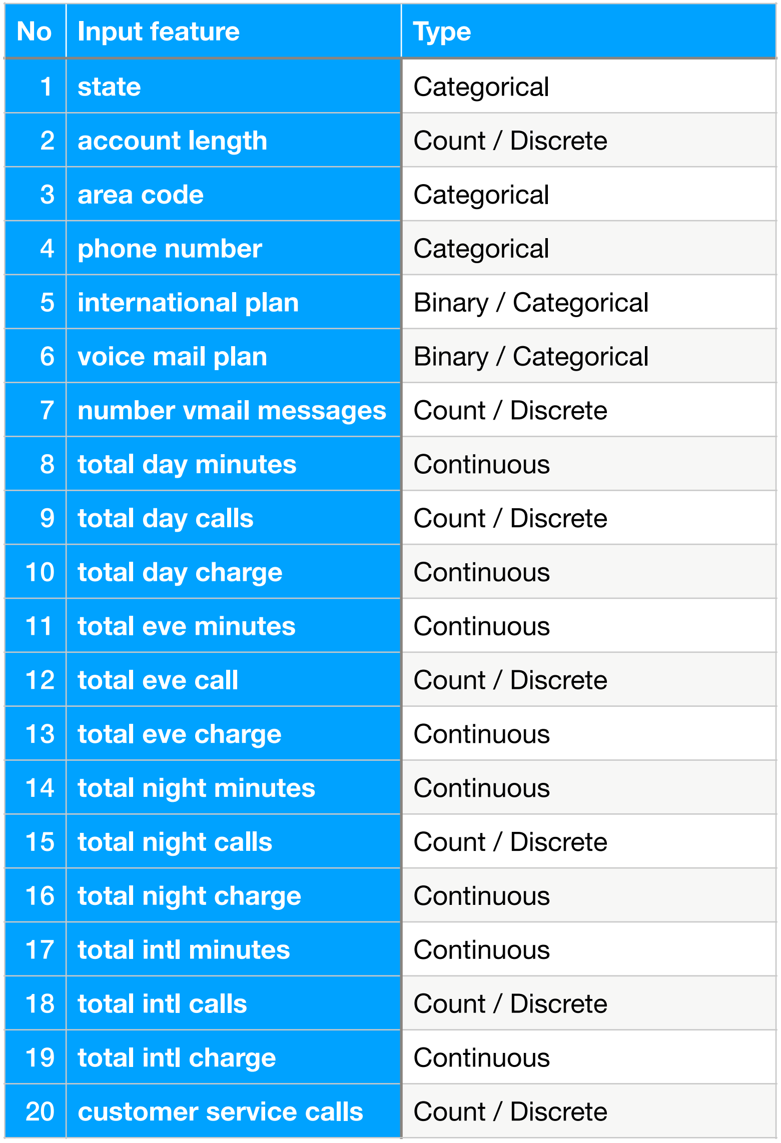
Data Analytics
Project 1

Problem 1

The files for this problem are under the Experiment 1 folder. Datasets to be used for experimentation: **telecom churn.csv**. Jupyter notebook to use as a starting point: **Exploratory data analysis.ipynb**. In this experiment, we will do exploratory data analysis to get a better sense of data. The dataset contains record of telecom customer along with the label “churn”. Churn = “true” signifies that the customer has left the company and churn = “false” signifies that the customer is still loyal to the company. Answer the following questions

1. How many records are there in the dataset?
   3333
2. How many input features are there for classification? Name each feature and assign it as categorical, count, or continuous.
   Number of Input features : 20 (‘churn’ is the predicting class)



No	Input feature	Type
1	state	Categorical
2	account length	Count / Discrete
3	area code	Categorical
4	phone number	Categorical
5	international plan	Binary / Categorical
6	voice mail plan	Binary / Categorical
7	number vmail messages	Count / Discrete
8	total day minutes	Continuous
9	total day calls	Count / Discrete
10	total day charge	Continuous
11	total eve minutes	Continuous
12	total eve call	Count / Discrete
13	total eve charge	Continuous
14	total night minutes	Continuous
15	total night calls	Count / Discrete
16	total night charge	Continuous
17	total intl minutes	Continuous
18	total intl calls	Count / Discrete
19	total intl charge	Continuous
20	customer service calls	Count / Discrete

3. For the continuous features, what is the average, median, maximum, minimum, and standard deviation values? Note that the 50 percentile value is same as the median.

