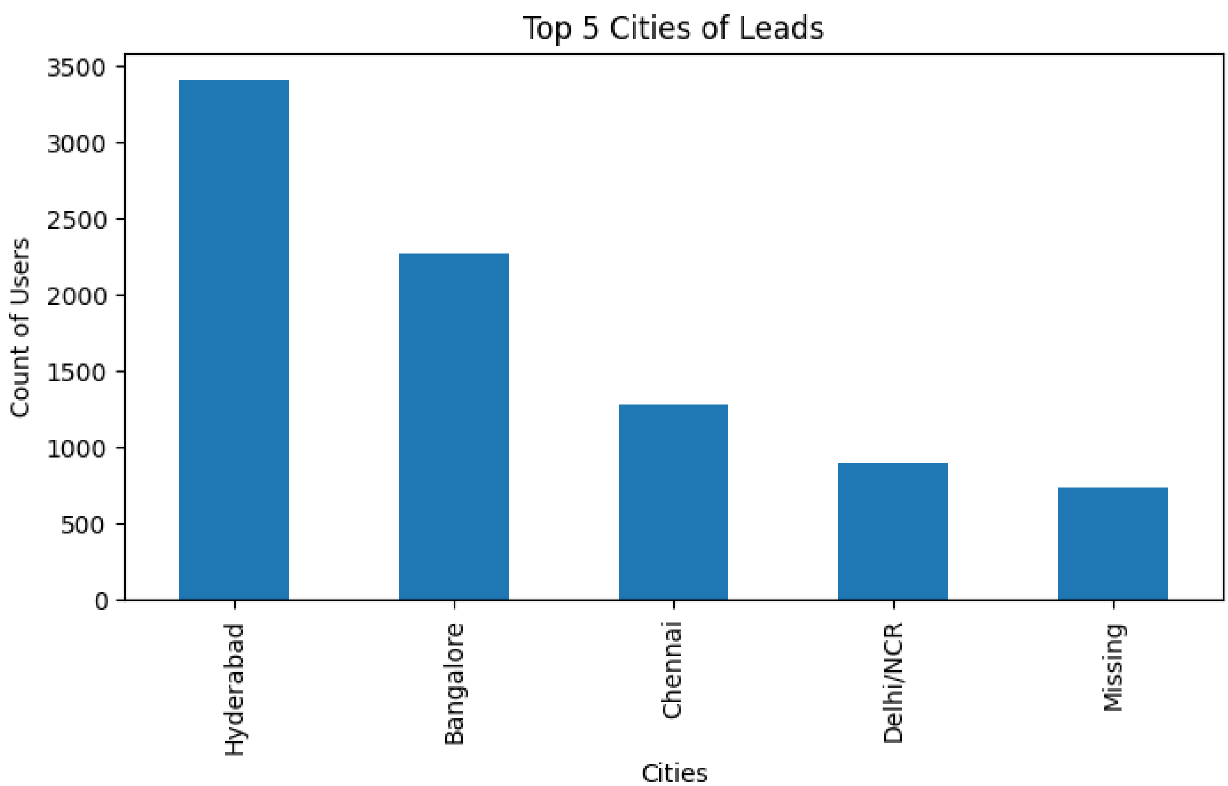
@ Vahan-Analyst/PM:-Take-Home Task

Applicant:- **Raushan Kumar Prakash**

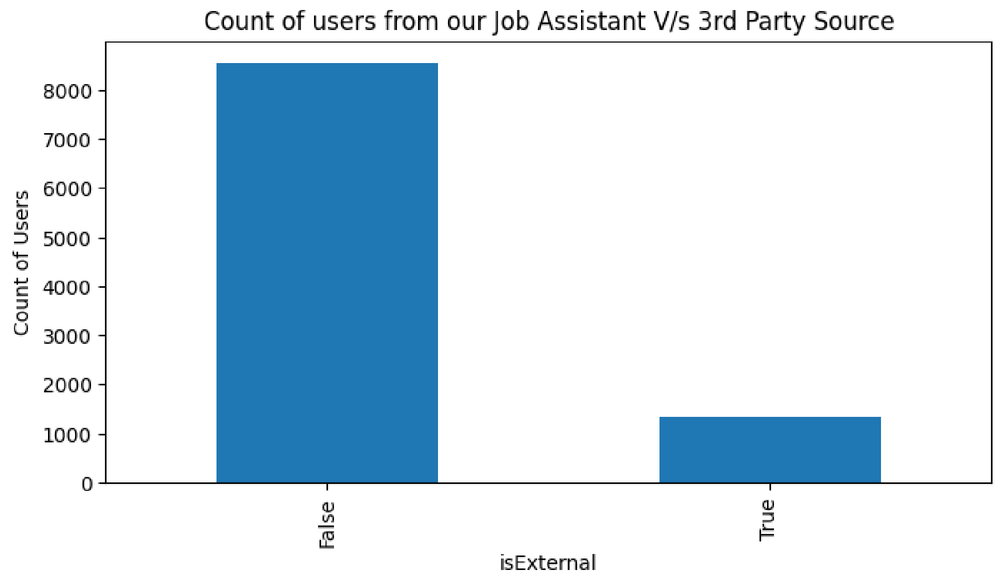
1. Insights from the Leads dataset:-

The following table gives the count of leads who belong to these cities:



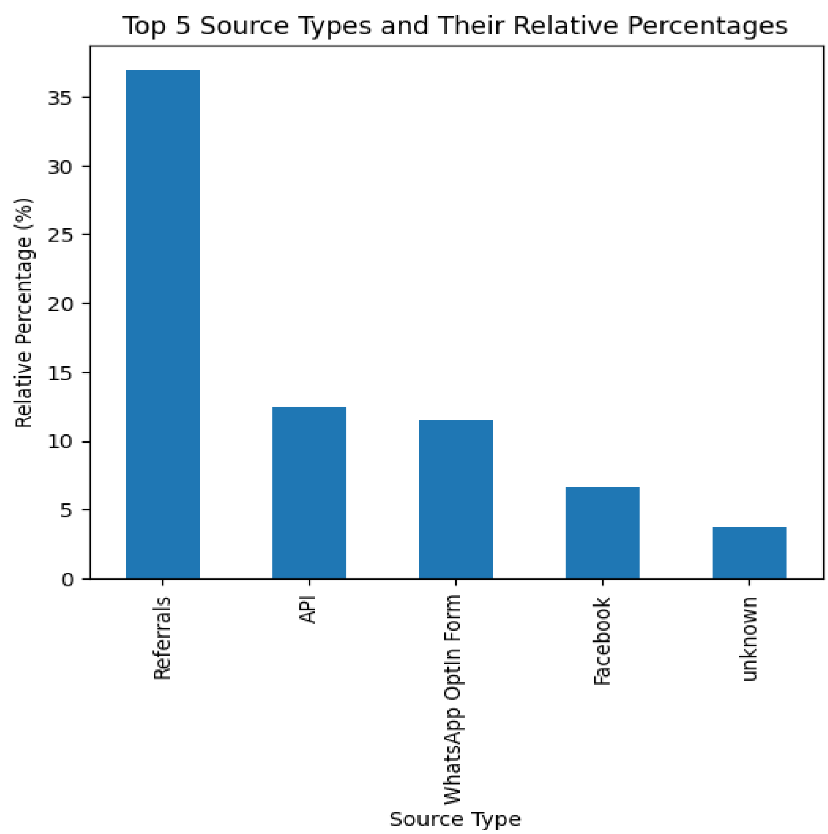
Clearly most of the leads have been generated in Southern States of India.

The following chart gives the count of users if they used our Job Assistant or Came through the 3rd party source.



