Mats Hansson – 830927

STAN41 HT2014
Examinator: Professor Björn Holmquist

Multivariate Analysis

Final project

Contents

Inledning

I detta projekt kommer jag att analysera om det finns någon skillnad och likheter mellan olika hundraser. Till mitt förfogande har jag 9 variabler, där variablerna innefattar storleken på olika käkben. För respektive variabler finns det 60 observationer. Två av dessa variabler är kategorivariabler, där den ena är hundraser, där jag har fem olika hundraser samt två hundar som är okända. För variabeln som innefattar hundraser finns det fyra raser som lever idag, vilka är Modern dog, Indian Wolf, Coens och Golden Jacka, där det finns 12 observationer för varje hundras. Den andra kategorivariabeln är kön, där vi har 6 stycken tikar och 6 stycken hanar för varje hundras. Inledningsvis kommer jag analysera data och därefter presenteras modeller.

Analys av data:

Inledningsvis kommer jag att analysera variablernas fördelning samt att ta fram relevanta nyckeltal. Därefter kommer det testas om variablerna är normalfördelade eller ej, dem variablerna som ej är normalfördelade kommer att logaritmeras. Här nedan presenteras mina medelvärden och standardavvikelser för variablerna samt för respektive hundras:

Medelvärden för varje variablerna								
x1	**x2**	**x3**	**x4**	**x5**	**x6**	**x7**	**x8**	**x9**
129.8	10.10	22.08	21.77	3.019	8.112	3.481	3.622	6.158

Medelvärden för hundraserna									
Hundraserna	**x1**	**x2**	**x3**	**x4**	**x5**	**x6**	**x7**	**x8**	**x9**
Cuons	134.28	10.91	24.31	23.70	3.07	8.50	3.37	3.63	6.66
Golden jackals	110.57	8.05	18.41	16.86	2.90	6.83	3.41	3.50	4.83
Indian wolves	156.13	11.45	25.89	24.45	3.20	9.43	3.69	3.69	7.29
Modern dogs	124.76	9.82	21.38	21.12	2.97	7.65	3.47	3.59	5.86
Prehistoric Thai dogs	122.79	10.33	19.91	22.98	2.95	8.19	3.49	3.58	6.17
unknown	124.30	9.95	23.00	21.50	2.99	7.95	3.40	3.61	6.10

Det verkar finnas en skillnad mellan hundraserna där vi kan se att Indian wolf har ett högre medelvärde i jämförelse till de andra hundraserna. För Golden Jackals verkar det vara tvärtom. Här nedan presenteras kovariansmatrisen:

Kovariansmatris för variablerna									
	x1	**x2**	**x3**	**x4**	**x5**	**x6**	**x7**	**x8**	**x9**
x1	309.07	19.34	54.43	45.69	40.17	15.72	55.20	76.25	14.71
x2	19.34	1.87	3.76	4.00	2.69	1.17	3.02	4.59	1.15
x3	54.43	3.76	13.00	8.83	7.05	2.67	6.67	11.89	2.49
x4	45.69	4.00	8.83	10.99	5.75	2.67	6.35	10.63	2.58
x5	40.17	2.69	7.05	5.75	6.10	2.17	7.79	10.14	2.09
x6	15.72	1.17	2.67	2.67	2.17	1.01	2.93	3.78	0.88
x7	55.20	3.02	6.67	6.35	7.79	2.93	17.74	15.15	2.55
x8	76.25	4.59	11.89	10.63	10.14	3.78	15.15	20.84	3.65
x9	14.71	1.15	2.49	2.58	2.09	0.88	2.55	3.65	0.93

Det som står i diagonalen är variansen för respektive variabel, de övriga värdena är kovariansen mellan variablerna. I övriga är det svårt att se några likheter i variansen för respektive variabler.

Standradavivkelsen för hundrasern									
Hundraser	**x1**	**x2**	**x3**	**x4**	**x5**	**x6**	**x7**	**x8**	**x9**
Cuons	6.58	0.61	1.62	1.66	0.04	0.39	0.05	0.06	0.49
Golden jackals	4.23	0.44	1.11	0.94	0.04	0.51	0.04	0.03	0.28
Indian wolves	13.11	0.90	4.01	1.98	0.05	0.50	0.05	0.07	0.66
Modern dogs	9.21	0.85	2.29	1.73	0.05	0.45	0.05	0.08	0.47
Prehistoric Thai dogs	8.50	0.76	1.89	2.87	0.05	0.74	0.07	0.05	0.48
unknown	15.13	0.78	2.83	4.24	0.08	0.64	0.06	0.08	1.13

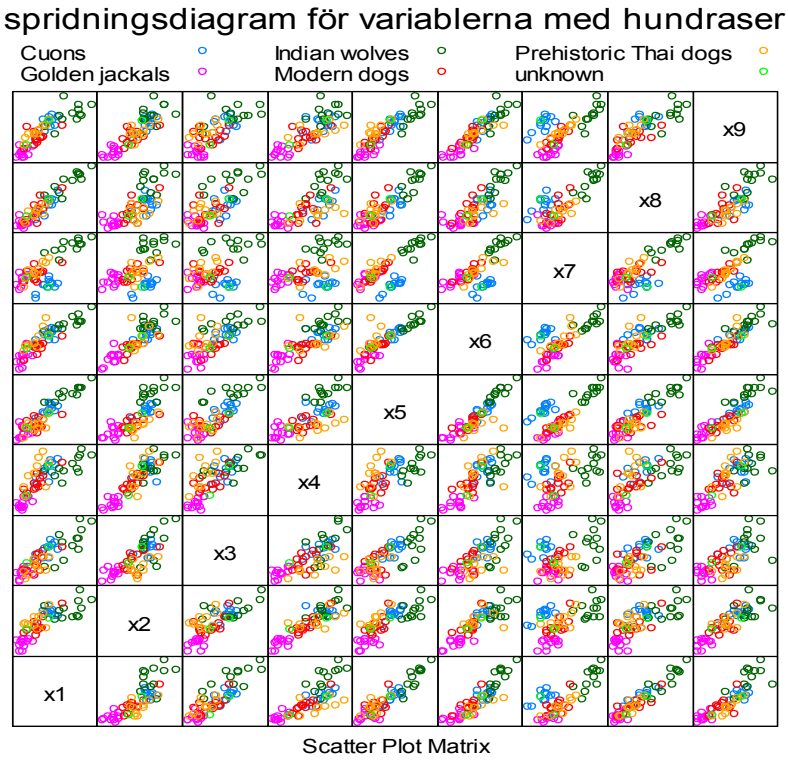
Vi kan även se att det finns en skillnad mellan standardavvikelserna, där vargen har en högre standardavvikelse än de övriga. Golden Jackals verkar även här har den lägsta standardavvikelse för hundraserna. I övriga är det svårt att urskilja något mönster. Korrelationsmatrisen ser ut följande:

Korrelationsmatrisen för variablerna

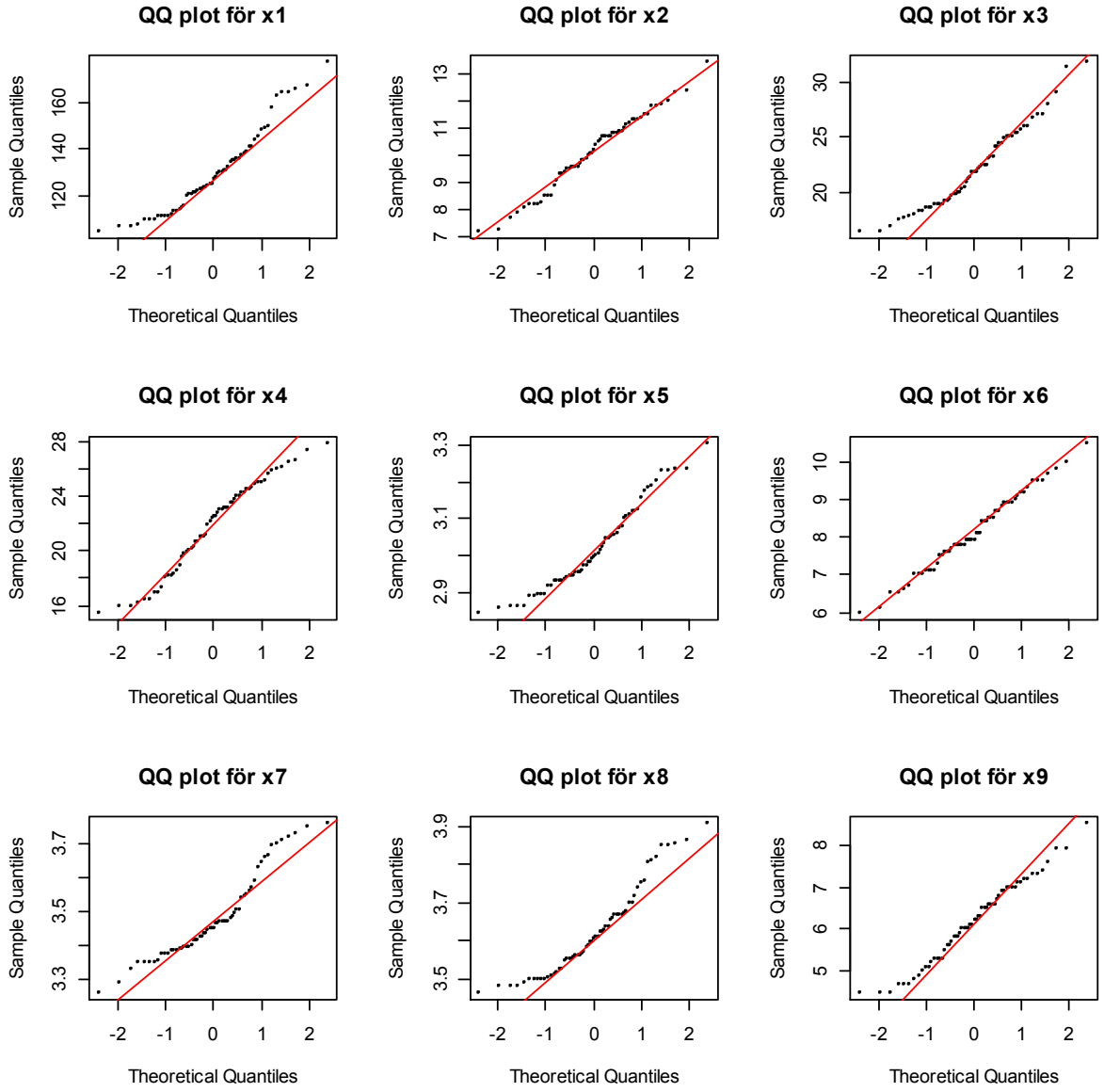
	x1	x2	x3	x4	x5	x6	x7	x8	x9
x1	1.00	0.80	0.86	0.78	0.93	0.89	0.75	0.95	0.87
x2	0.80	1.00	0.76	0.88	0.80	0.85	0.52	0.73	0.87
x3	0.86	0.76	1.00	0.74	0.79	0.74	0.44	0.72	0.72
x4	0.78	0.88	0.74	1.00	0.70	0.80	0.45	0.70	0.81
x5	0.93	0.80	0.79	0.70	1.00	0.88	0.75	0.90	0.88
x6	0.89	0.85	0.74	0.80	0.88	1.00	0.69	0.82	0.91
x7	0.75	0.52	0.44	0.45	0.75	0.69	1.00	0.79	0.63
x8	0.95	0.73	0.72	0.70	0.90	0.82	0.79	1.00	0.83
x9	0.87	0.87	0.72	0.81	0.88	0.91	0.63	0.83	1.00

Fördelen med att använda sig av korelationsmatrisen är att diagonalen är alltid 1. Enligt tabellen ovan verkar det finnas en stark korrelation mellan variablerna, där x1 och x8 har en korrelation på 0.95. Även de övriga variablerna har en stark korrelation.

Här nedanför presenteras mitt spridningsdiagram över hundraserna:

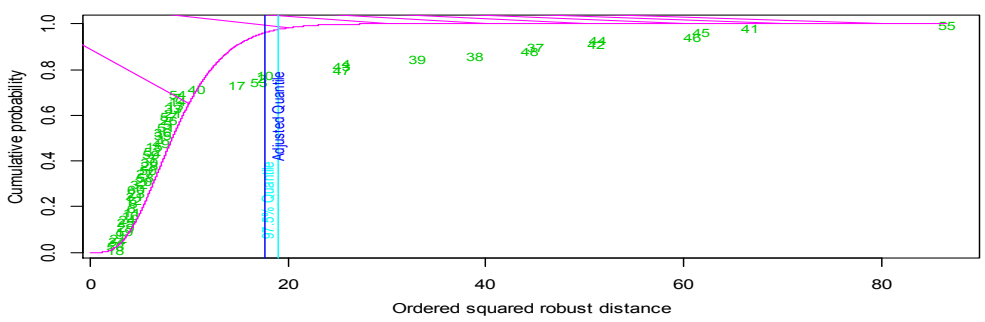


Bilden illusterar hur variablerna är korrelerade samt hur hundraserna är placerade för respektive variabel. Vi kan se i bilden att varje hundras har ett eget mönster, där vi kan se att vargen är placerat längt upp till höger och Modern dogs är placerat längst ner till vänster. Här nedan presenteras qq-plotar för respektive variabler.



Enligt jarque bera testet är x1,x2,x3,x4,x6 och x9 normalfördelade. De övriga är ej normalfördelade enligt Jarqu bera testet, jag valde att logatimera dessa variabler. Efter logaritmering blev samtliga variabler normalfördelade, visserligen fick jag ett lågt p-värde för x7 där p= 0.04738 jag valde att gå i min studie där jag har ett större urval än 30.

Vi kan se att alla variablerna har ett linjärt samband där vi även kan se att varje hundras är klustrade enskilt. Detta gäller för samtliga variabler. Här nedan presenteras våra utliggare för våra 60 observationer med dess 9 variabler.



obs	outliers											
[1]	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
[13]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
[25]	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
[37]	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
[49]	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE

Där: 1 – 12 är Modern dog, 13-24 är Golden jackals, 25-36 är Cuons
37 – 48 är Indian wolves, 49-58 är Prehistoric Thai dogs, 59-60 är unknown

Chi-två testet indikerade på att observationerna för vargen kan betraktas som uteliggare. Där även en förhistorisk hund kan betraktas som uteliggare. Jag valde att behålla dessa uteliggare med hänsyn till att alla uteliggare är för samma hundras

Analysdelen kommer att behandla både supervised och unsupervised modeller, där jag kommer att jämföra om modellerna predikterar samma för hundraserna samt för kön. Jag kommer först att presentera mitt MANOVA test för hundraserna och kön, därefter övergår jag till PCA analysen, faktor analys och LDA. Till sist kommer jag presentera hierarkiska träd över data.

