In [ ]:

Project Summary---

This project analyzes eye cancer data to understand demographic trends, treatment effectiveness, and survival outcomes. Key findings include identifying the most affected age groups and cancer types, evaluating the impact of different treatments and genetic markers, and highlighting geographical variations in patient survival. The analysis provides actionable insights for healthcare and public health efforts.

- Load necessary libraries

In [17]:
```
from datetime import datetime, time as dtime
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sqlalchemy import create_engine
import os
from scipy.stats import ttest_ind
```

In [ ]:

In [41]:
```
conn=create_engine('sqlite:///cancer.db')

# for files in os.listdir('eye_cancer'):
#     x=pd.read_excel('eye_cancer/'+files)
#     print(x.shape)
#     y=x.to_sql(con=conn,name='eye_cancer_analyes',if_exists='replace')
```

In [ ]:

In [ ]:

In [19]:
```
data=pd.read_sql_query('select * from eye_cancer_analyes',conn)
data.head()
```

Out[19]:

	index	Patient_ID	Age	Gender	Cancer_Type	Laterality	Date_of_Diagnosis	Stage_at_Diagnosis	Treatment_Type	Surgery_Statu
0	0	PID00062	30	Other	Melanoma	Left	2024-12-23 00:00:00.000000	Stage III	Chemotherapy	
1	1	PID00067	34	Other	Melanoma	Left	2022-04-17 00:00:00.000000	Stage IV	Surgery	
2	2	PID00171	66	Other	Melanoma	Left	2024-02-13 00:00:00.000000	Stage III	Surgery	
3	3	PID00295	68	Other	Melanoma	Left	2024-08-10 00:00:00.000000	Stage II	Surgery	
4	4	PID00321	45	Other	Melanoma	Left	2023-10-24 00:00:00.000000	Stage III	Surgery	

In [51]:
```
# Check for patients in the dataset
data.shape
```

Out[51]:
```
(5000, 20)
```

In [52]:
```
#Columns
data.columns
```

Out[52]:
```
Index(['index', 'Patient_ID', 'Age', 'Gender', 'Cancer_Type', 'Laterality',
       'Date_of_Diagnosis', 'Stage_at_Diagnosis', 'Treatment_Type',
       'Surgery_Status', 'Radiation_Therapy', 'Chemotherapy', 'Outcome_Status',
       'Survival_Time_Months', 'Genetic_Markers', 'Family_History', 'Country',
       'Day', 'Month', 'Year'],
      dtype='object')
```

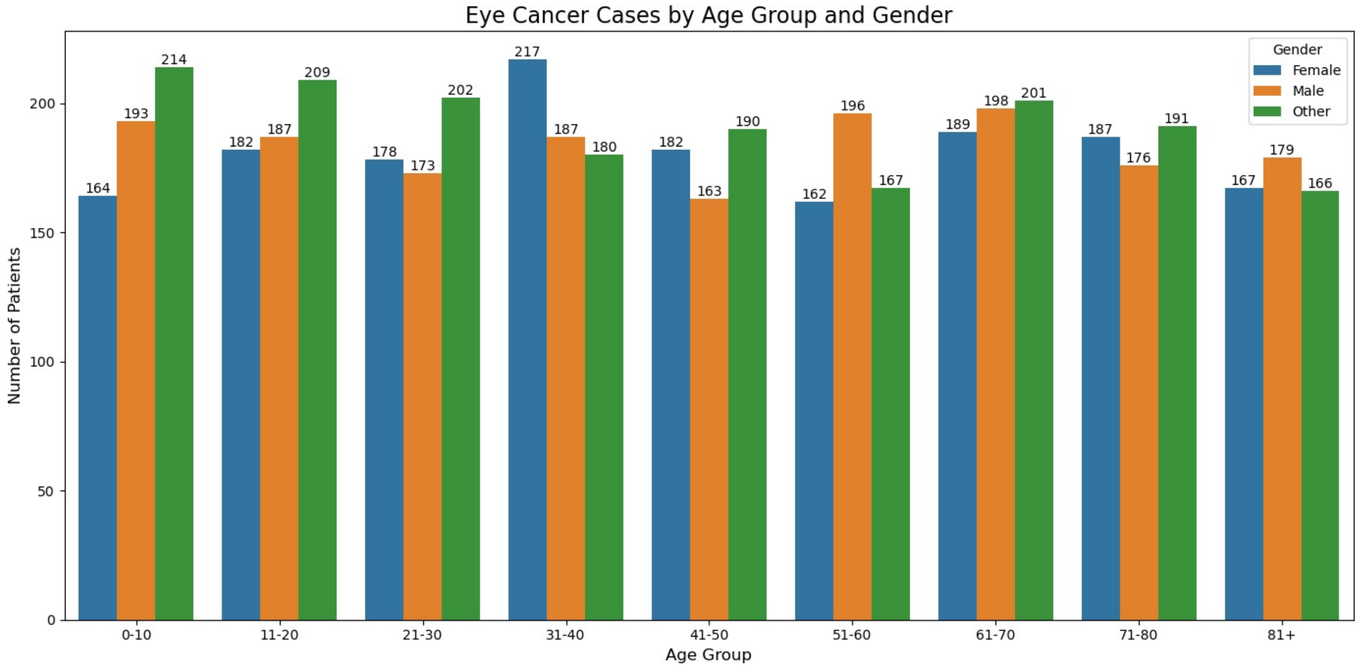
In [ ]:

-- Exploring some reserch question(EDA)--

1.Which age groups, categorized by Gender, are most affected by eye Cancer?

In [4]:
```
group_people= pd.read_sql_query('''
           SELECT Gender,CASE
        WHEN age BETWEEN 0 AND 10 THEN '0-10'
        WHEN age BETWEEN 11 AND 20 THEN '11-20'
        WHEN age BETWEEN 21 AND 30 THEN '21-30'
```

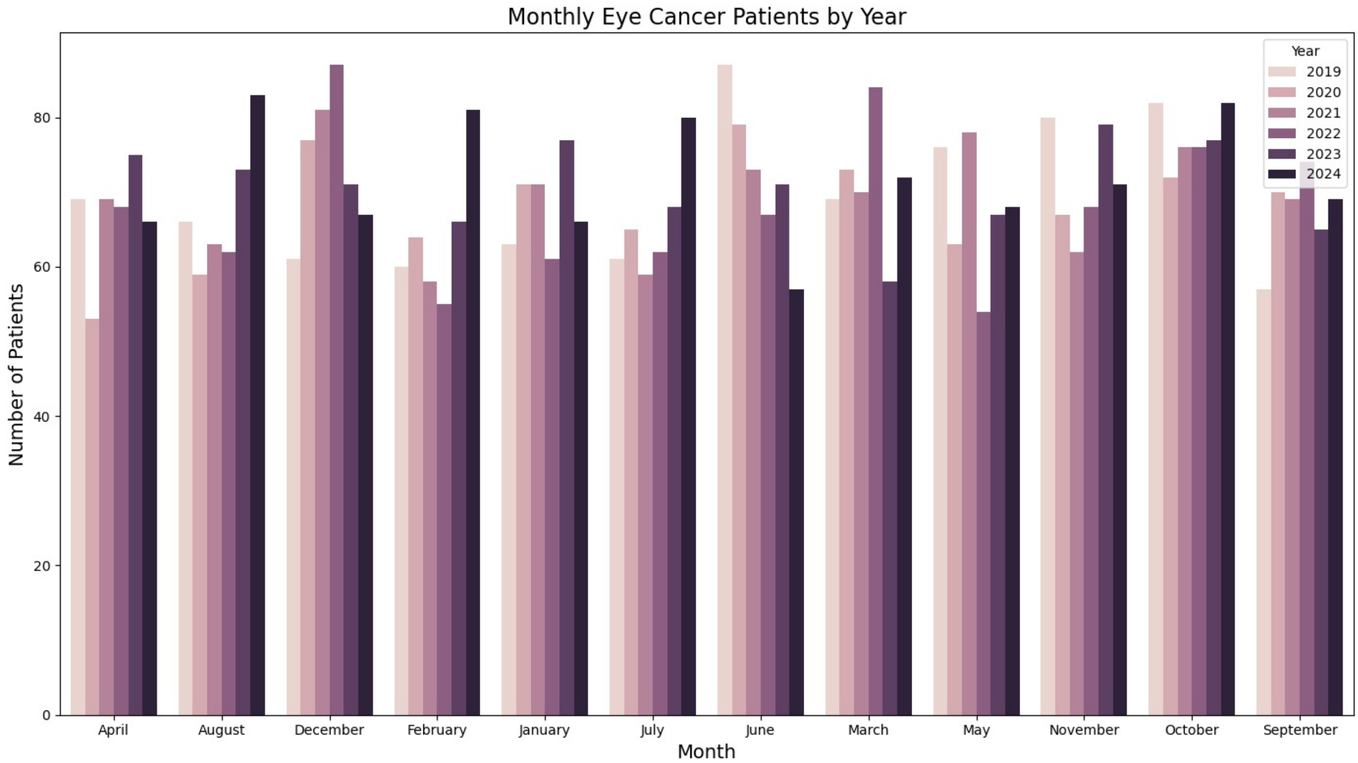
```
        WHEN age BETWEEN 31 AND 40 THEN '31-40'
        WHEN age BETWEEN 41 AND 50 THEN '41-50'
        WHEN age BETWEEN 51 AND 60 THEN '51-60'
        WHEN age BETWEEN 61 AND 70 THEN '61-70'
        WHEN age BETWEEN 71 AND 80 THEN '71-80'
        ELSE '81+'
        END AS age_group,
        COUNT(Gender) AS count,COUNT(Gender)over(partition by Gender)FROM  eye_cancer_analyes
        GROUP BY Gender,age_group ORDER BY age_group asc''',conn)

plt.figure(figsize=(14,7))
x=sns.barplot(data=group_people,x=group_people['age_group'],y=group_people['count'],hue=group_people['Gender'])
for i in x.containers:
    x.bar_label(i)
plt.title('Eye Cancer Cases by Age Group and Gender', fontsize=16)
plt.xlabel('Age Group', fontsize=12)
plt.ylabel('Number of Patients', fontsize=12)
plt.legend(title='Gender', loc='upper right')
plt.tight_layout()

plt.show()
# group_people
```




```
    *  The 61–70 age group has the highest number of eye cancer cases, closely followed by
31–40 and 11–20.
    * This indicates that older adults and middle-aged individuals are slightly more
affected,
           but eye cancer occurs across all age groups, including children
```

