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.”

```r
# Replace "file.txt" with the path to your text file
directory <- "C:/Users/kohya/OneDrive/CUNY/DATA 606/DATA 606 Spring/Project"
file_2020 <- "BREAST_2019-2020-updated.csv"
file_serv <- "BREAST_2011-2015.csv"
# Complete the file path
full_path_serv <- file.path(directory, file_serv)
full_path_eval<- file.path(directory, file_2020)


BREAST_DF_surv <- read.csv(full_path_serv, header = TRUE,
                      na.strings = "NA", check.names = FALSE)
BREAST_DF_eval <- read.csv(full_path_eval, header = TRUE,
                      na.strings = "NA", check.names = FALSE)

labels_of_interest <- c("Primary Site - labeled")

# View the first few rows of the data frame
kable(head(BREAST_DF_surv, 10))
```

Sex	Year of diagnosis	Race recode (W, B, AI, API)	Race and origin recode (NHW, NHB, NHAIAN, NHAPI, Hispanic)	Site recode ICD-O-3/WHO 2008	Primary Site - labeled	Grade Recode (thru 2017)	Grade Clinical (2018+)	Diagnostic Confirmation	Laterality	Chemotherapy recode (yes, no/unk)	Radiation recode	Months from diagnosis to treatment	Reason no cancer-directed surgery	Scope of reg lymph nd surg (1998-2002)	Survival months flag	Survival months	COD to site recode	First malignant primary indicator	Sequence number	Total number of in situ/malignant tumors for patient	Total number of benign/borderline tumors for patient	Patient ID	Marital status at diagnosis	Median household income inflation adj to 2021	Rural-Urban Continuum Code	Age recode (<60,60-69,70+)	Race and origin (recommended by SEER)	Year of follow-up recode	Year of death recode	SEER other cause of death classification	Tumor Size Summary (2016+)	RX Summ--Systemic/Sur Seq (2007+)	Origin recode NHIA (Hispanic, Non-Hisp)
Female	2015	White	Non-Hispanic White	Breast	C50.4-Upper-outer quadrant of breast	Moderately differentiated; Grade II	Blank(s)	Positive histology	Right - origin of primary	Yes	Beam radiation	002	Surgery performed	Blank(s)	Complete dates are available and there are more than 0 days of survival	0060	Alive	No	2nd of 2 or more primaries	02	0	309	Married (including common law)	$75,000+	Counties in metropolitan areas ge 1 million pop	50-54 years	All races/ethnicities	2020	Alive at last contact	Alive or dead due to cancer	Blank(s)	Systemic therapy after surgery	Non-Spanish-Hispanic-Latino
Female	2013	White	Non-Hispanic White	Breast	C50.9-Breast, NOS	Unknown	Blank(s)	Positive histology	Right - origin of primary	No/Unknown	None/Unknown	Blank(s)	Not recommended	Blank(s)	Complete dates are available and there are more than 0 days of survival	0028	Breast	No	3rd of 3 or more primaries	03	0	346	Divorced	$75,000+	Counties in metropolitan areas ge 1 million pop	40-44 years	All races/ethnicities	2015	2015	Alive or dead due to cancer	Blank(s)	No systemic therapy and/or surgical procedures	Non-Spanish-Hispanic-Latino

Sex	Year of diagnosis	Race recode (W, B, AI, API)	Race and origin recode (NHW, NHB, NHAIAN, NHAPI, Hispanic)	Site recode ICD-O-3/WHO 2008	Site recode ICD-O-3 2023 Revision	Primary Site – labeled	Grade Recode (thru 2017)	Grade Clinical (2018+)	Grade Pathological (2018+)	Diagnostic Confirmation	Laterality	Chemotherapy recode (yes, no/unk)	Radiation recode	Months from diagnosis to treatment	Reason no cancer-directed surgery	Scope of reg lymph nd surg (1998-2002)	Survival months flag	Survival months	COD to site recode	First malignant primary indicator	Sequence number	Total number of in situ/malignant tumors for patient	Total number of benign/borderline tumors for patient	Patient ID	Marital status at diagnosis	Median household income inflation adj to 2021	Rural-Urban Continuum Code	Age recode (<60,60-69,70+)	Race and origin (recommended by SEER)	Year of follow-up recode	Year of death recode	SEER other cause of death classification	Tumor Size Summary (2016+)	RX Summ--Systemic/Sur Seq (2007+)	Origin recode NHIA (Hispanic, Non-Hisp)
Female	2012	White	Non-Hispanic White	Breast	Breast	C50.2-Upper-inner quadrant of breast	Moderately differentiated; Grade II	Blank(s)	Blank(s)	Positive histology	Right - origin of primary	No/Unknown	None/Unknown	004	Surgery performed	Blank(s)	Complete dates are available and there are more than 0 days of survival	0099	Alive	No	2nd of 2 or more primaries	03	0	374	Widowed	$75,000+	Counties in metropolitan areas ge 1 million pop	80-84 years	All races/ethnicities	2020	Alive at last contact	Alive or dead due to cancer	Blank(s)	Systemic therapy before surgery	Non-Spanish-Hispanic-Latino
Female	2014	White	Non-Hispanic White	Breast	Breast	C50.8-Overlapping lesion of breast	Moderately differentiated; Grade II	Blank(s)	Blank(s)	Positive histology	Right - origin of primary	No/Unknown	None/Unknown	001	Surgery performed	Blank(s)	Complete dates are available and there are more than 0 days of survival	0084	Alive	No	2nd of 2 or more primaries	02	0	391	Married (including common law)	$75,000+	Counties in metropolitan areas ge 1 million pop	55-59 years	All races/ethnicities	2020	Alive at last contact	Alive or dead due to cancer	Blank(s)	Systemic therapy after surgery	Non-Spanish-Hispanic-Latino

Sex	Year of diagnosis	Race recode (W, B, AI, API)	Race and origin recode (NHW, NHB, NHAIAN, NHAPI, Hispanic)	Site recode ICD-O-3/WHO 2008	Primary Site - labeled	Grade Recode (thru 2017)	Grade Clinical (2018+)	Grade Pathological (2018+)	Diagnostic Confirmation	Laterality	Chemotherapy recode (yes, no/unk)	Radiation recode	Months from diagnosis to treatment	Reason no cancer-directed surgery	Scope of reg lymph nd surg (1998-2002)	Survival months flag	Survival months	COD to site recode	First malignant primary indicator	Sequence number	Total number of in situ/malignant tumors for patient	Total number of benign/borderline tumors for patient	Patient ID	Marital status at diagnosis	Median household income inflation adj to 2021	Rural-Urban Continuum Code	Age recode (<60,60-69,70+)	Race and origin (recommended by SEER)	Year of follow-up recode	Year of death recode	SEER other cause of death classification	Tumor Size Summary (2016+)	RX Summ--Systemic/Sur Seq (2007+)	Origin recode NHIA (Hispanic, Non-Hisp)
Female	2013	Black	Non-Hispanic Black	Breast	C50.9-Breast, NOS	Unknown	Blank(s)	Blank(s)	Direct visualization without microscopic confirmation	Left - origin of primary	No/Unknown	None/Unknown	Blank(s)	Not recommended	Blank(s)	Complete dates are available and there are more than 0 days of survival	0010	Breast	No	2nd of 2 or more primaries	02	0	5477	Widowed	$75,000+	Counties in metropolitan areas ge 1 million pop	85+ years	All races/ethnicities	2012	2012	Alive or dead due to cancer	Blank(s)	No systemic therapy and/or surgical procedures	Non-Spanish-Hispanic-Latino
Female	2013	White	Hispanic (All Races)	Breast	C50.9-Breast, NOS	Moderately differentiated; Grade II	Blank(s)	Blank(s)	Positive histology	Right - origin of primary	No/Unknown	Beam radiation	001	Surgery performed	Blank(s)	Complete dates are available and there are more than 0 days of survival	0086	Alive	No	2nd of 2 or more primaries	02	0	5677	Married (including common law)	$75,000+	Counties in metropolitan areas ge 1 million pop	70-74 years	All races/ethnicities	2020	Alive at last contact	Alive or dead due to cancer	Blank(s)	No systemic therapy and/or surgical procedures	Spanish-Hispanic-Latino

