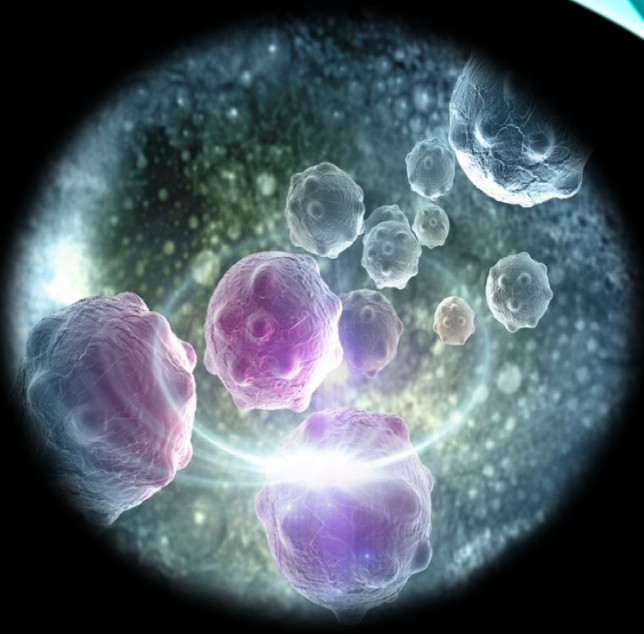
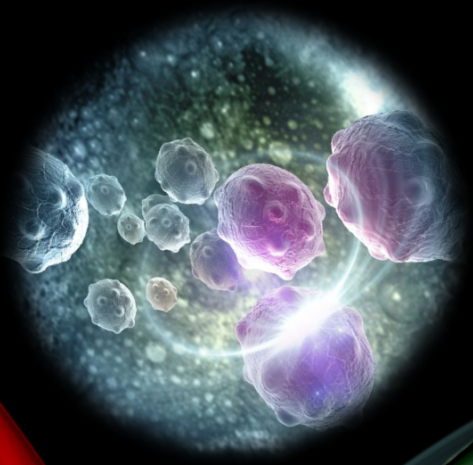
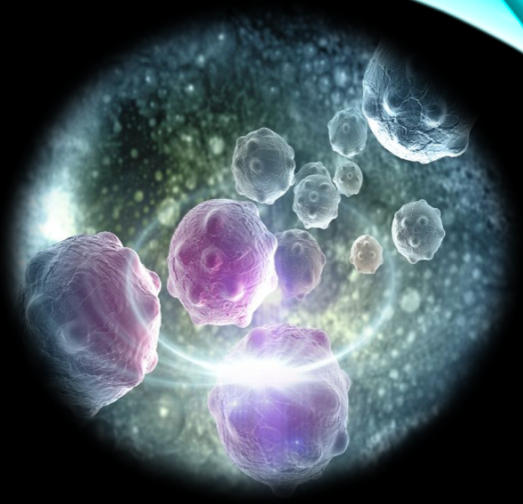


WISCONSIN BREAST CANCER CELL DIAGNOSTIC DATA

a case in Linear Discriminant Analysis | Artificial Intelligence via Machine Learning

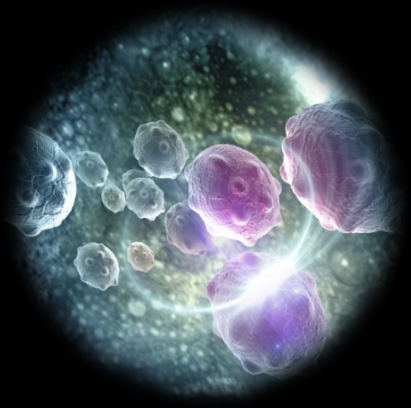




MSDS 6372 - 402 Applied Statistics: Inference and Modeling, Project II

TEAM 2

Ranga Nagarajan
Rob Rivens
Adam Tschannen



BREAST CANCER WISCONSIN (DIAGNOSTIC) DATA SET

When cells in the breast begin to grow out of control, this is typically a signal that breast cancer is beginning to form. The abnormal cells usually form a tumor that can often be detected via x-ray or felt as a lump. The tumor is malignant (or cancerous) if the cells invade surrounding tissues or spread to other areas of the body. In fact, cells in nearly any part of the body can become cancerous.

Research indicates that most breast lumps are benign and not cancer (or malignant). Non-cancerous breast cells (called tumors) are abnormal growths, but they do not spread outside of the breast and they are not life threatening. Some benign breast lumps can increase a woman's risk of getting breast cancer at a later date. As such, any breast lump or change needs to be checked by a health care professional to determine if it is benign or malignant (cancer) and if it might affect a person's future cancer risk[1].

Since malignant cells are typically malformed, research has shown that measuring various parts of the cell can aid in identifying malignant vs benign cells (Figure 1).

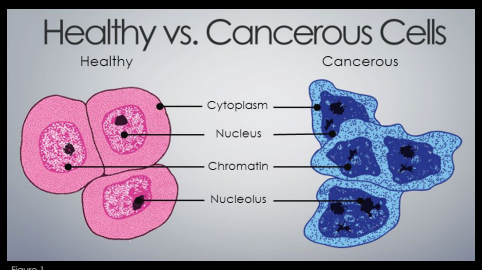


Figure 1

The cancer cell dataset obtained from Kaggle[3] features variables computed from a digitized image of a fine needle aspirate (FNA) of a breast mass, which describe characteristics of the cell nuclei present in the image.

Attribute Information:

1) ID number

2) Diagnosis (M = malignant, B = benign)

3-32) Ten features computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)

b) texture (standard deviation of gray-scale values)

c) perimeter

d) area

e) smoothness (local variation in radius lengths)

f) compactness (perimeter^2 / area - 1.0)

g) concavity (severity of concave portions of the contour)

h) concave points (number of concave portions of the contour)

i) symmetry

j) fractal dimension ("coastline approximation" - 1)

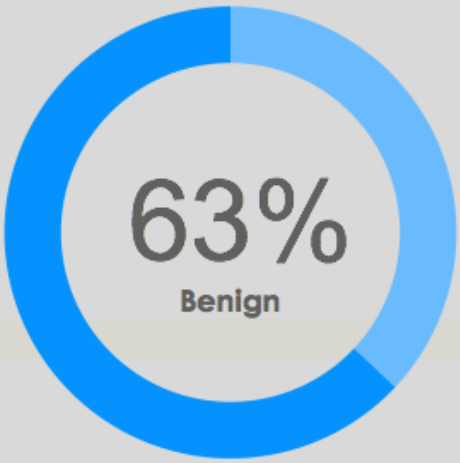
Note:

The mean, standard error and largest (mean of the three largest values – called "Worst") of these cell nucleus characteristics were computed for each image, resulting in 30 total features.

QUICK LOOK: DATASET ATTRIBUTES[2]

Dataset Columns
ID
Diagnosis
radius_mean
texture_mean
perimeter_mean
area_mean
smoothness_mean
compactness_mean
concavity_mean
concave_points_mean
symmetry_mean
fractal_dimension_mean
radius_se
texture_se
perimeter_se
area_se
smoothness_se
compactness_se
concavity_se
concave_points_se
symmetry_se
fractal_dimension_se
radius_worst
texture_worst
perimeter_worst
area_worst
smoothness_worst
compactness_worst
concavity_worst
concave_points_worst
symmetry_worst
fractal_dimension_worst

Diagnosis Distribution



Summary

Malignant Observations

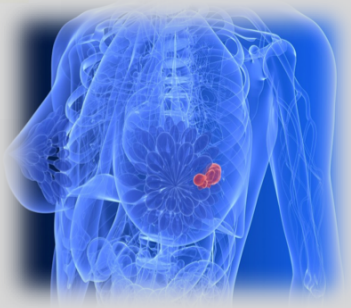
212

Benign Observations

357

Total Observations

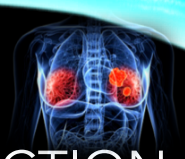
569



Summary of Measurements

Measurement	Malignant Mean	Malignant Worst *	Malignant Standard Error	Benign Mean	Benign Worst *	Benign Standard Error	Measurement Definition
a) radius	17.46	21.13	0.609	12.15	13.38	0.284	mean of distances from center to points on the perimeter
b) texture	21.60	29.32	1.211	17.91	23.52	1.220	standard deviation of gray-scale values
c) perimeter	115.37	141.37	4.324	78.08	87.01	2.000	
d) area	978.38	1,422.29	72.672	462.79	558.90	21.135	
e) smoothness	0.10	0.14	0.007	0.09	0.12	0.007	local variation in radius lengths
f) compactness	0.15	0.37	0.032	0.08	0.18	0.021	perimeter^2 / area - 1.0
g) concavity	0.16	0.45	0.056	0.05	0.17	0.057	severity of concave portions of the contour
h) concave points	0.09	0.18	0.015	0.03	0.07	0.010	number of concave portions of the contour
i) symmetry	0.19	0.32	0.020	0.17	0.27	0.021	
j) fractal dimension	0.06	0.09	0.004	0.06	0.08	0.004	coastline approximation - 1

Figures 2 - 5



STATE OF THE DATA | MODEL SELECTION

There were no missing attribute values. We had to clean-up some of the variable names, as some of them had spaces, which caused initial SAS import to account for more variables than were actually in the data. Adding an underscore to connect variables titles resolved this issue. We also had to add a binary indicator for "Malignant" vs "Benign", as the Principal Component Analysis requires numerical values in order to function properly.

Method Selection[a]

We decided to use Linear Discriminant Analysis (LDA) to model the data. LDA is used to determine which continuous variables discriminate between two naturally occurring groups. In this case, the cell characteristics are the variables and "Benign" vs "Malignant" are the groups. The fact that we only had two categorical states -- Benign and Malignant – made our dataset perfect for using LDA. First, we needed to ensure our data met the requirements for LDA. Categorically:

- Sample size: Unequal sample sizes acceptable, therefore we will proceed, despite our 63%/37% benign vs malignant samples.
- Normal distribution: due to the number of observations, under the Central Limit Theorem, it is assumed that the data represent a sample from a multivariate normal distribution.
- Homogeneity of variances: LDA is very sensitive to heterogeneity of variance-covariance matrices. Homoscedasticity is evaluated through scatterplots and corrected by transformation of variables. We will evaluate the *Simple* (*summary) Statistics* and run a log transformation, if necessary.
- Non-multicollinearity: If one of the independent variables is very highly correlated with another, or one is a function of other independents, then the tolerance value for that variable will approach 0 and the matrix will not have a unique discriminant solution. Since Principles Component Analysis alleviates collinearity, we will proceed with a PCA, then LDA.

The "training data set" features 30 variables, with the following Class distribution: 357 benign and 212 malignant cells. We ran a FULL PRINCIPLE COMPONENTS ANALYSIS, which we intended to use to reduce the number of variables from 30 to a more manageable number, and also to explore the relationship between the variables.

