Real Estate Analysis

Please use the following structure as a guide :

- All scripts are in the script folder.
- All visualizations are in the viz folder.
- data folder contains the dataset.
- assets folder contains miscellaneous screenshots for the report.

Motivation

The aim of this project is to analyse the given real estate dataset and provide a scoring criteria to measure an agents performance.
In this analysis, I have tried to wrangle the dataset provided from the CRM platform and have used this dataset to come up with an agent performance scoring criteria and use this algorithm to choose the Top 3 agents in the brokerage firm.

I have also included suggestions for the brokerage firms to improve their agent's performances.

Background

A Real estate agent works under a licensed broker (or brokerage). The brokerage provides the real estate agent with leads (leads are people who are looking to buy, sell or rent), and the agent's job is to help them in their home buying, selling or renting process.
They do this by regularly following-up with them using calls, text messages or emails. After every follow-up with a lead, the activity is recorded on the CRM(CRM is a Customer Relationship Manager, which is a tool that agents use to log data for a particular lead).

Exploring the Data

The dataset provided comes from a CRM platform used by the brokerage firm. The dataset has a CSV file format. The following is a screenshot of data along with the definitions of each column in the dataset :

Example (from dataset) - An agent with agentId 10 made 2 calls to a lead with leadId 20293

id	followup_date	lead_created_at	leadId	followup_type	agentId	additional_data
1	2020-03-05 14:38:54	2019-11-04 16:12:00-05:00	20293	calls	10	{'duration': 0, 'is_incoming': False}
2	2020-03-05 14:38:55	2019-11-04 16:12:00-05:00	20293	calls	10	{'duration': 0, 'is_incoming': False}

- **id**: Unique followup identifier
- **followup_date**: Timestamp when the followup happened
- **lead_created_at**: Lead creation timestamp on the CRM
- **leadId**: Unique id of the lead with whom followup happened
- **followup_type**: Type of followup - Call/ Email/ Text
- **agentId**: Agent who followed up with the lead
- **additional_data**: Additional data associated with the followup
    1. **duration**: Time for which call lasted in seconds
    2. **isIncoming**: False means agent made the Call/ Email/ Text

The columns follow_up_date and lead_created_at should be **datetime**. Columns id, leadId, agentId are **integers** and rest are **strings**.

The dataset provided has no null values and hence means it is a clean dataset with no data imputation needed initially.
But, data has been manipulated as per the analysis performed on it.