Sex	Year of diagnosis	Race recode (W, B, AI, API)	Race and origin recode (NHW, NHB, NHAIAN, NHAPI, Hispanic)	Site recode ICD-O-3/WHO 2008	Site recode ICD-O-3 2023 Revision	Primary Site - labeled	Grade Recode (thru 2017)	Grade Clinical (2018+)	Grade Pathological (2018+)	Diagnostic Confirmation	Laterality	Chemotherapy recode (yes, no/unk)	Radiation recode	Months from diagnosis to treatment	Reason no cancer-directed surgery	Scope of reg lymph nd surg (1998-2002)	Survival months flag	Survival months	COD to site recode	First malignant primary indicator	Sequence number	Total number of in situ/malignant tumors for patient	Total number of benign/borderline tumors for patient	Patient ID	Marital status at diagnosis	Median household income inflation adj to 2021	Rural-Urban Continuum Code	Age recode (<60, 60-69, 70+)	Race and origin (recommended by SEER)	Year of follow-up recode	Year of death recode	SEER other cause of death classification	Tumor Size Summary (2016+)	RX Summ--Systemic/Sur Seq (2007+)	Origin recode NHIA (Hispanic, Non-Hisp)
Female	2015	White	Non-Hispanic White	Breast	Breast	C50.8-Overlapping lesion of breast	Unknown	Blank(s)	Blank(s)	Positive histology	Left - origin of primary	Yes	None/Unknown	001	Not recommended	Blank(s)	Complete dates are available and there are more than 0 days of survival	0017	Breast	No	2nd of 2 or more primaries	02	0	760	Widowed	$75,000+	Counties in metropolitan areas ge 1 million pop	75-79 years	All races/ethnicities	201[illegible]	2016	Alive or dead due to cancer	Blank(s)	No systemic therapy and/or surgical procedures	Non-Spanish-Hispanic-Latino
Female	2015	White	Hispanic (All Races)	Breast	Breast	C50.4-Upper-outer quadrant of breast	Poorly differentiated; Grade III	Blank(s)	Blank(s)	Positive histology	Right - origin of primary	No/Unknown	None/Unknown	001	Surgery performed	Blank(s)	Complete dates are available and there are more than 0 days of survival	0007	Other Cause of Death	No	2nd of 2 or more primaries	02	0	941	Widowed	$75,000+	Counties in metropolitan areas ge 1 million pop	85+ years	All races/ethnicities	201[illegible]	2015	Dead (attributable to causes other than this cancer dx)	Blank(s)	No systemic therapy and/or surgical procedures	Spanish-Hispanic-Latino

Sex	Year of diagnosis	Race recode (W, B, AI, API)	Race and origin recode (NHW, NHB, NHAIAN, NHAPI, Hispanic)	Site recode ICD-O-3/WHO 2008	Site recode ICD-O-3 2023 Revision	Primary Site - labeled	Grade Recode (thru 2017)	Grade Clinical (2018+)	Grade Pathological (2018+)	Diagnostic Confirmation	Laterality	Chemotherapy recode (yes, no/unk)	Radiation recode	Months from diagnosis to treatment	Reason no cancer-directed surgery	Scope of reg lymph nd surg (1998-2002)	Survival months	Survival months flag	COD to site recode	First malignant primary indicator	Sequence number	Total number of in situ/malignant tumors for patient	Total number of benign/borderline tumors for patient	Patient ID	Marital status at diagnosis	Median household income inflation adj to 2021	Rural-Urban Continuum Code	Age recode (<60, 60-69, 70+)	Race and origin (recommended by SEER)	Year of follow-up recode	Year of death recode	SEER other cause of death classification	Tumor Size Summary (2016+)	RX Summ--Systemic/Sur Seq (2007+)	Origin recode NHIA (Hispanic, Non-Hisp)
Female	2015	White	Non-Hispanic White	Breast	Breast	C50.9-Breast, NOS	Poorly differentiated; Grade III	Blank(s)	Blank(s)	Positive histology	Right - origin of primary	No/Unknown	Beam radiation	001	Surgery performed	Blank(s)	0043	Complete dates are available and there are more than 0 days of survival	Cerebrovascular Diseases	No	2nd of 2 or more primaries	02	0	2056	Widowed	$75,000+	Counties in metropolitan areas ge 1 million pop	80-84 years	All races/ethnicities	2019	2019	Dead (attributable to causes other than this cancer dx)	Blank(s)	Systemic therapy after surgery	Non-Spanish-Hispanic-Latino
Female	2015	Black	Non-Hispanic Black	Breast	Breast	C50.8-Overlapping lesion of breast	Poorly differentiated; Grade III	Blank(s)	Blank(s)	Positive histology	Right - origin of primary	No/Unknown	None/Unknown	001	Surgery performed	Blank(s)	0070	Complete dates are available and there are more than 0 days of survival	Alive	No	3rd of 3 or more primaries	04	0	2605	Divorced	$75,000+	Counties in metropolitan areas ge 1 million pop	60-64 years	All races/ethnicities	2020	Alive at last contact	Alive or dead due to cancer	Blank(s)	No systemic therapy and/or surgical procedures	Non-Spanish-Hispanic-Latino

```r
kable(head(BREAST_DF_eval, 10))
```