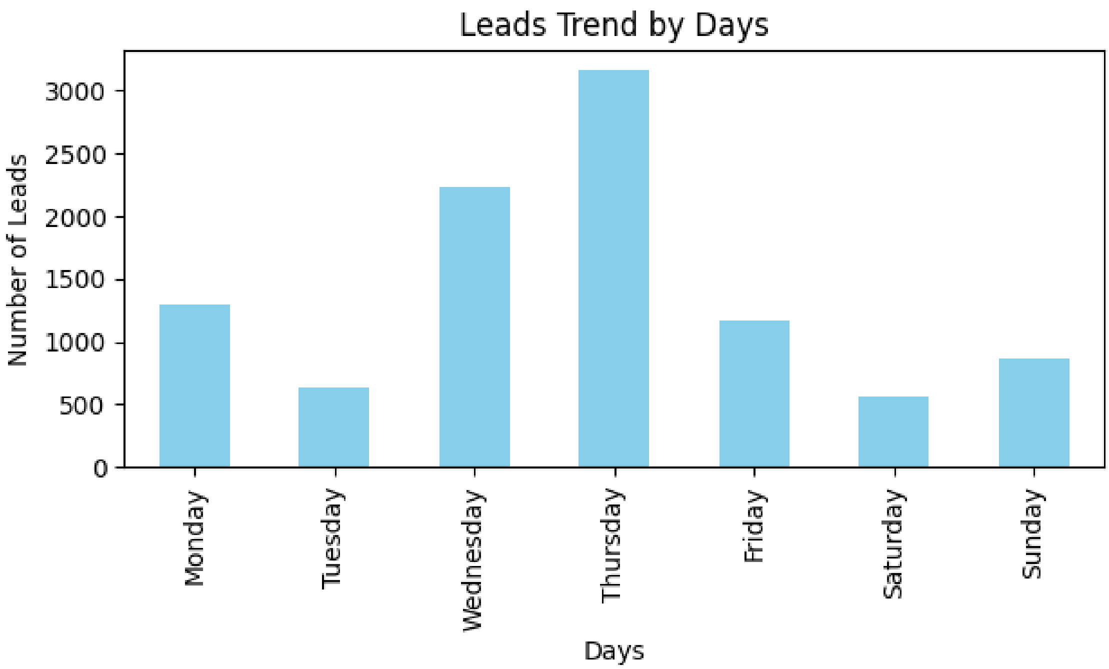
We See that majority of the users came from our Job Assistant

Plot of the top 5 types of sources of leads generation and their relative %.:-



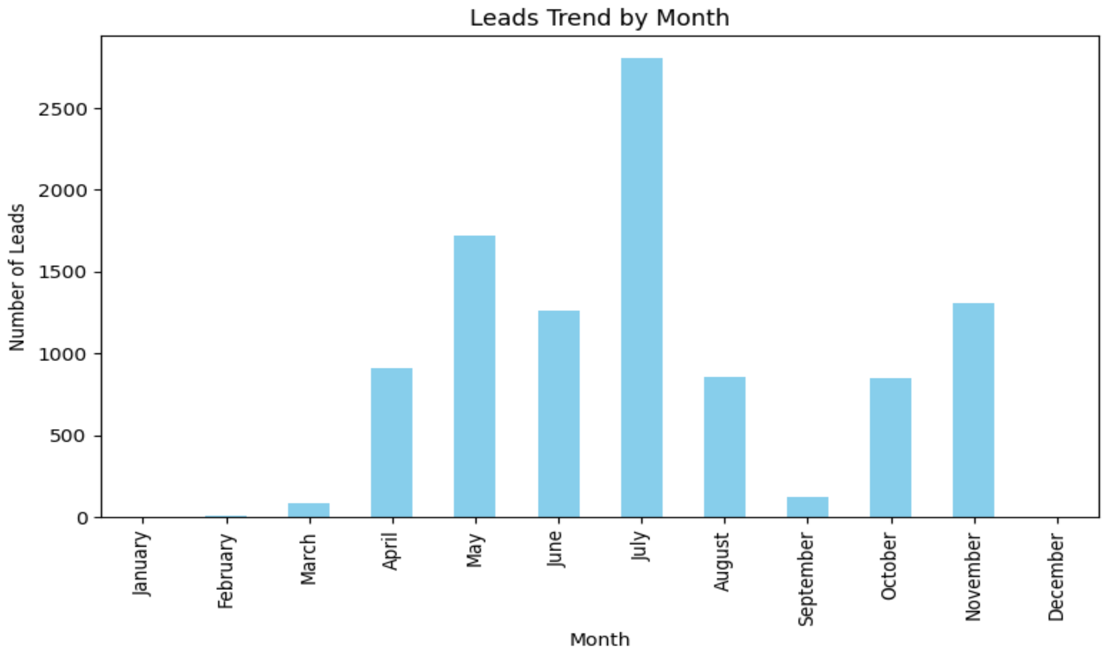
We see that most of the leads came through:- **referrals**, **api**, **whatsapp**, facebook

Trend in the number of leads generated on daily basis:-



Key findings:-
Most of the calls have been made on the weekdays.

Trend in the number of leads generated on monthly basis:-

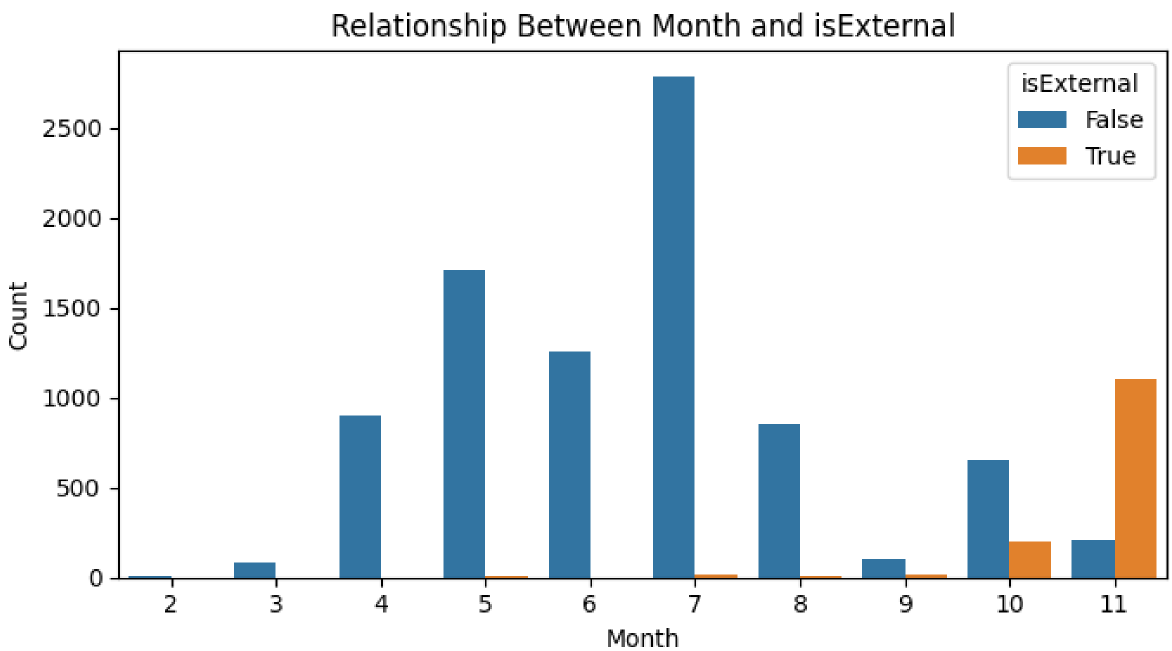


Key Findings:-

Most of the users came to use in the months of May, June, july.
Actually an increasing pattern was observed from January till july.
And again it started increasing from September till November.

