Delivery Business Case Study

Ibrahim Israfilov

Index:

- Introduction

Introduction

This report is done based on the datasets provided by #### and analysis executed by me on the R Programming Language. The aim of the report is successful completion of the recruitment.

```
Rows: 190,708
Columns: 12
$ partner_id                <int> 9              |0~
$ date                      <chr> "             |~
$ business_segment          <chr> "             |d~
$ country                   <chr> "             |"~
$ avg_order_cost            <dbl> 9              |0~
$ avg_order_revenue         <dbl> 2              |0~
$ is_integrated             <chr> "     #####    |"~
$ avg_delivery_time_min     <dbl> 4              |2~
$ avg_preparation_time_min  <dbl> 2              |0~
$ orders_daily              <dbl> 2              |~
$ connected_hours           <int> 1              |4~
$ vertical                  <chr> "             |o~
```

This is the structure of the table after joining the datasets. We have tot: 190,708 observations. Obviously to make the dataset more usable we will explore and modify the dataset for out needs.

```
   partner_id              date               business_segment
 Min.    :                Length:190708        Length:190708
 1st Qu.:    ###          Class :character     Class :character
 Median  :                Mode  :character     Mode  :character
 Mean    :     ##
 3rd Qu.:
 Max.    :
   country                 avg_order_cost       avg_order_revenue
 Length:190708            Min.   :  0.000       Min.   :
 Class :character         1st Qu.:  5.834       1st Qu.:    ###
 Mode  :character         Median : 10.076       Median :
                          Mean   :  9.680       Mean         ##
                          3rd Qu.: 12.979       3rd Qu.:
                          Max.   :198.463       Max.   :
 is_integrated            avg_delivery_time_min
 Length:190708            Min.   :   0.48
 Class :character         1st Qu.:  25.77
 Mode  :character         Median :  32.63
                          Mean   :  44.12
                          3rd Qu.:  44.15
                          Max.   : 321.57
 avg_preparation_time_min  orders_daily        connected_hours
 Min.   :                  Min.   :  1.000     Min.   :  0.000
 1st Qu.:    ###           1st Qu.:  1.000     1st Qu.:  0.000
 Median :                  Median :  2.000     Median :  6.000
 Mean          ##          Mean   :  6.617     Mean   :  6.521
 3rd Qu.:                  3rd Qu.:  6.000     3rd Qu.: 10.000
 Max.   :                  Max.   :626.000     Max.   :142.000
   vertical
 Length:190708
 Class :character
 Mode  :character
```

It's always nice to have (but optional) a statistical summary of the variables so we beforehand know what to expect in the future.

Q1: How many active partners do we have in our dataset?

Active partners represent the unique number of partners occurring in the dataset. **They are #####**
However, as you can see on right in our final dataset one partner can occur several times in a daily metrics. Here is the quick histogram of the distribution of partners. As you see on Y axis we have count of the rows where the partners are located. Some of them are observed more than #### times and some of them less than ####.

