



T.C

MİMAR SİNAN GÜZEL SANATLAR ÜNİVERSİTESİ
FEN-EDEBİYAT FAKÜLTESİ
İSTATİSTİK BÖLÜMÜ

ZEYNEP ECEM GÜNEŞ
20181101017

Verinin linki:

<https://www.kaggle.com/amandam1/breastcancerdataset>

Bu veri seti, tümörlerini çıkarmak için ameliyat olan bir grup meme kanseri hastasından oluşmaktadır. Analizimiz açıklanan çoğu değişkenimizi kullanarak ile gerçekleşecektir

Age – Tanı yaşı

Gender - Cinsiyet

Protein1,2,3,4 - İfade seviyeleri

Tumour stage - Tümör evresi

Histology – Dokubilim (hastalığın dokusal evreleri)

ER status – (Östrojen reseptörlerine sahip olan bir kansere östrojen reseptörü pozitif (veya ER+) adı verilir. Bu, normal meme hücreleri gibi kanser hücrelerinin, büyümelerini destekleyebilecek östrostojenden sinyaller alabileceğini düşündürmektedir.)

PR status – (PR-pozitif: Progesteron reseptörleri olan meme kanserlerine PR-pozitif (veya PR+) kanserler denir. Hormon reseptörü pozitif: Kanser hücresinde yukarıdaki reseptörlerden biri veya her ikisi varsa, hormon pozitif (hormon pozitif veya HR+ olarak da adlandırılır) meme kanseri terimi kullanılabilir.)

HER2 Status – Her2 durumu (HER2-pozitif meme kanseri, insan epidermal büyümeye faktörü reseptörü 2 (HER2) adı verilen bir protein için pozitif test eden bir meme kanseridir. Bu protein kanser hücrelerinin büyümeyi destekler. Her 5 meme kanserinden yaklaşık 1'inde kanser hücreleri, HER2 proteinini yapan genin fazladan kopyalarına sahiptir.)

Surgery_type – Ameliyat tipi

Patient_Status - Hastanın durumu

Normalilik testi

H0: Normal dağılıma uymaktadır.

HA: Normal dağılıma uymamaktadır.

Tests of Normality

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Age	.067	321	.001	.983	321	.001
Protein1	.038	321	.200*	.986	321	.004
Protein2	.059	321	.008	.980	321	.000
Protein3	.077	321	.000	.969	321	.000
Protein4	.052	321	.034	.984	321	.001

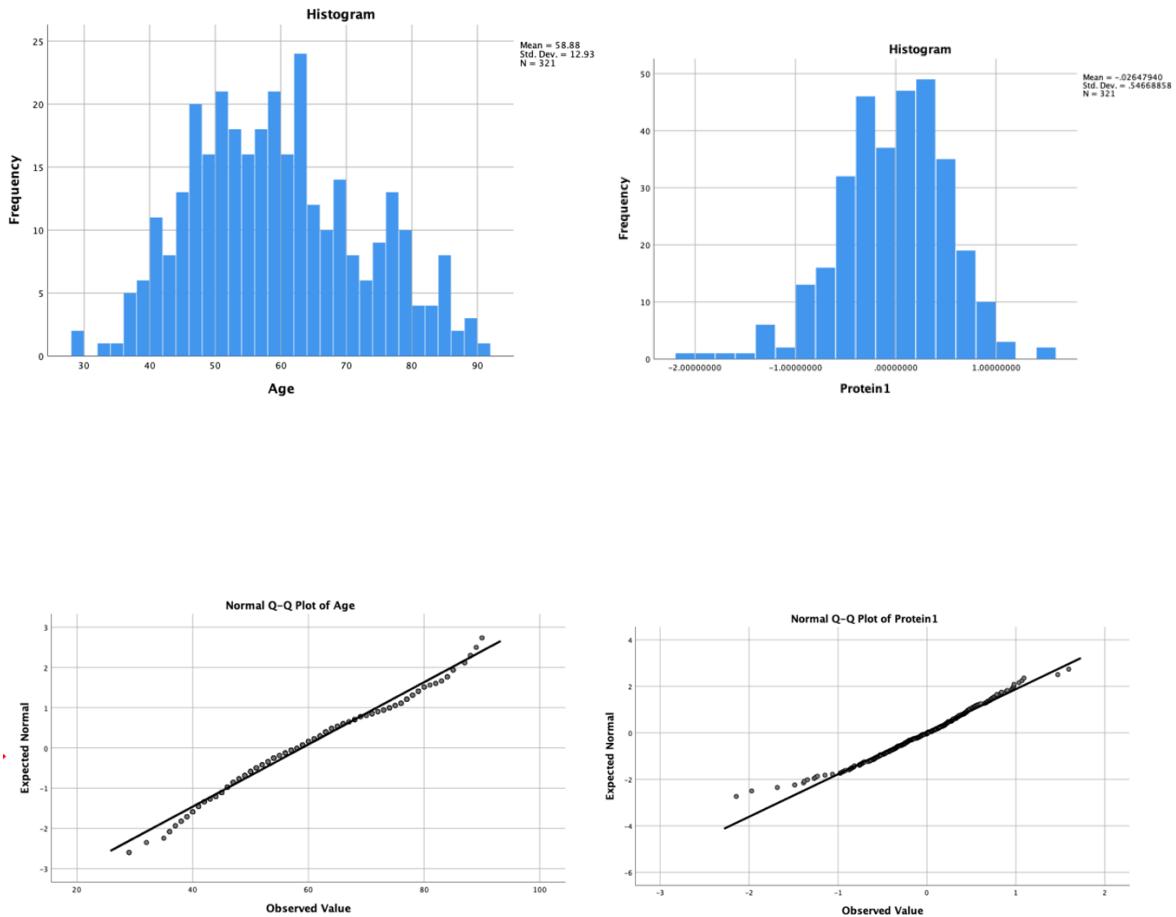
*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Tüm sig değerleri $< 0,05$ olduğu için H0 red yani verimiz normal dağılıma uymamaktadır.

Dönüştüm yaparak normalilik sağlanabilir miyiz bakıyoruz.

Gözlemlediğimiz değişkenlerin grafiklerinin bazılarını analizimize ekliyoruz. Grafikler normal dağılıyor gibi görünse de sigma değerlerimizden dolayı normalleştirilmeye devam ediyoruz.



LN DÖNÜŞÜMÜ İLE NORMALLİK TESTİ

Tests of Normality

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
LN_Age	.151	27	.114	.965	27	.470
LN_Protein1	.197	27	.008	.828	27	.000
LN_Protein2	.201	27	.006	.901	27	.014
LN_Protein3	.226	27	.001	.866	27	.002
LN_Protein4	.171	27	.042	.907	27	.019

a. Lilliefors Significance Correction

LN_Age > 0,05 ama

LN_Protein1 < 0,05

LN_Protein2 < 0,05

LN_Protein3 < 0,05

LN_Protein4 < 0,05 olduğu için normallik sadece age değişkeninde sağlanıyor ve yaptığımız ln dönüşümü normallik için yeterli olmuyor.

Case Processing Summary

	Valid		Cases		Total	
	N	Percent	N	Percent	N	Percent
LN_Age	27	8.4%	294	91.6%	321	100.0%
LN_Protein1	27	8.4%	294	91.6%	321	100.0%
LN_Protein2	27	8.4%	294	91.6%	321	100.0%
LN_Protein3	27	8.4%	294	91.6%	321	100.0%
LN_Protein4	27	8.4%	294	91.6%	321	100.0%

Aynı zamanda sağladığımız dönüşüm ile analizi anlamsızlaşdıracak büyülükte veri kaybı olduğunu gözlemlemiş oluyoruz.

KÖK DÖNÜŞÜMÜ İLE NORMALLİK TESTİ

Tests of Normality

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
KOK_Age	.130	27	.200*	.977	27	.786
KOK_Protein1	.083	27	.200*	.982	27	.903
KOK_Protein2	.147	27	.138	.941	27	.128
KOK_Protein3	.126	27	.200*	.956	27	.297
KOK_Protein4	.141	27	.177	.948	27	.190

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

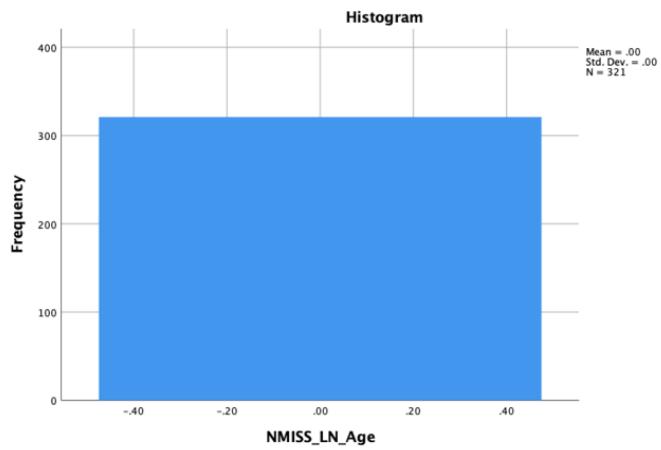
Kök dönüşümü ile test ettiğimiz normalilik için ise tüm değişkenlerimizin significance değerlerinin 0,05'ten büyük olduğunu görsek de ln dönüşümündeki gibi veri kaybı olduğunu gözlemliyoruz.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
KOK_Age	27	8.4%	294	91.6%	321	100.0%
KOK_Protein1	27	8.4%	294	91.6%	321	100.0%
KOK_Protein2	27	8.4%	294	91.6%	321	100.0%
KOK_Protein3	27	8.4%	294	91.6%	321	100.0%
KOK_Protein4	27	8.4%	294	91.6%	321	100.0%

Not:

Ln ve kök dönüşümü yapılrken oluşan veri kaybı, veri doldurma ile çözülmeye çalışılmış fakat paket programlar boşluklara 0.00 atadığı için analiz anlamsızlaşmıştırından analize dönüşümsüz veri ile devam edilmiştir.

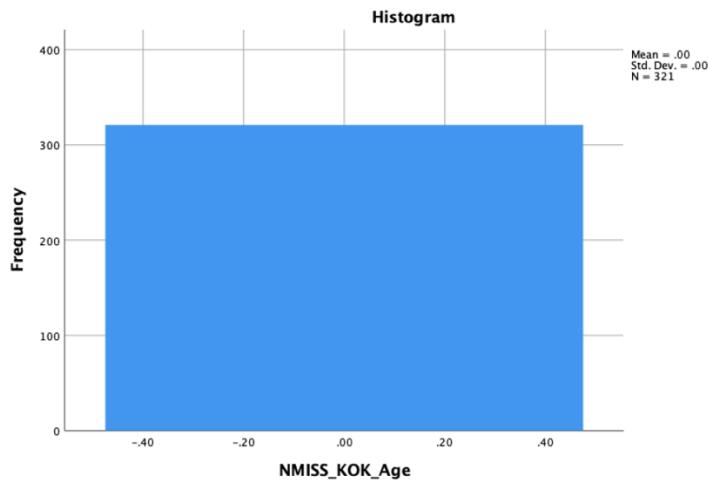


NMISS_LN_Age Stem-and-Leaf Plot

Frequency Stem & Leaf

Stem width: 10.00
Each leaf: 4 case(s)

NMISS_LN_Age	NMISS_LN_Protein1	NMISS_LN_Protein2	NMISS_LN_Protein3	NMISS_LN_Protein4
.00	.00	.00	1.00	1.00
.00	.00	1.00	1.00	.00
.00	.00	.00	.00	1.00
.00	1.00	.00	1.00	.00
.00	1.00	.00	1.00	.00
.00	1.00	1.00	.00	.00
.00	1.00	.00	1.00	.00
.00	1.00	.00	1.00	1.00
.00	1.00	.00	1.00	1.00
.00	.00	.00	1.00	.00
.00	1.00	.00	.00	1.00
.00	.00	.00	1.00	.00
.00	.00	.00	.00	1.00



NMISS_KOK_Age Stem-and-Leaf Plot

Frequency Stem & Leaf

Stem width: 10.00
Each leaf:

Each leaf: 4 case(s)

NMISS_KOK_Age	NMISS_KOK_Protein1	NMISS_KOK_Protein2	NMISS_KOK_Protein3	NMISS_KOK_Protein4
.00	.00	.00	.00	.00
.00	1.00	.00	.00	1.00
.00	.00	.00	1.00	1.00
.00	.00	1.00	1.00	.00
.00	.00	.00	.00	1.00
.00	1.00	.00	1.00	.00
.00	1.00	.00	1.00	.00
.00	1.00	1.00	.00	.00
.00	1.00	.00	1.00	.00
.00	1.00	.00	1.00	.00
.00	1.00	.00	1.00	.00
.00	.00	.00	1.00	.00
.00	1.00	.00	.00	1.00
.00	.00	.00	1.00	.00
.00	.00	.00	.00	1.00

ÇOK DEĞİŞKENLİ NORMALLİK TESTİ

H0: Çok değişkenli normal dağılıma uygundur.

HA: Çok değişkenli normal dağılıma uymamaktadır.

A tibble: 1 × 2

statistic	p.value
<dbl>	<dbl>
0.9746956	1.987585e-05

$1.987585e-05 > 0.05$ olduğu için H0 reddediliyor. Normal şartlarda analize devam edebilmemiz için normalliğin sağlanması gerekiyor fakat tek değişkenli normalliğe bakarken sağlanmayan koşullar burada da sağlanmadığından dolayı uygulama için verimizin çok değişkenli normalliğe uyduğunu varsayıyoruz ve analize devam ediyoruz.

TEK YÖNLÜ MANOVA

MANOVA, birden fazla bağımlı değişkenin bulunduğu deneylerde varyans analizi için kullanılan bir tekniktir. Tek yönlü ve çift yönlü olmak üzere kendi içinde ikiye ayrılır ve bağımlı değişkenlerin bağımsız değişken veya değişkenler üzerindeki etkisini ölçer.

Analiz öncesinde:
Varsayımlar kontrol edilir
Hesaplamalar yapılır
Gruplar arası fark varsa 2'li post hoc'lara bakılır

Between-Subjects Factors

		N
Surgery_type	Lumpectomy	66
	Modified Radical Mastectomy	92
	Other	98
	Simple Mastectomy	65

İncelemekte olduğum veride 4 ameliyat türünden kaç gözlem olduğunu bu çıktıda görebiliyoruz. Birbirlerine yakın olduklarından dolayı dengeli olduklarını ve box testinde sıkıntı çıkarmayacağını öngörerek analize devam edebiliyorum.

Descriptive Statistics

	Surgery_type	Mean	Std. Deviation	N
Protein4	Lumpectomy	.129899130	.617969236	66
	Modified Radical Mastectomy	.033978876	.591202793	92
	Other	-.04447607	.612840699	98
	Simple Mastectomy	-.06941286	.675890225	65
	Total	.008812798	.622491434	321
Protein3	Lumpectomy	-.10978308	.519168350	66
	Modified Radical Mastectomy	.020077202	.664300542	92
	Other	-.14568172	.577184929	98
	Simple Mastectomy	-.15795412	.543386151	65
	Total	-.09327853	.587980743	321
Protein2	Lumpectomy	1.13204979	.958542920	66
	Modified Radical Mastectomy	.799581001	.952171680	92
	Other	1.02498864	.879188883	98
	Simple Mastectomy	.887526769	.814980624	65
	Total	.954563443	.909635098	321

Descriptive statistics çıktısından sadece genel bir yorum yapabiliriz. Örneğin:

Protein4 ($0,129 + 0,6$) ve protein2 ($1,132 + 0,66$) kaynaklı kanser teşhisi konulan hastalar için en çok tercih edilen ameliyat tipi lumpectomy ve

Protein3 ($0,02 + 0,92$) kaynaklı kanser teşhisi konulan hastalar için en çok tercih edilen ameliyat tipi ise modified radical mastectomydir diyebiliriz

Box's Test of Equality of Covariance Matrices^a

Box's M	24.455
F	1.334
df1	18
df2	275198.521
Sig.	.154

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.

