# Marketing_Analysis

Fauzan Rizky Ramadhan

2025-07-16

## 1."Marketing Performance Analysis: Lead Conversion & Revenue Insights (2017-2018)

This project analyzes marketing lead performance from 2017 to 2018, focusing on lead-to-won conversions, top-performing origins, and overall conversion rates. It also highlights the distribution of leads with zero revenue ( both pending and recently converted ) and examines positive revenue by range and lead type. These insights aim to support data-driven decisions for improving marketing strategies and driving revenue growth. This analysis uses a dataset obtained from Kaggle, titled "Marketing Funnel by Olist". The dataset was published by Olist. and consists of two files:

1. olist_closed_deals_dataset.csv (year 2017-2018)
2. olist_marketing_qualified_leads_dataset.csv (year 2017-2018)

click here

### 2. Setup and Packages

The following R packages are used in this project to perform data manipulation, visualization, and analysis. Please ensure they are installed before running the code.

library(tidyverse) # Includes dplyr, tidyr, ggplot2, etc.

library(janitor) # For cleaning column names

library(lubridate) # For date/time cleaning

library(stringr) # For text cleaning

library(forcats) # For factor variables

library(tidyr) # For data manipulation

library(dplyr) # For data manipulation

### 3. Dataset Overview and Cleaning Steps

After importing the datasets from 2017 and 2018, we begin by exploring the structure and content of the combined data. This step helps us identify any inconsistencies, missing values, or formatting issues that may affect our analysis.

Once we understand the layout of the dataset, we proceed with basic cleaning such as renaming columns, handling missing values, and ensuring consistent data types to prepare the data for accurate and meaningful insights.

```r
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.4.2
```

```r
Marketing_Qualified_Leads_Dtaaset <- read_excel("C:/Users/Solit/Downloads/Marketing Qualified Leads Dtaa
head(Marketing_Qualified_Leads_Dtaaset)
```

```
## # A tibble: 6 x 4
##   mql_id                           first_contact_date  landing_page_id    origin
##   <chr>                            <dttm>              <chr>              <chr>
## 1 dac32acd4db4c29c230538b72f8dd87d 2018-02-01 00:00:00 88740e65d5d6b056e~ social
## 2 8c18d1de7f67e60dbd64e3c07d7e9d5d 2017-10-20 00:00:00 007f9098284a86ee8~ paid_~
## 3 b4bc852d233dfefc5131f593b538befa 2018-03-22 00:00:00 a7982125ff7aa3b20~ organ~
## 4 6be030b81c75970747525b843c1ef4f8 2018-01-22 00:00:00 d45d558f0daeecf3c~ email
## 5 5420aad7fec3549a85876ba1c529bd84 2018-02-21 00:00:00 b48ec5f3b04e90684~ organ~
## 6 28bdfd5f057764b54c38770f95c69f2f 2018-01-14 00:00:00 22c29808c4f815213~ organ~
```

```r
Marketing_Qualified_Leads_Dtaaset <- data.frame(lapply(Marketing_Qualified_Leads_Dtaaset,function(x) { 
dataset_clean <- Marketing_Qualified_Leads_Dtaaset
summary(dataset_clean)
```

```
##     mql_id          first_contact_date              landing_page_id
##  Length:8000        Min.   :2017-06-14 00:00:00.0   Length:8000
##  Class :character   1st Qu.:2017-12-31 00:00:00.0   Class :character
##  Mode  :character   Median :2018-02-25 00:00:00.0   Mode  :character
##                     Mean   :2018-02-05 14:19:51.5
##                     3rd Qu.:2018-04-15 06:00:00.0
##                     Max.   :2018-05-31 00:00:00.0
##     origin
##  Length:8000
##  Class :character
##  Mode  :character
##
##
##
```

```r
summary(is.na(dataset_clean))
```

```
##    mql_id        first_contact_date landing_page_id   origin
##  Mode :logical   Mode :logical      Mode :logical   Mode :logical
##  FALSE:8000      FALSE:8000         FALSE:8000      FALSE:7940
##                                                     TRUE :60
```

```r
dataset_clean$origin[is.na(dataset_clean$origin)] <- "not recorded"
```

### 4. Lead Distribution by Source: Organic Search Leads the Way

This analysis shows how leads are distributed across different sources. Organic search brings in the most leads, followed by paid search and social media. This means that SEO and organic content are key drivers for lead generation

```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.4.3

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.3
```

```r
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.4.2
```

```r
library(janitor)
```

```
## Warning: package 'janitor' was built under R version 4.4.3

##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```r
Marketing_Qualified_Leads_Dtaaset <- read_excel("C:/Users/Solit/Downloads/Marketing Qualified Leads Dta
head(Marketing_Qualified_Leads_Dtaaset)
```

```
## # A tibble: 6 x 4
##   mql_id                           first_contact_date  landing_page_id    origin
##   <chr>                            <dttm>              <chr>              <chr>
## 1 dac32acd4db4c29c230538b72f8dd87d 2018-02-01 00:00:00 88740e65d5d6b056e~ social
## 2 8c18d1de7f67e60dbd64e3c07d7e9d5d 2017-10-20 00:00:00 007f9098284a86ee8~ paid_~
## 3 b4bc852d233dfefc5131f593b538befa 2018-03-22 00:00:00 a7982125ff7aa3b20~ organ~
## 4 6be030b81c75970747525b843c1ef4f8 2018-01-22 00:00:00 d45d558f0daeecf3c~ email
## 5 5420aad7fec3549a85876ba1c529bd84 2018-02-21 00:00:00 b48ec5f3b04e90684~ organ~
## 6 28bdfd5f057764b54c38770f95c69f2f 2018-01-14 00:00:00 22c29808c4f815213~ organ~
```

```r
Marketing_Qualified_Leads_Dtaaset <- data.frame(lapply(Marketing_Qualified_Leads_Dtaaset,function(x) {
dataset_clean <- Marketing_Qualified_Leads_Dtaaset
summary(dataset_clean)
```

