Zadania rekrutacyjne
na stanowisko Data Analyst w firmie —

1. **Korzystając ze zbioru danych dotyczących podróży taksówką (link do pliku poniżej, ok. 20MB, Plik CSV) spróbuj ustalić jak najwięcej faktów na ten temat: sprawdź co zawiera zbiór, oceń jakość tych danych i wyciągnij z tych danych informacje oraz wnioski.**

https://drive.google.com/file/d/1fP4_LrFToQXF0VYsj7hPV_0MQ4gfRdn0/view?usp=sharing

Link do skryptu, w którym pracowałem nad uzyskaniem wszystkich poniższych informacji znajduje się tutaj.

Dane zawarte w pliku CSV składają się z 23 kolumn i 200 000 wierszy.
Poniżej przedstawiono nagłówki poszczególnych kolumn z krótkim opisem.

	Nagłówek	Opis
1	unique_key	Unikalny identyfikator kursu
2	taxi_id	Numer ID taksówki
3	trip_start_timestamp	Data i godzina(zaokrąglona do najbl. 15min) startu kursu
4	trip_end_timestamp	Data i godzina(zaokrąglona do najbl. 15min) końca kursu
5	trip_seconds	Czas trwania kursu [sek.]
6	trip_miles	Dystans [mile]
7	pickup_census_tract	-
8	dropoff_census_tract	-
9	pickup_community_area	-
10	dropoff_community_area	-
11	fare	Opłata za kurs
12	tips	Napiwek
13	tolls	Opłata drogowa
14	extras	Opłata dodatkowa
15	trip_total	Suma opłat
16	payment_type	Metoda płatności
17	company	Firma przewozowa
18	pickup_latitude	Szerokość geogr. punktu początkowego
19	pickup_longitude	Długość geogr. punktu początkowego
20	pickup_location	Lokalizacja punktu początkowego
21	dropoff_latitude	Szerokość geogr. punktu końcowego
22	dropoff_longitude	Długość geogr. punktu końcowego
23	dropoff_location	Lokalizacja punktu końcowego

```
2014    93668
2016    14721
2017    90280
2018     1331
Name: year, dtype: int64
```

Zbiór zawiera szczegółowe dane kursów taksówek w mieście **Chicago** w latach **2014 - 2018**.

Dane nie są jednak kompletne, gdyż ilości rekordów dla poszczególnych lat znacząco się różnią, co więcej - **w ogóle nie zawierają informacji o kursach wykonanych w 2015 roku.**

```
 #   Column                  Non-Null Count
---  ------                  --------------
 0   unique_key              200000 non-null
 1   taxi_id                 200000 non-null
 2   trip_start_timestamp    200000 non-null
 3   trip_end_timestamp      200000 non-null
 4   trip_seconds            199957 non-null
 5   trip_miles              199989 non-null
 6   pickup_census_tract     88947 non-null
 7   dropoff_census_tract    77952 non-null
 8   pickup_community_area   170983 non-null
 9   dropoff_community_area  126673 non-null
 10  fare                    199988 non-null
 11  tips                    199988 non-null
 12  tolls                   169863 non-null
 13  extras                  199988 non-null
 14  trip_total              199988 non-null
 15  payment_type            200000 non-null
 16  company                 106332 non-null
 17  pickup_latitude         170998 non-null
 18  pickup_longitude        170998 non-null
 19  pickup_location         170998 non-null
 20  dropoff_latitude        126673 non-null
 21  dropoff_longitude       126673 non-null
 22  dropoff_location        126673 non-null
dtypes: float64(15), object(8)
```

Zaledwie 5 z 23 kolumn zawiera kompletne dane:
‚unique_key'
‚taxi_id'
'trip_start_timestamp'
'trip_end_timestamp'
'payment_type'.

Przedstawione dane mogłyby posłużyć do filtrowania/grupowania niezbędnych rekordów, np. takich które posiadają informację o interesującym Przewoźniku lub określonej Metodzie płatności.

WNIOSEK:
Zwłaszcza pierwsza tabela pozwala stwierdzić iż dane zawarte w pliku są niekompletne, dlatego należałoby dokonać innego grupowania niż na podstawie lat.

	number_of_trips
Monday	26950
Tuesday	28630
Wednesday	28589
Thursday	31822
Friday	31941
Saturday	27709
Sunday	24359

Załączona tabela przedstawia ujęcie statystyczne dla kolumn czasu podróży oraz długości kursu.

