02450 PROJECT 1: ABALONES

Alexander Philip Hoffmann Laukamp (s204092), Mikkel Albeck-Madsen (s204132),Vidisha Sinha (s204081)

GROUP 125

Distribution of work	Vidisha	Alexander	Mikkel
1. Description of data set	60%		40%
2. Description of attributes in the data		100%	
3. Data Visualization and PCA (data visualization)	100%		
3. Data Visualization and PCA (PCA)			100%
4. Discussion		100%	
Exam Problem 1 and 2			100%
Exam Problem 3 and 4	100%		
Exam Problem 5 and 6		100%	

1. DESCRIPTION OF DATA SET

Abalones are a type of marine snail which are considered to be a rare and expensive seafood delicacy around the world. The majority of this species is found in cold waters, such as the coast of New Zealand, South Africa, Australia, Western North America, and Japan. They are considered to be an endangered species in many countries, and therefore their documentation is important.

The age of an abalone is determined by "cutting the shell through the cone, staining it, and counting the number of rings through a microscope" [1]. In the same way, the sex of the abalone has to be individually observed, which is a time and effort consuming task. Therefore, the goal of this paper is to use machine learning techniques and a data-set containing information about different characteristics of an abalone - including its sex, length, diameter, height, number of rings (representative of age), and weight in order to predict an abalone's age or sex based on its physical attributes.

The data-set used was originally owned by the Marine Resources Division in Tasmania and it collects its values from Abalones found in Tasmania. It was first published on UCI Machine Repository on $1995 - 12 - 01$ when there was a non-ML research conducted to study the "population biology of backlip abalone in Bass Strait". The paper "the Population Biology of Abalone (Haliotis species) in Tasmania" surveyed biological information of backlip abalones to determine a legal minimum size of stocks such that the egg production of the population is not overly affected. Residual analysis showed that 'fecundity', or fertility, is positively correlated with the size of the abalone and negatively correlated with its age [2].

As mentioned before, this paper will use supervised machine-learning techniques of regression and classification on the data-set to predict the number of rings and sex of an abalone, respectively, from its physical characteristics. Regression will use all attributes except 'rings' as an independent variable, and classification will use all attributes except 'sex' as independent variables.

A number of data transformations will be performed for the machine learning tasks. First off, each observation $\boldsymbol{x}$ is standardized by substracting the mean observation, $\boldsymbol{\mu_x}$, and dividing by the vector containing each attribute's standard deviation, $\boldsymbol{\sigma}$. $\tilde{\boldsymbol{x}} = (\boldsymbol{x} - \boldsymbol{\mu})/\boldsymbol{\sigma}$. Standardizing the data is generally a good practice, but it especially helps if we wish to reduce the dimensionality of the data via PCA (which requires standardization) and base our ML models on the reduced data.

A second data transformation is the transformation of the nominal data attribute "Sex" via one-out-of-K coding. Instead of having each sex represented by a single scalar $x_1 \in \{1, 2, 3\}$, each sex will be represented a 3-dimensional binary vector, where the i'th entry is 1 if $x_1 = i$ and the remaining elements is zero. For example, $x = [2]$ will be transformed to $\tilde{x} = [0, 1, 0]$.

2. DESCRIPTION OF ATTRIBUTES IN THE DATA

The attributes in the data are listed in the table below:

Attribute	Type of attribute	Unit
Sex	Discrete and Nominal	–
Length	Continuous and Ratio	mm
Diameter	Continuous and Ratio	mm
Height	Continuous and Ratio	mm
Whole weight	Continuous and Ratio	grams
Shucked weight	Continuous and Ratio	grams
Viscera weight	Continuous and Ratio	grams
Shell weight	Continuous and Ratio	grams
Rings	Discrete and Ratio	–

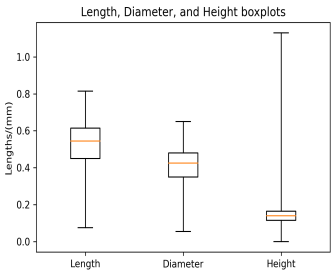
It should be noted that the data set isn't missing any values, i.e., for each abalone there is a value for each attribute.

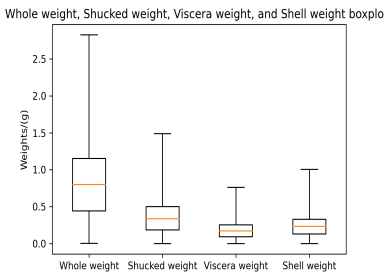
In the following table are the attributes with their respective mean $\bar{x}$ and standard deviation s_x:

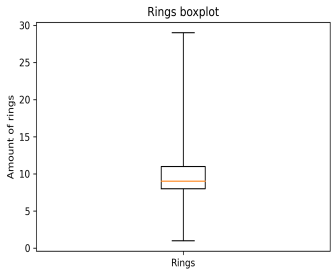
Attribute	$\bar{x}$	s_x
Length	0.5240 mm	0.1201 mm
Diameter	0.4079 mm	0.0992 mm
Height	0.1395 mm	0.0418 mm
Whole weight	0.8287 g	0.4904 g
Shucked weight	0.3594 g	0.2220 g
Viscera weight	0.1806 g	0.1096 g
Shell weight	0.2388 g	0.1392 g
Rings	9.9337	3.2242

It should be noted that the continuous values of the data have all previously been scaled by dividing by 200 (by the original owners of the data-set), so low values are in reality much greater. Looking at the table, it can be seen that the weight attributes have a large standard deviation - they're more than half of the value of their respective mean. Furthermore, it should be noted that 36.58% of the abalones are male, 31.29% are female, and 32.13% are infants.

The following are boxplots of each of the attributes except the Sex attribute:



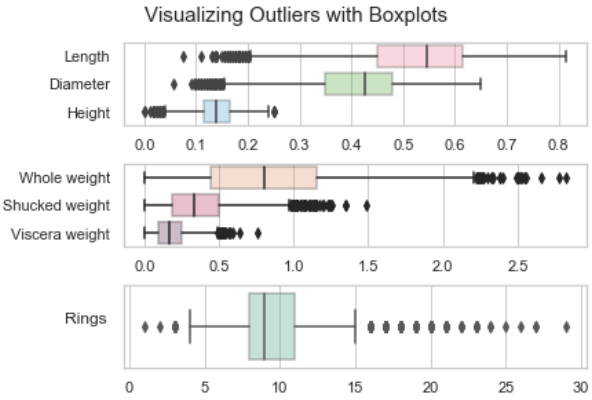




An interesting thing to note here is that despite having the lowest mean and the lowest standard deviation, the Height attribute, manages to have the highest maximum value of the three attributes that make up the dimensions of the abalone. This could possibly suggest that there are some outliers within this attribute. It's also interesting to note, that all the attributes that make up the weights of the abalone have a longer upper whisker, whilst the Length and Diameter attributes both have longer lower whiskers.

3. DATA VISUALIZATION AND PCA

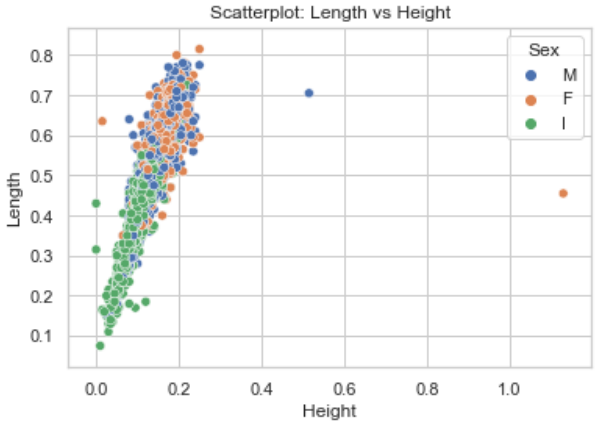
This section uses visualization techniques to further analyse the attributes in the data. The following modified box-plot can be used to assess if there are problems with outliers in the data. Modified box-plots define maximum whisker length (MWL) as $1.5 \times IQR$ (Inter-Quartile Range $= Q_3 - Q_1$), and any value outside the range $[Q_1 - MWL, Q_3 + MWL]$ is considered to be an outlier.



The above figure shows that all attributes have some outliers. 'Length' and 'diameter' attributes only have outliers before the first quartile, which shows that unlikely abalones tend to be smaller and thinner (in terms of diameter).

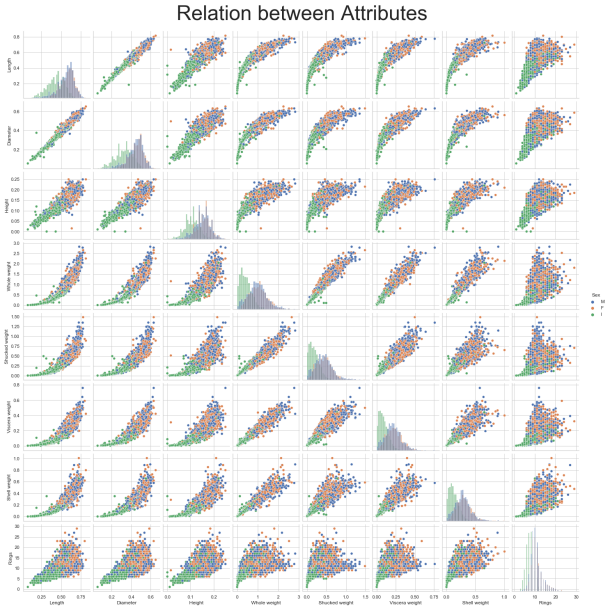
Outlier observations in weight related attributes exist only after the third quartile, which shows that some unlikely abalones exist - which have a slightly higher weight.

