% sure if meaning found by me is correct in case of our data. I have found information about this error on Mailgun and Current RMS websites, both related to mail sending so I believe it is applicable in case of our data. Description in both cases points that email address is suspended from delivery attempts due to previous failure. Mentioned failure may be caused by various factors so it's difficult to provide successful solution. Anyway I would recommend to verify if email address is correct (ex. with no spelling mistake) because it is the type of mistake which leads to 605 error in long run.

On the other hand we can look at this error from another perspective:

*select * from datatable where code in ('605','200') order by code, total desc*



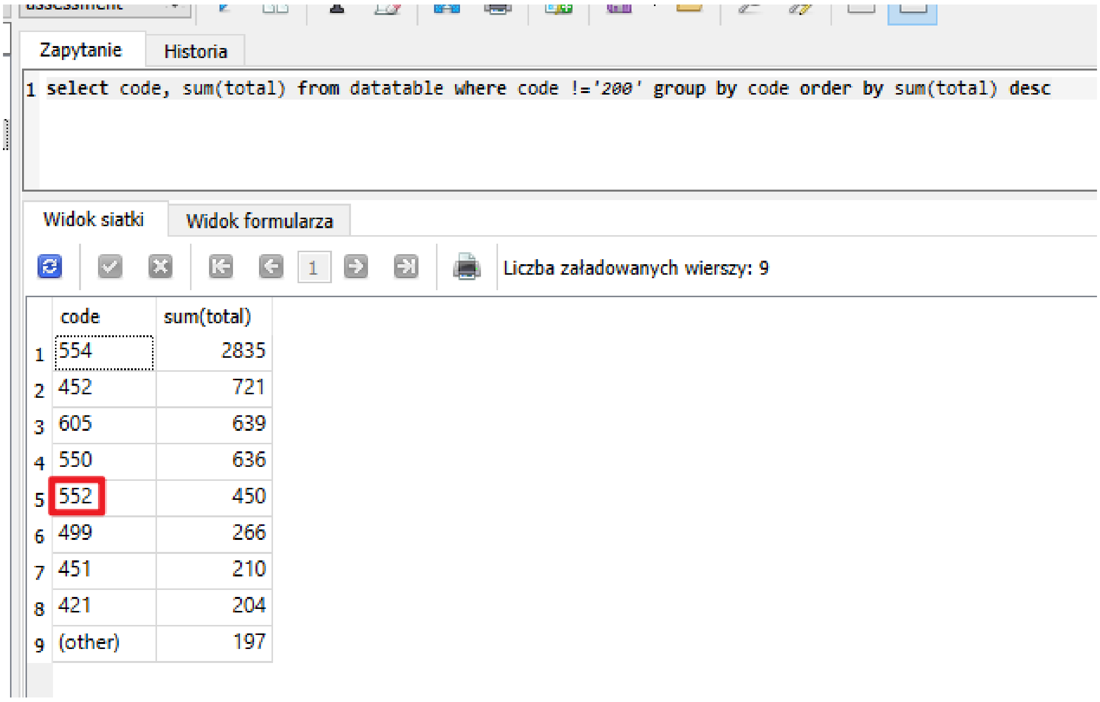
```
1 select * from datatable where code in ('605','200') order by code, total desc
```

Liczba załadowanych wierszy: 20

	code	domain	total	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	XVI	XVII	XVIII
5	200	hotmail.com	57267	3846	34	3828	15	19	3775	75	3754	86	3814	3824	41	23	3880	3825	80	3079	
6	200	seznam.cz	33210	2162	8	2163	8	2	2165	8	2136	39	2164	2167	6	14	2186	2165	33	2035	
7	200	stonline.sk	12894	840	14	836	8	6	838	11	833	9	831	828	9	16	849	833	40	826	
8	200	centrum.sk	9169	1048	14	1055	8	8	1058	20	625	19	641	352	6	24	375	290	12	2298	
9	200	zoznam.sk	7111	358	5	360	1	3	359	10	361	5	364	364	7	22	386	369	41	880	
10	200	icloud.com	6844	442	6	444	4	2	447	10	450	3	446	448	9	5	455	449	22	502	
11	605	(other)	243	4	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	231	
12	605	gmail.com	173	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	173	
13	605	azet.sk	172	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	172	
14	605	centrum.sk	18	1	0	0	0	0	0	0	0	0	0	1	0	0	0	3	0	10	
15	605	yahoo.com	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18	
16	605	hotmail.com	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	
17	605	zoznam.sk	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	
18	605	icloud.com	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
19	605	seznam.cz	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
20	605	stonline.sk	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Almost every single example of 605 code occurred at the same date for every domain. In this case it is difficult to point root cause and possible solution but I would point next step of investigation – we should take a closer look into emails sent in that specific date.