Relationship across features:-

Relationship between the month of the leads data received and the platform used [i.e., Our Job Assistant/External Source]

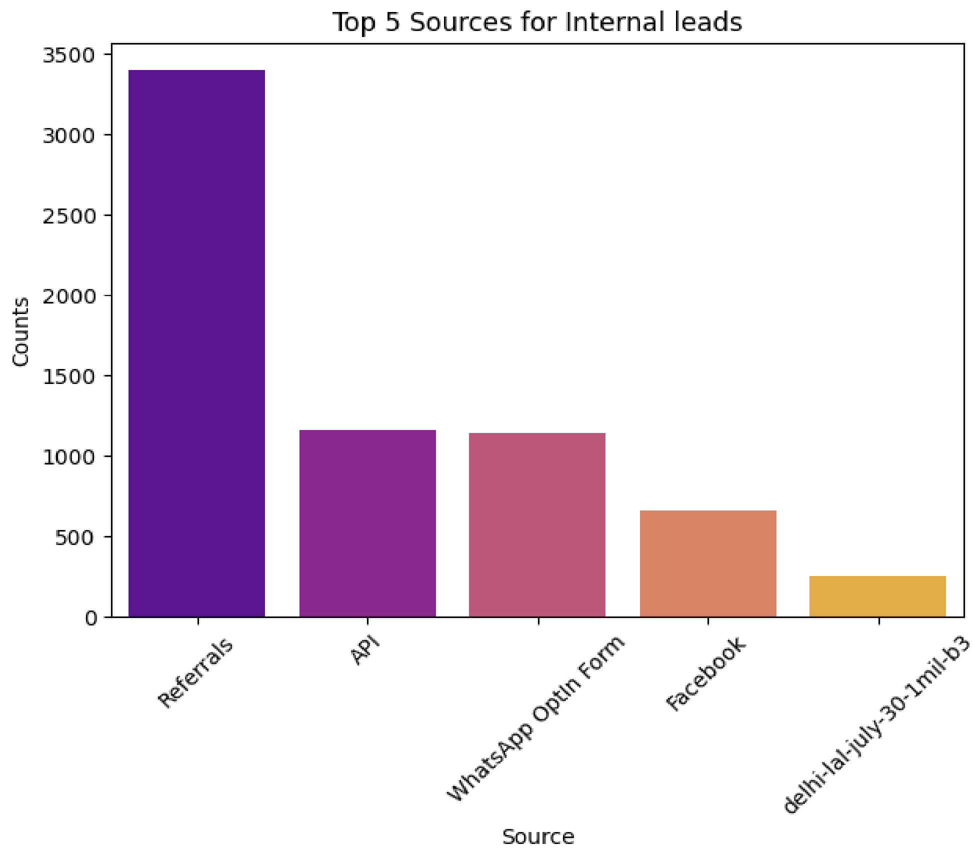


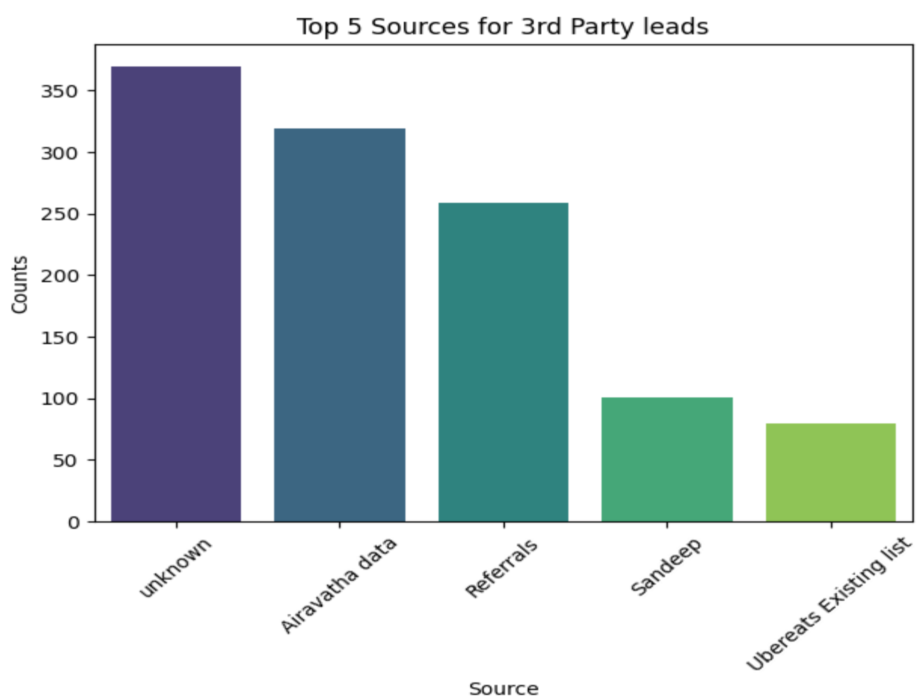
Key Findings:--
we can see that most of the 3rd party sources leads are coming at the end of months.
This might indicate the following possibilities:-

The 3rd party sources are particularly effective in these intensified periods, possibly due to promotional activities or special offers designed to meet monthly quotas.
This could ensure optimal engagement with these type of leads when they are most likely to convert.

Understanding this pattern can help in planning the lead nurturing process. For instance, if leads tend to cool off at the beginning of a month, strategies can be developed to maintain interest and engagement during these quieter periods.

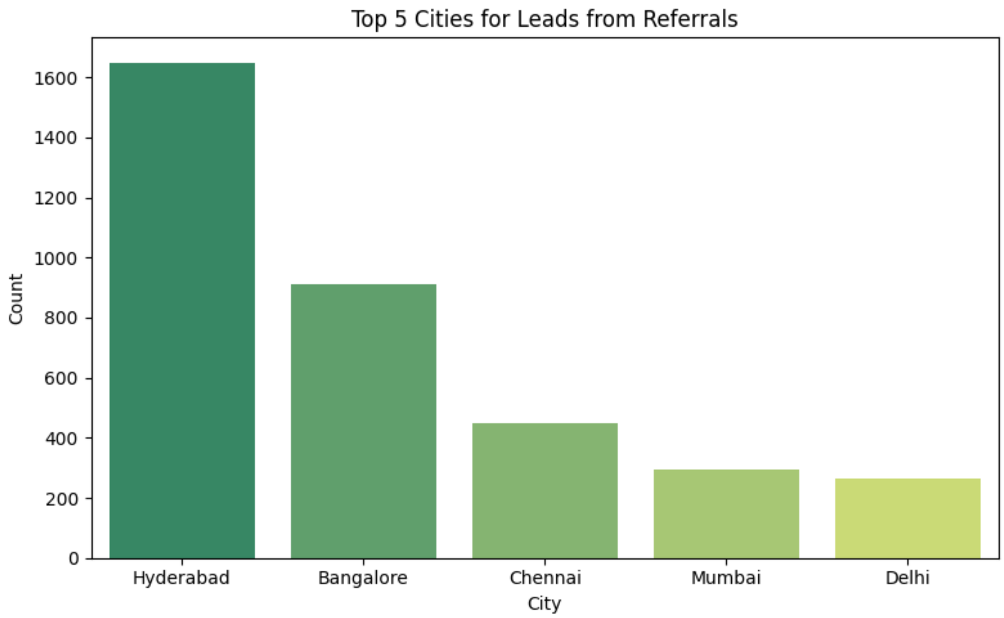
Third_party_leads V/s Source types && internal_leads V/s Source types





From above charts we find most of the internal leads are coming through **'referrals'**.

Top 5 Cities for Leads for source type:- Referrals



Findings:- Most of them are coming from **Hyderabad**

Date column

The trend analysis for the **"receivedAt"** column has yielded the following results:

Trends by Month:

The leads received vary by month, with the highest in July (2,801 leads) and the lowest in September (122 leads).

Trends by Day of the Month:

The first day of the month has the highest number of leads received (3,177 leads), with other days varying significantly.

Trends by Day of the Week:

Wednesday (3,160 leads) has the highest number of leads received, followed by Tuesday (2,229 leads).

The weekend days (Saturday and Sunday) have fewer leads received compared to weekdays.

Trends by Hour of the Day:

The majority of leads were received at midnight (00:00), with 7,885 leads, indicating a possible batch processing or system update rather than actual lead generation activity.

