Predicting Shelf Life of Blueberries

Group D - Bryce Russell Davis, Jordan Klustner, Keshob Sharma,
Megan O'Reilly, Twinkle Roy

STAT 8000: Statistical Collaboration

Client: Dr. Savithri Nambeesan
University of Georgia

November 8, 2021

Contents

List of Figures

List of Tables

1 Introduction

Blueberries have long been recognized for their health benefits and their contributions to a nutritious diet. Within the state of Georgia, the farm gate value, which indicates the market value of a product minus its selling costs, has surpassed $300 million for blueberries. After harvest and during storage, blueberries often experience negative side effects such as loss of firmness, shriveling, and loss of flavor, all of which can ultimately decrease the shelf-life of the fruit. Due to the loss incurred from the negative effects during storage, understanding the determinants of blueberry shelf-life is important to many.

Our client, Dr. Savithri Nambeesan, is one such person who wishes to explore the possibilities of predicting or prolonging the shelf-life of blueberries. Dr. Nambeesan is an assistant research scientist in the Department of Horticulture at the University of Georgia. Her research focuses primarily on exploring the various elements that determine the postharvest shelf-life of fruits and vegetables. Currently, Dr. Nambeesan is exploring the shelf-life of blueberries harvested within the state of Georgia. Her main objective is to determine whether there are certain predictors present at harvest that can be used to determine the final shelf life of blueberries. The research questions proposed by our client that guided our data analysis and modeling are as follows:

- Are there any significant correlations between physical and chemical traits at harvest?
- Are there any significant correlations between traits (both physical and chemical) at harvest and final shelf life of blueberries?
- Can we find significant predictors of final shelf life and make predictions about it based on the constructed model?

2 Data

2.1 Experimental Design

As part of her research project, Dr. Nambeesan collected data on blueberries from two different farms in Georgia. Alapaha Farm is solely used for research purposes and uses no extra treatments on their blueberries. Manor Farm is a commercial farm that uses pesticides and good management practices in an attempt to prolong their blueberries' shelf-life. Additionally, our client focused on two species of blueberries, Southern Highbush and Rabbiteye, and within each of those species there are several cultivars. The cultivars studied are shown in Table 1.

While many blueberries are machine harvested, the blueberries used in this project were hand-harvested to preserve their natural state and avoid unnecessary bruises or blemishes. The ripe fruits were packed in clamshells and transported to Athens by car, where they were stored at low temperatures ($4°C$) and high humidity (90% relative humidity). At regular

Table 1: Cultivars Included in the Dataset

Southern Highbush	**Rabbiteye**
Emerald	Alapaha
Farthing	Brightwell
Legacy	Krewer
Miss Alice Mae	Powderblue
Miss Jackie	Premier
MissLilly	Titan
Rebel	Vermon
Star	
Suzieblue	

intervals, the blueberries were brought to room temperature and fruit quality traits were evaluated. For each measurement, 4 replications were performed which is shown in the Figure 1.

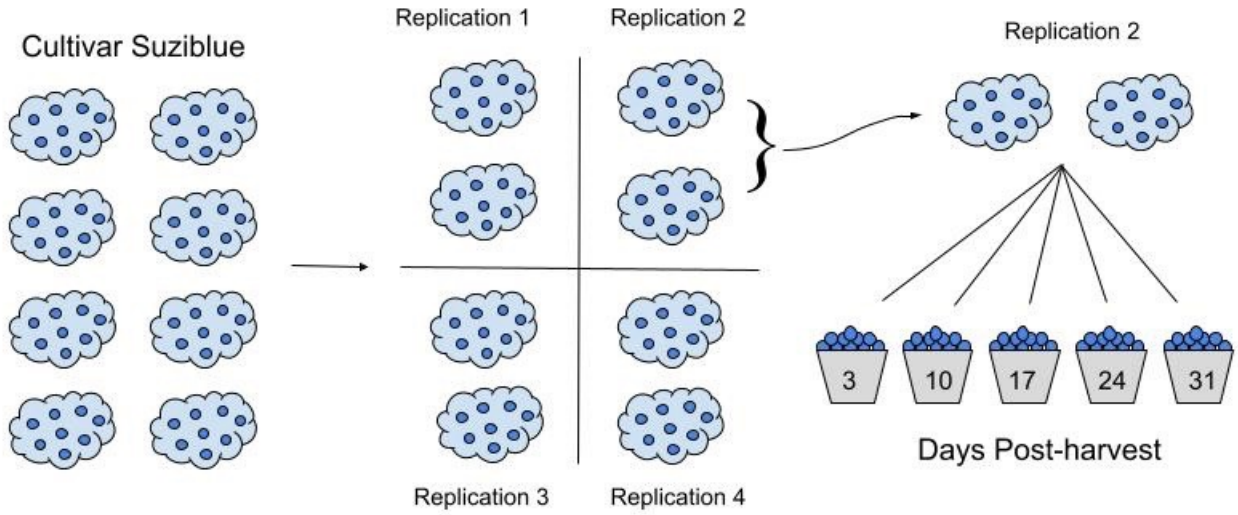

Figure 1: The Harvest and Sorting Process Used to Collect the Measurements of Fruit Traits

2.2 Batch-Time Data

Looking across the replications per blueberry, we can create a batch-time data, where we observe that when picking the blueberries at the fields, the number of plants present at the field per cultivar were counted. For each cultivar, the corresponding plants were then randomly divided into 4 groups to represent the number of replications performed for each measurement. Blueberries from each of these groups were picked and then separated into clamshells. As shown in Figure 1, one clamshell was collected for each "Days Post Harvest". For example, for a given year, there were 8 plants from cultivar Suziblue present in the field. Blueberries from 2 plants were picked to be used in Replication 2, and then these blueberries

were separated into 5 clamshells to represent the number of time points beyond the harvest dates where measurements were to be performed. The same process was used for the remaining 3 replications. For each measurement date post-harvest, blueberries were measured from four clamshells, where the data values obtained from a given clamshell represent one replication.

2.3 Physical/Chemical Measurements

The fruit quality traits measured included both physical and chemical traits (see Table 2). The physical traits measured were Visual Assessment, Compression, and Puncture. Visual Assessment finds whether any fruits had tears, shriveling, molding, dents, or any other physical defect that would classify it as unmarketable. Compression is done with a texture analyzer instrument and involves measuring the force required to compress the fruit by 1 mm. The higher this measurement, the firmer the fruit. Puncture is performed with an instrument that measures how much pressure is required to puncture the skin of the fruit. The higher this measurement, the tougher the fruit skin.

Table 2: Blueberry Trait Component Included in the Dataset

Trait	Chemical/Physical	Replications	Per Replication	Units
Visual Asses.	P	4	30 blueberries	% defect free
Compression	P	4	12 blueberries	kgf
Puncture	P	4	12 blueberries	kgf
Tot. Soluble Solids(TSS)	C	4	30-40 grams	oBrix
Titratable Acidity(TA)	C	4	30-40 grams	% juice
Weight	C	4	20 blueberries	grams
pH	C	4	30-40 grams	n/a

Following the measurements of the physical traits, the chemical traits were measured. The first chemical trait measured was Weight. Fruit Weight helps assess whether blueberries lost water weight during post-harvest storage, which can occur if the fruit is not stored properly. The remaining three chemical traits measured were Total Soluble Solids (TSS), Titratable Acidity (TA), and pH. These chemical traits required the destruction of the blueberries and were therefore measured last. The selected blueberries were pureed and the juice was extracted using a cheesecloth. The resulting blueberry juice was used to measure the remaining traits. The Total Soluble Solids (TSS) trait measures the total soluble solids (oBrix) of fruit juice, which is an indicator of sweetness. Titratable Acidity (TA) is an indicator of sourness, where a higher reading indicates more sour blueberries. The final chemical factor measured was the pH of the juice of the blueberries. The pH factor was important, as our client is highly interested in the Titratable Acidity (TA) trait, and TA is related to pH as they are both measurements of fruit acidity. All of these factors were recorded by our client and provided to us in a data set.

2.4 Explanation of Data-Set Organization

The data set given to us by our client contained valuable information relevant to our client's research. Including all the replications that were performed, there were 697 observations in our data set, with each observation being the complete results for one sample of blueberries. The first few columns provided details about the fruit samples, including Year, Farm, Harvest Date, Species, Cultivar, Postharvest Stage, and Replication. The next few columns provided data on the fruit attributes measured, including Defect Free, Compression, Puncture, TSS, TA, Weight, and pH. Relevant information about these factors is shown in Table 2.

