

Dear Candidate,

Thank you for your interest in joining our team as a Data Intern. We are looking for curious, analytical, and detail-oriented individuals who are eager to learn and contribute to meaningful projects in the exciting world of AI-based mobile applications. As part of our data team, you'll have the opportunity to work with real product data, support decision-making processes, and gain hands-on experience in a fast-paced tech environment.

## Question 1: User Acquisition and Monetization Analysis

You work at a mobile app company that tracks daily new user installs, ad spend, and revenue by platform and channel. The data file `Question1_Analysis.csv` contains three months of daily metrics with columns: **date**, **platform** (iOS or Android), **channel** (e.g. Social, Search, Email, Referral), **new\_installs**, **ad\_cost**, and **revenue**. Using Python perform the following analysis:

- **Derived Metrics:** For each platform, calculate ARPU (average revenue per user). For each channel, calculate ROI (return on ad spend) and CAC (Customer Acquisition Cost). Which platform has the higher ARPU? Which channel has the highest ROI? Which channel has the lowest CAC?
- **Time-Series Trend:** Group the data by month (extract the month from the date) and platform to compute monthly installs. Plot a line chart of monthly installs for each platform on the same axes (X-axis: month, Y-axis: installs). What trend do you observe over time for iOS vs Android installs?
- **Visualization:** Create a grouped bar chart to visualize ROI by channel for each month. The X-axis should represent months, and within each month, bars should represent different marketing channels (distinguished by color). The Y-axis should show ROI values. Based on this chart, identify which channels are the most cost-effective over time and highlight any monthly trends or anomalies.
- **Interpretation:** Based on your analysis and the charts, what patterns or differences stand out (e.g. between iOS vs Android or among channels)? What might be the causes of these patterns? Based on your findings, what recommendations would you give to the marketing or growth team to improve user acquisition or monetization efficiency?

Use clear and well-commented code for data preparation, and use Matplotlib or Seaborn for visualization. Write clean code with comments explaining each step, and feel free to extend the analysis with additional insights if you'd like.

## Question 2: Reasoning

In this question, we're primarily interested in understanding your analytical judgement. Don't hesitate to think out loud and share your reasoning. Use concrete examples from the `Question1_Analysis.csv` dataset (e.g., daily installs, ad cost, revenue split by platform and channel) to support your insights.

- Suppose your Question 1 charts showed that **Email** delivers high ROI but very low absolute installs. Describe a **data-driven experiment** you would run to test whether scaling Email spend is worthwhile. Include success metric(s) and guard-rails.
- Name **two additional raw metrics** (not in `Question1_Analysis.csv`) that you would most want to collect next. Sketch a project or analysis that would leverage it (e.g. creative-level optimisation) and outline the business value.

- Let's assume that on Mid-March, you notice a sharp drop in iOS revenue across all channels, but install volume remains steady. List a few specific metrics or events you would examine (e.g., conversion rates, in-app purchase activity, pricing changes, user locations, etc.) and explain why they might be relevant.

### Question 3: Campaign Performance Modeling

You have been given a CSV file `Question3_CampaignData.csv` containing summary data for 50 marketing campaigns. Columns include **campaign\_id**, **channel** (e.g. Email, Social, Search, Display), **impressions**, **clicks**, **conversions**, **spend**, and **revenue**. The goal is to build a model that predicts whether a campaign will be **profitable** (i.e. generate more revenue than it costs).

#### Step 1: SQL Feature Extraction.

Write a SQL-style query (as a text code block) to calculate the following features from `Question3_CampaignData`:

- **CTR**
- **Conversion Rate**
- **ROI**
- **Is Profitable** -> True, if revenue is greater than spend

*Include this query in your answer as a code block.*

#### Step 2: Model Building in Python.

Load `Question3_CampaignData.csv` into Python (e.g., with pandas). Use the features from your SQL query.

Split the data into a training set and test set. Build a basic classification model to predict `is_profitable`.

Report model performance on the test set using **accuracy**, **precision**, and **recall**. (You can use scikit-learn or similar libraries.)

- Load `Question3_CampaignData.csv` into Python.. Use the features from your SQL query.
- Split the data into a training set and test set. Build a basic classification model to predict `is_profitable`.
- Report model performance on the test set using **accuracy**, **precision**, and **recall**. (You can use scikit-learn or similar libraries.)

**Interpretation:** Explain your choice of evaluation metric for this business problem. For example, if profitable campaigns are rare, precision/recall might be more relevant than accuracy. Discuss whether there is any sign of overfitting (compare train vs test performance or comment on model complexity). Finally, interpret the results: what would you do with this model in a business context?

The goal here is *not* to optimize for a perfect model, but to demonstrate your ability to reason about machine learning metrics and business impact. what matters more is how clearly you explain the tradeoffs and risks. Focus more on **how you evaluate model performance**, **your metric choice**, and **what the results mean for the business**.

Please complete all sections and submit your responses within **two days**. Good luck, and we look forward to your insights!

**Links:**

Question 1 Analysis [Drive Link](#)

Question 3 [Campaign Data Link](#)

Any issues with the links, don't hesitate to reach out to us.