

```
In [ ]: # Initialize Otter
import otter
grader = otter.Notebook("cancer_mutations.ipynb")
```

Investigating Cancer Genomics: Mutation Patterns and Patient Outcomes

Created and developed by [Suparna Kompalli](https://www.linkedin.com/in/suparna-kompalli-79463b229/) and [Brandon Concepcion](https://www.linkedin.com/in/brandonconcepcion/), with assistance and supervision by [Jonathan Ferrari](https://www.linkedin.com/in/jonathanferrari/), [Professor Darcie McClelland](https://www.linkedin.com/in/darcie-mcclelland-descalzo-56796b1b/), and [Professor Eric Van Dusen](https://www.linkedin.com/in/ericvd/) as part of our work with UC Berkeley's [College of Computing, Data Science and Society](https://cdss.berkeley.edu/) as well as [El Camino College](https://www.elcamino.edu/)

Breast Cancer

Cancer is a disease of the genome, and every cancer originates from a series of **mutations in DNA** causing cells to grow uncontrollably. But not all cancers are the same — and neither are the mutations that drive them.

In this notebook, you will explore real breast cancer patient data to understand how different gene mutations relate to cancer subtypes and patient outcomes. You'll investigate questions like:

- Which gene mutations are most common in breast cancer?
- How do mutation rates vary across molecular subtypes
- Are certain mutations associated with more aggressive disease or lower survival?
- Do mutations correlate with how patients are treated (e.g., chemotherapy)?

Understanding these patterns can help doctors better **stratify risk**, **choose therapies**, and even **design new treatments**.

Run the cell below to import all of the necessary libraries for this assignment!

```
In [1]: import pandas as pd
from utils import *
from mut_widget import *
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

The METABRIC Dataset:

We'll be using real patient data from the **METABRIC (Molecular Taxonomy of Breast Cancer International Consortium)** study. We found this dataset on [Kaggle \(https://www.kaggle.com/\)](https://www.kaggle.com/), a widely-used, data-driven platform that hosts thousands of high-quality datasets. The link to our dataset can be accessed [here!](https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric) (<https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric>)

An excerpt from the dataset's Kaggle description provides a detailed overview of the data and its context.

"The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) database is a Canada-UK Project which contains targeted sequencing data of 1,980 primary breast cancer samples. Clinical and genomic data was downloaded from cBioPortal. The dataset was collected by Professor Carlos Caldas from Cambridge Research Institute and Professor Sam Aparicio from the British Columbia Cancer Centre in Canada and published on Nature Communications (Pereira et al., 2016). It was also featured in multiple papers including Nature and others:

- [Associations between genomic stratification of breast cancer and centrally reviewed tumor pathology in the METABRIC cohort \(https://www.nature.com/articles/s41523-018-0056-8\)](https://www.nature.com/articles/s41523-018-0056-8)
- [Predicting Outcomes of Hormone and Chemotherapy in the Molecular Taxonomy of Breast Cancer International Consortium \(METABRIC\) Study by Biochemically-inspired Machine Learning \(https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5461908/\)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5461908/)

Important Note: Every patient in this dataset already has breast cancer.

We'll start by using the `pandas` (<https://pandas.pydata.org/>) library. `pandas` is a powerful and widely-used Python tool for working with tabular data. It allows us to read, organize, and manipulate datasets efficiently, especially when dealing with spreadsheets or CSV (Comma Separate Value) files. Let's load our data using `pandas` :

```
In [2]: df = pd.read_csv("METABRIC.csv")
df.head()

/var/folders/rk/wdhf611s3h56skfqfbbm_25m0000gn/T/ipykernel_48413/2670141452.py:
1: DtypeWarning: Columns (678,688,690,692) have mixed types. Specify dtype optio
n on import or set low_memory=False.
df = pd.read_csv("METABRIC.csv")
```

```
Out[2]:
```

	patient_id	age_at_diagnosis	type_of_breast_surgery	cancer_type	cancer_type_detailed	cellularity	chemoth
0	0	75.65	MASTECTOMY	Breast Cancer	Breast Invasive Ductal Carcinoma	NaN	
1	2	43.19	BREAST CONSERVING	Breast Cancer	Breast Invasive Ductal Carcinoma	High	
2	5	48.87	MASTECTOMY	Breast Cancer	Breast Invasive Ductal Carcinoma	High	
3	6	47.68	MASTECTOMY	Breast Cancer	Breast Mixed Ductal and Lobular Carcinoma	Moderate	
4	8	76.97	MASTECTOMY	Breast Cancer	Breast Mixed Ductal and Lobular Carcinoma	High	

5 rows x 693 columns

```
In [3]: print(f"There are {len(df.columns)} columns in our dataset!")

There are 693 columns in our dataset!
```

Question 0.1 Wow, that’s quite a few columns! Can you think of a reason why this dataset might include so many??

Type your answer here, replacing this text.

SOLUTION: The METABRIC dataset brings together a wide range of biomedical information about each patient. While some columns describe clinical characteristics like age, tumor size, treatment, and survival, the majority of the columns represent whether specific genes are mutated or not. Since there are hundreds of genes that can be relevant to cancer development and progression, each tracked mutation gets its own column.

Below we’ll define some of the most relevant columns to focus on, these will be the primary features we use throughout the rest of the notebook. Feel free to reference the Kaggle [dataset \(https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric\)](https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric) on your own to explore other interesting aspects of the dataset!