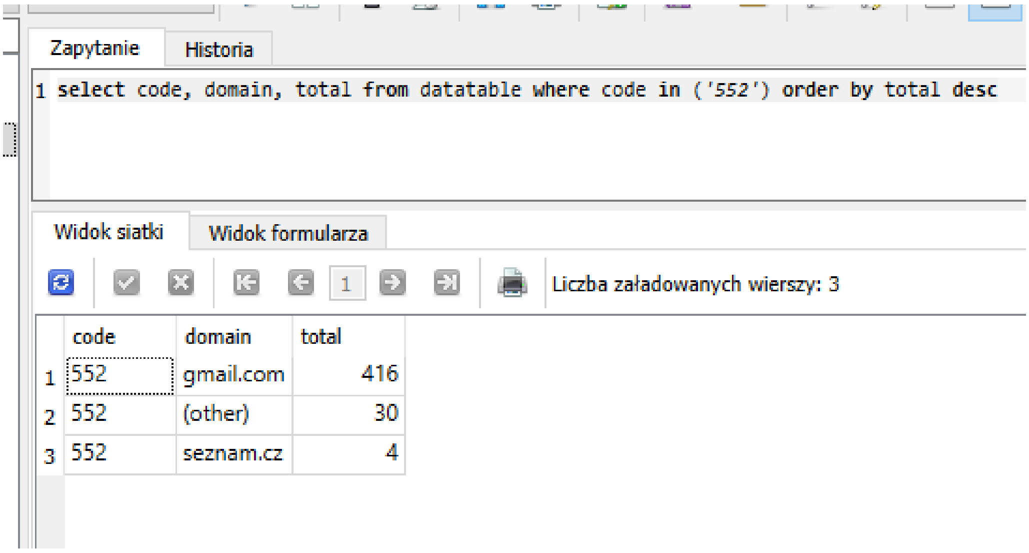
Code 552



```
select code, sum(total) from datatable where code !='200' group by code order by sum(total) desc
```

	code	sum(total)
1	554	2835
2	452	721
3	605	639
4	550	636
5	552	450
6	499	266
7	451	210
8	421	204
9	(other)	197

552 code is related to mailbox restrictions. Customers can adjust maximum size of received messages. If our mail exceeds this value it is rejected with code 552 ('5.3.4 Message size exceeds fixed maximum message size'). Another reason for 552 code reception could be the type of attachment (error 552 – '5.7.0 Our system detected an illegal attachment on your message'). Let's check which domains are affected:

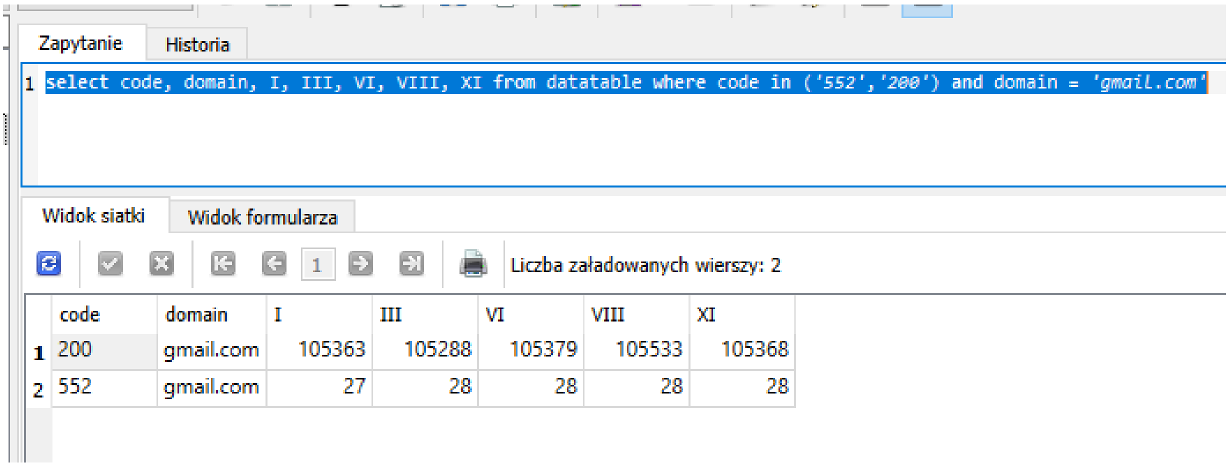


```
select code, domain, total from datatable where code in ('552') order by total desc
```

