Alex Teboul

DSC 441

Assignment 1

Due: Friday, September 21, 2018 by 11:59:59pm

<u>Problem 1 (5 points):</u>

Differentiate between the following terms:

a. classification and clustering

- The text differentiates between classification and clustering by stating that classification is the process of predicting categorical labels while clustering is for grouping data object together based on similarity – without using class-labels.
- A classifier, or classification model, can be used to determine what class a data object belongs to. For example, is this a picture of a 'cat' or 'not a cat'.
- Clustering can be used to generate class labels for a group of data, by grouping based on maximal intraclass similarly and minimal interclass similarity – basically similar things get grouped together and different things don't go together. Clustering is partitioning a group of data objects into subsets (clusters).

b. classification and prediction

- The text, 'Data Mining: Concepts and Techniques', distinguishes between classification and prediction by arguing that classification predicts categorical (discrete/unordered) labels, while prediction is for predicting numeric values (continuous, ordered).
- Classification is for predicting what class or category a new data point belongs to. For example, is this transaction 'Fraud' or 'Not Fraud', or should a new patient receive 'Treatment A', 'Treatment B', or 'Treatment C' based on their symptoms and medical history.
- Prediction is for predicting a missing or unknown value of interest, usually involving some form of regression analysis. For example, can you predict how much a home will sell for based on a variety of factors like number of rooms, square footage, proximity to a

school, etc. The result is a numeric value (could be in $ USD), that was likely arrived at using some form of regression analysis. Another example of a prediction could be predicting customer spend during an online sale.

c. feature selection and feature extraction

- Feature selection and extraction are two means of dimensionality reduction, which is necessary for model efficiency. Feature selection involves selecting a subset of features from the data for use in the model. Feature extraction transforms the features from the data into a lower dimensional space. Both reduce the dimensionality but selection 'selects' some features to use while leaving out others, and extraction 'extracts/ transforms' all the features such that their dimensionality is reduced.

d. data mining and SQL

- Data mining is the process of discovering interesting, hidden, and non-trivial patterns and knowledge from huge amounts of data. In data mining, value comes from learning the relationships between different attributes of database records. SQL is just a programming language used for querying and extracting obvious information from relational databases. Ultimately, data mining is a process that can uncover valuable nuggets of information/ knowledge for decision makers in a company, while SQL is a query programming language that can help reveal explicit information held in databases.

e. data warehouses and data marts

- Data warehouses are used to store huge amounts of an organization's historical data in order to facilitate decision making. Data warehouses differ from operational databases in that the records within them are never updated, just used for querying. A large organization that operates in multiple regions may have multiple, smaller data warehouses based on location or department – these comparatively smaller warehouses are called data marts.

<u>Problem 2 (5 points):</u>

Discuss whether or not each of the following activities is a data mining task.

(a) Monitoring the heart rate of a patient for abnormalities.

- Yes

- This can be considered a data mining task because it can involve the detection of unusual patterns from large amounts of data. In a sophisticated heart rate monitoring system, a variety of factors could be considered and context could be important as well for making the determination of whether or not the heart rate is abnormal. For example, factors like age, sex, level of fitness, medication, sleep, medical conditions, current activity, and time of the day can all affect heart rate. Making the determination over what is a 'normal' heart rate would not be as simple as defining thresholds for what is considered 'normal' versus 'abnormal'. An element of learning and non-trivial signal analysis would be necessary for the monitoring system to be optimally accurate. There may also be certain rhythms and signal variations that occur prior to heart attacks and other heart conditions that the model would need to recognize. Ultimately, as an outlier/anomaly detection and classification problem it qualifies as a data mining task.

(b) Computing the total number of courses offering by an university.

- No
- This is not a data mining task because it only involves finding an explicit piece of information from the data. The total number of courses will either be given already in the database or can be calculated easily by adding up all the courses.

(c) Sorting a student database based on student identification numbers.

- No
- This is not a data mining task because the task does not involve discovering any interesting pattern from the data. Sorting the student database based on IDs is explicit and be accomplished without the use of any data mining tools/techniques.

(d) Predicting the outcomes of tossing a (fair) pair of dice.

- No
- Again, not a data mining task because interesting patterns are not being discovered, in the sense that the outcomes of tossing a fair pair of dice are known to follow basic probability rules. This is trivial, explicit, and doesn't require massive amounts of data to figure out – it's just a probability calculation.

(e) Monitoring seismic waves for earthquake activities.

- Yes
- Monitoring seismic waves for earthquake activities could be considered a data mining task because it would involve the classification of seismic wave behavior linked to earthquake activities. The model would have to be capable of extracting some non-trivial, hidden knowledge from a large amount of seismic data.

Problem 3 (15 points):

Fisher's iris data (download the IRIS dataset from http://archive.ics.uci.edu/ml/datasets/Iris) consists of measurements on the sepal length, sepal width, petal length, and petal width of 150 iris specimens. There are 50 specimens from each of three species. Use SPSS to answer the following questions:

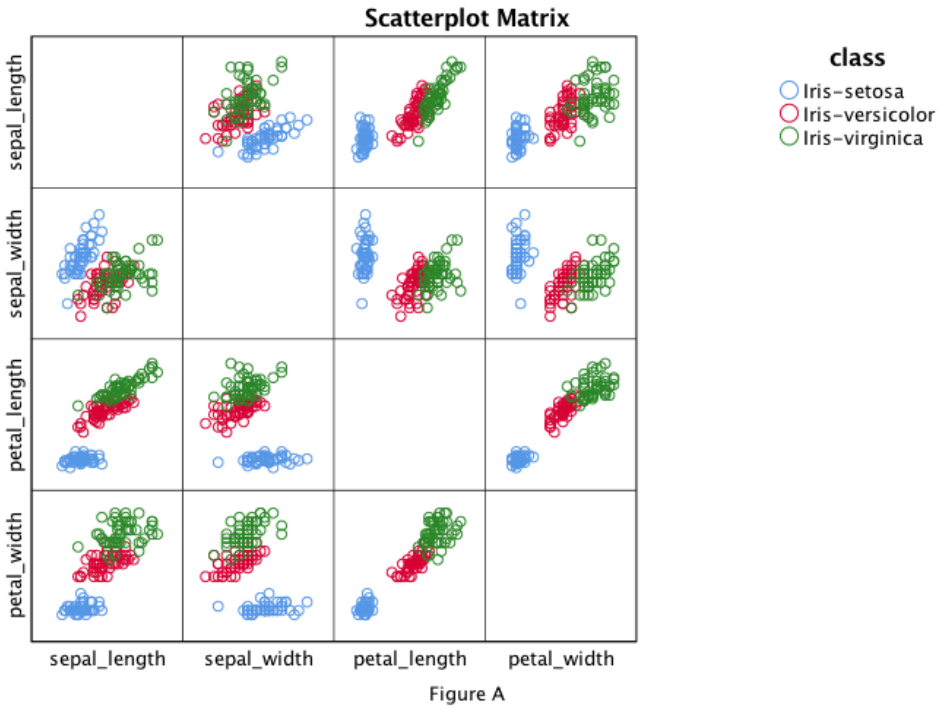


Figure A

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean		Std. Deviation	Variance
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic
sepal_length	150	3.6	4.3	7.9	5.843	.0676	.8281	.686
sepal_width	150	2.4	2.0	4.4	3.054	.0354	.4336	.188
petal_length	150	5.9	1.0	6.9	3.759	.1441	1.7644	3.113
petal_width	150	2.4	.1	2.5	1.199	.0623	.7632	.582
Valid N (listwise)	150							

- The scatterplot matrix helps identify interesting attributes to look at. For the purpose of this question though, only the sepal length v. sepal width and petal length v. petal width scatter plots are analyzed further. Descriptive statistics are reported to help with analysis.

a. Visualize and interpret the relationship between the two sepal variables, sepal length and sepal width. Provide the scatterplot that you created to visualize the data along with your interpretation. When you plot the data, you may want to use different colors/signs for representing the data points belonging to the different three class species. Do you think that a classification algorithm will be successful in classifying the data with respect to these two variables? Justify your answer.

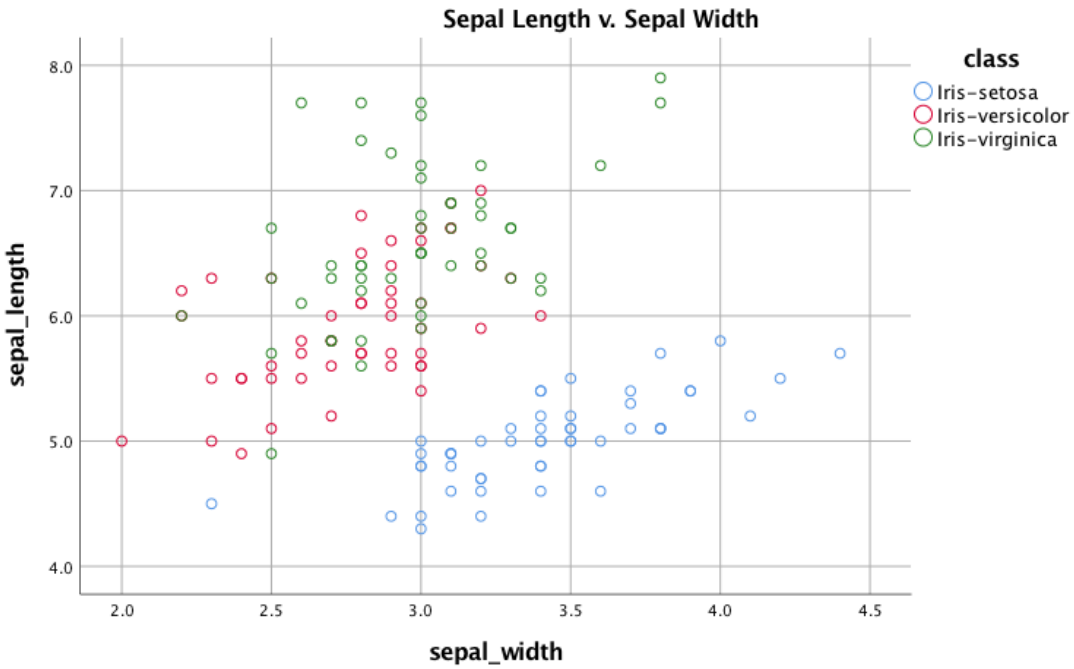


Figure 1: A scatterplot of sepal length versus sepal width.

- No - The attributes sepal length and sepal width are not optimal for classifying the three different class species. The class Iris-setosa, seen in blue, can be differentiated from the other two species as its values do not overlap with the values for the other two classes. There is not a good distinction between the classes Iris-versicolor and Iris-virginica. The points for sepal length v. sepal width for these two species overlap heavily, making classification difficult. Even after outliers are removed, classification is still difficult. Potential analysis methods to classify the points include using regression, support vector machines, k-means, k-nearest neighbors, etc. Other combinations seen above in my Figure A Scatterplot Matrix have far better distinction between the three classes.

b. Repeat part a. for the petal variables.

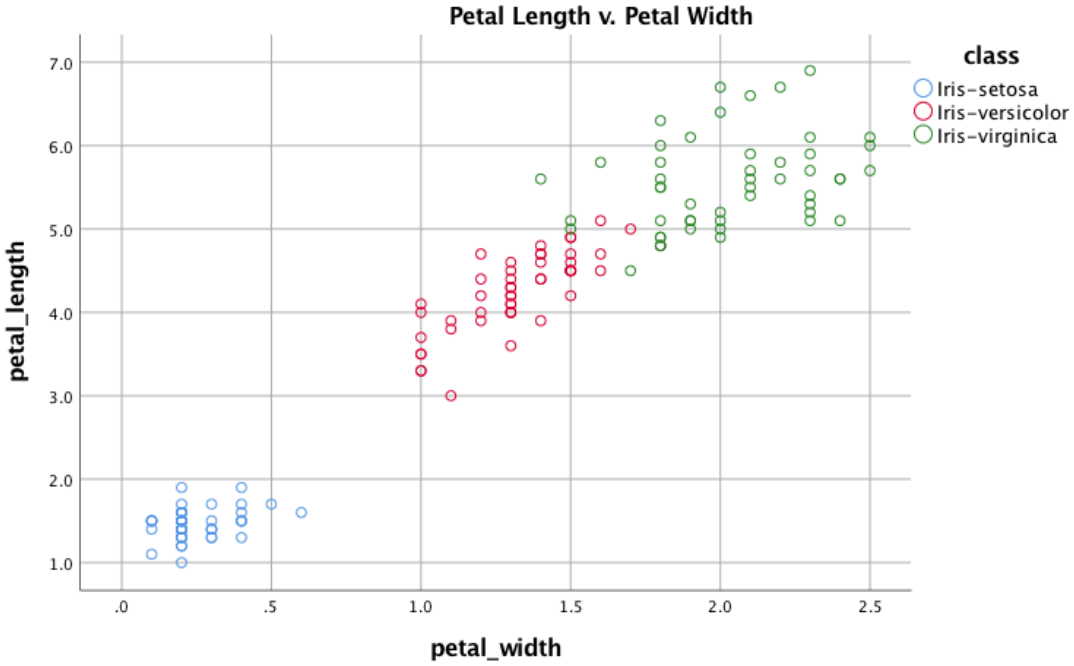


Figure 2: A scatterplot of petal length versus petal width.

