## Data Exploration

The Tickets and the Orders table contain identifiers for the orders, tickets, riders and drivers. But tickets do not contain any identifiers that will allow to match (join) them to the orders and determine the order, driver or rider involved in the issue.

There are also time fields in those tables that cover the same period (3 months). From the natural human behavior, we can expect that the time of a ticket is close to the time of an order and the “health of the market” can be studied base on the time “anonimized” with different time granularities.

There are 24 911 unique drivers and 869 596 unique riders in Orders table.

Top-10 active Riders

| rider_id | Orders |
|---|---|
| 2.7886486322534154e+17 | 1180 |
| 1.8892669804918825e+18 | 922 |
| 1.3306215521164321e+18 | 815 |
| 2.028041591563518e+18 | 651 |
| 2.056657841207826e+18 | 521 |
| 1.282729784074065e+18 | 517 |
| 1.198114796372781e+18 | 503 |
| 7.191135413425664e+17 | 499 |
| 1.5255663158908874e+18 | 478 |
| 1.530688250440306e+18 | 431 |

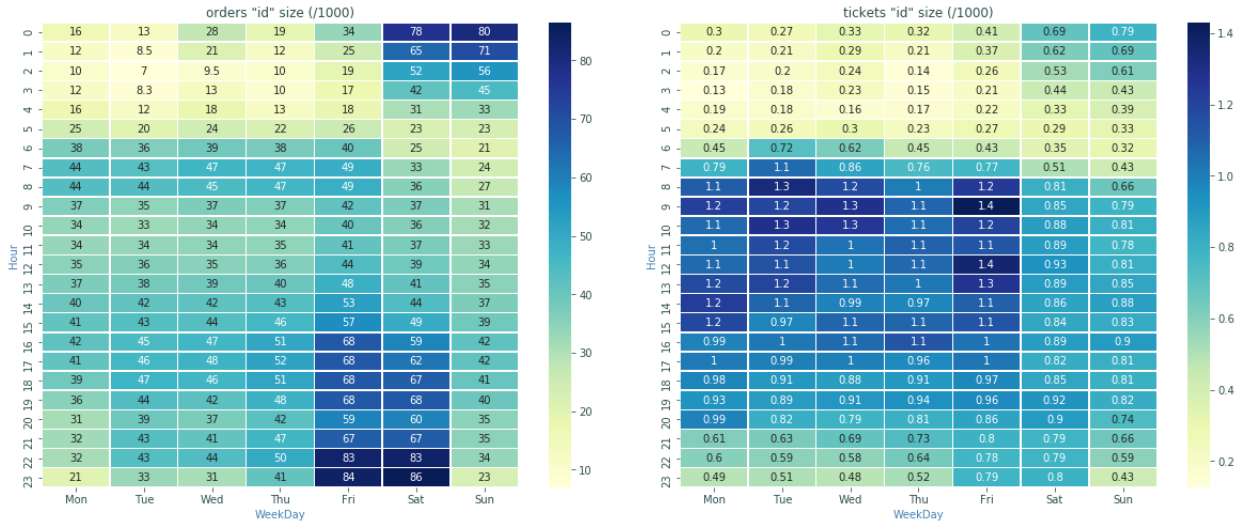
Top-10 active Drivers

| driver_id | Orders |
|---|---|
| 5.88235151312348e+17 | 3616 |
| 1.5627684533225088e+18 | 3241 |
| 1.6037248622321295e+18 | 2738 |
| 9.294467319079633e+17 | 2685 |
| 1.3594598411200812e+18 | 2659 |
| 1.825835975086783e+18 | 2655 |
| 7.280965789695789e+17 | 2496 |
| 5.95662008642942e+17 | 2490 |
| 1.7639176137739246e+18 | 2486 |
| 7.275374256221981e+17 | 2419 |

