Wrangling and Exploring Data on a Path to Understanding and Hypotheses

Jim Grayson, PhD, Professor

Houssein Sater, MD, Translational Immunotherapy Fellow

John Janik, MD, Director of Cancer Center Immune Therapy Clinical Trial Program

Augusta University

Georgia’s second-oldest and second-largest city, Augusta, is situated on the southern banks of the storied Savannah River.

AUGUSTA UNIVERSITY

Offering undergraduate programs in the liberal arts and sciences, business and education as well as a full range of graduate programs and hands-on clinical research opportunities, Augusta University is Georgia’s innovation center for education and health care.

The combination of nationally ranked business and nursing schools as well as the state’s flagship public medical school and only dental school makes Augusta University a destination of choice for the students of today and the leaders of tomorrow.

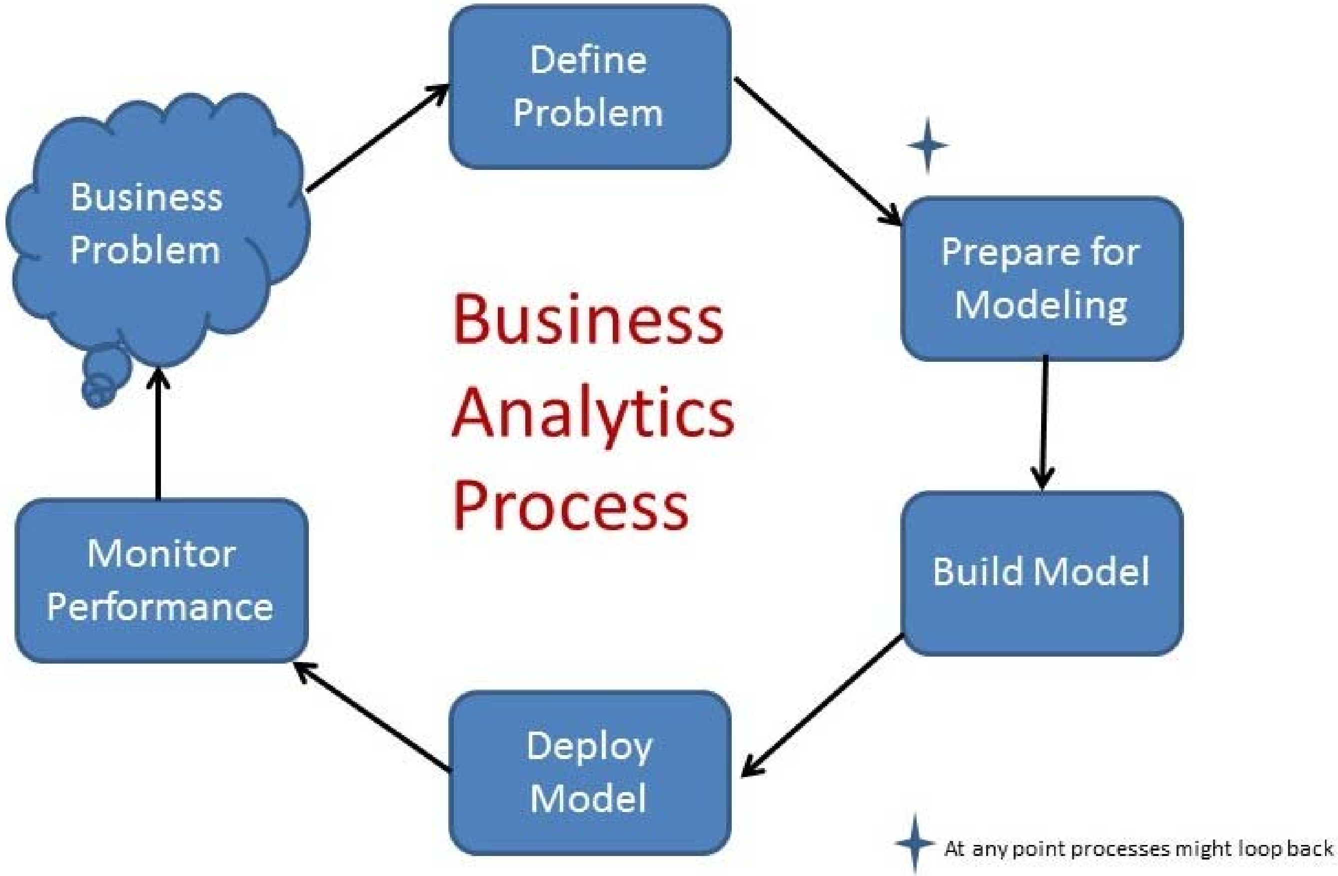
Presentation Goal

Developing models is not easy … It is **messy, iterative and multi-pathed process** often circling back many times.

Our data is a **small cancer data set** with a small number of observations and many potential predictors of which the majority are categorical variables.

We will show the iterative process of exploring, understanding and eventually coming to an understanding of our predictors and hypotheses for further study.

Analytics Process



Wrangling and Exploring

"Begin with the end in mind" (Stephen Covey)

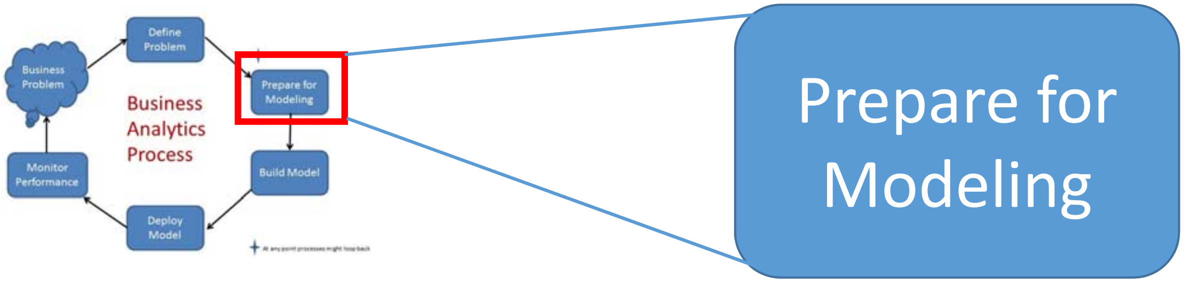
- What is the business goal?
- What is success?
- How will the model be used?

Data Wrangling Goal

Explore the differences in outcomes based on several parameters, clinical and pathologic, that are known or suspected to shape clinical outcomes.

Primary outcomes were considered to be overall survival, and overall response at time of last follow up.

Want to develop an understanding of data and relationships to propose hypotheses for next step.



