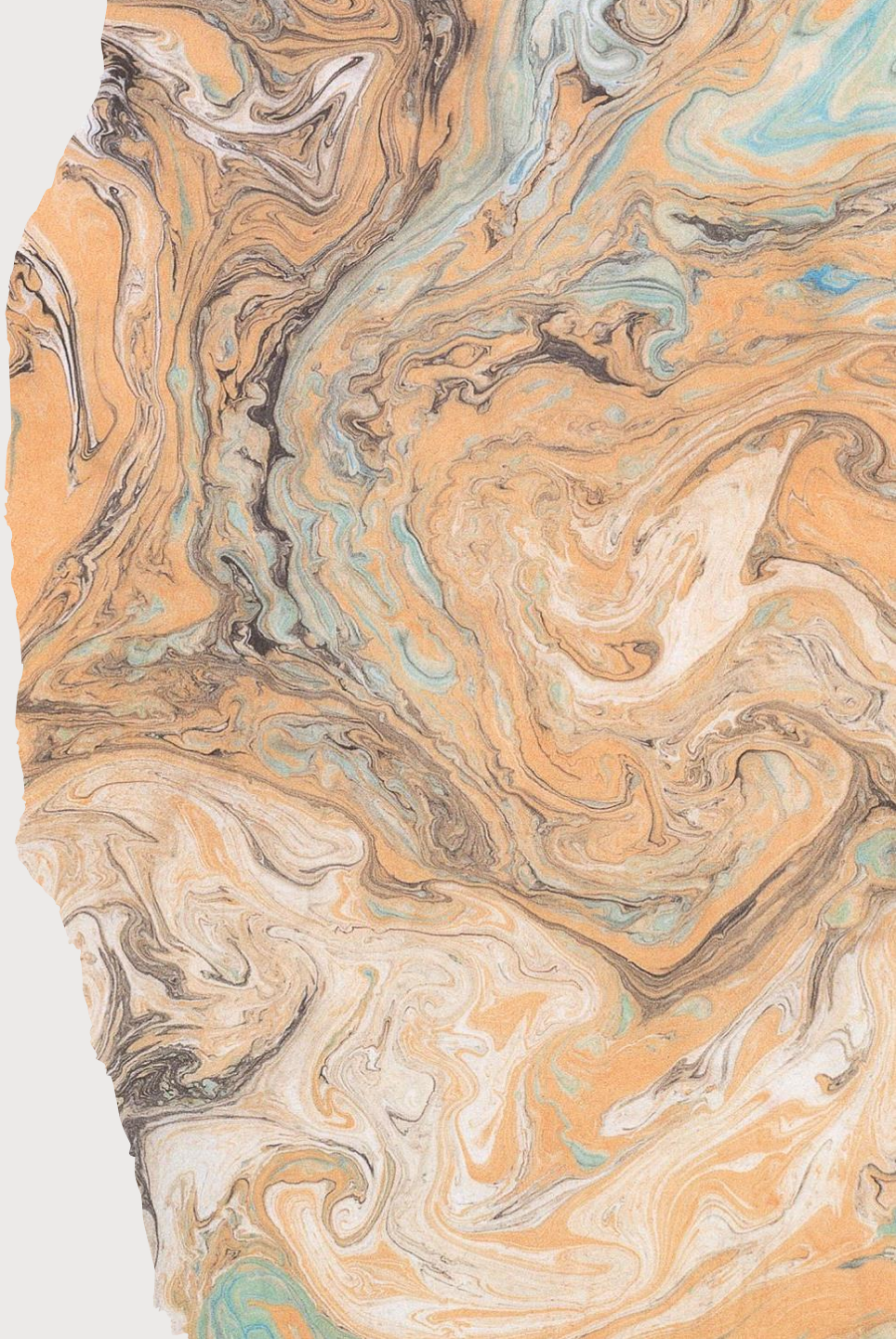
PROJECT 5: JOHNSON AND JOHNSON

Lung Cancer Trial Simulation

Team Members:

Samarth Kumar (210911)
Gaurav Tomar (231080039)
Sunita Kulariya (221106)
Jiyanshu Dhaka (220481)