	code	domain	total
1	552	gmail.com	416
2	552	(other)	30
3	552	seznam.cz	4

We can check file types blocked by Gmail here but in my opinion probably it is not a problem due to the fact that so many messages were delivered successfully at the same date and I suppose that content of both delivered and rejected messages were the same:

select code, domain, I, III, VI, VIII, XI from datatable where code in ('552','200') and domain = 'gmail.com'



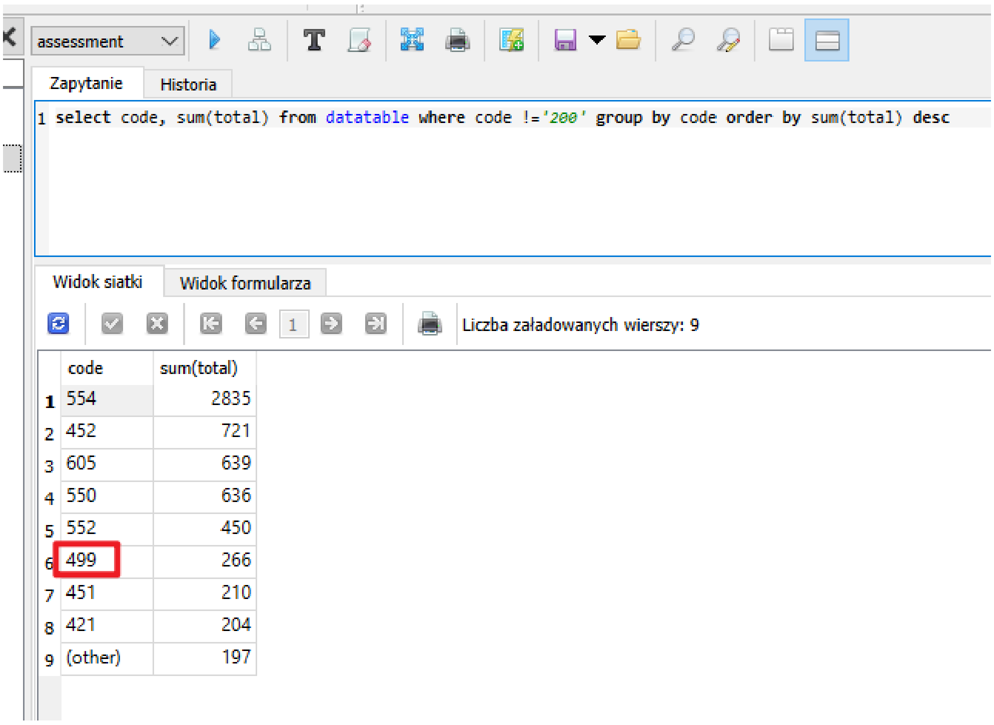
```
select code, domain, I, III, VI, VIII, XI from datatable where code in ('552','200') and domain = 'gmail.com'
```

Liczba załadowanych wierszy: 2

	code	domain	I	III	VI	VIII	XI
1	200	gmail.com	105363	105288	105379	105533	105368
2	552	gmail.com	27	28	28	28	28

Moreover we can see that number of returned 552 codes is more less the same for different dates so it convinces that particular email recipients set message size restrictions. It is possible to set in Gmail admin console. In May maximum number of rejected messages with 552 code was 31. It is not big number taking into consideration that hundreds of thousands messages are delivered successfully to Gmail at the same time. But if we want to reach affected recipients we could prepare dedicated message for them, with limited size.

