The question asked is if students who took the CURES class perform better or differently in the lecture than students who did not take the CURES class. The response for this test is letter grades for the lecture. Letter grades are an ordinal categorical response. The categorical binary predictor is whether the student took the CURES class or not. Logistic ordinal regression was used to determine a relationship exists between taking the CURES class and achieving a higher letter grade in the lecture. Before I ran the data analysis, I cleaned the data by removing students with a W or an I as the researcher mentioned he did not want to include the W's or I in the analysis.

$H_0 : B_K = 0$
$H_A: B_K \neq 0$

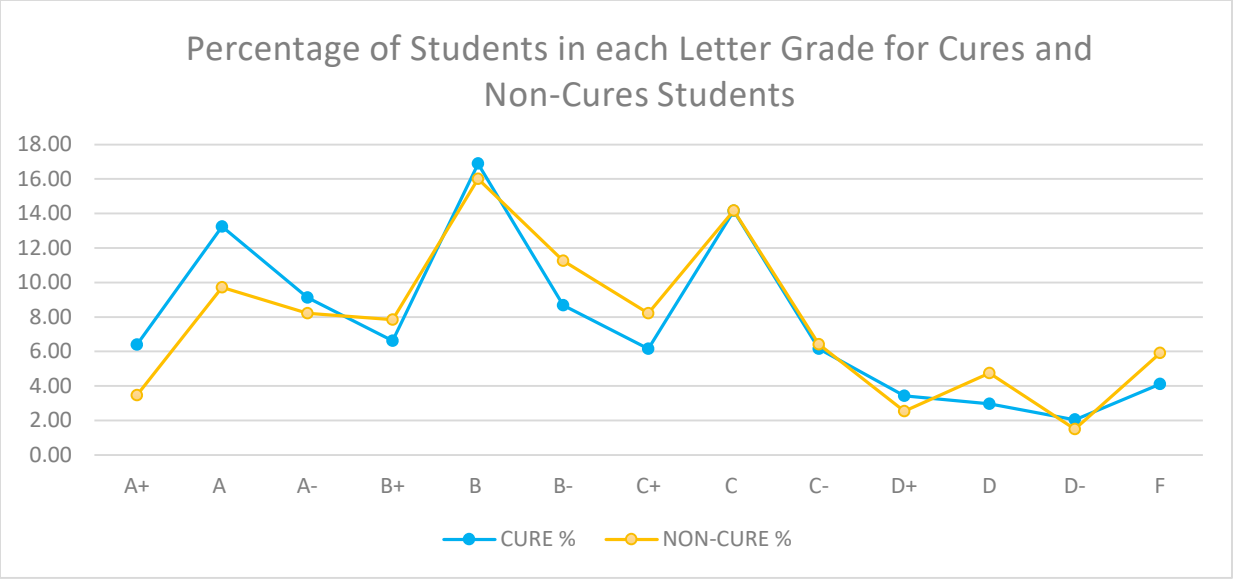
Response Information

Variable	Value	Count
Lecture_Grade	A+	111
	A	291
	A-	237
	B+	217
	B	458
	B-	308
	C+	224
	C	402
	C-	181
	D+	76
	D	127
	D-	45
	F	160
	I	1
	W	357
	Total	3195

Response Information

Variable	Value	Count
Lecture_Grade	A+	111
	A	291
	A-	237
	B+	217
	B	458
	B-	308
	C+	224
	C	402
	C-	181
	D+	76
	D	127
	D-	45
	F	160
	Total	2837

Grade	CURE	NON-CURE	CURE %	NON-CURE %
A+	28	83	6.39	3.46
A	58	233	13.24	9.71
A-	40	197	9.13	8.21
B+	29	188	6.62	7.84
B	74	384	16.89	16.01
B-	38	270	8.68	11.25
C+	27	197	6.16	8.21
C	62	340	14.16	14.17
C-	27	154	6.16	6.42
D+	15	61	3.42	2.54
D	13	114	2.97	4.75
D-	9	36	2.05	1.50
F	18	142	4.11	5.92
total	**438**	**2399**		



Logistic Regression Table

						95% CI	
Predictor	**Coef**	**SE Coef**	**Z**	**P**	**Odds Ratio**	**Lower**	**Upper**
Const(1)	-2.9750	0.1222	-24.35	0.0000			
Const(2)	-1.5729	0.0924	-17.03	0.0000			
Const(3)	-1.0057	0.0878	-11.45	0.0000			
Const(4)	-0.6086	0.0862	-7.06	0.0000			
Const(5)	0.0843	0.0853	0.99	0.3231			
Const(6)	0.5217	0.0859	6.07	0.0000			
Const(7)	0.8550	0.0868	9.85	0.0000			
Const(8)	1.5726	0.0906	17.35	0.0000			
Const(9)	2.0173	0.0947	21.30	0.0000			
Const(10)	2.2545	0.0976	23.09	0.0000			
Const(11)	2.7864	0.1068	26.09	0.0000			
Const(12)	3.0513	0.1130	26.99	0.0000			
Course_Type							
Non_Cure	-0.2721	0.0907	-3.00	0.0027	0.76	0.64	0.91

The logistic regression table above provides a significant **p-value** of 0.0027. This tells us there is a relationship that exists between the response letter grades and the binary predictor on whether a student took the cures class. Any p-value less than 0.05 would indicate significance with a 5% alpha level.

The **odds ratio** of 0.76 tells us (1 - 0.76 = 0.24) tells us the odds of achieving a higher grade in the lecture is 24% lower for non-cures students when compared with students who took the CURES class.

Test of All Slopes Equal to Zero

DF	G	P-Value
1	8.728	0.003

The test above tells us our slope is not zero. Which indicates we are seeing a change or a difference when we compare NON-CURES to CURES students. We will accept the alternative hypothesis H_A: $B_K \neq 0$ and reject the null hypothesis.

Goodness-of-Fit Tests

Method	Chi-Square	DF	P
Pearson	15.8610	11	0.146
Deviance	15.3069	11	0.169

With 11 degrees of freedom, we have a critical value of 19.675. Since the deviance is 15.861, we would conclude that the model is fitting!

Next, I wanted to look at the other factors to determine which other factors are significant in predicting the outcome of a student's final grade. When I ran the ordinal logistic regression with gender as the binary predictor I got a p-value of 0.684. This value tells us there is no association between gender and the outcome of the grade. The odds ratio of 1 tells us students are equally as likely to score the same grades regardless of gender.

Logistic Regression Table

						95% CI	
Predictor	**Coef**	**SE Coef**	**Z**	**P**	**Odds Ratio**	**Lower**	**Upper**
Const(1)	-3.21002	0.0993218	-32.32	0.000			
Const(2)	-1.81013	0.0581761	-31.11	0.000			
Const(3)	-1.24425	0.0500403	-24.87	0.000			
Const(4)	-0.847926	0.0464244	-18.26	0.000			
Const(5)	-0.156438	0.0435396	-3.59	0.000			
Const(6)	0.280050	0.0437595	6.40	0.000			
Const(7)	0.613179	0.0449903	13.63	0.000			
Const(8)	1.33048	0.0511107	26.03	0.000			
Const(9)	1.77510	0.0577125	30.76	0.000			
Const(10)	2.01207	0.0622790	32.31	0.000			
Const(11)	2.54370	0.0756602	33.62	0.000			
Const(12)	2.80850	0.0842004	33.35	0.000			
Gender							
M	0.0288191	0.0708229	0.41	0.684	1.03	0.90	1.18

The next factor I analyzed was ethnicity and from the p-values associated with each race you will see that ethnic background is not statistically significant.

Logistic Regression Table

Predictor	**Coef**	**SE Coef**	**Z**	**P**	**Odds Ratio**
Const(1)	-3.3140	0.5587	-5.93	0.000	
Const(2)	-1.9034	0.5528	-3.44	0.001	
Const(3)	-1.3284	0.5520	-2.41	0.016	
Const(4)	-0.9235	0.5516	-1.67	0.094	
Const(5)	-0.2137	0.5514	-0.39	0.698	
Const(6)	0.2366	0.5514	0.43	0.668	
Const(7)	0.5815	0.5515	1.05	0.292	
Const(8)	1.3263	0.5520	2.40	0.016	
Const(9)	1.7857	0.5527	3.23	0.001	
Const(10)	2.0283	0.5532	3.67	0.000	
Const(11)	2.5706	0.5550	4.63	0.000	
Const(12)	2.8398	0.5562	5.11	0.000	
Ethnicity					
Asian	0.4757	0.5653	0.84	0.400	1.61
Black or African American	-0.9173	0.5594	-1.64	0.101	0.40
Did not respond	-0.2670	0.6337	-0.42	0.674	0.77
Hispanic	-0.2844	0.5664	-0.50	0.616	0.75
Multi-Racial	-0.1767	0.5886	-0.30	0.764	0.84
Native Hawaiian Islander	1.1059	1.1506	0.96	0.336	3.02
White	0.2190	0.5524	0.40	0.692	1.24

	95% CI	
Predictor	**Lower**	**Upper**
Const(1)		
Const(2)		
Const(3)		
Const(4)		
Const(5)		
Const(6)		
Const(7)		
Const(8)		
Const(9)		
Const(10)		
Const(11)		
Const(12)		
Ethnicity		
Asian	0.53	4.87
Black or African American	0.13	1.20
Did not respond	0.22	2.65
Hispanic	0.25	2.28
Multi-Racial	0.26	2.66
Native Hawaiian or Other Pacific Islander	0.32	28.82
White	0.42	3.68

The next factor, age, is a continuous factor. The age of the student is statistically significant with a p-value of 0.026. The coefficient for age is negative so we would expect the lecture grade of a student to decrease as age increases.

Logistic Regression Table

						95% CI	
Predictor	**Coef**	**SE Coef**	**Z**	**P**	**Odds Ratio**	**Lower**	**Upper**
Const(1)	-2.41303	0.367592	-6.56	0.000			
Const(2)	-1.01231	0.358887	-2.82	0.005			
Const(3)	-0.445984	0.357844	-1.25	0.213			
Const(4)	-0.0495730	0.357534	-0.14	0.890			
Const(5)	0.642076	0.357587	1.80	0.073			
Const(6)	1.07905	0.357926	3.01	0.003			
Const(7)	1.41252	0.358331	3.94	0.000			
Const(8)	2.13060	0.359710	5.92	0.000			
Const(9)	2.57616	0.361026	7.14	0.000			
Const(10)	2.81364	0.361941	7.77	0.000			
Const(11)	3.34595	0.364790	9.17	0.000			
Const(12)	3.61081	0.366780	9.84	0.000			
Age (as of 6_13_2021)	-0.0366487	0.0165077	-2.22	0.026	0.96	0.93	1.00

The factor for residency was analyzed. It is a binary predictor as we have two options either LA or OOS (out of state). The factor is statistically significant when inferring lecture grades. The p-value is 0.00001.