PRINCIPLE COMPONENT ANALYSIS

Upon examination of the Simple Statistics matrix, we saw that there were a few Means which were drastically different in scale from the others. Since LDA is sensitive to scale, we decided we needed to further explore the possibility of transforming the data (Figure 6).

Reviewing the Eigenvalues of the Covariance Matrix, we discovered that the large-scaled values were causing the cumulative Principle Components composition to reflect 98% of the proportion of differences to be reflected in the first Component (Figure 7). Examination of the Scree Plot and Variance Explained graph confirmed this effect (Figure 8). The Component Scores Matrix (Figure 9) was difficult to analyze due to the number of components, but we felt confident that given the aforementioned effect of the scale issue, we needed to perform a log transformation on the data.

Cancer Data Set (Raw) Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
radius_mean	569	14.12729	3.52405	8038	6.98100	28.11000
texture_mean	569	19.28965	4.30104	10976	9.71000	39.28000
perimeter_mean	569	91.96903	24.29898	52330	43.79000	188.50000
area_mean	569	654.88910	351.91413	372632	143.50000	2501
smoothness_mean	569	0.09636	0.01406	54.82900	0.05263	0.16340
compactness_mean	569	0.10434	0.05281	59.37002	0.01938	0.34540
concavity_mean	569	0.08880	0.07972	50.52681	0	0.42680
concave_points_mean	569	0.04892	0.03880	27.83499	0	0.20120
symmetry_mean	569	0.18116	0.02741	103.08110	0.10600	0.30400
fractal_dimension_mean	569	0.06280	0.00706	35.73184	0.04996	0.09744
radius_se	569	0.40517	0.27731	230.54290	0.11150	2.87300
texture_se	569	1.21685	0.55165	692.38960	0.36020	4.88500
perimeter_se	569	2.86606	2.02185	1631	0.75700	21.98000
area_se	569	40.33708	45.49101	22952	6.80200	542.20000
smoothness_se	569	0.00704	0.00300	4.00632	0.00171	0.03113
compactness_se	569	0.02548	0.01791	14.49706	0.00225	0.13540
concavity_se	569	0.05651	0.00177	32.15191	0.01860	0.05661
concave_points_se	569	0.01180	0.00617	6.71200	0	0.05279
symmetry_se	569	0.02054	0.00827	11.68857	0.00788	0.07895
fractal_dimension_se	569	0.00379	0.00265	2.15930	0.0008948	0.02984
radius_worst	569	16.26919	4.83324	9257	7.93000	36.04000
texture_worst	569	25.67722	6.14626	14610	12.02000	49.54000
perimeter_worst	569	107.26121	33.60254	61032	50.41000	251.20000
area_worst	569	880.58313	569.35699	501052	185.20000	4254
smoothness_worst	569	0.13237	0.02283	75.31773	0.07117	0.22260
compactness_worst	569	0.25427	0.15734	144.67681	0.02729	1.05800
concavity_worst	569	0.27219	0.20862	154.87525	0	1.25200
concave_points_worst	569	0.11461	0.06573	65.21094	0	0.29100
symmetry_worst	569	0.29008	0.06187	165.05300	0.15650	0.66380
fractal_dimension_worst	569	0.08395	0.01806	47.76517	0.05504	0.20750

Figure 6

Eigenvalues of the Covariance Matrix
Cancer Data Set (Raw)

	Eigenvalue	Difference	Proportion	Cumulative
1	443782.605	436472.505	0.9820	0.9820
2	7310.100	6606.266	0.0162	0.9982
3	703.834	649.185	0.0016	0.9998
4	54.649	14.759	0.0001	0.9999
5	39.890	36.885	0.0001	1.0000
6	3.005	1.189	0.0000	1.0000
7	1.815	1.444	0.0000	1.0000
8	0.371	0.216	0.0000	1.0000
9	0.155	0.071	0.0000	1.0000
10	0.084	0.053	0.0000	1.0000

Figure 7

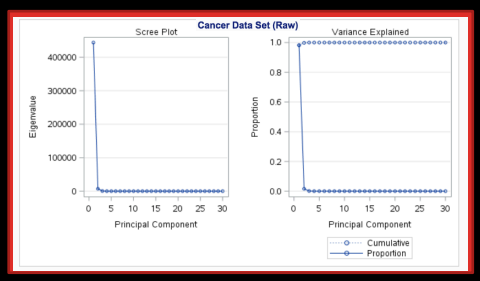


Figure 8

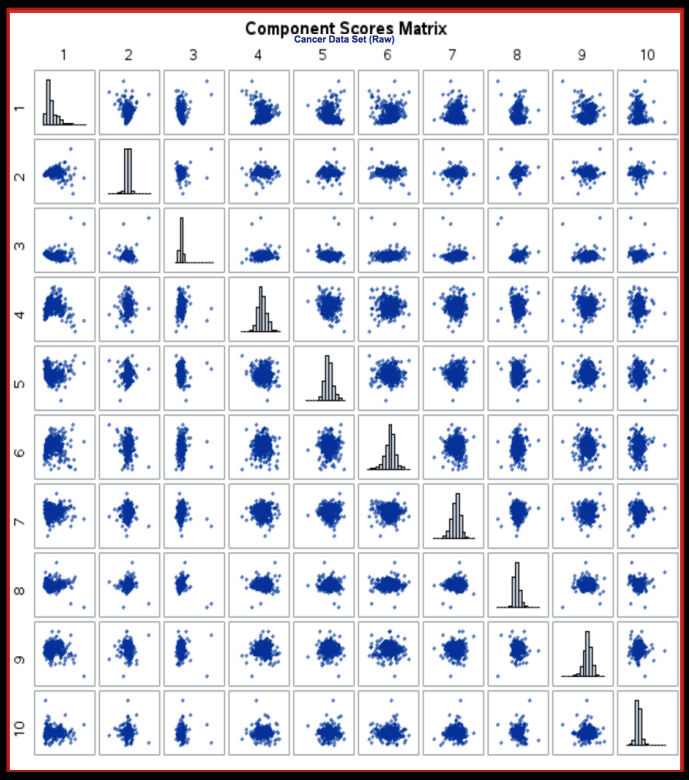


Figure 9

PRINCIPLE COMPONENT ANALYSIS

We then realized that we needed to look at the Eigenvalues of the Correlation Matrix, as well. Due to different scaling of variables, we knew that mathematically, the Covariance Matrix would be heavily influenced by variables with large variances.

Reviewing the Eigenvalues of the Correlation Matrix, we discovered that the large-scaled values didn't have much of an impact, as we'd need to consider nine (9) Principle Components to reflect 93% of the cumulative differences. Examination of the Scree Plot and Variance Explained graph confirmed this effect (Figures 10 and 11).

Eigenvalues of the Covariance Matrix
Cancer Data Set (Raw)

	Eigenvalue	Difference	Proportion	Cumulative
1	443782.605	436472.505	0.9820	0.9820
2	7310.100	6606.266	0.0162	0.9982
3	703.834	649.185	0.0016	0.9998
4	54.649	14.759	0.0001	0.9999
5	39.890	36.885	0.0001	1.0000
6	3.005	1.189	0.0000	1.0000
7	1.815	1.444	0.0000	1.0000
8	0.371	0.216	0.0000	1.0000
9	0.155	0.071	0.0000	1.0000
10	0.084	0.053	0.0000	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	12.9933384	7.5022140	0.4331	0.4331
2	5.4911244	2.7482457	0.1830	0.6161
3	2.7428787	0.7574262	0.0914	0.7076
4	1.9854525	0.5360424	0.0662	0.7738
5	1.4494101	0.2468090	0.0483	0.8221
6	1.2026012	0.2096780	0.0401	0.8622
7	0.9929232	0.3442027	0.0331	0.8953
8	0.6487205	0.1733436	0.0216	0.9169
9	0.4753769	0.0943472	0.0158	0.9327
10	0.3810297	0.0416627	0.0127	0.9454

Figure 10

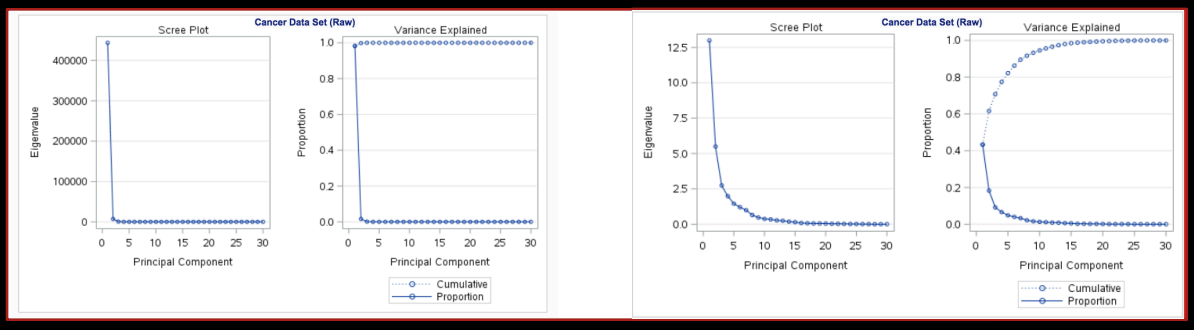


Figure 11

We decided to continue with our Log Transformation, to see of this would help reduce the number of Principle Components…

PRINCIPLE COMPONENT ANALYSIS_v2.0

Upon examination of the *log transformed* Simple Statistics matrix, we saw that the Means were now roughly on the same scale for all variables (Figure 12). This will now satisfy the scale guideline for arriving at a more reasonable output from our PCA.

Reviewing the Eigenvalues of the log transformed Correlation Matrix, we discovered that the cumulative Principle Components composition to reflect 93% of the proportion of differences requires at least eight (8) Components – a slight improvement vs the previous results (Figure 13). Examination of the Scree Plot and Variance Explained graph confirmed this effect (Figure 14). Although the Component Scores Matrix (Figure 15) is still a bit of an eye-chart due to the number of components, we can confirm the normality of histograms and lack of linearity in the scatter plots due to the Central Limit Theorem and the effect of running the PCA (uncorrelated data).

