Final Exam - EECS 398, Spring 2025

Full Name:

Uniqname:

UMID:

Starting Time: ◯ 1:30PM (standard) ◯ 10:30AM (alternate)

Instructions:

- You have 120 minutes to complete this exam.
- This exam consists of 8 questions, worth a total of 90 points.
- Write your uniqname in the top right corner of each page in the space provided.
- Please write **clearly** in the provided answer boxes; we will not grade work that appears elsewhere. Completely fill in bubbles and square boxes; if we cannot tell which option(s) you selected, you may lose points.

  ◯ A bubble means that you should only **select one choice**.

  ☐ A square box means you should **select all that apply**.
- You may refer to a **single** two-sided handwritten notes sheet. Other than that, you may not refer to any other resources or technology during the exam (no phones, watches, or calculators).

You are to abide by the University of Michigan/Engineering Honor Code. To receive a grade, please sign below to signify that you have kept the honor code pledge.

I have neither given nor received aid on this exam, nor have I concealed any violations of the Honor Code.

Signature:

Data Overview: Coffee Shops in Ann Arbor

In this exam, we'll work with the DataFrame `shops`, which contains information about coffee shops in Ann Arbor.

The first few rows of `shops` are shown below, but `shops` has many more rows than are shown.

	name	neighborhood	daily_sales	rating	seats	open_late
0	Zingerman's Coffee Company	Kerrytown	3450	4.6	40	1
1	RoosRoast Coffee	South University	2100	4.2	25	1
2	Literati Coffee	Downtown	1750	4.8	30	0
3	Comet Coffee	Downtown	2900	4.5	20	0
4	Sweetwaters Coffee & Tea	Old West Side	1200	4.1	15	0
5	Sweetwaters Coffee & Tea	Plymouth Road	1300	4.2	18	0
6	The Espresso Shop	Downtown	4200	4.7	50	1

Each row in `shops` contains information about a single coffee shop. The columns in `shops` are as follows:

- `"name" (str)`: The name of the coffee shop. Note that names are not unique, since some chains (like Sweetwaters Coffee & Tea) have multiple locations.
- `"neighborhood" (str)`: The neighborhood the coffee shop is in.
- `"daily_sales" (int)`: The coffee shop's average daily sales, in dollars.
- `"rating" (int)`: The coffee shop's rating on Google Maps, between 1 and 5.
- `"seats" (int)`: The number of seats the coffee shop has for customers.
- `"open_late" (int)`: `1` if the coffee shop is open late (after 8PM), `0` otherwise.

Throughout the exam, assume we have already run all necessary import statements.

uniqname: _____

Question 1 (15 pts)

Suppose we'd like to predict the rating of a coffee shop given various features. To do so, we build several different linear regression models, each with intercept terms but slightly different sets of features.

In parts (a) and (b), consider the expression and output below.

```
>>> shops["neighborhood"].value_counts()
Plymouth Road       40
Downtown            30
Old West Side       20
Kerrytown           5
South University    5
```

a) (3 pts) First, we build **Model A**, which:

- Uses `OneHotEncoder(drop="first")` on neighborhood.
- Uses `PolynomialFeatures(degree=2, include_bias=False)` on daily sales.
- Uses seats as-is.

How many columns are there in the design matrix created by Model A, **including** the column of all 1s? Give your answer as an integer.

b) (3 pts) Next, we build **Model B**, which:

- Uses `OneHotEncoder(drop="first")` on neighborhood, **then** `StandardScaler()` on each of the resulting one hot encoded columns.
- Uses daily sales as-is.

In the design matrix created by Model B, the column corresponding to one of the one hot encoded neighbors, after standardization, contains the unique values $\frac{\sqrt{6}}{2}$ and $-\frac{\sqrt{6}}{3}$. Which neighborhood does this column correspond to?

◯ Plymouth Road ◯ Downtown ◯ Old West Side ◯ Kerrytown

Hint: The standard deviation of a sequence of 0s and 1s is $\sqrt{p(1-p)}$, where p is the proportion of 1s. It is possible to answer the question without using this fact, but it may come in handy.

Models A and B from the previous page are repeated again below, along with two other models.

