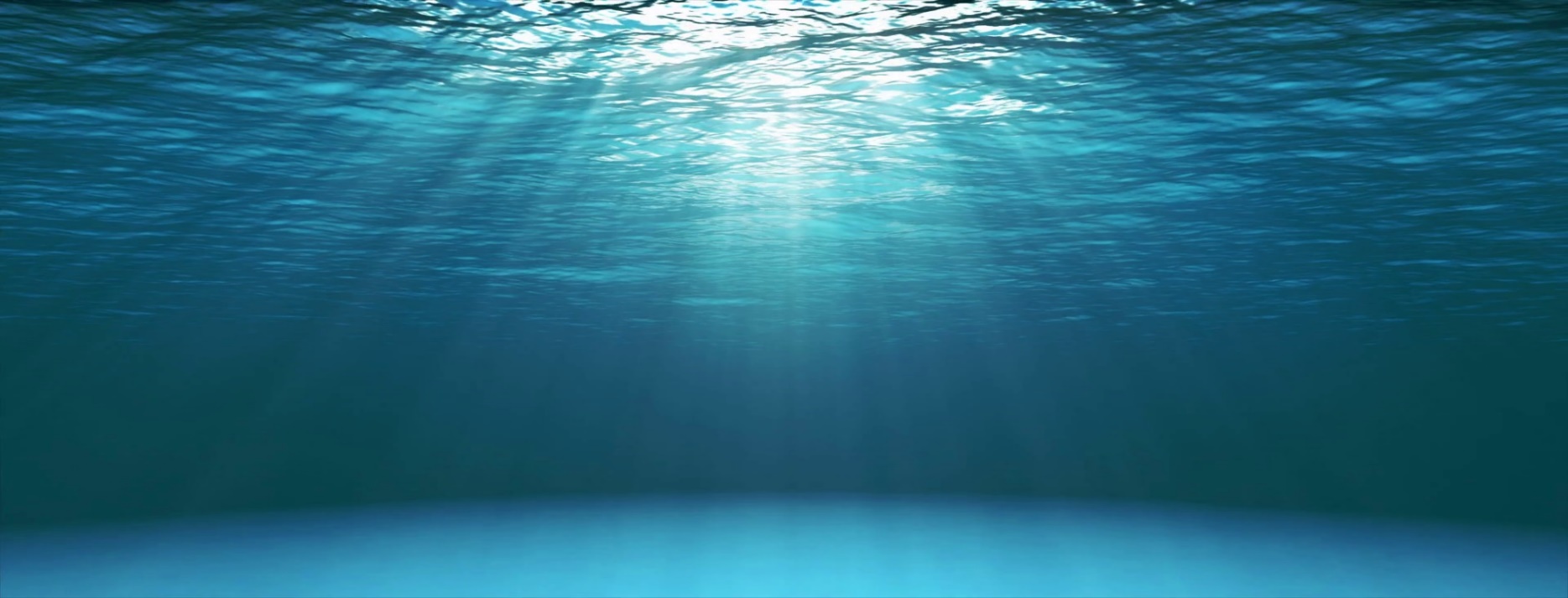
Salmon sales

Addressee: Whole Foods Management, Regional Supervisors and Seafood Team Leaders

Statistical analysis procedures aimed at reducing waste.



|  |  |  |
| --- | --- | --- |
| **Using Analytics to Reduce Waste** | | |
|  |  |  |

**overview**

This project was inspired by the fact that shrink numbers were abnormally high in the seafood department. As this became an ongoing issue, interest was taken in alternative solutions. The one outlined here uses statistical methods to turn raw data into knowledge and understanding. The project was hindered by a lack of data variables combined with a limited span of time. However, conclusions were still significant.

Overall, the project was extensive in nature and required a great deal of planning and reorganization throughout. Below are terms that will be used throughout this summary:

**Data-Join:** a connection between two tables, spreadsheets, or matrixes. The connection is established by aligning the rows/observations based on a shared column/variable.

**Inner-Join:** a type of data-join where there is a one-to-one relationship between the datasets.

**Outer-Join:** a type of data-join where some fields from one table are included even if they don’t have a matching counterpart.

**Query:** queries are used in Microsoft Access to evaluate and display relationships among tables. It is used as a pivot-table for multiple tables.

**RStudio:** a widely-recognized and supported statistical analysis tool.

**RStudio (packages):** RStudio is an open source application that has little to no use until packages are read-in. These packages can be thought of as “functions”. The packages used in this project are among the most commonly deployed and are sound/reliable.

**Regression:** a type of analysis that uses one or more variables to predict the outcome of another.

**General summary**

Not uncommon to analysis work, data-cleaning for this project was intensive and accounted for the vast majority of time spent. It began with visiting the store and taking a look at their files within a secured server that is inaccessible outside of the store. The data was held in a Tableau hosted environment, and was organized in such a way that would be confusing to an ordinary person outside of the company. Time was spent understanding the data at hand, and how to extract something concrete from it. After some time, it was discovered that the significant data could all be saved into multiple CSV files (Microsoft Excel). The data was saved and taken offsite.

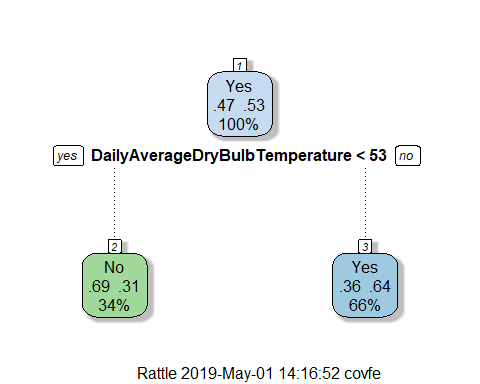
Additional weather data was requested from the NOAA. It was also included in this project.

Private analysis began with viewing the data in excel. All tables were eventually opened and sorted through. Almost one thousand columns were deleted, new variables were created, formatting was done to catalyze data-joins, and extensive notes were taken throughout the process to ensure that a proper understanding of the data was maintained throughout the cleaning. The largest table had close to three hundred columns, and was reduced to thirty nine. Overall there were eight different types of tables, each with multiple instances for different time periods. Unfortunately, most of the data that was removed was specific to Tableau operations and visualizations. What was left was data from observations, which was limited in both scope and period of time. Once cleaning was done, plans were made to join the tables.

There were two inner joins and three outer joins. Formatting issues were encountered throughout which required the data to be readjusted -- using a trial and error process of reading the data in and out of Access. The order and structure of the data was created by sorting all of the rows by date. All of the days’ information is listed with an average of thirty observations per day. Once fully joined, the data was exported back to excel to be saved as a CSV file once again. Some relative variables were sales amount, sales units, current price, average price, average cost, GIG, GIG%, category, subcategory, temperature, humidity, precipitation and windspeed.

The decision was made to only focus on sales data regarding Atlantic Salmon, Farm Raised. This is by far the highest grossing item, as well as the item with the greatest shrink amount. The completed table was filtered to only show the aforementioned data.

Finally, it was time to read the data into RStudio. This was done through standard commands, and the packages used in this project are shown below. Missingness is a term used to describe absent data within a table; this was shown to have a material effect over the integrity of our data. Less-important columns were deleted, and the others were imputed. Imputation is a process of estimating a value based on surrounding values. This filled in the missingness, leaving a full dataset with which to run regression against. With a clean dataset, a package was ran in R to determine to greatest combination of variables in predicting daily sales for Atlantic Salmon. This function revealed that the combination of humidity, temperature, windspeed, and weekend index lead to the most accurate predictions of salmon sales. The model is strong enough to indicate that elevated temperatures lead to higher sales for this time period. The strength of the model is inferred from the relative proximity of the testing accuracy and training accuracy. While the test accuracy was only .6, it still merits significance. The decision tree shows in greater detail, that if the temperature is greater than 53 degrees (NOT less than 53), there is a 66% percent chance that we will sell more than $231 in salmon, and a 34% chance of selling less than that amount. This model is strong, but could be improved greatly if there were more store data to include in the analysis, such as genre of music playing etc. Nevertheless, since this model is dependent on weather, we can still conclude that it is valuable in predicting salmon sales for the time period of February 4th to March 10th.



Final thoughts: the model could be slightly improved if average temperatures were smoothed out by being compared to historical averages, but it would be better to find a way to get twelve months of store data and use the corresponding weather data. Additionally, finding a way to include sales prices would make for an optimal model.

The following pages contain the R-markdown report

Code runs in order from top to bottom.

