Sample Data: Users session activity for May and Jun' 2022 is provided.

Problem Statement:

- Report on the data provided, illustrated with visualisations of data.
- Study on clickout ratio.
- KPI definition and chart or table showing the development over time along with descriptions and critical analysis of what you see. Are there any recommendations you can give the product teams?

The information regarding the provided data is as follows.
900000 observations with detailed information about user sessions that were seen on the website make up the data. The columns that are included in the data are listed below.
The xlsx format contains the raw data. Before continuing, I transformed the data into a standard table format.

Column Name	**Description**
Ymd	Date of session in YYYYMMDD format
Session_id	Unique identifier for the visit
Tracking_id	Unique identifier of the user's cookie
Platform	Country code of the platform of the session
Is_app	Flag sessions coming from mobile apps
Is_repeater	Flag sessions that are from a cookie that has visited before
Traffic_type	Coded categorization of session marketing channel
Country_name	Name of the country where the user IP is located
Agent_id	Coded categorization of session device type
Clickouts	Number of clicks to partner sites made by the session
Bookings	Number of bookings made by the session
Session_duration	Length of the session in seconds
Entry_page	First action of the user on our site
Total_ctp	Total number of content items viewed by the session
Arrival_day	Arrival date from search dates
Departure_day	Departure date from search dates

Before working on it, the data set needs to be cleaned up because it is a little disorganized and inconsistent. The following are some preliminary findings on the data:
I have carried out the descriptive statistics separately because the data is both numerical and categorical.

Descriptive statistics for categorical variables:

index	count	unique	top	freq
ymd	900000	61	20220501	18332
session_id	900000	900000	2022062620046057322	1
tracking_id	900000	892416	L47NTB4H8A	9
platform	900000	55	US	117589
is_app	900000	1	0	900000
is_repeater	900000	2	0	537261
traffic_type	900000	5	6	298502
country_name	899576	252	United States	106830
agent_id	900000	19	20	354645
entry_page	900000	980	2116	315613
arrival_day	900000	552	nan	547389
departure_day	900000	458	nan	549817

- Date: The user's session data activity spans from May 1, 2022, to June 30, 2022, with the majority of sessions (18332 occurrences) being recorded on May 1, 2022.
- Session Id: Each of the 900000 entries represents a distinct user session.
- Tracking ID: 7583 repeated IDs out of 900000 tracking records suggest that the same user may have interacted with a different kind of device or that there may be a mechanism that gathers all user activity under a single tracking ID. L47NTB4H8A Id is the most frequently repeated user, nine times.
- Platform: The platform US has been mostly used through which the sessions were logged in.
- Is_app: Value 0 in this column indicate users have not logged into the session by this device type at all.
- Is repater: Assuming 0 being non repeated and 1 being repeated. There are 537261 visitors who have not returned to the site.
- Traffic_type: Five distinct traffic types are denoted by numbers; the most common traffic type is number 6, which occurs 298502.
- Country: The USA is the most frequently occurring of 252 distinct nations where the user's IP location has been traced. The 252 distinct records indicate that there are inconsistencies in the data or that some are being null represented by a different character, which will be discovered later. There are 195 countries in the globe.
- Agent Id: Of the 19 distinct agents, agent ID 20 has appeared the most frequently, indicating that this gadget type has been utilized the most.
- Arrival, Departure Day: The days of arrival and departure have the highest percentage of missing values; this could be because either reservation was made (where the booking >0) or travel was not completed.

