GENERAL OUTLINE OF PROJECT

Tech-online shops analysis

The project's main goal was to compare prices of given electronic devices in biggest polish shops specializing in electronics. We wanted to show the clearly visible differences in prices of:

Macbooks **Airpods** **Motorolas**

The data we needed were - prices in PLN and a string with shop's name or url address in which that price was applied to that particular searched item (including currently applied discounts made by those shops). It can be very useful for people searching for a well priced devices for themselves, their family etc. or for small businesses that want to increase their resources (people or just emergency devices in case any breaks) while not having to pay the biggest possible price. It might help them in for example todays financially hard times regarding to currently prevailing pandemic situation. We wanted to automate the process of searching to make it as least time consuming as possible and to make it possible to find the lowest possible price of a certain item.

Data we used was mainly from website https://www.ceneo.pl/ . Since it's already gathered there from shop's websites that offer these particular products, we would call it **secondary data**.

https://www.ceneo.pl/99105003 https://www.ceneo.pl/87644960 https://www.ceneo.pl/95623431

The data is as comparable, accessible and relevant as it could be, since the website we used gets it directly from websites of shops that offer these products. Until it doesn't go out of stock, we won't have to worry that it will not be available in the future. The data is also not ambigous and it's not so complex too, we gather it in arrays of prices (number elements parsed to string) and shop names (strings) and for example the price of index = 1 is from the shop of that exact index in the second array.

About usability we can consider additional resources or risk analysis:

- Improving new skills needed to use and analyse the data,
- Accumulate a great amount size of the dataset of different products.

PROJECT IMPLEMENTATION

The data was processed using **web scraping** method, we gathered it from a website that shows selected item in various shops with their price of that item.

It resulted in two lists of data - an array of shop names and a second array of prices, where every string of a certain index in the first array regarded the price of the product in the shop with that name with the same index in the second array, for example:

['mediaexpert.pl', 'lantre.pl', 'pcoutlet.pl', 'morele.net', 'fotoplus.pl', 'allegro.pl', 'x-kom.pl', 'neo24.pl', 'idream.pl', 'al.to', 'switchit.pl', 'acom.pl', 'vobis.pl', 'neonet.pl', 'komputronik.pl']
['5199', '5149', '5149', '5149', '5199', '5199', '5199', '5199', '5199', '5199', '5199', '5199', '5199', '5290', '5290']

['allegro.pl', 'mediaexpert.pl', 'mediamarkt.pl', 'x-kom.pl', 'komputronik.pl', 'oleole.pl', 'al.to', 'avans.pl', 'electro.pl', 'euro.com.pl', 'tele-port.com.pl', 'sferis.pl', 'deluxury.pl', 'neonet.pl', 'vobis.pl']
['851', '919', '919', '919', '919', '919', '919', '919', '919', '919', '939', '959', '965', '969', '970']

['premium-gsm.pl', 'net-s.pl', 'sferis.pl', 'rk-technology.com.pl', 'morele.net', 'wellsamed.pl', 'fatpc.pl', 'alo.com.pl', 'siglo.pl', 'domsary.eu', 'proshop.pl']
['1079', '1097', '1145', '1202', '1209', '1209', '1218', '1260', '1295', '1380', '1599']

Wordcloud

110 ['ams-region-start', '1248', 'By', 'continuing', 'to', 'use', 'AliExpress', 'you', 'accept', 'our']
109 ['ams-region-start', '1248', 'by', 'continuing', 'to', 'use', 'aliexpress', 'you', 'accept', 'our', 'use', 'of', 'cookies', 'view', 'more', 'on', 'our', 'privacy', 'policy', 'you', 'can', 'adjust', 'your', 'cookie', 'preferences', 'at', 'the', 'bottom', 'of', 'this', 'page', 'close', 'cookie', 'preferences', 'help', 'customer', 'service', 'disputes', 'reports', 'report', 'ipr', 'infringement', 'buyer', 'protection', 'app', 'switcher', 'start', 'ship', 'to', 'language', 'currency', 'save', 'switcher', 'end', 'cart', 'wish', 'list', 'account', 'welcome', 'to', 'aliexpress', 'welcome', 'back', 'sign', 'out', 'join', 'sign', 'in', 'my', 'orders', 'message', 'center', 'wish', 'list', 'my', 'favorite', 'stores', 'my', 'coupons', 'invite', 'friends', 'to', 'give', 'us', '24', 'ams-region-end', '1248', 'categorias', 'aliexpresssmarter', 'shopping', 'better', 'living', 'ams-region-start', '1261', '0', 'ams-region-end', '1261', 'all', 'categories', 'harajuku']

