P06-2008

MASTECTOMY VERSUS LUMPECTOMY IN BREAST CANCER TREATMENT

Beatrice Ugiliweneza, University of Louisville, Louisville, KY

ABSTRACT

Objective: To extract information and analyze the cost of mastectomy and lumpectomy as breast cancer treatments using SAS. **Methods**: The data used are from the National Inpatient Sample (NIS). It contains a stratified sample of all hospital patient visits from 37 participating states. First, we extract breast cancer cases among all the data and then focus on those treated by mastectomy and lumpectomy. Then, data analysis techniques are used to examine and compare these two major surgical treatments. We used linear models in SAS/STAT to examine the data, and also PROC GPLOT methods. **Results**: For the data used, the study shows that the cost of mastectomy treatment is lower than the cost of lumpectomy treatment. Moreover, the analysis shows that mastectomy is more used than lumpectomy. **Conclusion**: SAS is a good tool for statistical data analysis, data mining and data visualization. Further study will include claims data to investigate longitudinal patient outcomes.

INTRODUCTION

Surgery is the most used and trusted treatment to control and cure breast cancer. There are two types of surgical treatments for breast cancer: the mastectomy and the lumpectomy. Lumpectomy is surgery that targets the cancer cells. Mastectomy is a more general surgery that removes the whole infected breast. Both of these surgeries are used to treat breast cancer and to try to cure the patient when possible. In this paper, we analyze these two treatments individually and then compare them. We study their use: which one is the most used, the most chosen and we analyze the cost, the total charges of the whole treatment sequence and finally, we look at the length of stay at the hospital for each of those treatments. The data used are the NIS (National Inpatient Survey) for the year 2005. We use SAS as a statistical tool for the analysis and SAS Enterprise Guide 4 as our interface. The data are analyzed by summary statistics, graphs, kernel density estimation and linear models.

METHOD

The data used are from the NIS (National Inpatient Sample) records of 2005. The data contain millions of records on patients from 37 participating states and a stratified sample of hospitals. The identities of the patients are respected in these records; for this reason, it is impossible to perform a longitudinal analysis. However, the data contain detailed diagnosis and procedure codes.

First, the breast cancer cases are extracted from other records. In the NIS data, diagnoses are recorded in the diagnosis columns DX1 thru DX15. The method we use is the ICD9-CM diagnosis coding. ICD9-CM stands for International Classification of Disease, 9th Division- Clinical Modification classification system. We consider the codes 199.0 (Disseminated cancer unspecified site (primary)), 199.1 (Disseminated cancer unspecified site (secondary)) and 174 (Malignant neoplasm of female breast). These codes define cases with general cancer and general breast problems. The next thing to do is to make sure that we consider only breast cancer. We extract two tables from this one: one containing the patient treated by lumpectomy and the other containing the patients treated by mastectomy. In the NIS, the procedures are recorded through the variables, PR1 thru PR15 and the coding method used is the ICD9-CM procedure codes. We consider the codes, 85.41, 85.43, 85.44, 85.45, 85.46, 85.47, 85.48, 85.33, 85.34, 85.36, 85.23, which represent the different types of mastectomies and the code, 85.21, which is the code for lumpectomy. We are now sure that these results reflect the breast cancer cases and we proceed to the analysis of each one of them. For each table of data (Mastectomy data and Lumpectomy data), we produce:

- Summary statistics giving the number of patients , the average age of patients , average length of stay, the average number of diagnoses per patient, average number of procedures per patient and the average total charges.
- A chart representing the frequency of the total charges to see how many people are charged how much money
- A chart representing the frequency of the length of stay to have an idea of how many people stay at the hospital during treatment and for how long.
- A scatter plot of the total charges versus the length of stay to analyze whether the length of stay affects in any way the total charges.
- Finally, we use the built-in Regression linear models analysis in Enterprise Guide 4 to study deeply the relation between total charges and length of stay.

After analyzing each one of the Mastectomy and Lumpectomy procedures, we compare them. In order to perform the

comparison, we first apply kernel density estimation to each one of the tables on the variables, total charges (totchg) and length of stay (los) individually. Then, we plot the compared graphs for mastectomy versus lumpectomy for total charges and length of stay.

RESULTS

To extract cancer cases and breast problem cases among others, we use the following code:

```
libname NIS2005 'E:\NIS2005';
data core1;
set NIS2005.NIS_2005_CORE;
if (dx1='1740' or dx2='1740' or … dx15='1740' or
dx1='1741' or… or dx15='1741' or dx1='1742' or … or dx15='1742' or
dx1='1743' or … or dx15='1743' or dx1='1744' or … or dx15='1744' or dx1='1745' or …
or dx15='1745' or dx1='1746' or … or dx15='1746' or
dx1='1747' or … or dx15='1747' or dx1='1748' or … or dx15='1748' or
dx1='1749' or … or dx15='1749' or dx1='1990' or … or dx15='1990' or
dx1='1991' or … or dx15='1991')
then code=1;
else code=0;
data NIS2005.BREAST_CANCER;
set core1;
where code=1;
run;
```

As a result, we obtain a table of 55,950 patients. These cases contain breast cancer cases as well as other unspecified cancers and other breast diseases. To extract cases treated by lumpectomy and mastectomy, we use the codes below:

```
data LUMP1;
set NIS2005.BREAST_CANCER;
if (pr1='8521' or … or pr15='8521' )
then code2=2;
else code2=0;
data NIS2005.LUMPECTOMY;
set LUMP1;
where code2=2;
run;

data MAST1;
set NIS2005.BREAST_CANCER;
if (pr1='8541' or … or pr15='8541' or
pr1='8543' or … or pr15='8543' or pr1='8544' or … or pr15='8544' or
pr1='8534' or … or pr15='8534' or
pr1='8533' or … or pr15='8533' or pr1='8536' or … or pr15='8536' or
pr1='8547' or … or pr15='8547' or pr1='8548' or … or pr15='8548' or
pr1='8545'
or … or pr15='8545' or pr1='8546' or … or pr15='8546' or
pr1='8523' or … or pr15='8523')
then code1=1;
else code1=0;
data NIS2005.MASTECTOMY;
set MAST1;
where code1=1;
run;
```

These codes produce a table of 1,413 patients treated by lumpectomy and a table of 12,139 patients treated by mastectomy.