Numerical Data's statistics:

index	count	mean	std	min	25%	50%	75%	max
clickouts	900000.0	0.9066166666666666	2.092823445788683	0.0	0.0	0.0	1.0	86.0
bookings	900000.0	0.01236888888888889	0.14754408492865528	0.0	0.0	0.0	0.0	18.0
session_duration	900000.0	390.9212533333333	987.9590271562944	0.0	12.0	64.0	285.0	83335.0
total_ctp	900000.0	14.700294444444445	134.90739241777317	0.0	0.0	0.0	1.0	3662.0

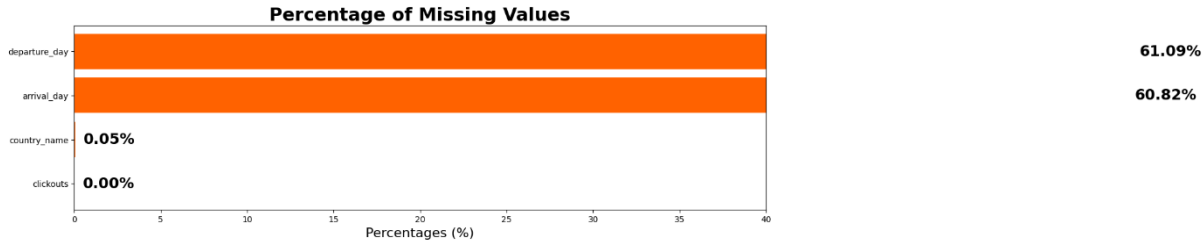
- The majority of the sessions did not result in a booking, as indicated by the average clicks per session of 0.9 and the booking average of 0.012.
- There are 900000 bookings in total, with the minimum and maximum values of 0 and 18 indicating that hardly any bookings are made during a session. Because the average value is around zero as well.
- session duration: The data is somewhat dispersed from its average, as indicated by the slight increase in the standard deviation number. The average session duration is approximately 391 seconds. Users are more involved in the session when the maximum value of 83355 is reached, and the ratio is lower at 285 seconds. There are outliers, as indicated by the significant discrepancy between the maximum and 75th percentile values.
- Total_cp: The total number of items viewed by the session indicates that, on average, 14 items were viewed. When the maximum and 75th percentile values are compared, the outlier's presence is clearly apparent. The range of connect numbers, from 0 to 3662, is rather wide.

Data anomalies and inconsistent data:

observed data has some anomalies and inconsistency which needs to be clarified and investigated before handling:

- Clickout is inconsistent with few data points being recorded as 'and saba'.
- A few entries in the agent_id column are in string format, such as republic of '*, which is unclear at the time because the majority of the data is in numeric form. I assume this would be misinterpreted, but it can be clarified when working with data engineers on the actual data.
- There are 292 unique countries, which is unusual given that there are 195 countries in the world. As a result, the remaining data may be inconsistent and was removed because the percentage is 0.05, which is extremely little.
- A couple records have zero clickouts but a number of bookings, which is technically strange because how could a booking have occurred if no ads were clicked? This has to be made clear.
- In a few cases, at least one or more bookings and clickouts were made, but the session time seems to be zero. This might be due to a data entry error or the session activity was completed in a very short amount of time—zero seconds—which is why it's necessary to check with the data team.
- In certain instances where bookings were made, the absence of an arrival column and the presence of a departure column, and vice versa, also indicates a missing entry from the logs. This is ideal when booking tickets, as either entry may be blank, but in this case, clarification is required. Since the majority of the data follows a similar pattern, it must be processed and cannot be dropped.
- The datatype is also inconsistent, with some columns containing date values and others containing values in the number of days.
- In some cases where the clickout and booking were made the total_ctp , no of items viewed are 0 , which is unlikely to happen which needs to be clarified. For inconsistent type agents like republic of * the no of convent viewed seems higher than the other agent types.

Assumptions for Handling Null



- If the column contains less than 10% missing data, I have dropped those rows as it will not create much impact being less frequent in the data.
- In many instances, over 60% of the data is missing; therefore, it cannot be eliminated because doing so could skew the results, hence I have replaced it with average value.
- According to the information provided, clickout is major matric, so it is assumed that the arrival and departure days won't provide much value. Hence, the value in both columns will be replaced by the value that occurs the most. The business and data teams can provide clarification on this.

