Telecom:
Finding Clients With Non-Optimal Plans

By: Vadim Fridman

Practicum 100 Germany
Data Analysis Course

Data

There were two datasets provided:

- On the top right a dataset containing user information such as: the user ID, the tariff plan they are subscribed to and the date of registration

	user_id	tariff_plan	date_start
0	166713	A	2019-08-15
1	166901	A	2019-08-23
2	168527	A	2019-10-29
3	167097	A	2019-09-01
4	168193	A	2019-10-16
...	...	...	...
727	166554	B	2019-08-08
728	166911	B	2019-08-23
729	167012	B	2019-08-28
730	166867	B	2019-08-22
731	166565	B	2019-08-08

732 rows × 3 columns

- On the bottom right is a dataset containing call information such as: the user ID that called, the date they called, the operator ID, the call duration, the waiting time etc.

```
In [127]: display(data)
```

	user_id	date	call_out	internal	operator_id	is_missed_call	calls_count	call_duration	total_call_duration
0	166377	2019-08-03 21:00:00	False	False	0	True	2	0	4
1	166377	2019-08-04 21:00:00	True	True	880022	True	3	0	5
2	166377	2019-08-04 21:00:00	True	True	880020	True	1	0	1
3	166377	2019-08-04 21:00:00	True	True	880020	False	1	10	18
4	166377	2019-08-04 21:00:00	True	False	880022	True	3	0	25
...	...	...	...	...	...	...	...	...	...
53897	168606	2019-11-09 21:00:00	True	True	957922	True	1	0	38
53898	168606	2019-11-10 21:00:00	True	True	957922	False	2	479	501
53899	168606	2019-11-14 21:00:00	True	True	957922	False	4	3130	3190
53900	168606	2019-11-14 21:00:00	True	True	957922	False	4	3130	3190
53901	168606	2019-11-18 21:00:00	False	False	0	True	2	0	64

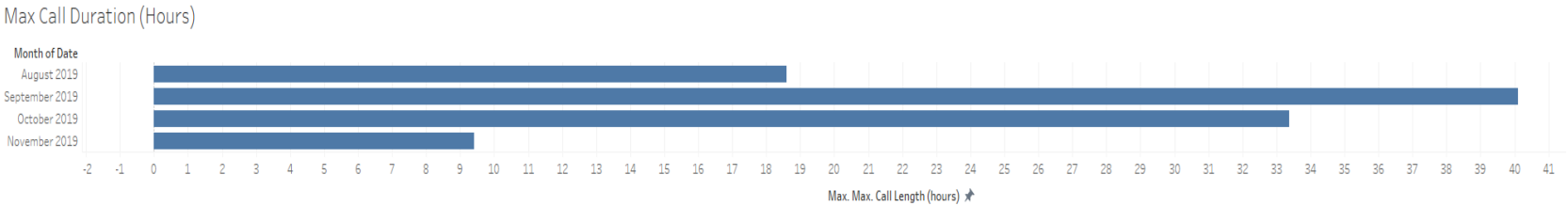
53902 rows × 9 columns

Data Preprocessing

- The given data had contained some untrustworthy information
- E.g. length of some calls were more than 40 hours (sic!)
- We filtered out untrustworthy information
- ca. 160 call logs with length 6+ hours were deleted (less than 1% of data we had)

