DATA 606 Data Project Proposal

KoohPy <- Koohyar Pooladvand

2024-04-20

Data Preparation

In this project, I have chosen to work on breast cancer. There are various resources available regarding this particular topic, with the SEER being the most reliable one.

The Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute (NCI) collects and publishes cancer data through a coordinated system of strategically placed cancer registries, which cover nearly 30% of the US population.

Currently, there are 18 SEER registries in the USA. This information can be found on the following website: https://seer.cancer.gov/data/access.html.

I have also used the following repository to assist me with this project: https://github.com/kohyarp/SEER_solid_tumor. The Database contains tons of data, the goal of my investigation will be focused only on BREAST cancer for 2011-2015 and 2019-2020. SEER has a software *STAT that I have used to import the data to a test that will be stored and used on my local computer. Additionally there is a GITHUB repository that I have used to some extent in this project. The repository is focused on all type of cancer, but my study is focused on BREAST, and I aim different question to answer. https://github.com/zgalochkina/SEER_solid_tumor

Research question

The primary question I aim to address is the survival rate of breast cancers and the influence of factors such as age, type, sex, and other parameters on this rate. Notably, a five-year threshold is commonly used to determine survival rates. Although my understanding of the rationale behind this five-year benchmark is limited, recognizing its significance has led me to divide the data into two separate datasets.

The dataset spanning from 2011-2015 assumes that the status of all patients within that period is known up to the database's current date in 2022. Additionally, I have selected the most recent data from 2019-2020 as my target years for potential correlation and regression studies to estimate survival rates.

This analysis is not scientific but rather a straightforward statistical exercise with no purpose beyond this course. However, I find the subject intriguing to investigate. I am uncertain if I will discover any significant relationships or correlations, and if found, whether they will be relevant, as I am not an expert in the field of breast cancer. My choice of topic is personal, as I have witnessed immediate family members diagnosed with this cancer, and I wish to gain a deeper understanding.

The database for 2011-2015 contains approximately 303,000 rows with 36 selected columns. I have chosen to focus solely on the 2019-2020 data, which comprises about 131,000 rows for prediction purposes. The question at hand is complex, and while I do not anticipate a definitive answer, I hope to uncover some patterns and test hypotheses, as well as engage in general data work, from tidying to cleaning.

Furthermore, I plan to explore regression analysis to determine if I can identify any linear or non-linear relationships among the critical parameters.

My knowledge of the subject is not extensive, but I am eager to learn as I progress.

Some of the general parameters to consider are as follows: * Years of diagnoses; * Age groups at diagnosis; * Cancer type (BREAST);

Some other parameters are also available to be edited, but they are secondary.

"to be added: adding a brief literature review to provide context for my research questions and hypotheses. This could include previous studies on breast cancer survival rates, factors affecting survival, and methods used for analysis."