- Yes – I believe a classification algorithm would do a good job of classifying these points with respect to class species. There is much better differentiation between the Iris-versicolor and Iris-virginica than there was in the sepal length v. sepal width scatterplot. Here, Iris-setosa has low petal length and width, Iris-versicolor has medium values of petal length and width, and Iris-virginica has high values of petal length and width.

c. Draw the histograms of the four variables and interpret the distributions of each one of the four variables.

Frequencies

Statistics

		sepal_length	sepal_width	petal_length	petal_width
N	Valid	150	150	150	150
	Missing	0	0	0	0
Median		5.800	3.000	4.350	1.300
Minimum		4.3	2.0	1.0	.1
Maximum		7.9	4.4	6.9	2.5
Percentiles	25	5.100	2.800	1.575	.300
	50	5.800	3.000	4.350	1.300
	75	6.400	3.300	5.100	1.800

- First, the 5 number summary is helpful to get a basic idea of what the Iris data set looks like. The median, min, max, Q1, and Q3 reported here help to define the distribution of the data. It is important to note that this is before any outliers have been thrown out. Additionally, this is not looking specifically at class. Boxplots later in this question help further describe the distribution of the data with respect to class and identifying outliers.

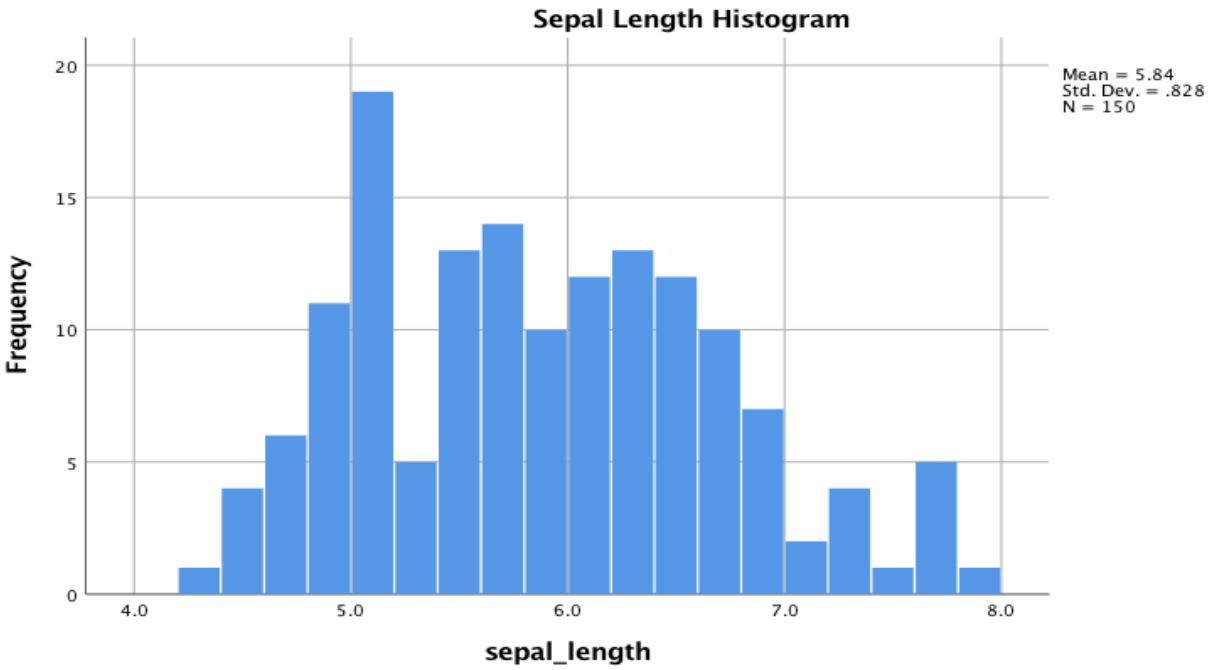
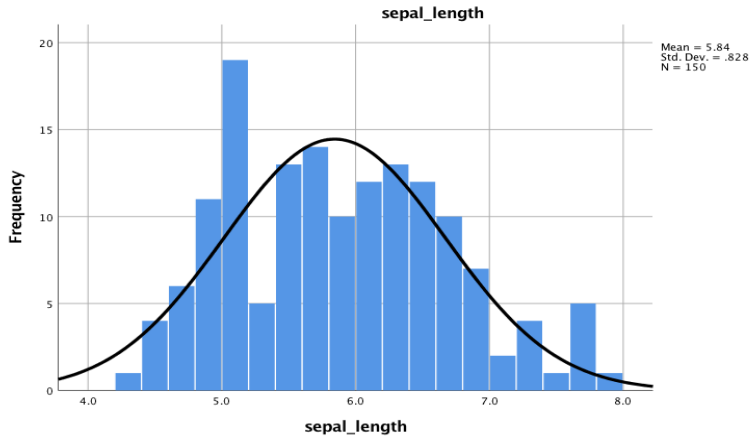


Figure 3: A histogram of sepal length.

- The sepal length histogram shows lengths in range 4.3cm to 7.9cm, a mean of 5.84cm, std of 0.828, median of 5.8cm, Q1 at 5.1cm and Q3 at 6.4cm for 150 plant samples.



- The data can fall under a normal distribution, but it would be more accurate to describe this data with multiple histograms – one for each class. This is shown below.