Sex	Year of diagnosis	Race recode (W, B, AI, API)	Race and origin recode (NHW, NHB, NHAIAN, NHAPI, Hispanic)	Site recode ICD-O-3/WHO 2008	Primary Site - labeled	Grade Recode (thru 2017)	Grade Clinical (2018+)	Grade Pathological (2018+)	Diagnostic Confirmation	Laterality	Chemotherapy recode (yes, no/unk)	Radiation recode	Months from diagnosis to treatment	Reason no cancer-directed surgery	Scope of reg lymph nd surg (1998-2002)	Survival months flag	Survival months	COD to site recode	First malignant primary indicator	Sequence number	Total number of in situ/malignant tumors for patient	Total number of benign/borderline tumors for patient	Patient ID	Marital status at diagnosis	Median household income inflation adj to 2021	Rural-Urban Continuum Code	Age recode (<60,60-69,70+)	Race and origin (recommended by SEER)	Year of follow-up recode	Year of death recode	SEER other cause of death classification	Tumor Size Summary (2016+)	RX Summ--Systemic/Sur Seq (2007+)	Origin recode NHIA (Hispanic, Non-Hisp)
Female	2019	Asian or Pacific Islander	Non-Hispanic Asian or Pacific Islander	Breast	C50.8-Overlapping lesion of breast	Unknown	1	1	Positive histology	Right - origin of primary	No/Unknown	None/Unknown	002	Surgery performed	Blank(s)	Complete dates are available and there are more than 0 days of survival	0019	Alive	No	2nd of 2 or more primaries	02	0	2750	Divorced	$75,000+	Counties in metropolitan areas ge 1 million pop	65-69 years	All races/ethnicities	2020	Alive at last contact	Alive or dead due to cancer	8	Systemic therapy after surgery	Non-Spanish-Hispanic-Latino
Female	2020	Asian or Pacific Islander	Non-Hispanic Asian or Pacific Islander	Breast	C50.8-Overlapping lesion of breast	Unknown	2	9	Positive histology	Right - origin of primary	No/Unknown	None/Unknown	000	Recommended, unknown if performed	Blank(s)	Complete dates are available and there are more than 0 days of survival	0000	Alive	No	2nd of 2 or more primaries	02	0	2870	Married (including common law)	$75,000+	Counties in metropolitan areas ge 1 million pop	75-79 years	All races/ethnicities	2020	Alive at last contact	Alive or dead due to cancer	50	No systemic therapy and/or surgical procedures	Non-Spanish-Hispanic-Latino

Sex	Year of diagnosis	Race recode (W, B, AI, API)	Race and origin recode (NHW, NHB, NHAIAN, NHAPI, Hispanic)	Site recode ICD-O-3/WHO 2008	Site recode ICD-O-3 2023 Revision	Primary Site - labeled	Grade Recode (thru 2017)	Grade Clinical (2018+)	Grade Pathological (2018+)	Diagnostic Confirmation	Laterality	Chemotherapy recode (yes, no/unk)	Radiation recode	Months from diagnosis to treatment	Reason no cancer-directed surgery	Scope of reg lymph nd surg (1998-2002)	Survival months flag	Survival months	COD to site recode	First malignant primary indicator	Sequence number	Total number of in situ/malignant tumors for patient	Total number of benign/borderline tumors for patient	Patient ID	Marital status at diagnosis	Median household income inflation adj to 2021	Rural-Urban Continuum Code	Age recode (<60, 60-69, 70+)	Race and origin recode (SEER)	Year of follow-up recode	Year of death recode	SEER other cause of death classification	Tumor Size Summary (2016+)	RX Summ--Systemic/Sur Seq (2007+)	Origin recode NHIA (Hispanic, Non-Hisp)
Female	2020	White	Non-Hispanic White	Breast	Breast	C50.4-Upper-outer quadrant of breast	Unknown	1	2	Positive histology	Right - origin of primary	No/Unknown	None/Unknown	000	Surgery performed	Blank(s)	Complete dates are available and there are more than 0 days of survival	0007	Alive	No	2nd of 2 or more primaries	02	0	3067	Divorced	$75,000+	Counties in metropolitan areas ge 1 million pop	70+ years	All races/ethnicities	2020	Alive at last contact	Alive or dead due to cancer	18	No systemic therapy and/or surgical procedures	Non-Spanish-Hispanic-Latino
Female	2020	White	Non-Hispanic White	Breast	Breast	C50.5-Lower-outer quadrant of breast	Unknown	2	9	Positive histology	Right - origin of primary	Yes	None/Unknown	001	Surgery performed	Blank(s)	Complete dates are available and there are more than 0 days of survival	0010	Alive	No	2nd of 2 or more primaries	02	0	3365	Widowed	$75,000+	Counties in metropolitan areas ge 1 million pop	70+ years	All races/ethnicities	2020	Alive at last contact	Alive or dead due to cancer	60	Systemic therapy both before and after surgery	Non-Spanish-Hispanic-Latino

Sex	Year of diagnosis	Race recode (W, B, AI, API)	Race and origin recode (NHW, NHB, NHAIAN, NHAPI, Hispanic)	Site recode ICD-O-3/WHO 2008	Primary Site - labeled	Grade Recode (thru 2017)	Grade Clinical (2018+)	Grade Pathological (2018+)	Diagnostic Confirmation	Laterality	Chemotherapy recode (yes, no/unk)	Radiation recode	Months from diagnosis to treatment	Reason no cancer-directed surgery	Scope of reg lymph nd surg (1998-2002)	Survival months flag	Survival months	COD to site recode	First malignant primary indicator	Sequence number	Total number of in situ/malignant tumors for patient	Total number of benign/borderline tumors for patient	Patient ID	Marital status at diagnosis	Median household income inflation adj to 2021	Rural-Urban Continuum Code	Age recode (<60, 60-69, 70+)	Race and origin (recommended by SEER)	Year of follow-up recode	Year of death recode	SEER other cause of death classification	Tumor Size Summary (2016+)	RX Summ--Systemic/Sur Seq (2007+)	Origin recode NHIA (Hispanic, Non-Hisp)
Female	2019	White	Non-Hispanic White	Breast	C50.8-Overlapping lesion of breast	Unknown	2	2	Positive histology	Right - origin of primary	No/Unknown	Radioactive implants (includes brachytherapy) (1988+)	000	Surgery performed	Blank(s)	Complete dates are available and there are more than 0 days of survival	0016	Alive	No	3rd of 3 or more primaries	03	0	36779	Divorced	$75,000+	Counties in metropolitan areas ge 1 million pop	75-79 years	All races/ethnicities	2020	Alive at last contact	Alive or dead due to cancer	10	No systemic therapy and/or surgical procedures	Non-Spanish-Hispanic-Latino
Female	2019	Asian or Pacific Islander	Non-Hispanic Asian or Pacific Islander	Breast	C50.9-Breast, NOS	Unknown	2	2	Positive histology	Right - origin of primary	No/Unknown	None/Unknown	004	Surgery performed	Blank(s)	Complete dates are available and there are more than 0 days of survival	0014	Alive	No	3rd of 3 or more primaries	04	0	37771	Married (including common law)	$75,000+	Counties in metropolitan areas ge 1 million pop	55-59 years	All races/ethnicities	2020	Alive at last contact	Alive or dead due to cancer	30	Systemic therapy after surgery	Non-Spanish-Hispanic-Latino