Variable Matrix of Cancer Data Set (Log Transformed)

The MEANS Procedure

Variable	N	Mean	Median	Std Dev	Variance	Minimum	Maximum
radius_mean	569	2.62	2.59	0.24	0.06	1.94	3.34
texture_mean	569	2.94	2.94	0.22	0.05	2.27	3.67
perimeter_mean	569	4.49	4.46	0.25	0.06	3.78	5.24
area_mean	569	6.36	6.31	0.48	0.23	4.97	7.82
smoothness_mean	569	-2.35	-2.34	0.15	0.02	-2.94	-1.81
compactness_mean	569	-2.38	-2.38	0.49	0.24	-3.94	-1.06
concavity_mean	556	-2.84	-2.73	1.08	1.16	-7.28	-0.85
concave_points_mean	556	-3.31	-3.36	0.86	0.73	-6.29	-1.60
symmetry_mean	569	-1.72	-1.72	0.15	0.02	-2.24	-1.19
fractal_dimension_mean	569	-2.77	-2.79	0.11	0.01	-3.00	-2.33
radius_se	569	-1.07	-1.13	0.54	0.29	-2.19	1.06
texture_se	569	0.10	0.10	0.43	0.18	-1.02	1.59
perimeter_se	569	0.89	0.83	0.54	0.29	-0.28	3.09
area_se	569	3.38	3.20	0.73	0.53	1.92	6.30
smoothness_se	569	-5.03	-5.05	0.37	0.14	-6.37	-3.47
compactness_se	569	-3.88	-3.89	0.65	0.42	-6.10	-2.00
concavity_se	569	-2.87	-2.87	0.05	0.00	-3.98	-2.87
concave_points_se	556	-4.53	-4.50	0.48	0.23	-6.29	-2.94
symmetry_se	569	-3.95	-3.98	0.34	0.12	-4.84	-2.54
fractal_dimension_se	569	-5.73	-5.75	0.53	0.28	-7.02	-3.51
radius_worst	569	2.75	2.71	0.28	0.08	2.07	3.58
texture_worst	569	3.22	3.24	0.24	0.06	2.49	3.90
perimeter_worst	569	4.63	4.58	0.29	0.08	3.92	5.53
area_worst	569	6.62	6.53	0.55	0.31	5.22	8.36
smoothness_worst	569	-2.04	-2.03	0.17	0.03	-2.64	-1.50
compactness_worst	569	-1.55	-1.55	0.62	0.38	-3.60	0.06
concavity_worst	556	-1.63	-1.46	0.98	0.97	-6.30	0.22
concave_points_worst	556	-2.32	-2.29	0.64	0.41	-4.74	-1.23
symmetry_worst	569	-1.26	-1.27	0.20	0.04	-1.85	-0.41
fractal_dimension_worst	569	-2.50	-2.53	0.20	0.04	-2.90	-1.57

Figure 12

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	4.81748264	3.84052124	0.6522	0.6522
2	0.97696140	0.29279296	0.1323	0.7844
3	0.68416844	0.48883054	0.0926	0.8770
4	0.19533790	0.02819454	0.0264	0.9035
5	0.16714336	0.04714839	0.0226	0.9261
6	0.11999497	0.02177921	0.0162	0.9424
7	0.09821576	0.01476196	0.0133	0.9556
8	0.08345380	0.02718361	0.0113	0.9669
9	0.05627019	0.00156050	0.0076	0.9746
10	0.05470970	0.01275821	0.0074	0.9820

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	13.1624350	7.4396623	0.4387	0.4387
2	5.7227727	2.9851604	0.1908	0.6295
3	2.7376123	0.8315369	0.0913	0.7208
4	1.9060754	0.4185222	0.0635	0.7843
5	1.4875532	0.3047502	0.0496	0.8339
6	1.1828030	0.1876345	0.0394	0.8733
7	0.9951685	0.2625884	0.0332	0.9065
8	0.7325801	0.2927487	0.0244	0.9309
9	0.4398314	0.0732847	0.0147	0.9456
10	0.3665466	0.0623673	0.0122	0.9578

Figure 13

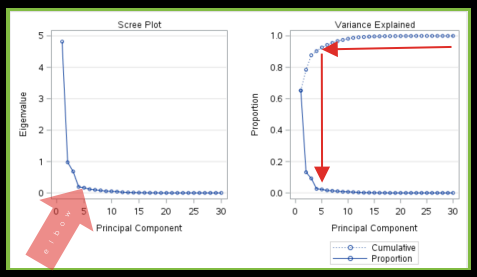


Figure 14

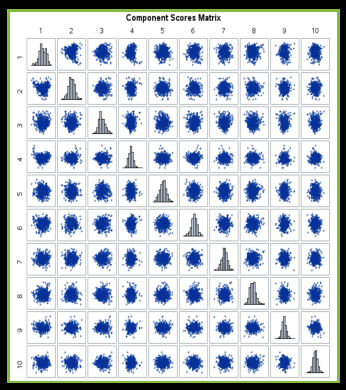


Figure 15

PRINCIPLE COMPONENT ANALYSIS_v2.0

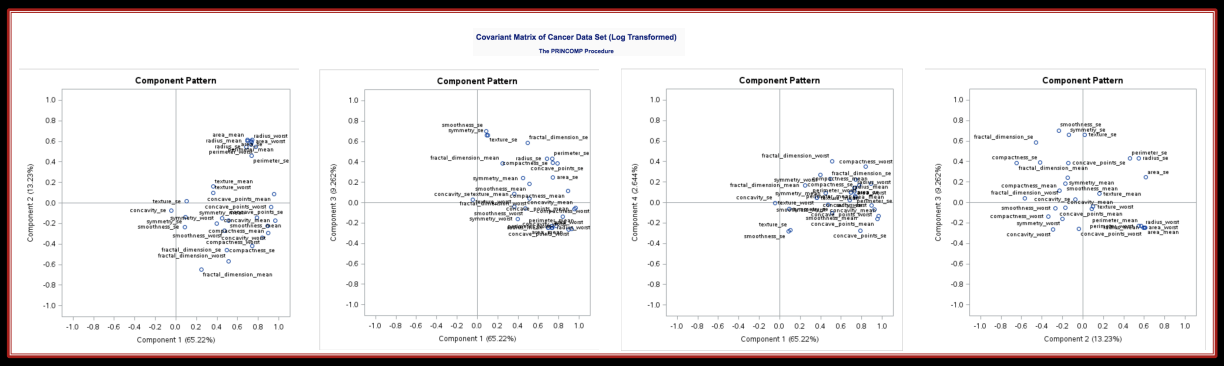


Figure 16

The Eigenvectors used to determine which components we used to create our new Explanatory Variables are depict the loading graphically (Figures 16 - 18).

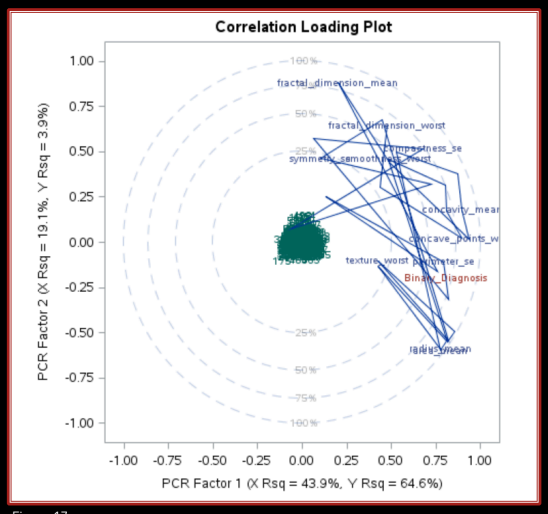


Figure 17

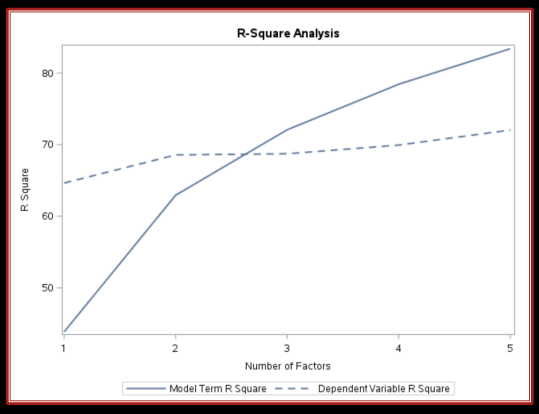


Figure 18

Figure 19 shows the delineation between Benign (blue) and Malignant (red) cancer cells in the dataset. There is a clear distinction between the groups, with only a minimal overlap.

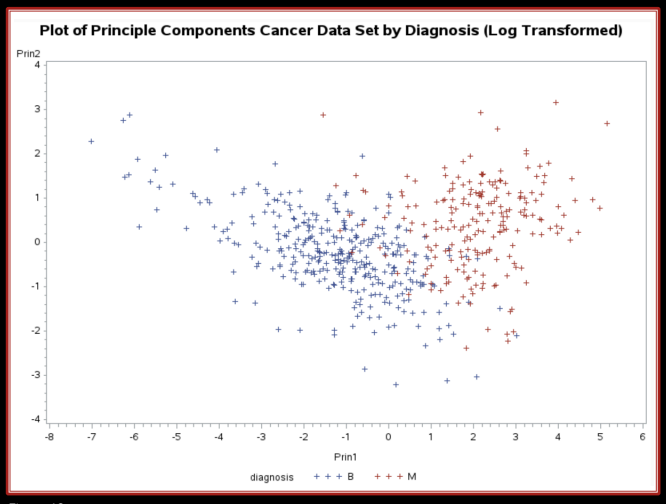
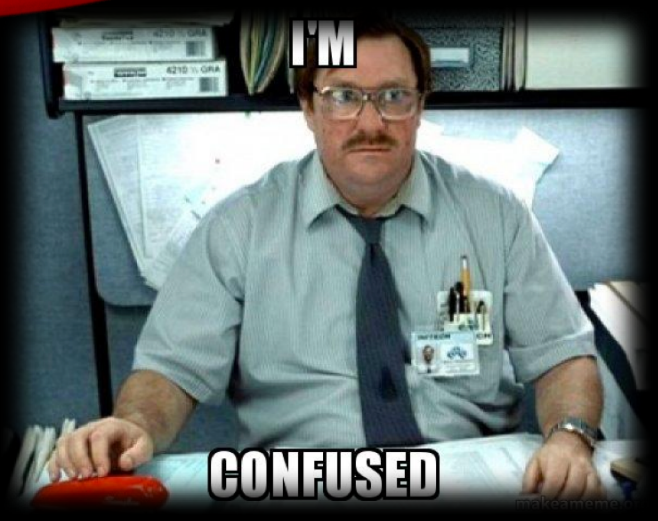


Figure 19

BUT WAIT…



Plot of Principle Components Cancer Data Set by Diagnosis (Log Transformed)

The PLS Procedure

Data Set	WORK.CANCERDATAPRINCOMP
Factor Extraction Method	Principal Components Regression
Number of Response Variables	1
Number of Predictor Parameters	30
Missing Value Handling	Exclude
Number of Factors	5

Number of Observations Read	569
Number of Observations Used	556

Figure 20

We noticed that not all the data was being used (Figure 20). Upon further discussion, we identified that using a log function to transform the data meant any negative or zero values would result in missing data. As a result, we used the following selected adjustments to standardize the data (Figure 21):

radius_mean = (radius_mean/10)
texture_mean = (texture_mean/10)
perimeter_mean = (perimeter_mean/100)
area_mean = (area_mean/1000)
texture_se = (texture_se/10)
perimeter_se = (perimeter_se/10)
area_se = (area_se/100)
radius_worst = (radius_worst/100)
texture_worst = (texture_worst/100)
perimeter_worst = (perimeter_worst/100)
area_worst = (area_worst/1000)

Figure 21

PRINCIPLE COMPONENT ANALYSIS_v3.0

Upon examination of the *standardized* Simple Statistics matrix, we saw that the Means were now roughly on the same scale for all variables (Figure 22). This will now satisfy the scale guideline for arriving at a more reasonable output from our PCA.