Measure	mean	median	std	min	25%	75%	max
total day minutes	179.78	179.40	54.47	0.00	143.70	216.40	350.80
total day charge	30.56	30.50	9.26	0.00	24.43	36.79	59.64
total eve minutes	200.98	201.40	50.71	0.00	166.60	235.30	363.70
total eve charge	17.08	17.12	4.31	0.00	14.16	20.00	30.91
total night minutes	200.87	201.20	50.57	23.20	167.00	235.30	395.00
total night charge	9.04	9.05	2.28	1.04	7.52	10.59	17.77
total intl minutes	10.24	10.30	2.79	0.00	8.50	12.10	20.00
total intl charge	2.76	2.78	0.75	0.00	2.30	3.27	5.40

4. What is the average number of customer service calls made by a customer to the company?
Average number of customer service calls = 1.57

5. What is the distribution of the class variable, "churn"? Calculate the probability of P(churn = True) and P(churn = False).
The values of the variable 'churn' are 'true' and 'false'.

churn	Count
True	480
False	2853

	Probability Value
P(churn = True)	0.14
P(churn = False)	0.86

6. What is the distribution of the feature, "international plan"? Calculate the probability of P(international plan = 'yes') and P(international plan = 'no').
The values of feature 'international plan' are 'yes' and 'no'.

international plan	Count
Yes	329
No	3004

	Probability Value
P(international plan = yes)	0.1
P(international plan = no)	0.9

7. Assume you have devised a classification model that states that if “international plan” = ’no’, then the customer will not churn (i.e., churn = False). Report the accuracy of this classification model on the given dataset.

$$Accuracy = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}} = \frac{2665 + 141}{3333} = 0.842$$

8. Calculate the following conditional probabilities:

churn	False	True	All
international plan			
No	2665	339	3004
Yes	188	141	329
All	2853	480	3333

- P(churn = True | international plan = ‘yes’) = $\frac{141}{141 + 188} = 0.429$
- P(churn = False | international plan = ‘yes’) = $\frac{188}{141 + 188} = 0.571$
- P(churn= True | international plan = ‘no’) = $\frac{339}{339 + 2665} = 0.113$
- P(churn = False | international plan = ‘no’) = $\frac{2665}{339 + 2665} = 0.887$

Based on the probabilities computed above and those computed in parts 5 and 6, answer the following question using the Bayes theorem: “Given that a customer has churned (churn = True), what are the probabilities that the customer has opted/not-opted for the international plan? Similarly, given that the customer has not churned (churn = False), what are the probabilities that the customer has opted/not-opted for the international plan?”

According to Naive Bayes:

$$P(A\,|\,B) = \frac{P(B\,|\,A)\;P(A)}{P(B)}$$

P(international plan = ‘yes’ | churn = True) $= \dfrac{0.429 * 0.1}{0.14} = 0.306$

P(international plan = ‘no’ | churn = True) $= \dfrac{0.113 * 0.9}{0.14} = 0.726$

P(international plan = ‘yes’ | churn = False) $= \dfrac{0.571 * 0.1}{0.86} = 0.066$

P(international plan = ‘no’ | churn = False) $= \dfrac{0.887 * 0.9}{0.86} = 0.928$

9. Assume you have devised a classification model which states that if “international plan” = “yes” and the number of calls to the service center is greater than 3, then the customer will churn (i.e., “churn” = True). Report the accuracy of this classification model on the given dataset.

churn	False	True	All
international plan = ‘yes’ and no of service calls > 3			
False	2841	460	3301
True	12	20	32
All	2853	480	3333