Logistic Regression Table

						95% CI	
Predictor	**Coef**	**SE Coef**	**Z**	**P**	**Odds Ratio**	**Lower**	**Upper**
Const(1)	-3.150	0.097	-32.33	0.00000			
Const(2)	-1.748	0.055	-31.77	0.00000			
Const(3)	-1.181	0.046	-25.42	0.00000			
Const(4)	-0.784	0.043	-18.36	0.00000			
Const(5)	-0.091	0.040	-2.27	0.02309			
Const(6)	0.347	0.040	8.59	0.00000			
Const(7)	0.681	0.042	16.24	0.00000			
Const(8)	1.401	0.049	28.74	0.00000			
Const(9)	1.849	0.056	33.14	0.00000			
Const(10)	2.087	0.061	34.46	0.00000			
Const(11)	2.621	0.074	35.26	0.00000			
Const(12)	2.886	0.083	34.76	0.00000			
Residency							
OOS	-0.427	0.094	-4.55	0.00001	0.65	0.54	0.78

The factor 1st Generation is found to be statistically significant. The p-value for this factor is 0.0002. The coefficient for 1st generation students is negative so we would expect lower grade levels for 1st generation students when compared with non-1st generation students.

Logistic Regression Table

						95% CI	
Predictor	**Coef**	**SE Coef**	**Z**	**P**	**Odds Ratio**	**Lower**	**Upper**
Const(1)	-3.156	0.097	-32.37	0.0000			
Const(2)	-1.754	0.055	-31.83	0.0000			
Const(3)	-1.187	0.047	-25.49	0.0000			
Const(4)	-0.789	0.043	-18.44	0.0000			
Const(5)	-0.095	0.040	-2.38	0.0175			
Const(6)	0.343	0.040	8.48	0.0000			
Const(7)	0.677	0.042	16.13	0.0000			
Const(8)	1.395	0.049	28.61	0.0000			
Const(9)	1.841	0.056	33.00	0.0000			
Const(10)	2.078	0.061	34.33	0.0000			
Const(11)	2.611	0.074	35.13	0.0000			
Const(12)	2.876	0.083	34.64	0.0000			
1st Gen							
Y	-0.351	0.093	-3.78	0.0002	0.70	0.59	0.84

The next factor Pell grant recipient also proves to be statistically significant with a p-value of <0.001. Since the coefficient is negative for students who have a Pell grant, we would expect these students to perform lower when compared to non-Pell grant recipients.

Logistic Regression Table

						95% CI	
Predictor	**Coef**	**SE Coef**	**Z**	**P**	**Odds Ratio**	**Lower**	**Upper**
Const(1)	-3.05722	0.0977814	-31.27	0.000			
Const(2)	-1.64833	0.0558802	-29.50	0.000			
Const(3)	-1.07622	0.0476937	-22.57	0.000			
Const(4)	-0.674388	0.0442449	-15.24	0.000			
Const(5)	0.0278904	0.0421496	0.66	0.508			
Const(6)	0.472217	0.0430966	10.96	0.000			
Const(7)	0.812357	0.0449139	18.09	0.000			
Const(8)	1.54633	0.0521134	29.67	0.000			
Const(9)	1.99985	0.0590850	33.85	0.000			
Const(10)	2.24015	0.0637381	35.15	0.000			
Const(11)	2.77806	0.0771276	36.02	0.000			
Const(12)	3.04563	0.0856149	35.57	0.000			
Pell							
Y	-0.731439	0.0765342	-9.56	0.000	0.48	0.41	0.56

As expected, high school GPA is a statistically significant predictor when comparing lecture grades. With a positive coefficient will find as the high GPA increases so does the grade for the lecture. The odds ratio of 27.90 tells us students with higher high school GPA's are 27.9 times more likely to obtain higher lecture scores than students with lower GPA scores!

Logistic Regression Table

						95% CI	
Predictor	**Coef**	**SE Coef**	**Z**	**P**	**Odds Ratio**	**Lower**	**Upper**
Const(1)	-15.6957	0.460854	-34.06	0.000			
Const(2)	-14.2083	0.450634	-31.53	0.000			
Const(3)	-13.5635	0.447080	-30.34	0.000			
Const(4)	-13.0975	0.444124	-29.49	0.000			
Const(5)	-12.2723	0.437604	-28.04	0.000			
Const(6)	-11.7306	0.432688	-27.11	0.000			
Const(7)	-11.3141	0.428743	-26.39	0.000			
Const(8)	-10.4226	0.420298	-24.80	0.000			
Const(9)	-9.89411	0.416180	-23.77	0.000			
Const(10)	-9.61109	0.414416	-23.19	0.000			
Const(11)	-8.99549	0.411956	-21.84	0.000			
Const(12)	-8.68659	0.411756	-21.10	0.000			
HS GPA	3.32847	0.119056	27.96	0.000	27.90	22.09	35.23

Eqv Comp is statistically significant with a p-value of <0.001.

Logistic Regression Table

						95% CI	
Predictor	**Coef**	**SE Coef**	**Z**	**P**	**Odds Ratio**	**Lower**	**Upper**
Const(1)	-10.789	0.311	-34.69	0.00000			
Const(2)	-9.245	0.294	-31.46	0.00000			
Const(3)	-8.590	0.288	-29.82	0.00000			
Const(4)	-8.126	0.284	-28.62	0.00000			
Const(5)	-7.300	0.276	-26.44	0.00000			
Const(6)	-6.770	0.271	-24.94	0.00000			
Const(7)	-6.369	0.268	-23.74	0.00000			
Const(8)	-5.532	0.263	-21.00	0.00000			
Const(9)	-5.041	0.262	-19.24	0.00000			
Const(10)	-4.783	0.262	-18.26	0.00000			
Const(11)	-4.216	0.264	-15.99	0.00000			
Const(12)	-3.925	0.266	-14.78	0.00000			
Eqv Comp	0.270	0.010	26.34	0.00000	1.31	1.28	1.34

Eqv English is statistically significant with a p-value <0.001.

Logistic Regression Table

						95% CI	
Predictor	**Coef**	**SE Coef**	**Z**	**P**	**Odds Ratio**	**Lower**	**Upper**
Const(1)	-8.583	0.253	-33.98	0.00000			
Const(2)	-7.083	0.235	-30.13	0.00000			
Const(3)	-6.451	0.230	-28.06	0.00000			
Const(4)	-6.004	0.226	-26.55	0.00000			
Const(5)	-5.212	0.219	-23.78	0.00000			
Const(6)	-4.703	0.215	-21.87	0.00000			
Const(7)	-4.317	0.212	-20.34	0.00000			
Const(8)	-3.501	0.208	-16.82	0.00000			
Const(9)	-3.018	0.208	-14.54	0.00000			
Const(10)	-2.762	0.208	-13.29	0.00000			
Const(11)	-2.198	0.211	-10.43	0.00000			
Const(12)	-1.909	0.214	-8.93	0.00000			
Eqv English	0.180	0.008	23.65	0.00000	1.20	1.18	1.21

ACT Science is statistically significant with a p-value <0.001.

Logistic Regression Table

						95% CI	
Predictor	**Coef**	**SE Coef**	**Z**	**P**	**Odds Ratio**	**Lower**	**Upper**
Const(1)	-8.855	0.291	-30.44	0.00000			
Const(2)	-7.345	0.272	-26.96	0.00000			
Const(3)	-6.714	0.267	-25.19	0.00000			
Const(4)	-6.266	0.263	-23.87	0.00000			
Const(5)	-5.493	0.256	-21.46	0.00000			
Const(6)	-4.991	0.252	-19.78	0.00000			
Const(7)	-4.612	0.250	-18.43	0.00000			
Const(8)	-3.784	0.247	-15.31	0.00000			
Const(9)	-3.301	0.247	-13.36	0.00000			
Const(10)	-3.032	0.248	-12.24	0.00000			
Const(11)	-2.463	0.251	-9.81	0.00000			
Const(12)	-2.162	0.254	-8.51	0.00000			
ACT Science	0.205	0.010	21.32	0.00000	1.23	1.20	1.25

Semester hours carried is significant with a p-value <0.001. It shows that students with higher semester hours tend to perform better in lecture grade.

Logistic Regression Table

						95% CI	
Predictor	**Coef**	**SE Coef**	**Z**	**P**	**Odds Ratio**	**Lower**	**Upper**
Const(1)	-7.52770	0.257866	-29.19	0.000			
Const(2)	-6.09781	0.242856	-25.11	0.000			
Const(3)	-5.50799	0.239330	-23.01	0.000			
Const(4)	-5.09010	0.237013	-21.48	0.000			
Const(5)	-4.35283	0.232875	-18.69	0.000			
Const(6)	-3.87904	0.230287	-16.84	0.000			
Const(7)	-3.51363	0.228415	-15.38	0.000			
Const(8)	-2.73112	0.225097	-12.13	0.000			
Const(9)	-2.24980	0.224069	-10.04	0.000			
Const(10)	-1.99487	0.224005	-8.91	0.000			
Const(11)	-1.43318	0.225522	-6.35	0.000			
Const(12)	-1.15854	0.227439	-5.09	0.000			
SemHrCarr	0.296019	0.0161648	18.31	0.000	1.34	1.30	1.39

Semester hours earned is significant with a p-value <0.001. It shows that students with higher semester hours tend to perform better in lecture grade.

Logistic Regression Table

						95% CI	
Predictor	**Coef**	**SE Coef**	**Z**	**P**	**Odds Ratio**	**Lower**	**Upper**
Const(1)	-9.112	0.212	-43.07	0.000			
Const(2)	-7.651	0.193	-39.57	0.000			
Const(3)	-7.033	0.189	-37.24	0.000			
Const(4)	-6.587	0.186	-35.44	0.000			
Const(5)	-5.778	0.180	-32.09	0.000			
Const(6)	-5.239	0.176	-29.78	0.000			
Const(7)	-4.810	0.172	-27.88	0.000			
Const(8)	-3.845	0.164	-23.40	0.000			
Const(9)	-3.209	0.159	-20.13	0.000			
Const(10)	-2.855	0.157	-18.15	0.000			
Const(11)	-2.046	0.155	-13.20	0.000			
Const(12)	-1.626	0.157	-10.39	0.000			
SemHrsEarn	0.408	0.013	32.25	0.000	1.50	1.47	1.54

The remaining factors were not analyzed because those terms should not hold statistical significance. The chart below shows which factors are statistically significant in determining a student's lecture grade and which factors have no association with the lecture grade.