Reviewing the Eigenvalues of the log transformed Covariance Matrix, we discovered that the cumulative Principle Components composition to reflect 93% of the proportion of differences requires at least nine (9) Components (Figure 23). Examination of the Scree Plot and Variance Explained graph confirmed this effect (Figure 24). Although the Component Scores Matrix (Figure 25) is still a bit of an eye-chart due to the number of components, we can confirm the normality of histograms and lack of linearity in the scatter plots due to the Central Limit Theorem and the effect of running the PCA (uncorrelated data).

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
radius_mean	569	1.41273	0.35240	803.84290	0.69810	2.81100
texture_mean	569	1.92896	0.43010	1098	0.97100	3.92800
perimeter_mean	569	0.91969	0.24299	523.30380	0.43790	1.88500
area_mean	569	0.65489	0.35191	372.63190	0.14350	2.50100
smoothness_mean	569	0.09636	0.01406	54.82900	0.05263	0.16340
compactness_mean	569	0.10434	0.05281	59.37002	0.01938	0.34540
concavity_mean	569	0.08880	0.07972	50.52681	0	0.42680
concave_points_mean	569	0.04892	0.03880	27.83499	0	0.20120
symmetry_mean	569	0.18116	0.02741	103.08110	0.10600	0.30400
fractal_dimension_mean	569	0.06280	0.00706	35.73184	0.04996	0.09744
radius_se	569	0.40517	0.27731	230.54290	0.11150	2.87300
texture_se	569	0.12169	0.05516	69.23896	0.03602	0.48850
perimeter_se	569	0.28661	0.20219	163.07877	0.07570	2.19800
area_se	569	0.40337	0.45491	229.51798	0.06802	5.42200
smoothness_se	569	0.00704	0.00300	4.00632	0.00171	0.03113
compactness_se	569	0.02548	0.01791	14.49706	0.00225	0.13540
concavity_se	569	0.05651	0.00177	32.15191	0.01860	0.05661
concave_points_se	569	0.01180	0.00617	6.71200	0	0.05279
symmetry_se	569	0.02054	0.00827	11.68857	0.00788	0.07895
fractal_dimension_se	569	0.00379	0.00265	2.15930	0.0008948	0.02984
radius_worst	569	0.16269	0.04833	92.57169	0.07930	0.36040
texture_worst	569	0.25677	0.06146	146.10340	0.12020	0.49540
perimeter_worst	569	1.07261	0.33603	610.31630	0.50410	2.51200
area_worst	569	0.88058	0.56936	501.05180	0.18520	4.25400
smoothness_worst	569	0.13237	0.02283	75.31773	0.07117	0.22260
compactness_worst	569	0.25427	0.15734	144.67681	0.02729	1.05800
concavity_worst	569	0.27219	0.20862	154.87525	0	1.25200
concave_points_worst	569	0.11461	0.06573	65.21094	0	0.29100
symmetry_worst	569	0.29008	0.06187	165.05300	0.15650	0.66380
fractal_dimension_worst	569	0.08395	0.01806	47.76517	0.05504	0.20750

Figure 22

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	12.9933384	7.5022140	0.4331	0.4331
2	5.4911244	2.7482457	0.1830	0.6161
3	2.7428787	0.7574262	0.0914	0.7076
4	1.9854525	0.5360424	0.0662	0.7738
5	1.4494101	0.2468090	0.0483	0.8221
6	1.2026012	0.2096780	0.0401	0.8622
7	0.9929232	0.3442027	0.0331	0.8953
8	0.6487205	0.1733436	0.0216	0.9169
9	0.4753769	0.0943472	0.0158	0.9327
10	0.3810297	0.0416627	0.0127	0.9454
11	0.3393670	0.0657687	0.0113	0.9567
12	0.2735983	0.0303651	0.0091	0.9659
13	0.2432333	0.0590213	0.0081	0.9740
14	0.1842120	0.0406551	0.0061	0.9801
15	0.1435569	0.0596345	0.0048	0.9849

Figure 23

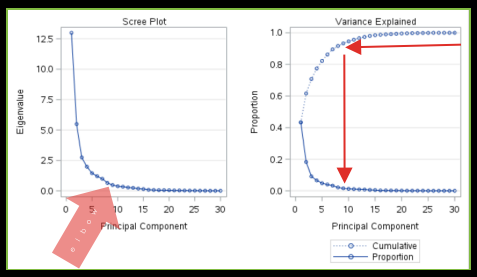


Figure 24

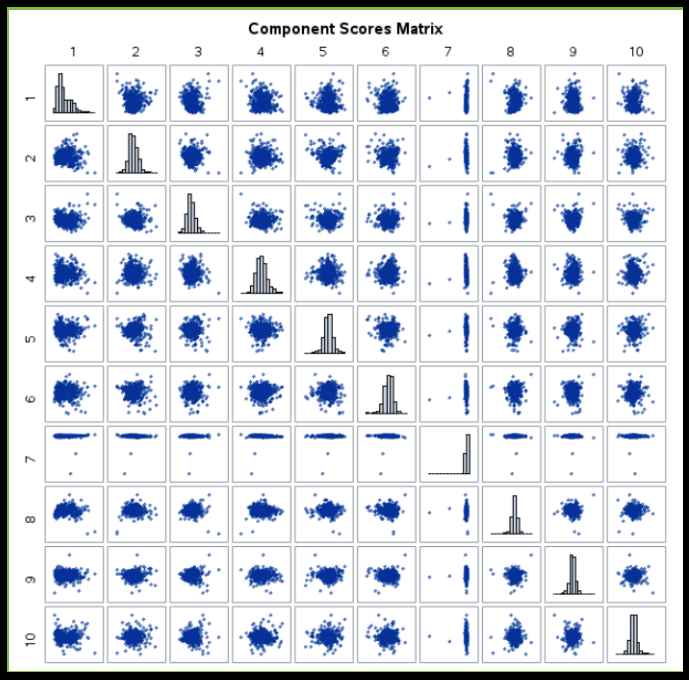


Figure 25

PRINCIPLE COMPONENT ANALYSIS_v3.0

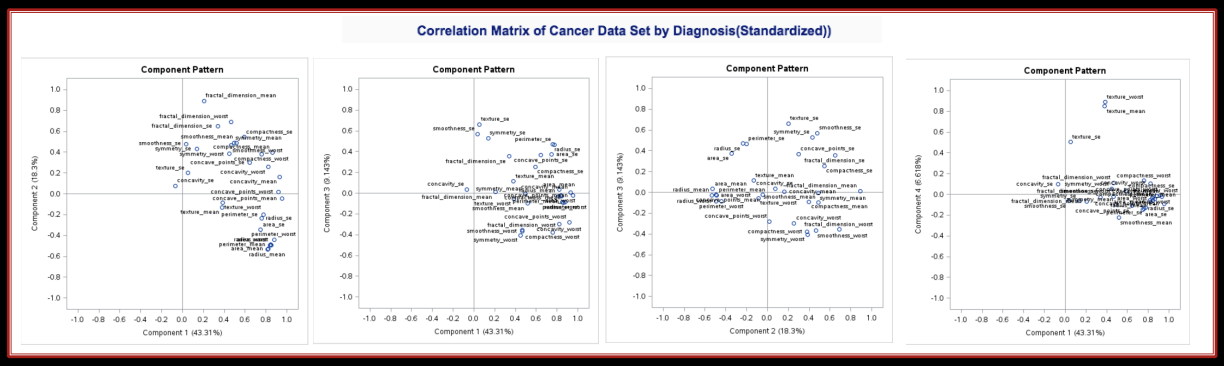


Figure 26

The Eigenvectors used to determine which components we used to create our new Explanatory Variables are depict the loading graphically (Figures 26 - 28).

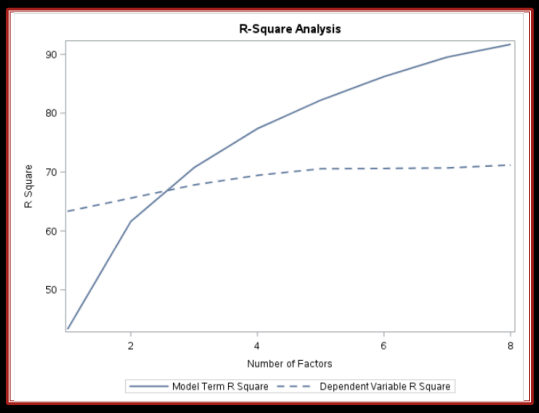


Figure 28

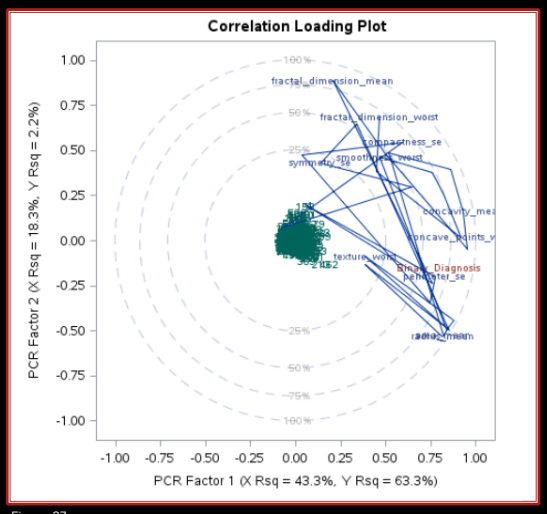


Figure 27

More specifically, Figure 27 depicts the Loading of variables included in the PCA, in a radial format. Figure 28 confirms the % of Variance – about 90% – explained by the first eight (8) factors in the model. Figure 29 shows the delineation between Benign (blue) and Malignant (red) cancer cells in the dataset for Principle Components 1 & 2. There is a clear distinction between the groups, with only a minimal overlap.

