17th Prof. Vladas Gronskas International Scientific Conference
Kaunas: Vilnius University Kaunas Faculty, 2nd of December, 2022

Big Data Application for Traffic Estimation on a Website: Big Daddy Case

Shubham Udairaj Singh[1], Rūta Banelienė[2]

[1] *Master's student, Vilnius Gediminas Technical University, Plytinės St. 25, Vilnius, Lithuania, shubham-udairaj.singh@stud.vilniustech.lt*
[2] *PhD, Vilnius Gediminas Technical University, Plytinės St. 25, Vilnius, Lithuania, ruta.baneliene@vilniustech.lt*

Abstract. While living under rapidly changing conditions innovation, flexibility and readiness to change are grounding prosperity of the firm. But any changes for companies should be reasoned and made on the basis of analytical approach. Big data usually could help in this situation without spending time and money on expensive research activities. Therefore, this paper is focused on big data application on customers' behavior switching from one product to new its look. Modeling is based on few months' daily data with application of regression analysis and the least squares method. The major finding comes up with the estimation output that the new webpage is more popular among IOS and WEB users, although Android systems showing negative impact on switching to new website.

Keywords: *big data, customers' behavior, new product.*

Introduction

While living under rapidly changing conditions innovation, flexibility and readiness to change are grounding prosperity of the firm. But any changes for companies should be reasoned and made on the basis of analytical approach. Big data usually could help in this situation without spending time and money on expensive research activities. By the definition, big data could be described as datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze (Manyika et al., 2014; Gentsch, 2019). Companies such as Google, Apple, Facebook, Amazon and others invested their knowledge, time and money for creation of such big databases and their application on particular situations analysis (Dash et al., 2019). Other companies realize the benefits of big database creation and usage for particular purpose inside of company for identification and solving many current internal and external problems, launching their products, improving communication with customers, sales volume prediction and for many other purposes whose, at final stage, are focused on profit and prosperity of firm.

In addition, COVID-19 pandemic situation in globe enhanced e-commerce opportunities to serve customers skipping few chains in supply chain or organize them in much smarter way by using big data.

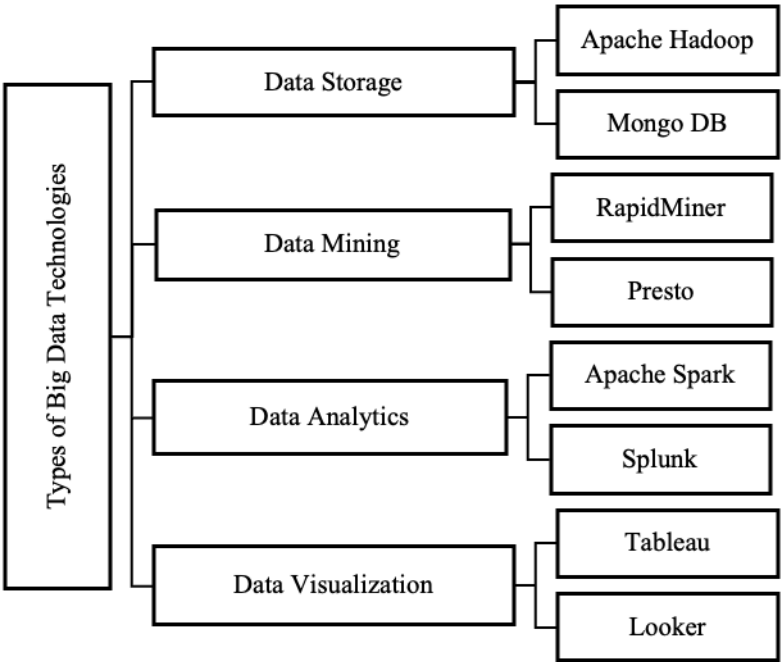
This paper is focused on big data application on customers' behavior switching from one product to new its look. The aim of this research is to investigate traffic on moving users to new website by using different devices. Our objectives are to evaluate the impact of switching users from old to new product – new website by focusing attention on users' devices – IOS, Android and WEB. Theoretical approach and empirical background overview options on usage of big data for chosen company, methodology uncovers model and data for particular case analysis, modeling output is presented in part of results, and part of conclusions and discussion outlines the research insights.

1. Theoretical approach

Big data could be integrated into the firm' activities via their connection to firm strategy, organizational activities and decisions making. Although big data without qualified interpretation could be just an additional virtual space for which firms pay.