2.what are the years , months are most seen ?

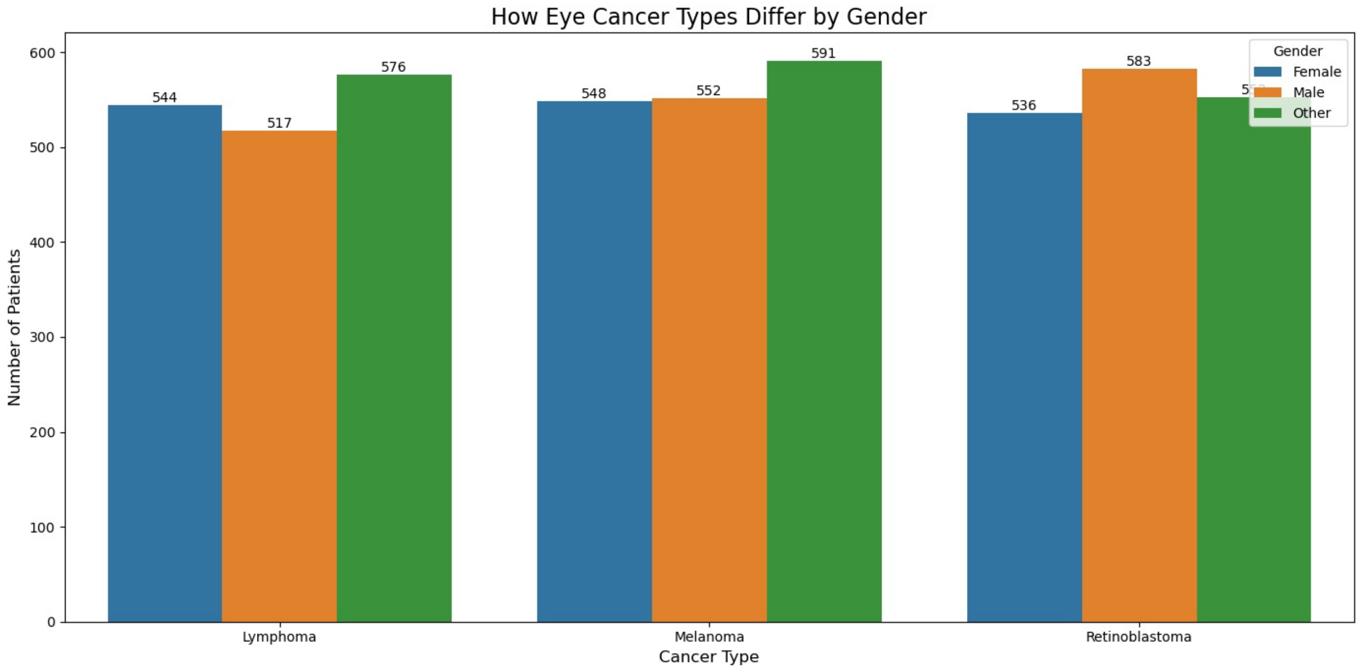
```
In [50]: patient=pd.read_sql_query('''
            select Year,Month,count(Month) month_by_patients
            from eye_cancer_analyes
            group by Year,Month
            ''',conn)


plt.figure(figsize=(14,8))
sns.barplot(data=patient,x=patient['Month'].sort_index(ascending=True),y=patient['month_by_patients'],hue=patie
plt.xlabel('Month', fontsize=14)
plt.ylabel('Number of Patients', fontsize=14)
plt.title('Monthly Eye Cancer Patients by Year', fontsize=16)
plt.legend(title='Year')
plt.tight_layout()
plt.show()
```




```
 Conclusion:
    The analysis indicates seasonal patterns in eye cancer cases, with certain months showing
higher patient counts.
    These insights can help in planning resources and focusing awareness efforts during peak
periods.
```

3.Which eye cancer types are most common among different genders?