(Only shown for proof-of-work)

See Excel “Dashboard” for visualization.

**WFM SFD ’19**

**SFD ’19**

Jake Stidham

4/21/2019

**library**(tidyverse)

## -- Attaching packages -------------------------------------------------------------------------------------------------------------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.1.1 v purrr 0.3.2 ## v tibble 2.1.1 v dplyr 0.8.0.1## v tidyr 0.8.3 v stringr 1.4.0 ## v readr 1.3.1 v forcats 0.4.0

## -- Conflicts ----------------------------------------------------------------------------------------------------------------------------------------------------------- tidyverse\_conflicts() --## x dplyr::filter() masks stats::filter()## x dplyr::lag() masks stats::lag()

**library**(GGally)

## ## Attaching package: 'GGally'

## The following object is masked from 'package:dplyr':## ## nasa

**library**(ggcorrplot)**library**(VIM)

## Loading required package: colorspace

## Loading required package: grid

## Loading required package: data.table

## ## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':## ## between, first, last

## The following object is masked from 'package:purrr':## ## transpose

## VIM is ready to use. ## Since version 4.0.0 the GUI is in its own package VIMGUI.## ## Please use the package to use the new (and old) GUI.

## Suggestions and bug-reports can be submitted at: https://github.com/alexkowa/VIM/issues

## ## Attaching package: 'VIM'

## The following object is masked from 'package:datasets':## ## sleep

**library**(ROCR)

## Loading required package: gplots

## ## Attaching package: 'gplots'

## The following object is masked from 'package:stats':## ## lowess

**library**(rpart)**library**(rattle)

## Rattle: A free graphical interface for data science with R.## Version 5.2.0 Copyright (c) 2006-2018 Togaware Pty Ltd.## Type 'rattle()' to shake, rattle, and roll your data.

**library**(RColorBrewer)**library**(e1071)**library**(caret)

## Loading required package: lattice

## ## Attaching package: 'caret'

## The following object is masked from 'package:purrr':## ## lift

**library**(dplyr)**library**(mice)

## ## Attaching package: 'mice'

## The following object is masked from 'package:tidyr':## ## complete

## The following objects are masked from 'package:base':## ## cbind, rbind

**library**(MASS)

## ## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':## ## select

SFD = **read.csv**("Salmon.csv")**summary**(SFD)

## ï..Calendar.Date Day.of.Week.Name Day.of.Week.Num Weekend.Ind ## 18-Feb-19: 2 Friday :5 Min. :1.000 Min. :0.0000 ## 19-Feb-19: 2 Monday :6 1st Qu.:2.000 1st Qu.:0.0000 ## 7-Feb-19 : 2 Saturday :5 Median :4.000 Median :0.0000 ## 1-Mar-19 : 1 Sunday :5 Mean :3.868 Mean :0.2632 ## 10-Feb-19: 1 Thursday :6 3rd Qu.:5.750 3rd Qu.:0.7500 ## 10-Mar-19: 1 Tuesday :6 Max. :7.000 Max. :1.0000 ## (Other) :29 Wednesday:5 ## DailyAverageDryBulbTemperature DailyAverageRelativeHumidity## Min. :39.00 Min. :31.00 ## 1st Qu.:49.25 1st Qu.:62.75 ## Median :54.50 Median :70.00 ## Mean :54.89 Mean :69.68 ## 3rd Qu.:60.00 3rd Qu.:81.25 ## Max. :71.00 Max. :91.00 ## NA's :10 ## DailyAverageStationPressure DailyAverageWindSpeed DailyPrecipitation## Min. :29.77 Min. : 3.200 0 :16 ## 1st Qu.:29.98 1st Qu.: 5.875 T : 6 ## Median :30.06 Median : 8.000 0.01 : 4 ## Mean :30.10 Mean : 8.103 0.03 : 2 ## 3rd Qu.:30.19 3rd Qu.:10.500 0.04 : 1 ## Max. :30.53 Max. :11.800 0.05 : 1 ## (Other): 8 ## ï..Projected.Sales Forecast.Sales ACTUAL.Sales X.NetProjAcc..## Min. :1100 Min. :1074 Min. :1060 13.68% : 2 ## 1st Qu.:1300 1st Qu.:1153 1st Qu.:1381 14.83% : 2 ## Median :1500 Median :1306 Median :1582 24.49% : 2 ## Mean :1601 Mean :1317 Mean :1662 -0.60% : 1 ## 3rd Qu.:1900 3rd Qu.:1456 3rd Qu.:1925 -1.07% : 1 ## Max. :2676 Max. :1748 Max. :2660 -10.90%: 1 ## (Other):29 ## X.NetFcstAcc.. UPC ## -21.62%: 2 Min. :2.92e+10 ## 12.03% : 2 1st Qu.:2.92e+10 ## 52.01% : 2 Median :2.92e+10 ## -14.91%: 1 Mean :2.92e+10 ## -2.30% : 1 3rd Qu.:2.92e+10 ## -21.18%: 1 Max. :2.92e+10 ## (Other):29 ## Item.Description Brand ## SALMON FILLET ATLANTIC FARM RAISED:38 Seafood (in general):38 ## ## ## ## ## ## ## Brand.Code Category.Description Subcategory.Description Expr1020 ## SFOOD:38 Fillet:38 Salmon:38 $9.99 :38 ## ## ## ## ## ## ## Average.Net.Retail.Price Average.Cost.w..SB GIG GIG.. ## Min. :9.98 Min. :5.36 Min. :4.62 46.29%:38 ## 1st Qu.:9.98 1st Qu.:5.36 1st Qu.:4.62 ## Median :9.98 Median :5.36 Median :4.62 ## Mean :9.98 Mean :5.36 Mean :4.62 ## 3rd Qu.:9.98 3rd Qu.:5.36 3rd Qu.:4.62 ## Max. :9.98 Max. :5.36 Max. :4.62 ## ## Current.Effective.Price Net.Unit.Sales.w..Shrink Net.Sales.USD.w..Shrink## $9.99 :38 Min. : 9.78 $212.48 : 2 ## 1st Qu.:20.09 $267.74 : 2 ## Median :23.41 $322.50 : 2 ## Mean :24.36 $117.88 : 1 ## 3rd Qu.:29.00 $153.75 : 1 ## Max. :38.67 $168.01 : 1 ## (Other) :29 ## HighSales Shrink.Units Shrink.USD Unit.Shrink.. Known.Shrink..## No :18 Min. : 0.460 Min. : 3.560 :18 :18 ## Yes:20 1st Qu.: 1.143 1st Qu.: 8.868 10.05% : 1 1.68% : 1 ## Median : 1.815 Median :14.048 11.77% : 1 1.74% : 1 ## Mean : 2.366 Mean :18.418 12.94% : 1 10.48% : 1 ## 3rd Qu.: 2.590 3rd Qu.:19.882 12.96% : 1 11.19% : 1 ## Max. :11.000 Max. :86.240 15.68% : 1 11.52% : 1 ## NA's :18 NA's :18 (Other):15 (Other):15 ## Shrink.Amount..fixed.....fixed.item.... Plan.Sales....month...## Min. :486 Min. :37876 ## 1st Qu.:486 1st Qu.:37876 ## Median :486 Median :37876 ## Mean :486 Mean :38397 ## 3rd Qu.:486 3rd Qu.:39361 ## Max. :486 Max. :39856 ## ## X.Plan.Shrink....Item.month.... Plan.Shrink......Month....## Min. :2053 5.42%:38 ## 1st Qu.:2053 ## Median :2053 ## Mean :2081 ## 3rd Qu.:2133 ## Max. :2160 ##

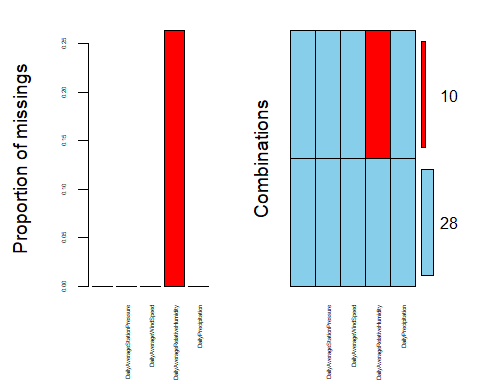
**str**(SFD)