ANALYSIS OF MASTECTOMY AND LUMPECTOMY INDIVIDUALLY

MASTECTOMY

Table1: Summary statistics of Mastectomy

Number of patients	12115
Average age of patients	62

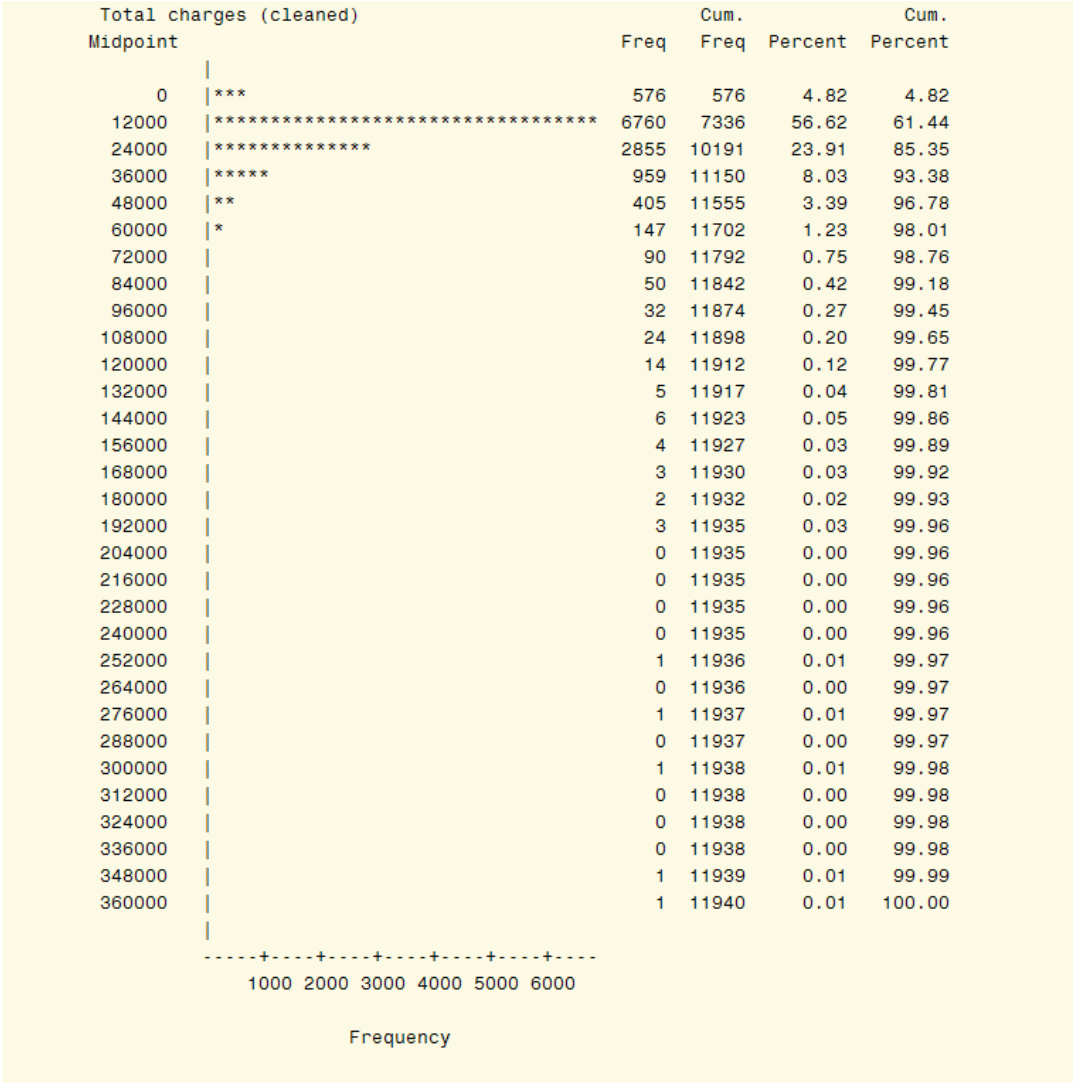
Average length of stay per patient (in days)	2
Average number of diagnoses per patient	4
Average number of procedures per patient	2
Average total charges	19475

This table was summarized from the result of the following code:

```
libname NIS2005 'E:\NIS2005';
data mast1;
set NIS2005.mastectomy;
keep age los ndx npr totchg;
run;

proc univariate data=mast1;
var age los ndx npr totchg;
output out=NIS2005.SummaryStat_mastectomy
    N=mastectomies
    mean=meanage meanlos meanndx meannpr meantotchg;
run;
```

The summary statistics table gives the averages of the considered variables. We are interested in the variables, total charges and length of stay. To view more clearly how they are in the mastectomy data, we produce charts.

