HarvardX PH125.9x Capstone - CYO - Report

b_woods

15/11/2020

Overview

Introduction

This report is created for the "HarvardX PH125.9x Data Science: Capstone" course on edX, functioning as the second of two capstone projects to complete the program. This report relates to the analysis and development of a machine learning model related to the Abalone Data Set provided to the UCI Machine Learning Repository by the Marine Research Laboratories in Tasmania, Australia. The model attempts to predict the of an Abalone, a type of marine snail, by a number of physical characteristics provided. This model could then be used to predict the age of other Abalone given these characters thereby saving essential research time of science teams.

This dataset was chosen for this report for two reasons:

- The dataset was sourced from a location geographically close to where I live and relates to my personal interest in ocean life.
- The process for determining age manually is a time consuming task therefore not an optimal use of a scientists time. This project provides an example of how laborious tasks can be semi-automated with the assistance of machine learning, therefore can serve as a reference point for future research.

This report contains five main sections:

- **Introduction:** An introduction to the report, dataset and project scope.
- **Analysis:** An analysis of the dataset, methods used to clean the dataset and development of a number of machine learning models for prediction.
- **Results:** Comparison of the machine learning models developed and selection of the most optimal model.
- **Conclusion:** A review of this project and assessment of areas of improvement.
- **Appendix:** References and tools used for this project.

This project was completed using the R programming language and a number of libraries relating to data science. This report focuses on the high level approach and outcomes and only shows the code where required - please refer to the accompanying R script to replicate the approaches taken. Alternatively the RMD version of this report also contains the code used to generate the results of this project.

_

Dataset

The Abalone dataset was sourced from the UCI Machine Learning Repository. It was originally compiled for a non-machine learning study by the Marine Research Laboratories - Taroona, Department of Primary Industry and Fisheries, Tasmania (Australia). The dataset describes a number of physical traits of Abalone. The details provided on UCI state:

Predicting the age of abalone from physical measurements. The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope – a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the age. Further information, such as weather patterns and location (hence food availability) may be required to solve the problem.

From the original data examples with missing values were removed (the majority having the predicted value missing), and the ranges of the continuous values have been scaled for use with an ANN (by dividing by 200). - https://archive.ics.uci.edu/ml/datasets/Abalone

The dataset contains `4177` rows and `9` columns. An example of the first few rows:

```
## # A tibble: 6 x 9
##   sex   length diameter height weight_Whole weight_shucked weight_viscera
##   <fct>  <dbl>    <dbl>  <dbl>        <dbl>          <dbl>          <dbl>
## 1 M      0.455    0.365  0.095        0.514         0.224          0.101
## 2 M      0.35     0.265  0.09         0.226         0.0995         0.0485
## 3 F      0.53     0.42   0.135        0.677         0.256          0.142
## 4 M      0.44     0.365  0.125        0.516         0.216          0.114
## 5 I      0.33     0.255  0.08         0.205         0.0895         0.0395
## 6 I      0.425    0.3    0.095        0.352         0.141          0.0775
## # ... with 2 more variables: weight_shell <dbl>, rings <int>
```

The descriptions for each variable is listed in the Abalone Names file.

Name	Data Type	Measurement	Description
Sex	Nominal	-	M, F, and I (infant)
Length	Continuous	mm	Longest shell measurement
Diameter	Continuous	mm	perpendicular to length
Height	Continuous	mm	with meat in shell
Whole weight	Continuous	grams	whole abalone
Shucked weight	Continuous	grams	weight of meat
Viscera weight	Continuous	grams	gut weight (after bleeding)
Shell weight	Continuous	grams	after being dried
Rings	Integer	-	+1.5 gives the age in years

Note that the dataset does not have an 'age' value. The age of an Abalone is determined by adding 1.5 to it's rings value. During data cleaning a new variable will be created for age, which will be the dependent variable in the machine learning models generated.

Project Structure and Goals

The aim of this project is to predict the age of an Abalone given a number of physical characteristics and to evaluate the accuracy of this prediction. There are two general methods for approaching this problem:

- **Classification** - Where the physical characteristics are used to classify the dataset into discrete age groups.

- **Regression** - Where the age is treated as a continuous value and the physical characteristics measured as factors that influence that continuous value.

While the project could appear as a classification problem due to a set number of categories, this project will instead treat this as a regression problem for a number of reasons:

- There are a total of "28" age values (i.e. categories). For a dataset with only "4177" rows, classification becomes more challenging. This is compounded by uneven age distribution, outlined in a later section.
- Age itself is a continuous variable and instead of expecting a prediction to match an exact value, having a continuous value means the predication can be measured within by how closely the range relates to the actual value.

As a regression problem, the accuracy of the prediction will be measured by the Root Mean Squared Error (RMSE) - which measures how much an average sample deviates from the predicted value. No target RMSE will be set, instead the aim is to compare RMSE among different machine learning models to determine the lowest value model, then evaluate the distribution and any patterns of the predicted age relative to the actual age value.

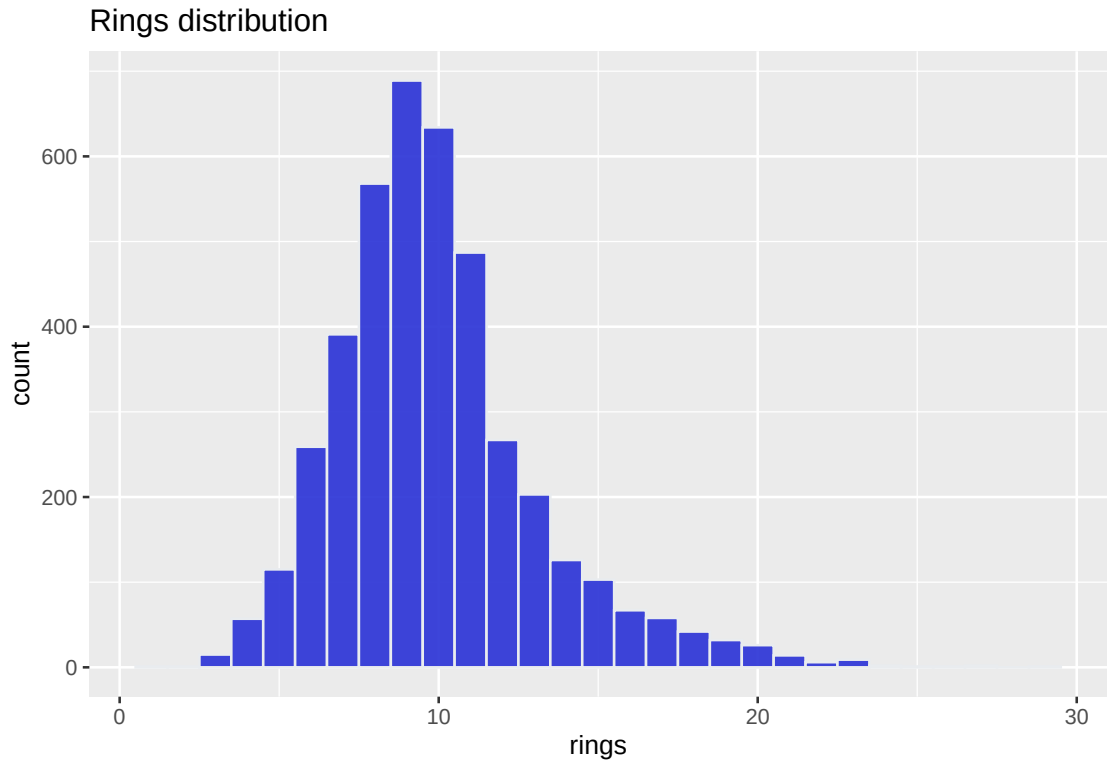
Analysis

Data Exploration

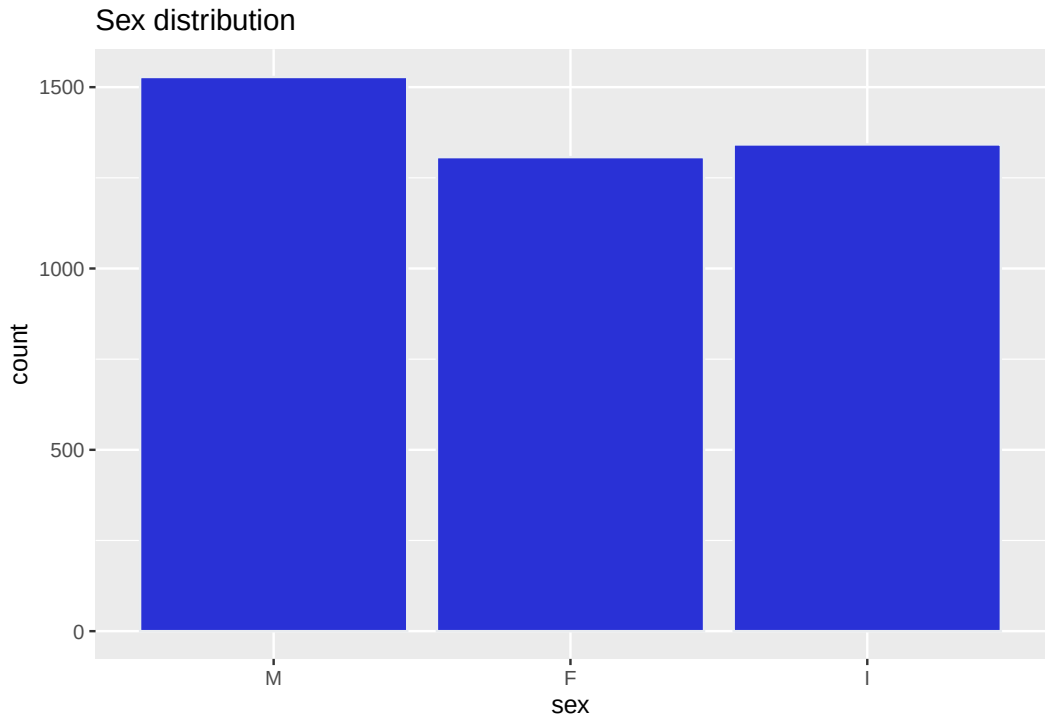
There are no NA values in the dataset. Summary statistics for the dataset:

```
##  sex          length         diameter          height        weight_Whole
##  M:1528   Min.   :0.075   Min.   :0.0550   Min.   :0.0000   Min.   :0.0020
##  F:1307   1st Qu.:0.450   1st Qu.:0.3500   1st Qu.:0.1150   1st Qu.:0.4415
##  I:1342   Median :0.545   Median :0.4250   Median :0.1400   Median :0.7995
##           Mean   :0.524   Mean   :0.4079   Mean   :0.1395   Mean   :0.8287
##           3rd Qu.:0.615   3rd Qu.:0.4800   3rd Qu.:0.1650   3rd Qu.:1.1530
##           Max.   :0.815   Max.   :0.6500   Max.   :1.1300   Max.   :2.8255
##  weight_shucked   weight_viscera    weight_shell        rings
##  Min.   :0.0010   Min.   :0.0005   Min.   :0.0015   Min.   : 1.000
##  1st Qu.:0.1860   1st Qu.:0.0935   1st Qu.:0.1300   1st Qu.: 8.000
##  Median :0.3360   Median :0.1710   Median :0.2340   Median : 9.000
##  Mean   :0.3594   Mean   :0.1806   Mean   :0.2388   Mean   : 9.934
##  3rd Qu.:0.5020   3rd Qu.:0.2530   3rd Qu.:0.3290   3rd Qu.:11.000
##  Max.   :1.4880   Max.   :0.7600   Max.   :1.0050   Max.   :29.000
```