Exploratory Data Analysis:

As I got some overview about the data now let's dig deep into finding the metric that are important to the business.

Top 20 Countries with frequent sessions:

About the chart:

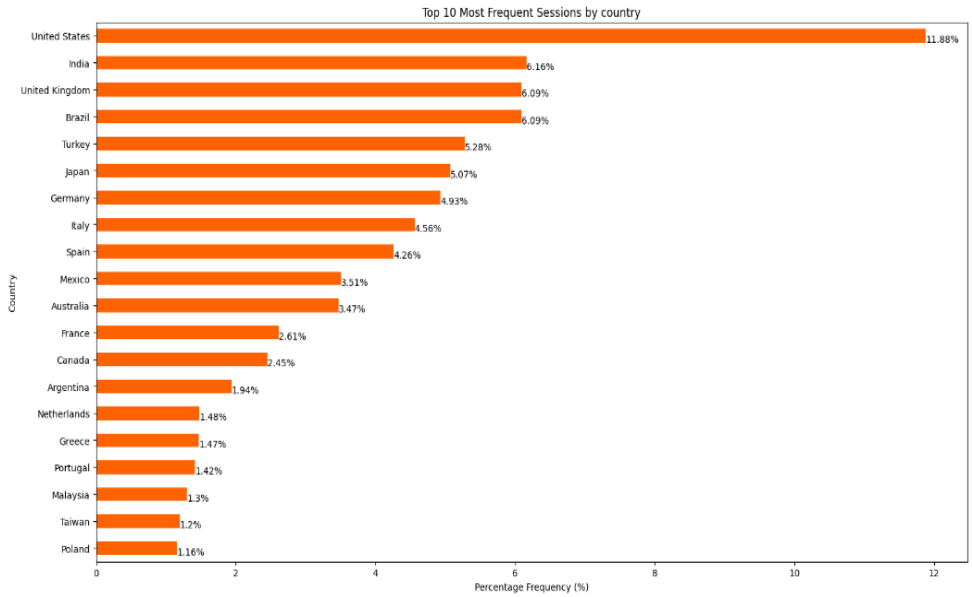
- The following chart shows country wise top session distribution

Findings:

- The following figure indicates that the United States is the most popular country, accounting for 11.82% of all sessions, while Poland and Taiwan are the least popular, accounting for only 1% of all data. Further investigation is necessary to determine the reason.

Recommendation:

- Need to investigate more for countries like Poland and Taiwan.



Top 10 highest booking countries.

About the Chart:

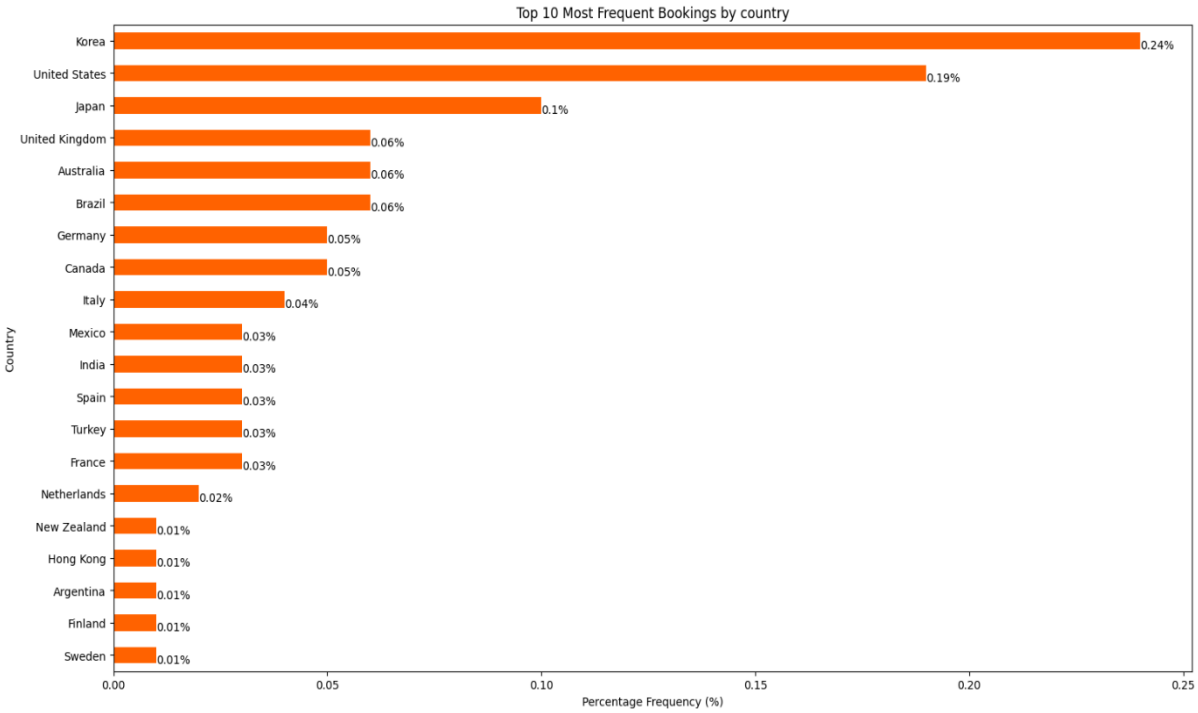
- These are the countries where the most bookings were made.

Findings:

- We can see, the average number of reservations is extremely low. Korea had the highest percentage (0.24%), while Sweden had the lowest (0.01%).
- According to a comparison of the booking chart and session frequency, Korea has the highest booking rate but is not among the top countries with the most frequent sessions, indicating that the number of sessions has no bearing on bookings because a single session may have the most bookings.

Recommendation:

- Countries like Sweden, Finland, New Zealand, Argentina seems to have lower booking rate in this particular season, need to investigate if any factors like weather are responsible for this decay in these months.



Trend Analysis:

Now analysing the Clickout and booking trend for given 2 months.

About the Chart:

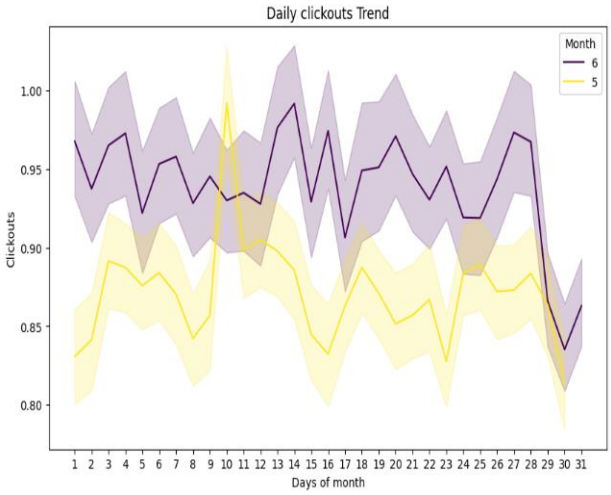
- Following chart illustrate the trend analysis for booking and clickouts.

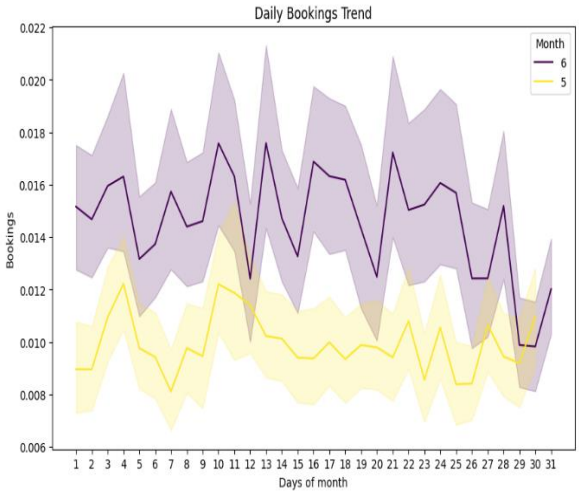
Findings:

- The chart below demonstrates that there were notable swings in the clickout and booking ratios during the month of May.
- It indicates that there was a notable decline in clickouts and bookings during the final week and rise after the first week, which may be related to customers' tendency to book hotels after receiving their pay checks and paying their bills.