Big Data is a pool of massive information which is structured, unstructured and semi-structured gathered by organizations to classify and identify patters of the audience in order to make effective decisions of an organization's future (Laney, 2001). Big Data is effectively monitored by organizations by choosing the type of Big Data Technologies that exists based on the necessities, requirements and industry demands of the respective organization.

https://doi.org/10.15388/VGISC.2023.14



Source: based on Osadchuk (2022).

Fig. 1. Big Data Technologies

The Big Data of all the organization is segregated into any one of the data categories as shown in Fig. 1. and covers data storage, data mining, data analytics and data visualization. Banking and health care sector majorly relies on data storage as it consists of highly confidential information and needs the collected reports and documents to be stored and required with high level of monitoring to avoid security breaches and documentation theft (RapidMiner, 2022). E-commerce organizations highly rely on data mining as it repetitively advertises any new sale; discount offers, frequently searched items and reminds the customer with the help of search engine data.

There is a wide range of software's in the market to process and model the data available in the organization servers, search engines storage cloud. The software's used differ from organization to organization based on the industry type, needs, targets and goals. Few examples of big data analyzing software's Apache Hadoop and MongoDB are provided. Apache Hadoop is an open platform source which is highly scalable and reliable for processing large data sets using basic programming models (Ahmed et al., 2020). Hadoop is a cost-effective solution for processing huge amounts of unstructured data and does not have any format requirements (Hadoop, 2018). MongoDB is a horizontal scale out software that is used for data storing. MongoDB which is found in the year 2007 is an open-source document database system which is very flexible and allows variations in data gathered (Kinsta, 2023). MongoDB allows developers and analysts to analyze data very quickly in a scalable way (MongoDB, 2017).

2. Empirical background and data analysis

The organization chosen for the research paper is a Danish multinational banking and financial services corporation. The organization is a retail bank that operates in Northern European region and has more than 5 million retail customers and 22,376 employees as of the year 2020 (Danske bank, 2022). The organization chosen has a history of 150 years and is the largest bank in Denmark with a net worth of 4.589 trillion DKK.

Danica pension in daily activities uses these types of Big Data technologies for data storage: Apache Hadoop and Mongo DB; for data mining: RapidMiner, Presto; for data analytics: Apache Spark, Splunk; and for data visualization: Tableau, Looker.

The structured data used for the research paper is quantitative which has the numerical data of number of people who visited the Danica pension website through various operating devices like

IOS, Android and Webpage. Among the huge data stored in the dedicated databases system of the organization, the "Health Insurance" tab has been chosen for research and modeling purposes, as it has the higher range of numerical data compared to other tabs in the organizations website. The Big Data gathered by using Apache Hadoop for the storage, Presto for data mining and Splunk for data analytics, and is a collection of facts such as words, measurements, observations that provides more information about a customer observation.

BIG DADDY daily data was chosen for particular case study for July-September 2022. During the mentioned period of time BIG DADDY webpage has been visited for 156 156 times: 96 048 (61.5%) – the 'new' webpage version and 60 108 (38.5%) – the 'old' webpage version. Particular 651 340 clicks by devices in numbers: IOS 404 775, ANDROID 57 691, WEB 188 874 (see Fig. 2).

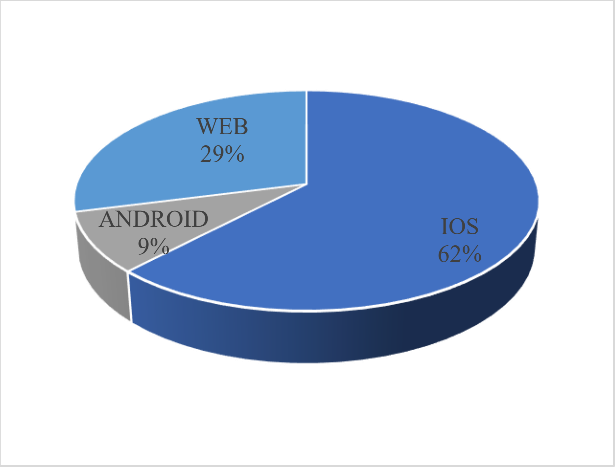


Fig. 2. Usage of the 'new' webpage by devices: accumulated data for July–September 2022

As per the BIG DATA collected from the chosen organization, it can be observed that the amount of Android users is very small compared to other IOS devices and web users (see Fig. 2 and Fig. 3).