- **Model A**:
  - Uses `OneHotEncoder(drop="first")` on neighborhood.
  - Uses `PolynomialFeatures(degree=2, include_bias=False)` on daily sales.
  - Uses seats as-is.
- **Model B**:
  - Uses `OneHotEncoder(drop="first")` on neighborhood, **then** `StandardScaler()` on each of the resulting one hot encoded columns.
  - Uses daily sales as-is.
- **Model C**:
  - Uses `OneHotEncoder()` on neighborhood and name, separately.
  - Uses daily sales and number of seats as-is, and no other features.
- **Model D**:
  - Uses `OneHotEncoder()` on neighborhood.
  - Uses daily sales as-is.

We train all four models on the same training set by minimizing mean squared error (MSE). Let $T(A)$ be the **training** MSE of Model A, $T(B)$ be the training MSE of Model B, and so on.

c) (9 pts) In each subpart below, fill in the blank with the inequality that describes the relationship between the pair of values. If it is impossible to guarantee the relationship between the values provided just by using the information above, select "N/A".

(i) $T(A)$ ____ $T(B)$ ◯ $\geq$ ◯ $=$ ◯ $\leq$ ◯ N/A

(ii) $T(A)$ ____ $T(C)$ ◯ $\geq$ ◯ $=$ ◯ $\leq$ ◯ N/A

(iii) $T(A)$ ____ $T(D)$ ◯ $\geq$ ◯ $=$ ◯ $\leq$ ◯ N/A

(iv) $T(B)$ ____ $T(C)$ ◯ $\geq$ ◯ $=$ ◯ $\leq$ ◯ N/A

(v) $T(B)$ ____ $T(D)$ ◯ $\geq$ ◯ $=$ ◯ $\leq$ ◯ N/A

(vi) $T(C)$ ____ $T(D)$ ◯ $\geq$ ◯ $=$ ◯ $\leq$ ◯ N/A

uniqname: _____

Question 2 (13 pts)

Suppose we'd like to fit a linear regression model with an intercept term by minimizing mean squared error, using the full rank design matrix $X \in \mathbb{R}^{n \times (d+1)}$ and observation vector $\vec{y} \in \mathbb{R}^n$.

Consider the matrix $\boxed{F = X(X^T X)^{-1} X^T}$.

a) (3 pts) How many rows and columns does F have? Give both answers as expressions involving n, d, and/or constants.

rows in $F =$ [ ] # columns in $F =$ [ ]

In each of the following parts, select one of following options:

1. The design matrix
2. A vector orthogonal to the span of the design matrix
3. The optimal hypothesis vector, $\vec{h}^*$
4. The optimal parameter vector, $\vec{w}^*$
5. A matrix, vector, or scalar that is none of these options
6. An invalid operation

b) (2 pts) What is XF?

◯ 1 ◯ 2 ◯ 3 ◯ 4 ◯ 5 ◯ 6

c) (2 pts) What is FX?

◯ 1 ◯ 2 ◯ 3 ◯ 4 ◯ 5 ◯ 6

d) (2 pts) What is $F\vec{y}$?

◯ 1 ◯ 2 ◯ 3 ◯ 4 ◯ 5 ◯ 6

e) (2 pts) What is $(I - F)\vec{y}$, where I is a square identity matrix with the same dimensions as F?

◯ 1 ◯ 2 ◯ 3 ◯ 4 ◯ 5 ◯ 6

f) (2 pts) What is $((I - F)\vec{y}) \cdot \vec{1}_n$, where $\vec{1}_n$ is a vector in $\mathbb{R}^n$ containing all 1s?