Średni czas trwania kursu to **1017.4 sek**, a **średnia długość kursu** to **5.54 mile**. Poza wartościami średnimi, minimalnymi i maksymalnymi przedstawiono także o**dchylenie standardowe** oraz r**ozkład danych**.

WNIOSEK:
Największy popyt na usługi taksówkarskie występuję w czwartki i piątki.

Chicago Carriage Cab Corp	18457
303 Taxi	17705
City Service	9989
Medallion Leasin	9306
Taxi Affiliation Service Yellow	8937
Sun Taxi	8625
Globe Taxi	6387
Metro Group	5986
Yellow Cab	3148
Nova Taxi Affiliation Llc	3003
Patriot Taxi Dba Peace Taxi Associat	2803
Norshore Cab	1939
24 Seven Taxi	1878
Checker Taxi Affiliation	1802
Chicago Independents	1241
Flash Cab	1057
Chicago Taxicab	911
Blue Diamond	792
Gold Coast Taxi	763
Service Taxi Association	462
Setare Inc	288
Metro Jet Taxi A	287
Checker Taxi	190
Leonard Cab Co	176
American United Taxi Affiliation	130
American United	41
Chicago Star Taxicab	23
5 Star Taxi	6

Tabela przedstawia wykaz firm przewozowych wraz z ilością wykonanych kursów zawartych w dostępnym zbiorze danych.

WNIOSEK:
Zdecydowanymi liderami na rynku są Chicago Carriage Cab Corp oraz 303 Taxi. Warto dodać, że 16 firm odnotowało ponad 1000 wykonanych kursów.

Cash	130485
Credit Card	68920
Prcard	431
Mobile	74
Pcard	65
Split	25

Kolumna ‚payment_type' pozwala zweryfikować metody płatności oraz to jak często z nich korzystano rozliczając kurs taksówką.

WNIOSEK:
Zdecydowana większość kursów była rozliczana gotówką. Należy jednak zadać sobie pytanie w ilu przypadkach była to jedyna opcja rozliczenia za kurs.

Poniżej coś co może wydać się oczywiste, jednak chciałem się upewnić:) Mianowicie, **maksymalna wysokość oraz średnia wysokość napiwku** przy płatności gotówką oraz innych formach płatności:

GOTÓWKA

count	237.000000
mean	5.656540
std	9.758828
min	0.500000
25%	2.000000
50%	3.000000
75%	6.730000
max	126.050000

INNE FORMY PŁATNOŚCI

count	62192.000000
mean	4.743952
std	4.766535
min	0.010000
25%	2.000000
50%	3.000000
75%	5.810000
max	199.000000

WNIOSEK:

Gotówką płacono 130 485 razy, jednak tylko w 237 przypadkach zostawiano napiwek. Maksymalnie wynosił on 126.05, a średnio 5.65. Innych form płatności użyto 69 515 razy i aż 62 192 razy pojawił się przy tym napiwek. Pomimo, że maksymalny wyniósł 199, to średnio był on niższy niż w przypadku gotówki i wyniósł - 4.74