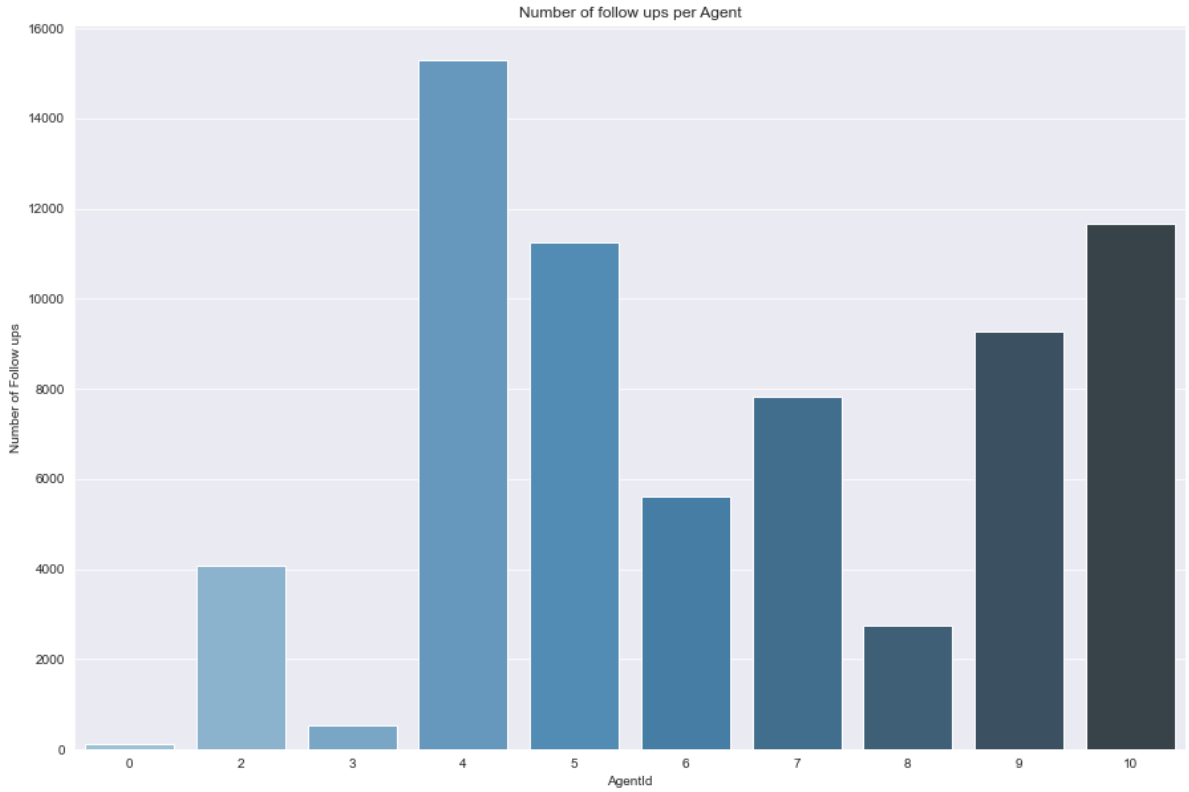
Analysis

The following libraries have been used for the analysis process :

- Pandas
- Seaborn
- Matplotlib

At first glance of the data, we see that each row corresponds to a follow up made by an agent to their leads. There can be multiple rows where one agent makes a follow up to the same lead over and over again. This is normal in real estate scenarios as an agent keeps going back and forth with their potential customers.

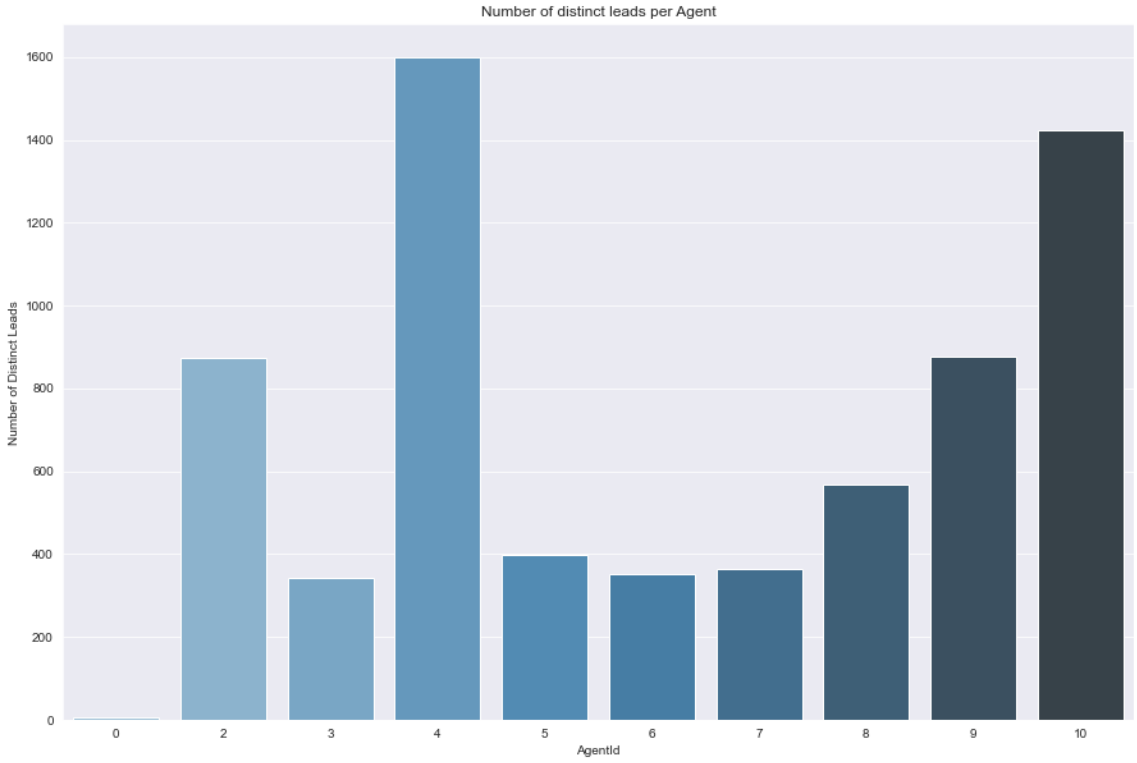
Thus, the first step in my analysis was to count the number of follow ups done by an agent. On its own, the count of follow ups of each agent doesn't necessarily mean that they are better agents. It could be just that they made multiple calls/mail/text to a single lead.