```
In [4]: # Manually defining important clinical and outcome columns
clinical_columns = [
    'patient_id',
    'age_at_diagnosis',
    'type_of_breast_surgery',
    'cancer_type_detailed',
    'pam50+_claudin-low_subtype',
    'chemotherapy',
    'hormone_therapy',
    'radio_therapy',
    'tumor_size',
    'tumor_stage',
    'overall_survival',
    'overall_survival_months'
]

# Select all columns that end with '_mut'
mutation_columns = [col for col in df.columns if col.endswith('_mut')]

# Combine the clinical + mutation columns
relevant_columns = clinical_columns + mutation_columns

# Filter the dataframe
filtered_df = df[relevant_columns]

# Display the shape to confirm how much we reduced it
print("Filtered Dataframe shape:", filtered_df.shape[0], "rows and", filtered_df.shape[1], "columns")
filtered_df.head()
```

Filtered Dataframe shape: 1904 rows and 185 columns

Out[4]:

	patient_id	age_at_diagnosis	type_of_breast_surgery	cancer_type_detailed	pam50+_claudin-low_subtype	chemotherapy
0	0	75.65	MASTECTOMY	Breast Invasive Ductal Carcinoma	claudin-low	0
1	2	43.19	BREAST CONSERVING	Breast Invasive Ductal Carcinoma	LumA	0
2	5	48.87	MASTECTOMY	Breast Invasive Ductal Carcinoma	LumB	1
3	6	47.68	MASTECTOMY	Breast Mixed Ductal and Lobular Carcinoma	LumB	1
4	8	76.97	MASTECTOMY	Breast Mixed Ductal and Lobular Carcinoma	LumB	1

5 rows × 185 columns

Python Widgets allow us to build interactive elements such as dropdowns, sliders, and checkboxes. They're particularly useful for dynamically visualizing and filtering datasets without having to rewrite your code.

Question 0.2. Below, we've created a widget for selecting certain columns from our dataset. Use the dropdown menu to choose a column and describe how it relates to a concept we've covered in this course!

```
In [5]: # Link dropdown selection to function
dropdown.observe(show_description, names='value')

# Display the widget
display(dropdown, output)

Dropdown(description='Column:', layout=Layout(width='60%'), options=('patient_id', 'age_at_diagnosis', 'type_o...

Output()
```

Type your answer here, replacing this text.

SOLUTION: Answers will vary, but any reasonable explanation/justification will receive credit!

Section 1: Frequency

Imagine for a second you're a doctor, or some kind of biomedical researcher, trying to understand the genetic factors behind breast cancer. One of your goals might be to identify which genes are most commonly mutated among patients, as this information can guide diagnoses, help risk assessment, or even formulate treatment strategies.

In our dataset, we've **modified** the mutation data to make it easier to analyze. For each breast cancer patient, *every gene is represented using binary values*:

- 0 – The patient does not have a mutation in that gene
- 1 – The patient has at least one type of mutation in that gene

For example, if a patient has **any kind** of mutation in their TP53 gene, but not in BRCA1, their data would show:

- TP53 = 1
- BRCA1 = 0

This simplification allows us to more easily quantify and compare mutation frequencies across the entire patient population.

Run the cell below to generate a bar chart showing the proportion of patients who carry mutations in several key cancer-related genes. Afterward, you'll answer a few questions based on the patterns you observe in the chart.

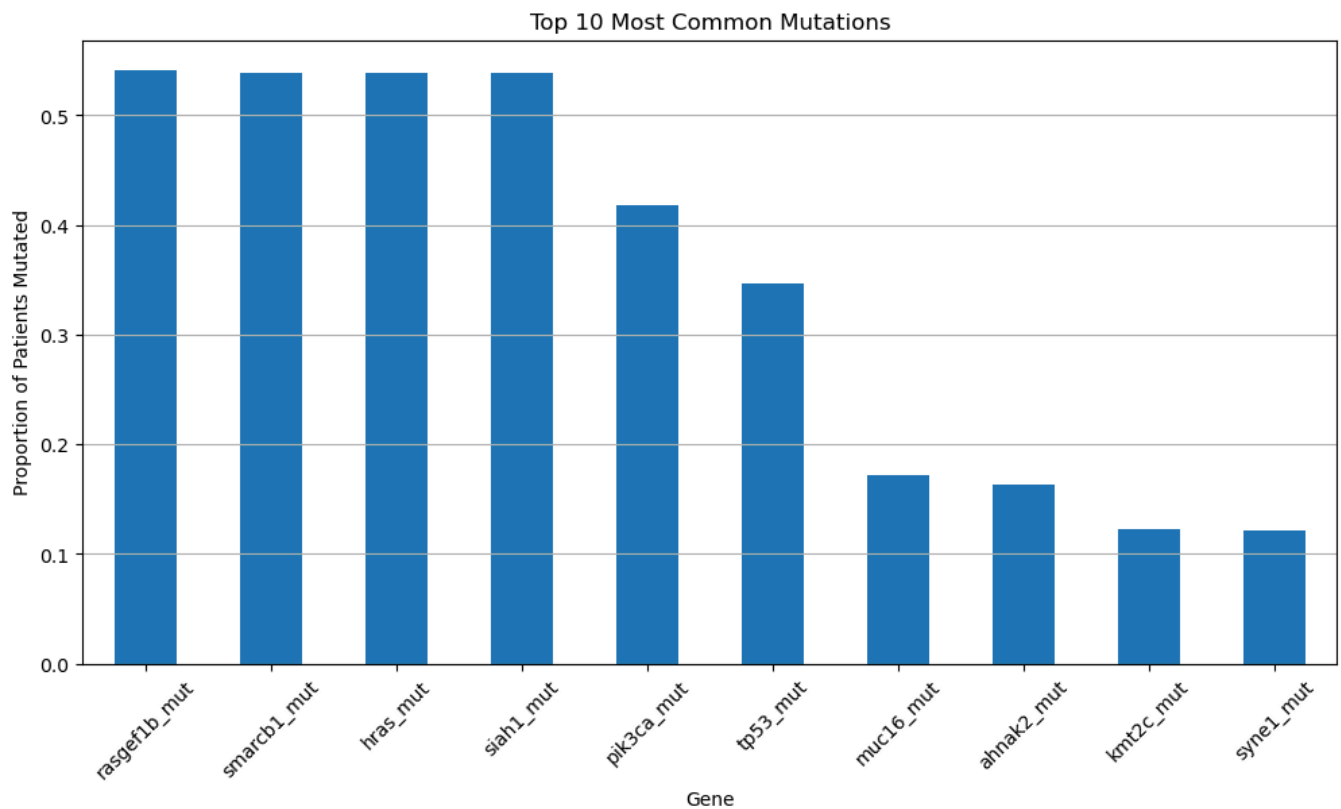
```
In [6]: mutation_cols = [col for col in filtered_df.columns if col.endswith('_mut')]