Box m testi, gruplar arası varyans kovaryans eşitliği varsayımini test eder.

H0: gruplar arası varyans- kovaryans matrisleri eşittir.

Ha: gruplar arası varyans-kovaryans matrisleri eşit değildir.

Sig değeri alfa değerinden büyük olduğu için ($0,15 > 0,05$), H0 kabul edilir. Yani gruplar arası varyans- kovaryans matrisleri eşittir.

Multivariate Tests

Multivariate Tests ^a							
Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Intercept	Pillai's Trace	.544	125.113 ^b	3.000	315.000	.000	.544
	Wilks' Lambda	.456	125.113 ^b	3.000	315.000	.000	.544
	Hotelling's Trace	1.192	125.113 ^b	3.000	315.000	.000	.544
	Roy's Largest Root	1.192	125.113 ^b	3.000	315.000	.000	.544
Surgery_type	Pillai's Trace	.042	1.500	9.000	951.000	.143	.014
	Wilks' Lambda	.958	1.498	9.000	766.778	.144	.014
	Hotelling's Trace	.043	1.493	9.000	941.000	.146	.014
	Roy's Largest Root	.022	2.288 ^c	3.000	317.000	.078	.021

a. Design: Intercept + Surgery_type

b. Exact statistic

c. The statistic is an upper bound on F that yields a lower bound on the significance level.

Tek yönlü anova yaptığımız için intercept kısmına bakmadıysak surgery type kısmından incelemeye devam ediyoruz.

H0: Gruplar arasında bağımlı değişkenler açısından anlamlı bir farklılık yoktur

(M1 = M2 = M3= M4)

Ha: Gruplar arasında bağımlı değişkenler açısından anlamlı bir farklılık vardır ((M1 = M2 = M3= M4)eşit değildir koy birine

Bu hipotezimizle aslında bağımlı değişkenlerimizden en az birinin farklı olup olmadığına bakıyoruz. Yani burada M1,M2,M3 ve M4 düzeylerimiz diyebiliriz.

Kullandığımız metodlardan:ya da testlerden

Roys Largest root'u verimizde her şey düzenli ise üç değer yoksa vb durumlarda kullanıyoruz. Pillai's Trace'e örneklem sayısı küçükken varyans kovaryans varsayıımı bozulduğunda ve grup değişkenleri çok değişken kullanıyoruz.

Hotelling' trace'i genelde sağlıklı sonuç alamadığımızdan tercih etmiyoruz.

Wilk's Lambda'yı temel varsayımlar (normallik, ortak varyans-kovaryans matrisinin eşitliği) sağlanıyorsa özellikle de risksiz olduğu için kullanıyoruz.

İncelemekte olduğumuz verimizde de willks lambda kullanmayı tercih ediyoruz. Bu noktada hangi significant değerine bakacağımızı bilmek çok önemli. Willksin karşısındakine bakıyoruz ve 0,144 olduğunu görüyoruz

$0,144 > 0,05$ olduğu için H_0 kabul yani Gruplar arasında bağımlı değişkenler açısından anlamlı bir farklılık yoktur diyebiliriz.

Eğer reddetseydik post hoc çıktımız aşağıdaki gibi olmazdı ve bu kez de 2'li ANOVA ile hangi değişkenler arasında farklılık olduğunu bulmamız gerekiirdi.

Çıktıda gördüğümüz partial eta squared değerimiz bizim etki genişliğimizdir. Bu genişliğe bakarak kullandığımız teste yani Willks Lambdaya göre bağımlı değişkenlerdeki değişimin % 0,014'ü grup değişkenleri tarafından açıklanmaktadır yorumunu yapabiliriz.

Levene's Test of Equality of Error Variances ^a					
		Levene Statistic	df1	df2	Sig.
Protein4	Based on Mean	.543	3	317	.653
	Based on Median	.537	3	317	.657
	Based on Median and with adjusted df	.537	3	315.609	.657
	Based on trimmed mean	.548	3	317	.650
Protein3	Based on Mean	1.514	3	317	.211
	Based on Median	1.436	3	317	.232
	Based on Median and with adjusted df	1.436	3	309.851	.232
	Based on trimmed mean	1.499	3	317	.215
Protein2	Based on Mean	1.346	3	317	.259
	Based on Median	1.374	3	317	.251
	Based on Median and with adjusted df	1.374	3	314.385	.251
	Based on trimmed mean	1.356	3	317	.256

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + Surgery_type

Bu çıktımızda gruplar arası varyansın eşitliğini test ediyoruz. Bağımlı değişkenimiz için tek tek varyans homojenliğine bakıyoruz. Eğer varyans eşitliği sağlanmazsa dönüşüm yapıyoruz.

H0: Gruplar arası varyanslar eşittir.

Ha: Gruplar arası varyanslar eşit değildir.

Hipotezimizi üç protein için de ayrı ayrı significant değerleri ile test edelim.

Protein 4 → ($0,05 < 0,653$) H0 kabul, gruplar arası varyanslar eşittir.

Protein 3 → ($0,05 < 0,211$) H0 kabul, gruplar arası varyanslar eşittir.

Protein 2 → ($0,05 < 0,259$) H0 kabul, gruplar arası varyanslar eşittir.

H0 hipotezini kabul ettiğimiz için TUKEY alanını yorumlayacağız. Reddetseydik tamhane alanını yorumlamamız gereklidir.

Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	Protein4	1.702 ^a	3	.567	1.471	.222	.014
	Protein3	1.741 ^b	3	.580	1.690	.169	.016
	Protein2	5.067 ^c	3	1.689	2.062	.105	.019
Intercept	Protein4	.048	1	.048	.126	.723	.000
	Protein3	2.998	1	2.998	8.727	.003	.027
	Protein2	286.331	1	286.331	349.490	.000	.524
Surgery_type	Protein4	1.702	3	.567	1.471	.222	.014
	Protein3	1.741	3	.580	1.690	.169	.016
	Protein2	5.067	3	1.689	2.062	.105	.019
Error	Protein4	122.297	317	.386			
	Protein3	108.890	317	.344			
	Protein2	259.712	317	.819			
Total	Protein4	124.024	321				
	Protein3	113.424	321				
	Protein2	557.272	321				
Corrected Total	Protein4	123.999	320				
	Protein3	110.631	320				
	Protein2	264.780	320				

a. R Squared = .014 (Adjusted R Squared = .004)

b. R Squared = .016 (Adjusted R Squared = .006)

c. R Squared = .019 (Adjusted R Squared = .010)

Bu çıktımız bize tek değişkenli ANOVA sonuçları sunuyor. Yine bu çıktımızda da grup değişkenimizin olduğu kısmı ele alıyoruz.

H0: Bağımlı değişken açısından gruplar arasında anlamlı bir fark yoktur.

Ha: Bağımlı değişken açısından gruplar arasında anlamlı bir fark vardır.

Hipotezimizi test ederken surgery type değişkenimizdeki 3 düzey için de (protein2,3,4) ayrı ayrı sonuçlara bakıyoruz. Yani hipotezi üçü için de ayrı ayrı deniyoruz diyebiliriz.

Protein 1: ($0,222 > 0,05$) → H0 kabul, bağımlı değişken açısından gruplar arasında anlamlı bir fark yoktur.

Protein 2: ($0,169 > 0,05$) → H0 kabul, bağımlı değişken açısından gruplar arasında anlamlı bir fark yoktur.

Protein 3: ($0,105 > 0,05$) → H0 kabul, bağımlı değişken açısından gruplar arasında anlamlı bir fark yoktur.

H_0 'ları kabul ettiğimiz için, protein 2, protein 3 ve protein 4 değişkenlerinin ortalamaları surgery type'a göre istatistiksel olarak anlamlı bir fark göstermemektedir kararına varabiliriz.

Bu çıktımda da Multivariate'deki gibi etki genişliği görüyoruz. Bu değerleri yorumlayacak olursak;

- Protein 2'deki değişimin %19'u grup değişkeni olan surgery type tarafından açıklanmaktadır.
 - Protein 3 'deki değişimin %16'sı grup değişkeni olan surgery type tarafından açıklanmaktadır.
 - Protein 4' deki değişimin %14'ü grup değişkeni olan surgery type tarafından açıklanmaktadır.
- diyebiliriz.

Post Hoc Tests

Surgery_type

		Multiple Comparisons						
Dependent Variable	(I) Surgery_type	(J) Surgery_type	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval		
						Lower Bound	Upper Bound	
Protein4	Tukey HSD	Lumpectomy	Modified Radical Mastectomy	.095920254	.100193653	.774	-.16285089	.354691393
			Other	.174375204	.098904139	.293	-.08106550	.429815904
			Simple Mastectomy	.199311985	.108538684	.258	-.08101195	.479635919
	Modified Radical Mastectomy	Lumpectomy	-.09592025	.100193653	.774	-.35469139	.162850886	
		Other	.078454951	.090166954	.820	-.15442013	.311330035	
		Simple Mastectomy	.103391731	.100641426	.734	-.15653588	.363319340	
	Other	Lumpectomy	-.17437520	.098904139	.293	-.42981590	.081065496	
		Modified Radical Mastectomy	-.07845495	.090166954	.820	-.31133004	.154420134	
		Simple Mastectomy	.024936781	.099357724	.994	-.23167540	.281548959	
	Simple Mastectomy	Lumpectomy	-.19931199	.108538684	.258	-.47963592	.081011948	
		Modified Radical Mastectomy	-.10339173	.100641426	.734	-.36331934	.156535877	
		Other	-.02493678	.099357724	.994	-.28154896	.231675397	
Tamhane	Lumpectomy	Modified Radical Mastectomy	.095920254	.097904493	.909	-.16546255	.357303055	
		Other	.174375204	.098074136	.384	-.08739310	.436143507	
		Simple Mastectomy	.199311985	.113200124	.396	-.10321283	.501836800	
	Modified Radical Mastectomy	Lumpectomy	-.09592025	.097904493	.909	-.35730305	.165462548	
		Other	.078454951	.087358591	.938	-.15383626	.310746161	
		Simple Mastectomy	.103391731	.104054098	.903	-.17474619	.381529653	
	Other	Lumpectomy	-.17437520	.098074136	.384	-.43614351	.087393098	
		Modified Radical Mastectomy	-.07845495	.087358591	.938	-.31074616	.153836260	
		Simple Mastectomy	.024936781	.104213731	1.000	-.25356659	.303440156	
	Simple Mastectomy	Lumpectomy	-.19931199	.113200124	.396	-.50183680	.103212829	
		Modified Radical Mastectomy	-.10339173	.104054098	.903	-.38152965	.174746190	
		Other	-.02493678	.104213731	1.000	-.30344016	.253566594	

Protein3	Tukey HSD	Lumpectomy	Modified Radical Mastectomy	-.12986028	.094542354	.517	-.37403576	.114315194
			Other	.035898640	.093325574	.981	-.20513425	.276931524
			Simple Mastectomy	.048171041	.102416694	.966	-.21634157	.312683652
Modified Radical Mastectomy			Lumpectomy	.129860281	.094542354	.517	-.11431519	.374035756
			Other	.165758921	.085081199	.210	-.05398113	.385498976
			Simple Mastectomy	.178031322	.094964872	.241	-.06723539	.423298037
Other			Lumpectomy	-.03589864	.093325574	.981	-.27693152	.205134245
			Modified Radical Mastectomy	-.16575892	.085081199	.210	-.38549898	.053981134
			Simple Mastectomy	.012272402	.093753575	.999	-.22986589	.254410689
Simple Mastectomy			Lumpectomy	-.04817104	.102416694	.966	-.31268365	.216341569
			Modified Radical Mastectomy	-.17803132	.094964872	.241	-.42329804	.067235393
			Other	-.01227240	.093753575	.999	-.25441069	.229865886
Tamhane			Modified Radical Mastectomy	-.12986028	.094236736	.673	-.38101122	.121290656
			Other	.035898640	.086506000	.999	-.19476448	.266561763
			Simple Mastectomy	.048171041	.092878780	.996	-.20001745	.296359534
Modified Radical Mastectomy			Lumpectomy	.129860281	.094236736	.673	-.12129066	.381011218
			Other	.165758921	.090532313	.348	-.07507444	.406592283
			Simple Mastectomy	.178031322	.096639948	.342	-.07959454	.435657188
Other			Lumpectomy	-.03589864	.086506000	.999	-.26656176	.194764484
			Modified Radical Mastectomy	-.16575892	.090532313	.348	-.40659228	.075074442
			Simple Mastectomy	.012272402	.089117928	1.000	-.22549398	.250038784
Simple Mastectomy			Lumpectomy	-.04817104	.092878780	.996	-.29635953	.200017452
			Modified Radical Mastectomy	-.17803132	.096639948	.342	-.43565719	.079594544
			Other	-.01227240	.089117928	1.000	-.25003878	.225493980

Protein2	Tukey HSD	Lumpectomy	Modified Radical Mastectomy	.332468787	.146008936	.106	-.04462994	.709567513
			Other	.107061145	.144129770	.880	-.26518424	.479306529
			Simple Mastectomy	.244523019	.158169877	.411	-.16398389	.653029926
Modified Radical Mastectomy			Lumpectomy	-.33246879	.146008936	.106	-.70956751	.044629939
			Other	-.22540764	.131397355	.317	-.56476889	.113953609
			Simple Mastectomy	-.08794577	.146661462	.932	-.46672978	.290838242
Other			Lumpectomy	-.10706115	.144129770	.880	-.47930653	.265184239
			Modified Radical Mastectomy	.225407642	.131397355	.317	-.11395361	.564768893
			Simple Mastectomy	.137461874	.144790764	.778	-.23649067	.511414415
Simple Mastectomy			Lumpectomy	-.24452302	.158169877	.411	-.65302993	.163983889
			Modified Radical Mastectomy	.087945768	.146661462	.932	-.29083824	.466729778
			Other	-.13746187	.144790764	.778	-.51141442	.236490668
Tamhane			Modified Radical Mastectomy	.332468787	.154194567	.181	-.07905530	.743992875
			Other	.107061145	.147677897	.978	-.28743081	.501553098
			Simple Mastectomy	.244523019	.155369370	.529	-.17076446	.659810493
Modified Radical Mastectomy			Lumpectomy	-.33246879	.154194567	.181	-.74399287	.079055301
			Other	-.22540764	.133199716	.441	-.57966689	.128851611
			Simple Mastectomy	-.08794577	.141679371	.990	-.46571660	.289825069
Other			Lumpectomy	-.10706115	.147677897	.978	-.50155310	.287430808
			Modified Radical Mastectomy	.225407642	.133199716	.441	-.12885161	.579666894
			Simple Mastectomy	.137461874	.134557946	.891	-.22148992	.496413671
Simple Mastectomy			Lumpectomy	-.24452302	.155369370	.529	-.65981049	.170764456
			Modified Radical Mastectomy	.087945768	.141679371	.990	-.28982507	.465716605
			Other	-.13746187	.134557946	.891	-.49641367	.221489923

Based on observed means.
The error term is Mean Square(Error) = .819.

Bu çıktıyı, farklılığın hangi grup ya da grplardan kaynaklandığını bulmak için kullanıyoruz. Mean difference kolonunda yıldızlılar bize hangi grplar arasında farklar olduğunu söylüyor. Yani significant değerlerine bakmamıza gerek yok. Fark olanları seçmenin bir diğer yöntemi ise güven aralıklarına bakmak. Eğer güven aralığımızın (lower bound ve upper bound) değerleri (+,+) ve (-,-) ise aralarında kesinlikle fark var diyebiliyoruz. (-,+) ve (+,-) olduğu durumlarda ise aralarında fark yoktur olarak yorumluyoruz.