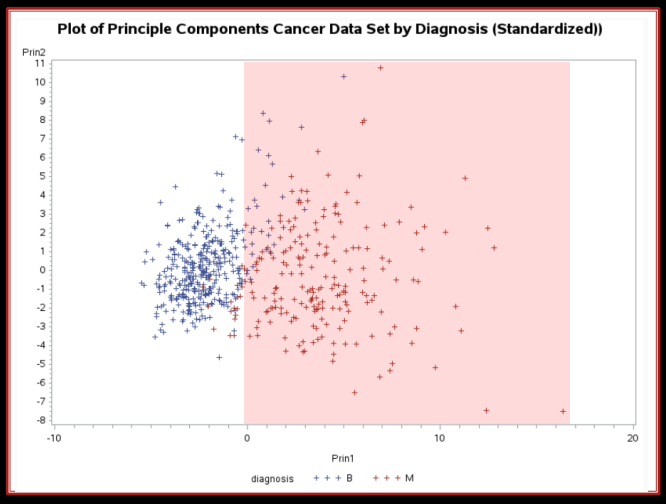
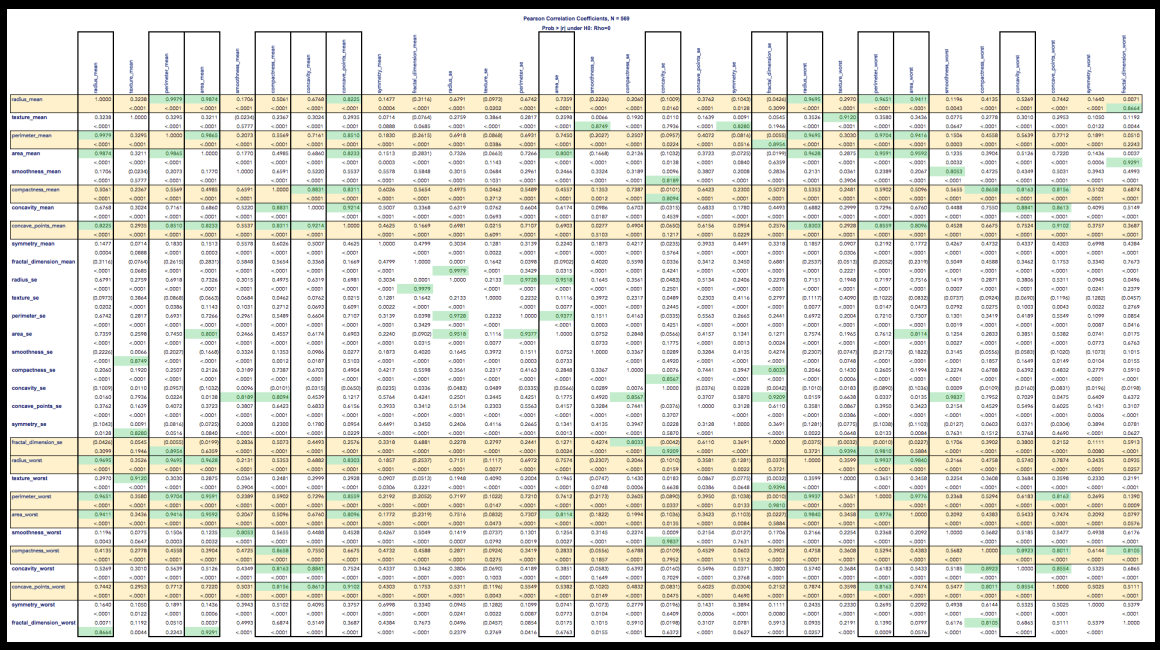