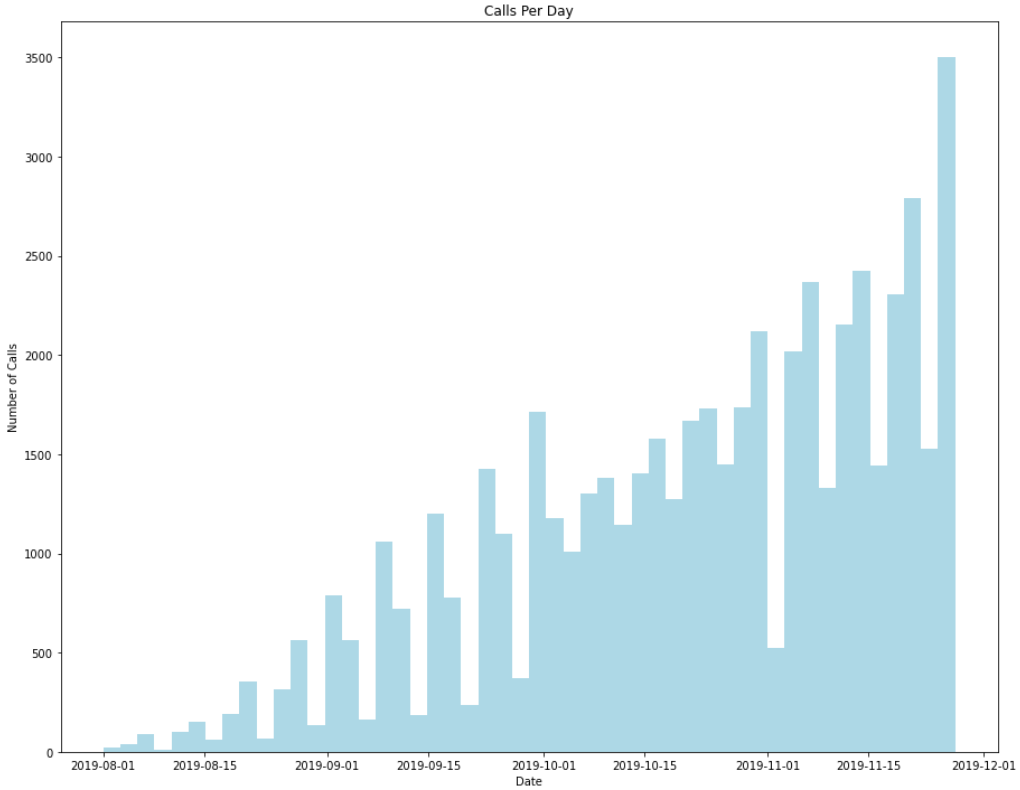
Max Call Duration (Hours)

Month of Date			
August 2019	September 2019	October 2019	November 2019
18.60	40.11	33.38	9.42



Number of calls per day (1)

- The histogram shows the distribution of number of calls (internal and external) per day within a given period
- The graph shows an upward trend in the number of calls which could be because user registration increased over the same time period, thus increasing phone calls in order to keep up with demand
- Another interesting point to notice about the graph is that there are dates where the number of calls is quite high(workdays) and then the day or few days after, it drops (weekends)



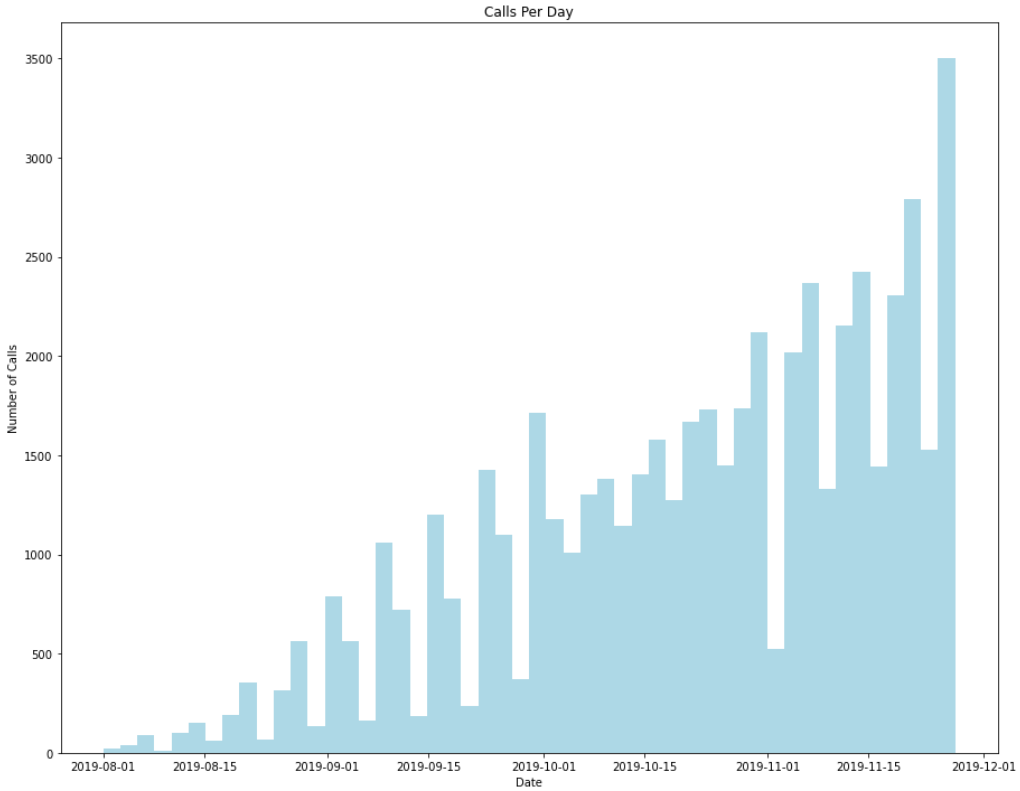
Number of calls per day (2)