Levene's çıktısında verdiği karardan sonra TUKEY için incelememize devam ediyoruz.

İncelediğimiz üç protein tipinin de TUKEY kısmında yıldız yok. Dikkatli bakıldığından da lower ve upper boundların hep farklı işarette olduğunu görebiliyoruz. Güven aralıklarından bu çıkarımı yapıyor olmamızın mantığı da, lower ve upper bound değerlerimiz (-,+) iken arada 0 değeri bulunduğuundan sıfırlanıyor oluş.

Homogeneous Subsets

Protein2			Protein3				
		Subset			Subset		
	Surgery_type	N		Surgery_type	N		
Tukey HSD ^{a,b,c}	Modified Radical Mastectomy	92	.799581001	Tukey HSD ^{a,b,c}	Simple Mastectomy	65	-.15795412
	Simple Mastectomy	65	.887526769		Other	98	-.14568172
	Other	98	1.02498864		Lumpectomy	66	-.10978308
	Lumpectomy	66	1.13204979		Modified Radical Mastectomy	92	.020077202
	Sig.		.103		Sig.		.234

Means for groups in homogeneous subsets are displayed.
Based on observed means.
The error term is Mean Square(Error) = .819.
 a. Uses Harmonic Mean Sample Size = 77.505.
 b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.
 c. Alpha = .05.

Means for groups in homogeneous subsets are displayed.
Based on observed means.
The error term is Mean Square(Error) = .344.
 a. Uses Harmonic Mean Sample Size = 77.505.
 b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.
 c. Alpha = .05.

Protein4

Subset			
	Surgery_type	N	1
Tukey HSD ^{a,b,c}	Simple Mastectomy	65	-.06941286
	Other	98	-.04447607
	Modified Radical Mastectomy	92	.033978876
	Lumpectomy	66	.129899130
	Sig.		.191

Means for groups in homogeneous subsets are displayed.
Based on observed means.
The error term is Mean Square(Error) = .386.

- a. Uses Harmonic Mean Sample Size = 77.505.
- b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.
- c. Alpha = .05.

Homogeneous subsets çıktısı ise bize hangilerinin birlikte bir küme oluşturduğunu sunuyor. Post hoc'un bir özeti de diyebiliriz.

Her bir surgery type'ı alt alta aynı tabloya yazdırdığı için de aralarında fark olmadığını kontrol etmiş olduk diyebiliriz. Eğer aralarında fark olsaydı yan yana farklı tablolarda olurlardı.

ÇİFT YÖNLÜ MANOVA

Between-Subjects Factors

		N
Surgery_type	Lumpectomy	66
	Modified Radical Mastectomy	92
	Other	98
Tumour_Stage	Simple Mastectomy	65
	I	61
	II	182
	III	78

Bu analizimizde bir diğer grup değişkeni olarak üç düzeyi olan tumorun aşamasını yani tumor stage değişkenimizi seçiyoruz. Analizin daha sağlıklı sonuçlara ulaşabilmesi için aslında gözlem değerlerimizin birbirine daha yakın olması gerekiyor fakat günlük hayatı da karşılaşacağımız verilerde de aynı durumu göreceğimiz için bu analize de devam etmekten çekinmiyoruz.

Box's Test of Equality of Covariance Matrices^a

Box's M	79.471
F	1.104
df1	66
df2	10717.080
Sig.	.264

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.

a. Design:
Intercept +
Surgery_type
+
Tumour_Stage
+
Surgery_type *
Tumour_Stage

box M testi, gruplar arası varyans kovaryans eşitliği varsayımini test eder.

H0: gruplar arası varyans- kovaryans matrisleri eşittir.

Ha: gruplar arası varyans-kovaryans matrisleri eşit değildir.

Sig değeri alfa değerinden büyük olduğu için ($0,26 > 0,05$), H0 kabul edilir. Yani gruplar arası varyans- kovaryans matrisleri eşittir.

MULTIVARIATE TEST

Multivariate Tests^a

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Intercept	Pillai's Trace	.444	81.701 ^b	3.000	307.000	.000	.444
	Wilks' Lambda	.556	81.701 ^b	3.000	307.000	.000	.444
	Hotelling's Trace	.798	81.701 ^b	3.000	307.000	.000	.444
	Roy's Largest Root	.798	81.701 ^b	3.000	307.000	.000	.444
Surgery_type	Pillai's Trace	.053	1.844	9.000	927.000	.057	.018
	Wilks' Lambda	.948	1.856	9.000	747.308	.055	.018
	Hotelling's Trace	.055	1.863	9.000	917.000	.054	.018
	Roy's Largest Root	.046	4.759 ^c	3.000	309.000	.003	.044
Tumour_Stage	Pillai's Trace	.005	.262	6.000	616.000	.954	.003
	Wilks' Lambda	.995	.262 ^b	6.000	614.000	.955	.003
	Hotelling's Trace	.005	.261	6.000	612.000	.955	.003
	Roy's Largest Root	.003	.326 ^c	3.000	308.000	.807	.003
Surgery_type * Tumour_Stage	Pillai's Trace	.099	1.763	18.000	927.000	.025	.033
	Wilks' Lambda	.902	1.784	18.000	868.812	.023	.034
	Hotelling's Trace	.106	1.803	18.000	917.000	.021	.034
	Roy's Largest Root	.083	4.286 ^c	6.000	309.000	.000	.077

a. Design: Intercept + Surgery_type + Tumour_Stage + Surgery_type * Tumour_Stage

b. Exact statistic

c. The statistic is an upper bound on F that yields a lower bound on the significance level.

Tek yönlü manovada surgery type değişkenimizi yorumladığımız için burada sadece tumour stage ve surgery type*tumour stage değişkenlerimizi yorumluyoruz. Tek yönlüde de açıkça belirtildiği gibi willks lambda üzerinden test ediyoruz.

Tumour stage için:

H0: Gruplar arasında bağımlı değişkenler açısından anlamlı bir farklılık yoktur

Ha: Gruplar arasında bağımlı değişkenler açısından anlamlı bir farklılık vardır

(0,955 > 0,05) olduğu için H0 kabul yani Gruplar arasında bağımlı değişkenler açısından anlamlı bir farklılık yoktur diyebiliriz.

Partial Eta Squared genişliğine bakarak kullandığımız teste yani Willks Lambdaya göre bağımlı değişkenlerdeki değişimin % 0,003'ü grup değişkenleri tarafından açıklanmaktadır yorumunu yapabiliriz

Surgery type*tumour stage için:

H0: surgery type ve tumour stage etkileşiminin bağımlı değişkenler üzerinde anlamlı bir etkisi yoktur.

Ha: surgery type ve tumour stage etkileşiminin bağımlı değişkenlerin en az biri üzerinde anlamlı bir etkisi vardır.

($0,023 < 0,05$) olduğu için H0 red. Yani, surgery type ve tumour stage etkileşiminin bağımlı değişkenlerin en az biri üzerinde anlamlı bir etkisi vardır.

Partial Eta Squared genişliğine bakarak kullandığımız teste yani Willks Lambdaya göre bağımlı değişkenlerdeki değişimin % 0,034'ü grup değişkenleri tarafından açıklanmaktadır yorumunu yapabiliriz.

LEVENE'S TEST

Levene's Test of Equality of Error Variances^a

		Levene Statistic	df1	df2	Sig.
Protein2	Based on Mean	1.636	11	309	.088
	Based on Median	1.235	11	309	.262
	Based on Median and with adjusted df	1.235	11	280.543	.263
	Based on trimmed mean	1.641	11	309	.086
Protein3	Based on Mean	1.093	11	309	.366
	Based on Median	.960	11	309	.483
	Based on Median and with adjusted df	.960	11	272.107	.484
	Based on trimmed mean	1.072	11	309	.383
Protein4	Based on Mean	.484	11	309	.913
	Based on Median	.464	11	309	.924
	Based on Median and with adjusted df	.464	11	289.502	.924
	Based on trimmed mean	.469	11	309	.922

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + Surgery_type + Tumour_Stage + Surgery_type * Tumour_Stage

Tek yönlü manovadan farklı olarak bu kez Levene's testimizde, grubu alt gruba ayırip varyans eşitliklerini test ediyoruz.

H0: Gruplar arası varyanslar eşittir.

Ha: Gruplar arası varyanslar eşit değildir.

Hipotezimizi üç protein için de ayrı ayrı significant değerleri ile test edelim.

Protein 4 → ($0,05 < 0,913$) H₀ kabul, gruplar arası varyanslar eşittir.

Protein 3 → ($0,05 < 0,366$) H₀ kabul, gruplar arası varyanslar eşittir.

Protein 2 → ($0,05 < 0,088$) H₀ kabul, gruplar arası varyanslar eşittir.

H₀ hipotezini kabul ettiğimiz için TUKEY alanını yorumlayacağız. Reddetseydik tamhane alanını yorumlamamız gerekiyor. Fakat normal şartlarda varyans eşitliği sağlanmıyor ise tamhane'den bakmamız gerekiyor ama spss kaynaklı bir durum olduğundan, bunu yapamazdık.

TEST OF BETWEEN SUBJECT EFFECTS

Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	Protein2	23.226 ^a	11	2.111	2.701	.002	.088
	Protein3	4.483 ^b	11	.408	1.186	.295	.041
	Protein4	4.256 ^c	11	.387	.998	.448	.034
Intercept	Protein2	179.528	1	179.528	229.656	.000	.426
	Protein3	1.919	1	1.919	5.587	.019	.018
	Protein4	.079	1	.079	.203	.652	.001
Surgery_type	Protein2	9.598	3	3.199	4.093	.007	.038
	Protein3	.805	3	.268	.781	.505	.008
	Protein4	1.560	3	.520	1.342	.261	.013
Tumour_Stage	Protein2	.067	2	.033	.043	.958	.000
	Protein3	.191	2	.095	.278	.758	.002
	Protein4	.256	2	.128	.330	.719	.002
Surgery_type * Tumour_Stage	Protein2	18.079	6	3.013	3.855	.001	.070
	Protein3	2.393	6	.399	1.161	.327	.022
	Protein4	2.259	6	.376	.971	.445	.019
Error	Protein2	241.554	309	.782			
	Protein3	106.148	309	.344			
	Protein4	119.742	309	.388			
Total	Protein2	557.272	321				
	Protein3	113.424	321				
	Protein4	124.024	321				
Corrected Total	Protein2	264.780	320				
	Protein3	110.631	320				
	Protein4	123.999	320				

a. R Squared = .088 (Adjusted R Squared = .055)

b. R Squared = .041 (Adjusted R Squared = .006)

c. R Squared = .034 (Adjusted R Squared = .000)

Tumor stage değişkenimiz için:

H0: Bağımlı değişkenler açısından tumor aşamaları arasında anlamlı bir fark yoktur.

Ha: Bağımlı değişkenler açısından tumor aşamaları arasında anlamlı bir fark vardır.

Protein 2:

$0,958 > 0,05$ yani H0 kabul, bağımlı değişkenler açısından tumor aşamaları arasında anlamlı bir fark yoktur.

Protein 3:

$0,758 > 0,05$ yani H0 kabul , bağımlı değişkenler açısından tumor aşamaları arasında anlamlı bir fark yoktur.

Protein 4:

$0,719 > 0,05$ yani H0 kabul , bağımlı değişkenler açısından tumor aşamaları arasında anlamlı bir fark yoktur.

Surgery type ve tumor stage etkileşim değişkenimiz için:

H0: Surgery type ve tumor stage etkileşimlerinin bağımlı değişken üzerinde anlamlı bir etkisi yoktur.

H1: Surgery type ve tumor stage etkileşimlerinin bağımlı değişken üzerinde anlamlı bir etkisi vardır.

Protein 2:

$0,01 < 0,05$ yani H0 red, Surgery type ve tumor stage etkileşimlerinin protein 2 değişkeni üzerinde anlamlı bir etkisi vardır.

Protein 3:

$0,327 > 0,05$ yani H₀ kabul , Surgery type ve tumor stage etkileşimlerinin protein 3 değişkeni üzerinde anlamlı bir etkisi yoktur.

Protein 4:

$0,445 > 0,05$ yani H₀ kabul , Surgery type ve tumor stage etkileşimlerinin protein 4 değişkeni üzerinde anlamlı bir etkisi yoktur.

Post Hoc Tests

Surgery_type

Multiple Comparisons

Tukey HSD

Dependent Variable	(I) Surgery_type	(J) Surgery_type	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Protein2	Lumpectomy	Modified Radical Mastectomy	.332468787	.142623239	.093	-.03593638	.700873949
		Other	.107061145	.140787648	.872	-.25660256	.470724854
		Simple Mastectomy	.244523019	.154502189	.390	-.15456625	.643612284
	Modified Radical Mastectomy	Lumpectomy	-.33246879	.142623239	.093	-.70087395	.035936375
		Other	-.22540764	.128350476	.297	-.55694532	.106130038
		Simple Mastectomy	-.08794577	.143260634	.928	-.45799736	.282105826
	Other	Lumpectomy	-.10706115	.140787648	.872	-.47072485	.256602564
		Modified Radical Mastectomy	.225407642	.128350476	.297	-.10613004	.556945322
		Simple Mastectomy	.137461874	.141433315	.766	-.22786964	.502793383
	Simple Mastectomy	Lumpectomy	-.24452302	.154502189	.390	-.64361228	.154566247
		Modified Radical Mastectomy	.087945768	.143260634	.928	-.28210583	.457997362
		Other	-.13746187	.141433315	.766	-.50279338	.227869636

Protein3	Lumpectomy	Modified Radical Mastectomy	-.12986028	.094544970	.517	-.37407613	.114355565
		Other	.035898640	.093328156	.981	-.20517410	.276971375
		Simple Mastectomy	.048171041	.102419528	.966	-.21638530	.312727384
Modified Radical Mastectomy	Lumpectomy	.129860281	.094544970	.517	-.11435556	.374076126	
	Other	.165758921	.085083553	.210	-.05401746	.385535306	
	Simple Mastectomy	.178031322	.094967499	.241	-.06727594	.423338588	
Other	Lumpectomy	-.03589864	.093328156	.981	-.27697137	.205174096	
	Modified Radical Mastectomy	-.16575892	.085083553	.210	-.38553531	.054017465	
	Simple Mastectomy	.012272402	.093756169	.999	-.22990592	.254450722	
Simple Mastectomy	Lumpectomy	-.04817104	.102419528	.966	-.31272738	.216385302	
	Modified Radical Mastectomy	-.17803132	.094967499	.241	-.42333859	.067275943	
	Other	-.01227240	.093756169	.999	-.25445072	.229905919	

Protein4	Lumpectomy	Modified Radical Mastectomy	.095920254	.100417029	.775	-.16346350	.355304012
		Other	.174375204	.099124640	.295	-.08167023	.430420638
		Simple Mastectomy	.199311985	.108780665	.260	-.08167559	.480299562
Modified Radical Mastectomy	Lumpectomy	-.09592025	.100417029	.775	-.35530401	.163463505	
	Other	.078454951	.090367976	.821	-.15497145	.311881347	
	Simple Mastectomy	.103391731	.100865801	.735	-.15715123	.363934696	
Other	Lumpectomy	-.17437520	.099124640	.295	-.43042064	.081670230	
	Modified Radical Mastectomy	-.07845495	.090367976	.821	-.31188135	.154971446	
	Simple Mastectomy	.024936781	.099579237	.994	-.23228290	.282156466	
Simple Mastectomy	Lumpectomy	-.19931199	.108780665	.260	-.48029956	.081675591	
	Modified Radical Mastectomy	-.10339173	.100865801	.735	-.36393470	.157151233	
	Other	-.02493678	.099579237	.994	-.28215647	.232282905	

Based on observed means.
The error term is Mean Square(Error) = .388.