Figure 29

PRINCIPLE COMPONENT ANALYSIS_v3.0



Pearson Correlation Coefficients, N = 569
Prob > |r| under H0: Rho=0

	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave_points_mean	symmetry_mean	fractal_dimension_mean	radius_se	texture_se	perimeter_se	area_se	smoothness_se	compactness_se	concavity_se	concave_points_se	symmetry_se	fractal_dimension_se	radius_worst	texture_worst	perimeter_worst	area_worst	smoothness_worst	compactness_worst	concavity_worst	concave_points_worst	symmetry_worst	fractal_dimension_worst
radius_mean	1.0000	0.3238	0.9979	0.9874	0.1704	0.5361	0.6746	0.8223	0.1477	(0.3116)	0.6791	(0.0973)	0.6742	0.7339	(0.2224)	0.2040	(0.1009)	0.3762	(0.1043)	(0.0424)	0.9465	0.2970	0.9455	0.9411	0.1196	0.4130	0.3269	0.7442	0.1640	0.0071
		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0004	<.0001	<.0001	0.0203	<.0001	<.0001	<.0001	<.0001	0.0140	<.0001	0.0128	0.3099	<.0001	<.0001	<.0001	<.0001	0.0043	<.0001	<.0001	<.0001	0.8644
texture_mean	0.3238	1.0000	0.3295	0.3311	0.0224	0.2342	0.3024	0.2913	0.0714	(0.0746)	0.2728	0.3884	0.2811	0.2598	0.0546	0.192	0.0110	0.1429	0.0975	0.0543	0.3316	0.9120	0.3498	0.3436	0.0772	0.2776	0.3010	0.2953	0.1050	0.1192
		<.0001	<.0001	<.0001	0.5777	<.0001	<.0001	<.0001	0.0888	0.0685	<.0001	<.0001	<.0001	<.0001	0.8749	<.0001	0.7936	<.0001	0.8245	0.1964	<.0001	<.0001	<.0001	<.0001	0.0647	<.0001	<.0001	<.0001	0.0132	0.0044
perimeter_mean	0.9979	0.3295	1.0000	0.9865	0.2073	0.5549	0.7141	0.8510	0.1830	(0.2615)	0.6918	(0.0848)	0.6901	0.7430	(0.2037)	0.3507	(0.0947)	0.4073	(0.0816)	(0.0055)	0.9435	0.3030	0.9704	0.9416	0.1506	0.4558	0.3639	0.7712	0.1891	0.0510
	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0396	<.0001	<.0001	<.0001	<.0001	0.0239	<.0001	0.0516	0.8954	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.2243
area_mean	0.9874	0.3311	0.9865	1.0000	0.1770	0.4985	0.6860	0.8223	0.1513	(0.2831)	0.7326	(0.0663)	0.7264	0.8001	(0.1648)	0.2126	(0.1032)	0.3723	(0.0720)	(0.0199)	0.9428	0.2873	0.9591	0.9592	0.1233	0.3904	0.3134	0.7220	0.1436	0.0037
	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	0.0003	<.0001	<.0001	0.1143	<.0001	<.0001	<.0001	<.0001	0.0138	<.0001	0.0840	0.6359	<.0001	<.0001	<.0001	<.0001	0.0032	<.0001	<.0001	<.0001	0.0006	0.9291
smoothness_mean	0.1704	(0.0224)	0.2073	0.1770	1.0000	0.6591	0.5208	0.5537	0.5478	0.5848	0.3014	0.3484	0.3811	0.3568	0.4268	0.2464	0.3324	0.3189	0.3807	0.2832	0.2851	0.0584	0.2661	0.2744	0.8033	0.4724	0.3692	0.5301	0.3943	0.4769
	<.0001	0.5777	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.1031	<.0001	<.0001	<.0001	<.0001	0.8189	<.0001	<.0001	<.0001	<.0001	0.3904	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
compactness_mean	0.5361	0.2342	0.5549	0.4985	0.6591	1.0000	0.8831	0.8311	0.4224	0.3454	0.4973	0.2642	0.5489	0.4537	0.1353	0.7387	(0.0107)	0.6423	0.2302	0.3073	0.5353	0.2481	0.5092	0.5096	0.3455	0.8638	0.8143	0.8156	0.5102	0.4874
	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	0.0012	<.0001	<.0001	0.0012	<.0001	0.7806	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
concavity_mean	0.6746	0.3024	0.7141	0.6860	0.5208	0.8831	1.0000	0.9214	0.3007	0.3398	0.4319	0.0762	0.6404	0.4174	0.0966	0.6703	(0.0315)	0.6833	0.1780	0.4493	0.6882	0.2999	0.7296	0.6760	0.4488	0.7150	0.8841	0.8613	0.4095	0.5149
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	0.0693	<.0001	<.0001	0.0167	<.0001	0.4539	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
concave_points_mean	0.8223	0.2913	0.8510	0.8223	0.5537	0.8311	0.9214	1.0000	0.3029	0.3548	0.4091	0.2115	0.6090	0.4693	0.2377	0.4958	(0.0450)	0.6156	0.2954	0.3576	0.8303	0.2935	0.8591	0.8076	0.4328	0.6675	0.7524	0.9110	0.3757	0.3487
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	0.6091	<.0001	<.0001	0.3103	<.0001	0.1217	<.0001	0.0329	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
symmetry_mean	0.1477	0.0714	0.1830	0.1513	0.5478	0.4224	0.3007	0.3029	1.0000	0.4798	0.3026	0.1331	0.3139	0.2240	0.1873	0.4217	(0.0315)	0.3913	0.4441	0.3318	0.1857	0.0901	0.2192	0.1772	0.4267	0.4372	0.4337	0.4303	0.6998	0.4384
	<.0001	0.0746	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	0.0022	<.0001	<.0001	<.0001	<.0001	0.5764	<.0001	<.0001	<.0001	<.0001	0.0036	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
fractal_dimension_mean	(0.3116)	(0.0746)	(0.2615)	(0.2831)	0.5848	0.3454	0.3398	0.3548	0.4798	1.0000	0.0001	0.1642	0.0199	(0.0002)	0.4020	0.5599	0.0336	0.3412	0.3455	0.6881	(0.2537)	(0.0513)	(0.2052)	(0.2319)	0.5549	0.4568	0.3442	0.1753	0.3340	0.7673
	<.0001	0.0685	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		0.9979	<.0001	0.6429	0.3313	<.0001	<.0001	0.4244	<.0001	<.0001	<.0001	0.2231	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
radius_se	0.6791	0.2728	0.6918	0.7326	0.3015	0.4973	0.4319	0.4091	0.3034	0.0001	1.0000	0.2313	0.9729	0.9918	0.1645	0.3561	(0.0483)	0.3134	0.2405	0.2278	0.7151	0.1948	0.7197	0.7516	0.1419	0.2871	0.3806	0.5311	0.0945	0.0496
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.9979	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	0.2351	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0007	<.0001	<.0001	<.0001	0.0241	0.2379
texture_se	(0.0973)	0.3884	(0.0848)	(0.0663)	0.0484	0.0642	0.0762	0.2115	0.1331	0.1642	0.2313	1.0000	0.2230	0.1116	0.3972	0.2317	0.0489	0.2303	0.4116	0.2797	(0.1117)	0.4090	(0.1022)	(0.0853)	(0.0737)	(0.0592)	(0.1342)	(0.1182)	(0.1282)	(0.0407)
	0.0203	<.0001	0.0396	0.1143	0.1031	0.2712	0.0693	0.6091	0.0022	<.0001	<.0001		0.0077	<.0001	<.0001	<.0001	0.2445	<.0001	<.0001	<.0001	0.0057	<.0001	0.0147	0.0473	0.0792	0.0270	0.1003	0.0043	0.0032	0.2769
perimeter_se	0.6742	0.2811	0.6901	0.7264	0.3811	0.5489	0.6404	0.7107	0.3139	0.3398	0.9729	0.2232	1.0000	0.9377	0.1511	0.4163	(0.0335)	0.3563	0.2845	0.2441	0.6972	0.2004	0.7210	0.7307	0.1301	0.3419	0.4189	0.5549	0.1099	0.0854
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	0.0003	<.0001	0.4231	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0019	<.0001	<.0001	<.0001	0.0087	0.0416
area_se	0.7339	0.2598	0.7430	0.8001	0.3568	0.4537	0.4174	0.4693	0.2240	0.0199	0.9918	0.1116	0.9377	1.0000	0.0752	0.2849	(0.0556)	0.4517	0.1341	0.1271	0.7574	0.1965	0.7613	0.8114	0.1254	0.2820	0.3851	0.5382	0.0741	0.0175
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.9377	<.0001	0.0003	<.0001		0.0733	<.0001	0.2611	<.0001	0.0024	<.0001	<.0001	<.0001	<.0001	<.0001	0.0037	<.0001	<.0001	<.0001	0.0773	0.6763
smoothness_se	(0.2224)	0.0546	(0.2037)	(0.1648)	0.4268	0.1353	0.0966	0.3227	0.1873	0.4020	0.1645	0.3972	0.1511	0.0752	1.0000	0.3367	0.0489	0.3384	0.4135	0.4274	0.2207	0.0745	0.1842	0.1745	0.8019	0.5768	0.5548	0.4837	0.5350	0.1075
	<.0001	0.2192	<.0001	<.0001	<.0001	0.0012	0.0167	0.1193	<.0001	<.0001	<.0001	<.0001	0.0003	0.0733		<.0001	0.4925	<.0001	<.0001	<.0001	<.0001	0.0749	<.0001	<.0001	0.1857	0.1449	0.2149	0.1004	0.0115
compactness_se	0.2040	0.1920	0.2207	0.2126	0.3189	0.7387	0.4539	0.4958	0.4217	0.5599	0.3561	0.2317	0.4163	0.2849	0.3367	1.0000	0.0376	0.7441	0.3942	0.8021	0.2054	0.1450	0.2455	0.1994	0.3274	0.5788	0.6832	0.4292	0.3179	0.5910
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		0.8547	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
concavity_se	(0.1009)	0.0110	(0.0947)	(0.1032)	0.0094	(0.0107)	(0.0315)	(0.0450)	(0.0315)	0.0336	(0.0483)	0.0489	(0.0335)	(0.0556)	0.0289	0.0376	1.0000	(0.0376)	0.0228	(0.0042)	(0.1010)	0.0183	(0.0890)	(0.1036)	0.0009	(0.0105)	(0.0140)	(0.0831)	(0.0196)	(0.0198)
	0.0140	0.7936	0.0224	0.0138	0.8189	0.8039	0.4539	0.1217	0.2544	0.4244	0.2351	0.2445	0.4231	0.2611	0.4925	0.8547		0.3707	0.5875	0.9209	0.9109	0.6436	0.5337	0.0138	0.9837	0.7952	0.7929	0.0373	0.7853	0.7853
concave_points_se	0.3762	0.1429	0.4073	0.3723	0.3807	0.6423	0.6833	0.6156	0.3913	0.3412	0.3134	0.3303	0.2845	0.4517	0.3384	0.7441	(0.0376)	1.0000	0.3128	0.4110	0.3581	0.0861	0.3952	0.3423	0.2154	0.4529	0.5496	0.6025	0.1431	0.3107
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.3707		<.0001	<.0001	<.0001	0.0386	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0006	<.0001
symmetry_se	(0.1043)	0.0975	(0.0816)	(0.0720)	0.2832	0.2481	0.1780	0.2954	0.4441	0.3455	0.2405	0.4116	0.1341	0.4135	0.3947	0.3942	0.2228	0.3128	1.0000	0.3491	0.2350	0.0470	0.1938	(0.1128)	0.0273	0.1706	0.2146	0.2359	0.0426	0.0327
	0.0128	0.2245	0.0516	0.0840	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.1875	<.0001		<.0001	0.0023	0.0468	0.0133	0.0084	0.7631	0.1512	0.3748	0.4490	<.0001	0.0427
fractal_dimension_se	(0.0424)	0.0543	(0.0055)	(0.0199)	0.2851	0.3073	0.4493	0.3576	0.3318	0.6881	0.2278	0.2797	0.2441	0.1271	0.4274	0.8021	(0.0042)	0.4110	0.3491	1.0000	(0.0375)	(0.0032)	(0.0010)	(0.0227)	0.1706	0.3962	0.3850	0.2132	0.1111	0.5913
	0.3099	0.1964	0.8954	0.6359	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.9209	<.0001	<.0001		0.3701	0.9391	0.8410	0.5884	<.0001	<.0001	<.0001	<.0001	0.0080	<.0001
radius_worst	0.9465	0.3316	0.9435	0.9428	0.2851	0.5353	0.6882	0.8303	0.1857	(0.2537)	0.7151	(0.1117)	0.6972	0.7574	(0.2207)	0.3042	(0.1010)	0.3581	(0.1281)	(0.0375)	1.0000	0.3599	0.9937	0.9840	0.2166	0.4758	0.5740	0.7874	0.2433	0.0935
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0077	<.0001	<.0001	<.0001	<.0001	0.0149	<.0001	0.0023	0.3701		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0252
texture_worst	0.2970	0.9120	0.3030	0.2873	0.0584	0.2481	0.2999	0.2935	0.0901	0.0513	0.1948	0.4090	0.2004	0.1965	(0.0745)	0.1450	(0.0183)	0.0861	(0.0470)	(0.0032)	0.3599	1.0000	0.3651	0.3436	0.2314	0.3405	0.3484	0.3398	0.3332	0.2191
	<.0001	<.0001	<.0001	<.0001	0.3904	<.0001	<.0001	<.0001	0.0036	0.2231	<.0001	<.0001	<.0001	<.0001	0.0748	0.0006	0.6438	0.0386	0.0468	0.9391	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
perimeter_worst	0.9631	0.3498	0.9704	0.9591	0.2661	0.5092	0.7296	0.8591	0.2192	(0.2052)	0.7197	(0.1022)	0.7210	0.7613	(0.2177)	0.2455	(0.0890)	(0.1938)	(0.0228)	(0.0010)	0.9937	0.3651	1.0000	0.9772	0.2098	0.4369	0.5332	0.7964	0.2188	0.0820
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0341	<.0001	0.5875	0.9209	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
area_worst	0.9411	0.3436	0.9416	0.9592	0.2744	0.5096	0.6760	0.8076	0.1772	(0.2319)	0.7516	(0.0853)	0.7307	0.8114	(0.1842)	0.1994	(0.1036)	0.3423	(0.1128)	(0.0227)	0.9840	0.3436	0.9772	1.0000	0.2092	0.4383	0.5433	0.7474	0.2092	0.0797
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0473	<.0001	<.0001	<.0001	<.0001	0.0135	<.0001	0.0084	0.5884	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	0.0576
smoothness_worst	0.1196	0.0772	0.1506	0.1233	0.8033	0.3455	0.4488	0.4328	0.4267	0.5549	0.1419	(0.0737)	0.1301	0.1254	0.8019	0.2820	0.1857	0.4529	0.2154	(0.0227)	0.1706	0.2314	0.2098	0.2092	1.0000	0.5682	0.5188	0.5477	0.4938	0.4176
	0.0043	0.0647	0.0032	0.0199	<.0001	<.0001	<.0001	<.0001	<.0001	0.0007	0.0792	0.0019	0.0037	<.0001	<.0001	0.0019	0.9837	<.0001	0.7631	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001
compactness_worst	0.4130	0.2776	0.4558	0.3904	0.4724	0.8638	0.7150	0.6675	0.4372	0.4568	0.2871	0.1342	0.3419	0.2820	0.5768	0.5788	0.0140	0.5496	0.4229	0.1512	0.4758	0.3405	0.4369	0.4383	0.5682	1.0000	0.8923	0.8011	0.6144	0.4180
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.8831	<.0001	<.0001	0.3748	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001
concavity_worst	0.3269	0.3010	0.3639	0.3134	0.4349	0.8143	0.8841	0.7524	0.4337	0.3442	0.3806	(0.0490)	0.4189	0.3851	(0.0583)	0.6832	(0.0140)	0.5496	0.5371	0.3850	0.5740	0.3484	0.4183	0.5433	0.5188	0.8923	1.0000	0.8554	0.5325	0.4865
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.1003	<.0001	<.0001	0.1449	<.0001	0.7929	<.0001	0.3748	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001
concave_points_worst	0.7442	0.2953	0.7712	0.7220	0.5301	0.8156	0.8613	0.9110	0.4303	0.1753	0.3311	(0.1196)	0.5549	0.5382	(0.1230)	0.4832	(0.0831)	0.6025	(0.0304)	0.2132	0.7874	0.3398	0.8163	0.7474	0.5477	0.8011	0.8554	1.0000	0.5025	0.3111
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0043	<.0001	<.0001	<.0001	<.0001	0.0149	<.0001	0.0473	<.0001	0.4490	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001
symmetry_worst	0.1640	0.1050	0.1891	0.1436	0.3943	0.5102	0.4095	0.3757	0.6998	0.3340	0.0945	0.1282	0.1099	0.0741	(0.1075)	0.3179	(0.0196)	0.1431	0.3894	0.1111	0.2433	0.2350	0.3459	0.2092	0.4938	0.6144	0.5325	0.5025	1.0000	0.5378
	<.0001	0.0132	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0241	0.0022	0.0087	0.0773	0.0104	<.0001	0.4409	0.0006	<.0001	0.0080	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001
fractal_dimension_worst	0.0071	0.1192	0.0510	0.0037	0.4769	0.4874	0.5149	0.3487	0.4384	0.7673	0.0496	(0.0407)	0.0854	0.0175	0.1075	0.5910	(0.0198)	0.3107	0.0781	0.5913	0.0935	0.2191	0.1390	0.0797	0.4176	0.4180	0.4865	0.3111	0.5378	1.0000
	0.8644	0.0044	0.2243	0.9291	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.2379	0.2479	0.0416	0.6763	0.0115	<.0001	0.8531	<.0001	0.8531	<.0001	0.0252	<.0001	0.0576	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	

Figure 30

The Correlation Coefficients show significant relationships in certain variables (highlighted green in Figure 30). Since this table is intimidating, we rely on the cumulative coefficients identified in the Principle Components Analysis. Therefore, are satisfied with our visual and statistical analyses and will proceed in running the LDA with nine (9) principle components.

