Codes:

Age chart:  
  
import pandas as pd

import matplotlib.pyplot as plt

# Load the dataset

df = pd.read\_csv('Breast\_Cancer.csv')

# Categorize patients based on Age

Age\_bins = [0, 20, 30, 40, 50, 60, 70, 80, 90, 100]

Age\_labels = ['0-20', '21-30', '31-40', '41-50', '51-60', '61-70', '71-80', '81-90', '91-100']

df['Age\_Category'] = pd.cut(df['Age'], bins=Age\_bins, labels=Age\_labels, right=False)

# Count the number of patients in each age category

age\_counts = df['Age\_Category'].value\_counts().sort\_index()

# Plot the chart

plt.bar(age\_counts.index, age\_counts.values, color='skyblue')

plt.title('Distribution of Patients Across Age Categories')

plt.xlabel('Age Category')

plt.ylabel('Number of Patients')

plt.xticks(rotation=45)

plt.grid(axis='y', linestyle='--', alpha=0.7)

plt.tight\_layout()

plt.show()

===  
  
Age and race

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

from scipy import stats

# Load the dataset

df = pd.read\_csv('Breast\_Cancer.csv')

# Plot the correlation between age and race using a box plot

plt.figure(figsize=(10, 6))

sns.boxplot(x='Race', y='Age', data=df, palette='Set3')

plt.title('Correlation between Age and Race')

# Perform t-tests for pairwise comparisons

white\_age = df[df['Race'] == 'White']['Age']

black\_age = df[df['Race'] == 'Black']['Age']

other\_age = df[df['Race'] == 'Other']['Age']

white\_black\_pval = stats.ttest\_ind(white\_age, black\_age).pvalue

white\_other\_pval = stats.ttest\_ind(white\_age, other\_age).pvalue

black\_other\_pval = stats.ttest\_ind(black\_age, other\_age).pvalue

# Annotate with p-values and significance levels

plt.text(0, white\_age.max() + 2, f'p = {white\_black\_pval:.3f}', ha='center')

plt.text(1, black\_age.max() + 2, f'p = {white\_other\_pval:.3f}', ha='center')

plt.text(2, other\_age.max() + 2, f'p = {black\_other\_pval:.3f}', ha='center')

plt.xlabel('Race')

plt.ylabel('Age')

plt.show()

===  
  
marital status  
  
import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

from scipy import stats

# Load the dataset

df = pd.read\_csv('Breast\_Cancer.csv')

# Categorize marital status into 'married' and 'currently single'

df['Marital\_Status\_Category'] = df['Marital Status'].apply(lambda x: 'married' if x == 'Married' else 'currently single')

# Plot the distribution of patient data across marital status categories

plt.figure(figsize=(8, 6))

sns.countplot(x='Marital\_Status\_Category', data=df, palette='Set2')

plt.title('Distribution of Patients Across Marital Status Categories')

plt.xlabel('Marital Status')

plt.ylabel('Number of Patients')

# Perform the T-test for significance

married\_age = df[df['Marital\_Status\_Category'] == 'married']['Age']

single\_age = df[df['Marital\_Status\_Category'] == 'currently single']['Age']

p\_val = stats.ttest\_ind(married\_age, single\_age).pvalue

# Annotate with p-value and significance level

plt.text(0, married\_age.count() + 2, f'p = {p\_val:.3f}', ha='center')

plt.text(1, single\_age.count() + 2, '', ha='center') # No p-value for single group

# Determine significance based on p-value

if p\_val < 0.05:

plt.text(0.5, married\_age.count() + 10, 'Significant', ha='center', color='red')

else:

plt.text(0.5, married\_age.count() + 10, 'Not Significant', ha='center', color='green')

plt.show()

===  
Boxplot (strogen and survival month  
**Distribution of Survival Months:**

* The boxplot reveals the distribution of survival months for patients categorized by their estrogen status (Positive vs. Negative).

**Median Survival Month:**

* The horizontal line within each box represents the median survival month. This indicates that for each estrogen status group, half of the patients have survival months that fall above the median and the other half below.