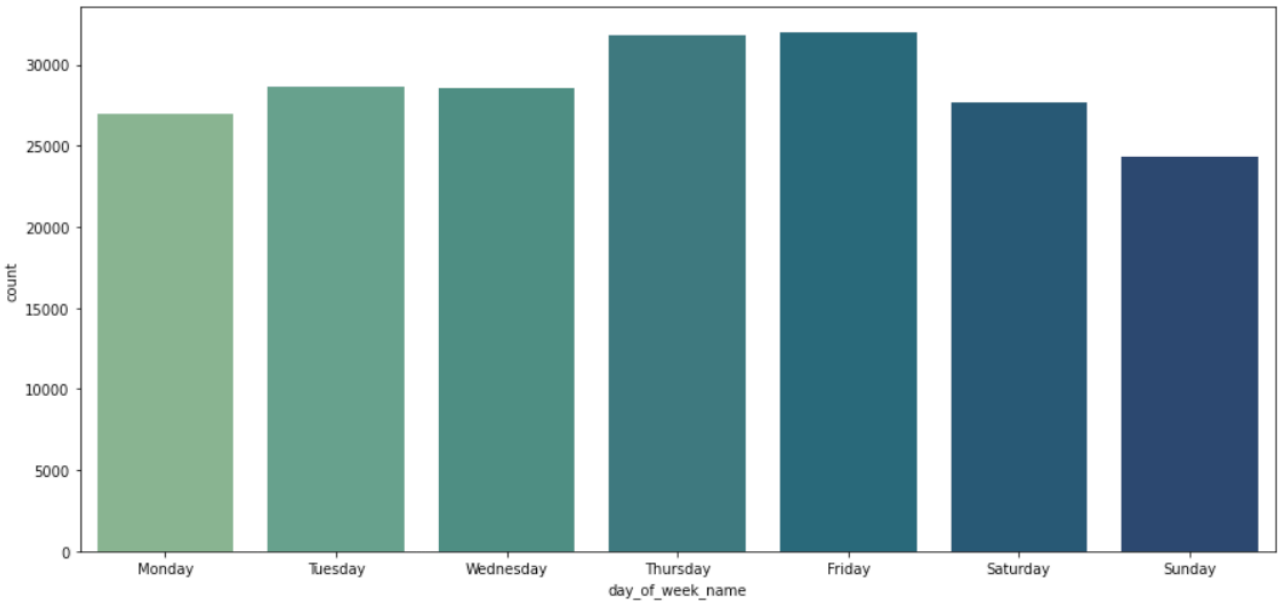
Poniższa tabela przedstawia ujęcie statystyczne dla danych wyłącznie z 2017 roku dla przewoźnika - Chicago Carriage Cab Corp, który odnotował największą ilość kursów.

	trip_seconds	trip_miles	fare	tips	trip_total
count	18441.000000	18456.000000	18457.000000	18457.000000	18457.000000
mean	1151.353723	5.344078	16.828813	1.791114	21.437260
std	2887.912883	7.028603	16.641203	3.723539	23.729959
min	1.000000	0.000000	0.000000	0.000000	0.000000
25%	491.000000	1.470000	7.750000	0.000000	9.000000
50%	740.000000	2.410000	10.500000	0.000000	12.500000
75%	1212.000000	6.410000	19.500000	2.000000	22.750000
max	85633.000000	156.210000	475.000000	155.000000	475.500000

2. Plik z ćwiczenia 1 zawiera sporą ilość danych, co zrobisz jeśli tych danych będzie 10 razy więcej?

W zależności od problemu biznesowego lub oczekiwań klienta/managementu wybrałbym jak najbardziej reprezentatywną część danych. Mógłby to być jeden pełny rok, zawierający największą ilość rekordów, lub konkretny miesiąc/kwartał z każdego roku dostępnego w zestawie danych, co dodatkowo pozwoliłoby na analizę porównawczą w odniesieniu rok-do-roku. Bądź też, tak jak starałem się uchwycić to w powyższej analizie - wykorzystać wszystkie dostępne dane przypisując je do konkretnych dni tygodnia.

Natomiast gdyby zadanie polegało na wykonaniu predykcji na podstawie dostępnych danych to z pewnością podzieliłbym je na zbiór treningowy oraz testowy, a dodatkowo wyodrębnił zbiór walwidacyjny z pierwszego z nich.



Wykres przedstawia ilość wykonanych kursów taxi w Chicago w poszczególne dni tygodnia, w oparciu o wszystkie kursy zawarte w dostępnym źródle.

3. Korzystając z dostępu do zbioru danych w Google Bigquery dotyczący podróży taksówkami w Chicago przeprowadź analogiczną analizę jak z w zadaniu 1, ale tym razem oprócz wyników i wniosków, do rozwiązania dołącz swoje zapytania SQL.
https://console.cloud.google.com/marketplace/product/city-of-chicago-public-data/chicago-taxi-trips -> dostęp jest darmowy pod warunkiem posiadania konta gmail.

```
SELECT
  EXTRACT(YEAR FROM trip_start_timestamp) AS YEAR,
  COUNT(1) AS rides
FROM
  `bigquery-public-data.chicago_taxi_trips.taxi_trips`
GROUP BY
  YEAR
ORDER BY
  YEAR
```