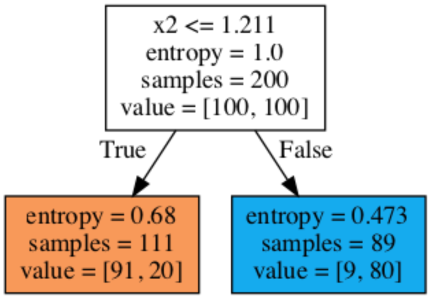
$$Accuracy = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}} = \frac{2841 + 20}{3333} = 0.858$$

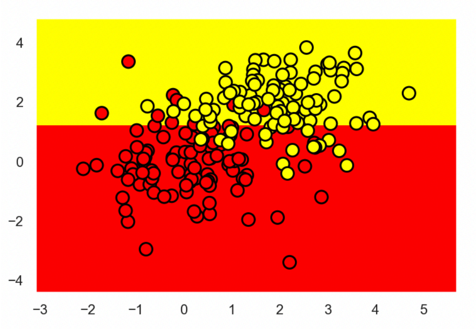
Problem 2

The files for this problem are under the Experiment 2 folder. Datasets to be used for experimentation: **telecom churn.csv**. Jupyter notebook to be used as a starting point: **Decision Trees and kNN.ipynb**. In this experiment we will apply and visualize decision trees and kNN, fine tune their hyper-parameters and learn about k-fold cross validation. To visualize decision tree we need additional packages to be installed i.e., **Graphviz** and **pydotplus** (check the Anaconda.org page for instructions on how to install them using conda). Answer the following questions:

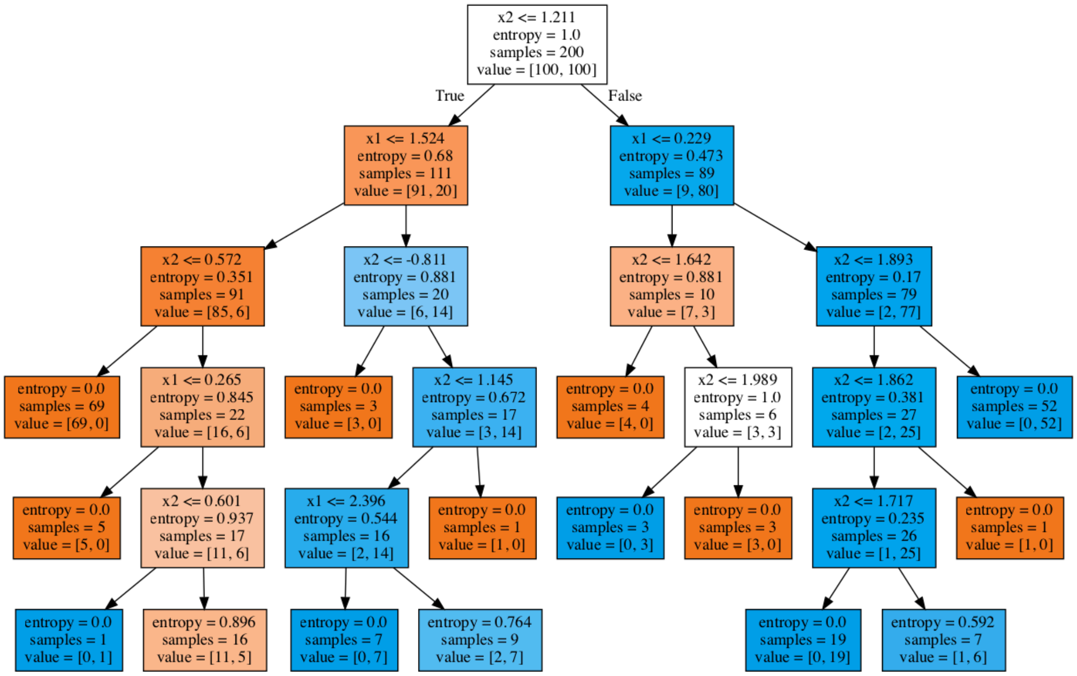
1. Consider training decision trees for the synthetic dataset involving two classes. How does the decision boundary look like when we overfit (max depth ≥ 4) and underfit (max depth = 1) the decision tree on the given data? For both cases, paste the decision tree and the decision boundary from Jupyter notebook output.

