Women in Data Science Worldwide | Afghanistan

WOMEN in tech® Afghanistan

Upskill Workshop I

June 24, 2024

Table of Content

- Technical Part
- WiDS Dataset 2024 Overview
- Dataset Availability
- Datasets Size
- Analyzing Relationships with the Target Variable
- Identifying the Most Related Columns to the Target Variable
- Summary Statics
- Heatmap of Missing Values
- Distribution of Numerical Data
- Bar Chart of Categorical Data
- Analysis of Target Column
- Analysis of Age
- Analysis of Region
- Analysis of Patient Race
- Socioeconomic Indicators
- Analysis of Cancer Codes Data
- Word Cloud of Caner Code Description
- Coding file

Technical Part

Environment:

- Google Colab

Language:

- Python

Libraries:

- numpy
- pandas
- scipy.stats
- matplotlib.pyplot
- seaborn
- math
- re
- nltk
- wordcloud

WiDS Dataset 2024 Overview

Dataset Source:

- Provided by Gilead Sciences through Health Verity, featuring patients diagnosed with metastatic triple negative breast cancer in the US.

Dataset Enrichment:

- Includes demographics (age, race, BMI), diagnosis details (cancer codes, treatment), and insurance information.
- Geo-demographic data enriched from US Zip Codes Database for socio-economic insights.
- Zip code level climate data added to explore climate's impact on healthcare access.
- Target Column: The dataset includes a target variable metastatic_diagnosis_period, which measures the period (in days) between the initial breast cancer diagnosis and the subsequent diagnosis of metastatic cancer.

Dataset Availability

The dataset is provided in two sets:

- **Training Dataset:** Named train.csv, this dataset contains observed values of the metastatic diagnosis period for model training.

- **Testing Dataset:** Named test.csv, this dataset withholds the observed values and is used for evaluating model predictions.

Datasets are available at :
https://www.kaggle.com/competitions/widsdatathon2024-challenge2/data

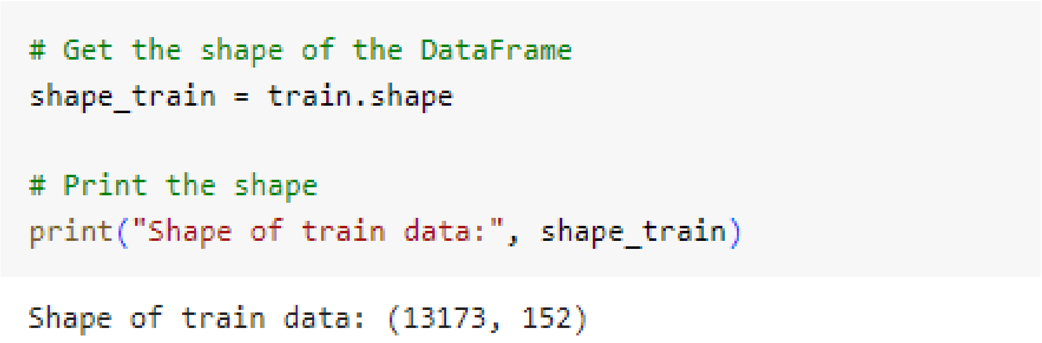
Datasets Size

Train Data Shape:

- Number of Rows: 13173
- Number of Columns: 152
- Numerical Columns: 141
- Categorical Columns: 11

Test Data Shape:

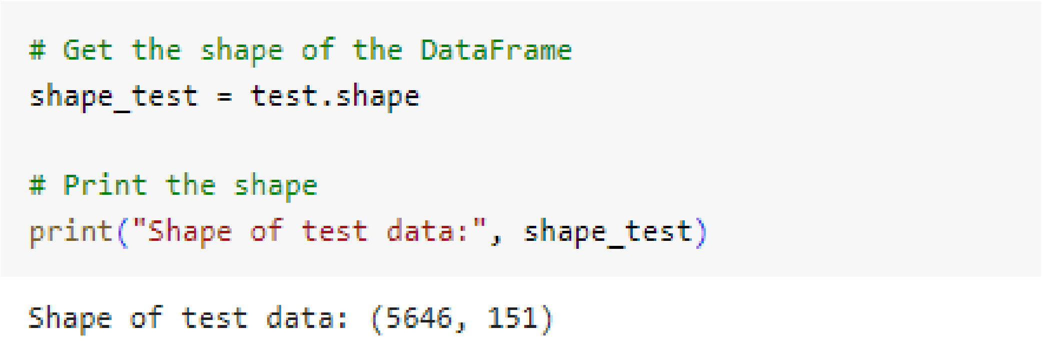
- Number of Rows: 5646
- Number of Columns: 151
- Numerical Columns: 140
- Categorical Columns: 11



```
# Get the shape of the DataFrame
shape_train = train.shape

# Print the shape
print("Shape of train data:", shape_train)
```

```
Shape of train data: (13173, 152)
```



```
# Get the shape of the DataFrame
shape_test = test.shape

# Print the shape
print("Shape of test data:", shape_test)
```

```
Shape of test data: (5646, 151)
```

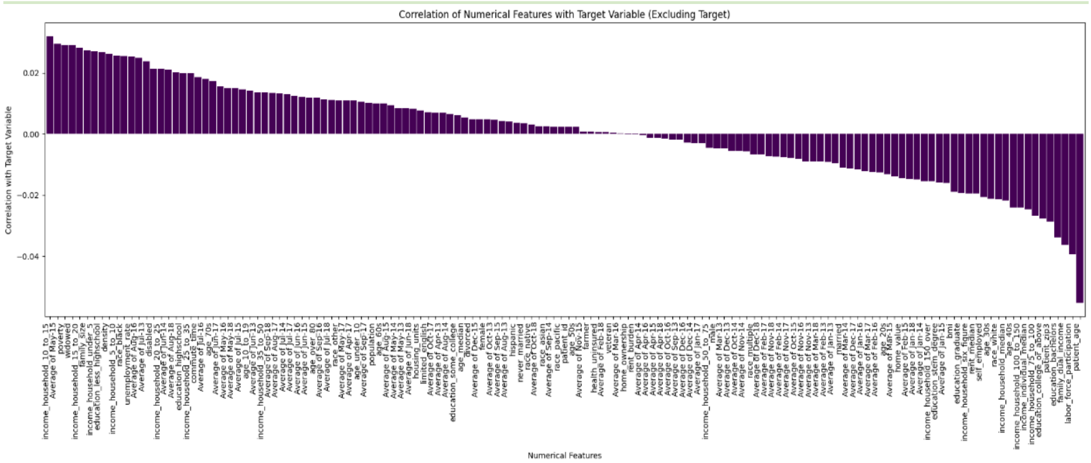
Analyzing Relationships with the Target Variable

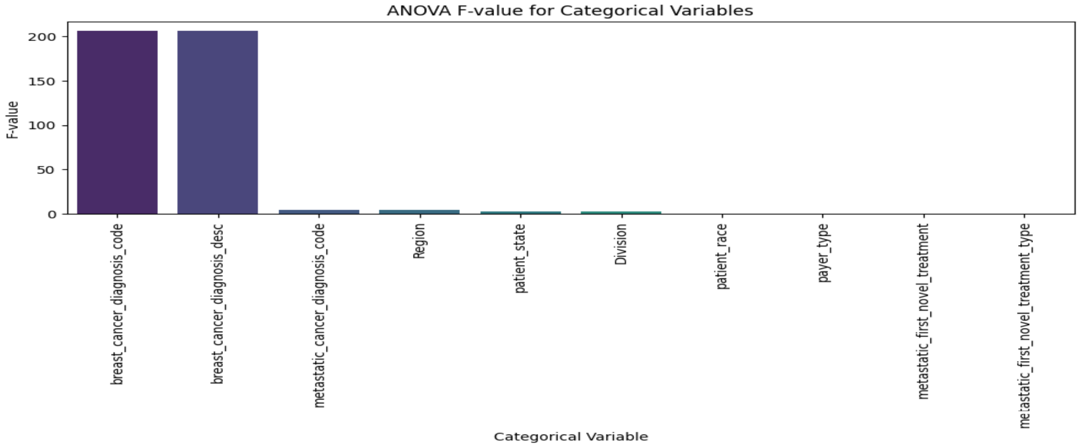
Numerical Columns

- Pearson correlation coefficient used to measure linear correlation with the target variable.

Categorical Columns:

- One-way ANOVA employed to assess statistical significance of categorical variables concerning the target variable.

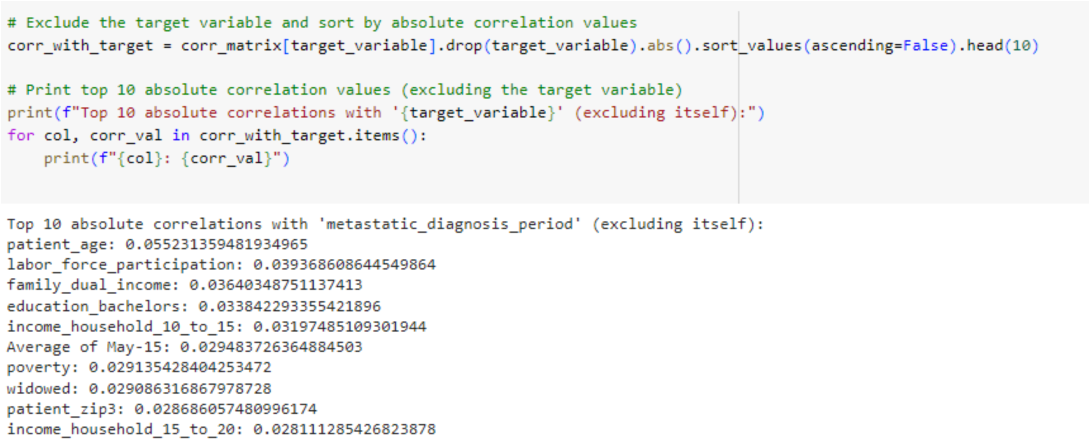




Identifying the Most Related Columns to the Target Variable

Subset Selection Method:

