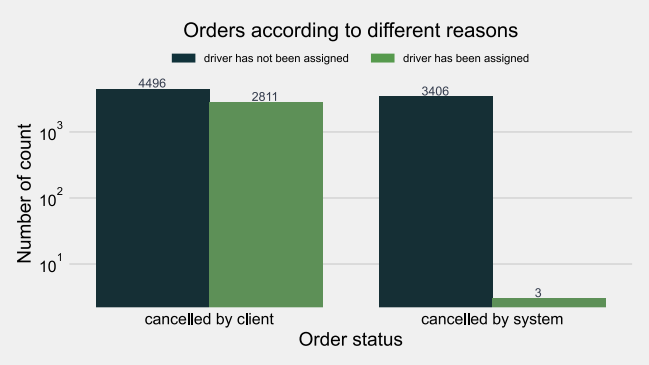
Data Science Practice Assignment 2 report

SID: 12110821
Name: Zhang Chi
Pull Request: 12110821 张弛

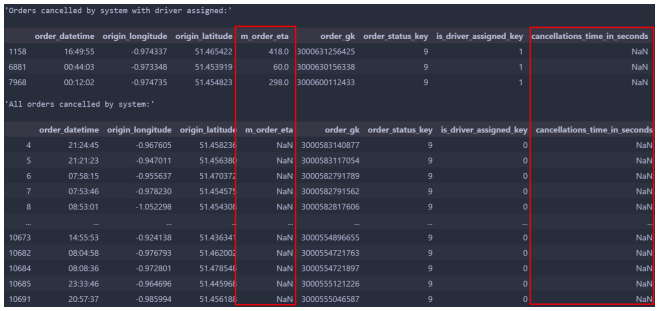
Q1

In this bar chart, there are two categories of order status (the order canceled by the client or the system), which are set in the x-axis. In each category, separated into two groups (whether has assigned a driver or not), the number of orders is counted and displayed on the top of each bar.



According to the image above, **the number of orders canceled by the client is larger than the number of orders canceled by the system**, with the order that no driver assigned accounts for more in both categories. Besides, **most orders are canceled by the client without assigning a driver, which is 4496.**

However, there are 3 orders canceled by the system even though the driver has been assigned, which is the least among all the categories. I get details and display below, from which we can find that these three `cancellations_time_in_seconds` are Nan while `m_order_eta` are not Nan. Perhaps the driver got to the pick-up point but the client was gone, so the system cancelled the order.



'Orders cancelled by system with driver assigned:'

	order_datetime	origin_longitude	origin_latitude	m_order_eta	order_gk	order_status_key	is_driver_assigned_key	cancellations_time_in_seconds
1158	16:49:55	-0.974337	51.465422	418.0	3000631256425	9	1	NaN
6881	00:44:03	-0.973348	51.453919	60.0	3000630156338	9	1	NaN
7968	00:12:02	-0.974735	51.454823	298.0	3000600112433	9	1	NaN

'All orders cancelled by system:'

	order_datetime	origin_longitude	origin_latitude	m_order_eta	order_gk	order_status_key	is_driver_assigned_key	cancellations_time_in_seconds
4	21:24:45	-0.967605	51.458236	NaN	3000583140877	9	0	NaN
5	21:21:23	-0.947011	51.456380	NaN	3000583117054	9	0	NaN
6	07:58:15	-0.955637	51.470372	NaN	3000582791789	9	0	NaN
7	07:53:46	-0.978230	51.454575	NaN	3000582791562	9	0	NaN
8	08:53:01	-1.052298	51.454308	NaN	3000582817606	9	0	NaN
...	...	...	...	...	...	...	...	...
10673	14:55:53	-0.924138	51.436341	NaN	3000554896655	9	0	NaN
10682	08:04:58	-0.976793	51.462002	NaN	3000554721763	9	0	NaN
10684	08:08:36	-0.972801	51.478548	NaN	3000554721897	9	0	NaN
10685	23:33:46	-0.964696	51.445966	NaN	3000555121226	9	0	NaN
10691	20:57:37	-0.985994	51.456188	NaN	3000555046587	9	0	NaN

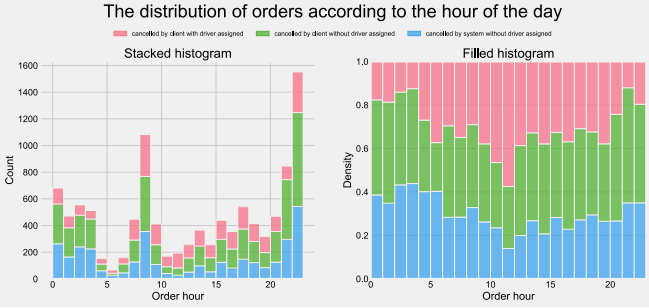
As only 3 orders are canceled by the and system after a driver has been assigned, I will drop this category in the following analysis.