White outstely tiveori- di- formediates 2 (in- in yearsic- lastdead the Hispanic-quadif- his-gin a- are or clud- metropioli- condue apyLatino ranfer- tol-of tion avail- more ing tan ties tacto afford of en- ogypri- able pri- com- ar- can- ter bretist mary and maries mon eas cer surgery ated; there law) ge Grade are 1 II more million 0 pop days of sur-vival													
and ori-													
Origin Site First of of Mahouse SEER		Race						tal tal					
Site		and						num u m	1- M	e-			
Rack SMACD		ori-							dia	an			
Coole Cool		gin Si	te				First	of of	Malno	use-		SEEF	t
Rack EMM-CD - ag		re- re	-		Months		ma-	in be-	i- ho	dd	Race	other	Ori-
Rack ISMMCD-		code co	$_{ m ode}$	Di-	from Se	cope	lig-	situn/ingua	a/btealrin-	_	and	cause	$_{ m gin}$
Year-NHB; O- Pri- Graute Chêtneyshteers pri- nanine tusin-Rural-gin/Year death- code of codn/Hedde mayer.death- code of codn/Hedde mayer.death- code to in-Sompress tusin- codn/Hedde tusin-grade/Hedde tusin- Rural-gin/Year death- code of codn/Hedde mayer.death- code to in-Sompress di-tion/Come-omfollowis- SizSun/Hisa- tusin- code	R	a¢NBM(CD-	ag-	di- of			lig-der-	staco	me	ori-	of	_
CODATY tu-tu-					RaathBreae	eg		-		Rural	-ginYea	ar dea Th	ı- code
di-(WANC)2028it4Re-ClilegCorlatre-a-sis no nd Sur- to in-Se-morsors di-tiorCome-omfollofvsi-Siz8ut(klisag-B, NHG)PRe-code i- fir-er-cockiono cansurgivaBusiteli-quénéeror Pasg-adjtin-codemendebath Suñsyspamiic/Suno-AIII-lisá/WiH60-(truel.cal.manal-(yese-trouble-de)46988ntiwa-ca-numpa-pa-tieinut-o-uni/ Gesis APhn2098oibel201/20/2008bity no/soudrestrig060)gmorather bertientient D siz 202Co69, \$EERthockion(20/2009Bisp) Fe-201W1NionBrcBst6550AiBlahlchbyRjyKeBe00cSunRingia(690MiNo 2nd)2 0 White outstely tivori- di- formediates 2 (in- in yeamsc- lastead theftispanic- qualif- his-gin a- are or clud- metrojioli- conduc apyLatino able pri- com- ar- can- ter bretist mary and maries mon eas cer surgery ated; Grade are II more than lion pop days of sur- vival Fe-201W1NionBrcBst650DaBlahlchbyRjgNo/NobitsWisher(690BrcNos3cr03 0 340D-875(0000-All201201&livelanish)Npn- dates 3 in yeamsc- dead tenHispanic- dead tenHispanic- days of sur- vival Fe-201W1NionBrcBst650DaBlahlchbyRjgNo/NobitsWisher(690BrcNos3cr03 0 340D-875(0000-All201201&livelanish)Npn- dates 3 in yeamsc- dead tenHispanic- dead tenHispanic- days of sur- wind maries cas cer sur- ar- can- and/or avail- more tito-of avail	of co	odNeHeAodde								-UrbAng	néreco-f	Yeadasmo	oRXNHIA
ag-B, NHG+Re— code i fir-er-codionto camergivaliusitali-quance de Paag-adjtin-codenende de l'Assis y spennier, Su no-Alliss/Willio (theal cal manal- (vese-treibir de 1984 de 1984) (vese-treibir de 1984) de 1985 de 1984 (1985) de 1985 (1984) de 19			·	1.0						_	- (
The content of the	,			~									,
Sexis AP\$\times in in years to lead 1 (20(20) in its in years to lead 1 (20(20) in its in years to lead to lea													
The color Norm No													
Hispanic Upper- i- ra- per plete of ried ties54 races/ather tenfipanish- white outsetely tivori- quadif- hissin a- are or clud- metropioli- conduc apyLatino ranfer- tol-of tion avail- more ing tan ties tacto af- of en- ogypri- breisit mary and maries mon eas cer surgery ated; Grade III — more law) ge III — more law) ge Grade III — more law) ge III — more law	-	-, ,		, ,, ,	, ,	. , .							
White outstely tiveri di-formediates 2 (in- in yeamsic-quadif-hisgin a- are or clud-metropoli-conduc application and the fispanic policy of enopypri- able pri- com- ar- can-ter bretist mary and maries mon eas cer surgery ated; ated; are 1 more mil-lion 0 pop days of survival White NOS tiveri- om- dates 3 in yeamsic- dead terminisgin mendeatre or metropoli- due the Eatino application avail- more than ties tacto af- terminisgin mendeatre or metropoli- due the Eatino application avail- more than ties to appropriate the first area and/or avail- more than ties to appropriate the pri- ar- can- and/or appropriate there are 1 cal-more than ties to appropriate the pri- ar- can- and/or appropriate there are 1 cal-more than lion ce- days of sur- cal-more than lion ce- days of sur- cal-more more mil- pro- than lion ce- days of sur- cal-more more mil- pro- days of sur- cal-more more mil- pro- days of sur- ce- days of sur-				() () -	_	,							,
ranfer- tol-of tion avail- more ing tan ties tacto affer- of en- ogypri- able pri- com- ar- can- ter surgery ated; ated; there are III more ing tan ties tacto affer- of en- ogypri- able pri- com- ar- can- ter surgery ated; there law ge are III more ing tan ties can- ter surgery ated; there are III more IIII more III more III more III more III more III more III more IIII more III more III more III more III more III more III more IIII more III more III more III more III more III more III more IIII more III more III more III more III more III more III more IIII more III more III more III more III more III more III more IIII more III more III more III more III more III more III more IIII more III more III more III more III more III more III more IIII more III more III more III more III more III more III more IIII more III more III more III more III more III more III more IIII more III more III more III more III more III more III more IIII more III	$_{\mathrm{male}}$	-			_	-							-
		White	v		di- form	eddates	2		,				-
of en- bretist mary and maries mon eas cer surgery ated; there law) ge Grade are 1 II more mil- than lion 0 days of sur- vival Fe-201W MoraBrdase 650 9 Blank 160 (2) gNo Nobel (Nobel) a Glob (6) 2Brolis (3) a Glob (6) 2Brolis			-	_	ો-	are	or		clud-	metro	pioli-	condue	
bretist mary and maries mon eas cer surgery ated; there law) ge Grade are 1 II more than lion 0 ays of sur- vival Fe-201W/NooBrets1635000Blanks005000FBlanks0050000FAll201201AlivBanks0000- hale Hispanic Bretstown i- kndwnowrrec- plete of vorcedties44 races/ether syssSpanish- his-gin mendodre or metropioli- due theLatino tol-of avail- more tan ties to apy ogypri- able pri- ar- can- and/or mary and maries eas cer sur- there are 1 cal more mil- lion ce- ge gi- are 1 cal more mil- lion ce- pop dures				tol-of t	ion	avail-	\mathbf{m}	ore	$_{ m ing}$	\tan	ties	tac t o	af-
ated; there law) ge 1 1 1 1 1 1 1 1 1			of en-	ogypri-		able	pr	i-	com-	ar-		can-	ter
Grade III more million 0 pop days of survival Fe-201W/NimBrdast6509-BlaBlaBlaBlaBlaBlaBlaBlaBlaBlaBlaBlaBlaB			bretaist	mary		and	\mathbf{m}	aries	mon	eas		cer	surgery
Fe-2017V NtorBrdsstastOyn BlankstastOyn			ated;			there			law)	ge			
than lion pop days of sur- vival Fe-20 IV/INGEBreBse656U9-BlaBlableBseByNo Notation May Non- male Hispanic Brewn i knewnownrec- plete of vorcedties44 races/ether sysSpanish- White NOS tivori- his-gin mendeare or metropioli- due theHatino tol-of avail- more tan ties to apy ogypri- able pri- mary and maries eas cer sur- there ge gi- are 1 cal more than lion pop in yeamsc- dead tenfitispanic- tan ties to apy and/or and/or ar- can and/or ge gi- are 1 cal more mil- lion ce- of pop dures days of sur-			Grade			are				1			
Ge-201W1NtomBreBste65tU9hBlahlands(R)gNo Nobel Nobel Survival Fe-201W1NtomBreBste65tU9hBlahlands(R)gNo Nobel Nobel Survival Fe-201W1NtomBreBste65tU9hBlahlands(R)gNo Nobel Nobel Survival Fe-201W1NtomBreBste65tU9hBlahlands(R)gNo Nobel Nobel Survival Fe-201W1NtomBreBste65tU9hBlahlands(R)gNo Nobel Nobel Survival Hispanic Bredstown i kndwnownec plete of vorcedties44 races/ether sysSpanish- there with the survival and maries of the survival and the s			II			more				mil-			
days of sur- vival Fe-201W/NicoBrdsc65U9Blainless(R)gNo North North and series of sur- wival Hispanic Brdastown i knownownec- hisgin mendoure or metropoli- due the tatino tol-of avail- more tan ties to apy ogypri- able pri- ar- can- and/or mary and maries eas cer sur- there ge gi- are 1 cal more than lion ce- than lion ce- of gi- sur- wival 44 Parexis 3 of 34 Di-\$75 (0000- All 2012) 15 li Wank (N) on- was Sys Spanish- tenffits panic- the tenffits panic- the tan ties to apy ogypri- able pri- ar- can- and/or are 1 cal more mil- lion ce- of pop dures						than				lion			
days of sur- vival Fe-201W/NicoBreasea50U9-Blairle (No. No. No. No. No. No. No. No. No. No.						0				pop			
of survival Fe-201W/NomBrdasc656U9-BlaBlaBlasceR)gNo/NomBrdasc8180Bya6k(8)gNo/NomBrdasc856U9-BlaBlasceR)gNo/NomBrdasc8180Bya6k(8)gNo/NomBrdasc818Bya6k(8)gNo/NomBrdasc818Bya						days							
Fe-201WhitenBreaseG5tU9hBlahletshek)gNotNombiaNhotBlack(R)gNotNomb													
Fe-201WhitenBreaseG5tU9hBlahletshek)gNotNombiaNhotBlack(R)gNotNomb													
Fe-201W histor Breast as 0.9 Bland 1994; No. 1 - known owner plete of vorced ties 44 races/ether sys Spanish White NOS tive ori- om- dates 3 in years; c dead tentific spanic his-gin mende or metropioli- due the Eatino tol-of avail- more tan ties to apy ogypri- able pri- ar- can- and/or mary and maries eas cer surthere are 1 cal more than lion cells are 1 cal more than lion cells are 1 proposed days of sur-													
Hispanic Brekastym i knokumownrec- plete of vorcedties44 races/ether sysSpanish-White NOS tiveori- om- dates 3 in yeamsc- dead tenfitispanic- his-gin mendeadre or metropoli- due the Latino tol-of avail- more tan ties to apy ogypri- able pri- ar- can- and/or mary and maries eas cer sur-there are 1 cal more than lion ce-than lion ce-days of sur-	Fe-201V	V hNtoerBreB:	s#G 5(1 19-BlaF	la Media (Sental			reNet3r	d03_0_3	34 0 Di-\$7	5CO600	-A11201	2601 & 1i√Rel	aNkoNon-
White NOS tiveorion—dates 3 in yeamic—dead tenfficispanic—his-gin mendeadre or metropoli—due the Latino tol-of avail—more tan ties to apy ogypri—able pri—ar—can—and/or mary and maries eas cer sur—there ge gi—are 1 cal more mil—pro—than lion—ce—days—of—sur—of—sur—sur—sur—sur—sur—sur—sur—sur—sur—sur	male												` '
his-gin mendeare or metropoli- due the Latino tol-of avail- more tan ties to apy ogypri- able pri- ar- can- and/or mary and maries eas cer sur- there are 1 cal more than lion ce- than lion ce- days of sur-	111010	-				-			,0100		,		v -
tol-of avail- more tan ties to apy ogypri- able pri- ar- can- and/or mary and maries eas cer sur- there ge gi- are 1 cal more mil- than lion ce- 0 pop dures days of sur-		***11100	1105										_
ogypri- mary and maries eas cer sur- there ge gi- are 1 more mil- than lion ce- 0 pop dures days of sur-				_	mon	_					-		
mary and maries eas cer surthere ge gi- are 1 cal more mil- than lion ce- 0 pop dures days of sur-											uics		
there ge gi- are 1 cal more mil- than lion ce- 0 pop dures days of sur-						_							,
are 1 cal more mil- pro- than lion ce- 0 pop dures days of sur- sur-				шагу			111	aries				cei	
more mil- pro- than lion ce- 0 pop dures days of sur-													_
$\begin{array}{ccc} & & & & & & \\ & & & & \\ 0 & & & & \\ & & & \\ days & & \\ of & & \\ sur- & & \\ \end{array}$													
$\begin{array}{c} 0 & \text{pop} & \text{dures} \\ \text{days} & \\ \text{of} & \\ \text{sur-} \end{array}$													-
days of sur-													
of sur-										pop			aures
sur-													
vival													
						vival							

	Race and ori-					To-To- tal tal numum- berber	di	an			
	gin Si	te			First	of of	Maln∈	use-		SEEI	
	re- re	-		Months	ma-	in be-	i- ho	ld	Race	other	Ori-
	code co		Di-	from Scope	lig-	situn/ingna,	htarin		and	cause	gin
	a¢ NISM		ag-	di- of		lig-der-	staeo		ori-	of	re-
of co di-(\' ag-B no-A	WJANCD(5, NH@PR, 1,His3/Wi	ma GraGrad 2 3 it Re-Clin e- code i- HO- (th c alca	chetthic apyd og Cont-atre- a fir-er-code al ma-al- (yeso	taethReareg i-no-sonlymph - sis no nd Sur- toto cannergiva e-treatir(d1998)nt ude)entr20092)gm	COMary to in-Se usitedi-qu kmd-ca-nu	e-mo ra ors nef oc efor P n pa -pa-ti	at fla di-tio aag-ad emmo-to	-UrbAn onCome jtin-co uun(r	gerecof - omfo odmenç <60,60	ll o fwsi- Si beddeafth Su -re-ca-m	oRXNHIA zSumHis- uSyspemiic aSeqNon-
-	-, ,	.,,	, ,, ,,	, , , , ,						, ,	
male	u n uom ereta. Hispanio		. , . , - ,	lobol4SinnSenGk(ol) nownper-plete					⊬Ali20 4 races		tem Sp ani
maic	White	inn a tely	tiveri-	formedates			owca		eamisc-	lastdead	-
	***11100	quadif-	his-gin	are	or				opoli-	condue	apyLatin
		ranfer-	tol-of	avail-		ore		tan	ties	tacto	be-
		of en-	ogypri-	able	pr			ar-	0100	can-	fore
		bretaist	mary	and	-	aries		eas		cer	surgery
		ated;	v	there				ge			0 ,
		Grade		are				1			
		II		more				mil-			
				than				lion			
				0				pop			
				days							
				of							
				sur-							
E 0010	WIND D	ACTION DITT	וא עוור ארם ארם ווי	vival	ON I'N O	DO 0 20	ንብሊ በ ው <u></u>	F/00 0 F1	1 4 1100	00011 A 11 D1	L C 1 (h .t)
male				lo n of Singlands (19) Innownper-plete			ried		⊢Ali20 9 races		la Sıks(N) on- tem Sp ani
шаю	White	le- ately	tiveri-	formedates			(in-				theHispa
	VV III UC	siondif-	his-gin	are	or		(condue	apyLatin
		of fer-	tol-of	avail-		ore			_	tacto	af-
		breast	ogypri-	able	pr		com-		0105	can-	ter
		ti-	mary	and	-	aries	mon			cer	surgery
		ated;	v	there			law)	ge			0 ,
		Q 1		are			,	1			
		Grade						mil-			
		Grade II		more							
				more than				lion			
				0 than							
				$\begin{array}{c} \text{than} \\ 0 \\ \text{days} \end{array}$				lion			
				than 0 days of				lion			
				$\begin{array}{c} \text{than} \\ 0 \\ \text{days} \end{array}$				lion			