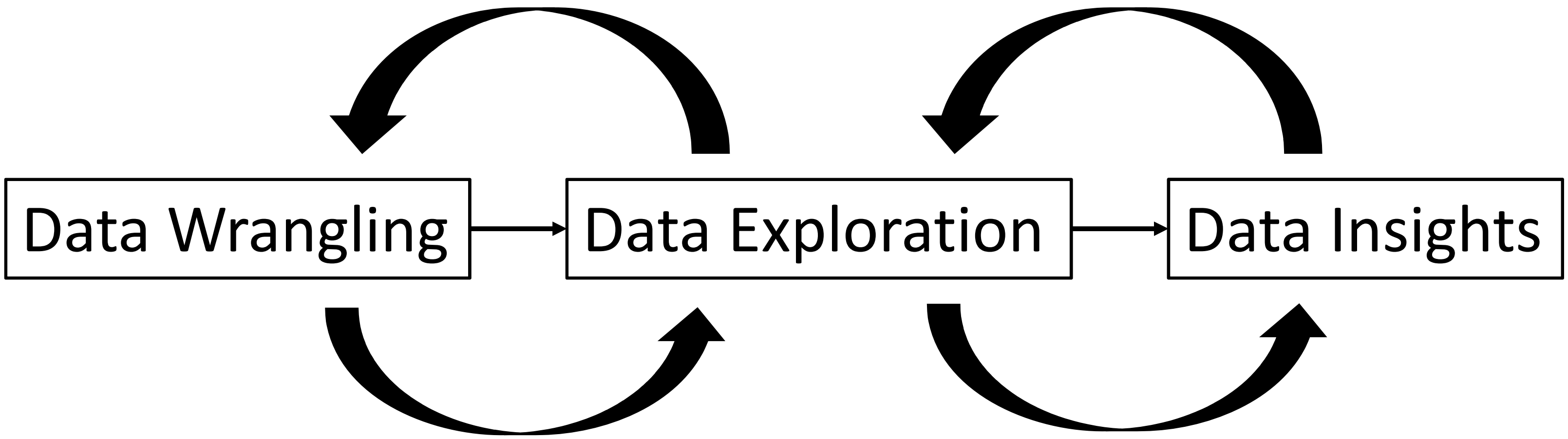
Define/Acquire Data
- Compile
- Combine
- Structure

Understand Data
- Explore
- Examine
- Characterize

Assess Data Quality
- Missing
- Outliers
- Potential Issues

Restructuring Data
- Recode
- Transform
- Features

Dimension Reduction
- Predictor Screening
- Graphical Exploration for Insights

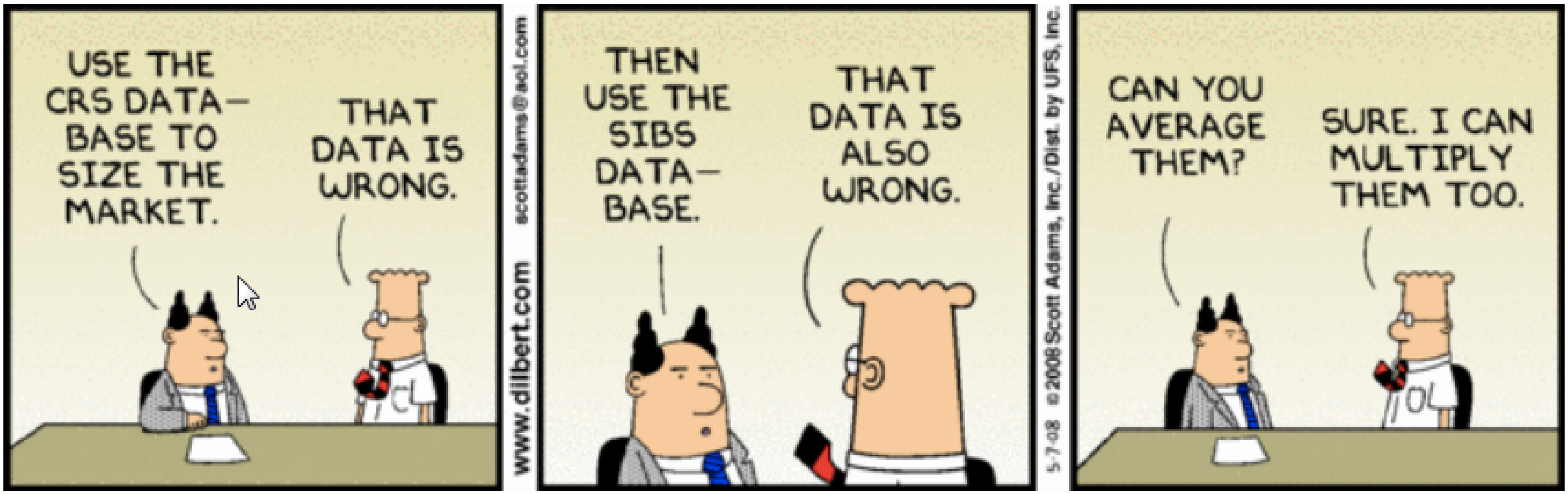


- Missing Data
- Outliers
- Recoding Variables
- Data Features

- One Way Analysis
- Two Way Analysis
- Multiple Variables

- Local Filter
- Global Filter
- Graph Builder

Suitable Data?



Georgia Cancer Center Research Programs

Methods

All female patients with invasive breast cancer treated at GCC from 2005-2010 were chart analyzed retrospectively. Data initially pulled from GCC registry included demographics as age at diagnosis, DOB, gender, ethnicity, as well as diagnostic and treatment data on these patients.

These data were further screened during chart review phase. Missing data were filled mostly by doing in-depth chart review. Despite this effort, many charts continued to have missing data.

Variables by Category

Demographic Data: Data known about patient before start of treatment

Sex: Female
Race: Black, White or other
Age at Diagnosis: in years
Weight (kg) at time of diagnosis
Height (cm) at time of diagnosis
BMI (Body Mass Index): A number calculated based on weight and height indicating how lean or obese a person is.