We can see that Agent 4 made the most number of follow ups. This also comes out to be around 22% of the total follow ups. On the other hand, Agent 0 made the least number of follow ups which was around 0.17%.

```
4     22.37
10    17.05
5     16.45
9     13.58
7     11.44
6      8.20
2      5.98
8      4.00
3      0.77
0      0.17
Name: agentId, dtype: float64
```

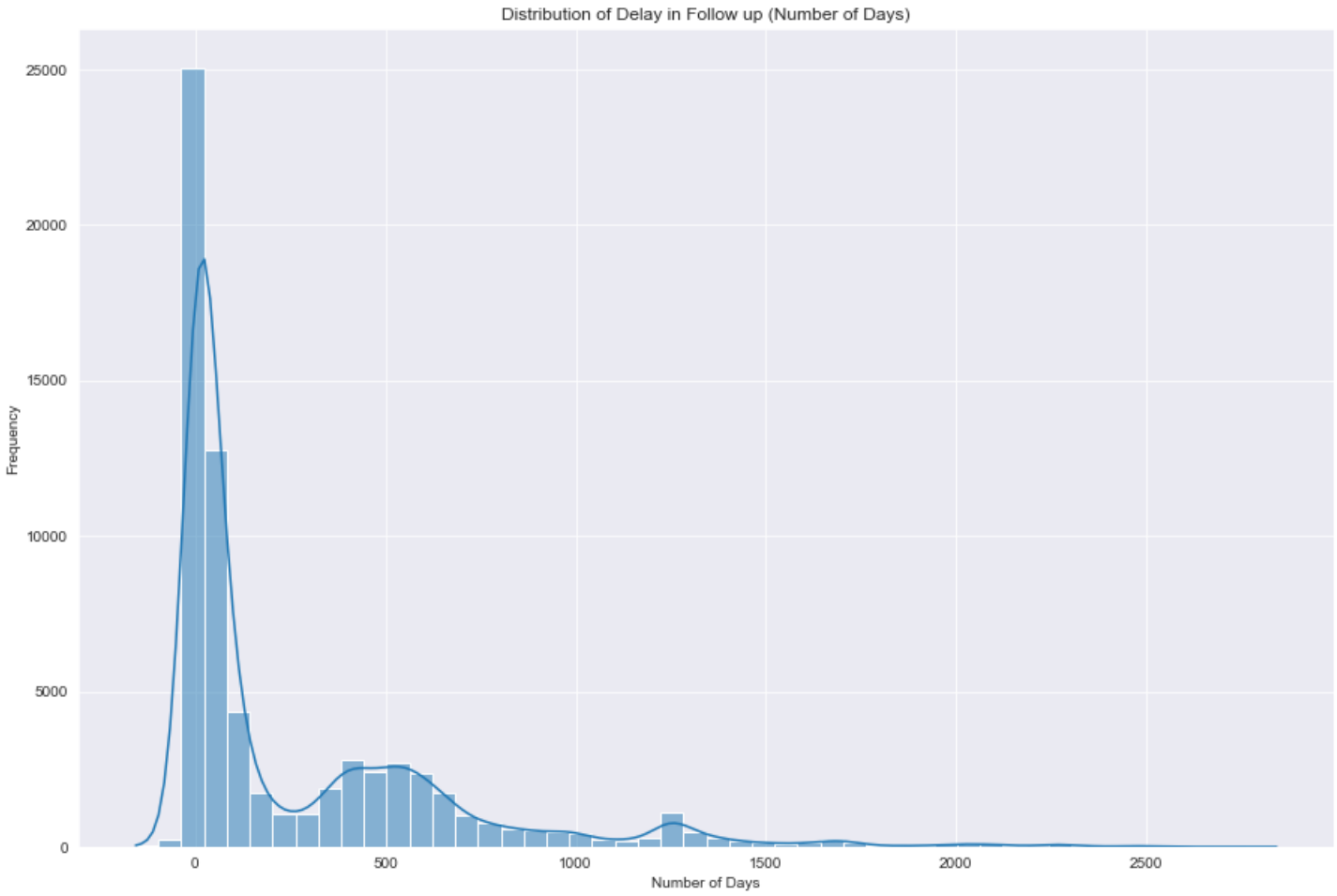
But, as discussed above, it doesn't make sense to show a number of follow ups on its own. Another performance indicator could be the unique number of leads each agent has.