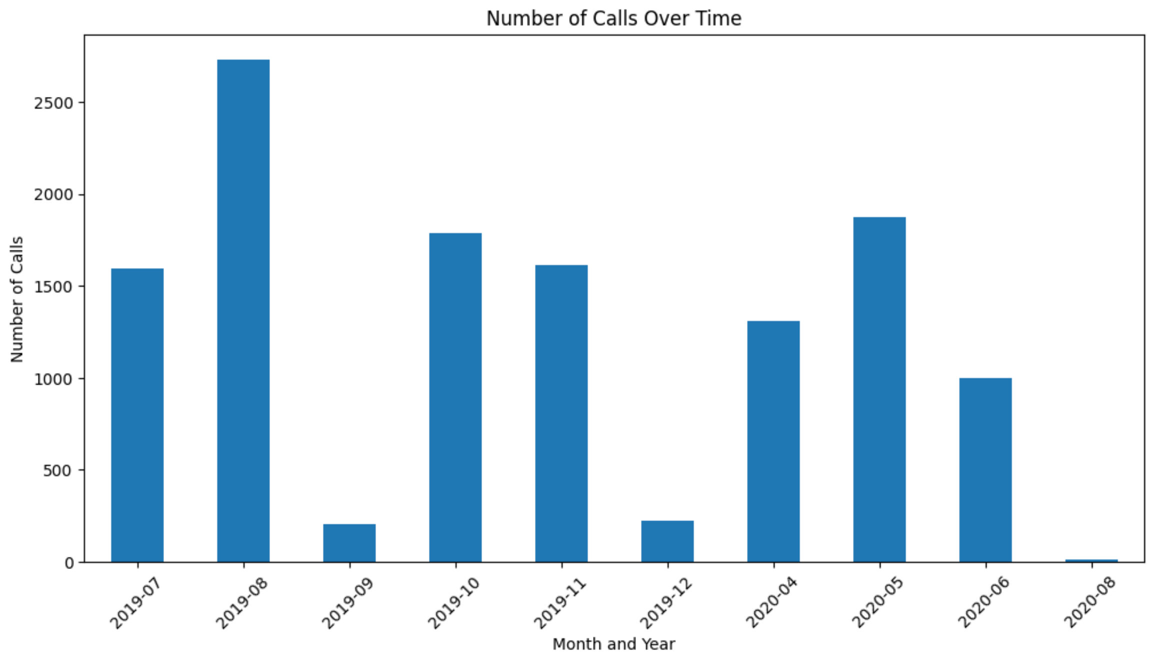
Other hours of the day show much smaller numbers, with some increased activity around 5 AM and in the early evening.

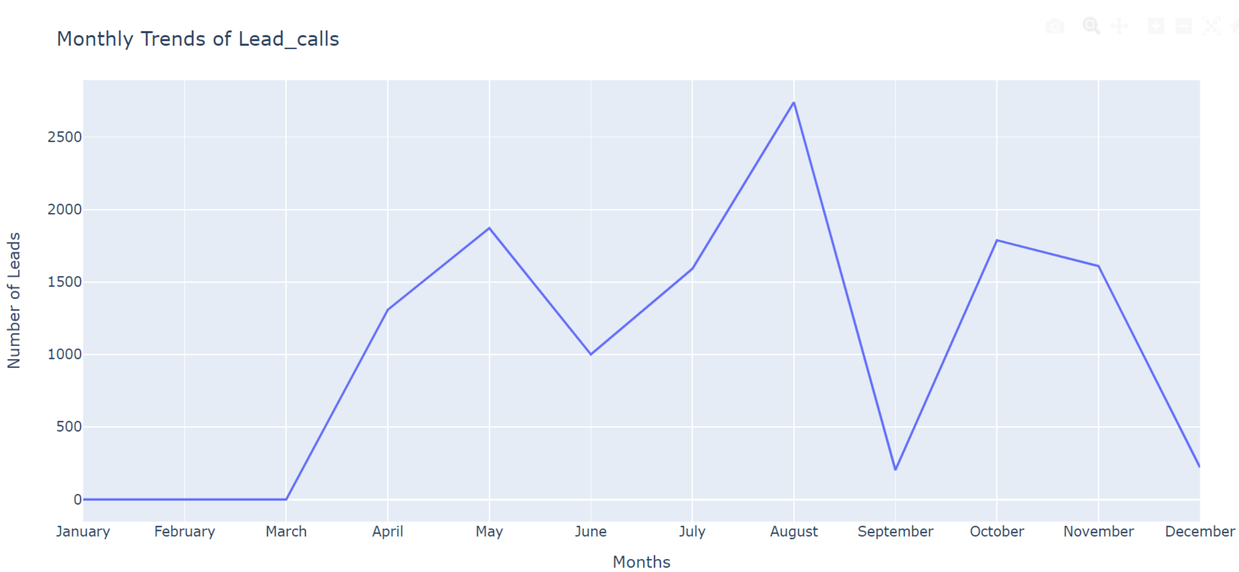
Weekends vs. Weekdays:

A significant majority of the leads were received on weekdays (8,487 leads) compared to weekends (1,429 leads).

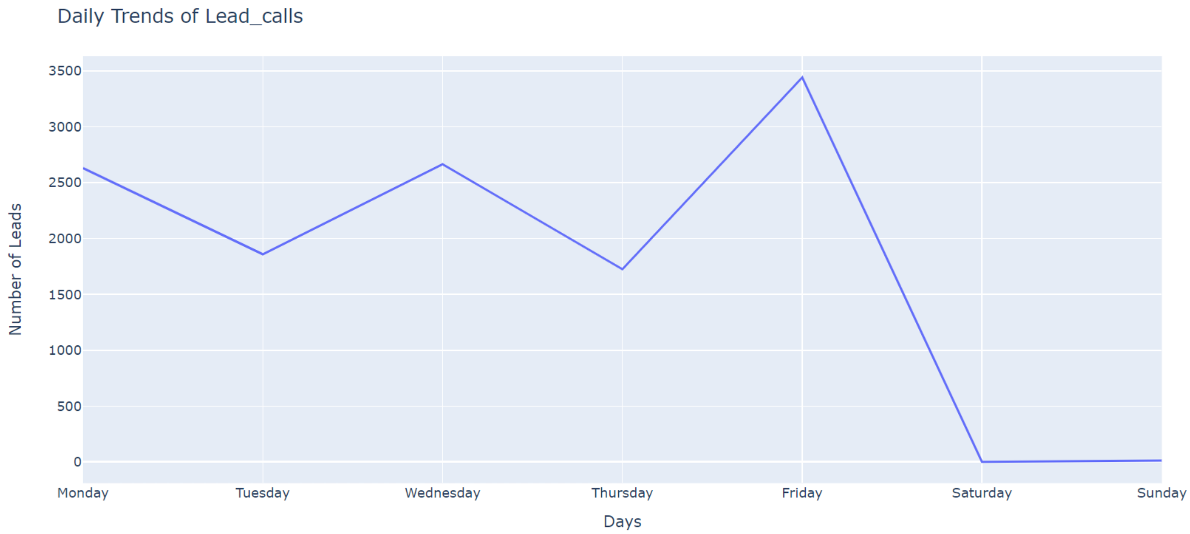
2. Lead_calls dataset:-

Year /month wise analysis of lead_calls.





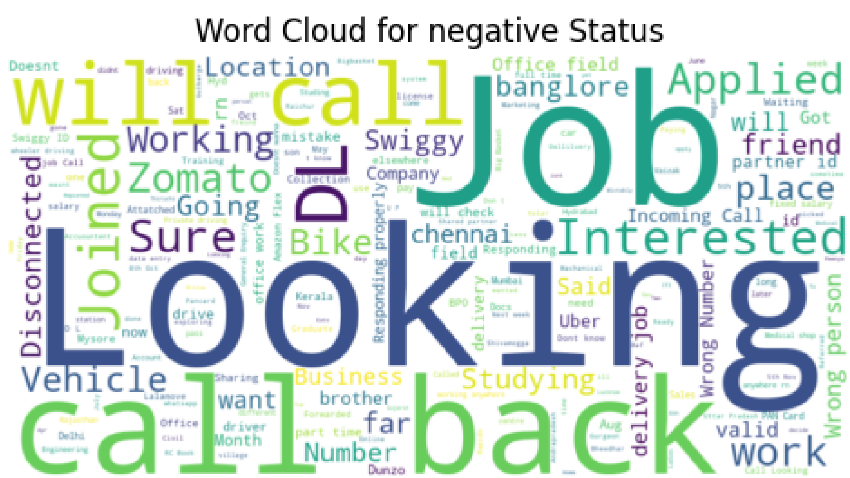
Most of the calls have been done in the month ranging **[March to November]** by our telecallers with August having the maximum of all



