Breast Cancer Data Insights: A Rigorous Analysis of SEER Records (1992-2020)

By:

Keerthi Dwaraka Kovelamudi

Teja Katikam

Tiffany Neza

Mayuresh Abhay Shastri

Agenda

Introduction

Background

Dataset Background

Research question

Variables

Data Analysis

Conclusion

References

Introduction

Breast Cancer is the second leading cause of death among women.

Breast Cancer is the one of the most common diagnosable cancers in women in the United States.

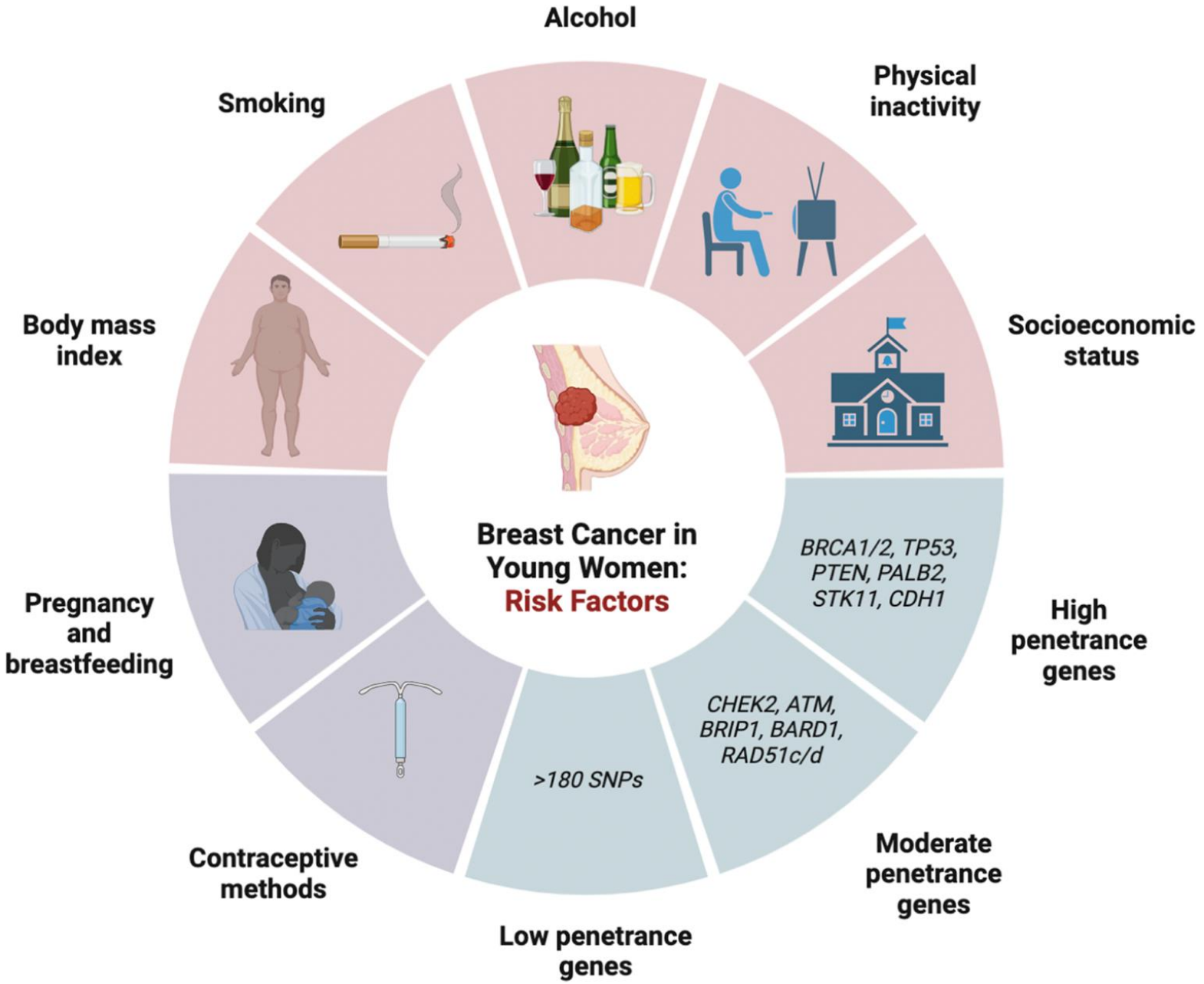
4,100,000 Women who are living with or have a history of breast cancer.

287,850 Invasive breast cancer cases are estimated to be diagnosed this year.

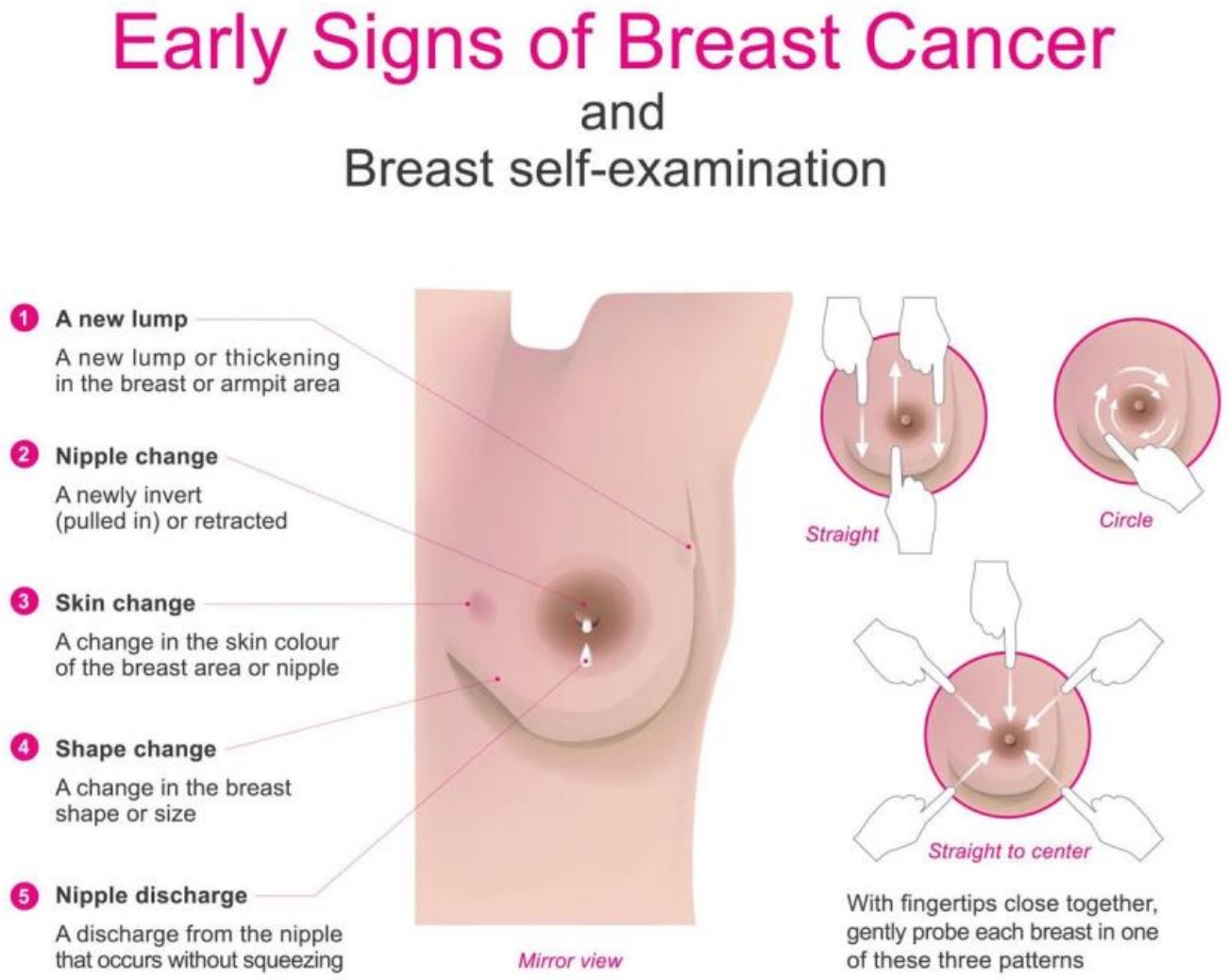
43,250 Women are estimated to die from breast cancer this year.