Chart1: Chart representing frequencies of the total charges in Mastectomy



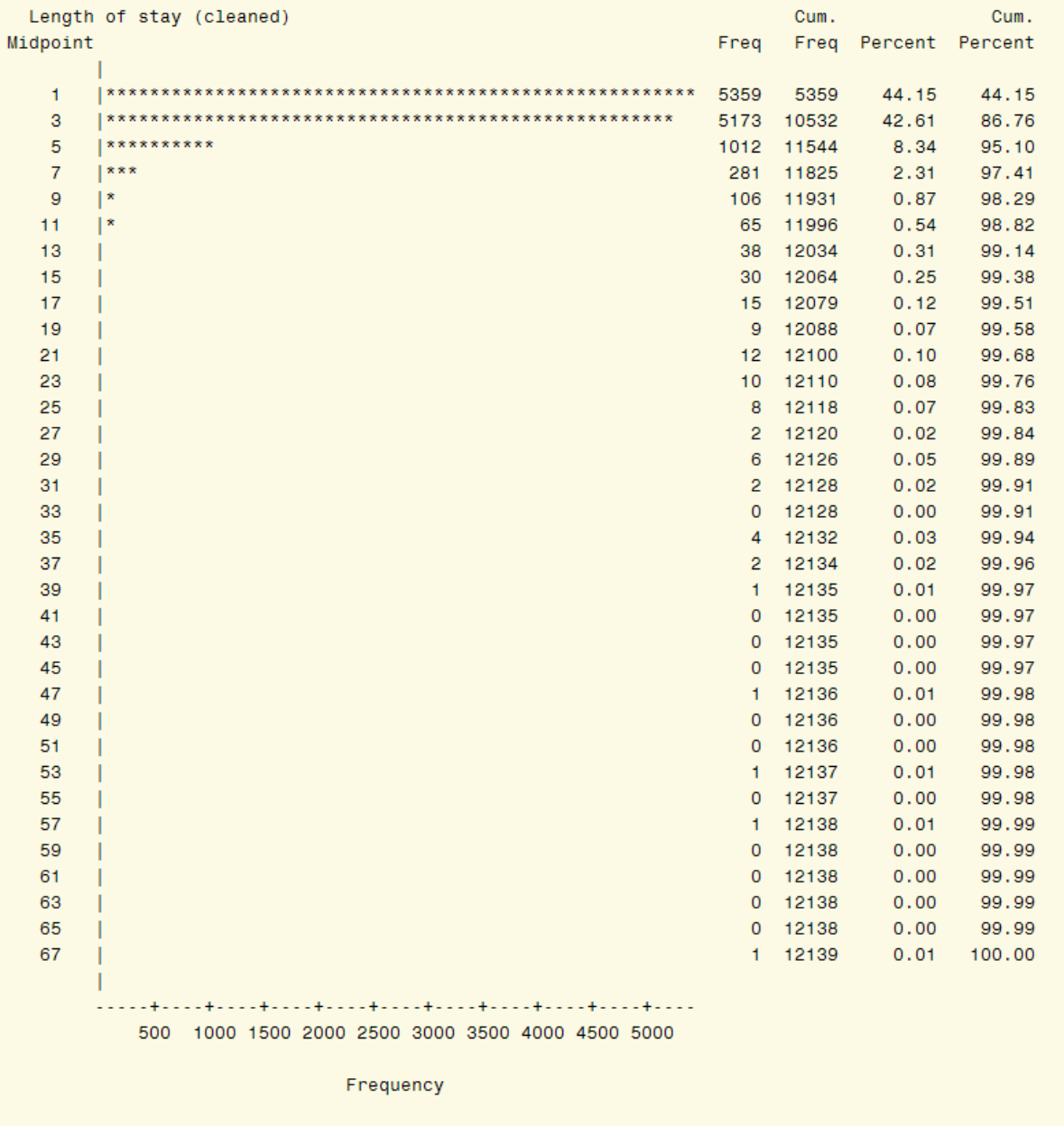
Total charges (cleaned) Midpoint		Freq	Cum. Freq	Percent	Cum. Percent
0	***	576	576	4.82	4.82
12000	**********	6760	7336	56.62	61.44
24000	**********	2855	10191	23.91	85.35
36000	*****	959	11150	8.03	93.38
48000	**	405	11555	3.39	96.78
60000	*	147	11702	1.23	98.01
72000		90	11792	0.75	98.76
84000		50	11842	0.42	99.18
96000		32	11874	0.27	99.45
108000		24	11898	0.20	99.65
120000		14	11912	0.12	99.77
132000		5	11917	0.04	99.81
144000		6	11923	0.05	99.86
156000		4	11927	0.03	99.89
168000		3	11930	0.03	99.92
180000		2	11932	0.02	99.93
192000		3	11935	0.03	99.96
204000		0	11935	0.00	99.96
216000		0	11935	0.00	99.96
228000		0	11935	0.00	99.96
240000		0	11935	0.00	99.96
252000		1	11936	0.01	99.97
264000		0	11936	0.00	99.97
276000		1	11937	0.01	99.97
288000		0	11937	0.00	99.97
300000		1	11938	0.01	99.98
312000		0	11938	0.00	99.98
324000		0	11938	0.00	99.98
336000		0	11938	0.00	99.98
348000		1	11939	0.01	99.99
360000		1	11940	0.01	100.00

Code to obtain the chart above:

```
data mast2;
set NIS2005.mastectomy;
keep totchg;
run;
proc chart data=mast2;
title1'Chart representing the frequency of'
title2 ' the total charges in mastectomy';
hbar totchg;
run;
```

From the table of summary statistics for Mastectomy, the average cost is $19,475, but looking at the chart, we see that the charges for most people (56.62%) are $12,000. Moreover, about 80.53% are charged between $12,000 and $24,000. From this, we conclude that the charges for Mastectomy are about $18,000.

Chart2: Chart with frequencies for the Length Of Stay in Mastectomy



Code to obtain the chart above:

```
data mast2;
set NIS2005.mastectomy;
keep los;
run;
proc chart data=mast2;
title'Chart representing the frequency';
title2 'of the length of stay in mastectomy';
hbar los;
run;
```

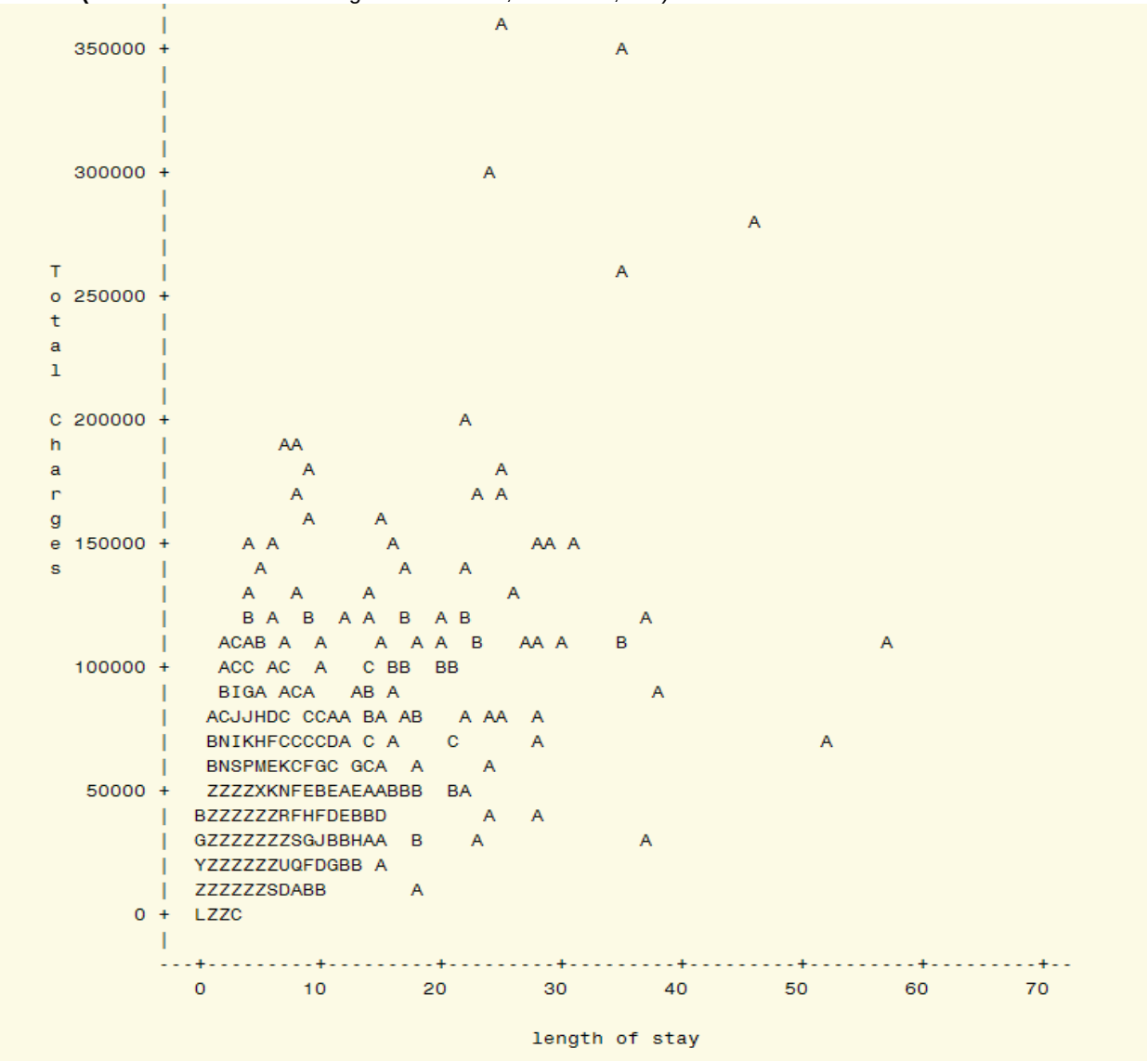
For the length of stay, the summary statistics table and these charts say the same thing. The average number of nights spent at the hospital during treatment is 2 to 3, and is at least 1.