Recommendation:

- The experiment needs to be conducted to see if the percentage of bookings and clickouts increases beyond the first week

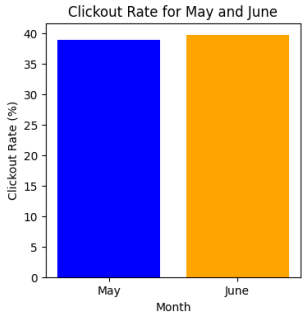




KEY Metrics:

Since the Clickout ratio is the primary revenue-generating KPI, let's examine it along with the other factors that have been presented to see how it affects the company's growth and performance.

1.Clickout Ratio= Total number of clickout/ Total number of sessions *100



- COR for May 2022: 38.81%
- COR for June 2022: 39.74%S
- MOM Change in COR: 2.34%

Bivariate Analysis:

Let’s analyse COR for different parameters to see if getting impacted by these factors which can be addressed later:

1.COR by Agents:

About the Chart:

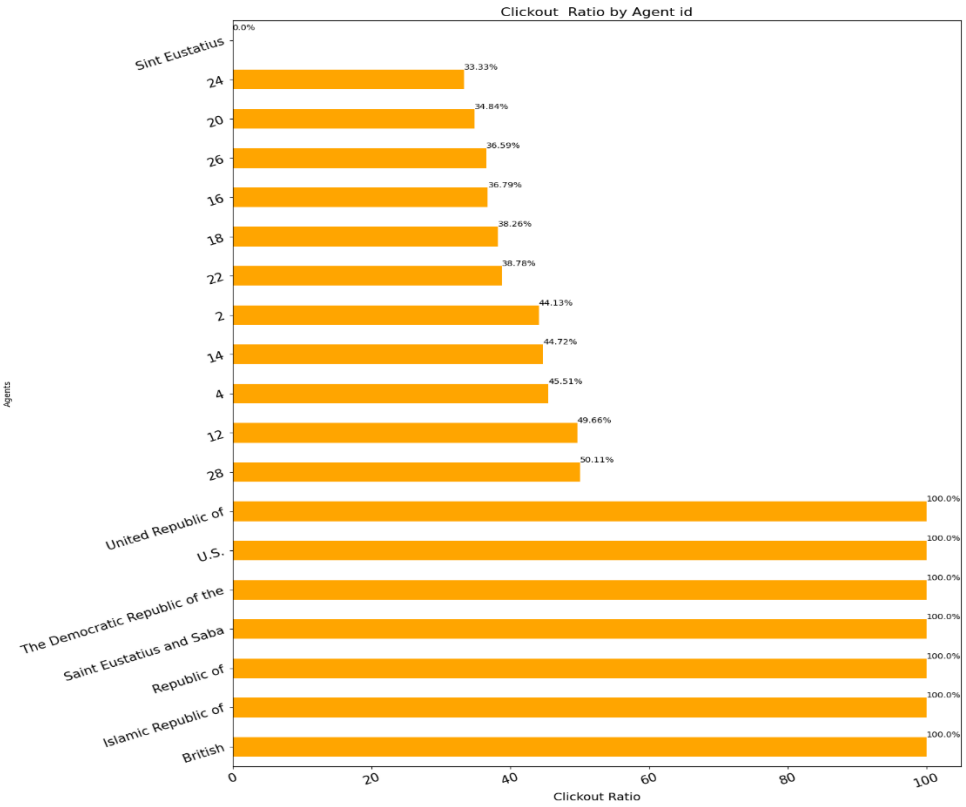
- The chart shows the COR distribution over agents.