Wiersz	YEAR	rides
1	2013	27217300
2	2014	37395079
3	2015	32385527
4	2016	31756403
5	2017	24979611
6	2018	20731105
7	2019	16476440
8	2020	3888831
9	2021	3947677
10	2022	6369867
11	2023	24

Zbiór danych w Google Bigquery jest znacznie obszerniejszy i obejmuje kursy taksówek w Chicago od roku 2013 do 2023

WNIOSEK:
Na przedstawionej tabeli można zaobserwować jak spadł popyt na tę usługę w latach 2020-2021, na co zapewne olbrzymi wpływ miał wybuch pandemii COVID-19.

```
SELECT
  EXTRACT(DAYOFWEEK FROM trip_start_timestamp) AS day,
  COUNT(1) AS rides
FROM
  `bigquery-public-data.chicago_taxi_trips.taxi_trips`
GROUP BY
  day
ORDER BY
  day
```

Wiersz	day	rides
1	1	24705537
2	2	26397866
3	3	28547744
4	4	29893339
5	5	31696392
6	6	33997131
7	7	29909855

Tabela przedstawia ilość kursów wykonanych w poszczególne dni tygodnia(cyfry 1-7 odpowiadają kolejno: pon, wt, śr, …, pt).

WNIOSEK:
Podobnie jak w wybranych danych z pliku CSV największe zainteresowanie kursami taksówką następowało w piątki i soboty. Bez wątpienia jest to najlepszy czas dla spotkań towarzyskich, a więc lepiej nie wsiadać za kółko na podwójnym gazie.

```sql
SELECT
  EXTRACT(year FROM trip_start_timestamp) AS year,
  MAX(trip_seconds) AS maximum_trip_time,
  FORMAT('%3.2f', AVG(trip_seconds)) AS avg_trip_time,
  MAX(trip_miles) AS maximum_trip_distance,
  FORMAT('%3.2f', AVG(trip_miles)) AS avg_trip_distance,
  MAX(trip_total) AS maximum_total_cost,
  FORMAT('%3.2f', AVG(trip_total)) AS avg_total_cost,
  COUNT(1) AS trips,
FROM
  `bigquery-public-data.chicago_taxi_trips.taxi_trips`
GROUP BY
  year
ORDER BY
  year
```

Wiersz	year	maximum_trip_time	avg_trip_time	maximum_trip_distance	avg_trip_distance	maximum_total_cost	avg_total_cost	trips
1	2013	86340	737.97	1998.1	2.18	9999.99	14.27	27217300
2	2014	86340	738.35	1530.4	2.81	9999.82	14.15	37395079
3	2015	86392	745.72	3460.0	3.19	9966.66	14.83	32385527
4	2016	86399	783.36	3353.1	3.90	9999.0	16.03	31756403
5	2017	86389	812.01	2151.86	3.61	9999.99	15.96	24979611
6	2018	86340	856.57	1841.96	3.67	9975.32	16.80	20731105
7	2019	86400	899.40	1428.97	3.69	9900.54	18.18	16476440
8	2020	86398	874.79	993.6	3.68	9955.55	18.37	3888831
9	2021	86382	1145.30	3430.53	5.70	9975.25	25.09	3947677
10	2022	86341	1198.62	2967.54	6.19	9999.75	26.83	6369867
11	2023	3294	995.46	18.7	6.06	86.5	26.16	24

Na powyższej tabeli można zaobserwować jak w latach 2021-2022 znacząco wzrosły wartości średniej dla czasu podróży, dystansu oraz całkowitego kosztu kursu.

WNIOSEK:

Zaobserwowane zmiany w latach 2021-2022 powiązałbym ze zmianą wyboru środka transportu przez mieszkańców Chicago. Być może to pandemia COVID-19 spowodowała, że chętniej decydowano się na ‚odizolowany' środek transportu.

```sql
SELECT EXTRACT(year FROM trip_start_timestamp) AS year,
  COUNT ( DISTINCT company) AS company,
FROM
  `bigquery-public-data.chicago_taxi_trips.taxi_trips`
GROUP BY
  year
ORDER BY
  year
```

Wiersz	year	company
1	2013	37
2	2014	106
3	2015	67
4	2016	89
5	2017	85
6	2018	73
7	2019	58
8	2020	52
9	2021	40
10	2022	37
11	2023	9

W tabeli obok przedstawiono ilość firm przewozowych dla poszczególnych lat.