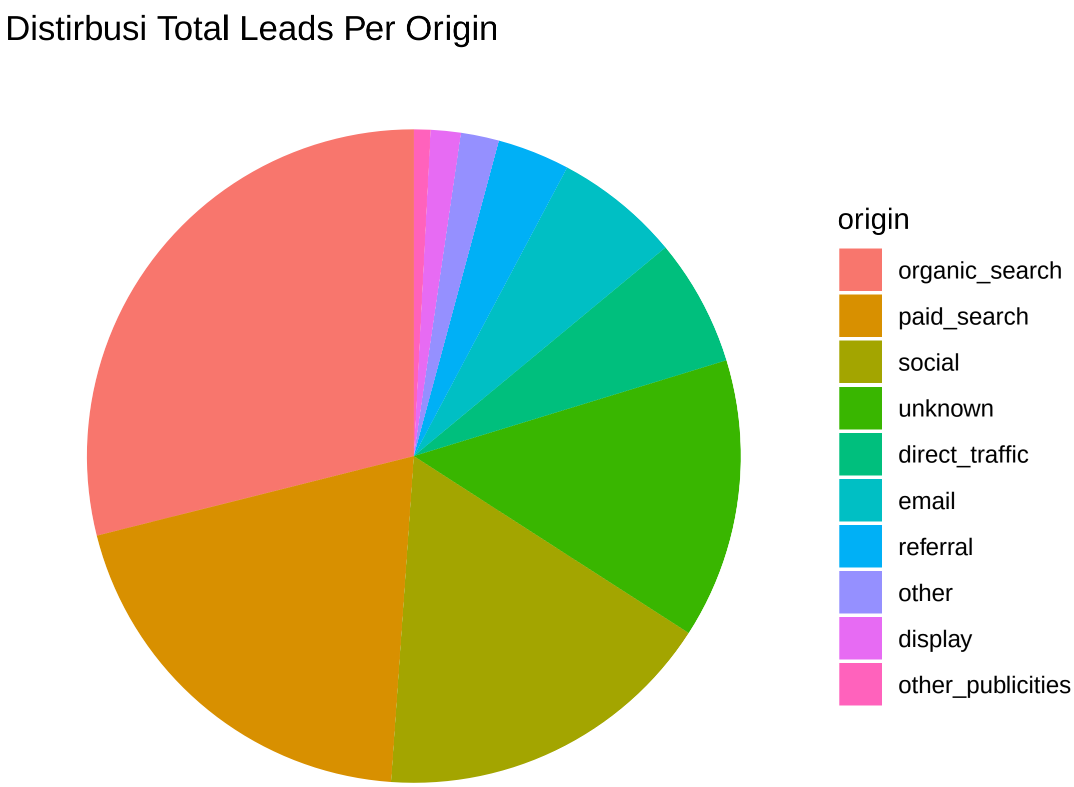
```
##     mql_id          first_contact_date              landing_page_id
##  Length:8000        Min.   :2017-06-14 00:00:00.0   Length:8000
##  Class :character   1st Qu.:2017-12-31 00:00:00.0   Class :character
##  Mode  :character   Median :2018-02-25 00:00:00.0   Mode  :character
##                     Mean   :2018-02-05 14:19:51.5
##                     3rd Qu.:2018-04-15 06:00:00.0
##                     Max.   :2018-05-31 00:00:00.0
##     origin
##  Length:8000
##  Class :character
##  Mode  :character
##
##
##
```

```r
summary(is.na(dataset_clean))
```

```
##    mql_id        first_contact_date landing_page_id   origin
##  Mode :logical   Mode :logical      Mode :logical   Mode :logical
##  FALSE:8000      FALSE:8000         FALSE:8000      FALSE:7940
##                                                     TRUE :60
```

```r
dataset_clean$origin[is.na(dataset_clean$origin)] <- "not recorded"

dataset_clean <- Marketing_Qualified_Leads_Dtaaset
Total_leads_per_origin <-dataset_clean %>%
  group_by(origin) %>%
  summarize( Total_leads = n_distinct(mql_id)) %>%
  arrange(desc(Total_leads))
Total_leads_per_origin <- na.omit(Total_leads_per_origin)
Total_leads_per_origin$origin <- factor(Total_leads_per_origin$origin,
                                        levels =Total_leads_per_origin$origin)
ggplot(Total_leads_per_origin) + geom_bar(mapping= aes (x = "", y = Total_leads , fill= origin), stat=
  coord_polar(theta = "y") + theme_void() + labs( title = "Distirbusi Total Leads Per Origin")
```




```r
head(Total_leads_per_origin)
```

```
## # A tibble: 6 x 2
##   origin         Total_leads
##   <fct>                <int>
## 1 organic_search        2296
## 2 paid_search           1586
## 3 social                1350
## 4 unknown               1099
## 5 direct_traffic         499
## 6 email                  493
```

### 5. Annual Lead Conversion Analysis: Key Trends and Opportunities (2017-2018)

Before analysis, we standardized the dataset to ensure accuracy. Our data reveals critical patterns in lead conversion performance across years and lead types. By identifying these trends, we can strategically allocate resources to maximize ROI. Below are the key insights from this analysis :