MANOVA

I den här delen kommer jag testa om det finns skillnader mellan hundraser samt kön med MANOVA och ANOVA tester. Först kommer jag att testa om hela modellen är signifikant skild från noll, därefter testar jag vilka hundraser som skiljer sig åt. Här nedanför presenteras mitt MANOVA test

MANOVA test för hundraser som faktor			
	DF	**Pillai**	**P-värde**
	4	2.6309	<0.001
Residualer	53		

Vi kan se att hela modellen är skild noll med ett lågt p-värde där jag testar om det finns en skillnad mellan hundraserna. Nedanför presenteras ANOVA tester för respektive variabeler för hundraserna.

P-värden för ANOVA med hundraser som faktorer										
Respons variabel	**M/G**	**M/C**	**M/I**	**M/P**	**G/C**	**G/I**	**G/P**	**C/I**	**C/P**	**I/P**
x1	<0.001	0.0080	<0.001	0.6111	<0.001	<0.001	<0.001	<0.001	0.0019	<0.001
x2	<0.001	0.0016	<0.001	0.160	<0.001	<0.001	<0.001	0.0995	0.061	0.005
x3	<0.001	0.0014	0.0026	0.122	<0.001	<0.001	0.0310	0.2175	<0.001	<0.001
x4	<0.001	0.0012	<0.001	0.122	<0.001	<0.001	<0.001	0.3254	0.470	0.171
x5	0.0014	0.0014	<0.001	0.512	<0.001	<0.001	0.0172	<0.001	<0.001	<0.001
x6	<0.001	<0.001	<0.001	0.047	<0.001	<0.001	<0.001	<0.001	0.222	<0.001
x7	0.0120	<0.001	<0.001	0.446	0.0361	<0.001	0.0070	<0.001	<0.001	<0.001
x8	<0.001	0.1983	<0.001	0.543	<0.001	<0.001	<0.001	<0.001	0.030	<0.001
x9	<0.001	<0.001	<0.001	0.138	<0.001	<0.001	<0.001	0.0141	0.029	<0.001

M=Modern dogs
G=Golden jackals
C=Cuons
I=Indian wolves
P=Prehistoric Thai dogs

Resultatet från MANOVAN indikerade på ett signifikant resultat, vilket indikera på att det finns någon skillnad mellan hundraserna, problemet är att det ej går att se vilken ras som skiljer sig åt.
Vid en närmare analys mellan respektive hundras så finns det ingen skillnad mellan Modern dogs och Prehistoric Thai dogs, vilket indikera på att Modern dogs är nära besläktat med Prehistoric Thai dogs. Det verkar även finnas en del likheter mellan Couns och Prehistoric dog enligt ovanstående tabell, där vi kan se att p-värden är markant högre mellan dessa hundarter.

MANOVA Full model för kön			
	DF	**Pillai**	**P-värde**
		0.0869	0.9188
Residuals			

P-värden för ANOVA med kön som faktor				
		Male/Female		
Respons variabel	**Modern dog**	**Golden jackals**	**Cuons**	**Indian wolves**
x1	0.0895	0.0583	0.4184	0.1701
x2	0.0229	0.1617	0.4054	0.0091
x3	0.3299	0.8273	0.6722	0.0545
x4	0.0134	0.3985	0.7132	0.0173
x5	0.0134	0.0684	0.3579	0.0225
x6	0.0136	0.0054	0.3222	0.090
x7	0.0065	0.0063	0.2266	0.5605
x8	0.1048	0.6902	0.5824	0.247
x9	0.1259	0.0194	0.3745	0.0418

Enligt MANOVAN finns det ingen skillnad mellan kön där mitt test indikera på ett högt p-värde. Vid en närmare analys mellan kön där hundrasen är inkluderat finns det en signifikant skillnad för Modern dog för variablerna x2,x4,x5,x6 och x7. I övrigt tyder det på att det finns knappt någon skillnad för kön.

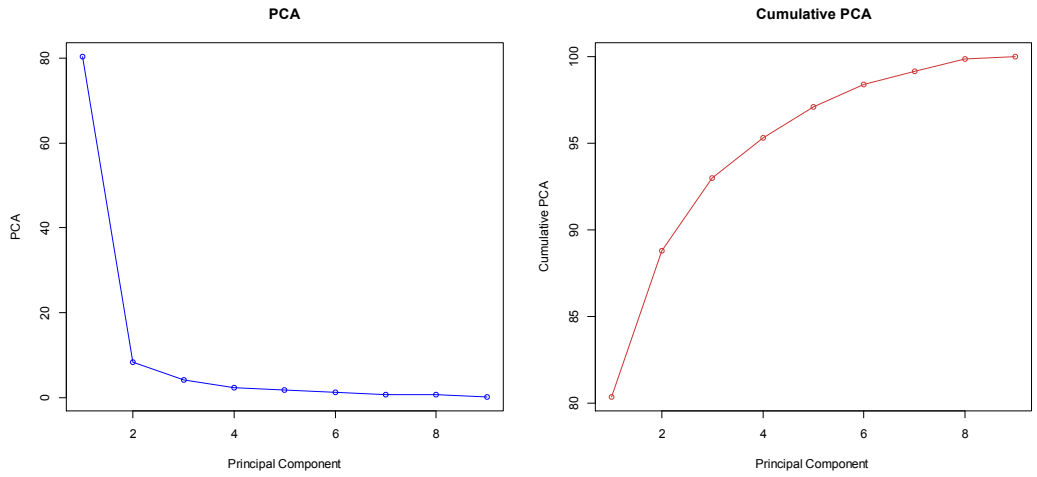
PCA OCH FAKTOR ANALYA:

ANALYS AV HUNDRASER

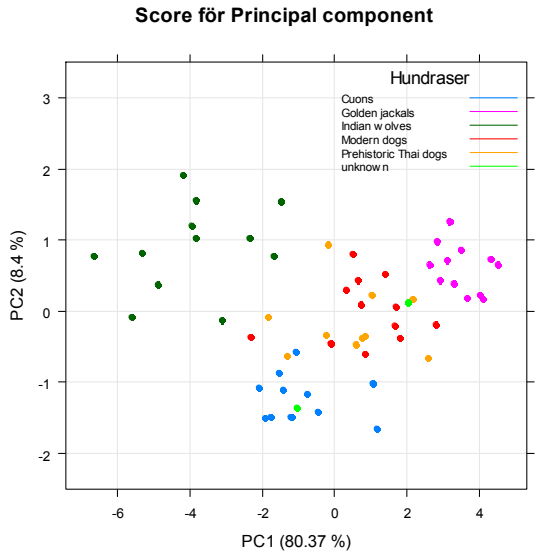
I denna del kommer vi att försöka förklara variationen i modellen samt att analysera korrelationen mellan variablerna. Jag kommer första att definiera mina Principal component. Här nedan presenteras mina principal component, där jag kommer att använda mig av standard deviation för att avgöra hur många PCA komponenter som kommer att innefattas i analysen.