The 'height' attribute shows that unlikely abalones have slightly smaller or slightly larger height values. However, there are two outliers which are really far out (greater than 0.4×200 mm). If a scatter plot is created with 'height' and other variables, the two outliers stand really far out from the general cluster created. As an example, the scatter-plot for 'Height' vs 'Length' is shown below.



The outliers above affect the variance and skew the data-distribution towards the right - which makes it difficult to assume normality, and affects multiple linear regression negatively because the process of finding the general pattern can be highly impacted by extreme values. Therefore, the two observations with a 'height' greater than 0.4 are removed from the data-set.

To further visualize the effect of two variables on each other, scatter-plots for different combinations of attributes are plotted. The diagonal of the following 'pair-plot' contains histograms which show the distribution of the attribute based on the sex of the abalones (green for infants, blue for males, and orange for females). A larger graph is present in Appendix A at the end of this paper for ease of readability.



The scatter-plots conveys some interesting conclusions. It can be seen that infants (green points) occupy the lower left corner of all attributes - which confirms the logical idea that infants have smaller values when it comes to physical characteristics as compared to more matured abalones. Males and female abalones tend more towards the upper right corner - therefore more matured abalones have greater lengths, heights, etc. All attributes seem to follow a general normal distribution as shown by the histograms. Moreover, it can be generally seen that the distributions for males and females overlap each other, and those for infants is shifted towards the left of them.

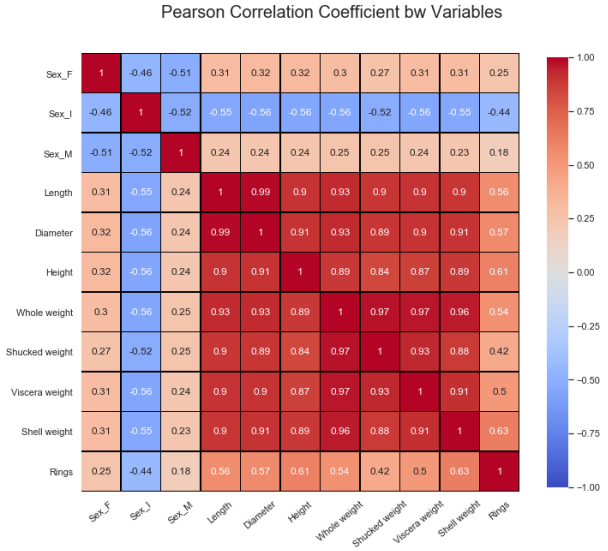
Attributes of 'diameter', 'length', and 'height' are more or less

normally distributed for infants, but are skewed a little more towards the left for males and females - which again shows that there are smaller-valued outliers present in these attributes. On the other hand, attributes of weight show distributions which are more skewed towards the right - for infants way more than males and females (the distributions of males and females are very alike). The total distribution of number of 'rings' is skewed towards the right, and infants make up the left side of the values.

It is possible to see how different attributes vary with each other by the scatter-plots. However, to get an idea of the difference in magnitude of relation, the Pearson correlation coefficients for the different combinations are found. The Pearson coefficient ($-1 \leq \rho_x y \leq 1$) measures the linear relation between two variables (x and y), and it can be found using the following formula.

$$\rho_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

A heat-map is made with the Pearson coefficients found, as shown below. A process of 'one-hot-encoding' is done for the 'Sex' attribute, so it is possible to see how different types of 'Sex' vary with the different attributes. A larger heat-map is present in Appendix B for easier reading.



The heat-map above shows that infants vary moderately negatively with all other attributes. Females seem to have a stronger positive relation with all other attributes than males do - however, the magnitude of these relations are only weak.

The number of rings moderately positively varies with the length, diameter, height, and weight of the abalones. This means that generally, higher values of the mentioned attributes points towards a greater number of rings.

Finally, it is possible to see a very strongly positive relation between attributes of length, diameter, height, total weight, and different forms of weights. This of course gives logical conclusions like smaller abalones are lighter and heavier abalones tend to be larger.

The above visualizations show that the ML models chosen by this paper are feasible. For example, while doing multiple linear regression, it is important to keep in mind that this supervised ML technique works best when the relationship between training data is linear. The values of Pearson coefficients show a generally linear trend amongst variables, which means that multiple linear regression will work well.

Very extreme outliers were present in the 'height' attribute only, which have been removed. This improves the sensitive linear regression task. Furthermore, if a Bayes' classifier is chosen for the classification task, then the removal of extreme outliers (that increases normality of data) would be a good thing.

A topic of concern is the fact that many independent variables are heavily correlated with each other ($|\rho| > 0.9$) - also known as multicollinearity. Multicollinearity can lead to misleading or skewed results and usually some of these attributes should be dropped. Instead of doing this, Principle Component Analysis can be performed on the data-set.