These factors are the explanatory variables. For each blueberry sample at a given postharvest date, four replications were performed and all the traits were recorded for each replication. To simplify our data, we took the averages of the trait values across the four replications. We first found the coefficient of variation (standard deviation / mean) for each of the sets of replications to decide whether the mean or median was more appropriate. As the average ratio of standard deviation / mean was less than 10%, the variations among replications in the samples are small. Thus, we used the mean values in this project to conduct our analysis. In Table 3, presents the summary statistics of the raw dataset. It is evident that this is an **unbalanced design** and there is issue of **confounding variables** with the factor variables (Year, Farm and Species) which can potentially be an issue. There is evidence of missingness in the data-set as well which we will analyze in the subsequent subsection 2.5.

Table 3: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Median	Max
DefectFreePct	696	75.297	21.394	10.000	80.000	100.000
Compression	695	0.232	0.048	0.096	0.232	0.368
Pucture	696	0.158	0.036	0.072	0.154	0.291
TSS	695	12.761	1.865	6.900	12.800	17.000
pH	680	3.816	0.300	2.900	3.800	4.957
TA	684	0.326	0.153	0.094	0.280	1.010
Weight	696	2.049	0.500	0.915	2.016	3.431

2.5 Missingness

The original data set contained missing values for the pH and TA factors, which caused issues when averaging those values over the replications. In total, there were 37 missing values in our data set. There was 1 missing value for Defect Free %, 1 missing value for Puncture, 1 missing value for Weight, 2 missing values for Compression, 2 missing values for TSS, 13 missing values for TA, and 17 missing values for pH. Most of the missing values were from the Year-2015. During that year, our client used different instruments to measure pH and TA than in the other years, and since she was not able to guarantee the reliability and accuracy of those measurements, she removed them from the data set. Figure 5 shows the distribution of missing values in our data set across the various factors. Since 97.27% of our data rows were complete and had no missing data, we decided on using data interpolation and extrapolation to predict the gaps if necessary and predict the Shelf-Life of Blueberries.

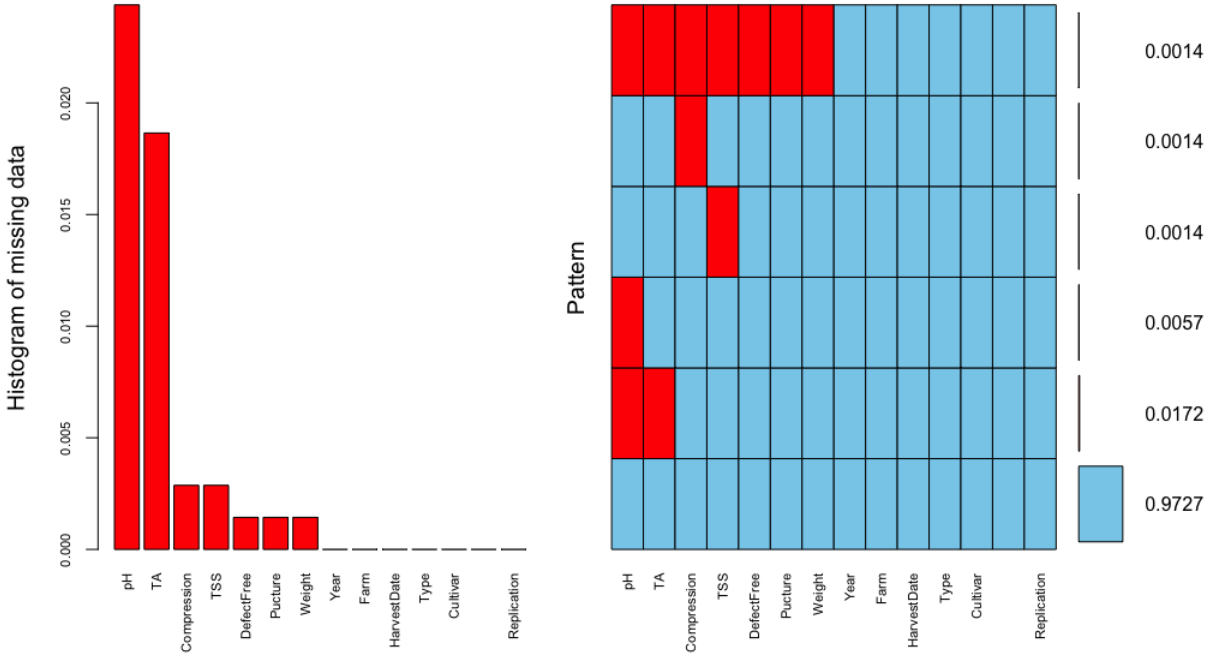

Figure 2: Analysis of Missing Values in Raw Dataset

3 Exploratory Data Analysis

In our further data analysis, we decided to use the blueberries dataset with measurements recorded directly post-harvest days. Most of the observations were collected either 3 or 4 days after harvest and a few were collected 2 or 5 days after harvest. Based on our client's questions, we first looked at whether any traits directly post-harvest correlate with one another. Then, to further explore whether there are any useful predictors that could be used to predict the Final Shelf Life, for those quantitative traits, we explored if there were any significant correlations between traits at harvest. Since our data could also be further classified by Year, Farm, Species and Cultivar, we explored if these qualitative traits could be helpful in determining the Final Shelf Life of blueberries. Based on the analysis, we find the four strongest correlations were found between TSS and Puncture, Puncture and Compression, pH and Titratable Acidity, and TSS and pH as seen on Figure 3 and conclude that there could be a potential **problem of multi-collinearity** if we were to include all the chemical/physical traits into our final model, hence some kind of transformation and variable selection would be required.

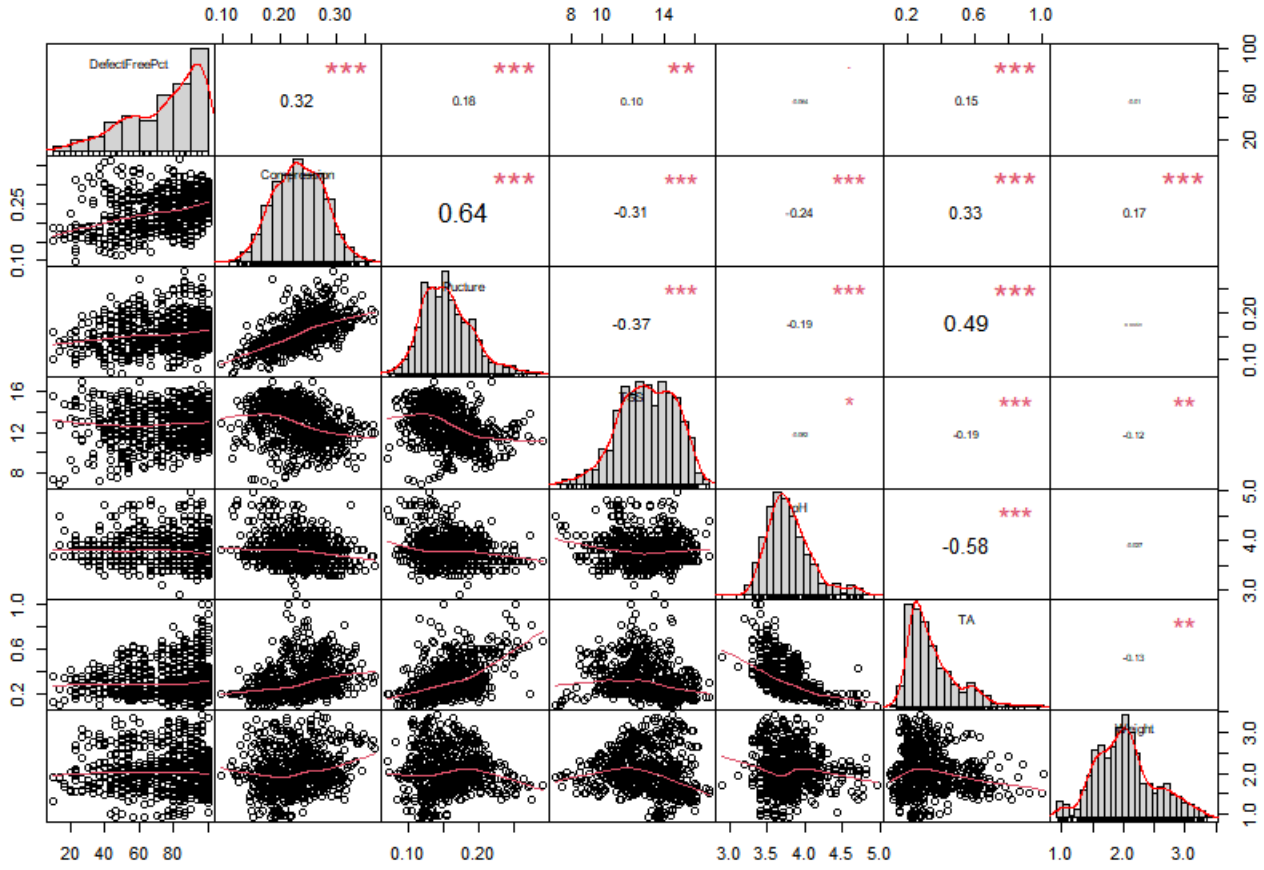

Figure 3: Correlation Matrix in Raw Dataset

Based on the scatterplot in Figure 4, we can see a moderate positive linear association between Compression and Puncture. Since these are both measurements of blueberry tough-