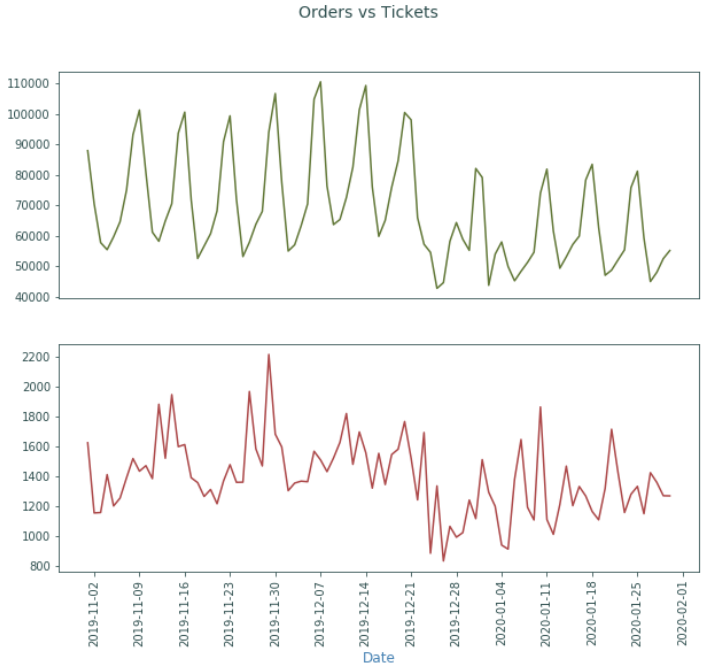
As for the time of creation:

- Tickets were created from 2019-11-01 00:00:16 to 2020-01-30 23:58:35
- Orders were created from 2019-11-01 00:00:00 to 2020-01-31 00:00:00

So the data presents the information for 3 consecutive months. To keep the data for the same periods cases from Orders table from the Jan 31 were filtered out.



From the plots we can observe that mostly tickets are generated during the daytime from 8am to 8 pm, while the most orders appear at the evenings. Different patterns are observed also over weekdays - workdays until Friday evening are different from weekends.




Correlation between number of Orders and number of created Tickets on a daily basis is 0.399, which means that the number of orders is not the only feature that can help to define the number of tickets in the future.

The Orders table contains the driver and rider IDs, but the Tickets table's field "user_type" shows the following user types: 'rider', 'driver', 'scooter_rider', 'fleet_supply', 'business_user', 'eater', 'provider', 'courier', 'courier_supply'. This demonstrate that the Tickets table contains all the tickets for the period for different businesses (like taxi, food delivery, scooter-sharing, etc.) but at the same time it is not clear if Orders table contains only "car-rides" (which is probably not true due to "Nans" in driver_id) or not. Or in other words, it is not clear if Tickets table needs to be filtered or not before matching with the orders table. But without an extra knowledge I can only guess that

a) from one side - we need to keep 'rider' and 'driver' from Tickets and maybe fleet_supply,
b) from the other - there are almost 5% of Nans in Orders "driver_id" field and this part of the data may represent the 'non-taxi' activity. Without additional information I'll proceed with non-filtered (as is) data.

1. **Forecast the daily number of agents that will be needed to work on that market to cover the expected volume of tickets for a minimum of 31 days.**

In the tickets data there are fields that represent the time for response and the time for the ticket to be solved and values vary a lot for different tickets.

For a great portion (52.85%) of tickets the time from “first response” to “solved” (*response_solved_time*) is equal to 0. The cause is not obvious for me as any action (like “reading” and “clicking”) takes time, so it may be explained only if the tickets were closed (or solved) by some automated solution and, therefore, no Customer support specialist needed for this part of tickets in case if it’s true.
It also can be supposed that if the time to solve an issue is from 0 to some number of minutes it may tell us that the tickets original issue was solved by the Customer Support specialist (during the conversation or a call). But if the issue is more serious it can take days to find the solution and will need to include other teams (Departments) to handle it.
For 0.497% of the tickets the time from “first response” to “solved” was in the range (0;1] hours and 11.6% in the range (0;8] hours (as 8 hours is a standard duration of the shift (working day)).

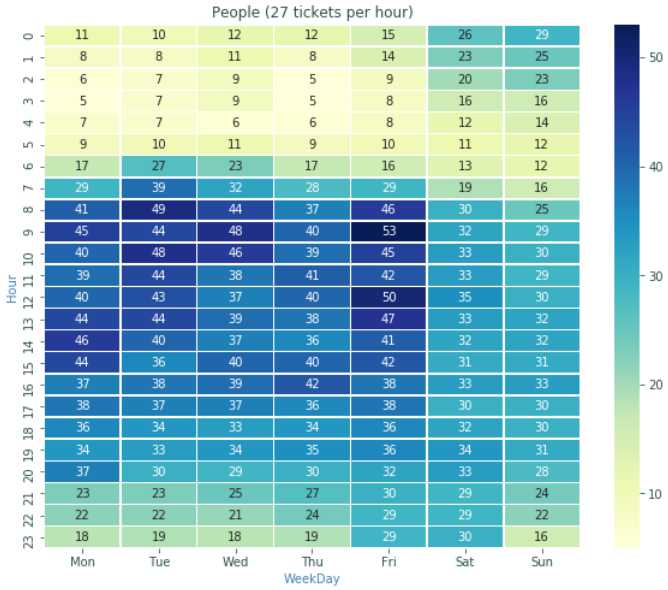
So there are several types of issues (by time and difficulty):