Partners Distrivution (Not unique)
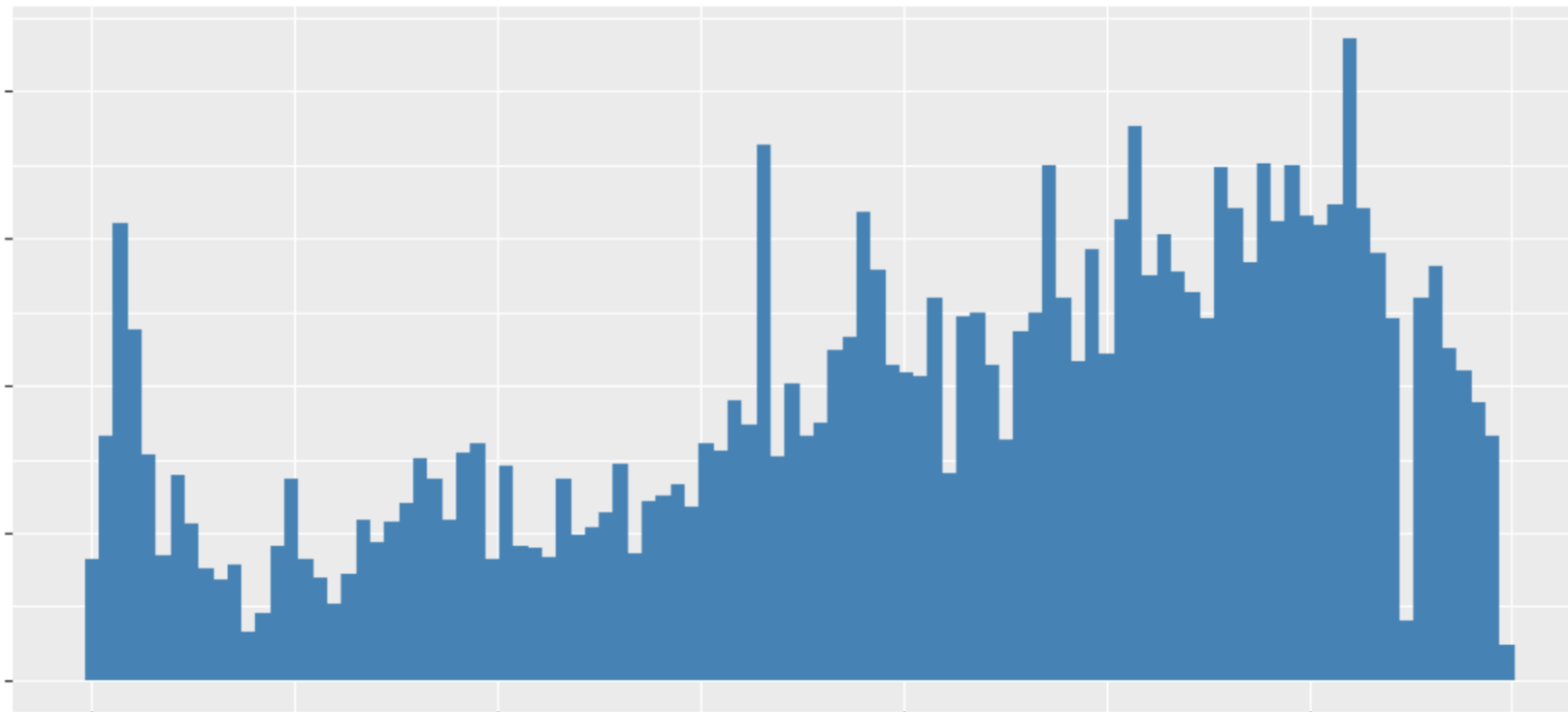
Q2: What is the breakdown per country? And per business segment?

Initially it was hard to figure out what action is intended as a breakdown. Intuitevely I have imagined that it could be the least earning in a certain observation. So that's why I have decided to do some feature engineering first and to calculate the profit for each observation with formula :
profit = (avg_revenue*orders – avg.cost*orders).

I have obtained the a new variable called Profit and done a quick statistical summary.

```
      Min.    1st Qu.    Median      Mean    3rd Qu.
 -26169.636      4.121    14.399    36.967    39.343
      Max.
  7924.597
```

This is what results we get if grouping by countries and finding the minimum of profit: **ES -#####** curious to know what happened that day

country <chr>	min <dbl>
ES	#####
IT	#####
PT	#####
UA	#####

While in business segments we have NA segment which is Because of the 2 missing values in business segment.

partner_id <int>	date <chr>	business_segment <chr>	country <chr>	avg_order_cost <dbl>	avg_order_revenue <dbl>
190512	##### 021-08-18		PT		#####
190664	##### 021-08-18		PT		#####

The observation made before -26170 is related with Q-Commerce segment.

business_segment <chr>	min <dbl>
Big Chain	-1047.11748
Chain	-193.65864
Local Hero	-198.84903
Long Tail	-800.04345
Q-Commerce	-26169.63603
Sidekick	-177.13683
NA	0.03135

Q3: What percentage of partners have delivered 80% of the orders?

Also here I have inferred the answer through intuitive approach. I would probably say that 80% stands for 80/20% law of Pareto.
Indeed, by glancing at the quick bar plot we can see that the 20% of the partners have done 80% of the orders.

To confirm the hypothesis I have calculated the found out actually 9873 unique partners have delivered 80% of the total orders.