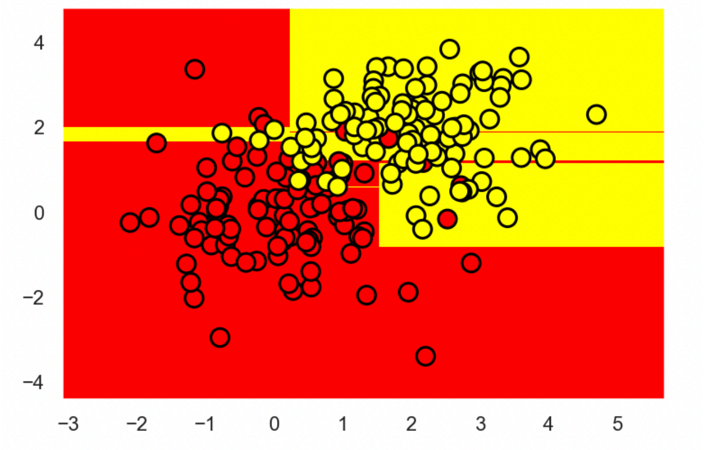
   The decision boundary is simple when we underfit (max depth = 1) the decision tree



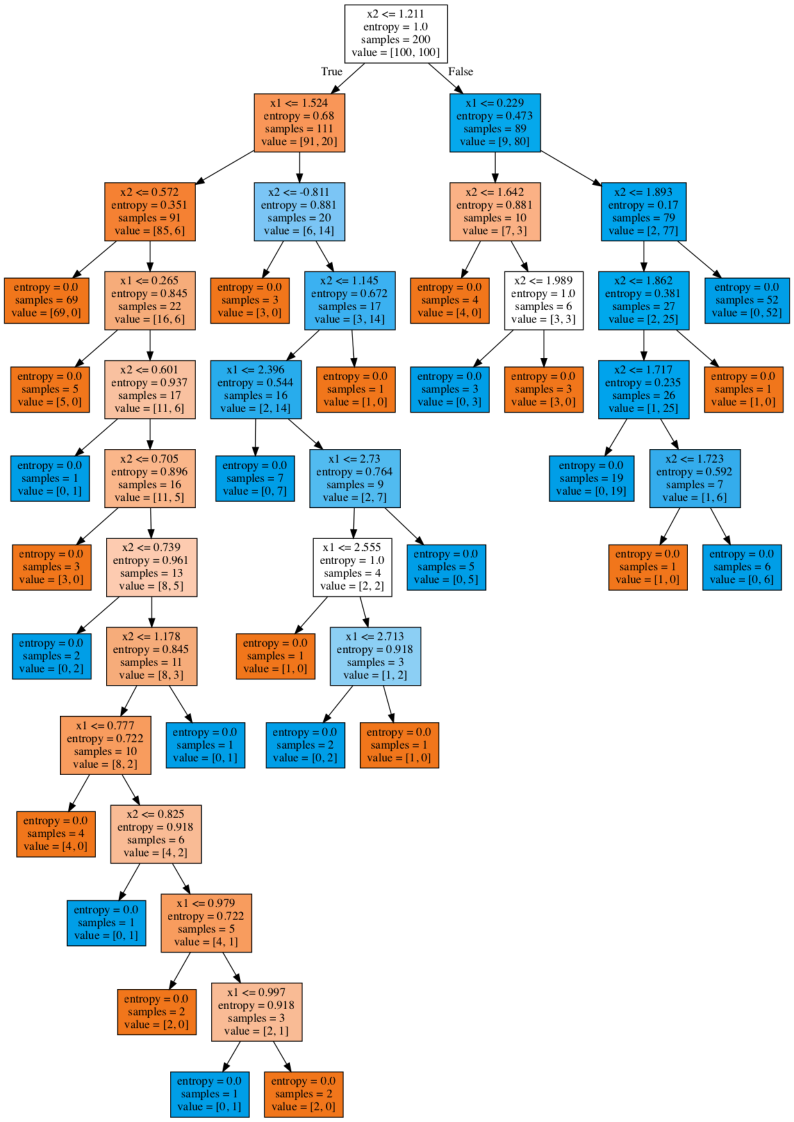


   The decision boundary is complex with lots of boundaries when we overfit (max depth = 5) the decision tree.