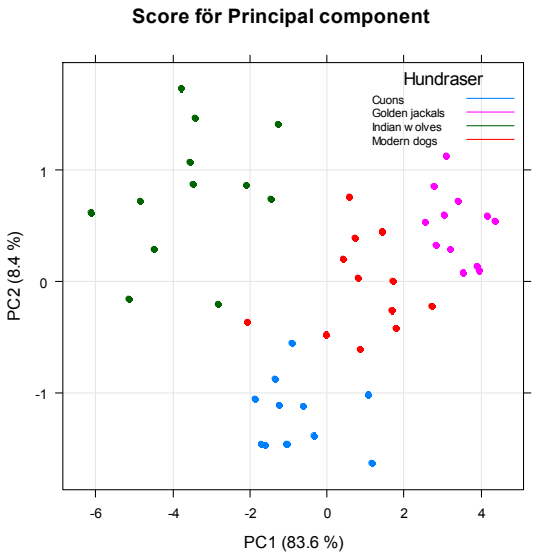
	Principal component								
	Comp.1	**Comp.2**	**Comp.3**	**Comp.4**	**Comp.5**	**Comp.6**	**Comp.7**	**Comp.8**	**Comp.9**
Standard deviation	2.689	0.870	0.612	0.459	0.399	0.339	0.263	0.251	0.119
Proportion ofVariance	0.803	0.084	0.041	0.023	0.017	0.012	0.007	0.007	0.002
Cumulative Proportion	0.803	0.888	0.929	0.953	0.970	0.983	0.991	0.998	1

	Loadings:								
	Comp.1	**Comp.2**	**Comp.3**	**Comp.4**	**Comp.5**	**Comp.6**	**Comp.7**	**Comp.8**	**Comp.9**
x1	-0.360		-0.293		0.199	0.153	-0.175		0.822
x2	-0.335	-0.312	0.328		-0.313	-0.614	-0.426	-0.106	
x3	-0.312	-0.331	-0.696	0.158	-0.338			0.265	-0.305
x4	-0.319	-0.405	0.342	0.559	0.257	0.199	0.441		
x5	-0.352	0.122	-0.165	-0.393	-0.113	-0.291	0.614	-0.452	
x6	-0.351		0.229	-0.262	-0.261	0.658	-0.305	-0.366	-0.165
x7	-0.268	0.739	0.148	0.415	-0.352			0.222	
x8	-0.345	0.239	-0.160		0.677	-0.185	-0.314	-0.143	-0.431
x9	-0.348		0.280	-0.503	0.140		0.128	0.712	



Vi kan se att den första komponenten förklara 80,3 % av variationen medan den andra komponenten förklara 8,4 %, jag väljer att gå vidare med 2 principal komponent där dem förklarar 88,8 % av variationen. Här nedan presenteras mina scorepoäng för modellen.

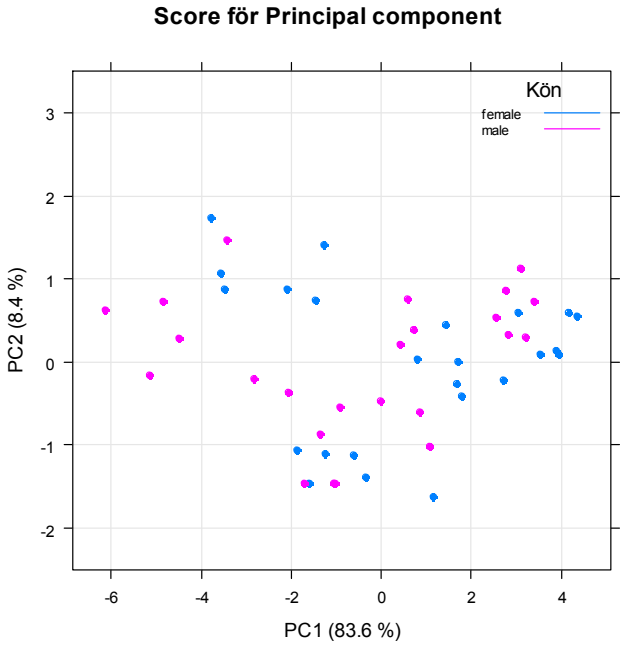


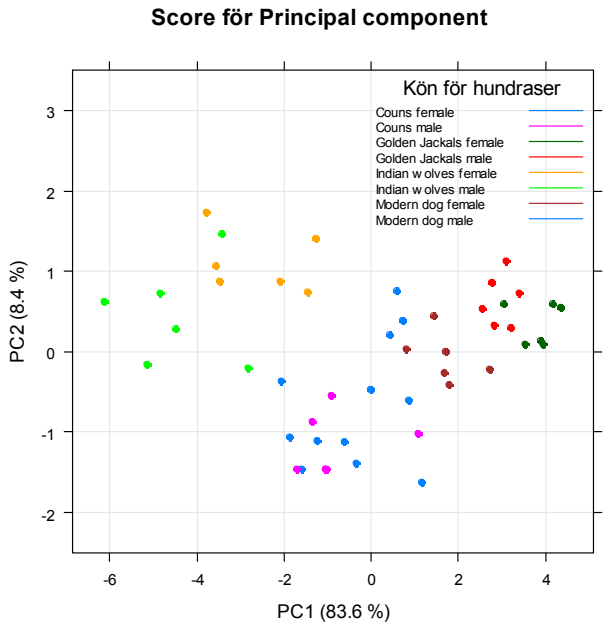


Varje färg är för respektive hundras, där vi kan se att det bildas tydliga grupper för hundraserna. I den vänstra bilden är alla hundraserna med inklusive de okända hundarna, där vi kan se att den förhistoriska hunden är klustrad med Modern dog. Om vi väljer att exkludera den förhistoriska hunden och de okända hundraserna så kan vi se att vår första komponent är starkare än den tidigare, vilket betyder att den förhistoriska hunden kan förklara 3.3 procentenheter. Den andra komponenen är nästan identisk.

ANALYS AV KÖN

I den här delen kommer jag att analysera om det går att urskilja något mönster baserat på kön. Jag valde att endast inkludera Modern dogs, Couns, Golden Jackals och Indian wolf i min modell. Här nedan visualiseras könet baserat på hundrasen.





Vid en visuell analys verkar det finnas en skillnad mellan kön när hundrasen är inkluderat. Där vi kan se att varje hanhund har ett högre snittvärde för PC1. För Hundrasen Couns verkar dessa kriterier ej uppfyllas.

Faktor analys

I den här delen kommer jag att använda mig av faktor analys, vilket betyder att jag kommer att analysera förhållandet mellan de uppenbara variablerna utan att göra några antagande om vilka uppenbara variabler som är relaterade till vilka faktorer.

Jag kommer första att testa hur många faktorer som ska ingå i modellen genom ett chi-två test, där testet indikerar på att 5 faktorer ska inkluderas i modellen. Modellen ser ut följande:

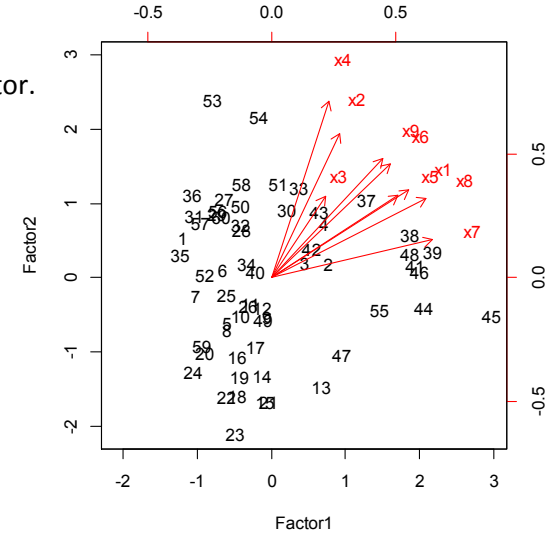
Uniquenesses								
x1	**x2**	**x3**	**x4**	**x5**	**x6**	**x7**	**x8**	**x9**
0.005	0.110	0.005	0.009	0.083	0.065	0.278	0.017	0.076

The Uniqunesses är den unika variansen och avser variabiliteten för respektive variabel, där den ej delas med de övriga variablerna. Den kommer senare att användas för att definiera residualerna för modellen.