1. **Online leads** dominate conversions, particularly the `online_top` category (highlight specific metrics from the plot).
2. **2018** saw a **X% increase** in conversions compared to 2017 (derive from data).
3. **"Not converted" leads** remain high at **Y%**, indicating opportunities for process improvement.

```r
# Load semua library yang dibutuhkan
library(readxl)
library(dplyr)
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.4.2
```

```r
library(janitor)
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 4.4.2
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```r
library(ggplot2)
Marketing_Closed_Leads_Dataset <- read_excel("C:/Users/Solit/Downloads/Marketing Closed Leads Dataset.xl
head(Marketing_Closed_Leads_Dataset)
```

```
## # A tibble: 6 x 14
##   mql_id   seller_id sdr_id sr_id won_date            business_segment lead_type
##   <chr>    <chr>     <chr>  <chr> <dttm>              <chr>            <chr>
## 1 5420aad~ 2c43fb51~ a8387~ 4ef1~ 2018-02-26 19:58:54 pet              online_m~
## 2 a555fb3~ bbb7d789~ 09285~ d3d1~ 2018-05-08 20:17:59 car_accessories  industry
## 3 327174d~ 612170e3~ b90f8~ 6565~ 2018-06-05 17:27:23 home_appliances  online_b~
## 4 f5fee8f~ 21e1781e~ 56bf8~ d3d1~ 2018-01-17 13:51:03 food_drink       online_s~
## 5 ffe6401~ ed8cb7b1~ 4b339~ d3d1~ 2018-07-03 20:17:45 home_appliances  industry
## 6 b94fba7~ 1c742ac3~ fdb16~ 495d~ 2018-02-07 18:04:05 health_beauty    online_m~
## # i 7 more variables: lead_behaviour_profile <chr>, has_company <lgl>,
## #   has_gtin <lgl>, average_stock <chr>, business_type <chr>,
## #   declared_product_catalog_size <chr>, declared_monthly_revenue <chr>
```

```r
Marketing_Closed_Leads_Dataset$declared_monthly_revenue <- as.numeric(Marketing_Closed_Leads_Dataset$dec
Joined_Dataset <-left_join(x = Marketing_Qualified_Leads_Dtaaset, y= Marketing_Closed_Leads_Dataset, by=
head(Joined_Dataset)
```

```
##                             mql_id first_contact_date
## 1 dac32acd4db4c29c230538b72f8dd87d         2018-02-01
## 2 8c18d1de7f67e60dbd64e3c07d7e9d5d         2017-10-20
## 3 b4bc852d233dfefc5131f593b538befa         2018-03-22
## 4 6be030b81c75970747525b843c1ef4f8         2018-01-22
## 5 5420aad7fec3549a85876ba1c529bd84         2018-02-21
## 6 28bdfd5f057764b54c38770f95c69f2f         2018-01-14
##                    landing_page_id         origin
## 1 88740e65d5d6b056e0cda098e1ea6313         social
## 2 007f9098284a86ee80ddeb25d53e0af8    paid_search
```

```
## 3 a7982125ff7aa3b2054c6e44f9d28522 organic_search
## 4 d45d558f0daeecf3cccdffe3c59684aa          email
## 5 b48ec5f3b04e9068441002a19df93c6c organic_search
## 6 22c29808c4f815213303f8933030604c organic_search
##                          seller_id                           sdr_id
## 1                             <NA>                             <NA>
## 2                             <NA>                             <NA>
## 3                             <NA>                             <NA>
## 4                             <NA>                             <NA>
## 5 2c43fb513632d29b3b58df74816f1b06 a8387c01a09e99ce014107505b92388c
## 6                             <NA>                             <NA>
##                              sr_id            won_date business_segment
## 1                             <NA>                <NA>             <NA>
## 2                             <NA>                <NA>             <NA>
## 3                             <NA>                <NA>             <NA>
## 4                             <NA>                <NA>             <NA>
## 5 4ef15afb4b2723d8f3d81e51ec7afefe 2018-02-26 19:58:54              pet
## 6                             <NA>                <NA>             <NA>
##       lead_type lead_behaviour_profile has_company has_gtin average_stock
## 1          <NA>                   <NA>          NA       NA          <NA>
## 2          <NA>                   <NA>          NA       NA          <NA>
## 3          <NA>                   <NA>          NA       NA          <NA>
## 4          <NA>                   <NA>          NA       NA          <NA>
## 5 online_medium                    cat          NA       NA          <NA>
## 6          <NA>                   <NA>          NA       NA          <NA>
##   business_type declared_product_catalog_size declared_monthly_revenue
## 1          <NA>                          <NA>                       NA
## 2          <NA>                          <NA>                       NA
## 3          <NA>                          <NA>                       NA
## 4          <NA>                          <NA>                       NA
## 5      reseller                          <NA>                        0
## 6          <NA>                          <NA>                       NA
```

```r
Joined_Dataset <- data.frame(lapply(Joined_Dataset,function(x) { if (is.character(x)) trimws(x) else x}
Joined_Dataset <- clean_names(Joined_Dataset)
Joined_Dataset$won_date <-as_datetime(Joined_Dataset$won_date)
class(Joined_Dataset$declared_product_catalog_size)
```

```
## [1] "character"
```

```r
class(Joined_Dataset$average_stock)
```

```
## [1] "character"
```

```r
Joined_Dataset$declared_product_catalog_size <- as.numeric(Joined_Dataset$declared_product_catalog_size
Joined_Dataset$average_stock <- as.numeric(Joined_Dataset$average_stock)
```