Levene's çıktısında verdigimiz karardan sonra TUKEY için incelememize devam ediyoruz.(Tamhane ile devam etmemiz gerekseydi de spss kaynaklı tukey'e bakacaktık.)

İncelediğimiz üç protein tipinin de mean difference değerlerinde yıldız yok. Dikkatli bakıldığından da lower ve upper boundların hep farklı işarette olduğunu görebiliyoruz.

H0: Lumpectomy ve simple mastectomy arasında protein 2 açısından istatistiksel bir anlamlı fark yoktur.

Ha: Lumpectomy ve simple mastectomy arasında protein 2 açısından istatistiksel bir anlamlı fark vardır.

$0,390 > 0,05$ H0 kabul, yani Lumpectomy ve simple mastectomy arasında protein 2 açısından istatistiksel bir anlamlı fark yoktur.

H0: Lumpectomy ve modified radical mastectomy arasında protein 3 açısından istatistiksel bir anlamlı fark yoktur.

Ha: Lumpectomy ve modified radical mastectomy arasında protein 3 açısından istatistiksel bir anlamlı fark vardır.

$0,517 > 0,05$ H0 kabul, yani Lumpectomy ve modified radical mastectomy arasında protein 3 açısından istatistiksel bir anlamlı fark yoktur.

H0: Modified radical mastectomy ve simple mastectomy arasında protein 4 açısından istatistiksel bir anlamlı fark yoktur.

Ha: Modified radical mastectomy ve simple mastectomy arasında protein 4 açısından istatistiksel bir anlamlı fark vardır.

$0,735 > 0,05$ H0 kabul, yani Modified radical mastectomy ve simple mastectomy arasında protein 4 açısından istatistiksel bir anlamlı fark yoktur.

Protein3

Tukey HSD^{a,b,c}

Surgery_type	N	Subset	
		1	
Simple Mastectomy	65	-.15795412	
Other	98	-.14568172	
Lumpectomy	66	-.10978308	
Modified Radical Mastectomy	92	.020077202	
Sig.		.234	

Means for groups in homogeneous subsets are displayed.

Based on observed means.

The error term is Mean Square(Error) = .344.

a. Uses Harmonic Mean Sample Size = 77.505.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

c. Alpha = .05.

Protein4

Tukey HSD^{a,b,c}

Surgery_type	N	Subset	
		1	
Simple Mastectomy	65	-.06941286	
Other	98	-.04447607	
Modified Radical Mastectomy	92	.033978876	
Lumpectomy	66	.129899130	
Sig.		.193	

Means for groups in homogeneous subsets are displayed.

Based on observed means.

The error term is Mean Square(Error) = .388.

a. Uses Harmonic Mean Sample Size = 77.505.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

c. Alpha = .05.

Homogeneous Subsets

Protein2

Tukey HSD^{a,b,c}

Surgery_type	N	Subset	
		1	
Modified Radical Mastectomy	92	.799581001	
Simple Mastectomy	65	.887526769	
Other	98	1.02498864	
Lumpectomy	66	1.13204979	
Sig.		.091	

Means for groups in homogeneous subsets are displayed.

Based on observed means.

The error term is Mean Square(Error) = .782.

a. Uses Harmonic Mean Sample Size = 77.505.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

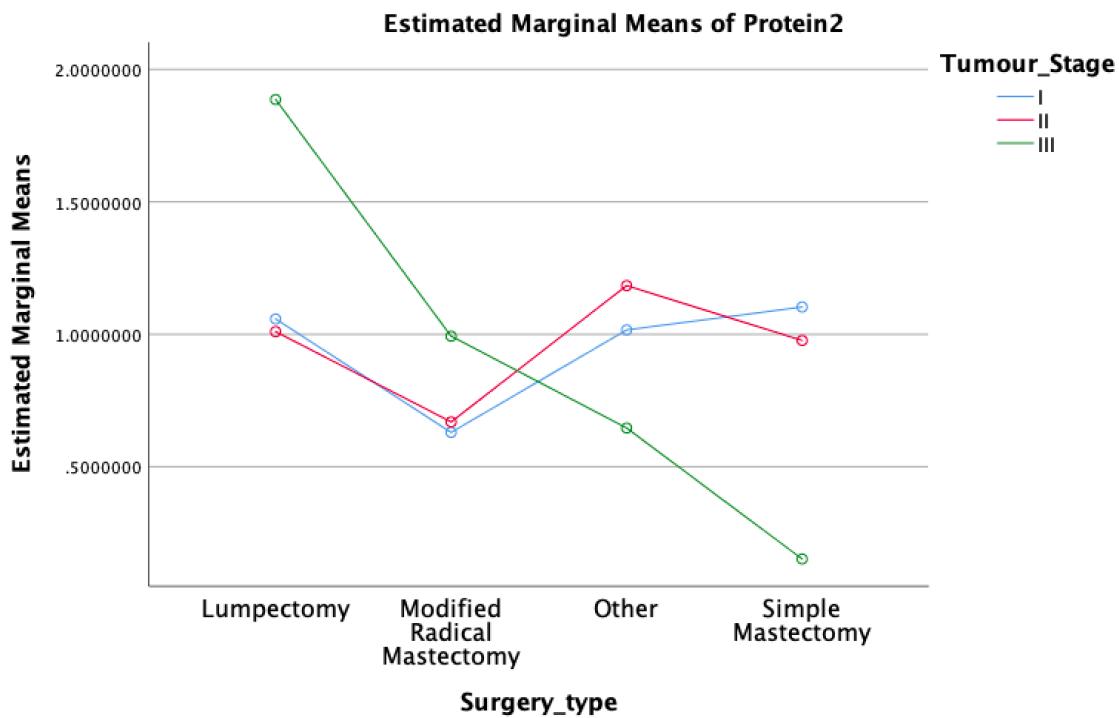
c. Alpha = .05.

Yine post hoc'un bir özetini gördüğümüz bu çıktılarla, farklı olmayan alt kümelerin bir arada olması gerektiğini ve post hoc'ta da yorumladığımız gibi farklılık olmadığını teyid etmiş oluyoruz.

Her bir bağımlı değişken için çıkardığımız grafikleri yorumluyoruz.

Profile Plots

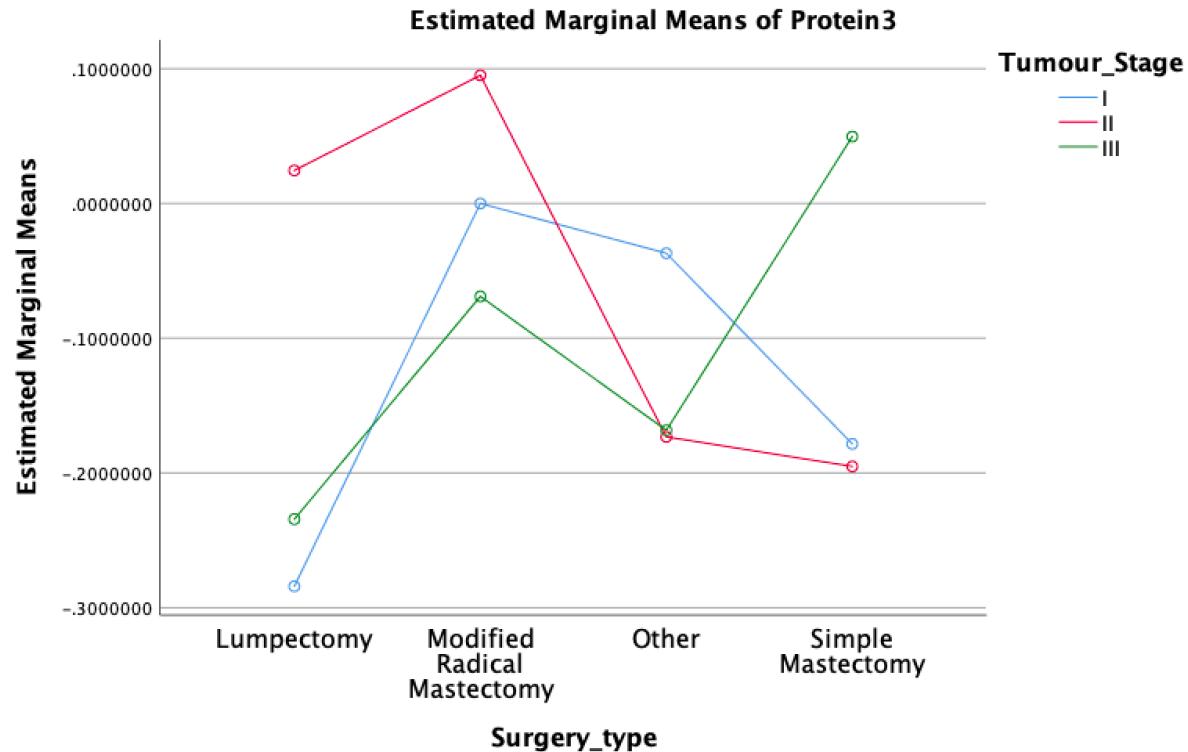
Protein2



Her bir bağımlı değişken için çıkardığımız grafikleri yorumluyoruz.

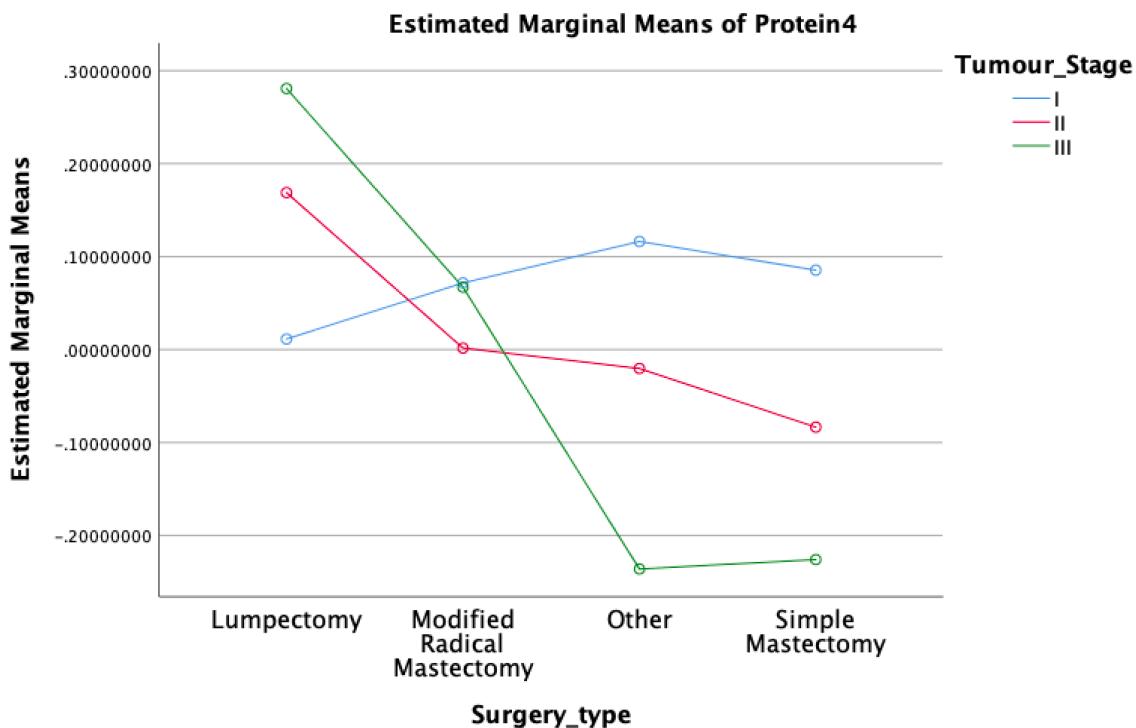
Kesişen kısımlarda etkileşim olduğunu gördüğümüzü söyleyebiliriz. Bu kısımlardan hayali bir dikey çizgi çekmişiz gibi düşünürsek protein 2 kaynaklı oluşan tumor için modified radical mastectomy ameliyatı ile çözülmesinde oluşan tumorun 3.aşamada olması 1 ve 2.aşamada olmasından daha yaygındır diyebiliriz. Aynı şekilde küçük bir fark olsa da 2.aşamanın da 1.aşamaya göre daha yaygın olduğunu söyleyebiliriz.

Protein3



Yine kesişen kısımdan bir çizgi çiztiğimizi düşünüyoruz. Bu kez test ettiğimiz ameliyatlardan başka ameliyat seçeneklerini belirten other düzeyi için konuşalım. Protein 3 kaynaklı oluşan tumor için bu(other) tip ameliyatlar ile çözülmlesi için tumorun 1. aşama olmasa 2. ve 3. aşamada olmasından daha yaygın diyebiliriz. bu grafiğimizde de 2 ve 3. aşama arasında çok küçük bir fark olsa da biz bu kez bahsedilen üç ameliyattan başka ameliyat olunduğu gözlemler için 3. aşamanın 2. aşamadan daha yaygın olduğunu söyleyebiliriz.

Protein4



Kesişen nokta karşımıza yine modified radical mastectomy ameliyatı çıkıyor. Bu kez de protein 4 için 1.ve 3. Tumor aşaması 2.ye göre daha yaygın diyebiliriz.

TEMEL BİLEŞENLER

Temel bileşenler ve faktör analizi, boyut indirmek için kullanılır.
Çok değişkenliliği koruyarak veri kümesinin boyutunu azaltmak
için gerçekleştirilir. Bunu 30 değişkeni 5 değişkene indirgerek
%20 bilgi kaybı ile çalışmak gibi örnekler ile açıklayabiliriz.
Amacımız verisetindeki maksimum varyansı açıklayan yeni bir
değişken oluşturmak. (maksimizasyon da denebilir)

Temel bileşen analizimizi metrik değişkenlerimiz ile
gerçekleştireceğiz. Bu analizi yapabilmemiz için korelasyonumuzun
0.3'ten büyük olması gerekiyor. Eğer 0.90'dan büyük ise çoklu
bağlantı problemi olduğunu söyleyebiliriz.

H0: Korelasyon katsayıları anlamlı değildir.

HA: Korelasyon katsayıları anlamlıdır.

