Decision Analysis Report
Analysis of Location and Type of Bambola that Yield the Largest Profit

Minjiao Yang

Contents

1 Summary

The main purpose of this report is to find the best combination of Location and Variety in planting Bambola seeds, such that Bambola Corporation could make the most profit from 10000. This report consists of 4 major sections. The introduction gives some background information for this experiment and raises several questions providing a direction for a further study. Data Summary part studies the distribution of Color; Germination rate under various Locations and Variety; transformation for two response Height and Total. Then, the Analysis section. Firstly, we try to predict log(Height) by Variety, Location, and Color. Started with the Full Model including 3 variables and all interactions, through a sequential elimination procedure, the final model only includes location and variety. RMSE is 0.419, all assumptions are met, and (FL: B) yields the biggest height. Secondly, we try to predict total profit with our model and conclude that (FL, D) is the most profitable combination.

The last is the conclusion part. It provides the answer to several questions in which we are interested in the early analysis. We conclude that: the distribution of COLOR in our data-set does not have a significant difference with White 57%, Pink 36%, the 7% Red; Germination rate is not significantly different among varieties but significantly different between FL and LA; (FL, C) produced significantly more Red canes than the others; (FL, B) is the most favorable condition for canes to get height; (FL, D) is the most profitable combination. So I recommend plant type D bambola in Florida will yield a total profit with 95%CI in (1,012,687, 1,556,313) dollars.

2 Introduction

The Bambola Corporation aims to find the most profitable plan to plant 10,00 Bambola seeds. Specifically, the client wants to know which combination of Location and Variety is the best. Four varieties of Bambolas were planted in the past, where 256 seeds of each (1024 seeds in all), half of each variety in the FL plot and half in the LA plot. Each planted seed may or may not germinate, but most of them germinated and grew into a cane, with the distribution believed to be about 57% White, 36% Pink, and 7% Red. The value of each cane is determined by both color and height. If the cane was 1 meter or longer in height, the price is 0.20/cm, 0.30/cm, and 0.50/cm (in dollars) for White, Pink, and Red canes, respectively. If the cane was less than 1 meter in height, then was chopped into 10 cm slices that were made into whistles to be sold with the price of 1, 2, and 4 (dollars) per whistle for White, Pink, and Red whistles, respectively. Whistles should be at least 10 cm, so a cane slice shorter than 10 cm is valueless. Except for the most profitable Variety and Location combination of Bambola planting, the client is also interested in the following questions:

1. For the overall experiment (not controlling for Variety or Location) is there any evidence that the [W: P: R] color distribution is significantly different than stated before?

2. Non-germination rate for each one of the varieties. Is the probability of non-germination significantly different between varieties A and D?

3. Is there a difference between the germination rates in the two locations?

4. Is there one Variety and/or Location that produces significantly more Red canes than the others?

5. Find a model for cane Height (given that the plant germinates) as a function of Variety, Location, Color, and/or their interactions. And find the most favorable condition.

6. For each combination of Variety and Location, compute the expected total value for a crop of 10,000 seeds.

3 Data Summary

Name	Explanation
SEQ	Plant Sequence ID Number(1-1024)
LOC	Location(FL or LA)
VAR	Variety (A, B, C, D)
LIVE	Indicator of Germination of Seed (1=Yes, 0=No)
CLR	Color of Cane [White(W), Pink(P), Red(R)]
HT	Height (in cm.) of Cane
WHIST	Value (in US dollar) of Cane Whistles
POLES	Value (in US dollar) of Cane Pole
TOTAL	Total Value (in US dollar) of Cane

Table 1: Explanation of variables in Bambola experiment

There are 1024 lines of data with each line stands for a seed in the experiment. For each line, nine columns are recorded: SEQ for plant sequence number; factor LOC for the planted location (FL or LA); factor VAR for variety (A, B, C, D); factor LIVE for whether the seed germinated; factor CLR for the color of Cane (W=White, P=Pink, R=Red); HT for the height of cane in cm; WHIST for the value of cane whistles in the dollar; POLES for the value of cane pole in the dollar; TOTAL is the total value of cane in the dollar. Here, WHIST, POLES, and TOTAL can be calculated by CLR and HT. And the trail was run for 2 years which is a normal time for bambola cane to reach maturity. For each variety, there are 256 seeds and planted half in the FL plot and half in the LA plot. All the seeds of one variety were randomly planted in the plot. So, it is a balanced design. Namely, each combination of VAR and LOC has the same number of experimental observations. However, some of the seeds didn't germinate, which means no color and height, thus no profit. So we create a new dataset(dat) in which LIVE=1 for all seeds.