Findings:

- As mentioned in the anomalies section, these agents tend to exhibit a 100% clickout rate, which is not the best scenario and is quite rare to occur.
- Agent ID 24 appears to have the lowest clickout rate at 33%, while ID 28 appears to have the highest at 50%.

Recommendation:

- Investigate the strategies to boost up the rate for agents 23,20,26.



2. COR by Traffic Type:

About the Chart:

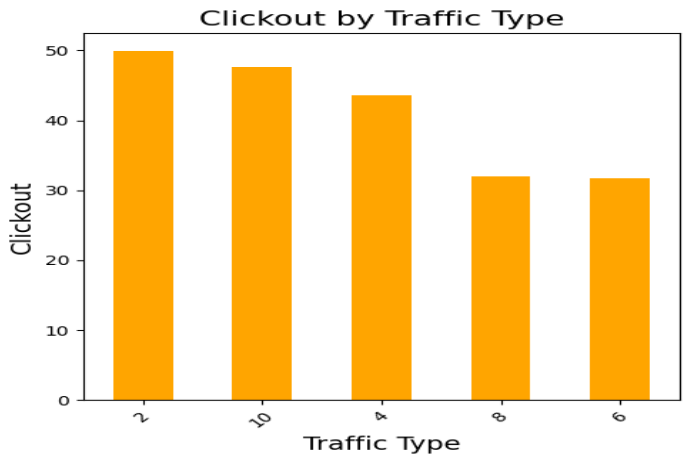
- Shows the distribution of COR by traffic type

Findings:

- Traffic type indicated with code Code 2 traffic types typically have more CORs (almost 50%) than code 6 traffic types (31%).
- To generate more revenue through these traffic type like 2,10 we need to investigate what are these traffic types and the strategy so can serve them better and those which are not need to optimise the process and strategy like 6.

Recommendation:

- Investigate the traffic types 6 and 8 by performing A/B tastings for different user segments.



3. COR by Platform:

About the chart:

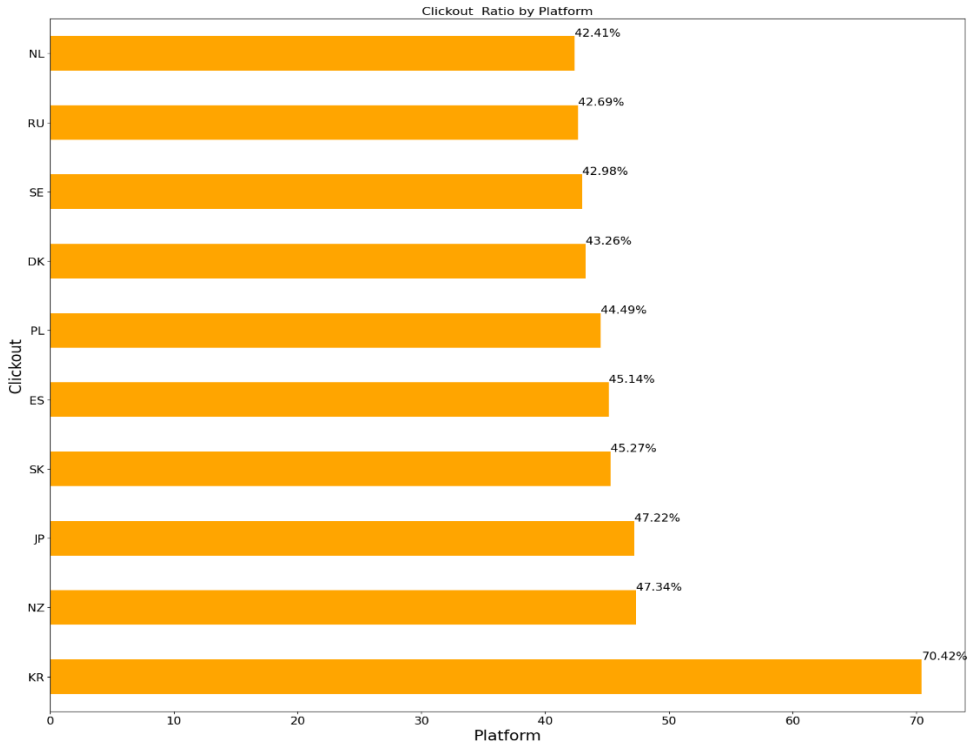
- Shows the distribution of COR by platform.