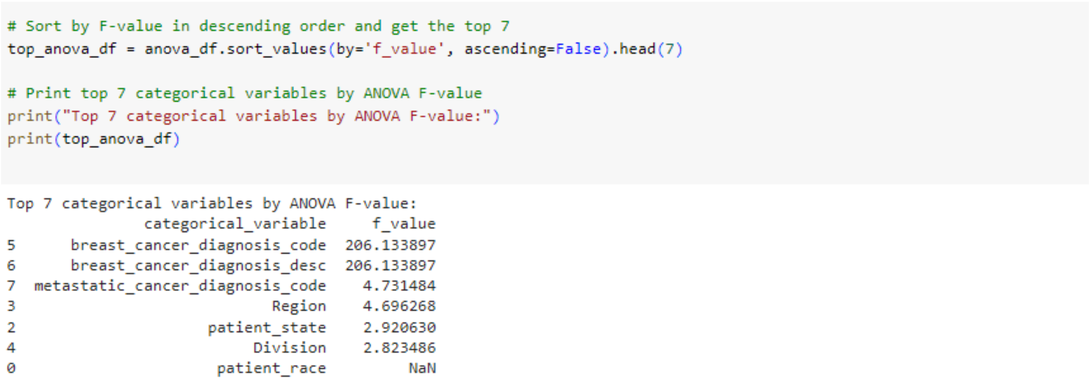
- Utilized Pearson correlation for numerical columns to identify the top 10 most correlated features with the target variable.
- Employed one-way ANOVA for categorical columns to select the 7 most relevant features in relation to the target variable.



```
# Exclude the target variable and sort by absolute correlation values
corr_with_target = corr_matrix[target_variable].drop(target_variable).abs().sort_values(ascending=False).head(10)

# Print top 10 absolute correlation values (excluding the target variable)
print(f"Top 10 absolute correlations with '{target_variable}' (excluding itself):")
for col, corr_val in corr_with_target.items():
    print(f"{col}: {corr_val}")
```

```
Top 10 absolute correlations with 'metastatic_diagnosis_period' (excluding itself):
patient_age: 0.055231359481934965
labor_force_participation: 0.039368608644549864
family_dual_income: 0.03640348751137413
education_bachelors: 0.033842293355421896
income_household_10_to_15: 0.031974851093019 44
Average of May-15: 0.029483726364884503
poverty: 0.029135428404253472
widowed: 0.029086316867978728
patient_zip3: 0.028686057480996174
income_household_15_to_20: 0.028111285426823878
```



```
# Sort by F-value in descending order and get the top 7
top_anova_df = anova_df.sort_values(by='f_value', ascending=False).head(7)

# Print top 7 categorical variables by ANOVA F-value
print("Top 7 categorical variables by ANOVA F-value:")
print(top_anova_df)
```

```
Top 7 categorical variables by ANOVA F-value:
             categorical_variable     f_value
5      breast_cancer_diagnosis_code  206.133897
6      breast_cancer_diagnosis_desc  206.133897
7  metastatic_cancer_diagnosis_code    4.731484
3                            Region    4.696268
2                     patient_state    2.920630
4                          Division    2.823486
0                      patient_race         NaN
```

Descriptions of Columns in the Provided Subset

- patient_age: Age of the patient derived from their year of birth.
- labor_force_participation: Percentage of residents aged 16 and older participating in the labor force.
- family_dual_income: Percentage of families with dual income earners.
- education_bachelors: Percentage of residents with a bachelor's degree (or equivalent) but no more.
- Average of May-15: Average temperature for the patient’s zip code in May 2015.
- poverty: Median value of owner-occupied homes.
- patient_zip3: Zip code of the patient's residence (first three digits).
- Region: Region of the patient's location.
- patient_state: State abbreviation of the patient's residence.
- Division: Division of the patient's location.
- patient_race: Race of the patient.
- breast_cancer_diagnosis_code: ICD10 or ICD9 diagnoses code for breast cancer.
- breast_cancer_diagnosis_desc: Description of the breast cancer diagnosis code.
- metastatic_cancer_diagnosis_code: ICD10 diagnoses code for metastatic cancer.
- metastatic_diagnosis_period: Period (in days) between breast cancer diagnosis and metastatic cancer diagnosis.

Before We Start…

Checklist for Dataset Analysis

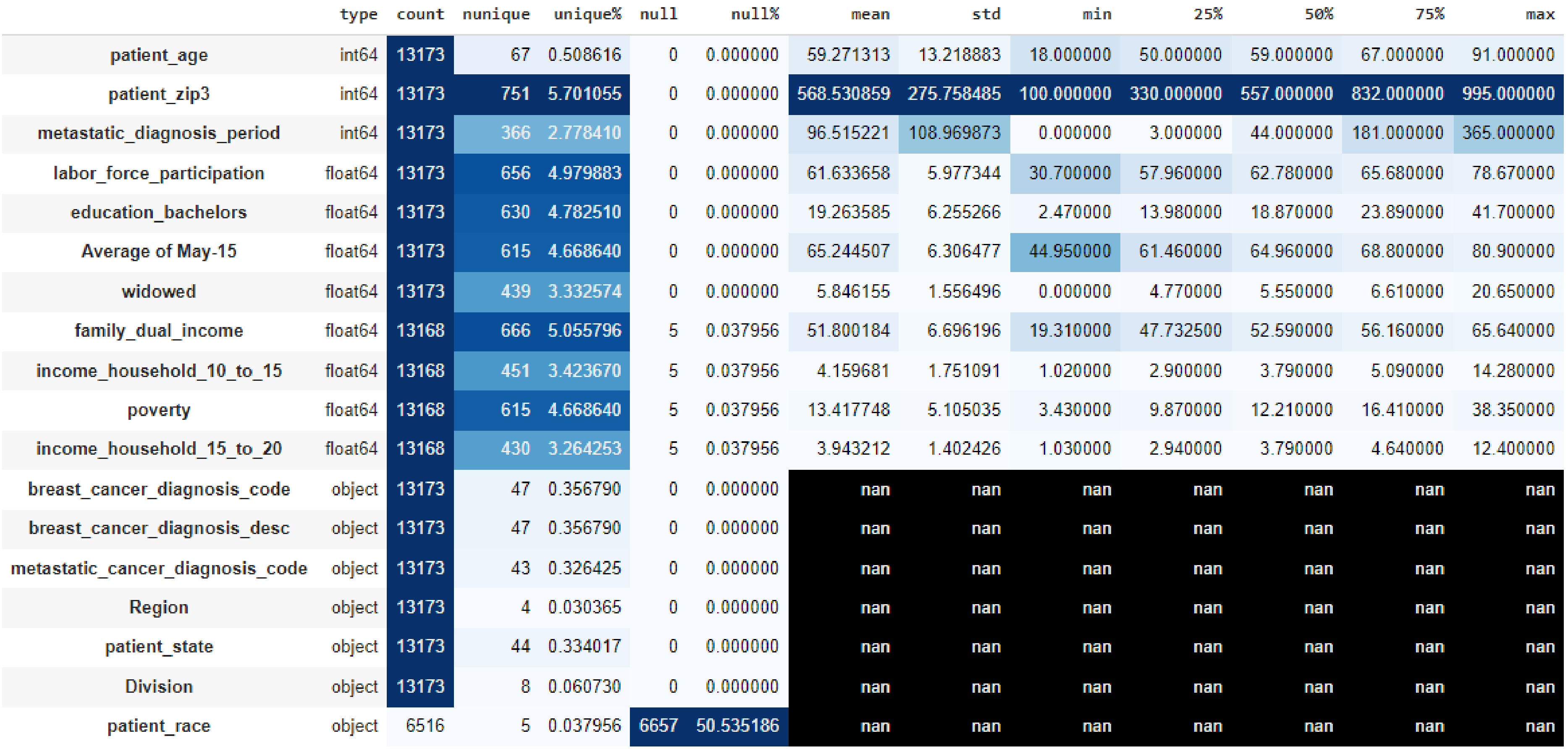
Key Questions:

What specific **questions** should this dataset answer?

Tracking Parameters:

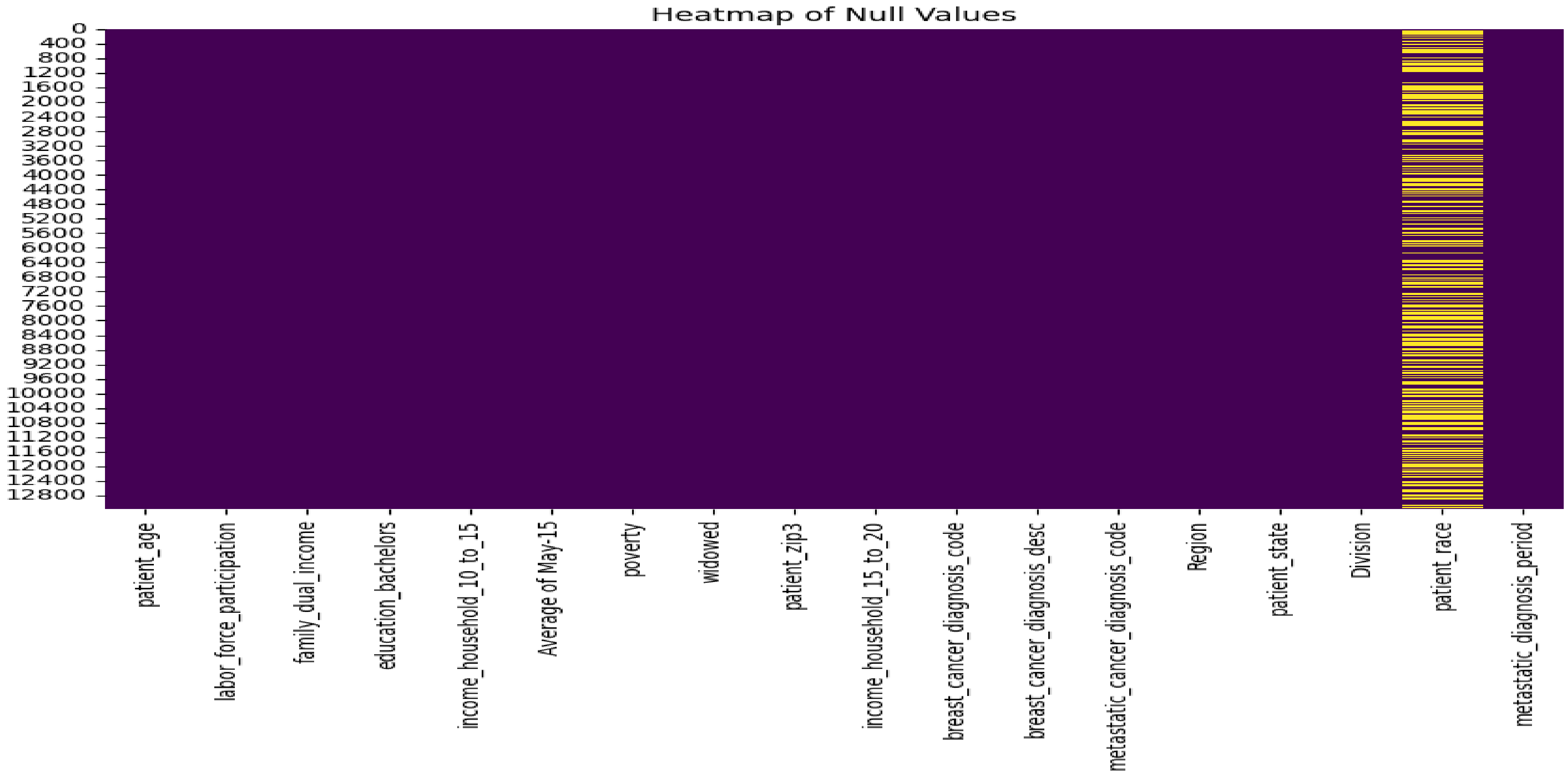
Which **parameters** do we need to monitor to answer these questions?