Sex	Year of diagnosis	Race recode (W, B, AI, API)	Race and origin recode (NHW, NHB, NHAIAN, NHAPI, Hispanic)	Site recode ICD-O-3/WHO 2008	Primary Site - labeled	Grade Recode (thru 2017)	Grade Clinical (2018+)	Grade Pathological (2018+)	Diagnostic Confirmation	Laterality	Chemotherapy recode (yes, no/unk)	Radiation recode	Months from diagnosis to treatment	Reason no cancer-directed surgery	Scope of reg lymph nd surg (1998-2002)	Survival months flag	Survival months	COD to site recode	First malignant primary indicator	Sequence number	Total number of in situ/malignant tumors for patient	Total number of benign/borderline tumors for patient	Patient ID	Marital status at diagnosis	Median household income inflation adj to 2020	Rural-Urban Continuum Code	Age recode (<60, 60-69, 70+)	Race and origin (recommended by SEER)	Year of follow-up recode	Year of death recode	SEER other cause of death classification	Tumor Size Summary (2016+)	RX Summ--Systemic/Sur Seq (2007+)	Origin recode NHIA (Hispanic, Non-Hisp)
Female	2019	Asian or Pacific Islander	Non-Hispanic Asian or Pacific Islander	Breast	C50.4-Upper-outer quadrant of breast	Unknown	1	1	Positive histology	Left - origin of primary	No/Unknown	None/Unknown	004	Surgery performed	Blank(s)	Complete dates are available and there are more than 0 days of survival	0014	Alive	No	4th of 4 or more primaries	04	0	3771	Married (including common law)	$75,000+	Counties in metropolitan areas ge 1 million pop	55-59 years	All races/ethnicities	2020	Alive at last contact	Alive or dead due to cancer	4	Systemic therapy after surgery	Non-Spanish-Hispanic-Latino
Female	2020	White	Non-Hispanic White	Breast	C50.8-Overlapping lesion of breast	Unknown	2	9	Positive histology	Right - origin of primary	No/Unknown	None/Unknown	001	Surgery performed	Blank(s)	Complete dates are available and there are more than 0 days of survival	0003	Alive	No	2nd of 2 or more primaries	02	0	6501	Married (including common law)	$75,000+	Counties in metropolitan areas ge 1 million pop	80-84 years	All races/ethnicities	2020	Alive at last contact	Alive or dead due to cancer	36	Systemic therapy both before and after surgery	Non-Spanish-Hispanic-Latino

Sex	Year of diagnosis	Race recode (W, B, AI, API)	Race and origin recode (NHW, NHB, NHAIAN, NHAPI, Hispanic)	Site recode ICD-O-3/WHO 2008	Primary Site - labeled	Grade Recode (thru 2017)	Grade Clinical (2018+)	Grade Pathological (2018+)	Diagnostic Confirmation	Laterality	Chemotherapy recode (yes, no/unk)	Radiation recode	Months from diagnosis to treatment	Reason no cancer-directed surgery	Scope of reg lymph nd surg (1998-2002)	Survival months	Survival months flag	COD to site recode	First malignant primary indicator	Sequence number	Total number of in situ/malignant tumors for patient	Total number of benign/borderline tumors for patient	Patient ID	Marital status at diagnosis	Median household income inflation adj to 2020	Rural-Urban Continuum Code	Age recode (<60, 60-69, 70+)	Race and origin recommended by SEER	Year of follow-up recode	Year of death recode	SEER other cause of death classification	Tumor Size Summary (2016+)	RX Summ--Systemic/Sur Seq (2007+)	Origin recode NHIA (Hispanic, Non-Hisp)
Female	2020	White	Non-Hispanic White	Breast	C50.3-Lower-inner quadrant of breast	Unknown	1	1	Positive histology	Left - origin of primary	No/Unknown	None/Unknown	002	Surgery performed	Blank(s)	0007	Complete dates are available and there are more than 0 days of survival	Alive	No	3rd of 3 or more primaries	03	0	7723	Married (including common law)	$75,000+	Counties in metropolitan areas ge 1 million pop	70-74 years	All races/ethnicities	2020	Alive at last contact	Alive or dead due to cancer	6	No systemic therapy and/or surgical procedures	Non-Spanish-Hispanic-Latino
Female	2019	White	Non-Hispanic White	Breast	C50.4-Upper-outer quadrant of breast	Unknown	2	9	Positive histology	Right - origin of primary	Yes	None/Unknown	002	Surgery performed	Blank(s)	0021	Complete dates are available and there are more than 0 days of survival	Alive	No	2nd of 2 or more primaries	02	0	8406	Unmarried or Domestic Partner	$75,000+	Counties in metropolitan areas ge 1 million pop	55-59 years	All races/ethnicities	2020	Alive at last contact	Alive or dead due to cancer	19	Systemic therapy both before and after surgery	Non-Spanish-Hispanic-Latino

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.

“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.")
}
```

```
## [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'.")
}
```

```
## [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()
```

```r
# Get the names of columns with all cells containing "Blank(s)"
blank_column_names <- names( BREAST_DF_surv)[Empty_column]

# 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+)"
```

```r
#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
```

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

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
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, NHAIAN, NHAPI, Hispanic)	0	1	18	42	0	6	0
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

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
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 2021	0	1	8	38	0	11	0
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

skim_variable	n_missing	complete_rate	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	32479919.67	17852417.08	309	16624928	35389654	49353656	63287749	
Year of follow-up recode	0	1.00	2018.90	2.14	2011	2019	2020	2020	2020	