	Factor1	**Factor2**	**Factor3**	**Factor4**	**Factor5**
SS loadings	3.114	2.745	1.908	0.500	0.084
Proportion var	0.346	0.305	0.212	0.056	0.009
Cumulative var	0.346	0.651	0.863	0.919	0.928

Tabellen ovan förklarar standardavvikelsen för respektive faktor.

Loadings:	**Factor1**	**Factor2**	**Factor3**	**Factor4**	**Factor5**
x1	0.694	0.442	0.555		
x2	0.346	0.727	0.384	0.305	
x3	0.271	0.413	0.858	0.126	
x4	0.289	0.889	0.340		
x5	0.640	0.414	0.468	0.336	
x6	0.601	0.575	0.345	0.322	-0.145
x7	0.811	0.190	0.125		
x8	0.782	0.401	0.388		0.228
x9	0.561	0.601	0.313	0.386	



Alla loadings för den första faktorn är positiva, vilket betyder att det finns en korrelation mellan variablerna. Våra variabler innefattar storleken på käkben, vilket kan tolkas som att vår första faktor kan vara storleken på käkbenen. Den andra faktorn har positiva värden för variablerna, vilket betyder att pilarna kommer att riktas uppåt åt höger, där x2 och x4 påverkar mer av den andra faktorn. Faktor 3 lyder under samma villkor som de två första, fast med lägre positiva värden. Faktor 4 påverkas av x2,x3,x5,x6 och x9, vilket betyder att de övriga variablerna understiger värdet 0.1.

Vi kan även se att x5 och x1 är nästan perfekt korrelerade, vilket kan även ses i korrelationsmatrisen. Detta kan betyda att jag bör exkludera antingen variabel x1 eller x5. Med faktor analysen kan vi analysera korrelation mellan respektive variabel, detta är ett bra verktyg för att se vilka variabler som bör ingå i den fulla modellen.

Här nedan presenteras residualmatrisen för modellen:

Residualmatris för modellen.

	x1	x2	x3	x4	x5	x6	x7	x8	x9
x1	0.000	-0.001	0.000	0	-0.001	0.000	0.000	0.000	0.001
x2	-0.001	0.000	0.000	0	0.002	0.000	0.021	0.000	-0.002
x3	0.000	0.000	0.000	0	0.000	0.000	-0.001	0.000	0.000
x4	0.000	0.000	0.000	0	0.000	0.000	0.000	0.000	0.000
x5	-0.001	0.002	0.000	0	0.000	-0.003	0.026	0.000	0.001
x6	0.000	0.000	0.000	0	-0.003	0.000	0.001	0.000	0.002
x7	0.000	0.021	-0.001	0	0.026	0.001	0.000	-0.001	-0.030
x8	0.000	0.000	0.000	0	0.000	0.000	-0.001	0.000	0.000
x9	0.001	-0.002	0.000	0	0.001	0.002	-0.030	0.000	0.000

Där den kvadratiska residualsumman är 0.004, och summan av de övriga faktorerna skattas till 0.567. Vilket betyder att modellen som innefattar 5 faktorer är en bra modell, där vi kan se att den kvadratiska residualsumman är mindre än dem uteslutande faktorerna som är kvadrerade.

Linear discriminant analysis

I den här delen analyseras vilken hundras som är nära besläktat med den förhistoriska hunden, samt att analysera hur könet klassificeras. Först kommer jag att klassificera hundraserna genom PCA - komponenterna och därefter kommer jag analysera hur de vanliga variablerna kan klassificera hundraserna.

Linear discriminant analysis för hundraser

Vi kan se i bilden från sida 7 att varje hundras klustras tillsammans, vilket bör indikera på att vi kan klassificera våra hundraser. I den vänstra bilden innehåller samtliga hundraser, där vi kan se att Modern dog bör härstammar från den förhistoriska hunden. Vid en visuell analys kan vi säga att modern dog härstammar från den förhistoriska hunden, fast utan något statistikstöd.

Jag kommer att använda mig av två mått när jag avgör hur bra modellen är dessa är $\text{Count } R^2 = \frac{number\ of\ correct\ predictions}{total\ number\ of\ observation}$ vilket betyder hur många observationer som klassificerar sig rätt, samt $\text{Count cross } R^2 = \frac{number\ of\ correct\ predictions}{total\ number\ of\ observation}$ för att krosvalidera modellen. Krosvalidering bygger på att vi utesluter en observation och testar den med de övriga observationer. Tabellen nedanför är för mina PCA modeller samt min variabelmodell.

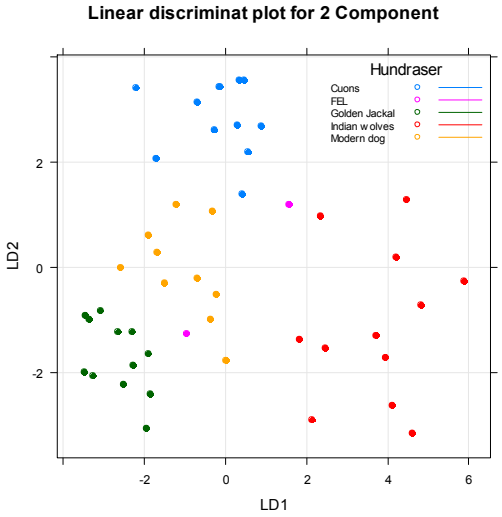
	Linear discriminant analysis	
	Count R^2	**Count cross R^2**
2 PCA	95,8%	95,8%
3 PCA	100 %	95,8%
9 PCA	100 %	97,9 %
Variablerna	100 %	97,9 %

	Proportion of trace:		
	LD1	**LD2**	**LD3**
2 PCA	0.6386	0.3614	-
3 PCA	0.5697	0.4303	0
9 PCA	0.6812	0.2781	0.0407
Variablerna	0.6812	0.2781	0.0407

Modellen som inkludera två PCA komponenter har en hög prediktions grad, där krossvalideringen indikerar på att modellen är robust.

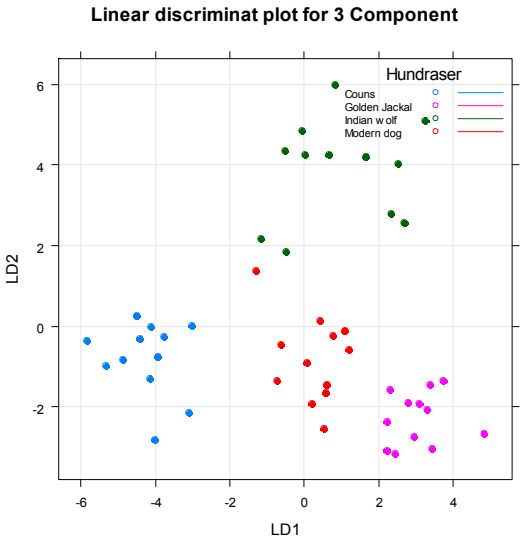
I tabellen kan vi också se att 3 PCA komponenter förklarar med samma tillförlitlighet som för alla 9 variabler. Krosvalideringen innehåller 1 felklassificering (47/48), vilket indikerar på en robust modell, där båda modellerna innehöll samma fel. Modellen som inkludera alla principal component har ett $\text{Count } R^2 = 100\ \%$ och korsvalideringen på $\text{Count } R^2 = \frac{47}{48}$,vilket betyder att modellen är lite robustare än den föregående. Vidare kan vi se att modellen för variablerna har samma proportion of trace som för 9 PCA, vilket betyder att dem kan förklara samma variation.