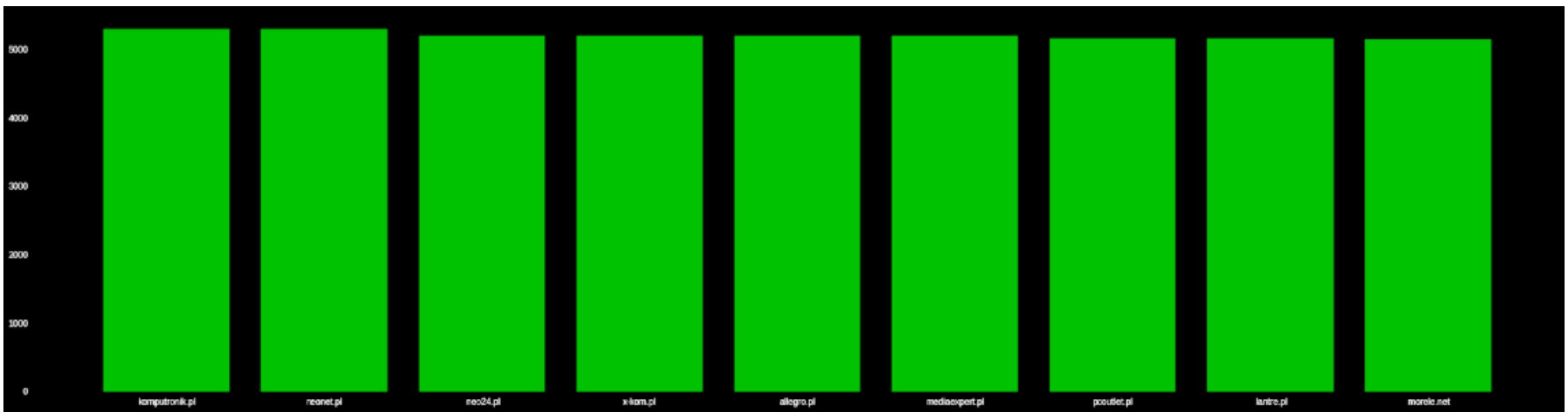
88 [('to', 4), ('my', 3), ('ams-region-start', 2), ('1248', 2), ('use', 2), ('aliexpress', 2), ('you', 2), ('our', 2), ('of', 2), ('cookie', 2)]

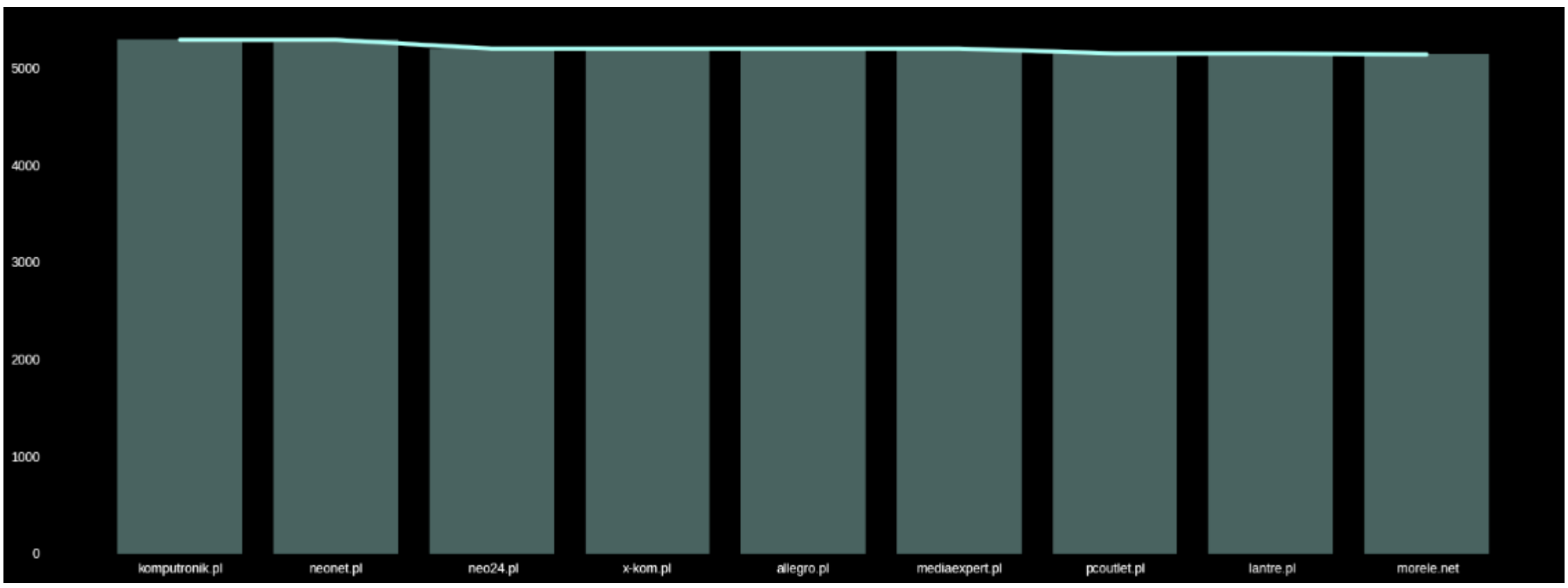
During the process, no major problems arose that would stop further work.