# Create a binary version of mutation columns
binary_mutations = filtered_df[mutation_cols].map(lambda x: 0 if x == "0" else 1)

# Merge back with the clinical columns
clinical_cols = [col for col in filtered_df.columns if col not in mutation_cols]
eda_df = pd.concat([filtered_df[clinical_cols], binary_mutations], axis=1)

# Mutation frequencies
mutation_rates = eda_df[mutation_cols].mean().sort_values(ascending=False)

# Plot
plt.figure(figsize=(12,6))
mutation_rates[:10].plot(kind='bar')
plt.title('Top 10 Most Common Mutations')
plt.ylabel('Proportion of Patients Mutated')
plt.xlabel('Gene')
plt.xticks(rotation=45)
plt.grid(axis='y')
plt.show()
```



Question 1.1. Which six genes have the highest mutation rates in the dataset? How can you tell?

SOLUTION:

1. sgef1b_mut
2. marcb1_mut
3. hras_mut
4. siah1_mut
5. pik3ca_mut
6. tp53_mut

We can tell which genes have the highest mutation rates by looking at the height of the bars in the chart. Each bar represents the proportion of patients who have a mutation in a specific gene. Since mutation columns are binary (0 = no mutation, 1 = mutation), the average value of each column tells us how frequently that gene is mutated across the dataset. The six genes with the highest bars—sgef1b_mut, marcb1_mut, hras_mut, siah1_mut, pik3ca_mut, and tp53_mut—therefore have the highest mutation rates.

Question 1.2. Do any of these genes look familiar, based on what we've discussed in class? Explain what you remember about them and why they might be important in the context of cancer.

Type your answer here, replacing this text.

SOLUTION: Answers will vary, but any reasonable explanation/justification will receive credit!

Question 1.3. Choose one of the top six most frequently mutated genes and propose a hypothesis why this gene might be so commonly mutated in breast cancer patients. Why do you think a mutation in this particular gene could be especially indicative of cancer development?

Type your answer here, replacing this text.

SOLUTION: Answers will vary, here are some possible hypotheses

- **TP53:** TP53 is a tumor suppressor gene that protects the genome by regulating cell division and repairing DNA. If TP53 is mutated, cells lose this checkpoint and can grow uncontrollably. Its high mutation rate in breast cancer likely reflects its key role in preventing tumors.