Task 1

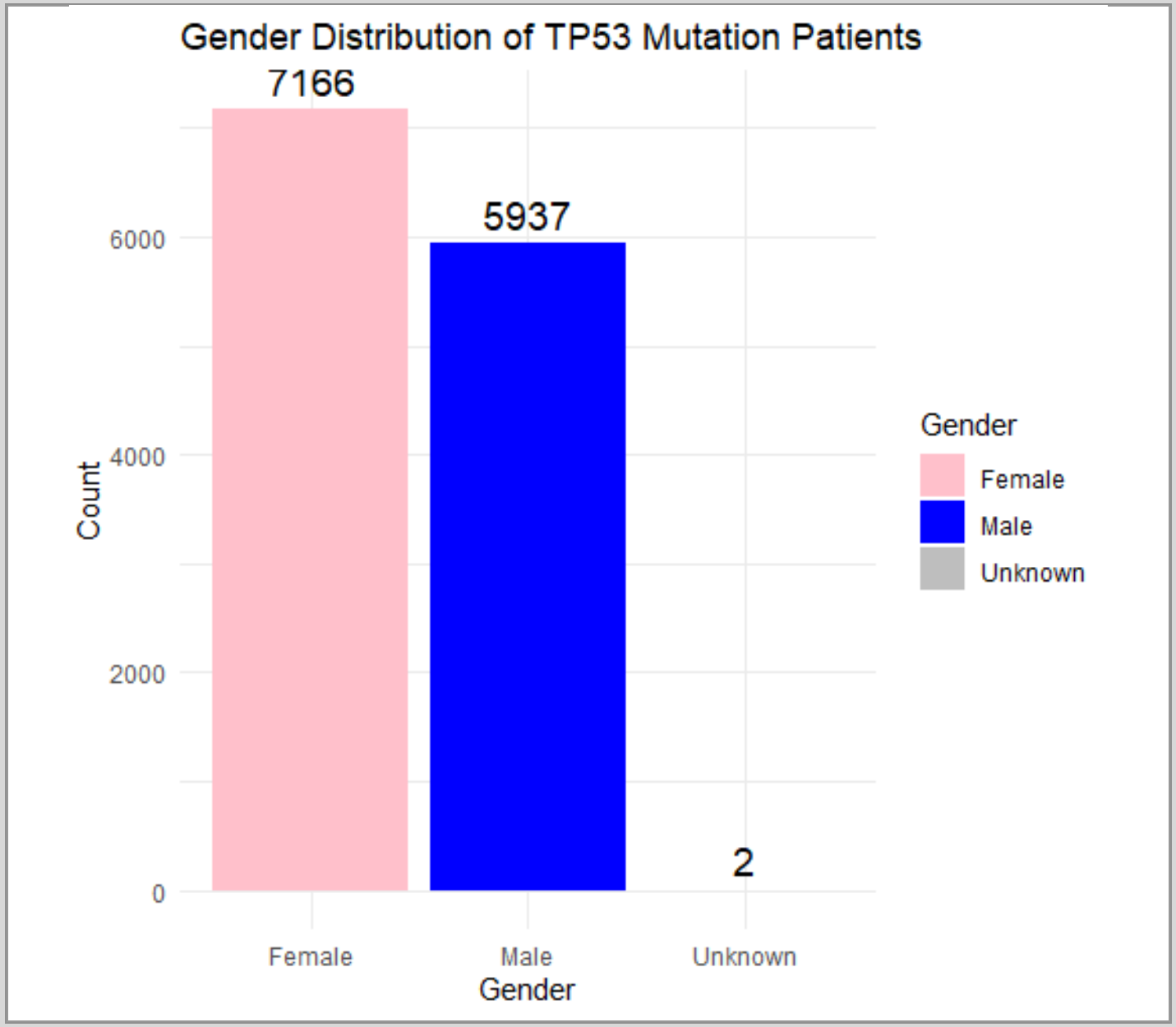
IDENTIFY A MUTATION PREVALENT IN AT LEAST 2 CANCER TYPES