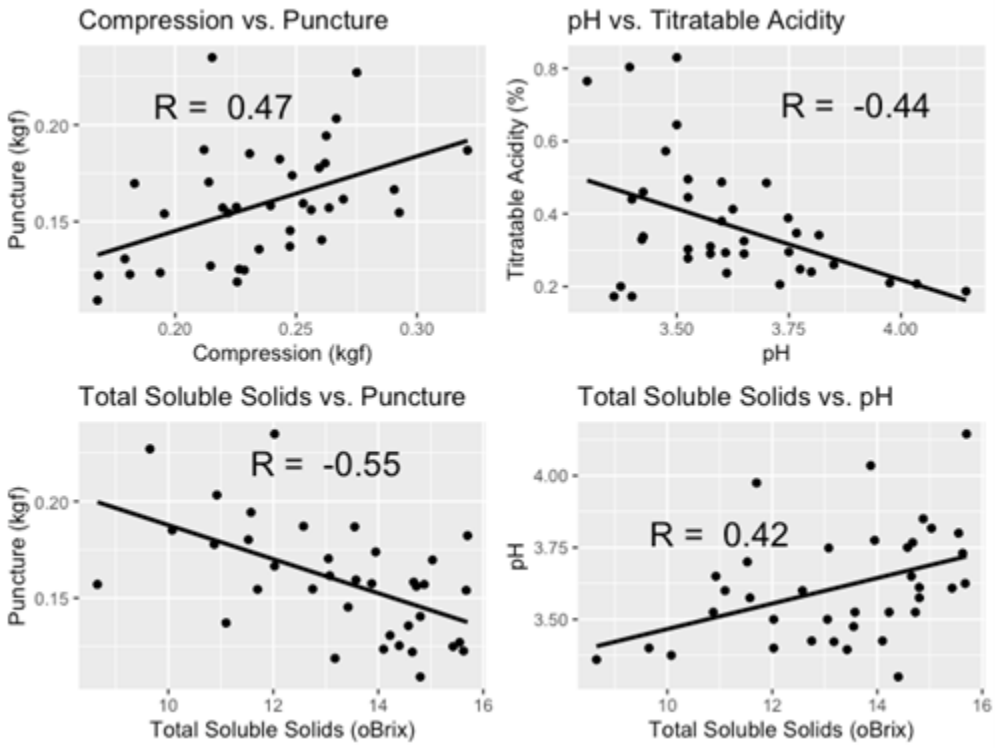

Figure 4: Analysis of Correlation in Raw Dataset

ness, it makes sense that there is a correlation between these two factors. Mushy blueberries that require less force to compress also likely have skin that requires less force to puncture. In the top right scatterplot, there is a moderate negative linear association between pH and TA. Since these two factors are both measures of acidity, the strong correlation makes sense. In the bottom left scatterplot, there is a moderate negative linear association between TSS and Puncture. In the bottom right scatterplot, there is a moderately strong negative linear association between TSS and pH. Identifying these significant correlation values will help us during the later stages of our project when we attempt to create a model to predict Final Shelf Life based on the initial fruit traits at harvest.

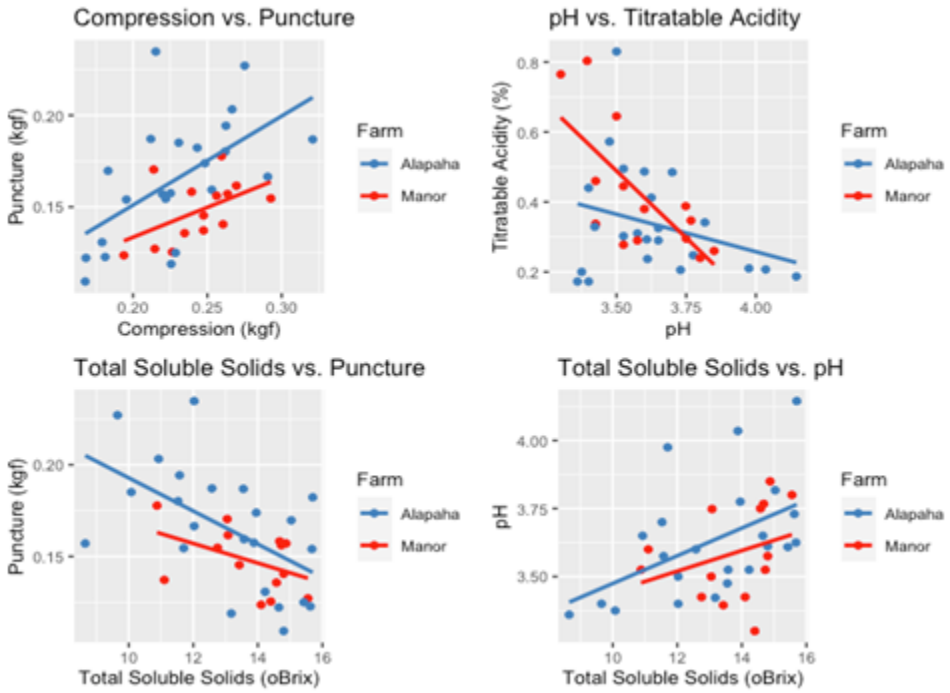

Figure 5: Analysis of Correlation in Raw Dataset Farm Wise

In addition, we looked into the same correlations above, but separated them by farm, as shown in Figure 5. These differences are likely because of the extra treatments that are used at Manor Farm and we would expect the batches of blueberries to have more shelf-life for that farm.

4 Model Set-Up

4.1 Creating Batch Data Set

The data set given to us has information on 37 batches of blueberries as mentioned on preceding sections. Each batch was measured at different time periods. When observing the measurements for each batch at each of these time periods, we observe 175 measurements. Additionally, for each measurement there were 4 replications of measurements leading to close to 700 (4x175) total observations in the data set presented to us.

We begin formatting the data by taking the average of all replicates for each of the variables (i.e. defect free percentage, pH, weight, etc.). When we do this we obtain a data set with 175 observations that contain information about the measurements of each batch at different batch times, or each time a given batch was measured.

Table 4: Table of Data Set with Averaged Replicates

	Year	Farm	Type	Cultivar	PostHarvestDays	Avg.ND Pct	avg.Comp
1	2015.00	Alapaha	Rabbiteye	Powderblue	3.00	47.50	0.20
2	2015.00	Alapaha	Rabbiteye	Powderblue	8.00	39.15	0.18
3	2015.00	Alapaha	Rabbiteye	Powderblue	13.00	30.85	0.19
4	2015.00	Alapaha	Rabbiteye	Powderblue	21.00	32.48	0.16
5	2015.00	Alapaha	Rabbiteye	Powderblue	28.00	20.00	0.17
6	2015.00	Alapaha	Rabbiteye	Premier	3.00	74.17	0.18
7	2015.00	Alapaha	Rabbiteye	Premier	8.00	72.50	0.18
8	2015.00	Alapaha	Rabbiteye	Premier	13.00	40.00	0.19
9	2015.00	Alapaha	Rabbiteye	Premier	21.00	38.33	0.18
10	2015.00	Alapaha	Rabbiteye	Premier	28.00	22.50	0.17

Table 3 shows the first ten observations in this new data set with a subset of variables. The first 5 observations are associated with the first of 37 batches and provide information about the measurements at each of the measurement times indicated by "Post Harvest Days". Also notice how the variable names for the traits have an "avg" prefix. This indicates an average of the four replicate measurements for each of the blueberry trait variables. For example, the first observation has an average defect free percentage of 47.5. This value is the average of the four replicate measurements of defect free percentage taken 3 days after harvest.