WNIOSEK:
Spadkowa tendencja w tym obszarze to być może efekt pojawienia się na rynku nieoficjalnych przewoźników typu Uber, bądź też przewaga wiodących firm i efektywniejszego wykorzystania przez nich kanałów dotarcia do potencjalnych klientów - programy lojalnościowe, intuicyjna aplikacja, skrócony czas oczekiwania na transport, znajomość ceny usługi jeszcze przed jej zamówieniem.

```sql
SELECT
  EXTRACT(year FROM trip_start_timestamp) AS year,
  FORMAT('%3.2f',(SUM(trip_total)/SUM(trip_miles))) AS
price_of_mile,
  FORMAT('%3.5f',((SUM(trip_total)/SUM(trip_seconds))*60)) AS
price_of_minute,
FROM
  `bigquery-public-data.chicago_taxi_trips.taxi_trips`
GROUP BY
  year
ORDER BY
  year
```

Wiersz	year	price_of_mile	price_of_minute
1	2013	6.54	1.21037
2	2014	5.03	1.15418
3	2015	4.66	1.19376
4	2016	4.11	1.22785
5	2017	4.42	1.17958
6	2018	4.58	1.17664
7	2019	4.93	1.21261
8	2020	4.99	1.26021
9	2021	4.40	1.31469
10	2022	4.33	1.34268
11	2023	4.32	1.57704

W tabeli obok przedstawiłem cenę za milę oraz cenę za minutę kursu w poszczególnych latach.

WNIOSEK:
Spadek ceny za milę, przy wzroście ceny za minutę kursu to być może reakcja branży taksówkarskiej na coraz większe natężenie ruchu i korki, które są nieodzownym elementem pracy taksówkarzy.

4. Korzystając z dostępu do Demo Account dla Google Analytics (https://support.google.com/analytics/answer/6367342?hl=en) ustal:

Do realizacji zadania czwartego wybrałem usługę - **GA4 - Google Merchandise Store** oraz dane z okresu **1 - 31 grudnia 2022r.**

- **jakie są najczęstsze źródła ruchu,**

Najczęstszymi źródłami ruchu w/w okresie są:

1. **direct - 52 390**
2. **google - 43 140**
3. **unattributable - 12 457**

	Źródło	↓Całkowita liczba użytkowników
	Razem	107 939 100,0% całości
1	(direct)	52 390
2	google	43 140
3	(unattributable)	12 457
4	Newsletter_November_2022_2	2 748
5	art-analytics.appspot.com	2 636
6	youtube.com	1 782
7	analytics.google.com	1 093
8	baidu	990
9	perksatwork.com	838
10	sites.google.com	808

- **jakie kanały/źródła (z wyłączeniem direct) najlepiej konwertują,**

Kanały/źródła, które najlepiej konwertują to:

1. **google / organic**
2. **google / cpc**
3. **unattributable / unattributable**
   ……..
4. **art.analytics.appspot.com / referral**

Z uwagi na brak możliwości uzyskania szczegółowych danych dla **Unattributable/ Unattributable** zdecydowałem się wymienić czwartą pozycję oraz użyć ją w kolejnym kroku.

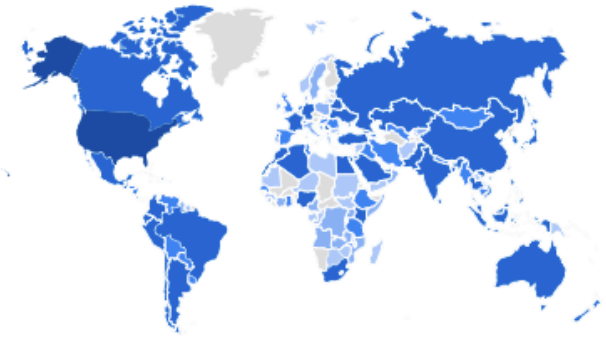
	Źródło/medium	↓Konwersje
	Razem	425 380 100,0% całości
1	(direct) / (none)	167 983
2	google / organic	132 244,73
3	google / cpc	44 761,55
4	(unattributable) / (unattributable)	22 901
5	art-analytics.appspot.com / referral	13 844,47
6	Newsletter_November_2022_2 / email	10 920,7
7	perksatwork.com / referral	4 957,94
8	sites.google.com / referral	4 774,27
9	analytics.google.com / referral	3 304,48
10	support.google.com / referral	2 172,71