4 Exploratory Data Analysis

4.1 Estimate Germination

To estimate germination for each color, variety of bambola, and planting site, a chi-square independence test and logistic regression on germination rate and occurrence rate of red bambola was conducted. The results are shown in the following tables (see table 2 - table8).

4.1.1 Germination by Location

To see whether the germination rate of different variety of bambola is the same, a Chi-square test with a null hypothesis that all variety has the same germination rate was performed. And a resulting p-value of 0.4193 (appendix table 14) indicating that there is no significant difference in germination rate between the 4 varieties. In another word, all 4 types of bombola have the same germination rate.

Variety	Germinated	Non-germinated	Total	P_{germ}	$P_{non-germ}$
A	228	28	256	89.1%	10.9%
B	231	25	256	90.2%	9.8%
C	225	31	256	87.9%	12.1%
D	236	20	256	92.2%	7.8%
Total	920	104	1024	89.8%	10.2%

Table 2: Germination for each variety

4.1.2 Germination by Location

To learn that if the germination rate in different planting sites is the same. Another Chi-square test with a null hypothesis that all planting site has the same germination rate was conducted. A resulting p-value of 0.00013 (appendix table 15) implies that there is no significant difference in germination rate between the 2 planting sites. Simply put, bambolas in Florida have a higher germination rate than in LA.

Location	germinated	non-germinated	Total	P_{germ}	$P_{non-germ}$
FL	441	71	512	86.1%	13.9%
LA	479	33	512	93.6%	6.4%
Total	920	104	1024	89.8%	10.2%

Table 3: Germination for each location

Besides, the probability of germination rate for each planting site and variety was calculated from constructed logistic models. The outcomes are consistent with previous testing results. In the conclusion, the germination rate of bambolas of different types is the same, but Florida is a better planting site for bambolas growth.

4.2 Investigate Color Distribution

Naturally occurring canes have three equally strong colors, but because of scarcity, the rarer colors are more valued. All these colored canes are from the dataset(dat) and are independent of each other. People believed that the distribution of color of cane is (White, Pink, Red)=(57% ,36% ,7%), the p-value of 0.0013 of the chi-square test of goodness of fit with a null hypothesis that color distribution of bambolas followed the distribution of (57% ,36% ,7%) indicating that color distribution of bambolas is different from believed distribution. After further investigation, the real color distribution is 62% white, 30% pink, and 8%red. This is obviously to see in the Mosaic plot (figure 1).

Color	Believed Distribution	Real Distribution
White	57%	62%
Pink	36%	30%
Red	7%	8%

Table 4: Color Distribution of Cane

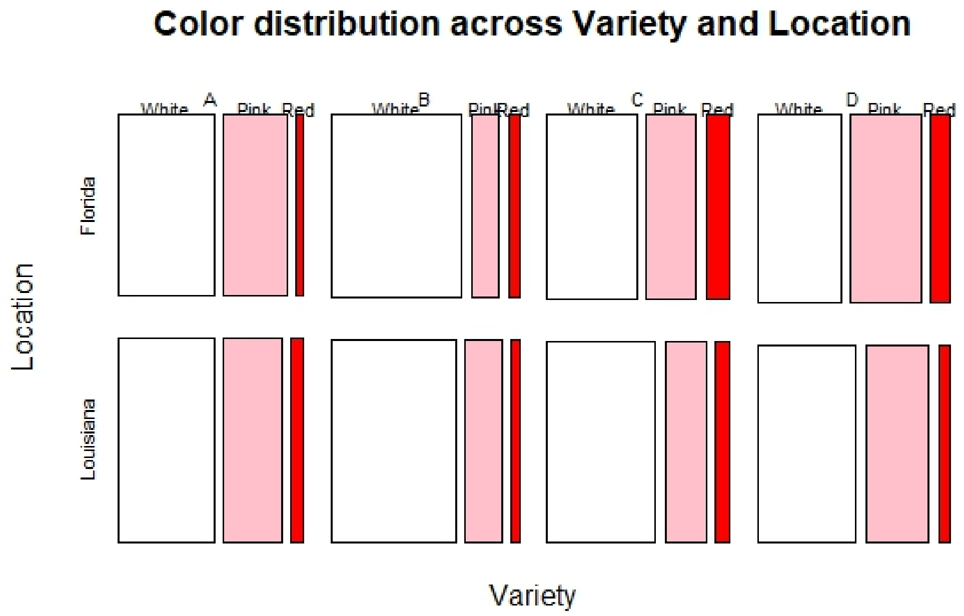

Figure 1: Mosaic plot of color distribution

In Figure 1, Florida and variety C is produced more red. But we need further investigation with chi-square independence and logistic regression tests. the p-value of 0.2726 (Appendix table 17) of A chi-square independence test with null hypothesis all locations with same red appearance rate show that there's no significant difference of red bambolas appearance rate in two planting sites.