Findings:

- Platform KR appears to have a higher COR (71%), indicating that its users are very engaged, while ID is 20%. Since the names of these platforms are not entirely apparent, we will make assumptions based on the information provided and work with the data team to define them later.

Recommendation:

- Investigate the Platform with the lower COR rate as ID

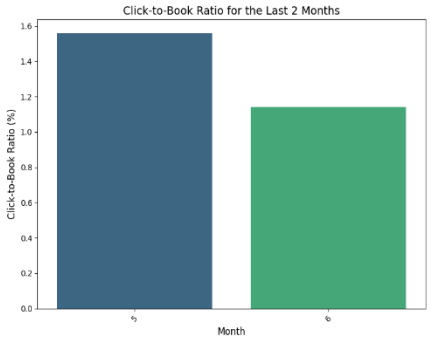


Metric 2: Click To Booking Ratio

- Does this KPI helps in producing revenue? As its mentioned on the 's website and the business model it drives based on Clickout ratio.
- Since the number of bookings actually occurred through the website, it is assumed that this KPI can also be regarded as a success ratio.

Conversion Rate= No of bookings/ No of Clickouts*100

Findings:



- Conversion for May 2022: 1.5%
- Conversion for June 2022: 1.2%
- MOM Change in Conversion: -25.1%

Click per booking ratio by Agents:

About the Chart:

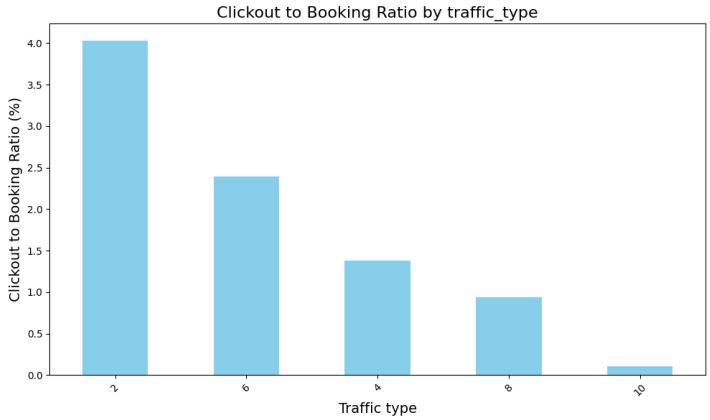
- Show the conversion ratio by different traffic type.

Findings:

- As we saw initially the overall click to booking is low, but with that traffic type 2 contributing more to this conversion rate and the 10 being the low.

Recommendation:

- Traffic type like 10 need to be investigated.



Similarly, this can be seen for different parameters as agents, parameters once confirmed if the KPI is contributing to the value.

Repeated Customers:

Given the information about repeated customers I will analyse it has any impact on the business metrics.

About the Chart:

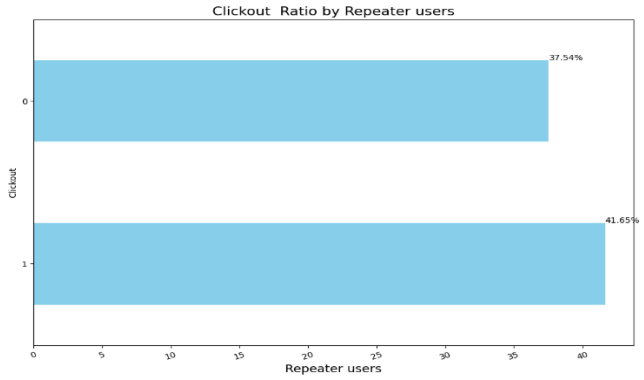
- Analysing now if the repeated customers tend to contribute in the COR and Conversion rate.