4.2 Interpolation and Extrapolation in Batch Data-Set

Next we will create a data set with information solely on the 37 batches of blueberries for use in our regression analysis. We have two main objectives for formatting this final data set. First, we will find a common initial measurement for"Defect Free Percentage". For the 37 batches, the most common initial measurement was taken 3 days after harvest. Many of the batches, however, have an initial measurement taken anywhere from 2 to 5 days after

harvest. Consequently, we want to find an initial measurement of "Defect Free Percentage" that is consistent for all 37 batches. In order to do this, we perform linear interpolation to find an approximate value of "Defect Free Percentage" if the initial measurement was taken on day 3.

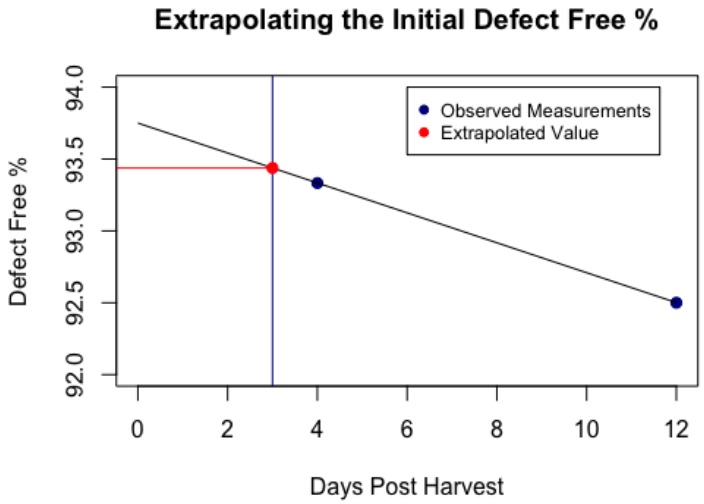

Figure 6: Example of Linear Extrapolation

Linear interpolation involves using two observations to construct a new observation that lies between the two observed points. In some cases extrapolation is necessary. Like linear interpolation, extrapolation uses two observations to construct a new observation outside the range of the two observations. We must be extra cautious using extrapolation since we are predicting a value outside of our range of values. As it pertains to our analysis, we will use the first two measurement times of each blueberry batch to approximate the defect free percentage at 3 days post harvest. Extrapolation in this situation would be appropriate since the values we are estimating are close to our observed range. Figure 2 helps visualize this process. In this example we use the defect free percentage of a batch of blueberries (from a 2017 Emerald Cultivar) at the first two measurements. We see that on day 4 after harvest (initial measurement) the defect free percentage was approximately 93.33. Similarly, the day 12 measurement (second measurement) yielded a defect free percentage of 92.5. We fit a line between these two data points and find what the approximate defect free percentage will be at 3 days post harvest. We perform this calculation for all initial measurements that are not taken 3 days post harvest. As a result, all initial measurements have a "Days Post Harvest" value of 3 and an interpolated value of "Defect Free Percentage" associated with 3 days post harvest.

After the interpolation and extrapolation process described above, we notice that some values of "Defect Free Percentage" are predicted by the extrapolation to be 100. Since this is representative of a percentage, we replace these values with 100 and proceed. Also note that the example above is an instance when we used extrapolation. Interpolation would occur when the desired point is in between the two observed points (i.e. when the initial

measurement was taken on day 2)

Our next step before beginning our analysis is to calculate shelf life for each of the 37 batches. We will calculate these shelf lives in a similar way to the defect free percentage. We want to find the number of days after harvest it takes until exactly 75 percent of the blueberries in the batch are defect free. It is at this point that we consider the batch to have gone bad. It is also important to note that some batches have an initial defect free percentage of less than 75. In these cases we will use the first two measurements to retroactively extrapolate the value at which the defect free percentage is 75. We will refer to the aforementioned process as "left extrapolation". Additionally, some batches never reach the 75 percent threshold within the completion of all measurements. As a result, we will extrapolate these values using the last two measurements and refer to this process as "right extrapolation". The rest of the batches we will interpolate the shelf life with the observations taken directly before and after the 75 percent threshold was reached.

Method	Frequency
Interpolation	25
Left Extrapolation	5
Right Extrapolation	7
Total	37

The table above shows the three methods of finding shelf life and the number of batches for which the respective method was used.

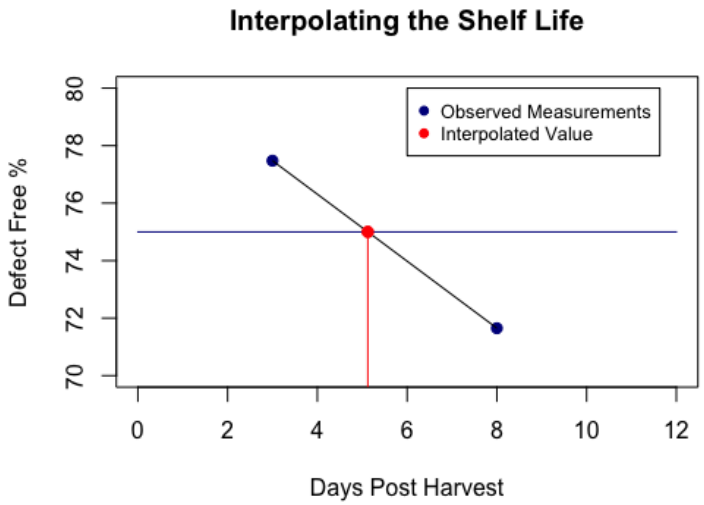

Figure 7: Interpolating Shelf Life Example

Figure 3 shows an example of the process of interpolation. The blue points represent the observed measurements of defect free. The red point is the estimated days post harvest when the defect free percentage reaches 75. The extrapolation process is similar to the example presented in figure 2. We use the two closest observations to estimate the post harvest days

when the defect free percentage is 75. In two cases the right extrapolations yielded shelf lives well outside our observed range. These values were 125 and 84.8. To adjust for this we need to choose a value that is greater than the next highest shelf life but isn't unreasonable in terms of the shelf lives we observed. We determined that this value should be 70 since the next highest shelf life was 67.625.

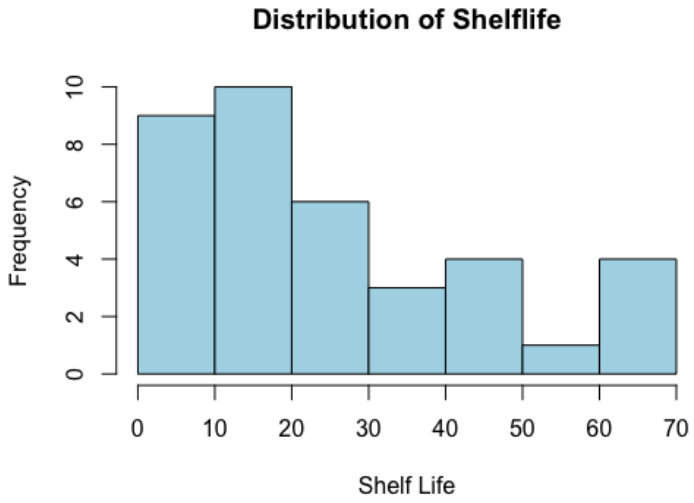

Figure 8: Distribution of Shelf Life

Table 5: Summary of Shelf Life

Mean	Std. Dev.	Min	Q1	Median	Q3	Max
25.62	21.13	0	12.67	18.59	38.53	70

Our final step is to create a data set of the 37 batches with all physical/chemical measurements and each batch's respective shelf life. The first 6 observations with a subset of all variables is shown in the table below.