## 'data.frame': 38 obs. of 37 variables:## $ ï..Calendar.Date : Factor w/ 35 levels "1-Mar-19","10-Feb-19",..: 21 14 28 26 29 3 32 17 20 1 ...## $ Day.of.Week.Name : Factor w/ 7 levels "Friday","Monday",..: 7 7 7 6 7 4 1 3 6 1 ...## $ Day.of.Week.Num : int 3 3 3 2 3 7 5 6 2 5 ...## $ Weekend.Ind : int 0 0 0 0 0 1 0 1 0 0 ...## $ DailyAverageDryBulbTemperature : int 59 51 62 58 39 67 68 49 50 57 ...## $ DailyAverageRelativeHumidity : int 75 90 76 68 NA NA 62 91 58 NA ...## $ DailyAverageStationPressure : num 30.1 30.2 30.1 30 30.2 ...## $ DailyAverageWindSpeed : num 5.5 7.2 5.5 4.1 8.7 10.8 11.8 7.6 3.2 6.9 ...## $ DailyPrecipitation : Factor w/ 14 levels "0","0.01","0.03",..: 1 10 1 1 1 1 1 3 1 12 ...## $ ï..Projected.Sales : int 1300 1389 1300 1300 1200 1900 1800 1895 1500 1800 ...## $ Forecast.Sales : int 1167 1108 1137 1148 1074 1394 1382 1367 1105 1363 ...## $ ACTUAL.Sales : int 1391 1237 1122 1500 1446 1728 1588 1530 1208 1858 ...## $ X.NetProjAcc.. : Factor w/ 35 levels "-0.60%","-1.07%",..: 35 4 8 24 26 15 6 10 11 31 ...## $ X.NetFcstAcc.. : Factor w/ 35 levels "-14.91%","-2.30%",..: 18 11 9 5 26 22 1 12 35 27 ...## $ UPC : num 2.92e+10 2.92e+10 2.92e+10 2.92e+10 2.92e+10 ...## $ Item.Description : Factor w/ 1 level "SALMON FILLET ATLANTIC FARM RAISED": 1 1 1 1 1 1 1 1 1 1 ...## $ Brand : Factor w/ 1 level "Seafood (in general)": 1 1 1 1 1 1 1 1 1 1 ...## $ Brand.Code : Factor w/ 1 level "SFOOD": 1 1 1 1 1 1 1 1 1 1 ...## $ Category.Description : Factor w/ 1 level "Fillet": 1 1 1 1 1 1 1 1 1 1 ...## $ Subcategory.Description : Factor w/ 1 level "Salmon": 1 1 1 1 1 1 1 1 1 1 ...## $ Expr1020 : Factor w/ 1 level "$9.99 ": 1 1 1 1 1 1 1 1 1 1 ...## $ Average.Net.Retail.Price : num 9.98 9.98 9.98 9.98 9.98 9.98 9.98 9.98 9.98 9.98 ...## $ Average.Cost.w..SB : num 5.36 5.36 5.36 5.36 5.36 5.36 5.36 5.36 5.36 5.36 ...## $ GIG : num 4.62 4.62 4.62 4.62 4.62 4.62 4.62 4.62 4.62 4.62 ...## $ GIG.. : Factor w/ 1 level "46.29%": 1 1 1 1 1 1 1 1 1 1 ...## $ Current.Effective.Price : Factor w/ 1 level "$9.99 ": 1 1 1 1 1 1 1 1 1 1 ...## $ Net.Unit.Sales.w..Shrink : num 9.78 11.82 15.39 16.82 17.16 ...## $ Net.Sales.USD.w..Shrink : Factor w/ 35 levels "$117.88 ","$153.75 ",..: 35 1 2 3 4 5 6 7 8 9 ...## $ HighSales : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...## $ Shrink.Units : num NA NA 0.6 2.5 NA 2.86 NA NA NA NA ...## $ Shrink.USD : num NA NA 4.64 19.35 NA ...## $ Unit.Shrink.. : Factor w/ 21 levels "","10.05%","11.77%",..: 1 1 13 4 1 5 1 1 1 1 ...## $ Known.Shrink.. : Factor w/ 21 levels "","1.68%","1.74%",..: 1 1 12 6 1 5 1 1 1 1 ...## $ Shrink.Amount..fixed.....fixed.item....: num 486 486 486 486 486 ...## $ Plan.Sales....month... : num 37876 37876 37876 37876 39856 ...## $ X.Plan.Shrink....Item.month.... : num 2053 2053 2053 2053 2160 ...## $ Plan.Shrink......Month.... : Factor w/ 1 level "5.42%": 1 1 1 1 1 1 1 1 1 1 ...

**names**(SFD)[1]<-"CalendarDate"**names**(SFD)[10]<-"ProjectedSales"**names**(SFD)[13]<-"NetProjAcc"**names**(SFD)[14]<-"NetFcstAcc"**names**(SFD)[21]<-"CurrentPrice"**names**(SFD)[35]<-"PlanShrinkUSDMonth"**names**(SFD)[36]<-"PlanShrinkFixedPercentofSales"**names**(SFD)[37]<-"Plan.Shrink.Month"**str**(SFD)

## 'data.frame': 38 obs. of 37 variables:## $ CalendarDate : Factor w/ 35 levels "1-Mar-19","10-Feb-19",..: 21 14 28 26 29 3 32 17 20 1 ...## $ Day.of.Week.Name : Factor w/ 7 levels "Friday","Monday",..: 7 7 7 6 7 4 1 3 6 1 ...## $ Day.of.Week.Num : int 3 3 3 2 3 7 5 6 2 5 ...## $ Weekend.Ind : int 0 0 0 0 0 1 0 1 0 0 ...## $ DailyAverageDryBulbTemperature : int 59 51 62 58 39 67 68 49 50 57 ...## $ DailyAverageRelativeHumidity : int 75 90 76 68 NA NA 62 91 58 NA ...## $ DailyAverageStationPressure : num 30.1 30.2 30.1 30 30.2 ...## $ DailyAverageWindSpeed : num 5.5 7.2 5.5 4.1 8.7 10.8 11.8 7.6 3.2 6.9 ...## $ DailyPrecipitation : Factor w/ 14 levels "0","0.01","0.03",..: 1 10 1 1 1 1 1 3 1 12 ...## $ ProjectedSales : int 1300 1389 1300 1300 1200 1900 1800 1895 1500 1800 ...## $ Forecast.Sales : int 1167 1108 1137 1148 1074 1394 1382 1367 1105 1363 ...## $ ACTUAL.Sales : int 1391 1237 1122 1500 1446 1728 1588 1530 1208 1858 ...## $ NetProjAcc : Factor w/ 35 levels "-0.60%","-1.07%",..: 35 4 8 24 26 15 6 10 11 31 ...## $ NetFcstAcc : Factor w/ 35 levels "-14.91%","-2.30%",..: 18 11 9 5 26 22 1 12 35 27 ...## $ UPC : num 2.92e+10 2.92e+10 2.92e+10 2.92e+10 2.92e+10 ...## $ Item.Description : Factor w/ 1 level "SALMON FILLET ATLANTIC FARM RAISED": 1 1 1 1 1 1 1 1 1 1 ...## $ Brand : Factor w/ 1 level "Seafood (in general)": 1 1 1 1 1 1 1 1 1 1 ...## $ Brand.Code : Factor w/ 1 level "SFOOD": 1 1 1 1 1 1 1 1 1 1 ...## $ Category.Description : Factor w/ 1 level "Fillet": 1 1 1 1 1 1 1 1 1 1 ...## $ Subcategory.Description : Factor w/ 1 level "Salmon": 1 1 1 1 1 1 1 1 1 1 ...## $ CurrentPrice : Factor w/ 1 level "$9.99 ": 1 1 1 1 1 1 1 1 1 1 ...## $ Average.Net.Retail.Price : num 9.98 9.98 9.98 9.98 9.98 9.98 9.98 9.98 9.98 9.98 ...## $ Average.Cost.w..SB : num 5.36 5.36 5.36 5.36 5.36 5.36 5.36 5.36 5.36 5.36 ...## $ GIG : num 4.62 4.62 4.62 4.62 4.62 4.62 4.62 4.62 4.62 4.62 ...## $ GIG.. : Factor w/ 1 level "46.29%": 1 1 1 1 1 1 1 1 1 1 ...## $ Current.Effective.Price : Factor w/ 1 level "$9.99 ": 1 1 1 1 1 1 1 1 1 1 ...## $ Net.Unit.Sales.w..Shrink : num 9.78 11.82 15.39 16.82 17.16 ...## $ Net.Sales.USD.w..Shrink : Factor w/ 35 levels "$117.88 ","$153.75 ",..: 35 1 2 3 4 5 6 7 8 9 ...## $ HighSales : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...## $ Shrink.Units : num NA NA 0.6 2.5 NA 2.86 NA NA NA NA ...## $ Shrink.USD : num NA NA 4.64 19.35 NA ...## $ Unit.Shrink.. : Factor w/ 21 levels "","10.05%","11.77%",..: 1 1 13 4 1 5 1 1 1 1 ...## $ Known.Shrink.. : Factor w/ 21 levels "","1.68%","1.74%",..: 1 1 12 6 1 5 1 1 1 1 ...## $ Shrink.Amount..fixed.....fixed.item....: num 486 486 486 486 486 ...## $ PlanShrinkUSDMonth : num 37876 37876 37876 37876 39856 ...## $ PlanShrinkFixedPercentofSales : num 2053 2053 2053 2053 2160 ...## $ Plan.Shrink.Month : Factor w/ 1 level "5.42%": 1 1 1 1 1 1 1 1 1 1 ...