Code 499



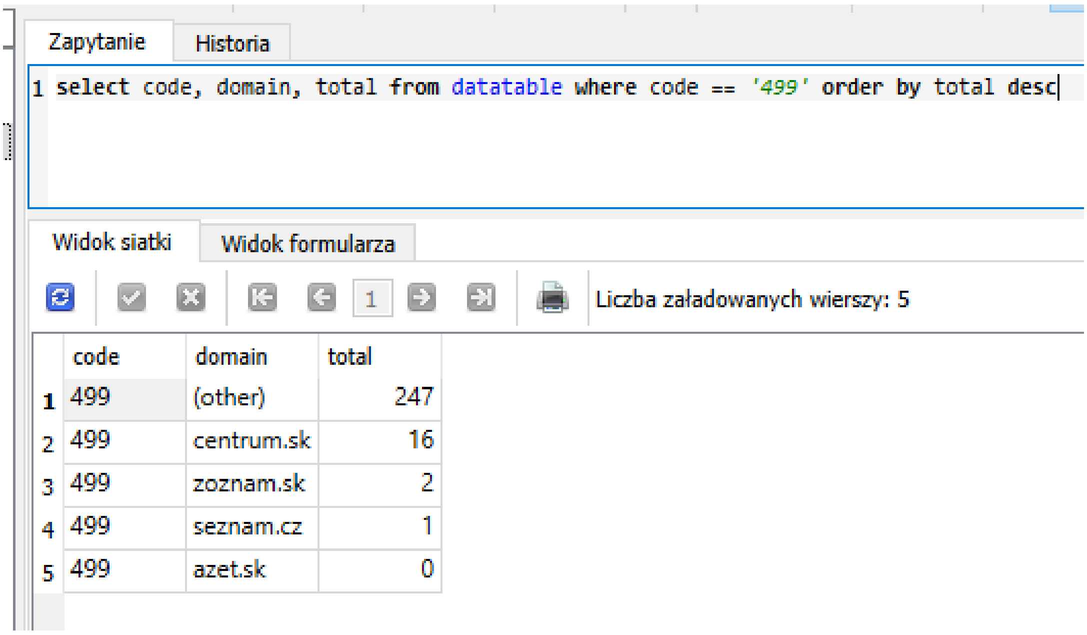
```
select code, sum(total) from datatable where code !='200' group by code order by sum(total) desc
```

Liczba załadowanych wierszy: 9

	code	sum(total)
1	554	2835
2	452	721
3	605	639
4	550	636
5	552	450
6	499	266
7	451	210
8	421	204
9	(other)	197

As in case of code 605 it is difficult to find meaning of this code. So let's try to check if it is typical for some particular domain:

select code, domain, total from datatable where code == '499' order by total desc



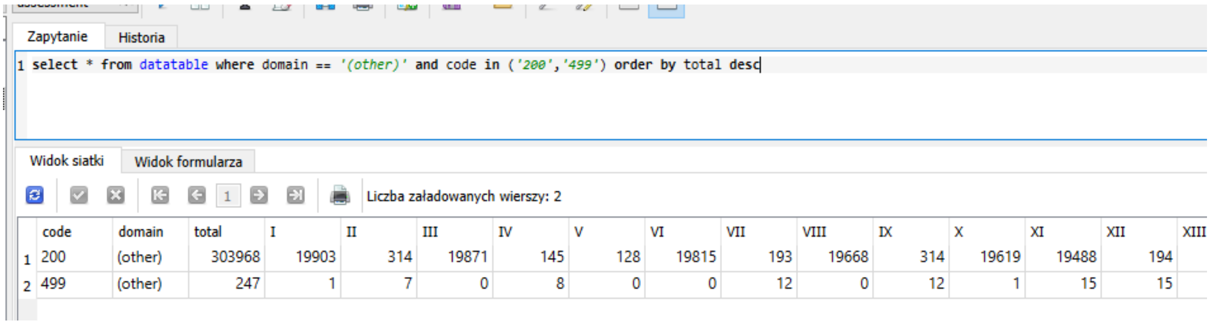
```
select code, domain, total from datatable where code == '499' order by total desc
```

Liczba załadowanych wierszy: 5

	code	domain	total
1	499	(other)	247
2	499	centrum.sk	16
3	499	zoznam.sk	2
4	499	seznam.cz	1
5	499	azet.sk	0

As we see it is impossible to point any domain in this case. So last thing we can do is to check how it occurred during entire month:

*select * from datatable where domain == '(other)' and code in ('200','499') order by total desc*