Summary Statics

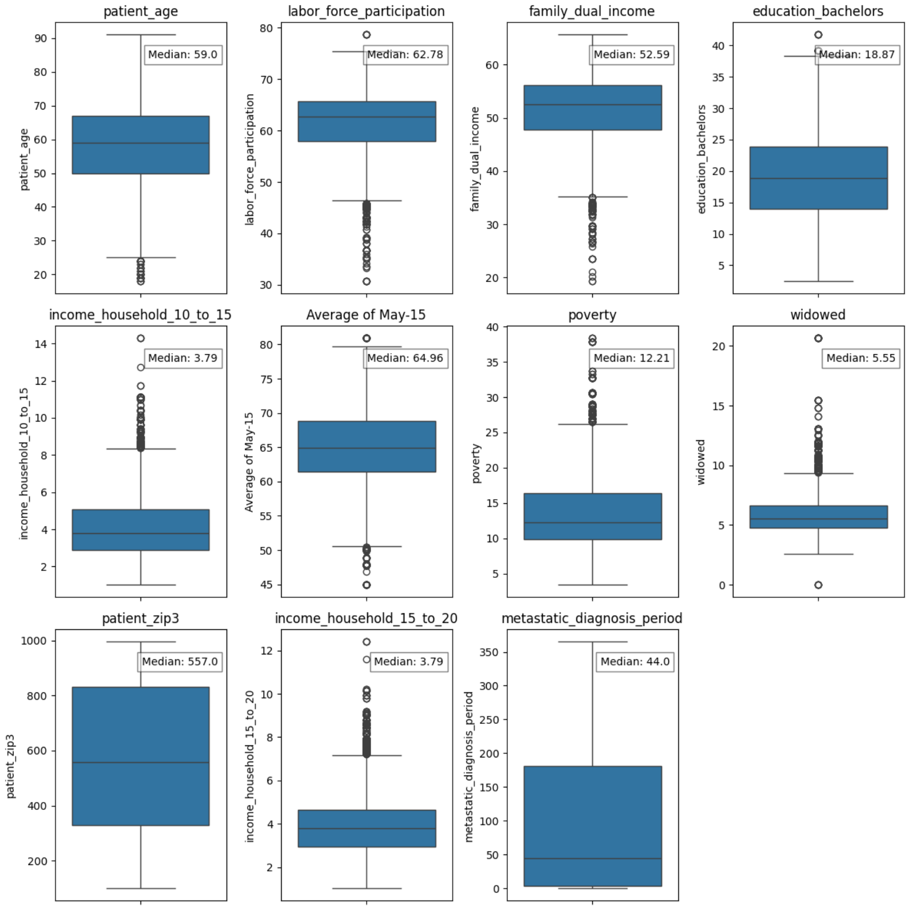


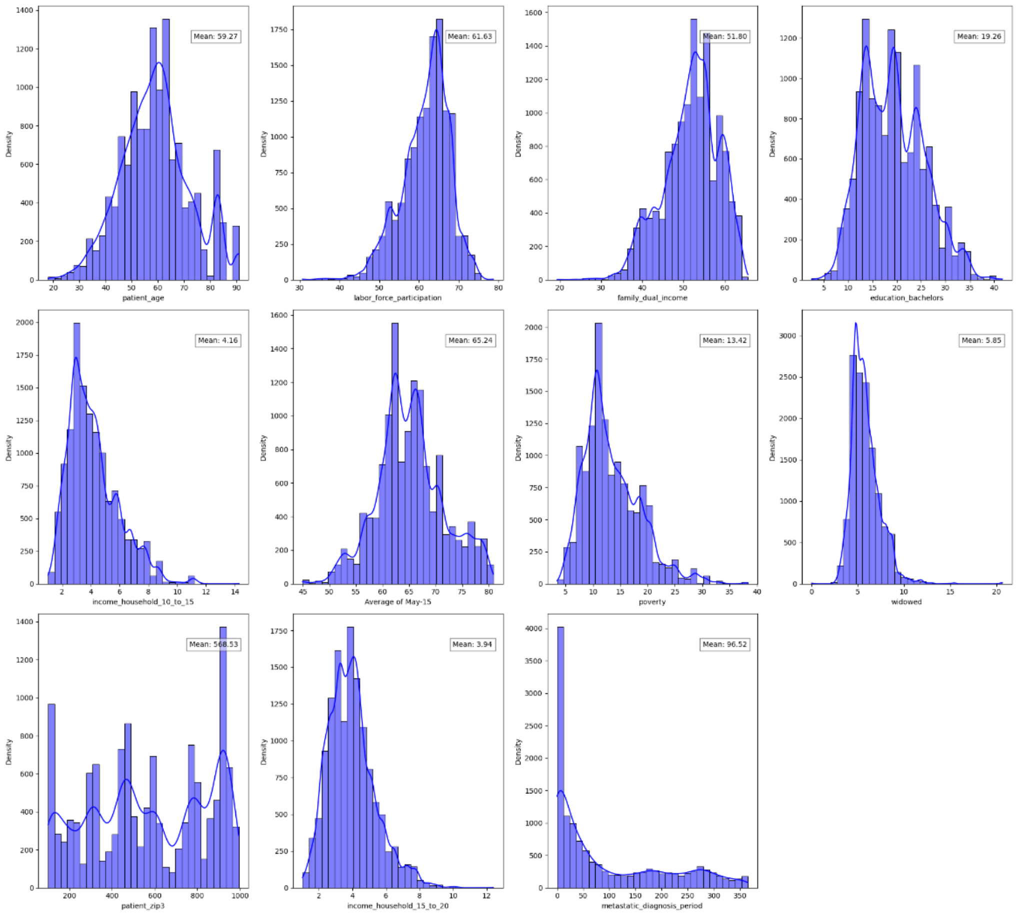
	type	count	nunique	unique%	null	null%	mean	std	min	25%	50%	75%	max
patient_age	int64	13173	67	0.508616	0	0.000000	59.271313	13.218883	18.000000	50.000000	59.000000	67.000000	91.000000
patient_zip3	int64	13173	751	5.701055	0	0.000000	568.530859	275.758485	100.000000	330.000000	557.000000	832.000000	995.000000
metastatic_diagnosis_period	int64	13173	366	2.778410	0	0.000000	96.515221	108.969873	0.000000	3.000000	44.000000	181.000000	365.000000
labor_force_participation	float64	13173	656	4.979883	0	0.000000	61.633658	5.977344	30.700000	57.960000	62.780000	65.680000	78.670000
education_bachelors	float64	13173	630	4.782510	0	0.000000	19.263585	6.255266	2.470000	13.980000	18.870000	23.890000	41.700000
Average of May-15	float64	13173	615	4.668640	0	0.000000	65.244507	6.306477	44.950000	61.460000	64.960000	68.800000	80.900000
widowed	float64	13173	439	3.332574	0	0.000000	5.846155	1.556496	0.000000	4.770000	5.550000	6.610000	20.650000
family_dual_income	float64	13168	666	5.055796	5	0.037956	51.800184	6.696196	19.310000	47.732500	52.590000	56.160000	65.640000
income_household_10_to_15	float64	13168	451	3.423670	5	0.037956	4.159681	1.751091	1.020000	2.900000	3.790000	5.090000	14.280000
poverty	float64	13168	615	4.668640	5	0.037956	13.417748	5.105035	3.430000	9.870000	12.210000	16.410000	38.350000
income_household_15_to_20	float64	13168	430	3.264253	5	0.037956	3.943212	1.402426	1.030000	2.940000	3.790000	4.640000	12.400000
breast_cancer_diagnosis_code	object	13173	47	0.356790	0	0.000000	nan	nan	nan	nan	nan	nan	nan
breast_cancer_diagnosis_desc	object	13173	47	0.356790	0	0.000000	nan	nan	nan	nan	nan	nan	nan
metastatic_cancer_diagnosis_code	object	13173	43	0.326425	0	0.000000	nan	nan	nan	nan	nan	nan	nan
Region	object	13173	4	0.030365	0	0.000000	nan	nan	nan	nan	nan	nan	nan
patient_state	object	13173	44	0.334017	0	0.000000	nan	nan	nan	nan	nan	nan	nan
Division	object	13173	8	0.060730	0	0.000000	nan	nan	nan	nan	nan	nan	nan
patient_race	object	6516	5	0.037956	6657	50.535186	nan	nan	nan	nan	nan	nan	nan

Heatmap of Missing Values



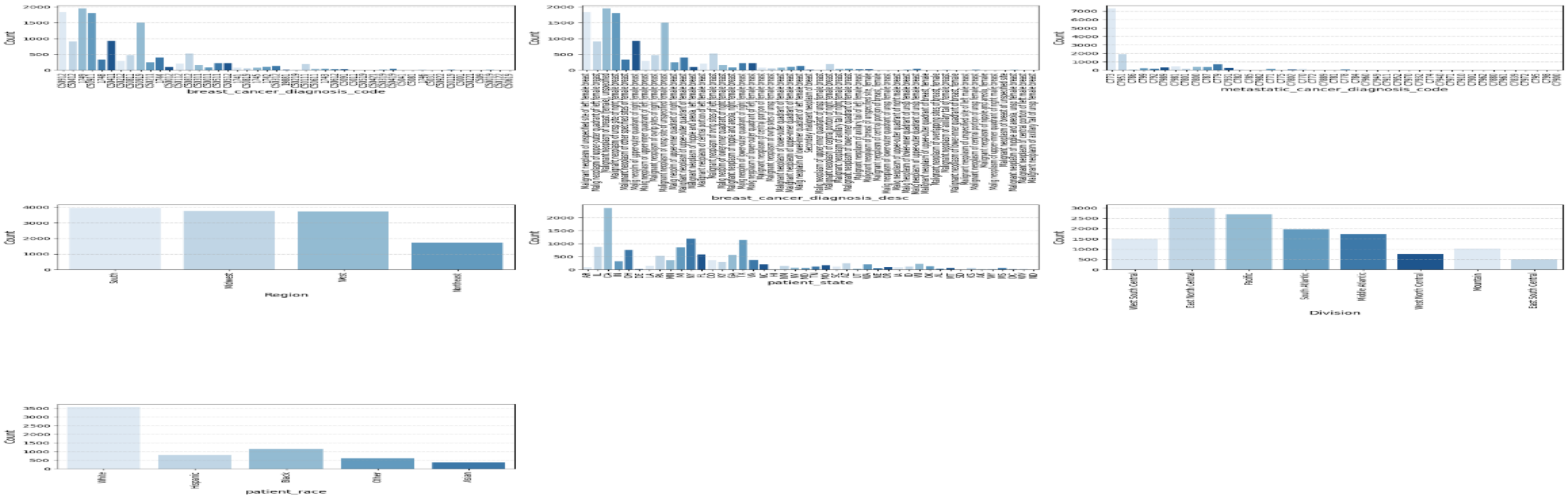
Distribution of Numerical Data





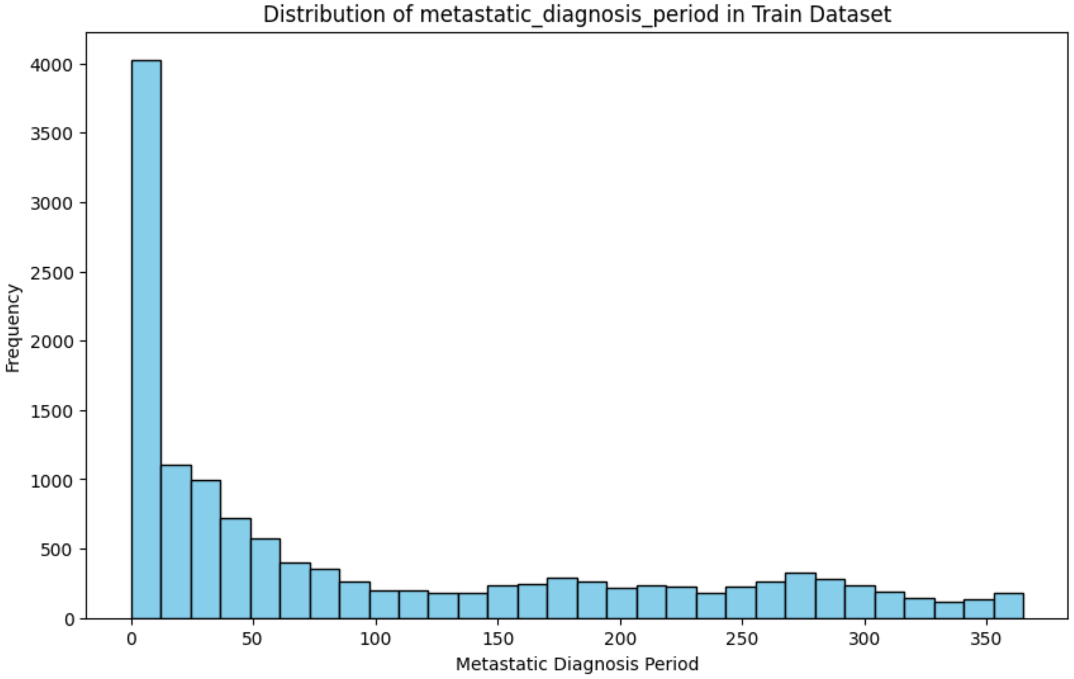
How are the numerical variables distributed across the dataset, and what insights can be derived from their distributions?