This also can be proved by table 5, only a 2.2% distinction between two sites. And a resulting p-value of 0.0498 (Appendix table 18) of a chi-square test for variety comparison and a logistic regression analysis (table 6 and 8) reveals that the germination rate of red bambolas is significantly higher than others. Moreover, a test result of p-value 0.089 (Appendix table 19) for comparing each variety in each location demonstrate that red bambolas germination rate is a significant difference for different type in a different location, and variety C in Florida has a higher production rate of red bambolas (see table 7 and Figure 2).

Location	Red	Total	P_{red}
FL	41	441	9.3%
LA	34	479	7.1%

Table 5: Red Cane for each location

Variety	Red	Total	P_{red}
A	14	228	6.1%
B	13	231	5.6%
C	27	225	12%
D	21	236	8.9%

Table 6: Red Cane for each variety

Coefficients	Estimate	Std.Error	z value	Pr(>|z|)
(Intercept)	-2.72692	0.27586	-9.885	<2e-16
Var(B)	-0.09263	0.39700	-0.233	0.8155
Var(C)	0.73449	0.34379	2.136	0.0326
Var(D)	0.40080	0.35829	1.119	0.2633

Table 7: Red Cane for each variety

Variety	Location	
	FL	LA
A	4.7%	7.4%
B	6.4%	5.0%
C	14.8%	9.4%
D	11.2%	6.7%

Table 8: Red Cane for each variety in each location

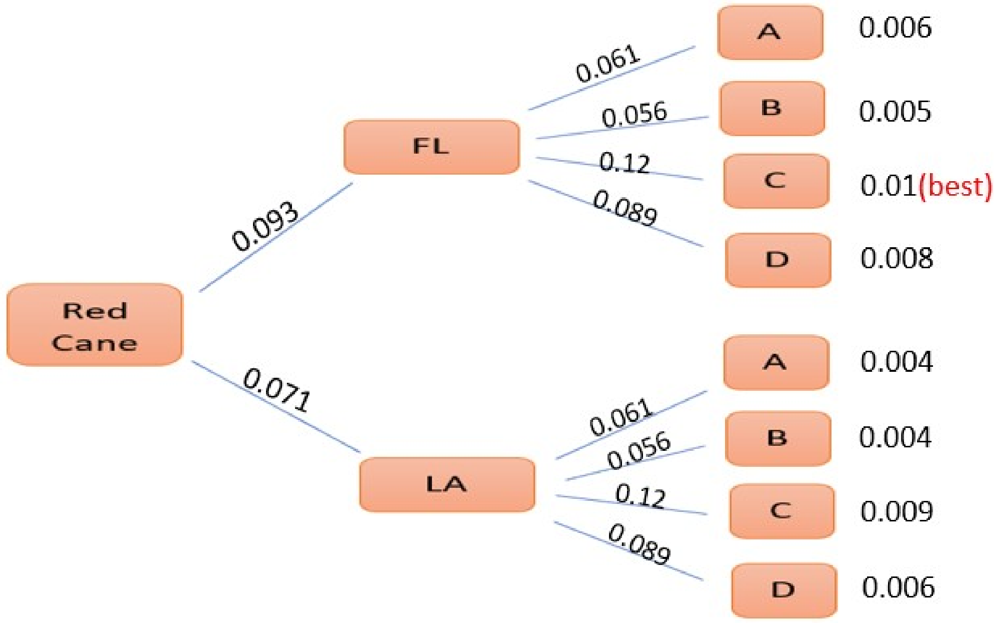

Figure 2: Decision Tree of red bambolas germination rate

5 Model for Cane Height

Since Height is a response which might be predicted by location (LOC), Varitey(VAR) and color(CLR), so, it should satisfy the assumptions for 3-way ANOVA model that the response variable should be normally distributed. But it is apparently skewed with longer tail(Figure 3). Thus, we could consider Box-Cox transformation to find a transformed Height such that the transformed Height subject to normal. From the box-cox transformation plot for Height (Appendix figure 5), 95% CI for λ includes 0, so we do log transformation on Height.After transformation, the distribution of cane height is normally distributed.