Significant	**No Association**
Non-Cure Class	Gender
Age	Ethnicity
Residency	
1st Gen	
Pell Grant	
High School GPA	
Eqv Comp	
Eqv English	
ACT Science	
Semester Hours Carried	
Semester Hours Earned	

The next question asked by the researcher was would we see the same results if they removed the 1207 or the honors students from the population. They wanted to make sure they were not creating bias with an honors class. I removed the 133 students who were part of the honors class. It should be noted that all the honors students did take the cures and most of these students scored a B level grade or better in the biology course. The results below show a p-value of 0.047. Since 0.047 is less than 0.05 I would conclude again the factor CURES is statistically significant in determining performance in the biology course.

Logistic Regression Table

						95% CI	
Predictor	**Coef**	**SE Coef**	**Z**	**P**	**Odds Ratio**	**Lower**	**Upper**
Const(1)	-3.51304	0.141804	-24.77	0.000			
Const(2)	-2.10133	0.111077	-18.92	0.000			
Const(3)	-1.51518	0.105980	-14.30	0.000			
Const(4)	-1.10344	0.103793	-10.63	0.000			
Const(5)	-0.398036	0.101834	-3.91	0.000			
Const(6)	0.0495159	0.101549	0.49	0.626			
Const(7)	0.386586	0.101833	3.80	0.000			
Const(8)	1.11054	0.104254	10.65	0.000			
Const(9)	1.55429	0.107453	14.46	0.000			
Const(10)	1.78834	0.109865	16.28	0.000			
Const(11)	2.32113	0.117860	19.69	0.000			
Const(12)	2.58158	0.123377	20.92	0.000			
Course_Type							
Non_Cure	0.210884	0.105967	1.99	0.047	1.23	1.00	1.52

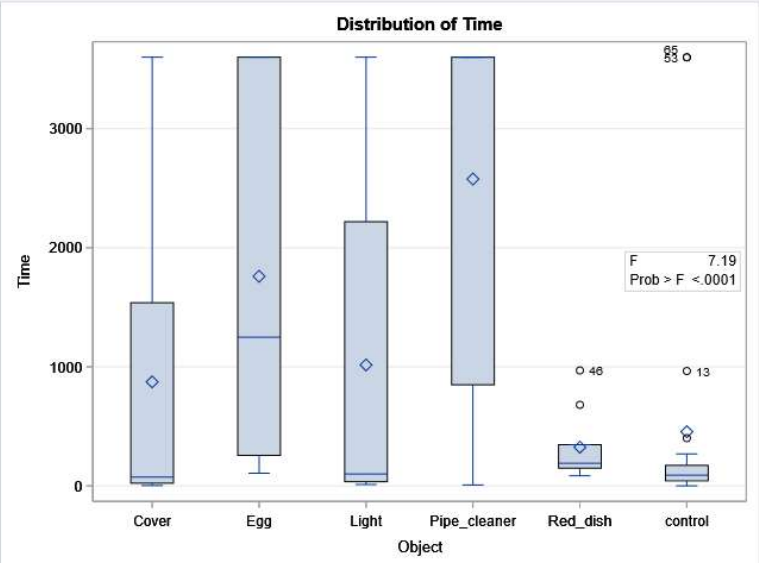
After our preliminary discussion, I wanted to examine the food and object test variables to determine if the mean time to the dish is equal for various objects and foods. After our discussion with Dr. Li, I believe we learned that for the time distribution only two groups exist. However, you want to have three groups so that you can remove the intermediate noise "group". After running an ANOVA on each of the variables included in the habituation stacked spreadsheet it was determined the objects used in this experiment are not equal. The data can be found below. From the data below, we find the red dish is very close to the control so I would recommend that we eliminate the red dish from the experiment. With the cover and the light variable, we do not see many neophobic birds as this object is not too severe. Yet with the pipe cleaner almost all birds are neophobic. I would also recommend removing these 3 variables from the experiment too and only use the egg an object. The egg includes the entire sample space and gives us both neophobic birds and non-neophobic birds while reducing the variables in the model. The egg provides the best differentiation between neophobic and non-neophobic birds.

Dependent Variable: Time

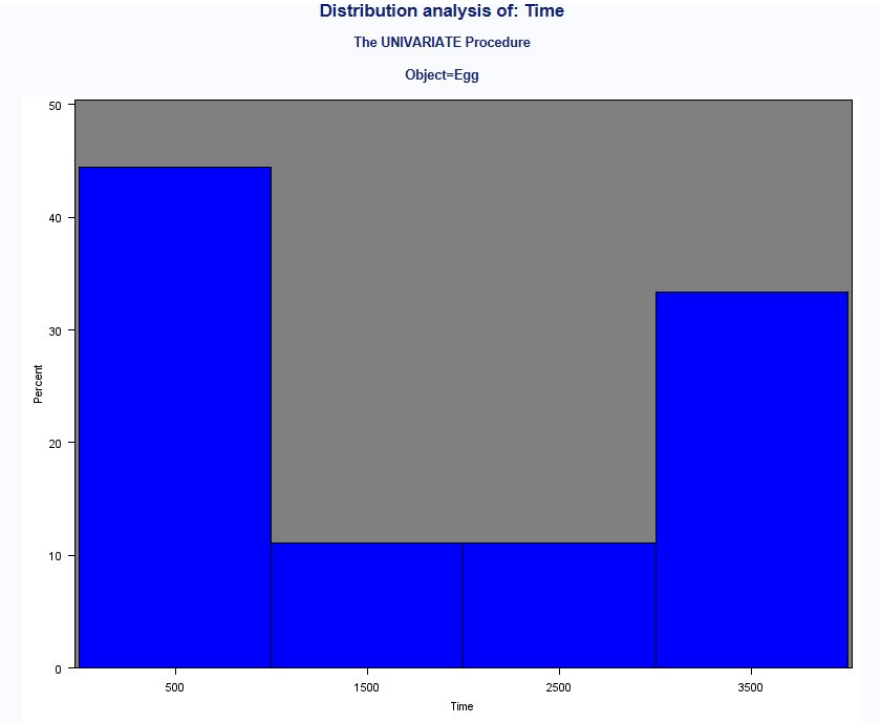
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	63706607.3	12741321.5	7.19	<.0001
Error	82	145319180.3	1772185.1		
Corrected Total	87	209025787.6			

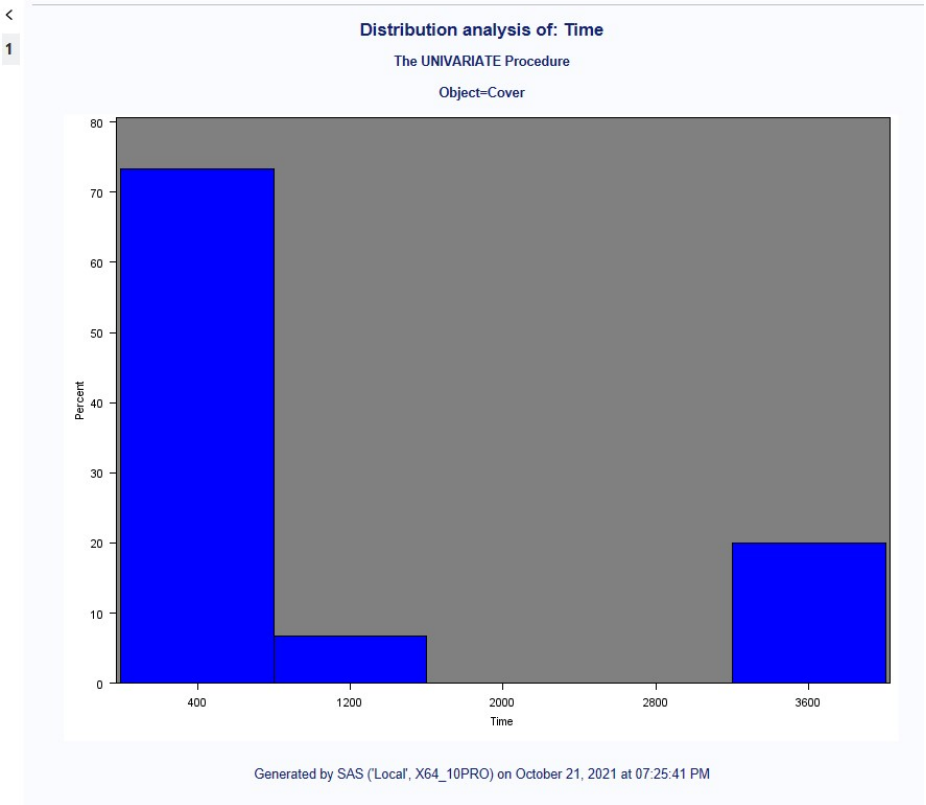
R-Square	Coeff Var	Root MSE	Time Mean
0.304779	108.2685	1331.234	1229.568

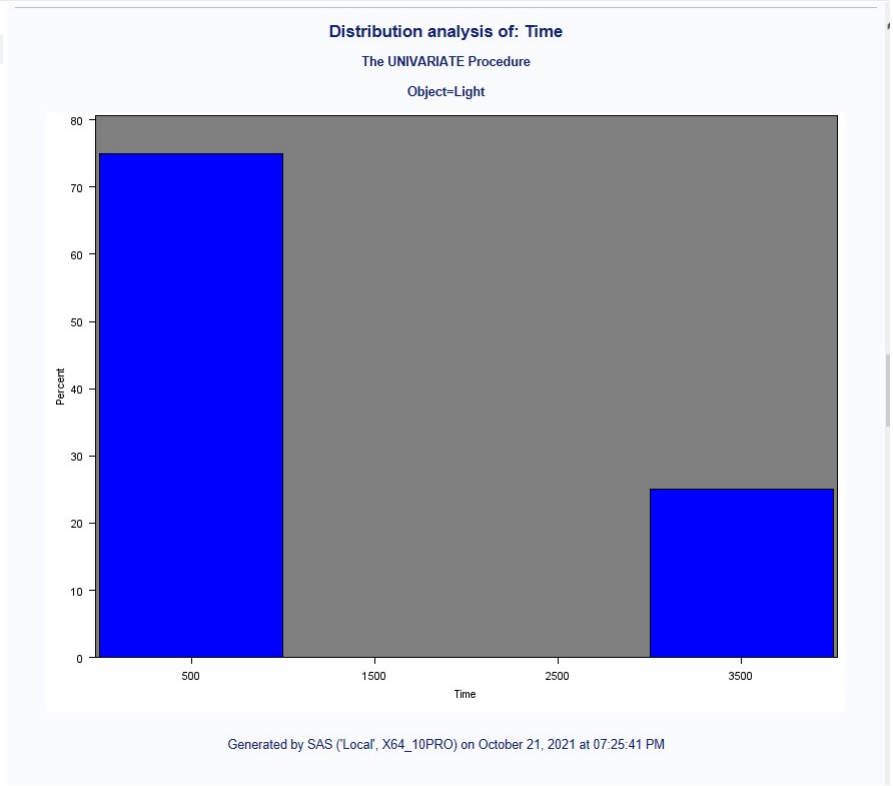
Source	DF	Anova SS	Mean Square	F Value	Pr > F
Object	5	63706607.27	12741321.45	7.19	<.0001