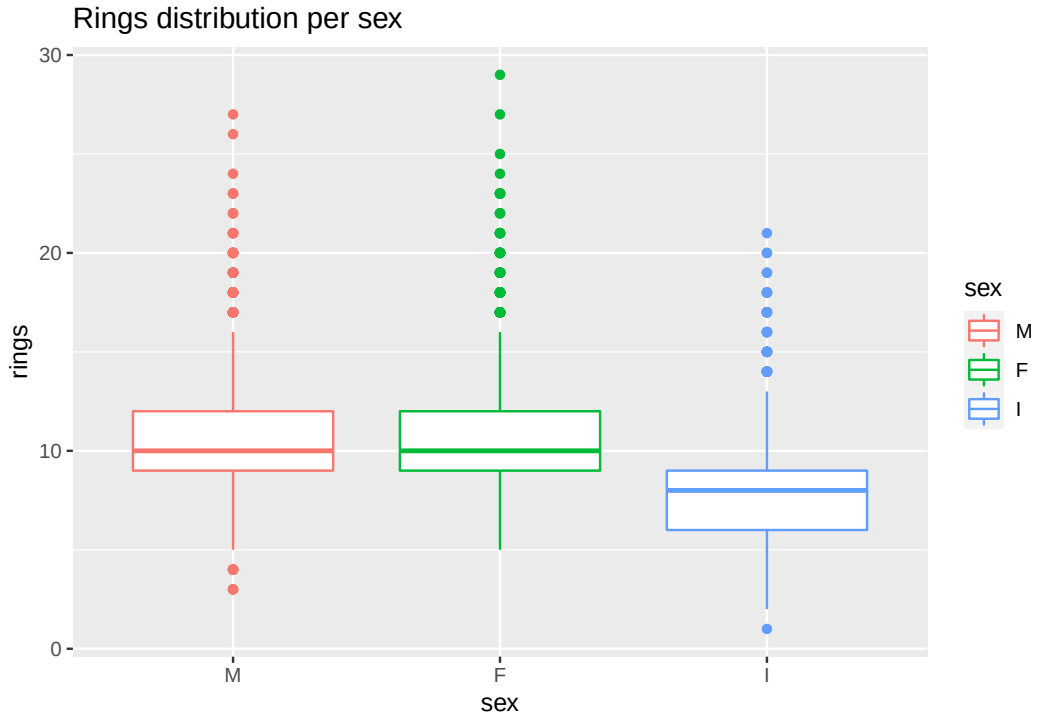
Distribution of rings is roughly normally distrusted with a mean around 10 rings, though the tail to the right extends into much larger values.



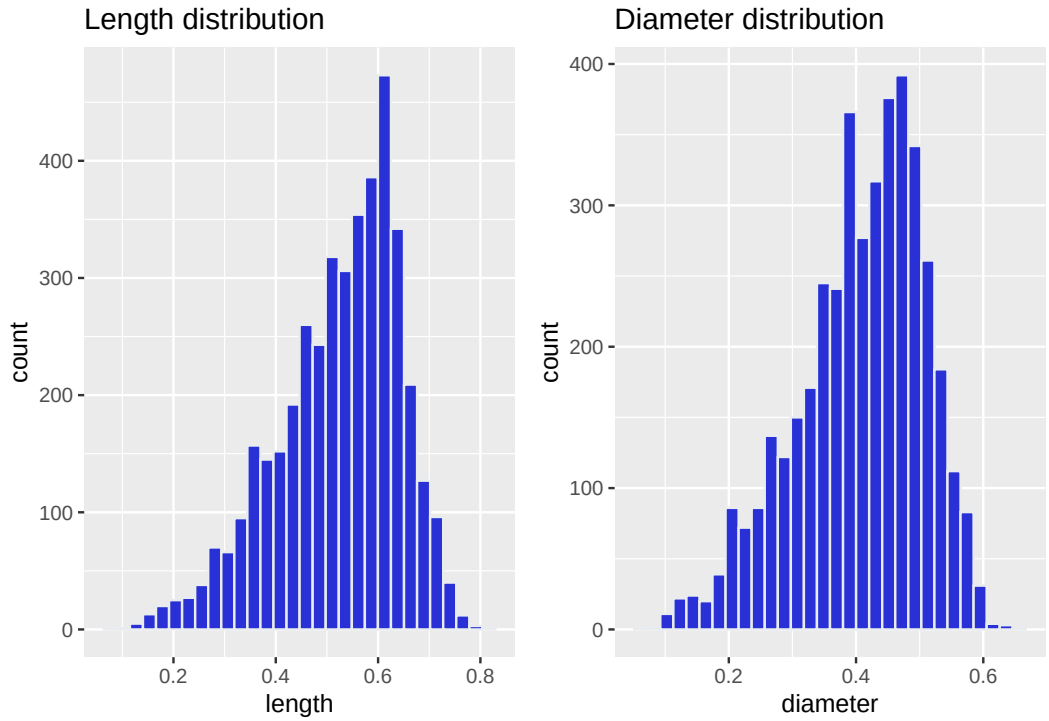
There are three categories of Sex - Male, Female and Infant. Female and infant have similar record counts with the male count being a little larger.



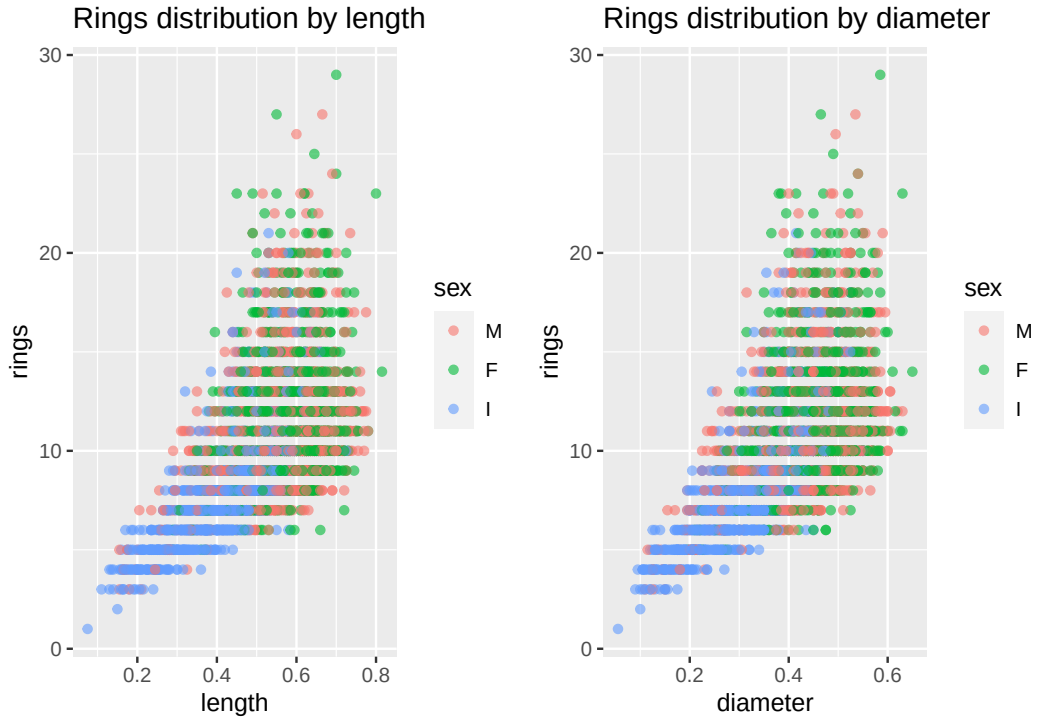
Distribution of rings per sex show male and female having similar distributions, with infant being smaller, which is expected as rings is related to age. There are a large number of outliers relating to high values while few smaller outliers.



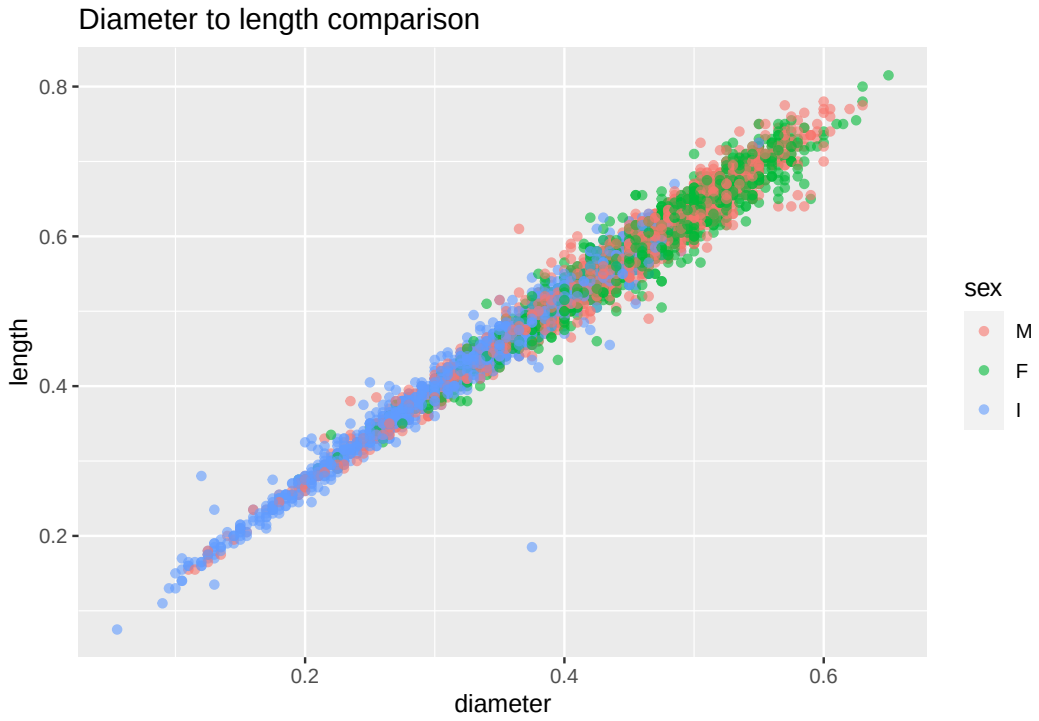
Length and diameter have a similar distribution; both seem to be roughly normally distributed and share a right skew.



