# **Data Science Screening**

Given the specific objective of forecasting the earnings for a publicly traded company with stores in malls, along with the theoretical dataset, complete the following questionnaire. This should not take more than 30 minutes.

# **Question 1**

You have access to the following data sources:

* A table with the prices of all store items (date, storeID, productID, description, price) going back 2 years.
* Hourly satellite images of each mall parking lot (going back 1 year).
* Email receipts for online purchases, collected from 10,000 inboxes (with some inboxes missing company receipts) over the course of one year.
* Daily counts of customers who visited the store next door in the same mall (going back 2 years, with 60% missing data).
* A list of all other stores in the same malls as the company in question.

Describe your approach to preprocessing this data for forecasting the company's earnings. Consider how you would integrate all these data sources and what feature engineering steps you might take.

# **Question 2**

In the dataset provided, you find that the satellite image dataset has 75% missing instances. How would you handle this missing data specifically for the earnings forecast? Explain your strategy.

# **Question 3**

Considering the nature of the forecasting task, what three machine learning models would you select to test for forecasting the company's earnings, and why? Discuss the suitability of these models for time-series data, incorporating various data sources, and dealing with seasonality.

# **Question 4**

In Question 3, you selected three models for forecasting earnings. How would you compromise between bias and variance during the model tuning process, considering the unique challenges posed by this dataset and forecasting objective? Describe the trade-offs and considerations specific to this task.

# **Question 5**

You now have two new raw data sources:

* Social media sentiment data related to the company's products.
* Economic indicators for the regions where the malls are located.

Explain how you would change your approach to incorporate these new data sources into your analysis and forecasting models. Describe the steps you would take to integrate these additional data sources and discuss how they can improve the accuracy of the earnings forecast.