- **PIK3CA:** PIK3CA is part of a signaling pathway that promotes cell survival and growth. Mutations in this gene may keep the pathway turned "on," leading to excessive cell division — a hallmark of cancer. This makes it a common driver mutation in breast cancer.
- **SIAH1:** SIAH1 helps degrade damaged or unnecessary proteins. If it's mutated, that system might fail, leading to a buildup of proteins that disrupt normal cell function. This disruption can trigger pathways that support cancer development.
- **HRAS:** HRAS is involved in sending signals for cells to grow. When mutated, it can act like a stuck accelerator pedal — causing cells to grow continuously. This uncontrolled growth is one of the key features of cancer.

Section 2: Aggressive

Breast Cancer Subtypes and Mutation Patterns

Breast cancer is not a single disease — it includes many **molecular subtypes**. One common classification is called **PAM50**, which includes:

- **Luminal A** – less aggressive, hormone-positive
- **Luminal B** – more aggressive than Luminal A
- **HER2-enriched** – driven by HER2 amplification
- **Basal-like** – most aggressive, often triple-negative
- **Normal-like** – rare, resembles healthy breast tissue

In this section, we'll see how gene mutations vary across these subtypes.

We'll begin by looking at **TP53**, which is often referred to as the “*guardian of the genome*.” This gene plays a critical role in regulating cell growth and initiating repair when DNA damage is detected.

When TP53 is mutated, these protective functions become disrupted. Thus, damaged cells can begin to grow and divide unchecked. This loss of control is a hallmark of more aggressive forms of cancer, which is why TP53 mutations are so frequently observed in cancer patients.

We encourage you to learn more about TP53 [here!](https://medlineplus.gov/genetics/gene/tp53/) (<https://medlineplus.gov/genetics/gene/tp53/>)

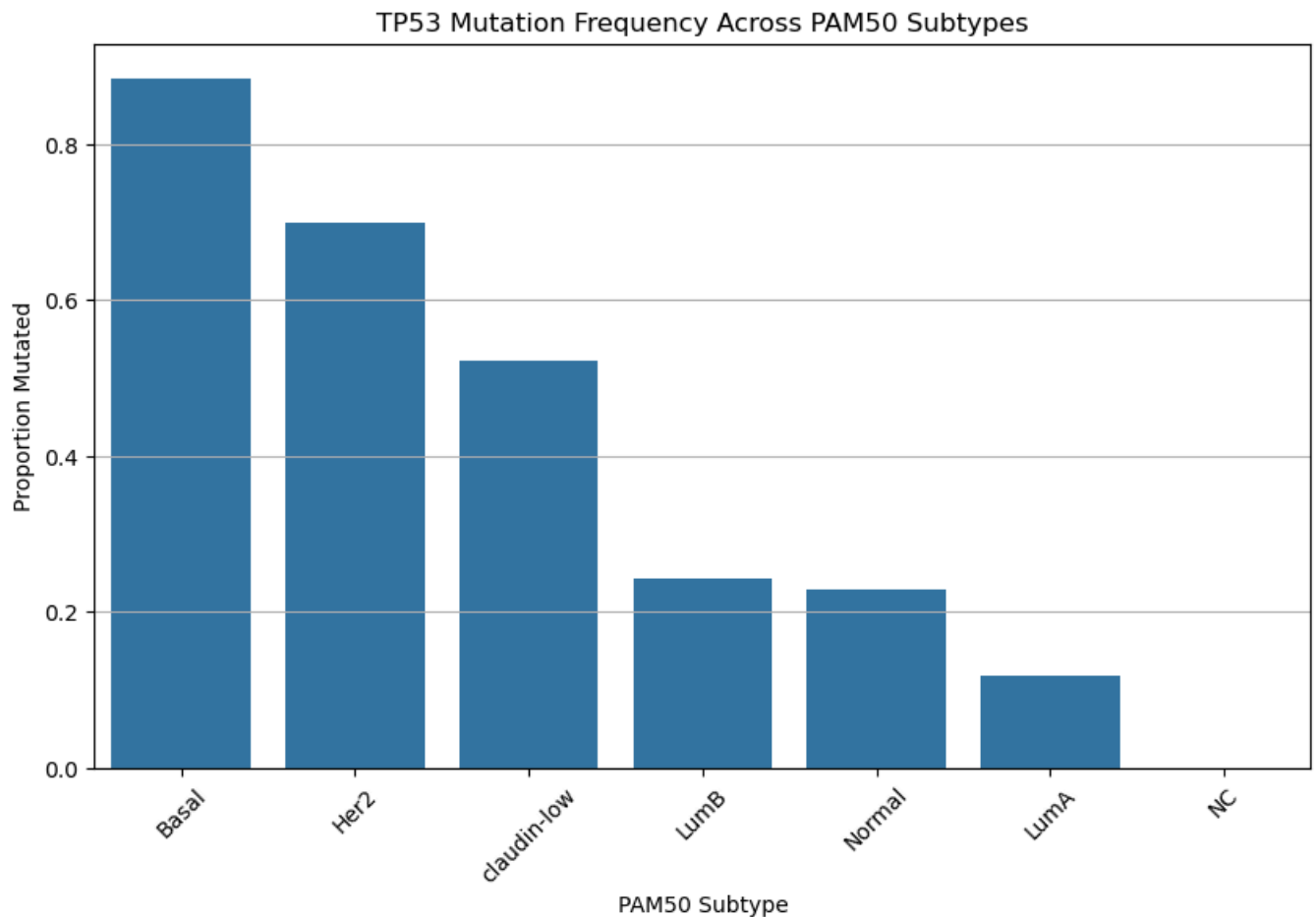
Bar Charts

Run the cell below. Please do not modify any of the lines—this cell has been pre-written to generate a specific visualization you'll use to answer the following questions.