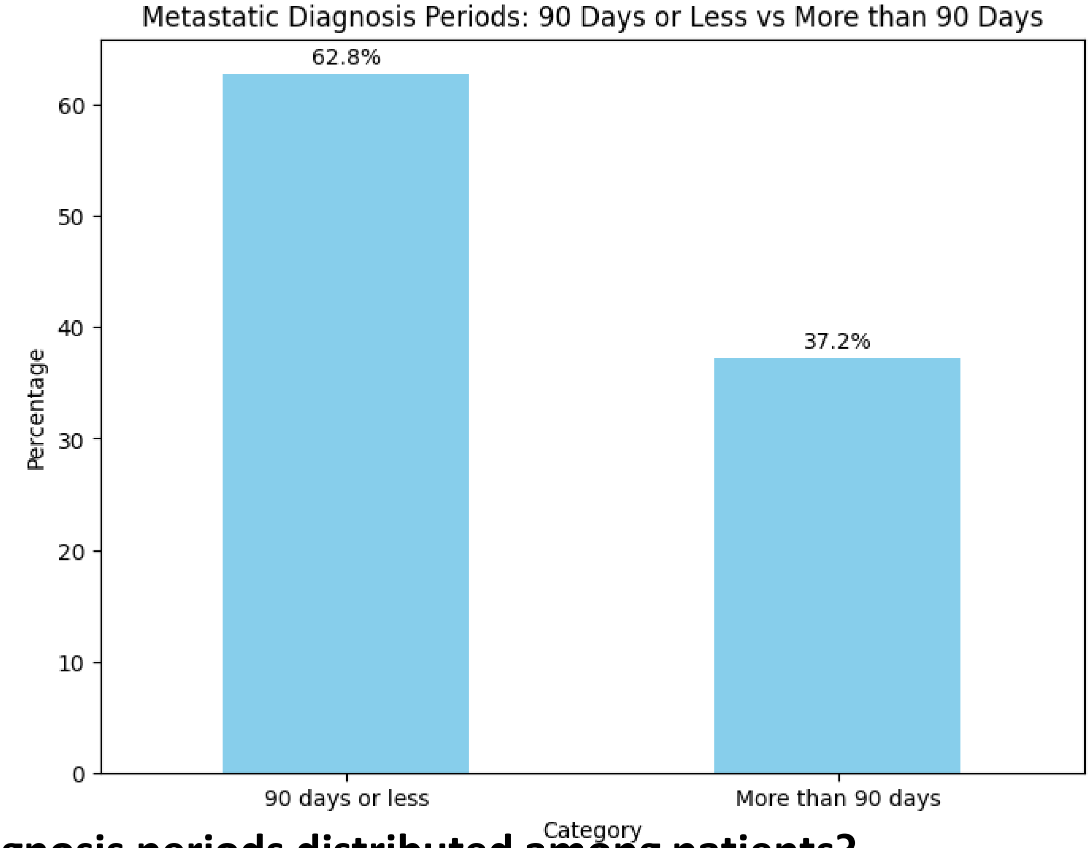
Bar Chart of Categorical Data



How do different categories within each categorical variable compare in terms of their frequency or distribution?

Analysis of Target Column

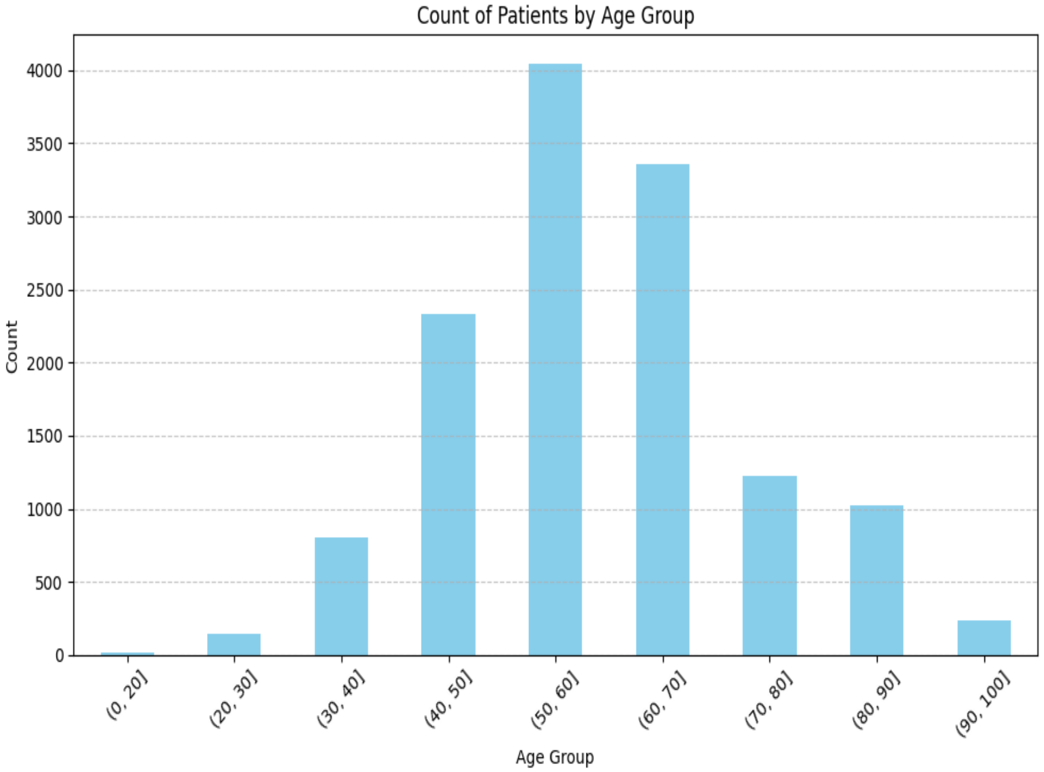


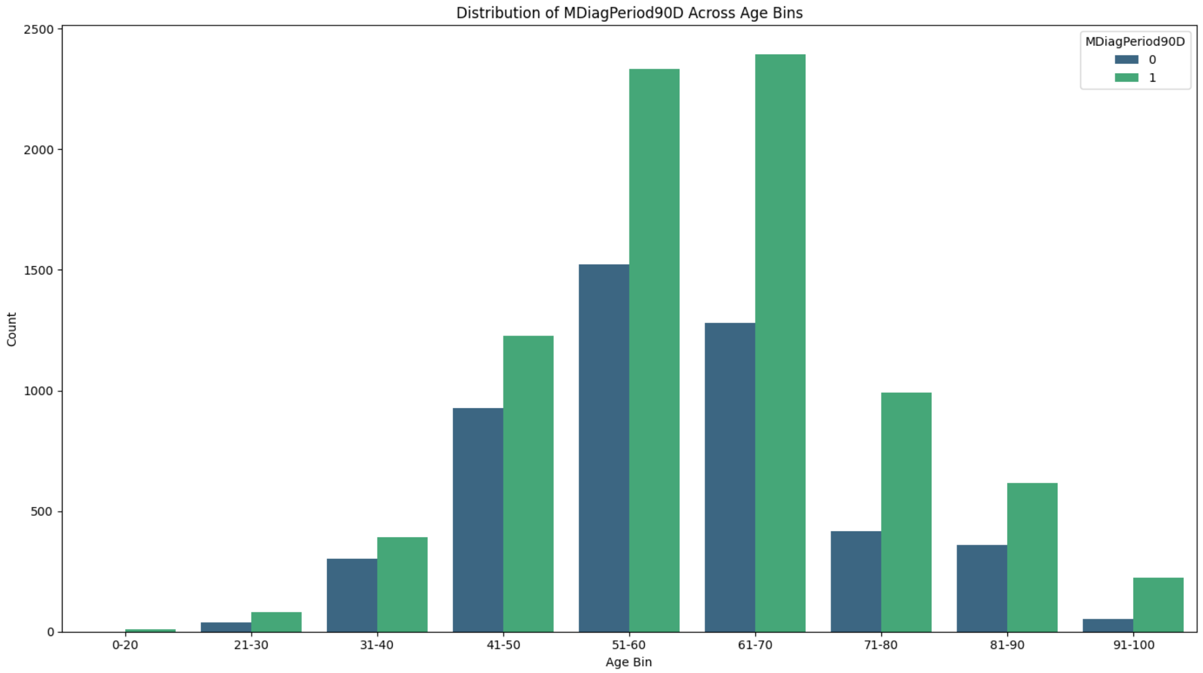


- **How is the frequency distribution of metastatic diagnosis periods distributed among patients?**
- **How is the distribution of metastatic diagnosis periods categorized as 90 days or less versus more than 90 days among patients?**