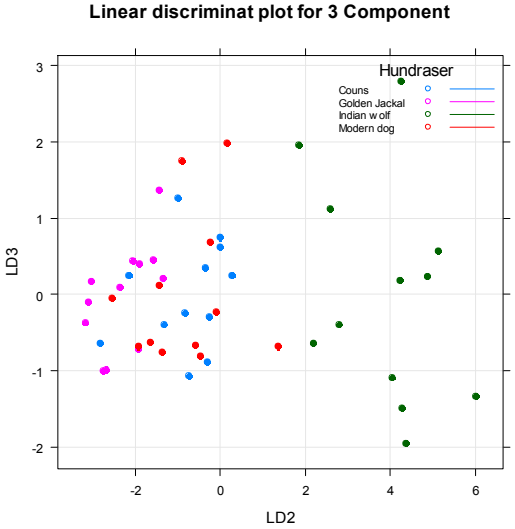
Visualisering för LDA med komponenterna 2:

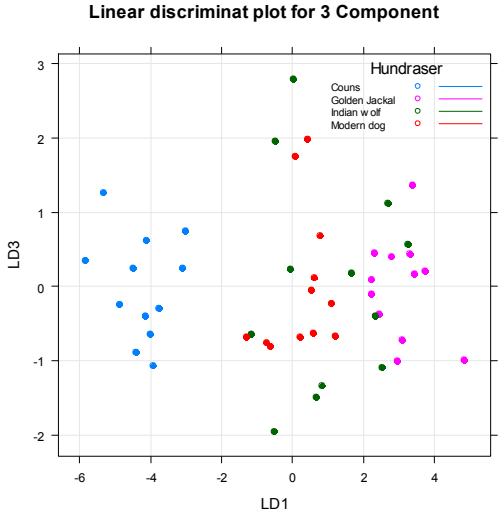


Enligt bilden ovan kan vi lätt se att två komponenter kan klassificera Modern dog, Couns, Golden Jackal och Indian wolf, vilket betyder att det bör räcka med två komponenter för att klassificera dessa hundraser. Det finns visserligen två prediktions fel, vilket betyder att det finns möjligheter att prediktera bättre när fler komponenter är inkluderade i modellen. Min korsvalidering tyder på att det finns två klassificeringsfel, vilket betyder att klustringen för hundraserna är ej robust.

Visualisering för LDA med komponenterna 3:

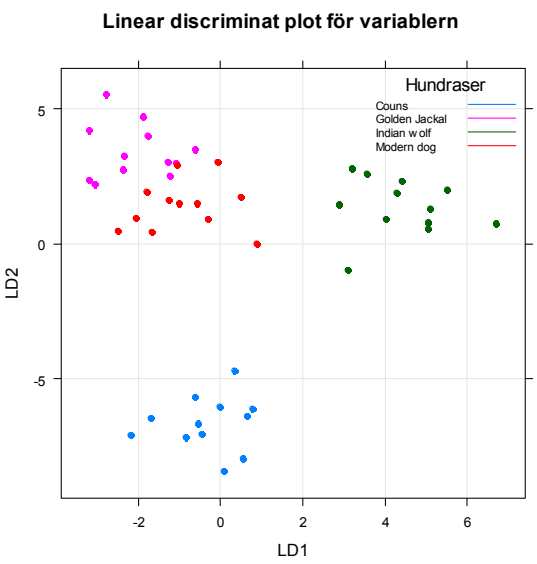


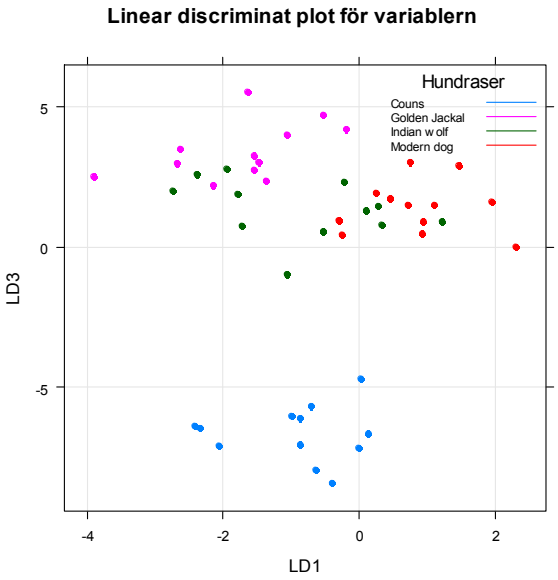


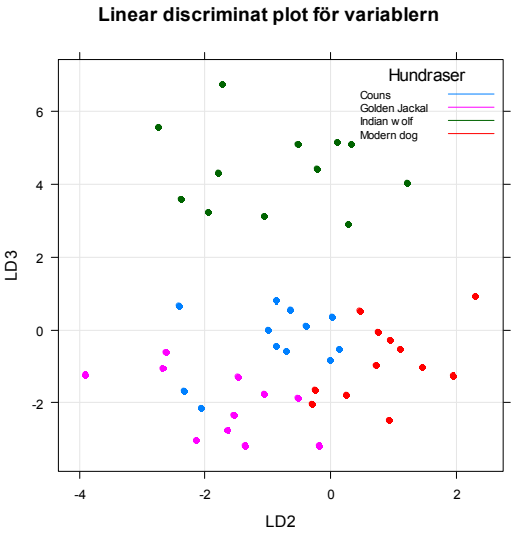


När tre komponenter är inkluderade i modellen får vi en bättre prediktion, vid en okulär besiktning för LD1 mot LD2, så bildas tydliga mönster för respektive hundras. Vi vet även att denna modell klassificera rätt 48/48.

Visualisering för LDA med variablerna:







När variablerna är inkluderade i modellen finns det tydliga kluster för respektive hundras, vilket betyder att vi kan dra slutsatsen att det finns skillnader mellan hundraserna. Visserligen verkar modellen ha problem med hitta skillnaden mellan Golden Jackals och Modern dog när vi granskar LD1 mot LD2 för variablerna.

Här nedan presenteras min prediktion för den förhistoriska hunden och den okända hundarten.

Prediktion för den förhistoriska hunden samt den okända								
	2 PCA		3 PCA		9 PCA		Variablerna	
	Prehistoric Thai dogs	Unknown	Prehistoric Thai dogs	Unknown	Prehistoric Thai dogs	Unknown	Prehistoric Thai dogs	Unknown
Cuons	1	1	4	1	6	1	0	1
Golden jackals	0	0	0	0	0	1	0	1
Indian wolves	0	0	0	0	0	0	0	0
Modern dogs	9	1	6	1	4	0	10	0

Modellen som är baserat på mina 9 variabler indikerar på att Modern dog är nära besläktat med den förhistoriska hunden. När prediktionen genomförs på PCA komponenterna förändrades resultatet beroende på hur många komponenter som innefattas i modellen. För den okända hundarten verkar samtliga modeller indikera på att den ena hunden är en Couns och den andra kan antingen vara en Modern dog eller Golden Jackals.