```
In [7]: mutation = 'tp53_mut'

mutation_by_subtype = eda_df.groupby('pam50+_claudin-low_subtype')[mutation].mean().sort_values(ascending=False)

plt.figure(figsize=(10,6))
sns.barplot(x=mutation_by_subtype.index, y=mutation_by_subtype.values)
plt.title('TP53 Mutation Frequency Across PAM50 Subtypes')
plt.ylabel('Proportion Mutated')
plt.xlabel('PAM50 Subtype')
plt.xticks(rotation=45)
plt.grid(axis='y')
plt.show()
```



Based on the chart above, do your best to answer the following questions:

*Hint: TP53 mutations are often found in more **aggressive** cancers.*

Question 2.1. Which cancer subtype shows the highest frequency of TP53 mutation? Which shows the lowest?

Type your answer here, replacing this text.

SOLUTION: The Basal subtype shows the highest frequency of TP53 mutation, with a mutation proportion above 0.85. The LumA (Luminal A) subtype shows the lowest frequency, with a mutation proportion just above 0.1.

Question 2.2. TP53 mutations are highly frequent in some breast cancer subtypes but almost absent in others. Why do you think TP53 mutations might be more common in certain subtypes? What does this suggest about the biology of these cancers?

Type your answer here, replacing this text.

SOLUTION: TP53 mutations are more common in certain breast cancer subtypes, such as like Basal and Her2, because these subtypes tend to be more aggressive and less differentiated. TP53 is a tumor suppressor gene that plays a critical role in regulating cell division and preventing the formation of tumors. When TP53 is mutated, cells can grow uncontrollably, which may contribute to the rapid progression and poor prognosis seen in these aggressive subtypes. The lower frequency of TP53 mutations in subtypes like LumA suggests that these cancers may rely on different molecular pathways for tumor development and progression.

Question 2.3. Based on the plot alone, can we conclude that if TP53 mutation exists, it causes a subtype to be more aggressive? Why or why not?

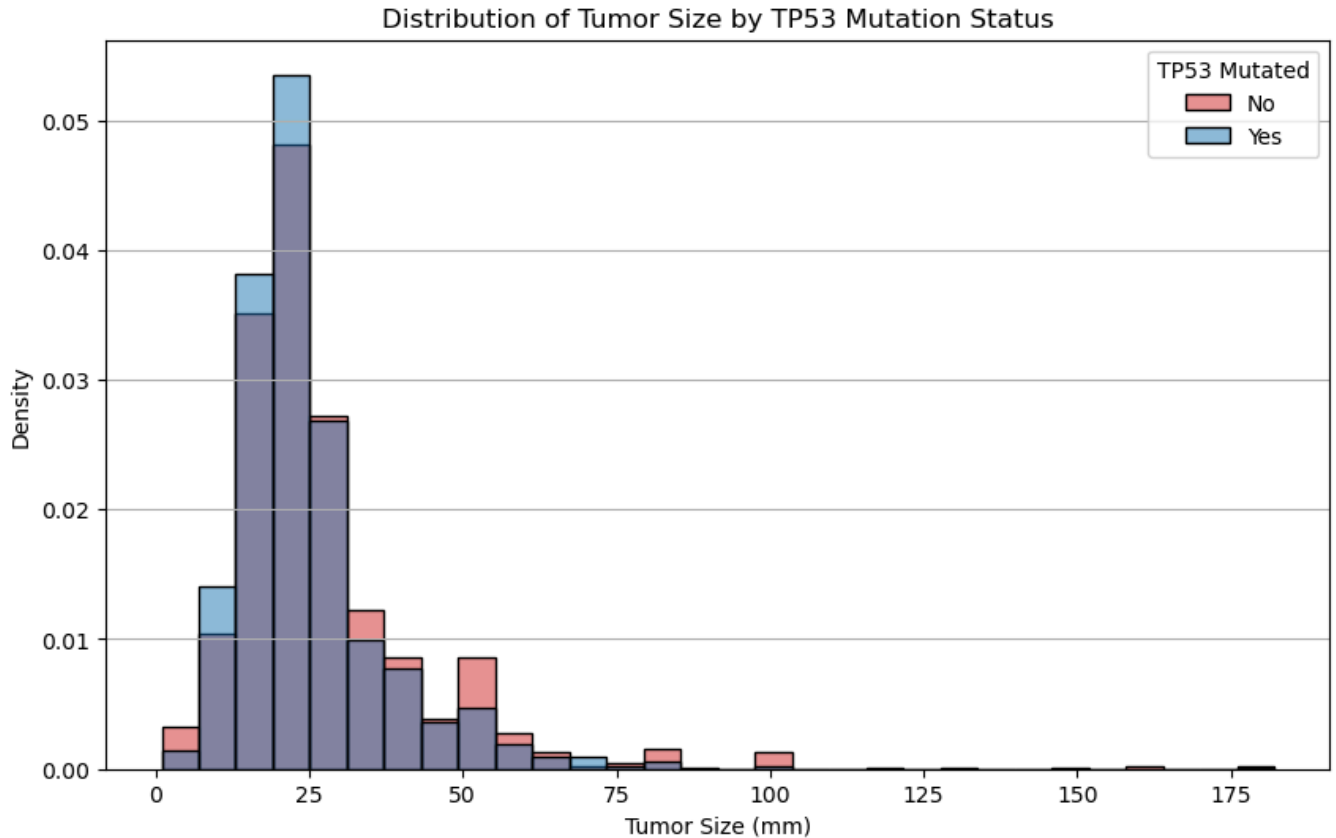
Type your answer here, replacing this text.

SOLUTION: No, we cannot conclude causation based on the plot alone. The plot shows a correlation between TP53 mutation frequency and certain aggressive cancer subtypes, but it does not provide evidence that TP53 mutations cause the aggressiveness. To determine causation, we would need additional experimental or longitudinal data—such as functional studies showing how TP53 mutations directly lead to aggressive behavior in cells, or time-series data showing changes in tumor progression after a TP53 mutation occurs.

Histograms

Run the cell below. Please do not modify any of the lines—this cell has been pre-written to generate a specific visualization you'll use to answer the following questions.