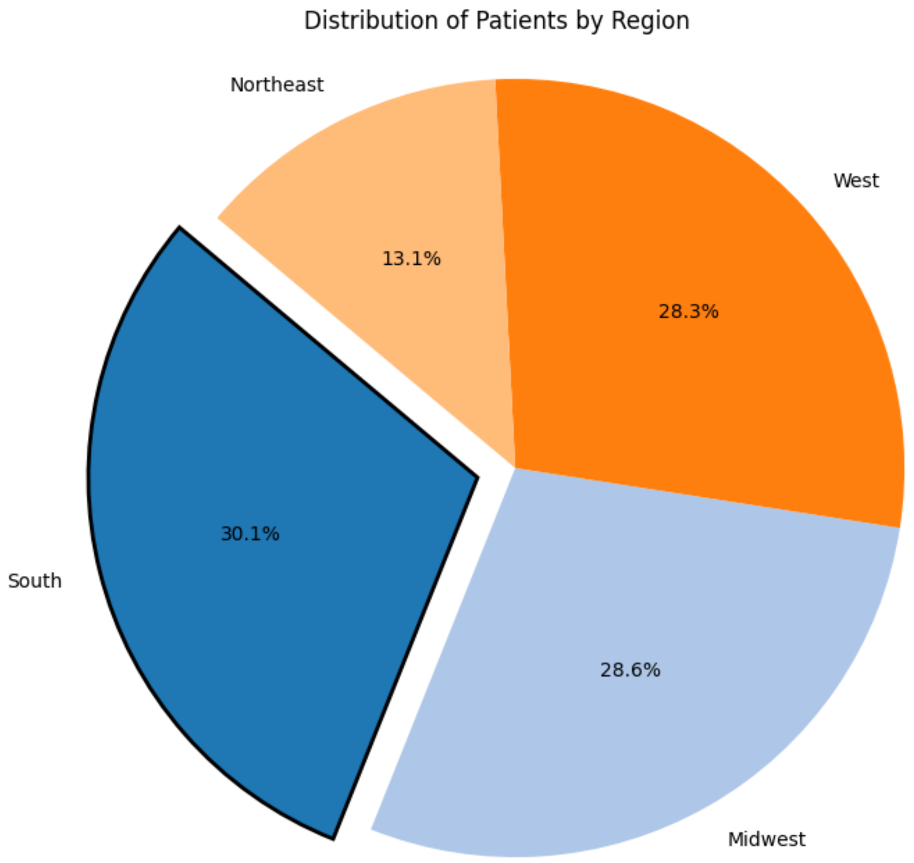
Analysis of Age

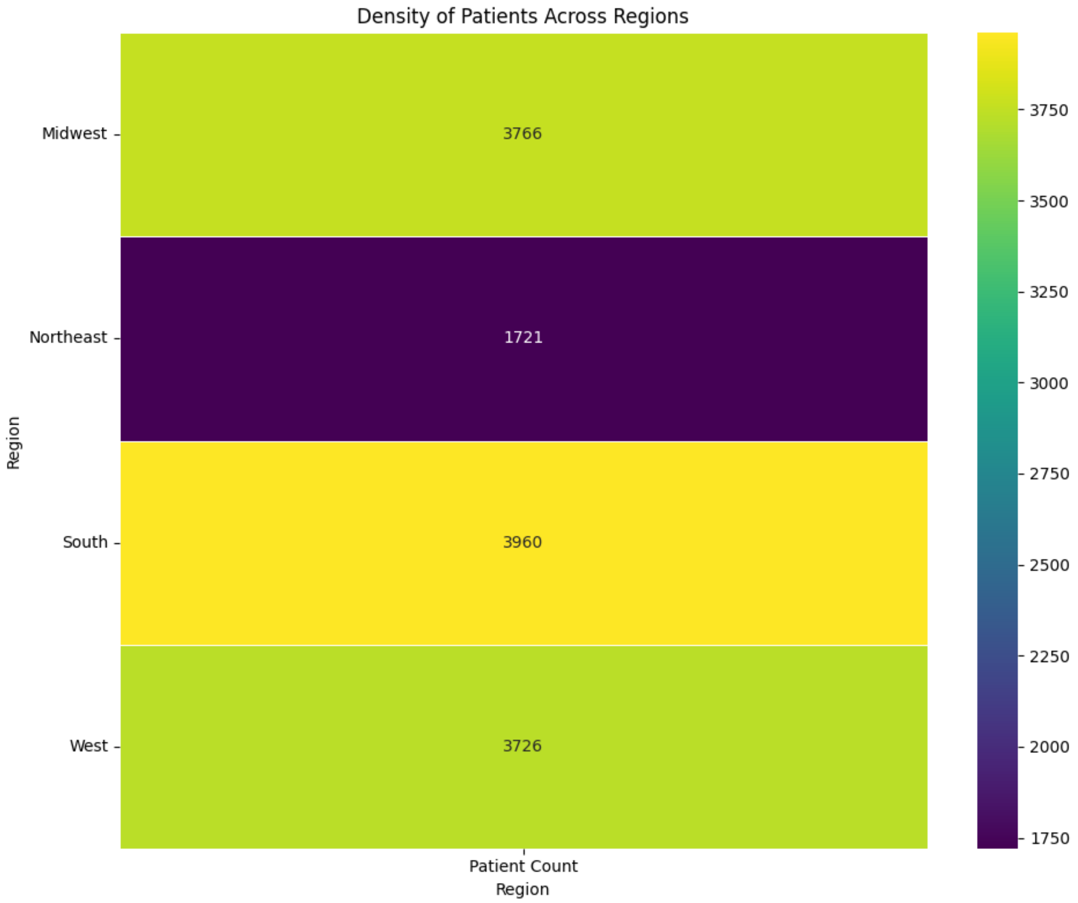
- **What is the distribution of patients across different age groups in the dataset?**
- **How does the distribution of `MDiagPeriod90D` (Metastatic Diagnosis Period within 90 days) vary across different age groups?**





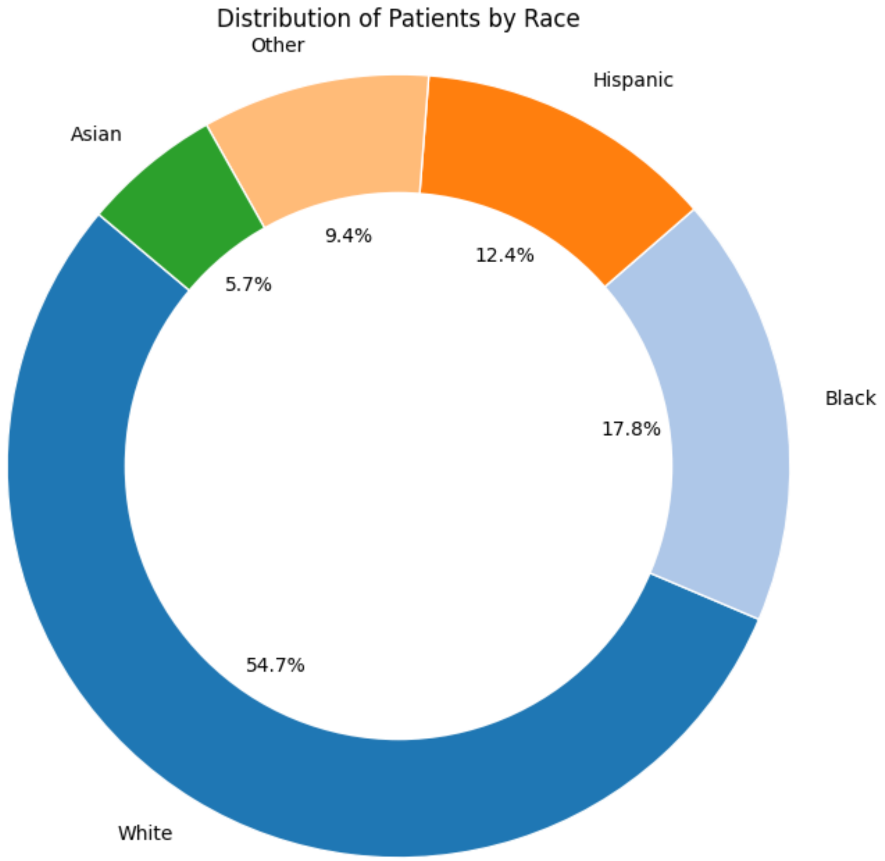
Analysis of Region

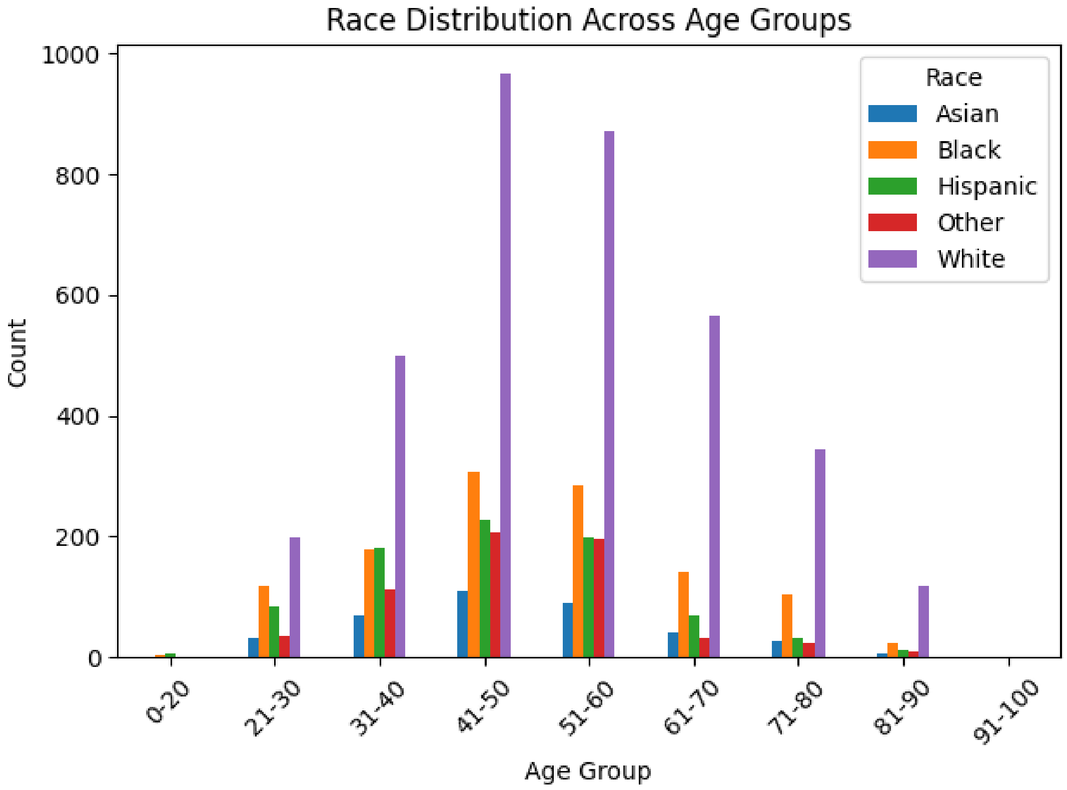




How are patients distributed across various regions?

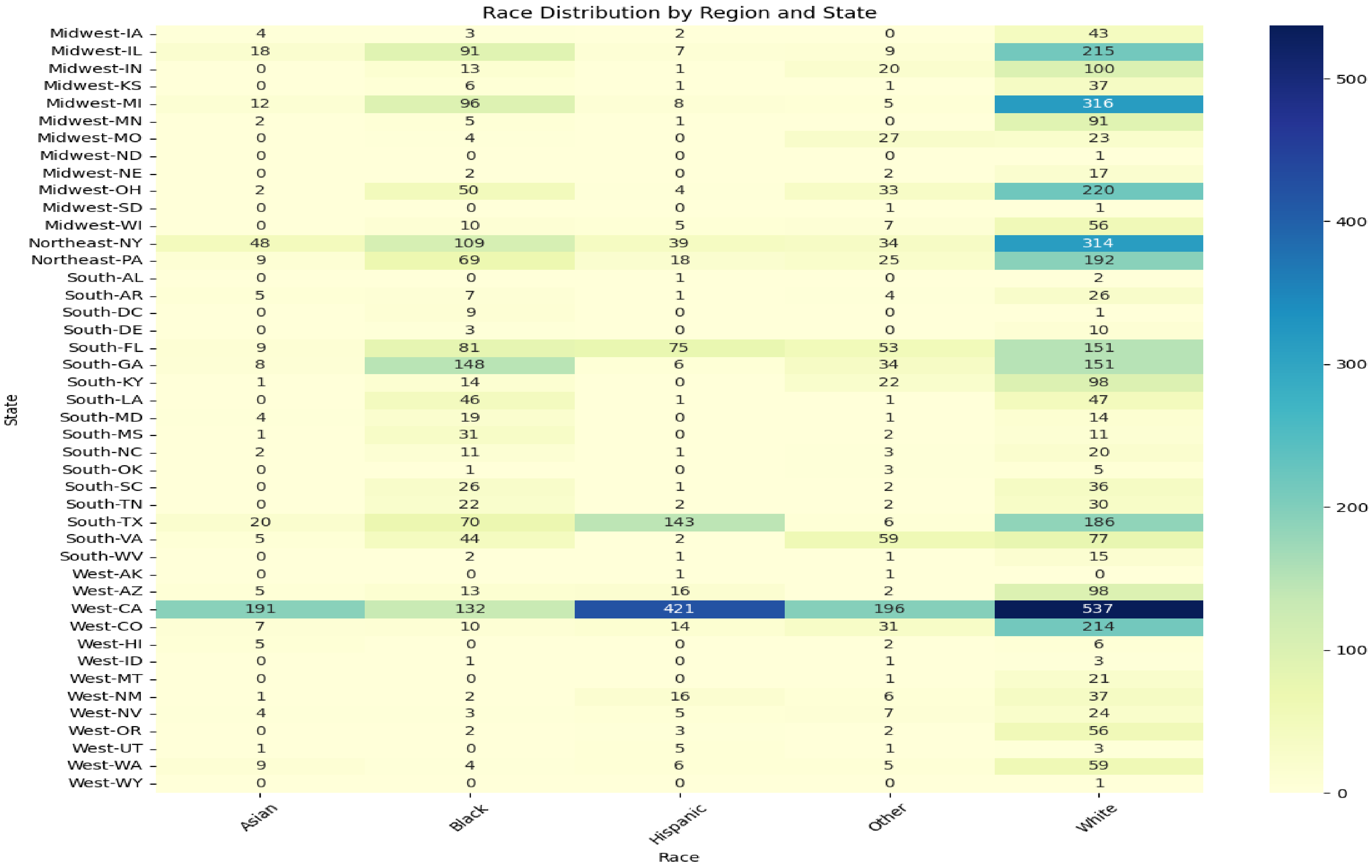
Analysis of Patient Race





- **What is the distribution of patients across different racial groups in the dataset?**
- **How does the distribution of patient races vary across different age groups?**

Cont.