Q2

There are two subplots in this figure.

On the left is a stacked histogram showing the number of canceled orders per hour of the day, with colors to distinguish between different cancellation reasons. We can see that **the number of canceled orders is the largest at 23:00 than the other time**. This may indicate that customers are more likely to cancel orders at the end of the day. It is speculative as fewer drivers are working at night or the safety can not be guaranteed for customers. Notably, the number is also much larger at 9:00, when people are in the morning rush hour and may have less time to wait for the order.

On the right is a filled histogram, presenting a proportion of different reasons in each hour. It is abnormal that **consumers are more likely to cancel an order even though the driver has been assigned at 11:00, which is above 50%**. A reasonable explanation is that the client can easily change his/her mind (hard to decide where to have lunch).



Q3

Checking missing values

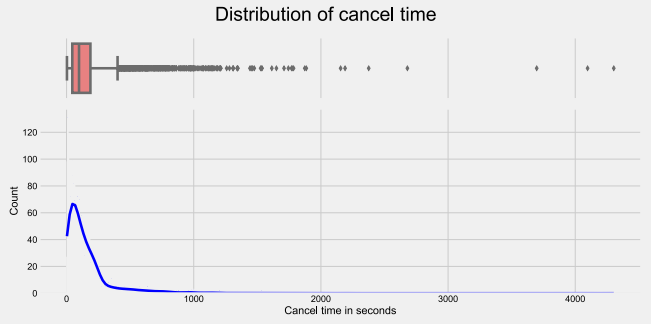
Based on the table information (presented below), all missing values of `cancellations_time_in_seconds` are caused by the reason that the order was canceled by system, which is a regular missing type.

```
The number of records with 'cancellations_time_in_seconds' is Nan: 3409
The number of records cancelled by system: 3409
The number of records satisfying both conditions: 3409
```

	Number of missing values
order_datetime	0
origin_longitude	0
origin_latitude	0
m_order_eta	7902
order_gk	0
order_status_key	0
is_driver_assigned_key	0
cancellations_time_in_seconds	3409
order_hour	0
category	0

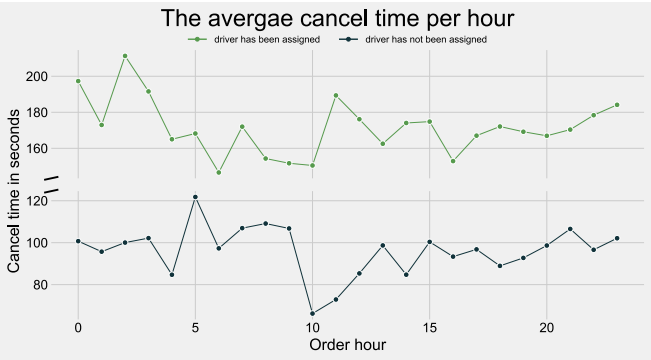
What's more, when using the `groupby` and `agg` functions, if there are missing values in the data, the pandas library automatically ignores those missing values and calculates the average of the non-missing values. Therefore, we can feel free to make the analysis.

Before deleting outliers



As shown in the left figure, it is clear that the distribution of cancel time is **extremely skewed** to the right as well as **many outliers plotted** in the boxplot. So I will filter them by using the `1.5 IQR` criteria. In addition, as the mean is not the same in the two groups, I will make it separately.

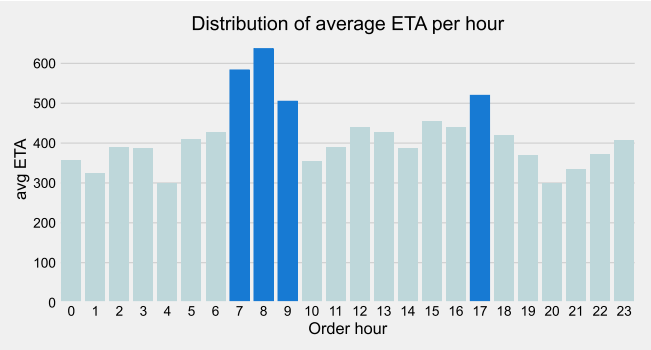
After deleting outliers



A tick is added in the middle of the y-axis, representing the change of interval as the two categories' values are not at the same level.

Examining the diagram of the lineplot, we can find that the **cancel time of driver assigned order is longer than that of no diver assigned**. It illuminates that the customer will be more patient to wait if a driver has been assigned as soon as possible.

Q4



The chart placed above reveals the average waiting time of customers before drivers arrive at the pick-up point, and **bars in dark blue are significantly high at 7:00-9:00 and 17:00 among all hours, happening to be the morning rush hour and the evening rush hour.

Q5

Inspecting the picture below, a red dot is a single order and 8 hexes are placed to cover over 80% number of orders, with the color representing the number of orders in each hex (the more the redder). The hex in the middle is the reddest, which is reasonable as the city center is the most crowded place.