```
In [8]: plt.figure(figsize=(10,6))
sns.histplot(data=eda_df, x='tumor_size', hue='tp53_mut', bins=30, palette=['#1f77
7b4', '#d62728'], stat='density', common_norm=False)
plt.title('Distribution of Tumor Size by TP53 Mutation Status')
plt.xlabel('Tumor Size (mm)')
plt.ylabel('Density')
plt.legend(title='TP53 Mutated', labels=['No', 'Yes'])
plt.grid(axis='y')
plt.show()
```



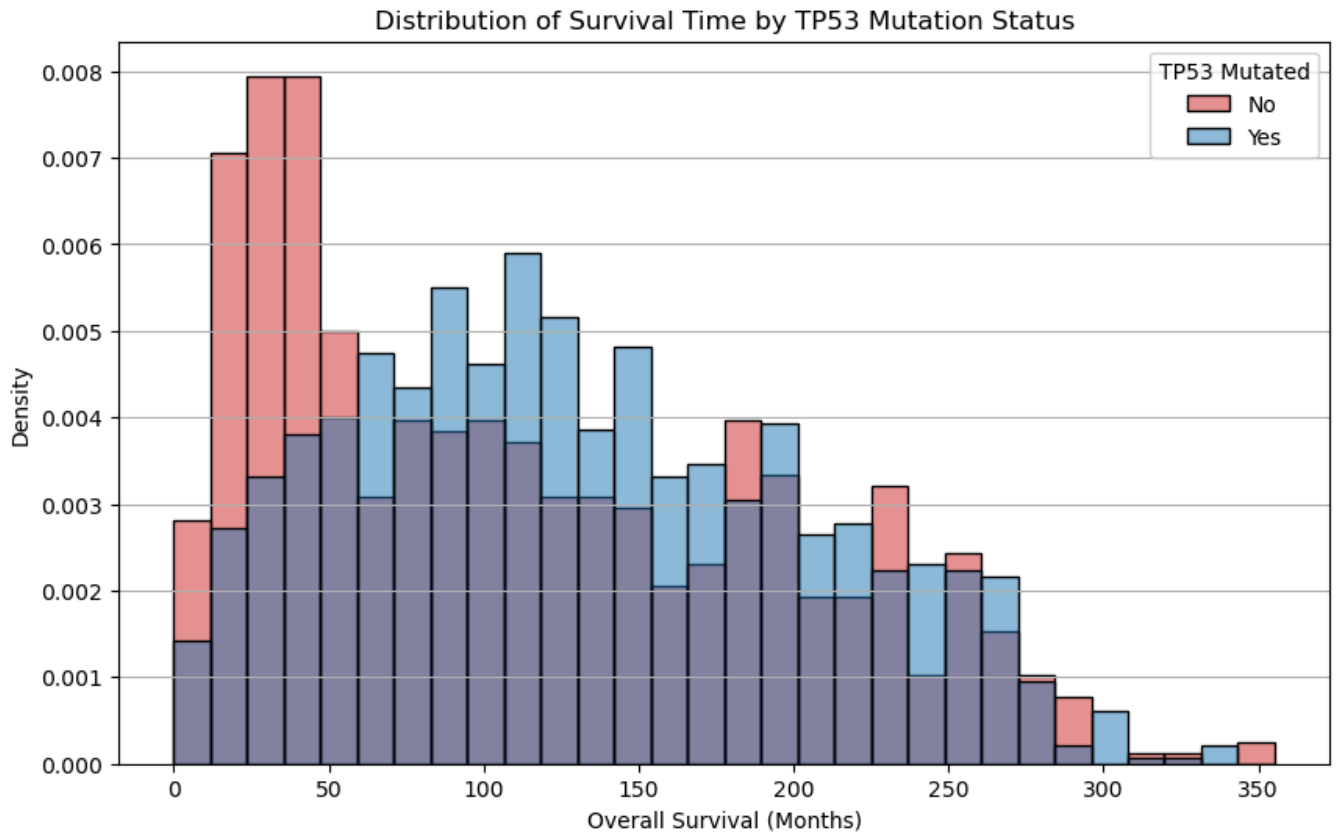
Question 2.4. Compare the shapes of the overlaid distributions. Among all patients *with* breast cancer, do patients with TP53 mutations tend to have larger tumors?

Type your answer here, replacing this text.

SOLUTION: No, based on this plot alone, we cannot conclude that patients with TP53 mutations tend to have larger tumors. The distributions are quite similar, with no clear shift suggesting that one group consistently has larger tumors than the other. Any observed differences are minor and may not be statistically significant without further analysis.

Run the cell below. Please do not modify any of the lines—this cell has been pre-written to generate a specific visualization you'll use to answer the following questions.