P

	Protein1	Protein2	Protein3	Protein4
Protein1	0.0000	0.0405	0.0000	
Protein2	0.0000		0.0000	0.1340
Protein3	0.0405	0.0000		0.1669
Protein4	0.0000	0.1340	0.1669	

Korelasyon tablomuzdan:

Protein4 ve Protein2 arasında korelasyon katsayıları anlamlı değildir. ($0.13 > 0.05$, H0 Kabul)

Protein4 ve Protein3 arasında korelasyon katsayıları anlamlı değildir. ($0.16 > 0.05$, H0 Kabul)

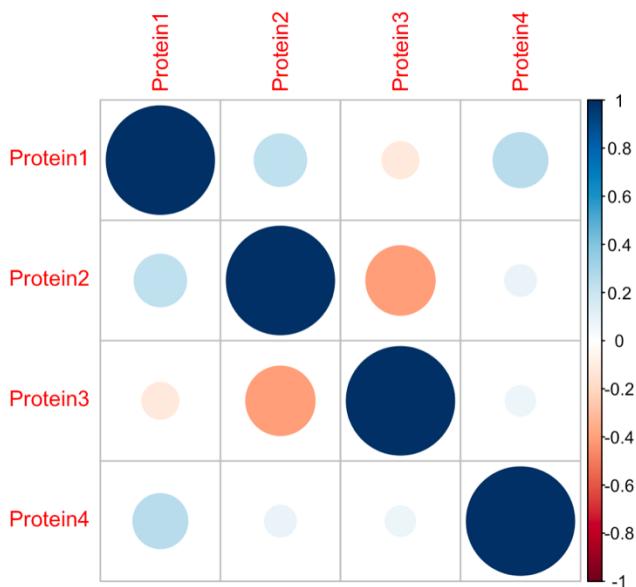
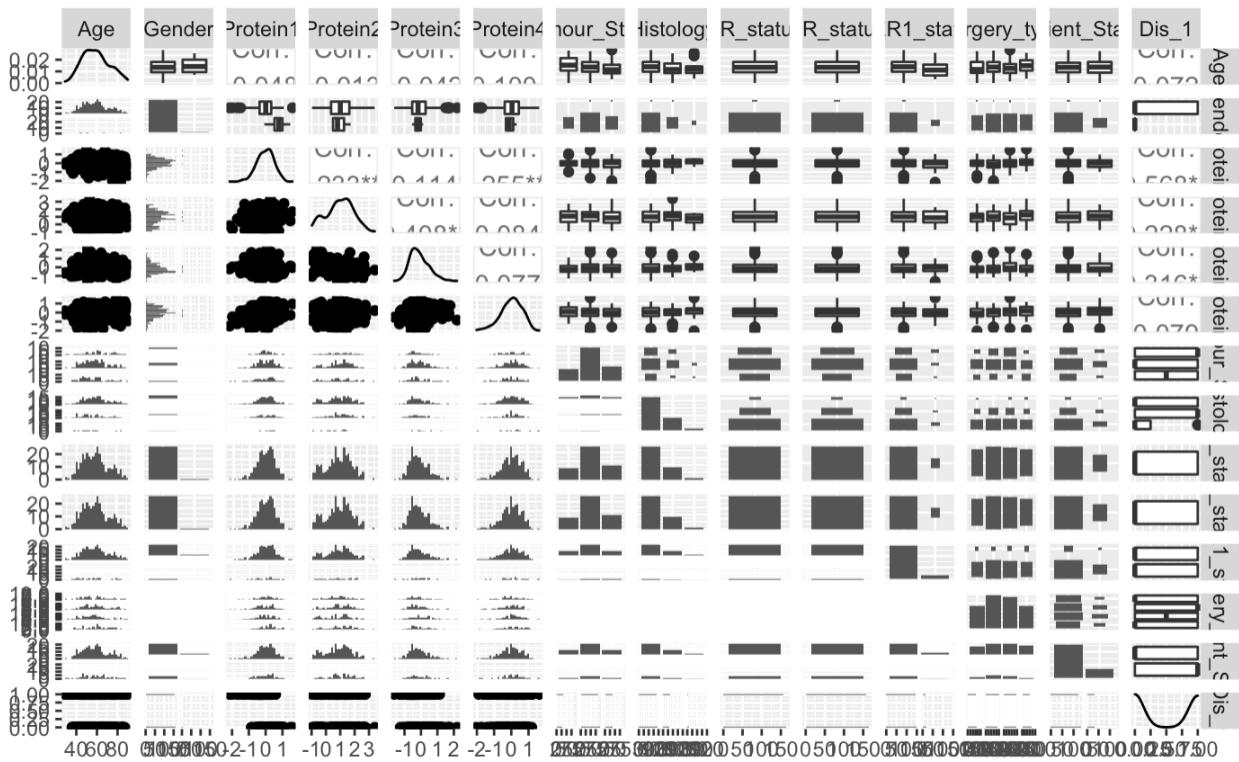
Protein1 ve Protein3 arasında korelasyon katsayıları anlamlı değildir. ($0.40 > 0.05$, H0 Kabul)

Protein1 ve Protein2 arasında korelasyon katsayıları anlamlıdır. ($0.00 < 0.05$, H0 Red)

Protein1 ve Protein4 arasında korelasyon katsayıları anlamlıdır. ($0.00 < 0.05$, H0 Red)

Protein3 ve Protein2 arasında korelasyon katsayıları anlamlıdır. ($0.00 < 0.05$, H0 Red)

Korelasyon matrisi grafiklerimiz şekildeki gibidir.



KMO:

Örneklem yeterliliğimizin ölçütü olarak kullandığımız bu test bize değişkenlerimizin önemli olup olmadığını sunar.

Eğer KMO değeri 0.50 den küçükse Temel Bileşenler Analizi uygulanamaz.

KMO Değerimiz:

- 0.5 ile 0.6 aralığında ise geçerli,
- 0.6 ile 0.7 aralığında ise orta düzey,
- 0.7 ile 0.8 aralığında ise iyi,
- 0.8 ile 0.9 aralığında ise mükemmel kabul edilir.

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = breast_cancer_data)
Overall MSA =  0.53
MSA for each item =
Protein1 Protein2 Protein3 Protein4
  0.58      0.53      0.51      0.48
```

Protein 4 değişkenimizin KMO değeri 0.4 yani 0.5'ten küçük bir değer olduğu için analizimize Protein4 değişkenimizi çıkararak devam ediyoruz.

BARTLETT KÜRESELLİK TESTİ:

H0: Korelasyon matrisimiz birim matristir.

Ha: Korelasyon matrisimiz birim matris değildir.