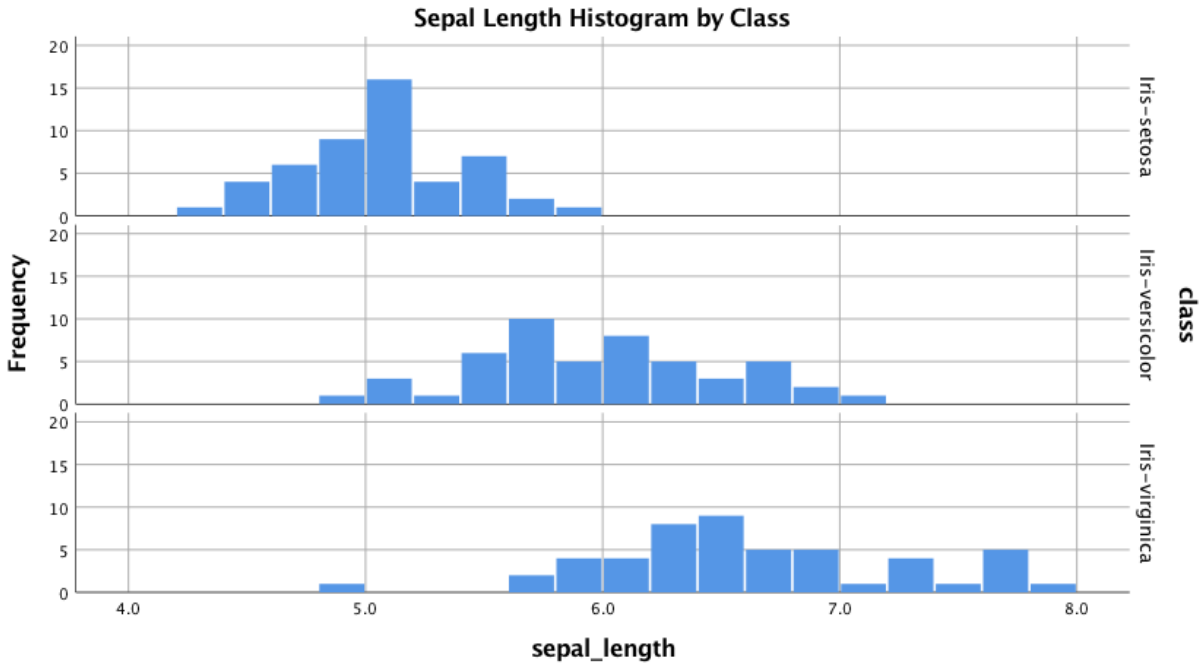


Figure 3.1: A histogram of sepal length panelled by class.

- Separating out the histograms based on class shows three different distributions in the data. This makes sense given that the three species of Iris generally have different sepal length characteristics.

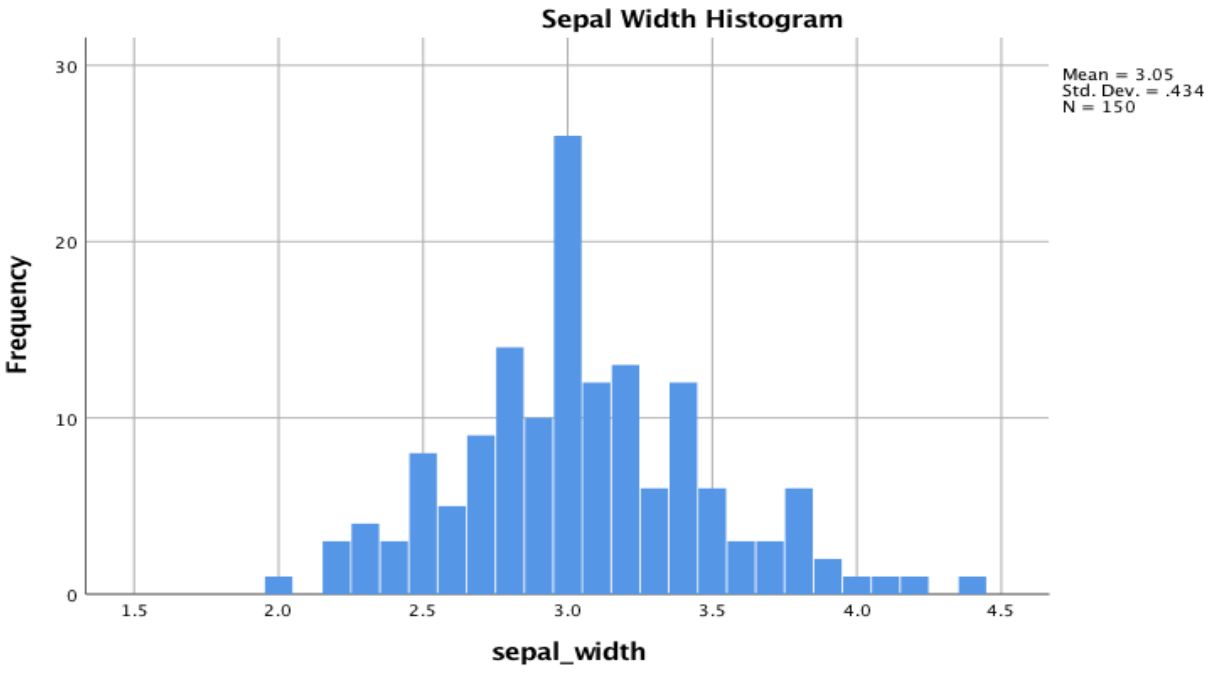
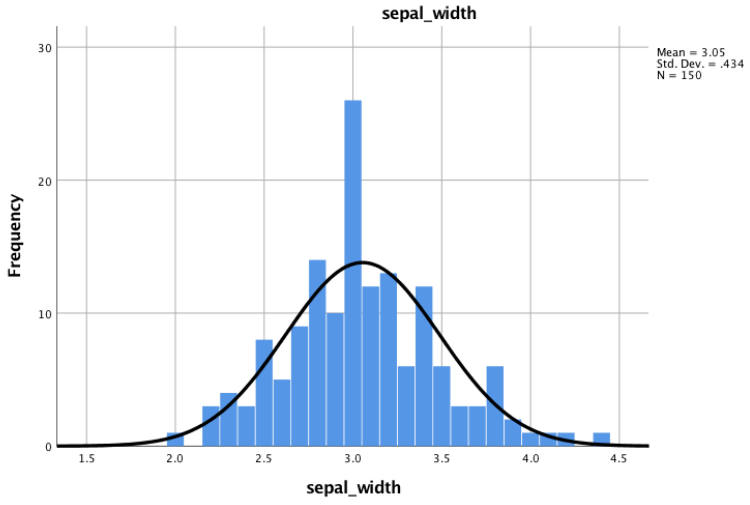


Figure 4: A histogram of sepal width.

- The sepal width histogram shows lengths in range 2cm to 4.4cm, a mean of 3.05cm, std of 0.434, median of 3cm, Q1 at 2.8cm and Q3 at 3.3cm for 150 plant samples.



- Sepal width falls more clearly under a normal distribution bell curve than sepal length did. Now to examine the histograms by class to see if they are all correlated.