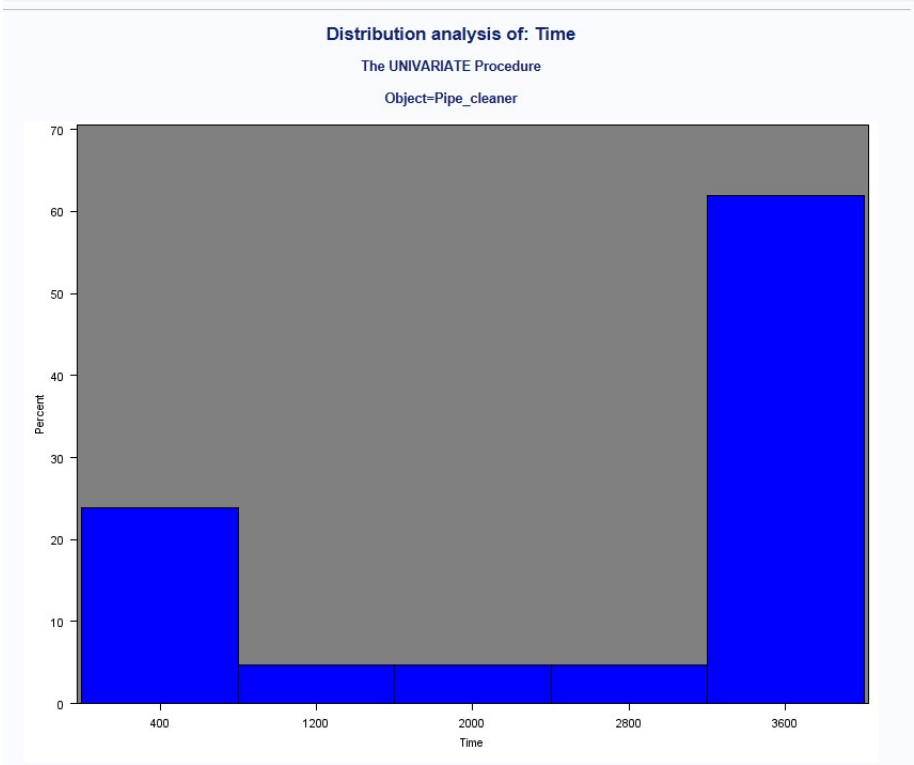


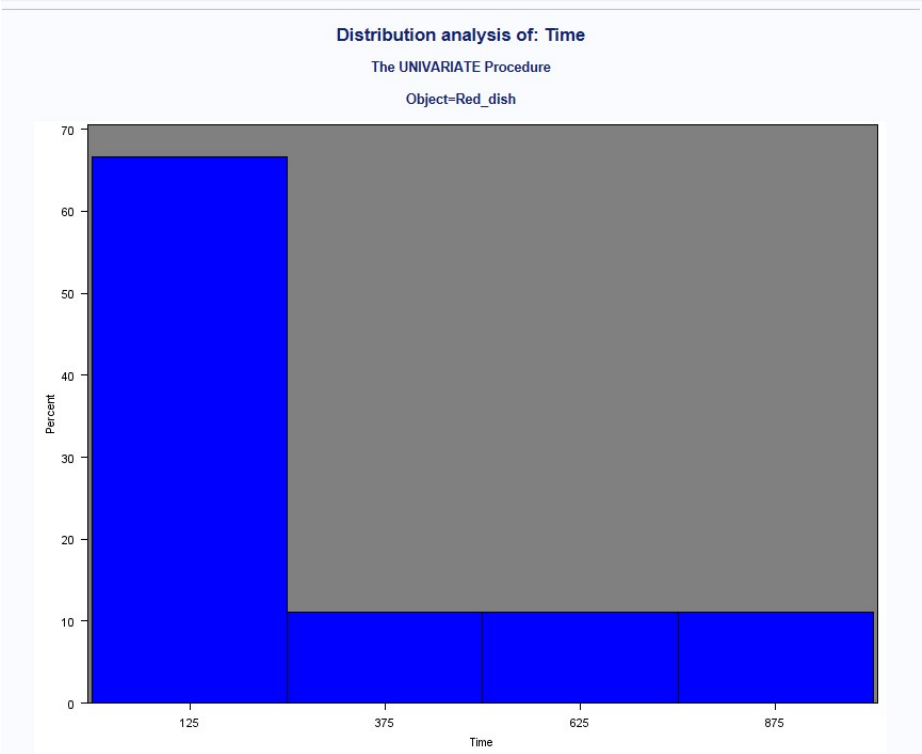
HISTOGRAMS OF OBJECTS

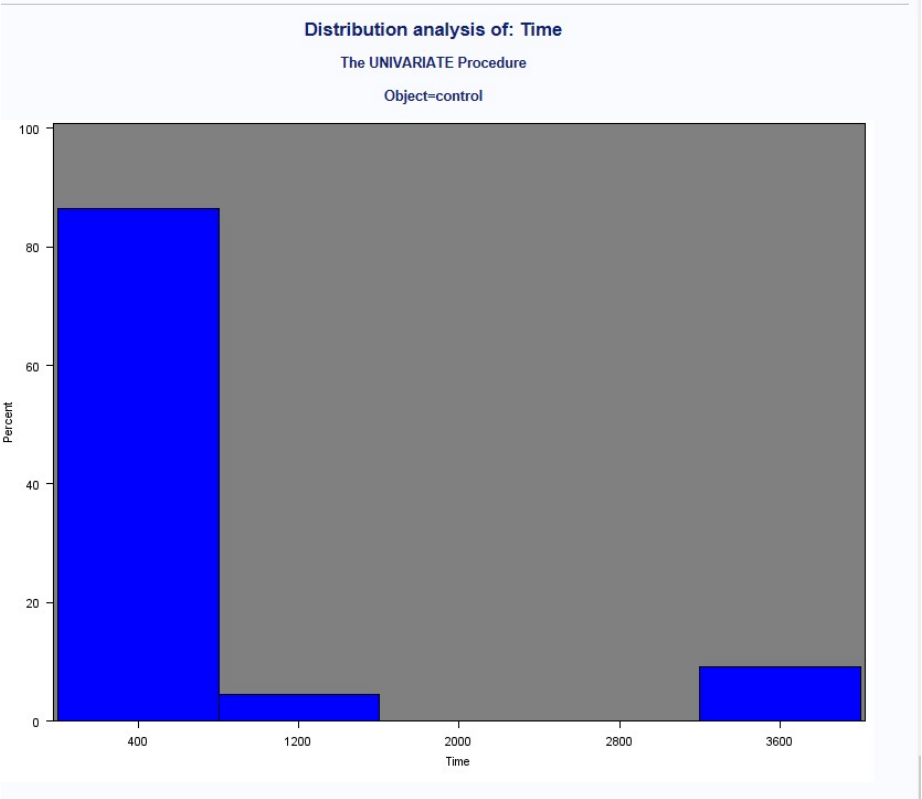





Distribution analysis of: Time
The UNIVARIATE Procedure
Object=Light
Percent
0
10
20
30
40
50
60
70
80
500
1500
2500
3500
Time
Generated by SAS ('Local', X64_10PRO) on October 21, 2021 at 07:25:41 PM


Distribution analysis of: Time
The UNIVARIATE Procedure
Object=Pipe_cleaner
Percent
0
10
20
30
40
50
60
70
400
1200
2000
2800
3600
Time


Distribution analysis of: Time
The UNIVARIATE Procedure
Object=Red_dish
Percent
70
60
50
40
30
20
10
0
125
375
625
875
Time


Distribution analysis of: Time
The UNIVARIATE Procedure
Object=control
Percent
100
80
60
40
20
0
400
1200
2000
2800
3600
Time

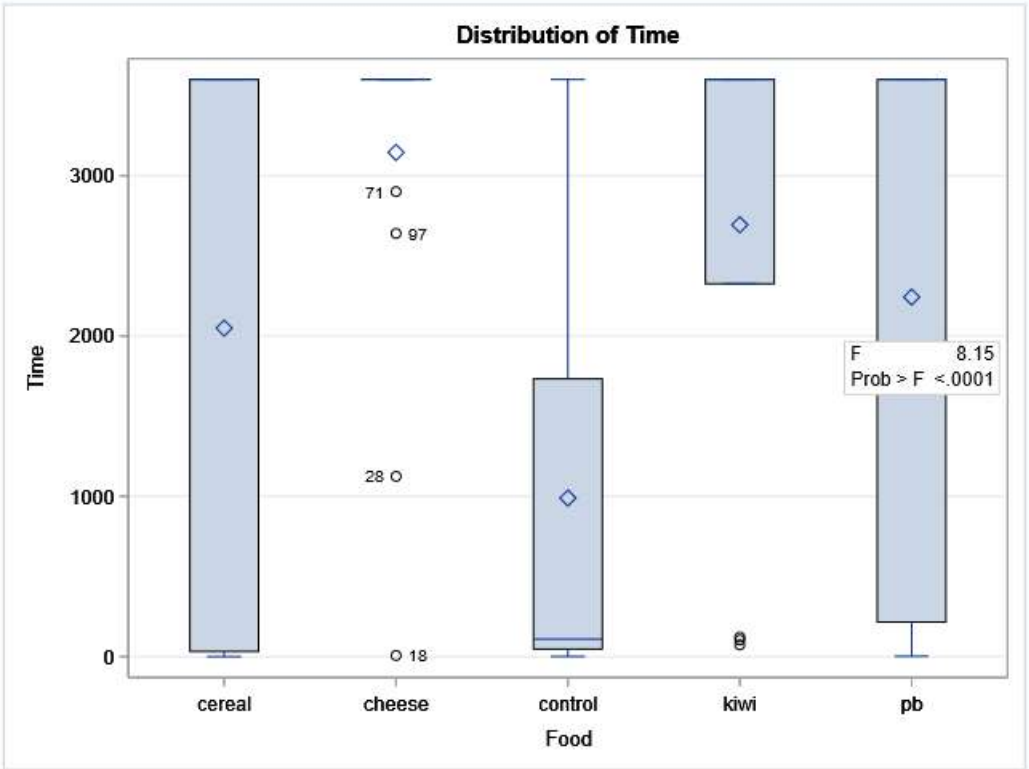
Once again, I used the ANOVA test to compare the mean times to the dish of the various foods from the stacked foods sheet. The means were not found to be equivalent for the different foods used in this experiment. From the data below, you will find the best two foods that differentiate the neophobic birds from the non-neophobic birds are cereal and peanut butter.

