Dataset 1

Q1. How many unique device IDs are there in this dataset?
There are 1169 unique device ids

Q2. You are asked to do data analysis. What will you find?

The data analysis is performed in the code, and a more descriptive set of results are in there as well. Here some overall observations:

- There are overwhelmingly more successes than failures
- metric1 has the largest values and the most frequency as well
- metric2, 3, 4, 7, 8, and 9 all have huge, “spike” clusters of similar numbers
- metric5 and 6 appear to have more normally distributed metric values
- All the metrics appear to have little/no correlation, however metrics 7 and 8 appear to have perfect positive correlation with each other as they have correlation coefficients of 1 with each other
- The correlation matrix is also perfectly symmetrical along the diagonal axis running from the top left to the bottom right

Q3. You are asked to build a prediction model. Which kind of machine learning will be used and why? Supervised learning or Unsupervised learning? Regression, classification, or clustering? Which model will you use?

Overall, we should build a prediction model that uses supervised learning with classification. We should perform supervised learning because the data is already labeled, we have an item (in the form of the ‘failure’ column) that we can predict, and we can predict this outcome based on the other metrics. We should use classification when the target variable is categorical. Since “failure” is a binary classification problem, classification is the appropriate choice for the model. Regression is not suitable as the target variable is not continuous, and clustering is not suitable as it takes similar data points without predefined data points and groups them. I chose to do a logistic regression model as well as a random forest model.

Q4. Can you find the important features, informative features, or coefficient values? (Note: the answer will depend on your selected machine learning model.)

Yes! Here they are for the two models I chose

Logistic Regression Coefficients:

	Feature	Coefficient
0	metric1	0.123192
1	metric2	0.114720

4 metric5 0.107849
5 metric6 0.072002
3 metric4 0.064073
6 metric7 0.036494
7 metric8 0.036494
8 metric9 -0.022129
2 metric3 -0.099016

Random Forest Feature Importances:

	Feature	Importance
0	metric1	0.356292
5	metric6	0.261757
4	metric5	0.087470
1	metric2	0.085766
3	metric4	0.083531
7	metric8	0.046101
6	metric7	0.038544
8	metric9	0.029323
2	metric3	0.011215

Q5. There are two data from two devices, please predict the corresponding failure values.
The failure values are [0,0] for both models I made.

Dataset 3

Q1. How many columns are there in this csv file and what are these columns' names?

There is technically only one column in the data, however it looks like the data is supposed to have several columns as there are several comma separated items in each row. If we properly split the data, it looks like there are 7 columns in the file with the names: "Text, Sentiment, Source, Date/Time, User ID, Location, Confidence score"

Q2. Do the following steps and show the results.
● Clean data
● Drop nan values
I chose to remove duplicate rows in this case because of the context of the data being "reviews". In real life, it is highly unlikely to have multiple of the *exact same review* (based off the 7 column names, which are the review parameters) done by people. If this happens, I can assume the activity was generated by robots or error and can be deleted.

● Convert data and time to datetime
● Create new features - month, day, and hour
● Create new features again - Total Words, Total Chars, and Total Words After Transformation of "Text," where Total Words After Transformation means The natural logarithm of the word count of the "Text".

Q3. Do the following visualizations, eg., histplot, displot, barplot, kdeplot, etc., and write your findings.

1) by Sentiment 'Positive' and 'Negative'
There are overall more positive than negative sentiments

2) by 'Source' and 'Sentiment'
The most overall sentiments come from online stores, with the reviews being overwhelmingly negative

3) by 'Location' and 'Sentiment'
Every location has a mix of both good and bad sentiments, but NY and Orlando have only positive sentiments while Chicago has only negative ones (that hurts considering Chi-town is my hometown)

4) by 'Confidence Score' and 'Sentiment'
Positive sentiments all have lower confidence scores than negative ones

5) by 'Month' and 'Sentiment'
The vast majority of both positive and negative sentiments appear in month 7

6) by 'Day' and 'Sentiment'
Sentiments are relatively evenly distributed by day, although days 15 and 16 have a bit more sentiments (particularly positive ones) compared to the others

7) by 'hour' and 'Sentiment'
More sentiments tend to appear around the middle of the day, although the beginning of the day has more sentiments than the end of the day

8) by 'Total Words' and 'Sentiment'
Positive sentiments tend to have more total words than negative ones

9) by 'Total Chars' and 'Sentiment'
Positive sentiments tend to have more total chars than negative ones, although the distribution for both are quite close (especially considering positive sentiments have more outliers)

10) Wordcloud by Sentiment = Negative
Looks the common language used by an angry customer anywhere

11) Wordcloud by Sentiment = Positive
Looks words from a happy customer

12) by Top 25 Negative Words
13) by Top 25 Positive Words
The positive words seem to appear in “bins” of several words having the same frequency, whereas the negative words also have bins but of much smaller size

Q4. Build eight classification models taught in the class and find their own best parameters’ settings with the dataset, where "sentiment" is set as the target.

I started by loading and preprocessing the data. I then trained the models by performing hyperparameter tuning to determine the best parameters and scores for each classification model. I also went ahead and coded the evaluation of each classification model and their performances on the test set, although I wasn’t sure if I needed it for the sake of this question.

Q5. Show the confusion matrix and classification report of your eight models, compare these models and write your findings.

All of the confusion matrices and classification reports are generated in the plots section and console of the IDE respectively, however here were some general findings I observed:

Gaussian Naive Bayes, SVC, Multinomial Naive Bayes, and Logistic Regression seem to have the best overall results, with the Best Scores and accuracy values for all these models being the highest among the 8 models tested.

It appears that the confusion matrices for all 8 classification models are identical.

Dataset 4

Q1. Which kind of data selection method will you use to split csv data to training and testing datasets? sequential or random? WHY?

Random selection. This is done to mitigate bias, ensure representativeness, and avoid temporal patterns. Sequential selection may be appropriate in some special cases, like for time series data or for controlled experiments where data points are known to be independent and identically distributed.

Q2. In class, we learned many model evaluation methods, such as confusion matrix, accuracy score, precision score, recall score, and so on. In addition to the confusion matrix and accuracy score, which must be used in the Q3 and Q4, if you were to choose two evaluation metrics/scores, which two would you choose? Why?

I would use precision, for its relevance in the high cost of false positives, and recall, for its relevance in the high cost of false negatives. Using these two methods will also allow me to obtain the f1 score, which is the harmonic mean of precision and recall.

Q3. Please use eight classification models taught in the class and find their own best parameters' settings.

Unable to effectively test the classification models due to the massive load on my computer from the code.

Q4. Which prediction model is the best between these eight classification models? WHY?

Q5. Insurance_validation.csv is a validation dataset. Please use your best prediction model to get the "Response" and output the results to a csv file. The format of the csv file is the following sample.