**Interquartile Range (IQR):**

* The box itself represents the interquartile range (IQR). This signifies the middle 50% of the data within each estrogen status group.
  + The top of the box shows the third quartile (Q3), indicating the survival month above which 75% of patients in that group lived.
  + The bottom of the box shows the first quartile (Q1), indicating the survival month below which 25% of patients in that group lived.

**Whiskers and Outliers:**

* The lines extending from the box are called whiskers. They represent the range of data points that fall within 1.5 times the IQR from the quartiles.
  + The upper whisker extends to the highest data point within this range for patients with positive/negative estrogen status.
  + The lower whisker extends to the lowest data point within this range.
* Data points beyond the whiskers are considered outliers and are typically plotted as individual circles. These outliers represent patients with significantly shorter or longer survival months compared to the majority in their respective groups.

**Observations and Potential Insights:**

* **Median Survival Month:** From the boxplot, it appears that the median survival month is likely higher for patients with negative estrogen status compared to positive. This suggests that on average, patients with negative estrogen status might tend to survive longer.
* **Interquartile Range (IQR):** The IQR seems wider for patients with positive estrogen status. This indicates a greater spread in survival months within this group. This wider spread suggests that for patients with positive estrogen status, some individuals might have very short survival months while others might live for a longer time.
* **Outliers:** The presence of outliers on both sides for both positive and negative estrogen status indicates that there are patients whose survival months deviate significantly from the majority in their respective groups. These outliers could be further investigated to understand potential contributing factors.

**Additional Considerations:**

* It's important to acknowledge that this analysis is based solely on the boxplot. Statistical tests would be necessary to determine if the observed differences in median survival months between the two groups are statistically significant.
* Other factors beyond estrogen status might influence survival months. Future analysis could explore additional variables in the dataset to gain a more comprehensive understanding of factors affecting survival rates.

Overall, the boxplot provides a valuable visual summary of the distribution of survival months based on estrogen status. The potential differences in median survival and the wider IQR for the positive group warrant further investigation.

===  
  
  
T stage and survival month  
**Observations from the Box Plot:**

* The median survival month appears to decrease across tumor stages (T1 to T4). This suggests that on average, patients with higher tumor stages tend to have shorter survival months.
* The interquartile range (IQR) seems to widen as the tumor stage increases (T1 to T4). This indicates a greater spread in survival months for patients with more advanced tumor stages. There might be a larger variation in survival outcomes for patients with higher tumor stages.

**T-Test Results:**

* Text annotations are displayed above each box, indicating the p-value from a t-test comparing the survival months of that particular tumor stage with all other stages combined.
* A p-value less than 0.05 (shown in red) is typically considered statistically significant, indicating that the difference in survival months between that stage and the others is unlikely to be due to chance.
* Based on the image, all the p-values are significant (less than 0.05), suggesting that the observed differences in median survival months between tumor stages are statistically significant.

**Overall Interpretation:**

The box plot and t-test results provide strong evidence that tumor stage is correlated with survival months in this breast cancer dataset. Patients with higher tumor stages (more advanced tumors) tend to have lower median survival months and a wider range of survival outcomes compared to those with lower tumor stages. The statistically significant t-test results further reinforce the notion that these observed differences are not random but likely reflect a true association between tumor stage and survival months.

===  
scattered plot  
tumor size and survival month  
  
Based on the image you attached, the scatter plot shows a weak negative relationship between tumor size and survival months in the breast cancer dataset. Here's a more detailed explanation:

**Observations:**

* There's a slight downward trend in the data points, suggesting that larger tumor sizes (on the x-axis) tend to be associated with shorter survival months (on the y-axis). This aligns with the expectation that more advanced tumors might lead to decreased survival times.
* The data points are scattered, indicating a weak negative relationship. This means there's not a very precise or strong correlation between tumor size and survival months. There's a considerable amount of variation in survival months observed across different tumor sizes.
* There are some outliers, which are data points that fall far away from the main cluster. These outliers could represent patients with exceptional circumstances or potential data errors that warrant further investigation.

pen\_spark