2. Decision tree classifier *sklearn.tree.DecisionTreeClassifier* has parameter “max depth” which defines the maximum depth of the tree. What happens if we don’t specify any value for this parameter? Paste the decision tree and the decision boundary you will obtain for this default case from Jupyter notebook output.
   If the max depth is not specified, the decision is built until all the leaf nodes are pure.


x2 <= 1.211 entropy = 1.0 samples = 200 value = [100, 100]
True
False
x1 <= 1.524 entropy = 0.68 samples = 111 value = [91, 20]
x1 <= 0.229 entropy = 0.473 samples = 89 value = [9, 80]
x2 <= 0.572 entropy = 0.351 samples = 91 value = [85, 6]
x2 <= -0.811 entropy = 0.881 samples = 20 value = [6, 14]
x2 <= 1.642 entropy = 0.881 samples = 10 value = [7, 3]
x2 <= 1.893 entropy = 0.17 samples = 79 value = [2, 77]
entropy = 0.0 samples = 69 value = [69, 0]
x1 <= 0.265 entropy = 0.845 samples = 22 value = [16, 6]
entropy = 0.0 samples = 3 value = [3, 0]
x2 <= 1.145 entropy = 0.672 samples = 17 value = [3, 14]
entropy = 0.0 samples = 4 value = [4, 0]
x2 <= 1.989 entropy = 1.0 samples = 6 value = [3, 3]
x2 <= 1.862 entropy = 0.381 samples = 27 value = [2, 25]
entropy = 0.0 samples = 52 value = [0, 52]
entropy = 0.0 samples = 5 value = [5, 0]
x2 <= 0.601 entropy = 0.937 samples = 17 value = [11, 6]
x1 <= 2.396 entropy = 0.544 samples = 16 value = [2, 14]
entropy = 0.0 samples = 1 value = [1, 0]
entropy = 0.0 samples = 3 value = [0, 3]
entropy = 0.0 samples = 3 value = [3, 0]
x2 <= 1.717 entropy = 0.235 samples = 26 value = [1, 25]
entropy = 0.0 samples = 1 value = [1, 0]
entropy = 0.0 samples = 1 value = [0, 1]
x2 <= 0.705 entropy = 0.896 samples = 16 value = [11, 5]
entropy = 0.0 samples = 7 value = [0, 7]
x1 <= 2.73 entropy = 0.764 samples = 9 value = [2, 7]
entropy = 0.0 samples = 19 value = [0, 19]
x2 <= 1.723 entropy = 0.592 samples = 7 value = [1, 6]
entropy = 0.0 samples = 3 value = [3, 0]
x2 <= 0.739 entropy = 0.961 samples = 13 value = [8, 5]
x1 <= 2.555 entropy = 1.0 samples = 4 value = [2, 2]
entropy = 0.0 samples = 5 value = [0, 5]
entropy = 0.0 samples = 1 value = [1, 0]
entropy = 0.0 samples = 6 value = [0, 6]
entropy = 0.0 samples = 2 value = [0, 2]
x2 <= 1.178 entropy = 0.845 samples = 11 value = [8, 3]
entropy = 0.0 samples = 1 value = [1, 0]
x1 <= 2.713 entropy = 0.918 samples = 3 value = [1, 2]
x1 <= 0.777 entropy = 0.722 samples = 10 value = [8, 2]
entropy = 0.0 samples = 1 value = [0, 1]
entropy = 0.0 samples = 2 value = [0, 2]
entropy = 0.0 samples = 1 value = [1, 0]
entropy = 0.0 samples = 4 value = [4, 0]
x2 <= 0.825 entropy = 0.918 samples = 6 value = [4, 2]
entropy = 0.0 samples = 1 value = [0, 1]
x1 <= 0.979 entropy = 0.722 samples = 5 value = [4, 1]
entropy = 0.0 samples = 2 value = [2, 0]
x1 <= 0.997 entropy = 0.918 samples = 3 value = [2, 1]
entropy = 0.0 samples = 1 value = [0, 1]
entropy = 0.0 samples = 2 value = [2, 0]