Family History of cancers

Alcohol Hx: Cumulative Amount of alcohol consumed in a subjective way of classification
Tobacco Use History: Based on period and quantity of cigarettes consumed over years
Past Medical Histories: Known medical problems diagnosed before diagnosis of cancer
Cardiac: Heart problems
DM: Diabetes Mellitus
Lung: Lung problems
CMI: Comorbidity index, sum of number of comorbid conditions that the patient has. The more sick a patient, the higher the number

Variables by Category (continued)

Clinical Data:
(data recorded for patient while on treatment)

Surgery: Indicates if tumor removed initially
Chemotherapy: Indicates whether patient received chemotherapy
Radiation: Indicates whether patient received radiation therapy
Hormone: Indicates whether patient received hormonal therapy
Recurrence Date: indicate when a tumor disappeared then came back
Progression Date: Indicates when did tumor continue to grow
Vital Status: Alive or Dead at last visit, censor
Response: Implies overall response at last follow up
Survival: years lived before last follow up or death

Pathologic Data:
<u>Grade/Differentiation</u>
Pathologic T: Size of tumor on dg
Pathologic N: lymph node status on diagnosis
Pathologic M: metastatic state of disease on diagnosis
Pathologic Stage Group Best CS/AJCC Stage: stage of tumor based on T,N, and M
Tumor Characteristics
ER %: percent expression of Estrogen receptor
PR %: percent expression of Progesterone receptor
HER2: Expression of human epidermal receptor 2
Ki67 %: proliferation index
Lymphovascular Invasion: Presence of cancer in lymphatic vessels

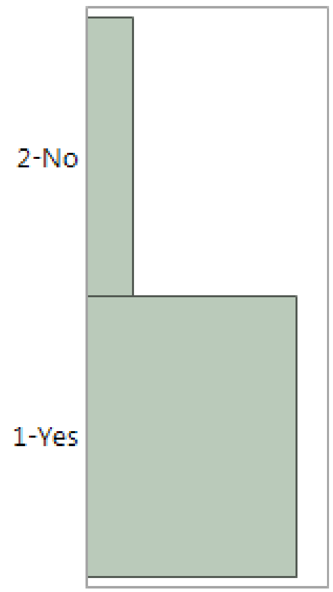
Mile Wide ... Inch Deep

Univariate Data Exploration

JMP Tools used to explore variables

Use **Analyze > Distribution** to look individually at variables

Response



Frequencies

Level	Count	Prob
1-Yes	307	0.82086
2-No	67	0.17914
Total	374	1.00000

N Missing 26

2 Levels

Two Levels:
Our target focus: Yes*

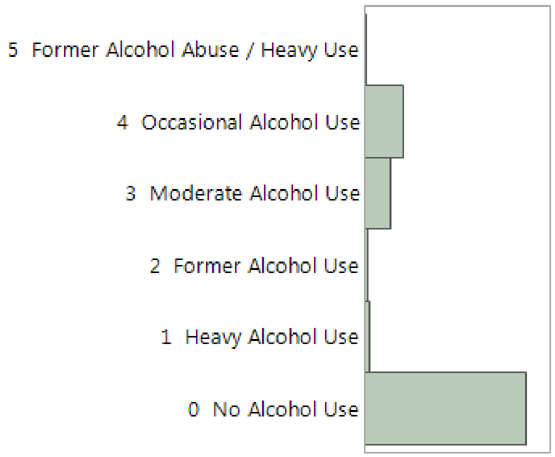
We wanted to understand the patients that survived -- these patients favorably responded to treatment

Use **Cols > Column Viewer** for snapshot of number of observations, missing values and characteristics of categorical and continuous variables

41 Columns Clear Select Distribution

Columns	N	N Missing	N Categories	Min	Max	Mean	Std Dev
BCSubtype	380	20	4	.	.	.	.
Race	400	0	3	.	.	.	.
Age_at_Diagnosis	400	0	.	25	91	59.5275	13.239238548
BMI	400	0	.	0	59.8	30.030264235	5.9366208298
Alcohol_Hx	400	0	6	.	.	.	.
Tobacco_Use_History	400	0	7	.	.	.	.
Tobacco	400	0	3	.	.	.	.
Alcohol	400	0	2	.	.	.	.
Cardiac	400	0	2	.	.	.	.
DM	400	0	2	.	.	.	.
Lung	400	0	2	.	.	.	.
CMI	400	0	.	0	3	0.905	0.7984791308
Family_Hx_br_cancer	339	61	3	.	.	.	.
FHx_Ovarian	337	63	2	.	.	.	.
Family_hx_of_other_cancers	337	63	2	.	.	.	.
Family_History	400	0	5	.	.	.	.
Family_Hx	400	0	2	.	.	.	.
Local_Recurrence	400	0	2	.	.	.	.
Distant_Recurrence	400	0	2	.	.	.	.
AJCC_Stage	400	0	11	.	.	.	.

Alcohol_Hx

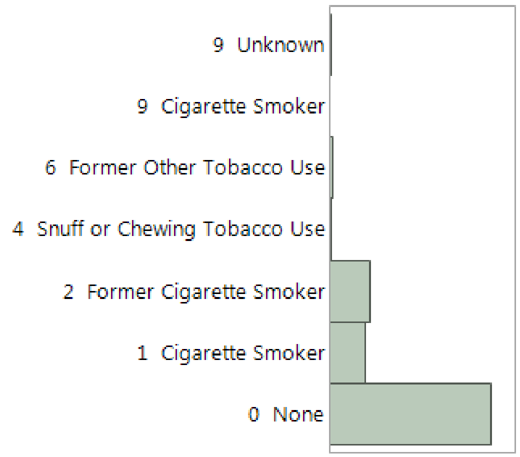


Frequencies

Level	Count	Prob
0 No Alcohol Use	275	0.68750
1 Heavy Alcohol Use	8	0.02000
2 Former Alcohol Use	5	0.01250
3 Moderate Alcohol Use	45	0.11250
4 Occasional Alcohol Use	65	0.16250
5 Former Alcohol Abuse / Heavy Use	2	0.00500
Total	400	1.00000

N Missing 0
6 Levels

Tobacco_Use_History



Frequencies

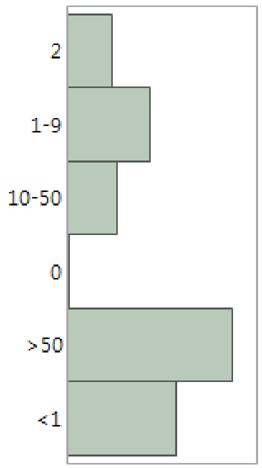
Level	Count	Prob
0 None	264	0.66000
1 Cigarette Smoker	59	0.14750
2 Former Cigarette Smoker	66	0.16500
4 Snuff or Chewing Tobacco Use	3	0.00750
6 Former Other Tobacco Use	4	0.01000
9 Cigarette Smoker	1	0.00250
9 Unknown	3	0.00750
Total	400	1.00000

N Missing 0
7 Levels

Use **Distribution** platform to identify variables with many levels or very low observations in a level

Use **Cols > Recode** to reduce many levels to fewer levels

ER_Pct



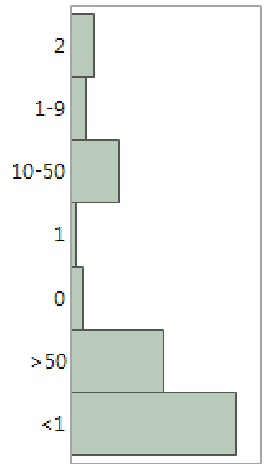
Frequencies

Level	Count	Prob
<1	96	0.24242
>50	145	0.36616
0	1	0.00253
10-50	43	0.10859
1-9	72	0.18182
2	39	0.09848
Total	396	1.00000