The histogram also shows a problem we have with data: It is obviously skewed :-(

There at least two possible reasons:

- Seasonal factor (August is traditionally a month of lowest business activity)
- Or we possibly dealing with data were collected since a start of some service (the number of calls in the beginning of a time period is slightly above zero, it continuously grows till the mid fall)

It would be better to analyse more data to avoid conclusions flawed by skewed data!



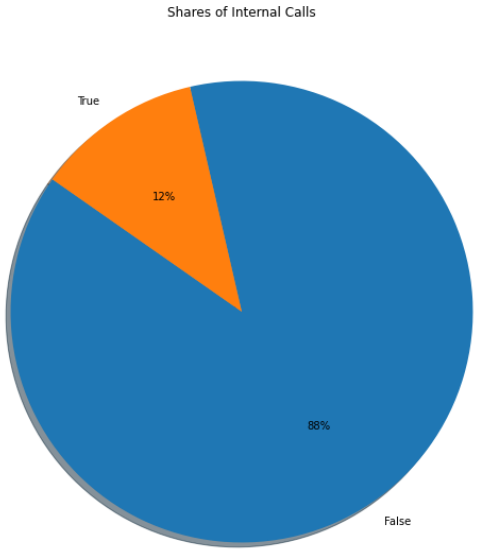
Skewed data

It would be better to analyse more data to avoid conclusions flawed by skewed data!

- But for this educational project we just using the data we have

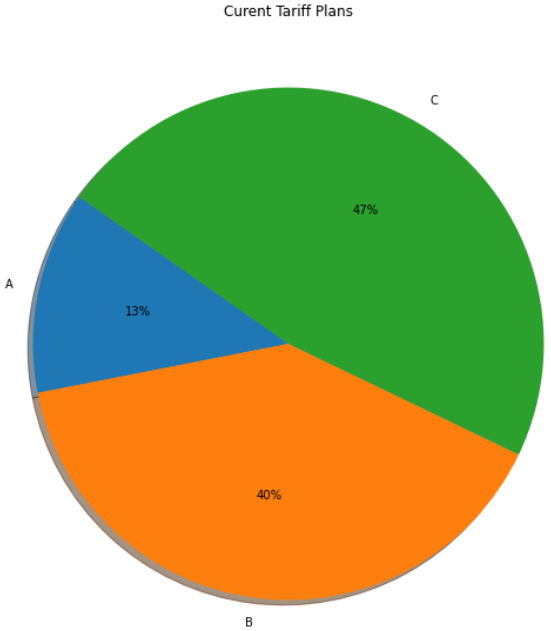
Shares of internal and external calls

- The adjacent pie chart shows the shares of calls found in the dataset, which contains internal calls (within the operator network) and external calls (from/to outside the operator network)
- According to the data, about 88% of calls found within the dataset are external calls.
- Most calls are from users to the companies