Table 6: Data Set of 37 Batches

	Year	Farm	Type	Cultivar	avg.NDpct	avg.Comp	SL
1	2015.00	Alapaha	Rabbiteye	Powderblue	47.50	0.20	0.00
2	2015.00	Alapaha	Rabbiteye	Premier	74.17	0.18	0.54
3	2015.00	Alapaha	Rabbiteye	Titan	56.67	0.32	0.00
4	2015.00	Alapaha	Rabbiteye	Vernon	75.83	0.25	3.38
5	2015.00	Alapaha	Southern highbush	Legacy	77.47	0.21	5.12
6	2015.00	Alapaha	Southern highbush	Rebel	63.85	0.22	0.00

4.3 General Linear Models for Predicting Shelf-Life

After we have our data formatted in the preceding section, we fit a multiple linear regression model to predict the shelf life of blueberries. A multiple linear regression model attempts to model the relationship of one response variable and several continuous or categorical explanatory variables. If any categorical predictors are used, such as Farm, Species, Cultivar, or Year, then we will use a General Linear Model instead. In our analysis, we will use a subset of the variables mentioned to predict shelf life. The line of best fit is determined by minimizing the distance from the line to the observations and has the following form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon$$

Here, y is the response variable we are trying to predict, the shelf life of a batch. The $x_i's$ are the predicting variables. The $\beta_i's$ are the coefficients of each predictor variable and ϵ is the error in prediction.

As it pertains to the analysis, we will begin by fitting a simple linear regression, which is a model with a response and only one predictor. We will do this for every physical/chemical variable in the batch data set. We will observe how well each of these variables perform at predicting shelf life. Next we will observe which categorical variables are good predictors of shelf life by fitting ANOVA models. Next, we will use an subset of all the variables to fit a general linear model. Finally we will select a final subset of variables that collectively predict shelf life well. Finally, we will diagnose our final model to make sure it is a good fit and make adjustments as needed.

5 Analysis/Results

5.1 Physical/Chemical Predictors

5.1.1 Simple Linear Regressions

After performing the interpolation and extrapolation, it is pertinent to check individual causal relationship between the response variable - **SL (Shelf Life)** and each of the 13 other predictors we have in the model. Firstly we observe if there is any significant correlation between the predictors and the response variable as depicted in Figure 9. We notice a very high significant positive correlation between Shelf Life and Avg Defect Free % which means that if we have more defect free blueberries in the dataset, we will have a longer prolonged Shelf Life which we would expect to vary across the Farm and Species of blueberry we look at. Additionally, although other chemical and physical traits exhibit correlations to a marginally lesser degree we do not observe any significant relationship with the Shelf-Life. However, this cannot potentially rule out individual causal relationship which these other predictors can have with the Shelf Life and hence we need to conduct Simple Linear Regression with the response variable in a step-wise manner and we also check for STEP AIC criterion for our variable selection of the final model.

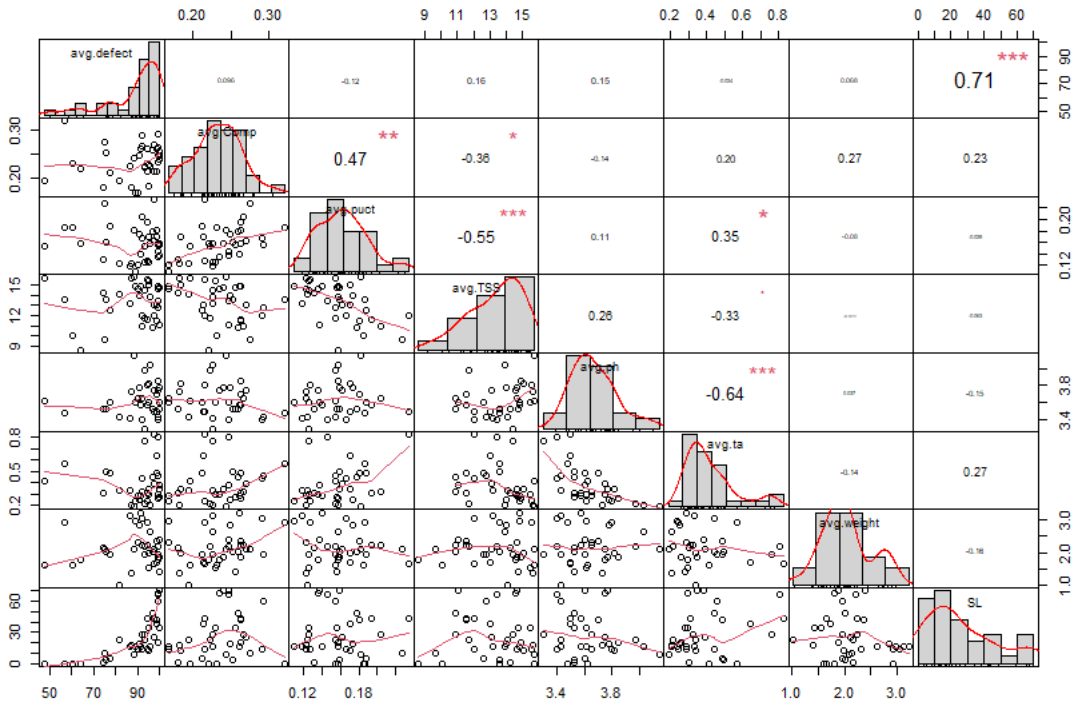

Figure 9: Correlation Matrix of Shelf Life and Other Potential Predictors

We fit a simple linear regression with every blueberry attribute for each batches on its own and find that none of these are good predictors of shelf life on their own. Next we fit

the categorical predictors to the model individually to see how they perform. We find that the Farm, Species, and Year were good at explaining the variation of Shelf Life but we only present the results for the ones which have the highest R^2 i.e gives us the most variation. The regression model is as mentioned below:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i$$

From Table 7, Simple Linear Regression results are presented across 37 batches of blueberry

Table 7: Individual Predictor Results

	Dependent variable:				
	SL				
	(1)	(2)	(3)	(4)	(5)
Farm-Manor	25.668***				
	(5.827)				
Type-Southern highbush		18.404***			
		(6.467)			
Average NDPct			1.121***		
			(0.190)		
Year-2016				25.833***	
				(8.649)	
Year-2017				35.874***	
				(8.038)	
Year-2018				25.908***	
				(8.205)	
Avg.Weight					-6.494
					(6.723)
Constant	15.905***	14.674***	-73.307***	2.365	39.532**
	(3.584)	(4.987)	(16.974)	(6.116)	(14.818)
Observations	37	37	37	37	37
R^2	0.357	0.188	0.498	0.386	0.026
Residual Std. Error	17.190 (df = 35)	19.313 (df = 35)	15.185 (df = 35)	17.298 (df = 33)	21.151 (df = 35)
F Statistic	19.404*** (df = 1; 35)	8.099*** (df = 1; 35)	34.717*** (df = 1; 35)	6.908*** (df = 3; 33)	0.933 (df = 1; 35)

at their initial Shelf-Life.We can clearly confirm a significant causal relationship between Farm, Species and Year with the response variable Shelf-Life. We check the p-value of the coefficients and it indicates whether or not you can reject or accept a hypothesis. The hypothesis, in this case, is that the predictor is not meaningful for our model. The p-value for the predictors mentioned in Table 7, are all less that 0.05 which means that Species, Farm, Year and Average NPDefect % are an excellent addition to the model. The p-value for the average weight is 0.85. In other words, there's 85% chance that this predictor is not meaningful for the regression as per Table 7. Also, we also check the STEP AIC criterion for selecting a subset of predictor variables from a larger set (e.g., stepwise selection). We perform stepwise selection (both) using the stepAIC( ) function from the MASS package in R software. We conclude stepAIC( ) performs stepwise model selection by exact AIC and we find the Specie, Farm and Avg Defect Free Pct to be most relevant for our model (we have not presented those results in the report as we conclude the same from Table 7 as well).