N Missing 4

6 Levels

PR_Pct



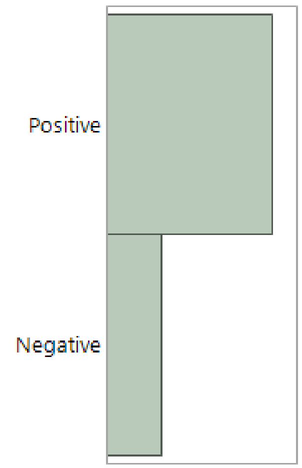
Frequencies

Level	Count	Prob
<1	182	0.45844
>50	102	0.25693
0	12	0.03023
1	6	0.01511
10-50	52	0.13098
1-9	17	0.04282
2	26	0.06549
Total	397	1.00000

N Missing 3

7 Levels

ER_PR 2



Frequencies

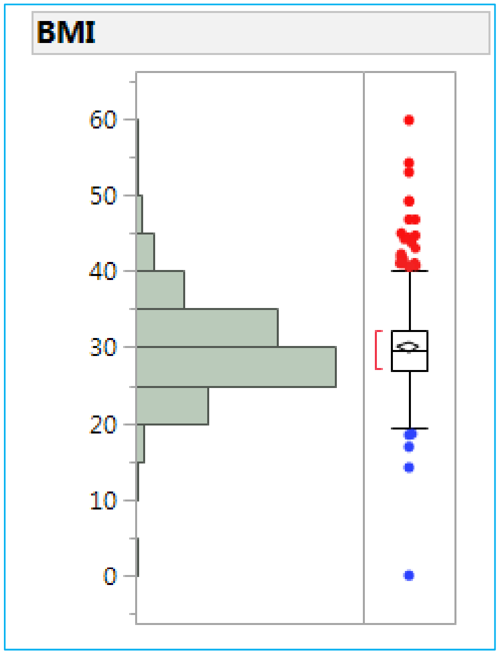
Level	Count	Prob
Negative	99	0.24750
Positive	301	0.75250
Total	400	1.00000

N Missing 0

2 Levels

Use **Formula** tool (Comparison and Conditional) to create a feature variable

Exploring Potential Outliers



- Use **Lasso tool** to select outliers
- Use **Name Selection in Column** to Mark Yes or No
- Use **Selected | Data View** to examine subset of variables

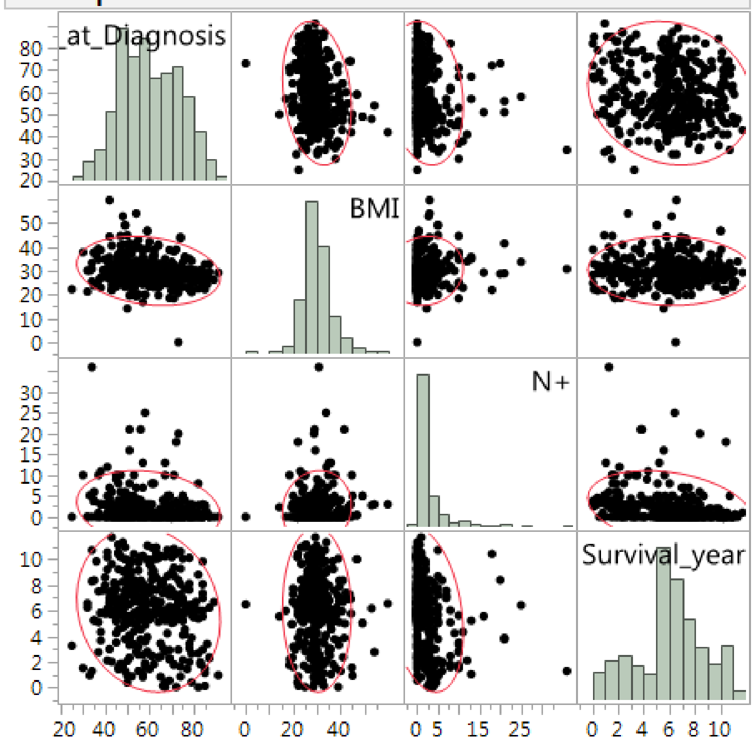
Bivariate and Multivariate Data Exploration

JMP Tools used to explore association between variables

Correlations

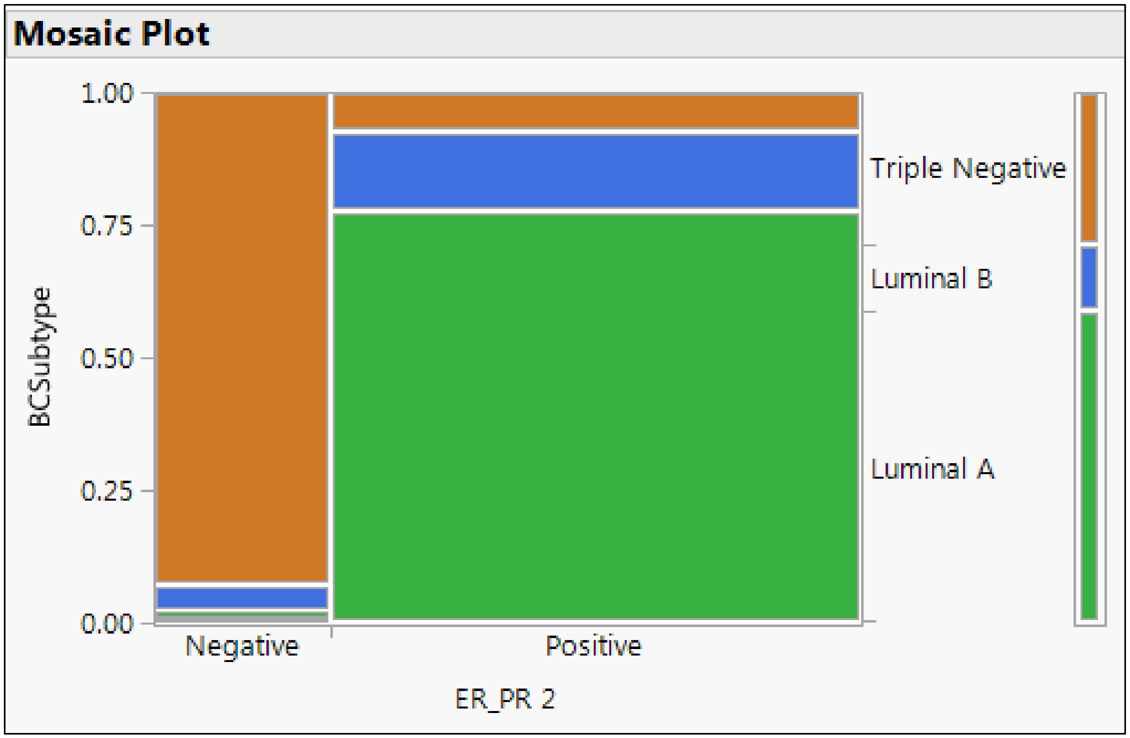
	Age_at_Diagnosis	BMI	N+	Survival_year
Age_at_Diagnosis	1.0000	-0.2047	-0.1787	-0.1338
BMI	-0.2047	1.0000	0.0784	-0.0323
N+	-0.1787	0.0784	1.0000	-0.2850
Survival_year	-0.1338	-0.0323	-0.2850	1.0000