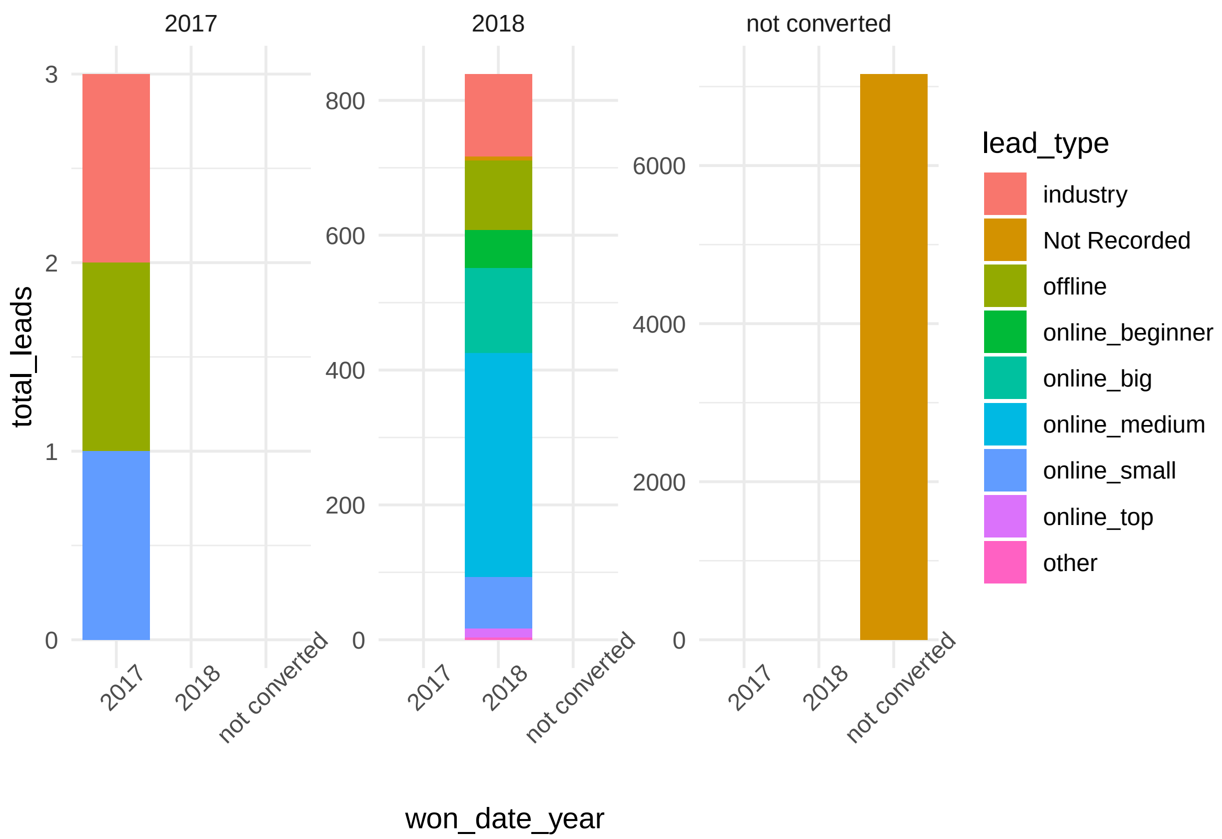
```
## Warning: NAs introduced by coercion
```

```r
Joined_Dataset <-Joined_Dataset %>%
  mutate(across(where(is.logical), as.character)) %>%
```

```r
  mutate(across(where(is.character), ~replace_na(., "Not Recorded"))) %>%
  mutate(across(where(is.numeric), ~replace_na (. , 0))) %>%
  mutate(won_date_year = ifelse(is.na(won_date),"not converted",as.character(year(won_date))))
leads_peryear<- Joined_Dataset %>%
  group_by(won_date_year, lead_type) %>%
  summarize( total_leads = n_distinct(mql_id), .groups="drop")
head(leads_peryear)
```

```
## # A tibble: 6 x 3
##   won_date_year lead_type    total_leads
##   <chr>         <chr>              <int>
## 1 2017          industry               1
## 2 2017          offline                1
## 3 2017          online_small           1
## 4 2018          Not Recorded           6
## 5 2018          industry             122
## 6 2018          offline              103
```

```r
ggplot(leads_peryear) + geom_bar(mapping = aes( x=won_date_year, y = total_leads, fill = lead_type), sta
```



## 6.Traffic Source Analysis: Lead Volume, Conversion Efficiency, and Strategic Implications (2017–2018)

This analysis compares total leads, conversions, and conversion rates across traffic sources, revealing that organic search leads in volume (2,296 leads, 271 conversions), but not in efficiency because its conversion rate is 11.8%. Surprisingly, the "Not Recorded" origin, though small in total leads (60), has the highest conversion rate at 23.33%, signaling a potentially overlooked quality source. Meanwhile, social and unknown traffic also perform well in terms of conversion efficiency (16.29% and 11.22% respectively), despite moderate total leads. On the other hand, display and email underperform across all metrics. As always, these findings rely on properly cleaned and standardized data, as shown in the preprocessing steps. **Actionable Insights:**

1. Invest in conversion optimization for high-traffic sources like organic search—improve ad targeting, messaging, or landing pages to increase efficiency.
2. Identify and track "Not Recorded" leads,they may represent an overlooked campaign or channel with exceptional conversion potential.

```
convertion_leads <-Joined_Dataset %>%
  filter(won_date_year!="not converted") %>%
  group_by(origin) %>%
  summarize(converted_leads= n(), converted_leads = sum(converted_leads, na.rm = TRUE), .groups= "drop"
  arrange(desc(converted_leads))
head(convertion_leads)
```

```
## # A tibble: 6 x 2
##   origin          converted_leads
##   <chr>                     <int>
## 1 organic_search              271
## 2 paid_search                 195
## 3 unknown                     179
## 4 social                       75
## 5 direct_traffic               56
## 6 referral                     24
```