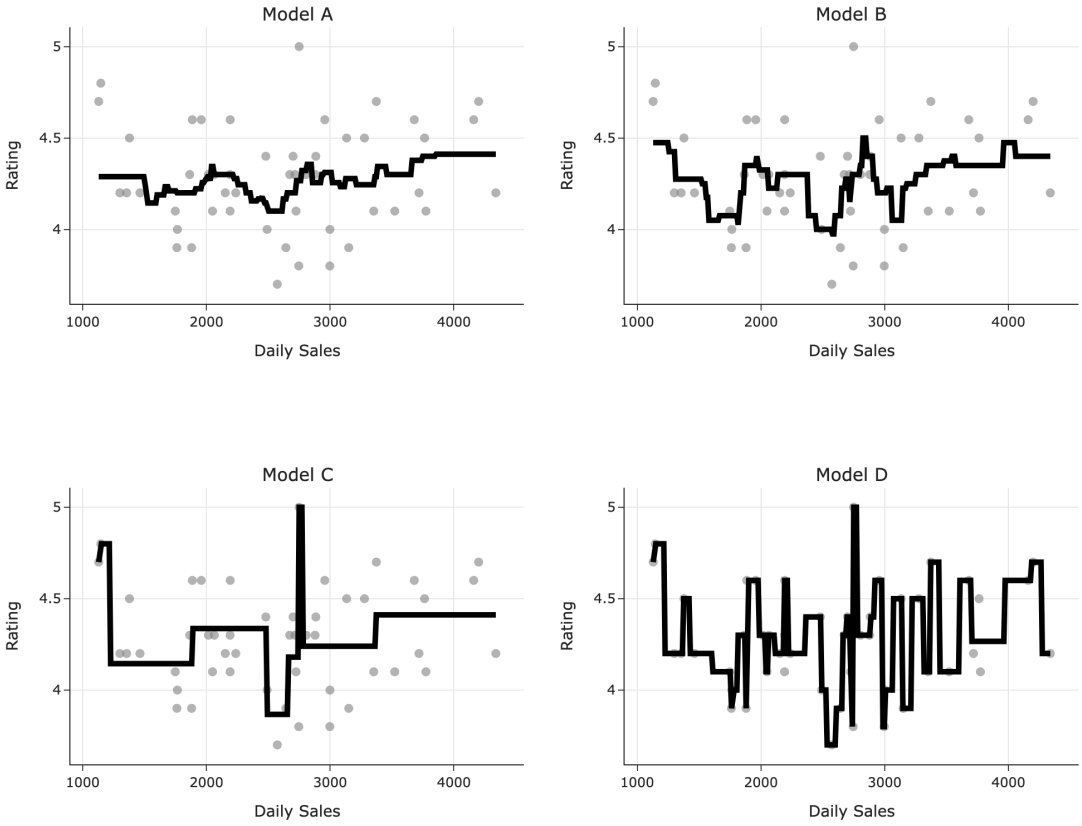
◯ 1 ◯ 2 ◯ 3 ◯ 4 ◯ 5 ◯ 6

Question 3 (8 pts)

Like k-nearest neighbors, decision trees can be used for regression as well as classification.

To predict the y_i-value for an input $\vec{x}_i$, a decision tree regressor asks up to d yes/no questions (where d is the tree's maximum depth), and predicts the **mean** y_i-value among all points in the training set with the same sequence of answers.

Below, we visualize the predictions of four different models that predict a coffee shop's rating given its daily sales.



a) (2 pts) Which model is a decision tree regressor with a maximum depth of 4?

◯ A ◯ B ◯ C ◯ D

b) (2 pts) Which model is a k-nearest neighbors regressor with $k = 4$?

◯ A ◯ B ◯ C ◯ D

c) (2 pts) Which model is a decision tree regressor with a maximum depth of 9?

◯ A ◯ B ◯ C ◯ D

d) (2 pts) Which model is a k-nearest neighbors regressor with $k = 9$?