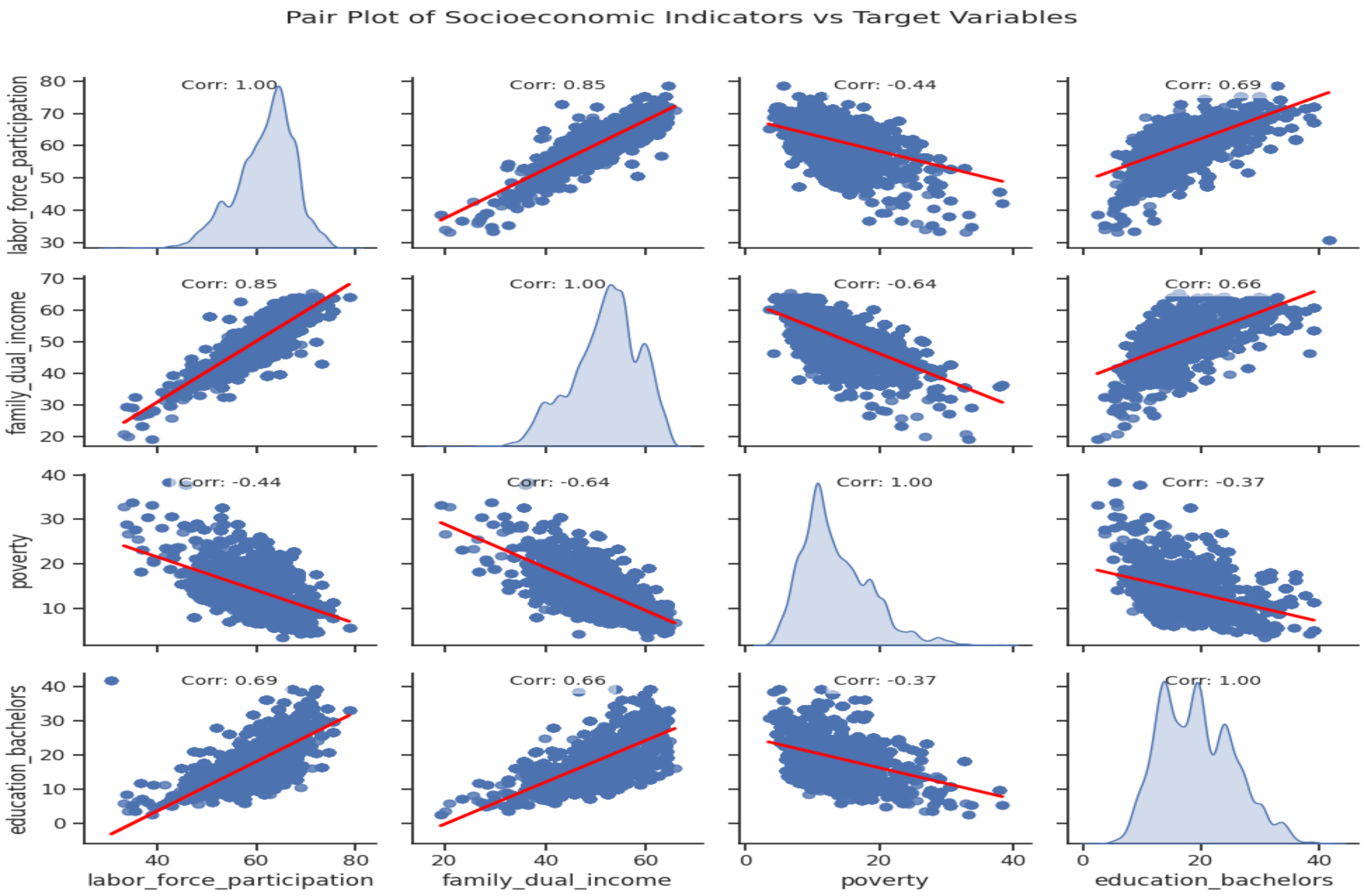


Race Distribution by Region and State

State	Asian	Black	Hispanic	Other	White
Midwest-IA	4	3	2	0	43
Midwest-IL	18	91	7	9	215
Midwest-IN	0	13	1	20	100
Midwest-KS	0	6	1	1	37
Midwest-MI	12	96	8	5	316
Midwest-MN	2	5	1	0	91
Midwest-MO	0	4	0	27	23
Midwest-ND	0	0	0	0	1
Midwest-NE	0	2	0	2	17
Midwest-OH	2	50	4	33	220
Midwest-SD	0	0	0	1	1
Midwest-WI	0	10	5	7	56
Northeast-NY	48	109	39	34	314
Northeast-PA	9	69	18	25	192
South-AL	0	0	1	0	2
South-AR	5	7	1	4	26
South-DC	0	9	0	0	1
South-DE	0	3	0	0	10
South-FL	9	81	75	53	151
South-GA	8	148	6	34	151
South-KY	1	14	0	22	98
South-LA	0	46	1	1	47
South-MD	4	19	0	1	14
South-MS	1	31	0	2	11
South-NC	2	11	1	3	20
South-OK	0	1	0	3	5
South-SC	0	26	1	2	36
South-TN	0	22	2	2	30
South-TX	20	70	143	6	186
South-VA	5	44	2	59	77
South-WV	0	2	1	1	15
West-AK	0	0	1	1	0
West-AZ	5	13	16	2	98
West-CA	191	132	421	196	537
West-CO	7	10	14	31	214
West-HI	5	0	0	2	6
West-ID	0	1	0	1	3
West-MT	0	0	0	1	21
West-NM	1	2	16	6	37
West-NV	4	3	5	7	24
West-OR	0	2	3	2	56
West-UT	1	0	5	1	3
West-WA	9	4	6	5	59
West-WY	0	0	0	0	1

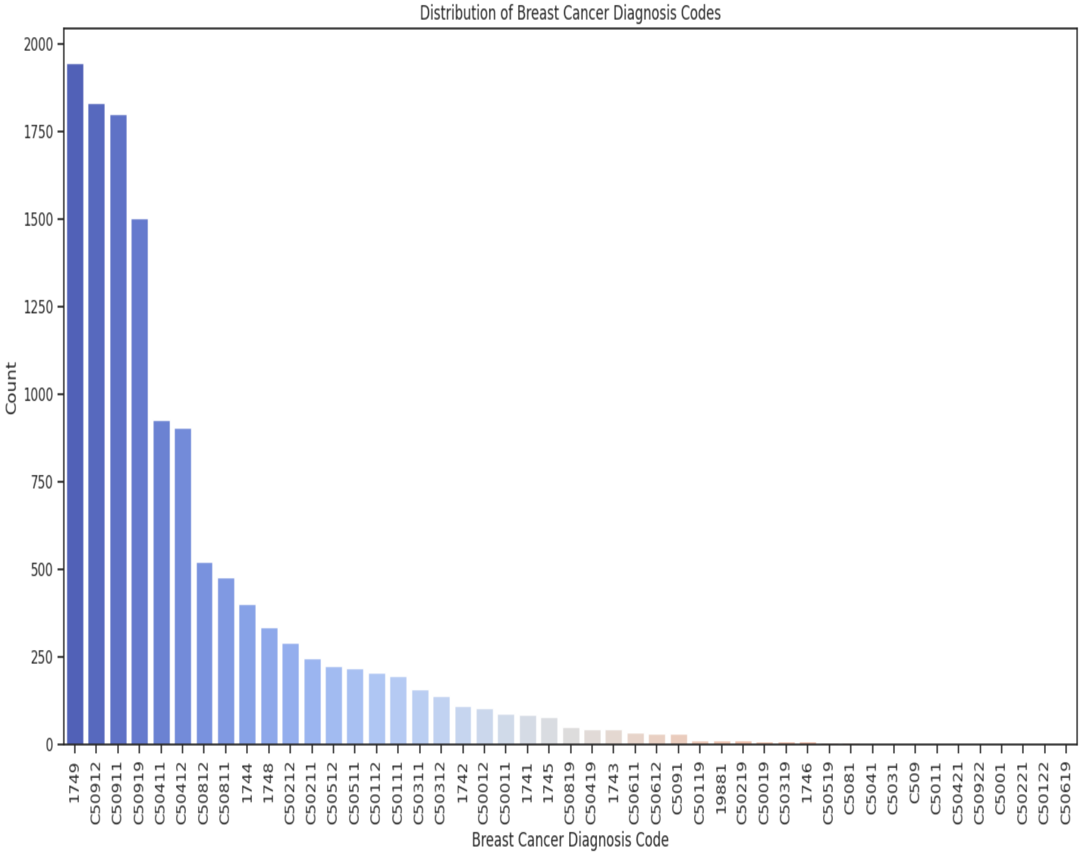
How does the distribution of patient races vary across different regions and states?