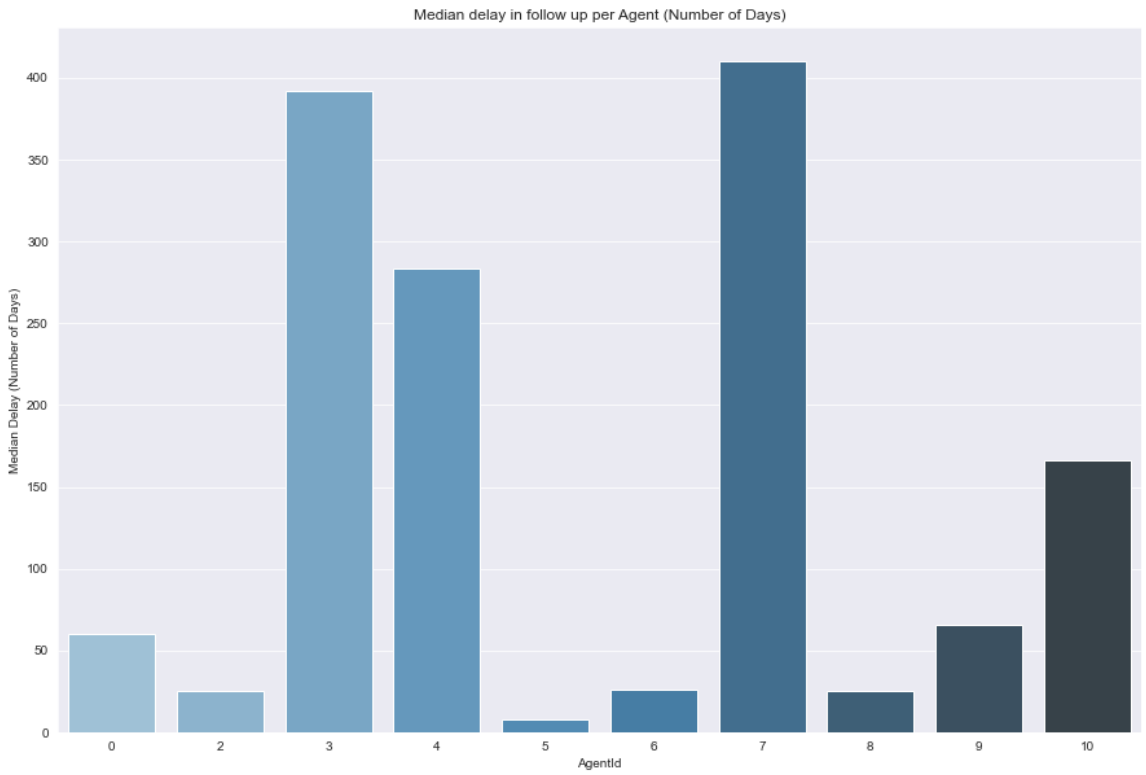
Again, the graph shows that Agent 4 has the most number of leads and Agent 0 has the least number of leads. This still doesn't guarantee a great performance indicator. Instead, it could just mean that the brokerage has put more trust on Agent 4 as they have provided them with the most leads and Agent 0 might be a new agent hence has the least number of leads.

I went deeper into the data and found that there is a Date when the lead is created and a follow up date. This could provide an idea about how long does each Agent take to follow up to their lead. A higher value means they are taking a lot of days to follow up and the lead might not be satisfied with this agent.

Since, the follow_up_delay (days) is a highly right skewed distribution, it would make sense to look at the median rather than mean as an indicator because median is not affected by outliers.