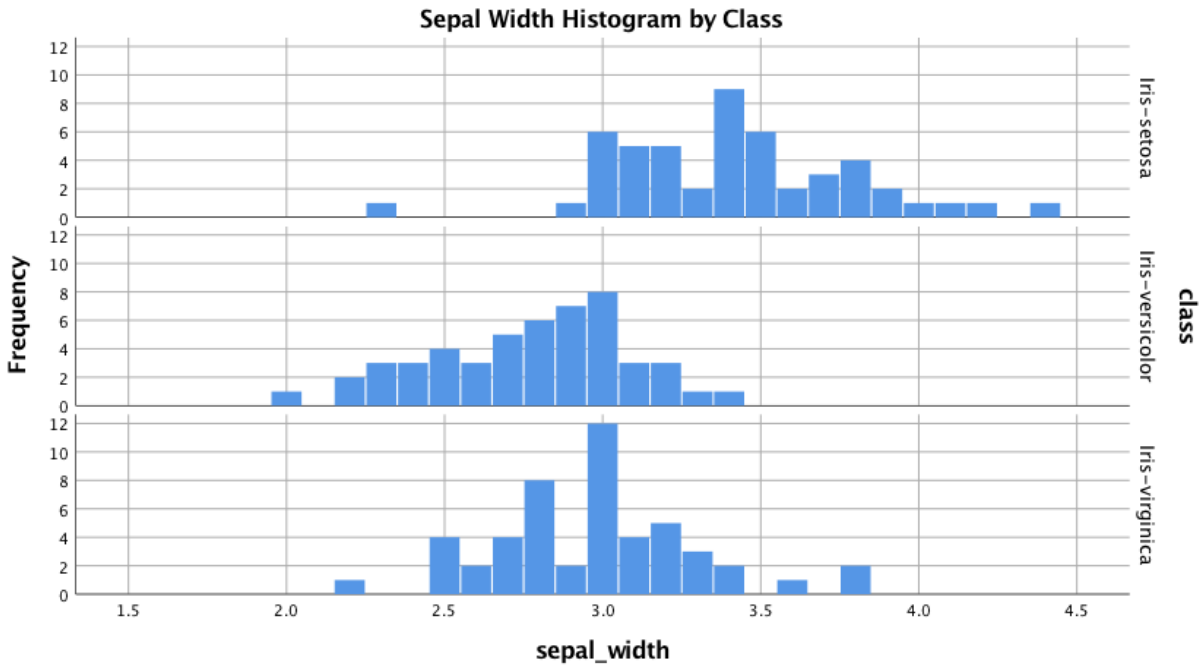


Figure 4.1: A histogram of sepal width panelled by class.

- Separating out the histograms by class shows less variability than in the sepal length attribute. Here virginica has a fairly normal distribution with a mode of 3, versicolor has a mode of 3 as well but is skewed to the left, and setosa has a mode of 3.4 and an unclear skew. Of all the four attributes, this one most clearly composes to a single normal distribution.

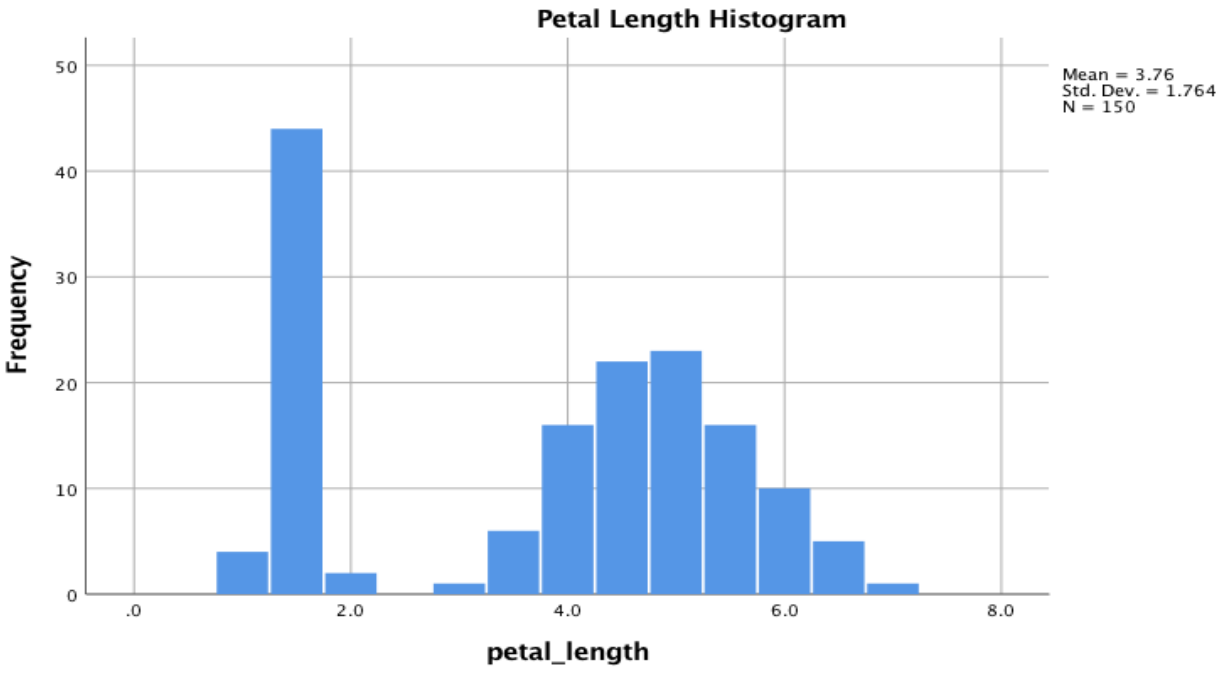
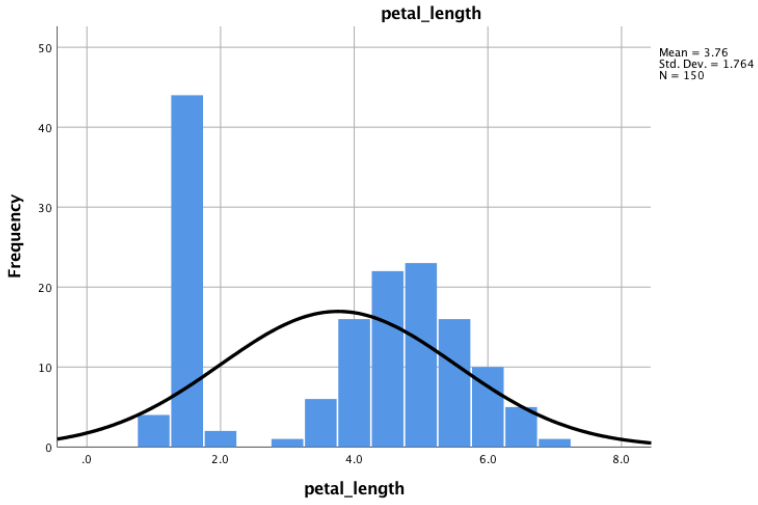


Figure 5: A histogram of petal length.