5.1.2 Multiple Regression

The dataset is an *imbalanced dataset* since the distribution of the response variable (Shelf Life) is heavily skewed and hence we conclude that a transformation would be required to conduct further analysis. A simple multivariate linear regression model was first constructed using the response variable 'Final Shelf Life in Days' and only the main effects (Defect Free, Farm and Type) as the potential predictors for our model listed below in the regression model respectively where Average NPDefect Free % is a trait and Farm, and Species are factor variables which have a significant impact on the shelf life. The regression results for Model 1 is as:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i$$

Table 8: Non Transformed Multiple Linear Model Results

	Dependent variable:
	SL
Farm-Manor	13.835**
	(5.279)
Type-Southern highbush	9.757**
	(4.773)
Average NDPct %	0.754***
	(0.201)
Constant	-52.342***
	(16.544)
Observations	37
R^2	0.623
Residual Std. Error	13.566 (df = 33)
F Statistic	18.212*** (df = 3; 33)

Table 8 indicates that all predictors, % Defect Free , Type of Farm and Species of Blueberry are significant factors as their corresponding p-values were less than 0.05. We could interpret their coefficient estimates as the following: as % Defect Free increases by one unit, Final Shelf Life of the blueberries will increase by nearly 0.754 days. Similarly, as Farm type is Manor, Final Shelf Life of the blueberries will increase by nearly 13.8 days which is expected because Manor Farm belongs to the treatment group. Multiple R-squared is a statistical measure of how close the data are to the fitted regression line, so it is typically used to evaluate the goodness of fit of the model. According to Table 8, our explains 62.3% of the variability of Final Shelf Life.

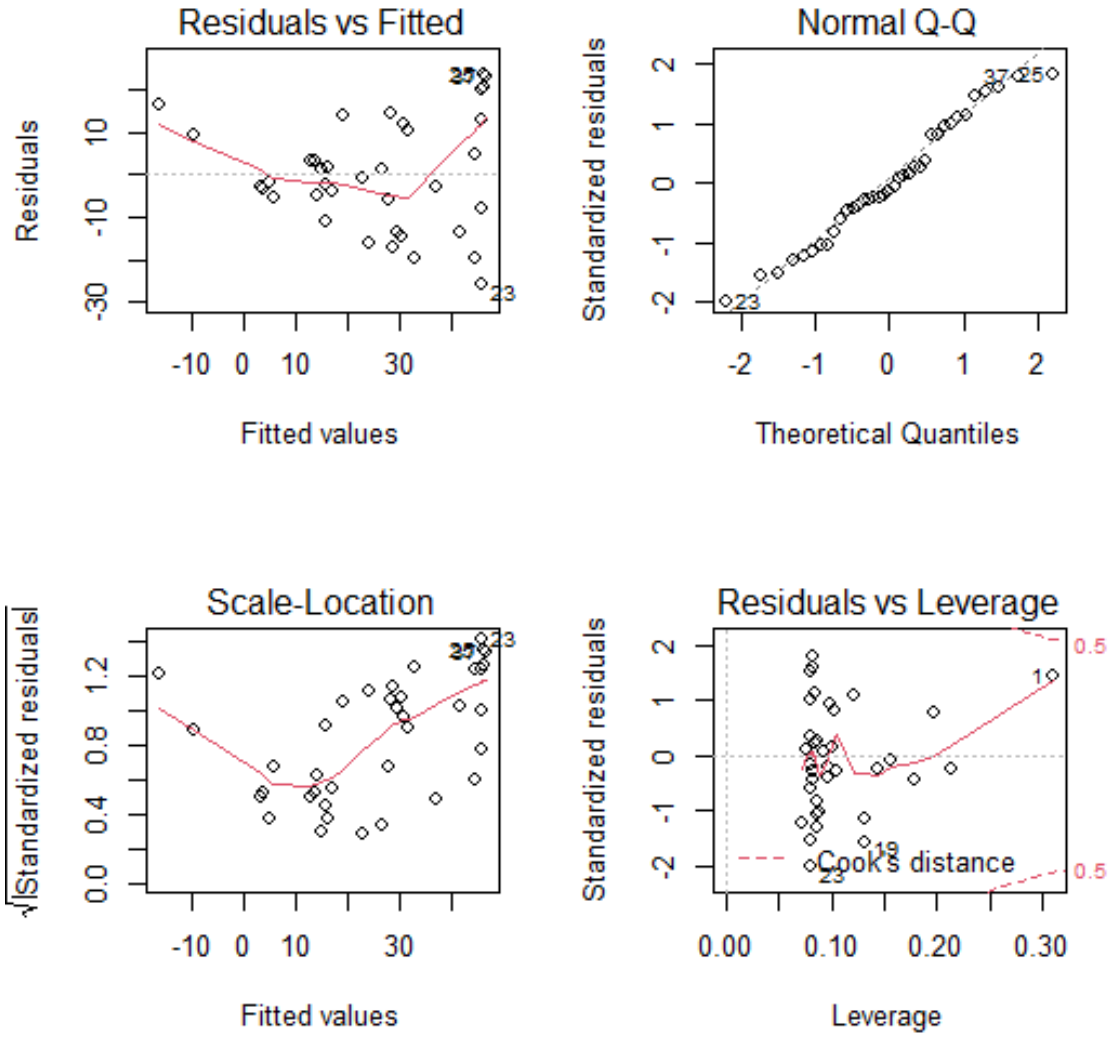

Figure 10: Residual Plot and QQ Plot of Model 1

Figure 9 shows the residual vs. fitted values plot, with an obvious U-shape pattern which indicates that we need to have a transformation of the response variable to the model to account for that curvature. The QQ plot shows a long tail to the right, which indicates that the normal assumption may be violated in this case.

5.2 Class Variable Predictors

5.2.1 One-WAY ANOVA Models

We could further classify the dataset based on Farm and Blueberry Species. Intuitively, these qualitative variables could also have potential effects on % Defect Free and therefore affect Final Shelf Life. We are interested in investigating whether % Defect Free is significantly different between different levels of Farm and Blueberry Species. We do recognize the presence of Confounders in the dataset.

Table 9: Anova Table

	Effect	df	MSE	F	ges	p.value
1	Species	1, 33	237.94	9.81 **	.229	.004
2	Farm	1, 33	237.94	12.71 **	.278	.001
3	Species:Farm	1, 33	237.94	3.29 +	.091	.079

Observed from Table 9, the p-values for treatment factors 'Species', 'Farm' are all significant except for interaction variable between Farm and Species significant at 10%. We conclude that there is significant variation between the levels of the four treatment factors, and hence all the treatment factors are effective. Thus, the four non-trait variables were all significant and should be considered to include in the final model.

5.2.2 Multiple Predictor ANOVA

For the given model, in order to conduct a more comprehensive analysis, we also look into multi-factor ANOVA to determine if numeric (Average Defect Free %) or categorical predictors (Farm and Type) explain variation in a Shelf Life outcome. A multi-factor ANOVA is similar to a one-way ANOVA, where a F-statistic is calculated to measure the amount of variation accounted for by each predictor relative to the left-over error variance. Based on the results of Table 10, we conclude that while controlling for Farm Type, Shelf Life does not significantly differ across groups (F=5.24, df=(1239), p=0.01). While controlling for Blueberry Species, Shelf Life does not significantly differ between the kind of blueberry grown in the farm (F=6.85, df=(751), p=0.05). Also, there is no interaction between Blueberry Species and Farm at 10% significance level. Similar conclusion can be drawn for the Average Defect Free % on Shelf Life.

Table 10: Multiple Predictor ANOVA

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	948.95	1	5.24	0.0286
Farm	1239.58	1	6.85	0.0133
Species	751.80	1	4.15	0.0497
Average NDPct %	2659.04	1	14.68	0.0005
Residuals	5976.03	33		

5.3 Final General Linear Model

Based on the analysis in the preceding sections, we conclude that for the final model, the data set contains the initial values of the fruit traits at harvest, thus it is a relatively small data set with sample size of only 37 after interpolation and extrapolation. We would also like to mention that we chose not to complicate our analysis as we did not want interpretibility to be lost. Keeping in mind about the client's statistical background, we fit a model with **Farm, Species and average defect-free %** as our predictors which we choose based on our analysis from the simple linear regressions in Section 5.1.2. and find that all three predictors remain significant. We also conduct a VIF (multi-collineaity check) check on the predictors and we find there is no multi-collinearity present in the data set. Additionally, the model yields an R^2 of .8 for Final Model we think fit, with a square-root transformation. We try to add Cultivar as a predictor the Adjusted R^2 decreases to .44 most likely due to the addition of every factor of cultivar and hence we decide to exclude Cultivar Type completely from the model and not present those results as that was insignificant to our conclusion. Based on the analysis conducted from Figure 9, we do conclude presence of non-linearity in the model and the residuals and fitted plot on the top left panel of Figure 9 depict a fanning out pattern. We also acknowledge that a square-root transformation could be difficult to interpret and there could be potential problems with explainability, we still further decide to transform our response variable Shelf Life to fit a better model. The model is as below:

$$\sqrt{\text{ShelfLife}} = \beta_0 + \beta_1 \text{Factor(Farm)} + \beta_2 \text{Factor(Species)} + \beta_3 \text{Average NDPct} + \epsilon \qquad (1)$$

