Time Constraint: 1d

Instructions:

1. Attempt all questions. Recommended that you read through the questions first and attempt them in order of increasing perceived difficulty.
2. The analysis should be done in Python
3. The submission for each problem should be in the form of a Jupyter notebook with associated .py script files. All documentation should be within the submitted notebook.
4. Please include as much information as possible in the text fields of your notebook to show your thought process while attempting each question.
5. Libraries required to replicate the results should be clearly communicated as a requirements.txt file.

Task A: Data Analysis

The dataset ProductSales.csv contains monthly sales for 2 products. Product 1 is a lotion and Product 2 is a detergent. The company has performed sub-par on fulfillment on average during the time period of the dataset. Product 1 was hit by a supply shortage in 2005, shortage periods for Product 2 need to be estimated from the data. Sales values are in tons.

Answer the following questions using the dataset. The questions are in no particular order.

1) Based on the nature of the products, hypothesize for each if it is seasonal or not. Analyze the sales numbers to test your hypothesis.

2) Estimate the months in 2005 for Product 1 which were affected by the supply disruption.

3) Isolate any data points for Product 2 that are likely to have been affected by similar shortfalls.

4) Estimate the expected sales in the affected months for each product.

7) Determine the average percentage loss per year for each product as a result of the supply shortage.

8) For each product, estimate if it is growing, declining or stable by the end of 2006.

9) Develop a monthly sales projection for the next 12 months for each product, assuming no further supply shortages.

Task B: Classification

The dataset Class.csv contains 20 features that may or may not be useful for predicting the labels column in Class.csv. No further information is available about the process that generated the data.

1) Put together an analysis of the dataset which identifies the following:

a) Outliers

b) Uninformative features

c) Redundant features

d) Imbalance

Justify the methods used to answer the above question.

2) Visualize the relationship between the informative features and the class labels

3) Train a classifier to predict the labels from the provided features, assuming that each type of misclassification error is equally costly. Clearly label each step of the process wrt your findings from 1 above.

4) Analyze the predictive accuracy of the trained classifier using an appropriate metric.

5) Modify the classifier from problem 3 above, based on the following cost structure

		Predicted		
		0	**1**	**2**
Actual	**0**	0.0	1.0	1.0
	1	1.0	0.0	1.0
	2	10000.0	10000.0	0.0

Task C: Probability

CallCounts.csv contains incoming call counts in hourly buckets over a year for a call center. It is known that the incoming calls are from a global audience, and the call center experiences a fairly constant stream of calls regardless of time-of-day, day-of-week and date. Fit a named

statistical distribution to the data (e.g. Gaussian, Beta etc.). Clearly state any assumptions made and justify your choice of distribution logically.

Task D: Optimization

There are 3 factories F_1, F_2, F_3 which can produce a particular good, and 4 cities C_1, C_2, C_3, C_4 to which the item must be supplied to be sold.

Assume that the production schedule is decided on a weekly basis and all deliveries are made over the weekend.

Capacity = the maximum qty of the item that a warehouse can produce over the current week. (in Tons)

Demand = the expected maximum qty of the item that can be sold in each city over the following week. (in Tons)

Shipping cost = the cost of shipping one ton of finished goods from a factory to a city. (in million Rs)

	C1	**C2**	**C3**	**C4**	**Factory Capacity**
F1	8	6	10	9	35
F2	9	12	13	7	50
F3	14	9	16	5	40
City Demand	45	20	30	30	

Determine the optimal amount of material to supply on each route to minimize the cost.