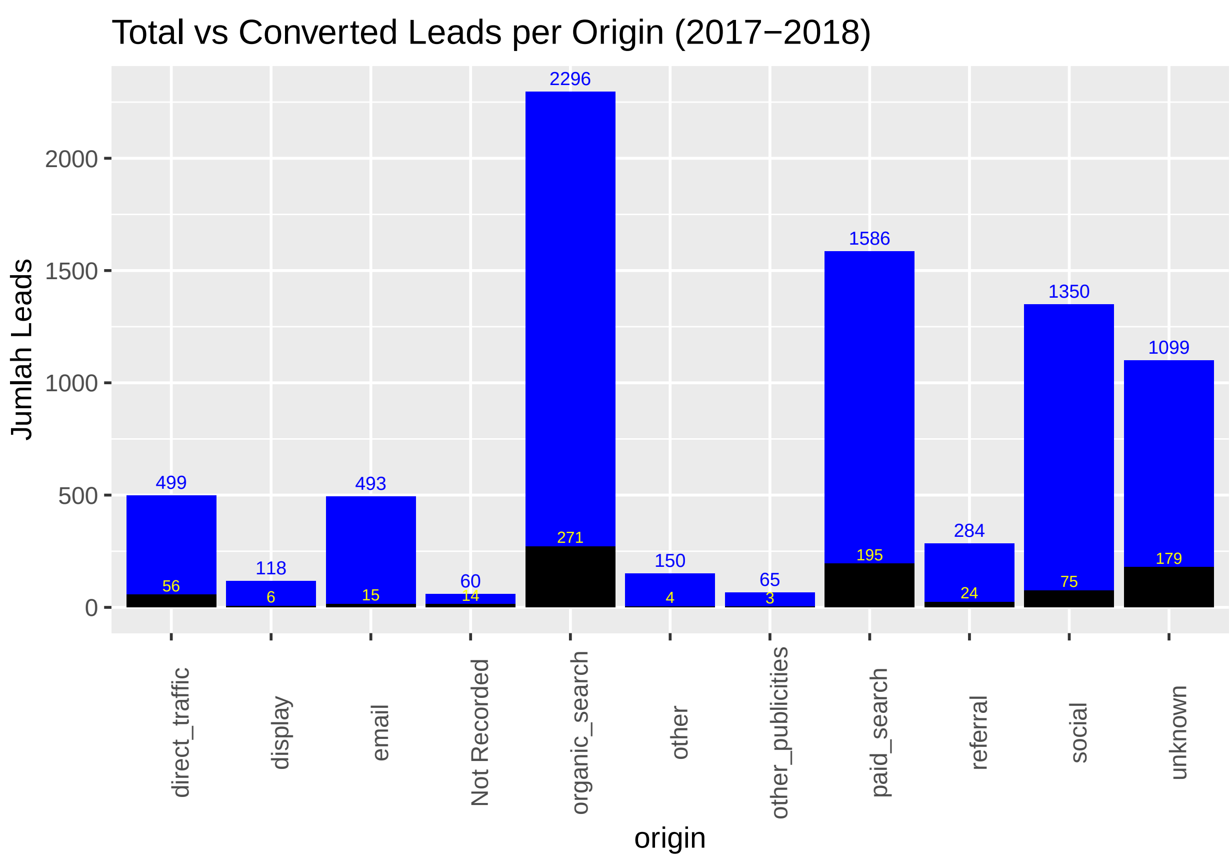
```
total_leads_per_origin <-Joined_Dataset %>%
  group_by(origin) %>%
  summarize(total_leads = n(), .groups= "drop" ) %>%
  arrange(desc(total_leads))
head(total_leads_per_origin)
```

```
## # A tibble: 6 x 2
##   origin          total_leads
##   <chr>                 <int>
## 1 organic_search         2296
## 2 paid_search            1586
## 3 social                 1350
## 4 unknown                1099
## 5 direct_traffic          499
## 6 email                   493
```

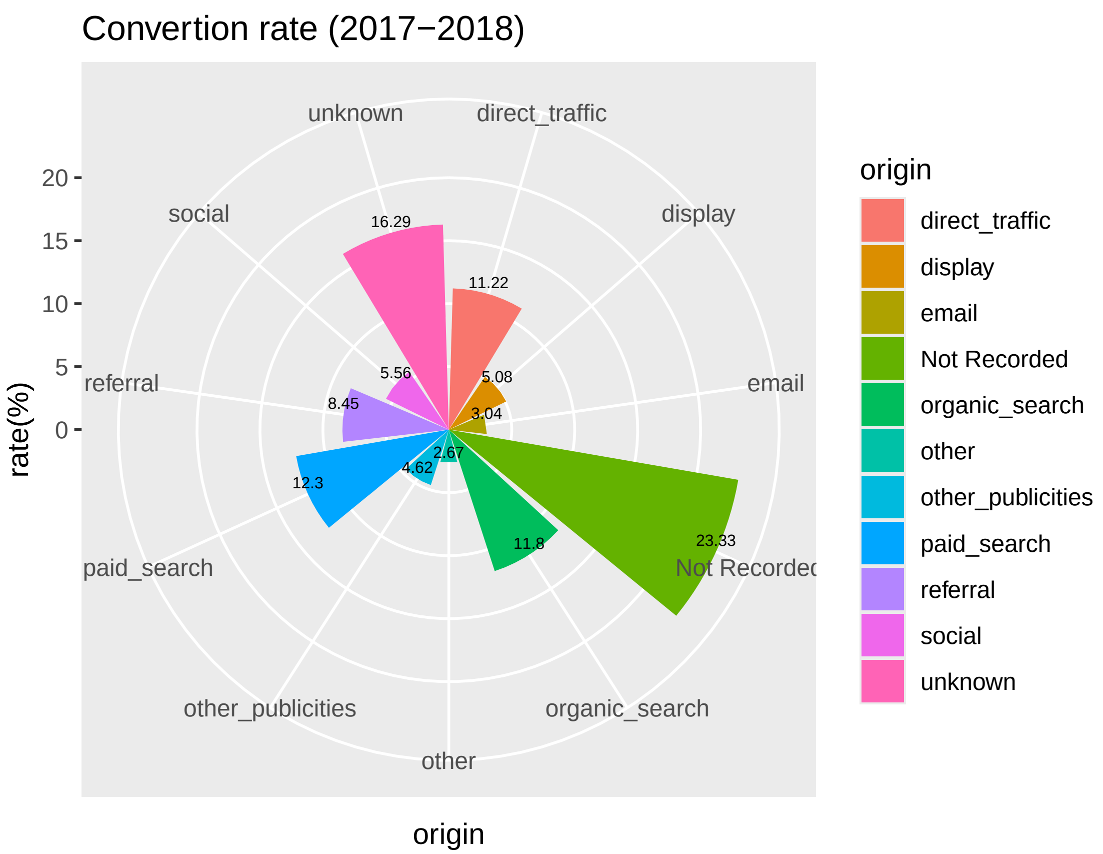
```
convertion_rate <-left_join(x= convertion_leads, y= total_leads_per_origin, by= "origin")
head(convertion_rate)
```

```
## # A tibble: 6 x 3
##   origin          converted_leads total_leads
##   <chr>                     <int>       <int>
## 1 organic_search              271        2296
## 2 paid_search                 195        1586
## 3 unknown                     179        1099
## 4 social                       75        1350
## 5 direct_traffic               56         499
## 6 referral                     24         284
```

```
convertion_rate <-convertion_rate %>%
  mutate(rate = round((converted_leads/total_leads)*100, 2))
longdata_convertionrate <- convertion_rate%>%
  pivot_longer(cols = c(converted_leads, total_leads),
               names_to = "metric",
               values_to = "value")
ggplot(convertion_rate, aes(x = origin)) +
  geom_col(aes(y = total_leads), fill = "blue")+  geom_col(aes(y = converted_leads), fill = "black") +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(y = "Jumlah Leads", title = "Total vs Converted Leads per Origin (2017-2018)") +  geom_text(aes(
  geom_text(aes( y = converted_leads, label = converted_leads), vjust = -0.3, colour = "yellow", size =
```




```
ggplot(convertion_rate) + geom_col(mapping = aes(x = origin, y = rate, fill = origin)) + coord_polar()
  geom_text(aes(x = origin, y = rate, label = rate), size = 2, vjust = -0.5)+
  labs( y = "rate(%)", x = "origin", title = "Convertion rate (2017-2018)")
```