Table 11: Square-rootTransformed Final Linear Model Results

	Dependent variable:
	SLˆ1/2
Farm-Manor	6.850**
	(2.618)
Type-Southern highbush	4.823**
	(2.367)
Average NDPct	0.382***
	(0.100)
Constant	-26.362***
	(8.206)
Observations	37
R^2	0.829
Residual Std. Error	6.729 (df = 33)
F Statistic	18.590*** (df = 3; 33)

Table 11 indicates that all predictors, % Defect Free , Type of Farm and Species of Blueberry are significant factors as their corresponding p-values were less than 0.05. We could interpret their coefficient estimates as the following: as % Defect Free increases by 1 unit, Final Shelf Life of the blueberries will increase by nearly $(0.382) \times 2 \times 1 \times 1 = 0.764$ times the square of the current defect free rate (1 unit here we consider- not in % for easier interpretibility). For instance, if the current defect free rate is 10, then a unit increase in shelf life is associated with an increase $(0.764 \times 10 \times 10)$ days in Shelf Life and so on. Similarly, as Farm type is Manor, Final Shelf Life of the blueberries will increase by nearly 13.7 times the square of the Farm Type in days which is expected because Manor Farm belongs to the treatment group The key takeaway would . Multiple R-squared is a statistical measure of how close the data are to the fitted regression line, so it is typically used to evaluate the goodness of fit of the model. According to Table 8, our explains 82.9% of the variability of Final Shelf Life.

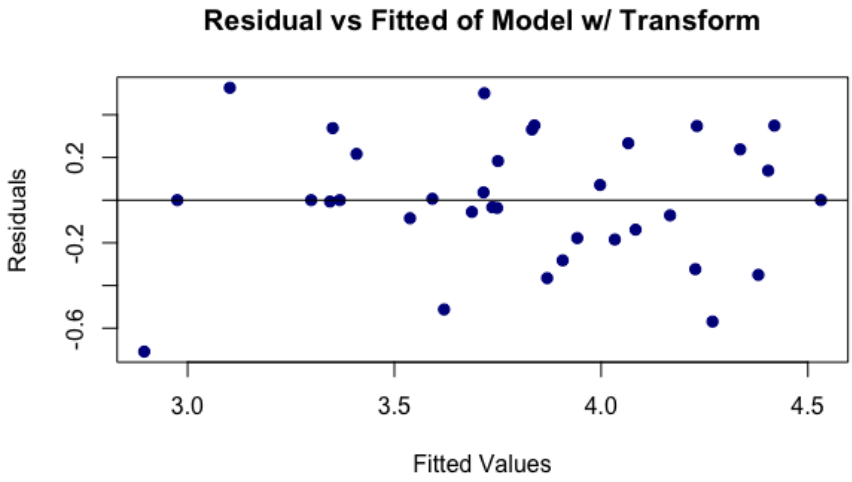

Figure 11: Residual vs Fitted Plot of Transformed Model

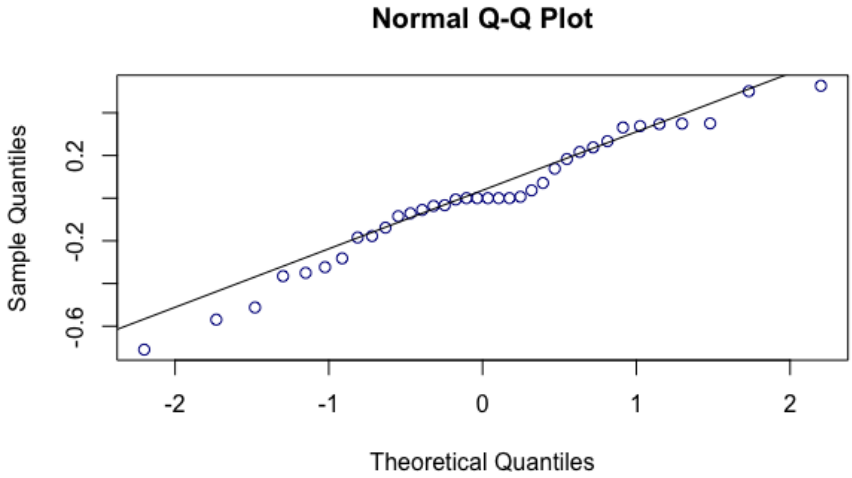

Figure 12: Residual vs Fitted of Transformed Model

There doesn't seem to be any significant non-linear trends however there is a fanning out pattern which indicates non-homogeneity in the variance as per Figure 11. In other words, the variance changes as the fitted values increase. As a result, a square-root transform of the response variable was indeed beneficial. However this would not have been possible if our response had negative values. We discovered that the shelf life interpolation we performed resulted in negative numbers due to batches with defect rates already under 75 percent during initial measurement which we had to truncate to 0 through interpolation and we also had experienced extreme values of 125 days and 92 days for some batch times of blueberries which was unreasonable beyond the scope of our analysis. We also see a massive improvement in our R^2 after the transformation in our model. Additionally, 80% percent of the variation of Shelf Life can be explained by our model. The residuals appear to have no obvious pattern. The QQ Plot also indicates that the residuals are approximately normal. We choose this as our final linear regression model. To further conclude our analysis, we also include

added variable plots which are individual plots that display the relationship between a response variable (SL) and each of predictor variable in a multiple linear regression model, while controlling for the presence of other predictor variables in the model as depicted in Figure 12. We can notice the slope of the regression line is identical with the slope of the focal predictor xi in the full model in Figure 12 for each of the predictors. In contrast to the partial-residual plot, the residuals of the regression line in the added-variable plot are identical with the residuals of the full model. Because the values on the x-axis show values of the focal predictor xi conditional on the other predictors, points far to the left or right are cases for which the value of xi is unusual given the values of the other predictors. Hence, influential data values can be easily seen and helps to detect nonlinearity, heteroscedasticity and unusual patterns which we do not see so much in Figure 12 due to the transformation of the response variable.

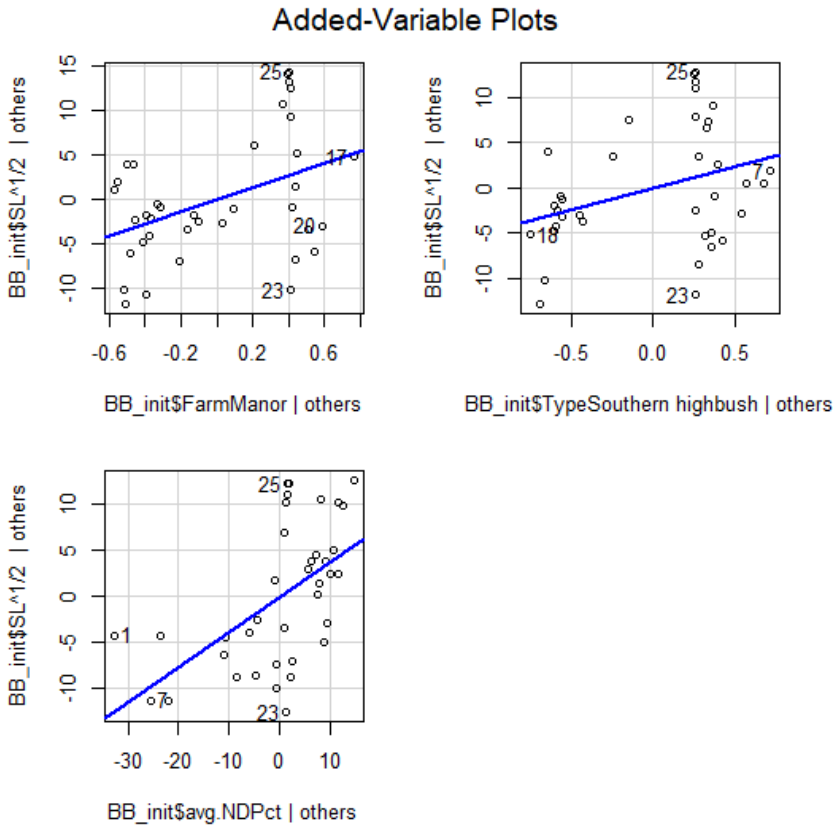

Figure 13: Added Variable Plot For Final Model

6 Conclusion

After some exploratory analysis, we start our main analysis by creating a batch-data set. We calculate the initial measurement of the defect free percentage for 3 days after harvest which is most common measurement day. To do this, we use linear interpolation and extrapolation to create a measure of effective defect free percent that will be on 3 days of harvest. We also calculate the shelf life for all the 37 batches that is consistent with the client's request of 75% threshold. Using the consistent data set that we create, we fit several models to investigate the physical chemical predictors of shelf life. The individual predictor models suggests that Type, Farm, Year,and Average Defect% are good predictors of the shelf life.