Distribution of Delay in Follow up (Number of Days)
Frequency
25000
20000
15000
10000
5000
0
0
500
1000
1500
2000
2500
Number of Days

The following is the median number of days taken by each agent to follow up to their leads.

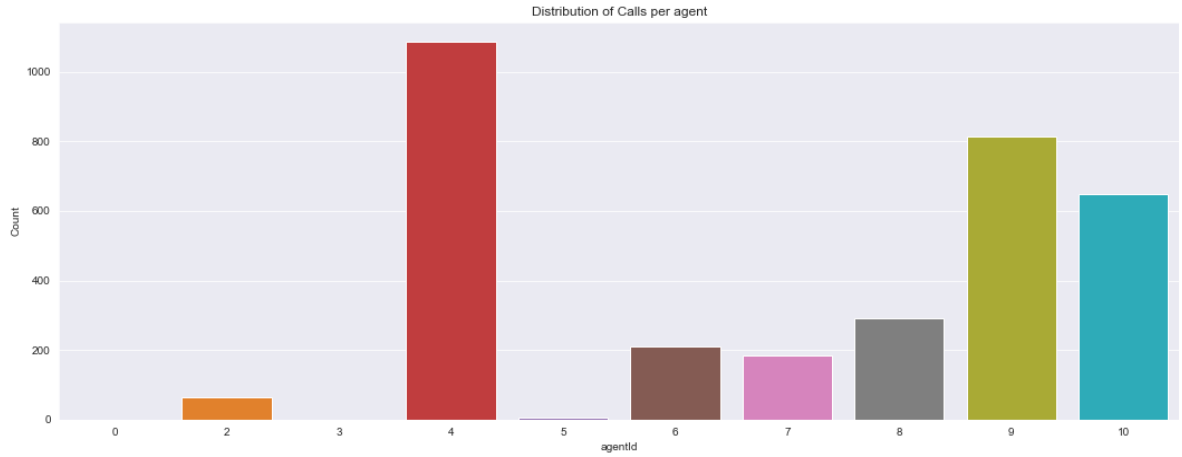


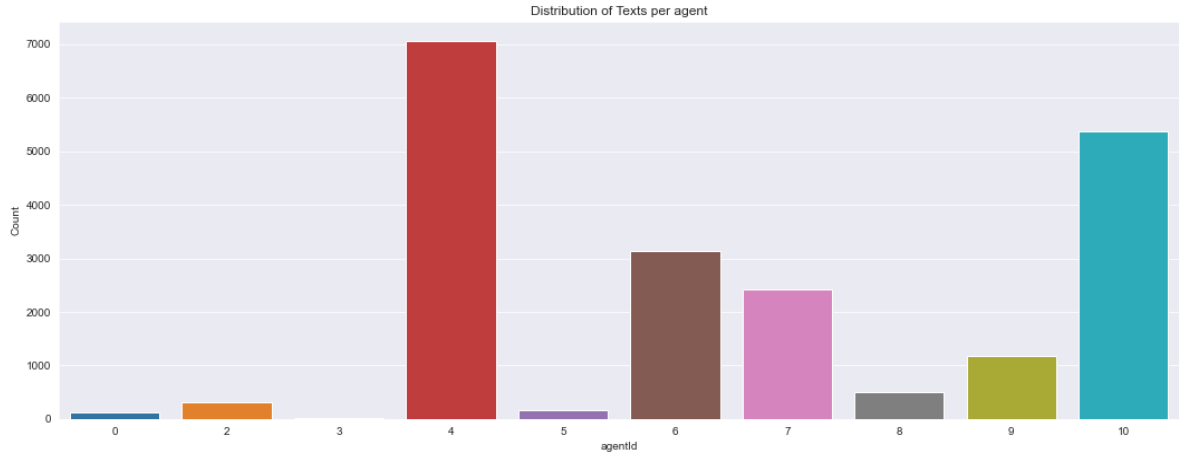
As you can see, Agent 5 takes only 8 days to get back to their leads whereas Agents 3 and 7 take more than a year to get back to their leads. This might mean that the potential customers aren't happy with agents 3 and 7 as they don't want to be kept waiting for such a long time.