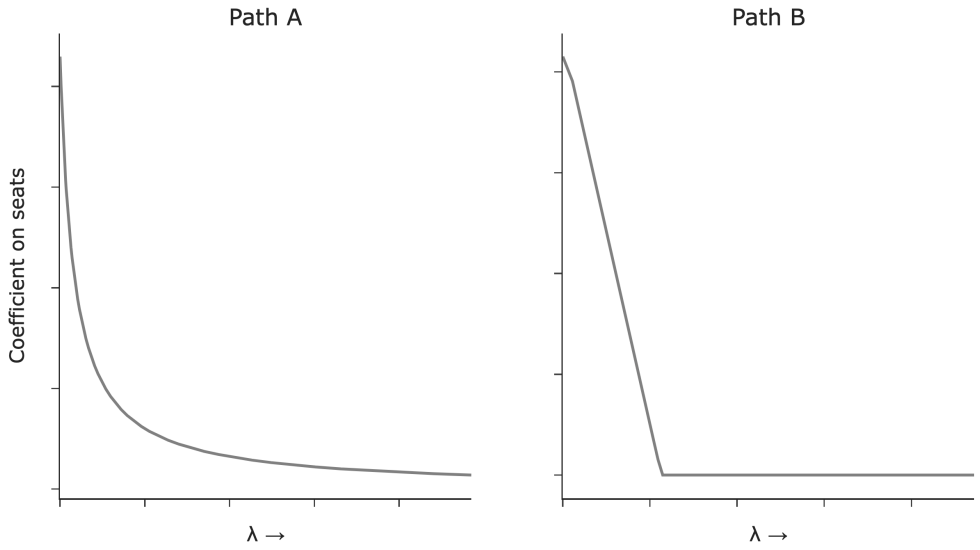
◯ A ◯ B ◯ C ◯ D

uniqname: _____

Question 4 (8 pts)

Suppose we'd like to fit a linear regression model with an intercept term to predict a coffee shop's rating given various features, including its number of seats.

a) (4 pts) First, we consider ridge regression and LASSO regression. The two plots below show the "path" of the optimal coefficient on seats, w^*_{seats}, as the regularization hyperparameter, λ, increases.



One plot above corresponds to ridge regression, while the other corresponds to LASSO regression. Which plot corresponds to **ridge regression**? Select an answer, and justify your answer in two English sentences.

◯ Path A corresponds to ridge regression

◯ Path B corresponds to ridge regression

b) (4 pts) Now, suppose we'd like to fit the model using L_2-regularized mean **absolute** error. In other words, we'd like to minimize the objective function:

$$R(\vec{w}) = \frac{1}{n}\sum_{i=1}^{n} |y_i - \text{Aug}(\vec{x}_i) \cdot \vec{w}| + \lambda \sum_{j=1}^{d} w_j^2$$

where $\vec{x}_i \in \mathbb{R}^d$ is a feature vector, $\vec{w} \in \mathbb{R}^{d+1}$ is a parameter vector, and λ is a non-negative regularization penalty hyperparameter.

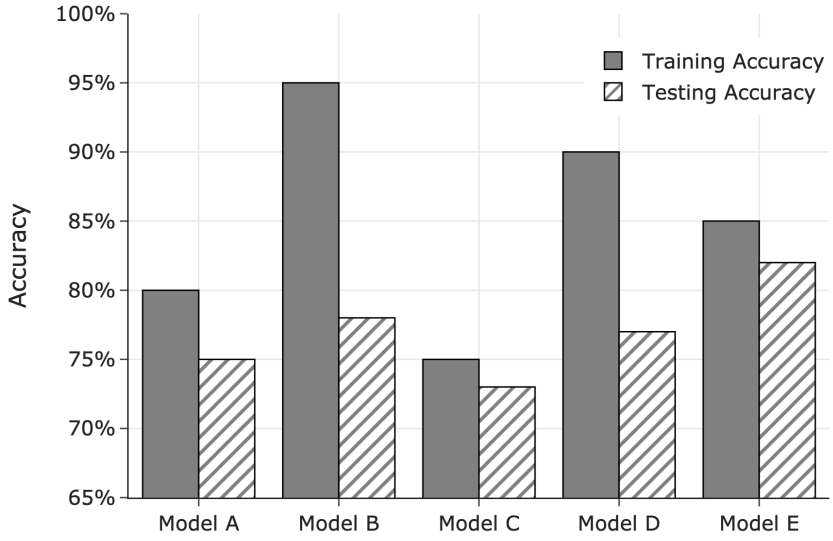
As $\lambda \to \infty$, what do the predictions of the resulting model approach? In two English sentences, describe the nature of the resulting predictions and briefly justify your answer.

uniqname: _____

Question 5 (16 pts)

Suppose we build 5 different classifiers, named Model A through Model E, to predict whether a coffee shop is open late (class 1) or not (class 0).

The performance of all 5 models on both our training set and test set are given below.



First, we start by using **just** the information above to choose between the 5 models.

a) (2 pts) Which model has the highest model bias?