Our multiple regression model indicates that %defect free, type of farm, and type of blueberry are significant in predicting shelf life. The model explains 62% of the variation in final shelf life as shown by the R^2 value of 62.3%. Our Q-Q model and the residual plot suggest transformation of the outcome variables. We use, a square root transformation of shelf life in our final model with predictors chosen from our simple and multiple linear models. Our final model performs pretty-well explaining 82.9% of the variation in the response variable. We suggest that the farm in which blueberry was grown, type of blueberry, and average defect free percent are strong predictors of blue-berry shelf life. This implies that special care needs to be taken to get defect free blueberry while harvesting, so that the shelf life of blueberry increases.

We acknowledge that there may be confounding variables that can potentially be an issue. If confounding variable are not addressed than our estimates can be biased. The estimates using our models may not reflect the true relationship among the predictor and outcome variables in the presence of confounding variables. We have tried to avoid the confounding variation by estimating models with cultivar and farm types that could be driving variation in the shelf life.

7 Appendix/R-Code

7.1 R-Code

```
library(readxl)
Blueberries = read_excel("Downloads/Blueberry postharvest_2015-2018.xlsx")


View(Blueberries)

# New Data Set with replicate values averaged

BB = Blueberries %>% group_by(Year, Farm, Type, Cultivar, PostHarvestDays) %>%
  summarize(avg.defect = mean(DefectFreePct), avg.Comp = mean(Compression),
            avg.puct = mean(Pucture), avg.TSS = mean(TSS),
            avg.ph = mean(pH), avg.ta = mean(TA), avg.weight = mean(Weight))


BB %>% drop_na()


# export data frame

head((data.frame(BB))[,1:7], 10)

print(xtable(head(data.frame(BB)[,1:7], 10), type = "latex"), file = "/Users/jordanklust

write.csv(BB,"/Users/jordanklustner/Desktop/STAT8000/BB.csv", row.names = FALSE)

# Create new variable to determine the initial measurement

BB$IM = rep(0, times = length(BB$Year))

for(i in 1:length(BB$Year)){
  if(BB$PostHarvestDays[i] <= 5){
    BB$IM[i] = 1
  }
}

# If the first measurement is not 3 days, we interpolate DefectFreePct for day 3
```

```
for(i in 1:length(BB$Year)){
  x = BB$PostHarvestDays[BB$Year == BB$Year[i] & BB$Cultivar == BB$Cultivar[i]][1:2]
  y = BB$avg.defect[BB$Year == BB$Year[i] & BB$Cultivar == BB$Cultivar[i]][1:2]
  if(BB$IM[i] == 1 & BB$PostHarvestDays[i] != 3){

    BB$PostHarvestDays[i] = 3
    BB$avg.defect[i] = predict(lm(y~x), data.frame(x = 3))
    if(BB$avg.defect[i] > 100){
      BB$avg.defect[i] = 100
    }
  }
}

j = 0
batch = rep(0, times = length(BB$Year))

for(i in 1:length(BB$Year)){
  if(BB$IM[i] == 1){
    j = j+1
  }
  batch[i] = j
}


BB$batch = batch
BB = BB[-109,]


Shelflife = rep(100, times = length(BB$Year))

for(i in 1:length(BB$Year)){

  if(BB$avg.defect[i] >= 75 & BB$avg.defect[i+1] <= 75 & BB$batch[i] == BB$batch[i+1] &
    y = BB$PostHarvestDays[i:(i+1)]
    x = BB$avg.defect[i:(i+1)]
    Shelflife[i] = predict(lm(y~x), data.frame(x = 75))
  }
  if(BB$avg.defect[i] > 75 & (BB$batch[i] != BB$batch[i+1]) & (i != length(BB$Year))){
    y = BB$PostHarvestDays[(i-1):(i)]
    x = BB$avg.defect[(i-1):(i)]
    Shelflife[i] = predict(lm(y~x), data.frame(x = 75))
  }
  if(BB$avg.defect[i] < 75 & BB$PostHarvestDays[i] == 3 & (i != length(BB$Year))){
    y = BB$PostHarvestDays[(i):(i + 1)]
    x = BB$avg.defect[(i):(i + 1)]
```

```
    Shelflife[i] = predict(lm(y~x), data.frame(x = 75))
  }
}

Shelflife = Shelflife[Shelflife != 100]
Shelflife

x = BB$avg.defect[173:174]
y = BB$PostHarvestDays[173:174]
Shelflife[37] = predict(lm(y~x), data.frame(x = 75))
```

```
# Create new data set with 37 batches

BB_init = BB[(BB$PostHarvestDays == 3),]
Shelflife = ifelse(Shelflife < 0, 0, Shelflife)
BB_init$SL = Shelflife
BB_init$SL[25] = 70
BB_init$SL[37] = 70

write.csv(BB_init,"/Users/jordanklustner/Desktop/STAT8000/BB_init.csv", row.names = FALS


# Fitting individual predictors against shelf life

g_farm = lm(BB_init$SL ~ factor(BB_init$Farm))
summary(g_farm)

g_type = lm(BB_init$SL ~ BB_init$Type)
summary(g_type)

g_year = lm(BB_init$SL ~ factor(BB_init$Year))
summary(g_year)

g_ph = lm(BB_init$SL ~ BB_init$avg.ph)
summary(g_ph)
```

```r
g_ta = lm(BB_init$SL ~ BB_init$avg.ta)
summary(g_ta)

g_comp = lm(BB_init$SL ~ BB_init$avg.Comp)
summary(g_comp)

g_TSS = lm(BB_init$SL ~ BB_init$avg.TSS)
summary(g_TSS)

g_w = lm(BB_init$SL ~ BB_init$avg.weight)
summary(g_w)

g_def = lm(BB_init$SL ~ BB_init$avg.defect)
summary(g_def)

# Variable Selection


g = lm(BB_init$SL^(1/2) ~ BB_init$Farm + BB_init$Type + BB_init$avg.defect)
summary(g)
plot(g$fitted.values, g$residuals)

hist(BB_init$SL)

# Final Model

g_1 = lm((BB_init$SL)^(1/2) ~ BB_init$Farm + BB_init$avg.weight + BB_init$avg.defect)
summary(g_1)

plot(g_1$fitted.values, g_1$residuals, pch = 19,
     main = 'Residual vs Fitted Values of Model with Transform',
     xlab = 'Fitted Values', ylab = 'Residuals',
     col = 'darkblue')
abline(h=0)
par(mfrow = c(2, 2))
plot(g_1)

qqnorm(g_1$residuals)
qqline(g_1$residuals)

# Other Useful Plots


# Interpolating initial DF% (Southern Highbush, Emerald, 2017)
```

```r
plot(c(4,12), c(93.333,92.5), pch = 19, col = 'darkblue',
     xlim = c(0,12), ylim = c(92,94),
     main = 'Interpolating the Initial Defect Free %',
     xlab = 'Days Post Harvest', ylab = 'Defect Free %')
lines(c(0,12), c(93.75,92.5))
points(3, 93.4372, pch = 19, col = 'red')
legend(6, 94, legend=c("Observed Measurements", "Interpolated Value"),
       col=c("darkblue", "red"), pch = 19, cex=0.8)

# Finding Shelf Life

x = BB$avg.defect[BB$Year == '2016' & BB$Cultivar == 'Miss Jackie' ]
y = BB$PostHarvestDays[BB$Year == '2016' & BB$Cultivar == 'Miss Jackie']

plot(x,y, pch = 19, col = 'darkblue',
     main = 'Finding Shelf Life',
     xlab = 'Defect Free %', ylab = 'Days Post Harvest')
abline(lm(y~x))
points(75, 41.897, col = 'red', pch = 19)
legend(72, 15, legend=c("Observed Measurements", "Estimated Shelf Life"),
       col=c("darkblue", "red"), pch = 19, cex=0.8)
length(BB$Year)
```