From the line chart we can see that mostly the calls made by the telecaller are on the weekdays, and around 0 calls on weekends

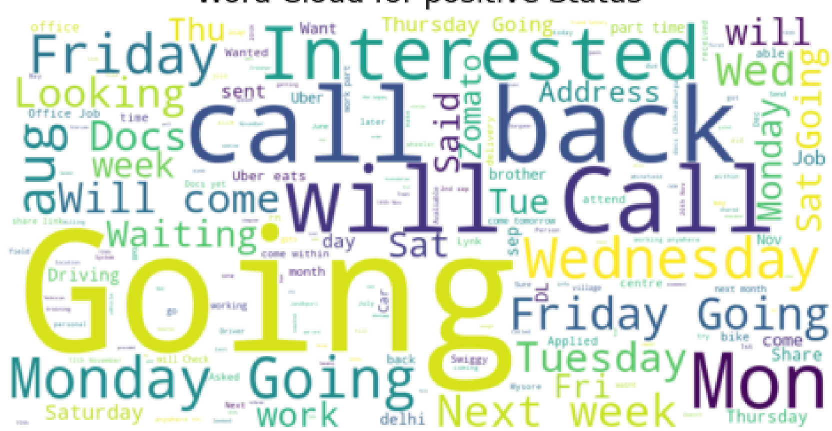
Analysis of the comments given by the leads to our telecallers:-

Word clouds for different status types:-

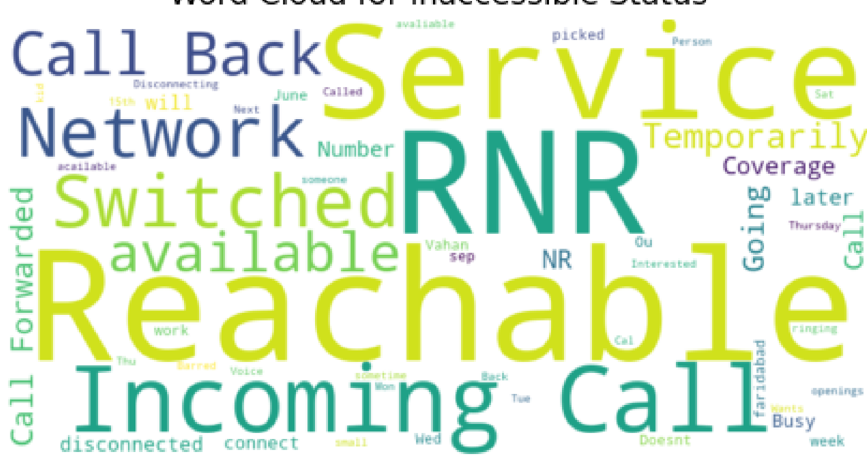


Word Cloud for rnr Status

Word Cloud for positive Status



Word Cloud for inaccessible Status



Findings from the Word cloud of different status types.
Negative Status:

Prominent words like "Looking", "Job", and "call back" suggest that many calls may relate to job inquiries or job-related discussions where the outcome was not positive.

RNR Status (Ring No Response):

The word "busy" is prominent, which could indicate that calls are often not answered because the recipient is busy.

Other terms like "office" and "Number" might imply that calls during working hours or issues with contact information contribute to the RNR status.

Inaccessible Status:

Words such as "Service", "Reachable", and "Network" suggest that calls often cannot be completed due to service issues or network problems.

"Incoming", "Call Forwarded", and "Switched" might indicate that technical reasons are a common cause of inaccessibility.

Positive Status:

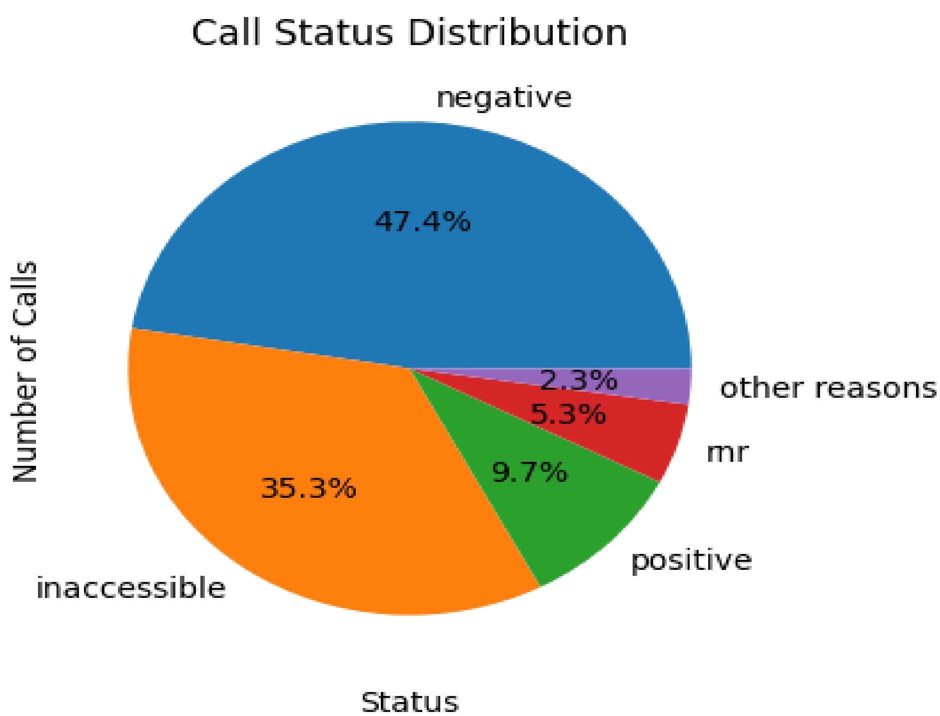
The word "Interested" appears again, likely indicating that interest expressed on calls correlates with a positive outcome.

Other Reasons Status:

"Repeated number" and variations like "Repeated" and "number" are very prominent, hinting at issues with contact numbers being a common reason for calls being categorized under 'other reasons'.

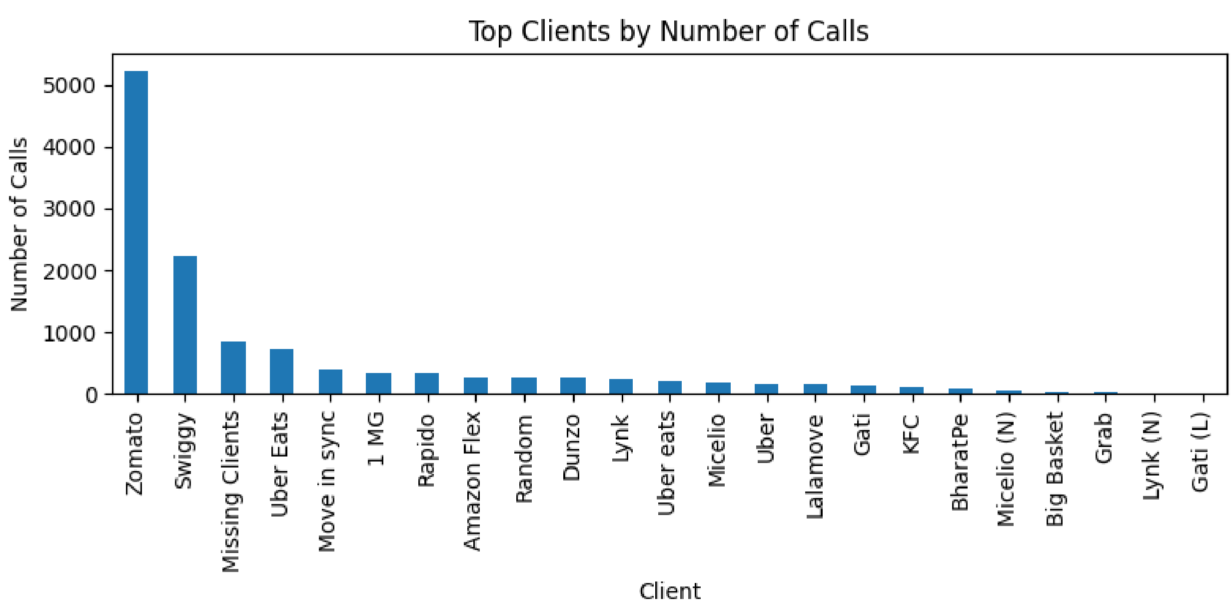
"Working" and "Interested" suggest that the status might be related to employment status or job interest, but not fitting neatly into the positive or negative categories.

