1st project:

Assessing low-income families and affordable housing.

Datasets: Census 5% sample from New Jersey.

We will find out the proportion of families that live housing that are not affordable.
The important thing is to find segments of the population that are not leaving in
affordable housing.

Important variables for affordability are
SMOCAPI
GRAPI

Important variables for low income are: HINC
For a family of 4 Low income is below 80% of median
For 3 lower this by 10%
For 5 higher by 8%

This link has the meaning of the variables in the file:
http://www.census.gov/prod/cen2000/doc/pums.pdf

2nd Project

Loan applications

The data set below consists of a set of variables selected from a database of 768 Loan applications at a bank that have been decided by loan officers.
We want to find to what extent is possible to define some objective rules that define the

You have one week to prepare a short report on this study
(No more than 4 pages of text).

Variables:

JOB:	Number of years in current job (or last job)
CAR:	Owns a car 1:no 2:yes
RACE:	2:Caucasian, 1:Other
SALARY:	Current monthly salary in thousands. Zero means missing or unemployed.
GENDER:	1:Female, 2:Male
SAV:	Savings, Assets in US$1000
OFFICER:	One of four loan officers working for the bank
AGE:	In years
RESPONSE:	1: Got Loan, 0:Not

3rd Project

Market segmentation.

Using our dataset orthopedic you need to find out market segments and hospitals that are likely customers, but where my sales are low. (See textbook)

4th Project Microarray Data: Khan Data

DNA Microarray data. See DNAMR for more details and my book "Exploration and analysis of DNA microarray and Protein array data."

5th Project Microarray Data: Tissue Data

Another DNA Microarray data. See DNAMR package on my website for more details.

6th Project Compound classification: Active or Inactive

7th Project Plastic explosives detection.

Data Set : Pex23

The data comes from a study for the detection of plastic explosives in suitcases using X-ray signals.

The 23 variables are the discrete xcomponents of the xray absorption spectrum.

The response is the last variable in the dataset. It takes two values:

0: There is explosive

1: There is not.

The objective is to detect the suitcases with explosives. (See textbook)

8th Project

Shopping Patterns of TV viewers of ER, Friends, Ally McBill, Fraiser, Jesse.

EXHIBIT 1. Types of merchandise bought from catalogs in the last 12 months by demographic variables and the TV Shows –
Based on Nielsen Media Research and Simmons Market Research Bureau data.

	V 1	V 2	V 3	V 4	V 5	V 6	V 7	V 8
		Total	Clothing	Electro.	Home	Houseware	Sport	Toys/
		US						
		in '000			Furnishing		Goods	Games
1	18-24	23965	3757	560	1211	1009	1005	1263
2	25-34	42832	9668	1706	3422	2459	1798	3852
3	35-44	39908	12381	1970	4641	2732	2441	4377
4	45-54	27327	8500	1563	2578	2293	1363	1788
5	55-64	21238	6001	684	2088	1782	885	1571
6	65 -over	30552	7443	944	2220	1424	436	1243
7	Graduated Collage	36463	13249	2177	3703	2330	2276	3796
8	Attended Collage	44294	12881	2016	4839	3040	1936	3669

9	High School	66741	15820	2182	5408	4682	2806	4883
10	Did not Grad.Hsch.	38324	5802	1053	2210	1647	911	1748
11	T. Male	88956	16824	4231	4789	3861	4442	5339
12	T. Female	96866	30928	3196	11370	7838	3486	8757
13	Employed Male	65500	12746	3232	3548	2786	3610	4227
14	Employed Female	55910	20363	2023	7836	5039	2539	5874
15	Full Time Employed	110363	29409	4926	10149	6965	5546	8830
16	Part-Time Employed	11047	3702	329	1187	860	604	1271
17	Not Employed	64412	14641	2172	4775	3874	1779	3994
18	Single	41284	6962	1485	2135	1752	1487	1550
19	Married	109023	32641	4912	11614	7878	5586	10473
20	Div./Sep./Wid.	35515	8149	1030	2411	2069	856	2073
21	Parents	62342	18215	2968	6702	3988	3523	7714
22	Inc.75,000-more	24165	8600	1067	2679	1941	1209	2705
23	60,000-more	40979	13440	1986	4045	2952	2100	4015
24	50,000-more	57996	18943	2653	6064	4100	3216	5561
25	40,000-more	80078	25274	3726	8441	5561	4488	7420
26	30,000-more	106838	32712	4962	10840	7167	5819	9582
27	20,000-29,000	30669	6539	1200	2497	1842	964	1870
28	10,000-19,999	29083	5594	695	1794	1850	771	1698
29	under 10,000	19232	2907	570	1028	840	374	946
30	**E.R.**	19,640	6216	828	2090	1418	977	1881
31	**Friends**	16,000	3828	563	1207	937	720	903
32	**Frasier**	14,840	4605	590	1560	1224	692	1263
33	**Jesse**	13,550	2594	511	879	674	575	716
34	**Ally McBeal**	10,190	2221	245	654	644	467	680

