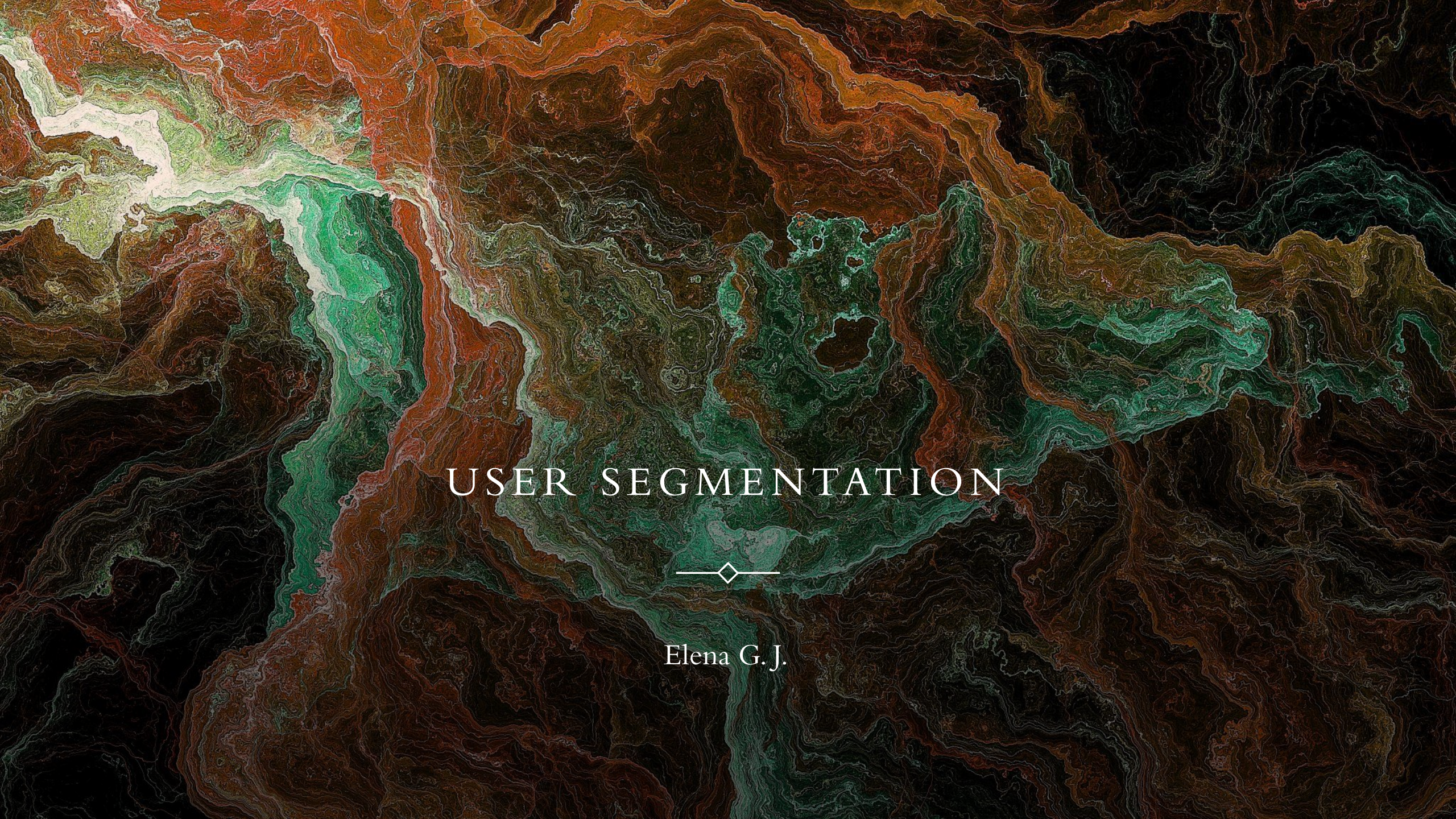

USER SEGMENTATION
Elena G. J.

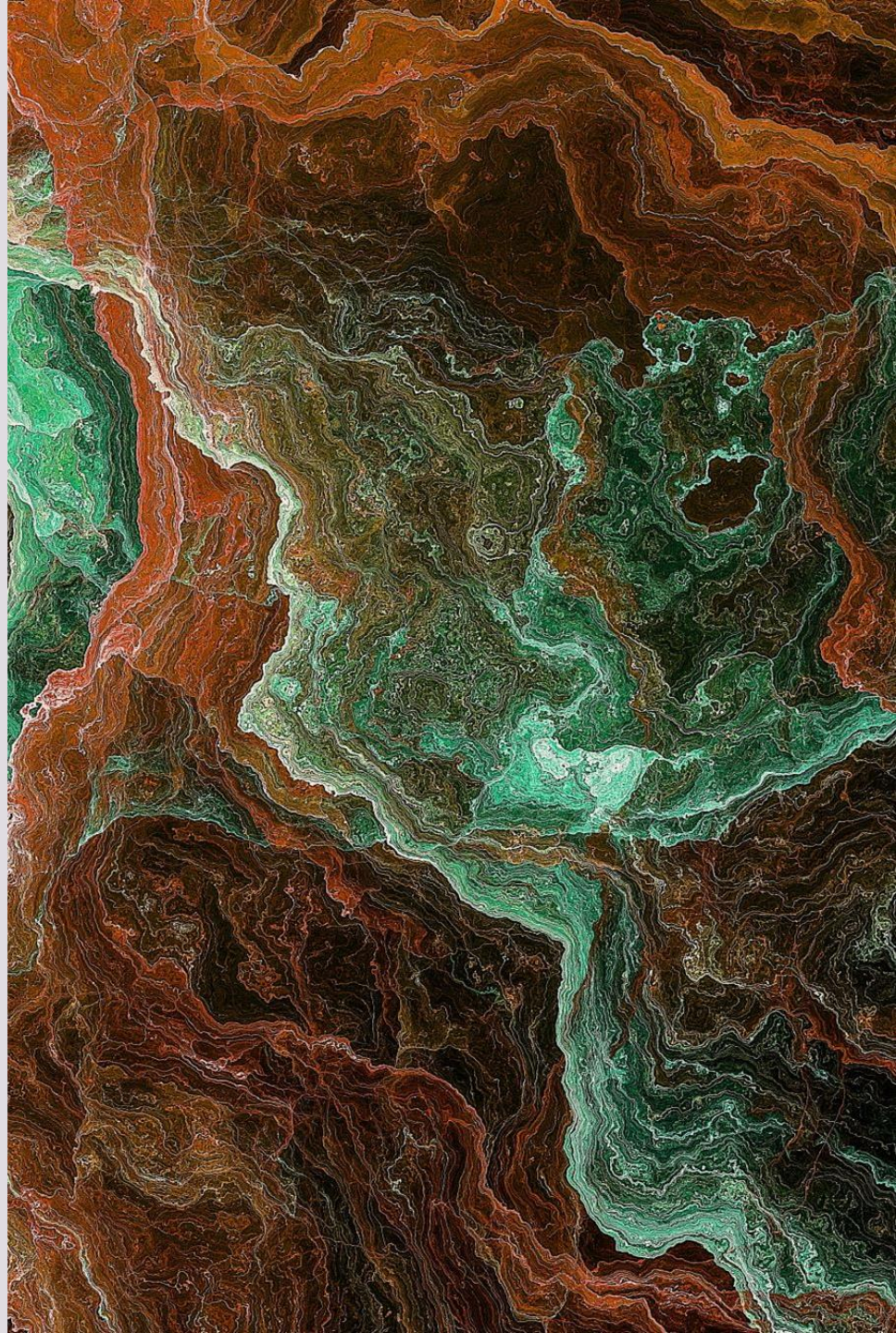
Data exploration and cleaning I

1. Delete columns that have null values
    - All the columns within the dataset seem to have values, even though some of them can be null sometimes. That is why we kept all columns for future analysis
2. Transform date columns (in string format) to date values
    - Columns transformed: registration_date, first_purchase_day and last_purchase_day
3. Delete columns that have redundant information
    - Based on the statistics, name of the variables and in the values that they can adopt (examined thoroughly in the section "Check for congruent information across the different columns"), we can conclude that there aren't variables that contain redundant information.