The ##% is the percentage of partners which have delivered 80% of orders.
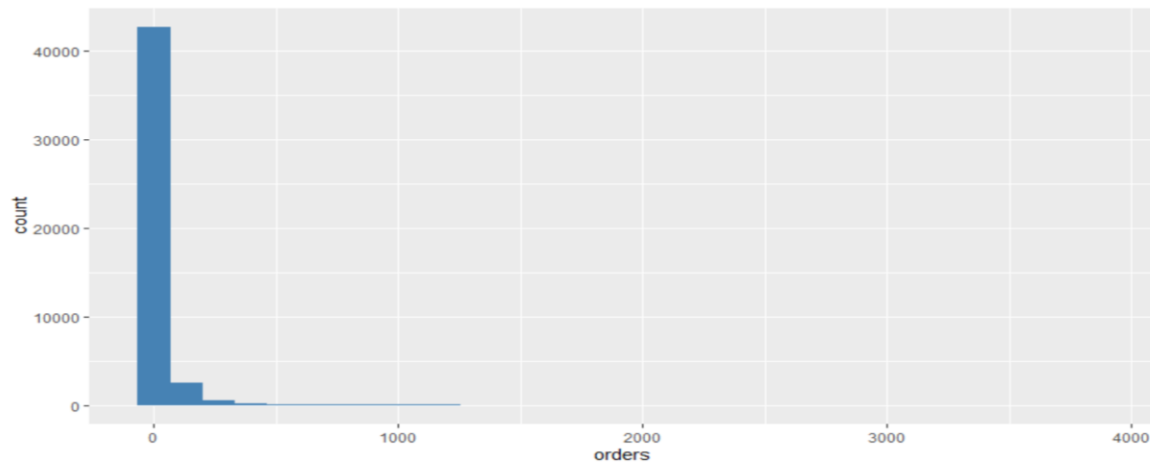
Just to understand whether I'm in a right way I have done a quick calculation and found out that ## out of ## partners have done ##k orders.

Q4: What is the average delivery time in Portugal (PT)?

By using the formula of the avg(avg_time_delivery) by Portugal which gave us the result of
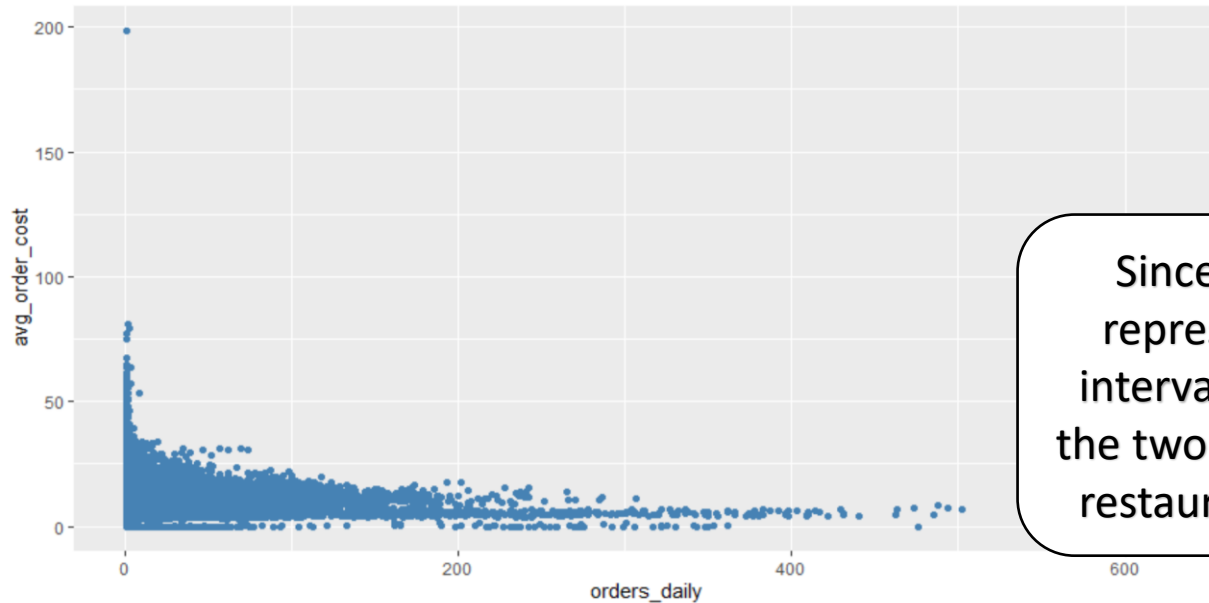
Average ## min

Q5: What is the share of orders that integrated partners delivered?