## 7. Zero-Revenue Lead Breakdown by Lead Type: Identifying Waste and Conversion Gaps

This section focuses on leads that brought no declared revenue,a critical blind spot in marketing ROI. Using a filtered dataset (declared_monthly_revenue == 0), we break down zero-revenue cases by lead type and conversion status. The use of head() after grouping is essential to preview the structure and confirm that the filtering logic has segmented the data correctly before plotting. The chart reveals that an overwhelming 7,158 non-converted leads fall under the "Not Recorded" lead type, raising serious concerns about data tracking and lead quality. Even more concerning, structured lead types like "online_medium" show a high number of converted leads with zero revenue, suggesting a monetization, pricing, or post-conversion engagement issue. It's also important to consider that some leads may still be within the early stages of the sales funnel, meaning they haven't had sufficient time to convert—highlighting the need to factor in lead age or conversion window in future analyses.

Actionable key:

1. Audit and enhance lead tracking systems—particularly for "Not Recorded" leads—to ensure attribution accuracy, eliminate funnel blind spots, and improve campaign-level reporting.
2. Investigate zero-revenue conversions in structured lead types like online_medium to uncover gaps in monetization, pricing strategy, or customer nurturing after conversion.

The summary statistics show that most values are zero across all three variables, including the 25th percentile (Q1), median, and 75th percentile (Q3). This suggests a high concentration of leads with no declared economic value. Despite a few extremely high values (outliers), the overall average remains low — the mean declared monthly revenue is approximately $7,723 USD.

This heavily skewed distribution highlights the importance of segmentation:

Only a small fraction of leads present significant business potential, while the majority may be either inactive, not yet monetized, or poorly recorded.

To avoid distortion caused by the overwhelming number of zero values, we limit the current descriptive analysis to general distribution patterns. A more focused analysis on leads with non-zero declared revenue is presented in the following section to better understand their characteristics and revenue potential.

```
Joined_Dataset%>%
  group_by(declared_monthly_revenue) %>%
  summarize(total = n())
```

```
## # A tibble: 27 x 2
##    declared_monthly_revenue total
##                       <dbl> <int>
##  1                        0  7955
##  2                        6     1
##  3                     1000     1
##  4                     4000     1
##  5                     5000     2
##  6                     6000     1
##  7                     8000     1
##  8                    10000     3
##  9                    15000     2
## 10                    20000     3
## # i 17 more rows
```

```
Joined_Dataset <-Joined_Dataset %>%
  mutate(convertion_status = ifelse(won_date_year =="2017" | won_date_year == "2018", "converted", "not
head(Joined_Dataset)
```