```r
skimr::skim(BREAST_DF_eval)
```

Table 6: Data summary

Name	BREAST_DF_eval
Number of rows	131395
Number of columns	32
Column type frequency:	
character	25
numeric	7
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
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, NHAIAN, NHAPI, Hispanic)	0	1	18	42	0	6	0
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 2021	0	1	8	38	0	11	0
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

skim_variable	n_missing	complete_rate	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	33137047.98	20379812.73	750	16896696	36734404	49994276	63289421	
Year of follow-up recode	0	1.00	2019.98	0.14	2019	2020	2020	2020	2020	

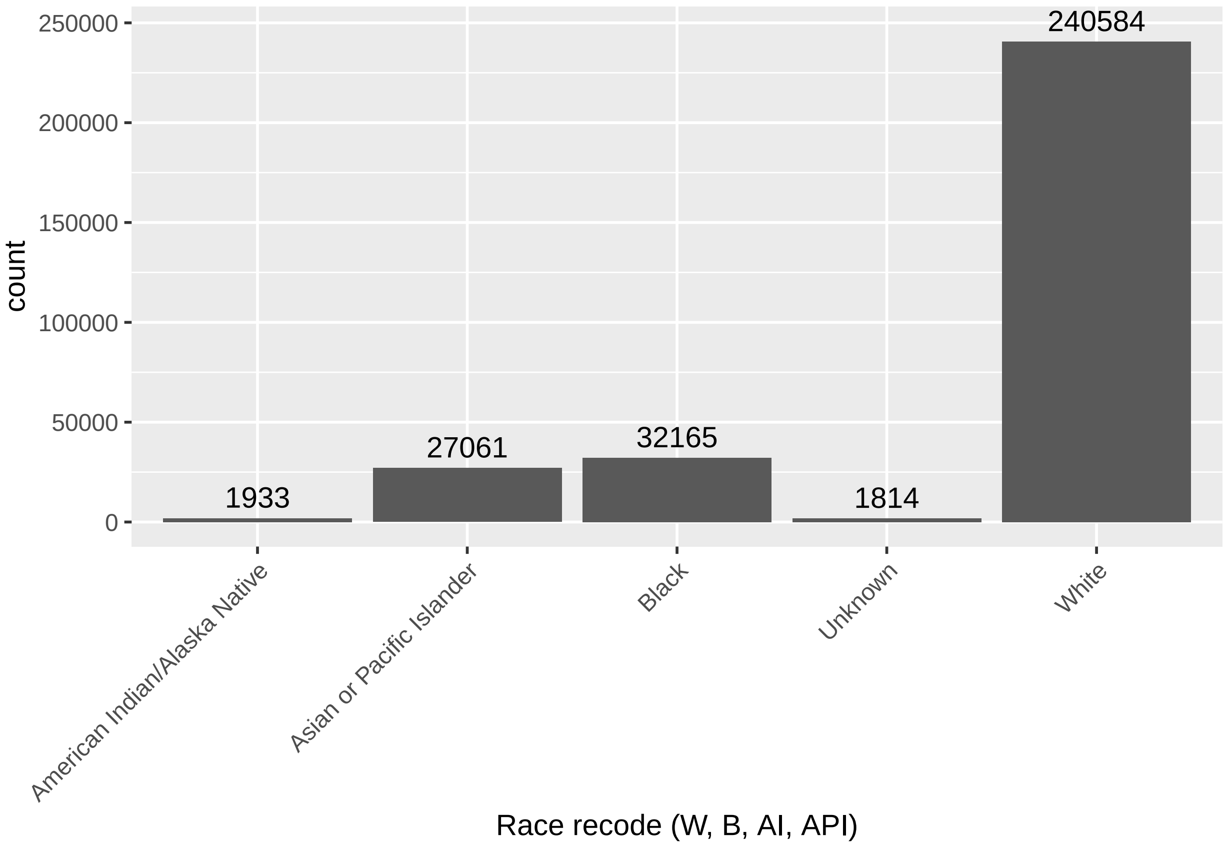
What is the response variable? Is it quantitative or qualitative?

Independent Variable(s)

Relevant summary statistics

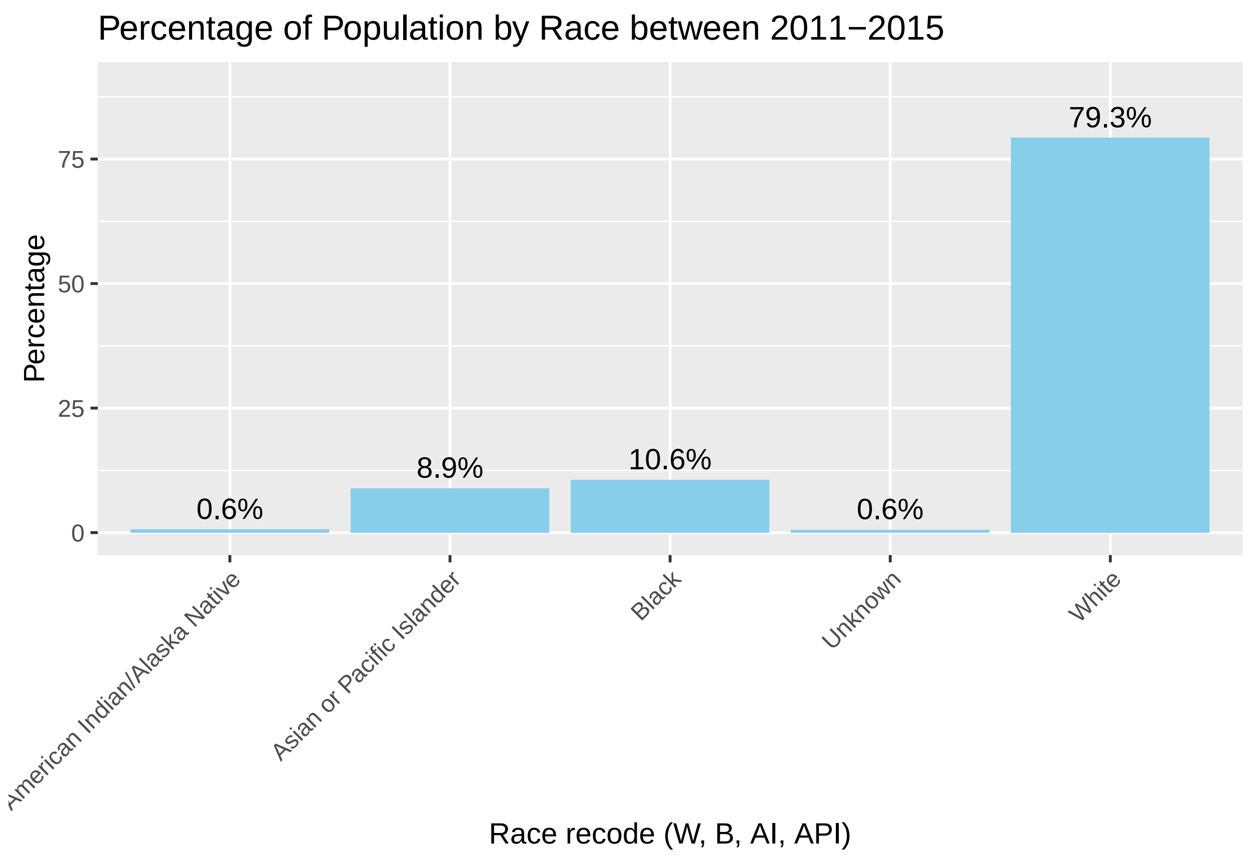
Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

```
#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)
```

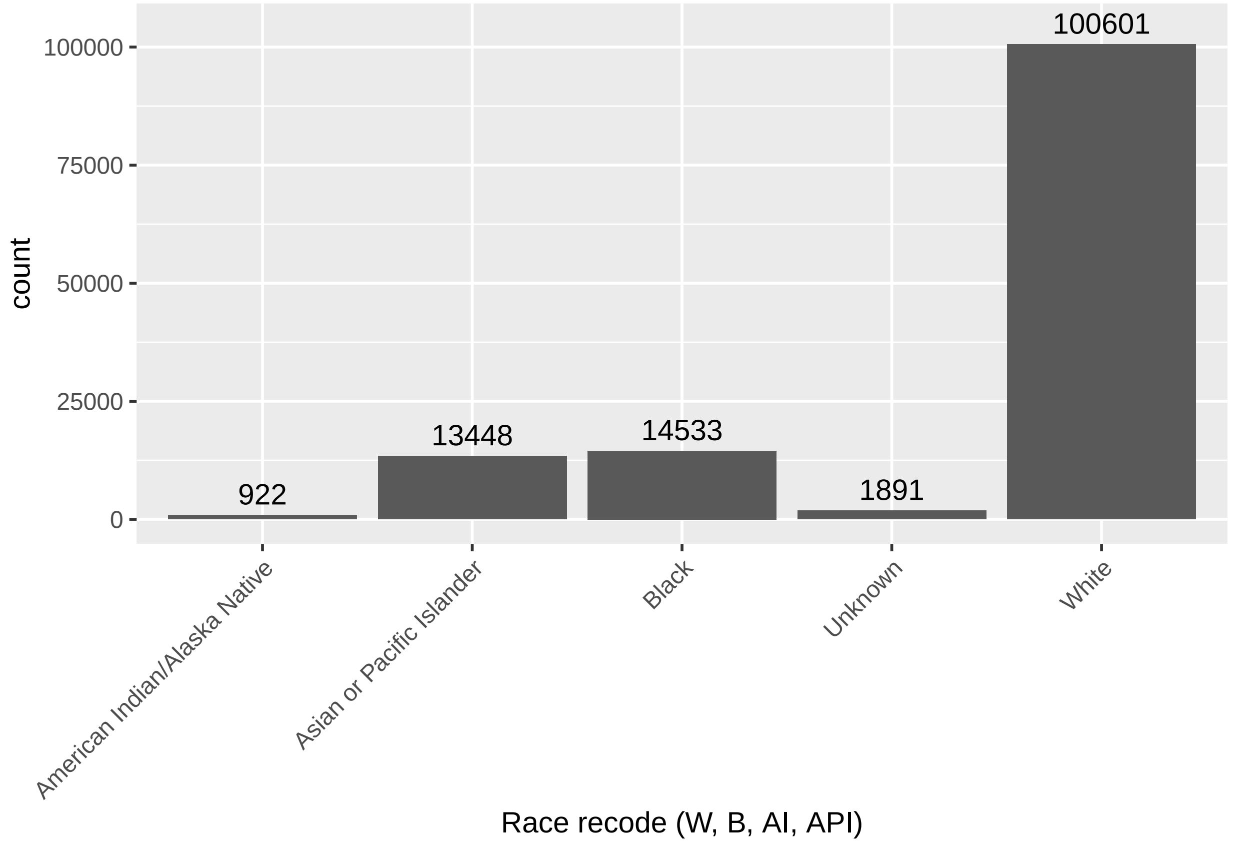