Scatterplot Matrix



Use **Analyze > Multivariate Methods > Multivariate** to examine associations between continuous variables

Cancer Subtype vs Triple Negative Feature Variable



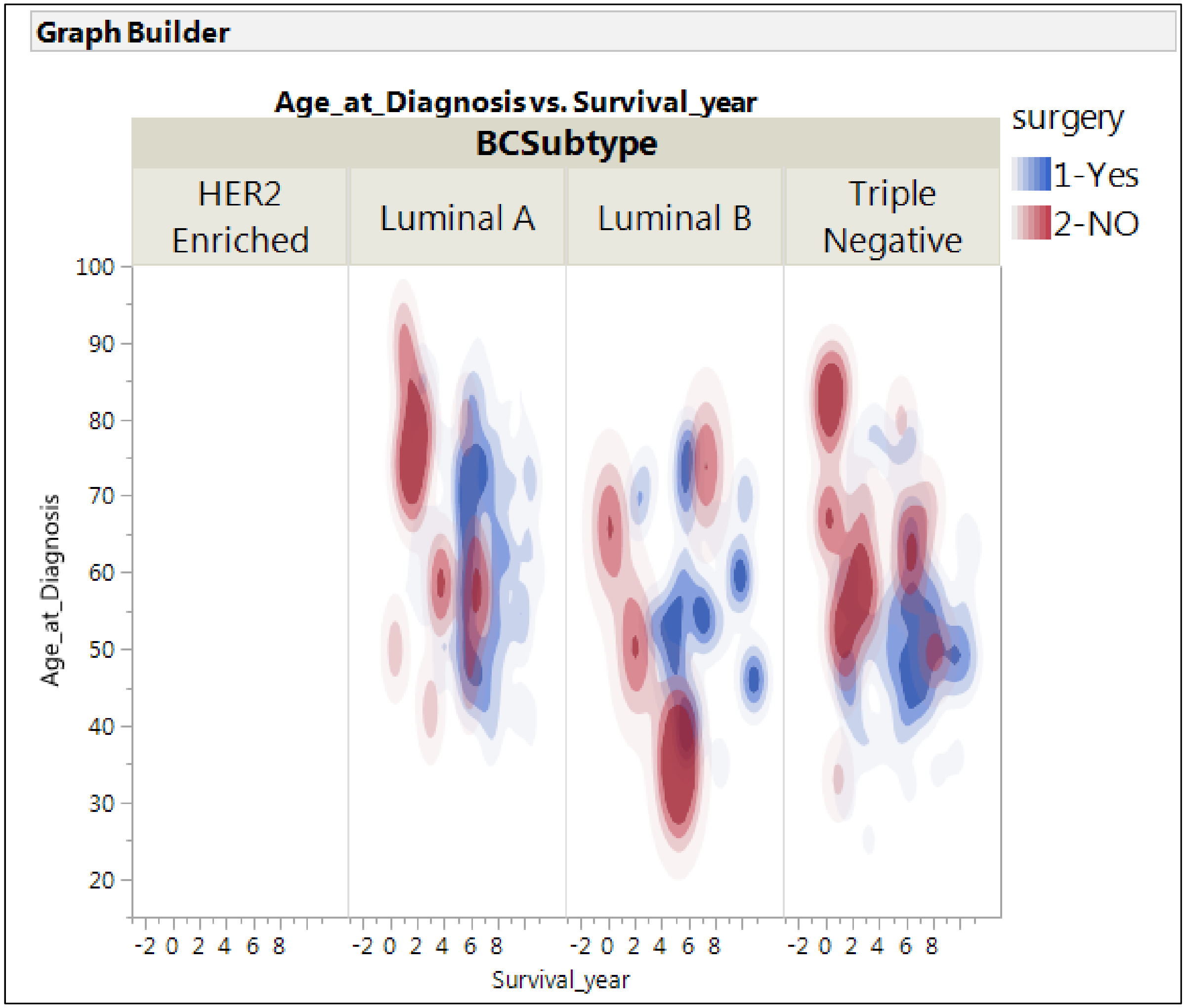
Use **Fit Y by X (Mosaic Plot)** to examine categorical response and categorical variable for association.