Integrated partners have delivered about **###k** orders and it makes **##%** of our total orders.

Q6: What is the distribution of the cost per order? Does it follow any known distribution? Is there anything odd in the distribution?

In order to see the relationship between the orders and the cost I have first used the scatterplot. The more the volume of orders are the less it costs.
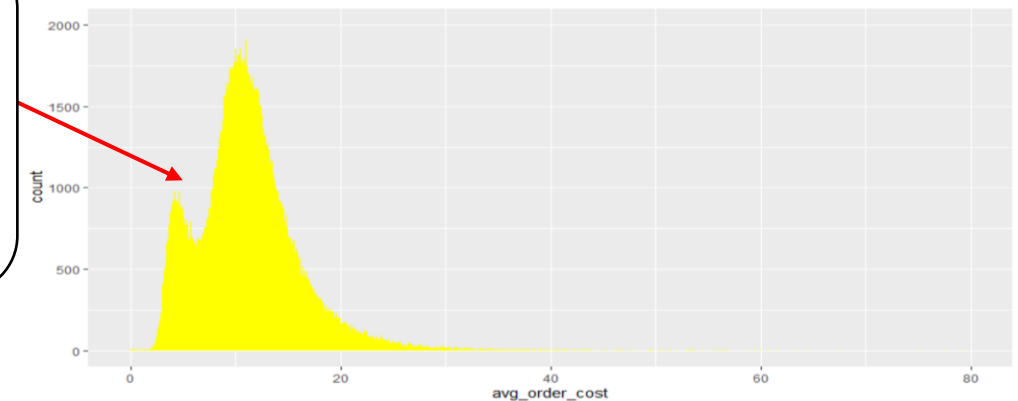

Next I have ran a quick statistical summary. To see the distribution.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.     Max.
  0.000   5.834  10.076   9.680  12.979  198.463
```

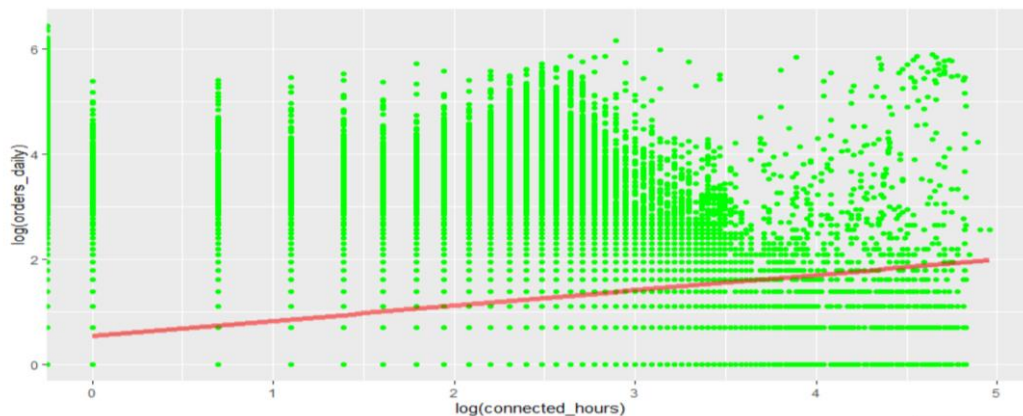We **don't expect** it to be normally distributed. Indeed it's a right skewed distribution meaning that the most of our observations are near to 0.

Since this dip represents the interval between the two type of the restaurant types.

Q7: What is the number of orders compared to connected time? Is there a correlation between the two?
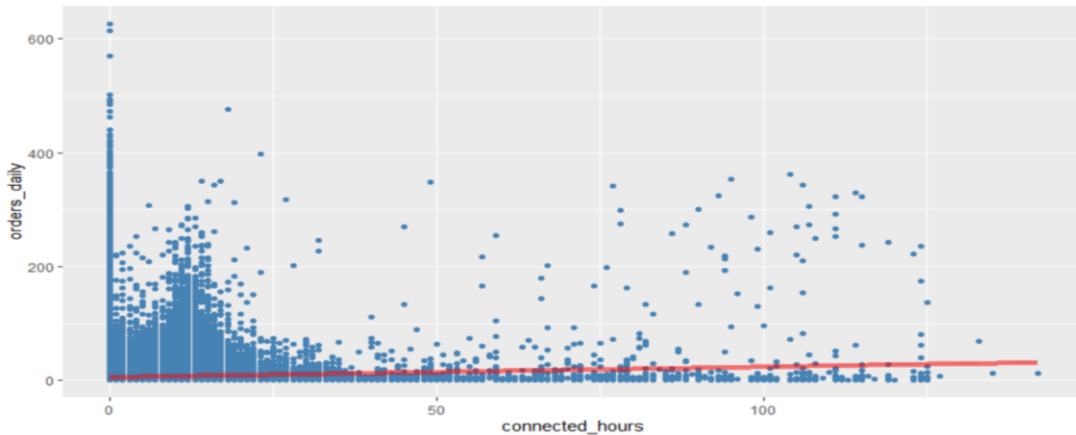
Graphical representation of two variables and the correlation line.

As seen from the plot there is a **slight positive correlation** between 2 variables.

According to the Pearson's correlation coefficient there a little correlation between connected time and orders. P<0.005 allows us to reject the Null Hypothesis and accept this finding.
So 1 hour increase in connected hours will increase the orders to 0.09