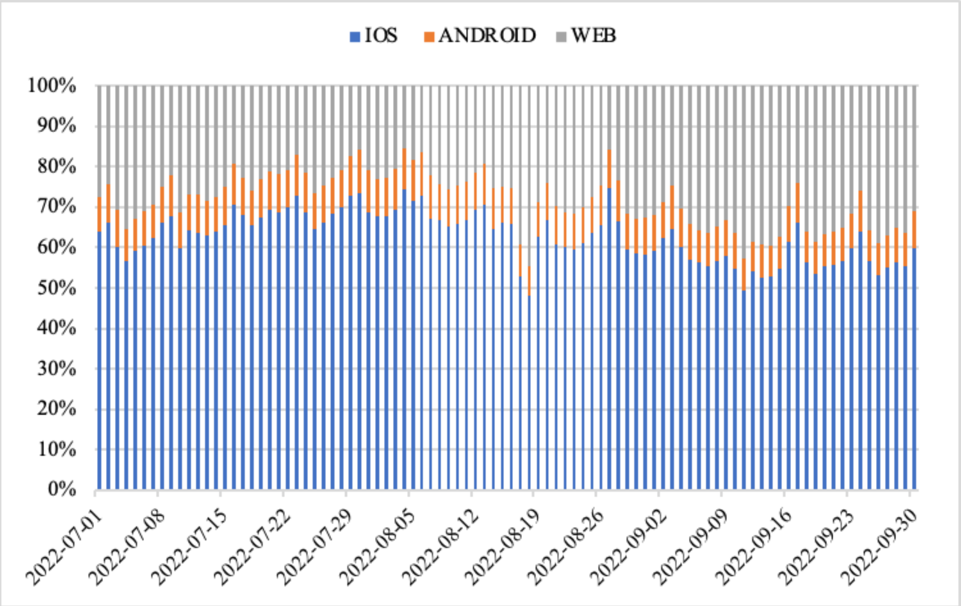


Fig. 3. Usage of the 'new' webpage by devices: daily data for July-September 2022

Also, we have focused on the number of errors that occurred to various customers due to different reasons while they were on the webpage. We have divided these errors into two different segments, Front end/Midrange errors and Backend Errors. The main reason for collecting this data is to avoid these scenarios for the best customer performance and also, the Development team can fix the bugs before they release the next version of this webpage.

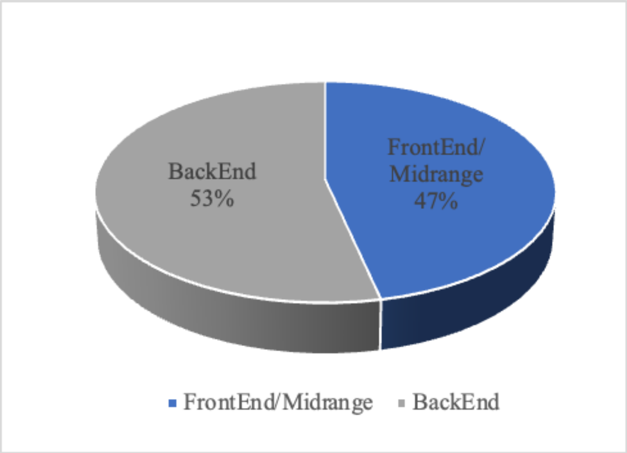


Fig. 4. Reasons of 'Oops' pages

There might be a possibility that organization may decide to switch completely from Android devices to WEB and IOS devices due to the number of errors (Front end, Backend and Midrange) that are occurring due to various reasons on Android devices. So, to increase the traffic on the website, the organizations can choose better customer performance and better customer service if they want to continue with the Android devices. In other words, there might be a possibility that as the complexity of the webpage increases the performance of the Android operating system starts decreasing significantly so if the organizations want to keep the Android users they have to start focusing on the back end and front end development of those websites and have to monitor where the errors are occurring. They should find the root cause of these error messages and fix them permanently.

BIG DADDY data defines the situation where data sets have grown to such enormous sizes that a traditional information technology cannot effectively handle the size of the data set, the volume and growth of the data set. Even if the data was somehow handled without the use of BIG DADDY data, but the amount of time it will take to produce the results it will be not an appropriate choice for multinational companies and it will be extremely hard to determine the accuracy of the data.