◯ A ◯ B ◯ C ◯ D ◯ E

b) (2 pts) Which model has the highest model complexity?

◯ A ◯ B ◯ C ◯ D ◯ E

c) (2 pts) If the only information we have access to is the plot above, which model would we select to make predictions moving forward?

◯ A ◯ B ◯ C ◯ D ◯ E

Now, suppose we instead use k-fold cross-validation on our training set to choose between the 5 models.

d) (2 pts) After using k-fold cross-validation, is the model we select guaranteed to be the same as the model you selected in part (c)? Select the **best** answer.

◯ Yes, the selected model is guaranteed to be the same

◯ No, but the selected model is likely to be the same

◯ No, and the selected model is most likely different

◯ No, and the selected model is guaranteed to be different

e) (4 pts) Abhi claims that **it's possible** that the model that k-fold cross-validation selects **never** has the highest validation accuracy in any one fold. Is he correct?

- If the answer is **Yes**, select the **Yes** bubble **and** fill in the grid below with 10 example accuracies proving him correct, using $k = 2$ as an example.
- If the answer is **No**, select the **No** bubble and fill in the response box beneath it with a brief English explanation as to why this is not possible.

◯ Yes

	Model A	**Model B**	**Model C**	**Model D**	**Model E**
Fold 1					
Fold 2					

*Write in this grid **only** if you selected **Yes**.*

◯ No

*Write in this box **only** if you selected **No**.*

f) (4 pts) Suppose we perform **10**-fold cross-validation to choose between the **5** models. Each time a model is trained, **20** rows are used for validation.

Each time a model is trained, we **write down the number of rows used to train that specific model** (that is, not including the validation fold). Only one number is written down each time a model is trained.

In both boxes below, give your answers as integers.

(i) **How many** numbers are written down in total?

(ii) What is the **sum** of all numbers written down?

uniqname: _____

Question 6 (13 pts)

Suppose we use logistic regression to predict the probability that a coffee shop is open late (class 1) or not (class 0), given various features. Our model's predicted probabilities on a 12-row test set are given below, along with the true classes for each coffee shop.

True Class	0	0	0	1	0	???	1	0	1	1	1	0
$\boldsymbol{P(y_i = 1 \mid \vec{x}_i)}$	0.19	0.24	0.25	0.33	0.55	0.55	0.55	0.62	0.62	0.85	0.85	0.90

The true class for one coffee shop is unknown, and is marked by ???.

In order to convert predicted probabilities into predicted classes, we apply a threshold T, where $0 \leq T \leq 1$. Once we've chosen a threshold, if $P(y_i = 1|\vec{x}_i) \geq T$, we predict 1; otherwise, we predict 0.

a) (6 pts) Suppose we choose a threshold of $T = 0.53$.

In this part, give both answers as fractions or decimals.

(i) If the precision of the resulting predictions is 1/2, what is the recall of the resulting predictions?

(ii) The false negative rate (FNR) of a binary classifier is the proportion **of** truly positive data points **that** are incorrectly classified as negative.

If the false negative rate of the resulting predictions is 1/6, what is the accuracy of the resulting predictions?

b) (3 pts) Suppose we choose a threshold of $T = 0.4$.

Among all 12 points in the test set, what is the value of the **largest** cross-entropy loss incurred by any one point? Give your answer as an expression involving the log function and/or constants.