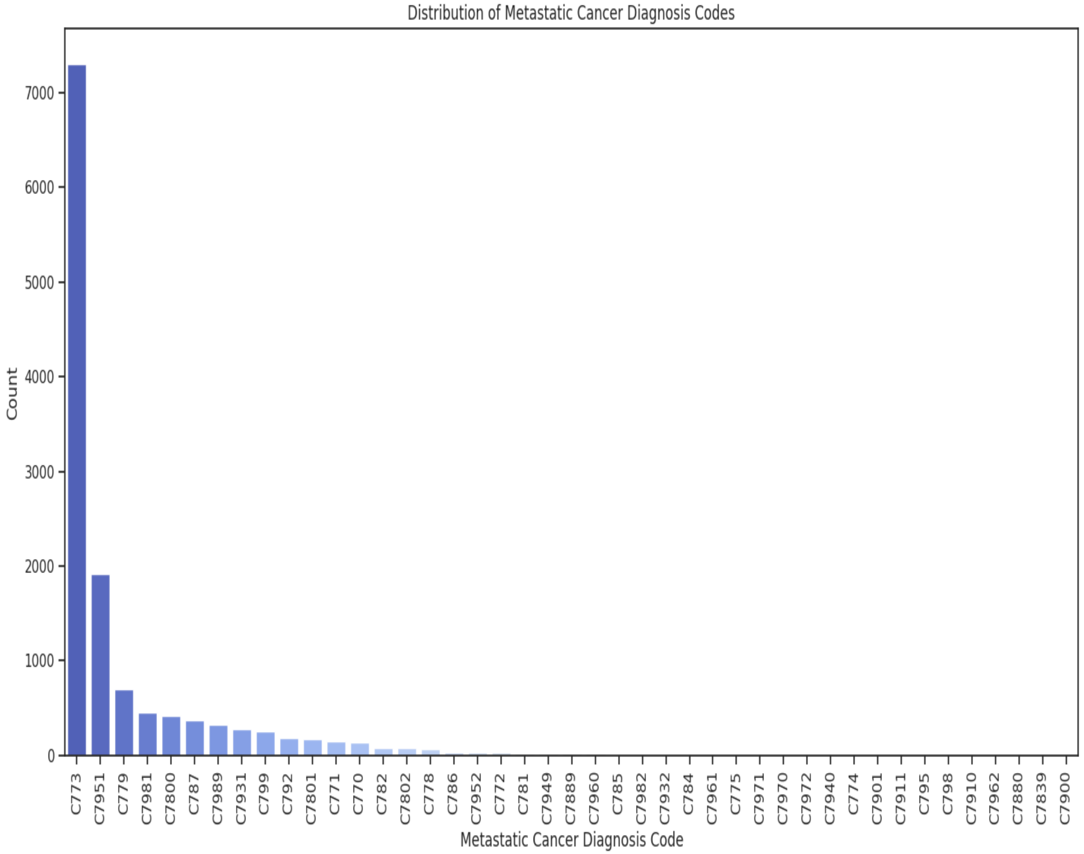
Socioeconomic Indicators



How can we explore the relationships among socioeconomic indicators?

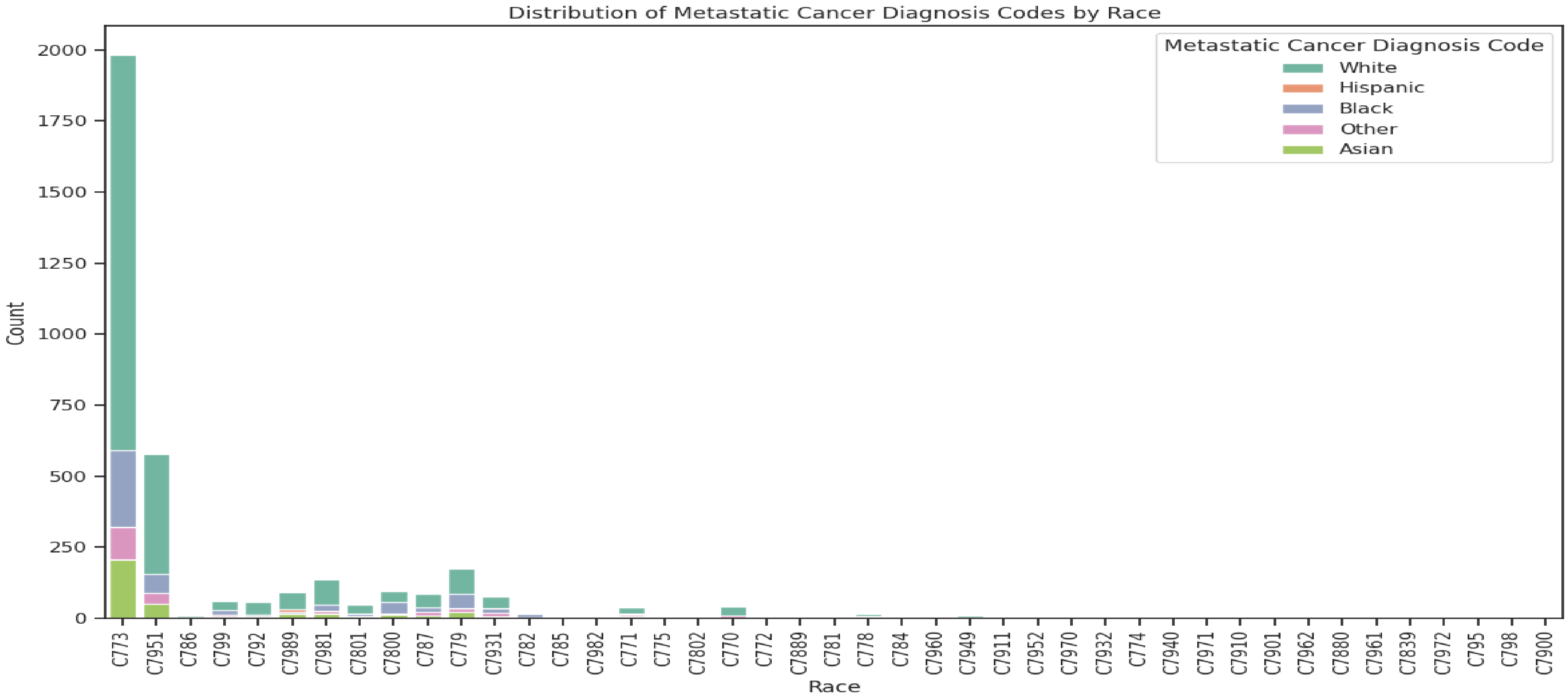
Analysis of Cancer Codes Data





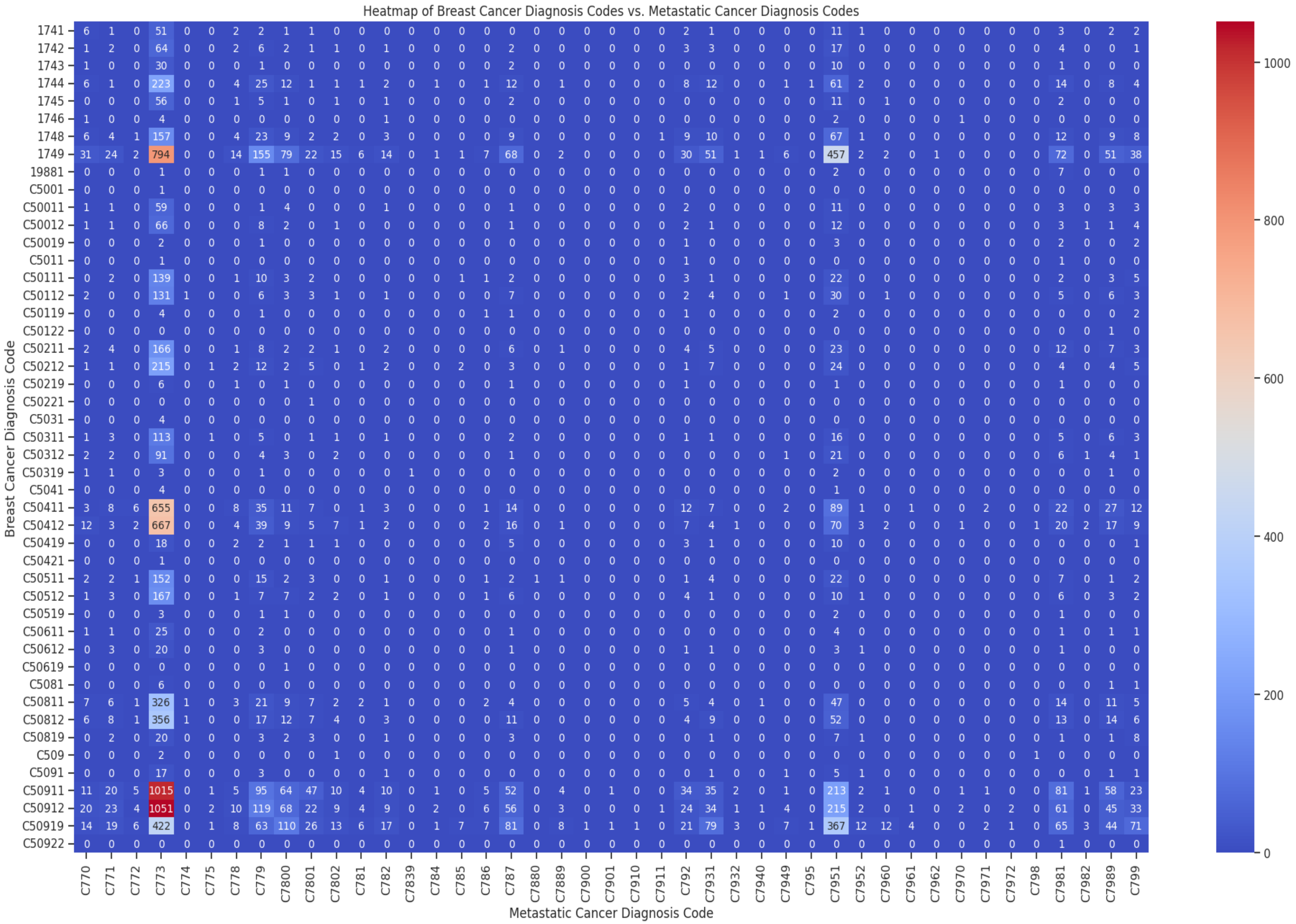
What are the distributions of breast cancer diagnosis codes and metastatic cancer diagnosis codes in the dataset?

Cont.



What is the distribution of metastatic cancer diagnosis codes among different races in the dataset?

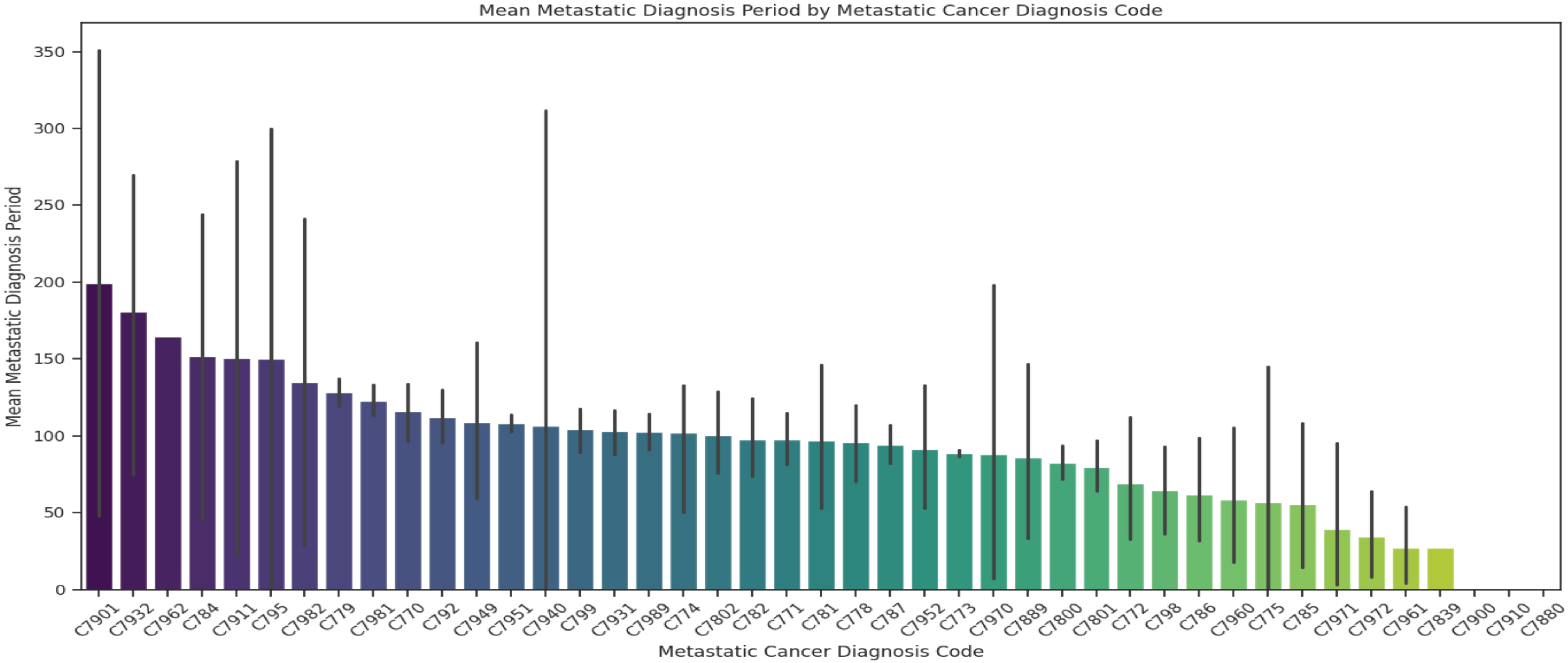
Cont.