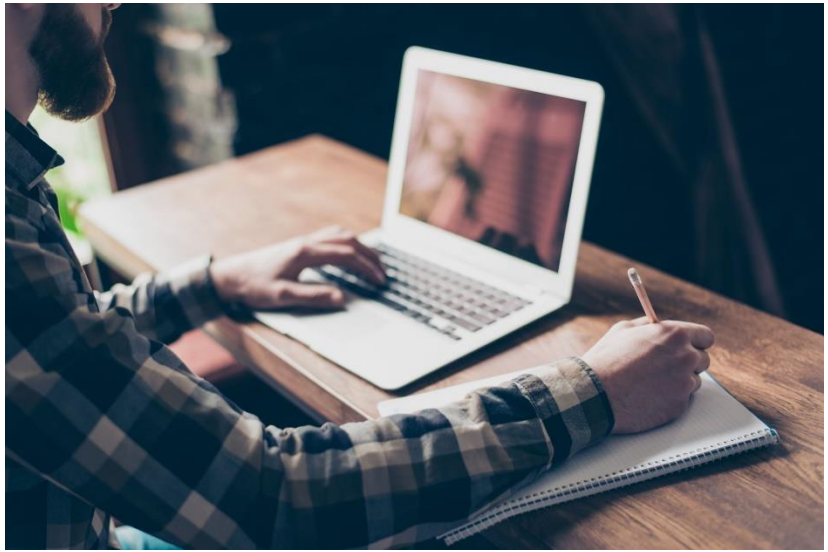


Tariff Plans

- The adjacent pie chart shows the distribution of tariff plans that users are subscribed to within the dataset
- Out of the three available plans, almost the half of the users are subscribed to the tariff C (smallest basic price/biggest minute price)
- The next most popular plan is tariff B and the least popular plan is tariff A with the biggest basic price/smallest minute price



Task



Telecom: Finding Clients With Non-Optimal Plans

Task

Find clients who:

a) Overpay for their current plans
b) Need bigger plans, since they'll be more advantageous

- Carry out exploratory data analysis.
- Find clients with non-optimal plans. Only look at active clients. Think about how you can offset the risk of a considerable reduction in revenue should all users switch to plans that are more profitable for them.
- Test statistical hypotheses.

Task:solution

To solve the task we wrote a function in Python to analyse data we have and calculate the optimal tariff plan for each user using price list we had:

Pricing

All prices are given in generic mo

`Basic price of plan A` — 5000
`Basic price of plan B` — 2000
`Basic price of plan C` — 1000
`Incoming calls for all plans` –

`Internal calls for all plans` –

`Plan A` — 0.1 units/minute
`Plan B` — 0.15 units/minute
`Plan C` — 0.3 units/minute

Outgoing (not internal) calls:

`Plan A` — 0.4 units/minute
`Plan B` — 0.5 units/minute
`Plan C` — 0.7 units/minute

```python
def calc(row):
    '''Calculating of potential price for the each customer\'s comsumer profile with every tariff plan'''

    # Tariff details

    counter_internal = 0

    internal_call_price_A = 0.1
    internal_call_price_B = 0.15
    internal_call_price_C = 0.3

    call_price_A = 0.4
    call_price_B = 0.5
    call_price_C = 0.7

    # check if more then 2000 minutes per month
    if row['Aug'] > 2000:
        # if True, calculate the price of internal calls
        counter_internal += (row['Aug'] - 2000)
    if row['Sep'] > 2000:
        # if True, calculate the price of internal calls
        counter_internal += (row['Sep'] - 2000)
    if row['Oct'] > 2000:
        # if True, calculate the price of internal calls
        counter_internal += (row['Oct'] - 2000)
    if row['Nov'] > 2000:
        # if True, calculate the price of internal calls
        counter_internal += (row['Nov'] - 2000)

    # calculate the monthly costs for each user and plan with all the plans and save the results in separat columns

    row['plan_A'] = 4*5000 + counter_internal*internal_call_price_A + row['call_duration']*call_price_A
    row['plan_B'] = 4*2000 + counter_internal*internal_call_price_B + row['call_duration']*call_price_B
    row['plan_C'] = 4*1000 + counter_internal*internal_call_price_C + row['call_duration']*call_price_C
    return row
```

