

# Breast Exploratory Analysis

Jonathan Ma

2025-12-05

## Table of Contents

1	Introduction.....	2
2	Data Structure and Levels.....	2
3	Variable-Specific Treatment.....	6
3.1	Complete Summary of available data.....	7
4	Hierarchal Encoding.....	7

# 1 Introduction

This exploratory data analysis (EDA) is the first step in our project pipeline:

1. Bayesian Hierarchical Modeling
2. Posterior Inference & Diagnostics
3. Clustering of Individual Risk Profiles

We use data from the [SEER Program](#) focused on breast cancer cases, structured across three levels:

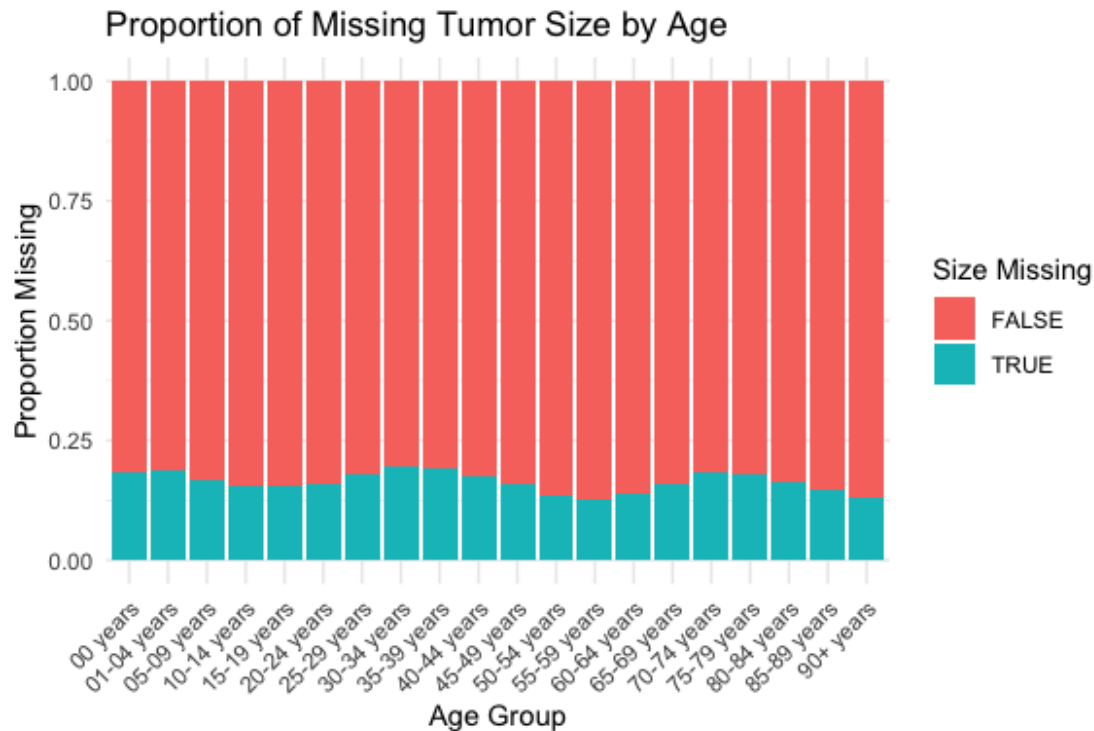
- **Level 1 (Tumor):** Tumor-specific details (e.g., stage, size, grade)
- **Level 2 (Patient):** Demographics and marital status
- **Level 3 (Region):** County-level metro/rural designations

## 2 Data Structure and Levels

Level	Variable	Description
Tumor	stage	SEER stage: Localized, Regional, Distant
Tumor	grade	Tumor grade: 1–4 or unknown
Tumor	size	EOD Tumor Size (mm)
Tumor	site	Tumor site (ICD code)
Patient	id	Patient identifier
Patient	age	Age group at diagnosis
Patient	sex	Biological sex
Patient	raceth	Race/ethnicity
Patient	year	Year of diagnosis
Patient	marry	Marital status
Region	region	County-level metro/rural code

We replace “Unknown” and “Blank(s)” with NA, and then analyse missing data. We want to see if missing data is specific to a group so we can determine if the missingness is MCAR or not.

Is missing data more common in older patients?