Data exploration and cleaning II

4. Check for congruent information across the different columns

- Total purchases is the sum of purchases delivered and purchases from takeaway
- Total purchases match the purchases by store type
- First purchase day is always lower or equal to last purchase date
- Total purchases do not match the breakfast, lunch, evening, dinner and late night purchases in all cases (211 users out of 21983 users)
- Preferred devices available in 99.67% of cases
- Total purchases information is congruent with the purchases by device type



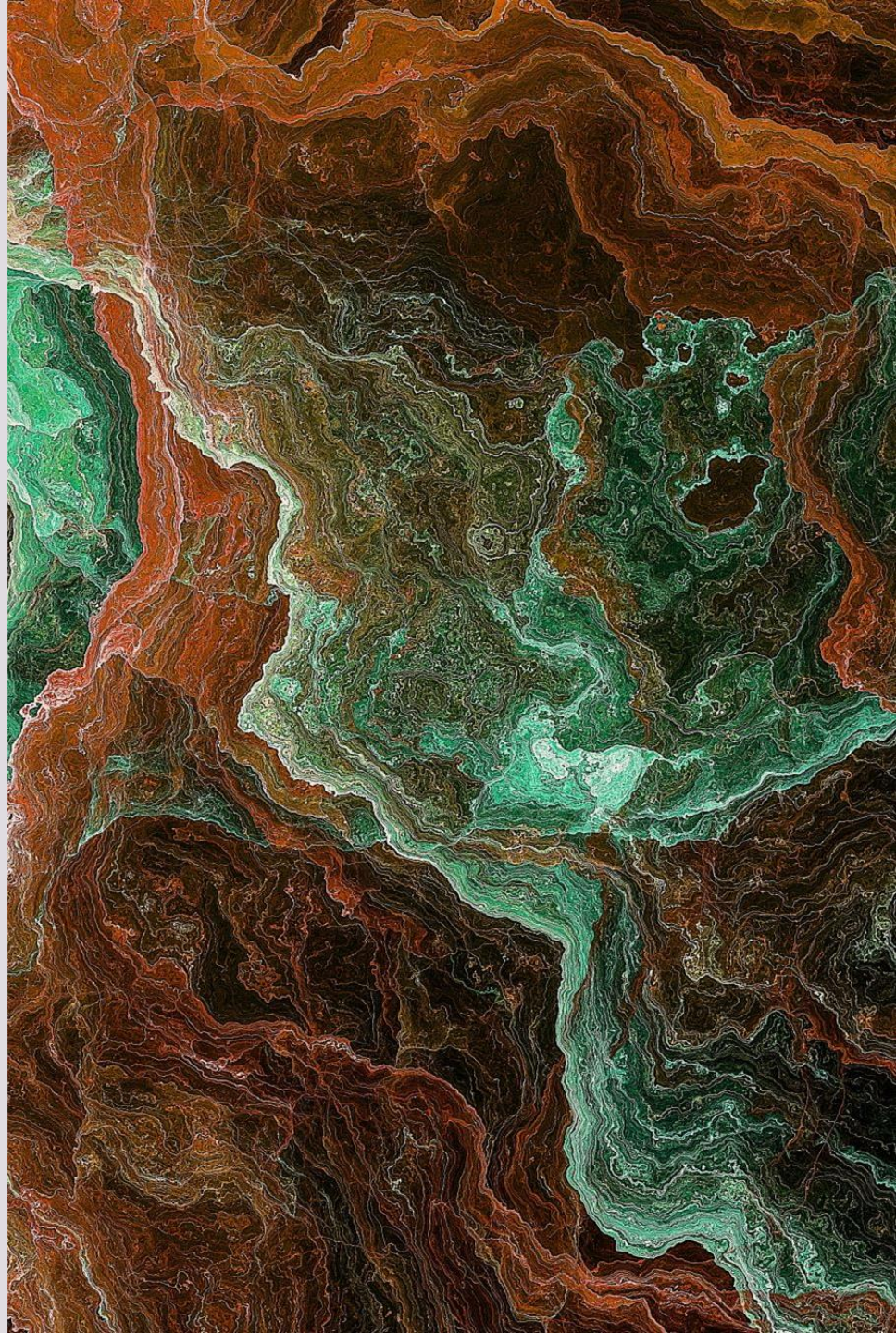
Data exploration and cleaning III

4. Check for congruent information across the different columns

- Preferred restaurant types: only 2694 users have completed this information out of 21983 users (i.e. 12.25% of total users). Also, a total of 2666 users have both fulfilled this information and made a purchase, representing the 22.16% of users who made a purchase
- Users who have at least made a purchase have also available information about most common hour of day and weekday of purchase
- Average and median days between purchases: this information is missing in the 34.88% of cases when users have at least made a purchase. Most of those cases have the same value for first and last purchase date, but in 46 cases, the information is missing because the first or last purchase date is missing, or is not calculated
- The information about average distance in km is fulfilled in all cases where at least a purchase has been made

Data exploration and cleaning IV

5. Drop duplicates: there were not duplicates in data



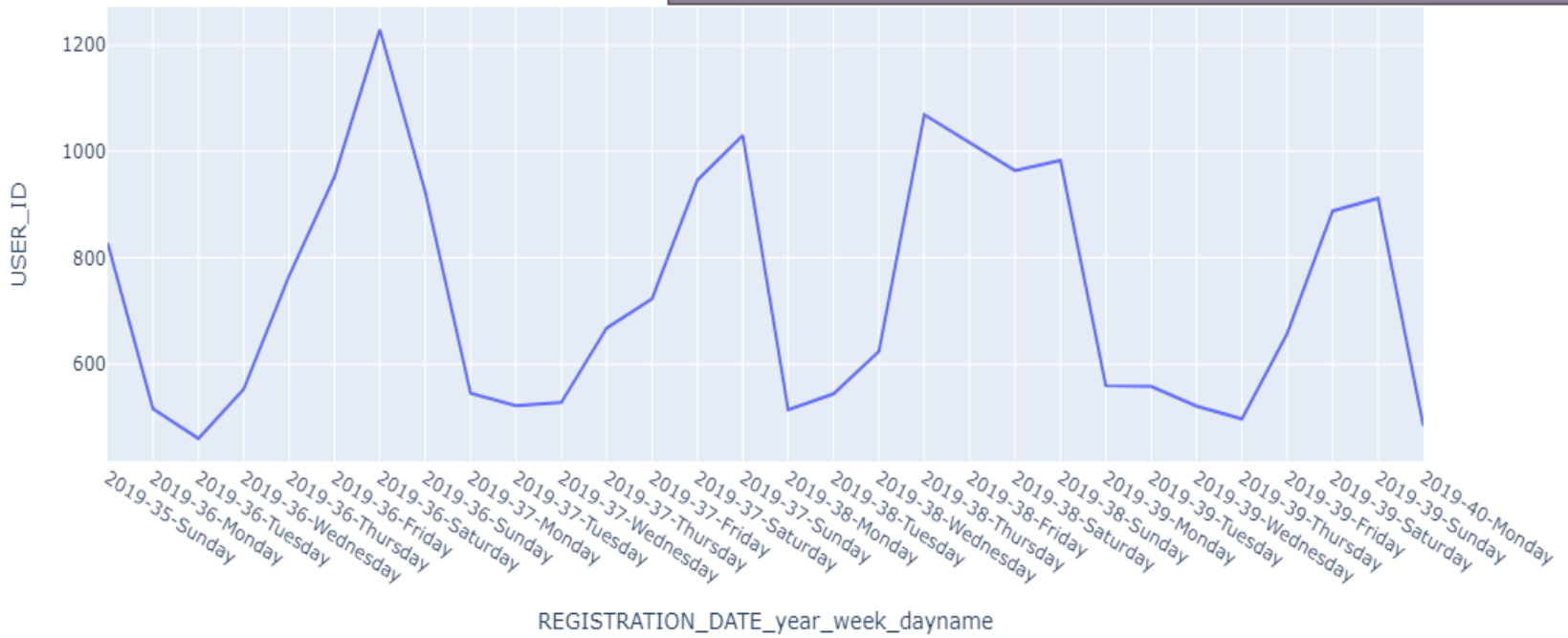
User segmentation based on registration date I