```
In [9]: plt.figure(figsize=(10,6))
sns.histplot(data=eda_df, x='overall_survival_months', hue='tp53_mut', bins=30, p
alette=['#1f77b4', '#d62728'], stat='density', common_norm=False)
plt.title('Distribution of Survival Time by TP53 Mutation Status')
plt.xlabel('Overall Survival (Months)')
plt.ylabel('Density')
plt.legend(title='TP53 Mutated', labels=['No', 'Yes'])
plt.grid(axis='y')
plt.show()
```



Question 2.5. Which group seems to have longer average overall survival? Note key parts from the graph when writing your response.

Type your answer here, replacing this text.

SOLUTION: Patients with TP53 mutations (blue) appear to have longer average overall survival, as their distribution is more centered around higher survival times (75–150 months). In contrast, the non-mutated group (red) has a sharp peak in early survival (0–50 months), suggesting more early deaths that lower the overall average.

Heat Map

The heatmap below displays the *proportion of patients* within each breast cancer subtype who carry a mutations in the various genes. This plot shows **how frequently** each gene is mutated **within each subtype**.

- Each row corresponds to a gene (e.g., TP53, PIK3CA)
- Each column corresponds to a cancer subtypes (e.g., Basal, Luminal A)
- Each value in a particular cell corresponds to the proportion of patients **with that mutation, among those in the subtype**
 - For example: A value of 0.33 for **AKT1 in NC** means that **33% of NC-like patients** have an AKT1 mutation.

Important Note: These are **not correlations** between genes.

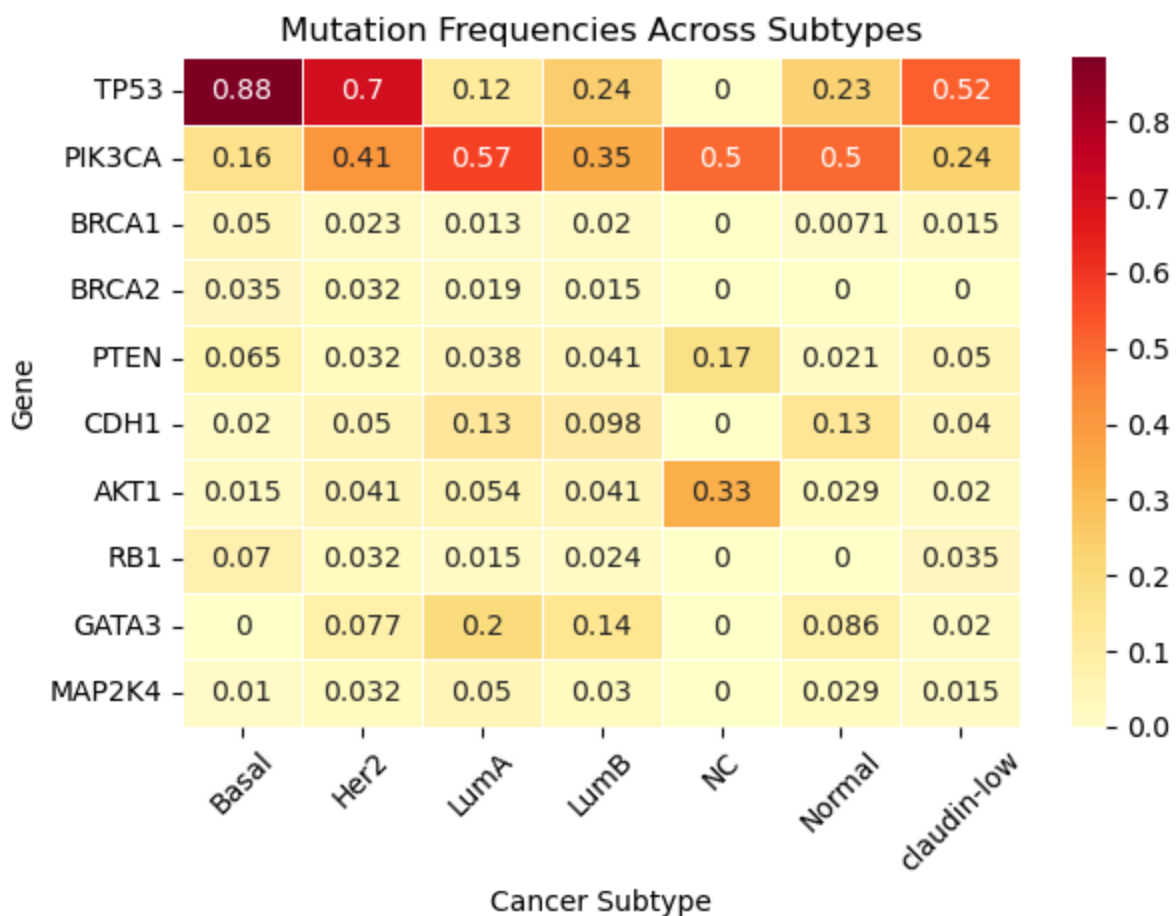
Run the cell below. Please do not modify any of the lines—this cell has been pre-written to generate a specific visualization you'll use to answer the following questions.