weather = SFD **%>%** dplyr**::select**(**c**(DailyAverageDryBulbTemperature,DailyAverageStationPressure,DailyAverageWindSpeed,DailyAverageRelativeHumidity,DailyPrecipitation)) vim\_plot = **aggr**(weather, numbers = TRUE, prop = **c**(TRUE, FALSE),cex.axis=.4)



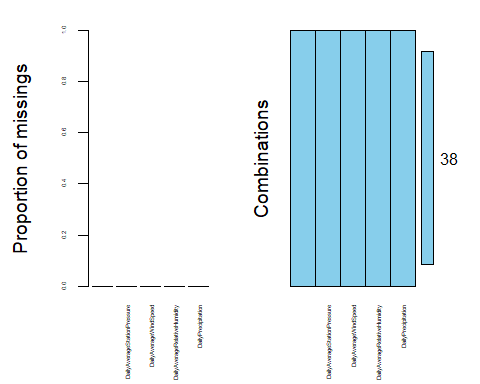
**set.seed**(1234) imp\_sfd = **mice**(weather, m=1, method='pmm', printFlag=FALSE)

## Warning: Number of logged events: 5

weather\_complete = **complete**(imp\_sfd) **summary**(weather\_complete)

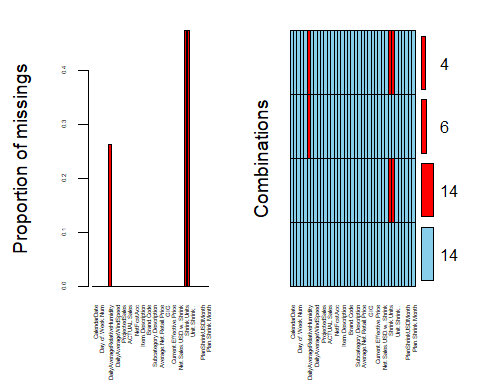
## DailyAverageDryBulbTemperature DailyAverageStationPressure## Min. :39.00 Min. :29.77 ## 1st Qu.:49.25 1st Qu.:29.98 ## Median :54.50 Median :30.06 ## Mean :54.89 Mean :30.10 ## 3rd Qu.:60.00 3rd Qu.:30.19 ## Max. :71.00 Max. :30.53 ## ## DailyAverageWindSpeed DailyAverageRelativeHumidity DailyPrecipitation## Min. : 3.200 Min. :31.00 0 :16 ## 1st Qu.: 5.875 1st Qu.:55.75 T : 6 ## Median : 8.000 Median :69.00 0.01 : 4 ## Mean : 8.103 Mean :66.87 0.03 : 2 ## 3rd Qu.:10.500 3rd Qu.:76.00 0.04 : 1 ## Max. :11.800 Max. :91.00 0.05 : 1 ## (Other): 8

vim\_plot = **aggr**(weather\_complete, numbers = TRUE, prop = **c**(TRUE, FALSE),cex.axis=.4)



Appending Tables

vim\_plot = **aggr**(SFD, numbers = TRUE, prop = **c**(TRUE, FALSE),cex.axis=.4)



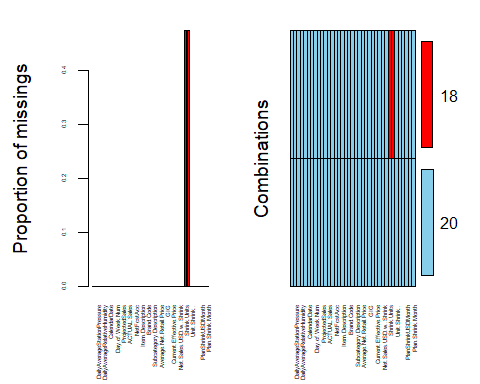
SFD = SFD **%>%** dplyr**::select**(**c**(**-**DailyAverageDryBulbTemperature,**-**DailyAverageStationPressure,**-**DailyAverageWindSpeed,**-**DailyAverageRelativeHumidity,**-**DailyPrecipitation)) **str**(SFD)

## 'data.frame': 38 obs. of 32 variables:## $ CalendarDate : Factor w/ 35 levels "1-Mar-19","10-Feb-19",..: 21 14 28 26 29 3 32 17 20 1 ...## $ Day.of.Week.Name : Factor w/ 7 levels "Friday","Monday",..: 7 7 7 6 7 4 1 3 6 1 ...## $ Day.of.Week.Num : int 3 3 3 2 3 7 5 6 2 5 ...## $ Weekend.Ind : int 0 0 0 0 0 1 0 1 0 0 ...## $ ProjectedSales : int 1300 1389 1300 1300 1200 1900 1800 1895 1500 1800 ...## $ Forecast.Sales : int 1167 1108 1137 1148 1074 1394 1382 1367 1105 1363 ...## $ ACTUAL.Sales : int 1391 1237 1122 1500 1446 1728 1588 1530 1208 1858 ...## $ NetProjAcc : Factor w/ 35 levels "-0.60%","-1.07%",..: 35 4 8 24 26 15 6 10 11 31 ...## $ NetFcstAcc : Factor w/ 35 levels "-14.91%","-2.30%",..: 18 11 9 5 26 22 1 12 35 27 ...## $ UPC : num 2.92e+10 2.92e+10 2.92e+10 2.92e+10 2.92e+10 ...## $ Item.Description : Factor w/ 1 level "SALMON FILLET ATLANTIC FARM RAISED": 1 1 1 1 1 1 1 1 1 1 ...## $ Brand : Factor w/ 1 level "Seafood (in general)": 1 1 1 1 1 1 1 1 1 1 ...## $ Brand.Code : Factor w/ 1 level "SFOOD": 1 1 1 1 1 1 1 1 1 1 ...## $ Category.Description : Factor w/ 1 level "Fillet": 1 1 1 1 1 1 1 1 1 1 ...## $ Subcategory.Description : Factor w/ 1 level "Salmon": 1 1 1 1 1 1 1 1 1 1 ...## $ CurrentPrice : Factor w/ 1 level "$9.99 ": 1 1 1 1 1 1 1 1 1 1 ...## $ Average.Net.Retail.Price : num 9.98 9.98 9.98 9.98 9.98 9.98 9.98 9.98 9.98 9.98 ...## $ Average.Cost.w..SB : num 5.36 5.36 5.36 5.36 5.36 5.36 5.36 5.36 5.36 5.36 ...## $ GIG : num 4.62 4.62 4.62 4.62 4.62 4.62 4.62 4.62 4.62 4.62 ...## $ GIG.. : Factor w/ 1 level "46.29%": 1 1 1 1 1 1 1 1 1 1 ...## $ Current.Effective.Price : Factor w/ 1 level "$9.99 ": 1 1 1 1 1 1 1 1 1 1 ...## $ Net.Unit.Sales.w..Shrink : num 9.78 11.82 15.39 16.82 17.16 ...## $ Net.Sales.USD.w..Shrink : Factor w/ 35 levels "$117.88 ","$153.75 ",..: 35 1 2 3 4 5 6 7 8 9 ...## $ HighSales : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...## $ Shrink.Units : num NA NA 0.6 2.5 NA 2.86 NA NA NA NA ...## $ Shrink.USD : num NA NA 4.64 19.35 NA ...## $ Unit.Shrink.. : Factor w/ 21 levels "","10.05%","11.77%",..: 1 1 13 4 1 5 1 1 1 1 ...## $ Known.Shrink.. : Factor w/ 21 levels "","1.68%","1.74%",..: 1 1 12 6 1 5 1 1 1 1 ...## $ Shrink.Amount..fixed.....fixed.item....: num 486 486 486 486 486 ...## $ PlanShrinkUSDMonth : num 37876 37876 37876 37876 39856 ...## $ PlanShrinkFixedPercentofSales : num 2053 2053 2053 2053 2160 ...## $ Plan.Shrink.Month : Factor w/ 1 level "5.42%": 1 1 1 1 1 1 1 1 1 1 ...

