**Introduction**: In this era of modern science and technology, data has become a very powerful tool to solve real life problems. With the advancement in the data science domain like implementing machine learning and artificial intelligence tool in order to solve critical problems, it is now easier to accelerate business decisions and processes. It is even possible to predict the future outcome by analyzing and studying the historical data. That’s why advanced data science techniques are being practice in different industries to improve the business prospects.

In this report, a real life business problem is identified and analyzed in order to understand how data science techniques can help in business decision making, what could be the challenges implementing them in real life business scenario where too many factors are involved, as well as some recommendations are proposed for the entrepreneurs who eagerly want to develop efficient solutions for their business. In the following sections, the business cases and problems are discussed based on a real life dataset, then analyzed with different techniques and a set of challenges and recommendations are presented later in the report. Finally, based on the findings a conclusion wraps up the report.

**Dataset**: The dataset is collected from UCI Machine Learning Repository [1]. The data is revised by removing data examples with missing values and scaling ranges of continuous values from the original study [2] which was done based on the population biology of abalone (Haliotis species) and published by Marine Research Laboratories Taroona, Department of Primary Industry and Fisheries, Tasmania, Australia in 1994. Later the dataset became popular among data science and machine learning practitioners as it is a very useful dataset and can be used to develop advanced algorithms.

The dataset contains different physical attributes of abalones such as their weights in different combinations, sex, length, diameter etc. It has 4177 data instances which is sufficient for data science experiments. The dataset can be used to predict the age of abalone based on their physical structures and measurements. Different relationships can be found in the dataset which require further study. But from different study, we can determine the age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope [2]. Of course, this can be implemented with statistical analysis. But that can become cumbersome and complex. So, if we can implement any data science tool to automate this process it can be a very good contribution to solve a real life business problem.

**Business Prospect**: Abalones can be found in the cold coastal areas like New Zealand, South Africa, Australia, Western North America, and Japan [3]. These marine snails are considered as highly nutritious food and very popular around the world. So the farmers and sellers profit a worldwide business by importing and exporting it. As found in studies [4], the price of an abalone is positively correlated to its age. Best way to estimate an abalone’s age is to cut the shells and count the rings through microscopes. It is very boring, time consuming and sensitive task and nonetheless it is related to business profit. So, if an algorithm can predict the age of an abalone by simple measuring the weights, it would very useful and efficient solution for the business. In this report, we are going to demonstrate that it can be possible with data science techniques.

**Data Analysis**: The chosen dataset contains 4177 observations with nine different features such as variables, namely, Sex, Length, Diameter, and Height, Whole weight, Shucked weight, Viscera weight, Shell weight and Rings. The variable Rings is linearly related to the age of an abalone, as age equals to number of rings plus 1.5. So, we can divide them in 8 descriptive features and 1 target feature. The target feature is the rings of abalone. It is an integer to describe the age of abalone, number of rings add 1.5 gives the age in years of them.

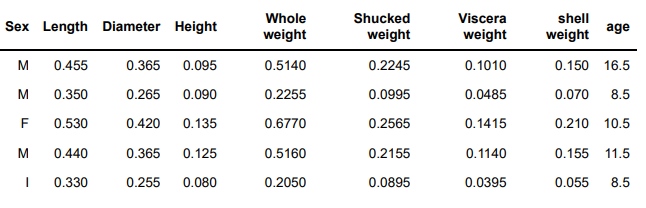


Figure 1: Sample rows from the dataset.

As we can see in the figure 1, the dataset have all numerical features as they represent physical measurements of abalone in float values except the feature Sex. Sex is represented by characters where ‘M’ stands for Male, ‘F’ stands for Female and ‘I’ stands for Immature abalones. The data types always help to know the data before analyzing and studying it.

Before starting with the dataset preprocessing, it is better to see the data distribution for each feature as we want to know that each feature has got equally distributed amount of data so that it is sufficient for our analysis. Figure 2 shows how our data is distributed in each feature which exactly tells us that our data is normally distributed.

