Antarctica Global – Data Analyst Assignment

MV AMOGH | Code Appendix Via Github Repository |

Date of submission – 16/07/2025

DATA SUMMARY and INITIAL OBSERVATIONS

(a) The **dataset** included the daily performance logs for each associate with the following –

- day and date (**datetime**),
- associate_name (**varchar**),
- leads generated (**int**)
- time_spent (**int**)
- time_per_lead (**int**)
- daily_team_review (**varchar**)
- incomplete_leads (**int**)

(b) I used **Power BI** for the initial data cleanup.

(c) The data **was split** into 3 tables, each table **including data** of one associate, I merged all 3 associates data into one table and **removed blank rows and errors**.

(d) The data was analysed using a combination of **SQL** and **Python (**pandas,matplotlib,sklearn,numpy)

BUSINESS QUESTIONS (SQL , Python – *Appendix to the code)*

I used SQL for the majority part of the business questions, and python for visualisations. The code explanation is included on the github repository.

(a) Lead Generation Efficiency – *Ratio of the leads generated to the total time spent*

- Upon analysing the data, we see that Arya has the highest efficiency at 8.51 leads per 100 minutes
- We calculate the efficiency by dividing the total leads generated by the total time spent (x100 to express it as leads per 100 minutes).
- However, ratio is 0.09. So, Arya produced 0.09 leads every minute.
- The result was scaled to leads per 100 minutes to improve interpretability.
- This indicates Arya is generating more leads per minute than her peers.
- This suggests a strong output for time invested.

(b) Daily Performance Variability – *which associate is deviating more from their average*?

- We calculate this using the standard deviation function.
- Ali seems to be deviating highest compared to his peers, and is inconsistent.
- Arya is stable, evident when we look at Arya's efficiency

(c) Time Management Analysis – *determining correlation between time per lead and leads generated*

- This helps identify if spending more or less time per lead influences overall productivity
- The correlation was found **python (pandas library)**
- The higher the number, the direct the relationship between time and leads
- **Raj** has a correlation at **-0.334**. That means the less time he spends per lead, the more leads he generates.
- Arya is on the same boat. Working faster leads to more output, in both their case.

(d) Impact of Daily Team Reviews – *is performance affected by missing team reviews?*

- Average leads generated were compared between days when the **review** was **attended and missed**
- **Arya never missed** a single day of team review. **Raj** showed a 3.9% decrease in performance upon missing team reviews

- **Ali** showed a **7.9%** decrease in performance, suggesting he benefits more from attending team reviews.

(e) **Incomplete Leads Trend** – *using linear regression model*

- Python – **matplotlib** was used to build a **linear regression model.**
- Incomplete leads were analysed for each associate, and it was found that it is indeed on a downward trend.
- It signals adaptability and learning from past mistakes, and improved follow-up, better process understanding and developing experience over time.

(f) **Performance Consistency** – *Co-efficient of variation*

- I used MySQL to determine the **co-efficient of variation** by **dividing Deviation** and **Average Mean.**
- Overall, the CV for the 3 associates is in good figures **(<0.5)**
- It signals steady performance and consistency.

(g) **High Performance Days** - *identifying the top days of performance and the leads on those days*

- I used **window function (dense_rank)** on SQL to easily filter out the top 10% of days using **ranks**
- The average time spent by **Raj** is highest standing at **258 minutes** spent, **Ali** is at **135 minutes. Arya** has been the most efficient at **135 minutes**

(h) **Impact of Longer Lead Generation Time** - *optimal 'sweet spot' for time spent* in lead generation

- Time segmented was created into 4 buckets of time using **CASE statements,** where in **avg_leads** fell under the following – **0-60 min**, **60-120 min**, **120-180min** and **180+ min**.
- Upon analysing, I found that the optimal time where associates start generating leads without diminishing leads is between **60-120min**, beyond which they are able to maximize output.

(i) **Comparative Day Analysis** – *does day of the week influence lead generation?*

- I used SQL to determine average leads during particular times of the week – weekday, midweek and weekend.
- **Raj and Ali** performed best on **weekends**. **Arya** seems to be more energetic in the **middle of the week,** slightly dropping off by the end of the week.
- Individual productivity patterns vary and are subjective in nature, possibly influenced by their lifestyle and routines.