- Can be solved automatically (like with chat-bots) – people are not involved
- Can be solved by Customer support specialist – a person is involved (time to solve is relatively short)
- Should be described and rerouted to other specialists or teams (billing, data warehouse, etc.) – many people are involved

There is no data in the tables provided that will allow to calculate the (average) time needed to deal with the ticket as well as the information about the tickets that need to be reviewed manually, therefore it is hard to estimate the amount of manual work needed.

If we assume that all the provided tickets need to be manually reviewed then we can utilize only the information about the number of tickets to assess people-resources needed.

As an estimate for the time needed for a Customer Support specialist to deal with relatively simple tickets or to reroute them to other Teams I will proceed with the average *response_solved_time* for the tickets within an Hour range ([0;1]) and the estimate is 2.2 minutes which makes around 27 tickets per hour for 1 person. Taking this number as a reference we can estimate the number of people needed according to the distribution of tickets by day and time (“tickets load”).



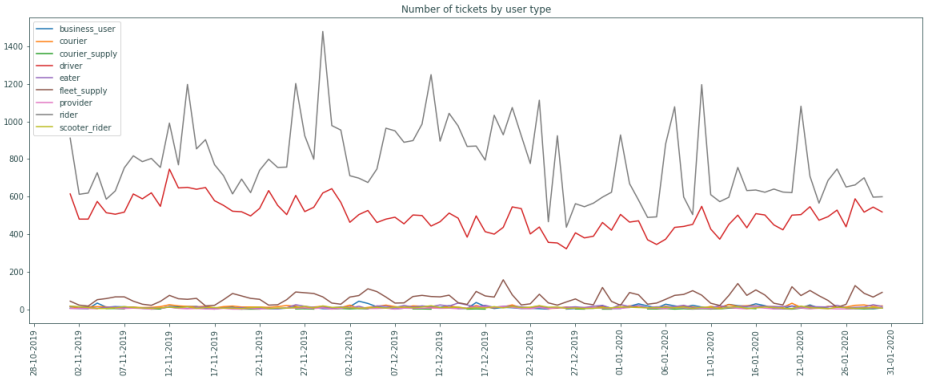

This general solution may be a benchmark for further improvements and better scheduling the “supply” based on the ticket “demand”.

Ideas for improvements:

- Number of tickets depends on number of orders (but the data shows that the correlation is quite low (0.4))
- Number of orders depends on the season and a lot of other features (there were Christmas and New Year holidays in the provided period, there are less scooter-rides during winter in some countries, etc.)
- Number of tickets depends on the day and hour.
- Number of tickets depends on the “user_type” (some users create more time-consuming tickets)
- In-coming and out-coming tickets take in average different time to resolve (first response time in average for 'is_incoming'=False is 10 times less than for a True value)
- Number of tickets depends on the maturity of the market (in more mature markets some solutions were developed and implemented earlier)
- Some data may need to be filtered out (for example if automated solution is applied to deal with some simple tickets)
- Data for relatively similar markets may be included as the reference
- Number of agents depends on the speed - the “average” time to close a ticket

2. **First, provide a formal definition for "TPR". Then analyze market health, provide insights and conclusions. Explain these insights and conclusions both as you would present to another technical analyst, and to a stakeholder not familiar with the technicalities.**

- The definition needs to include the size of the business (market), so the change in numbers could be comparable and relative to the number of orders (or rides) and representing the relative quality (which also can be derived from naming of the ratio which is "Tickets per Ride").
- The data from the "Data Exploration" section shows us that the number of tickets and the number of orders vary a lot during the day and during the week. That is why the definition of "Tickets Per Ride" (TPR) should represent the longer period to absorb this "seasonality" in the data taking the period of several consecutive weeks (1w,2w,…).
- The definition also depends on the speed of reaction to the changes, and taking into account that the reactive measures may take time to show an effect. So it may be excessive to provide a more volatile hour or daily moving average rather than a "weekly moving average" estimate.
- The definition could be business-specific – some tickets reasons can be common to different businesses while other will be specific (the case with "dead" battery on a scooter is different than complains about the cleanness of the car interior or food quality).