- The petal length histogram shows lengths in range 1cm to 6.9cm, a mean of 3.76cm, std of 1.764, median of 4.25cm, Q1 at 1.575cm and Q3 at 5.1cm for 150 plant samples.



- This clearly should not fall under a single bell curve. It would be better described by two curves. The histogram by class analysis below should reveal by the data seems to be composed of multiple curves.

Figure 5.1: A histogram of petal length panelled by class.

- Here we see that the Iris-setosa is distinct from the other two species with a peak at 1.6, while versicolor and virginica have peaks at 4.4 and 4.8 respectively. It is the combination of the versicolor and virginica curves that produce the right curve seen in the combined histogram and the setosa is that left curve. This makes me think that maybe setosa is less related to versicolor and virginica than versicolor and virginica are to each other.

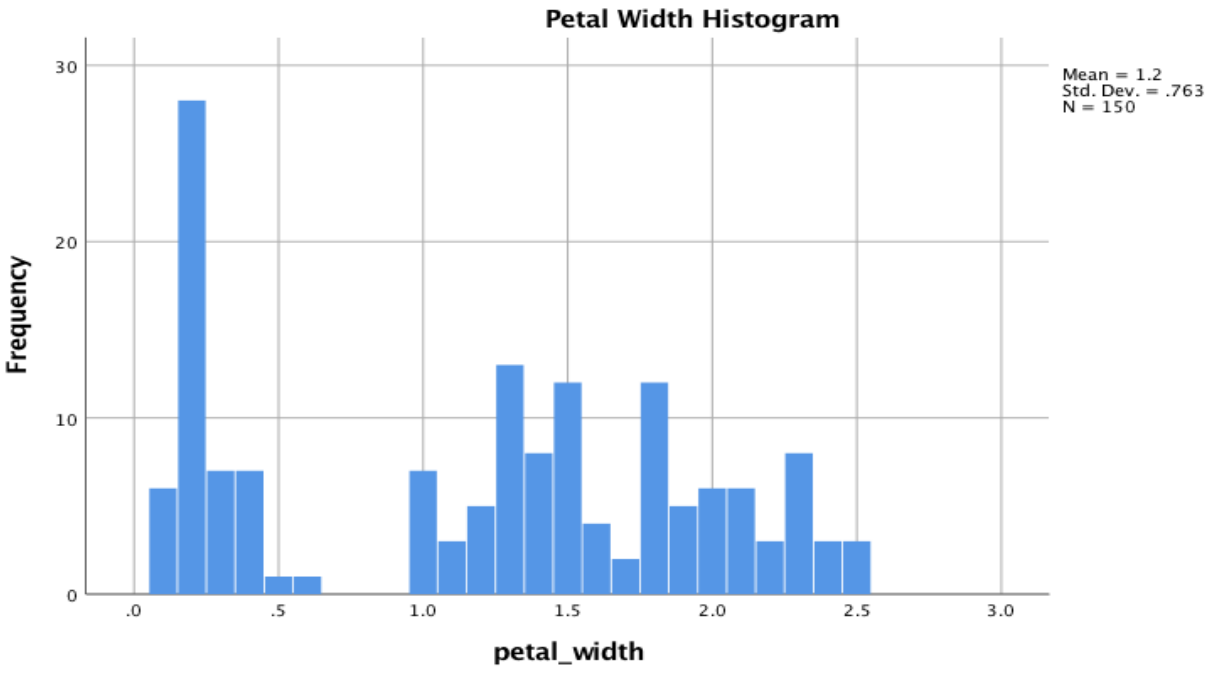
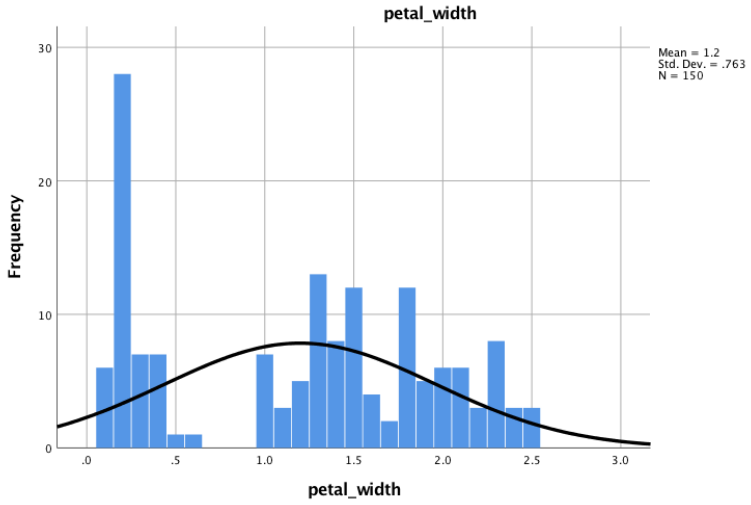


Figure 6: A histogram of petal width.

- The petal width histogram shows lengths in range 0.1cm to 2.5cm, a mean of 1.2cm, std of 0.763, median of 1.3cm, Q1 at 0.3cm and Q3 at 1.8cm for 150 plant samples.



- Again the histogram for petal width does not seem to be composed of a single bell curve. There is variation amongst the classes. This variation is interpreted below.