Split based on year, week of year and if it's weekend or not

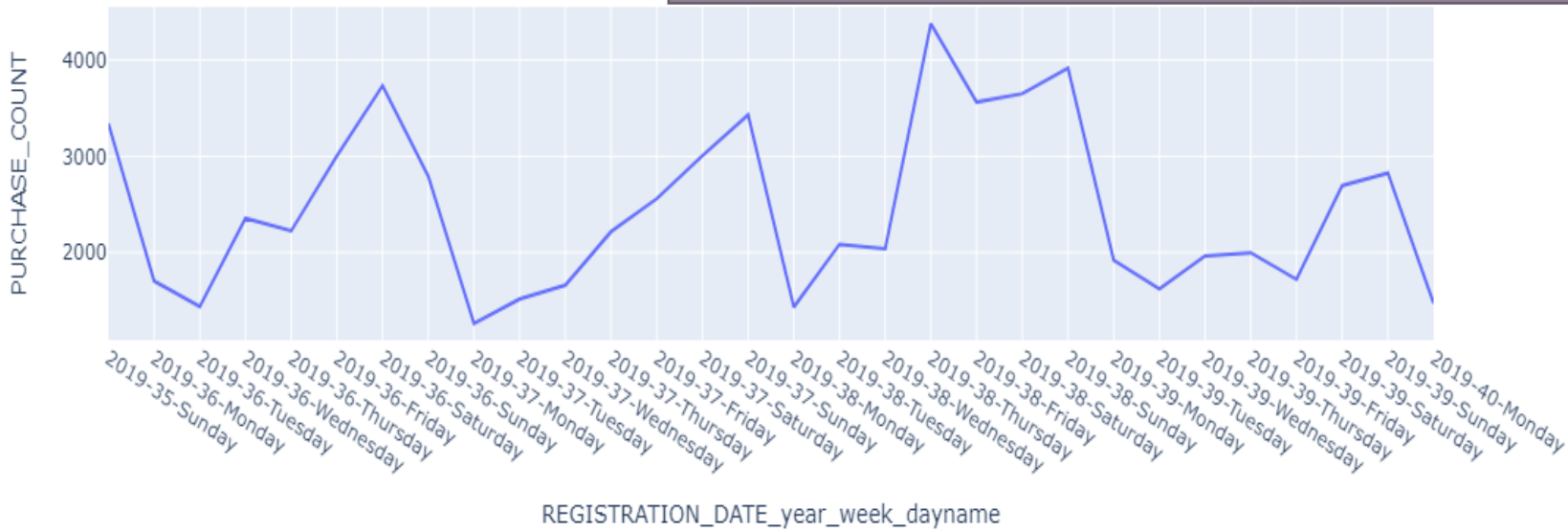
Weekend: Friday, Saturday, Sunday

Not weekend: Monday, Tuesday, Wednesday, Thursday

Number of users registered per registration date, year, week and weekend / not weekend: non statistical significant difference



Number of purchases per registration date, year, week and weekend / not weekend: statistical significant difference



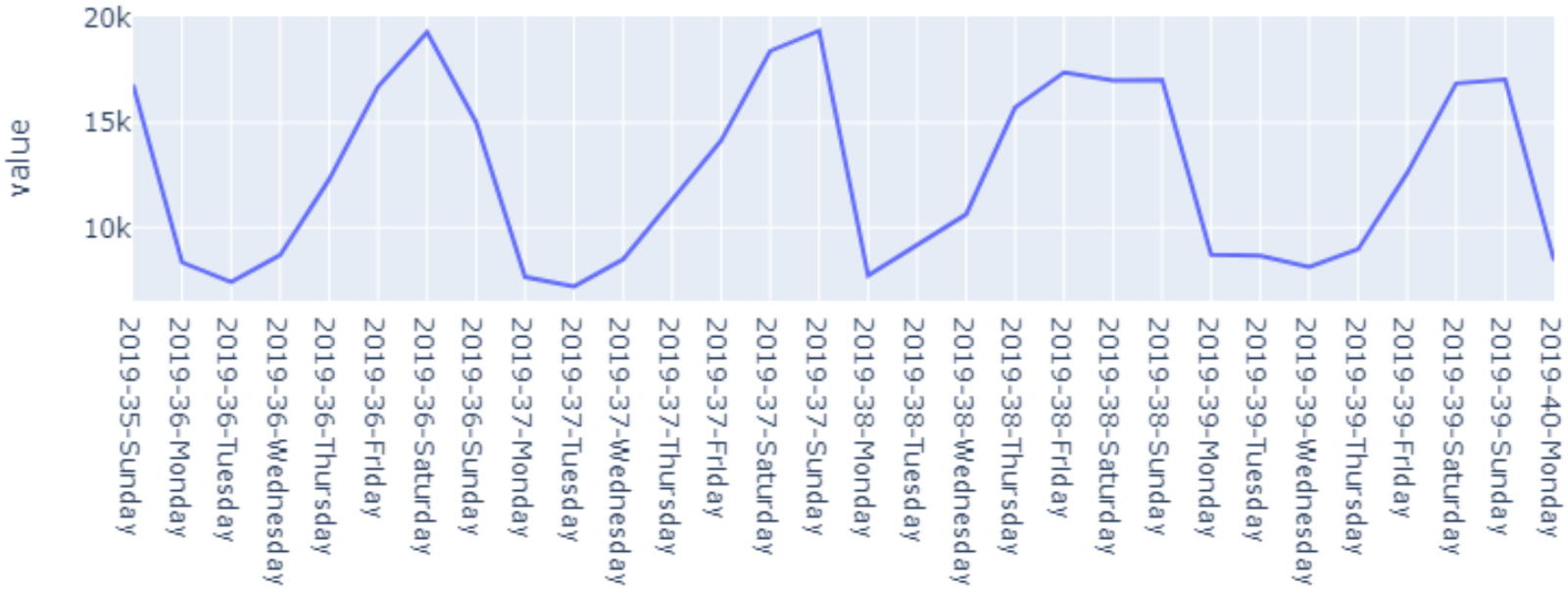
User segmentation based on registration date II