Before PCA is carried out, the data set is transformed. Since PCA is not appropriate for nominal data, the attribute "Sex" is ignored for PCA. This reduced data set is represented in the data matrix, $\mathbf{X}$. The data will now be standardized. Let $\boldsymbol{\mu}$ denote the mean observations, where μ_j denotes the mean value for attribute j across all observations, and let $\boldsymbol{\sigma}$ denote the standard deviation vector, where σ_j is the standard deviation for attribute j across all observations. Then every observation, $\mathbf{x}_i$, is standardized by:

$$\mathbf{y}_i = \frac{\mathbf{x}_i - \boldsymbol{\mu}}{\boldsymbol{\sigma}}$$

where division between vectors is understood as element-wise division. All $\mathbf{y}_i$ is collected in the matrix, $\mathbf{Y}$.
PCA is carried out using Singular Value Decomposition to obtain the eigenvectors and eigenvalues.

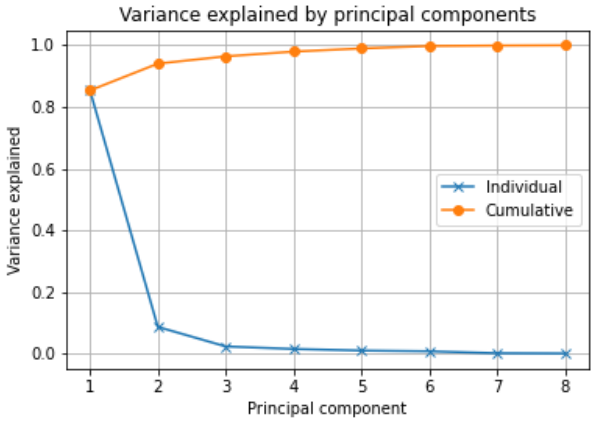
$$\mathbf{Y} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$$

Where $\mathbf{V}$ is a matrix, where the i'th column, $\mathbf{v}_i$, is the eigenvector with the i'th greatest eigenvalue. $\boldsymbol{\Sigma}$ is a diagonal matrix containing the eigenvalues in descending order (such that the i'th diagonal element is the eigenvalue corresponding to $\mathbf{v}_i$.)

Now, the explained variance of each principal component is examined. The cumulative variance explained by the k'th first principal components is computed by:

$$\text{Variance Explained} = \frac{\sum_{i=1}^{k} \sigma_i^2}{\sum_{i=1}^{M} \sigma_i^2}$$

Below, the explained variance by each PC, as well as the cumulative explained variance is shown as a function of the number of PCs



I can be seen that the variance explained by PC1 is 85.4% and that variance explained by PC2 is 8.7%. So a one-dimensional subspace through the transformed data explains the vast majority of variance

in the data, and by just using PC1 and PC2, the number of dimensions in the data can be reduced from 8 to 2 while retaining 94.1% of the data. All this could indicate strong linear relations between all attributes, which is consistent with the pair plot and the pearson correlation table.

The coefficients of the first principal components can reveal more about these relations. The coefficient of PC1 and PC2 is seen below

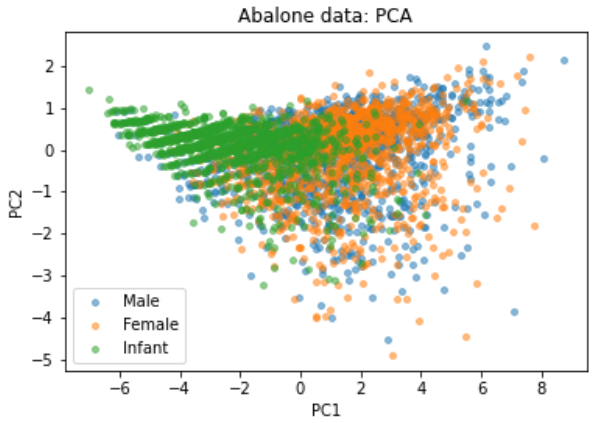
Attributes	Coef. of PC1	Coef. of PC2
Length	0.369	0.069
Diameter	0.370	0.041
Height	0.359	-0.068
Whole weight	0.375	0.138
Schucked weight	0.359	0.300
Viscera weight	0.366	0.173
Shell weight	0.368	-0.045
Rings	0.241	-0.921

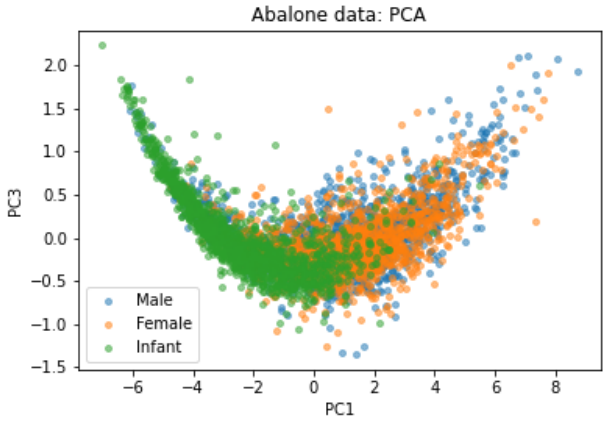
All PC1's coefficients except "Rings" are around ≈ 0.360, which indicates that all these attributes have some fundamental linear relationship - consistent with the pair plot. The "Rings"-coefficient is slighty lower at 0.241, which indicates that this linear relation is not quite as strong, which also correlates with pair plot. In the pair plot, the "Rings"-attribute forms wider clusters compared to other attributes' relations. This is also seen by PC2's "Rings" coefficient of -0.921, whose magnitude is by far the greatest. This means that the second direction of most variance in the data relates the variation given by the "Rings" attributes. So PC1 seems to capture the general size of the abalones, while PC2 seems to capture the number of rings of the abalones