Distribution of lead calls by status type



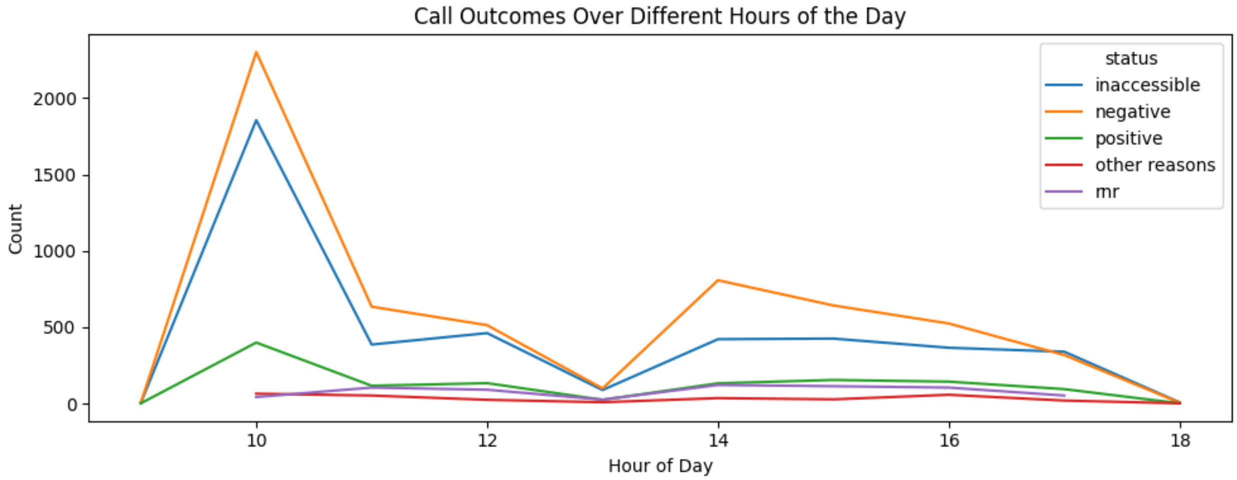
Only around **10%** of the calls are turning out to be **+ve**

Analysis of calls per client



We find that most of the calls have been made targeting **Zomato** and **Swiggy** with a total of 60%.

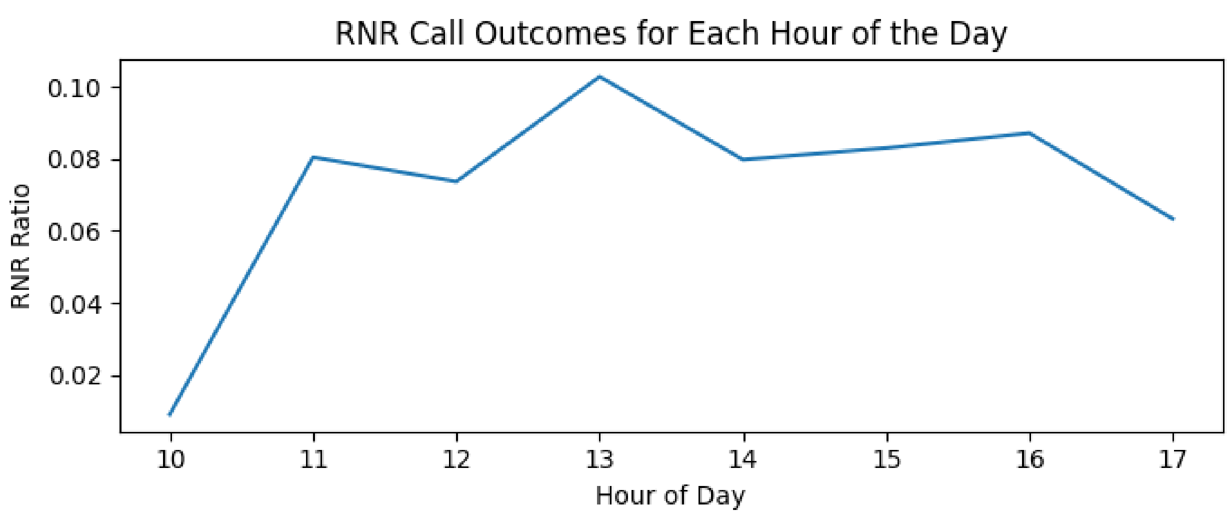
Call Outcomes over Different Hours of the Day

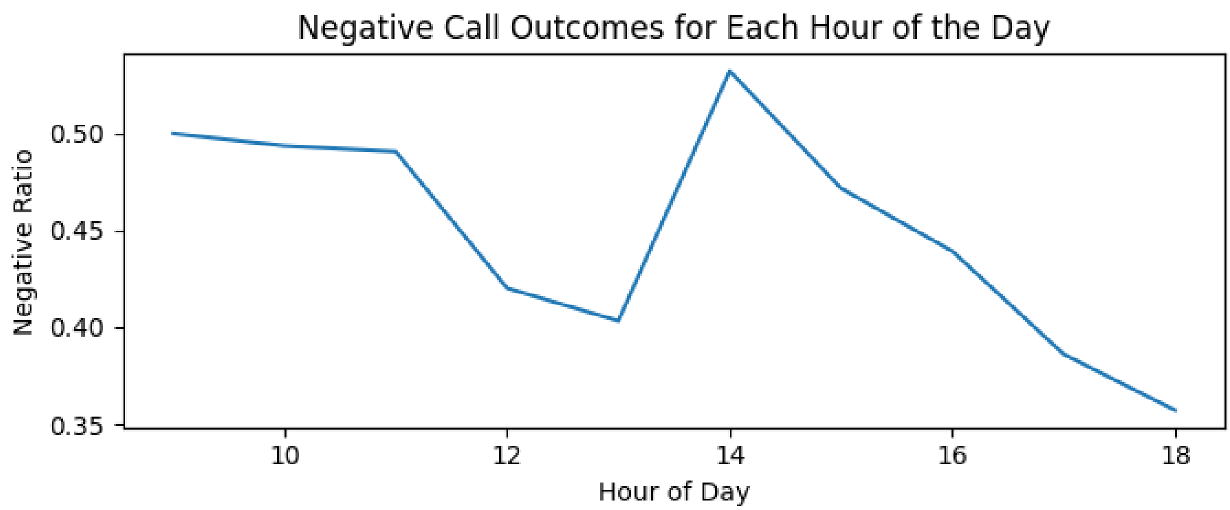


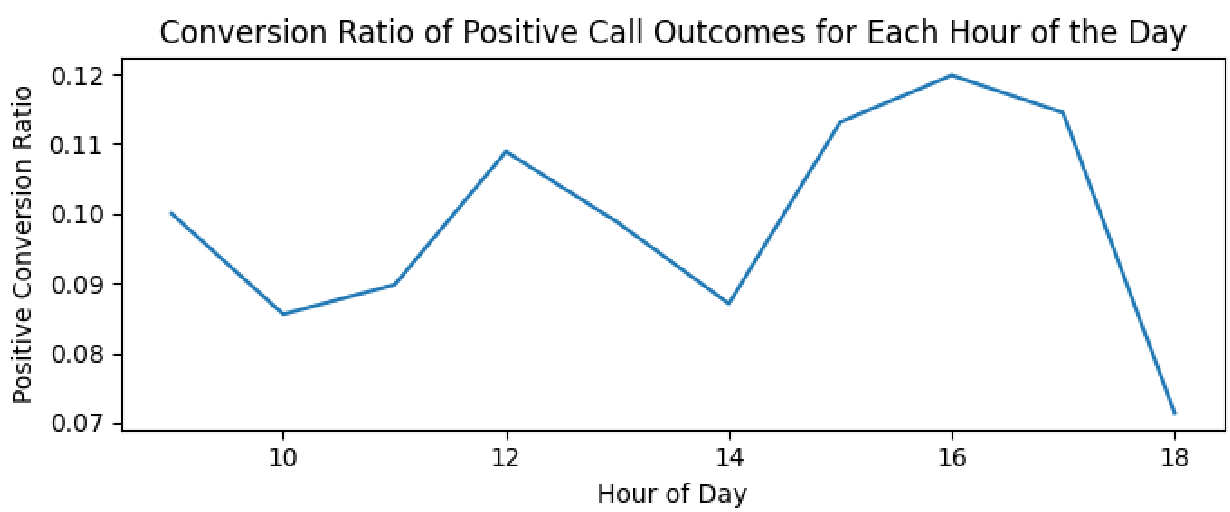
Key Findings:-

Most of the calls are made in the initial hours of the day around 10.00 am morning.
And it's obvious from the graph that during lunch and post evening the call made become too less or 0

Line chart for the types of call outcomes over different hours of day.