Dependent Variable: Time

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	73646999.0	18411749.8	8.15	<.0001
Error	105	237289638.3	2259901.3		
Corrected Total	109	310936637.4			

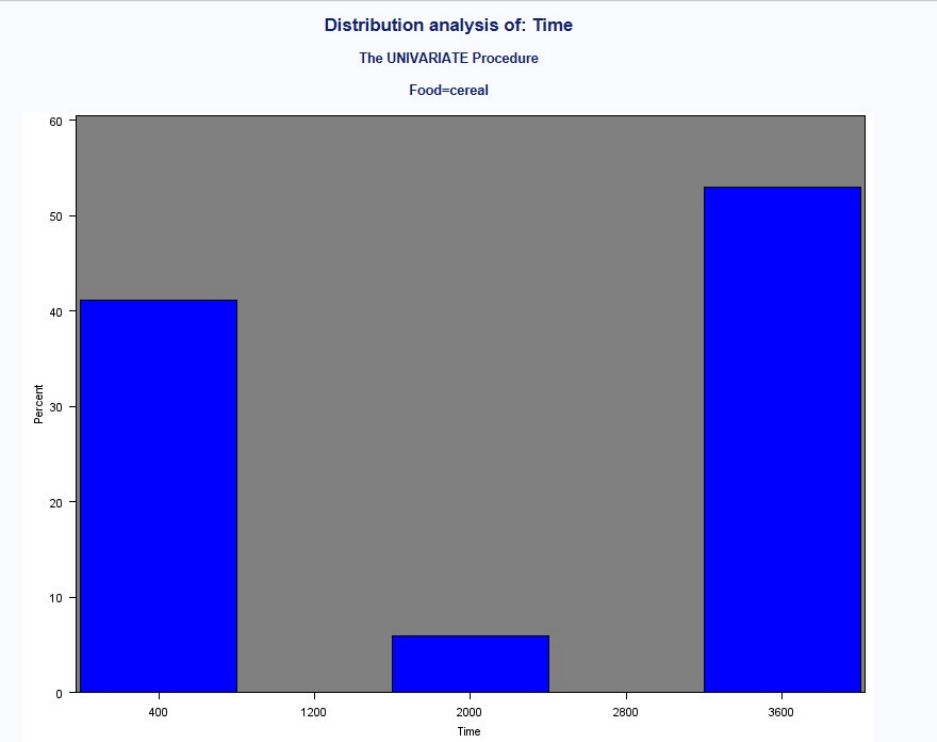
R-Square	Coeff Var	Root MSE	Time Mean
0.236855	78.92414	1503.297	1904.736

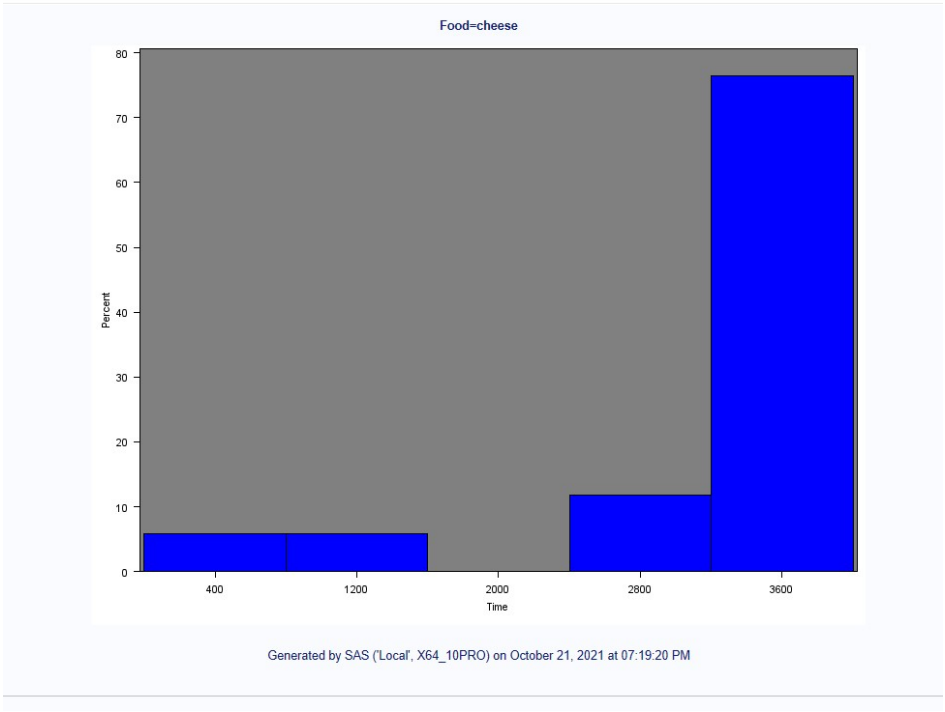
Source	DF	Anova SS	Mean Square	F Value	Pr > F
Food	4	73646999.03	18411749.76	8.15	<.0001

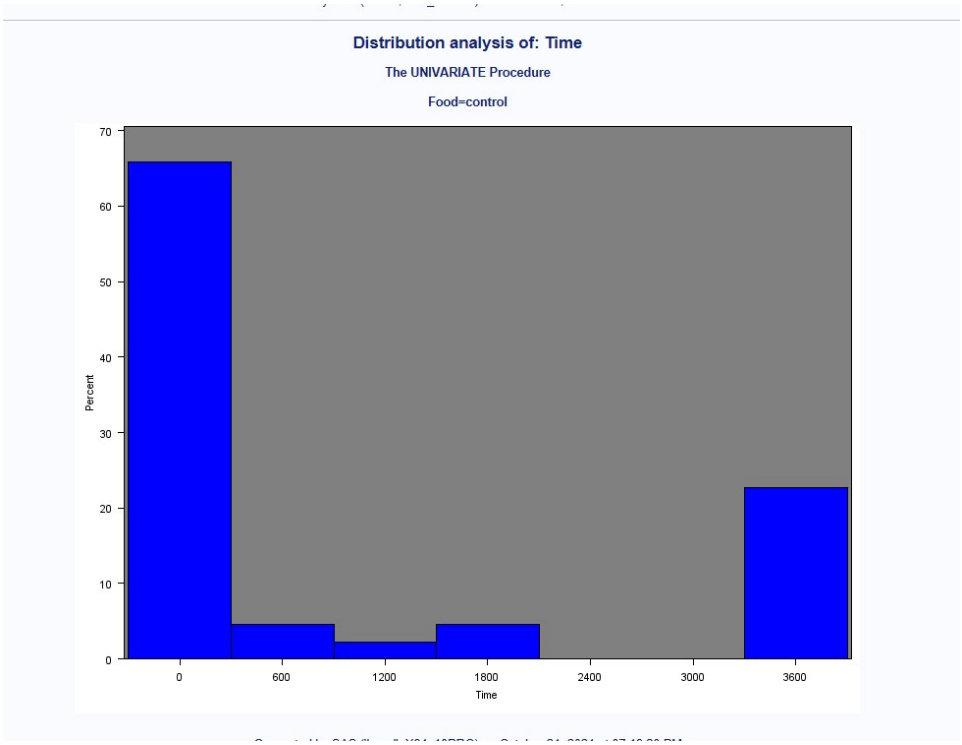


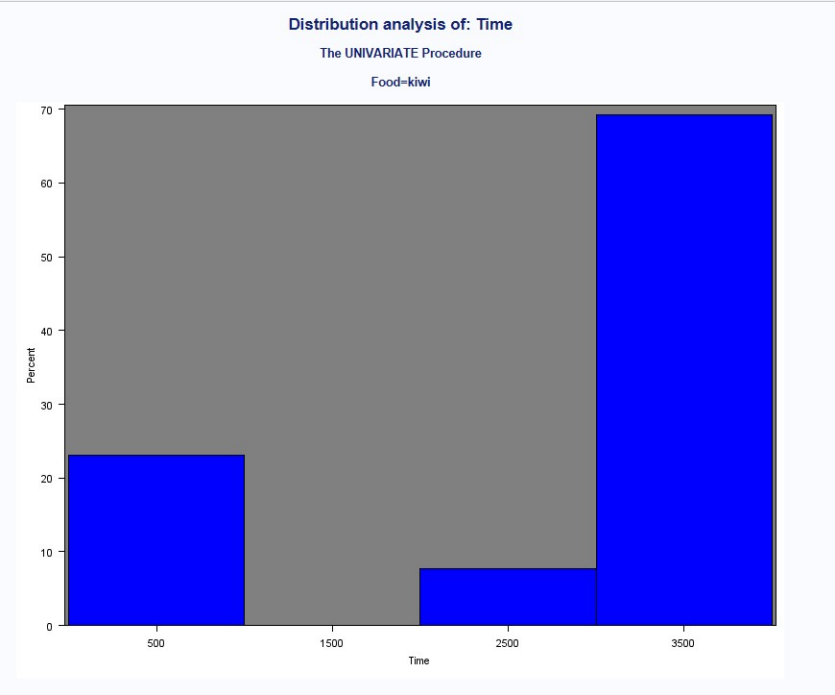
Generated by SAS ('Local' X64_10PRO) on October 21, 2021 at 07:17:14 PM