```
        Pearson's product-moment correlation

data:  joined_data$orders_daily and joined_data$connected_hours
t = 38.261, df = 190706, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.08282524 0.09173309
sample estimates:
      cor
0.08728091
```


To see the line better I have created also logarithmic plot

Q8: What are the differences in the metrics for food vs Q-commerce?

Here is the tabular representation of the vertical variable. The main difference is in the delivery time. The Q-Commerce from first glance looks to be more efficient. Moreover the Q-Commerce is a separate business segment.
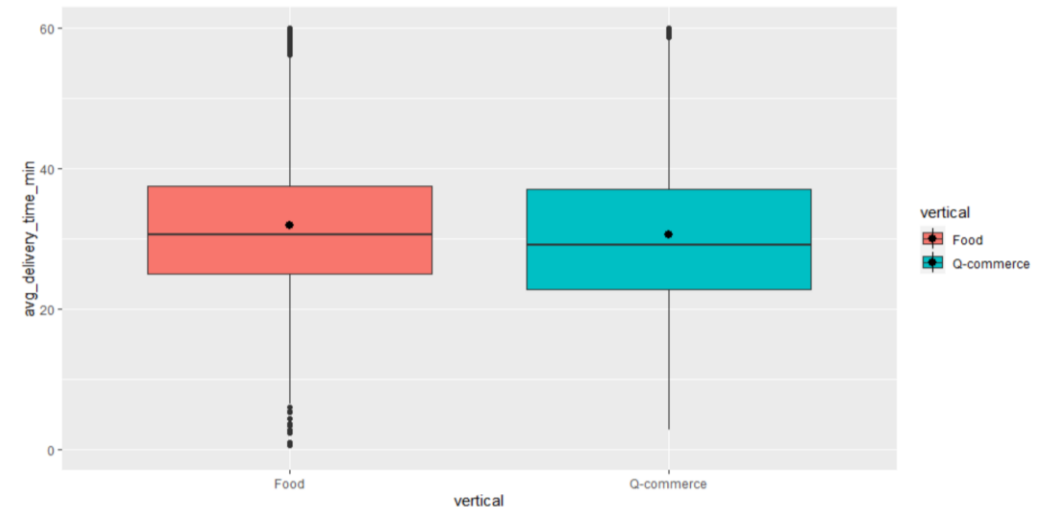
vertical <chr>	mean(orders_daily) <dbl>	mean(avg_order_revenue) <dbl>	mean(avg_order_cost) <dbl>	mean(avg_delivery_time_min) <dbl>	mean(avg_preparation_time_min) <dbl>	n() <int>	mean(connected_hours) <dbl>
Food	6.925353	17.32937	9.819183	45.88916	11.27222	164240	6.429664
Q-commerce	4.702093	11.15118	8.816125	33.12379	12.58992	26468	7.089655

Let's visualize it.
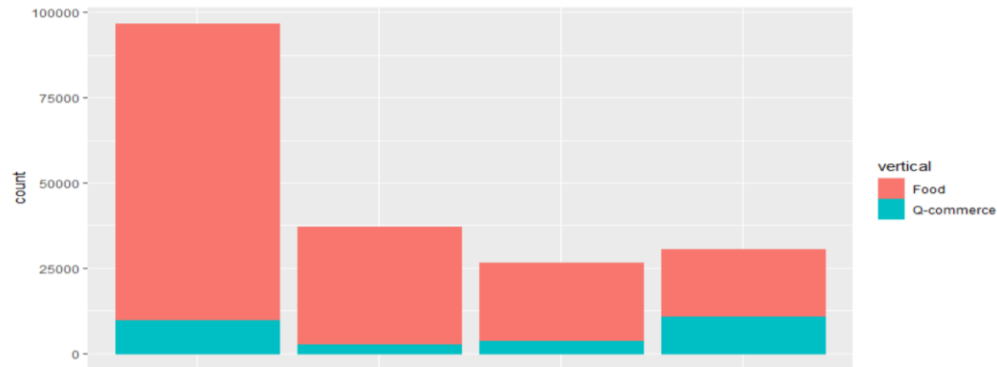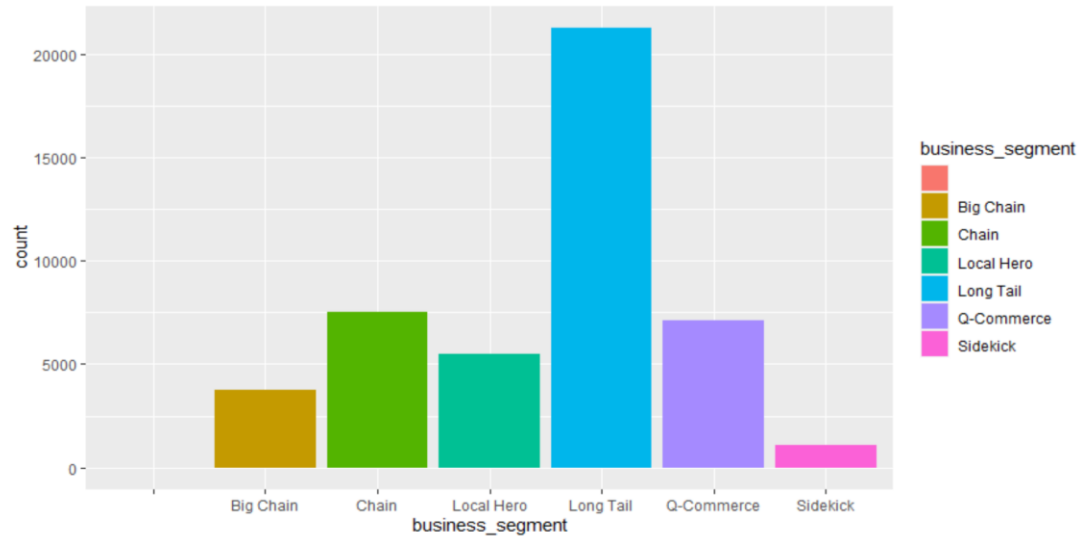