	C770	C771	C772	C773	C774	C775	C778	C779	C7800	C7801	C7802	C781	C782	C7839	C784	C785	C786	C787	C7880	C7889	C7900	C7901	C7910	C7911	C792	C7931	C7932	C7940	C7949	C795	C7951	C7952	C7960	C7961	C7962	C7970	C7971	C7972	C798	C7981	C7982	C7989	C799
1741	6	1	0	51	0	0	2	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	0	11	1	0	0	0	0	0	0	0	3	0	2	2
1742	1	2	0	64	0	0	2	6	2	1	1	0	1	0	0	0	0	2	0	0	0	0	0	0	3	3	0	0	0	0	17	0	0	0	0	0	0	0	0	4	0	0	1
1743	1	0	0	30	0	0	0	1	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	1	0	0	0
1744	6	1	0	223	0	0	4	25	12	1	1	1	2	0	1	0	1	12	0	1	0	0	0	0	8	12	0	0	1	1	61	2	0	0	0	0	0	0	0	14	0	8	4
1745	0	0	0	56	0	0	1	5	1	0	1	0	1	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	11	0	1	0	0	0	0	0	0	2	0	0	0
1746	1	0	0	4	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	1	0	0	0	0	0	0	0
1748	6	4	1	157	0	0	4	23	9	2	2	0	3	0	0	0	0	9	0	0	0	0	0	1	9	10	0	0	0	0	67	1	0	0	0	0	0	0	0	12	0	9	8
1749	31	24	2	794	0	0	14	155	79	22	15	6	14	0	1	1	7	68	0	2	0	0	0	0	30	51	1	1	6	0	457	2	2	0	1	0	0	0	0	72	0	51	38
19881	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	7	0	0	0
C5001	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C50011	1	1	0	59	0	0	0	1	4	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	2	0	0	0	0	0	11	0	0	0	0	0	0	0	0	3	0	3	3
C50012	1	1	0	66	0	0	0	8	2	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	2	1	0	0	0	0	12	0	0	0	0	0	0	0	0	3	1	1	4
C50019	0	0	0	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	3	0	0	0	0	0	0	0	0	2	0	0	2
C5011	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
C50111	0	2	0	139	0	0	1	10	3	2	0	0	0	0	0	1	1	2	0	0	0	0	0	0	3	1	0	0	0	0	22	0	0	0	0	0	0	0	0	2	0	3	5
C50112	2	0	0	131	1	0	0	6	3	3	1	0	1	0	0	0	0	7	0	0	0	0	0	0	2	4	0	0	1	0	30	0	1	0	0	0	0	0	0	5	0	6	3
C50119	0	0	0	4	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	2
C50122	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
C50211	2	4	0	166	0	0	1	8	2	2	1	0	2	0	0	0	0	6	0	1	0	0	0	0	4	5	0	0	0	0	23	0	0	0	0	0	0	0	0	12	0	7	3
C50212	1	1	0	215	0	1	2	12	2	5	0	1	2	0	0	2	0	3	0	0	0	0	0	0	1	7	0	0	0	0	24	0	0	0	0	0	0	0	0	4	0	4	5
C50219	0	0	0	6	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
C50221	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C5031	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C50311	1	3	0	113	0	1	0	5	0	1	1	0	1	0	0	0	0	2	0	0	0	0	0	0	1	1	0	0	0	0	16	0	0	0	0	0	0	0	0	5	0	6	3
C50312	2	2	0	91	0	0	0	4	3	0	2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	21	0	0	0	0	0	0	0	0	6	1	4	1
C50319	1	1	0	3	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	1	0
C5041	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
C50411	3	8	6	655	0	0	8	35	11	7	0	1	3	0	0	0	1	14	0	0	0	0	0	0	12	7	0	0	2	0	89	1	0	1	0	0	2	0	0	22	0	27	12
C50412	12	3	2	667	0	0	4	39	9	5	7	1	2	0	0	0	2	16	0	1	0	0	0	0	7	4	1	0	0	0	70	3	2	0	0	1	0	0	1	20	2	17	9
C50419	0	0	0	18	0	0	2	2	1	1	1	0	0	0	0	0	0	5	0	0	0	0	0	0	3	1	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	1
C50421	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C50511	2	2	1	152	0	0	0	15	2	3	0	0	1	0	0	0	1	2	1	1	0	0	0	0	1	4	0	0	0	0	22	0	0	0	0	0	0	0	0	7	0	1	2
C50512	1	3	0	167	0	0	1	7	7	2	2	0	1	0	0	0	1	6	0	0	0	0	0	0	4	1	0	0	0	0	10	1	0	0	0	0	0	0	0	6	0	3	2
C50519	0	0	0	3	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	1	0	0	0
C50611	1	1	0	25	0	0	0	2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	1	0	1	1
C50612	0	3	0	20	0	0	0	3	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	3	1	0	0	0	0	0	0	0	1	0	0	0
C50619	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C5081	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
C50811	7	6	1	326	1	0	3	21	9	7	2	2	1	0	0	0	2	4	0	0	0	0	0	0	5	4	0	1	0	0	47	0	0	0	0	0	0	0	0	14	0	11	5
C50812	6	8	1	356	1	0	0	17	12	7	4	0	3	0	0	0	0	11	0	0	0	0	0	0	4	9	0	0	0	0	52	0	0	0	0	0	0	0	0	13	0	14	6
C50819	0	2	0	20	0	0	0	3	2	3	0	0	1	0	0	0	0	3	0	0	0	0	0	0	0	1	0	0	0	0	7	1	0	0	0	0	0	0	0	1	0	1	8
C509	0	0	0	2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
C5091	0	0	0	17	0	0	0	3	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	5	1	0	0	0	0	0	0	0	0	0	1	1
C50911	11	20	5	1015	0	1	5	95	64	47	10	4	10	0	1	0	5	52	0	4	0	1	0	0	34	35	2	0	1	0	213	2	1	0	0	1	1	0	0	81	1	58	23
C50912	20	23	4	1051	0	2	10	119	68	22	9	4	9	0	2	0	6	56	0	3	0	0	0	1	24	34	1	1	4	0	215	2	0	1	0	2	0	2	0	61	0	45	33
C50919	14	19	6	422	0	1	8	63	110	26	13	6	17	0	1	7	7	81	0	8	1	1	1	0	21	79	3	0	7	1	367	12	12	4	0	0	2	1	0	65	3	44	71
C50922	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0

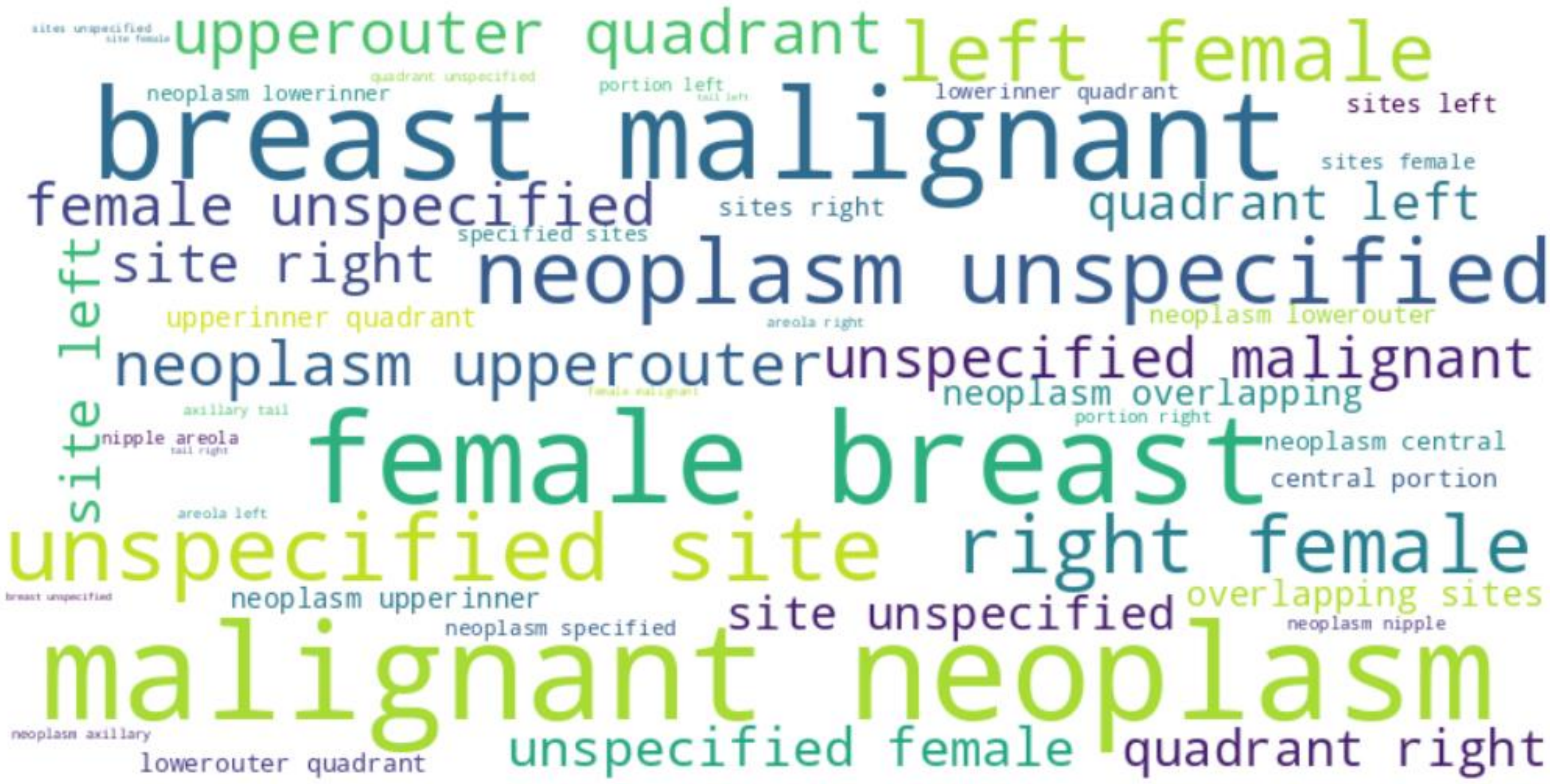
What is the co-occurrence pattern between breast cancer diagnosis codes and metastatic cancer diagnosis codes based on the dataset?

Cont.



What is the mean metastatic diagnosis period for different metastatic cancer diagnosis codes in the dataset?

Word Cloud of Caner Code Description



THANK YOU