Now, we look at the scatter plot of total charges versus length of stay. This scatter plot will tell us whether or not the

two variables are related in mastectomy.

Graph1: Scatter plot of total charges versus length of stay in Mastectomy

(Plot of TOTCHG*LOS. Legend: A = 1 obs, B = 2 obs, etc.)



The scatter plot was obtained with the use of the following code:

```
proc plot data=mast2;
title1'Scatter plot of the total charges';
title2 'vs the length of stay in mastectomies';
plot totchg*los;
label totchg='Total Charges' los='length of stay';
run;

proc plot data=lump2;
title1'Scatter plot of the total charges';
title2 'vs the length of stay in lumpectomies';
plot totchg*los;
label totchg='Total Charges' los='length of stay';
run;
```

From the scatter plot, we can conclude that in the mastectomy cases, the length of stay affects the total charges. The more the patient stays at the hospital, the higher the charges.

We next analyze the total charges as a function of the length of stay. According to the scatter plot, there is a linear relation between total charges and length of stay. We use the built-in linear regression functions in SAS Enterprise Guide 4 (EG4) to analyze this relationship.

LINEAR MODELS APPLIED ON THE MASTECTOMY DATA

The following results are obtained

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.190639E12	1.190639E12	6792.50	<.0001
Error	11938	2.092579E12	175287214		
Corrected Total	11939	3.283218E12			

Root MSE	13240	R-Square	0.3626
Dependent Mean	19473	Adj R-Sq	0.3626
Coeff Var	67.99029		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	10415	163.58024	63.67	<.0001
LOS	Length of stay (cleaned)	1	3980.52145	48.29754	82.42	<.0001

From the first table (ANOVA table), we have the p.value<.0001, which means that Length of stay affects total charges. The second table confirms this fact: the large value of the R-square means that the model found is good. The third table (Parameter estimates table) gives the estimated equation of the model relation, which is:

Total charges= 3980.52145*Length of stay + 10,415.

The positive slope, 3980.52145, of the linear regression line suggests that the longer the patient stays at the hospital, the higher the total charges (same conclusion as from the scatter plot). In addition, one additional night at the hospital during mastectomy treatment should increase the charges by almost $4,000.

LUMPECTOMY

Tabe2: Summary statistics of Lumpectomy

Number of patients	1412
Average age of patients	64
Average length of stay per patient (in days)	3
Average number of diagnoses per patient	5

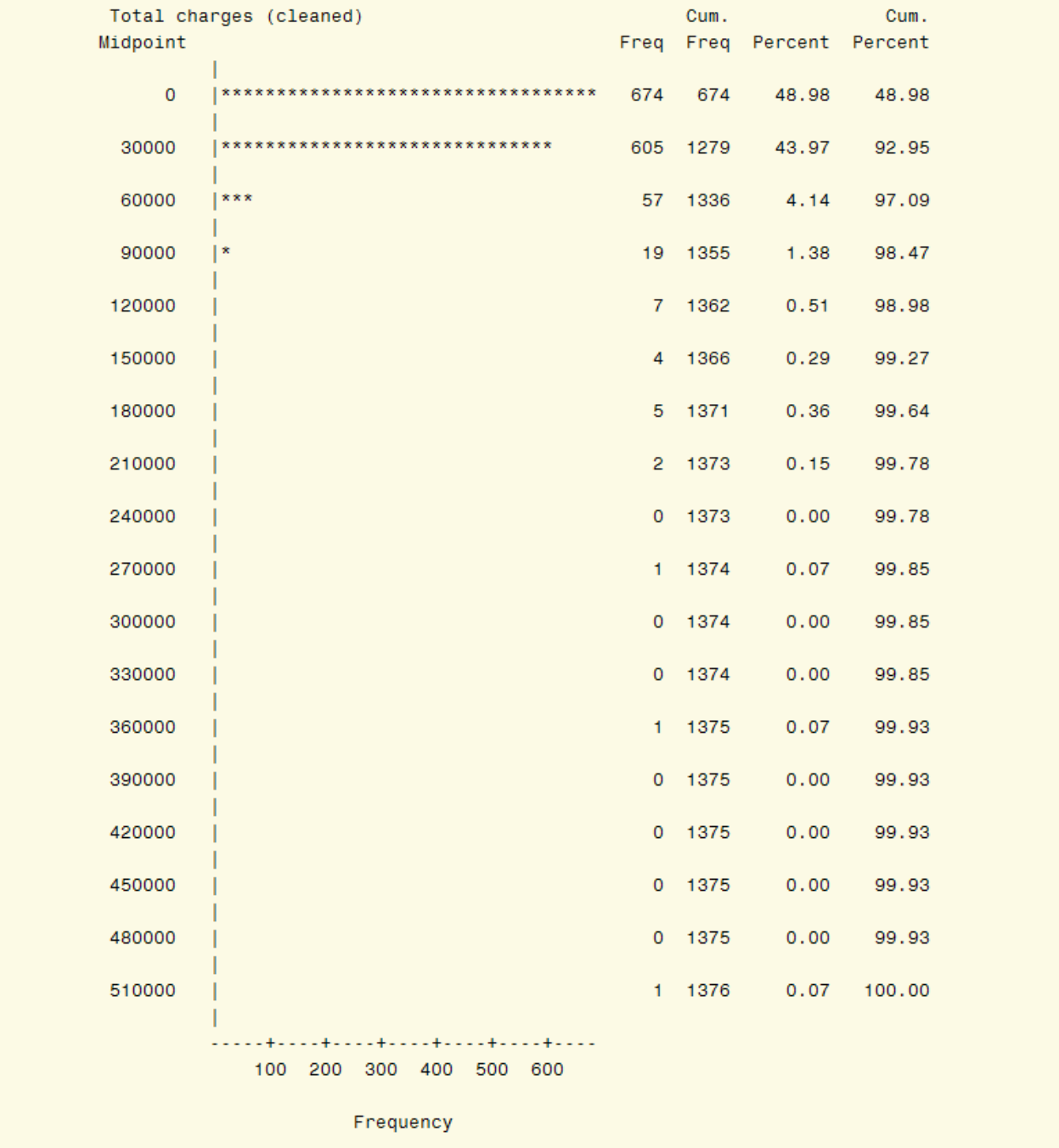
Average number of procedures per patient	3
Average total charges	21584.77