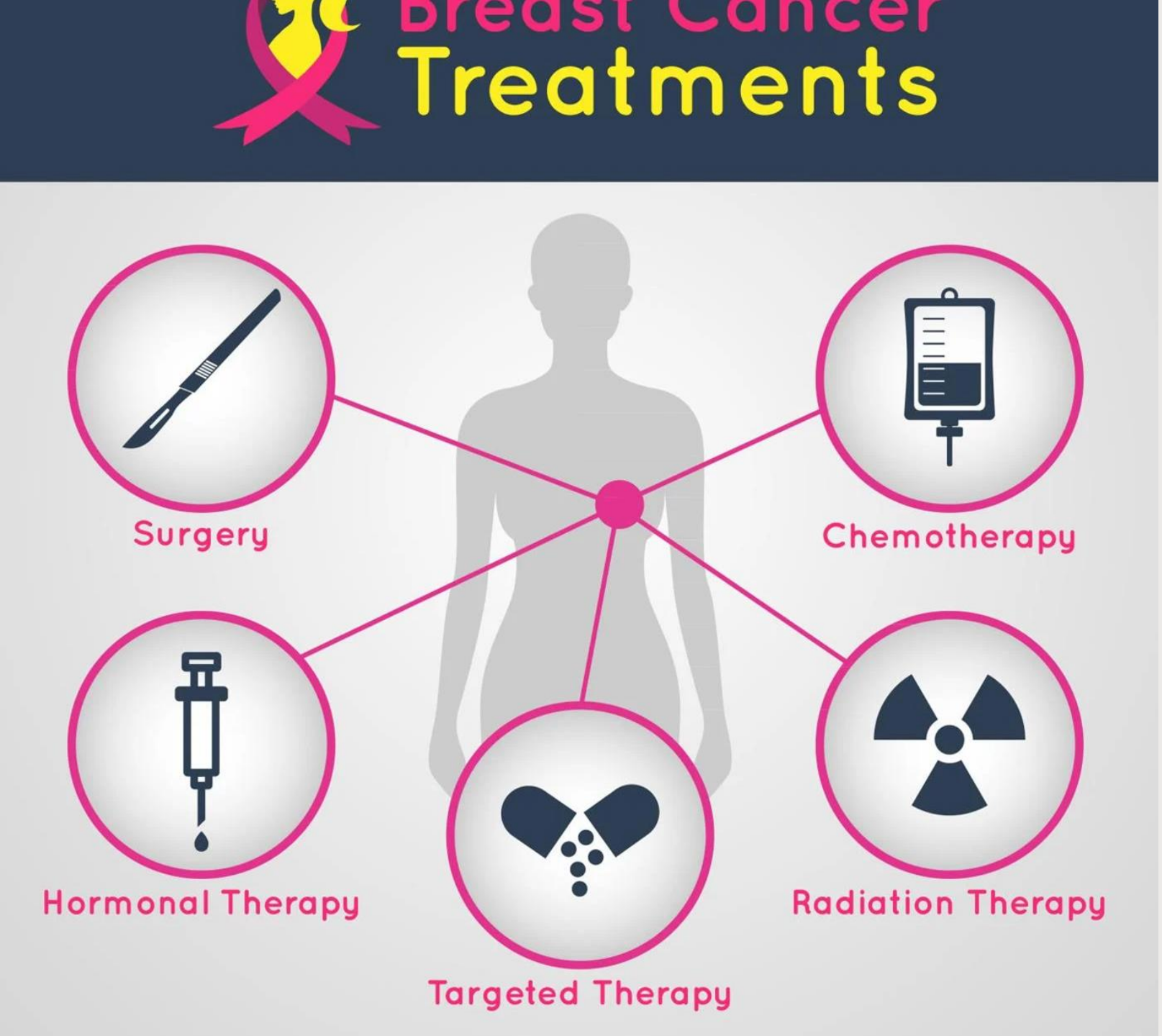
Background

Risk Factors of Breast Cancer



Key Symptoms of Breast Cancer





Treatment Options

Prevention


PREVENTION BREAST CANCER
BREAST-FEED
EXERCICISE REGULARLY
DON'T SMOKE
KEEP YOUR WEIGHT DOWN
LIMIT USING HORMONE MEDICATION
EAT HEALTHY FOOD
LIMIT EXPOSURE TO POLLUTION
WATCH OUT FOR CHANGES TO YOUR BREASTS
LIMIT YOUR ALCOHOL CONSUMPTION
GO FOR A MAMMOGRAM
you can exercise breast cancer prevention by doing these

Dataset Background

- Data was obtained from the National Cancer Institute of Surveillance, Epidemiology, and End Results Program (SEER).

- Our dataset from the SEER Registry, includes individuals who were diagnosed with breast cancer between 1992 and 2020.

- The dataset contains information on age, sex, main cancer site, year of diagnosis, ICD codes, months from diagnosis to therapy, race, cause of death, course of treatment, and other facts.

- SEER now collects and disseminates data on cancer incidence and survival from population-based cancer registries, which cover approximately 48.0% of the US population.

- SEER receives death data from the National Center for Health Statistics. The demographic data required to calculate cancer rates are frequently provided by the Census Bureau. The registration is updated annually.

Focus of the study:

Survival length

Overall average length of survival of breast cancer and survival rates.

Research Question:

What is the impact of delay in treatment on survival months?

Variables

VARIABLES	TYPE	VARIABLE NAME
GENDER OF PATIENT	NOMINAL	SEX
RACE OF PATIENT	NOMINAL	RACE_RECODE__WHITE__BLACK__OTHE
AGE RANGE OF PATIENT	CATEGORICAL	AGE_RECODE_WITH__1_YEAR_OLDS
MONTHS OF SURVIVAL	CONTINUOUS	SURVIVAL_MONTHS
YEAR OF DIAGNOSIS	CONTINUOUS	YEAR_OF_DIAGNOSIS
DEATH ASSOCIATED WITH CANCER	BINARY	SEER_CAUSE_SPECIFIC_DEATH_CLASS
MONTHS BETWEEN DIAGNOSIS AND TREATMENT	CONTINUOUS	MONTHS_FROM_DIAGNOSIS_TO_TREATMENT

Handling Data

Missing Data

- The term "unknown" is used to remove missing data as “Blanks” from records.