Is missing data associated with Race/Ethnicity?

```
##
##
##      FALSE      TRUE
## American Indian/Alaska Native  40048    4775
## Asian or Pacific Islander      453003    58740
## Black                          369571    62192
## White                         3427119    685078
```

Is stage more likely to be missing in nonmetro rural areas?

```
##
##
##      FALSE
## Counties in metropolitan areas ge 1 million pop  0.5234788
## Counties in metropolitan areas of 250,000 to 1 million pop  0.5292829
## Counties in metropolitan areas of lt 250 thousand pop  0.5255872
## Nonmetropolitan counties adjacent to a metropolitan area  0.5250781
## Nonmetropolitan counties not adjacent to a metropolitan area  0.5329171
## Unknown/missing/no match (Alaska or Hawaii - Entire State)  0.4969591
## Unknown/missing/no match/Not 1990-2023  0.6902941
##
##
##      TRUE
## Counties in metropolitan areas ge 1 million pop  0.4765212
## Counties in metropolitan areas of 250,000 to 1 million pop  0.4707171
## Counties in metropolitan areas of lt 250 thousand pop  0.4744128
## Nonmetropolitan counties adjacent to a metropolitan area  0.4749219
## Nonmetropolitan counties not adjacent to a metropolitan area  0.4670829
```

##	Unknown/missing/no match (Alaska or Hawaii - Entire State)	0.5030409
##	Unknown/missing/no match/Not 1990-2023	0.3097059

Although we did not do a statistical test for MCAR, visual and tabular inspection clearly show patterns in missingness. Tumor size by age is non-uniform, with younger and older ends having higher missing rates. Missing rates also vary by racial group, with whites having notably higher missingness.

Logistic regression is a standard method for exploring missing data mechanisms. By modeling the missingness indicator (e.g., stage\_miss) as a binary outcome and regressing it on observed variables, we can test whether missingness is related to any known covariates.

If none of the predictors are significant, this supports the MCAR assumption. But if missingness is significantly associated with observed variables (as seen here), MCAR is rejected, and MAR becomes a more reasonable assumption for downstream modeling.

#### *Significant Predictors of Missingness*

term	estimate	std.error	statistic	p.value
(Intercept)	-3.3421234	0.1087065	-30.744454	0.0000000
sexMale	0.1130525	0.0082547	13.695489	0.0000000
racethBlack	-0.4788434	0.0480318	-9.969298	0.0000000
Region250k1m	-0.0416125	0.0096221	-4.324685	0.0000153
regionAdjacent	-0.0353161	0.0165871	-2.129134	0.0332432
age05-09 years	-0.6052225	0.1070662	-5.652788	0.0000000
age10-14 years	-0.6300932	0.1017166	-6.194596	0.0000000
age15-19 years	-0.8424653	0.0979111	-8.604393	0.0000000
age20-24 years	-1.6771137	0.0978460	-17.140335	0.0000000
age25-29 years	-2.1683220	0.0947807	-22.877256	0.0000000
age30-34 years	-2.6078400	0.0926182	-28.156878	0.0000000
age35-39 years	-2.7019504	0.0903702	-29.898703	0.0000000
age40-44 years	-2.7800923	0.0887195	-31.335753	0.0000000
age45-49 years	-2.7567900	0.0876996	-31.434472	0.0000000
age50-54 years	-2.8570329	0.0872416	-32.748516	0.0000000
age55-59 years	-2.8805064	0.0870325	-33.096915	0.0000000
age60-64 years	-2.9813609	0.0869839	-34.274843	0.0000000
age65-69 years	-3.0790547	0.0869820	-35.398756	0.0000000
age70-74 years	-3.1090831	0.0870408	-35.719855	0.0000000
age75-79 years	-3.1206521	0.0871319	-35.815275	0.0000000
age80-84 years	-3.0736712	0.0873639	-35.182385	0.0000000

term	estimate	std.error	statistic	p.value
age85-89 years	-3.0133837	0.0880146	-34.237313	0.0000000
age90+ years	-2.9132745	0.0898371	-32.428399	0.0000000
marryMarried	-0.1079600	0.0132770	-8.131365	0.0000000
marryWidowed	0.0556292	0.0167833	3.314558	0.0009179
gradeModDif2	3.0275322	0.0454699	66.583279	0.0000000
gradeNKcell	1.3816412	0.1920227	7.195198	0.0000000
gradeNull	1.9072555	0.3334825	5.719207	0.0000000
gradePoorDif3	3.3996482	0.0453239	75.007824	0.0000000
gradeT-cell	1.3353125	0.0538452	24.799082	0.0000000
gradeUndif4	3.8783633	0.0463071	83.753163	0.0000000
gradeWelldif1	3.2931637	0.0463463	71.055608	0.0000000
size10	1.0994725	0.1328726	8.274637	0.0000000
size13	-1.1778932	0.3582677	-3.287746	0.0010099
size14	-1.1490940	0.1784081	-6.440819	0.0000000
size15	-0.7816986	0.1989564	-3.928994	0.0000853
size21	-3.9884711	0.2430364	-16.411002	0.0000000
size22	-3.5769274	0.0718433	-49.787917	0.0000000
size23	-3.8274960	0.1328494	-28.810792	0.0000000
size24	-2.2211520	0.1011867	-21.951034	0.0000000
size25	0.4252906	0.0411808	10.327392	0.0000000
size28	-3.0998239	0.4087901	-7.582922	0.0000000
size30	0.1620699	0.0132818	12.202437	0.0000000
size31	3.6734760	0.0271003	135.550905	0.0000000
size32	-0.4748670	0.0689579	-6.886337	0.0000000
size33	-3.0787173	0.1339825	-22.978495	0.0000000
size34	1.1102713	0.1186365	9.358597	0.0000000
size35	-0.8325717	0.2475636	-3.363062	0.0007708
size36	-3.5255034	0.4096422	-8.606299	0.0000000
size40	0.4475888	0.0139823	32.011012	0.0000000
size41	-0.5123659	0.0299619	-17.100556	0.0000000
size42	1.2539769	0.0261819	47.894853	0.0000000
size43	2.3565217	0.0394405	59.748781	0.0000000
size45	0.5494725	0.0449365	12.227761	0.0000000
size47	-0.7049307	0.1450540	-4.859781	0.0000012