RESULTS OF ANALYSIS

Data visualisation distincted by the selected product:

- **Macbook**



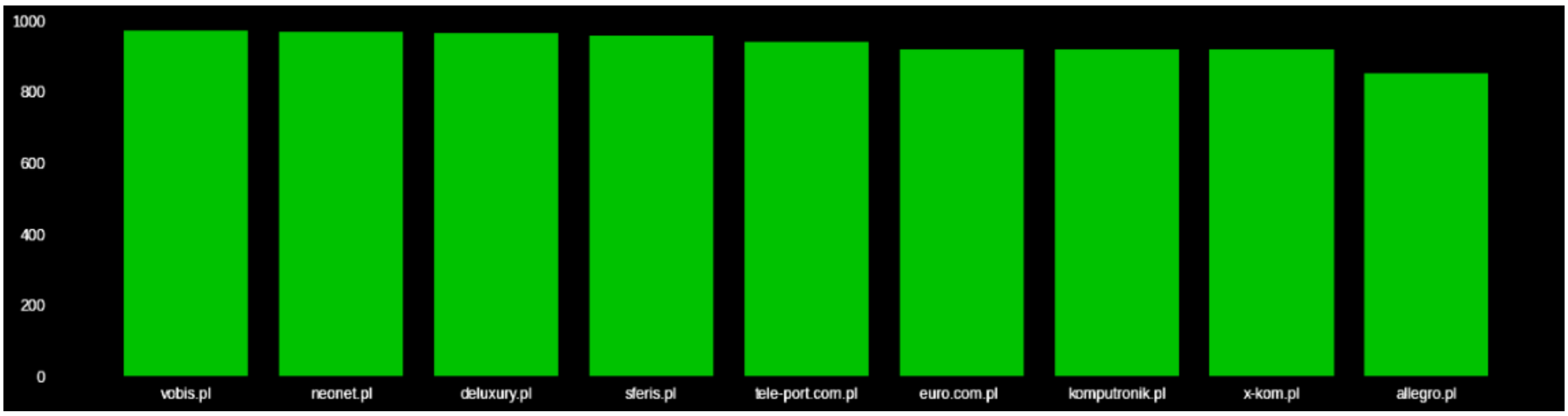


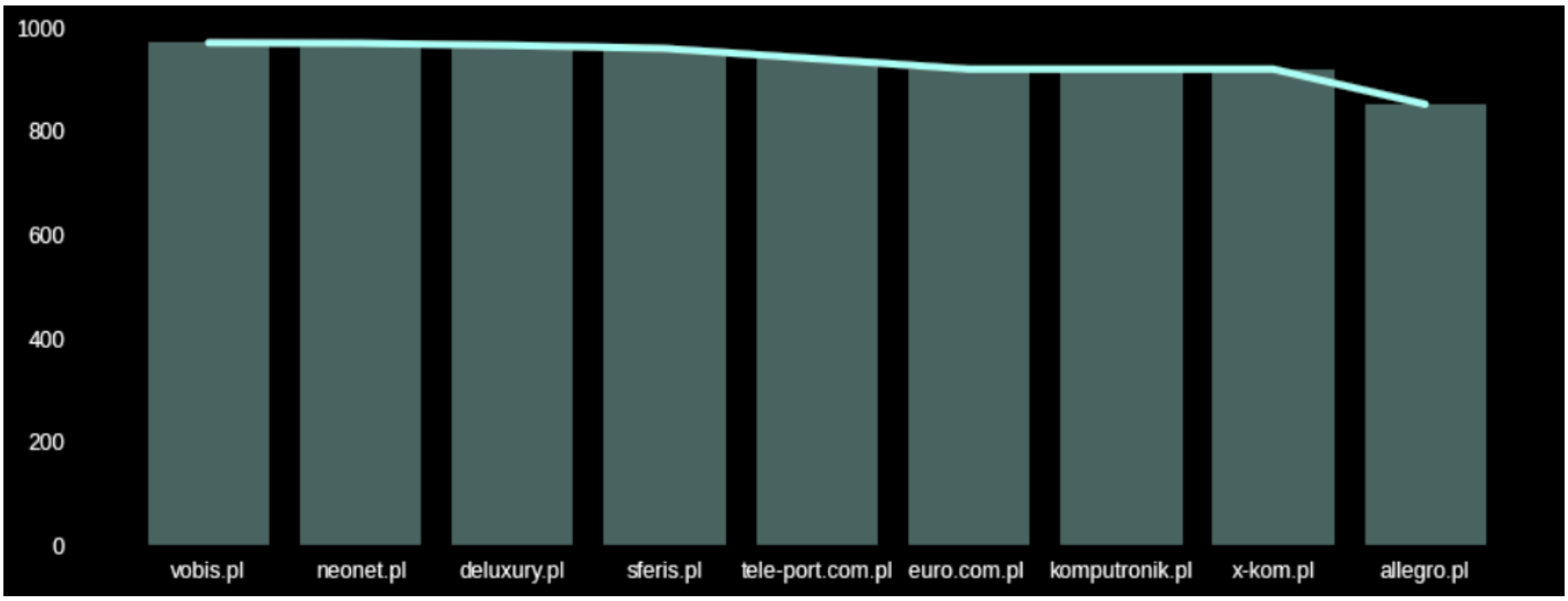
count	15,00
mean	5200,00
std	42,00
min	5139,00
25%	5199,00
50%	5199,00
75%	5199,00
max	5290,00