Split based on year, week of year and if it's weekend or not

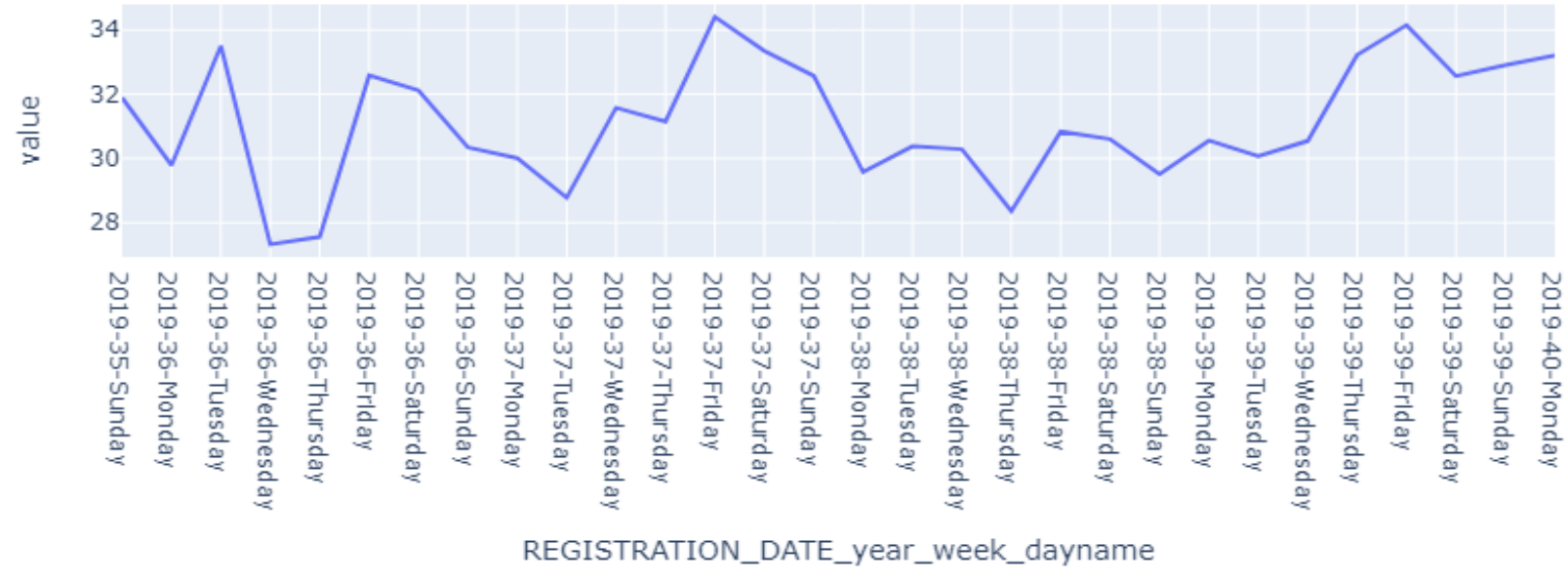
Weekend: Friday, Saturday, Sunday

Not weekend: Monday, Tuesday, Wednesday, Thursday

Sum of average purchases (eur) per user per registration date, year, week and weekend / not weekend: statistical significant difference



Average of average purchases (eur) per user per registration date, year, week and weekend / not weekend: non statistical significant difference

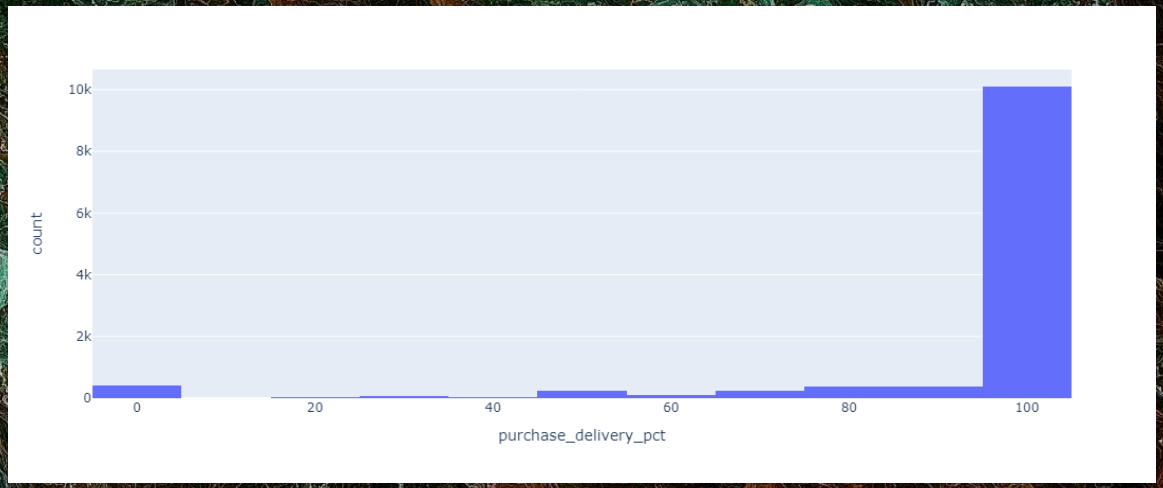


User segmentation based on country

REGISTRATION_COUNTRY	USER_ID
FIN	10277
DNK	8081
GRC	3042
USA	70
GBR	54
SWE	45

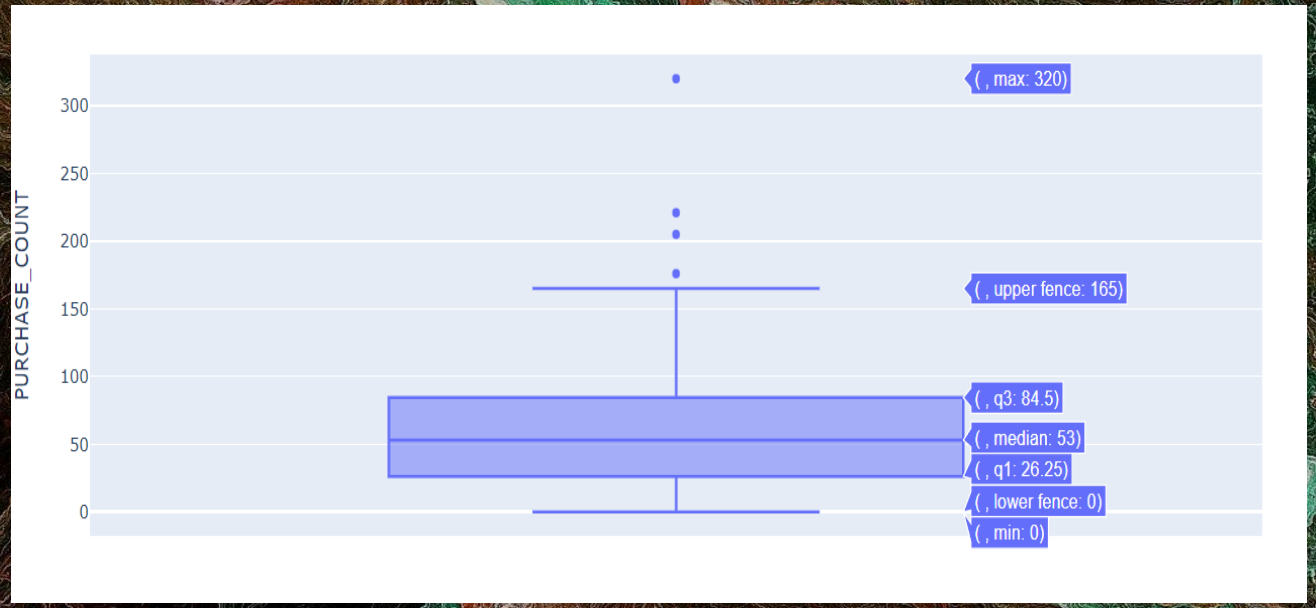
User segmentation based on purchase count delivery and total purchases

$$\%\,delivery = \frac{total\ purchases\ delivery}{total\ purchases} \cdot 100$$



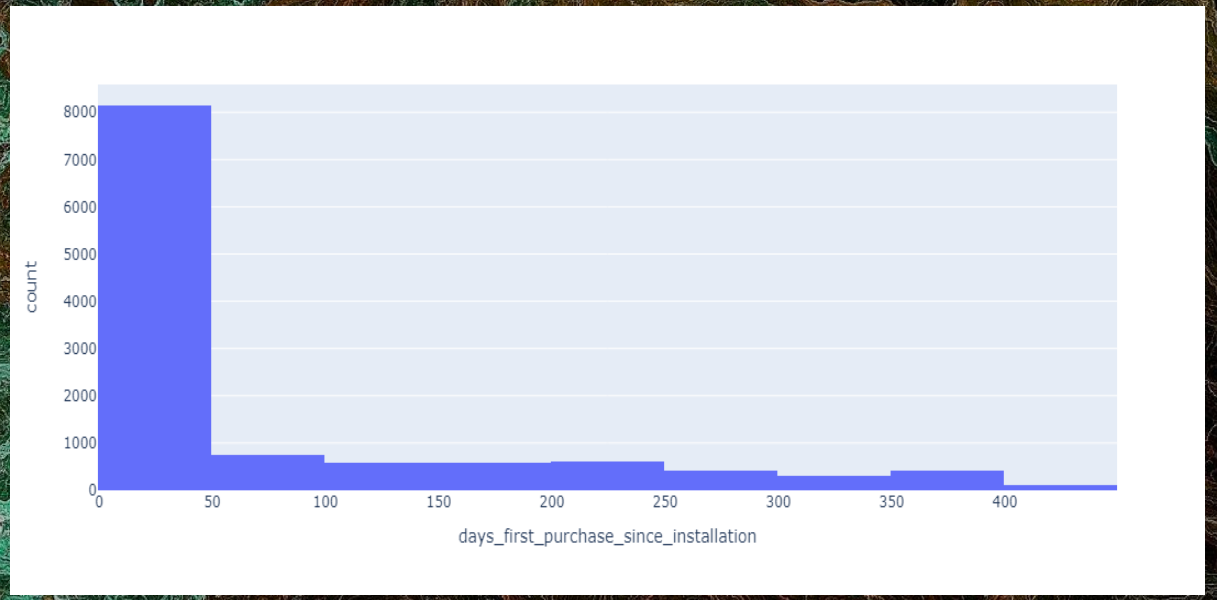
All_purchases_delivery	USER_ID
False	2035
True	9993