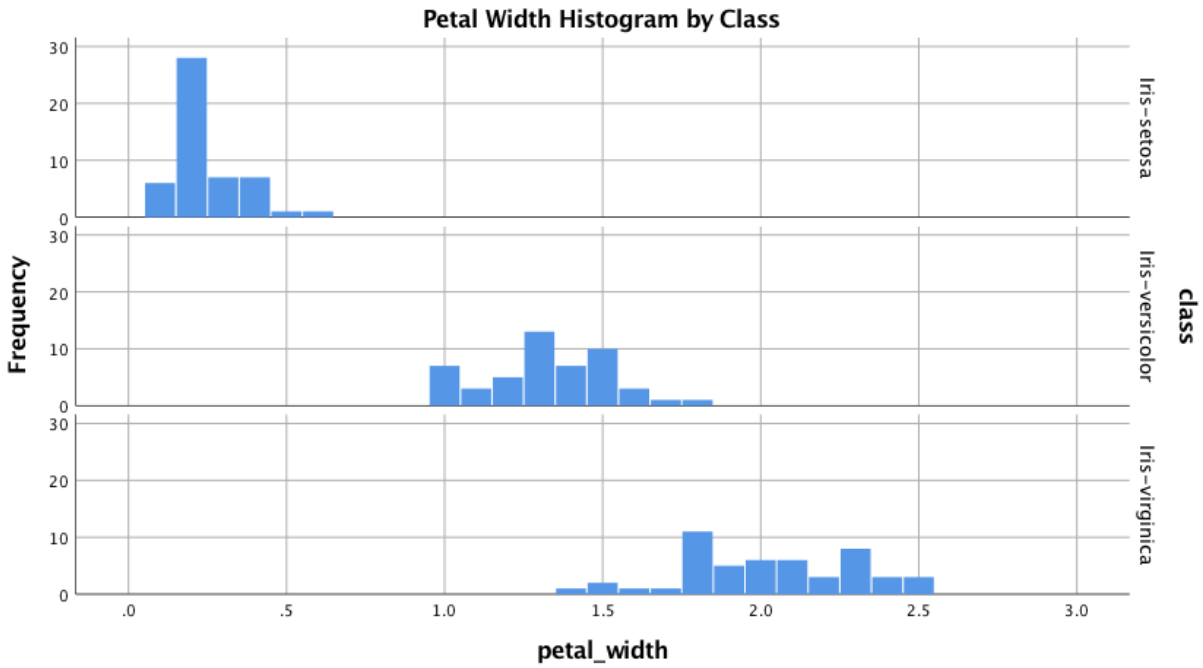


Figure 6.1: A histogram of petal width panelled by class.

- Here the three classes have distinct ranges of petal widths and unique distributions. Just examing at the histograms qualitatively it is clear that setosa have smaller petal widths, versicolor have medium width petals, and virginica generally have larger width petals. After eliminating outliers in the next step, this distinction will become more clear.

d. Determine if there are any outliers in the data with respect to the sepal length.

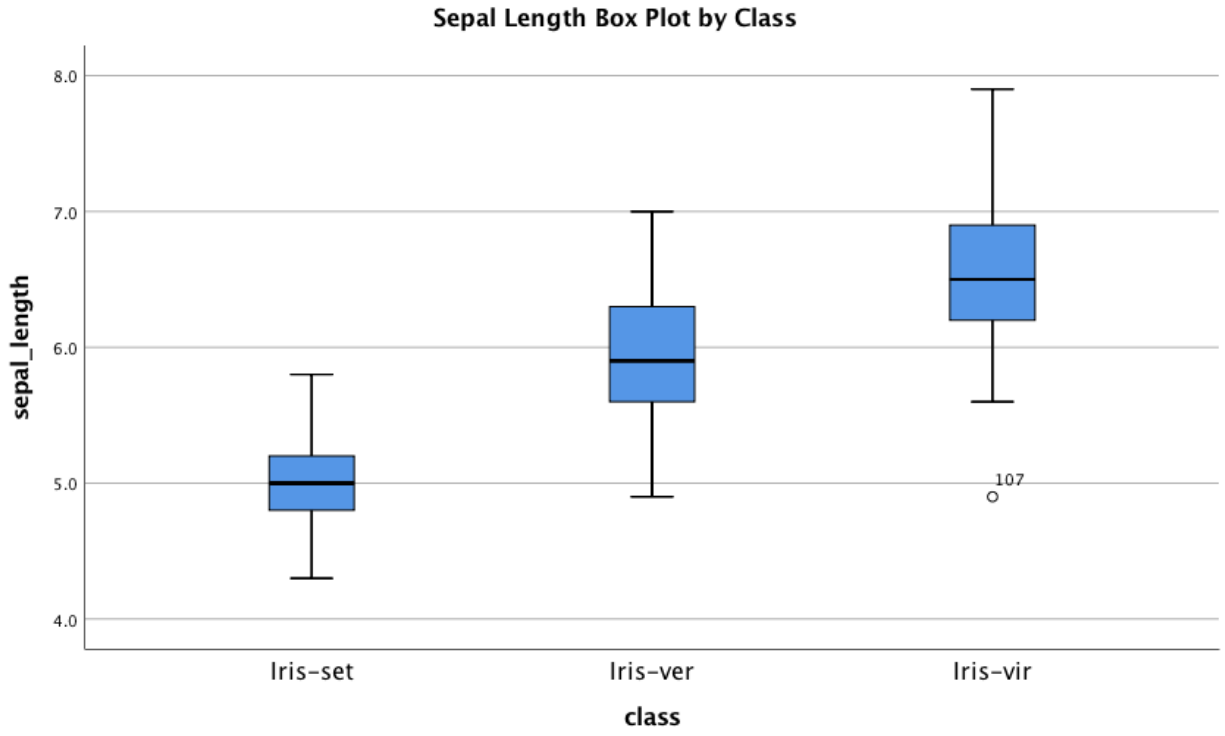


Figure 7: This is a box plot of the sepal length by class. Here we can see that there is an outlier in the class Iris-virginica.

- Record Number 107 is identified as an outlier: [**4.9**, 2.5, 4.5, 1.7, Iris-virginica]
    - It is important to note that this outlier is detected using the rule: values more than 1.5 IQR's but less than 3 IQR's from the end of the box are labelled as outliers. If there are values beyond 3 IQR's from the end of the box they will not show on this graph.

e. Repeat d. for the petal length.

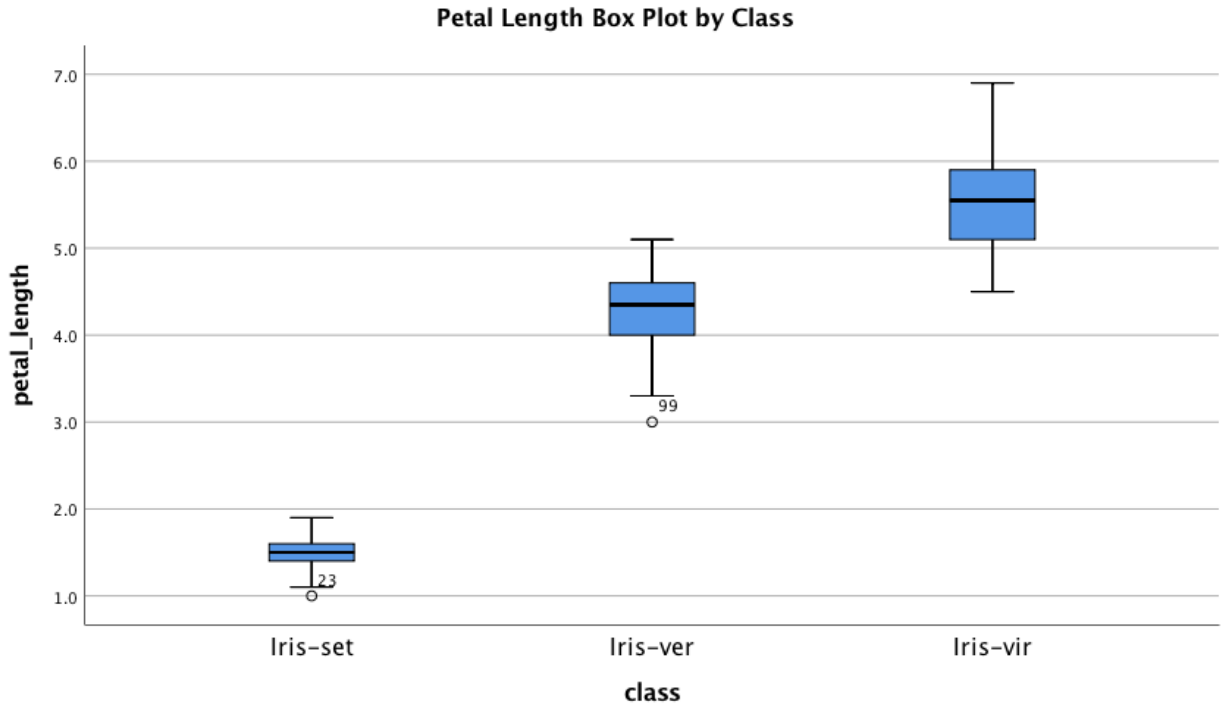


Figure 8: This is a box plot of the petal length by class. Note that outliers have been identified in the Iris-setosa and Iris-versicolor classes.

- Record Number 23 is identified as an outlier: [4.6, 3.6, **1.0**, .2, Iris-setosa]

    - Given the 1.5 IQR rule for determining outliers, the value of 1.0cm for petal length in the class Iris-setosa has been labelled an outlier. In a larger dataset, this record may not have been labelled as an outlier, as we could expect a wider range of petal lengths from the Iris-setosa class. The histogram for Iris-setosa petal length shows a fairly normal distribution, so this could be an extreme case. For the purpose of this assignment though it is still an outlier.

- Record Number 99 is identified as an outlier: [5.1, 2.5, **3.0**, 1.1, Iris-versicolor]
    - Again, this point falls outside of the 1.5 IQR's from Q1 so it is an outlier.

Problem 4 (5 points):

The following paper presented at the *ACM KDD 2017 Workshop on Machine Learning Meets Fashion* showcases an interesting application of data science to fashion and social media: "Identifying Fashion Accounts in Social Networks" by Doris Jung-Lin Lee, Jinda Han, Dana Chambourova, and Ranjitha Kumar:

https://kddfashion2017.mybluemix.net/final_submissions/ML4Fashion_paper_21.pdf

Read the paper and briefly answer the following questions:

1. **What was the data used for the study? Include descriptions on the type of data and the size of the data.**

    a. The authors collected an initial dataset of 10230 potentially fashion related twitter accounts. They started with 12 definitively fashion-related twitter accounts and then collected further accounts using a content-based, snowball-sampling technique. Their sampling algorithm identified these 10230 potentially fashion-related twitter accounts by checking to see if the number of fashion words in a user's tweets exceeds a threshold and if they have the word 'fashion' in their profile, and then expanding out to accounts they follow.

    b. For the classification step, they use features from each individual's most recent tweets (up to 3200 tweets) and account information. Three features were identified in the paper. The first is the normalized fashion counts – the number of fashion-related words found in all of an account's tweets divided by the total number of words in all of an account's tweets. The second is the number of tweets made by the account. The third is the user's profile description – which is checked to see if it contains the word 'fashion'.

    c. Ground-truth labels for the 10230 accounts were obtained via Amazon Mechanical Turk, with a total of 30510 respondents. If at least 2/3 of respondents called an account 'fashion', it was labelled as such.

    d. The data they used for the study came from Twitter accounts in the form of words (strings), as well as Amazon Mechanical Turk labels of the accounts as 'fashion'/ 'not fashion'/'inaccessible (Boolean or string).

2. **Was the data preprocessed or cleaned before applying any modeling techniques?**

    a. Yes, though the nitty-gritty details of how they stored and integrated the data has been left out of the paper.

b. Preprocessing and data reduction – They made use of a filtered fashion vocabulary, non-discriminative words, and stop words so that their model wouldn't need to look at every word, but only relevant ones. For the final classifier, it appears that they were only storing word counts, tweet counts, and label values for 'fashion' or 'not fashion' (probably either Boolean or integer, 1 or 0).

c. Data cleaning – Accounts labelled 'inaccessible' were removed from the dataset. Some LaPlace smoothing was done with the Naïve Bayes and they used grid search to determine optimal C and gamma values for the SVM.

3. **Did the authors solve a classification, a prediction, or a clustering problem as part of the pattern discovery stage? Justify your answer.**

    a. Classification

    b. They solved a classification problem. Specifically, they wanted to see with what precision they could classify a twitter account as 'fashion' or 'not-fashion' based on features of the accounts including the use of fashion-related words, having the word 'fashion' in their profile descriptions, and the number of total tweets the user posted. This is a classification task and solution not only because they described it as such, but also because it involves predicting class/categorical labels ('fashion'/ 'not-fashion').

    c. Ultimately, this technique may bring value to retailers who are trying to stay abreast of emerging fashion trends online, with the assumption that social media fashion accounts can be used to keep up with these trends and knowing which accounts are fashion accounts can streamline analysis.

4. **For the problem identified, which algorithm(s) the authors use to solve that problem?**

    a. Naïve Bayes and Support Vector Machines (SVM)

    b. Regularization was accomplished for NB using LaPlace smoothing and the SVM used an RBF kernel coefficient gamma equal to 0.05 with regularization constant C equal to 4. Optimal values were found using grid search.

c. The results of applying NB and SVM to a sample of 5500 non-fashion accounts and 2734 fashion accounts reveal a fairly high precision but a very low recall score. The authors explain this low recall by explaining that media heavy accounts sometimes had very low fashion-word counts. A user might post pictures of fashion items, include links to fashion sites and other fashion users' pages, etc. So despite an account being clearly fashion related, an account might not be classified as such by both the NB and SVM algorithms. They also identify fashion accounts posting about non-fashion related things as problematic for their classifier.