SFD\_Full = **cbind**(weather\_complete,SFD)**str**(SFD\_Full)

## 'data.frame': 38 obs. of 37 variables:## $ DailyAverageDryBulbTemperature : int 59 51 62 58 39 67 68 49 50 57 ...## $ DailyAverageStationPressure : num 30.1 30.2 30.1 30 30.2 ...## $ DailyAverageWindSpeed : num 5.5 7.2 5.5 4.1 8.7 10.8 11.8 7.6 3.2 6.9 ...## $ DailyAverageRelativeHumidity : int 75 90 76 68 31 63 62 91 58 62 ...## $ DailyPrecipitation : Factor w/ 14 levels "0","0.01","0.03",..: 1 10 1 1 1 1 1 3 1 12 ...## $ CalendarDate : Factor w/ 35 levels "1-Mar-19","10-Feb-19",..: 21 14 28 26 29 3 32 17 20 1 ...## $ Day.of.Week.Name : Factor w/ 7 levels "Friday","Monday",..: 7 7 7 6 7 4 1 3 6 1 ...## $ Day.of.Week.Num : int 3 3 3 2 3 7 5 6 2 5 ...## $ Weekend.Ind : int 0 0 0 0 0 1 0 1 0 0 ...## $ ProjectedSales : int 1300 1389 1300 1300 1200 1900 1800 1895 1500 1800 ...## $ Forecast.Sales : int 1167 1108 1137 1148 1074 1394 1382 1367 1105 1363 ...## $ ACTUAL.Sales : int 1391 1237 1122 1500 1446 1728 1588 1530 1208 1858 ...## $ NetProjAcc : Factor w/ 35 levels "-0.60%","-1.07%",..: 35 4 8 24 26 15 6 10 11 31 ...## $ NetFcstAcc : Factor w/ 35 levels "-14.91%","-2.30%",..: 18 11 9 5 26 22 1 12 35 27 ...## $ UPC : num 2.92e+10 2.92e+10 2.92e+10 2.92e+10 2.92e+10 ...## $ Item.Description : Factor w/ 1 level "SALMON FILLET ATLANTIC FARM RAISED": 1 1 1 1 1 1 1 1 1 1 ...## $ Brand : Factor w/ 1 level "Seafood (in general)": 1 1 1 1 1 1 1 1 1 1 ...## $ Brand.Code : Factor w/ 1 level "SFOOD": 1 1 1 1 1 1 1 1 1 1 ...## $ Category.Description : Factor w/ 1 level "Fillet": 1 1 1 1 1 1 1 1 1 1 ...## $ Subcategory.Description : Factor w/ 1 level "Salmon": 1 1 1 1 1 1 1 1 1 1 ...## $ CurrentPrice : Factor w/ 1 level "$9.99 ": 1 1 1 1 1 1 1 1 1 1 ...## $ Average.Net.Retail.Price : num 9.98 9.98 9.98 9.98 9.98 9.98 9.98 9.98 9.98 9.98 ...## $ Average.Cost.w..SB : num 5.36 5.36 5.36 5.36 5.36 5.36 5.36 5.36 5.36 5.36 ...## $ GIG : num 4.62 4.62 4.62 4.62 4.62 4.62 4.62 4.62 4.62 4.62 ...## $ GIG.. : Factor w/ 1 level "46.29%": 1 1 1 1 1 1 1 1 1 1 ...## $ Current.Effective.Price : Factor w/ 1 level "$9.99 ": 1 1 1 1 1 1 1 1 1 1 ...## $ Net.Unit.Sales.w..Shrink : num 9.78 11.82 15.39 16.82 17.16 ...## $ Net.Sales.USD.w..Shrink : Factor w/ 35 levels "$117.88 ","$153.75 ",..: 35 1 2 3 4 5 6 7 8 9 ...## $ HighSales : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...## $ Shrink.Units : num NA NA 0.6 2.5 NA 2.86 NA NA NA NA ...## $ Shrink.USD : num NA NA 4.64 19.35 NA ...## $ Unit.Shrink.. : Factor w/ 21 levels "","10.05%","11.77%",..: 1 1 13 4 1 5 1 1 1 1 ...## $ Known.Shrink.. : Factor w/ 21 levels "","1.68%","1.74%",..: 1 1 12 6 1 5 1 1 1 1 ...## $ Shrink.Amount..fixed.....fixed.item....: num 486 486 486 486 486 ...## $ PlanShrinkUSDMonth : num 37876 37876 37876 37876 39856 ...## $ PlanShrinkFixedPercentofSales : num 2053 2053 2053 2053 2160 ...## $ Plan.Shrink.Month : Factor w/ 1 level "5.42%": 1 1 1 1 1 1 1 1 1 1 ...

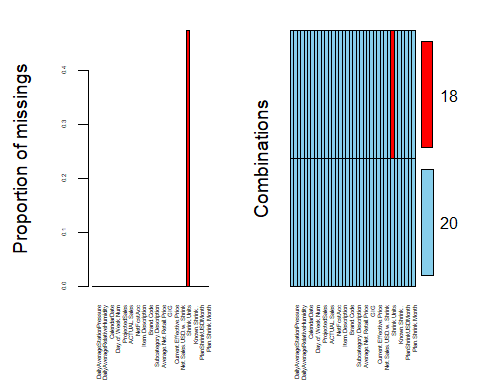
vim\_plot = **aggr**(SFD\_Full, numbers = TRUE, prop = **c**(TRUE, FALSE),cex.axis=.4)



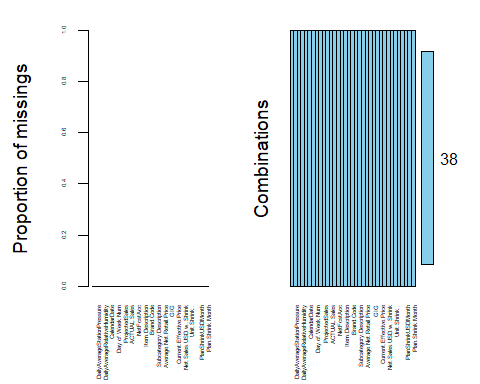
**str**(SFD\_Full)