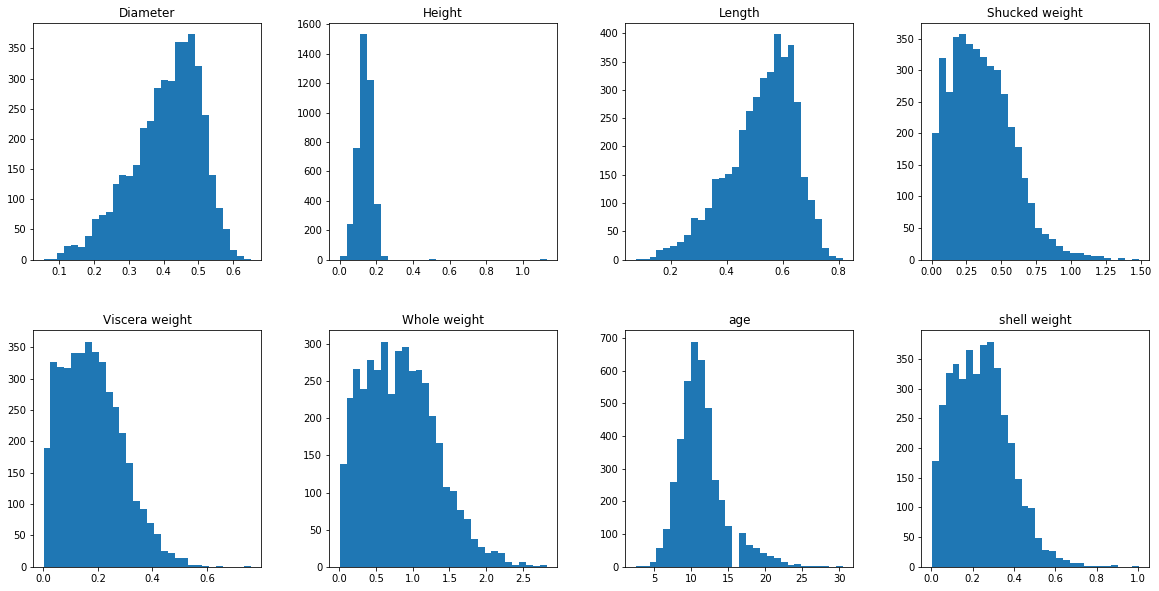


Figure 2: Feature data distribution

It is really good to have normally distributed dataset because it is more reliable in smaller samples. In this case we have 4177 examples and it is sufficient for our data analysis and to make predictions out of the data.

Cleaning the dataset is one of the most important pre-processing step for any data analysis task. By cleaning dataset means removing garbage or unwanted values from the dataset which might create biasness in the predicting model or simply create disturbance in the training process. First step would be check for missing values, missing row values. And removing them or add a random value to it. Fortunately, after studying the dataset, 0% missing data was found. So, the dataset is pretty clean.

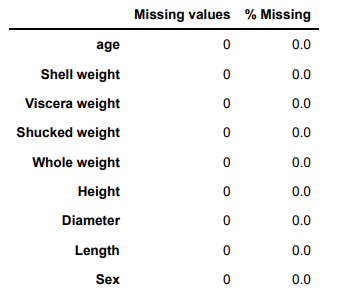


Figure 3: Finding missing data for different features

Another pre-processing technique is to remove outliers or biased data. Biased data points apparently represent very large or small value than the average value throughout all the data points. So, they are not very good example for a model to learn from. Also if majority of the data represents similar kind of data points, we can simply remove the outliers and rely on the majority of the data. In figure 4, we can see the scatter plots of different features except for Sex as it does not contain numerical data. Figure 4 has presented the data points of different features and Age. So we can easily find out the data which are minors and situated outside the average data point zones.

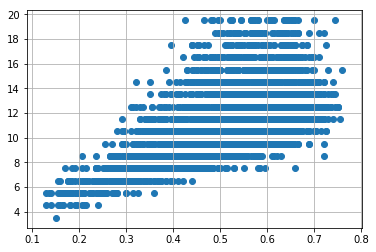
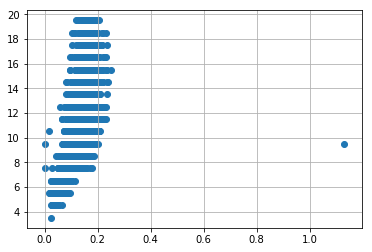
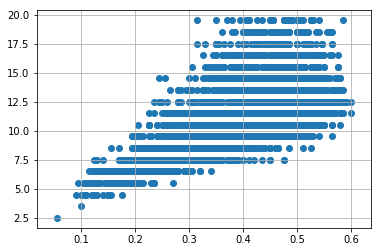
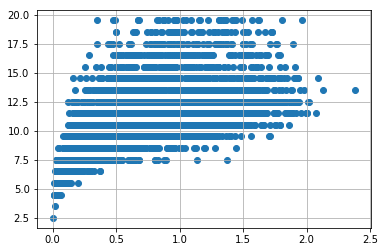
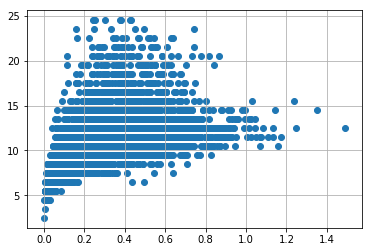
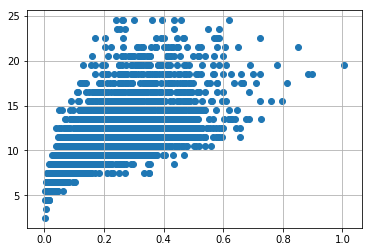
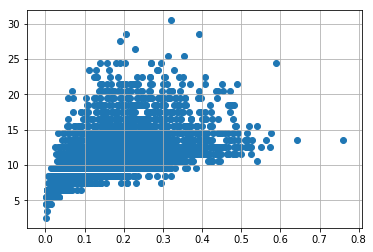


Figure 4: Scatter plots of different features and Age

After the pre-processing of the dataset, we need to find out the correlations between the features. Because the models we are train to predict outcomes will eventually find out these correlations. We are also supposed to see some correlations by just looking at the data so that we can say that the data is good for predicting.