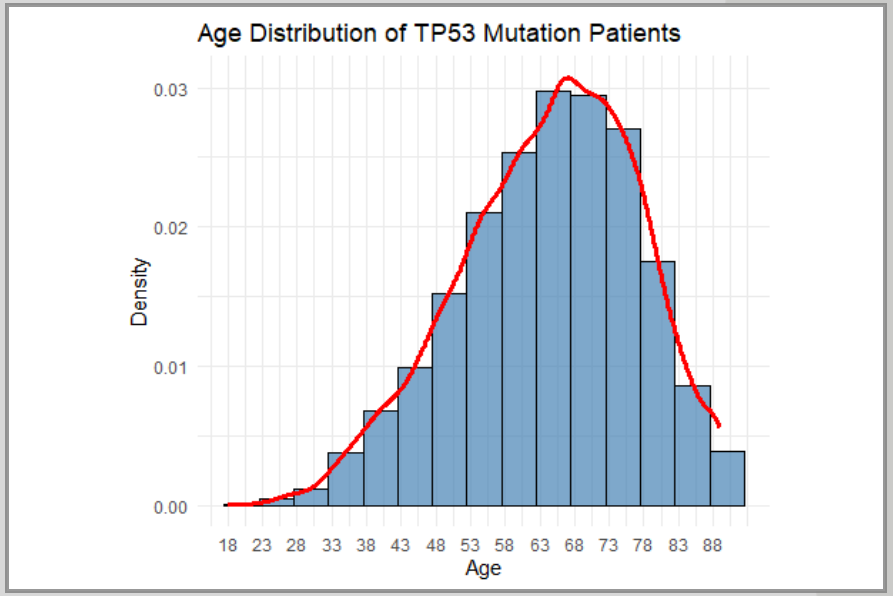
Methodology 1

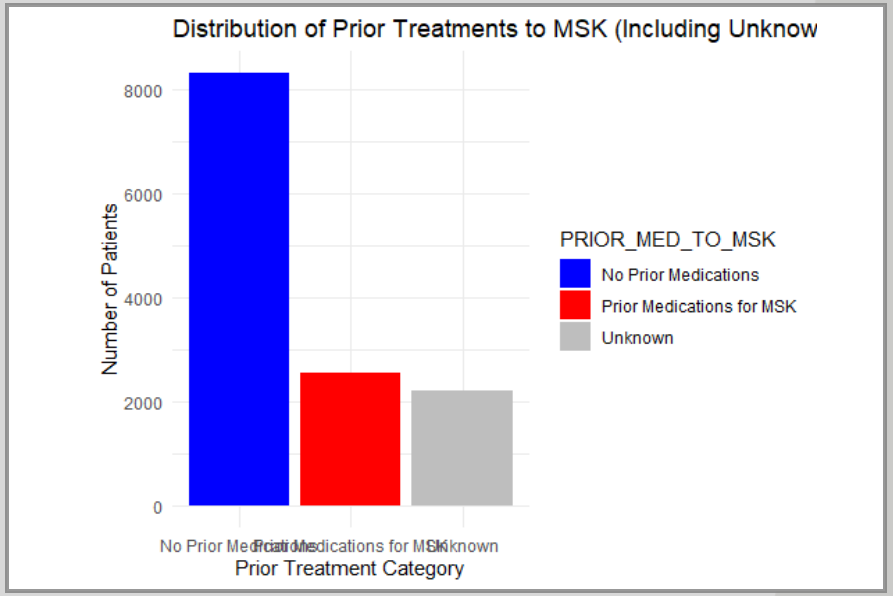
- **Counting Mutations by Cancer Type**
  - Group data by **gene mutation (Hugo_Symbol)** and **cancer type**, then count occurrences.
- **Identifying Cross-Cancer Mutations**
  - Summarize mutations found in at least **two different cancer types** and rank them by total mutation count.
- **Final Mutation Summary**
  - Filter and sort results to highlight frequently occurring mutations across multiple cancer types.