3. For Bank Dataset, what are the 5 different age values that the decision tree used to construct the splits of the tree? What is the significance of these 5 values?
   43.5, 19, 22.5, 30, 32 are the age values taken by the decision tree for the split.
   These age values are mean values between the ages where the target class is switching from 0 to 1 or 1 to 0.

4. For the customer churn prediction task, we show that the accuracy of the decision tree is 94% when max depth is set to 5. What happens to accuracy when we leave the value of max depth to its default value? Explain the rise/fall of accuracy.
   The accuracy when max_depth = 5 is 0.91.
   This is the accuracy of the test data. Due to over fitting, when the data is complex, the accuracy of the test data decreases.

5. Given a dataset d, with n sample and m continuous features, what does Standard Scaler *sklearn.preprocessing.StandardScaler* do? Given dataset d = [[0, 0], [0, 0], [1, 1], [1, 1]], write down its scaler transformation.
   The StandardScaler normalizes/standardizes the data. It is also called z-score transformation. This transformation changes the standard deviation of data to 1 and mean of the data = 0.
   Given dataset d = [[0, 0], [0, 0], [1, 1], [1, 1]], StandardScaler transforms the data to d_transformed = [[-1,-1], [-1,-1], [1, 1], [1, 1]]

6. How many decision trees do we have to construct if we have to search the two-parameter space, max depth[1-10] and max features[4-18]? If we consider 10-fold cross-validation with the above scenario, how many decision trees do we construct in total?
   Number of decision trees to construct in 5 fold cross validation = 630
   Number of decision trees to construct in 10 fold cross validation = 1260

7. For the customer churn prediction task, what is the best choice of k[1-10] in the k- nearest neighbor algorithm in the 10-fold cross-validation scenario?
   For k=7, in the 10-fold cross validation scenario is the best choice.

8. For MNIST dataset, what was the accuracy of the decision tree [max depth = 5] and K-nearest neighbor [K = 10]? What were the best hyper-parameter values and test accuracy for decision trees when we used GridSearchCV with 5 fold cross-validation?
   Accuracy of decision tree (max depth = 5) = 0.667
   Accuracy of K-nearest neighbor (k = 10) = 0.976

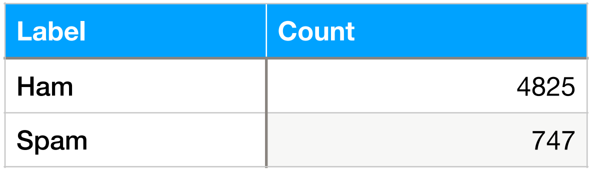
   For 5 fold cross validation scenario,
   Best max depth = 10
   Best max features = 50
   Test Accuracy = 0.843

Problem 3

The files for this problem are under Experiment 3 folder. Datasets to be used for experimentation: **spam.csv**. Jupyter notebook to be used as starting point: **Naive Bayes Spam.ipynb**. The dataset contains 5,574 messages tagged according to ham (legitimate) or spam. In this experiment we will learn about text features, how to convert them in matrix form, and apply the Naive Bayes algorithm. Answer the following questions:

1. What is the distribution of the "label" class. Is it skewed?

   The 'label' class has 2 unique values : 'ham' and 'spam'



Label	Count
Ham	4825
Spam	747

   Yes, it is skewed towards 'ham'.

2. How many unique values of SMS are there in the dataset? What is the SMS that occurred most frequently and what is its frequency?