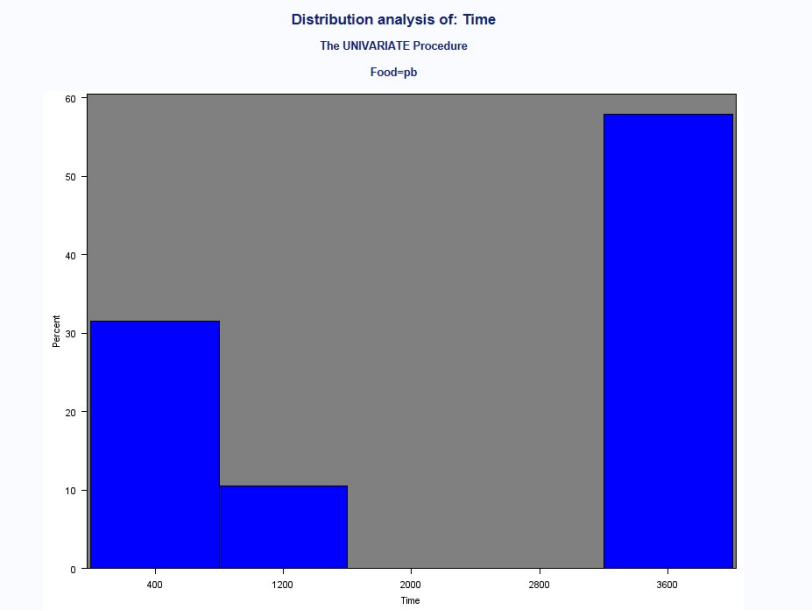
HISTOGRAMS OF FOOD




Food=cheese
Percent
Time
0
10
20
30
40
50
60
70
80
400
1200
2000
2800
3600
Generated by SAS ('Local', X64_10PRO) on October 21, 2021 at 07:19:20 PM


Distribution analysis of: Time
The UNIVARIATE Procedure
Food=control
Percent
Time
0
10
20
30
40
50
60
70
0
600
1200
1800
2400
3000
3600



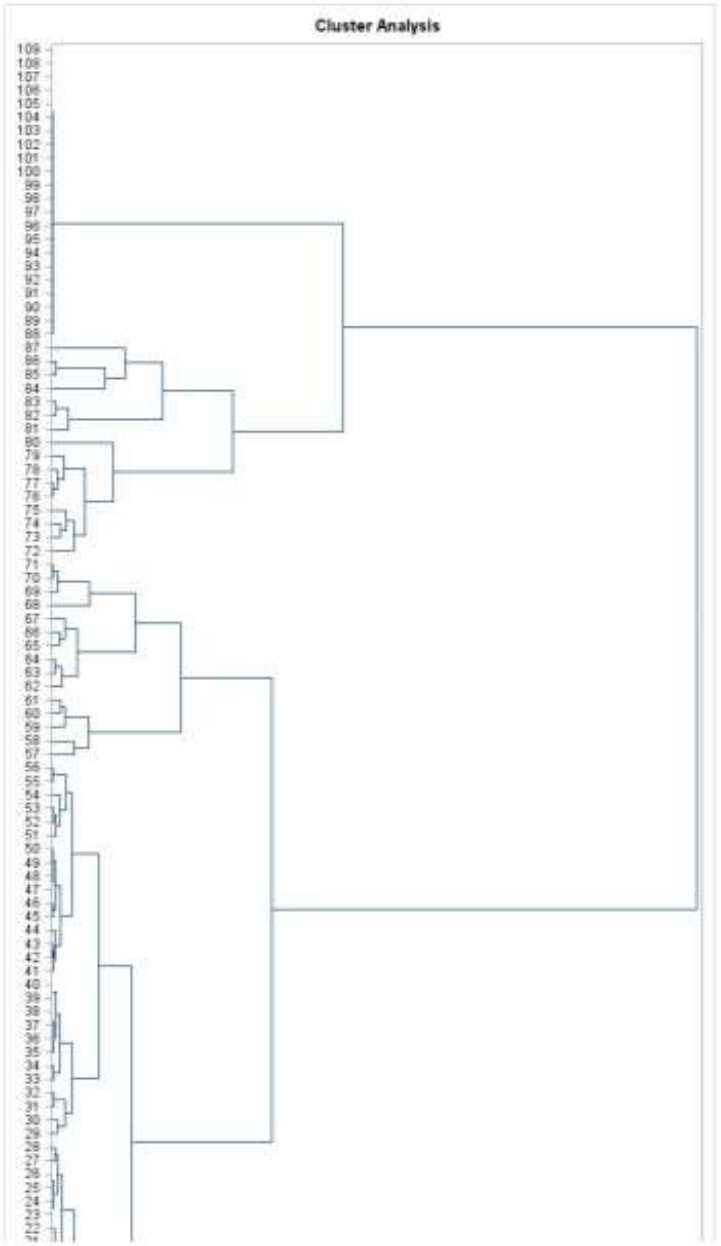


To simplify this experiment and reduce both the noise and the number of variables I would suggest using only a control dish and the egg as an object. Further, I would also recommend using only a control food and either peanut butter or a peanut butter and cereal blended or combo type of meal. An experiment designed as such could potentially reduce the intermediate noise and while still providing better differentiation between neophobic and non-neophobic birds.

To further understand the separation that exists between neophobic and non-neophobic birds I used the Hierarchical Cluster Analysis. From the output provided you find that we do have two groups or two clusters in this experiment. I then used SAS to separate the object feed time in seconds into two groups. From the data given you will find the highest feeding time for non-neophobic birds in cluster 1 is 1582 seconds. The shortest feeding time for cluster two for neophobic birds is 1910. The midpoint between these two clusters is 1746 and this feeding time could be another option for the point of separation between neophobic and non-neophobic birds. Dr. Li provided you with the cluster analysis for the Gaussian finite mixture method. There are about 10 different algorithms that can be used for cluster analysis. Neither method is right or wrong, but they all provide slightly different solutions. Below is additional information on cluster analysis that I found on the internet.

"There are many clustering algorithms to choose from and no single best clustering algorithm for all cases. Instead, it is a good idea to explore a range of clustering algorithms and different configurations for each algorithm. Clustering is a Machine Learning technique that involves the grouping of data points. Given a set of data points, we can use a clustering algorithm to classify each data point into a specific group. In theory, data points that are in the same group should have similar properties and/or features, while data points in different groups should have highly dissimilar properties and/or features. Clustering is a method of unsupervised learning and is a common technique for statistical data analysis used in many fields."

https://machinelearningmastery.com/clustering-algorithms-with-python/


Cluster Analysis

CLUSTER=1

Obs	Object_feeding_time_s
2	10
3	10
4	64
5	64
6	150
7	150
8	4
9	5
10	42
11	43
12	142
13	143
14	332
15	333
16	349
17	350
18	478
19	479
20	334
21	12
22	104
23	106
24	112
25	114
26	79
27	82
28	464
29	467
30	1582
31	1585
32	482
33	437
34	441
35	16
36	75
37	49
38	55
39	486
40	36
41	583
42	590
43	267
44	275
45	301
46	309
47	519
48	528
49	323
50	429
51	92
52	123
53	450
54	164
55	179
56	1148
57	1163
58	507
59	1560
60	205
61	229
62	1214
63	1245
64	549
65	862
66	899
67	1112
68	1283
69	824
70	664
71	757
72	1418

CLUSTER=2

Obs	Object_feeding_time_s
73	3600
74	3600
75	3600
76	3600
77	3600
78	3600
79	3600
80	3600
81	3600
82	3600
83	3600
84	3600
85	3600
86	3600
87	3600
88	3600
89	3600
90	3600
91	3600
92	3600
93	3600
94	3595
95	2090
96	2095
97	2574
98	2591
99	3006
100	3025
101	2116
102	1961
103	1998
104	2152
105	2040
106	2515
107	1910
108	2794
109	2298
110	3246

Bird Neophobia

EMILY CONRAD

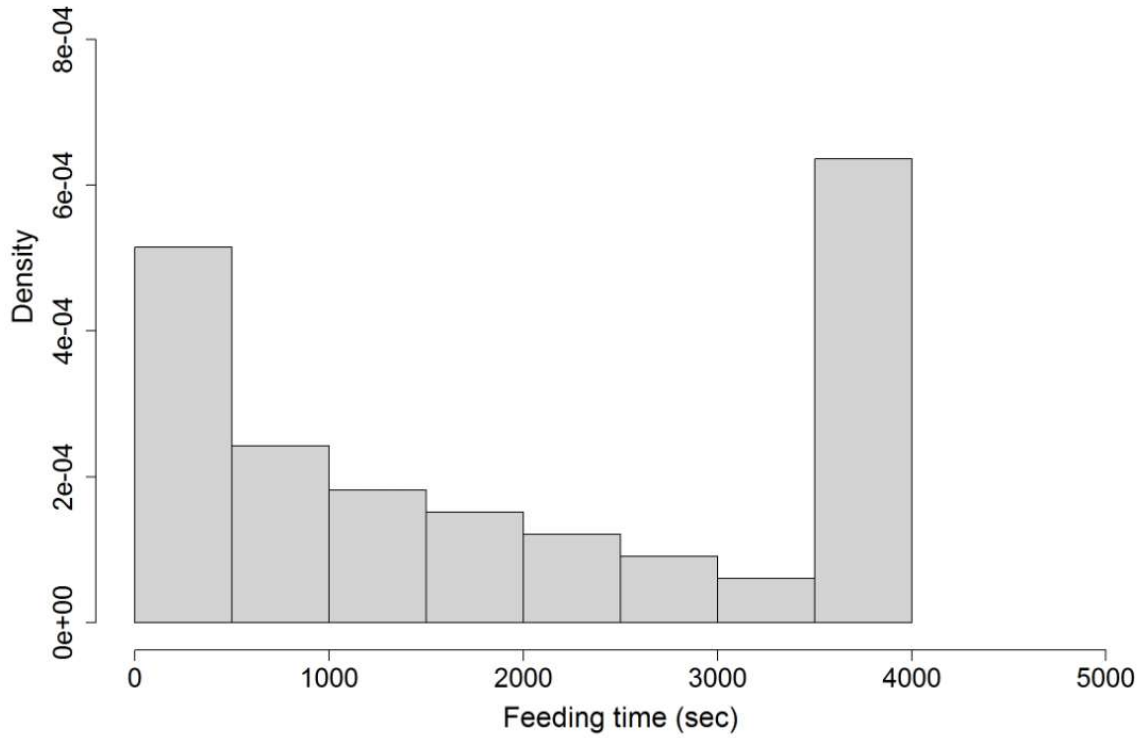
Studying Neophobia in House Sparrows

A research group at LSU is studying neophobia in house sparrows at LSU, which they quantify by measuring latency to approach their food dish after overnight fasting with a novel object and without as a control group. They consistently end up with 1/3 of birds being neophobic, 1/3 being intermediate, and 1/3 non-neophobic. After birds are assigned to groups, they use these groups as predictor variables in other analyses. They recently received a critique about splitting our data into 3 groups, because they could be introducing bias. They are seeking to find a consistent, quantitative way to assign group membership.