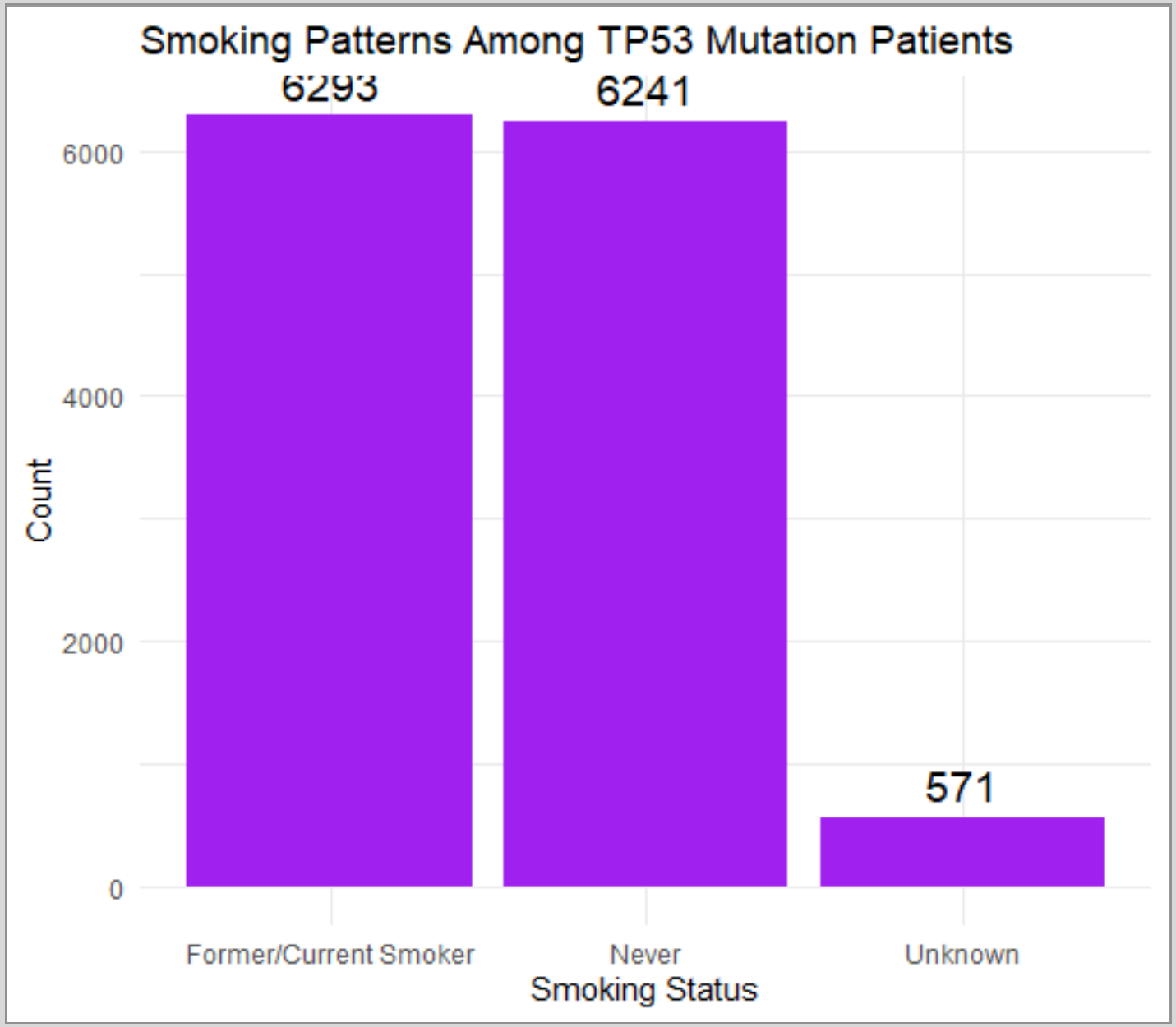
Methodology 2

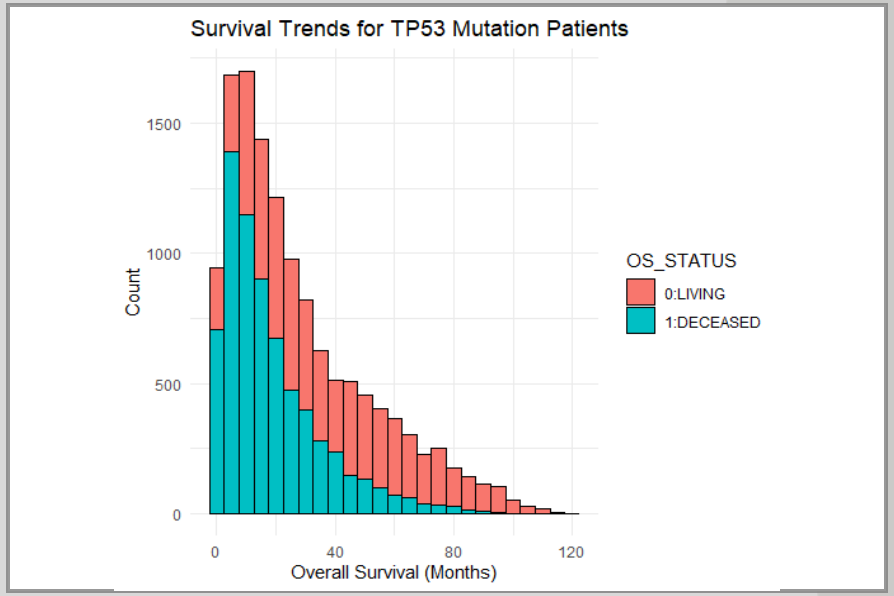
- **Summing & Sorting Mutations**
  - Compute the absolute sum of mutations for each gene and sort them in descending order.
- **Identifying Patients with Mutations**
  - Extract patient sample IDs for each gene with nonzero mutations.
- **Mapping Cancer Types to Mutations**
  - Retrieve cancer types associated with mutated genes using clinical sample data.