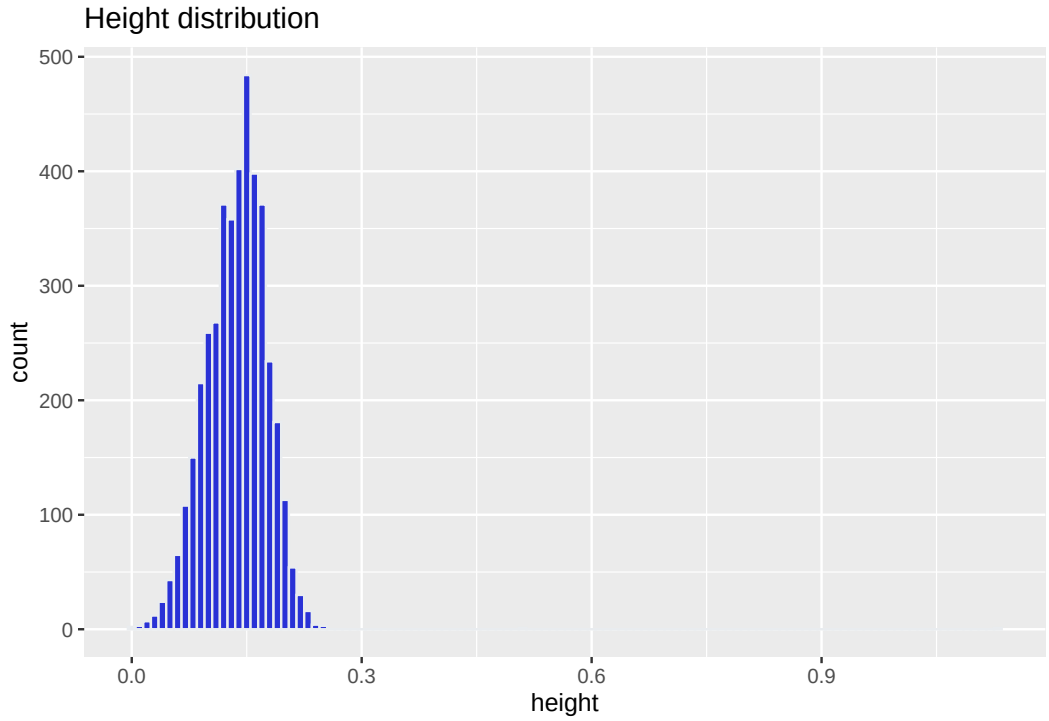
Rings distribution by length or diameter have a similar general shape. Infants take up the majority of the smaller range while the larger range is split between males and females. A small number of outliers in the upper ranges.



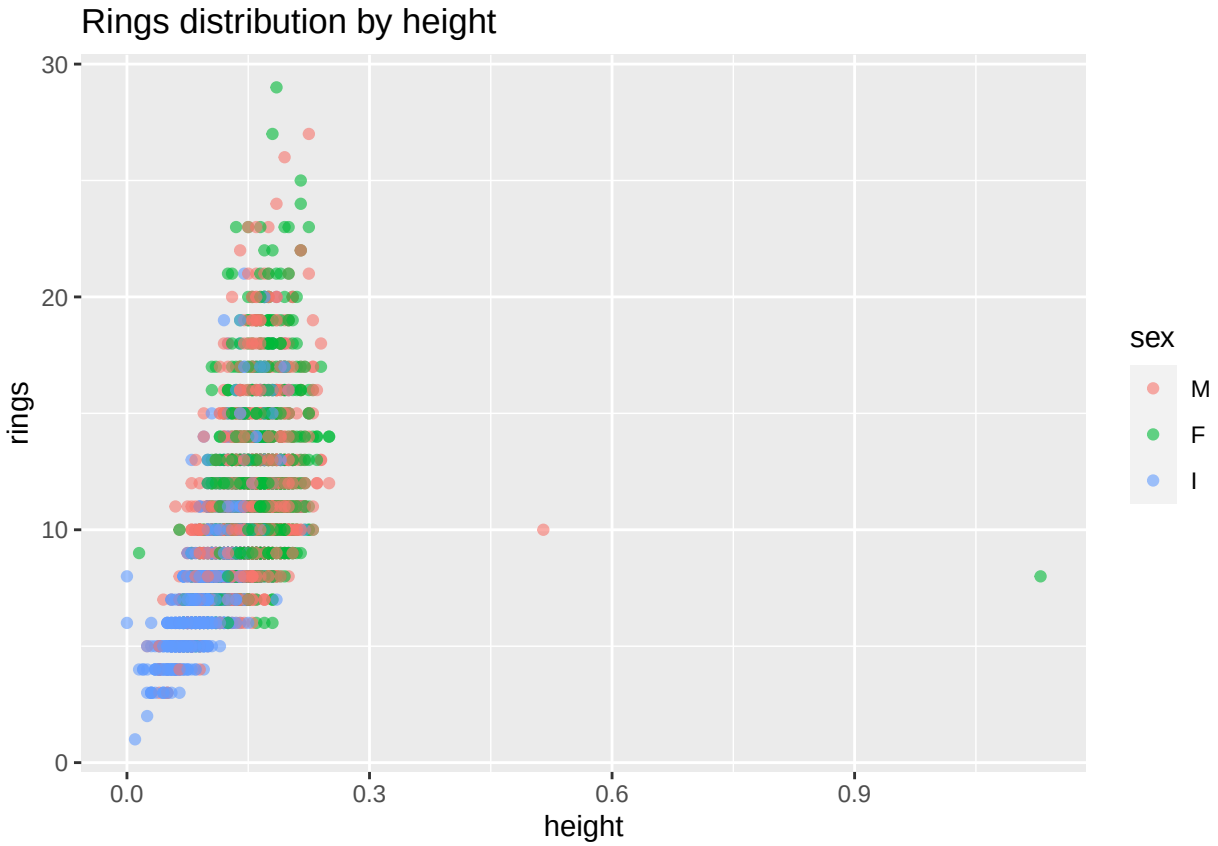
Diameter and length have a clear linear relationship. The range of deviation increases as the values increase. Infant abalone have a smaller range of deviation but more notable outliers than adult Abalone.



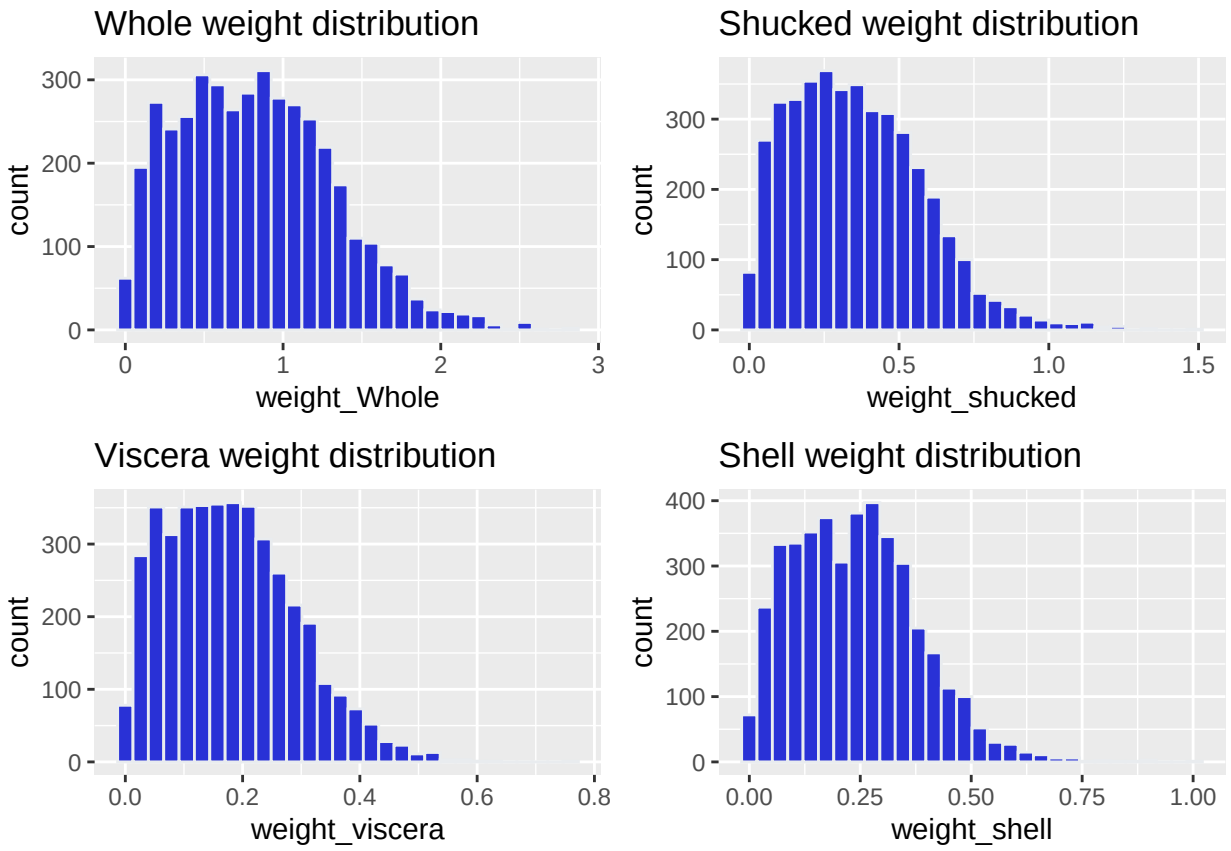
Abalone height distribution has a limited range of values having a roughly normal distribution centered around 0.2. Note there are 2 records with a height of 0, which will need to be removed as this is invalid data.



Comparing ring distribution by height shows a similar relationship - a relatively small range with a symmetric distribution that starts to break up in higher ring counts. Note the extreme height outliers in the 10 rings range.