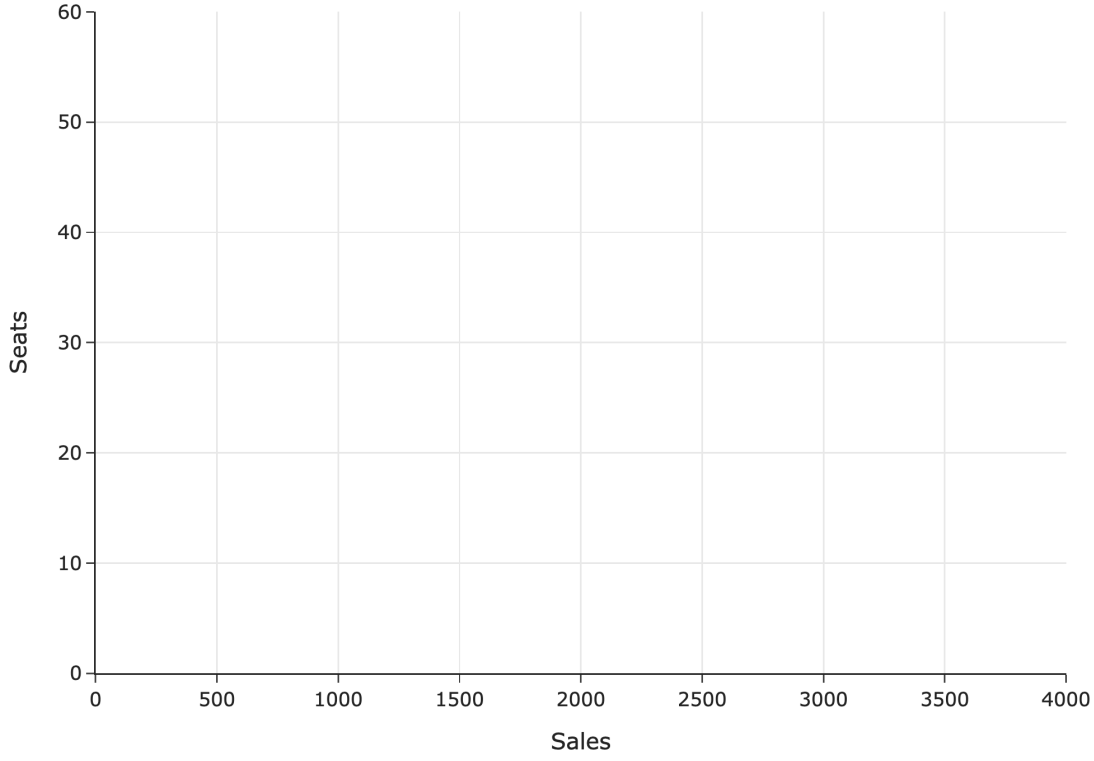
c) (4 pts) Suppose that our logistic regression model uses a coffee shop's daily sales and number of seats to predict the probability that it is open late. In other words:

$$P(y_i = 1|\vec{x}_i) = \sigma(w_0 + w_1 \cdot \text{sales}_i + w_2 \cdot \text{seats}_i)$$

After minimizing mean cross-entropy loss (without regularization), the model's optimal parameter vector $\vec{w}^*$ is:

$$\vec{w}^* = \begin{bmatrix} 1500 \\ 2 \\ -150 \end{bmatrix}$$

Given a threshold of $T = 0.5$, **draw** the decision boundary for the resulting classifier on the axes below. Label enough points on the boundary to clearly show that it is placed correctly.



uniqname: _____

Question 7 (9 pts)

Let $\vec{u} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}$. Consider the function q, defined below.

$$q(\vec{u}) = (u_1 + u_2 + u_3)^2 + (u_1 - u_2)^2 + (u_2 - u_3)^2$$

We write code that implements gradient descent in order to minimize q, using some initial guess, $\vec{u}^{(0)}$, and learning rate/step size, π. In our code, we add print statements that show us the values of $\vec{u}^{(t)}$ and $\nabla q(\vec{u}^{(t)})$ (the gradient of q) after each iteration.

Here's what we see:

After 1 iteration, $\vec{u}^{(1)} = \begin{bmatrix} 0 \\ -1 \\ -3 \end{bmatrix}$, $\nabla q(\vec{u}^{(1)}) = \begin{bmatrix} -6 \\ -6 \\ -12 \end{bmatrix}$, $q(\vec{u}^{(1)}) = 21$