Observations by country:
The most of the observations are in ##.

Observations by Delivery Averge:
Since on the table the delivery time was the main metrics was making the difference I have decided to plot the boxplot: We see the average delivery time which is longer for Food due to the outliers. And within the limits the delivery time average almost the same for both.

Q9:Among all the possible combinations of dimensions (segments), which one has the highest number of partners?

business_segment	n
<chr>	<int>
Long Tail	21273
Chain	7529
Q-Commerce	7110
Local Hero	5475
Big Chain	3730
Sidekick	1089
NA	2

Long Tail is the most populated**.**

Long Tail has 21k partners.

# Q10: Summarize your findings in an executive summary

- Takeaways:

1. Not all partners are contributing to the revenue on the same level. As we have seen 80/20 rule works also in the delivery business.

2. Since the average delivery time is for food 45' and for Q-Commerce 33' Portugal with ###' is showing nice results.

3. Q-Commerce is the fast delivery type rather than others.

4. The most of the partners are in long-tail segment

5. On the next slides: There is a positive correlation between the cost and preparation time (probably because of the workforce). And negative correlation between cost and delivery time (The longer it takes to deliver the more it costs)

Part 2: Now that you are familiar with Glovo's partners, imagine we want to run an experiment in Spain (ES). Create two groups of partners that behave similarly in terms of metrics. Make sure all types of partners are represented

There are two points can be extracted from the correlation heat map are two points. **1 positive correlation between the cost and the preparation time.** The more it takes to prepare the more the cost is (This is what we will focus on since we need to create 2 groups of partners)

2. Moreover, for my own curiosity I have discovered another pattern which I would love to draw your attention too. This is the negative correlation between the cost and the time delivery. The faster delivery costs more. It would be the nice argument to optimize by running an experiment.