Findings:

- Repeated clients have a greater clickout ratio, indicating that they typically make more clickouts but fewer bookings. This might be merely browsing, or if the user's session duration is also low, an incorrect website clicks.

Recommendation:

- Better recommendation engine



Repeated customers pattern:

About the chart:

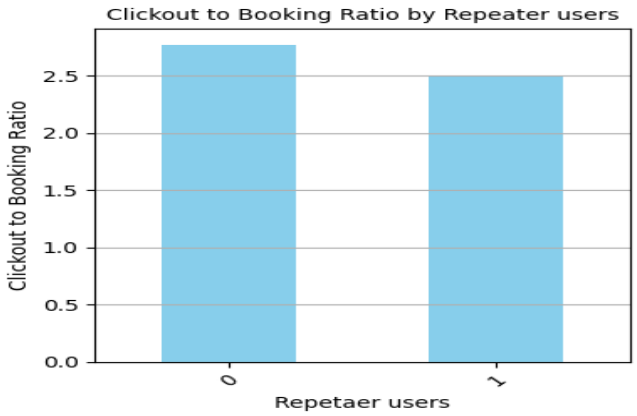
- Chart shows the behaviour of repeated customer on bookings

Findings:

- No of bookings and repeat customers have a negative relationship, indicating that repeat consumers are less likely to make reservations.

Recommendations:

- Needs to show the better recommendation options to repeated customers as it generates COR and overall success rate.



Metric 3: Session Duration

Give in the information about session lengths lets find out if the longer sessions tend to contribute in clickouts and bookings.

About the Chart:

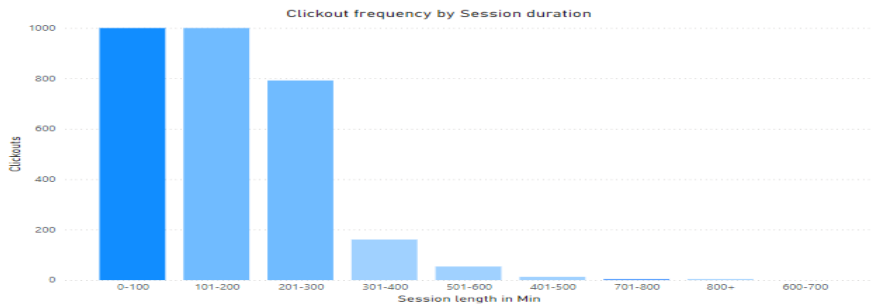
- Shows if the session duration is affecting COR ratio.

Findings:

- According to the chart, clickouts occurred up to the 200-minute session length, and longer sessions lead to decreased clickout rates.

Recommendations:

- Conducting more experiments to test these results as it can be partially true.
- By examining the session length trend and avoiding or focusing on those that might be outliers, the recommendation system can be made more enticing.



Final Recommendation:

- As per the given data it says that no of users tends to spend more time on exploring instead and not booking so need to understand more their expectations and offer better recommendations.
- To generate more revenue more strategies should be applied on traffic types which are contributing low.
- For platforms with lower COR, improving the user interface or analysing the user journey more deeply to identify drop-off points.
- Traffic types contributing low to COR, need to experiment A/B testing for different user segments, or optimizing ad campaigns for traffic sources that drive higher engagement.
- To better serve those with contributing session length patterns, more experiments are being conducted.
- Addressing whether a specific month of the year with more contributions can provide customers with greater opportunities for engagements during those days.
- Understanding the external factors for lowest contributing countries to increase user engagements