In this instance the plot revealed an internal inconsistency with our data

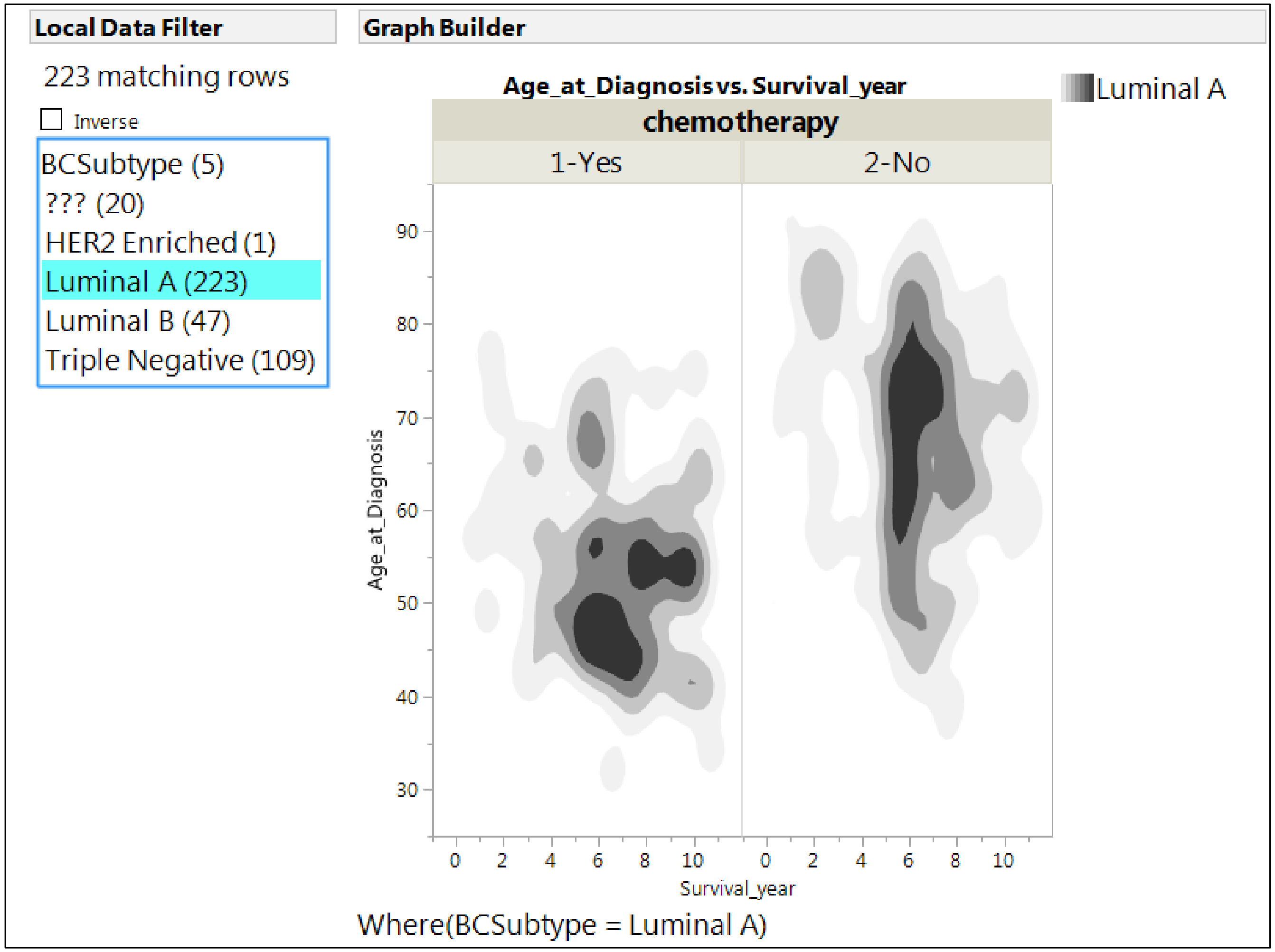
Use Graph Builder to Explore Relationships

JMP tools used to interactively explore multidimensional relationships

Use interactive **Graph Builder** to explore for multiple dimension .. In this instance looking at breast cancer subtypes by age and surgery (yes or no) to observe the impact of patients patient's survival year

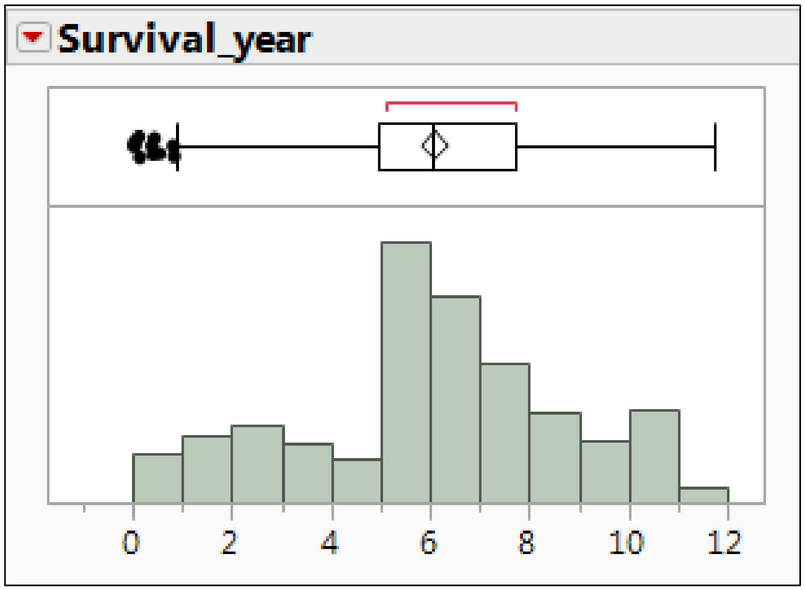


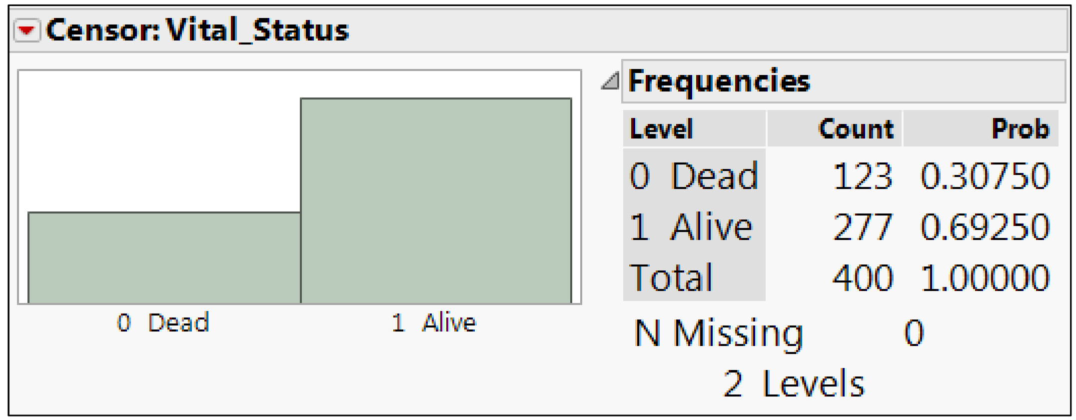
Use **Graph Builder** combined with **Local Data Filter** (breast cancer subtype) and **Column Switcher** (to one at a time observe each treatment type) to see associations between age at diagnosis and survival year



Circling Back to the Goal: Focusing on Those Who Survived

Two Ways to Distinguish Those Who Survive

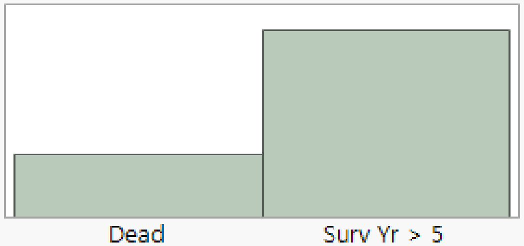




Level	Count	Prob
0 Dead	123	0.30750
1 Alive	277	0.69250
Total	400	1.00000

N Missing 0
2 Levels

Patient Survive > 5 years



Frequencies

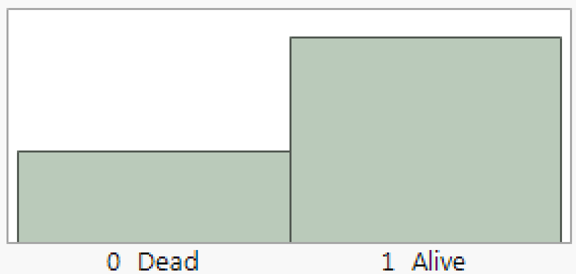
Level	Count	Prob
Dead	100	0.25000
Surv Yr > 5	300	0.75000
Total	400	1.00000