```
#we want to coampre the percentage of the diferent race in the eval and survival data, thus i use sumam
#find percentage of race for the survival
BREAST_DF_perc_surv <- BREAST_DF_surv %>%
```

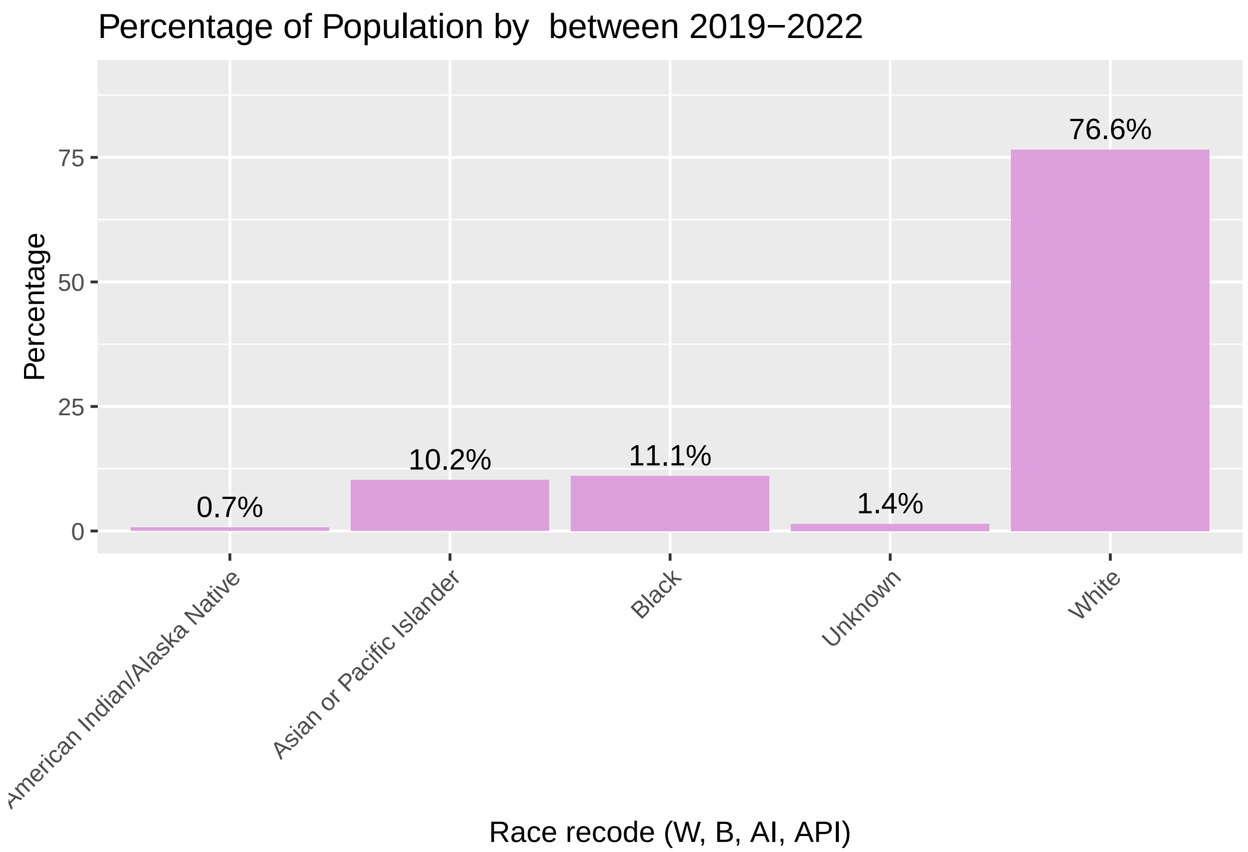
```r
  group_by(`Race recode (W, B, AI, API)`) %>%
  dplyr::summarise(count = dplyr::n()) %>%  # Calculate count per group
  ungroup() %>%  # Ungroup the data
  mutate(total_count = sum(count)) %>%  # Calculate total count
  mutate(percentage = count / total_count * 100)  # Calculate percentage using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



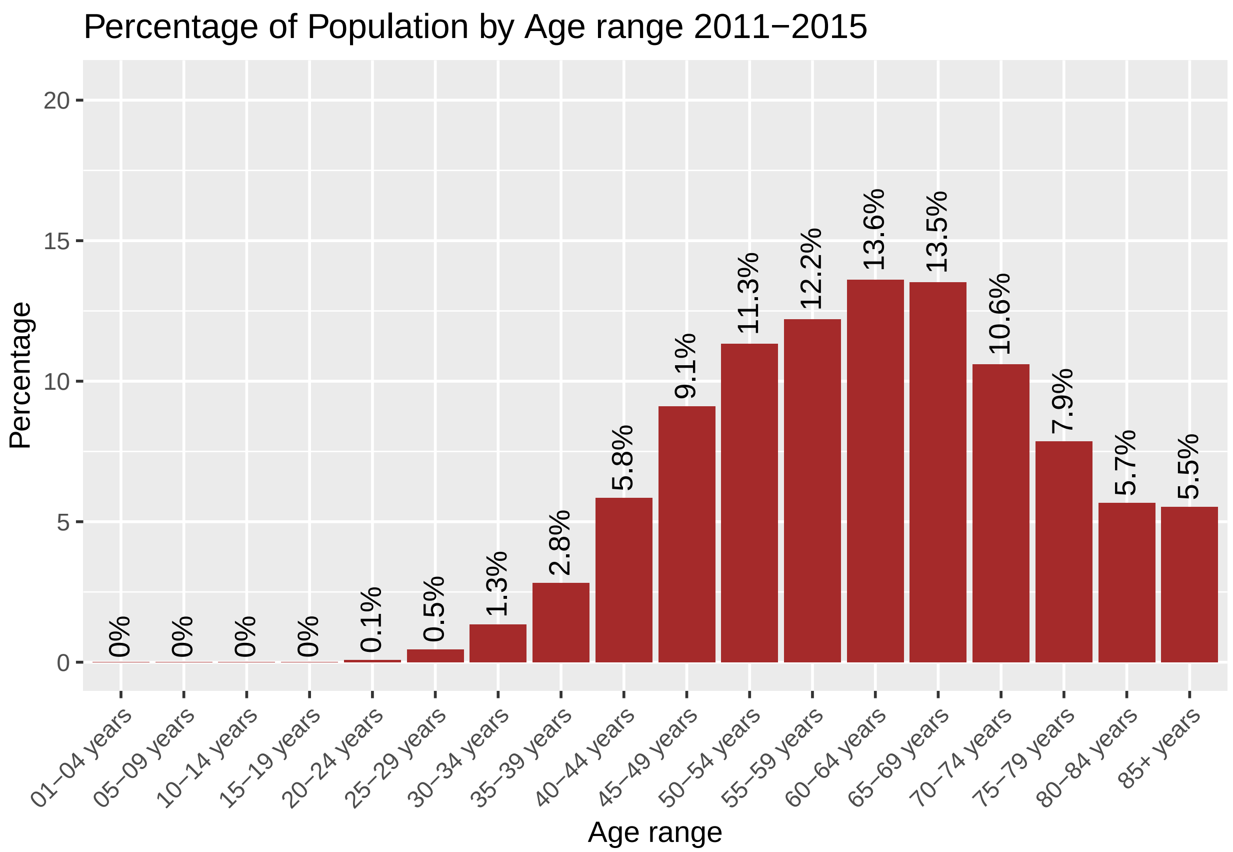
```r
BREAST_DF_eval |>
  ggplot(mapping = aes(x=`Race recode (W, B, AI, API)`)) +
  geom_bar(stat = "count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  geom_text(aes(label = after_stat(count)), stat = "count", vjust = -0.5) +
  ylim(0, 104000)
```