This table was summarized from the result of the following code:

```
libname NIS2005 'E:\NIS2005';
data lump1;
set NIS2005.lumpectomy;
keep age los ndx npr totchg;
run;

proc univariate data=lump1;
var age los ndx npr totchg;
output out=NIS2005.SummaryStat_lumpectomy
   N=lumpectomies
   mean=meanage meanlos meanndx meannpr meantotchg;
run;
```

This summary statistics table gives the averages of the considered variables. Since we are interested in the variables, total charges and length of stay, we produce charts to view more clearly how they are in the lumpectomy data.

Chart3: Chart representing the frequencies of the total charges in Lumpectomy



Total charges (cleaned)

Midpoint		Freq	Cum. Freq	Percent	Cum. Percent
0	**********	674	674	48.98	48.98
30000	**********	605	1279	43.97	92.95
60000	***	57	1336	4.14	97.09
90000	*	19	1355	1.38	98.47
120000		7	1362	0.51	98.98
150000		4	1366	0.29	99.27
180000		5	1371	0.36	99.64
210000		2	1373	0.15	99.78
240000		0	1373	0.00	99.78
270000		1	1374	0.07	99.85
300000		0	1374	0.00	99.85
330000		0	1374	0.00	99.85
360000		1	1375	0.07	99.93
390000		0	1375	0.00	99.93
420000		0	1375	0.00	99.93
450000		0	1375	0.00	99.93
480000		0	1375	0.00	99.93
510000		1	1376	0.07	100.00

100 200 300 400 500 600

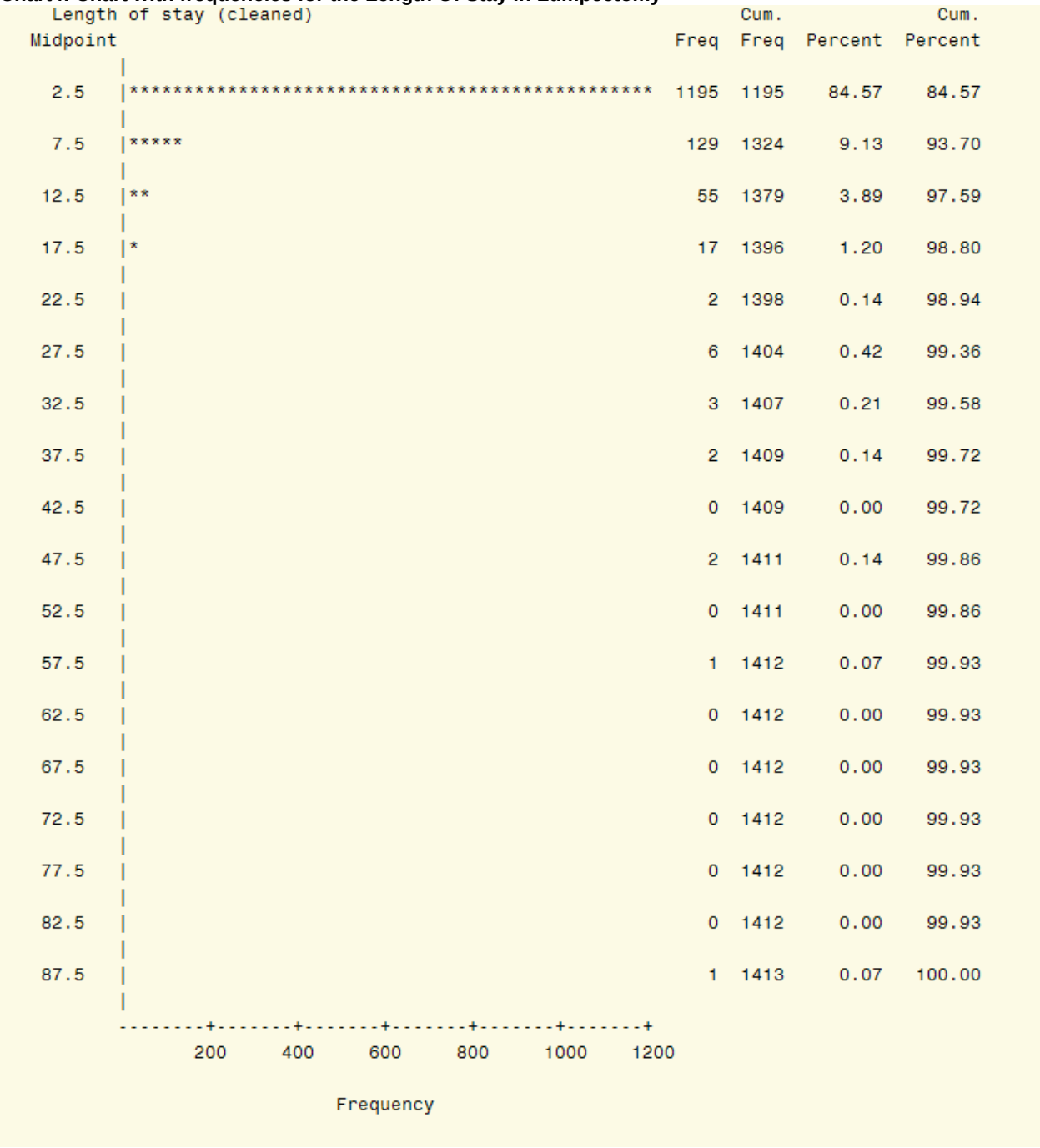
Frequency

Code to obtain the chart above:

```
data lump2;
set NIS2005.lumpectomy;
keep totchg;
run;
proc chart data=lump2;
title1'Chart representing the frequency of'
title2 ' the total charges in lumpectomies';
hbar totchg;
run;
```

The table of summary statistics for Lumpectomy gives an average amount of $21,584.77 for the total charges and the chart shows that about 43.97% of the patients are charged $30,000. If we choose not to take into consideration those with an average cost of $0.0, we can conclude that the charges for Lumpectomy are about $30,000.