Figure 5 shows a pair-plot for the features which can help us to perform bivariate analysis. Bivariate analysis tells us how each features are affected in presence of other features. It is very helpful for hyper parameter tuning and fine tuning the model for better accuracy. It also helps us understand and identify significance features, overcome multi-collinearity effect, inter dependency and thus, provides insights on hidden data noise pattern.

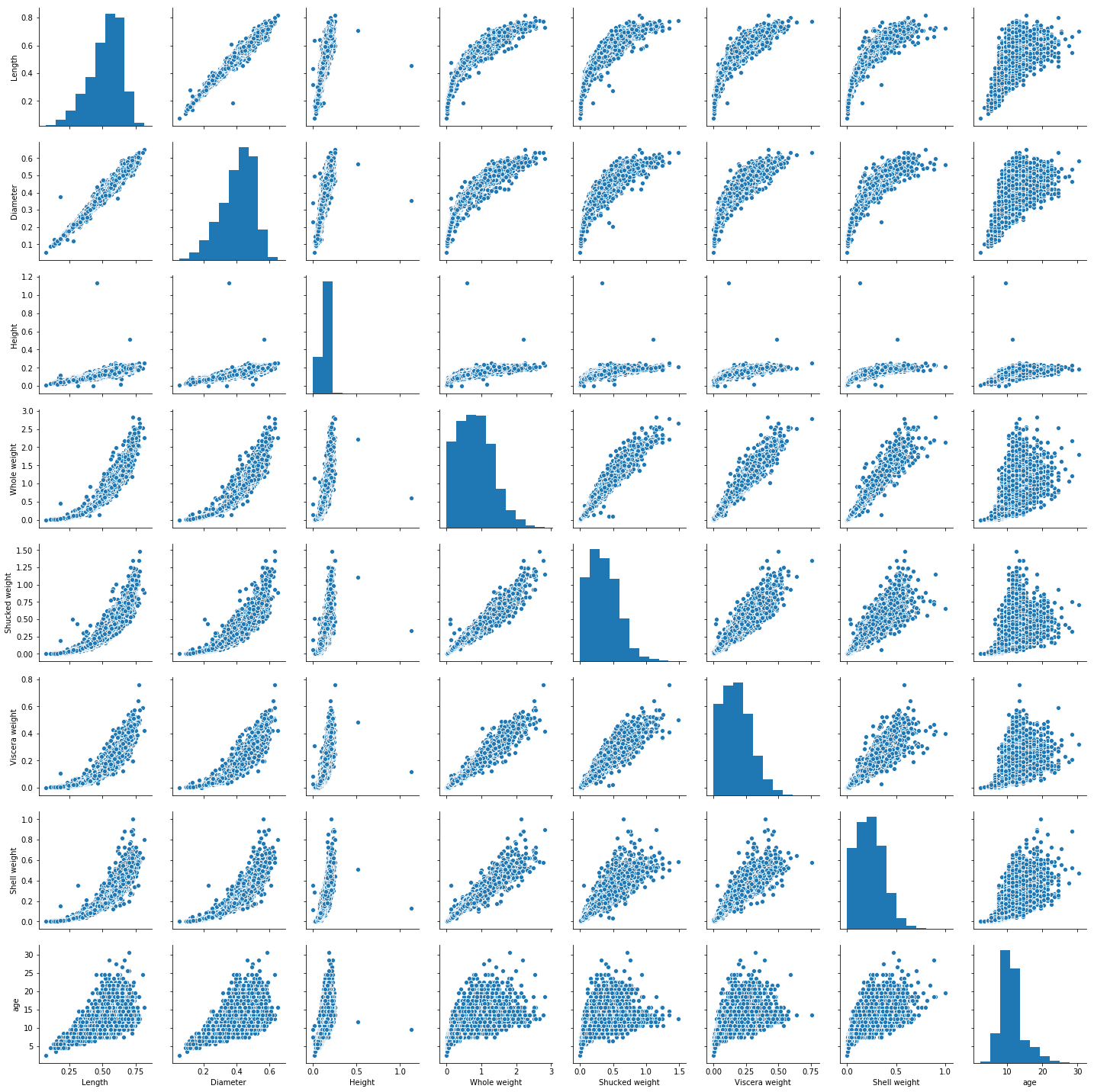


Figure 5: Pair-plot for bivariate analysis

A heat map also represent the correlations between the features with colors. So, it is easier to visualize with colors and numbers rather than scattered plots from pair-plot. They both are very useful visualization tools.