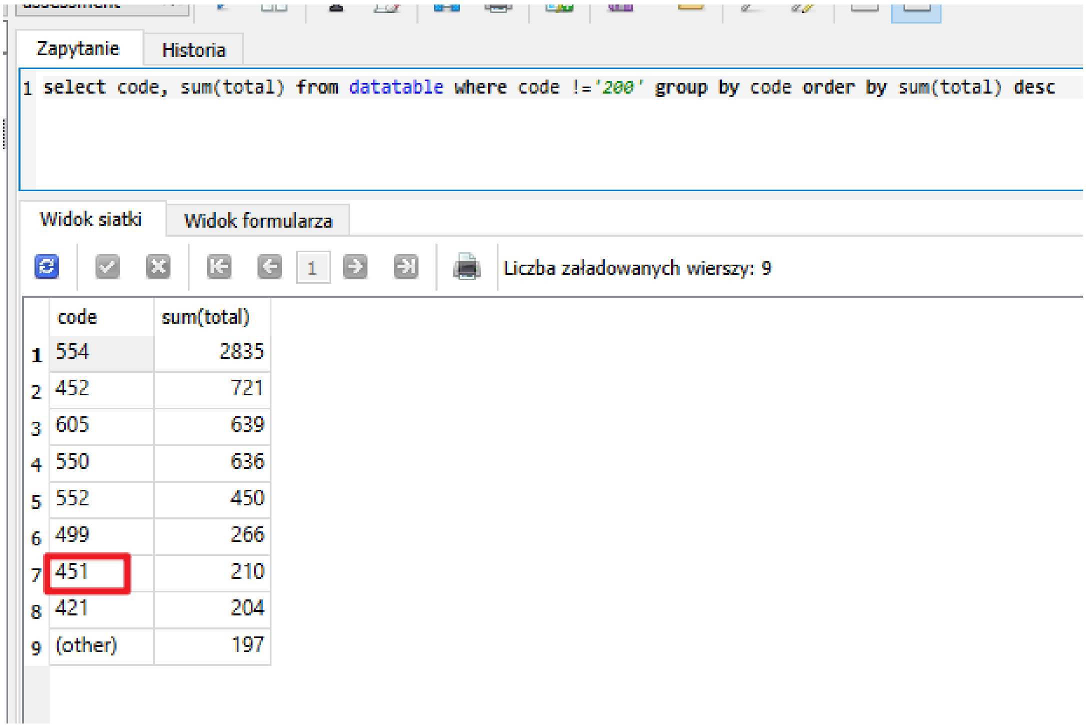
```
select * from datatable where domain == '(other)' and code in ('200','499') order by total desc
```

Liczba załadowanych wierszy: 2

	code	domain	total	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII
1	200	(other)	303968	19903	314	19871	145	128	19815	193	19668	314	19619	19488	194	
2	499	(other)	247	1	7	0	8	0	0	12	0	12	1	15	15	

As we can see the occurrence of errors seems not to be correlated to overall number of sent messages. Similar number of errors are returned when huge number of messages are sent and when only few messages are sent (see day VII and XI).

Code 451



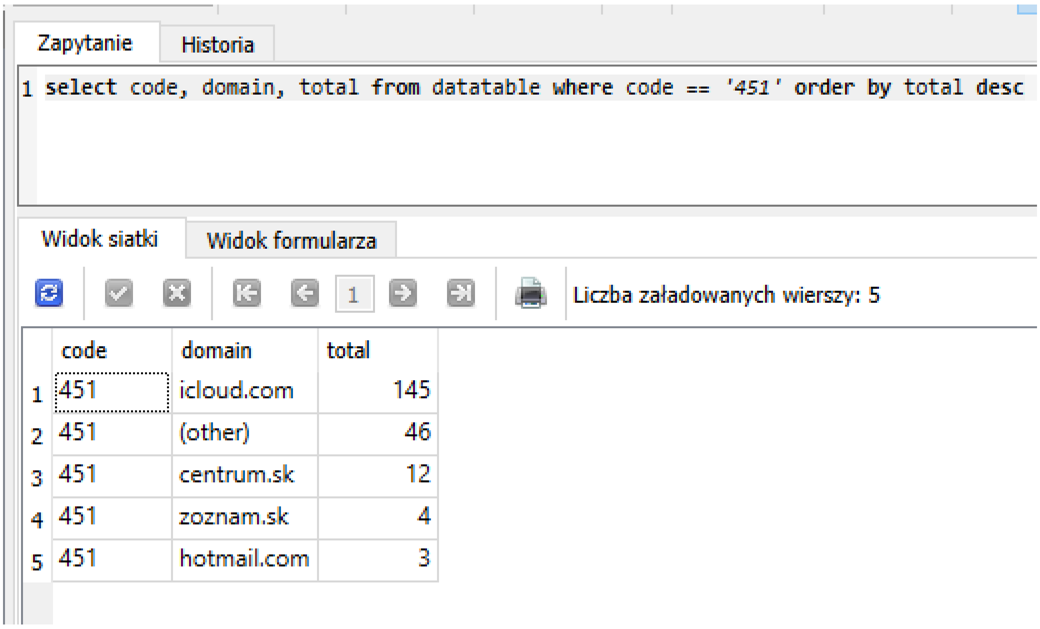
```
select code, sum(total) from datatable where code !='200' group by code order by sum(total) desc
```