User segmentation based on purchase count (percentiles and outliers)



PURCHASE_COUNT_categorical	USER_ID
Below or equal to Q1	21510
Above Q1 up to median	379
Above median up to Q3	63
Above Q3 up to upper fence	27
Outliers	4

User segmentation based on days between first purchase and installation date (percentiles and outliers)



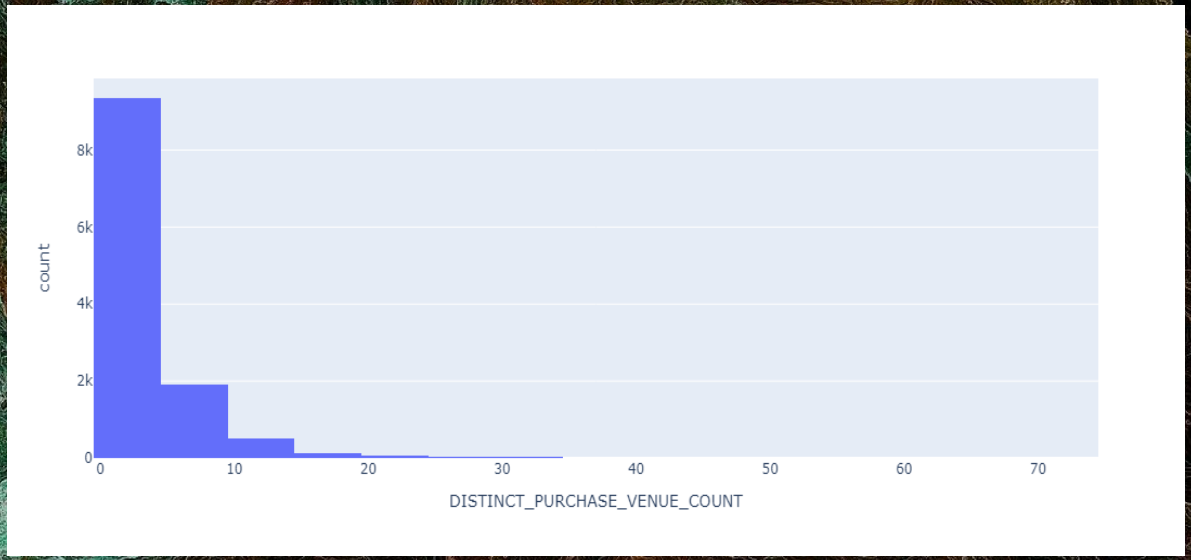
days_first_purchase_since_installation_categorical	USER_ID
First purchase less than 50 days after install...	8158
First purchase 50 days or later after installa...	3783

User segmentation based on breakfast, lunch, evening, dinner and late night purchases

Created the column "Preferred_order_type", to sort in descendent order (by purchases per type, in percentages) the different types of purchases.

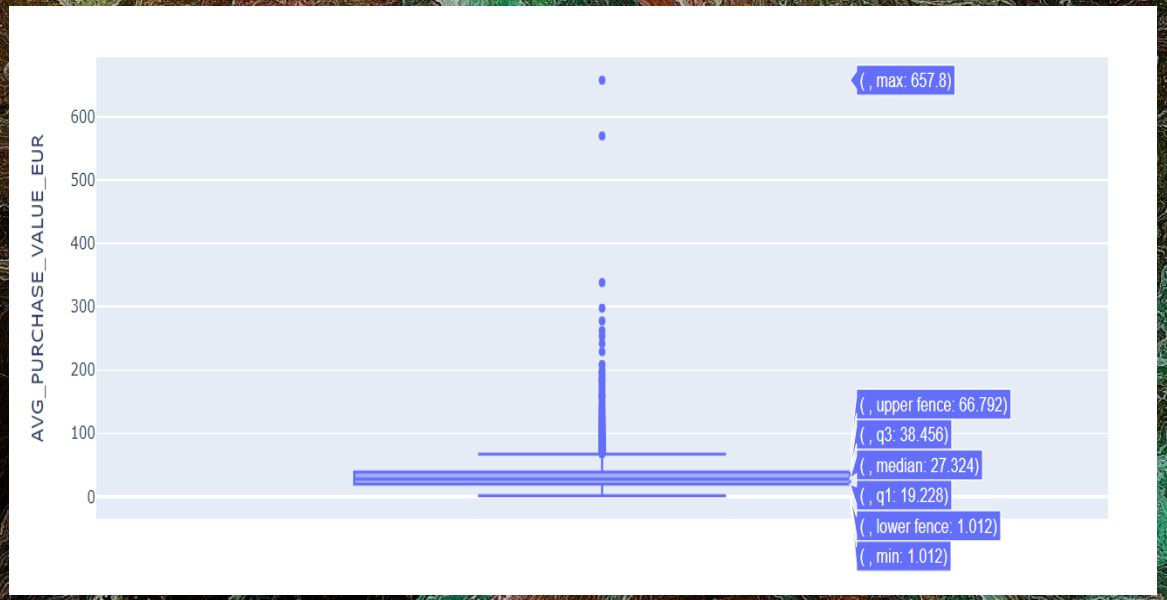
Preferred_order_type	USER_ID
{"DINNER_PURCHASES": 100.0}	3553
{"LUNCH_PURCHASES": 100.0}	2065
{"LUNCH_PURCHASES": 50.0, "DINNER_PURCHASES": 50.0}	740
{"EVENING_PURCHASES": 100.0}	454
{"DINNER_PURCHASES": 66.66666666666667, "LUNCH_PURCHASES": 33.333333333333336}	381
...	...

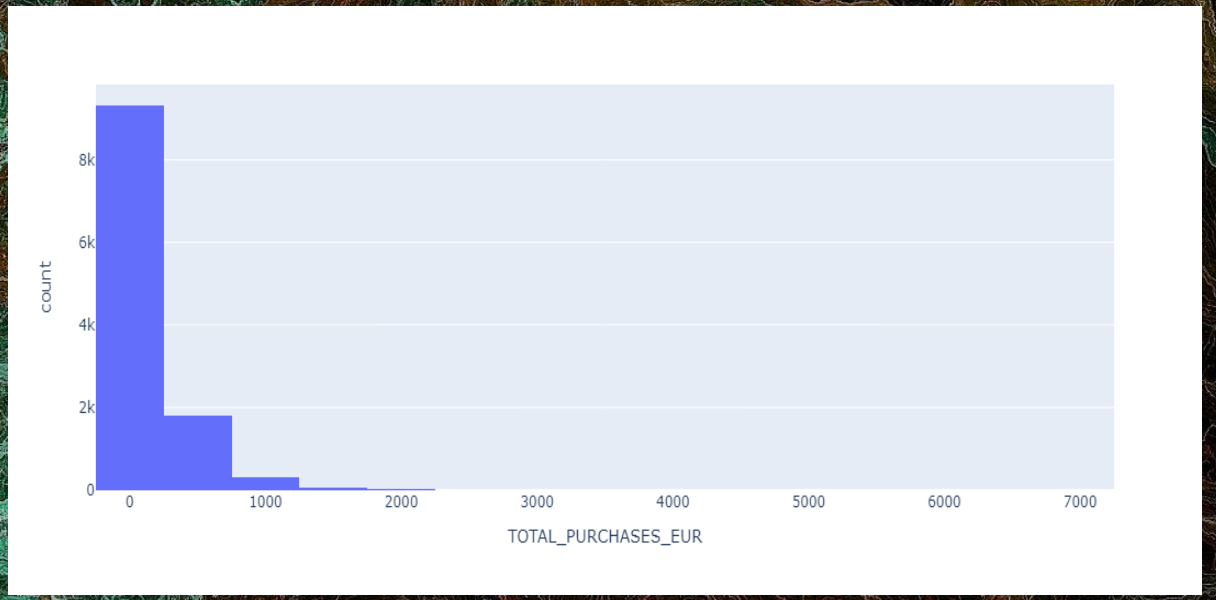
User segmentation based on distinct purchase venues