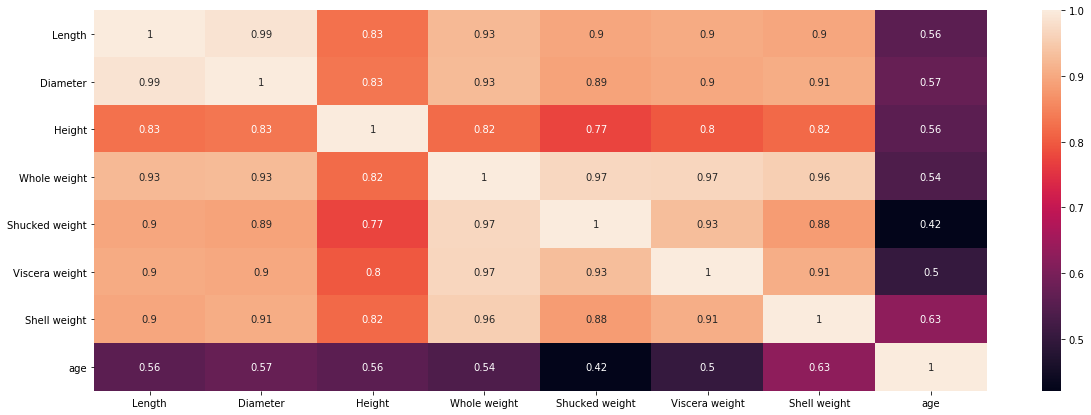


Figure 6: Heat map

From figure 5 & 6, we can identify a few key insights. Length is linearly correlated with diameter while, non-linear relation with height, whole weight, shucked weight, viscera weight and shell weight. Whole Weight is almost linearly varying with all other features except age. Height has least linearity with remaining features. Age is most linearly proportional with Shell Weight followed by Diameter and length. Age is least correlated with Shucked Weight. Such high correlation coefficients among features can result into multi-collinearity.

After doing all the pre-processing and visualization on the dataset, we are now ready to predict from them. We have enough idea of the data to make prediction and we can assume how the model can perform. So, for this case, we have selected 5 algorithms to make the predictions such as Linear Regression, Ridge Regression, Support Vector Regression, Random Forrest Regression, K-Nearest Neighbor Regression. All of them are regression algorithms because we are trying to predict continuous data. Our target feature is Age and we cannot classify age rather we can predict a continuous value.

First, we need to split the data into train and test set with an 80-20 split. That means we will have 80% data to learn from and 20% data to evaluate our prediction. This 20% test data will be completely unseen for our model because it will gather knowledge from the rest 80% of the data.

After the training and examining the data, we have found out that SVR performed best.

**Challenges**: Predicting from such dataset can be challenging because to get a high accuracy, performance of the model largely depend on the quality of data. We need more reliable data, hyper parameter tuning and fine tuning. Also there are other problems that can occur when the dataset is not sufficient such as overfitting, under fitting etc. More robust engineering and study is needed to implement this in real life scenario.

**Recommendations**: By using proper data to predict a business outcome is really helpful and recommended. But business owners also have to consider the risks and challenges. But as data science tools are completely based on statistics, it can be reliable to trust them in small scenario. But in a vast scale, more cross checking and assurance is required in order to deploy fully automated data science tool in business.

**Conclusion**: As our generation is fully digitized, data is generating and moving so fast. So it is high time for us to preserve the data and make good of it. But humans are not capable of controlling such huge amount of data by themselves. And data science has proven that it can share the burden with humans. So, business organizations should start implementing and experimenting with the tools. Data science and machine learning has a lot to offer in the real life industries.

**References**:

[1] “Abalone Dataset,” Available Online: *http://archive.ics.uci.edu/ml/datasets/Abalone*

[2] W. J. Nash, T. L. Sellers, S. R. Talbot, A. J. Cawthorn and W. B. Ford, “The Population Biology of Abalone (Haliotis species) in Tasmania. I. Blacklip Abalone (H. rubra) from the North Coast and Islands of Bass Strait,” *Sea Fisheries Division, Technical Report No. 48 (ISSN 1034-3288)*, 1994.

[3] “Abalone,” Available Online: *https://en.wikipedia.org/wiki/Abalone*

[4] M. Hossain and N. Chowdhury, “Econometric Ways to Estimate the Age and Price of Abalone,” *Department of Economics, University of Nevada*, 2019.

Appendices:

Appendix A: Python Code