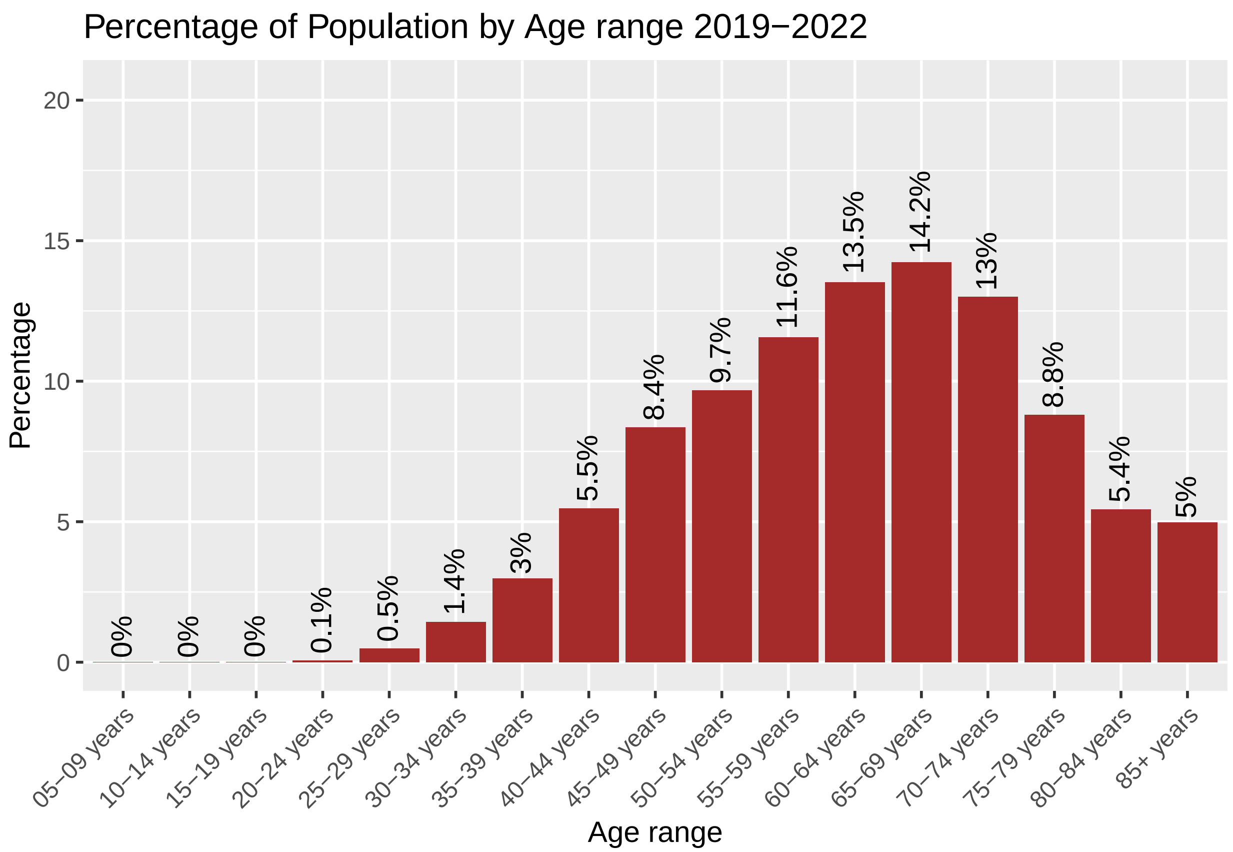
```
BREAST_DF_perc_eval <- BREAST_DF_eval %>%
  group_by(`Race recode (W, B, AI, API)`) %>%
  dplyr::summarise(count = dplyr::n()) %>%  # Calculate count per group
  ungroup() %>%  # Ungroup the data
  mutate(total_count = sum(count)) %>%  # Calculate total count
  mutate(percentage = count / total_count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



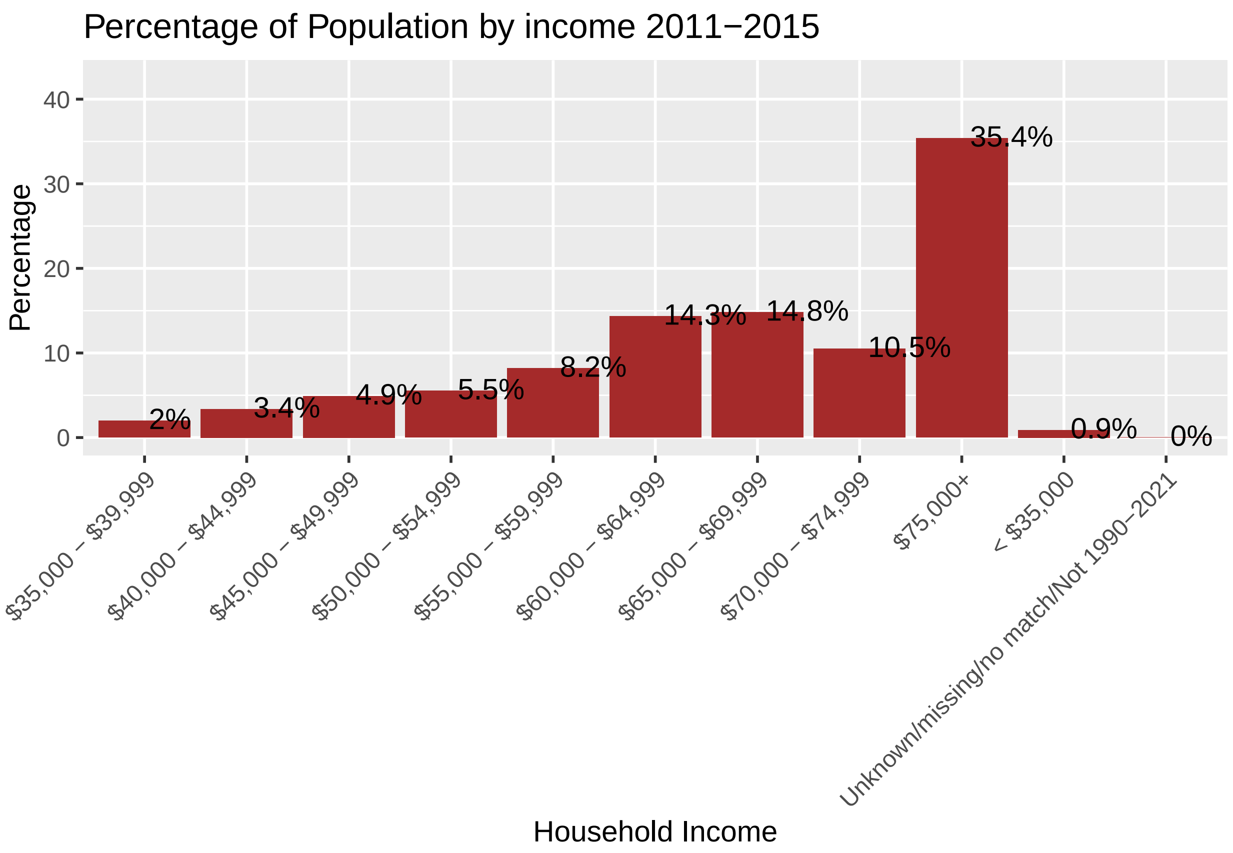
```
# In this section I want to focus on the age and see if age matetrs, same sets of data is going to be p
#find percentage of race for the survival
#find ubique values for column ratted to age
uniques_ages <- unique(BREAST_DF_surv[29])