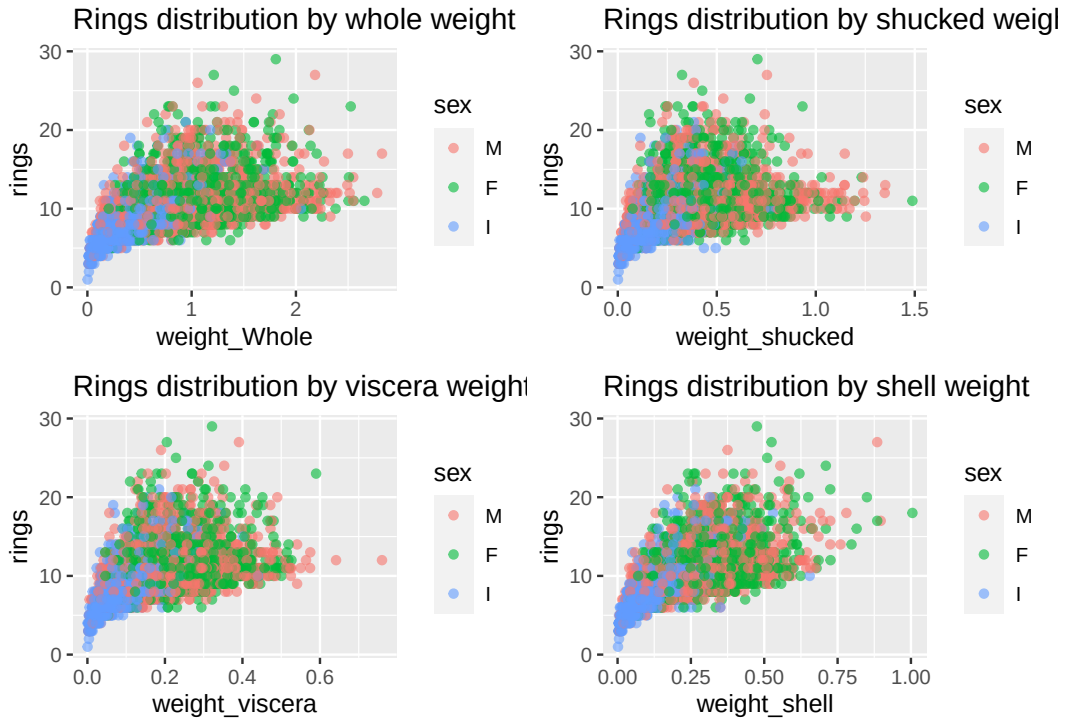


There are four different weight types - Whole, shucked, viscera and shell. The general shape of each distribution is similar with a clear right skew. Shucked weight seems to relate most directly to whole weight based on the shape and value similarity, however there is a clear distinction between each component.



The relationship between whole weight and the weight sub-components is confirmed by plotting these against each other. This is a clear linear relationship though the linearity starts to diverge when the values are larger, similar to the height vs diameter relationship. Note how the pattern of outliers in larger values varies considerably between different comparisons.

Comparing rings and the different weight types shows a roughly linear relationship, though the range of linearity has far more divergence than when comparing ring count with other values. The shape between the ring vs weight type distribution is roughly similar between all four types, though again the higher ring count values have much more variance.



Insights

From this data analysis, a number of insights can be drawn from the data:

- The data is generally normally distributed and has a linear relationship between the rings (i.e. age) and the physical characteristics.
- There are a number of records with 0 height which will need to be removed as this is invalid data.
- The structure of the data becomes less consistent at higher ring counts, which could lead to inconsistencies modeling across this range.
- Likewise while some values such as length and diameter are closely correlated, for records with a higher ring count the correlation is less consistent. For this reason all values should be used in the model.
- The dataset being a relatively small count (4000 rows) for a large prediction space could lower the accuracy of the model.
- As the variables are physical measurements of the Abalone, the model will fail to take into account other potential impacts such as differences in geography. This is outlined in the original dataset details and in the first section of this report.

Data Cleaning

Prior to building the model, a cleaned dataset is created with three changes:

- Records with a Height of 0 are removed.
- A new column is created for Age, which will be the dependent variable when modeling. Age is equivalent to rings + 1.5
- As Age is now within the dataset, Rings is no longer required, so is removed from the dataset.

```
abalone_clean <- abalone %>%
  filter(height != 0) %>%
  mutate(age = rings + 1.5) %>%
  select(-rings)
```

In order to test without overfitting, the data is split into a train and test set. The train set is used to develop the model while the test set is only used for testing each of the models. The split ration is 75% train and 25% test, which gives ~3000 records for training and ~1000 testing. As the dataset is quite small a larger ratio is preferred to ensure more accurate testing while still making the majority of records available for training.

```
set.seed(1, sample.kind="Rounding") # Use a common seed value for reproducibility
abalone_clean_test_index <- createDataPartition(y = abalone_clean$sex, times = 1,
                                                p = .25, list = FALSE)

abalone_clean_train <- abalone_clean[-abalone_clean_test_index,]
temp <- abalone_clean[abalone_clean_test_index,]

# Ensure age values in the test set also exist in the train set
abalone_clean_test <- temp %>%
  semi_join(abalone_clean_train, by = "age")

removed <- anti_join(temp, abalone_clean_test, by = "age")
abalone_clean_train <- rbind(abalone_clean_train, removed)
```

Model Development

A number of different machine learning algorithms have been used to create an optional model. This section outlines the creation of these models, training the model and testing using the test data set to determine the RMSE of that model. Model results are analyzed and compared in the Results section.

As a baseline, determine the mean age in the training dataset and calculate the RMSE of just using this single value. While this is not a detailed or reliable model, it provides a point of reference for comparing the RMSE values in more complex models.

```
train_mean_age <- mean(abalone_clean_train$age)
rmse_mean <- RMSE(train_mean_age, abalone_clean_test$age)
cat("RMSE for mean only:", rmse_mean)
```

```
## RMSE for mean only: 3.146942
```

```
# Create a storage table for tracking different RMSE values
rmse_results <- tibble(method = "Mean Only",
                       RMSE = rmse_mean)
```

This indicates that Abalone age values deviate from the age by an average of `3.1469417`. The following models attempt to improve this score.

1) Linear Model

A linear attempts to create a linear relationship between the dependent variable and the predictors. This is relevant where the data itself shows a linear relationship, which was observed in the Data Exploration section.

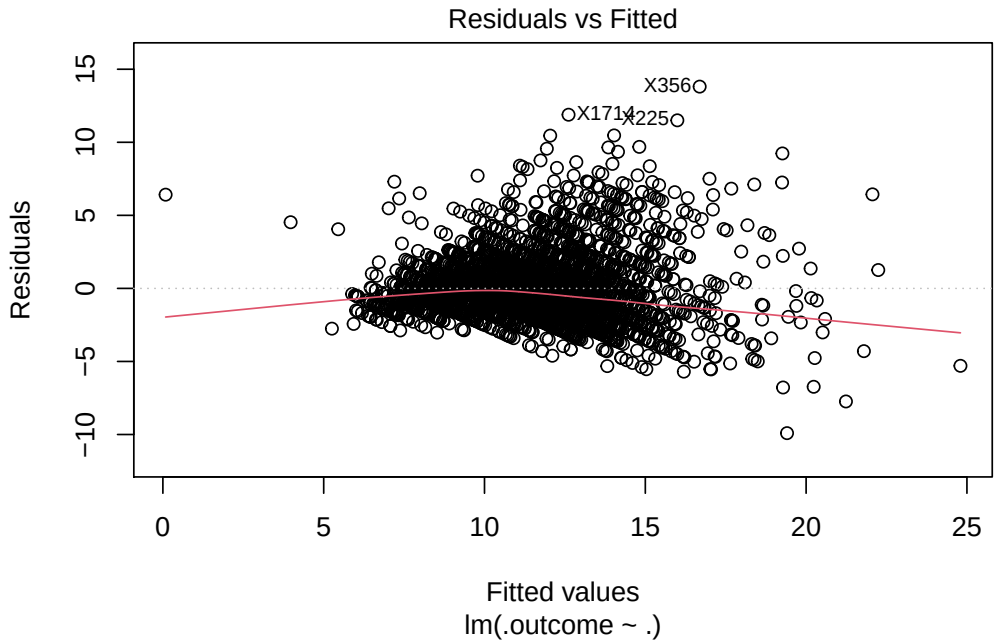
As with proceeding models, age is the predicted value using all other variables in the train dataset. All models will also use 10 fold cross validation to improve evaluation.

```
lm_control <- trainControl(method = "cv", number = 10) # Cross validation
train_lm <- train(age ~ .,
                  data = abalone_clean_train,
                  method = "lm",
                  trControl = lm_control)
```

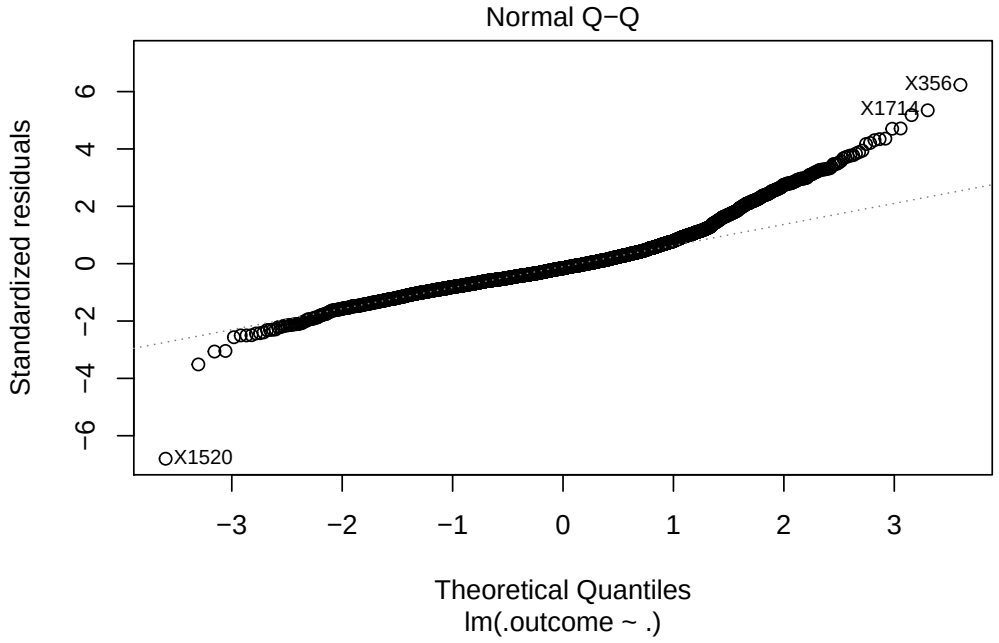
The model builds a linear model based on how the predictors relate to Age:

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Coefficients:
##    (Intercept)            sexF            sexI          length        diameter
##        5.47236        -0.06747        -0.90739         0.33528        10.12675
##         height    weight_Whole  weight_shucked  weight_viscera    weight_shell
##       10.19491         9.26951       -20.10310       -11.14317         8.95035
```

The Residuals vs Fitted graph shows how the model relates to the values in the training set.



The QQ plot of this confirms the earlier insight that the data is roughly normally distributed but diverges more in upper ranges.



The trained model is then applied to the test set to determine the RMSE.

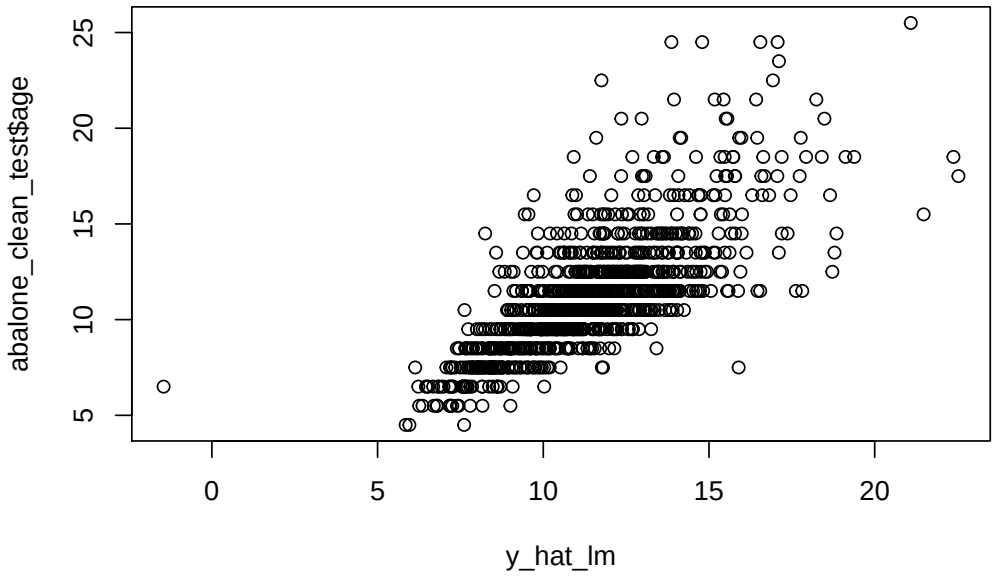
```
# Test the linear model against the test data set
y_hat_lm <- predict(train_lm, abalone_clean_test)
```

```r
# Calculate RMSE and add to reference table
rmse_lm <- RMSE(y_hat_lm, abalone_clean_test$age)
rmse_results <- add_row(rmse_results,
                        method = "Linear Model",
                        RMSE = rmse_lm)

cat("RMSE for Linear Model:", rmse_lm)
```

```
## RMSE for Linear Model: 2.103293
```

The RMSE of `2.1032926` is a clear improvement over the pure mean value of `3.1469417`. Plotting the actual age values in the test set against the predicted age values shows a linear pattern which again diverges in the upper ranges.

```r
plot(y_hat_lm, abalone_clean_test$age)
```



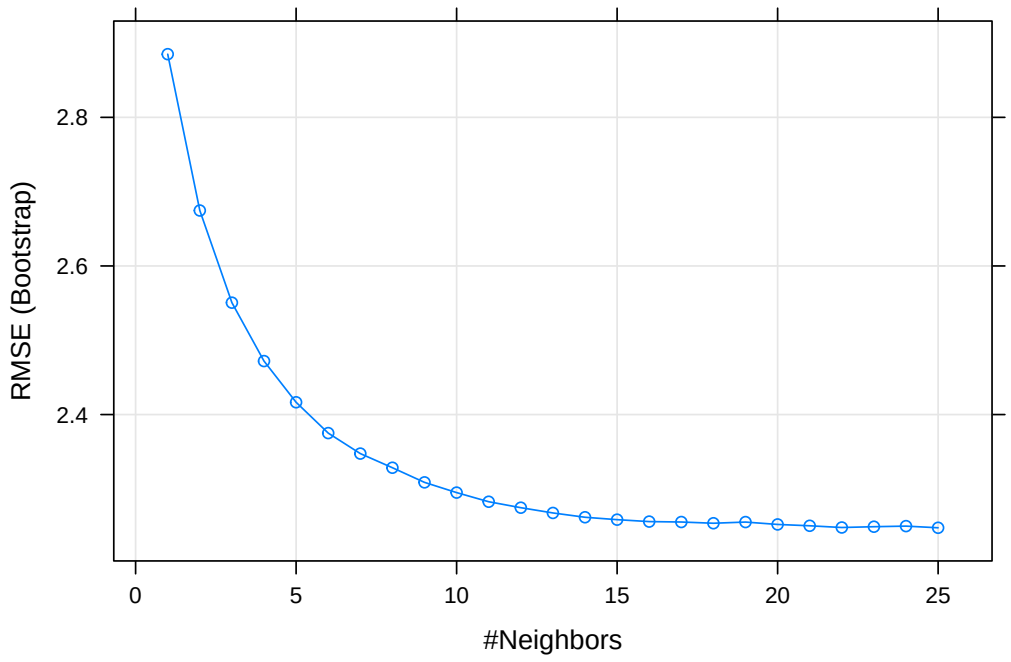
2) K Nearest Neighbors (KNN)

KNN uses the value of surrounding data points to determine a data points predicted value. This operates under the idea that similar values share characteristics and this relationship can be used for prediction. KNN has been chosen as the distribution examined in the Data Exploration section show a number of patterns between Abalone which share physical characteristics, therefore KNN could be used to determine an Abalone's age by examining similar Abalone.

As KNN references surrounding values, the number of surrounding values, referred to as K, must be provided. As the optimal K is unknown a series of values between 1 and 25 are provided to test with.

```
knn_control <- trainControl(method = "cv", number = 10, p = .8) # Cross validation
knn_tunegrid <- data.frame(k = seq(1, 25, 1)) # Determine optimal K between 1 and 25

train_knn <- train(age ~ .,
                   method = "knn",
                   data = abalone_clean_train,
                   tuneGrid = knn_tunegrid,
                   control = knn_control
                   )
```

Plotting the series of K values and the corrosponding RMSE within the train set shows an optimal K value of 25.



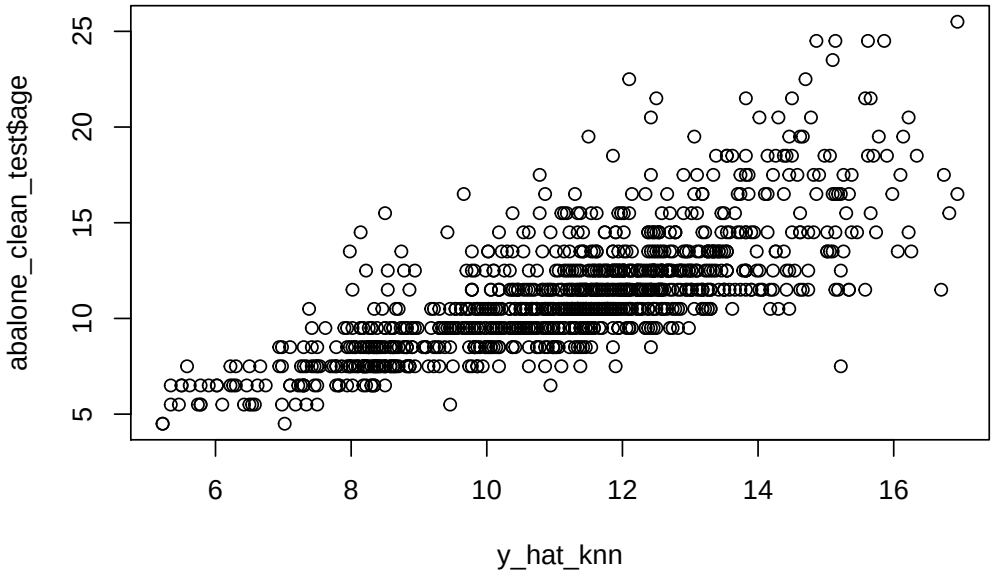
The trained model can then be applied to the test set to predict the Age value.

```
y_hat_knn <- predict(train_knn, abalone_clean_test)

# Calculate RMSE and add to reference table
rmse_knn <- RMSE(y_hat_knn, abalone_clean_test$age)
rmse_results <- add_row(rmse_results,
```

```
                          method = "KNN",
                          RMSE = rmse_knn)

cat("RMSE for KNN:", rmse_knn)
```

```
## RMSE for KNN: 2.108812
```

The RMSE of `2.1088121` is very close to the LM RMSE of `2.1032926`. Plotting the relationship between actual age and predicted KNN age shows a very different distribution - the values are far more spread out though a clear linear pattern is still shown.



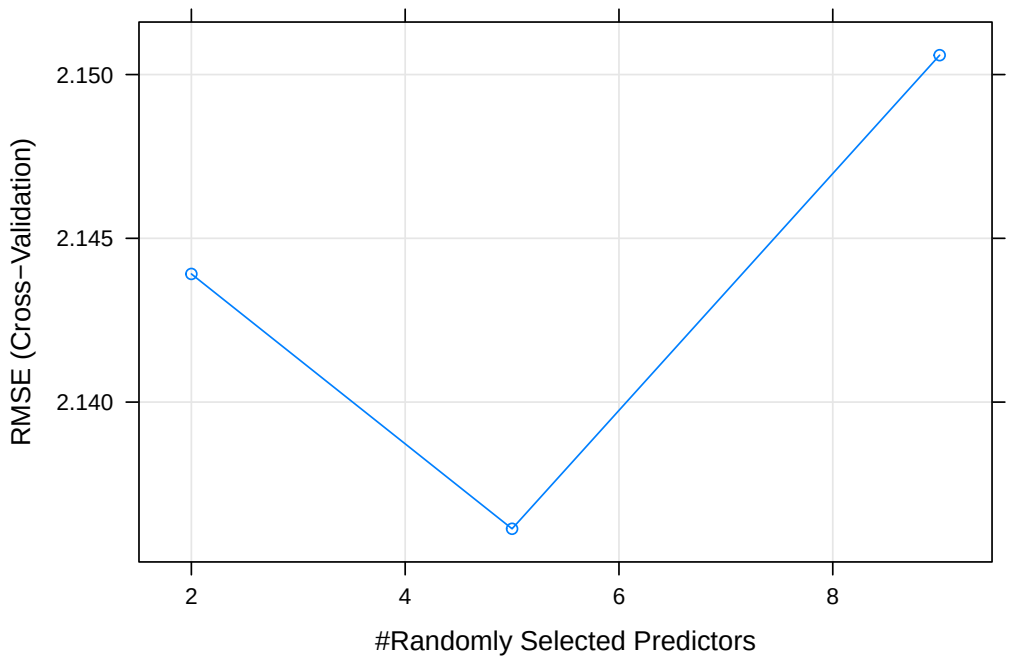
3) Random Forest

Random Forests work by creating a decision tree from a sample of the data and repeating this process a large number of times, essentially creating a collection of decision random decisions trees (hence the name, Random Forest). The values from these decision trees are then aggregated to create a final model for prediction. Random Forest is selected here as it offers high accuracy by determining in how different ranges of variables can be used for prediction.

Train a Random Forest model:

```
rf_control <- trainControl(method="cv", number = 10) # Cross validation

train_rf <- train(age ~ .,
                  method = "rf",
                  data = abalone_clean_train,
                  trControl = rf_control
)
```

The trained model shows the lowest RMSE within the train set when `5` randomly selected predictors are used for training the mode.



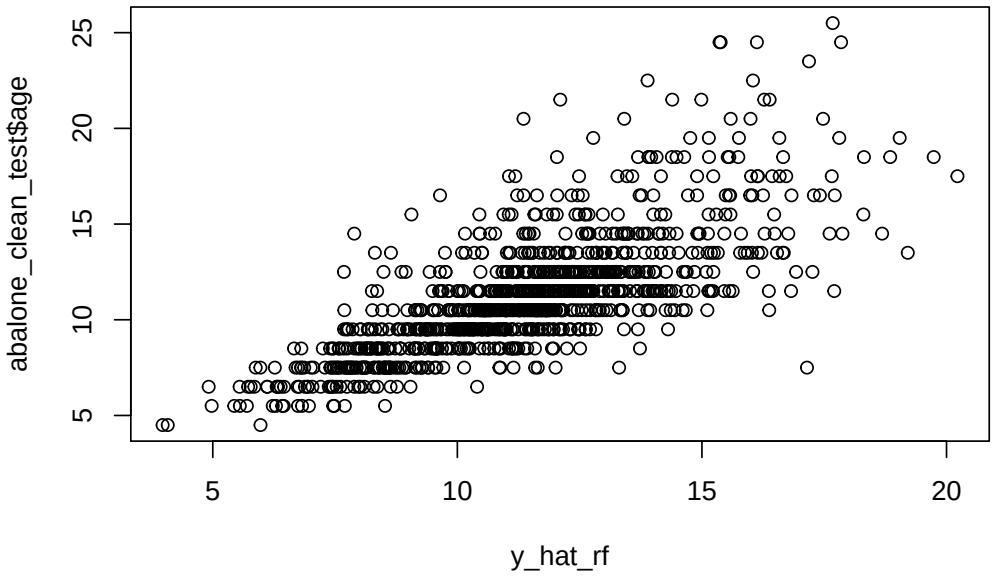
The Random Forest model can then be used for prediction:

```
y_hat_rf<- predict(train_rf, abalone_clean_test)

# calculate RMSE and add to reference table
rmse_rf <- RMSE(y_hat_rf, abalone_clean_test$age)
rmse_results <- add_row(rmse_results,
                        method = "RF",
                        RMSE = rmse_rf)
```

```
cat("RMSE for Random Forest:", rmse_rf)
```

```
## RMSE for Random Forest: 2.121503
```

The RMSE of `2.1215028` for the Random Forest model is similar to the RMSE of the KNN (`2.1088121`) and LM models (`2.1032926`). However the distribution of actual age to predicted age is again different - this is more spread out then LM but more condensed than KNN.



4) Ensemble Models

Each of the models produced have similar RMSE values when used for predictions on the test dataset, however the distributions of these predictions vary between models. This indicates each model has it's own strengths and weaknesses and combining these models could produce a model with further accuracy. This is referred to as an ensemble model and two different ensemble models will be built:

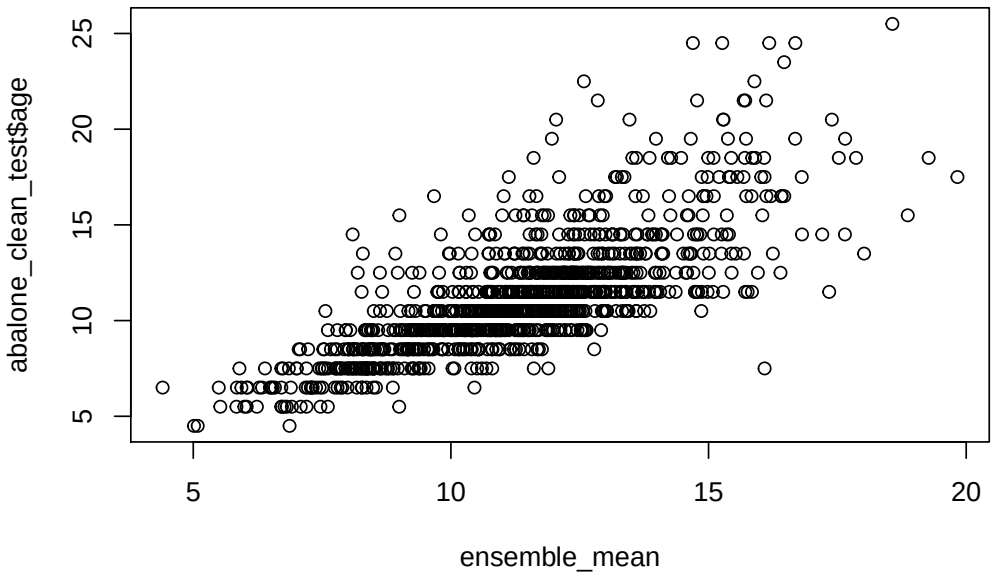
- **Mean Based Ensemble:** For each record, take the three predicted values from the other models and determine the mean value, which will be the predicted value. This ensemble is useful when the models predict a general range of values so the ensemble will take the middle value of that range.
- **Median Based Ensemble:** For each record, take the three predicted values from the other models and determine the median value, which will be the predicted value. This ensemble is useful when there are large variances in the prediction data therefore the middle value will be chosen as this is more likely to be a standard prediction.

Firstly create a dataset containing all three predictions then source the mean of each row, using this for prediction.

```
predctions_combined <-
  tibble(lm = y_hat_lm,
         knn = y_hat_knn,
         rf = y_hat_rf)


# First ensemble is the mean of all three predictions
ensemble_mean <- rowMeans(predctions_combined)


# calculate RMSE and add to reference table
rmse_ensemble_mean <- RMSE(ensemble_mean, abalone_clean_test$age)
rmse_results <- add_row(rmse_results,
                        method = "Ensemble - Mean",
                        RMSE = rmse_ensemble_mean)

cat("RMSE for Ensemble - Mean:", rmse_ensemble_mean)
```

```
## RMSE for Ensemble - Mean: 2.040929
```

The RMSE of the Mean Ensemble is `2.040929` which is far lower than the previous models, indicating the ensemble approach boosts the overall accuracy compared to individual models. The distribution plot of actual age to predicted age also looks like a combination of the previous models.

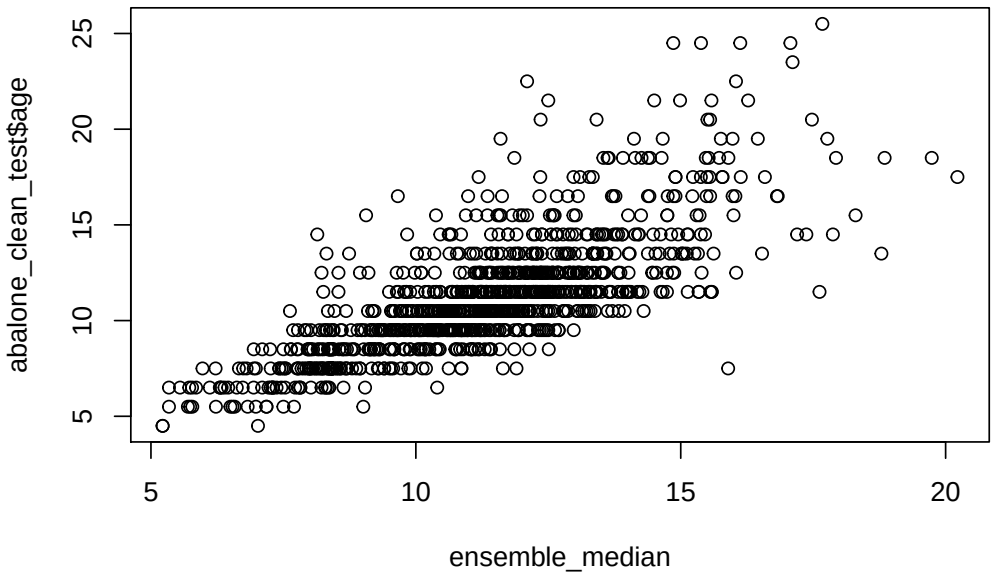


This method can is also applied for the median ensemble model approach:

```
ensemble_median <- rowMedians(as.matrix(predctions_combined))

# calculate RMSE and add to reference table
rmse_ensemble_median <- RMSE(ensemble_median, abalone_clean_test$age)
rmse_results <- add_row(rmse_results,
                        method = "Ensemble - Median",
                        RMSE = rmse_ensemble_median)