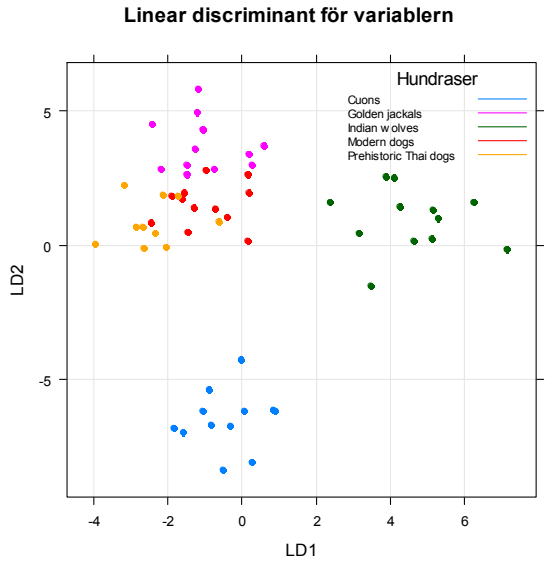
Unikheten hos den förhistoriska hunden

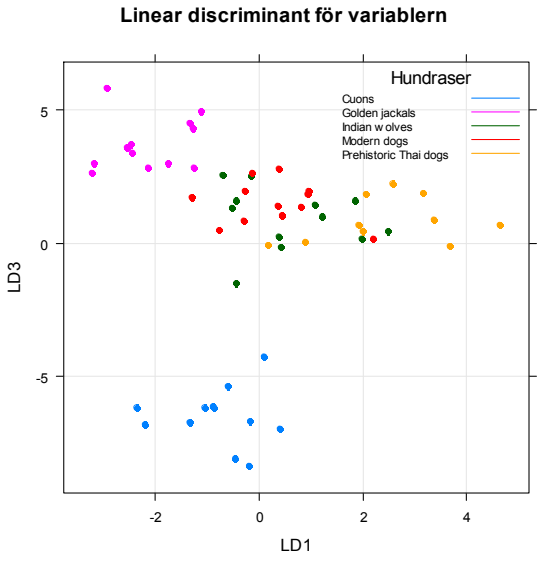
I den här delen kommer jag att analysera om den förhistoriska hunden har en egen klustrings förmåga, jag kommer att använda mig av variablerna för att skatta denna modell. Modell ser ut följande:

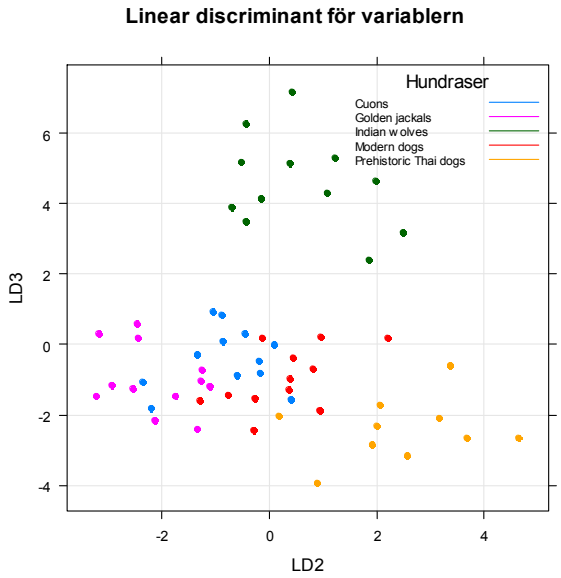
Linear discriminant analysis for variable				
	LD1	LD2	LD3	LD4
Proportion of trace	0.6055	0.2810	0.1068	0.0067

Count R^2	98,3 %
Count cross R^2	98,3 %

Vi kan se att modellen predikterar bra när vi ska klassificera alla hundraserna, där båda mina R värden är 98,3 %, vilket bör indikera på att variabelmodellen kan hitta en unik variation för den förhistoriska hunden. Här nedan illusterar min prediktion.







Den första bilden fångar upp variationen för Couns och Indian wolf. I den andra bilden är Golden Jackals för sig själv, vilket verkar stämma bra överens med tidigare studier. Det som är intressant är att den tredje bilden kan förklara den unika variationen hos den förhistoriska hunden. Detta kan förklaras med att den tredje bilden förklarar variationen på en lägre nivå än de övriga, där den hittar en unik variation för den förhistoriska hunden. Jag valde även att testa detta för PCA komponenterna, detta resulterade i ett sämre resultat.

Linear Linear discriminant för kön

Här nedan presenteras mina R^2 för samtliga modeller för kön där hundras är inkluderat.

Linear discriminant analysis för hundras som är baserat på kön				
	2 PCA	**3 PCA**	**9 PCA**	**9 variabler**
Count R^2	75 %	83,3 %	89,6 %	89.6%
Count kros R^2	64.6%	62.5%	60.4 %	60.4 %

Mina modeller har ett större prediktionsfel när vi ska klassificera kön för hundraserna. Mina PCA modeller verkar även ha ett sjunkande värde på krosvalideringen, vilket tyder på att osäkerheten i modellen ökar när flera PCA komponenter inkluderas i modellen. Enligt samtliga modeller verkar det finnas möjligheter att prediktera könet för samtliga hundraser, vilket bör indikera på att det finns en viss skillnad mellan hanar och honor. Här nedan presentera jag prediktionen för modellerna.

Prediktion för den förhistoriska hunden samt den okända								
	2 PCA		**3 PCA**		**9 PCA**		**Variablerna**	
	Prehistoric Thai dogs	**unknown**	**Prehistoric Thai dogs**	**unknown**	**Prehistoric Thai dogs**	**unknown**	**Prehistoric Thai dogs**	**unknown**
Cuons female	0	1	0	1	3	0	0	1
Cuons male	1	0	4	0	3	1	0	0
Golden jackals female	0	0	0	0	0	0	0	0
Golden jackals male	0	0	0	0	0	1	1	0
Indian wolves female	0	0	0	0	0	0	0	0
Indian wolves male	0	0	0	0	0	0	0	0

Modern dogs female	4	1	2	1	1	0	2	1
Modern dogs male	5	0	4	0	3	0	7	0

Prediktion för den förhistoriska hunden baserat på kön ger ett svårtolkat resultat. Dem två modellerna som verkar mest likartade är den med 2 PCA komponenter och den modellen som innefattar variablerna, vilket indikerar på att ett större antal komponenter bör icke inkluderas i modellen, där vi ser i tabellen att krosvaliderings värdet sjunker vid ett större antal komponenter. För de okända hundarna verkar samtliga modeller prediktera samma resultat, vilket bör betyda att den ena är Modern dogs female och den andra bör vara Couns female. För att testa om mina 2 PCA komponenter kan verkligen förklara skillnaden i kön, då borde MANOVA testet ge ett lågt p-värde. Här nedanför presenteras min MANOVA samt ANOVA för kön när hundrasen är inkluderad.