Liczba załadowanych wierszy: 9

	code	sum(total)
1	554	2835
2	452	721
3	605	639
4	550	636
5	552	450
6	499	266
7	451	210
8	421	204
9	(other)	197

This error in most cases is a result of server temporary problem (see here). As usual let’s check affected domains:

select code, domain, total from datatable where code == '451' order by total desc



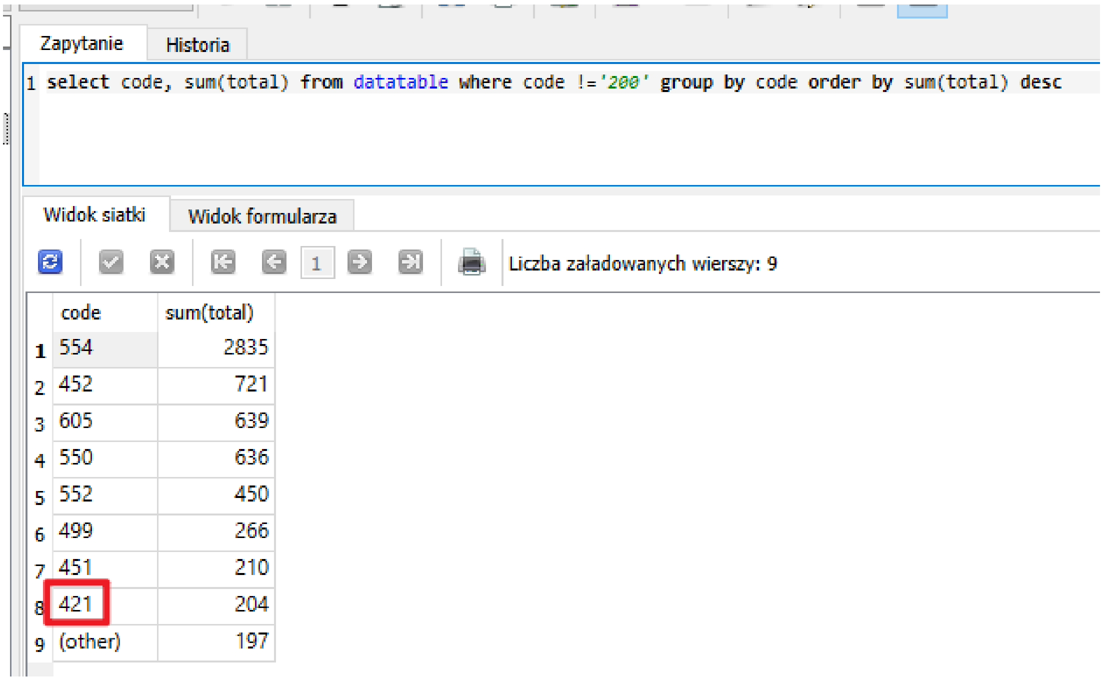
```
select code, domain, total from datatable where code == '451' order by total desc
```

Liczba załadowanych wierszy: 5

	code	domain	total
1	451	icloud.com	145
2	451	(other)	46
3	451	centrum.sk	12
4	451	zoznam.sk	4
5	451	hotmail.com	3

As we see icloud is domain with the biggest number of failures. As it is temporary unavailability problem I would recommend to do retry for failed messages.