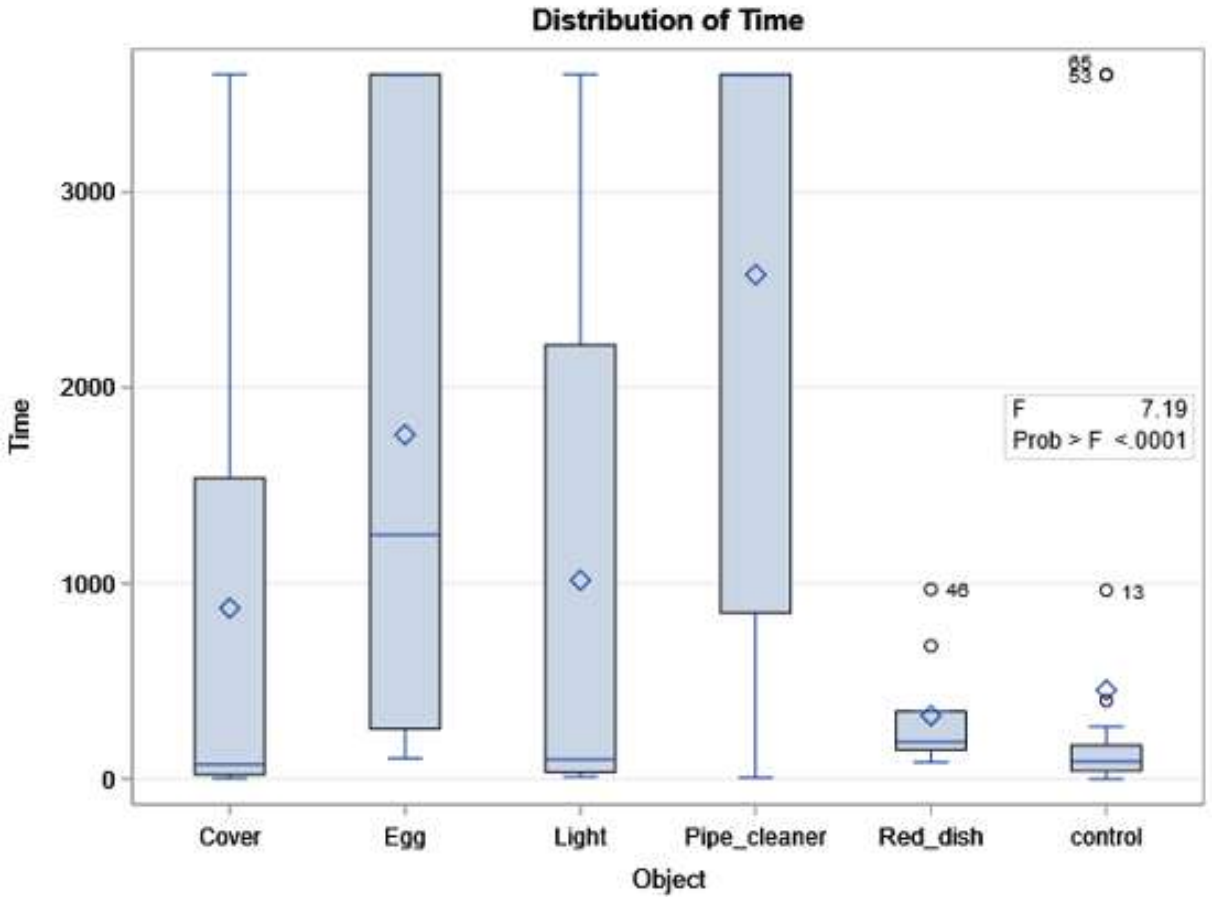
Feeding Times in Birds

Novel Objects Include: Control, Red Dish, Pipe Cleaner, Light, Egg, Cover

Food Variables Include: Control, Kiwi, Peanut Butter, Cereal, Cheese



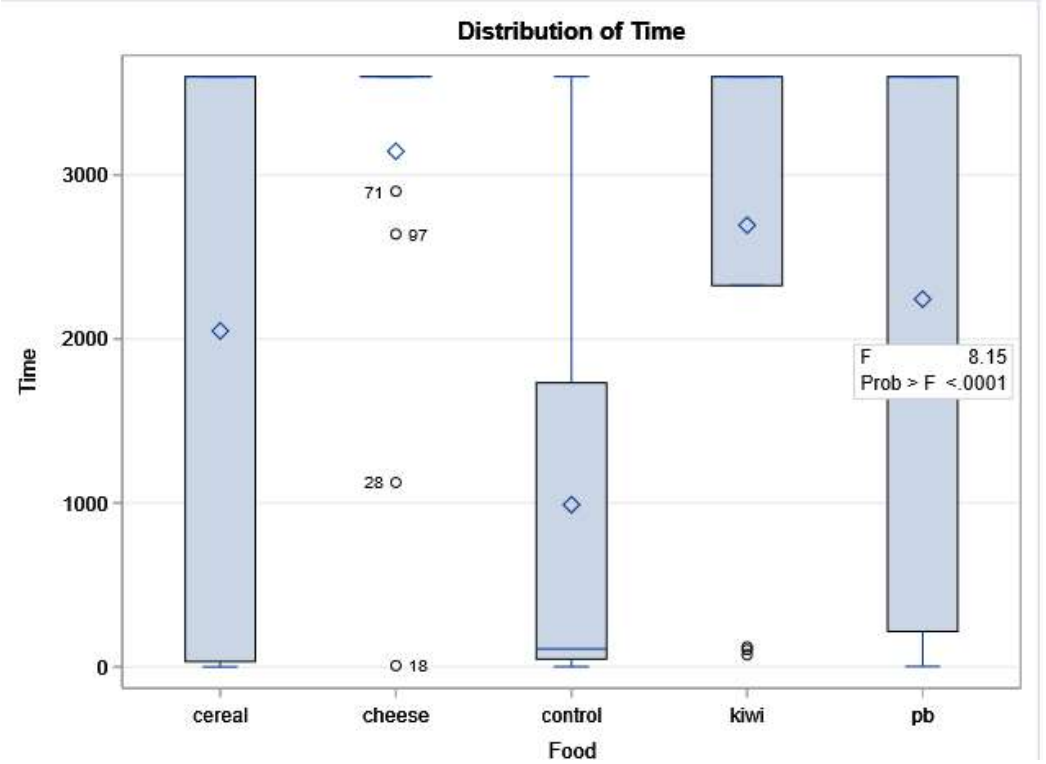
Using the ANOVA to determine if novel objects are equivalent



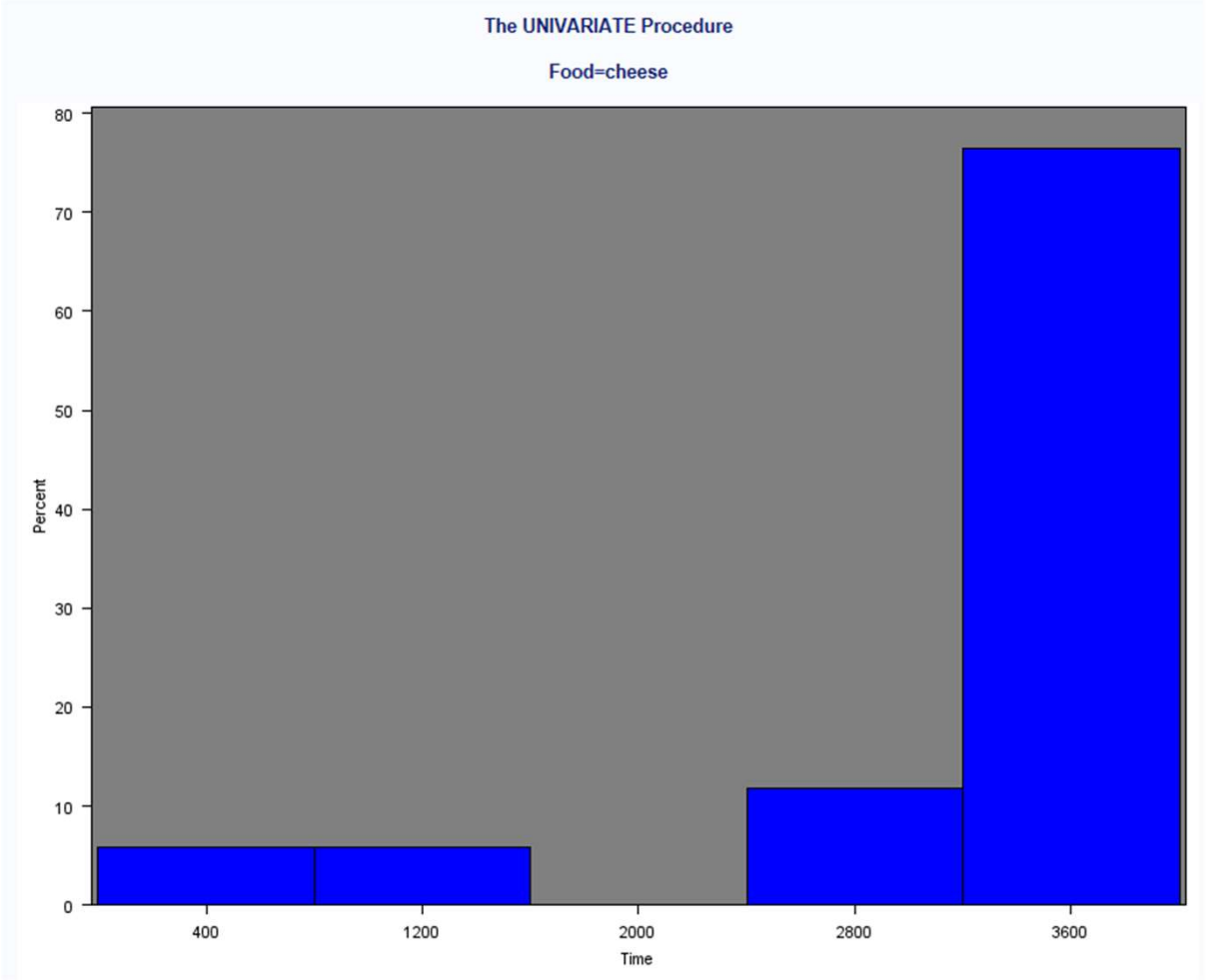
Source	DF	Anova SS	Mean Square	F Value	Pr > F
Object	5	63706607.27	12741321.45	7.19	<.0001