We decoded the Principle Components in the table, below (Figure 31):

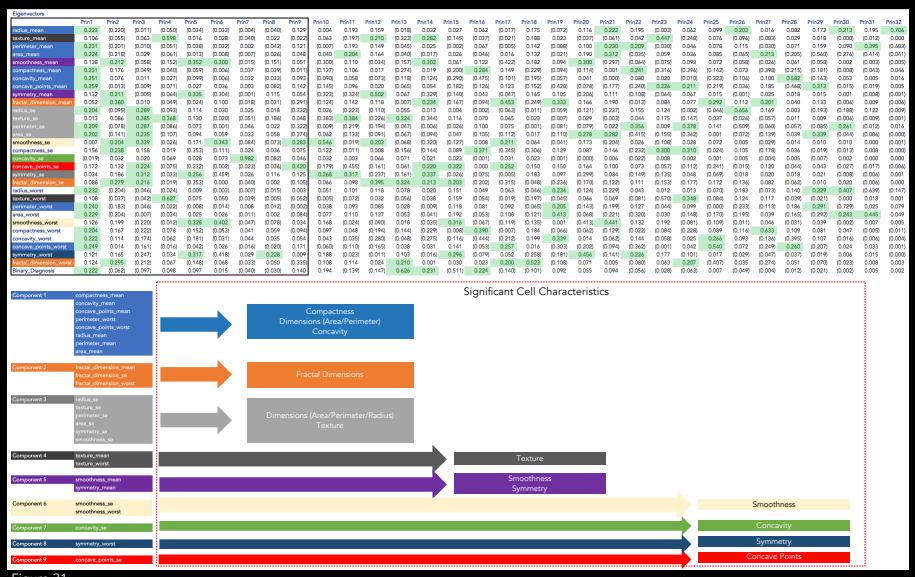
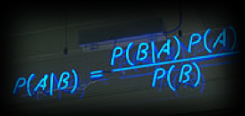


Figure 31



INTERPRETATION

In order to properly interpret our results, we review Bayes' Theorem[3]:

"…the probability of an event, based on prior knowledge of conditions that might be related to the event."

In our case, we expect our model to consider the probability of cells being classified as "Benign" or "Malignant", given the test model and variables chosen in the PCA model. We assumed a 50/50 chance of classification (although we could have used our knowledge of the data being 63%/37% Benign vs Malignant).

We allowed SAS to determine whether to use a Linear Discriminate Analysis or a Quadratic Discriminate Analysis. Since the P-Value for the Test of Homogeneity was significant, the model selected a Quadratic Discriminate Analysis (Covariance Matrices are not same for Benign and Malignant -- Figure 32).

Our SAS output reveals a split of **64.85%** Benign and **35.15%** Malignant, given priors of 50/50 (Figure 33).

The Cross-Validation Summary shows the error-rates by Diagnosis; about 3% for Benign and 8.5% for Malignant diagnoses, respectively. The average error-rate for both diagnoses is about **5.7%** (Figure 34).

In conclusion, based on *Occam's Razor*, our PCA model identified nine (9) principle components, which explained 93% of the cell characteristics needed to automatically determine the status of a Benign vs Malignant Cancer Cell. Our model – using 569 observations – made these predictions with a 5.7% error-rate.

Note to Professor: in the real world, in the case of a determination as serious as cancer, we discussed that we PERSONALLY would want our error rate to be as low as possible. To test this notion, we ran the PCA with all 15 components and only yielded a .7% improvement in the overall error rate. As a result. we feel good about our analysis from an academic and practical standpoint.

Linear Discriminate Analysis - Wisconsin Cancer Cells Diagnostic Data Set (Standardized))

The DISCRIM Procedure

Total Sample Size	569	DF Total	568
Variables	11	DF Within Classes	567
Classes	2	DF Between Classes	1

Number of Observations Read	569
Number of Observations Used	569

Class Level Information

diagnosis	Variable Name	Frequency	Weight	Proportion	Prior Probability
B	B	357	357.0000	0.627417	0.500000
M	M	212	212.0000	0.372583	0.500000

Within Covariance Matrix Information

diagnosis	Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
B	11	-86.73978
M	11	-77.66674
Pooled	11	-80.21811

Linear Discriminate Analysis - Wisconsin Cancer Cells Diagnostic Data Set (Standardized))

The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
1744.119240	66	<.0001

Figure 32

Linear Discriminate Analysis - Wisconsin Cancer Cells Diagnostic Data Set (Standardized))

The DISCRIM Procedure
Classification Summary for Test Data: WORK.STDCANCERDATA
Classification Summary using Quadratic Discriminant Function

Observation Profile for Test Data	
Number of Observations Read	569
Number of Observations Used	569

Number of Observations and Percent Classified into diagnosis

From diagnosis	B	M	Total
B	351 98.32	6 1.68	357 100.00
M	18 8.49	194 91.51	212 100.00
Total	369 64.85	200 35.15	569 100.00
Priors	0.5	0.5	

Error Count Estimates for diagnosis

	B	M	Total
Rate	0.0168	0.0849	0.0509
Priors	0.5000	0.5000	

Figure 33

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.CANCERDATAPRINCOMP
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into diagnosis

From diagnosis	B	M	Total
B	347 97.20	10 2.80	357 100.00
M	18 8.49	194 91.51	212 100.00
Total	365 64.15	204 35.85	569 100.00
Priors	0.5	0.5	

Error Count Estimates for diagnosis

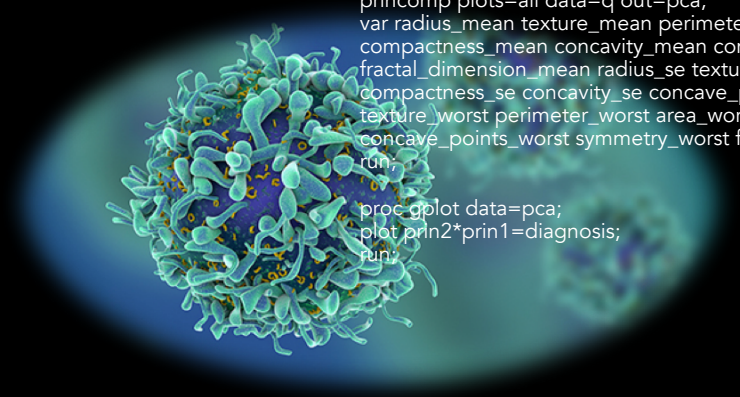
	B	M	Total
Rate	0.0280	0.0849	0.0565
Priors	0.5000	0.5000	

Figure 34

APPENDIX

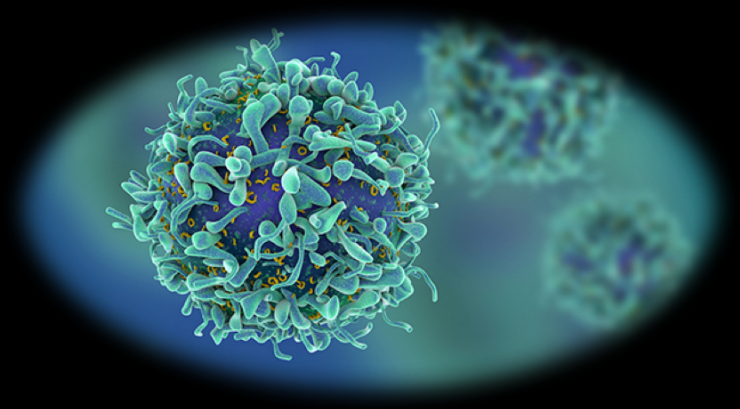
SAS CODE

```
/*Read in Cancer Data*/data CancerData;
infile '/home/rrivens1/MSDS 6372 Applied Statistics Inference and Modeling/CancerData.csv'
dlm=',' firstobs=2 missover;
input id diagnosis $ radius_mean texture_mean perimeter_mean area_mean smoothness_mean
compactness_mean concavity_mean concave_points_mean symmetry_mean
fractal_dimension_mean radius_se texture_se perimeter_se area_se smoothness_se
compactness_se concavity_se concave_points_se symmetry_se fractal_dimension_se radius_worst
texture_worst perimeter_worst area_worst smoothness_worst compactness_worst concavity_worst
concave_points_worst symmetry_worst fractal_dimension_worst Binary_Diagnosis;
*/proc print data=CancerData;
*/run;

/*Create Variable Matrix - Raw Data
*/title "Variable Matrix of Cancer Data Set (Raw)";
proc means data = CancerData;
run;


/*Check for normality - plots - Raw Data*/
title "Scatter Plots - Cancer Data Set (Raw)"
;proc corr data = CancerData plots=matrix(histogram);
var radius_mean texture_mean perimeter_mean area_mean smoothness_mean
compactness_mean concavity_mean concave_points_mean symmetry_mean
fractal_dimension_mean radius_se texture_se perimeter_se area_se smoothness_se
compactness_se concavity_se concave_points_se symmetry_se fractal_dimension_se radius_worst
texture_worst perimeter_worst area_worst smoothness_worst compactness_worst concavity_worst
concave_points_worst symmetry_worst fractal_dimension_worst;
run;

proc means data =CancerData n mean median std var min max maxdec=2;
var radius_mean texture_mean perimeter_mean area_mean smoothness_mean
compactness_mean concavity_mean concave_points_mean symmetry_mean
fractal_dimension_mean radius_se texture_se perimeter_se area_se smoothness_se
compactness_se concavity_se concave_points_se symmetry_se fractal_dimension_se radius_worst
texture_worst perimeter_worst area_worst smoothness_worst compactness_worst concavity_worst
concave_points_worst symmetry_worst fractal_dimension_worst;
run;

proc corr data=CancerData plots=matrix(histogram);
var radius_mean texture_mean perimeter_mean area_mean smoothness_mean
compactness_mean concavity_mean concave_points_mean symmetry_mean
fractal_dimension_mean radius_se texture_se perimeter_se area_se smoothness_se
compactness_se concavity_se concave_points_se symmetry_se fractal_dimension_se radius_worst
texture_worst perimeter_worst area_worst smoothness_worst compactness_worst concavity_worst
concave_points_worst symmetry_worst fractal_dimension_worst;
run;

proc princomp plots=all data=CancerData cov out=pca;
var radius_mean texture_mean perimeter_mean area_mean smoothness_mean
compactness_mean concavity_mean concave_points_mean symmetry_mean
fractal_dimension_mean radius_se texture_se perimeter_se area_se smoothness_se
compactness_se concavity_se concave_points_se symmetry_se fractal_dimension_se radius_worst
texture_worst perimeter_worst area_worst smoothness_worst compactness_worst concavity_worst
concave_points_worst symmetry_worst fractal_dimension_worst;
run;

princomp plots=all data=q out=pca;
var radius_mean texture_mean perimeter_mean area_mean smoothness_mean
compactness_mean concavity_mean concave_points_mean symmetry_mean
fractal_dimension_mean radius_se texture_se perimeter_se area_se smoothness_se
compactness_se concavity_se concave_points_se symmetry_se fractal_dimension_se radius_worst
texture_worst perimeter_worst area_worst smoothness_worst compactness_worst concavity_worst
concave_points_worst symmetry_worst fractal_dimension_worst;
run;

proc gplot data=pca;
plot prin2*prin1=diagnosis;
run;
```