Modeling for cane height was started with full models including locations, varieties, color, and their interactions. After backward model selection of eliminating the insignificant variables and interactions, the final model is:
$log10(Height) = 2.5315 - 0.1150 \times I(LA)0.0371 \times I(B) - 0.0721 \times I(C) - 0.0451 \times I(D)$
Summary of the model (table 10) shows that with baseline site FL and type A, the cane height in Florida is significantly higher than in LA, and type B is higher than type A, C, and D. Therefore, we

conclude that Variety B in Florida is the most favorable condition for growing bambolas and will end up with higher height.

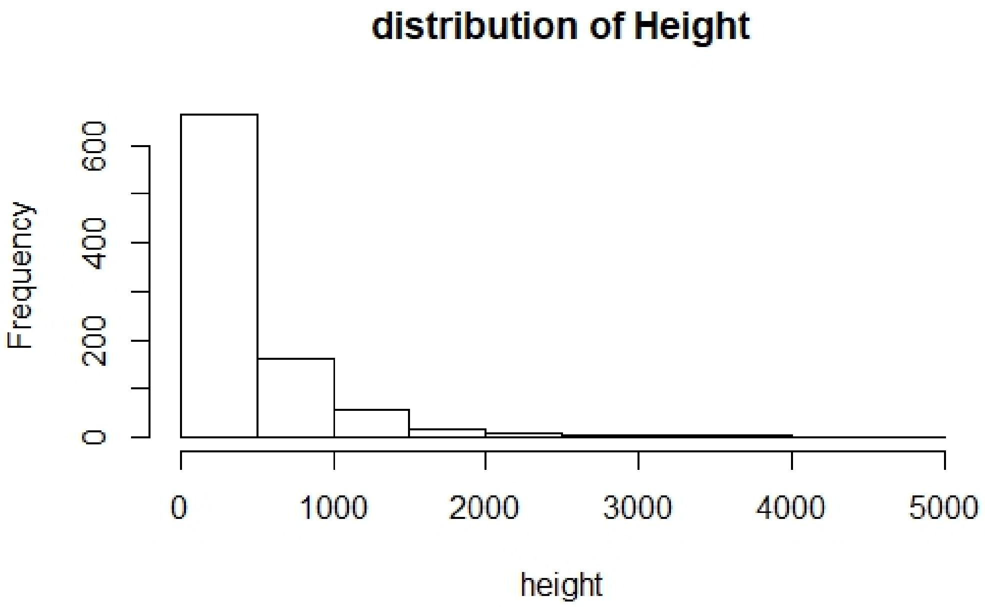

Figure 3: Distribution of Cane Height

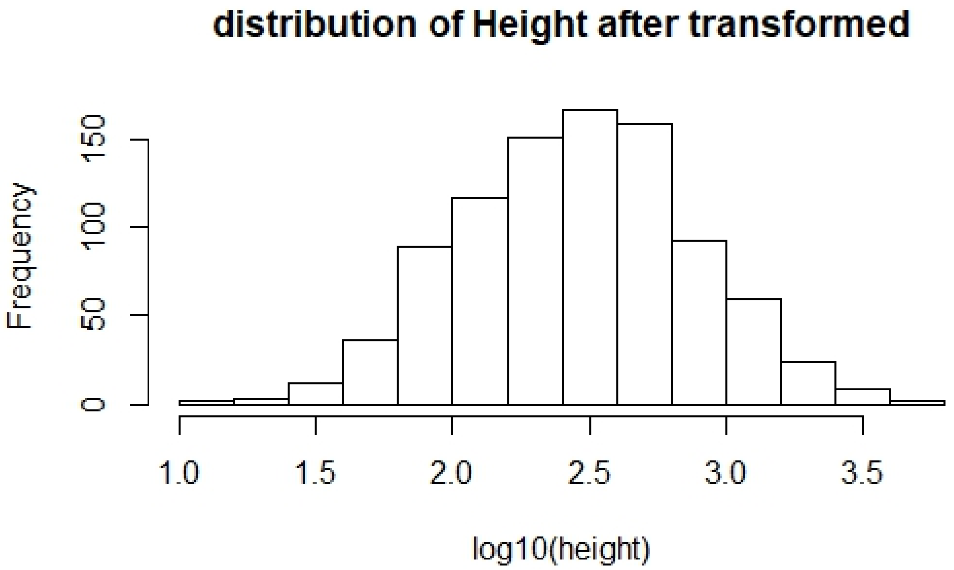

Figure 4: Distribution of Cane Height after Transformation

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.5315	0.0314	80.63	0.0000
factor(LOC)LA	-0.1150	0.0277	-4.16	0.0000
factor(VAR)B	0.0371	0.0391	0.95	0.3429
factor(VAR)C	-0.0721	0.0394	-1.83	0.0674
factor(VAR)D	-0.0451	0.0389	-1.16	0.2468