DISTINCT_PURCHASE_VENUE	USER_ID
10 places or more	740
Between 1 and 4 places	9369
Between 5 and 9 places	1919

User segmentation based on average purchase quantity (EUR): percentiles and outliers





AVG_PURCHASE_VALUE_EUR_qualitative	USER_ID
Above Q1 up to median	3022
Above Q3 up to upper fence	2424
Above median up to Q3	3054
Below or equal to Q1	3087
Outliers	441

User segmentation based on preferred device, most used device in purchases and congruence between them

PREFERRED_DEVICE	USER_ID
android	8448
ios	9747
web	3715

most_used_device	USER_ID
android	4086
ios	5678
web	2264

most_used_device_same_preferred_device	USER_ID
False	1314
True	10714

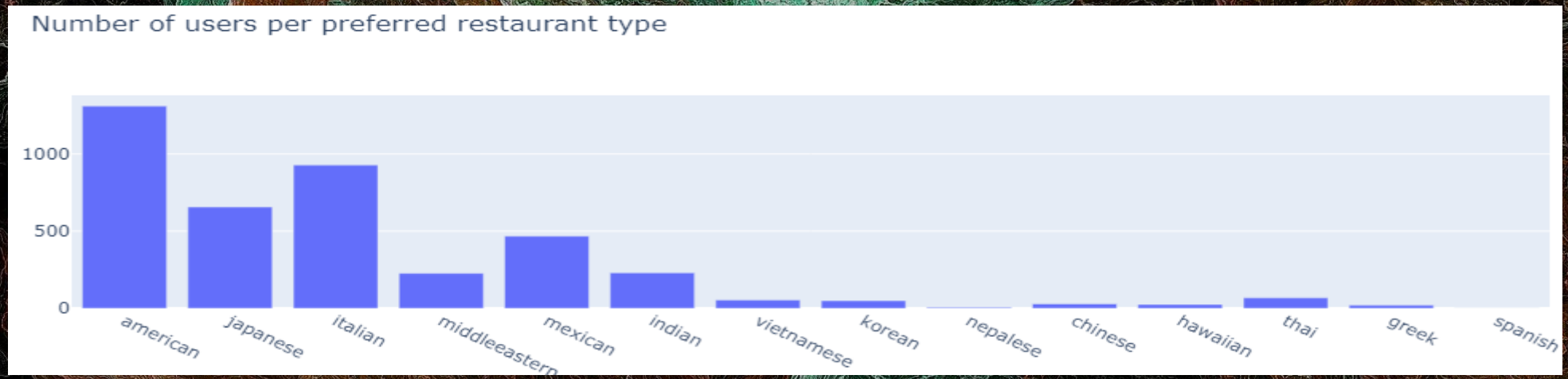
User segmentation based on number of preferred restaurant types

Preferred_distinct_restaurant_types	USER_ID
1.0	1797
2.0	557
3.0	234
4.0	77
5.0	23
6.0	5
7.0	1

User segmentation based on preferred restaurant types

Created one column per option of restaurant types.

index	PREFERRED_RESTAURANT_TYPES (USER_ID count)
[american]	658
[japanese]	367
[italian]	345
[mexican]	175
[american, italian]	131
...	...

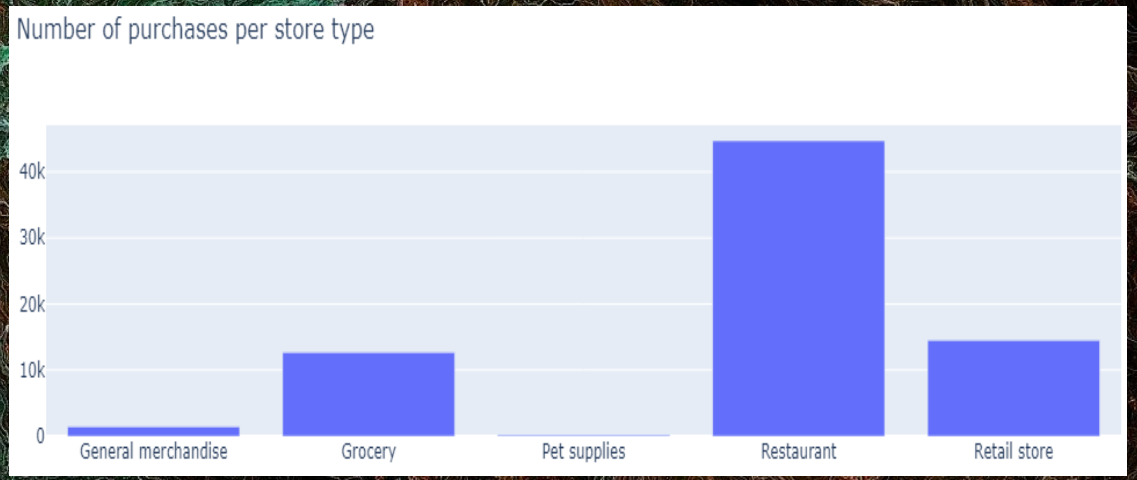


User segmentation based on valid payment method and purchases

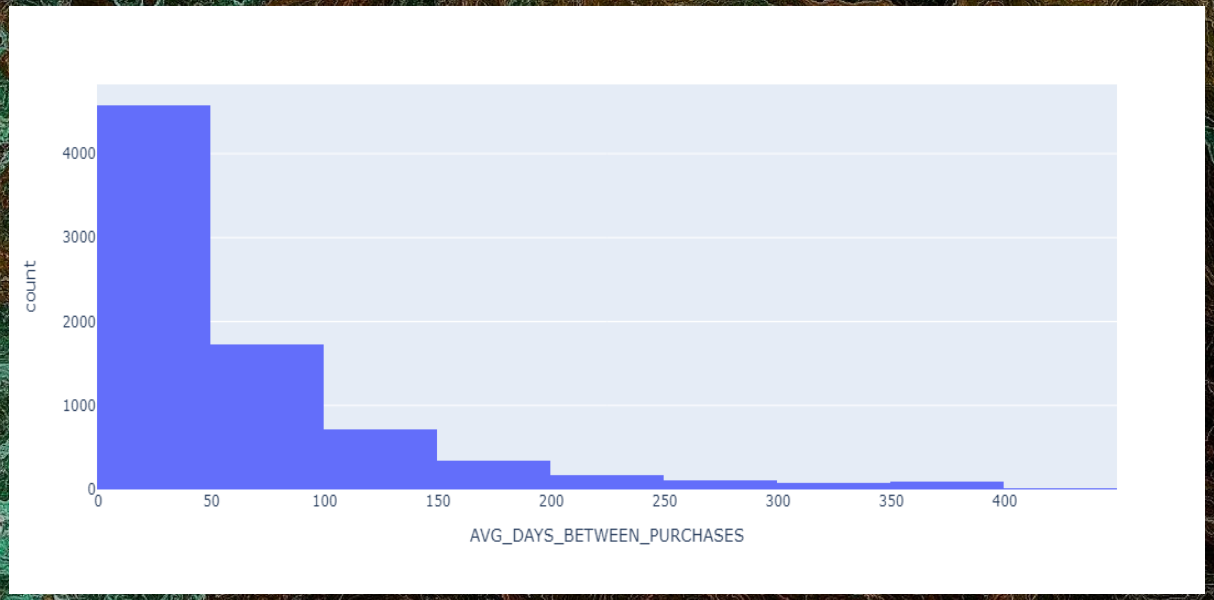
valid_payment_purchases	USER_ID
User has bought at least once and valid payment method	7117
User has bought at least once but not valid payment method	4911
User never bought and not valid payment method	9504
User never bought and valid payment method	451