									To-							
	Race								tal t	al						
	and									ı um-	M					
	ori-								berl			an				
	gin S	Site					Fi	irst	of o	of	Maln∈	ouse-		SEI	${}^{\mathrm{ER}}$	
	re- r	e-		M	onths		m	ıa-	in l	oe-	i- ho	old	Race	oth	er	Ori-
	code c	ode	Di-	fre	om So	cope	lig	g-	situ	yingma/	t tat in	-	and	cau	se	$_{ m gin}$
R	a¢N ISW €	CD-	ag-	di	- of	•	na	ant	lig-c	der-	sta€o	me	ori-	of		re-
Year	e-NH IB, (O- Pri-	Grandos-C	h ≠ t	h B areare	g	pr	ri-	nan	t ne	tusin	- Rura	l-ginYe	ear dea	tlhu-	code
of c	odNHAodi	e ma G ra Gha F	dattino- ap	oydi-no	-sonly	mph	COA	ary	tu-t	u-	at fla	-Urb	ve re c ∙f	Yearlas	moF	XNHIA
di- (WJANÇI2	023iteRe-Clih	egCorLatre	e-a-sis	s no no	l Sur-	to in	ı-Se	-moı	nors	di- tio	onCome	e-omfo	ll o fwsi-	Siz S	un(alHris-
ag-E	B, NH @ PI	R,e- codie i-	fir-er-co	odteionto	canso	e rg ivaSu	sitedi	i- qu	efoce	or Pa	ag-ac	ljtin-co	odmenuç	bedeafth	SunS	ys pemic ,/St
ag-B, NHQPR,e- code i- fir-er-codeiono caracergivaSursitedi-quencefor Pa-ag-adjtin-codene undedeath SurSyspenniic,/Su no-AI,His3/WiHQ- (thrulcal ma-al- (yes-e-treatir-ed-1940) envitand-ca-numa-pa-tienno-to uum/< 60,60-re-ca-ma-SyqNon-																
no-A1,H183/ where (threatealmaal- (yese-tream-eathmentumse-ca-numa-pa-tiemio-to uum/conyon-re-ca-manyeqnon- Sexsis APpan2008onbel201(220(22018) http://www.no./uodhenturg00022.gmocrotheorbertientionsis 202Code,50EE/Rhaddoon(20(1200018))																
Fe-201BlaNorBreaseasou9-Blankank (Seftno Norbelanko) and kan 18 presistand 2 0 547Wi&75(0006-All 201201Alivelank (Non-																
male		ic Br east ,wn				plete		of					earrasces			ys-Spanish-
	Black	NOS	vi- ori-		om-	dates		2				in	nic-			em lik ispanic
			su-gin		mend			or				metr	opoli-	due		he Ł atino
			al- of			avail-		mo	ore			tan	ties	to		ру
			iza-pri-			able		pr				ar-		can		$\frac{1}{nd}$
			tiomary	7		and			aries			eas		cer		ur-
			with-			there						ge			g	
			out			are						1				al
			mi-			more						mil-				ro-
			cro-			than						lion			-	e-
			scopic			0						pop				ures
			con-			days						r				
			fir-			of										
			ma-			sur-										
			tion			vival										
Fe-201V	VH illis Bre	BretG5tM9-BlaE	landak jejn	t Bend	hlSur B é		MailiNe	o2n	d 02 (56	7Ma\$7	'5COO	A 1120	2 0 lixAdir	BelaN	k (Sp)anish-
male	panic	Brænst,	i kı		_	plete		of			ried		4 races			ysHispanic
	(All	NOstely	tiveri-	di-	-	eddates		2			(in-				•	em lia tino
	Races)	dif-	his-gin	a-		are		or					opoli-	comue		her-
	,	fer-	tol-of	tion		avail-		mo	ore		ing	tan	ties	tac t o		ру
		en-	ogypri-			able		pr	i-		com-	ar-		can		nd/or
		ti-	mary	7		and		-	aries		mon	eas		cer		ur-
		ated;	v			there					law)	ge				i-
		$\overline{\text{Grade}}$				are					,	1			_	al
		II				more						mil-				ro-
						than						lion				e-
						0						pop				ures
						days						r · r			-	
						of										
						sur-										
						vival										
						, , ,										

	Race and						To-To- tal tal numur		Ле-			
	ori-						berber		lian			
	gin S	ite				First	of of		ouse-		SEEF	
	re- re			Months		ma-	in be-			Race	other	
	code co		Di-	from Sc	ope	lig-	situn/ing	,		and	cause	0
	a¢NBM		ag-	di- of			lig-der			ori-	of	re-
	e-NH I B, C			aoghBreareg		pri-	nankine			l-ginYear		
		maGraGhal		- no-sonlyr	-	COmbary				- (oRXNHIA
			ogCoiLatre- a-									zounums- nalyspomic,/Sur
no-A	I,His3/W	iHO-(th ca lc	almaal- (ye s e	-treadiredt	9 938 9 nvtil	kmade-ca-nu	ma-pa-	tiemnto-t	o uun(x	< 6 09,600-re	- ca- m	a S eqNon-
	-/ /		2011 8 11 t)y no/ac		.,.						`	
			akatadastseffesN	,	` '		.d02 0					
male	-	c Ov enlapppi	~	nownrec-	plete	of		owe		races/et		sysSpanish-
	White	le-	tiveri-	om-	dates	2			in ye			tem lik ispanic-
		sion	his-gin	mend		or			metro	-	due	theLatino
		of	tol-of		avail-		ore		\tan	ties	to	apy
		breast	ogypri-		able and	pr	ı- aries		ar-		can-	and/or
			mary		there	1116	aries		eas		cer	sur- gi-
					are				ge 1			cal
					more				mil-			pro-
					than				lion			ce-
					0				pop			dures
					days				F ~ F			
					of							
					sur-							
					vival							
	VhHditie-Breb∂	st65tP.4dBl3H	akanan jaka jaka ja				d02 0					a Nk(Sp)anish-
male	panic	Up plefr –	ikndwa	-	-			owe		earrasces/et		sysHispanic-
	(All	outfer-	tiveri-	forme	edates				$_{ m in}$	nic-	tribut	tabidia tino
	Races)	quadr-	his-gin		are	Deathor			metre	-	to	ther-
		ranti-	tol-of		avail-		ore .		\tan	ties	cause	
		of ated;	ogypri-		able	pr			ar-			and/or
		bre Ga tade	mary		and	ma	aries		eas		than	
		III			there				ge 1		this	gi-
					are						can-	cal
					more than				mil- lion		dx	pro- ce-
					0						ux)	dures
					days				pop			aures
					of							
					sur-							
					vival							

```
То-То-
                                                                     tal tal
         Race
                                                                     numum-
         and
                                                                                  Me-
                                                                     berber
                                                                                  dian
         ori-
                                                                                                      SEER
         gin
               Site
                                                               First of of
                                                                               Mahouse-
                                            Months
                                                                     in be-
                                                                               i- hold
                                                                                                     other
                                                                                                               Ori-
               re-
                                                               ma-
                                                                                            Race
         re-
         code code
                               Di-
                                            from Scope
                                                                     situn/ingna/bloak-in-
                                                               lig-
                                                                                            and
                                                                                                     cause
                                                                                                               gin
     RaceNISMACD-
                                                               nant lig-der-
                               ag-
                                            di-
                                                  of
                                                                               sta€ome
                                                                                            ori-
                                                                                                     of
                                                                                                               re-
  Yeare-NHR. O- Pri-
                            Grandos-
                                      Chelanth Reareg
                                                               pri-
                                                                     nankine
                                                                               tusin-Rural-ginYear deaThu-
                                                                                                               code
  of codyHAode maGraGraPathic-
                                     apydi-no-sonlymph
                                                            COmbary tu-tu-
                                                                               at fla-UrbAngerecof YearlasmoRXNHIA
  di-(WIANCE)02Sit&e-ClilegCorLatre-a- sis no nd Sur- to in-Se-morsors di-tiorCorre-omfollofwsi- Siz&ur(His-
  ag-B, NHAPRe- code i- fir-er-codeiono camaergivaSursitedi-quencefor Pa-ag-adjtin-codea enquede ath Sursiyspanniic/Sur
  no-AI, His 3/WiHIQ- (thrul cal maal- (vere-treatire delegation til med-ca-numa-pa-tien to-to-uum (<60); 60-re-ca-mas eq Non-
Sexis APJan200Sorbel201(720(120<del>418) inty</del> no /www.entr2002agmocrather bertientent sis 202Co. 60, SEER (20120-043)
Fe-201WhiteBreasta5tP.O-BlandardsRight Beach SurBlandk (00) 43en to 2nd 20
                                                                            2056Vi$7500666-All20120119eaRlaSiksAvon-
                               i- - known per- plete brovasef
male
         Hispanic Bredist.
                                                                               owed ties84 races/eth(at-tenSpanish-
         White
                  NOGr-
                               tiværi-
                                         di-
                                               formedates cu-
                                                                                      in yearsic-
                                                                                                      tributahleHispanic-
                                                            lar
                                                                                                            apyLatino
                     en-
                               his-gin
                                                     are
                                                                  or
                                                                                      metropoli-
                                                                                                     causesaf-
                     ti-
                               tol-of
                                         tion
                                                     avail- Dis-
                                                                  more
                                                                                      tan
                                                                                            ties
                     ated:
                               ogypri-
                                                     able
                                                            eases pri-
                                                                                      ar-
                                                                                                     other ter
                     Grade
                                                     and
                                                                  maries
                                                                                                      than surgery
                                  mary
                                                                                      eas
                     III
                                                     there
                                                                                                      this
                                                                                      ge
                                                                                      1
                                                     are
                                                                                                     can-
                                                     more
                                                                                      mil-
                                                                                                     cer
                                                                                      lion
                                                     than
                                                                                                     dx
                                                     0
                                                                                      pop
                                                     days
                                                     of
                                                     sur-
                                                     vival
Fe-201BlaNorBreasta50P.84BlaNalandsRjgNto Notae /SinBlandsk(1997AliNeo 3rd04 0
                                                                            26(15i-$75(00000-All2020lix4divBlaNk(Non-
         Hispanic Ovehifappingi- - kndwnownper- plete
                                                                               vorcedties64 races/athor
male
                                                                  of
                                                                                                            svs-Spanish-
         Black
                  le-fer-
                               tiveri-
                                               formediates
                                                                  3
                                                                                      in yearsic-
                                                                                                  lastdead temHispanic-
                  sionen-
                               his-gin
                                                     are
                                                                                      metropoli-
                                                                                                  condue
                                                                                                            the Latino
                                                                  or
                  of ti-
                               tol-of
                                                     avail-
                                                                  more
                                                                                      tan
                                                                                            ties
                                                                                                  tacto
                                                                                                            apy
                  breased;
                               ogypri-
                                                     able
                                                                                                            and/or
                                                                  pri-
                                                                                      ar-
                                                                                                     can-
                     Grade
                                  mary
                                                     and
                                                                  maries
                                                                                                            sur-
                                                                                      eas
                                                                                                     cer
                     III
                                                     there
                                                                                                            gi-
                                                                                      ge
                                                                                      1
                                                                                                            cal
                                                     are
                                                                                      mil-
                                                     more
                                                                                                            pro-
                                                     than
                                                                                      lion
                                                                                                            ce-
                                                     0
                                                                                                            dures
                                                                                      pop
                                                     days
                                                     of
                                                     sur-
                                                     vival
```