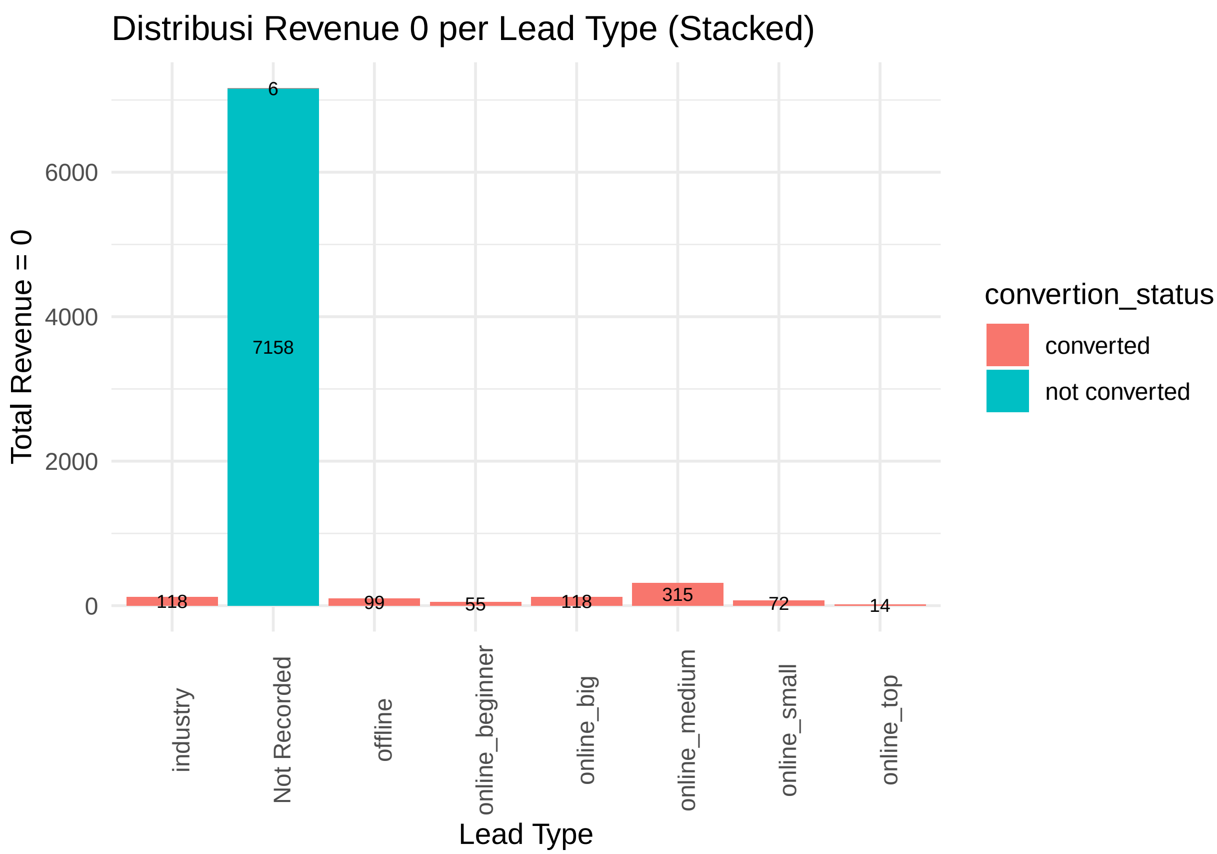
```
##                             mql_id first_contact_date
## 1 dac32acd4db4c29c230538b72f8dd87d         2018-02-01
## 2 8c18d1de7f67e60dbd64e3c07d7e9d5d         2017-10-20
## 3 b4bc852d233dfefc5131f593b538befa         2018-03-22
## 4 6be030b81c75970747525b843c1ef4f8         2018-01-22
## 5 5420aad7fec3549a85876ba1c529bd84         2018-02-21
## 6 28bdfd5f057764b54c38770f95c69f2f         2018-01-14
##                    landing_page_id         origin
## 1 88740e65d5d6b056e0cda098e1ea6313         social
## 2 007f9098284a86ee80ddeb25d53e0af8    paid_search
## 3 a7982125ff7aa3b2054c6e44f9d28522 organic_search
## 4 d45d558f0daeecf3cccdffe3c59684aa          email
## 5 b48ec5f3b04e9068441002a19df93c6c organic_search
## 6 22c29808c4f815213303f8933030604c organic_search
##                          seller_id                           sdr_id
## 1                     Not Recorded                     Not Recorded
## 2                     Not Recorded                     Not Recorded
## 3                     Not Recorded                     Not Recorded
## 4                     Not Recorded                     Not Recorded
## 5 2c43fb513632d29b3b58df74816f1b06 a8387c01a09e99ce014107505b92388c
## 6                     Not Recorded                     Not Recorded
##                              sr_id            won_date business_segment
```

```
## 1                     Not Recorded                <NA>     Not Recorded
## 2                     Not Recorded                <NA>     Not Recorded
## 3                     Not Recorded                <NA>     Not Recorded
## 4                     Not Recorded                <NA>     Not Recorded
## 5 4ef15afb4b2723d8f3d81e51ec7afefe 2018-02-26 19:58:54              pet
## 6                     Not Recorded                <NA>     Not Recorded
##      lead_type lead_behaviour_profile  has_company     has_gtin average_stock
## 1  Not Recorded           Not Recorded Not Recorded Not Recorded             0
## 2  Not Recorded           Not Recorded Not Recorded Not Recorded             0
## 3  Not Recorded           Not Recorded Not Recorded Not Recorded             0
## 4  Not Recorded           Not Recorded Not Recorded Not Recorded             0
## 5 online_medium                    cat Not Recorded Not Recorded             0
## 6  Not Recorded           Not Recorded Not Recorded Not Recorded             0
##   business_type declared_product_catalog_size declared_monthly_revenue
## 1  Not Recorded                             0                        0
## 2  Not Recorded                             0                        0
## 3  Not Recorded                             0                        0
## 4  Not Recorded                             0                        0
## 5      reseller                             0                        0
## 6  Not Recorded                             0                        0
##   won_date_year convertion_status
## 1 not converted     not converted
## 2 not converted     not converted
## 3 not converted     not converted
## 4 not converted     not converted
## 5          2018         converted
## 6 not converted     not converted
```

```r
monthly_revenue_0 <-  Joined_Dataset%>%
  filter( declared_monthly_revenue == 0) %>%
  group_by(lead_type, convertion_status) %>%
  summarize(total_0_revenue = n(), .groups = "drop")
total_leads<- n_distinct(Joined_Dataset$mql_id)
monthly_revenue_0 <- monthly_revenue_0 %>%
  mutate(total_allleads = total_leads)
head(monthly_revenue_0 )
```

```
## # A tibble: 6 x 4
##   lead_type      convertion_status total_0_revenue total_allleads
##   <chr>          <chr>                       <int>          <int>
## 1 Not Recorded   converted                       6           8000
## 2 Not Recorded   not converted                7158           8000
## 3 industry       converted                     118           8000
## 4 offline        converted                      99           8000
## 5 online_beginner converted                     55           8000
## 6 online_big     converted                     118           8000
```

```r
ggplot(monthly_revenue_0, aes(x = lead_type, y = total_0_revenue, fill = convertion_status)) +
  geom_col(position = "stack") + labs(title = "Distribusi Revenue 0 per Lead Type (Stacked)", x = "Lead
  theme_minimal() + geom_text(aes(x = lead_type, y = total_0_revenue, label = total_0_revenue),position
  theme(axis.text.x = element_text(angle = 90))
```




```
Stat_desc <- Joined_Dataset %>%
  select ( declared_monthly_revenue, declared_product_catalog_size, average_stock)
summary(Stat_desc)
```

```
##  declared_monthly_revenue declared_product_catalog_size average_stock
##  Min.   :       0         Min.   :   0.00               Min.   :    0.00
##  1st Qu.:       0         1st Qu.:   0.00               1st Qu.:    0.00
##  Median :       0         Median :   0.00               Median :    0.00
##  Mean   :    7723         Mean   :   2.01               Mean   :   57.22
##  3rd Qu.:       0         3rd Qu.:   0.00               3rd Qu.:    0.00
##  Max.   :50000000         Max.   :2000.00               Max.   :45778.00
```