Chart4: Chart with frequencies for the Length Of Stay in Lumpectomy



Code to obtain the chart above:

```
data lump2;
set NIS2005.lumpectomy;
keep los;
run;
```

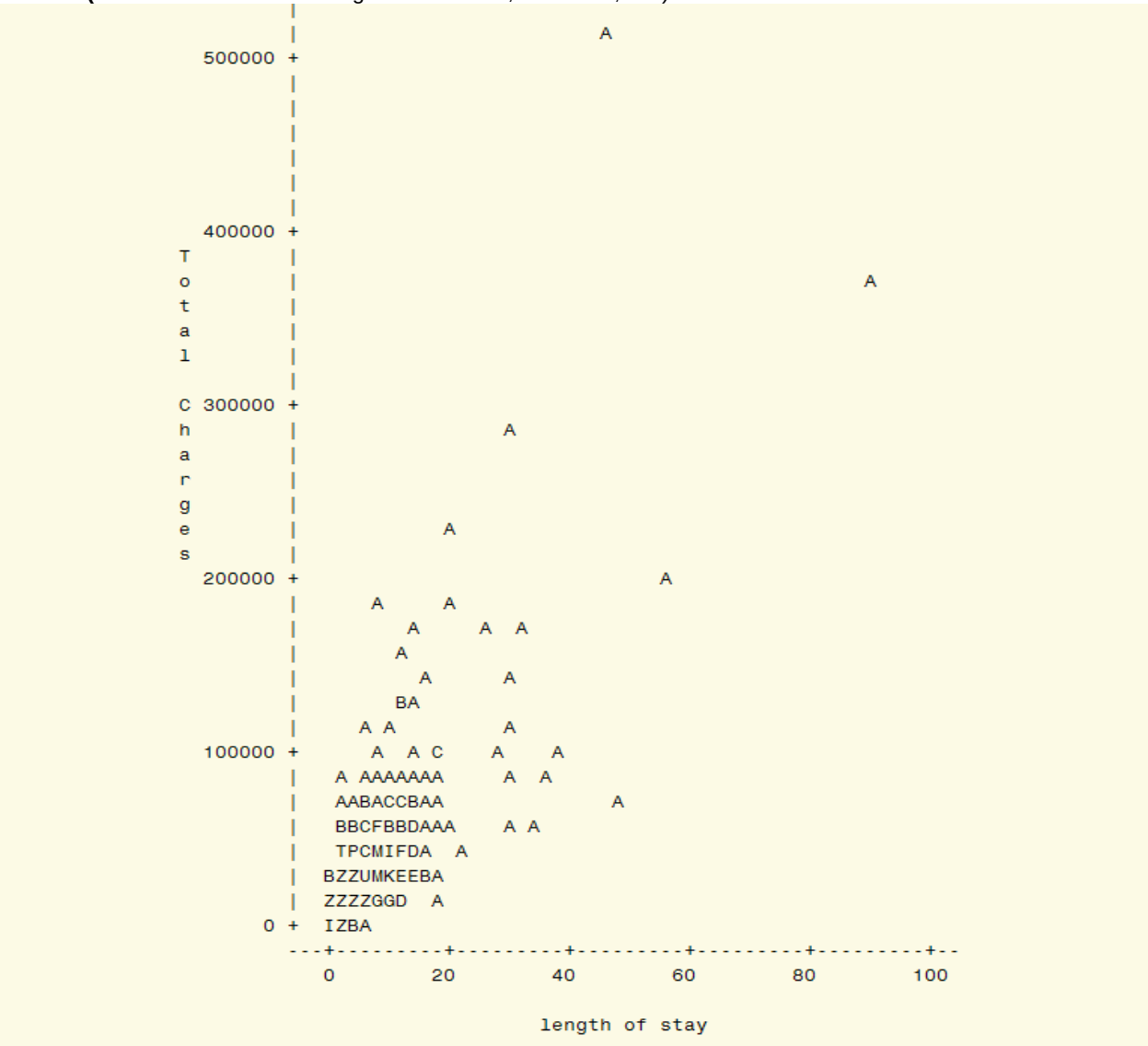
```
proc chart data=mast2;
title'Chart representing the frequency of';
title2 'the length of stay in lumpectomies';
hbar los;
run;
```

For the length of stay, the summary statistics table and these charts say the same thing. The average number of nights spent at the hospital during treatment is 2 to 3, and is at least 1.

Next, we look at the scatter plot of total charges versus length of stay. It will show us whether or not the two variables are related in lumpectomy and what kind of relation they have if there is any.

Graph2: Scatter plot of total charges vs length of stay in Lumpectomy

(Plot of TOTCHG*LOS. Legend: A = 1 obs, B = 2 obs, etc.)



The scatter plot was obtained with the use of the following code:

```
proc plot data=lump2;
title1'Scatter plot of the total charges';
title2 'vs the length of stay in lumpectomies';
```

```
plot totchg*los;
label totchg='Total Charges' los='length of stay';
run;
```

From the scatter plot, we can conclude that in the lumpectomy treatment, the length of stay affects the total charges. The more time the patient spends at the hospital, the more the charges.

This inspires us to analyze the total charges as a function of the length of stay. Looking at the scatter plots, the relation between total charges and length of stay is linear. Just like in the mastectomy case, we use the built-in linear regression functions in SAS Enterprise Guide 4 (EG4) to study this relationship.