(j) **Predictive Analysis**– *predicting lead generation and analysing effectiveness using*

- A simple linear regression model was built using Python to predict the number of leads generated based on time spent on lead generation.
- The model was evaluated using the R^2 score from sklearn.metrics, which measure how effective is the predictive model.
- The **R^2 scores** are as follows for **Raj, Ali and Arya**, respectively, **0.413, 0.272 and 0.100**
- These weak R^2 scores indicate that time spent is a weak standalone predictor of productivity, other factors matter more.

DASHBOARDING(Appendix to the Python code used – *for heatmap and boxplot*)

The dashboard was created using **both Power BI and Python**. **Boxplot and heatmap** already were included in **Python libraries**, which made it easier. The other charts were simple were chosen to be done on **Power BI**.

(a) **Linechart –** *Attendance with leads generated by associate*

- **Arya** has the **highest attendance** (100%) among her peers and consistently performs well. There are no gaps or anomalies.
- **Raj** missed 2 days. However, it doesn't look like it impacted his performance much. He could generate a **good average amount** (11 leads) even on his missed day.

- Ali missed 1 day, he also managed to get 11 leads.
- They are either handing off the tasks to their colleagues or actually working pretty efficiently on their missed days.

(b) **Heatmap** – *Leads generated VS Time buckets*

- Raj and Ali's productivity increases as their time spent increases, with their max being >= 11 leads after >180 min spent.
- **Arya** seems to **hit a wall** after 180 minutes, might signal **fatigue and stress.**
- But, **almost everyone** performs well between time buckets– middle of 60-120 minutes and 120-180 minutes. So, between **90 – 150 minutes** is the optimal time for productivity.

(c) **Bar Chart** – *Monthly Leads Total*

- **July is the strongest month** across the whole chart, and **Arya** seems to be **the highest performer** on that specific month.
- **Arya's performance** seems to **not** be the **highest in June and August**. There might have been **a motivation** or a **mood boost** during month of **July for Arya**.
- During months of **June and August**, however, **Ali** seems to be **performing the best** among his peers.
- The **numbers** may **signal towards Arya being good**, maybe **because her average** is good and she is **consistent** with the number range.

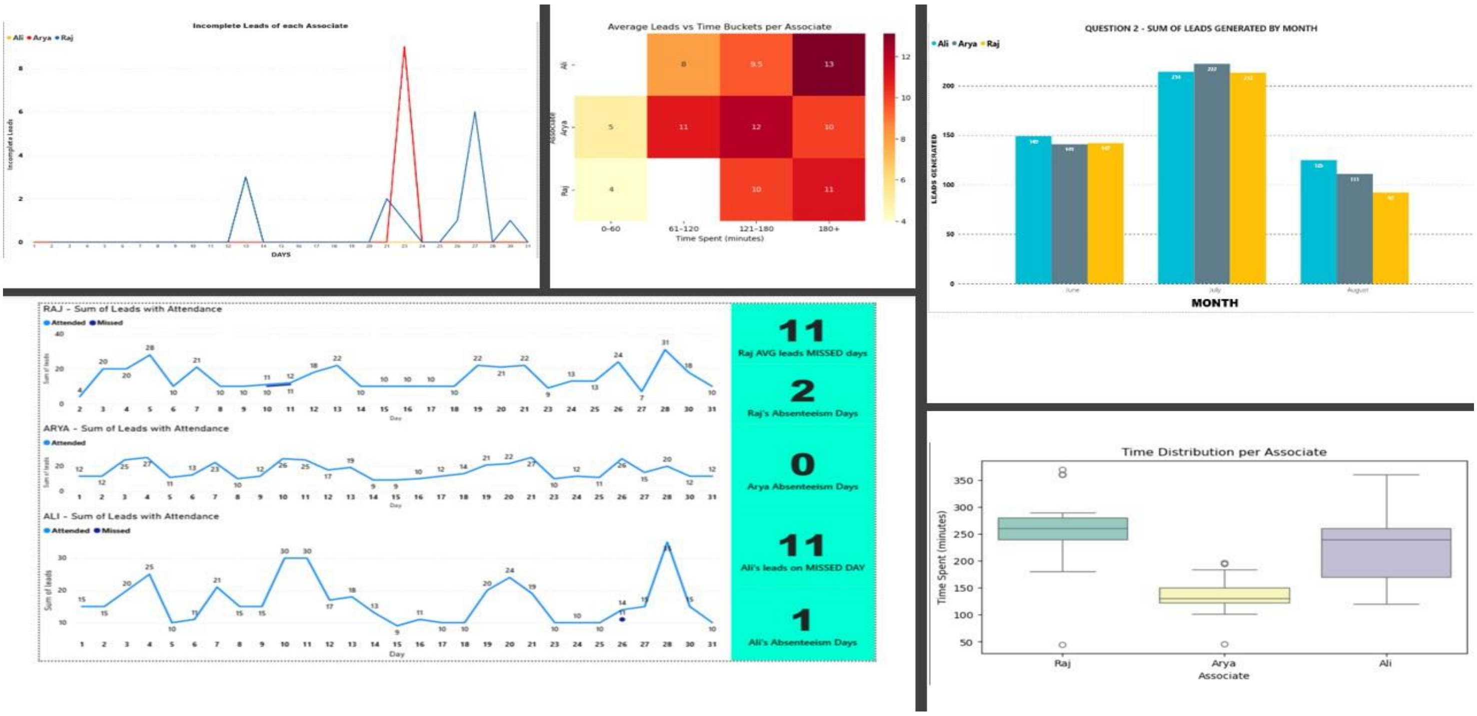
(d) **Trend Line** - Incomplete Leads trend