## 'data.frame': 38 obs. of 37 variables:## $ DailyAverageDryBulbTemperature : int 59 51 62 58 39 67 68 49 50 57 ...## $ DailyAverageStationPressure : num 30.1 30.2 30.1 30 30.2 ...## $ DailyAverageWindSpeed : num 5.5 7.2 5.5 4.1 8.7 10.8 11.8 7.6 3.2 6.9 ...## $ DailyAverageRelativeHumidity : int 75 90 76 68 31 63 62 91 58 62 ...## $ DailyPrecipitation : Factor w/ 14 levels "0","0.01","0.03",..: 1 10 1 1 1 1 1 3 1 12 ...## $ CalendarDate : Factor w/ 35 levels "1-Mar-19","10-Feb-19",..: 21 14 28 26 29 3 32 17 20 1 ...## $ Day.of.Week.Name : Factor w/ 7 levels "Friday","Monday",..: 7 7 7 6 7 4 1 3 6 1 ...## $ Day.of.Week.Num : int 3 3 3 2 3 7 5 6 2 5 ...## $ Weekend.Ind : int 0 0 0 0 0 1 0 1 0 0 ...## $ ProjectedSales : int 1300 1389 1300 1300 1200 1900 1800 1895 1500 1800 ...## $ Forecast.Sales : int 1167 1108 1137 1148 1074 1394 1382 1367 1105 1363 ...## $ ACTUAL.Sales : int 1391 1237 1122 1500 1446 1728 1588 1530 1208 1858 ...## $ NetProjAcc : Factor w/ 35 levels "-0.60%","-1.07%",..: 35 4 8 24 26 15 6 10 11 31 ...## $ NetFcstAcc : Factor w/ 35 levels "-14.91%","-2.30%",..: 18 11 9 5 26 22 1 12 35 27 ...## $ UPC : num 2.92e+10 2.92e+10 2.92e+10 2.92e+10 2.92e+10 ...## $ Item.Description : Factor w/ 1 level "SALMON FILLET ATLANTIC FARM RAISED": 1 1 1 1 1 1 1 1 1 1 ...## $ Brand : Factor w/ 1 level "Seafood (in general)": 1 1 1 1 1 1 1 1 1 1 ...## $ Brand.Code : Factor w/ 1 level "SFOOD": 1 1 1 1 1 1 1 1 1 1 ...## $ Category.Description : Factor w/ 1 level "Fillet": 1 1 1 1 1 1 1 1 1 1 ...## $ Subcategory.Description : Factor w/ 1 level "Salmon": 1 1 1 1 1 1 1 1 1 1 ...## $ CurrentPrice : Factor w/ 1 level "$9.99 ": 1 1 1 1 1 1 1 1 1 1 ...## $ Average.Net.Retail.Price : num 9.98 9.98 9.98 9.98 9.98 9.98 9.98 9.98 9.98 9.98 ...## $ Average.Cost.w..SB : num 5.36 5.36 5.36 5.36 5.36 5.36 5.36 5.36 5.36 5.36 ...## $ GIG : num 4.62 4.62 4.62 4.62 4.62 4.62 4.62 4.62 4.62 4.62 ...## $ GIG.. : Factor w/ 1 level "46.29%": 1 1 1 1 1 1 1 1 1 1 ...## $ Current.Effective.Price : Factor w/ 1 level "$9.99 ": 1 1 1 1 1 1 1 1 1 1 ...## $ Net.Unit.Sales.w..Shrink : num 9.78 11.82 15.39 16.82 17.16 ...## $ Net.Sales.USD.w..Shrink : Factor w/ 35 levels "$117.88 ","$153.75 ",..: 35 1 2 3 4 5 6 7 8 9 ...## $ HighSales : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...## $ Shrink.Units : num NA NA 0.6 2.5 NA 2.86 NA NA NA NA ...## $ Shrink.USD : num NA NA 4.64 19.35 NA ...## $ Unit.Shrink.. : Factor w/ 21 levels "","10.05%","11.77%",..: 1 1 13 4 1 5 1 1 1 1 ...## $ Known.Shrink.. : Factor w/ 21 levels "","1.68%","1.74%",..: 1 1 12 6 1 5 1 1 1 1 ...## $ Shrink.Amount..fixed.....fixed.item....: num 486 486 486 486 486 ...## $ PlanShrinkUSDMonth : num 37876 37876 37876 37876 39856 ...## $ PlanShrinkFixedPercentofSales : num 2053 2053 2053 2053 2160 ...## $ Plan.Shrink.Month : Factor w/ 1 level "5.42%": 1 1 1 1 1 1 1 1 1 1 ...

SFD\_Full = SFD\_Full **%>%** dplyr**::select**(**-**`Shrink.USD`)vim\_plot = **aggr**(SFD\_Full, numbers = TRUE, prop = **c**(TRUE, FALSE),cex.axis=.4)



SFD\_Full = SFD\_Full **%>%** dplyr**::select**(**-**`Shrink.Units`)vim\_plot = **aggr**(SFD\_Full, numbers = TRUE, prop = **c**(TRUE, FALSE),cex.axis=.4)



SFD\_Full = SFD\_Full **%>%** dplyr**::select**(**-**`Item.Description`)SFD\_Full = SFD\_Full **%>%** dplyr**::select**(**-**`Brand`)SFD\_Full = SFD\_Full **%>%** dplyr**::select**(**-**`Brand.Code`)SFD\_Full = SFD\_Full **%>%** dplyr**::select**(**-**`Category.Description`)SFD\_Full = SFD\_Full **%>%** dplyr**::select**(**-**`Subcategory.Description`)SFD\_Full = SFD\_Full **%>%** dplyr**::select**(**-**`CurrentPrice`)SFD\_Full = SFD\_Full **%>%** dplyr**::select**(**-**`GIG..`)SFD\_Full = SFD\_Full **%>%** dplyr**::select**(**-**`Current.Effective.Price`)SFD\_Full = SFD\_Full **%>%** dplyr**::select**(**-**`Plan.Shrink.Month`)SFD\_Full = SFD\_Full **%>%** **drop\_na**()**str**(SFD\_Full)

## 'data.frame': 38 obs. of 26 variables:## $ DailyAverageDryBulbTemperature : int 59 51 62 58 39 67 68 49 50 57 ...## $ DailyAverageStationPressure : num 30.1 30.2 30.1 30 30.2 ...## $ DailyAverageWindSpeed : num 5.5 7.2 5.5 4.1 8.7 10.8 11.8 7.6 3.2 6.9 ...## $ DailyAverageRelativeHumidity : int 75 90 76 68 31 63 62 91 58 62 ...## $ DailyPrecipitation : Factor w/ 14 levels "0","0.01","0.03",..: 1 10 1 1 1 1 1 3 1 12 ...## $ CalendarDate : Factor w/ 35 levels "1-Mar-19","10-Feb-19",..: 21 14 28 26 29 3 32 17 20 1 ...## $ Day.of.Week.Name : Factor w/ 7 levels "Friday","Monday",..: 7 7 7 6 7 4 1 3 6 1 ...## $ Day.of.Week.Num : int 3 3 3 2 3 7 5 6 2 5 ...## $ Weekend.Ind : int 0 0 0 0 0 1 0 1 0 0 ...## $ ProjectedSales : int 1300 1389 1300 1300 1200 1900 1800 1895 1500 1800 ...## $ Forecast.Sales : int 1167 1108 1137 1148 1074 1394 1382 1367 1105 1363 ...## $ ACTUAL.Sales : int 1391 1237 1122 1500 1446 1728 1588 1530 1208 1858 ...## $ NetProjAcc : Factor w/ 35 levels "-0.60%","-1.07%",..: 35 4 8 24 26 15 6 10 11 31 ...## $ NetFcstAcc : Factor w/ 35 levels "-14.91%","-2.30%",..: 18 11 9 5 26 22 1 12 35 27 ...## $ UPC : num 2.92e+10 2.92e+10 2.92e+10 2.92e+10 2.92e+10 ...## $ Average.Net.Retail.Price : num 9.98 9.98 9.98 9.98 9.98 9.98 9.98 9.98 9.98 9.98 ...## $ Average.Cost.w..SB : num 5.36 5.36 5.36 5.36 5.36 5.36 5.36 5.36 5.36 5.36 ...## $ GIG : num 4.62 4.62 4.62 4.62 4.62 4.62 4.62 4.62 4.62 4.62 ...## $ Net.Unit.Sales.w..Shrink : num 9.78 11.82 15.39 16.82 17.16 ...## $ Net.Sales.USD.w..Shrink : Factor w/ 35 levels "$117.88 ","$153.75 ",..: 35 1 2 3 4 5 6 7 8 9 ...## $ HighSales : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...## $ Unit.Shrink.. : Factor w/ 21 levels "","10.05%","11.77%",..: 1 1 13 4 1 5 1 1 1 1 ...## $ Known.Shrink.. : Factor w/ 21 levels "","1.68%","1.74%",..: 1 1 12 6 1 5 1 1 1 1 ...## $ Shrink.Amount..fixed.....fixed.item....: num 486 486 486 486 486 ...## $ PlanShrinkUSDMonth : num 37876 37876 37876 37876 39856 ...## $ PlanShrinkFixedPercentofSales : num 2053 2053 2053 2053 2160 ...

train.rows = **createDataPartition**(y = SFD\_Full**$**HighSales, p=0.6, list = FALSE) train = SFD\_Full[train.rows,] test = SFD\_Full[**-**train.rows,]

allmod = **glm**(HighSales **~** DailyAverageDryBulbTemperature **+** DailyAverageWindSpeed **+** DailyAverageRelativeHumidity **+** DailyPrecipitation **+** Day.of.Week.Num **+** Weekend.Ind, train, family = "binomial")