Table 9: model for cane height

5.1 The Variety-Location that yield the Most Profit

From table 11 and 12, we can see type D bambolas in Florida has highest average total value of profit, but $SE = 155.88/\sqrt{128} = 13.78$ is large, which is fall into \$ 11-15 , that is most of the total value for each type of bambolas in each location are not significant different from each other. Besides, average total value of profit standard error for type C and D in LA are small. So, they are significant worsen than others. Because (FL,D) yield the greatest profit, when plant 10,000 seeds, we will get the expected total value be:
$E(T) = 10,000 \times \$128.45 = \$1,284,500$
And calculation for 95% CI of total value of profit for type D in Florida is shown in details:
$SE(T) = \$13.78 \times 10,000 = \$13,7800$
$SD(T) = 100 \times \$155.88 = \$15,588$
$SD(T) = \sqrt{(\$15,588)^2 + (\$13,7800)^2} = \$138,680$
95% prediction Interval for total value of (FL, D) is
$(E(T) - 1.96 \times SD(T), E(T) + 1.96 \times SD(T)) = (\$1,012,687, \$1,556,313)$

So Florida and Variety D expect to yield total value between \$ 1,012,687 and \$ 1,556,313.

	FL	LA
A	128	128
B	128	128
C	128	128
D	128	128

Table 10: Size for total value

	FL	LA
A	102.49	121.80
B	108.15	106.99
C	110.70	70.29
D	128.45	76.66

Table 11: Mean for total value

	FL	LA
A	157.77	174.54
B	150.76	130.55
C	152.28	90.13
D	155.88	80.04

Table 12: SD for total value

6 Conclusion

Based on our analysis above, we can conclude that:

1. The distribution of COLOR in our data-set has not significant difference with White 57%, Pink 36%, the 7% Red.

2. Non-germination rate not significantly different between varieties A and D.

3. Non-germination rates are significantly different between location LA and FL.

4. The combination (FL, C) produced significantly more Red canes than the others.

5. The combination (FL, B) is the most favorable condition for canes to get height.

6. (FL, D) is most profitable combination. So I recommend plant type D bambola in Florida will yield a total profit with 95%CI in (1,012,687, 1,556,313) dollars.

7 Appendix

1. Germination for each variety

$X^2 = 2.8254$	df = 3	p-value = 0.4193

Table 13: Pearson's Chi-squared test

2. Germination for each Location

$X^2 = 14.652$	df = 1	p-value = 0.0001293

Table 14: Pearson's Chi-squared test

3. Color Distribution of Cane

$X^2 = 13.272$	df = 2	p-value = 0.001312

Table 15: Pearson's Chi-squared test for given probability

4. Red Cane for each location

$X^2 = 1.2036$	df = 1	p-value = 0.2726

Table 16: Pearson's Chi-squared test

5. Red Cane for each Variety

$X^2 = 7.8231$	df = 3	p-value = 0.04981

Table 17: Pearson's Chi-squared test

6. Red Cane for each variety in each location

$X^2 = 12.376$	df = 7	p-value = 0.08884

Table 18: Pearson's Chi-squared test

7. Box-cox Transformation plot for cane height

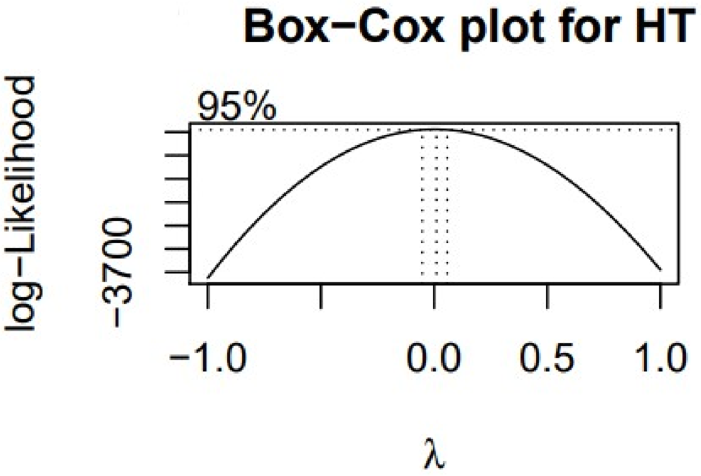

Figure 5: Box-cox Transformation plot for cane height