kable(head(BREAST DF eval, 10))

				То-То-			
Race				tal tal			
and				numur			
ori-				berber		~	
O .	lite			First of of	Malnouse-	SEI	
	e-		Months	ma- in be-	i- hold	Race oth	
code c		Di-	from Scope	, ,	na∤btoarin-	and cau	0
Ra ¢NI\$™		ag-	di- of	nant lig-der		ori- of	re-
Yeare-NHB,			acthdreaeg	pri- nankine		l-ginYear dea	
of codNHA	·	1.0	v -	Offiary tu-tu-		gerecof Yearlas	
, , ,		~		o in-Se-mo rs o			\
			onto camaergivaSursi				
			- treadired 1990 Penvilus				
Sexsis APphan200x	301bel 20 1(72))(12 011i)oitt)y no /ao	n da) e nt ır 2002 3gmoca	the rbertientier	£tDsis 202€Co 6€	, SŒE)Rd ædteior	(20(1 2 0 10 171stp))
Fe-201AsiAnonBre	BstG5tU8H	1 PosRigNto No	5m0e0/255um=13em978x(na)) 1.45	li N eo 2nd 02 0	27 5D i-\$75 ,0000 5	-All202 0 li A di	& SysNon-
male or Hispani				of		races/ at her	ten Sp anish-
Pa-Asian	le-	tiveri-	formedates	2	in ye	eannisc- ['] lastdea	d the Hispanic-
cifior	sion	hisgin	are	or	metro		-
Is- Pa-	of	tol-of	avail-	more	\tan	ties tacto	af-
landieirc	breast	ogypri-	able	pri-	ar-	can	- ter
Is-		mary	and	maries	eas	cer	surgery
lander		•	there		ge		~ ·
			are		1		
			more		mil-		
			than		lion		
			0		pop		
			days				
			of				
			sur-				
			vival				
Fe-2024)siAbonBreE	BstG5tU8+2	9 PosRigNto No	511049/01RenaBla661x(1919)04	lli N eo2nd02 0	28 701 a\$75 0066	⊢All202 0 li x hdi	50 No Non-
male or Hispani	ic Ov enlap y	ningikndwa	nownom-plete	of	ried ties79	races/ather	sysSpanish-
Pa-Asian	le-	tiveri-	mendedates	2	(in- in ye	:ammisc-lastdea	d tem lik ispanic-
$\operatorname{cifi}\mathbf{o}\mathrm{r}$	sion	$\operatorname{his}_{\operatorname{\mathbf{gin}}}$	un- are	or	clud- metro	opioli-com-lue	the Ł atino
Is- Pa-	of	tol-of	knownavail-	more	ing tan	ties tacto	apy
landific	breast	ogypri-	if able	pri-	com- ar-	can	- and/or
Is-		mary	per- and	maries	mon eas	cer	sur-
lander			formedhere		law) ge		gi-
			are		1		cal
			more		mil-		pro-
			than		lion		ce-
			0		pop		dures
			days				
			of				
			sur-				
			vival				

Race and ori- gin Site re- re- code code RaceNHSMECD- Yeare-NHBE, O- Pri- of codNHAOGE maggraghed di-(WIANCI2023itRe-Clift ag-B, NHQPRe- code i- no-AI,His3/WiHQ- (thrale	Di- f ag- c Grandes- Cheltase Stattino- apydi-n og-ConLatre-a-s - fir-er-codtaion	Months from Scope di- of thereogeno-sonlymph CC sis no nd Sur- to co caraergivaSursite	nant lig-der- pri- nankine Mary tu-tu- in-Se-morson di-quenceorl	dian Mahouse- i- hold Rac a/harin- and - stacome ori- tusin-Rural-gin at fla-Urbanere cs di-tionCome-ome	l cause of Year dea¶h of Yeadasmo follofwsi-Siz uptedleafth Su	Ori- gin re- code oRXNHIA zSur(iHis- uSyspeumic,/Sur
Sexsis APp)an2008onbel201(200	2011 (Sitt) no /aad	a)entirg@0@2jgmocotd	hor bertientien	HD sis 202Co 69,50 H	E)Robodie on (2	0(1 2 01 0) (1 2 010)
Fe-202WhMcmBrd3stG5tU41 2 male Hispanic Up serown White outer quad- rant of breast		wnper- plete formedlates are avail- able and there are more than 0 days of sur- vival	Neo 2nd02 0 3 of 2 or more pri- maries	3067i-\$75(2006i-All vorcedtiesyearscein nic-metropolitan ties ar-eas ge 1 million pop	es/ather - lastdead i- condue	No Non- sysSpanish- tenHispanic- theLatino apy and/or sur- gi- cal pro- ce- dures
Fe-202WhitemBreBrea50U5-2 9 male Hispanic Lookenewn White outer quad- rant of breast		wnper- plete formedlates are available and there are more than 0 days of survival	of 2 or more pri- maries	owed tiesyearsec in nic- metropolitan ties ar- eas ge 1 mil- lion pop	es/ather - lastdead i- condue	SysNon- tenSpanish- theHispanic- apyLatino both be- fore and af- ter surgery

Race and ori- gin Site re- re- code code RaceNISMECD- Yeare-NHB; O- Pri- of codNHAOGE macerached di-(WIANCE023iteRe-Clift ag-B, NHAPR,e- code i- no-AI,His3/WiHl@- (thralo Sexsis APD)an2008onbel@01(20)	Grands- Chetta anglette Rhetthic- apydi-no-son agColl-atre-a- sis no - fir-er-codeionto can cal maal- (yese- treathre	s ma Scope lig of na reg pri lymph COD nd Sur- to in- sucrgiv:Sursiteli- (d.1998-rvilvad-ca-	berber dirst of of Maha- a- in be- i- he- a- situyigua/badin nt lig-der- stacc i- nanktne tusin ary tu-tu- at fla - Se-moraors di-ti - quefoce or Paag-ac -numa-pa-tiemmo-to	old Race and ome ori- a- Rural-gin Yea a- Urb Angere of on Conre- omfold djtin-codmende o uun (< 60),60-	YeadasnoR lofwsi-SizSu adleafth Suusy re-ca-ma <i>s</i> y	ur(u Hi s- ys permic ,/Sur eqNon-
Fe-2010VhNtonBreBstG50U8-2				75 ,0005 -All202		
	ngi knodvionac-per	-		edties79 races/		sSpanish-
White le-		mediates	3	•	lastdead te	•
sion of	hisgin im- tol-of plants	are	or	•		.e L atino
breast	tol-of plants ogypri- (in-	avail- able	more pri-	tan ties ar-	tacto ap	oy nd/or
Dicast	mary cludes	and	maries	eas	cer su	
	brachythe		maries	ge	gi-	
	apy)	are		1	ca	
	(1988+)	more		mil-	pr	
	(====1)	than		lion	ce	
		0		pop		ires
		days		1 1		
		of				
		sur-				
		vival				
Fe-201 A si N nonBre Bse (3 50U9 -2	2 PosRigNto/Nonoe/45um	BekanGok (00) 1/41 iNec	o3rd04 0 377Ma\$-	75 ,000,5 -All202	2A0lixAdi&OSy	vsNon-
male or Hispanic Breast, wn	ikndwnnownper	- plete		ties59 races/	'et her te	n Sp anish-
Pa-Asian NOS	tiveri- form	med lates	3 (in-	*	lastdead th	
cifior	hisgin	are		*		yLatino (
Is- Pa-	tolof	avail-	more ing	tan ties	tacto af	
lan diéi c	ogypri-	able	pri- com-		can- te	
Is-	mary	and	maries mon		cer su	rgery
lander		there	law)	_		
		are more		1 mil-		
		than		lion		
		0		pop		
		$_{ m days}$		ьоь		
		of				
		sur-				
		vival				