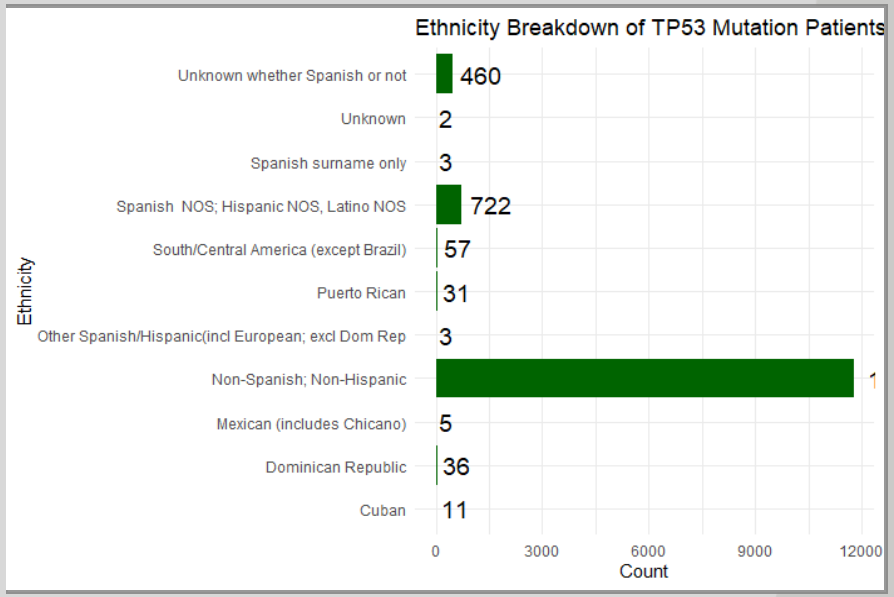

Gender Distribution of TP53 Mutation Patients
7166
5937
2
6000
4000
2000
0
Count
Female
Male
Unknown
Gender
Gender
Female
Male
Unknown


Age Distribution of TP53 Mutation Patients
0.03
0.02
0.01
0.00
Density
18 23 28 33 38 43 48 53 58 63 68 73 78 83 88
Age


Distribution of Prior Treatments to MSK (Including Unknow
8000
6000
4000
2000
0
Number of Patients
PRIOR_MED_TO_MSK
No Prior Medications
Prior Medications for MSK
Unknown
Prior Treatment Category


Smoking Patterns Among TP53 Mutation Patients
6293
6241
571
6000
4000
2000
0
Count
Former/Current Smoker
Never
Unknown
Smoking Status


Survival Trends for TP53 Mutation Patients
1500
1000
500
0
Count
0
40
80
120
Overall Survival (Months)
OS_STATUS
0:LIVING
1:DECEASED