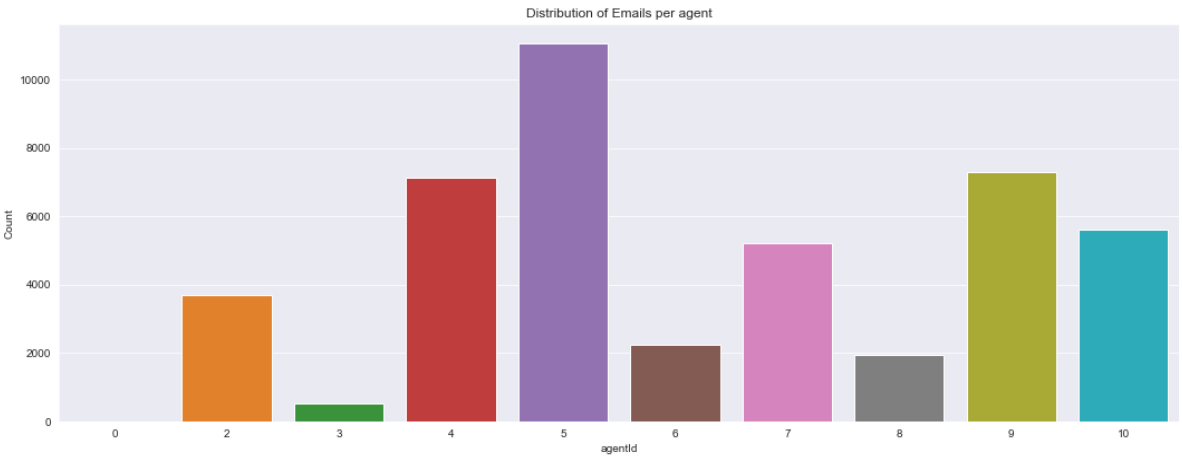
```
agentId
0      60.0
2      25.0
3     392.0
4     283.5
5       8.0
6      26.0
7     410.0
8      25.0
9      66.0
10    166.0
Name: follow_up_delay (days), dtype: float64
```

If the median delay is decreased then surely the potential customers would be satisfied with the agents and thus increasing the profits of the brokerage firm.

Additionally, the follow up type is a good indicator of performance as leads are more likely to reply to a call immediately rather than an email or a text message. Thus agents calling up their leads might have better success rate than the others.