LINEAR MODELS APPLIED ON THE LUMPECTOMY DATA

From EG, we get the following results:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5.802538E11	5.802538E11	1758.49	<.0001
Error	1374	4.53383E11	329973099		
Corrected Total	1375	1.033637E12			

Root MSE	18165	R-Square	0.5614
Dependent Mean	21585	Adj R-Sq	0.5611
Coeff Var	84.15729		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	9814.30932	564.43938	17.39	<.0001
LOS	Length of stay (cleaned)	1	3982.33597	94.96601	41.93	<.0001

Like in the mastectomy case, for the first table (ANOVA table), we get the p.value<.0001, which means that, here also, Length of stay affects total charges. The R-square value from the second table is a little higher than the one found in the Mastectomy data, so it also confirms the conclusion drawn from the ANOVA table. The third table (Parameter estimates table) gives the estimated equation of the model relation:

Total charges= 3982.33597*Length of stay + 9814.30932.

The positive slope, 3982.33597, of the linear regression line suggests that the longer the patient stays at the hospital, the higher the total charges, and one additional night at the hospital during lumpectomy treatment will increase the charges by almost $4,000 (same amount as in the mastectomy case).

COMPARISON OF MASTECTOMY AND LUMPECTOMY DATA

We first look at the compared summary statistics.

Table3: Summary statistics comparing the mastectomy and the lumpectomy

	Mastectomy	Lumpectomy
Number of patients	12115	1412
Average age of patients	62	64
Average length of stay per treatment	2	3
Average amount of total charges	19475	21584.77

The mastectomy is the most used compared to lumpectomy. The length of stay for both surgeries is almost the same. There is a major difference in the total charges for these two surgeries: the lumpectomy is the most expensive. This last observation given by the means of the two total charges is also confirmed by the frequencies (see chart1 and chart2); while most people are charged $18,000 for mastectomy, many people are charged $30,000 for lumpectomy.

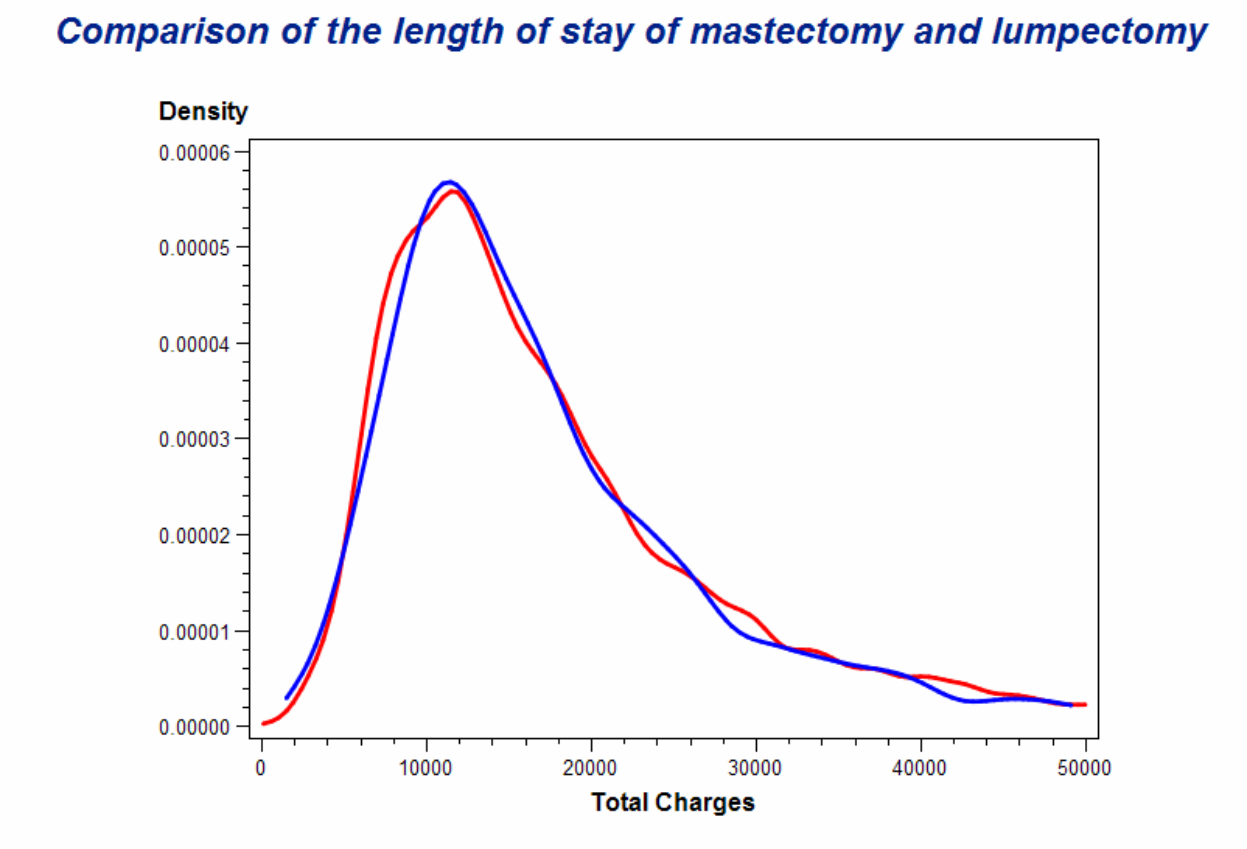
We then look at the densities of the total charges in both the mastectomy and the lumpectomy. For this, we use kernel density estimation. Kernel density estimation is a way to estimate the density probability function. We use PROC KDE to compute the estimated probabilities and PROC GPLOT to plot the functions. The code used is the following:

```
/*compute densities in mastectomy and in lumpectomy*/
libname NIS2005 'E:\NIS2005';
data mast1_totchg_density;
set NIS2005.mastectomy;
keep totchg;
run;
proc kde data=mast1_totchg_density out=mast2_totchg_density;
var totchg;
run;

data lump1_totchg_density;
set NIS2005.lumpectomy;
keep totchg;
run;
proc kde data=lump1_totchg_density out=lump2_totchg_density;
var totchg;
run;

/*Combine tables*/
data NIS2005.mast3_totchg_density (rename=(totchg=mast_totchg
density=mast_density count=mast_count));
set mast2_totchg_density;
run;

data NIS2005.lump3_totchg_density (rename=(totchg=lump_totchg
density=lump_density count=lump_count));
set lump2_totchg_density;
run;

Data NIS2005.mast_lump_totchg_density;
merge NIS2005.mast3_totchg_density NIS2005.lump3_totchg_density ;
run;

/*Graph the two densities in the same coordinate system* (red=mastectomy,
blue=lumpectomy*/
title'Comparison of the total charges of mastectomy and lumpectomy';
proc gplot data=NIS2005.Mast_lump_totchg_density;
```

```
symbol1 color=red i=spline w=3 v=none;
symbol2 color=blue i=spline w=3 v=none;
plot mast_density*mast_totchg lump_density*lump_totchg/overlay
                        haxis=0 to 50000 by 10000;
label mast_totchg='Total Charges'
      lump_totchg='Total Charges'
      mast_density='Density'
      lump_density='Density';
run;
```

This code produces the compared densities for total charges above.

Graph3: Comparison of the total charges densities in Mastectomy and Lumpectomy



Blue:Lumpectomy Red: Mastectomy

From this graph, we show that the estimates of charges for mastectomy and lumpectomy are the same. The probability of paying a certain amount of money is the same when treated by mastectomy or lumpectomy

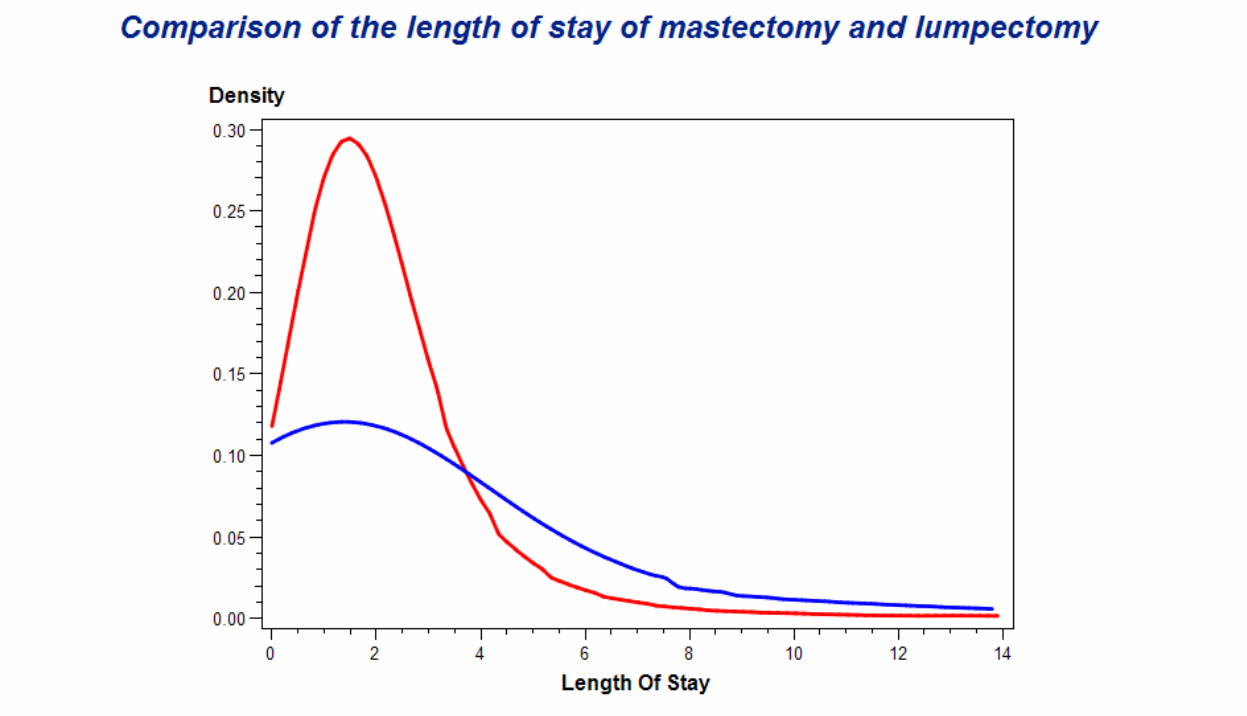
Finally, we look at the densities of the length of stay in the mastectomy and the lumpectomy. Again, we use the kernel density estimation. The code below is used:

```
/*compute densities in mastectomy and in lumpectomy*/
   data mast1_los_density;
   set NIS2005.mastectomy;
   keep los;
   run;
   proc kde data=mast1_los_density out=mast2_los_density;
   var los;
   run;
```

```
data lump1_los_density;
set NIS2005.lumpectomy;
keep los;
run;
proc kde data=lump1_los_density out=lump2_los_density;
var los;
run;

/*Combine tables*/
data NIS2005.mast3_los_density (rename=(los=mast_los density=mast_density
count=mast_count));
set mast2_los_density;
run;

data NIS2005.lump3_los_density (rename=(los=lump_los density=lump_density
count=lump_count));
set lump2_los_density;
run;

Data NIS2005.mast_lump_los_density;
merge NIS2005.mast3_los_density  NIS2005.lump3_los_density;
run;

/*Graph the two densities in the same coordinate system* (red=mastectomy,
blue=lumpectomy*/
title'Comparison of the length of stay of mastectomy and lumpectomy';
proc gplot data=NIS2005.Mast_lump_los_density;
symbol1 color=red i=spline w=3 v=none;
symbol2 color=blue i=spline w=3 v=none;
plot mast_density*mast_los lump_density*lump_los/overlay
                                                    haxis=0 to    14 by 2;
label mast_los='Length Of Stay'
     lump_los='Length Of Stay'
     mast_density='Density'
     lump_density='Density';
run;
```

As a result, we have the graph below:

Graph4: Comparison of the length of stay densities in Mastectomy and Lumpectomy



Blue:Lumpectomy Red: Mastectomy

This graph tells us that the patient is more likely to spend nights at the hospital when treated by mastectomy than when treated by lumpectomy.

CONCLUSION

Lumpectomy and mastectomy are important surgical treatments used for breast cancer. It is important for a patient to know the similarities and the differences, especially in total charges and length of stay at the hospital. Our study shows that lumpectomy is more expensive than mastectomy and the difference can be more than $10,000. The length of stay is almost the same, on average, 2 to 3 days, but one is more likely to spend nights at the treatment center when treated by mastectomy then when treated by lumpectomy. From our study, we can conclude that SAS is a powerful tool for data analysis, data visualization and data description.

ACKNOWLEDGMENTS

My acknowledgments go to Dr Patricia Cerrito who helped me and continue to help and support me.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Beatrice Ugiliweneza
University of Louisville
Department of Mathematics
328 Natural Science Building
Louisville, KY 40292
Work Phone: 502-852-6826
Fax: 502-852-7132
E-mail:b0ugil01@louisville.edu

Other brand and product names are trademarks of their respective companies.