Survival Months

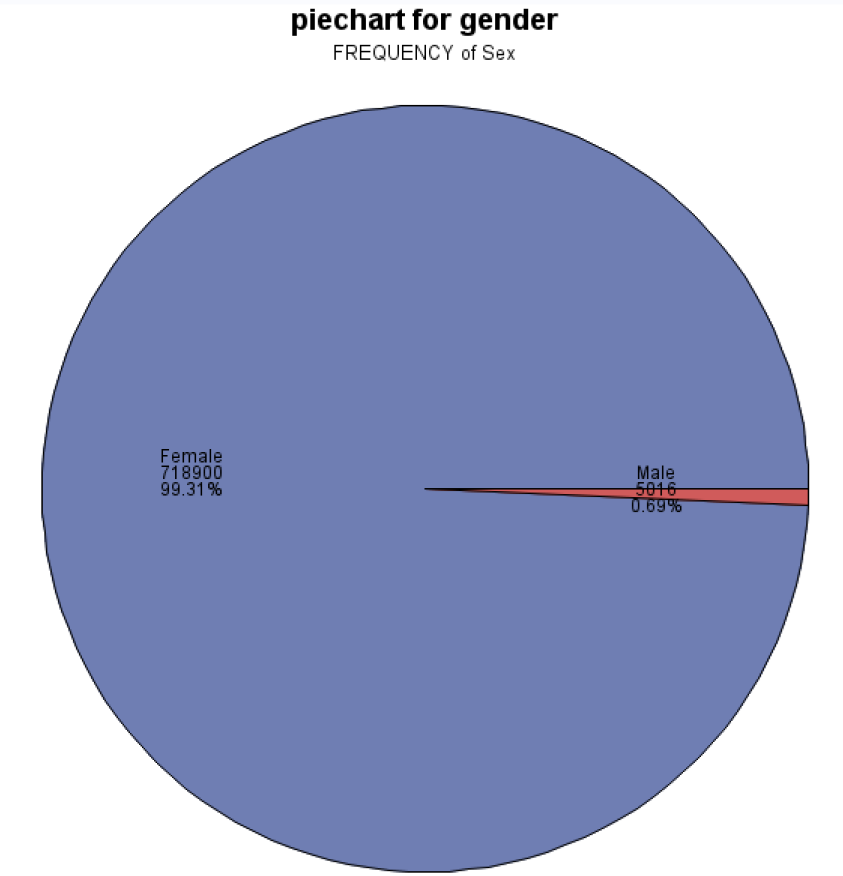
- survival months are recoded into four categories as survival code.
  - <5 years
  - 5-<10 years
  - 10-<15 years
  - 15-20 years

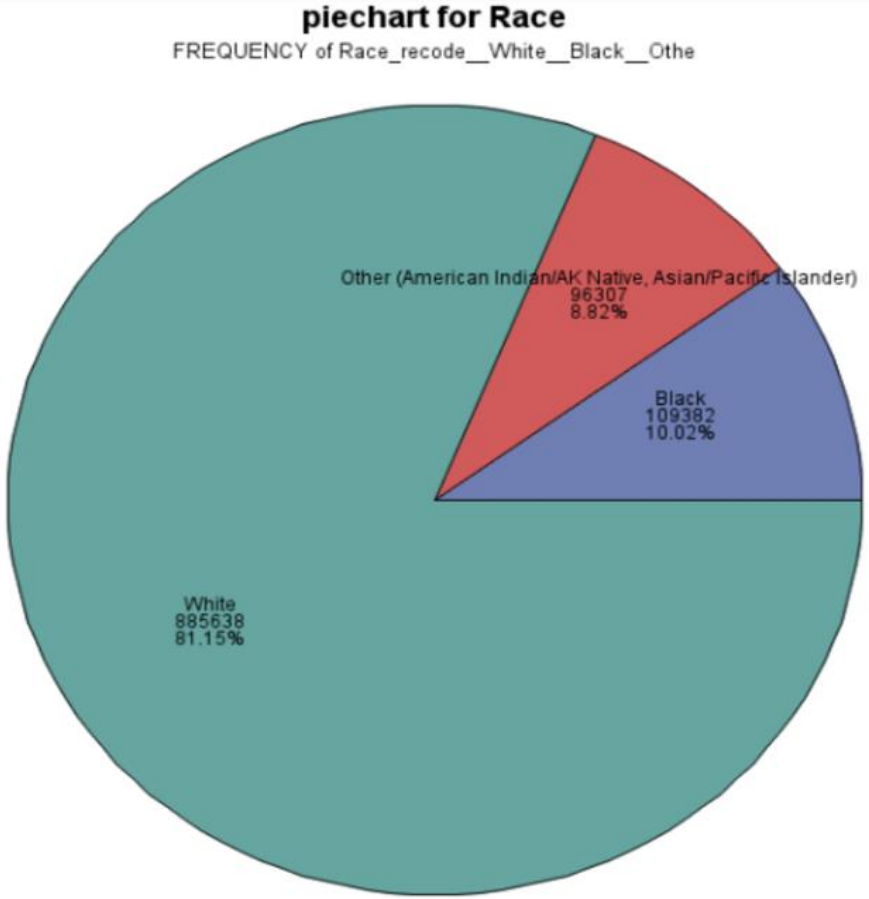
Data Analysis

Demographics

SEX:

- FEMALE : 99.31%
- MALE: 0.69%

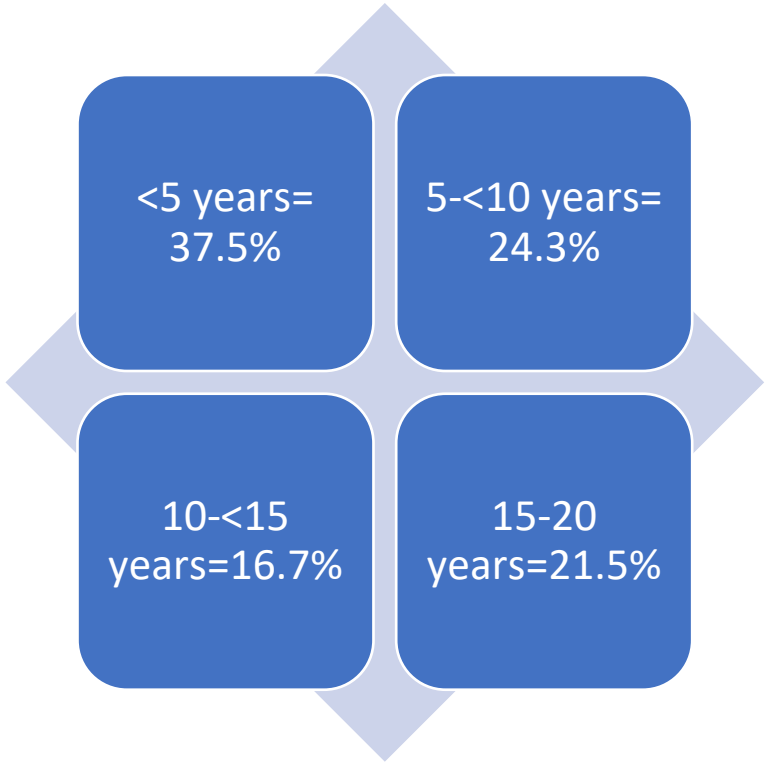


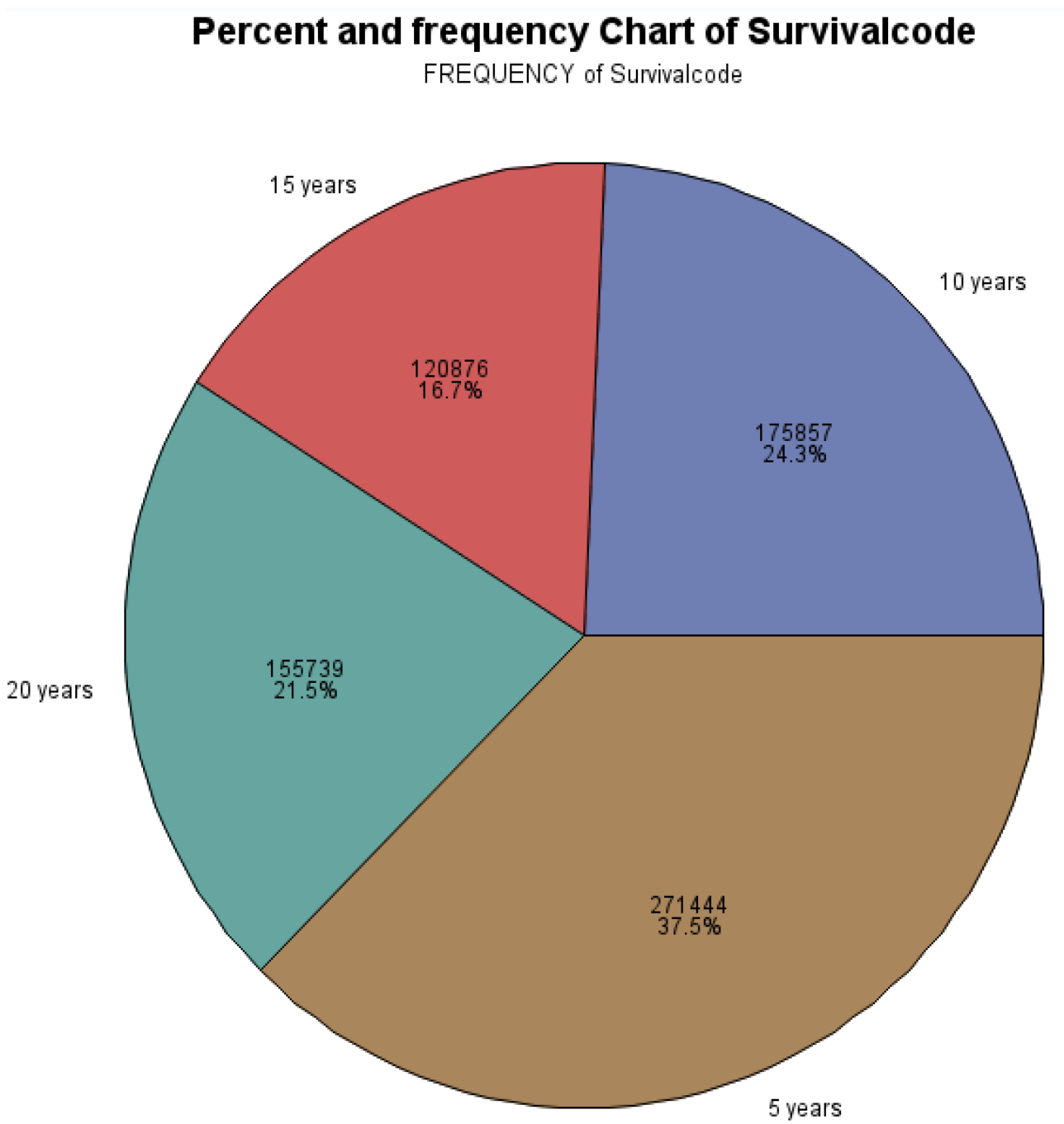


RACE

- WHITE= 81.15%
- BLACK= 10.02%
- OTHER= 8.82%

Survival Months





Descriptive Statistics and Tests of Normality

Mean Survival_months was 86.57 with SD(64.19)

- Median = 108.0
- Skewness = 0.76
- Kurtosis = -0.33
- IQR (Interquartile Range) = 129
- All tests for normality for Survival _months variable were <0.05.
- ⇒Shows non Normal distribution.

Descriptive statistics

The UNIVARIATE Procedure
Variable: Survival_months

Moments			
N	723916	Sum Weights	723916
Mean	108.405457	Sum Observations	78476445
Std Deviation	86.0216328	Variance	7399.7213
Skewness	0.7661484	Kurtosis	-0.3355452
Uncorrected SS	1.3864E10	Corrected SS	5356769248
Coeff Variation	79.3517549	Std Error Mean	0.1011029

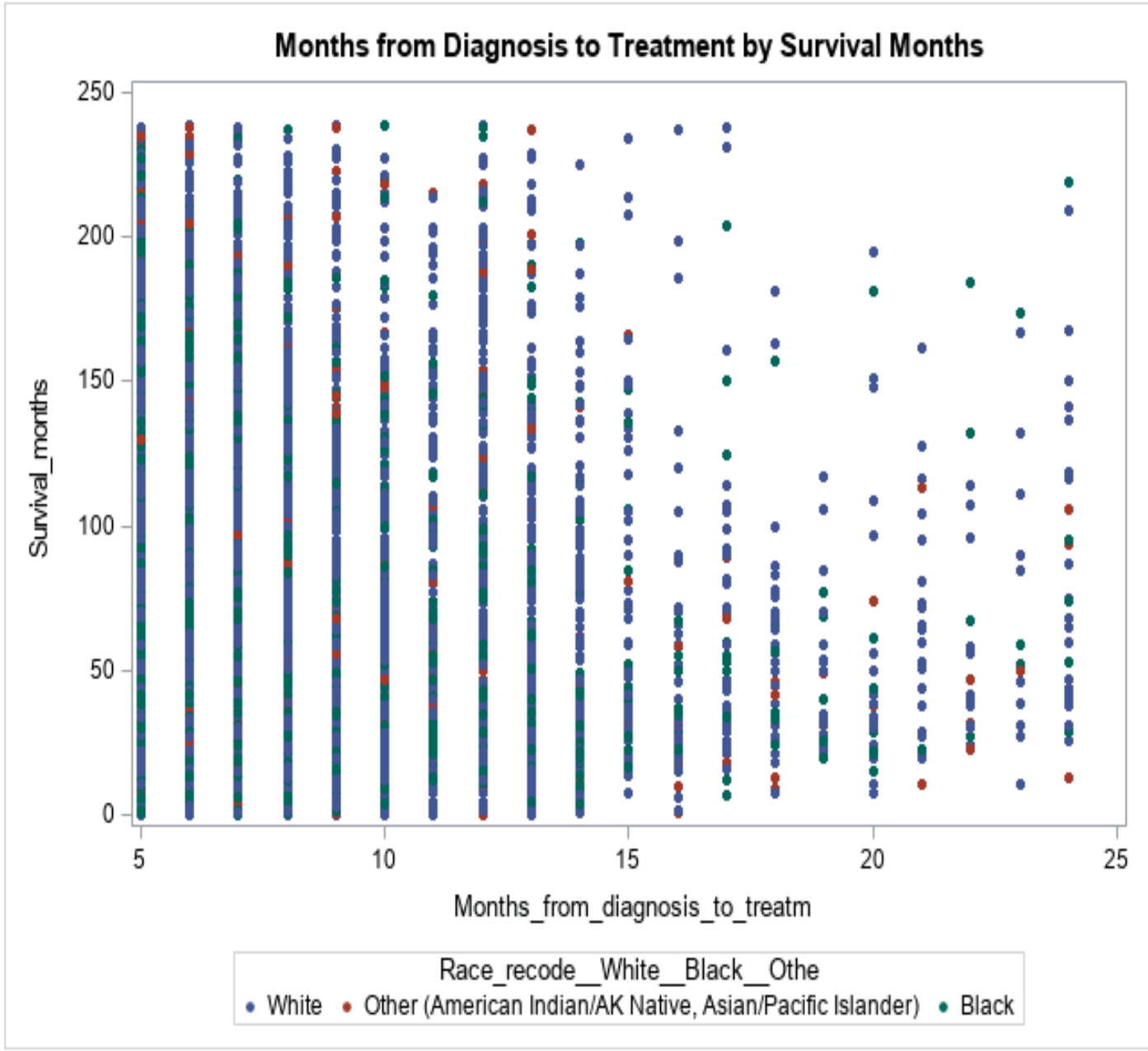
Basic Statistical Measures			
Location		Variability	
Mean	108.4055	Std Deviation	86.02163
Median	88.0000	Variance	7400
Mode	0.0000	Range	348.00000
		Interquartile Range	129.00000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	1072.229	Pr > |t|	<.0001
Sign	M	357830	Pr >= |M|	<.0001
Signed Rank	S	1.28E11	Pr >= |S|	<.0001

Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.103797	Pr > D	<0.0100
Cramer-von Mises	W-Sq	2650.471	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	16830.85	Pr > A-Sq	<0.0050

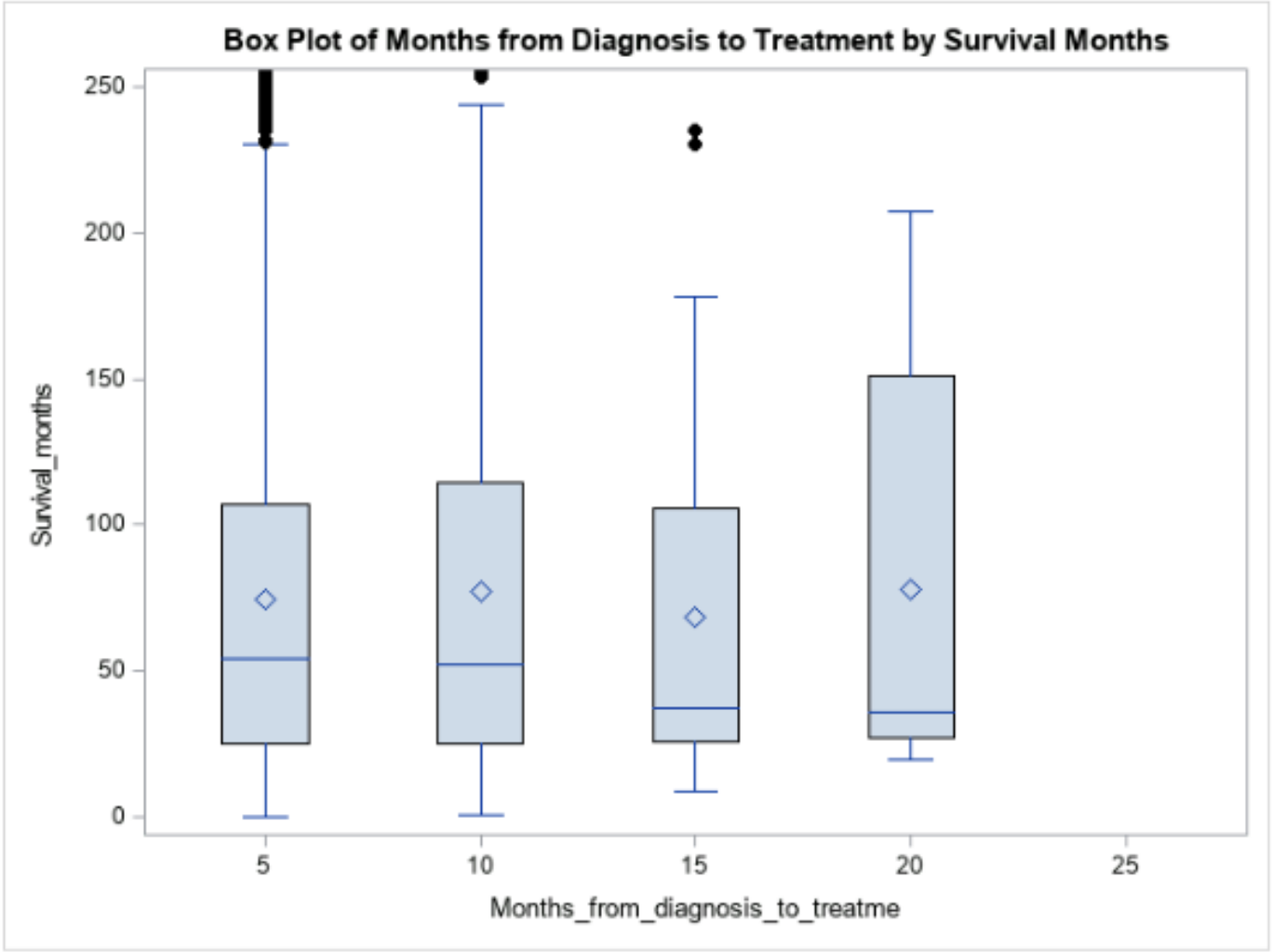
Scatter Plot

Grouped scatter plot displaying as months from diagnosis to treatment increases, length of survival decreases. Weak/negative correlation.



Box Plot

Similar to the scatter plot, this Box Plot is also displaying that as months from diagnosis increases, length of survival decreases.



5-year Survival Association with Race

Survival rate between race?

- Only 57% of breast cancer patients survived for at least 5 years.
- 5yr survival among Race
  - Black = 50.64%
  - Other Race = 55.24%
  - White = 58.62%

Black patients have lower 5 yr survival

chi-square analysis 5 years survival

The FREQ Procedure

Table of New_Survivalcode by Race_recode__White__Black__Othe

Frequency Expected Percent Row Pct Col Pct	Race_recode__White__Black__Othe			
New_Survivalcode	Black	Other (American Indian/AK Native, Asian/Pacific Islander)	White	Total
N	53992 48466 4.95 11.65 49.36	43107 40912 3.95 9.30 44.76	366502 376223 33.58 79.06 41.38	463601 42.48
Y	55390 62916 5.08 8.82 50.64	53200 55395 4.87 8.48 55.24	519136 509415 47.57 82.70 58.62	627726 57.52
Total	109382 10.02	96307 8.82	885638 81.15	1091327 100.00

Statistics for Table of New_Survivalcode by Race_recode__White__Black__Othe

Statistic	DF	Value	Prob
Chi-Square	2	2760.7342	<.0001
Likelihood Ratio Chi-Square	2	2738.1389	<.0001
Mantel-Haenszel Chi-Square	1	2748.9001	<.0001
Phi Coefficient		0.0503	
Contingency Coefficient		0.0502	
Cramer's V		0.0503	

Sample Size = 1091327

*New_Survivalcode Y = Patient survived for >=5 years

Chi Square Analysis

Only 40% patients started Chemotherapy.

(While only 51.6% of black patients, 38.08% white patients, and 44.08% others received chemotherapy.)