	_					То-То-					
	Race					tal tal					
	and					numum-					
	ori-					berber	di			onn.	
	gin Si			2.5		of of	Malne		_	SEEF	
	re- re			Months	ma-	in be-	i- ho		Race	other	Ori-
	code co		Di-	from Scope	lig-	situn/inguna/			and	cause	0
	¢N ISMĄC		ag-	di- of		lig-der-	sta€o		ori-	of	re-
	NHB, O			under de de la companya de la compan	pri-	nankine				ear dea Th	
		matGraCheR		no-sonlymph	COhary			,	- (oRXNHIA
				sis no nd Sur-		-mo rn ors					,
				onto canaergiva							
				trealired 1998 on							
Sexsis AP	фаn2008	onbel 260/1(72)0(13	80 4isoù ty no∕oo	de) e nt r 2002)gn	no cothe rbe	ertienttientt	$0 \sin 20$	2CO669	, 50±	dopodeion(2	0(1 26)1017 1s 1 p))
Fe-201 A si	iMonBreBa	st6350U4b-1	PosLefNo Nd	onoe/45un+3etanGk(6	M)1A4liNeo4ti	h04 0 37	7Ma\$7	5 (0605	-All20	240 lixAdiv4e	SysNon-
		Up ker wn		wwnper-plete			ried	ties59			ten Sp anish-
Pa-	-Asian	outer	tiveri-	formedate	s 4		(in-	in ye	a nis c-	lastdead	the Hispanic-
cifi	or	quad-	hisgin	are	or		clud-	metro	pioli-	condue	apyLatino
Is-	Pa-	rant	tol-of	avail	l- mo	ore	ing	\tan	ties	tacto	af-
lan	diefrc	of	ogypri-	able	pr	i-	com-	ar-		can-	ter
	Is-	breast	mary	and	ma	aries	mon	eas		cer	surgery
	lander			there	e		law)	ge			
				are				1			
				more	е			mil-			
				than	l			lion			
				0				pop			
				days	;						
				of							
				sur-							
				viva	l						
Fe-20 2W ł	NitomBreBa	stG5tU8+2 9	PosRigNto Nd	omoe/15/un=BearGilo(6	M)0A3liNeo2n	$d02 \ 0 \ 65$	6Ma\$7	5Ç06 66	-All20	2 0 li x di 36	SysNon-
male	Hispanio	:Ovenhapppin	ngikndwan	www.per-plete	e of		ried	ties84	races	/ether	ten Sp anish-
	White	le-	tiveri-	formed late	\sim 2		(in-	in ye	a mis c-	lastdead	$the {\bf H} is panic-$
		sion	hisgin	are	or		clud-	metro	pioli-	condue	apyLatino
		of	tol-of	avail	l- mo	ore	ing	\tan	ties	tacto	both
		breast	ogypri-	able	pr	i-	com-	ar-		can-	be-
			mary	and	ma	aries	mon	eas		cer	fore
				there	e		law)	ge			and
				are				1			af-
				more				mil-			ter
				than	l			lion			surgery
				0				pop			
				days	;						
				of							
				sur-							
				viva	l						

					To-To				
	Race				tal ta	l			
	and				nu m e	ım- Me-			
	ori-				berbe	er dian			
	gin Si	ite			First of of	Malnouse-		SEER	2
	re- re) -		Months	ma- in be	- i- hold	Race	other	Ori-
	code co	ode	Di-	from Scope	lig-situn/i	gna/bteat-in-	and	cause	$_{ m gin}$
R	a¢N ISWE	CD-	ag-	di- of	nant lig-de	~ ,	ori-	of	re-
	e-NHBe , Ó			noth-Reaeg	pri- nankir		l-ginYea		- code
	,	matyrache			Officery tu-tu				
			- 0	sis no nd Sur- to			~ (
				onto cancergivaSursi					
				- trealtired 1990 Penvilvad					
				nde)entr20024gmom					
		.,,							
				June 25 June Bergals (01) OA		77281a\$750000			
male	-	c Lovkenewn		nownper-plete	of		4 races/a		sysSpanish
	White	inner	tiveri-	formed ates	3	,			tem lik ispani
		quad-	hisgin	are	or	clud- metr	o p oli-c	ondue	theLatino
		rant	tol-of	avail-	more	ing tan	ties t	acto	apy
		of	ogypri-	able	pri-	com- ar-		can-	and/or
		breast	mary	and	maries	mon eas		cer	sur-
				there		law) ge			gi-
				are		1			cal
				more		mil-			pro-
				than		lion			ce-
				0		pop			dures
				days					
				of					
				sur-					
				vival					
Fe-201 VA	/hNtonBreB	stG5tW4 2 9	PosRigWesN	01000/25Jun=13eanGk (1910) 2A	liNao 2nd02 0	8406n\$7500 6	5-A112024	0liAdi40	SysNon-
nale		c Up kero wn	9	nownper- plete	of	mar- ties59			ten Sp anish
110110	White	outer	tiveri-	formedates	2		,		theHispanie
	***************************************	quad-	hisgin	are	or			ordue	apyLatino
		rant	tolof	avail-	more	Do- tan	-	acto	both
		of	ogypri-	able	pri-	mes- ar-	0105 0	can-	be-
		breast	mary	and	maries	tic eas		cer	fore
		bicast	шагу	there	manes	Part- ge		CCI	and
				are		ner 1			af-
				more		mil-			ter
				than		lion			surgery
				0		pop			
				days					
				of					
				sur-					
				vival					

Cases

What are the cases, and how many are there? There are 131,395 cases in the BREAST cancer list of 2019-2020. And There are 303557 in 2011-2015 dataset.

[&]quot;adding more exploratory data analysis (EDA) to understand the structure and distribution of variables in

your dataset. This could include summary statistics, histograms, scatter plots, or other visualizations."

By employing Exploratory Data Analysis (EDA) methods like summary statistics and graphical representations, we aim to reveal insights that will enhance our comprehension of breast cancer outcomes and therapeutic approaches. The dataset is rich with details, encompassing variables such as the patient's age at operation, operation year, count of positive axillary nodes detected, and survival status post-treatment.

https://medium.com/@navamisunil174/exploratory-data-analysis-of-breast-cancer-survival-prediction-dataset-c423e4137e38

Data collection

Describe the method of data collection. I used the SEER *STAT to collect the data and export it as a TXT to be able to import it to the R for analyses. How SEER collects the data is explained in the following page in summary:

- The SEER program collects cancer incidence data through a network of population-based cancer registries. These registries gather information on patient demographics, primary tumor site, tumor morphology, stage at diagnosis, and first course of treatment. They also follow up with patients for vital status.
- By law, these facilities are required to report new cancer cases to a central cancer registry, like a state cancer registry.
- The SEER program releases new research data annually, based on submissions from the previous year, and makes it available for public use through a data request process. This comprehensive approach ensures that the SEER database is a valuable resource for cancer research and surveillance.

https://training.seer.cancer.gov/registration/data/collection.html

Type of study

This will be an observational study, information is gathered for different patients and I will be evaluating the available data to present and evaluate.

"discussing potential limitations of observational studies, such as confounding variables and biases, and how you plan to address them in analysis."

What type of study is this (observational/experiment)?

Data Source

Data is collected from SEER program and I used SEER *STAT software to glean them in a format that can be used and imported as TXT to R (Surveillance, Epidemiology, and End Results Program 2023).

"providing additional details about the specific variables included in dataset and how they were collected"

If you collected the data, state self-collected. If not, provide a citation/link.

Dependent Variable

I am still looking into the data, it seems I will have a combination of both quantitative and qualitative data to work with. For example, while the number of tumors, and survival months are qualitative. Other like race, marital status, type of cancer are categorical. I am still looking to see if I can find any qualitative data.

Categorical features, such as 'Median household income ...' 'Marital Status,' 'Grade recode' 'laterality' and 'Radiatio recode' and so on are represented as objects (characters).

Integer data types (int64) are assigned to 'Patient ID,' 'Year of diagnosis,' 'total number of ...'.