Based on the rate charts of different call statuses across various hours of the day, we can derive several business insights:

Negative Call Outcomes:

The highest rate of negative outcomes seems to occur around midday. This could be due to various factors such as leads being busy during lunch hours or experiencing midday work pressure, which might make them less receptive to calls.
The rate of negative outcomes decreases as the day progresses, especially after peak business hours. This might indicate that leads are more available and potentially more receptive later in the day.

RNR (Ring No Response) Call Outcomes:

The RNR rate is relatively stable throughout the day with slight variations. However, there's a noticeable dip in the late afternoon. This might suggest that leads are more likely to answer calls during the early part of the day or later towards the evening.
The relative stability of the RNR rate might indicate that leads have consistent periods where they are not available to take calls, such as meetings or personal commitments.
Positive Call Outcomes:

The conversion ratio of positive outcomes shows fluctuations throughout the day, with peaks around mid-morning and mid-afternoon. This suggests that these may be the best times to reach out to leads for a positive response.

The drop in positive outcomes towards the end of the workday could indicate that leads are winding down and may not be interested in engaging in new conversations.

Strategic Business Actions:

Schedule Optimization: Adjust the timing of call activities based on these insights. For example, avoid calling during hours with the highest negative rates and focus on times with higher positive response rates.

Caller Training: Train telecallers to anticipate and handle midday negative responses more effectively, perhaps by preparing for quick, concise pitches during these hours.

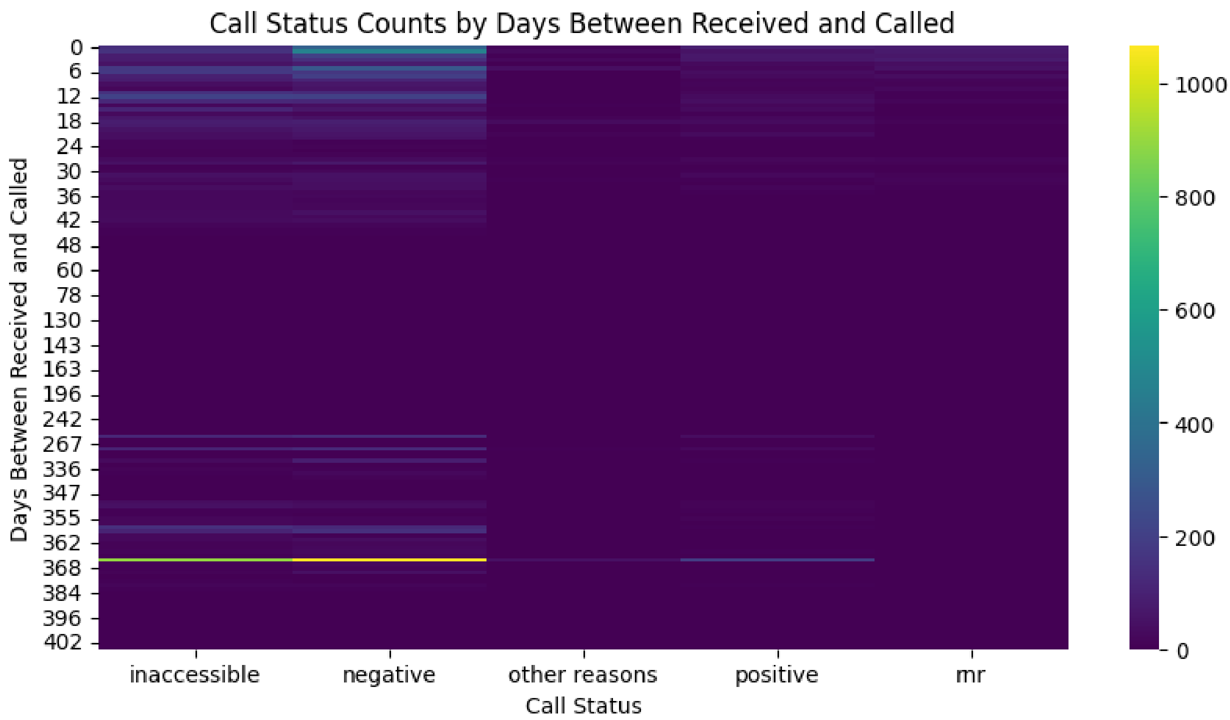
Call Routing: Implement a call routing strategy that considers the availability patterns suggested by the RNR rate. This could involve rerouting calls to times with historically better response rates.

Follow-Up Strategy: Develop a follow-up call strategy for the hours immediately following the positive peaks to capitalize on the higher engagement levels during these times.

These insights can be used to refine call scheduling, enhance training for call center staff, and improve overall call strategy to maximize successful outcomes and lead conversion.

3. Merged dataset findings:- (leads, lead_calls, telecallers)

Heat map of the day_difference column against all the different call status types.

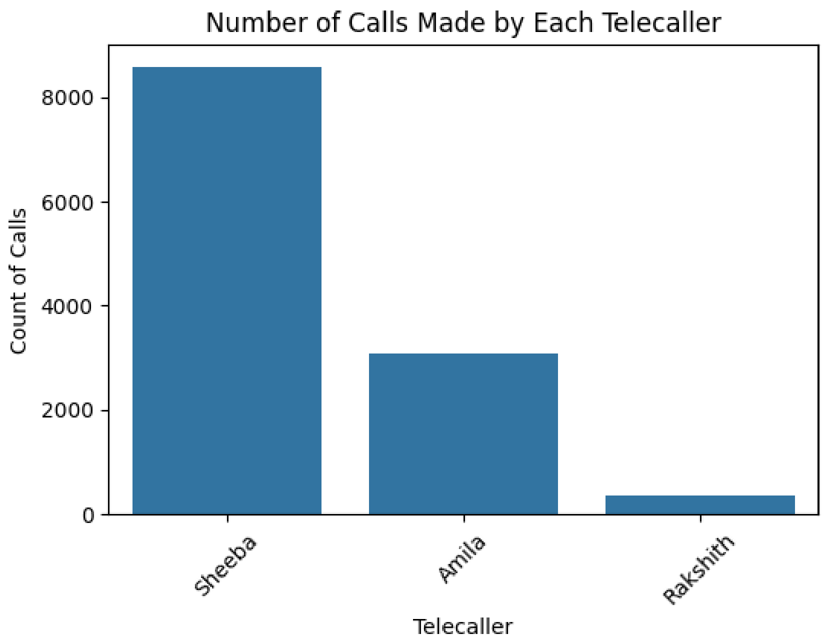


Key findings from this heatmap:-

Longer Delays and Outcomes:
For leads that were contacted after a longer delay, there is still a presence of 'positive' outcomes, but it's less frequent compared to negative outcomes, which could imply that the chances of a successful call decrease as the delay increases.