3. Methodology: model and data

Seeking to understand better the customers' behavior on launching marketing innovation – new BIG DADDY webpage, two steps model has been developed. First step was focused on the ways by which webpage was reached (see Fig. 5).

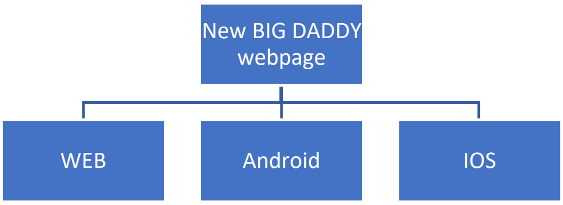


Fig. 5. Model on customers' behaviour switching from 'old' to 'new'

On the second step, regression analysis has been chosen for modeling where dependable variable is number of webpage visits and independent variables are the ways to open webpage:

$$BIG\ DADDY = c + IOS + ANDROID + WEB + \varepsilon \ (1)$$

The data that we have collected for this research is quantitative and consists of number of people in Denmark visiting to the Danica Pension website using different ways – for example IOS, Android and WEB. It can be either PC, Laptop Mobile or any IOS devices. There are different segments on this webpage and we focused on the segments where the traffic was a more as compared to others. In other words, we focused on collecting the data where customers have mostly clicked on Health & Insurances tab which is also known as BIG DADDY in technical terms.

According to statistical data which shows IOS users share 62%, WEB users 29% and Android users 9% shares of total clicks on new web page (see Fig. 2), we raised few related hypotheses which could help company to identify further steps related to new webpage traffic and development.

Hp1: Danica Pension website should be focused on IOS and WEB users due to their positive impact on clientele formation.

Hp2: Danica Pension website should be less focused on Android users due to their low impact on clientele formation.

The data for the months of July, August and September 2022 has been collected for studying and modeling. We have organized this data on the basis on number of logins on a daily basis for 3 months using different devices: 7/01/2022 – 9/30/2022 period data sample with 92 observations (see Annex 1). We have used EViews software for our modeling and its statistical representation. For modeling, the least squares method has been applied.

4. Results

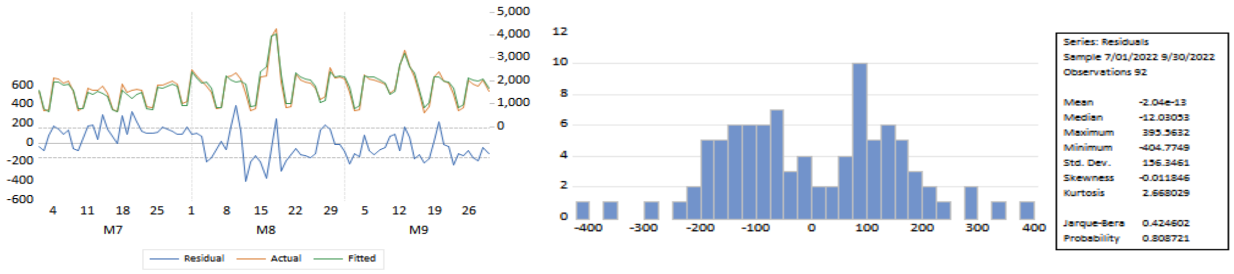
The estimation output by Equation 1 shows that the new webpage is more popular among IOS and WEB users, although Android systems showing negative impact on switching to new website (see Equation 2 and Table 1).

$$\mathit{BIG\ DADDY} = 218.30 + 0.27{*}\mathit{IOS} - 1.24{*}\mathit{ANDROID} + 0.51{*}\mathit{WEB} \quad (2)$$

Table 1

Estimation output

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	218.2985	50.25410	4.343895	0.0000
IOS	0.272161	0.070245	3.874458	0.0002
ANDROID	-1.241797	0.554260	-2.240459	0.0276
WEB	0.516477	0.025324	20.39470	0.0000
R-squared	0.947579	Mean dependent var		1697.348
Adjusted R-squared	0.945792	S.D. dependent var		682.8663
S.E. of regression	158.9887	Akaike info criterion		13.01805
Sum squared resid	2224412.	Schwarz criterion		13.12769
Log-likelihood	-594.8302	Hannan-Quinn criterion		13.06230
F-statistic	530.2424	Durbin-Watson stat		0.985519
Prob(F-statistic)	0.000000			



a) Residual, actual, fitted b) Normality test

Fig. 6. Modelling data analysis

The biggest negative impact on using new BIG DADDY web page was shown by Android users: one click on the new web page decreasing further intention to visit web site by 1.24 clicks.