User segmentation based on distinct store type in which users have bought

distinct_store_types_bought	USER_ID
0	9955
1	8554
2	1857
3	1617

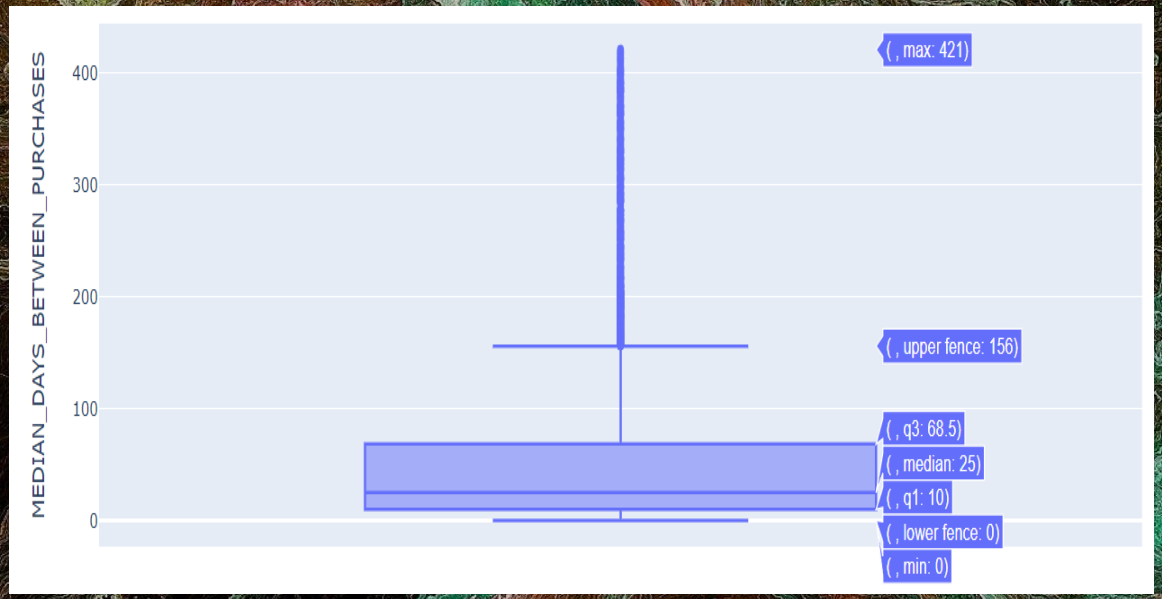


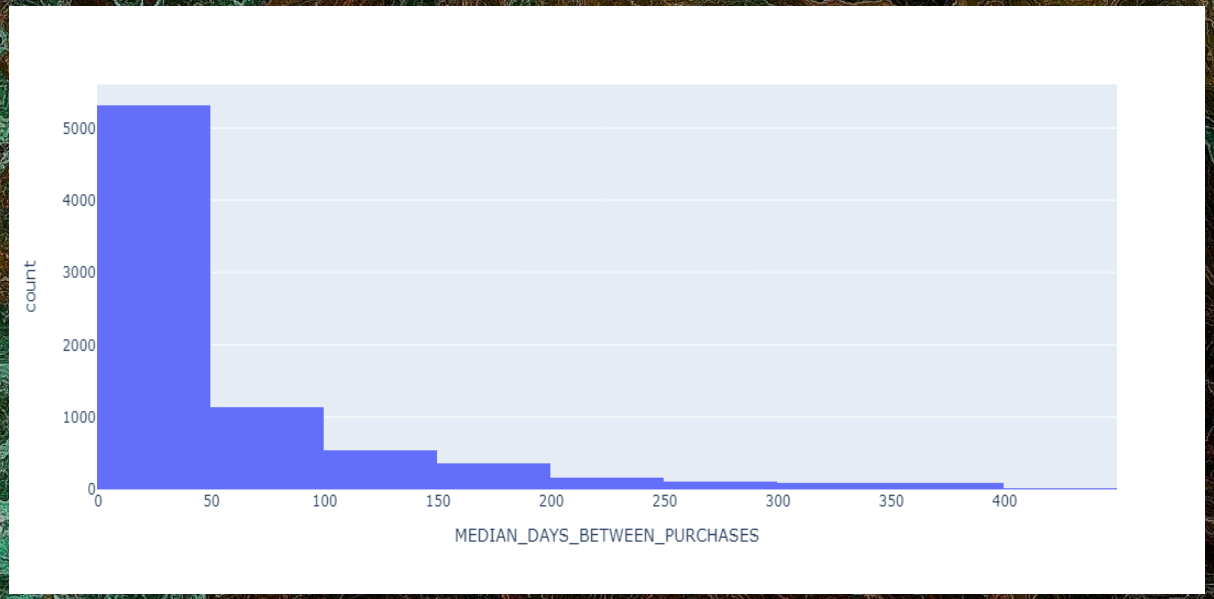
User segmentation based on average days between purchases (percentiles and outliers)



AVG_DAYS_BETWEEN_PURCHASES_qualitative	USER_ID
Above Q1 up to median	1915
Above Q3 up to upper fence	1305
Above median up to Q3	1966
Below or equal to Q1	2001
Outliers	645

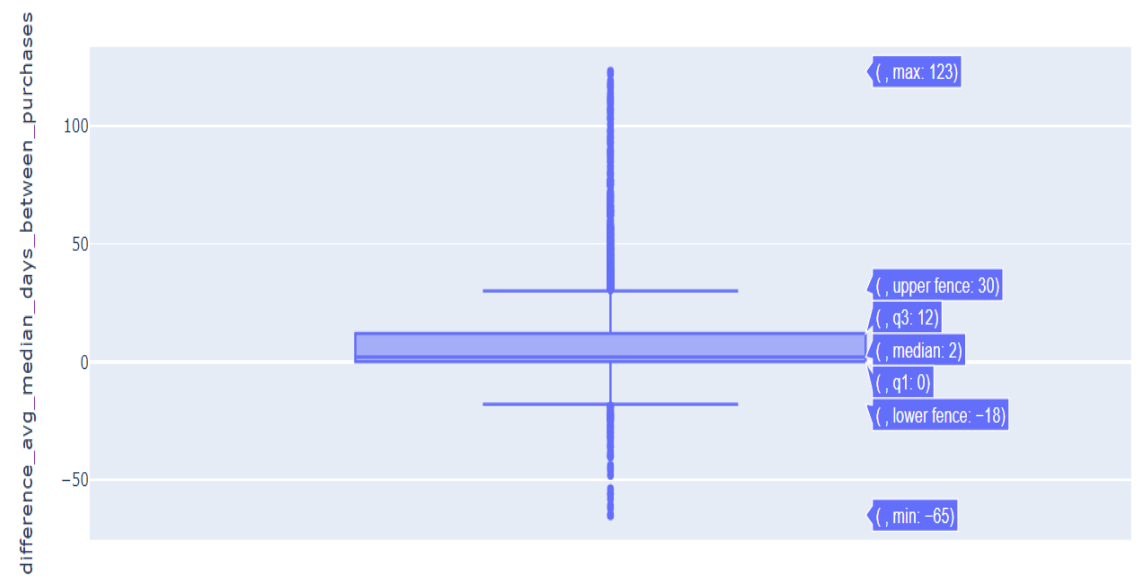
User segmentation based on median days between purchases (percentiles and outliers)