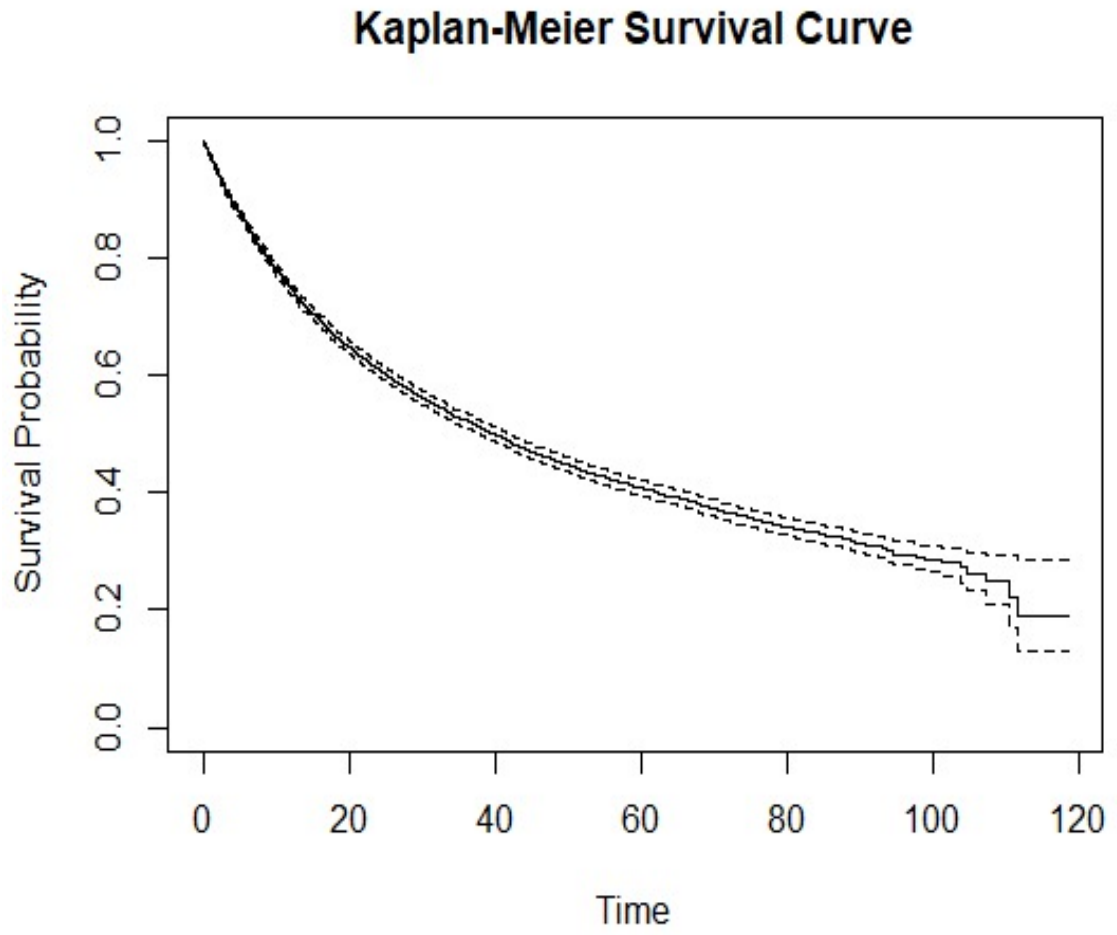

Ethnicity Breakdown of TP53 Mutation Patients
Unknown whether Spanish or not
460
Unknown
2
Spanish surname only
3
Spanish NOS; Hispanic NOS, Latino NOS
722
South/Central America (except Brazil)
57
Puerto Rican
31
Other Spanish/Hispanic(incl European; excl Dom Rep
3
Non-Spanish; Non-Hispanic
Mexican (includes Chicano)
5
Dominican Republic
36
Cuban
11
Ethnicity
0
3000
6000
9000
12000
Count

TASK 2

MEDIAN SURVIVAL TIME

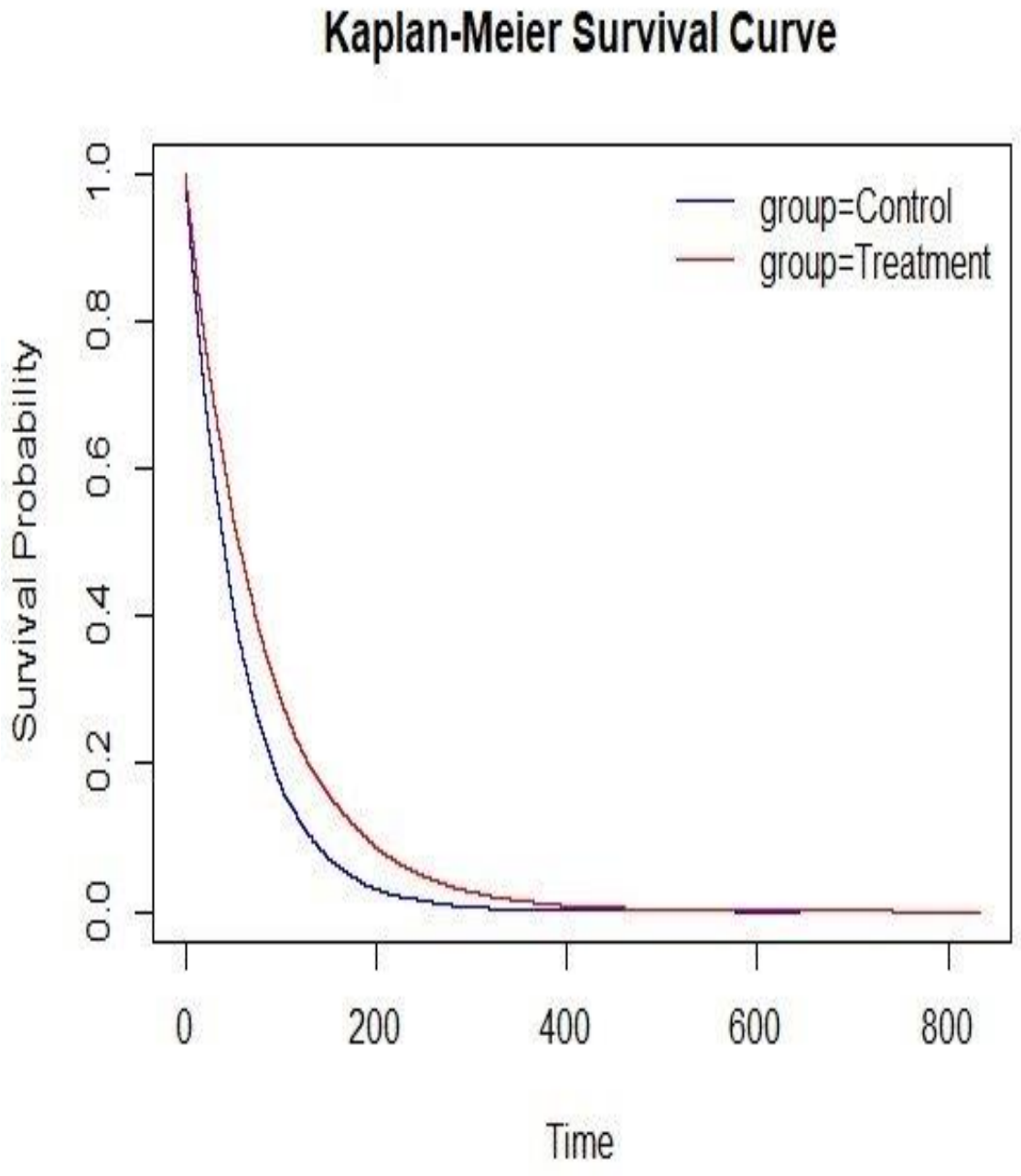
a) The **Kaplan-Meier analysis** estimated the **median survival time** for **Non-Small Cell Lung Cancer (NSCLC) patients** to be **39.95 months**.
   - This indicates that **50% of NSCLC patients survived for approximately 40 months** after diagnosis.