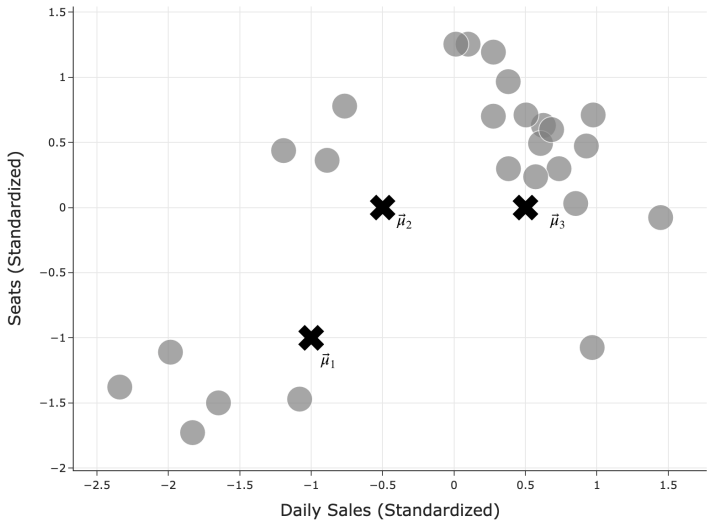
After 2 iterations, $\vec{u}^{(2)} = \begin{bmatrix} 1.2 \\ 0.2 \\ -0.6 \end{bmatrix}$, $\nabla q(\vec{u}^{(2)}) = \begin{bmatrix} 3.6 \\ 1.2 \\ 0 \end{bmatrix}$, $q(\vec{u}^{(2)}) = 2.28$

a) (3 pts) What is value of π? Give your answer as a fraction or decimal.

b) (6 pts) What is the value of $\vec{u}_2^{(0)}$, i.e. what is the second component of the **initial guess vector**, $\vec{u}^{(0)}$? Show your work, and put a $\boxed{\text{box}}$ around your final answer, which should be an expression in terms of π and/or constants. (If your answer does not involve π, we cannot give you partial credit in case your answer to (a) was incorrect.)

Question 8 (7 pts)

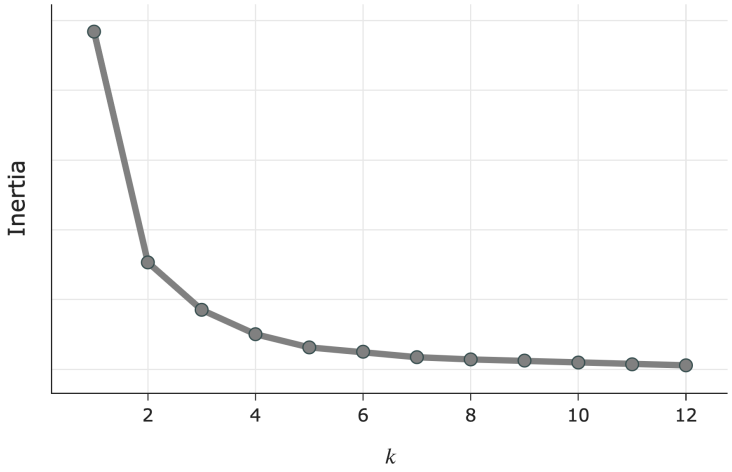
We'd like to use k-means clustering to place a dataset of $n = 25$ coffee shops into $k = 3$ clusters, using standardized versions of the the daily sales and number of seats features. In the scatter plot below, the points in gray correspond to the dataset, while the **X**s correspond to three initial centroids, $\vec{\mu}_1$, $\vec{\mu}_2$, and $\vec{\mu}_3$.



a) (2 pts) **Directly on the graph above**, **draw** an **X** (similar to the three shown above) at the approximate new location of $\vec{\mu}_1$ after one iteration of the k-means clustering algorithm. Don't draw anything else on the graph; if there are multiple markings on the graph, your answer will not receive credit.

b) (3 pts) When performing the first iteration of the k-means clustering algorithm using the initial centroids drawn above, we run into an issue when deciding where to move $\vec{\mu}_2$ and $\vec{\mu}_3$. In two English sentences, identify the issue and propose a solution.

c) (2 pts) Below is a plot of inertia vs. k, where k is the number of clusters used when running k-means clustering on the dataset of $n = 25$ points.



What is the y-coordinate of the point at $k = 1$ in the graph above? Give your answer as an integer. *Hint: Remember that both features have been standardized.*

Make sure you've written your uniqname in the space provided in the top right corner of every page of this exam.

Congrats on finishing the course — we'll miss you! Feel free to draw us a picture about Practical Data Science below :)

(1 pt) And here's a free point!