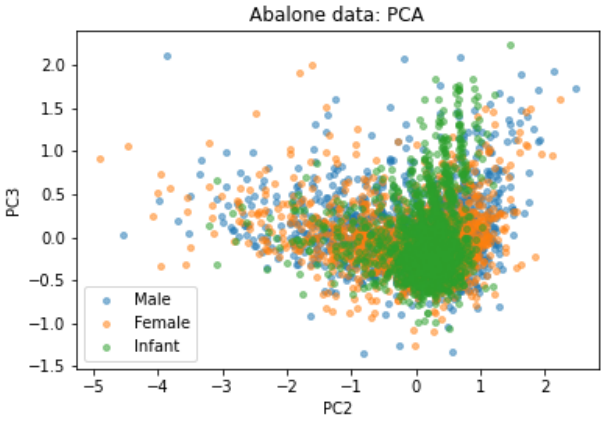
Now, the data is projected onto the principal component:

$$\mathbf{Z} = \mathbf{YV}$$

Where $\mathbf{Z}$ is the matrix containing the projected data. Below is three plots of the projections onto PC1 and PC2, PC1 and PC3, and PC2 and PC3:



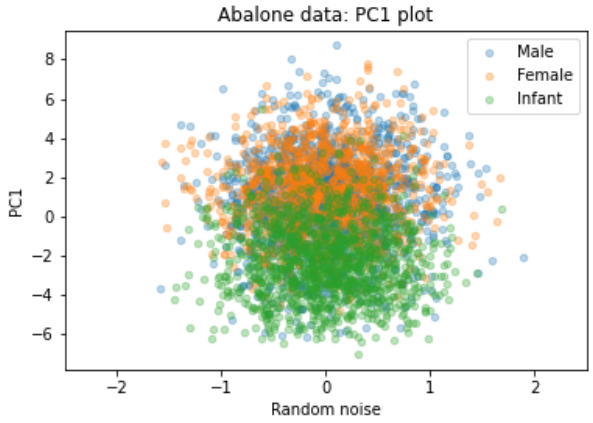




From the "PC1-PC2" plot, there is a lot of overlap between sexes, and every sex seems to be represented in every part of the plot. However, the left of the plot seems to contain more infants, the middle of the plot contains more females, and in the rightmost part of the plot, males might be more represented. In the left part of the plot, some horizontal lines can be seen in the data. This is most likely the "Rings"-attributes, which is discretely valued.

The "PC1-PC3" plot reveals an interesting "smiling mouth"-shape. The same overlap and general tendencies from the "PC1-PC2" can be seen in this plot. In the final "PC2-PC3"-plot, all data seems to be bunched up around the origin with very little patterns to be seen in data or between sexes, which makes sense since PC2 and PC3 only explains 11% of the total variance.

Since PC1 explains the vast majority of the variance, the projection of the data onto PC1 is plotted below. The projected value is seen along the y-axis, and random normal noise along the x-axis to make dots distinguishable.



The general trends from the "PC1-PC2" can still be seen here: Great overlap between sexes, and big variation for each sex. However, more infants in the bottom, more females in the middle, and more males at the very top is still the vague trend.

4. DISCUSSION

Using the Abalone data set, that has been cleared of missing values and that has had its continuous values scaled, we have visualized the data and done PCA on it. In the pairplot, it could be seen that the attributes are all highly correlated. This would also explain why the first principal component carries almost all of the variance (85.4%). Also in the pairplot, it can be seen that the attributes split by Sex all appear normally distributed, with the males and females being pretty much on top of each other and the infants being behind them. Looking at those bell curves and the boxplots, the skewness of the boxplots matches the skewness of their respective normal distribution. However, in the pair plot, it could also be seen that the Height attribute had two outliers, an issue that was already seen as being a possibility from the boxplots, and we therefore removed those points.

Since the attributes appear to be linearly correlated with one another, linear regression seems to be an obvious possibility for modeling the data. Furthermore, they all appear normally distributed, which fulfills the assumptions of many ML models, and it also helps when using probability based ML models, because we can confidently model probabilities on unseen data by fitting univariate or multivariate normal distributions on the training data set.

5. REFERENCES

[1] UCI Machine Learning Repository, "Abalone dataset," .

