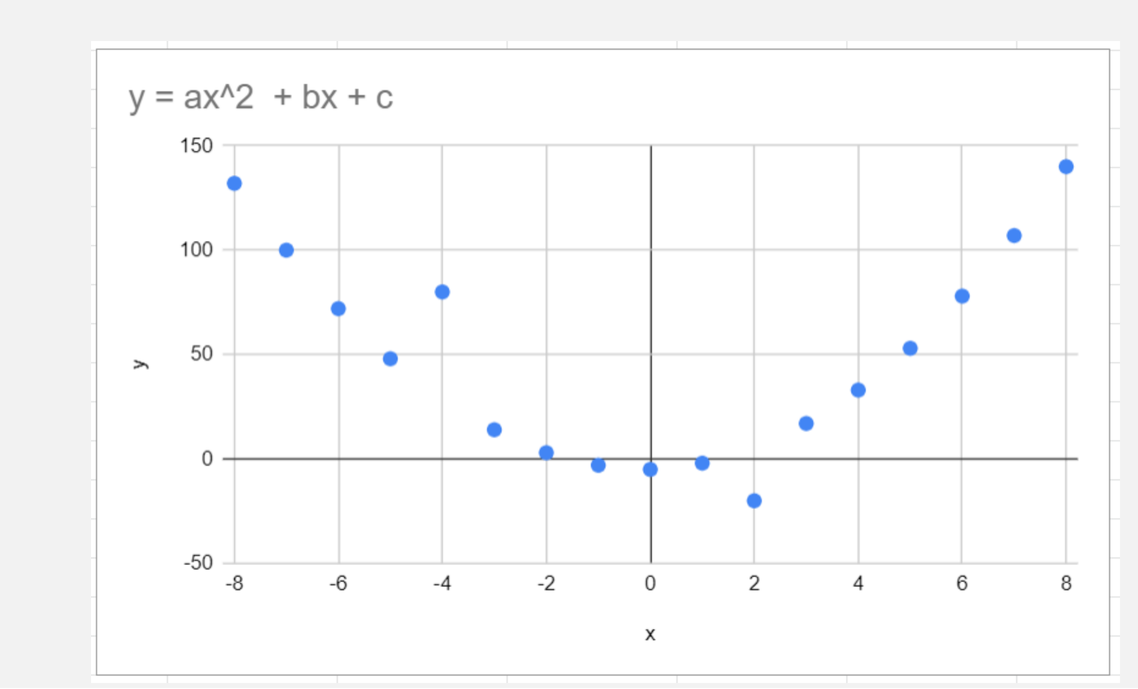
King Mongkut's University of Technology **Name: ________________**

Suthep Madarasmi, Ph.D. **I.D. Number: ___________**

Take Home Quiz 9 Due April 29, 2023. **Score: ______ / 95**

1. ***20 points.*** **1 hour. RANSAC Regression**. Use RANSAC to find a, b, c for the following dataset where points (x_i, y_i) are discrete samples from a function $f(x) = ax^2 + bx + c$ with 2 outliers. *Hint*: You should get a, b, and c close to 2.2, 0.5, -4.5, respectively.

x_i	y_i
-8	132
-7	100
-6	72
-5	48
-4	80
-3	14
-2	3
-1	-3
0	-5
1	-2
2	-20
3	17
4	33
5	53
6	78
7	107
8	140



2. Use K Means clustering on the IRIS dataset.

2.1 *10 points.* 0.5 hour. Using K = 3, cluster the entire dataset into 3 labels using only features 1 & 3; namely, sepal length and petal length (Note: the example in class used all 4 features for clustering). Show a scatter plot based on these 2 features using known training 3 classes using markers "<" for class 1 (Setosa), ">" for class 2 (Versicolor), and "^" for class 3 (Virginica) while also using colors based on the 3 computed clusters using colors of "pink" for cluster 1, "yellow" for cluster 2, and "cyan" for cluster 3.

2.2 *5 points.* Report based on known labels what percent is misclassified when using 2 features.

2.3 *10 points.* 0.5 hour. Plot the result of K Means clustering using all 4 features with K = 4.

2.4 *15 points.* 1 hour. Reduce the 4 features (sepal length, sepal width, petal length, petal width) into 2 PCA features (an example is also provided in class). Use K = 3 to cluster the entire dataset using these 2 PCA features. Show a scatter plot like in problem 2.1 along with percent misclassified as in problem 2.2.

2.5 *20 points.* Redo the example in class with all 4 features and K = 3, but using your own class or function **my_k_means** in Python that has initialization parameters: K, X, max_iterations, centroid_move_epsilon and returns y as a 1-D array of integer labels of 1, 2, …, K.. Each input N-dimensional data X[i] will have a 1-dimensional output label y[i] for i = 1..M where M is the number of data points. The algorithm should start by assigning K cluster centers based on random values from the (min, max) range of each dimension in the N-dimensional data X. It should stop when all centers have moved by less than the centroid_move_epsilon or when the max_iterations is reached. Make sure your results are similar to the K Means library class.

Algorithm 1 k-means algorithm

```
Specify the number $k$ of clusters to assign.
Randomly initialize $k$ centroids.
repeat
    expectation: Assign each point to its closest centroid.
    maximization: Compute the new centroid (mean) of each cluster.
until The centroid positions do not change.
```

3. ***15 points.*** **1 hour. Decision Trees.** Change the “IRIS Decision Tree.ipynb” shown in class, to use SKlearn’s Wine Recognition Dataset instead. Report the classification accuracy % for a single tree using 70% training samples and for a random forest with 100 estimators.