## 8.Lead Type Distribution by Revenue Tier: Prioritizing High-Value Segments

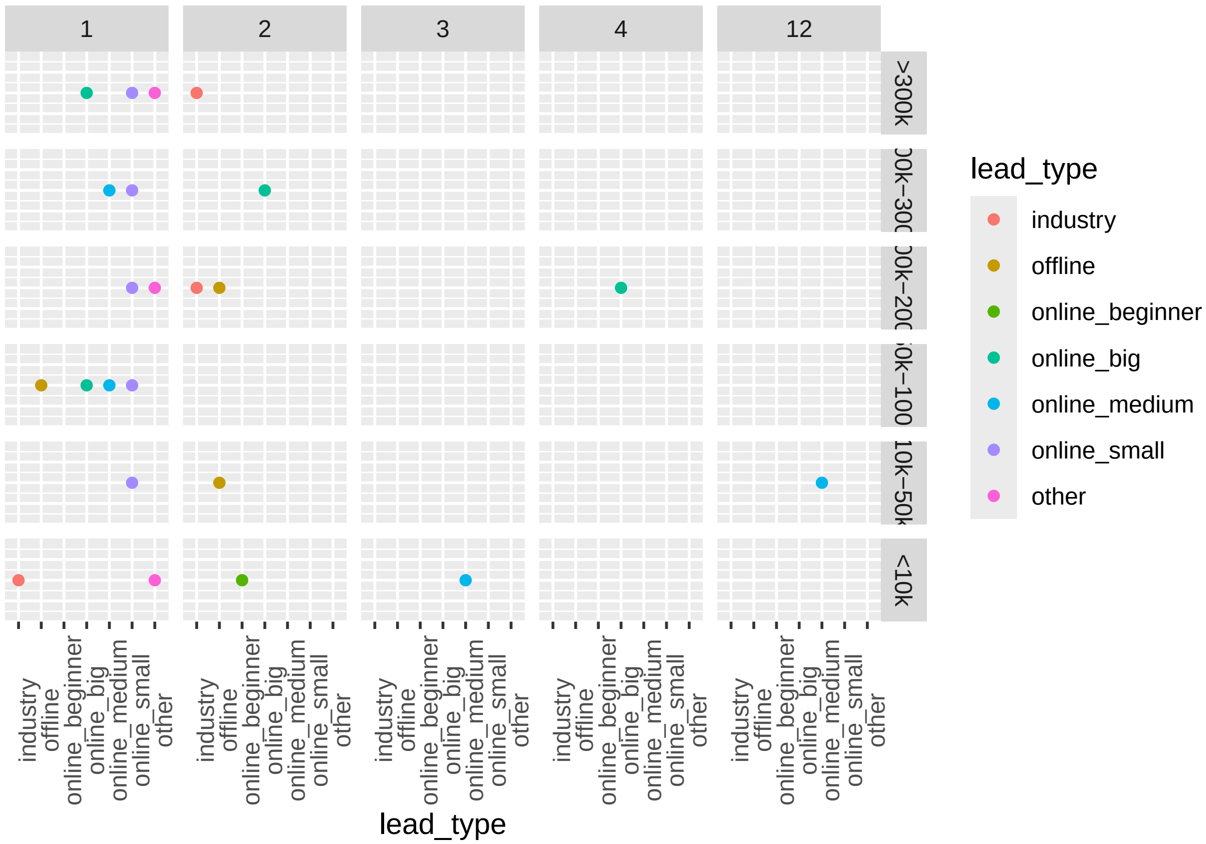
To better understand how different lead types relate to revenue generation, i categorized declared_monthly_revenue into clear tiers (e.g., <10K, 10K–50K, >300K) and factored them to visualize progression. The horizontal facet labels (1, 2, 3, 4, 12) show the number of leads per lead type, while the right-side labels represent the revenue range brackets. This structure helps identify which lead types consistently contribute to higher-value brackets.

The chart highlights a strong positive association between lead type and revenue potential—lead types like online medium and online big frequently appear in mid to high revenue ranges, suggesting they attract higher-CLV customers. This makes sense, as these lead types are often aligned with business maturity or budget size. In contrast, types like online_beginner or offline tend to dominate the lower tiers, indicating limited immediate value. Understanding this relationship is critical for marketers aiming to optimize acquisition spend and prioritize high-yield segments. Actionable key:

1. Focus acquisition efforts on high-CLV lead types like online medium and online_big, which show strong alignment with high revenue ranges and likely deliver better ROI over time.

2. Reassess targeting or positioning for low-revenue lead types such as offline or online_beginner, either to nurture them toward upsell paths or to optimize budget allocation away from low-return segments.

```
monthly_revenue <-  Joined_Dataset%>%
  filter( declared_monthly_revenue != 0) %>%
  group_by(lead_type, declared_monthly_revenue, convertion_status)%>%
  summarize(total_leads = n_distinct(mql_id), .groups = "drop")
monthly_revenue <- monthly_revenue %>%
  mutate(revenue_range = case_when(
    declared_monthly_revenue < 10000 ~ "<10k",
    declared_monthly_revenue < 50000 ~ "10k-50k",
    declared_monthly_revenue < 100000 ~ "50k-100k",
    declared_monthly_revenue < 200000 ~ "100k-200k",
    declared_monthly_revenue < 300000 ~ "200k-300k",
    TRUE ~ ">300k"
  ))
monthly_revenue$revenue_range <- factor(monthly_revenue$revenue_range,
                                        levels = c(">300k", "200k-300k", "100k-200k", "50k-100k", "10k-5
monthly_revenue <- monthly_revenue %>%
  group_by(revenue_range, lead_type) %>%
  summarize( total_leads_per_revenuerange = sum(total_leads), covertion_status = first(convertion_statu
  arrange(lead_type, revenue_range)
ggplot(monthly_revenue) + geom_point(mapping = aes (x = lead_type, y = total_leads, colour = lead_type)
  theme( axis.text.x = element_text(angle = 90),axis.text.y = element_blank(), axis.ticks.y = element_b
```


1
2
3
4
12
>300k
lead_type
industry
offline
online_beginner
online_big
online_medium
online_small
other
lead_type