MANOVA test för kön som faktor			
	DF	**Pillai**	**P-värde**
	7	2.9537	<0.001
Residualer	40		

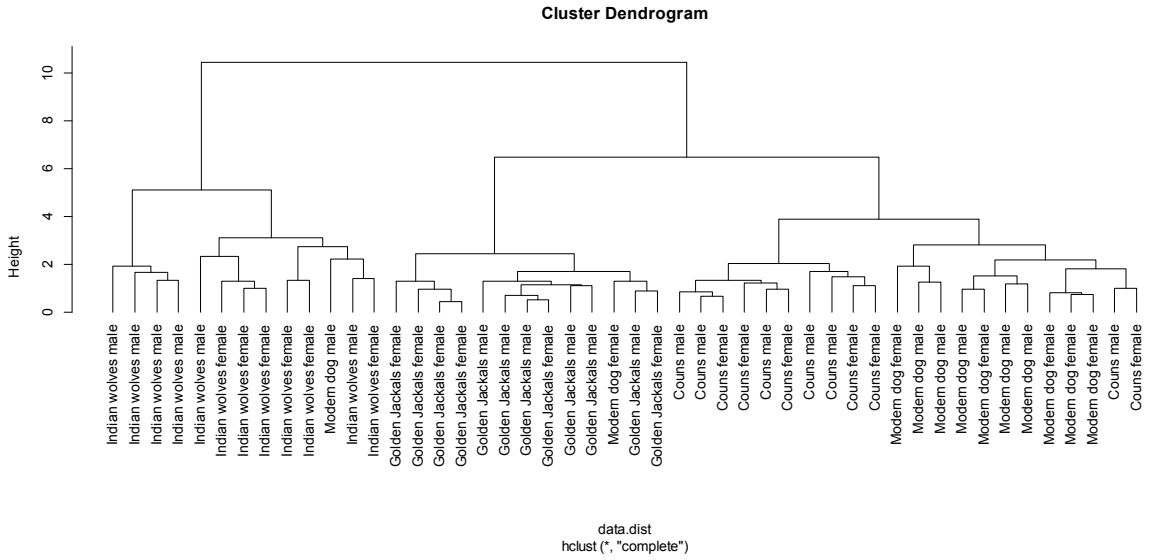
MANOVA testet indikerade på ett signifikant resultat, vilket tyder på att det kan finnas en skillnad mellan kön.

P-värden för ANOVA med kön som faktor				
	Male/Female			
Respons variabel	**Modern dog**	**Golden jackals**	**Cuons**	**Indian wolves**
Comp.1	0.0109	0.0047	0.9017	0.0190
Comp.2	0.841	0.1042	0.4054	0.0524

Enligt Pillai indikerade det på att det finns en skillnad för kön när vi inkluderar hundrasen i analysen. De hundraserna som skiljer sig för kön är Modern dog, Golden Jackals och Indian wolves. Couns fick ett högt p-värde, vilket indikera på att det finns ingen skillnad mellan hanar och honor.

Hierarchical clustering

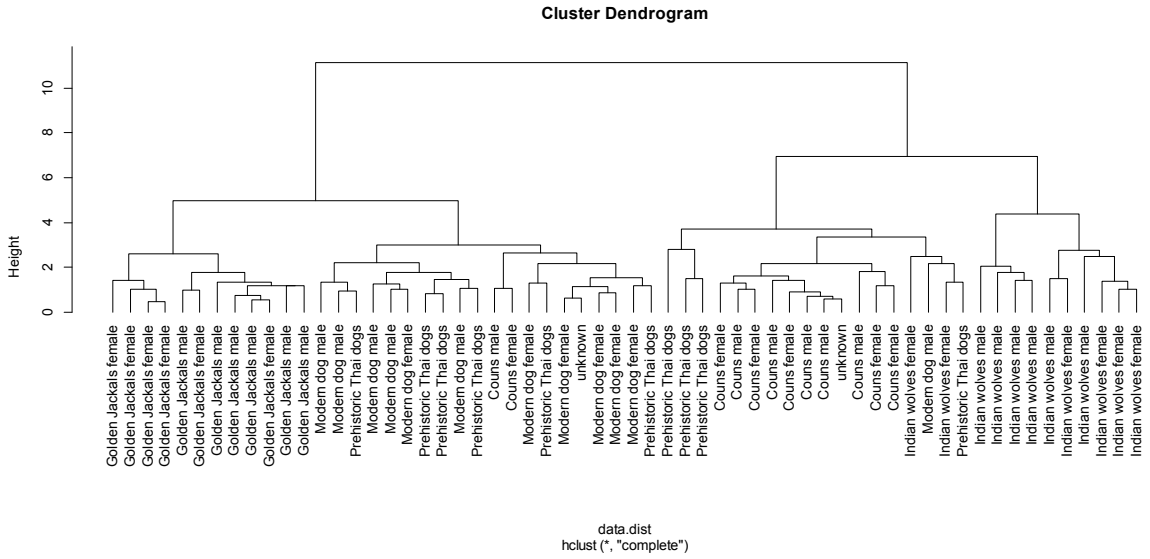
I den här delen kommer jag att använda mig hierarchical clustering. Jag kommer först att presentera mitt hierarkiska träd, där jag kommer att använda mig av ”complete method” som mäter distansen mellan observationerna. Här nedanför är trädet för hundraserna baserat på kön:



I Dendrogrammet ovan kan vi se att det bildas tydliga grupper. Om trädet beskärs när Height = 4, vilket betyder att grenen längst till vänster innehåller Indian wolf och en Modern dog, där vi kan se att kön är klustrat på en lägre gren.

Den andra grenen innehåller Golden Jackals och en Modern dog, det kan dock vara svårt att urskilja mönster för könet på nivån under. Medan den tredje grenen innehåller 10 Couns där det är svårt att urskilja kön på grenen under. Den fjärde grenen innehåller Modern dog och Couns och även här är det svårt att urskilja något mönster för könet.

Vi vet från mina score plottning för hundraser att det finns en Modern dog som placerar sig närma vargen, när jag utesluter denna observation från analysen blir även mitt Dendogram korrekt för grenen längst till vänster. Här nedanför presenteras mitt Dendogram för den förhistoriska hunden samt den okända:



Vi kan se att det finns 4 tydliga grupper i dendogrammet ovan, där samtliga hundraser har även tydliga grupper för könet på nivån under. När vi endast fokuserar på den förhistoriska hunden, är den grupperat med Modern dog (6/10) samt Couns (4/10). Detta kan betyda att den förhistoriska hunden är både besläktat med Modern dog och Couns. Problemet är att Modern dog och Couns skiljer sig kraftigt åt enligt dendrogrammet, där de delar sig redan på första grenen, en teori kan vara att alla hundarna härstammar från Prehistoric dog.

Dem okända hundrana placerade sig i grenen för Modern dog och Couns, vilket även styrker mina tidigare studier.

Slutsats:

Alla mina tester indikerar på samma resultat att det finns en skillnad mellan Modern dogs, Couns, Golden Jackals och Indian wolf. Där vi kan se att dem hundarna som skiljer sig mest åt är Golden Jackal och Couns. När vi sedan inkludera den förhistoriska hunden i analysen verkar samtliga tester indikera på att Modern dog är nära besläktat med den förhistoriska hunden. Enligt min PCA analys verkar även Couns också vara nära besläktat med den förhistoriska hunden, fast ej lika närma som Modern dogs. Medan Golden Jackal och Indian wolf är en mer avlägsnad släkting till den förhistoriska hunden. Vi bör även ha i åtanke att den förhistoriska hunden har ett eget unikt mönster enligt mina linear discriminant analys, vilket tyder på att Modern dog har också utvecklats en viss unik variation med tiden.

Det intressanta med det hierarkiska trädet är att den förhistoriska hunden är placerat tillsammans med Couns och Modern dogs. Detta betyder att de delar sig redan på den första ”avstyckningen”, vilket säger att Couns och Modern dog ej är nära besläktat.

Den enda tolkningen jag kan göra som statistiker är att alla dessa hundar har samma anfader, fast Modern dog har ej förändrats lika mycket som de övriga hundarna. Couns har också förändrats med tiden fast inte lika mycket som Indian wolf och Golden Jackals.