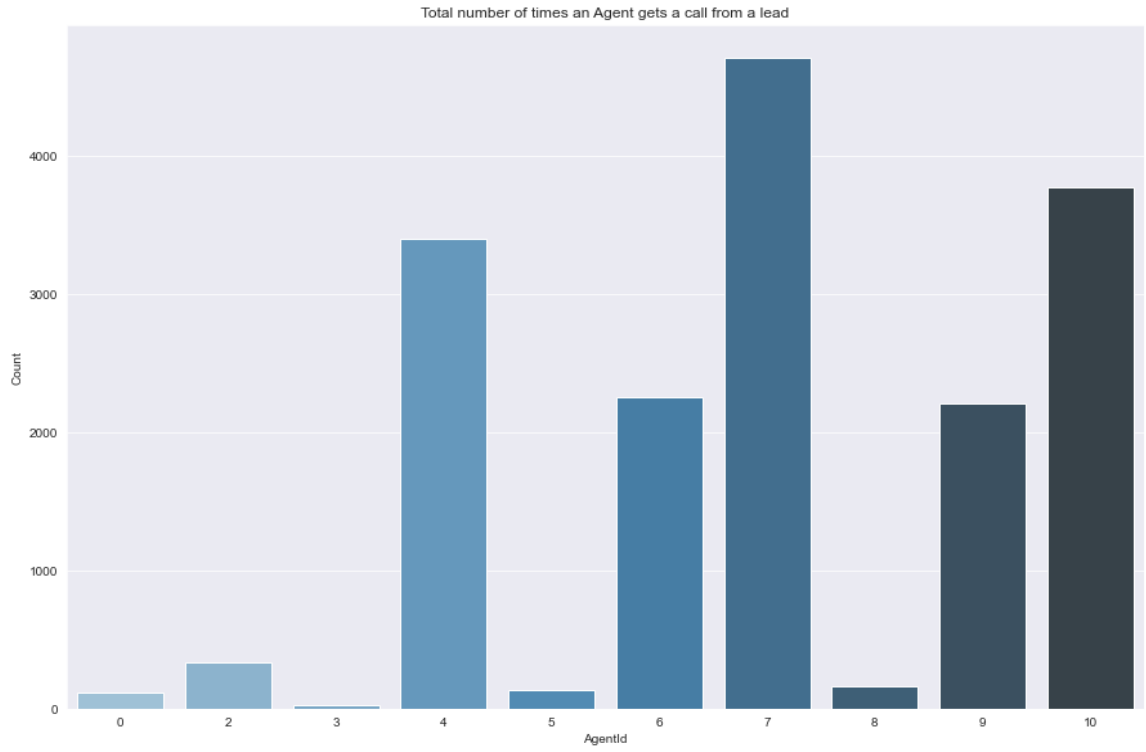

Distribution of Calls per agent
Count
agentId
0
2
3
4
5
6
7
8
9
10
0
200
400
600
800
1000


Distribution of Texts per agent
Count
agentId
0
2
3
4
5
6
7
8
9
10
0
1000
2000
3000
4000
5000
6000
7000


Distribution of Emails per agent
Count
agentId
0
2
3
4
5
6
7
8
9
10
0
2000
4000
6000
8000
10000

It is clear that Agent 4 spends a lot of time calling their leads and texting them whereas Agent 5 relies on waiting for replies on their mails.

Finally, if an agent receives a call from their lead, this means that the potential customer is interested in doing business with the brokerage firm and this shows a positive sign in agents performance.



Agent 7 gets a lot of calls from their leads and it might show that he is good at his job and converting leads to customers. On the other hand, agent 3 hardly receives any calls from their leads.

Scoring Criteria

For scoring the agents performances, I took the following into consideration :

- Number of follow ups per agent
- Number of leads per agent
- Median number of days delay in following up
- Number of times an agent gets a call from their leads
- Follow up type used by each agent

The idea behind the scoring criteria is to assign values from 0-10 based on each criteria individually where 0 means worst performance and 10 means best performance. After assigning ranks, I will total them up and find the top 3 agents having the most score.

For the median number of days delay in following up, a value of 10 would mean that the agent followed up quickly whereas a value of 0 would mean that agent took a long long time to follow up.

Also, since calls are most likely to be answered immediately by a lead, I assigned a value of 3 to calls and 2 to texts and 1 to emails as leads would take the most time to reply to an email.

Conclusion

Upon evaluating the agents based on the scoring index described above, it was found that Agent 4, Agent 10 and Agent 9 were the top 3 agents in that order. The following is a screenshot of the scoring values for each agent :

AgentID	Number of Follow Ups	Number of Leads	Median Delay	Incoming Calls	Total of Calls/Mails/Texts	Total
4	10.0	10.0	3.0	8.0	10.0	82.0
10	9.0	9.0	4.0	9.0	9.0	80.0
9	7.0	8.0	5.0	6.0	8.0	68.0
5	8.0	5.0	10.0	3.0	7.0	66.0
2	4.0	7.0	8.0	5.0	4.0	56.0
6	5.0	3.0	7.0	7.0	5.0	54.0
7	6.0	4.0	1.0	10.0	6.0	54.0
8	3.0	6.0	9.0	4.0	3.0	50.0
0	1.0	1.0	6.0	2.0	1.0	22.0
3	2.0	2.0	2.0	1.0	2.0	18.0

Suggestions

This was a quick analysis done on the dataset provided. But some of the following recommendations can be considered to improve:

- Get a dataset containing details whether or not a lead purchased a property and along with the cost of the property. This could help us understand how much sales an agent is bringing into the brokerage.

- To improve performance of other agents, the number of follow ups along with the number of distinct leads per agent should be increased as currently a lot of imbalanced data is present and currently Agent 4 has the most number of leads and consequently has the most number of follow ups.

- Another suggestion to improve performance would be to call the leads up rather than mailing them. Calling is effective as Agent 4 calls the most number of leads and has the highest performance score. Agent 5 should switch from mailing to calling as this could boost up their performance score a lot.

- Finally, agents should not take a long time to get back to their leads. Potential customers want agents focus on them and pay attention to them. As a result, if an agent takes a lot of time to get back to their leads, the leads would feel unsatisfied with the agents and the brokerage firm and might end up leaving and not doing business thus decreasing profits for the firm.