```
$chisq
[1] 75.88695
```

```
$p.value
[1] 2.338818e-16
```

```
$df
[1] 3
```

P value değerimiz ($2.338818e-16 < 0.05$) olduğu için H₀ Red, korelasyon matrisimiz birim matris değildir. Yani değişkenler korelemdir, değişkenler arasında anlamlı bir ilişki vardır. Artık bileşenlerin açıklama oranları yardımıyla yeterli bileşenleri seçeceğiz.

PRCOMP:

Importance of components:

	PC1	PC2	PC3
Standard deviation	1.2348	0.9503	0.7564
Proportion of Variance	0.5083	0.3010	0.1907
Cumulative Proportion	0.5083	0.8093	1.0000

Analizde görüldüğü üzere kümülatif değerler yardımıyla ilk bileşenin %50 açıkladığını görmekteyiz. Bu bileşenlerle beraber bu değer %80' e, çıkmıştır. Temel Bileşenler Analizi için 2/3 oranında açıklayıcılık yeterlidir. Bu sebeple 2 bileşen bizim için yeterlidir.

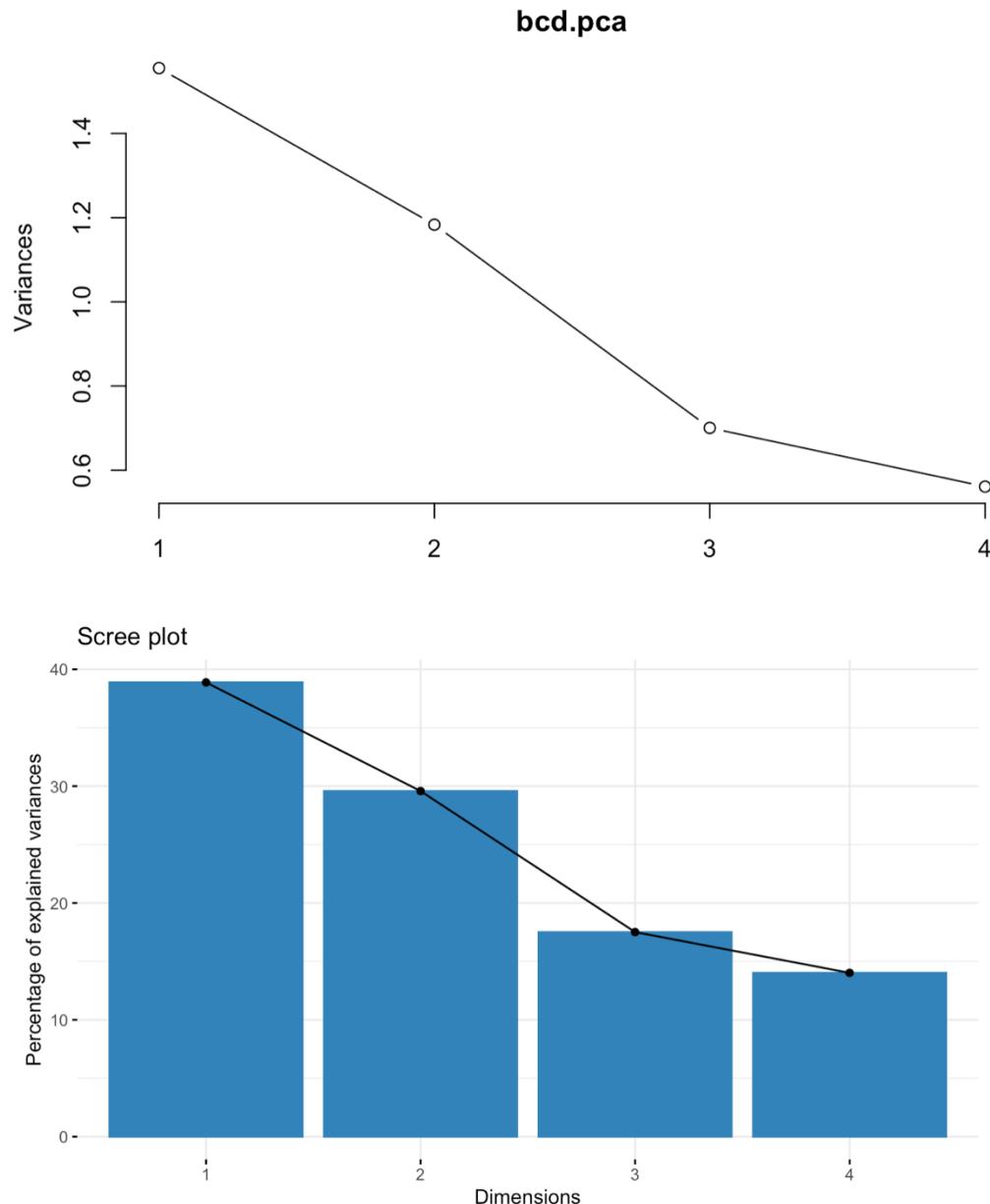
SUMMARY FIT.PCA:

[1] 1.5248515 0.9030325 0.5721160

Özdegeri 1'den büyük olan değişkenleri alabildiğimiz için sadece Protein1 değişkenimizin temel bileşenler analizine uyum sağladığı sonucuna ulaşsak da protein2 değişkenimizin de özdeğerinin 1'e yakın olmasından kaynaklı analizimizin anlamlı devam edebilmesi adına (normal şartlarda yapılması uygun değildir.) protein 2 değişkenimizi de alıyoruz.

SCREE PLOT:

Görsel olarak yeterli bileşen sayımızı gösterir. Ani düşüş gözlemlenir Kırılma noktası kadar kriterimiz olabilir diyebiliriz.



Kırılma noktası 3'e kadar hızlı bir düşüş; 3'ten sonra da düşüşte yavaşlama gözlemlendiği için 2 bileşene karar verilmiştir diyebiliriz.

FIT PCA ROTATION:

	PC1	PC2
Protein1	-0.4289239	0.8736968
Protein2	-0.6655009	-0.1338151
Protein3	0.6108460	0.4677044

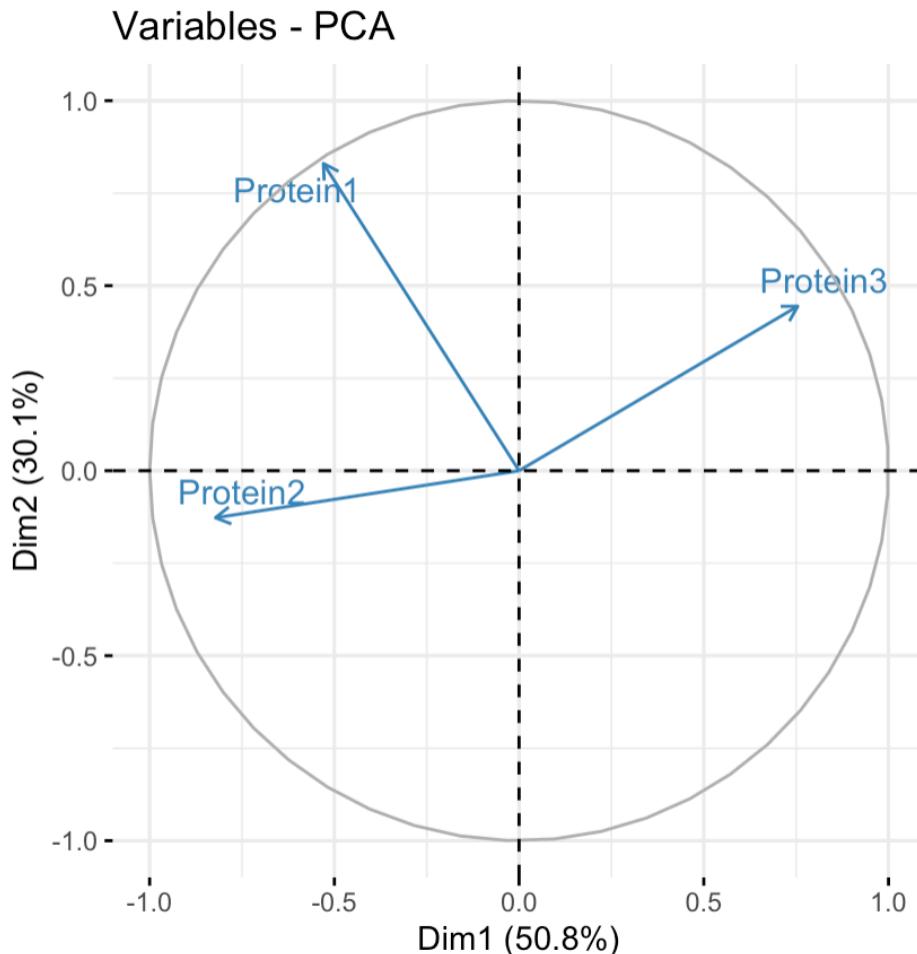
Y1= -0.42 Protein1 -0.66 Protein2 +0.61 Protein3

Y2= 0.87 Protein1 -0.13 Protein2 +0.46 Protein3

İNDEX OLUŞTURMA:

1.7019	2.1934	1.1352	1.6575	1.3379	1.7691
4.864137	4.598694	4.042734	3.976363	3.872520	3.438380

FACTOR EXTRA:



Hangi bileşenin hangi eksene yakın olduğunu bakmak için ve bu açı ne kadar küçük ise (eksene yakınsa) o kadar ilişkili olduğunu söyleyebildiğimiz bir çıktıdır.

X ekseni 1. Bileşene

Y ekseni ise 2. Bileşene yakındır.

1.bileşen toplam varyansın %30.1'ini açıklarken 2.bileşen %50.8'ini açıklamaktadır. İki bileşenimiz toplam %80 açıklayıcılık sağladığı için 2/3(%66) açıklanabilirlik oranı sağlanmıştır.

FAKTÖR ANALİZİ

Correlation Matrix^a

		Protein1	Protein2	Protein3	Protein4
Correlation	Protein1	1.000	.233	-.114	.255
	Protein2	.233	1.000	-.408	.084
	Protein3	-.114	-.408	1.000	.077
	Protein4	.255	.084	.077	1.000
Sig. (1-tailed)	Protein1		.000	.020	.000
	Protein2	.000		.000	.067
	Protein3	.020	.000		.083
	Protein4	.000	.067	.083	

a. Determinant = .723

Korelasyon matrisinin determinantı 0'a ne kadar yakınsa veriseti içindeki bağılilik o kadar fazladır. Bu tablomuzdan determinantımızın 0.72 olduğunu görebiliyoruz.

Protein1 değişkenimiz ile:

Protein2 (0.23) ve protein4 (0.25)'ün arasında pozitif yönde; protein 3,'ün (-0, 11) arasında negatif ilişki vardır.

Protein2 değişkenimiz ile:

Protein1 (0.23) ve protein4 (0.84)'ün arasında pozitif yönde; protein 3,'ün (-0, 40) arasında negatif ilişki vardır.

Protein3 değişkenimiz ile:

Protein2 (-0.40) ve protein1 (-0.11)'ün arasında negatif yönde; protein 4,'ün (0.77) arasında pozitif ilişki vardır.

Protein4 değişkenimiz ile:

Protein1 (0.25), protein 3,'ün (0.77) ve protein2 (0.84)'ün arasında pozitif yönde ilişki vardır.

SIG.(1-Tailed):

H0: Değişkenler arasında istatistiksel olarak anlamlı ilişki yoktur.

H1: Değişkenler arasında istatistiksel olarak anlamlı ilişki vardır.

Protein 2 ve protein 4 arasında anlamlı ilişki vardır ($0,67 > 0,05$) H0 KABUL.

Protein 3 ve protein 4 arasında anlamlı ilişki vardır ($0,83 > 0,05$) H0 KABUL.

Protein 1 ve protein 4 arasında anlamlı ilişki yoktur ($0,00 > 0,05$) H0 RED.

Protein 1 ve protein 2 arasında anlamlı ilişki yoktur ($0,00 > 0,05$) H0 RED.

Protein 1 ve protein 3 arasında anlamlı ilişki yoktur ($0,02 > 0,05$) H0 RED.

Inverse of Correlation Matrix

	Protein1	Protein2	Protein3	Protein4
Protein1	1.127	-.214	.063	-.275
Protein2	-.214	1.260	.497	-.089
Protein3	.063	.497	1.222	-.152
Protein4	-.275	-.089	-.152	1.089

Köşegen elemanları bize varyans şışme değerlerini verir(VIF). Eğer VIF > 5 ise çoklu bağlantı problemi vardır diyebiliriz. eğer çoklu bağlantı problemimiz varsa değişkenlerimiz birbirinden türemiş olabilir. Bu değişkenlerin çıkarılması gerekmektedir.

Hiçbir değerimiz 5'ten büyük olmadığı için şimdilik aynı değişkenlerimizle devam ediyoruz.

KMO VE BARTLETT'S TEST

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.	.527
Bartlett's Test of Sphericity	103.040
df	6
Sig.	.000

KMO değerimiz > 0.5 olduğu için bu kez kabul edilir diyebiliriz.

Bartlett's test p değerimizin (sig) ise <0.05 olduğunu görüyoruz.

H0: Korelasyon matrisi birim matristir.

Ha: Korelasyon matrisi birim matris değildir.

0,00 < 0,05 olduğu için H0 red, korelasyon matrisi birim matris değildir.

ANTI IMAGES MATRICES:

Bu çıktımızda köşegen elemanlarımız değişken bazında KMO verir. Köşegen dışı elemanlar ise kısmi korelasyon matrisinin negatif değerini verir ve ne kadar küçükse ilişki o kadar fazladır.

Anti-image Matrices

		Protein1	Protein2	Protein3	Protein4
Anti-image Covariance	Protein1	.887	-.151	.046	-.224
	Protein2	-.151	.794	.323	-.065
	Protein3	.046	.323	.819	-.114
	Protein4	-.224	-.065	-.114	.918
Anti-image Correlation	Protein1	.579 ^a	-.180	.053	-.248
	Protein2	-.180	.535 ^a	.400	-.076
	Protein3	.053	.400	.507 ^a	-.132
	Protein4	-.248	-.076	-.132	.480 ^a

a. Measures of Sampling Adequacy(MSA)

Tüm köşegen elemanlarımızın 0.5'ten büyük olduğunu görüyoruz ve hiçbir değişkenimizi çıkarmadan analizimize devam ediyoruz.

COMMUNALITIES

Communalities

	Initial	Extraction
Protein1	1.000	.605
Protein2	1.000	.685
Protein3	1.000	.733
Protein4	1.000	.715

Extraction Method: Principal Component Analysis.

Extraction kısmımızın %50'den büyük olmasını istiyoruz. Küçük olduğu durumlarda değişkeni analizden çıkarmamız gerekiyor. Tablomuzdan hiçbir değişkenin %50'den küçük olmadığını ve initial değerlerimizin birbirine eşit olduğunu görüyoruz ve değişkenlerimiz eksilmeden analizimize devam ediyoruz.

TOTAL VARIANCE EXPLAINED

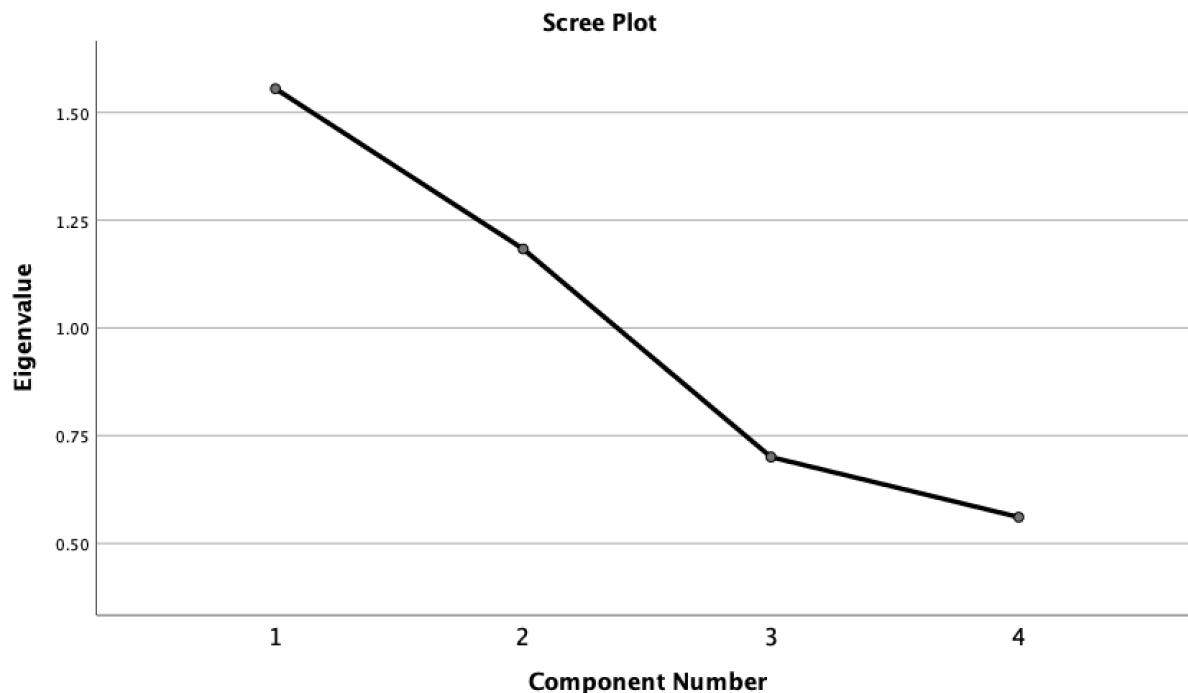
Total Variance Explained

Component	Total	Initial Eigenvalues		Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
		% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.555	38.882	38.882	1.555	38.882	38.882	1.455	36.363	36.363
2	1.183	29.583	68.465	1.183	29.583	68.465	1.284	32.102	68.465
3	.700	17.512	85.977						
4	.561	14.023	100.000						

Extraction Method: Principal Component Analysis.

Extraction kısmında 1.faktörümüzün %38 oranında, 2.faktörün ise %29 oranında açıklandığını gözlemliyoruz. Açıklanabilirliği artırmak için çeşitli rotasyonlar deneyeceğiz.

SCREE PLOT



Temel bileşenlerde de R üzerinden aldığımız scree plotu burada da görüyoruz. Yine kırılma noktası 3.noktada.

COMPONENT MATRIX

Component Matrix^a

	Component	
	1	2
Protein2	.796	
Protein3	-.669	.535
Protein1	.615	.477
Protein4	.310	.787

Extraction Method: Principal Component Analysis.

- a. 2 components extracted.

Değişkenlerimizin faktörlere yerleşme matrisini incelediğimizde protein2 dışındaki değişkenlerimizin yerleşirken sıkıntı yaşadığını gözlemliyoruz. Rotasyon işlemimizden sonra bu durumu tekrar inceleyeceğiz.

REPRODUCED CORRELATIONS

Reproduced Correlations

		Protein1	Protein2	Protein3	Protein4
Reproduced Correlation	Protein1	.605 ^a	.382	-.156	.566
	Protein2	.382	.685 ^a	-.653	.070
	Protein3	-.156	-.653	.733 ^a	.214
	Protein4	.566	.070	.214	.715 ^a
Residual ^b	Protein1		-.149	.042	-.310
	Protein2	-.149		.245	.014
	Protein3	.042	.245		-.137
	Protein4	-.310	.014	-.137	

Extraction Method: Principal Component Analysis.

a. Reproduced communalities

b. Residuals are computed between observed and reproduced correlations.
There are 4 (66.0%) nonredundant residuals with absolute values greater than 0.05.

Faktör analizimizde elde ettiğimiz ilk korelasyon matrisimizin üstüne rotasyondan sonra tekrar oluşturduğumuz matrisimizi inceliyoruz. Köşegen elemanlarının ilk matrisin köşegen elemanları ile arasındaki farkın minimum 0.05 daha az olması istenir. Bu azalma ne kadar yüksekse analiz o kadar sağlıklı denebilir. Bizim değerlerimizde ise 0.30-0.40 aralığında azalmalar gözleniyor. Döndürme işlemimiz istediğimiz sonuçları vermeye başladı diyebiliriz.

ROTATED COMPONENT MATRIX

Rotated Component Matrix^a

	Component	
	1	2
Protein3	-.849	
Protein2	.798	
Protein4		.834
Protein1		.726

Extraction Method: Principal Component Analysis.

Rotation Method: Quartimax with Kaiser Normalization.

- a. Rotation converged in 3 iterations.

Gerçekleştirdiğimiz quartimax rotasyonundan sonra bu kez değişkenlerimizin faktörlere rahatlıkla hiçbir karışıklık olmadan yerleşiklerini görüyoruz.

Faktör1: -0.84 Protein3 + 0.79 Protein2

Faktör2: 0.83 Protein3 + 0.72 Protein2

EQUAMAX ROTATION

Rotated Component Matrix^a

Component	
1	2
Protein3	-.849
Protein2	.797
Protein4	.833
Protein1	.727

Extraction Method: Principal Component Analysis.
Rotation Method: Equamax with Kaiser Normalization.

- a. Rotation converged in 3 iterations.