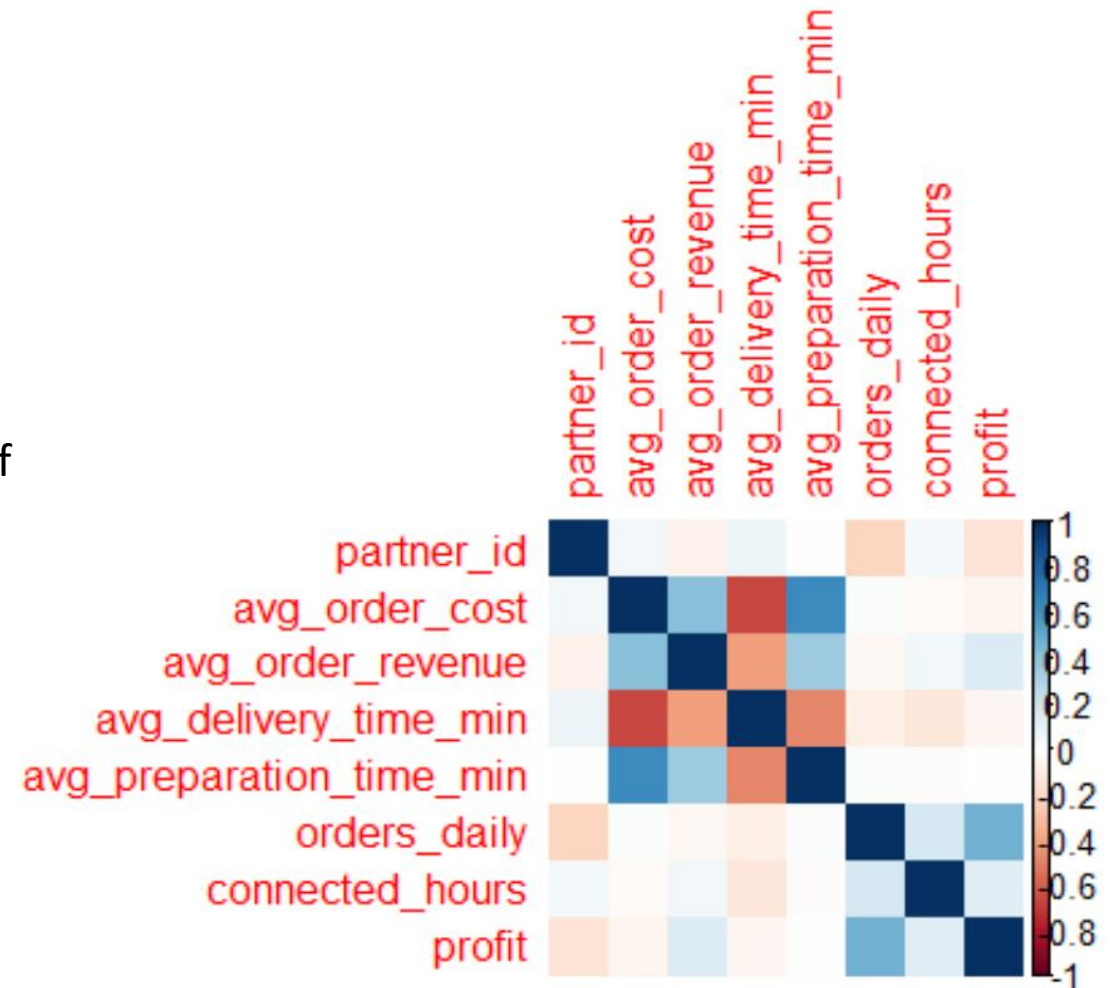
Let's deep dive.

Plot on right side reveals one more hint. The less it takes to prepare the more revenue we get. As you see the higher between 0 and 25 the cost is lesser and the revenue is higher.

We have found the pattern which allows us to develop the group criterions for an experiment in Spain. Now let's deep dive into the segments too.

What I have noticed are:
1. Long Tail and Q-Commerce has more deviation in terms of the preparation time.
2. Big Chain and Chain has nice models which could be used to optimize Long Tail.
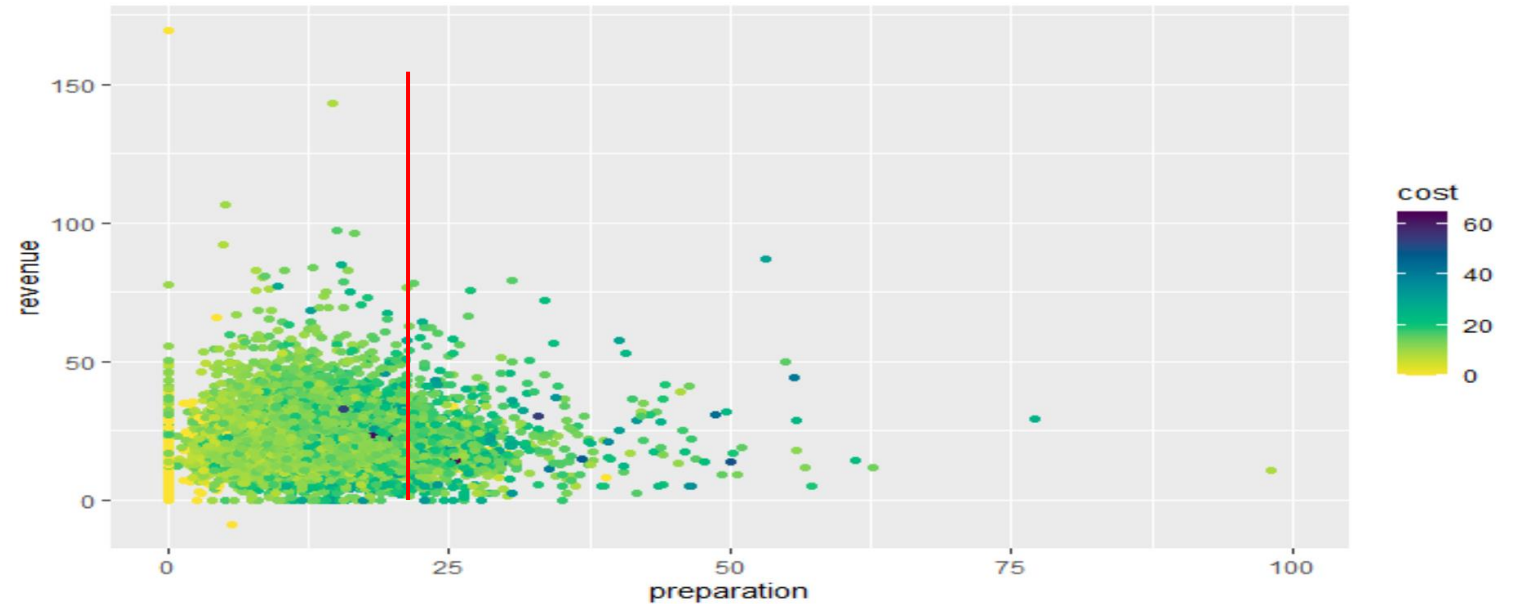
Developing two groups of partners

As I have mentioned before. The pattern of big chain and chain bring most revenue and reduces the costs. So we would love to set the threshold on 20 minutes as in these two and split the partners into 2 groups based on this threshold.

We get:
Group 1: Fast : 21613 partners
Group 2: Slow: 5372 partners

So we can run an experiment and find out the effect of the experiment by econometric models.

Thank you!!!