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

**summary**(allmod)

## ## Call:## glm(formula = HighSales ~ DailyAverageDryBulbTemperature + DailyAverageWindSpeed + ## DailyAverageRelativeHumidity + DailyPrecipitation + Day.of.Week.Num + ## Weekend.Ind, family = "binomial", data = train)## ## Deviance Residuals: ## Min 1Q Median 3Q Max ## -1.328e-05 -2.409e-06 2.110e-08 1.215e-06 1.315e-05 ## ## Coefficients:## Estimate Std. Error z value Pr(>|z|)## (Intercept) -1.882e+02 9.319e+05 0.000 1## DailyAverageDryBulbTemperature 1.200e+01 1.920e+04 0.001 1## DailyAverageWindSpeed -2.173e+01 8.085e+04 0.000 1## DailyAverageRelativeHumidity -7.766e+00 2.037e+04 0.000 1## DailyPrecipitation0.01 2.547e+02 1.357e+06 0.000 1## DailyPrecipitation0.05 2.687e+02 1.123e+06 0.000 1## DailyPrecipitation0.06 1.918e+02 5.995e+05 0.000 1## DailyPrecipitation0.13 -3.686e+01 4.690e+05 0.000 1## DailyPrecipitation0.24 2.893e+02 7.125e+05 0.000 1## DailyPrecipitation0.29 2.782e+02 9.796e+05 0.000 1## DailyPrecipitation0.37 1.992e+02 5.768e+05 0.000 1## DailyPrecipitation0.44 -1.023e+02 4.668e+05 0.000 1## DailyPrecipitationT 1.564e+01 6.015e+05 0.000 1## Day.of.Week.Num 4.227e+01 6.937e+04 0.001 1## Weekend.Ind 2.166e+01 2.852e+05 0.000 1## ## (Dispersion parameter for binomial family taken to be 1)## ## Null deviance: 3.1841e+01 on 22 degrees of freedom## Residual deviance: 7.2396e-10 on 8 degrees of freedom## AIC: 30## ## Number of Fisher Scoring iterations: 25

emptymod = **glm**(HighSales **~**1, train, family = "binomial") *#use ~1 to build an empty model***summary**(emptymod)

## ## Call:## glm(formula = HighSales ~ 1, family = "binomial", data = train)## ## Deviance Residuals: ## Min 1Q Median 3Q Max ## -1.215 -1.215 1.141 1.141 1.141 ## ## Coefficients:## Estimate Std. Error z value Pr(>|z|)## (Intercept) 0.08701 0.41742 0.208 0.835## ## (Dispersion parameter for binomial family taken to be 1)## ## Null deviance: 31.841 on 22 degrees of freedom## Residual deviance: 31.841 on 22 degrees of freedom## AIC: 33.841## ## Number of Fisher Scoring iterations: 3

Backward stepwise (reusing code from linear regression)

*#backward*backmod = **stepAIC**(allmod, direction = "backward", trace = TRUE)

## Start: AIC=30## HighSales ~ DailyAverageDryBulbTemperature + DailyAverageWindSpeed + ## DailyAverageRelativeHumidity + DailyPrecipitation + Day.of.Week.Num + ## Weekend.Ind

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC## - Weekend.Ind 1 0.0000 28.000## - DailyAverageWindSpeed 1 0.0000 28.000## - DailyAverageRelativeHumidity 1 0.0000 28.000## <none> 0.0000 30.000## - DailyPrecipitation 9 20.2184 32.218## - Day.of.Week.Num 1 7.7218 35.722## - DailyAverageDryBulbTemperature 1 15.9848 43.985

## Warning: glm.fit: algorithm did not converge## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## ## Step: AIC=28## HighSales ~ DailyAverageDryBulbTemperature + DailyAverageWindSpeed + ## DailyAverageRelativeHumidity + DailyPrecipitation + Day.of.Week.Num

## Warning: glm.fit: algorithm did not converge## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC## - DailyAverageWindSpeed 1 0.000 26.000## - DailyAverageRelativeHumidity 1 0.000 26.000## <none> 0.000 28.000## - DailyPrecipitation 9 23.768 33.768## - Day.of.Week.Num 1 13.783 39.783## - DailyAverageDryBulbTemperature 1 17.404 43.403

## Warning: glm.fit: algorithm did not converge## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## ## Step: AIC=26## HighSales ~ DailyAverageDryBulbTemperature + DailyAverageRelativeHumidity + ## DailyPrecipitation + Day.of.Week.Num## ## Df Deviance AIC## <none> 0.000 26.000## - DailyPrecipitation 9 24.321 32.321## - DailyAverageRelativeHumidity 1 12.925 36.925## - Day.of.Week.Num 1 13.988 37.988## - DailyAverageDryBulbTemperature 1 18.361 42.361

**summary**(backmod)

## ## Call:## glm(formula = HighSales ~ DailyAverageDryBulbTemperature + DailyAverageRelativeHumidity + ## DailyPrecipitation + Day.of.Week.Num, family = "binomial", ## data = train)## ## Deviance Residuals: ## Min 1Q Median 3Q Max ## -1.934e-05 -2.409e-06 2.110e-08 1.215e-06 1.652e-05 ## ## Coefficients:## Estimate Std. Error z value Pr(>|z|)## (Intercept) -599.970 641894.126 -0.001 0.999## DailyAverageDryBulbTemperature 13.308 14122.241 0.001 0.999## DailyAverageRelativeHumidity -4.943 5459.530 -0.001 0.999## DailyPrecipitation0.01 131.025 356901.672 0.000 1.000## DailyPrecipitation0.05 -11.605 368160.956 0.000 1.000## DailyPrecipitation0.06 122.485 373124.231 0.000 1.000## DailyPrecipitation0.13 -143.830 382855.892 0.000 1.000## DailyPrecipitation0.24 267.362 445327.121 0.001 1.000## DailyPrecipitation0.29 127.790 399525.213 0.000 1.000## DailyPrecipitation0.37 185.071 421484.101 0.000 1.000## DailyPrecipitation0.44 -231.666 421232.799 -0.001 1.000## DailyPrecipitationT -156.064 198211.373 -0.001 0.999## Day.of.Week.Num 70.597 76963.852 0.001 0.999## ## (Dispersion parameter for binomial family taken to be 1)## ## Null deviance: 3.1841e+01 on 22 degrees of freedom## Residual deviance: 1.3143e-09 on 10 degrees of freedom## AIC: 26## ## Number of Fisher Scoring iterations: 25

Forward stepwise (again reusing code)

*#forward*forwardmod = **stepAIC**(emptymod, direction = "forward", scope=**list**(upper=allmod,lower=emptymod), trace = TRUE)

## Start: AIC=33.84## HighSales ~ 1## ## Df Deviance AIC## + DailyAverageDryBulbTemperature 1 24.896 28.896## + Weekend.Ind 1 27.526 31.526## <none> 31.841 33.841## + DailyAverageWindSpeed 1 30.127 34.127## + Day.of.Week.Num 1 31.031 35.031## + DailyAverageRelativeHumidity 1 31.822 35.822## + DailyPrecipitation 9 20.683 40.683## ## Step: AIC=28.9## HighSales ~ DailyAverageDryBulbTemperature## ## Df Deviance AIC## + Weekend.Ind 1 22.128 28.128## <none> 24.896 28.896## + DailyAverageWindSpeed 1 24.154 30.154## + DailyAverageRelativeHumidity 1 24.363 30.363## + Day.of.Week.Num 1 24.888 30.888## + DailyPrecipitation 9 15.865 37.865## ## Step: AIC=28.13## HighSales ~ DailyAverageDryBulbTemperature + Weekend.Ind## ## Df Deviance AIC## <none> 22.128 28.128## + Day.of.Week.Num 1 21.235 29.235## + DailyAverageWindSpeed 1 21.277 29.277## + DailyAverageRelativeHumidity 1 21.522 29.522## + DailyPrecipitation 9 12.540 36.541

**summary**(forwardmod)

## ## Call:## glm(formula = HighSales ~ DailyAverageDryBulbTemperature + Weekend.Ind, ## family = "binomial", data = train)## ## Deviance Residuals: ## Min 1Q Median 3Q Max ## -1.6248 -0.7990 0.0001 0.8389 1.7449 ## ## Coefficients:## Estimate Std. Error z value Pr(>|z|) ## (Intercept) -7.8462 4.0184 -1.953 0.0509 .## DailyAverageDryBulbTemperature 0.1428 0.0744 1.920 0.0549 .## Weekend.Ind 18.2013 3701.8483 0.005 0.9961 ## ---## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1## ## (Dispersion parameter for binomial family taken to be 1)## ## Null deviance: 31.841 on 22 degrees of freedom## Residual deviance: 22.128 on 20 degrees of freedom## AIC: 28.128## ## Number of Fisher Scoring iterations: 17