Code 421



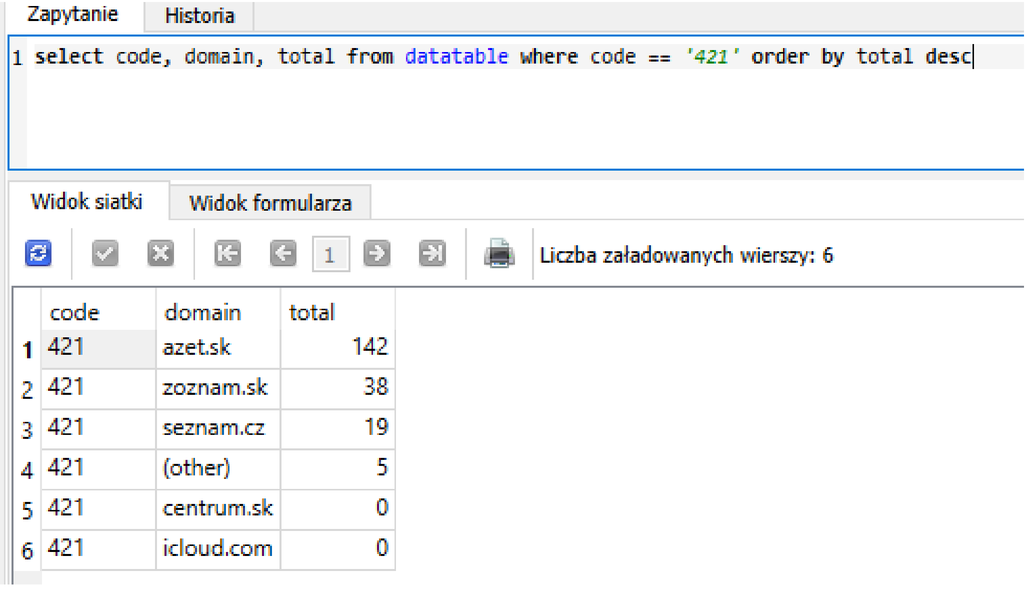
```
select code, sum(total) from datatable where code !='200' group by code order by sum(total) desc
```

Liczba załadowanych wierszy: 9

	code	sum(total)
1	554	2835
2	452	721
3	605	639
4	550	636
5	552	450
6	499	266
7	451	210
8	421	204
9	(other)	197

Last code is 421. It also related to temporary unavailability. It might be caused by events like too many connections, busy server or rebooting mail server. I suppose that cause of problem is the same as in case of 451 and we distinct both 421 and 451 because one domain in certain conditions returns 421 while other under the same conditions returns 451. Let's list domains affected with 421 and compare it with analogical list for 451:

select code, domain, total from datatable where code == '421' order by total desc



```
select code, domain, total from datatable where code == '421' order by total desc
```

Liczba załadowanych wierszy: 6

	code	domain	total
1	421	azet.sk	142
2	421	zoznam.sk	38
3	421	seznam.cz	19
4	421	(other)	5
5	421	centrum.sk	0
6	421	icloud.com	0

We can see that both lists contain different sets of domains. As previously I would recommend retry.

Step 5 – Final conclusions

It is not possible to point one general root cause and propose one general solution as we face variety of codes. General conclusion could be that if we want to catch the root cause more precisely we need to verify messages received with codes. But with already known set of data we can list following assumptions:

- most common code **554** is most probably caused by SPAM filter on recipients side – mailing tool could be reconfigured the way it avoids firewall
- **550** code is returned in case of invalid mail address – I would recommend to verify if involved addresses are the same as requested by client (e.g. no undesirable mark while copying added) and notify client with list of invalid addresses

In both cases I would propose to develop dedicated ETL tool to support and automatize recommended activities. In first case ETL can extract affected addresses and notify proper team to adjust the tool. In second case it can extract invalid addresses and automatically notify client with list ofrecipients.

- 684 messages did not reach destination due to overload of Gmail (code **552**) – we can reduce it by limiting sending ratio in case of this domain
- it is difficult to explain the meaning of **605** code – if my hypothesis about blocking of previously failed attempts is correct it might help when solutions from above points will be implemented
- we can explain to the client that some recipients did not get the message because they limited allowed size of incoming messages (code **552**) – in addition we can propose that new message fitting the restrictions will be prepared
- both **451** and **421** codes describe temporary unavailability – I would recommend to point affected addresses and do retry after the time span