N Missing 0

2 Levels

Recode Years Survival to Two Levels

Censor: Vital_Status



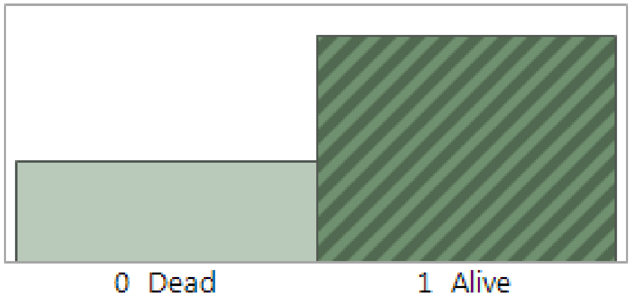
Frequencies

Level	Count	Prob
0 Dead	123	0.30750
1 Alive	277	0.69250
Total	400	1.00000

N Missing 0

2 Levels

Censor: Vital_Status



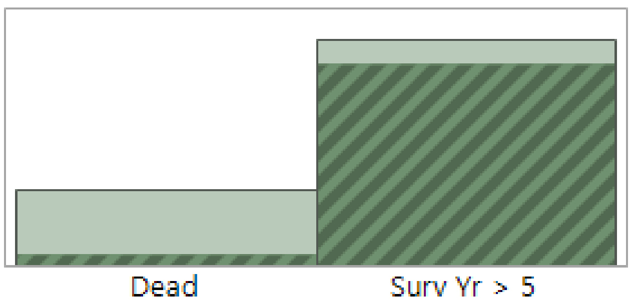
Frequencies

Level	Count	Prob
0 Dead	123	0.30750
1 Alive	277	0.69250
Total	400	1.00000

N Missing 0

2 Levels

Patient Survive > 5 years



Frequencies

Level	Count	Prob
Dead	100	0.25000
Surv Yr > 5	300	0.75000
Total	400	1.00000

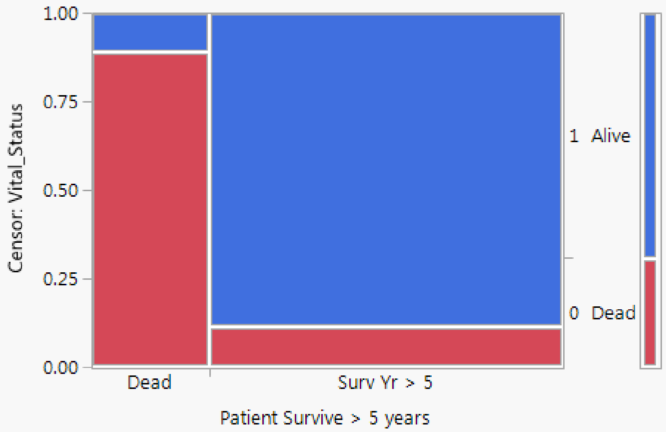
N Missing 0

2 Levels

Recode shows Alive and Patient Survival

Contingency Analysis of Censor: Vital_Status By Patient Survive > 5 years

Mosaic Plot



Contingency Table

Censor: Vital_Status

Patient Survive > 5 years	Count / Total % / Col % / Row %	0 Dead	1 Alive	Total
	Dead	89	11	100
		22.25	2.75	25.00
		72.36	3.97	
		89.00	11.00	
	Surv Yr > 5	34	266	300
		8.50	66.50	75.00
		27.64	96.03	
		11.33	88.67	
	Total	123	277	400
		30.75	69.25	

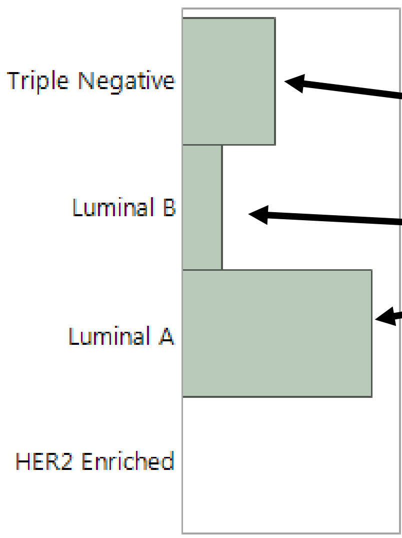
88.67% is the Conditional Probability of Patient Being Alive ***[Censor: Vital Status]***

given

Patient Survived More Than Five Years

[Patient Survive > 5 Years]

BCSubtype



Frequencies

Level	Count	Prob
HER2 Enriched	1	0.00263
Luminal A	223	0.58684
Luminal B	47	0.12368
Triple Negative	109	0.28684
Total	380	1.00000

N Missing 20
4 Levels

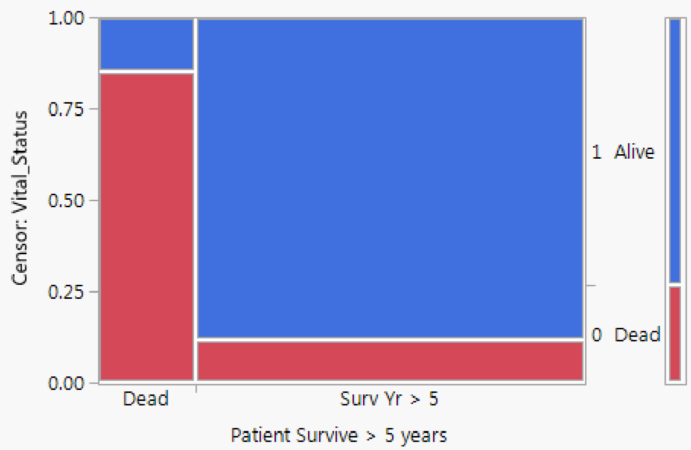
Triple Negative vs
Combined **Luminal A and B**