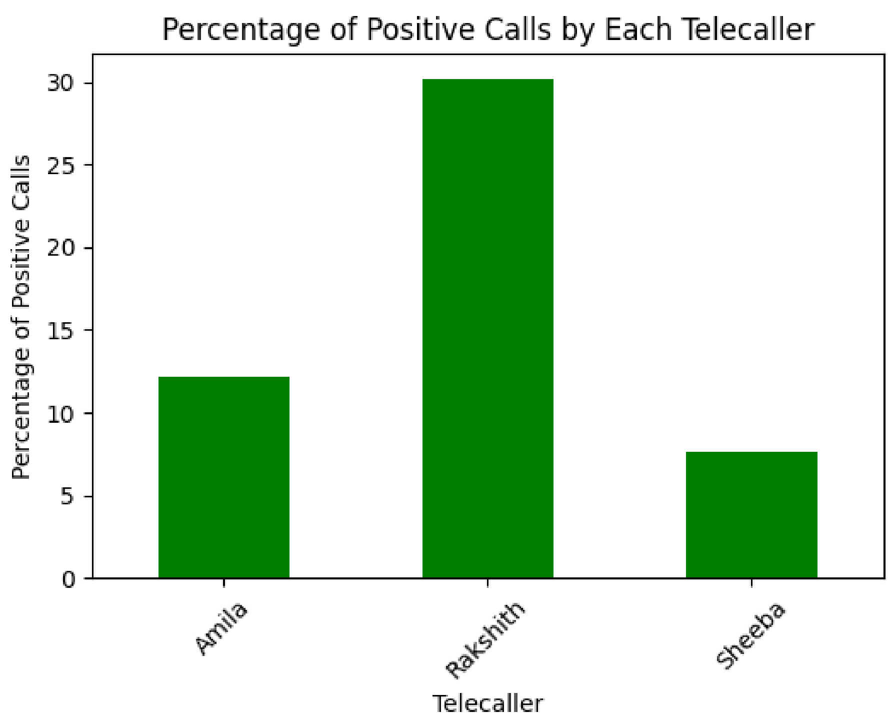
Positive Outcomes with Prompt Calling:
There is a visible band of 'positive' outcomes that seem to occur more frequently when the call is made within a short time frame after receiving the lead. This suggests that prompt calling may correlate with more successful outcomes.

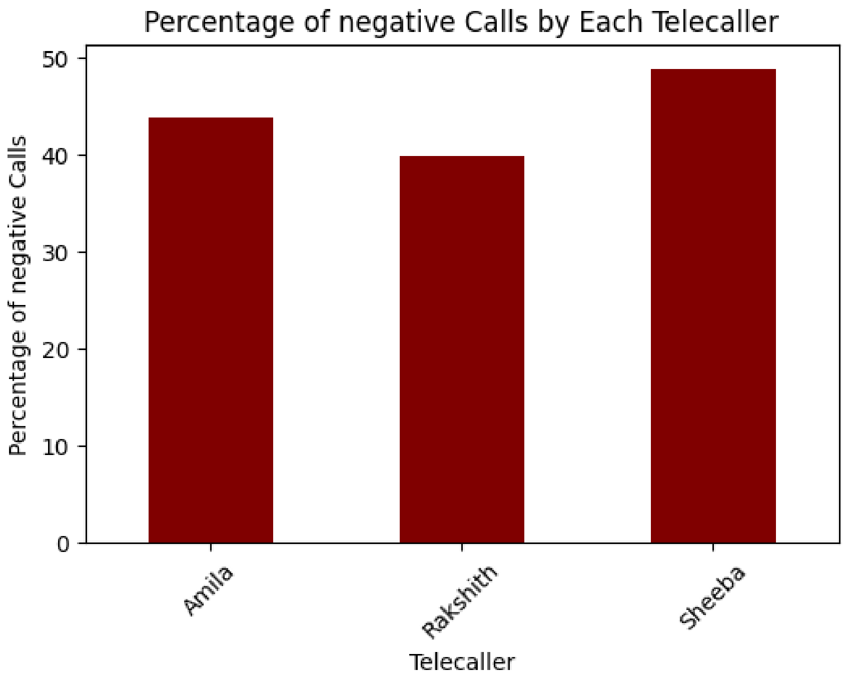
Lead Call analysis for each Telecaller



Sheeba has made the maximum number of lead calls.

% of positive calls for each telecaller





Key findings from the call analysis of the telecallers:-

Rakshith has the highest percentage of positive calls, suggesting a high level of effectiveness in engaging with leads and possibly a strong skill set in converting calls into favorable outcomes.

Sheeba, despite making the most calls, has the lowest percentage of positive calls, which may point to issues with lead quality, call approach, or possibly indicate a need for additional training or support.

Suggested Actions based on above analysis:-

Rakshith's techniques could be analyzed and potentially taught to other telecallers to improve their conversion rates.
Sheeba needs additional training on dealing with leads to get them converted to positives. Or maybe the number of calls allocated to her is disturbing the quality of calls.

Questions:

1. How efficient and effective are our tele-calling operations?

Ans:-

Efficiency and effectiveness can be inferred from the distribution of call outcomes and the insights drawn from the datasets.

Efficiency: The high volume of **'negative'** and **'inaccessible'** statuses suggests that a significant proportion of tele-calling efforts do not result in successful connections or positive outcomes. This may indicate inefficiencies in either the targeting process or the timing and method of calls.

Effectiveness: The effectiveness appears to be low, as evidenced by the low rate of **'positive'** statuses across clients. This implies that calls are not often resulting in a successful sale or positive interaction.

2. What are three things you would recommend we do to the efficiency and effectiveness of our tele-calling operations?

Based on the dataset analysis, the following recommendations can be made to improve tele-calling operations:-

Enhance Lead Qualification: Use data-driven strategies to improve lead scoring and qualification. By analyzing the common characteristics of calls that result in positive outcomes, the business can better target leads that are more likely to convert, thus improving both efficiency and effectiveness.

Optimize Call Timing: The 'days_difference' (difference b/w the telecaller calling the lead and lead registering on job assistant) feature suggests that timing plays a crucial role. Calls made too late or too early may not be effective. Analyzing the optimal timing for calls and adjusting the tele-calling schedule accordingly could yield better results.

Training and Script Refinement: From the word cloud analysis for each of the call status types, we found that some keywords are important. It may be beneficial to refine the tele-calling script and provide additional training to the telecallers to handle objections effectively and guide conversations towards a positive outcome.

3. What are some other ways in which we can utilize this dataset to add value to Business, Operations or Product?

The dataset can be leveraged in multiple ways to add value across various domains:

Customer Segmentation:
The data can help identify different segments of leads or customers, which can then be targeted with tailored strategies.

Product Feedback:
The dataset may contain implicit feedback on products or services, which can be valuable for the product development team.

Predictive Analytics:
Building predictive models to forecast call outcomes can help allocate resources more effectively and prioritize high-potential leads.

Process Optimization:
Analyzing patterns and correlations in the data can identify bottlenecks or inefficiencies in the tele-calling process.

Market Trends:
Analyzing call outcomes in relation to client types and sources can provide insights into market trends and customer preferences.

Time-based analysis:
Properly analyzing the trend followed over time in the nature of call status can really add value to the quality of the leads.

Model building:- features and techniques used

I tried building classification model on the final modified dataset:-

Key steps done:-

1. Checking the % of missing values in each of the features.
   Actions:- Dropped the comment column since it was having 71% of missing values.
2. Labeling the companies whose count was less than 5% as “Others”. Since there were around 50 different categories.
3. Put the similar meaning status under a single status type: - [‘positive’, ‘negative’, ‘inaccessible’, ‘other reason’, ‘Rnr’]
4. Similarly did the same for the source, city feature.
5. Checking the distribution of day_difference. (day_difference = calledAt- receivedAt).
   Action:- Applied square root/logarithmic transformation to handle the skewness in this feature.
6. Did all the basic model building steps like, train test split, encoding the categorical features.
7. In order to handle the imbalance target variable we used Smote technique(Synthetic Minority Over-sampling Technique), which does oversampling of the less frequent class and thus helping in handling the imbalance variable.
8. Used PCA (principal component analysis) to get an optimal number of features with 90% variance explained.
9. Applied various ml classification models like Logistic regression, KNN, Decision tree, random forest, support vector machine and checked their accuracy.

The model accuracy came out to be insignificant. There could have been different reasons behind this, maybe I missed some important feature which could have impacted the model significantly. Also some different feature selection and feature extraction techniques could be explored as per our business problem.