```
In [10]: # List of mutation columns you want to include (binary: 0 = no mutation, 1 = mutated)
selected_mutations = [
    'tp53_mut', 'pik3ca_mut', 'brca1_mut', 'brca2_mut',
    'pten_mut', 'cdh1_mut', 'akt1_mut', 'rb1_mut',
    'gata3_mut', 'map2k4_mut'
]

# Group by cancer subtype and compute the mean (i.e. mutation rate) for each gene
mutation_subtype_matrix = eda_df.groupby('pam50+_claudin-low_subtype')[selected_mutations].mean().T

# Optional: Make row labels cleaner
mutation_subtype_matrix.index = [gene.upper().replace('_MUT', '') for gene in mutation_subtype_matrix.index]

sns.heatmap(mutation_subtype_matrix, annot=True, cmap='YlOrRd', linewidths=0.5)
plt.title('Mutation Frequencies Across Subtypes')
plt.ylabel('Gene')
plt.xlabel('Cancer Subtype')
plt.xticks(rotation=45)
plt.yticks(rotation=0)
plt.tight_layout()
plt.show()
```



Question 2.6. Which gene is mutated most often across all subtypes? Be sure to calculate the average mutation frequency across all subtypes when providing your answer.

Type your answer here, replacing this text.

SOLUTION: To determine which gene is mutated most often across all breast cancer subtypes, we calculated the average mutation frequency for each gene based on the heatmap data. The average frequency for PIK3CA is approximately 0.39, while TP53 has a slightly lower average of 0.384. All other genes have considerably lower averages. Therefore, PIK3CA is the gene mutated most frequently on average across all subtypes.

Question 2.7. TP53 mutation is very common in Basal-like tumors but rare in Luminal A. Why could this be the case?

Type your answer here, replacing this text.

SOLUTION: TP53 mutations are much more common in Basal-like tumors because these subtypes tend to be more aggressive and less differentiated, relying heavily on disruptions in tumor suppressor genes like TP53. In contrast, Luminal A tumors are typically less aggressive, slower-growing, and often driven by hormone receptor pathways rather than TP53-related mechanisms—explaining why TP53 mutations are rare in that subtype.

Question 2.8. Are there genes that are highly specific to one or two subtypes? Provide at least two examples.

Type your answer here, replacing this text.

SOLUTION: Yes, the heatmap shows that some genes are highly specific to one or two subtypes. For example, TP53 has a very high mutation frequency in Basal (0.88) and Her2 (0.70) subtypes, but is nearly absent in LumA and NC. Another example is CDH1, which is highly mutated in Her2 (0.13) and Normal-like (0.13) subtypes, but has very low or zero mutation frequency in most other subtypes.


Question 2.9. Are there genes that are rarely mutated in any subtype?

Type your answer here, replacing this text.

SOLUTION: Yes, there are genes that are rarely mutated across all subtypes. For example, MAP2K4 shows consistently low mutation frequencies in every subtype, with values close to or below 0.05. Similarly, BRCA2 has very low mutation rates across the board, never exceeding 0.035.

Congratulations!

Cookie 🍪 congratulates you on finishing the Cancer & Mutations notebook!

 No description has been provided for this image

In this notebook, we

- Explored cancer mutation data to uncover how specific gene mutations, such as TP53, BRCA1/2, and PIK3CA, vary across different cancer subtypes
- Explored how these mutations might relate to tumor size and patient survival
- Utilized visual tools like heatmaps and histograms, to identify patterns in mutation frequency, and began to connect genetic changes to clinical outcomes

We hope you had fun learning throughout this notebook! If you're curious to explore some of these topics further, we encourage you to explore how statistical testing can be used to assess the significance of the patterns we observed. You might also try your hand at time-to-event models such as survival analysis, even incorporating clinical features such as treatment type or patient age. This is just the tip of the iceberg when it comes to the world of genomics!

Below are some opportunities to explore further:

- **El Camino College Research Club:** Participate in hands-on projects like DNA barcoding and species identification. Feel free to reach out to Professor McClelland to learn more!
- **[Mathematics, Engineering, Science Achievement \(MESA\) Program](https://www.elcamino.edu/support/resources/mesa/)** (<https://www.elcamino.edu/support/resources/mesa/>): Offers access to STEM internships, research opportunities, and academic support
- **[Honors Transfer Program \(HTP\)](https://www.elcamino.edu/support/resources/honors-transfer/index.php)** (<https://www.elcamino.edu/support/resources/honors-transfer/index.php>): Engage in research projects and present findings at honors conferences!
- **Internships through the Life Sciences Department:** Opportunities such as the SEA Lab internship focus on marine biology, environmental studies, and conservation!
- **The [Lundquist Institute](https://lundquist.org/)** (<https://lundquist.org/>): Located near El Camino College, this institute offers programs where students can gain firsthand experience in scientific research environments

Submission

Make sure you have run all cells in your notebook in order before running the cell below, so that all images/graphs appear in the output. The cell below will generate a zip file for you to submit. **Please save before exporting!**

These are some submission instructions.

```
In [ ]: # Save your notebook first, then run this cell to export your submission.
        grader.export(run_tests=True)
```