Comparing Subtypes

Triple Negative

Luminal A
Luminal B

Mosaic Plot



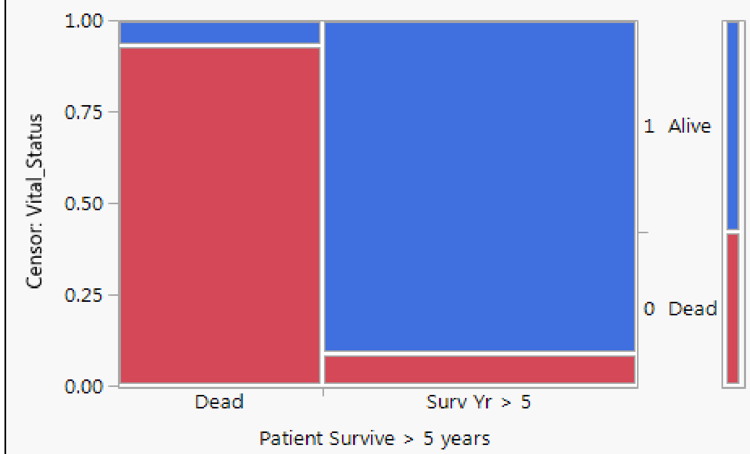
Contingency Table

Censor: Vital_Status

Patient Survive > 5 years	Count / Total % / Col % / Row %	0 Dead	1 Alive	Total
Dead	Count	46	8	54
	Total %	17.04	2.96	20.00
	Col %	63.89	4.04	
	Row %	85.19	14.81	
Surv Yr > 5	Count	26	190	216
	Total %	9.63	70.37	80.00
	Col %	36.11	95.96	
	Row %	12.04	87.96	
Total	Count	72	198	270
	Total %	26.67	73.33	

Cond Prob

Mosaic Plot



Contingency Table

Censor: Vital_Status

Patient Survive > 5 years	Count / Total % / Col % / Row %	0 Dead	1 Alive	Total
Dead	Count	40	3	43
	Total %	36.70	2.75	39.45
	Col %	86.96	4.76	
	Row %	93.02	6.98	
Surv Yr > 5	Count	6	60	66
	Total %	5.50	55.05	60.55
	Col %	13.04	95.24	
	Row %	9.09	90.91	
Total	Count	46	63	109
	Total %	42.20	57.80	

Cond Prob

Based on exploration decided to focus on <u>Triple Negative</u>

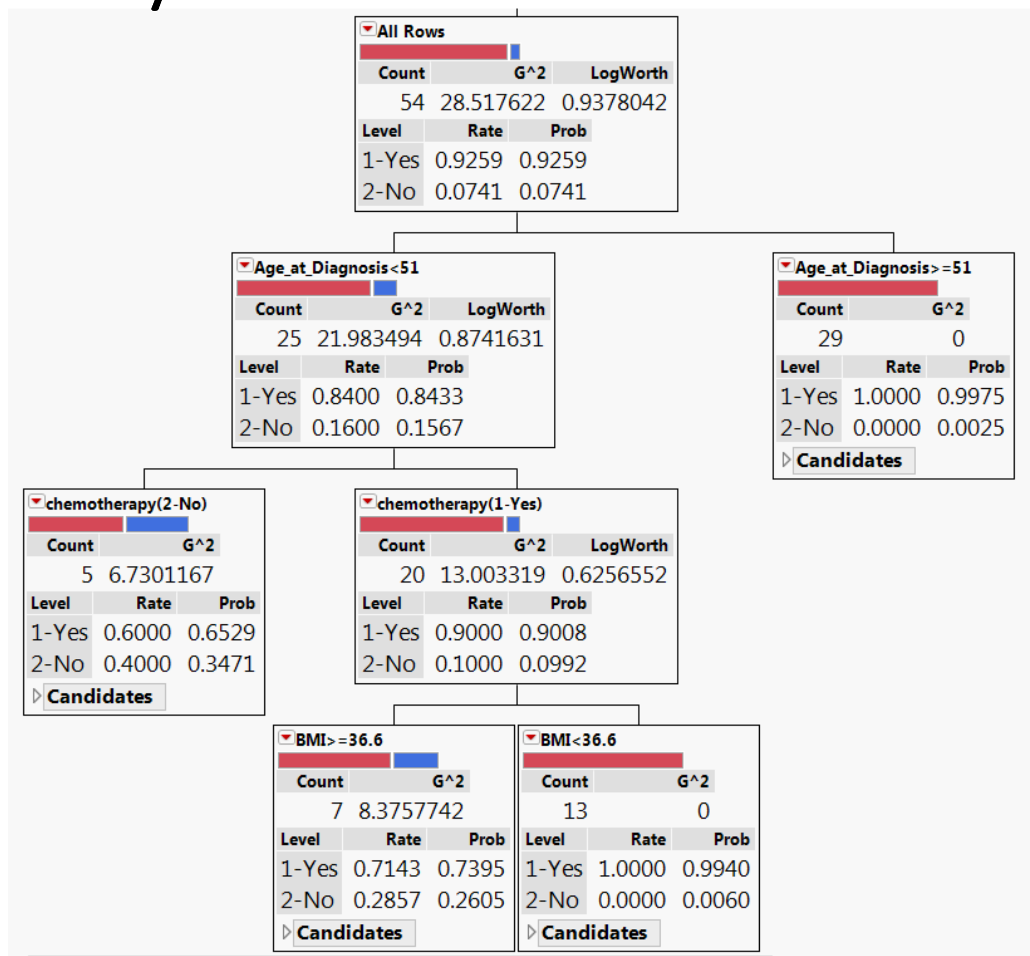
- More important clinically – no therapy except surgery
- Other subtypes have targeted therapies
- Clinically more relevant

Triple Negative Subset: A Cursory Look

```
Race,
Age at Diagnosis, BMI,
Family Hx br cancer,
AJCC Stage,
Chemotherapy
```

Column Contributions

Term	Number of Splits	G^2	Portion
Age_at_Diagnosis	1	6.53412763	0.4872
BMI	1	4.6275447	0.3450
chemotherapy	1	2.25005836	0.1678



Cursory Observations

Younger age, BMI and chemotherapy influence the outcome of breast cancer in Triple Negative disease.

BMI didn't affect outcomes when all types of breast cancer were analyzed, but it did when only Triple Negative disease was analyzed.

All other factors didn’t stand out as major players in TNBC.

Cursory Observations (con’t)

African American females with breast cancer will have higher percentage (almost double national average) of triple negative disease at Georgia Cancer Center.

Probability of being alive after surviving five years with breast cancer is (slightly) higher in Triple Negative breast cancer – a surprising result.

Where From Here? Future Research Goal

African American females with breast cancer will have higher percentage (almost double national average) of triple negative disease at Georgia Cancer Center.

More completely explore the differences in outcomes within the Triple Negative Breast Cancer group based on several parameters, clinical and pathologic, that are known or suspected to shape clinical outcomes.

Primary outcomes were considered to be overall survival, and overall response at time of last follow up.

Lessons Learned

Astoundingly easy to use **Graph Builder** “on the fly” to create multidimensional associations for domain experts.

Data mining is not an easy job in the medical field, it is time consuming and take a lot of steps to process data.

Medical data exploration is enhanced with a collaborative team of domain experts and trained analysts using data exploration skills.