   No of unique values of SMS : 5169

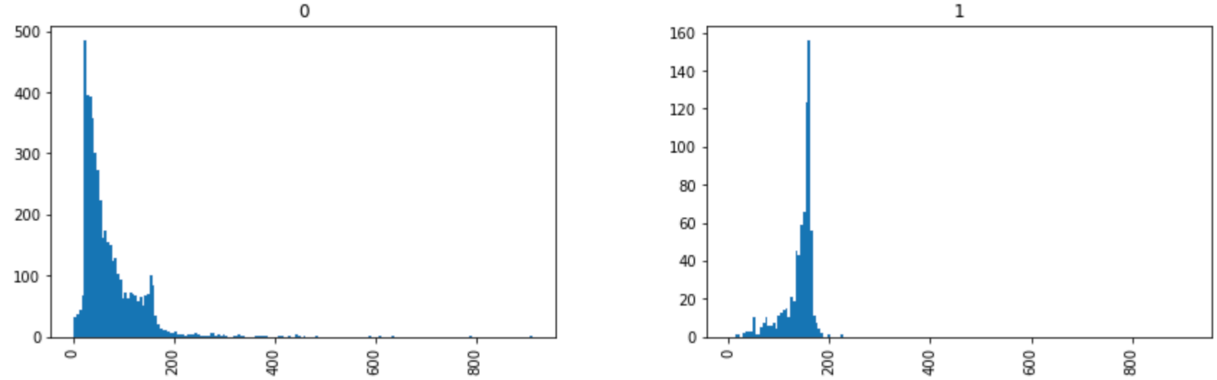
   'Sorry, I'll call later' is the most frequent SMS of frequency 30.

3. What is the maximum and minimum length of SMS present in the dataset?

   Maximum length of SMS present in Dataset : 910

   Minimum length of SMS present in Dataset : 2

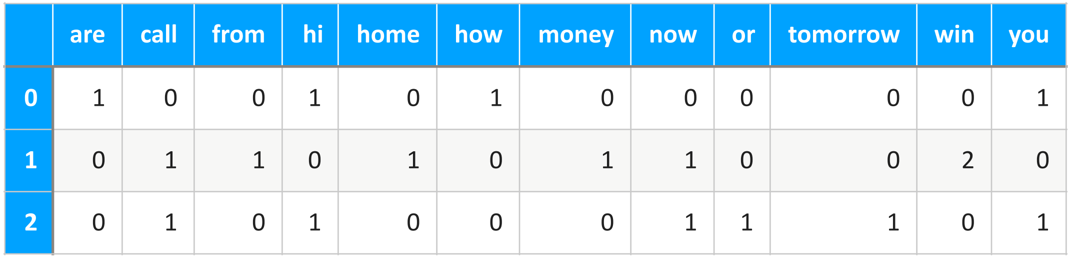
4. Plot the histogram of the length of SMS for both labels separately with bin size 5, i.e. histogram of the length of all ham SMS and histogram of the length of all spam SMS. What can you say about the difference in SMS lengths across the two labels after examining the plots?



   0 titled histogram is ham labeled data histogram whereas 1 titled histogram is spam labeled data histogram.

From the 2 different histograms, we can tell that
ham labeled histogram is right skewed and has many outliers.
spam labeled histogram is left skewed and has few outliers when compared with rev one.

5. Using bag of words approach, convert documents = [’Hi, how are you?’, ’Win money, win from home. Call now.’, ’Hi., Call you now or tomorrow?’] to its document-term matrix.



	are	call	from	hi	home	how	money	now	or	tomorrow	win	you
0	1	0	0	1	0	1	0	0	0	0	0	1
1	0	1	1	0	1	0	1	1	0	0	2	0
2	0	1	0	1	0	0	0	1	1	1	0	1

6. Report accuracy, precision, recall and F1 score for the spam class after applying Naive Bayes algorithm.
Accuracy score: 0.985
Precision score: 0.942
Recall score: 0.935
F1 score: 0.939