- **charakterystykę i rodzaj użytkowników pozyskiwanych przez te kanały/źródła,**

Poniższa charakterystyka dotyczy trzech źródeł/medium **(google/organic, google/cpc, art.analytics.appspot.com / referral)** dla nowych oraz powracających użytkowników:

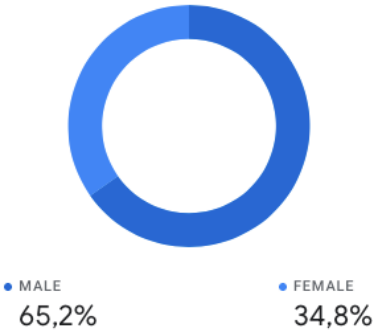
NOWI UŻYTKOWNICY

Nowi użytkownicy ▾ według: Kraj

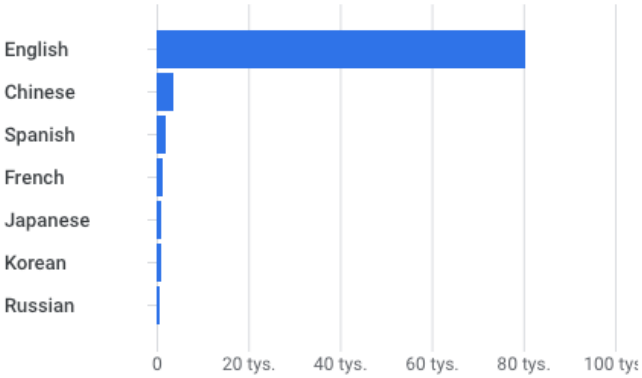


KRAJ	NOWI UŻYTKOWNI...
United States	64 tys.
Canada	6,2 tys.
India	6,2 tys.
China	2,2 tys.
Japan	1 tys.
South Korea	843
Taiwan	784