cat("RMSE for Ensemble - Median:", rmse_ensemble_median)
```

```
## RMSE for Ensemble - Median: 2.05307
```

The results and actual age to predicted age distribution are very similar to the mean ensemble.



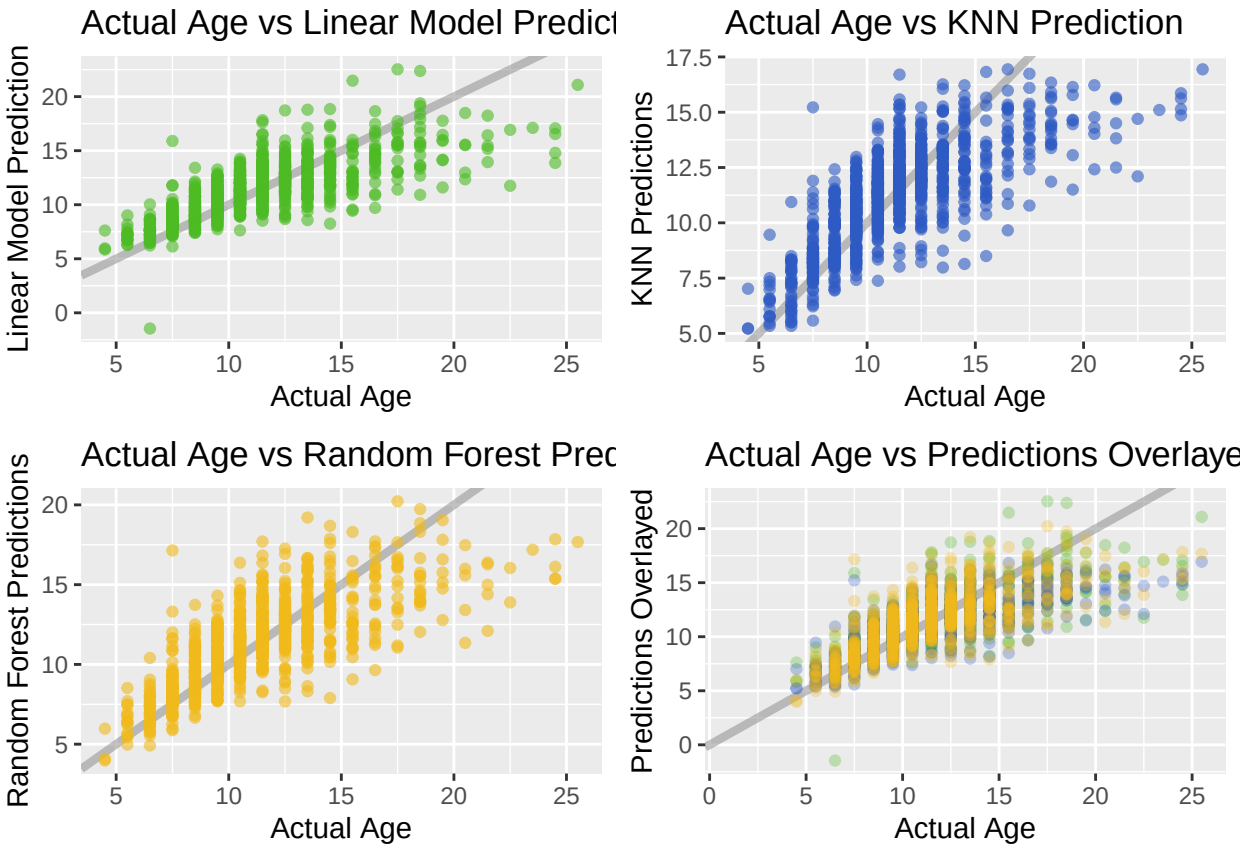
Results

Model Comparison

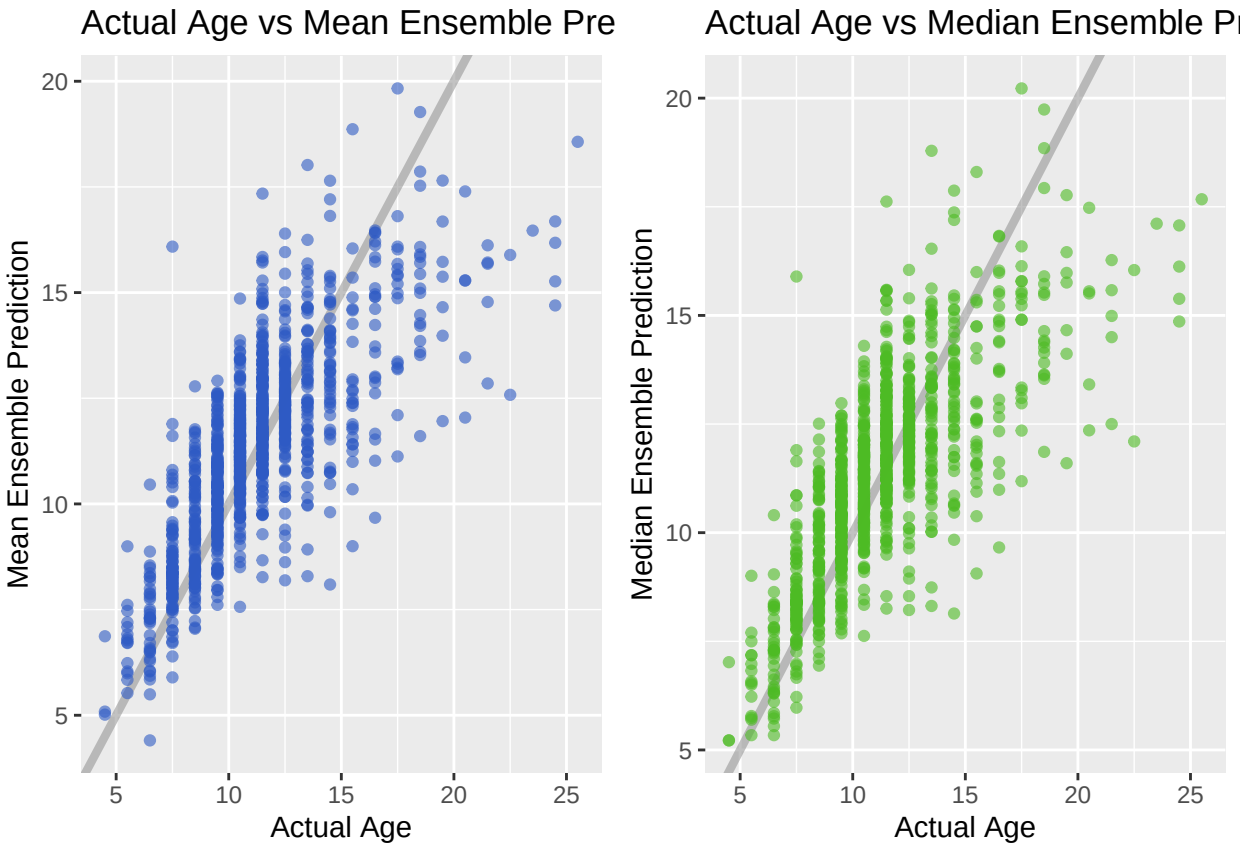
All the models developed have a lower RMSE than the mean, indicating they have potential for prediction. All three single models have very similar RMSE values, while the ensemble models both have lower RMSE values. The ensemble models are also very close together with the mean ensemble having a slightly lower RMSE.

```
## # A tibble: 6 x 2
##   method             RMSE
##   <chr>             <dbl>
## 1 Mean Only          3.15
## 2 Linear Model       2.10
## 3 KNN                2.11
## 4 RF                 2.12
## 5 Ensemble - Mean    2.04
## 6 Ensemble - Median  2.05
```

Comparing the individual models predicted age to the actual age side by side shows the difference in distribution. They all have similar general shapes but differ in how spread out the predictions are in different places - comparing the plotted values to the gray line of expected value confirms this. The fourth graph shows each of these models overlaid - the center of the model is shared among all the graphs but the outliers are handled differently. This comes back to the earlier point of how much the physical characters of Abalone differ in older Abalone and how this could lead to inaccuracy in the prediction model.

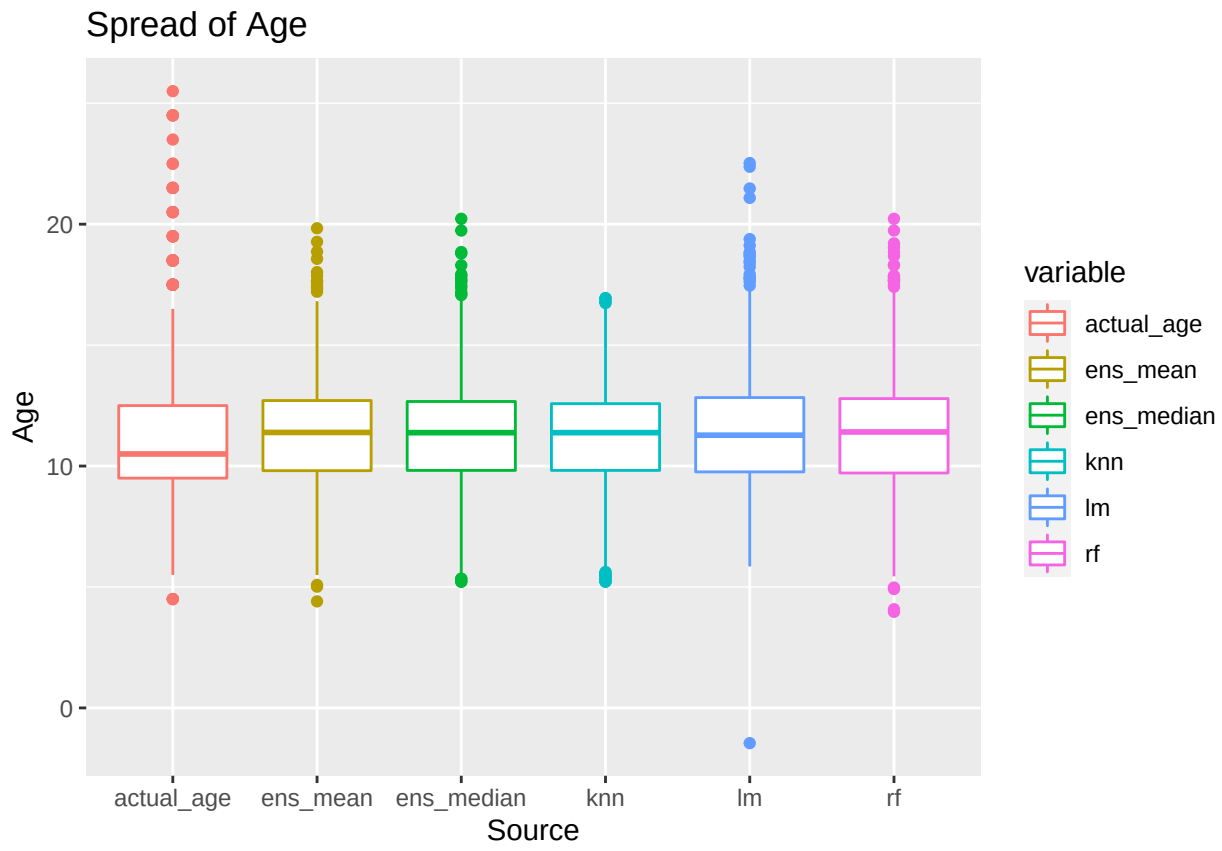


Plotting the ensemble models in the same method shows how similar their distributions are. The linear relationship here appears more clearly and there appears to be a little less scattering in the upper ranges.



The upper range outliers can be further analyzed by comparing the spread of range across the actual age and the predicted ages. The actual age shows a large number of high outliers which are not captured effectively by the other models - the closest is LM but this still does not cover the full range. LM also has large outliers in the lower ranges which are not present in the actual age or other models. KNN has the smallest range, possibly due to the large K value meaning less spread in the data. Random Forest condenses the upper outliers into a smaller range than the other models - this may be due to the decision trees focusing more on the center of the data so the outliers are pulled towards the center.

The ensemble models shape is also represented appropriately as these appear to capture characteristics of the spread of each of the other models. The median model has a little more upper range whereas the mean model captures the lower outliers more accurately. The median and quartiles of the models are all very close, though these are all a little higher than the actual age - possibly due to uneven spread of data causing the center to be off in the predicted models.



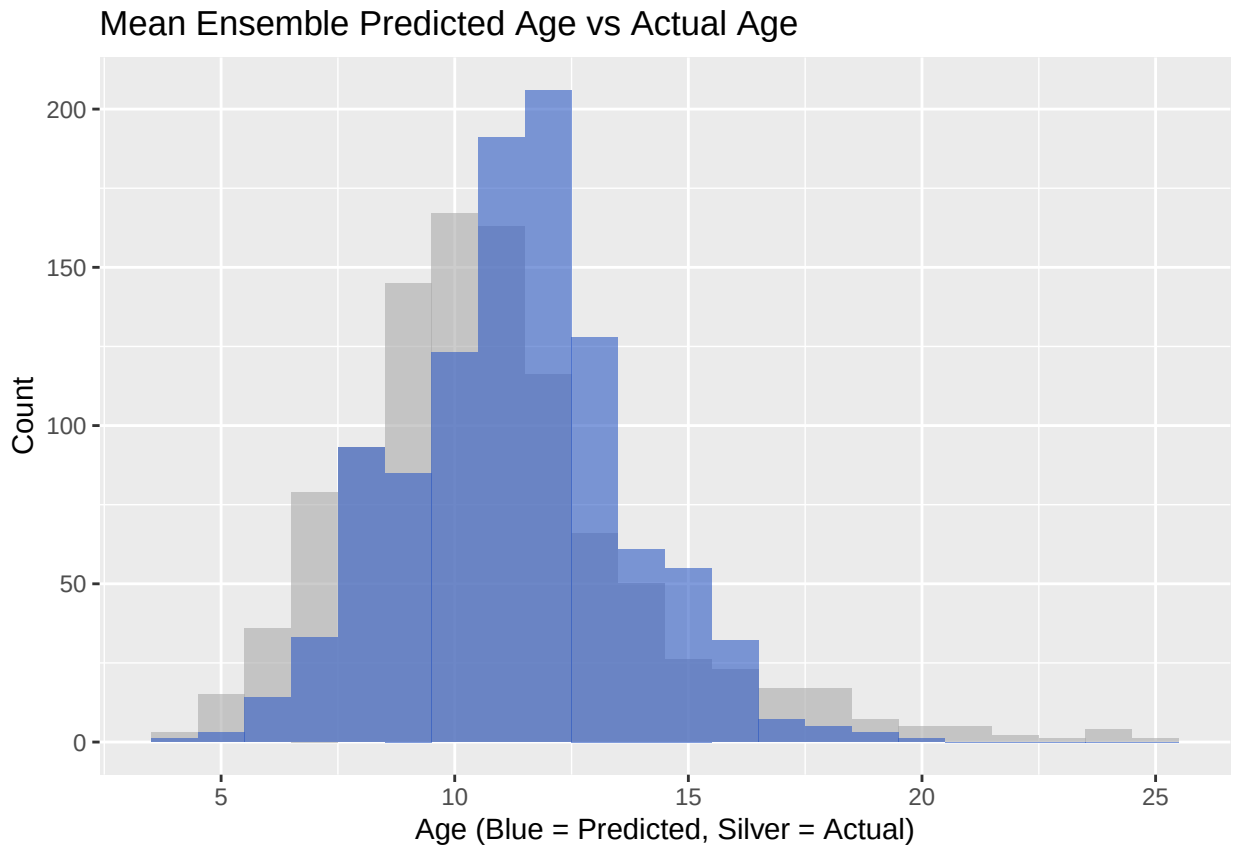
Both ensemble models are worthy choices and both have their advantages and disadvantages while being quite similar. Either could be chosen as the final and modifying the makeup of the ensemble or adding further data points could affect the decision. However given this dataset and results the final model selected is the mean model, as the RMSE is slightly lower and the age distribution seems a little more centered.

Final Model Distribtution

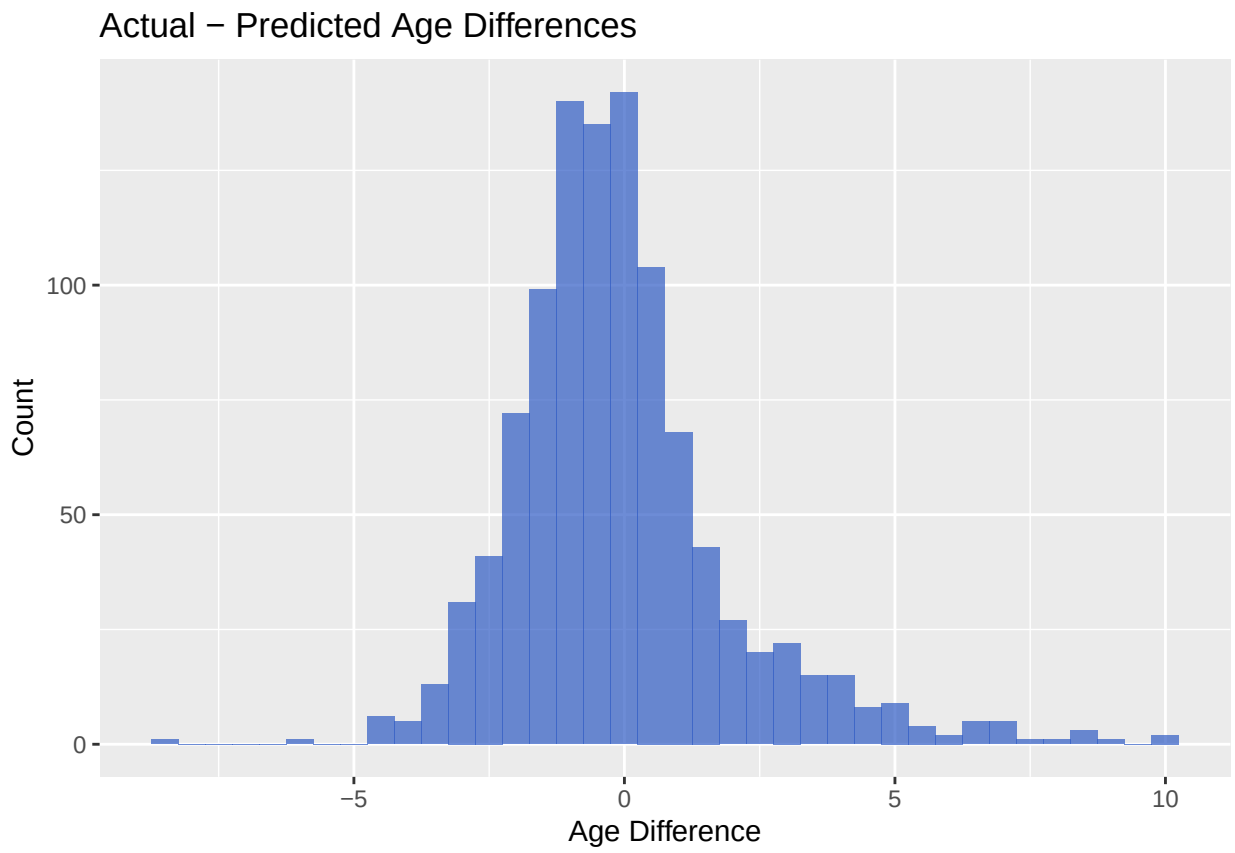
The summary statistics of the actual age, predicted age and difference between these two values are presented below.