**summary**(backmod)

## ## Call:## glm(formula = HighSales ~ DailyAverageDryBulbTemperature + DailyAverageRelativeHumidity + ## DailyPrecipitation + Day.of.Week.Num, family = "binomial", ## data = train)## ## Deviance Residuals: ## Min 1Q Median 3Q Max ## -1.934e-05 -2.409e-06 2.110e-08 1.215e-06 1.652e-05 ## ## Coefficients:## Estimate Std. Error z value Pr(>|z|)## (Intercept) -599.970 641894.126 -0.001 0.999## DailyAverageDryBulbTemperature 13.308 14122.241 0.001 0.999## DailyAverageRelativeHumidity -4.943 5459.530 -0.001 0.999## DailyPrecipitation0.01 131.025 356901.672 0.000 1.000## DailyPrecipitation0.05 -11.605 368160.956 0.000 1.000## DailyPrecipitation0.06 122.485 373124.231 0.000 1.000## DailyPrecipitation0.13 -143.830 382855.892 0.000 1.000## DailyPrecipitation0.24 267.362 445327.121 0.001 1.000## DailyPrecipitation0.29 127.790 399525.213 0.000 1.000## DailyPrecipitation0.37 185.071 421484.101 0.000 1.000## DailyPrecipitation0.44 -231.666 421232.799 -0.001 1.000## DailyPrecipitationT -156.064 198211.373 -0.001 0.999## Day.of.Week.Num 70.597 76963.852 0.001 0.999## ## (Dispersion parameter for binomial family taken to be 1)## ## Null deviance: 3.1841e+01 on 22 degrees of freedom## Residual deviance: 1.3143e-09 on 10 degrees of freedom## AIC: 26## ## Number of Fisher Scoring iterations: 25

salmon = **glm**(formula = HighSales **~** DailyAverageDryBulbTemperature **+** DailyAverageRelativeHumidity **+**  DailyPrecipitation **+** Day.of.Week.Num, family = "binomial",  data = train)

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

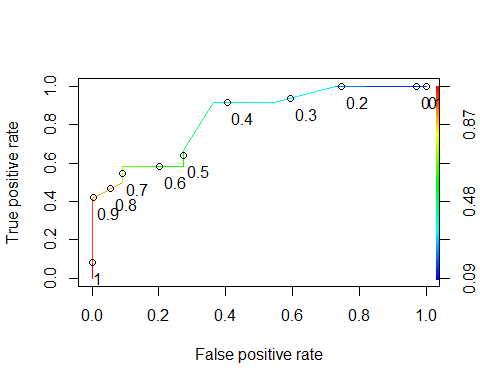
Develop predictions

predictions = **predict**(forwardmod, type = "response") **head**(predictions)

## 2 4 5 9 10 11 ## 0.36309565 0.60774551 0.09313922 0.33075294 0.57322393 0.09313922

Threshold selection

ROCRpred = **prediction**(predictions, train**$**HighSales) ROCRperf = **performance**(ROCRpred, "tpr", "fpr")**plot**(ROCRperf, colorize=TRUE, print.cutoffs.at=**seq**(0,1,by=0.1), text.adj=**c**(**-**0.2,1.7))



*### When we ask what is the "threshold" we are talking about the values on the line within the graph. Threshold of 0 = (1,1)  
### sensitivity is the y axis  
### (1-specificity) is the x-axis  
### "positives" are survived = yes  
### more specific = more negative*

Area under the curve (AUC). AUC is a measure of the strength of the model. Values closer to 1 are better. Can be used to compare models.

**as.numeric**(**performance**(ROCRpred, "auc")**@**y.values)

## [1] 0.8333333

*###>>>>>>> This line addresses how much of the "perfect right-angle" is under our curve.*

*#Determine threshold to balance sensitivity and specificity*opt.cut = **function**(perf, pred) {cut.ind = **mapply**(FUN=**function**(x, y, p) {d = (x **-** 0)**^**2 **+** (y-1)**^**2 ind = **which**(d **==** **min**(d)) **c**(sensitivity = y[[ind]], specificity = 1**-**x[[ind]],  cutoff = p[[ind]])}, perf**@**x.values, perf**@**y.values, pred**@**cutoffs)}**print**(**opt.cut**(ROCRperf, ROCRpred))

## [,1]## sensitivity 0.9166667## specificity 0.6363636## cutoff 0.4313620

Test thresholds to evaluate accuracy

*#confusion matrix...shows us where we accurately and innaccurately predicted results*t1 = **table**(train**$**HighSales,predictions **>** 0.4313620)t1

## ## FALSE TRUE## No 8 3## Yes 4 8

Calculate accuracy

(t1[1,1]**+**t1[2,2])**/nrow**(train) *### >>>>> this is the accuracy calculation*

## [1] 0.6956522

A naive prediction

t1 = **table**(train**$**HighSales,predictions **>** 1) t1

## ## FALSE## No 11## Yes 12

(t1[1])**/nrow**(train) *### >>> naive accuracy*

## [1] 0.4782609

Now that we are done with modeling (selected a “champion” logistic regression model AND a threshold), we can evaluate model performance on the testing set.

Start by developing predictions

predictions\_test = **predict**(forwardmod, newdata = test, type = "response") *#develop predicted probabilities on the training set*

Develop confusion matrix

t1 = **table**(test**$**HighSales,predictions\_test **>** 0.6)t1

## ## FALSE TRUE## No 1 6## Yes 0 8

Calculate accuracy

(t1[1,1]**+**t1[2,2])**/nrow**(test) *###strong dropoffs between trainin and testing data indicates overfitting*

## [1] 0.6

Compare to naive accuracy

t1 = **table**(test**$**HighSales,predictions\_test **>** 1) t1

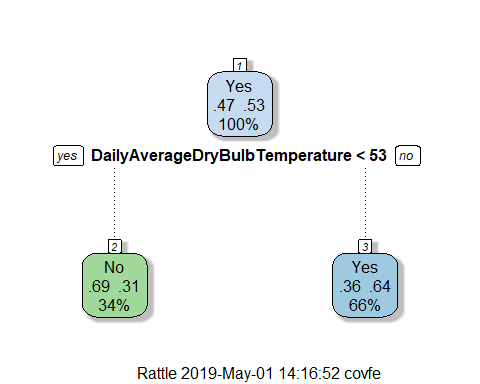
## ## FALSE## No 7## Yes 8

(t1[1])**/nrow**(test)

## [1] 0.4666667

Can attempt to “induce” a tree.

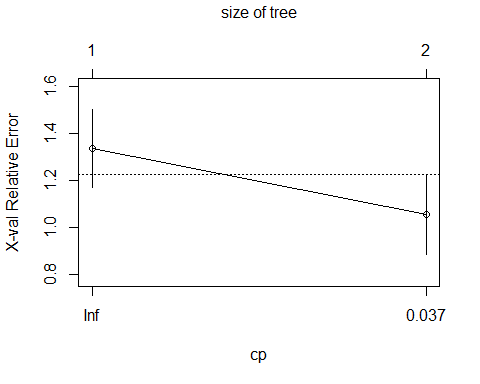
fit2 = **rpart**(HighSales **~** DailyAverageDryBulbTemperature **+** Weekend.Ind **+** Day.of.Week.Num, data = SFD\_Full, method = "class", control=**rpart.control**(cp=.005))**fancyRpartPlot**(fit2) *#plot tree*



**printcp**(fit2) *#see cp table*

## ## Classification tree:## rpart(formula = HighSales ~ DailyAverageDryBulbTemperature + ## Weekend.Ind + Day.of.Week.Num, data = SFD\_Full, method = "class", ## control = rpart.control(cp = 0.005))## ## Variables actually used in tree construction:## [1] DailyAverageDryBulbTemperature## ## Root node error: 18/38 = 0.47368## ## n= 38 ## ## CP nsplit rel error xerror xstd## 1 0.27778 0 1.00000 1.3333 0.16520## 2 0.00500 1 0.72222 1.0556 0.17123

**plotcp**(fit2) *#see cp plot*



*#as cp gets smaller, the tree will grow larger (inverse relationship)*

*This stepwise model shows that the combination of humiditiy, temperature, windspeed, and weekend index lead to the most accurate predictions of salmon sales. The model is strong enough to indicate that elevated temperatures lead to higher sales. The strength of the model is inferred from the relative proximity of the testing accuracy and training accuracy. While the test accuracy was only .6, it still merits significance. The decision tree shows in greater detail, that if the temperature is greater than 53 degrees (NOT less than 53), there is a 66% percent chance that we will sell more than $231 in salmon, and a 34% chance of selling less than that amount. This model is strong, but could be improved greatly if there were more store data to include in the analysis, such as genre of music playing etc. Additoinally, having more than 4 weeks of data to work with would be useful.* **Nevertheless, since this model is dependent on weather, we can still conclude that it is valuable in predicting sales for the time period of February 4th to March 10th.** *The model could be slightly improved if average temperatures were smoothed out by being compared to historical averages, but it would be better to find a way to get more store data overall, and use the corresponding weather data. A years worth of data would be optimal*