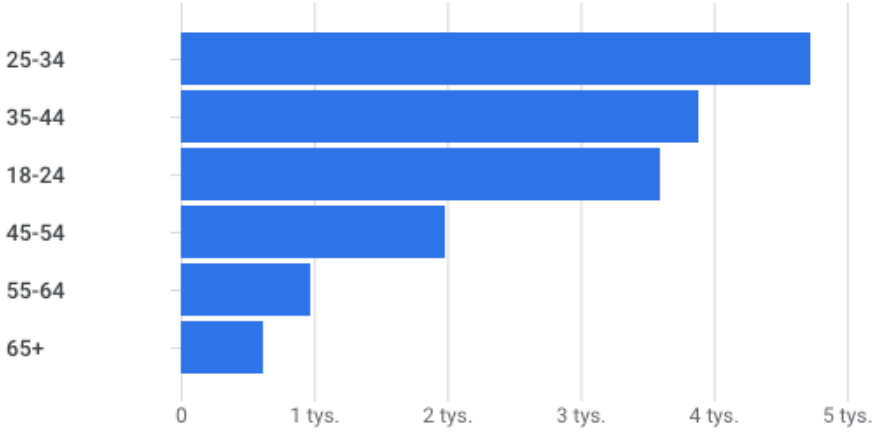
Nowi użytkownicy ▾ według: Płeć



Nowi użytkownicy ▾ według: Język

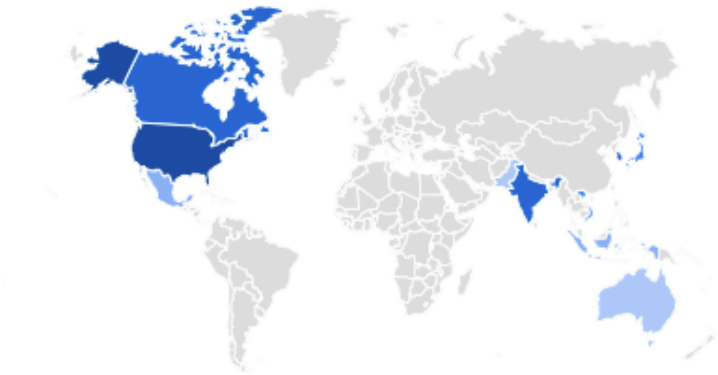


Nowi użytkownicy ▾ według: Wiek



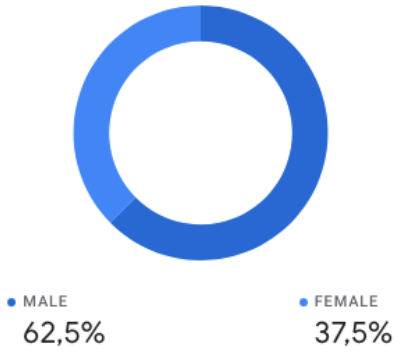
POWRACAJĄCY UŻYTKOWNICY

Powracający użytkownicy ▾ według: Kraj

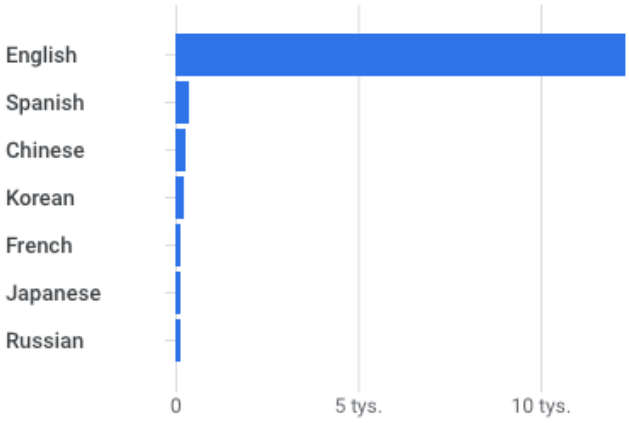


KRAJ	POWRACAJĄCY UŻ...
United States	3 tys.
India	573
Canada	410
South Korea	120
Singapore	119
Vietnam	69
Taiwan	67

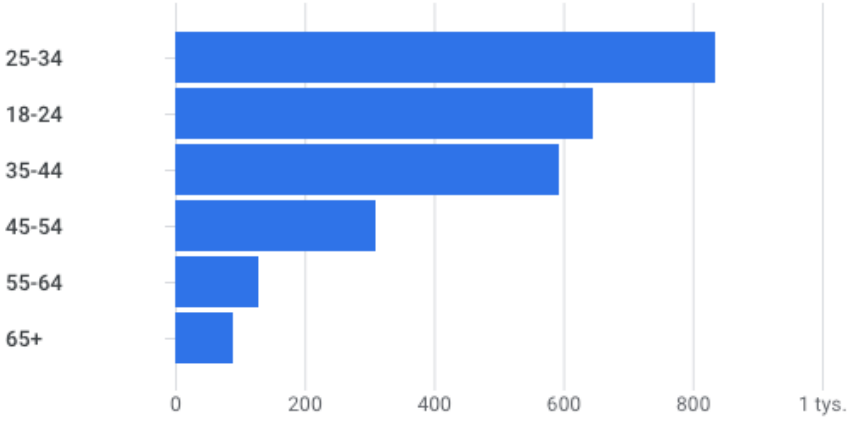
Powracający użytkownicy ▾ według: Płeć



Powracający użytkownicy ▾ według: Język



Powracający użytkownicy ▾ według: Wiek



- **czy atrybucja last non-direct click jest właściwa? Dlaczego?**

Z pewnością zależy to od polityki prowadzenia biznesu. W przypadku e-commerce jest to dobre rozwiązanie, szczególnie gdy stawia się na internetowe formy reklamy. Poznanie ‚last non-direct' medium pozwala na dodatkową analizę kroku, który mógł mieć bezpośredni wpływ na decyzję zakupową.

Mając na uwadze swoje dotychczasowe doświadczenia jakie kroki optymalizacyjne można podjąć przy zwiększaniu efektywności pozyskania wartościowego ruchu dla strony.

Bazując na zdobytej wiedzy zdecydowałbym się na analizę lejków sprzedażowych, skróceniu procesu zakupu(optymalizacja ilość kroków), pozycjonowaniu, a także reklamie w najbardziej „nośnych" social-mediach.