```
##    actual_age    predicted_age       age_diff
##  Min.   : 4.50   Min.   : 4.405   Min.   :-8.58760
##  1st Qu.: 9.50   1st Qu.: 9.809   1st Qu.:-1.28446
##  Median :10.50   Median :11.389   Median :-0.32693
##  Mean   :11.26   Mean   :11.328   Mean   :-0.07171
##  3rd Qu.:12.50   3rd Qu.:12.709   3rd Qu.: 0.69168
##  Max.   :25.50   Max.   :19.830   Max.   : 9.91888
```

Comparing the distribution of predicted age (blue) and actual age (silver) shows the main difference between the distributions - the actual values are more spread out while predicted values are more centered. The center for predicted is also 2 years higher, matching the RMSE value.



Graphing the counts of age difference shows a roughly normal distribution centered around 0, though the distribution is also left skewed, showing far more differences in the higher positive ranges - representing the large age outliers which were not accurately predicted by the mean ensemble model.



Conclusion

In conclusions, the analysis of Abalone data provided insight which lead to the development of a number of machine learning models. While each had their strengths and weaknesses, the Mean Ensemble model was the final choice as it was able to combine the strengths of the individual models to create the most reliable prediction model. The model did suffer from it's inability to handle outliers in the large range and having all the predictors focus on the physical characteristics of the Abalone rather than it's environment. The final RMSE of this model is `2.040929`, meaning that when provided with the physical characteristics of an Abalone, the age can be predicted within an average range of `2.040929`years. Such a model could reduce the amount of time scientists have to spend preparing samples for experiment data, or to act as a baseline for the development of a more complex mode.

There are a number of potential areas of improvement for this model:

- **Additional data sources:** The variables in this dataset relate to physical characteristics of the Abalone. Alone this may not be enough to accurately determine age - for example as outlined in the dataset details the geography or food sources could provide valuable insight for prediction.
- **Larger data set:** For a scientific dataset, 4000 records of Abalone physical characteristics is an impressive dataset that would have taken considerable resources to collate. However for a machine learning project, this may not be enough. This is particularly true with how the models developed struggled with estimates of Abalone with outlier characteristics, which could be better estimated with more data in that range.
- **Improved ML algorithms:** This project only tested three machine learning models which may not be the most optimal models for this dataset. More research into different models could lead to better model selection. Likewise in these actual models better tuning could improve the prediction capability.
- **More complex ensemble:** Further to the previous point, with better models or better tuning, the ensemble could better combine the strengths of performing models and minimize the weaknesses of models that under-perform in some situations.

As a personal reflection, this project was the first time I've built machine learning models under my own direction and I feel the outcome has been successful. I was fortunate to find a dataset related to my interests which also matched the areas studied in this course. The project highlighted the effectiveness of ensemble models and this is an area I would like to continue researching in the future. I feel my weakness when developing these models was a lack of tuning - testing different tuning methods often lead to under-performance compared to the default settings. I plan to spend more time learning how tuning options relate to the models better so future work can be more optimized towards the dataset being studied.

Appendix

References

Courses.edx.org. 2020. Course | PH125.8X | Edx. [online] Available at: https://courses.edx.org/courses/course-v1:HarvardX+PH125.8x+2T2020/course/ [Accessed 01 November 2020].

Courses.edx.org. 2020. Course | PH125.9X | Edx. [online] Available at: https://courses.edx.org/courses/course-v1:HarvardX+PH125.9x+2T2020/course/ [Accessed 01 November 2020].

Irizarry, R., 2020. Chapter 32 Machine learning in practice | Introduction To Data Science. [online] Rafalab.github.io. Available at: https://rafalab.github.io/dsbook/machine-learning-in-practice.html [Accessed 01 November 2020].

Archive.ics.uci.edu. 2020. UCI Machine Learning Repository: Abalone Data Set. [online] Available at: https://archive.ics.uci.edu/ml/datasets/Abalone [Accessed 1 November 2020].

Fish.wa.gov.au. 2020. Abalone. [online] Available at: https://www.fish.wa.gov.au/Species/Abalone/Pages/default.aspx [Accessed 1 November 2020].

Technical Tidbits From Spatial Analysis & Data Science. 2020. Predictive Modeling And Machine Learning In R With The Caret Package. [online] Available at: http://zevross.com/blog/2017/09/19/predictive-modeling-and-machine-learning-in-r-with-the-caret-package/ [Accessed 1 November 2020].

Libraries Used

- **tidyverse 1.3.0** Data manipulation and charting.
- **data.table 1.13.0** Data storage.
- **cowplot 1.1.0** Arranging charts.
- **matrixStats 0.57.0** Row based calculations.
- **caret 6.0-86** Machine learning.

Programming Environment

```
##                _
## platform       x86_64-w64-mingw32
## arch           x86_64
## os             mingw32
## system         x86_64, mingw32
## status
## major          4
## minor          0.2
## year           2020
## month          06
## day            22
## svn rev        78730
## language       R
## version.string R version 4.0.2 (2020-06-22)
## nickname       Taking Off Again
```