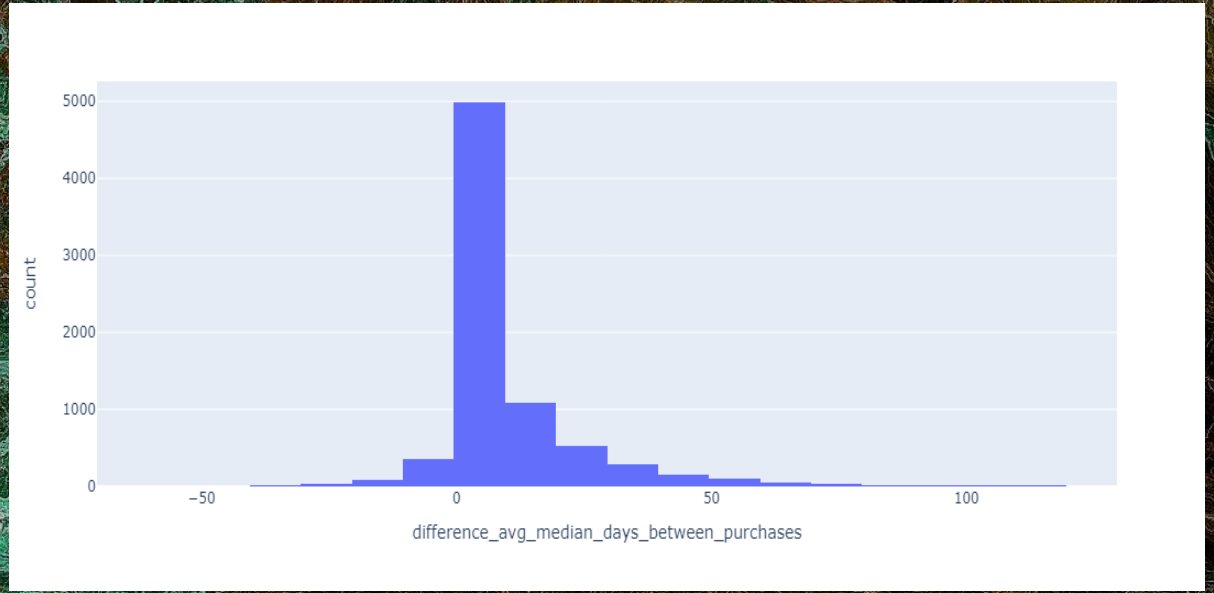




MEDIAN_DAYS_BETWEEN_PURCHASES_qualitative	USER_ID
Above Q1 up to median	1920
Above Q3 up to upper fence	1305
Above median up to Q3	1921
Below or equal to Q1	2033
Outliers	653

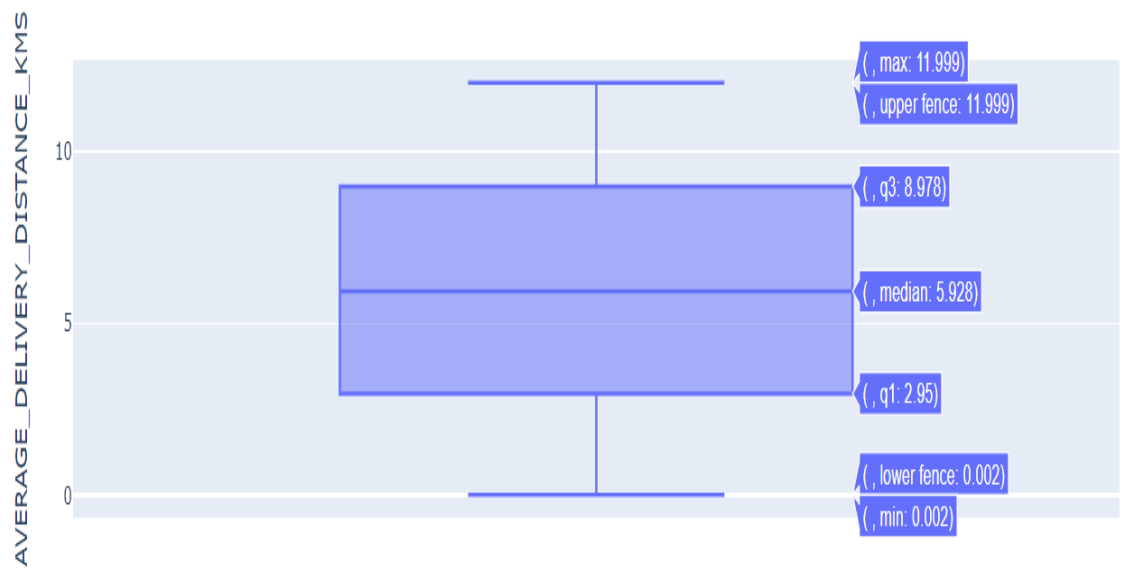
User segmentation based on difference between mean and median days between purchases (percentiles and outliers)

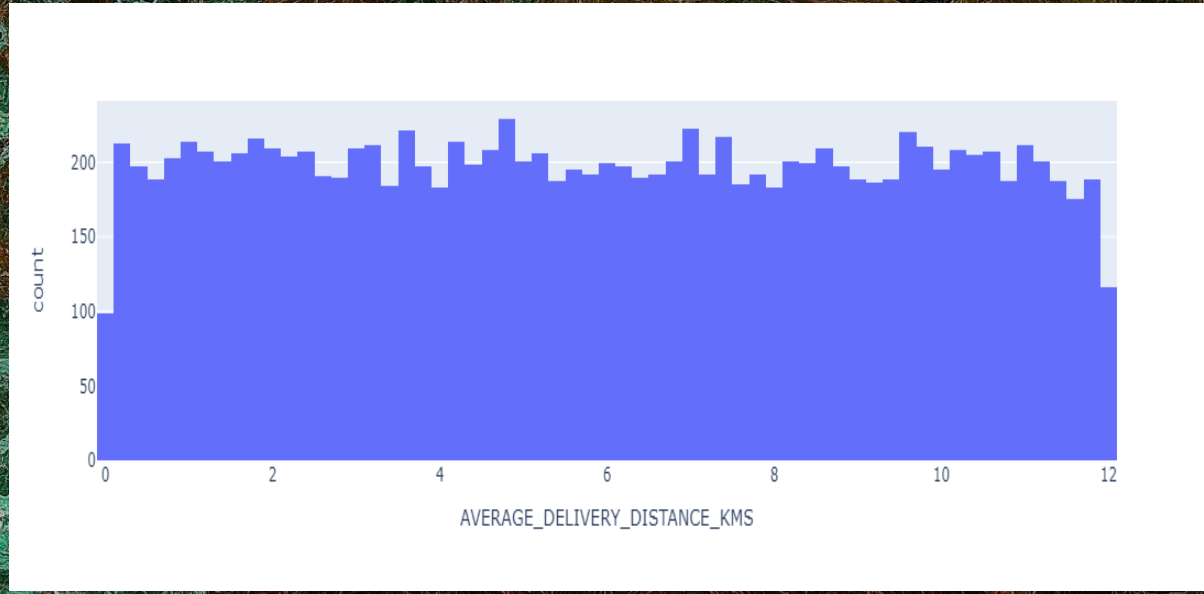




qualitative_difference_avg_median_days_between_purchases	USER_ID
Below Q1	435
Between Q1 and below median	3325
Between Q2 and below Q3	1973
Negative outlier	73
Positive outlier	2026

User segmentation based on average deliverance distance in km (percentiles and outliers)





AVERAGE_DELIVERY_DISTANCE_KMS_qualitative	USER_ID
Between Q1 and below median	3007
Between Q3 and max value	3008
Between median and below Q3	3006
Between min value and below Q1	3007