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)
# 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 Ag
       x = "Age range",
       y = "Percentage") +
  ylim(0, round(1.5 * perc_max, 1))
```



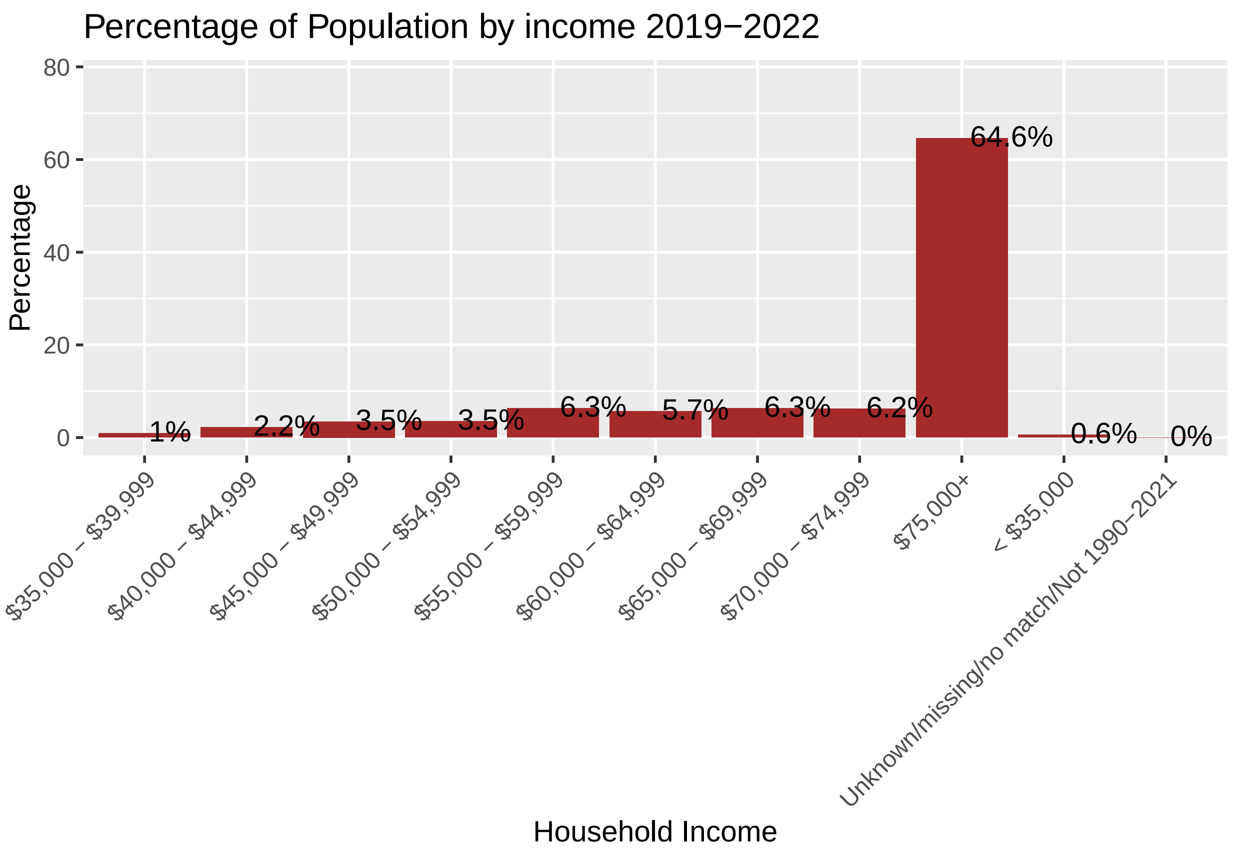
```
# In this section we do the same analyses for Eval dta based on age
BREAST_DF_age_perc_eval <- BREAST_DF_eval %>%
  dplyr::group_by(`Age recode (<60,60-69,70+)`) %>%
  dplyr::summarise(count = dplyr::n()) %>%  # Calculate count per group
  ungroup() %>%  # Ungroup the data
  mutate(total_count = sum(count)) %>%  # Calculate total count
  mutate(percentage = count / total_count * 100)  # Calculate percentage using total count

# Plot the percentages
ggplot(BREAST_DF_age_perc_eval, 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 Ag
       x = "Age range",
       y = "Percentage") +
  ylim(0, round(1.5 * perc_max, 1))
```



```
# In this section, we do the analyses on household income}
#find ubique values for column ratted to age
uniques_househosldes <- 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 = percenta
  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 2011-2015", x = "Household Income", y = "Percentage"
  ylim(0, 1.2*perc_max)
```



```
#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)
ggplot(BREAST_DF_income_perc_eval, aes(x = `Median household income inflation adj to 2021`, y = percenta
  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)
```



```
# 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])
DF_col_names[20]
```

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

```
# 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)
```

COD	count	Total Count	Population
Alive	228221	303557	75
Breast	38472	303557	13
Other	36864	303557	12

```
# Let's first group by the number of tumor and find hom many in the population have those and then amon
BREAST_DF_TNoT_perc_surv <- BREAST_DF_surv %>%
  dplyr::group_by(`Total number of in situ/malignant tumors for patient`) %>%
  dplyr::add_count() %>%
  filter(COD == "Breast") %>%
  dplyr::summarise(`Event Population` = n(),
           Population = dplyr::first(n))  # Use `first()` to extract the total count in each

# Do simple math to fidn the percentage of the groupn un the population and then the percentage of the

BREAST_DF_TNoT_perc_surv$`Group % in total` <- round(BREAST_DF_TNoT_perc_surv$Population/sum(BREAST_DF_

BREAST_DF_TNoT_perc_surv$`Death %` <- round(BREAST_DF_TNoT_perc_surv$`Event Population`/BREAST_DF_TNoT_


kable(BREAST_DF_TNoT_perc_surv)
```

Total number of in situ/malignant tumors for patient	Event Population	Population	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, unl

#use DPLYR to filter based on two parameters chemotheraphy and radiation therapy and evalaute 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

kable(BREAST_DF_RNC_perc_surv)
```

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