b) The **hazard rate (HZ)** is calculated using the formula:
   - HZ = ln(2)/Median Survival Time
     - **HZ=ln(2)/39.95 ≈ 0.01735**
   - This means that at any given month, an NSCLC patient has approximately a **1.735% chance of dying.**



c) Simulating Control and Treatment Group Survival Times

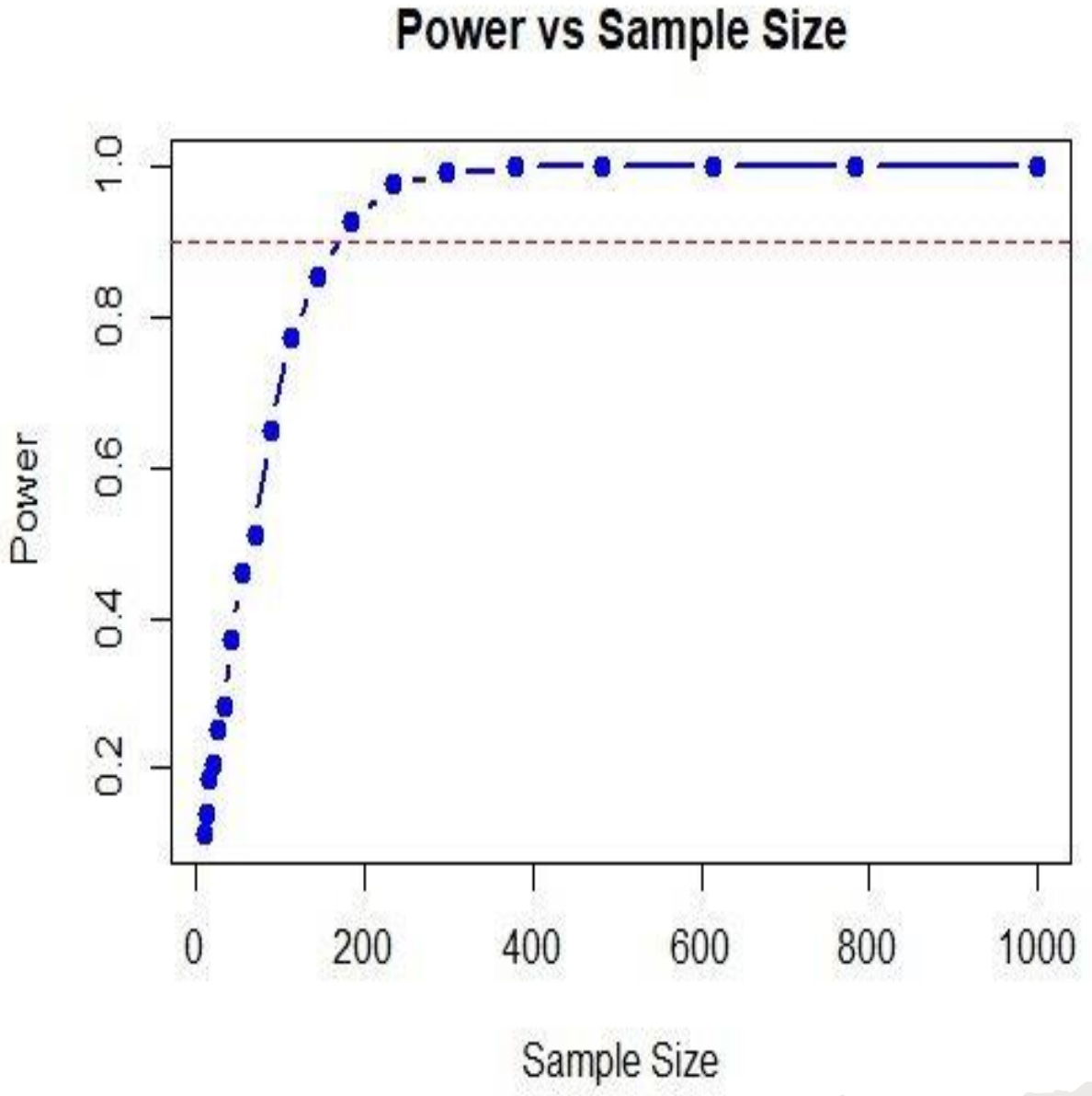
- Simulated **control and treatment group survival times** using an **exponential distribution** based on the estimated hazard rate (HZ).
- Conducted a **log-rank test** to compare survival distributions between the two groups for hypothesis Hazard Ratio < 1(Treatment improves survival)
- The test yielded a **highly significant p-value (2e-16)**, confirming a **statistically significant difference** in survival distributions between the control and treatment groups.