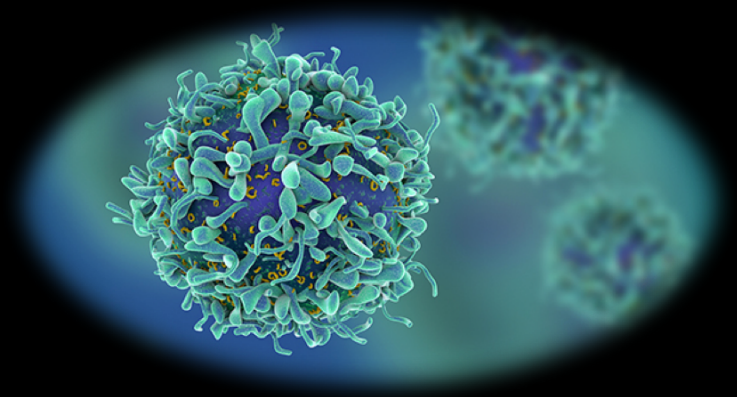
SAS CODE

```
/*Log Transform Variables*/
data CancerDatalog; set CancerData
;radius_mean = log(radius_mean);
texture_mean = log(texture_mean);
perimeter_mean = log(perimeter_mean);
area_mean = log(area_mean);
smoothness_mean = log(smoothness_mean);
compactness_mean = log(compactness_mean);
concavity_mean = log(concavity_mean);
concave_points_mean = log(concave_points_mean);
symmetry_mean = log(symmetry_mean);
fractal_dimension_mean = log(fractal_dimension_mean);
radius_se = log(radius_se);texture_se = log(texture_se);
perimeter_se = log(perimeter_se);area_se = log(area_se);
smoothness_se = log(smoothness_se);
compactness_se = log(compactness_se);
concavity_se = log(concavity_se);
concave_points_se = log(concave_points_se);
symmetry_se = log(symmetry_se);
fractal_dimension_se = log(fractal_dimension_se);
radius_worst = log(radius_worst);
texture_worst = log(texture_worst);
perimeter_worst = log(perimeter_worst);
area_worst = log(area_worst);
smoothness_worst = log(smoothness_worst);
compactness_worst = log(compactness_worst);
concavity_worst = log(concavity_worst);
concave_points_worst = log(concave_points_worst);
symmetry_worst = log(symmetry_worst);
fractal_dimension_worst = log(fractal_dimension_worst);
run;

/*Check for normality - plots - review outliers*/
title "Variable Matrix of Cancer Data Set (Log Transformed)";
proc means data =CancerDatalog n mean median std var min max maxdec=2;
var radius_mean texture_mean perimeter_mean area_mean smoothness_mean compactness_mean
concavity_mean concave_points_mean symmetry_mean fractal_dimension_mean radius_se texture_se
perimeter_se area_se smoothness_se compactness_se concavity_se concave_points_se symmetry_se
fractal_dimension_se radius_worst texture_worst perimeter_worst area_worst smoothness_worst
compactness_worst concavity_worst concave_points_worst symmetry_worst fractal_dimension_worst;
run;

proc corr data=CancerDatalog plots=matrix(histogram);
var radius_mean texture_mean perimeter_mean area_mean smoothness_mean compactness_mean
concavity_mean concave_points_mean symmetry_mean fractal_dimension_mean radius_se texture_se
perimeter_se area_se smoothness_se compactness_se concavity_se concave_points_se symmetry_se
fractal_dimension_se radius_worst texture_worst perimeter_worst area_worst smoothness_worst
compactness_worst concavity_worst concave_points_worst symmetry_worst fractal_dimension_worst;
run;
```



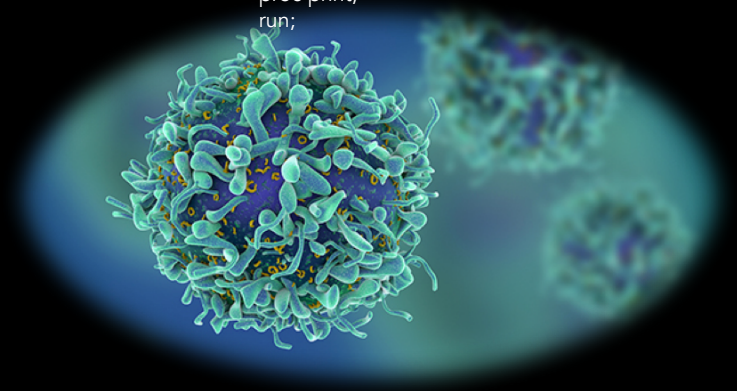
SAS CODE

```
Use corr or covar matrix?;title "Covariant Matrix of Cancer Data Set (Log Transformed)";
proc princomp plots=all data=CancerDatalog cov out=pca;
var radius_mean texture_mean perimeter_mean area_mean smoothness_mean compactness_mean
concavity_mean concave_points_mean symmetry_mean fractal_dimension_mean radius_se texture_se
perimeter_se area_se smoothness_se compactness_se concavity_se concave_points_se symmetry_se
fractal_dimension_se radius_worst texture_worst perimeter_worst area_worst smoothness_worst
compactness_worst concavity_worst concave_points_worst symmetry_worst fractal_dimension_worst;
run;

title "Correlation Matrix of Cancer Data Set by Diagnosis(Log Transformed)";
var radius_mean texture_mean perimeter_mean area_mean smoothness_mean compactness_mean
concavity_mean concave_points_mean symmetry_mean fractal_dimension_mean radius_se texture_se
perimeter_se area_se smoothness_se compactness_se concavity_se concave_points_se symmetry_se
fractal_dimension_se radius_worst texture_worst perimeter_worst area_worst smoothness_worst
compactness_worst concavity_worst concave_points_worst symmetry_worst fractal_dimension_worst;
id diagnosis;
run;

title "Plot of Principle Components Cancer Data Set by Diagnosis (Log Transformed)";
proc gplot data=pca;
plot prin2*prin1=diagnosis;
run;

/*Standardize Select Variables*/

data STDCancerData; set CancerData;
radius_mean = (radius_mean/10);
texture_mean = (texture_mean/10);
perimeter_mean = (perimeter_mean/100);
area_mean = (area_mean/1000);
texture_se = (texture_se/10);
perimeter_se = (perimeter_se/10);
area_se = (area_se/100);r
adius_worst = (radius_worst/100);
texture_worst = (texture_worst/100);
perimeter_worst = (perimeter_worst/100);
area_worst = (area_worst/1000);
run;

/*Check for normality - plots - review outliers*/
title "Variable Matrix of Cancer Data Set (Standardized)";
proc means data =STDCancerData n mean median std var min max maxdec=2;
var radius_mean texture_mean perimeter_mean area_mean smoothness_mean compactness_mean
concavity_mean concave_points_mean symmetry_mean fractal_dimension_mean radius_se texture_se
perimeter_se area_se smoothness_se compactness_se concavity_se concave_points_se symmetry_se
fractal_dimension_se radius_worst texture_worst perimeter_worst area_worst smoothness_worst
compactness_worst concavity_worst concave_points_worst symmetry_worst fractal_dimension_worst;
run;
```



SAS CODE

```
proc corr data=STDCancerData plots=matrix(histogram);
var radius_mean texture_mean perimeter_mean area_mean smoothness_mean compactness_mean concavity_mean concave_points_mean symmetry_mean fractal_dimension_mean radius_se texture_se perimeter_se area_se smoothness_se compactness_se concavity_se concave_points_se symmetry_se fractal_dimension_se radius_worst texture_worst perimeter_worst area_worst smoothness_worst compactness_worst concavity_worst concave_points_worst symmetry_worst fractal_dimension_worst;  run;

Use corr or covar matrix?;
title "Covariant Matrix of Cancer Data Set (Standardized))";
proc princomp plots=all data=STDCancerData cov out=pca;
var radius_mean texture_mean perimeter_mean area_mean smoothness_mean compactness_mean concavity_mean concave_points_mean symmetry_mean fractal_dimension_mean radius_se texture_se perimeter_se area_se smoothness_se compactness_se concavity_se concave_points_se symmetry_se fractal_dimension_se radius_worst texture_worst perimeter_worst area_worst smoothness_worst compactness_worst concavity_worst concave_points_worst symmetry_worst fractal_dimension_worst;  run;

title "Correlation Matrix of Cancer Data Set by Diagnosis(Standardized))";
proc princomp plots=all data=STDCancerData out=pca;
var radius_mean texture_mean perimeter_mean area_mean smoothness_mean compactness_mean concavity_mean concave_points_mean symmetry_mean fractal_dimension_mean radius_se texture_se perimeter_se area_se smoothness_se compactness_se concavity_se concave_points_se symmetry_se fractal_dimension_se radius_worst texture_worst perimeter_worst area_worst smoothness_worst compactness_worst concavity_worst concave_points_worst symmetry_worst fractal_dimension_worst;
run;

title "Plot of Principle Components Cancer Data Set by Diagnosis (Standardized))";
proc gplot data=pca;
plot prin2*prin1=diagnosis;
run;

/*Identify Principle Data - 9 Components*/
proc princomp data=CancerDatalog out=CancerDataPrinComp;run;
proc pls data = CancerDataPrinComp method = pcr nfact=9;
model Binary_Diagnosis = radius_mean texture_mean perimeter_mean area_mean smoothness_mean compactness_mean concavity_mean concave_points_mean symmetry_mean fractal_dimension_mean radius_se texture_se perimeter_se area_se smoothness_se compactness_se concavity_se concave_points_se symmetry_se fractal_dimension_se radius_worst texture_worst perimeter_worst area_worst smoothness_worst compactness_worst concavity_worst concave_points_worst symmetry_worst fractal_dimension_worst;
run;

/*Run LDA*/
title "Linear Discriminate Analysis - Wisconsin Cancer Cells Diagnostic Data Set (Log Transformed)";
proc discrim data = CancerDataPrinComp pool=test crossvalidate testdata=CancerDatalog testout=a;class diagnosis;var radius_mean area_mean smoothness_worst compactness_se concavity_mean concave_points_worst symmetry_se fractal_dimension_mean texture_worst fractal_dimension_worst perimeter_se;
run;
proc print;
run;
```



REFERENCES

[1] https://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html

[2] https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

[3] https://en.wikipedia.org/wiki/Bayes%27_theorem

Images:

http://slideplayer.com/slide/9171054/

http://userwww.sfsu.edu/efc/classes/biol

http://710/discrim/discrim.pdf

https://www.verywell.com/cancer-cells-vs-normal-cells-2248794

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3772701/figure/F3/

Project GitHub Link:

https://github.com/yupster2/Wisconsin-Cancer---PCA---LDA.git