term	estimate	std.error	statistic	p.value
size50	-1.4686617	0.0206932	-70.973031	0.0000000
size51	-3.8085600	0.1234997	-30.838625	0.0000000
size52	-1.9613675	0.1112684	-17.627349	0.0000000
size53	-1.8154795	0.1290900	-14.063671	0.0000000
size54	-1.2622363	0.1476548	-8.548560	0.0000000
size60	2.1628059	0.0163218	132.509978	0.0000000
size62	-4.7924637	1.0003017	-4.791018	0.0000017
size70	-1.0941376	0.0928675	-11.781711	0.0000000
size80	-0.6650624	0.1101807	-6.036105	0.0000000
size90	1.3585843	0.0473712	28.679569	0.0000000
size98	6.7787245	0.0452769	149.717226	0.0000000
size99	1.7228091	0.0690427	24.952801	0.0000000

The missingness in stage is likely not MCAR, because missingness depends significantly on several observed variables (e.g., sex, raceth, region, age, marry, grade, size). Therefore, it's reasonable to assume MAR (Missing at Random) for modeling purposes. We find we have 209123 cases to work with (4.06% of the whole thing) after removing "Unknown" and "Blank(s)" Values.

```
## Complete cases: 209123 out of 5149008 (4.06% complete)
```

### 3 Variable-Specific Treatment

Further cleaning is done via tabling all variables.

We want to filter for top ten cancer sites (breast, lung, etc). We exclude 998 and 999 because they are unknown.

```
##
## 010 012 015 018 020 025 030 035 040 050
## 4269 3940 5738 3284 5440 4994 6046 3768 4953 4415
```

Collapse and recode into ordered, interpretable regional factors, removes the Alaska/Hawaii "unknown" entries.

```
##
## Large Small Nonmetro
## 39835 2410 4497
```

Restrict to tumors that are invasive (drop In situ, Unknown/unstaged).

```
##
## Localized Regional Distant
## 26446 13382 3918
```

Extract numeric grade 1–4; drop cell-line entries (“B-cell”, “T-cell” etc.). Convert them to Start (1/2) and End (3/4) since our model is logistic.

```
##
## Start End
## 42877 3650
```

Making this numeric and then filtering implausible values.

```
## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
## 10.00 23.00 30.00 34.66 44.00 99.00 6305
```

Making this binary, only married or unmarried.

```
##
## Married Unmarried
## 28032 18815
```

Recode RACETH to be three levels, Black, White, and Other.

```
##
## W B O
## 36288 3978 6581
```

### 3.1 Complete Summary of available data

Seeing how many cases are available now:

```
## Remaining complete cases: 37530
```

## 4 Hierarchal Encoding

To do hierarchal models in R, identifiers must be encoded as numeric integers (e.g., 1, 2, ..., N) for levels like REGION and ID. We complete that, then set factor variables, and check resulting data structure.

```
## 'data.frame': 37530 obs. of 13 variables:
## $ id : int 1028 1034 1152 2172 4489 5518 9181 9314 9415 10092 ...
## $ region : Factor w/ 3 levels "Large","Small",...: 1 1 1 1 1 1 1 1 1 1 ...
## $ stage : Factor w/ 3 levels "Localized","Regional",...: 1 1 1 1 1 1 1 1 1 1 ...
## $ age : Factor w/ 20 levels "00 years","01-04 years",...: 14 16 17 19 15 13 17 18 18 17 ...
## $ sex : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 1 1 1 1 ...
## $ raceth : Factor w/ 3 levels "W","B","O": 1 2 1 3 1 2 3 1 3 1 ...
```

```
## $ grade      : Factor w/ 2 levels "Start","End": 1 1 1 2 1 1 1 1 2 1 ...
## $ size       : num  30 30 38 27 22 49 22 50 27 27 ...
## $ year       : int   2017 2016 2016 2017 2017 2016 2017 2017 2017 2016 ...
## $ marry      : Factor w/ 2 levels "Married","Unmarried": 1 2 1 2 2 2 1 1 2
2 ...
## $ site       : chr   "020" "025" "035" "020" ...
## $ regionid   : int    1 1 1 1 1 1 1 1 1 1 ...
## $ patientid  : int    1 2 3 4 5 6 7 8 9 10 ...
```