EXHIBIT 2. Types of merchandise bought from catalogs in the last 12 months by demographic variables and the TV shows as a percentage of the group.

	V 1	V 2	V 3	V 4	V 5	V 6	V 7	V 8
		Total US	**Clothing**	**Electro.**	**Home**	**Houseware**	**Sport**	**Toys/**
		in '000			**Furnishing**		**Goods**	**Games**
1	18-24	23965	15.68	2.34	5.05	4.21	4.19	5.27

2	25-34	42832	22.57	3.98	7.99	5.74	4.20	8.99
3	35-44	39908	31.02	4.94	11.63	6.85	6.12	10.97
4	45-54	27327	31.10	5.72	9.43	8.39	4.99	6.54
5	55-64	21238	28.26	3.22	9.83	8.39	4.17	7.40
6	65 –over	30552	24.36	3.09	7.27	4.66	1.43	4.07
7	Graduated Collage	36463	36.34	5.97	10.16	6.39	6.24	10.41
8	Attended Collage	44294	29.08	4.55	10.92	6.86	4.37	8.28
9	High School	66741	23.70	3.27	8.10	7.02	4.20	7.32
10	Did not Grud.Hsch.	38324	15.14	2.75	5.77	4.30	2.38	4.56
11	T.Male	88956	18.91	4.76	5.38	4.34	4.99	6.00
12	T.Female	96866	31.93	3.30	11.74	8.09	3.60	9.04
13	Employed Male	65500	19.46	4.93	5.42	4.25	5.51	6.45
14	Employed Female	55910	36.42	3.62	14.02	9.01	4.54	10.51
15	Full Time Employed	110363	26.65	4.46	9.20	6.31	5.03	8.00
16	Part-Time Employed	11047	33.51	2.98	10.74	7.78	5.47	11.51
17	Not Employed	64412	22.73	3.37	7.41	6.01	2.76	6.20
18	Single	41284	16.86	3.60	5.17	4.24	3.60	3.75
19	Married	109023	29.94	4.51	10.65	7.23	5.12	9.61
20	Div./Sep./Wid.	35515	22.95	2.90	6.79	5.83	2.41	5.84
21	Parents	62342	29.22	4.76	10.75	6.40	5.65	12.37
22	Inc.75,000-more	24165	35.59	4.42	11.09	8.03	5.00	11.19
23	60,000-more	40979	32.80	4.85	9.87	7.20	5.12	9.80
24	50,000-more	57996	32.66	4.57	10.46	7.07	5.55	9.59
25	40,000-more	80078	31.56	4.65	10.54	6.94	5.60	9.27
26	30,000-more	106838	30.62	4.64	10.15	6.71	5.45	8.97
27	20,000-29,000	30669	21.32	3.91	8.14	6.01	3.14	6.10
28	10,000-19,999	29083	19.23	2.39	6.17	6.36	2.65	5.84
29	under 10,000	19232	15.12	2.96	5.35	4.37	1.94	4.92
30	**E.R.**	19,640	31.65	4.22	10.65	7.22	4.97	9.58
31	**Friends**	16,000	23.92	3.52	7.54	5.86	4.5	5.64
32	**Freiser**	14,840	31.03	3.98	10.51	8.25	4.66	8.51
33	**Jesse**		19.14	3.77	6.49	4.97	4.24	5.28

34	**Ally McBill**	13,550 10,190	21.80	2.40	6.42	6.32	4.58	6.67

In order to determine the shopping profile of our target segment customer who watches a particular TV show we performed an analysis, which consists of a combination of factor, and cluster analyses. The cluster analysis produces an association between groups of viewers with certain demographic characteristics and specific TV shows.

9th project

Pima Indians: Very famous data set about ladies in a tribe of Pima Indians whso appear to have high incidence of diabetes. Watch out with zeros that are really missing values.

Compare the following methods:

GLM net
SVM
Random forest
Boosting