Equamax rotasyonumuzda da baştaki reproduced correlations ve component matrix tablolarımız aynı oranlarda değişim gösterdiginden, bu kısımları analizimizin sadece outputs kısmına ekliyoruz. Fakat promax rotasyonu sonrası pattern matriximizi rotasyon türü değiştiğinde sonucumuzun şimdilik değişmediğini gözlemleyebilmek için rapor kısmımıza ekliyoruz.

Faktör1: -0.84 Protein3 + 0.79 Protein2

Faktör2: 0.83 Protein3 + 0.72 Protein2

PROMAX ROTATION

Pattern Matrix^a

	Component	
	1	2
Protein3	-.860	
Protein2	.789	
Protein4		.847
Protein1		.713

Extraction Method: Principal Component Analysis.
Rotation Method: Promax with Kaiser Normalization.

- a. Rotation converged in 3 iterations.

Promax rotasyonumuzda da equamax rotasyonunda olduğu gibi belirli kısımları tekrar analizimize eklememize gerek kalmıyor. Promax rotasyonu sonrası pattern matriximizi rotasyon türü değiştiğinde sonucumuzun yine değişmediğini gözlemleyebilmek için rapor kısmımıza ekliyoruz.

Faktör1: -0.86 Protein3 + 0.78 Protein2

Faktör2: 0.84 Protein3 + 0.71 Protein2

DİSKRİMİNANT ANALİZİ

Diskriminant analizinde amaç veri kümesine yeni eklenen bir gözlemi var olan grplardan birine atamaktır.Çok değişkenli normallik ve ortak varyans kovaryans matrisini sağlamak diskriminant analizi yapabilmemiz için gerekli varsayımlardır. **Grup sayısına göre iki alt başlıkta incelenir grupların adına son parttan bi daha bak ($k=2$, $k >2$). Lojistik regresyondan farkı veri setindeki kategorik değişkenler yerine hipotatik değişken kullanılmasıdır. Fakat biz uygulamamız gereği veri setimizdeki kanser hücrelerinin büyümeyi destekleyen bir protein çeşidi olan HER2_status kategorik değişkenimizi kullanacağız.**

MANOVA kısmında verimizdeki tüm nicel değişkenlerin tek ve çift değişkenli normalliklerine bakmış olduğumuz için bu aşamada da aynı değişkenler ile analize devam edeceğimiz için tekrar normalliğe bakmıyoruz. Normal şartlarda normallik sağlanamazsa alternatif analiz olan lojistik regresyon analizi yapılmalı fakat uygulama gereği normallik sağlanmamış şekilde discriminant analizimize devam ediyoruz.

Correlations

			Protein1	Protein2	Protein3	Protein4
Spearman's rho	Protein1	Correlation Coefficient	1.000	.232 **	-.119 *	.223 **
	Protein2	Correlation Coefficient	.232 **	1.000	-.344 **	.094
	Protein3	Correlation Coefficient	-.119 *	-.344 **	1.000	.076
Protein4	Protein1	Correlation Coefficient	.223 **	.094	.076	1.000
	Protein2	Correlation Coefficient	.000	.	.000	.094
	Protein3	Correlation Coefficient	.033	.000	.	.174
N	Protein1	N	321	321	321	321
	Protein2	N	321	321	321	321
	Protein3	N	321	321	321	321
N	Protein4	N	321	321	321	321

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Korelasyon tablomuzdan değişkenlerimizin korelasyon katsayılarının düşük olduğunu gözlemliyoruz. Hiçbir değişkenimizi çıkarmadan analizimize devam ediyoruz.

ANALYSIS CASE PROCESSING SUMMARY

Analysis Case Processing Summary

Unweighted Cases		N	Percent
Valid		321	100.0
Excluded	Missing or out-of-range group codes	0	.0
	At least one missing discriminating variable	0	.0
	Both missing or out-of-range group codes and at least one missing discriminating variable	0	.0
Total		0	.0
Total		321	100.0

Bu çıktımızı incelediğimizde veri kümemizde eksik gözlemimizin olmadığını rahatlıkla söyleyebiliyoruz.

Group Statistics

		Valid N (listwise)	
	HER1_status	Unweighted	Weighted
0	Protein1	292	292.000
	Protein2	292	292.000
	Protein3	292	292.000
	Protein4	292	292.000
1	Protein1	29	29.000
	Protein2	29	29.000
	Protein3	29	29.000
	Protein4	29	29.000
Total	Protein1	321	321.000
	Protein2	321	321.000
	Protein3	321	321.000
	Protein4	321	321.000

N kısmından her2 değişkenimizin gözlem dağılımının dengesiz olduğunu görüyoruz. Normal şartlarda gözlem sayımızın dengeli olması gereklidir fakat biz uygulama gereği analizimize devam ediyoruz.

BOX M

Test Results

Box's M	10.252
F	Approx.
	.971
df1	10
df2	10277.826
Sig.	.466

Tests null hypothesis of equal population covariance matrices.

H0: Gruplar arası varyans- kovaryans matrisleri eşittir.

Ha: Gruplar arası varyans-kovaryans matrisleri eşit değildir.

$0.466 > 0.05$ H0 Kabul, gruplar arası varyans- kovaryans matrisleri eşittir.

SUMMARY OF CANONICAL DISCRIMINANT FUNCTIONS

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.002 ^a	100.0	100.0	.047

a. First 1 canonical discriminant functions were used in the analysis.

Özdeğer çıktılarımızda

eigenvalue(özdeğer) değerimiz ne kadar büyükse bağımlı değişkendeki varyans o kadar fazla açıklanır. kesin olmamakla birlikte bu oranın 0.40'tan büyük olması beklenir.

Canonical correlation kısmımızda ise diskriminant fonksiyon skorlarıyla gruplar arasındaki ilişkiyi ölçer bu değerimizin karesi bize varyansı verir.

Kurulan modelin grup değişkenindeki varyansın yaklaşık %5'ini açıkladığını gözlemlemiş oluyoruz. Sağlamasını yapmak için bir de Wilk' Lambda çıktıımızı inceleyelim.

WILK'S LAMBDA

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.998	.708	4	.950

Model tarafından açıklanamayan kısmı verir (Canonical correlation)² ile wilks lambda kısmının toplamı 1 eder. Yani birbirlerinin tümleyenleridir diyebiliriz.

Düzen sayımız 2 olduğu için wilks lambda çıktıımızdan bir fonksiyon elde ettik (2-1=1).

H0: Diskriminant fonksiyonu önemsizdir

H1: Diskriminant fonksiyonu önemsiz değildir.

Açıklanamayan kısımdan bahsettiğimiz için H0'ı reddetmek istiyoruz fakat maalesef $0.95 > 0.05$ olduğu için H0'ı kabul ediyoruz. Bu açıklanabilirlik oranına rağmen uygulama gereği analize devam edeceğiz.

STRUCTURE MATRIX

**Structure
Matrix**

	Function 1
Protein1	.711
Protein3	.392
Protein2	.321
Protein4	-.077

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

Diskriminant fonksiyonları ile değişkenlerin ilişkilerini gösterir. En yüksek değer en yüksek ilişkili olduğu için en çok açıklayandır şeklinde yorumlayabiliriz.

CANONICAL DISCRIMINANT FUNCTION COEFFICIENTS

Canonical Discriminant Function Coefficients

Function	
1	
Protein1	1.423
Protein2	.499
Protein3	1.181
Protein4	-.591
(Constant)	-.323

Unstandardized coefficients

Diskriminant fonksiyonumuzu çıkardığımız çıktıdır.

$$y = -0.323 - 0.591 \text{ Protein4} + 1.181 \text{ Protein3} + 0.499 \text{ Protein 2} + 1.423 \text{ Protein 1}$$

FUNCTIONS OF GROUP CENTROIDS

Functions at Group Centroids

Function	
HER1_status	1
0	.015
1	-.150

Unstandardized canonical discriminant functions evaluated at group means

Grup merkezlerini gösteriyor. Ortalama diskriminant fonksiyonlarını veriyor. Değerlerimizin (+) ve (-) olması iyi ayırtığını gösteriyor. Eğer düzey sayımız ikiden fazla olsaydı bu kez de. (- (+) (+) gibi durumlarda (+) (+) ayısamamış (-) olan ayırmış şeklinde yorumlardık.

CLASSIFICATION STATISTICS

Prior Probabilities for Groups

HER1_status	Prior	Cases Used in Analysis	
		Unweighted	Weighted
0	.500	292	292.000
1	.500	29	29.000
Total	1.000	321	321.000

Bu çıktımız düzeylerin total içindeki yüzdesini verir. Doğru atamanın yeterliliğine karar vermek için nisbi şans kriteri kullanılır. Prior probabilities tablomuzda yüzdelerimizi kontrol ettikten sonra classification results tablomuzdan diskriminant modelimizi değerler üzerinde inceliyoruz.

$P1^2 + P2^2 <$ dopru atama oranı ise iyi bir diskriminant modelidir deriz.

Classification Results^a

Original	Count	HER1_status	Predicted Group Membership		Total
			0	1	
	0	0	147	145	292
	1	1	17	12	29
	%	0	50.3	49.7	100.0
		1	58.6	41.4	100.0

a. 49.5% of original grouped cases correctly classified.

Bu kısımda köşegenlerimiz doğru atamaları gösterir (147, 12). Diğer iki değer de (145,17) yanlış atamaları gösterir. Tablodan doğru atama oranımızın %41 ; yanlış atama oranımızın da %58 olduğunu gözlemliyoruz. Doğru atama oranımız count kısmının köşegenleri toplamının total gözleme bölünmesi ile bulunur.

Diskriminant analizimizde asıl önemli ve ilgilenilmesi kısımlar yanlış atamalardır. Fakat biz örneğin bu veri ile çalışan bir doktor olmadığımız için bu yanlış atamaların modelin belki karışıklığa uğraması belki değer ezberlemesi vs olup olmadığını anlayamayacağımız için bu yanlış atamalar ile ilgili kesin yargılarda bulunamıyoruz.

LOJİSTİK

Lojistik regresyonda amaç bağımlı değişken kategorilerinden birine atanma olasılığı elde etmektir. Yani analizimizin sonucunda elde ettiğimiz, bir olasılık olacak. Diskriminantın farkı temel varsayımlar sağlanmadığında (özellikle normallik) lojistik regresyon kullanabiliyor olmamız. Diskriminanta göre çok daha rahat yorumlanır. Bağımsız değişkenler kategorik de olabilir. Lojistik regresyon analizinde ne üzerinden test yapıyorsak ona 1 değeri atanır. Örneğin bankalardaki batan ve batmayan müşteriler üzerinden bir analiz yapacak olursak batanları inceleyeceğimiz için 1 batanlar 0 batmayanlar olur. Matematiksel olarak esnektir.

Hatalar ve y değerleri binom dağılımlıdır.

Yine her analizimizde olduğu gibi çoklu bağlantı problemleri ve aykırı değerlerin olmamasını isteriz

Case Processing Summary

Unweighted Cases ^a		N	Percent
Selected Cases	Included in Analysis	321	100.0
	Missing Cases	0	.0
	Total	321	100.0
Unselected Cases		0	.0
Total		321	100.0

a. If weight is in effect, see classification table for the total number of cases.

Eksik gözlemimizin olmadığını ve 321 gözlemimizin olduğunu bu çıktıdan gözlemliyoruz.

Dependent Variable Encoding

Original Value	Internal Value
0	0
1	1

Spss bizim yerimize yüksek düzeyli değişkene 1 atıyor ve onu tahminleyeceğimiz olarak seçiyor.