- **Ali** has **0 spikes**, **no incomplete** leads at all. The numbers **may say otherwise**, in **favour of Arya**, maybe because the leads she **gets in-hand** is already at a **high number**, it **impacts her average** conversion rate.
- **Ali** raises **red flags** as well, pertaining to what **exactly is Ali** doing with the leads. Is he providing **unauthorized discounts**, **coercing** leads?
- **Arya** and **Raj** have had their share of spikes in incomplete leads, but they bounced back. It signals **adaptability and learnability.**

(e) **Box Plot** – *Time distribution per associate*

- **Raj** seems to be **consistent** with his **time usage**, hovering near the top end.
- **Arya** also seems to be time-efficient. Narrow spread, low time and steady output.
- **Ali** is **irregular**, the time frame **gap is huge** and his line is **wider.**

DASHBOARD (STATIC, made using Power BI and Python Seaborn package)



KEY INSIGHTS THROUGH THE DASHBOARD AND NUMBERS

- **Arya** has **perfect attendance** and a relatively **smooth** lead generation **curve.** She is consistent and a reliable performer.
- **Ali** has **0 incomplete leads**, this is **unrealistic**. Even **Arya or Raj**, who are **strong performers** have had **spikes** in incomplete leads.
- **Ali** missed 1 day, still managed to have **0 incomplete leads** raises red flags. **Handing off tasks**, **pre-filling** values on 'conversion sheet' or something similar might be the case for **Ali.**
- **Ali's standard deviation** in leads is the highest (-0.3), meaning his **performance is volatile**. If that is true, then **how** is his **conversion rate perfect** (0 incomplete leads)?
- Even the **boxplot** shows **Ali's time usage wider** than normal – the **widest** of all three. Again, this **doesn't sit right** when he has **perfect conversion scores**
- **Lastly,** his **weekday vs midweek vs weekend** is exactly the same, **unlike others. Either** he is **exceptionally disciplined** (rare case) or he is raising an **actual red flag**

BUSINESS RECOMMENDATIONS

- Considering **Ali's suspicious numbers**, a **manual examination** of **Ali's login** and **logout** or a **complete audit** of his work can be conducted to verify his activity and lead generation method.
- The **optimal time spent** for leads is **90-180 minutes**, beyond **which fatigue sets** in. **Not encouraging** long hours and **prioritizing health** should be ideal to **retain the associates.**
- **Ali and Raj** generated 11 leads on **missed days**. Does it correlate with their previous performance records? However, **missing team reviews** should not be encouraged.
- **Ali** must develop consistency in time and leads generated. His **boxplot shows a very wide timeframe** compared to his peers. A time **consistency session can be conducted**, not just for Ali but for **all associates**.
- **Arya** balances **time spent and lead generated**. Not to mention, she also has **perfect attendance**. She could be **used as an example during onboarding or training sessions**. But, need to lessen the tight working hours to prevent **mental fatigue for Arya.**
- **Spikes** in **incomplete leads** should be **tracked regularly**. An **automated incomplete leads** report could be **created via Python** that connects to the **main database**. Any **sudden spikes** should be **flagged and intervened immediately.**
- **July** was a **significant month** for **performance and productivity** among associates. Why is that? Were there any **community engagement sessions**, **new policy**, **new incentive**, **leadership activity**? The **factors** must be analyzed