The highest price was 5290 PLN, while the lowest - 5139. The std of that statistic equals 42 PLN, it tells by how much on average the values of the examined feature deviate from the arithmetic mean. The median 5199, the midpoint value is the point at which half the observations are above the value and half the observations are below the value.

RESULTS OF ANALYSIS

- **Airpods Pro**



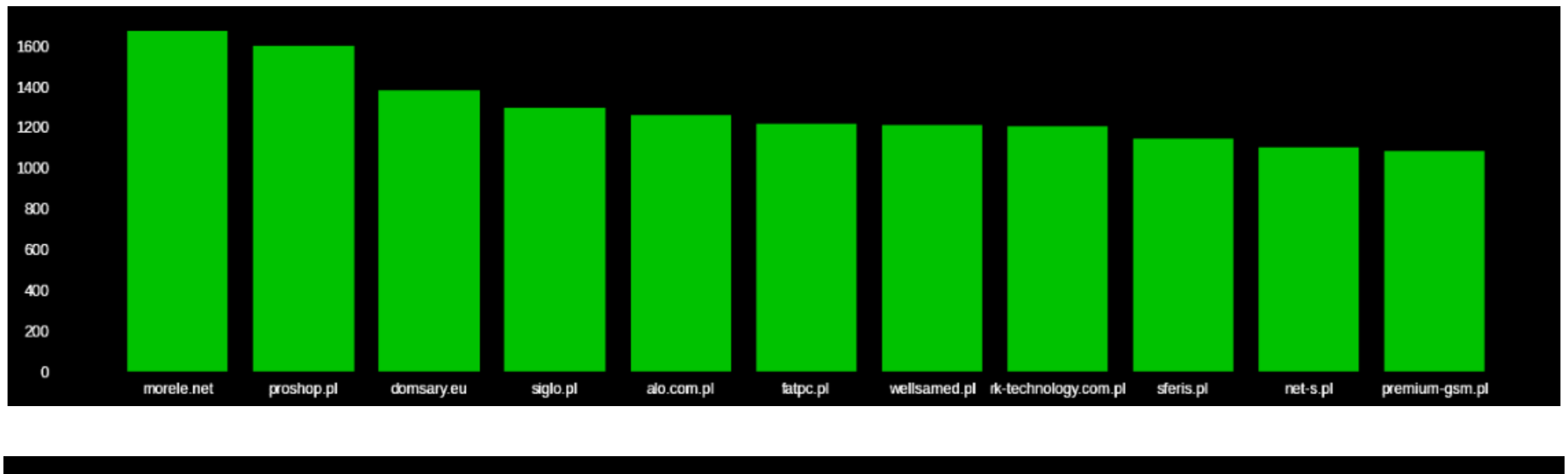


count	15,00
mean	928,00
std	30,00
min	851,00
25%	919,00
50%	919,00
75%	949,00
max	970,00

The highest price was 970 PLN, while the lowest - 851. The std of that statistic equals 30 PLN, it tells by how much on average the values of the examined feature deviate from the arithmetic mean. The median 919, the midpoint value is the point at which half the observations are above the value and half the observations are below the value. In case of the airpods the price varies in percentage were very similar to Macbook.

RESULTS OF ANALYSIS

- **Motorola**



count	11,00
mean	1245,00
std	145,00
min	1079,00
25%	1174,00
50%	1209,00
75%	1278,00
max	1599,00

The highest price was 1599 PLN, while the lowest - 1079. The std of that statistic equals 145 PLN, it tells by how much on average the values of the examined feature deviate from the arithmetic mean. The median 1209, the midpoint value is the point at which half the observations are above the value and half the observations are below the value. There occurred the biggest price fluctuations.

RESULTS OF ANALYSIS

Generating a word cloud

https://best.aliexpress.com/?lan=en

Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud.