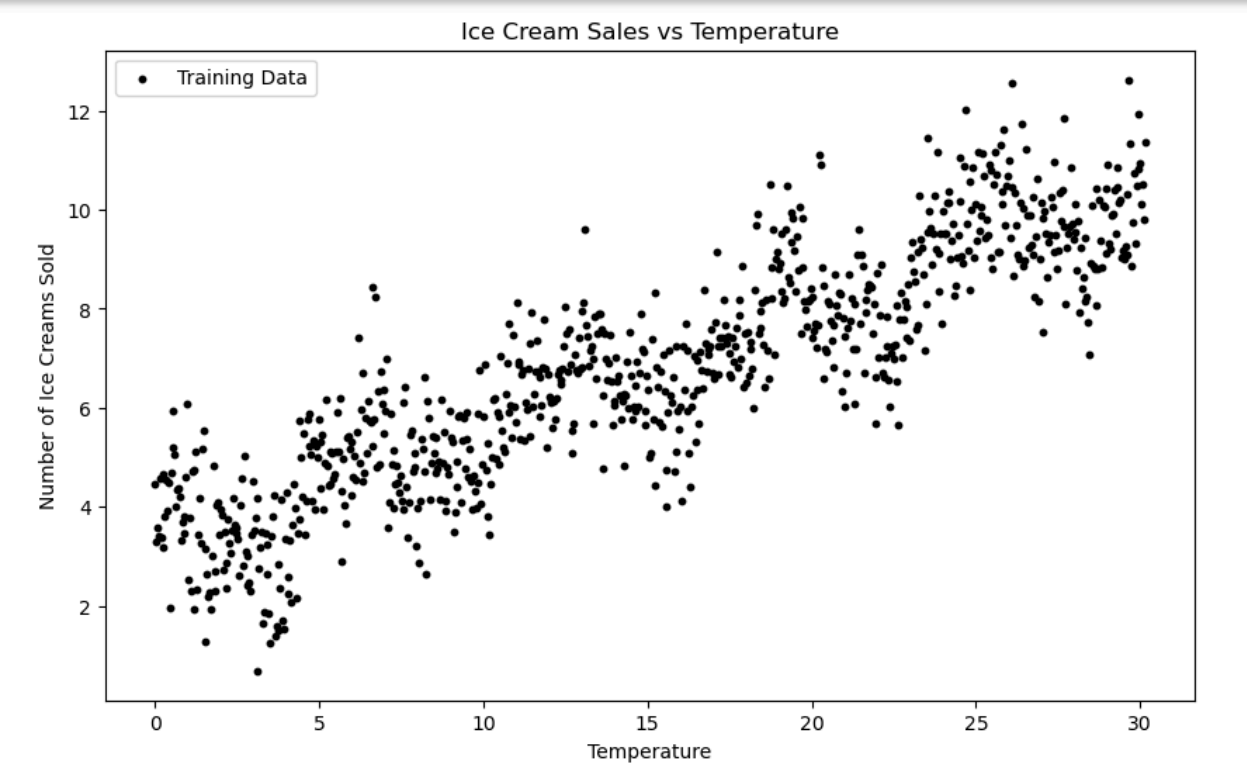
# 1. Introduction

- This report analyzes the relationship between temperature and ice cream sales based on data provided in `train.txt` and `test.txt`. Using various statistical and machine learning techniques, we build and evaluate models to predict ice cream sales based on temperature. The objective is to identify the best predictive model using both linear and non-linear approaches, including feature engineering and regularization techniques (Lasso and Ridge).

**Data Description**:

- The data is sourced from `train.txt`, and `test.txt` containing two columns: temperature (independent variable) and ice cream sales (dependent variable). We aim to find the best linear and non-linear models, including Lasso and Ridge regression, to maximize prediction accuracy on sales.

## 2. Initial Data Analysis and Plotting




Data plotting

- Figure above: Temperature vs Ice cream sales
- The plot suggests a positive trend between temperature and sales. This relationship will be explored through linear and non linear modeling techniques.

## 3. Linear Regression Model

- Model description
  - I start by fitting linear regression model y=β0+β1x to understand baseline relationship between temperature and ice cream sales (code for it)

- Results and Interpretation
  - Coefficients:
    - Estimated β̂0 (Intercept): 3.191037800253055
      Estimated β̂1 (Slope): 0.23839763045936505
  - Statistical Inference:
    - The OLS summary (Table 1) reveals a statistically significant positive relationship between temperature and ice cream sales, as both the intercept (β̂0) and slope (β̂1) have p-values less than 0.05, indicating they are significantly different from zero.

      This table below confirms the linear relationship, supporting the hypothesis that temperature positively influences ice cream sales.

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 3.1910 | 0.078 | 40.962 | 0.000 | 3.038 | 3.344 |
| x1 | 0.2384 | 0.004 | 53.289 | 0.000 | 0.230 | 0.247 |

## 4. Feature Engineering and Model Selection with Non-linear Features

- Feature Selection Approach:
  - To improve the model beyond a simple linear fit, I included non-linear transformations of temperature, such as cos(x), log(x), cos(4x), sin(3x), sin(5x), and sin(2x)×cos(2x). Given that the feature set was relatively small, I evaluated all possible combinations of these features to identify the best subset.

- Metric Used:
  - Adjusted $R$^2 was employed as the metric for feature selection. Unlike standard $R$^2 , adjusted $R$^2 accounts for the number of predictors in the model, ensuring that added features improve the model's fit without simply increasing complexity.
- **Selection Process and Results:**
  - The exhaustive feature selection process revealed that the optimal feature combination is: x,cos(x),sin(3x) with an adjusted R^2 of 0.852. This combination showed the highest adjusted R^2. (screenshot with results below)

```
Best feature combination: ('x', 'cos_x', 'sin_3x')
Highest Adjusted R^2: 0.852448879534917
Model with highest Adjusted R^2 uses features: ('x', 'cos_x', 'sin_3x')
```