The event indicator refers to the death and the time registered is either the time-to-event (when the individual eventually dies) or the time-to-censorship (the event is not observed), measured in months.

```
# Find unique values in each column
# Apply function to find unique values for each column
unique_values <- data.frame(unique = apply(BREAST_DF_surv, 2, function(x) length(unique(x))),colnames =
# Check for NULL values
any_null <- any(sapply(BREAST_DF_surv, is.null))
# Check for NA values
any_na <- any(sapply(BREAST_DF_surv, is.na))
# Check if there are any NULL or NA values
if (any_null || any_na) {
    print("The data frame contains NULL or NA values.")
} else {
    print("The data frame does not contain any NULL or NA values.")
}</pre>
```

[1] "The data frame does not contain any NULL or NA values."

```
has_na_character <- any(sapply(BREAST_DF_surv, function(x) any(x == "NA")))

if (has_na_character) {
   print("The data frame contains character values of 'NA'.")
} else {
   print("The data frame does not contain character values of 'NA'.")
}</pre>
```

[1] "The data frame does not contain character values of 'NA'."

Data tyding

Upon exploring the data, it seems data might have an empty column, in this data-based, the empty values are filled with "Blanks". Thus, in this section, I first explore if there is any column which is entirely empty, then will remove it and if there are others which have some empty values filled with "blancked" I will repalced them with "NA" whoch is handled better in dplyr and tydiverse.

```
# There are cells in the DF that contianes "Blank(s) which is literally NA, first I want to find if the
#look for columns with all "Blank(s)" values
Empty_column <- BREAST_DF_surv %>%
    dplyr::summarise(dplyr::across(everything(), ~all(. == "Blank(s)"))) %>%
    as.logical() %>%
    unlist()
```

```
# Get the names of columns with all cells containing "Blank(s)"
blank_column_names <- names( BREAST_DF_surv) [Empty_column]</pre>
# Print the column names with all cells containing "Blanks"
print(blank_column_names)
## [1] "Grade Clinical (2018+)"
## [2] "Grade Pathological (2018+)"
## [3] "Scope of reg lymph nd surg (1998-2002)"
## [4] "Tumor Size Summary (2016+)"
#remove those empty column from thr DF
BREAST_DF_surv <- BREAST_DF_surv[, !names(BREAST_DF_surv) %in% blank_column_names]
BREAST_DF_eval <- BREAST_DF_eval[, !names(BREAST_DF_eval) %in% blank_column_names]
#Then let's see if there is any cell in the remaining that migth still have "Blank(s)", if so repalce i
#This code first replaces all occurrences of "Blank(s)" with an empty string "", and then uses na_if()
BREAST DF surv <- BREAST DF surv %>%
  mutate if(is.character, ~ifelse(. == "Blank(s)", "", .)) %%  # For character columns
 mutate_if(is.numeric, ~ifelse(. == "", as.numeric(NA), .)) # For numeric columns
# Now, empty character cells are replaced with NA
BREAST_DF_surv <- BREAST_DF_surv %>%
 mutate_if(is.character, na_if, "")
#same to be done for eval dataset
BREAST_DF_eval <- BREAST_DF_eval %>%
 mutate_if(is.character, ~ifelse(. == "Blank(s)", "", .)) %%  # For character columns
  mutate_if(is.numeric, ~ifelse(. == "", as.numeric(NA), .)) # For numeric columns
# Now, empty character cells are replaced with NA
BREAST_DF_eval <- BREAST_DF_eval %>%
 mutate_if(is.character, na_if, "")
#Change characters to numerics
BREAST_DF_surv$`Months from diagnosis to treatment` <- as.numeric(BREAST_DF_surv$`Months from diagnosis
BREAST_DF_surv$`Survival months` <- as.numeric(BREAST_DF_surv$`Survival months`)
## Warning: NAs introduced by coercion
BREAST_DF_surv$`Total number of in situ/malignant tumors for patient` <-
 as.numeric(BREAST_DF_surv$`Total number of in situ/malignant tumors for patient`)
```

Warning: NAs introduced by coercion

```
BREAST_DF_surv$ Total number of benign/borderline tumors for patient` <-
as.numeric(BREAST_DF_surv$`Total number of benign/borderline tumors for patient`)

#Change the character to numeric in Eval dataset too

BREAST_DF_eval$ Months from diagnosis to treatment` <- as.numeric(BREAST_DF_eval$`Months from diagnosis

BREAST_DF_eval$ Survival months` <- as.numeric(BREAST_DF_eval$`Survival months`)

## Warning: NAs introduced by coercion

BREAST_DF_eval$ Total number of in situ/malignant tumors for patient` <- as.numeric(BREAST_DF_eval$`Total number of in situ/malignant tumors for patient`)

## Warning: NAs introduced by coercion

BREAST_DF_eval$ Total number of benign/borderline tumors for patient` <- as.numeric(BREAST_DF_eval$`Total number of benign/borderline tumors for patient`)

# View the structure of the data frame

# str(BREAST_DF_surv)

skimr: skim(BREAST_DF_surv)
```

Table 3: Data summary

Name Number of rows	BREAST_DF_surv 303557
Number of columns	32
Column type frequency:	
character	25
numeric	7
Group variables	None

Variable type: character

	n_miss-	com-					whites-
skim_variable	ing	$plete_rate$	\min	max	empty	n_unique	pace
Sex	0	1	6	6	0	1	0
Race recode (W, B, AI, API)	0	1	5	29	0	5	0
Race and origin recode (NHW, NHB,	0	1	18	42	0	6	0
NHAIAN, NHAPI, Hispanic)							
Site recode ICD-O-3/WHO 2008	0	1	6	6	0	1	0
Site recode ICD-O-3 2023 Revision	0	1	6	6	0	1	0
Primary Site - labeled	0	1	12	36	0	9	0
Grade Recode (thru 2017)	0	1	7	38	0	5	0
Diagnostic Confirmation	0	1	7	57	0	9	0
Laterality	0	1	24	53	0	5	0
Chemotherapy recode (yes, no/unk)	0	1	3	10	0	2	0
Radiation recode	0	1	12	53	0	8	0

	n_miss-	com-					whites-
skim_variable	ing	$plete_rate$	\min	max	empty	n _unique	pace
Reason no cancer-directed surgery	0	1	15	76	0	8	0
Survival months flag	0	1	61	73	0	5	0
COD to site recode	0	1	5	55	0	87	0
First malignant primary indicator	0	1	2	3	0	2	0
Sequence number	0	1	16	60	0	13	0
Marital status at diagnosis	0	1	7	30	0	7	0
Median household income inflation adj to	0	1	8	38	0	11	0
2021							
Rural-Urban Continuum Code	0	1	38	60	0	7	0
Age recode $(<60,60-69,70+)$	0	1	9	11	0	18	0
Race and origin (recommended by SEER)	0	1	21	21	0	1	0
Year of death recode	0	1	4	21	0	11	0
SEER other cause of death classification	0	1	16	55	0	4	0
RX Summ-Systemic/Sur Seq (2007+)	0	1	16	55	0	8	0
Origin recode NHIA (Hispanic, Non-Hisp)	0	1	23	27	0	2	0

Variable type: numeric

	n_miss	s- com-								
skim_variable	ing	plete_ra	te mean	sd	p0	p25	p50	p75	p100	hist
Year of diagnosis	0	1.00	2013.04	1.42	2011	2012	2013	2014	2015	
Months from diagnosis to treatment	15843	0.95	1.13	1.14	0	0	1	2	24	
Survival months	1290	1.00	74.22	29.88	0	62	78	97	119	
Total number of in situ/malignant tumors for patient	3	1.00	1.36	0.65	1	1	1	2	20	
Total number of benign/borderline tumors for patient	0	1.00	0.01	0.09	0	0	0	0	5	
Patient ID	0	1.00	32479919	167185241	73089	166249	923853890	65493530	6 562 3287′	749
Year of follow-up recode	0	1.00	2018.90	2.14	2011	2019	2020	2020	2020	

skimr::skim(BREAST_DF_eval)

Table 6: Data summary

Name Number of rows	BREAST_DF_eval
Number of columns	32
Column type frequency:	
character	25
numeric	7
Group variables	None

Variable type: character

	n_miss-	com-				7	whites-
skim_variable	ing	$plete_rate$	\min	max	empty	n _unique	pace
Sex	0	1	6	6	0	1	0
Race recode (W, B, AI, API)	0	1	5	29	0	5	0
Race and origin recode (NHW, NHB,	0	1	18	42	0	6	0
NHAIAN, NHAPI, Hispanic)							
Site recode ICD-O-3/WHO 2008	0	1	6	6	0	1	0
Site recode ICD-O-3 2023 Revision	0	1	6	6	0	1	0
Primary Site - labeled	0	1	12	36	0	9	0
Grade Recode (thru 2017)	0	1	7	7	0	1	0
Diagnostic Confirmation	0	1	7	57	0	9	0
Laterality	0	1	24	53	0	5	0
Chemotherapy recode (yes, no/unk)	0	1	3	10	0	2	0
Radiation recode	0	1	12	53	0	8	0
Reason no cancer-directed surgery	0	1	15	76	0	8	0
Survival months flag	0	1	61	73	0	5	0
COD to site recode	0	1	5	55	0	67	0
First malignant primary indicator	0	1	2	3	0	2	0
Sequence number	0	1	16	60	0	16	0
Marital status at diagnosis	0	1	7	30	0	7	0
Median household income inflation adj to	0	1	8	38	0	11	0
2021							
Rural-Urban Continuum Code	0	1	38	60	0	7	0
Age recode $(<60,60-69,70+)$	0	1	9	11	0	17	0
Race and origin (recommended by SEER)	0	1	21	21	0	1	0
Year of death recode	0	1	4	21	0	3	0
SEER other cause of death classification	0	1	16	55	0	4	0
RX Summ-Systemic/Sur Seq (2007+)	0	1	16	55	0	8	0
Origin recode NHIA (Hispanic, Non-Hisp)	0	1	23	27	0	2	0