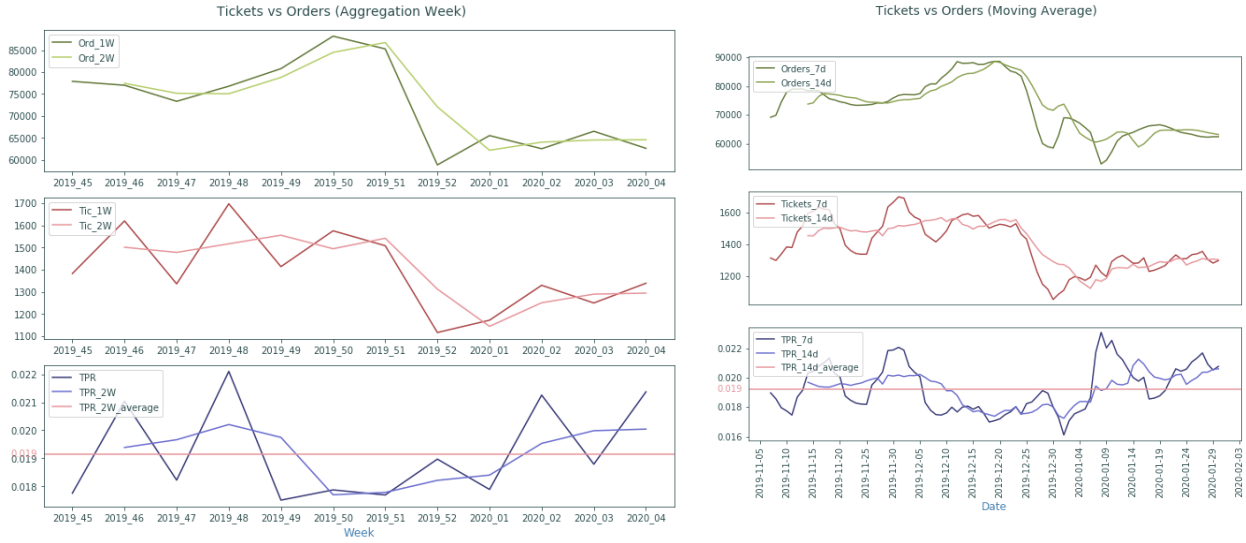



We can observe that 'rider', 'driver' and 'fleet_supply' users generate the greatest number of tickets in the data and effect from smaller in size businesses will be absorbed by a trend of a larger one.

If we look closer into the data then we will observe that the number of orders in January was lower than in previous months, but the relative number of tickets started to grow.

For the purpose of this analysis and to capture early market health I'd stay with 2 full weeks (Monday – Sunday) moving average number of tickets per orders.
More formal:

$$TpR = \left( \left( \frac{\sum Tickets}{\sum Orders} \right)_{(week\ -1)} + \left( \frac{\sum Tickets}{\sum Orders} \right)_{(week\ )} \right) / 2$$




As we can see from the left plot despite the fact that the number of orders decreased in January, the number of tickets started to grow and with relative indicator of market health starts to grow over the “average” level. There was a relatively better performance observed during December. From the graph for user_type distribution it is clear that the tickets were mainly created by riders and this number is volatile demonstrating huge spikes.

Both – the decrease in number of orders and relatively stable and high number of tickets generated during January reflects in the increased TPR as an indicator of worsening of market health.

A deeper understanding of tickets reasons and more careful business-specific analysis may shed light to the causes which in turn will allow to locate the problem and investigate for the solution. And I am ready to study and investigate later this with more detailed data and with better knowledge and the description of the fields.

**Finally, include with your submission a motivational letter covering why you would like to become a Business Analyst for Bolt Customer Support division.**

I am product and business-minded, pragmatic, self-starter analyst and mid-level data scientist with hands-on experience in finance, risk-management, e-commerce and marketing; oriented not only on results, but also on improving and optimizing processes, achieving business goals via experiments and getting extra value from data through custom analytics, programming, routine automation and machine learning.

My data analysis full-cycle hands-on experience (SQL, Python, Excel) allows me to toss the data in any possible way and find the hidden there answers. 10 Months as Data Scientist in FinTech Company and 6+ years in Banking risk management help me to understand risks and complex (interdisciplinary) processes. My personal projects (web, trading, analytics) also added a lot to my skillset.

It is important to mention that I like to organize, improve and simplify things, make them practical not only through analysis of the data but also by taking into account general picture of the business.

I have a number of compelling personal skills (humanity, empathy, honesty, humor) which makes me a good team-player.

All that makes me a proper candidate for this role.