SIMULATION FINDINGS :

- Estimated **statistical power** for detecting a **hazard ratio of** 0.7, resulting in 0.689.
- Computed the **minimum sample size** required to achieve **90% power**, which was **183** participants.
- Plotted **Power vs. Sample Size** to illustrate the relationship between sample size and statistical power.
- Simulated the **required sample sizes** for detecting different **hazard ratios** (0.6 – 1.0)

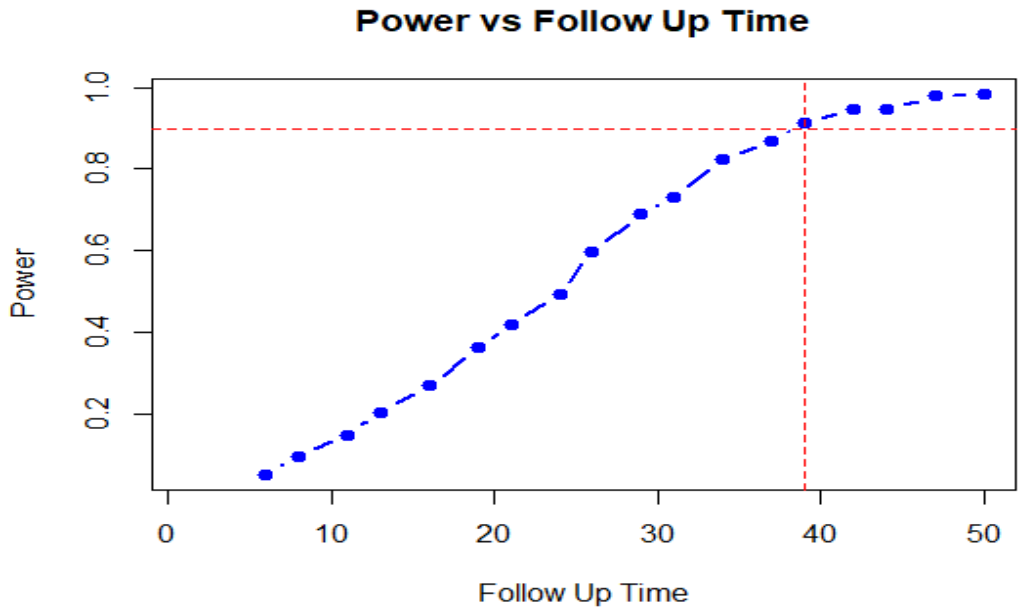
Hazard Ratio (HR)	0.60	0.64	0.69	0.73	0.78	0.82	0.87	0.91
Sample Size Required	166	166	166	464	464	1291	1291	3593



TASK 3
RECRUITMENT AND DROPOUT DYNAMICS

Uniform Recruitment:

A. Uniform recruitment of 20 new patients every month
B. Time till death is joining time + event time
   1. Event Time is measured from time of joining of patien
C. Time of Events are sampled using medium survival time.



Follow up Time:

I. Set follow up time before running simulation
II. If simulated even time exceeds follow up time: status = 0 (lived till end of trial), else set it to 1.

Uniform Dropout Rate:

A. Uniform dropout rate of 1 patient every month
B. Simulation: Randomly choose 1 patient that leaves every month, until the follow up time.
C. Delete patients that have left and calculate for remaining patient.
   1. When calculating Power vs. Follow up time, number of patients changes!

Hazard Ratio (HR)	0.60	0.64	0.69	0.73	0.78	0.82	0.87	0.91	0.96	1.00
follow up time	33	33	39	50	6	6	6	6	6	6

TASK 4

INTERIM ANALYSIS

Methodology:

I. Chose real response rates of both indications
II. Assumed **binomial distribution** and sampled response data, i.e. fraction of patients that have responded till that time.
III. Assumed a weak priori, U[0,1] for both indications
IV. Calculated the probability that the response rate is greater than 50% based on the given data
    - Done by assuming an independent beta-binomial framework and using its standard properties
V. Calculated this for 9 possible combinations of the true response times of indication one and two.

X1-> X2 | V	0.3	0.5	0.7
0.3	(0,0)	(1,0)	(1,0)
0.5	(0,1)	(1,1)	(1,1)
0.7	(0,1)	(1,1)	(1,1)

THANK YOU