Variable type: numeric

	n_miss	s- com-								
skim_variable	ing	plete_ra	te mean	sd	p0	p25	p50	p75	p100	hist
Year of diagnosis	0	1.00	2019.48	0.50	2019	2019	2019	2020	2020	
Months from diagnosis to treatment	6807	0.95	1.26	1.18	0	1	1	2	24	
Survival months	537	1.00	11.07	7.05	0	5	11	17	23	
Total number of in situ/malignant tumors for patient	11	1.00	1.31	0.62	1	1	1	1	50	
Total number of benign/borderline tumors for patient	0	1.00	0.01	0.09	0	0	0	0	2	
Patient ID	0	1.00	33137047	1982 03798	127350	168966	69 6 6734	40 6 9994	2 76 3289	421
Year of follow-up recode	0	1.00	2019.98	0.14	2019	2020	2020	2020	2020	

What is the response variable? Is it quantitative or qualitative?

Independent Variable(s)

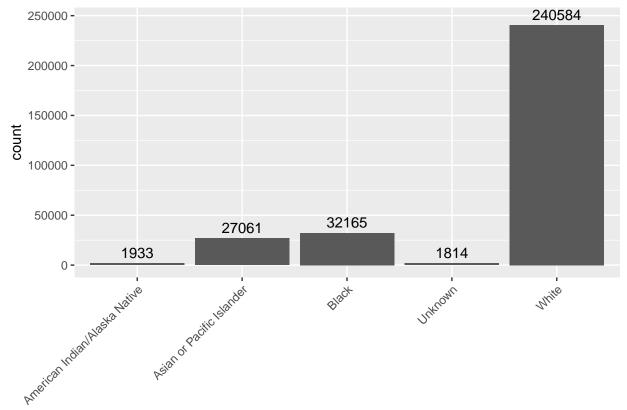
Relevant summary statistics

Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

```
#find column name to use later if needed
DF_col_names <- colnames(BREAST_DF_surv)

#Find unique values in `Race recode (W, B, AI, API)` column
uniques_races <- unique(BREAST_DF_surv$`Race recode (W, B, AI, API)`)

# use ggplot to plot the race information
BREAST_DF_surv |>
    ggplot(mapping = aes(x=`Race recode (W, B, AI, API)`)) +
    geom_bar(stat = "count") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    geom_text(aes(label = after_stat(count)), stat = "count", vjust = -0.5) +
    ylim(0, 246000)
```



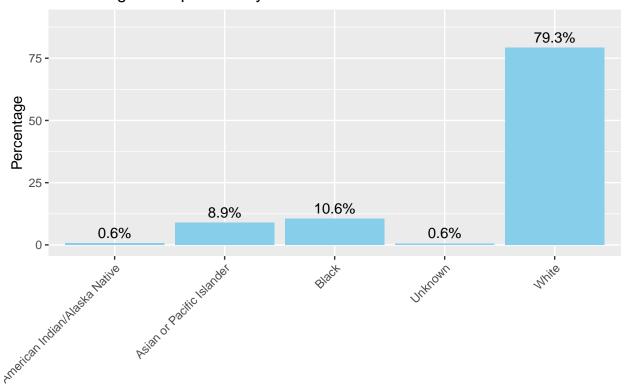
Race recode (W, B, AI, API)

#we want to coampre the percentage of the different race in the eval and survival data, thus i use sumam
#find percentage of race for the survival
BREAST_DF_perc_surv <- BREAST_DF_surv %>%

```
group_by(`Race recode (W, B, AI, API)`) %>%
    dplyr::summarise(count = dplyr::n()) %>%  # Calculate count per group
    ungroup() %>%  # Ungroup the data
    mutate(total_count = sum(count)) %>%  # Calculate total count
    mutate(percentage = count / total_count * 100)  # Calculate percentage using total count

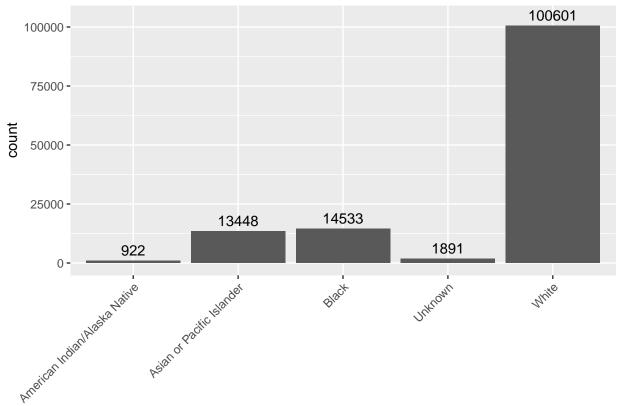
# Plot the percentages
ggplot(BREAST_DF_perc_surv, aes(x = `Race recode (W, B, AI, API)`, y = percentage)) +
    geom_bar(stat = "identity", fill = "skyblue") +
    geom_text(aes(label = paste0(round(percentage, 1), "%")), vjust = -0.5, color = "black") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    labs(title = "Percentage of Population by Race between 2011-2015", x = "Race recode (W, B, AI, API)",
```

Percentage of Population by Race between 2011–2015



Race recode (W, B, AI, API)

```
BREAST_DF_eval |>
    ggplot(mapping = aes(x=`Race recode (W, B, AI, API)`)) +
    geom_bar(stat = "count") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    geom_text(aes(label = after_stat(count)), stat = "count", vjust = -0.5) +
    ylim(0, 104000)
```

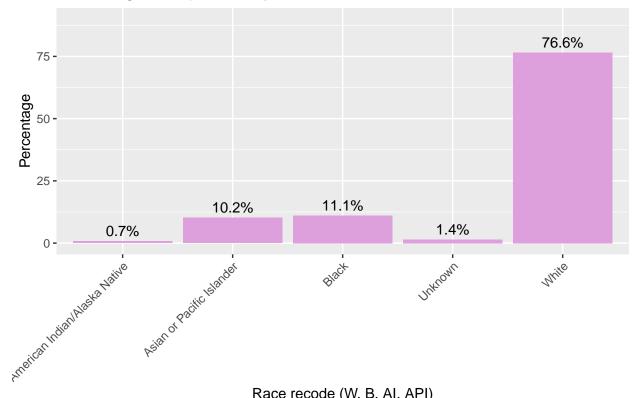


Race recode (W, B, AI, API)

```
BREAST_DF_perc_eval <- BREAST_DF_eval %>%
  group_by(`Race recode (W, B, AI, API)`) %>%
  dplyr::summarise(count = dplyr::n()) %>% # Calculate count per group
  ungroup() %>% # Ungroup the data
  mutate(total_count = sum(count)) %>% # Calculate total count
  mutate(percentage = count / total_count * 100) # Calculate percentage using total count

# Plot the percentages
ggplot(BREAST_DF_perc_eval, aes(x = `Race recode (W, B, AI, API)`, y = percentage)) +
  geom_bar(stat = "identity", fill = "plum") +
  geom_text(aes(label = paste0(round(percentage, 1), "%")), vjust = -0.5, color = "black") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Percentage of Population by between 2019-2022", x = "Race recode (W, B, AI, API)", y =
```

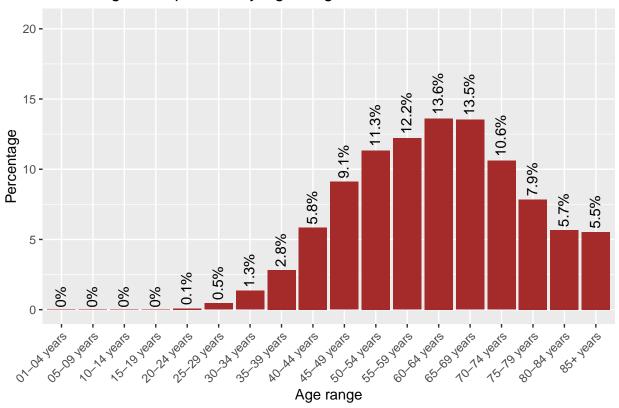
Percentage of Population by between 2019–2022