Bağımsız değişkenler eklenmeden Log likelihood değerimiz:

Block 0: Beginning Block

Iteration History^{a,b,c}

Iteration		-2 Log likelihood	Coefficients
			Constant
Step 0	1	209.005	-1.639
	2	195.274	-2.170
	3	194.739	-2.302
	4	194.738	-2.309
	5	194.738	-2.309

a. Constant is included in the model.

b. Initial -2 Log Likelihood: 194.738

c. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Block 1:

Iteration History^{a,b,c,d}

Iteration		-2 Log likelihood	Coefficients				
			Constant	Protein1	Protein2	Protein3	Protein4
Step 1	1	208.663	-1.621	-.077	-.027	-.064	.032
	2	194.634	-2.134	-.168	-.060	-.142	.071
	3	194.025	-2.259	-.224	-.081	-.192	.097
	4	194.022	-2.266	-.231	-.084	-.199	.100
	5	194.022	-2.266	-.231	-.084	-.199	.100

a. Method: Enter

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 194.738

d. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

-2 log likelihood bize bağımlı değişkenin açıklanamayan varyansını verir, doğal olarak biz de büyük bir değer çıkışmasını tercih etmeyiz.

Bağımsız değişkenler eklenmeden Log likelihood değerimiz, ekledikten sonraki değerimizden farklı çıkmıyor. Yani uyum anlamında bir farklılık yok diyebiliriz. iterasyon çıktımızdan parametre kestirimlerinin 5.adımda bittiğini gözlemliyoruz.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	.715	4	.949
	Block	.715	4	.949
	Model	.715	4	.949

Bu çıktımız bize bağımsız değişken eklenmesi şeklinde kurulan modelin anlamlılığını veriyor.

H0: Modele eklenen bağımsız değişkenler modele anlamlı bir katkı sağlamamaktadır. Yani constant değerimiz ile kurulamn model daha iyidir.

Ha: Modele eklenen bağımsız değişkenler modele anlamlı bir katkı sağlamaktadır. Yani constant değerimiz ile kurulamn model daha iyi değildir.

(Beklentimiz h0'ı reddetmek.)

Sig değeri $0.949 > 0.05$ olduğu için H0 kabul yani modele eklenen bağımsız değişkenler modele anlamlı bir katkı sağlamamaktadır. Yani constant değerimiz ile kurulamn model daha iyidir.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	194.022 ^a	.002	.005

- a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Modelin sonunda -2loglikelihoodumuzu yorumladığımız için tekrar üstinden geçmiyoruz. Bu çıktımızda gördüğümüz Cox & Snell R Square ve Nagelkerke R square için modelin uyumunu değerlendirmek için seçenek modelleri karşılaştırmada kullanıldıklarını söyleyebiliriz. Modelin açıklayıcıları da olsalar doğrusal regresyondaki gibi 1'e yakın olmalarını bekleyemeyiz. Fakat tabii ki hala 1'e ne kadar yakınsa o kadar açıklayıcıdır. Özellikle Nagelkerke R square'e bakmamız daha açıklayıcı bir yorum yapmamızı sağlayabilir.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	9.781	8	.281

H0: Model, veri kümesi ile uyumludur, gözlenen e yakındır.

Ha: Model, veri kümesi ile uyumlu değildir, gözlenen e yakın değildir.

(doğal olarak H0'ı kabul etmek istiyoruz.)

$0.281 > 0.05$ H0 kabul, model, veri kümesi ile uyumludur, gözlenen e yakındır.

Classification Table^a

Observed		Predicted		Percentage Correct
		HER1_status 0	1	
Step 1	HER1_status	0	292	0
		1	29	0
Overall Percentage				91.0

a. The cut value is .500

292 değişkenin %100'ü pozitife atanmış.

29 değişkenin tamamı ise pozitife atanmış. Gerçekte 1 iken 0'a atanarak yanlış atama gerçekleşmiştir.

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a	Protein1	-.231	.373	.383	1	.536
	Protein2	-.084	.244	.118	1	.731
	Protein3	-.199	.375	.281	1	.596
	Protein4	.100	.330	.092	1	.761
	Constant	-2.266	.291	60.555	1	.104

a. Variable(s) entered on step 1: Protein1, Protein2, Protein3, Protein4.

$$\ln\left(\frac{p}{1-p}\right)$$

Normal şartlarda sig değerleri > 0.05 olanlar anlamsız olduğu için yorumlamamıza gerek yok fakat uygulama gereği yorumluyoruz.

$$Y = -2.266 + 0.10 \text{ Protein4} - 0.19 \text{ Protein3} - 0.084 \text{ Protein2} - 0.231 \text{ Protein1}$$

Katsayılarımızı $\exp\beta$ 'den yorumluyoruz. $\exp\beta$ odds yani olabilirlik oranı:

$\exp\beta < 1$ ise negatif etki,

$\exp\beta > 1$ ise pozitif etki,

$\exp\beta = 1$ ise etki yoktur.

$\exp>1$ olduğu için protein4'ün çıkarılması gerekmektedir. Çünkü 1 birimlik artışa karşılık HER1_status kanser proteininin pozitif olma olasılığını negatif olma olasılığına göre 1.105 kat artırrır

KÜMELEME

Kümeleme analizi veri setini belirli gruplara ayırmaya yaramaktadır.

Benzerlikleri gözeterek yani uzaklıklarını kullanarak benzerlik dereceleri hesaplanıyor ve benzer olanlar aynı kümeye konuluyor. Analiz sonunda elde edilen kümelerin kendi içinde homojen, kendi aralarında ise heterojen bir yapıda olmaları beklenir.

Kümeleme analizimizin temel varsayımları normalilik, korelasyon ve çoklu bağlantı problemi olmamasıdır. Biz kümeleme yaptığımız değişkenlerimizin daha önce bu varsayımlarını inceleyip gerekli yorumları yaptığımız için bu aşamada da analiz gereği hiçbir değişkenimizi veriden çıkarmadan devam ediyoruz.

Hiyerarşik ve hiyerarşik olmayan kümeleme analizi ile verimizi kümelere ayıracagız.

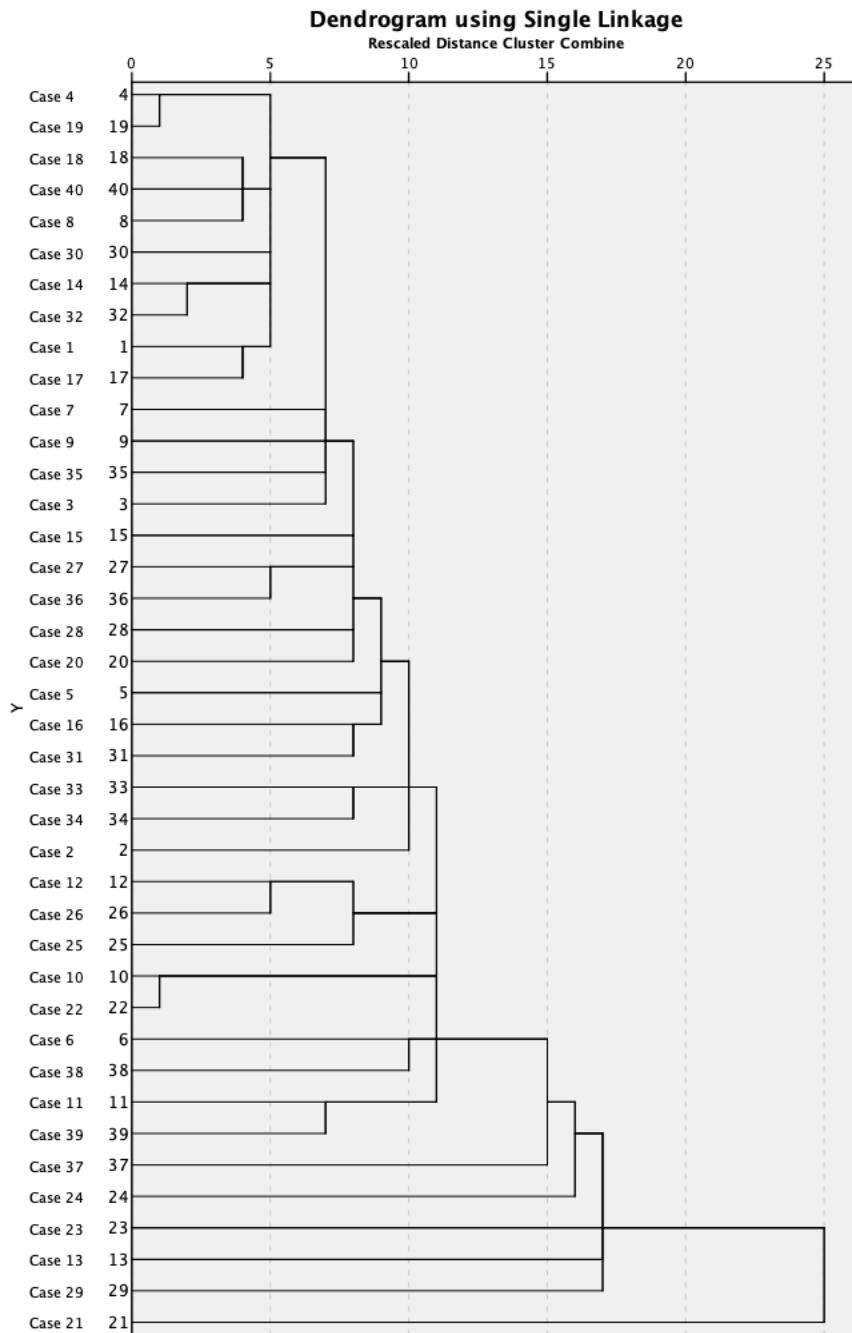
Hiyerarşik kümeleme analizinde dendogram kullanılır(bu grafik dallar ve çubuklar ile ifade eder.) hiyerarşik denmesinin sebebi bu adının aşama aşama gerçekleşmesidir. Veriyi kaç kümeye ayıracagımıza bu adımda karar veririz. Veri kümemiz fazla gözlemli olduğundan dendogramları ve diğer analizleri daha sağlıklı yorumlamak istiyoruz bu yüzden verimize %10'luk bir sampling yapıyoruz.

Hiyerarşik olmayan kümeleme analizinde ise verimiz toplamı minimum olacak k tane kümeye ayrıılır. Bu kümeler kendi aralarında homojen birbirleri arasında heterojen olmalıdır.

Her $n - k$ gözlem ortalaması en yakın olan kümeye atanır. Gözlemler tam olarak kümelere yerleşene kadar bu atama devam eder.

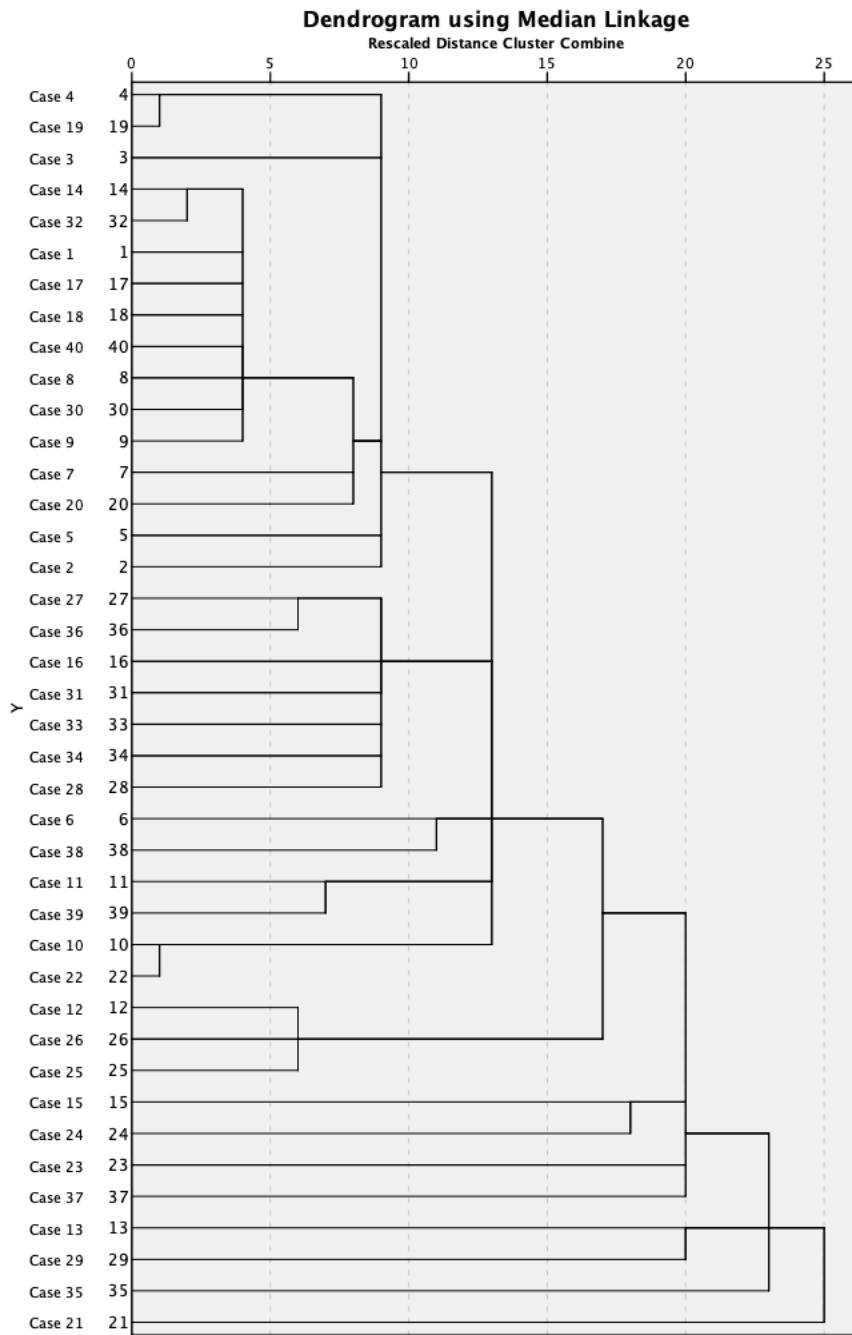
Hiyerarşik kümeleme analizi

Nearest Neighbor Kümeleme Methodu:



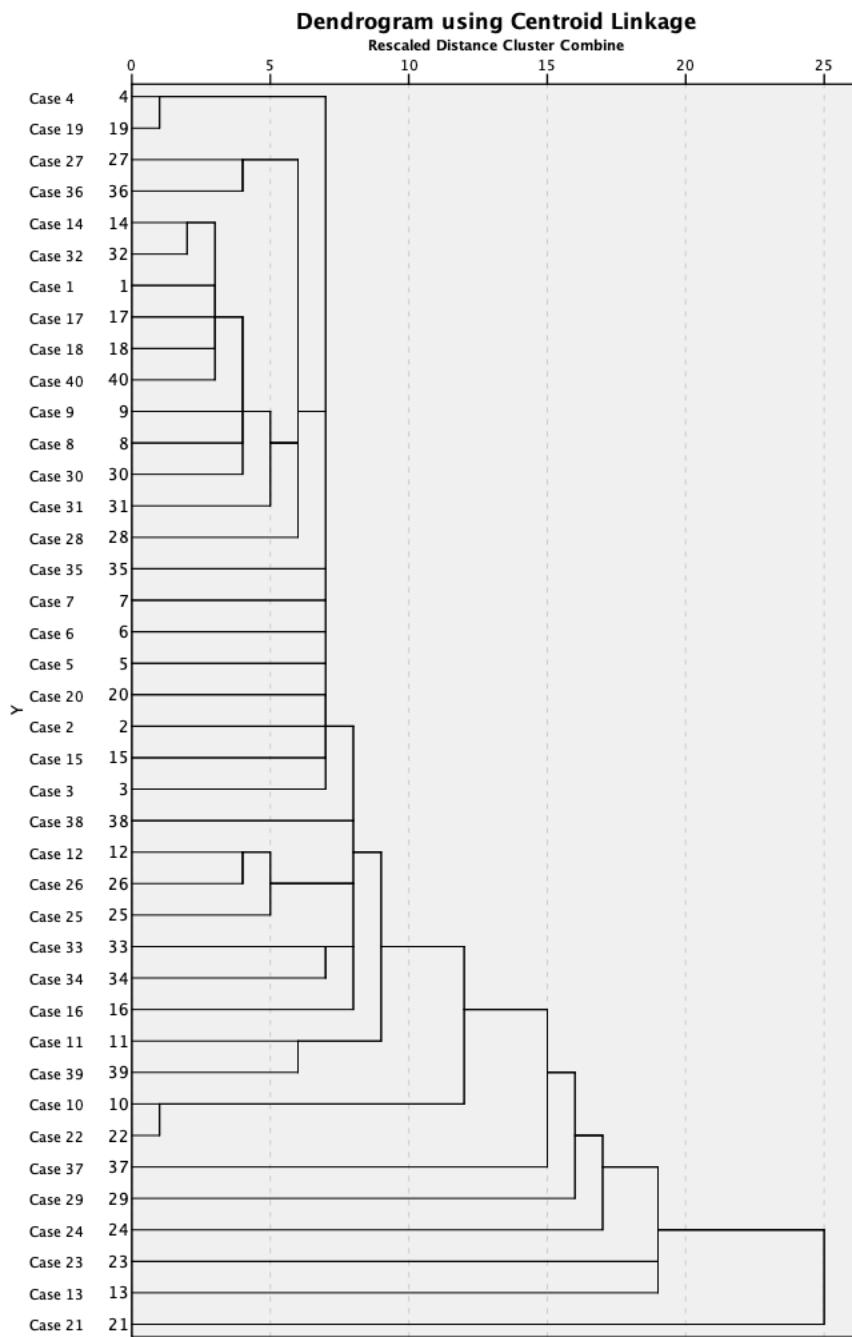
Bu kümeleme metodumuz ile açıklanabilirliği yüksek bir grafik elde edemediğimiz için tercih etmiyoruz.

MEDİAN CLUSTERING METHOD:



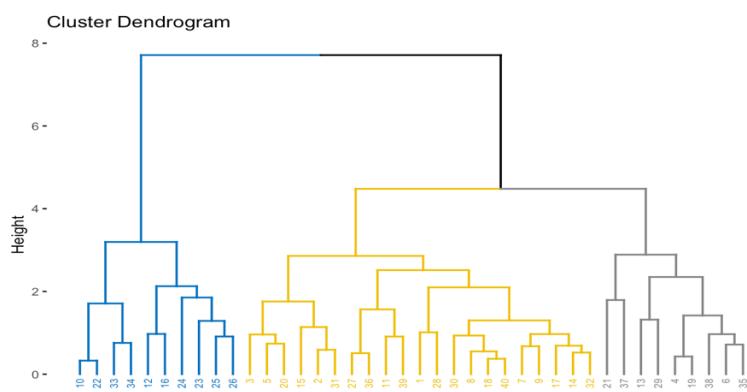
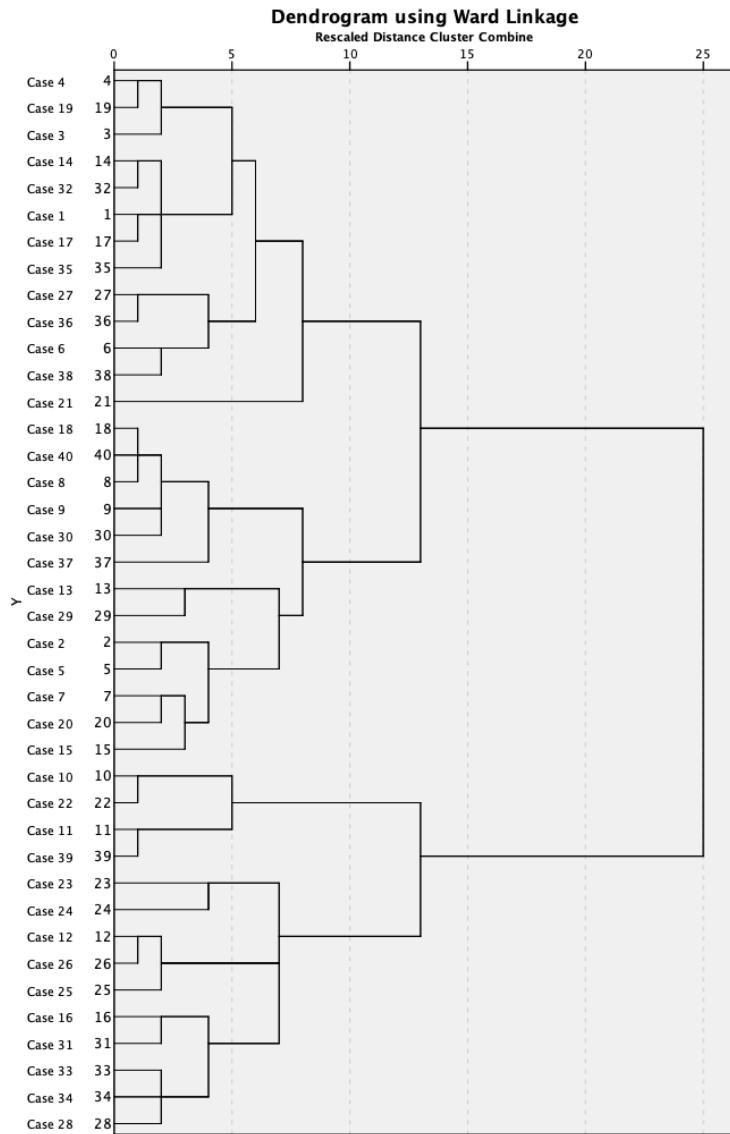
Bu kümeleme metodumuz ile açıklanabilirliği yüksek bir grafik elde edemediğimiz için tercih etmiyoruz.

CENTROID CLUSTERING METHOD



Bu kümeleme metodumuz ile açıklanabilirliği yüksek bir grafik elde edemediğimiz için tercih etmiyoruz.

WARD'S METHOD



Ward's method yöntemi ile 3 küme oluşturabildiğimizi rahatlıkla söyleyebiliyoruz.

Gerçekleştirdiğimiz hiyerarşik kümemeleme ile k-means değerimizin 3 olduğu sonucuna varıyoruz.

Hiyerarşik Olmayan Kümemeleme Analizi:

K means değerimizi 3 alarak aldığımız outputları yorumluyoruz.

Initial Cluster Centers

	Cluster		
	1	2	3
Protein1	-.64398000	.33912000	-1.2517000
Protein2	-.5936300	1.3193000	2.6739000
Protein3	-.1097400	.5874000	-.5358000
Protein4	-1.6028000	.35192000	-1.1071000

Rastgele seçilen kümelerin merkezlerini verir

Iteration History^a

	Change in Cluster Centers		
Iteration	1	2	3
1	.908	.909	1.121
2	.363	.041	.369
3	.219	.046	.197
4	.000	.000	.000

- a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 4. The minimum distance between initial centers is 2.785.

Merkezlerin kaç iterasyonda oturduğuna iteration history tablomuzdan bakıyoruz. Gözlemler yerleşene kadar bu işlem devam eder.

Final Cluster Centers

	Cluster		
	1	2	3
Protein1	-.21692560	-.04139062	-.09865226
Protein2	-.5536050	.9593474	2.0869889
Protein3	.2916350	-.2152076	-.4680714
Protein4	-.30604740	.25807705	-.30572744

4 farklı protein değişenimizin 3 kümeye dağıldığını gözlemlediğimiz bu tablodan,

1.kümemizin Protein3 değişkenimiz bakımından en zengin küme olduğunu görüyoruz.

2.kümemizde ise en çok bulunan proteinimizin Protein4 olduğunu söyleyebiliriz.

3.kümemizde en fazla bulunan değişkenimiz ise Protein2'dir.

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Protein1	.104	2	.177	37	.591	.559
Protein2	16.857	2	.203	37	82.942	.000
Protein3	1.475	2	.243	37	6.065	.005
Protein4	1.586	2	.388	37	4.090	.025

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Elde ettiğimiz 3 kümemiz artık bizim kategorik değişkenlerimiz olmuş oluyor.

H0: Kümelere göre değişkenlerinin ortalamaları anlamlı fark göstermemektedir.

Ha: Kümelere göre değişkenlerinin ortalamaları anlamlı fark göstermektedir.

Protein1 ve Protein4 değişkenimizin sig değerleri sırasıyla $0.559 > 0.05$ ve $0.025 < 0.05$ olduğu için, H₀ kabul yani Protein1 ve Protein4, kümelere göre değişkenlerinin ortalamaları anlamlı fark göstermemektedir.

Protein2 ve Protein3 değişkenlerimizin sig değerleri sırasıyla $0.00 < 0.05$ ve $0.005 < 0.05$ olduğu için, H₀ red yani Protein2 ve Protein3, kümelere göre değişkenlerinin ortalamaları anlamlı fark göstermektedir.

Yani Protein2 ve Protein3 kümelemede etkili birer değişkenken, Protein1 ve Protein4 değişkenlerimiz kümelememizde etkili değildir.

Number of Cases in each Cluster

Cluster	1	10.000
	2	21.000
	3	9.000
Valid		40.000
Missing		.000

```
DATASET DECLARE D0.588681832189169.  
PROXIMITIES Protein1 Protein2 Protein3 Protein4  
/MATRIX OUT(D0.588681832189169)  
/VIEW=CASE  
/MEASURE=EUCLID  
/PRINT NONE  
/STANDARDIZE=VARIABLE Z.
```

Kümelerdeki gözlem sayılarına bakmak için number of cases in cluster tablomuza bakıyoruz.

1.kümemizde 10

2.kümemizde 21

3.kümemizde 9 gözlem bulunmaktadır.