Although positive impact on switching to the new web page was shown by IOS and, especially, by WEB users which have the highest impact on traffic extension of the new website.

Modeling results show high reliability of our model due to 0.95 R-squared and p-values which are below 0.01 with one exemption for ANDROID users – $p<0.05$. Graphical view of our model' reliability is shown in Fig. 6 (a) where the actual data are sufficiently reflected by fitted data – calculated by our model. Normality tests of estimation (Fig. 6 (b)) supports the validity of our modeling results by showing a nonrejection hypothesis on the data used for our modeling normal distribution.

Conclusions

As per the Big Data collected from the chosen organization, it can be observed that the amount of Android users is very less compared to other IOS devices and web users. In addition, estimation based on big data proved hypothesis that company should be focused on IOS and web users and pay less attention to Android due to its negative impact on clientele formation. There might be a possibility that organization may decide to switch completely from Android devices to WEB and IOS devices due to the number of errors (Front end, Backend and Midrange) that are occurring due to various reasons on Android devices. So, to increase the traffic on the website, better customer performance and better customer service the organizations can choose if they want to continue with the Android devices.

References

1. Ahmed, N., Barczak, A. L. C., Susnjak, T., & Rashid, M. A. (2020). A comprehensive performance analysis of Apache Hadoop and Apache Spark for large scale data sets using HiBench. *Journal of Big Data*, *7*, 110. https://doi.org/10.1186/s40537-020-00388-5
2. Danske Bank (2020). Danske Bank in numbers. Retrieved on November 20, 2022, Retrieved from https://danskebank.com/
3. Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. *Journal of Big Data, 6,* 54. Retrieved from https://link.springer.com/article/10.1186/s40537-019-0217-0
4. Gentsch, P. (2019). *AI in marketing, sales and service. How marketers without a data science degree can use AI, big data and bots.* Palgrave Macmillan.
5. IBM. (2018). Apache Hadoop. Retrieved on November 20, 2022, Retrieved from https://www.ibm.com/au-en/analytics/hadoop
6. Kinsta®. (2023). What is MongoDB? All about the popular open-source database. Retrieved on February 6, 2023, Retrieved from https://kinsta.com/knowledgebase/what-is-mongodb/
7. Laney, D. (2001). 3D Data management: controlling data volume, velocity, and variety. *META Group Research Note, 6.*
8. Manyika, J. et al. (2014). *Small states: economic review and basic statistics*. *17*, Commonwealth Secretariat.
9. MongoDB. (2017). Why use MongoDB and when to use it? Retrieved on November 20, 2022, Retrieved from https://www.mongodb.com/why-use-mongodb
10. Osadchuk, S. (2022). Top big data technologies: Transform your business. DOIT Software. Retrieved from https://doi.software/blog/big-data-technologies#screen6
11. RapidMiner (2022). Data mining tools. Retrieved on November 20, 2022, from Retrieved from https://rapidminer.com/glossary/data-mining-tools