Using the ANOVA to determine if various food selections are equivalent

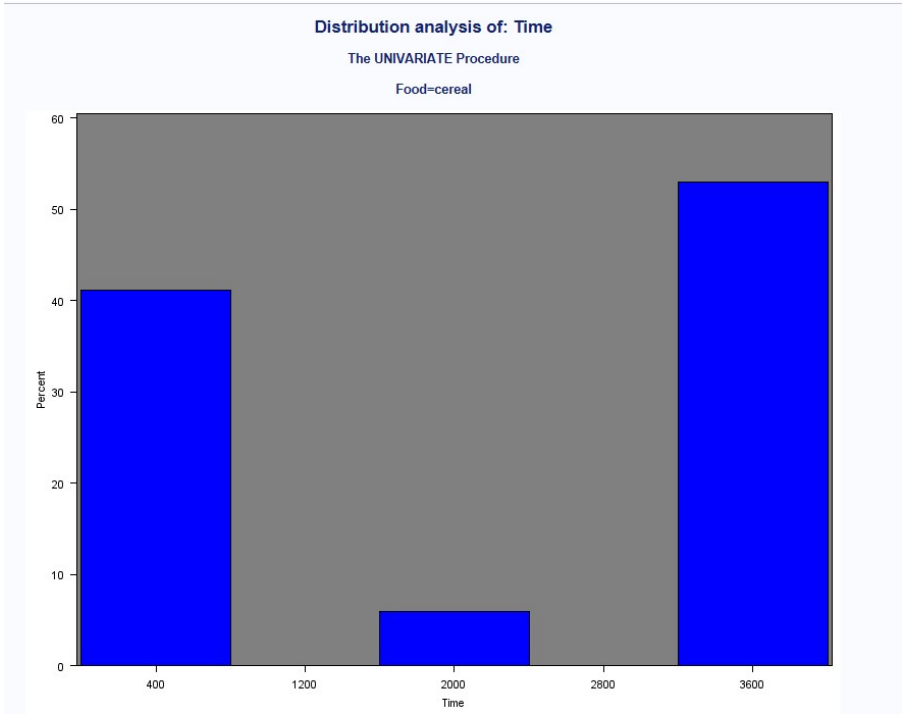
Source	DF	Anova SS	Mean Square	F Value	Pr > F
Food	4	73646999.03	18411749.76	8.15	<.0001

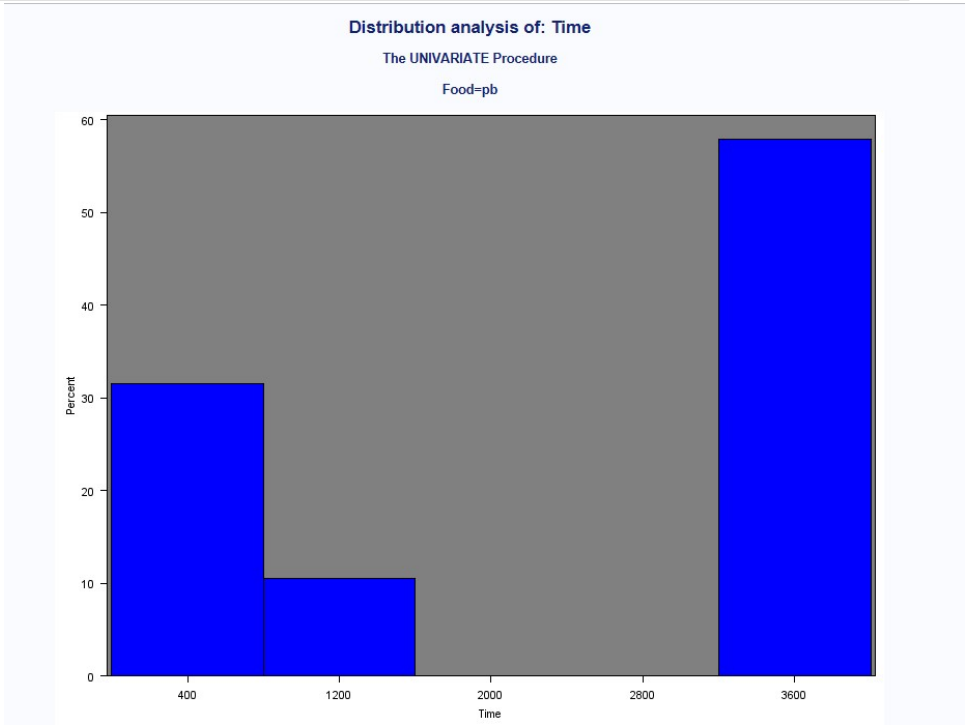


Cheese Histogram



Histograms of Peanut Butter & Cereal





Recommendations

It was recommended to the group studying bird neophobia to design an experiment that reduces variables but encompasses the entire sample space in order to achieve differentiation between neophobic and non-neophobic birds. To remove the noise of the “intermediate” birds I recommended they design the experiment and only use the egg as the novel object and to use either only cereal, only peanut butter, or only a mixture of cereal & peanut butter as the food option.

How do we define a feeding time that separates neophobic and non-neophobic birds?

Cluster Analysis

Cluster analysis, or clustering is an unsupervised machine learning task.

We used clustering to determine how many groups are present in the dataset. After the analysis it further confirmed that only two groups exist. It can also be used to define quantitatively the point at which the groups are different.

There are many types of cluster analysis algorithms.

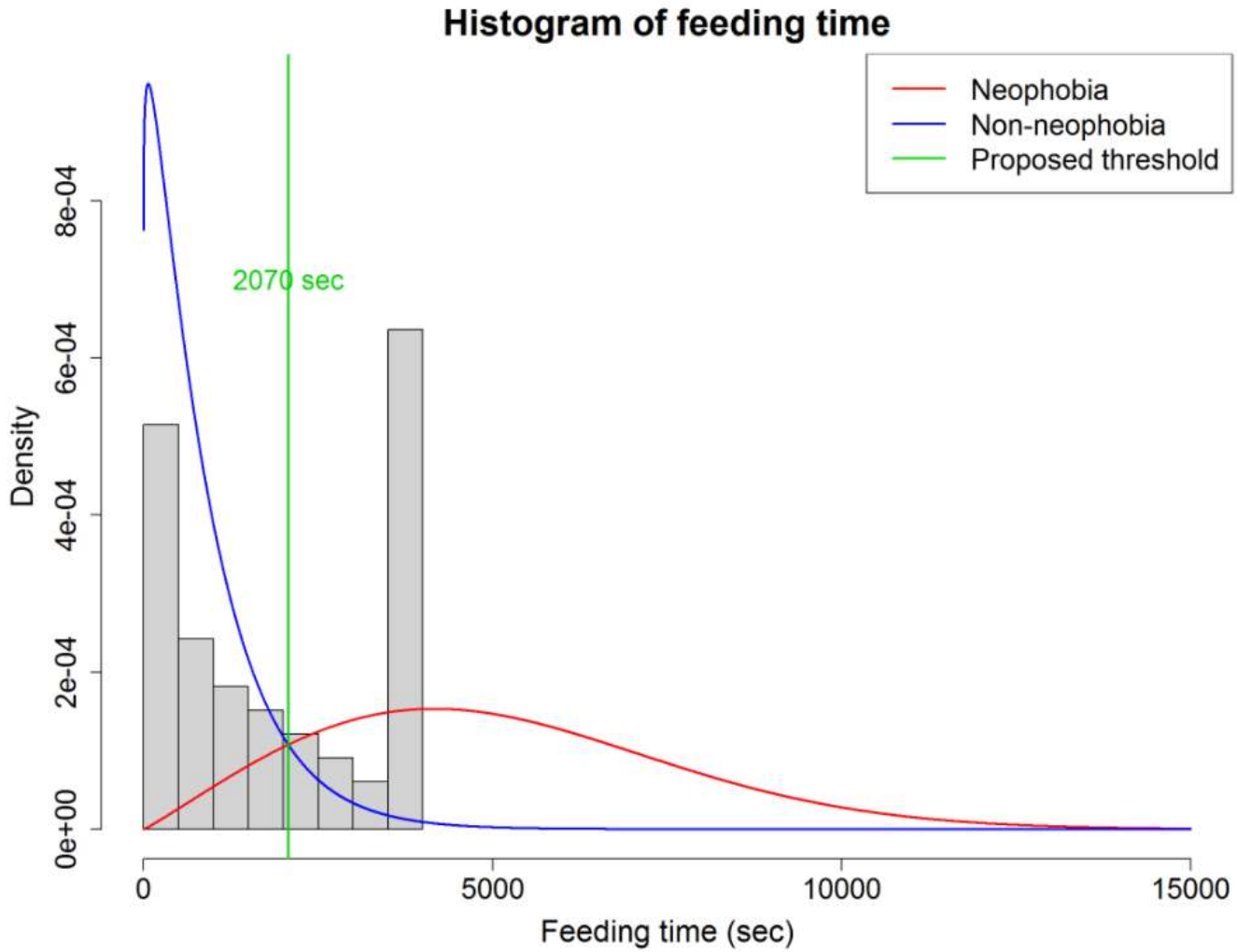
There is no best clustering algorithm, and no easy way to find the best algorithm for your data.

One strategy for determining the best algorithm to use for your data, other than experience, is to use a few algorithms and see which one is the best fit for your dataset and your output.

https://machinelearningmastery.com/clustering-algorithms-with-python/

Dr. Li's Gaussian Finite Mixture Model

- Right Censored Data
- Non-Parametric



Hierarchical Cluster Analysis

Finds the point with the largest separation within dataset.

CLUSTER=1

Obs	Object_feeding_time_s
2	10
3	10
4	64
5	64
6	150
7	150
8	4
9	5
10	42
11	43
12	142
13	143
14	332
15	333
16	349
17	350
18	478
19	479
20	334
21	12
22	104
23	106
24	112
25	114
26	79
27	82
28	464
29	467
30	1582
31	1585
32	482
33	437
34	441
35	16
36	75
37	49
38	55
39	486
40	36
41	583
42	590
43	267
44	275
45	301
46	309
47	619
48	528
49	323
50	429
51	92
52	123
53	450
54	164
55	179
56	1148
57	1163
58	507
59	1560
60	205
61	229
62	1214
63	1245
64	549
65	862
66	899
67	1112
68	1283
69	824
70	664
71	757
72	1418

CLUSTER=2

Obs	Object_feeding_time_s
73	3600
74	3600
75	3600
76	3600
77	3600
78	3600
79	3600
80	3600
81	3600
82	3600
83	3600
84	3600
85	3600
86	3600
87	3600
88	3600
89	3600
90	3600
91	3600
92	3600
93	3600
94	3595
95	2090
96	2095
97	2574
98	2591
99	3008
100	3025
101	2116
102	1961
103	1998
104	2152
105	2040
106	2515
107	1910
108	2794
109	2298
110	3246

Cluster 1
Max = 1582 sec

Cluster 2
Min = 1910 sec

Midpoint = 1746 sec

THANK YOU!