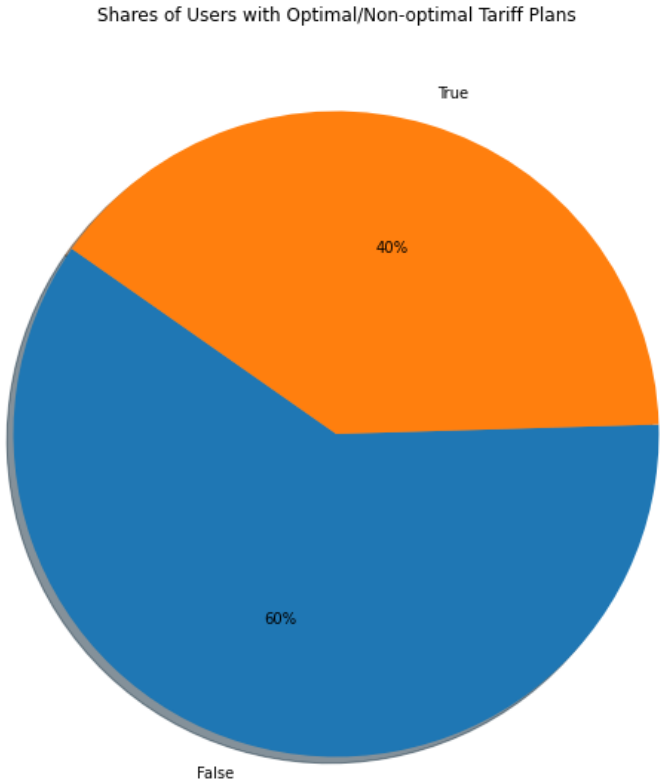
Optimal tariff plans (1)

- With our function we analysed the calling data of all active customers we had
- We calculated the price for the calls (each of available tariff plan for every user) for the given period off time
- We checked if each customer has an optimal tariff plan or not and got a result table (*part of the result is on the picture*)

current_plan	A	B	C	optimal_plan	is_optimal
C	20141.2	8176.50	4247.1	C	True
A	51398.0	47247.50	58946.5	B	False
C	20014.0	8017.50	4024.5	C	True
A	20818.8	9023.50	5432.9	C	False
C	20030.4	8038.00	4053.2	C	True
C	20101.6	8127.00	4177.8	C	True
B	20004.0	8005.00	4007.0	C	False
B	20106.8	8133.50	4186.9	C	False
B	21920.8	10401.00	7361.4	C	False
B	38606.0	31257.50	36560.5	B	True
B	20362.0	8452.50	4633.5	C	False
C	22108.4	10635.50	7689.7	C	True
C	367828.0	442785.00	612699.0	A	False

Optimal tariff plans (2)

- We also found out how many of the customers are subscribed to the optimal tariff plan: 40%
- As we can also find on the picture, 60% of customers could pay for their calls less
- The price difference can be up to a factor of 5 or even more (e.g., 5,000 instead of 20,000 and vice versa)



Hypothesis: The average call duration for customers who use optimal and non-optimal tariff plans are same

We also performed a statistical hypothesis test:

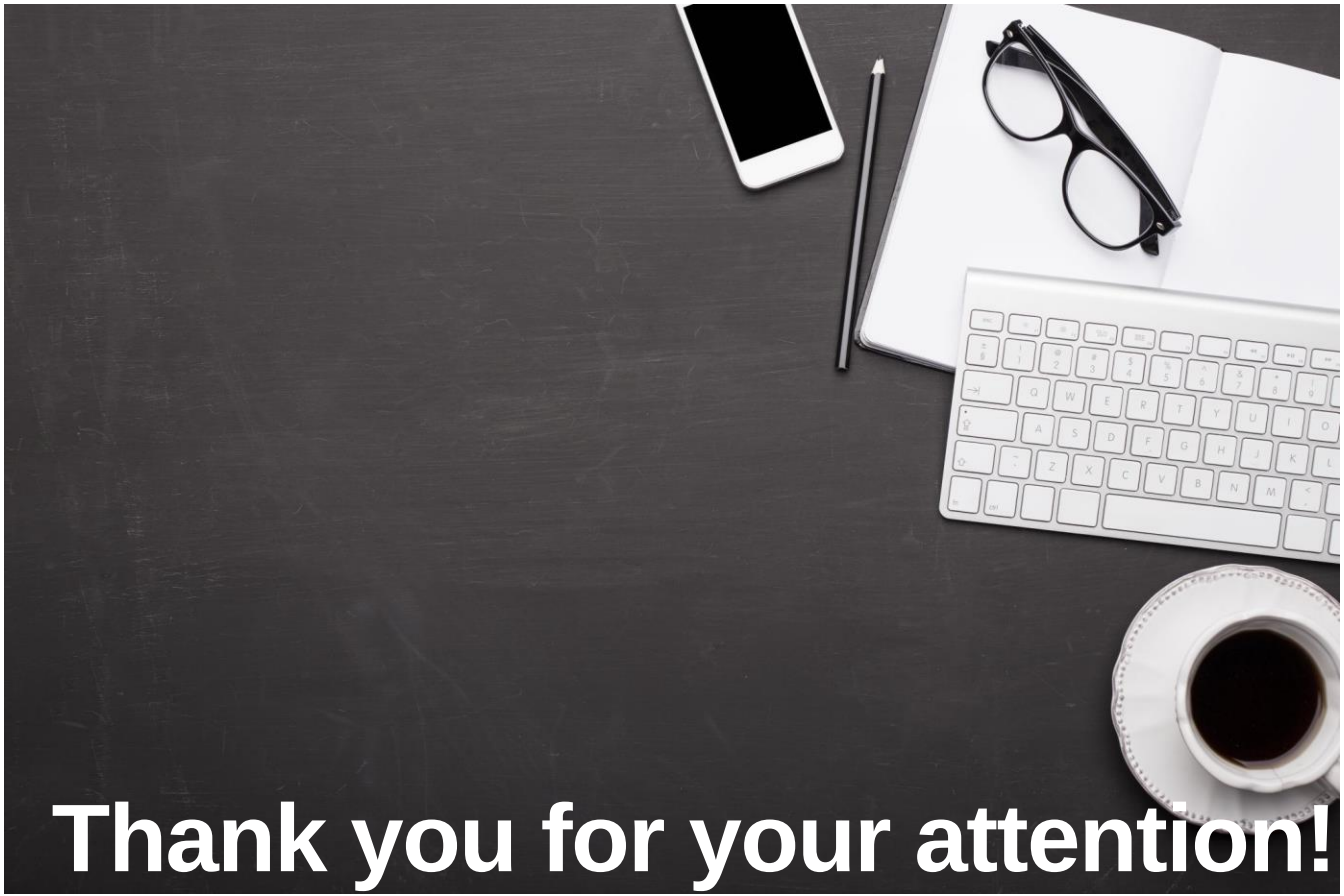
- To test the hypothesis that the average call duration are the same, an independent samples t-test to compare the means from the two groups (average call duration of customers who use optimal and non-optimal tariff plans) can be used

- The result of the t-test suggests that we should reject the hypothesis, thus indicating that the average call duration of customers who use optimal and non-optimal tariff plans do differ enough to be statistically significant.

Conclusions

- We found out, that more than the half of the customers (60%) could spent less for their calls during the given period
- We determined which tariff plan would be optimal for each of the active customers.
- If customers had access to this kind of data, many of them probably would change their tariff plan

- We are still can't understand why we as junior analysts are working on this task, since we don't know any telecom service provider in the world, which will voluntarily advice their customers to change their tariff plan to the more suitable *for the customer* one
- We assume that the research could be performed for internal use within the telecom service provider only to understand which tariff plans are most profitable *for provider* and maybe to use this information in ads and other marketing activities
- But yes, dear Anton, the junior data analyst had to do something :-)


Thank you for your attention!