Annex 1

Data used for modeling: numbers of clicks by devices and logins to web

	Devices			Logins to web		
Date	IOS	ANDROID	WEB	BIG DADDY DATA	NEMID LOGINS	OLD SOLUTION
2022-07-01	4091	535	1762	1538	889	649
2022-07-02	1890	268	699	684	367	317
2022-07-03	1245	190	637	731	332	399
2022-07-04	4110	589	2567	2111	1351	760
2022-07-05	4379	591	2419	2074	1301	773
2022-07-06	4349	604	2223	1889	1188	701
2022-07-07	4674	612	2205	2004	1292	712
2022-07-08	4500	599	1702	1524	933	591
2022-07-09	2272	339	737	711	380	331
2022-07-10	1511	222	787	804	419	385
2022-07-11	4039	549	1689	1681	897	784
2022-07-12	3888	585	1632	1577	865	712
2022-07-13	3936	531	1776	1580	985	595
2022-07-14	3734	488	1605	1750	857	893
2022-07-15	3730	537	1423	1450	741	709
2022-07-16	2142	312	585	782	286	496
2022-07-17	1621	219	544	664	285	379
2022-07-18	4364	584	1724	1850	949	901
2022-07-19	4220	582	1448	1486	746	740
2022-07-20	3937	538	1208	1574	628	946
2022-07-21	4508	623	1421	1629	770	859
2022-07-22	4868	646	1447	1607	672	935
2022-07-23	2548	356	591	874	309	565
2022-07-24	1960	279	613	825	330	495
2022-07-25	4737	655	1942	1811	967	844
2022-07-26	4898	668	1824	1826	925	901
2022-07-27	5298	669	1772	1888	896	992
2022-07-28	6212	825	1860	1965	950	1015
2022-07-29	6871	932	1638	1864	785	1079
2022-07-30	3408	497	728	1000	360	640
2022-07-31	2870	434	874	1074	445	629
2022-08-01	7599	1006	2600	2471	1341	1130
2022-08-02	6942	951	2335	2231	1218	1013
2022-08-03	6836	984	2028	1969	1046	923
2022-08-04	8261	1120	1704	1759	780	979
2022-08-05	6320	886	1610	1513	784	729
2022-08-06	3005	445	677	764	309	455
2022-08-07	2456	396	805	831	358	473
2022-08-08	6707	880	2449	2149	1207	942
2022-08-09	6055	839	2374	2224	1312	912
2022-08-10	6000	857	2239	2339	1157	1182
2022-08-11	6316	899	2231	2096	1184	912
2022-08-12	6127	807	1886	1453	955	498
2022-08-13	2847	412	782	686	370	316
2022-08-14	2415	375	951	769	452	317
2022-08-15	7212	970	2721	2183	1417	766
2022-08-16	7679	1003	2957	2222	1502	720
2022-08-17	7924	1163	5878	3904	2939	965
2022-08-18	6750	1011	6262	4284	3152	1132
2022-08-19	6064	844	2771	1953	1394	559
2022-08-20	2765	386	989	813	525	288
2022-08-21	2418	377	1179	892	572	320
2022-08-22	5934	869	3082	2292	1604	688
2022-08-23	5347	790	2845	2035	1449	586
2022-08-24	5390	794	2656	1939	1420	519
2022-08-25	5632	786	2428	1878	1291	587
2022-08-26	5532	828	2075	1655	1086	569
2022-08-27	3719	473	789	1182	396	786
2022-08-28	3519	540	1232	1327	630	697
2022-08-29	6076	906	3216	2548	1594	954
2022-08-30	5145	766	2880	2142	1452	690
2022-08-31	5338	821	2982	2174	1489	685
2022-09-01	5250	767	2835	2068	1415	653
2022-09-02	4394	618	2039	1479	1033	446
2022-09-03	2193	358	836	690	412	278
2022-09-04	2006	313	1012	755	490	265
2022-09-05	5063	789	3051	2275	1577	698
2022-09-06	4598	662	2912	2069	1474	595
2022-09-07	4510	655	2957	2035	1502	533
2022-09-08	4533	693	2783	1964	1453	511
2022-09-09	4494	692	2570	1857	1335	522
2022-09-10	2890	482	1917	1467	972	495
2022-09-11	2469	381	2134	1609	1071	538
2022-09-12	5255	700	3760	2643	1929	714
2022-09-13	6425	1019	4782	3333	2375	958
2022-09-14	4991	725	3717	2651	1925	726
2022-09-15	4701	667	3200	2156	1631	525
2022-09-16	4069	609	1963	1463	1031	432
2022-09-17	2091	302	760	598	365	233
2022-09-18	1930	268	1234	884	597	287
2022-09-19	4136	593	2980	2162	1549	613
2022-09-20	4480	648	2970	2386	1523	863
2022-09-21	4195	631	2720	1970	1371	599
2022-09-22	4206	606	2589	1913	1296	617
2022-09-23	3911	553	2056	1426	959	467
2022-09-24	2104	338	853	700	387	313
2022-09-25	1777	243	1117	842	528	314
2022-09-26	4074	616	2989	2030	1432	598
2022-09-27	4131	609	2778	1864	1292	572
2022-09-28	4254	626	2656	1782	1212	570
2022-09-29	4337	638	2855	2034	1396	638
2022-09-30	4168	648	2154	1547	1031	516

Source*: Danica Pension* data (2022).