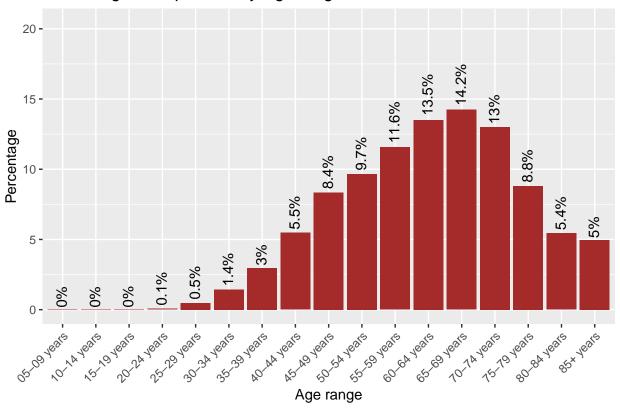
Race recode (W, B, AI, API)

```
# In this section I want to focus on the age and see if age matetrs, same sets of data is going to be p
#find percentage of race for the survival
#find ubique values for column ratted to age
uniques_ages <- unique(BREAST_DF_surv[29])</pre>
BREAST_DF_age_perc_surv <- BREAST_DF_surv %>%
  dplyr::group_by(`Age recode (<60,60-69,70+)`) %>%
  dplyr::summarise(count = dplyr::n()) %>% # Calculate count per group
  ungroup() %>% # Ungroup the data
  mutate(total_count = sum(count)) %>% # Calculate total count
  mutate(percentage = count / total_count * 100) # Calculate percentage using total count
perc_max <- max(BREAST_DF_age_perc_surv$percentage)</pre>
# Plot the percentages
ggplot(BREAST_DF_age_perc_surv, aes(x = `Age recode (<60,60-69,70+)`, y = percentage)) +
  geom_bar(stat = "identity", fill = "brown") +
  geom_text(aes(label = paste0(round(percentage, 1), "%")), hjust = -0.1 , vjust = 0.4, color = "black"
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +labs(title = "Percentage of Population by A
       x = "Age range",
       y = "Percentage") +
  ylim(0, round(1.5 * perc_max, 1))
```

Percentage of Population by Age range 2011–2015



Percentage of Population by Age range 2019–2022

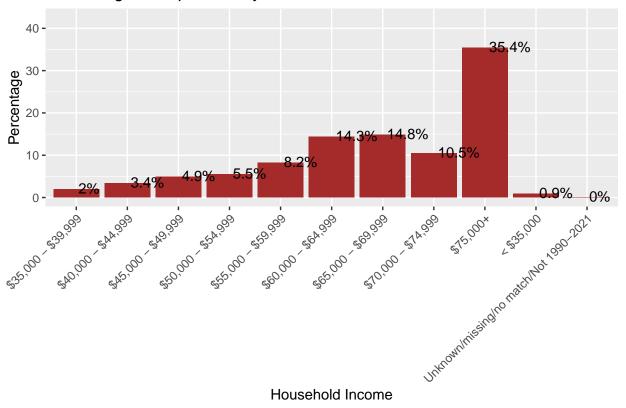


```
# In this section, we do the analyses on household income}
#find ubique values for column ratted to age
uniques_householdes <- unique(BREAST_DF_surv[27])

BREAST_DF_income_perc_surv <- BREAST_DF_surv %>% dplyr::group_by(`Median household income inflation adj
dplyr::summarise(count = dplyr::n()) %>% # Calculate count per group
ungroup() %>% # Ungroup the data
mutate(total_count = sum(count)) %>% # Calculate total count
mutate(percentage = count / total_count * 100) # Calculate percentage using total count

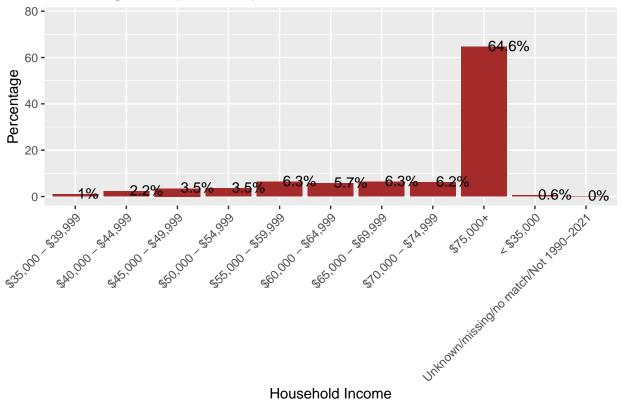
perc_max <- max(BREAST_DF_income_perc_surv$percentage) # Plot the percentages
ggplot(BREAST_DF_income_perc_surv, aes(x = `Median household income inflation adj to 2021`, y = percent
geom_bar(stat = "identity", fill = "brown") +
geom_text(aes(label = pasteo(round(percentage, 1), "%")), hjust = -0.1 , vjust = 0.4, color = "black"
theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
labs(title = "Percentage of Population by income 2011-2015", x = "Household Income", y = "Percentage"
ylim(0, 1.2*perc_max)
```

Percentage of Population by income 2011–2015



```
#In this section we do the same analyses for Eval data based on age
BREAST_DF_income_perc_eval <- BREAST_DF_eval %>%
  dplyr::group_by(`Median household income inflation adj to 2021`) %>%
  dplyr::summarise(count = dplyr::n()) %>% # Calculate count per group
  ungroup() %>% # Ungroup the data
  mutate(total_count = sum(count)) %>% # Calculate total count
  mutate(percentage = count / total_count * 100) # Calculate percentage using total count
#Plot the percentages
perc max <- max(BREAST DF income perc eval$percentage)</pre>
ggplot(BREAST_DF_income_perc_eval, aes(x = `Median household income inflation adj to 2021`, y = percent
  geom_bar(stat = "identity", fill = "brown") +
  geom_text(aes(label = paste0(round(percentage, 1), "%")), hjust = -0.1 , vjust = 0.4, color = "black"
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Percentage of Population by income 2019-2022", x = "Household Income", y = "Percentage"
 ylim(0, 1.2*perc_max)
```

Percentage of Population by income 2019–2022



In this section I want to focus on the cause of dead, COD, and investigate whether those who have had
#find percentage of deceased due to breast cancer
#find unique values for column ratted to age
uniques_CODs <- unique(BREAST_DF_surv[20])</pre>

[1] "Total number of in situ/malignant tumors for patient"

DF_col_names[20]

```
# check if the column `COD to site recode` has value of Alive or Breast meaning they are still alive or
BREAST_DF_surv <- BREAST_DF_surv %>%
    mutate(COD = ifelse(`COD to site recode` %in% c("Alive", "Breast"), `COD to site recode`, "Other"))

BREAST_DF_COD_perc_surv <- BREAST_DF_surv %>%
    dplyr::group_by(COD) %>%
    dplyr::summarise(count = dplyr::n()) %>% # Calculate count per group
    ungroup() %>% # Ungroup the data
    mutate(`Total Count` = sum(count)) %>% # Calculate total count
    mutate(Population = round(count / `Total Count` * 100),2) # Calculate percentage using total count
kable(BREAST_DF_COD_perc_surv)
```

opulation
75
13
12

Total number of in situ/malignant tumors for patient	Event Population	Popula- tion	Group % in total	Death %
1	27314	217122	71.53	12.58
2	8945	68082	22.43	13.14
3	1808	14579	4.80	12.40
4	322	2996	0.99	10.75
5	68	595	0.20	11.43
6	9	126	0.04	7.14
7	3	29	0.01	10.34
8	2	18	0.01	11.11
18	1	1	0.00	100.00

```
# Let' focus on the treatemnt, There are two type of treatment and can be a 4 combination ,as follows:
BREAST_DF_surv <- BREAST_DF_surv %>%
   mutate(Radiation = ifelse(`Radiation recode` %in% c("None/Unknown", "Refused (1988+)", "Recommended, und
#use DPLYR to filter based on two parameters chemotheraphy and radiation therapy and evaluate the death
BREAST_DF_RNC_perc_surv <- BREAST_DF_surv %>%
   dplyr::group_by(Radiation, `Chemotherapy recode (yes, no/unk)`) %>%
   dplyr::add_count() %>%
   filter(COD == "Breast") %>%
   dplyr::summarise(`Event Population` = n(),
        Population = dplyr::first(n)) # Use `first()` to extract the total count in each
```

^{## `}summarise()` has grouped output by 'Radiation'. You can override using the
`.groups` argument.

```
#knwoign the population calcualte the gorup rate and death rate in each group
BREAST_DF_RNC_perc_surv$`Group % in total` <- round(BREAST_DF_RNC_perc_surv$Population/sum(BREAST_DF_RNC
BREAST_DF_RNC_perc_surv$`Death %` <- round(BREAST_DF_RNC_perc_surv$`Event Population`/BREAST_DF_RNC_perc_surv$)</pre>
kable(BREAST_DF_RNC_perc_surv)
```

Radiation	Chemotherapy recode (yes, no/unk)	Event Population	Popula- tion	Group % in total	Death %
No/Un- known	No/Unknown	15684	107012	35.25	14.66
No/Un- known	Yes	9929	54966	18.11	18.06
Yes Yes	No/Unknown Yes	3731 9128	79926 61653	26.33 20.31	4.67 14.81

#next let's look into the surgery and the survival rate and whether it migth have been critical or not.

Results of the exploratory data analysis

In this section, we look into some exploratory data analysis such as

- Cause of death of those who have had cancer
- Total number of tumors (Malignant or Benign)
- Radiation and chemotherapy
- Marital Status

We looked into the population and then among the population how many survived the cancer. Later we will run some analyses to see whether those were important or deciding factors or not.

Surveillance, Epidemiology, and End Results Program. 2023. "SEER*stat Database: Incidence - SEER Research Data, 8 Registries, Nov 2021 Sub (1975-2020) - Linked to County Attributes - Time Dependent (1990-2020) Income/Rurality, 1969-2020 Counties." National Cancer Institute, DCCPS, Surveillance Research Program, released April 2023, based on the November 2022 submission. https://seer.cancer.gov/data/citation.html.