Patients who got chemotherapy have better 5 year survival (58.19%)

chi-square analysis chemotherapy with Race

The FREQ Procedure

Frequency Expected Percent Row Pct Col Pct	Table of Chemotherapy_recode__yes__no_un by Race_recode__White__Black__Othe			
	Race_recode__White__Black__Othe			
Chemotherapy_recode__yes__no_un	Black	Other (American Indian/AK Native, Asian/Pacific Islander)	White	Total
No/Unknown	52884 65666 4.85 8.07 48.35	53856 57816 4.93 8.22 55.92	548421 531679 50.25 83.71 61.92	655161 60.03
Yes	56498 43716 5.18 12.95 51.65	42451 38491 3.89 9.73 44.08	337217 353959 30.90 77.31 38.08	436166 39.97
Total	109382 10.02	96307 8.82	885638 81.15	1091327 100.00

Statistics for Table of Chemotherapy_recode__yes__no_un by Race_recode__White__Black__Othe

Statistic	DF	Value	Prob
Chi-Square	2	8222.9656	<.0001
Likelihood Ratio Chi-Square	2	8083.8640	<.0001
Mantel-Haenszel Chi-Square	1	8203.1146	<.0001
Phi Coefficient		0.0868	
Contingency Coefficient		0.0865	
Cramer's V		0.0868	

Sample Size = 1091327

*Missing data excluded

chi-square analysis 5 years survival by chemotherapy

The FREQ Procedure

Frequency Expected Percent Row Pct Col Pct	Table of New_Survivalcode by Chemotherapy_recode__yes__no_un		
	Chemotherapy_recode__yes__no_un		
New_Survivalcode	No/Unknown	Yes	Total
N	281234 278316 25.77 60.66 42.93	182367 185285 16.71 39.34 41.81	463601 42.48
Y	373927 376845 34.26 59.57 57.07	253799 250881 23.26 40.43 58.19	627726 57.52
Total	655161 60.03	436166 39.97	1091327 100.00

Statistics for Table of New_Survivalcode by Chemotherapy_recode__yes__no_un

Statistic	DF	Value	Prob
Chi-Square	1	133.1218	<.0001
Likelihood Ratio Chi-Square	1	133.1864	<.0001
Continuity Adj. Chi-Square	1	133.0762	<.0001
Mantel-Haenszel Chi-Square	1	133.1217	<.0001
Phi Coefficient		0.0110	
Contingency Coefficient		0.0110	
Cramer's V		0.0110	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	281234
Left-sided Pr <= F	1.0000
Right-sided Pr >= F	<.0001
Table Probability (P)	<.0001
Two-sided Pr <= P	<.0001

Sample Size = 1091327

*Missing data excluded

Analysis of variance:Months of survival and Age category

- Ho: No difference in means
- H1: there is difference

- Findings: P<0.05 suggests rejecting the null hypothesis, there is a statistically significant difference in survival months between age categories.

Analysis of variance in Surval months between Age groups

The ANOVA Procedure

Dependent Variable: Survival_months

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	17	196986548	11587444	2939.47	<.0001
Error	1.1E6	4322670133	3942		
Corrected Total	1.1E6	4519656680			

R-Square	Coeff Var	Root MSE	Survival_months Mean
0.043584	72.52531	62.78551	86.57049

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Age_recode_with__1_y	17	196986547.8	11587444.0	2939.47	<.0001

Independent T-Test: Survival Months by Gender

Means and Standard Deviations:

- Female group: Mean = 108.6, S.D = 86.08
- Male group: Mean = 83.55, S.D = 73.64

95% Confidence Intervals:

- Females = 108.4 - 108.8
- Males = 81.51 - 85.59
- Difference in means = 22.64 - 27.42

T-Test Results:

- The t-value is 20.54 with a very low p-value (< 0.0001), suggesting a statistically significant difference in survival months between males and females.

The TTEST Procedure

Variable: Survival_months1

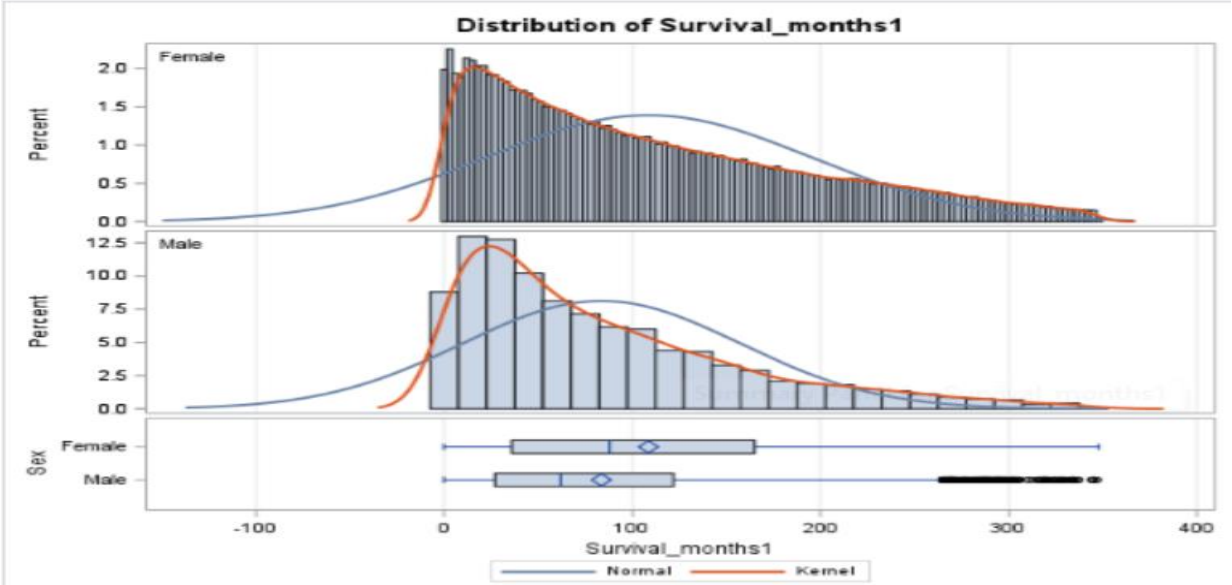
Sex	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
Female		718900	108.6	86.0766	0.1015	0	348.0
Male		5016	83.5486	73.6421	1.0398	0	347.0
Diff (1-2)	Pooled		25.0302	85.9966	1.2185		
Diff (1-2)	Satterthwaite		25.0302		1.0447		

Sex	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Female		108.6	108.4	108.8	86.0766	85.9361	86.2175
Male		83.5486	81.5102	85.5871	73.6421	72.2288	75.1122
Diff (1-2)	Pooled	25.0302	22.6421	27.4184	85.9966	85.8568	86.1369
Diff (1-2)	Satterthwaite	25.0302	22.9821	27.0784			

Equality of Variances:

- The test for equality of variances (Folded F) is statistically significant ($p < 0.0001$), indicating that the variances in survival months are significantly different between males and females.

In summary, females have a significantly higher and clinically meaningful survival rate than males, supported by both statistical tests and confidence intervals.



Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	723914	20.54	<.0001
Satterthwaite	Unequal	5111.1	23.96	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	718899	5015	1.37	<.0001

Box Plot : Survival Months by Gender

Females:

1. Central Tendency:
Mean survival month: ~91.24
Median (50th percentile): 60.0

2. Variability:
Standard deviation: ~85.42
Interquartile range (IQR): 118.0

3. Distribution Shape:
Positively skewed (Skewness = 1.12)
Kurtosis: 0.22 (relatively normal)

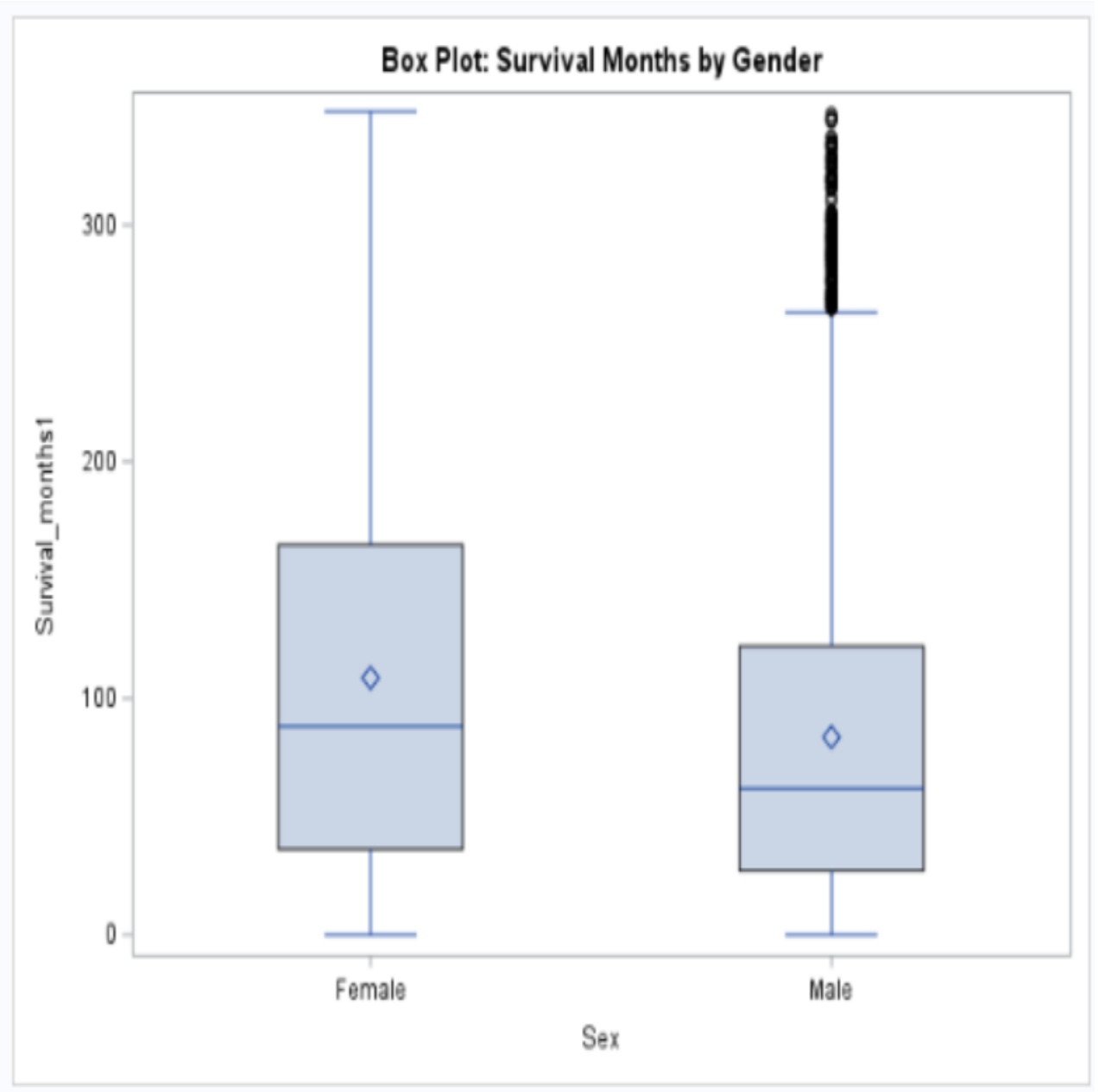
Males:

1. Central Tendency:
Mean survival month: ~108.41
Median (50th percentile): 88.0

2. Variability:
Standard deviation: ~86.02
Interquartile range (IQR): 129.0

3. Distribution Shape:
Positively skewed (Skewness = 0.77)
Kurtosis: -0.34 (flatter distribution)

- Both genders show positively skewed distributions with wide ranges.
- Extreme values impact the mean, which is higher than the median for both males and females.
- Females have a relatively lower median survival month compared to males.
- The distributions have varying shapes, with females showing a right-skewed distribution and males having a flatter distribution.



Contingency Table : Sex vs Survival

Cell Frequencies:

- The frequencies in the cells show the distribution of individuals who survived across different months for both males and females.

Chi-Square Test:

- The Chi-Square test indicates a statistically significant realation between "Sex" and "Survival_months1" (p-value < 0.05).

Observation:

- While a statistically significant relation between gender and survival months has been identified, the weak association indicates that gender might not be a strong predictor or determinant of survival months after diagnosis.

Contingency Table: Sex vs. Survival (2010-2020)

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Sex	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	...	125	126	127	128	129	130	131	Total
Female	5135 1.63 1.64 99.00	4173 1.32 1.33 99.33	4127 1.31 1.32 99.28	3752 1.19 1.20 98.89	3791 1.20 1.21 99.16	3618 1.15 1.16 98.93	3578 1.13 1.14 98.95	2505 0.79 0.80 98.93	2282 0.72 0.73 99.30	3560 1.13 1.14 99.16	3506 1.11 1.12 99.12	3789 1.20 1.21 99.16	3497 1.11 1.12 99.23	3447 1.09 1.10 99.28	3565 1.13 1.14 99.14	3447 1.09 1.10 99.42	3494 1.11 1.12 99.06	3344 1.06 1.07 99.26	3542 1.12 1.13 99.24	3174 1.01 1.01 98.94	3490 1.11 1.11 99.12	...	1374 0.44 0.44 99.42	1406 0.45 0.45 99.79	1379 0.44 0.44 99.35	1240 0.39 0.40 99.28	1337 0.42 0.43 99.48	1326 0.42 0.42 99.18	1174 0.37 0.37 99.75	313087 99.25
Male	52 0.02 2.20 1.00	28 0.01 1.18 0.67	30 0.01 1.27 0.72	42 0.01 1.78 1.11	32 0.01 1.35 0.84	39 0.01 1.65 1.07	38 0.01 1.61 1.05	27 0.01 1.14 1.07	16 0.01 0.68 0.70	30 0.01 1.27 0.84	31 0.01 1.31 0.88	32 0.01 1.35 0.84	27 0.01 1.14 0.77	25 0.01 1.06 0.72	31 0.01 1.31 0.86	20 0.01 0.85 0.58	33 0.01 1.40 0.94	25 0.01 1.06 0.74	27 0.01 1.14 0.76	34 0.01 1.44 1.06	31 0.01 1.31 0.88	...	8 0.00 0.34 0.58	3 0.00 0.13 0.21	9 0.00 0.38 0.65	9 0.00 0.38 0.72	7 0.00 0.30 0.52	11 0.00 0.47 0.82	3 0.00 0.13 0.25	2364 0.75
Total	5187 1.64	4201 1.33	4157 1.32	3794 1.20	3823 1.21	3657 1.16	3616 1.15	2532 0.80	2298 0.73	3590 1.14	3537 1.12	3821 1.21	3524 1.12	3472 1.10	3596 1.14	3467 1.10	3527 1.12	3369 1.07	3569 1.13	3208 1.02	3521 1.12	...	1382 0.44	1409 0.45	1388 0.44	1249 0.40	1344 0.43	1337 0.42	1177 0.37	315451 100.00