[2] Simon R. Talbot Andrew J. Cawthorn Warwick J. Nash, Tracy L. Sellers and Wes B. Ford, "The Population Biology of Abalone (Haliotis species) in Tasmania. I. Blacklip Abalone (H. rubra) from the North Coast and the Islands of Bass Strait, journal =," .

6. PROBLEMS

Problem 1

Here is a walk through of the attribute types in the data set:

x_1: Interval. Since real time life differences is proportional the the difference between their corresponding coded differences, x_1 is interval. And since $x_1 = 0$ is not defined and wouldn't represent an absence of time, x_1 cannot be ratio.

$x_2, x_3, x_4, x_5, x_6, x_7$: Ratio. Since every of these attributes represent a number of events, a value of 0 represents the absence of an event. It also makes sense to say that the number of accidents at some time day is 2.5 times greater than the number of accident at some other time. This means that these attributes are ratio.

y: ordinal. $y_i = 3$ represent a higher congestion level than $y_j = 1$, but it doesn't make sense to say that the congestion level is 3 times greater. $y = 0$ is undefined. And the difference in congestion level $y_i - y_j = 2$ doesn't contain any usable information. Therefore, y is ordinal.

The only answer corresponding to the assessments above is D. The answer is D.

Problem 2

The definition of the p-distance,$d_p(x, y)$ is given by:

$$d_p(\mathbf{x}, \mathbf{y}) = \begin{cases} (\sum_{i=1}^{M} |x_i - y_i|^p)^{1/p} & 1 \leq p \leq \infty \\ \max\{|x_1 - y_1|, ..., |x_M - y_M|\} & p = \infty \end{cases}$$

Using the definition, the following p-distances can be calculated on the 14th and 18th observation, $x_{14} = [26, 0, 2, 0, 0, 0, 0]$, $x_{18} = [19, 0, 0, 0, 0, 0, 0]$. Results is seen below:

$$d_{p=\infty}(x_{14}, x_{18}) = 7.000$$
$$d_{p=3}(x_{14}, x_{18}) = 7.054$$
$$d_{p=1}(x_{14}, x_{18}) = 9.000$$
$$d_{p=4}(x_{14}, x_{18}) = 7.012$$

Therefore, A is correct!

Problem 3

The explained variance of the first m PCs can be found by the following formula for a data-set with N attributes.

$$\text{variance explained by } PC_m = \frac{\sigma_1^2 + ... + \sigma_m^2}{\sigma_1^2 + \sigma_2^2 + ... + \sigma_N^2}$$

A. Variance explained by the first four PCs is found in the following way:

$$\text{variance explained by } PC_4 = \frac{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2}{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2} \approx 0.87$$

0.87 is greater than 0.8, therefore, A is correct .

B. Variance explained by last 3 PCs is

$$\frac{11.48^2 + 10.03^2 + 9.45^2}{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2} \approx 0.48$$

which is less than 0.51.

C. Variance explained by first 2 PCs is

$$1 - \text{Variance explained by last 3 PCs}$$
$$= 1 - 0.48 \approx 0.52$$

which is greater than 0.5.

D.

$$\frac{13.9^2 + 12.47^2 + 11.48^2}{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2} \approx 0.72$$

which is greater than 0.7.

Problem 4

For generally positive projections, we want greater values of features with positive coordinates and smaller values of features with negative coordinates in the PC (and vice versa for generally negative projections).

A. Incorrect, because described scenario leads to negative value of projection onto PC5.

B. Incorrect, because described scenario will typically lead to a negative value of projection onto PC3.

C. Incorrect, because coordinate for **broken truck** has a high magnitude so it will have a larger effect on value of projection - typically making it positive.

D. Correct , because there are greater values of features with positive coordinates and smaller values of features with negative coordinates.

Problem 5

The Jaccard similarity is defined as the following:

$$J(x, y) = \frac{f_{11}}{K - f_{00}}$$

Where J(x, y) is the Jaccard similarity, x and y are two documents, K is the total number of words, f_{11} is the number of words that both documents contain, and f_{00} is the number of words that the documents both don't contain.

For this specific problem, the vocabulary size is $M = 20000$, which would correspond to K. Both the documents $s1$ and $s2$ contain the words "the" and "words", meaning that $f_{11} = 2$. There is a total of 13 different words in the combination of $s1$ and $s2$, meaning that the number of words, that they both don't contain, is $f_{00} = 20000 - 13 = 19987$.

This means that the Jaccard similarity between document $s1$ and document $s2$ is as follows:

$$\text{J}(s1, s2) = \frac{2}{200000 - 19987} = \frac{2}{13} = 0.1538546$$

Therefore, option A is correct.