Phi Coefficient, Contingency Coefficient, Cramer's V:

- These coefficients measure the strength of association. In this case, they are all very close to zero (0.0233), suggesting a weak association between gender and survival months

Statistics for Table of Sex by Survival_months1

Statistic	DF	Value	Prob
Chi-Square	131	171.8578	0.0096
Likelihood Ratio Chi-Square	131	180.2993	0.0028
Mantel-Haenszel Chi-Square	1	60.8604	<.0001
Phi Coefficient		0.0233	
Contingency Coefficient		0.0233	
Cramer's V		0.0233	

Sample Size = 315451

Linear Regression Analysis

Linear regression analysis in SAS is used for modeling the relationship between a dependent variable and one or more independent variables. It helps in predicting outcomes and understanding the strength and nature of the relationship between variables.

Our R-Square value here is approximately 0.26, suggesting that around 26% of the variability in survival can be explained by our model.

The F-values are well above the threshold for statistical significance, showing that our model has a strong fit.

All p-values are <0.05→ all variables are significantly associated with survival months.

In conclusion, this regression model is a valuable tool for predicting outcomes and analyzing the interplay of various factors affecting survival time.

Linear Regression analysis

The GLM Procedure

Dependent Variable: Survival_months

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	118632839	29658210	7391.32	<.0001
Error	1.09E6	4379006311	4013		
Corrected Total	1.09E6	4497639149			

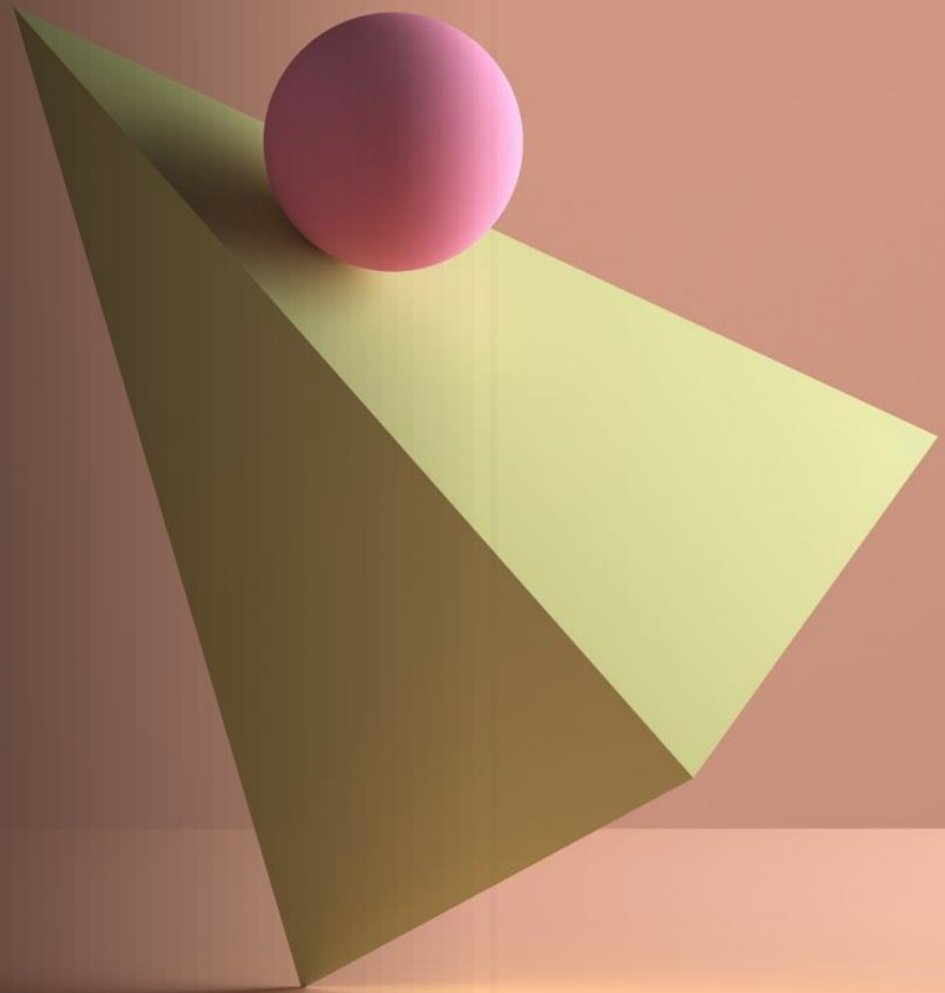
R-Square	Coeff Var	Root MSE	Survival_months Mean
0.026377	73.07641	63.34485	86.68304

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Months_from_diagnosi	1	104027994.3	104027994.3	25925.5	<.0001
Chemotherapy_recode_	1	2889798.8	2889798.8	720.19	<.0001
Race_recode__White__	2	11715045.5	5857522.8	1459.79	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Months_from_diagnosi	1	98671036.91	98671036.91	24590.5	<.0001
Chemotherapy_recode_	1	3989392.02	3989392.02	994.22	<.0001
Race_recode__White__	2	11715045.54	5857522.77	1459.79	<.0001

*New_Survivalcode Y= Patient survived for >=5 years

Conclusion



Insights from Breast Cancer Survival Data

- Survival Rates Following Diagnosis
- Length of Survival
- Differences in Survival Rates
- Chemotherapy differs based on groups

Methods Critique and Suggestions

- Analysis Approach
- Examination of Disparities
- Understanding Treatment Utilization

- Enhanced Reporting
- Exploring Various Factors
- Consider Generalizability

What Would We Have Done Differently?

- Consider performing more in-depth exploratory data analysis
- Evaluate the impact of outliers
- Analyze time-to-event data more effectively.

Enjoyable and Challenging Aspects

- Enjoyable: Creating visualizations like pie charts, bar charts, and scatter plots to understand the data distribution.
- Challenging: Dealing with missing or unknown data and deciding on appropriate strategies for handling them. Also, interpreting the results of statistical tests accurately.

References

Centers for Disease Control and Prevention. (2022, September 26). Breast cancer. Centers for Disease Control and Prevention. Retrieved December 1, 2022, from https://www.cdc.gov/cancer/breast/index.htm

Akram, M., Iqbal, M., Daniyal, M., & Khan, A. U. (2017). Awareness and current knowledge of breast cancer. Biological Research, 50(1). https://doi.org/10.1186/s40659-017-0140-9

Sun, Y.-S., Zhao, Z., Yang, Z.-N., Xu, F., Lu, H.-J., Zhu, Z.-Y., Shi, W., Jiang, J., Yao, P.-P., & Zhu, H.-P. (2017). Risk factors and preventions of breast cancer. International Journal of Biological Sciences, 13(11), 1387–1397. https://doi.org/10.7150/ijbs.21635

Wheeler, S. B., Reeder-Hayes, K. E., & Carey, L. A. (2013). Disparities in breast cancer treatment and outcomes: Biological, social, and Health System Determinants and opportunities for research. The Oncologist, 18(9), 986–993. https://doi.org/10.1634/theoncologist.2013-0243


THANK
YOU