Problem 6

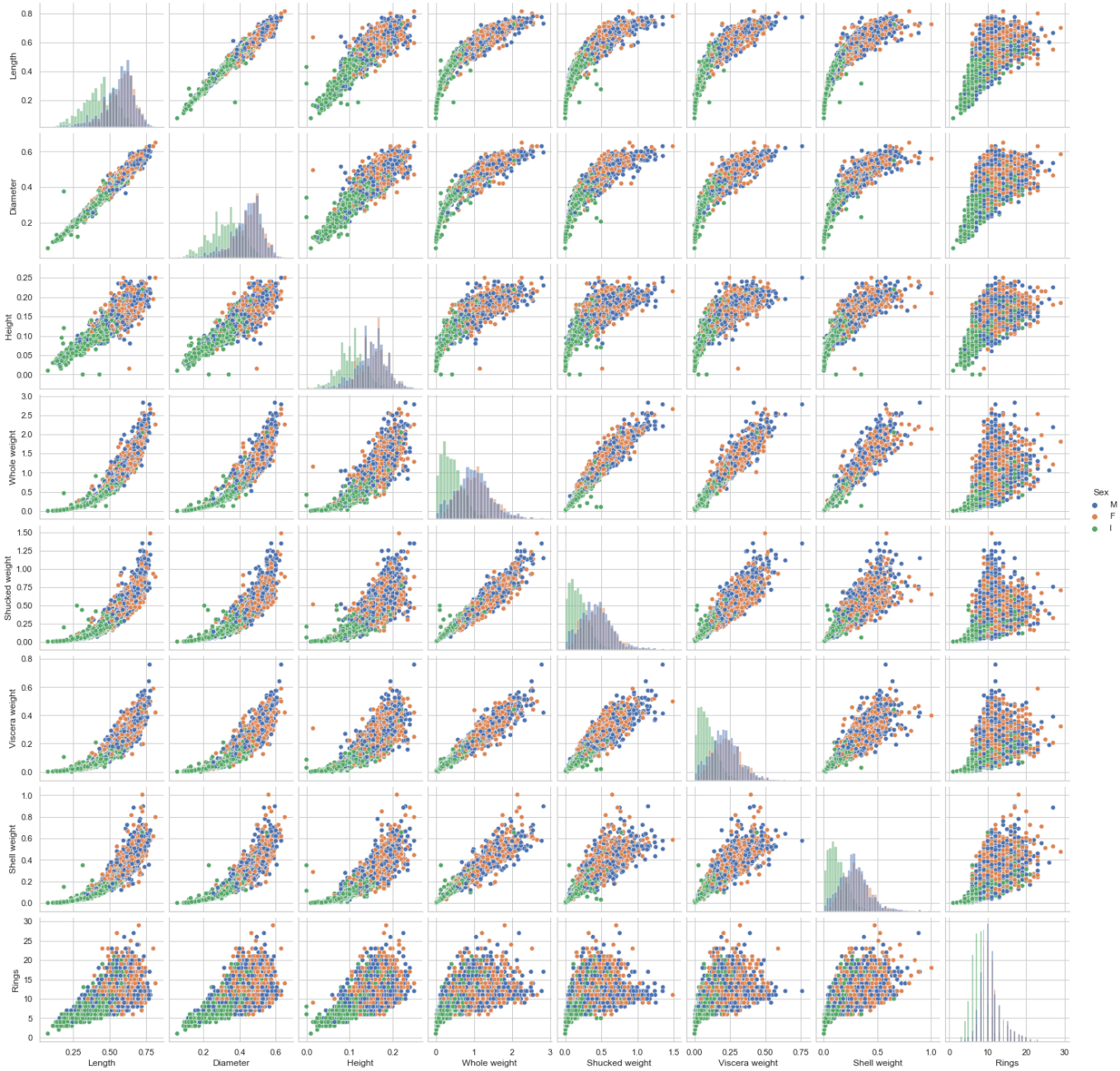
It can be read from table two that $p(\hat{x}_2 = 0, \hat{x}_7 = 0|y = 2) = 0.81$ and that $p(\hat{x}_2 = 0, \hat{x}_7 = 1|y = 2) = 0.03$. Since $\hat{x}_7$ can only take on the values 0 and 1, then

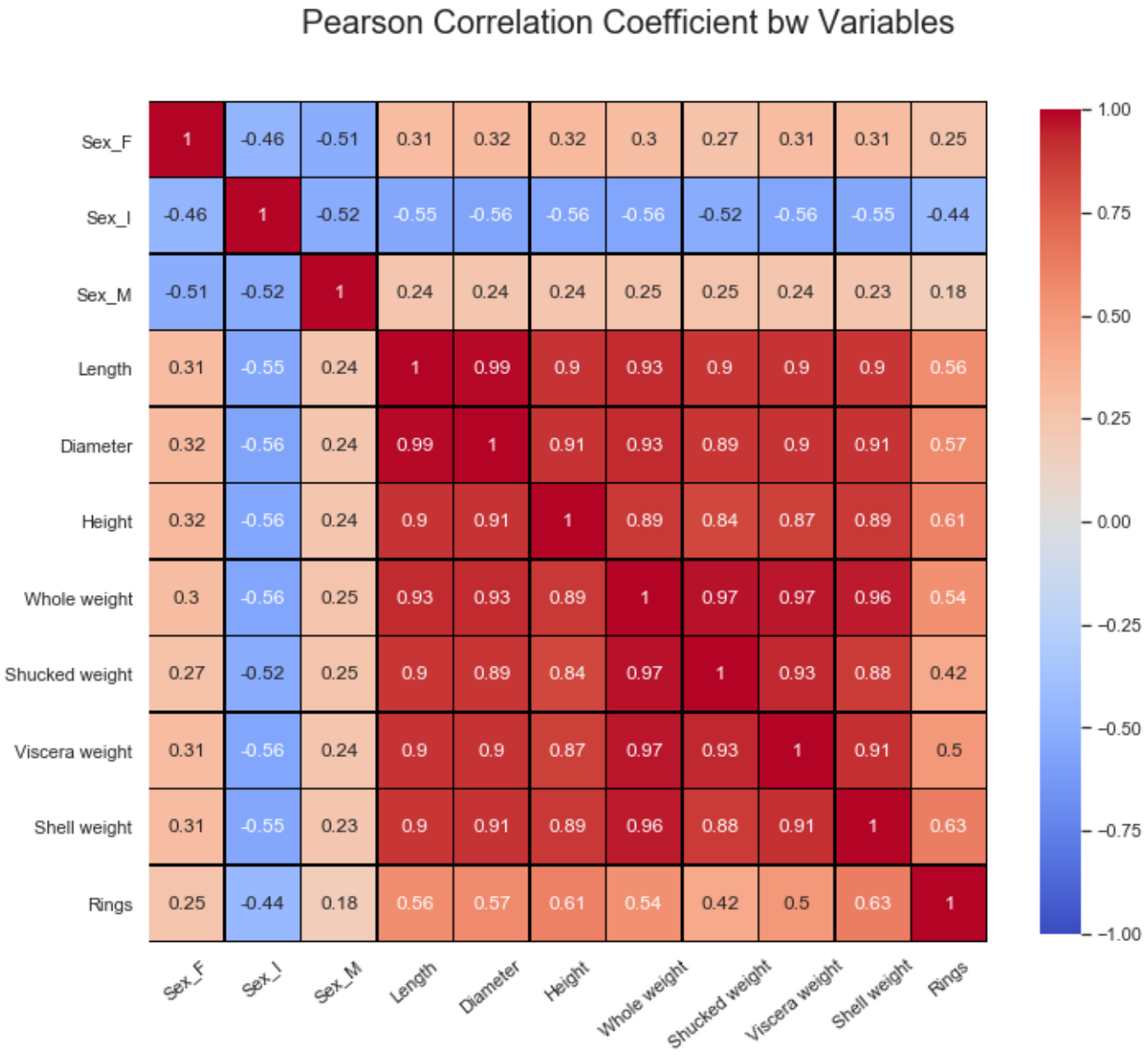
$$p(\hat{x}_2 = 0|y = 2) = p(\hat{x}_2 = 0, \hat{x}_7 = 0|y = 2) + p(\hat{x}_2 = 0, \hat{x}_7 = 1|y = 2)$$

$$= 0.81 + 0.03 = 0.84$$

Therefore, option B is correct.

Relation between Attributes




Pearson Correlation Coefficient bw Variables
Sex_F 1 -0.46 -0.51 0.31 0.32 0.32 0.3 0.27 0.31 0.31 0.25
Sex_I -0.46 1 -0.52 -0.55 -0.56 -0.56 -0.56 -0.52 -0.56 -0.55 -0.44
Sex_M -0.51 -0.52 1 0.24 0.24 0.24 0.25 0.25 0.24 0.23 0.18
Length 0.31 -0.55 0.24 1 0.99 0.9 0.93 0.9 0.9 0.9 0.56
Diameter 0.32 -0.56 0.24 0.99 1 0.91 0.93 0.89 0.9 0.91 0.57
Height 0.32 -0.56 0.24 0.9 0.91 1 0.89 0.84 0.87 0.89 0.61
Whole weight 0.3 -0.56 0.25 0.93 0.93 0.89 1 0.97 0.97 0.96 0.54
Shucked weight 0.27 -0.52 0.25 0.9 0.89 0.84 0.97 1 0.93 0.88 0.42
Viscera weight 0.31 -0.56 0.24 0.9 0.9 0.87 0.97 0.93 1 0.91 0.5
Shell weight 0.31 -0.55 0.23 0.9 0.91 0.89 0.96 0.88 0.91 1 0.63
Rings 0.25 -0.44 0.18 0.56 0.57 0.61 0.54 0.42 0.5 0.63 1
Sex_F Sex_I Sex_M Length Diameter Height Whole weight Shucked weight Viscera weight Shell weight Rings
1.00 0.75 0.50 0.25 0.00 −0.25 −0.50 −0.75 −1.00