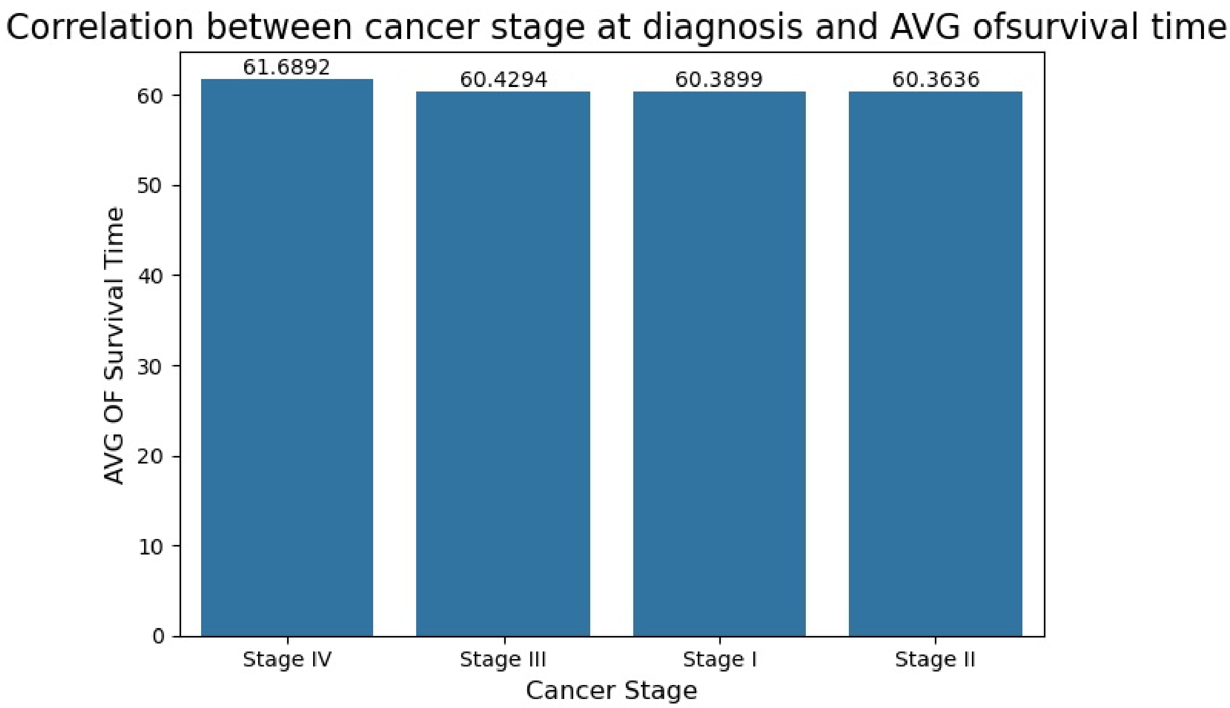
In [6]:
```
cancer_type=pd.read_sql('''
           SELECT Cancer_Type,  Gender,  COUNT(*) AS total_patients
           FROM eye_cancer_analyes
           GROUP BY Cancer_Type, Gender
           ORDER BY Cancer_Type, Gender
           ''',conn)
plt.figure(figsize=(14,7))
x=sns.barplot(data=cancer_type,x=cancer_type['Cancer_Type'],y=cancer_type['total_patients'],hue=cancer_type['Gen
for i in x.containers:
    x.bar_label(i)
plt.title('How Eye Cancer Types Differ by Gender', fontsize=16)
plt.xlabel('Cancer Type', fontsize=12)
plt.ylabel('Number of Patients', fontsize=12)
plt.legend(title='Gender', loc='upper right')
plt.tight_layout()

plt.show()
```




```
   * The most common eye cancer type across all genders is Retinoblastoma, followed by
Melanoma and Lymphoma.
   * Males show a slightly higher count in most cancer types compared to females and other
genders.
```

4.Is there a correlation between cancer stage at diagnosis and survival time?

In [7]:
```
relation=pd.read_sql_query('''
                  Select Stage_at_Diagnosis,avg(Survival_Time_Months) as avg_month from eye_cancer_analyes
                  Group by Stage_at_Diagnosis
                  Order by avg_month DESC
                  ''',conn)

x=sns.barplot(data=relation,x=relation['Stage_at_Diagnosis'],y=relation['avg_month'])
for i in x.containers:
    x.bar_label(i)
plt.title(' Correlation between cancer stage at diagnosis and AVG ofsurvival time', fontsize=16)
plt.xlabel('Cancer Stage', fontsize=12)
plt.ylabel('AVG OF Survival Time', fontsize=12)
plt.tight_layout()
plt.show()
# relation
```

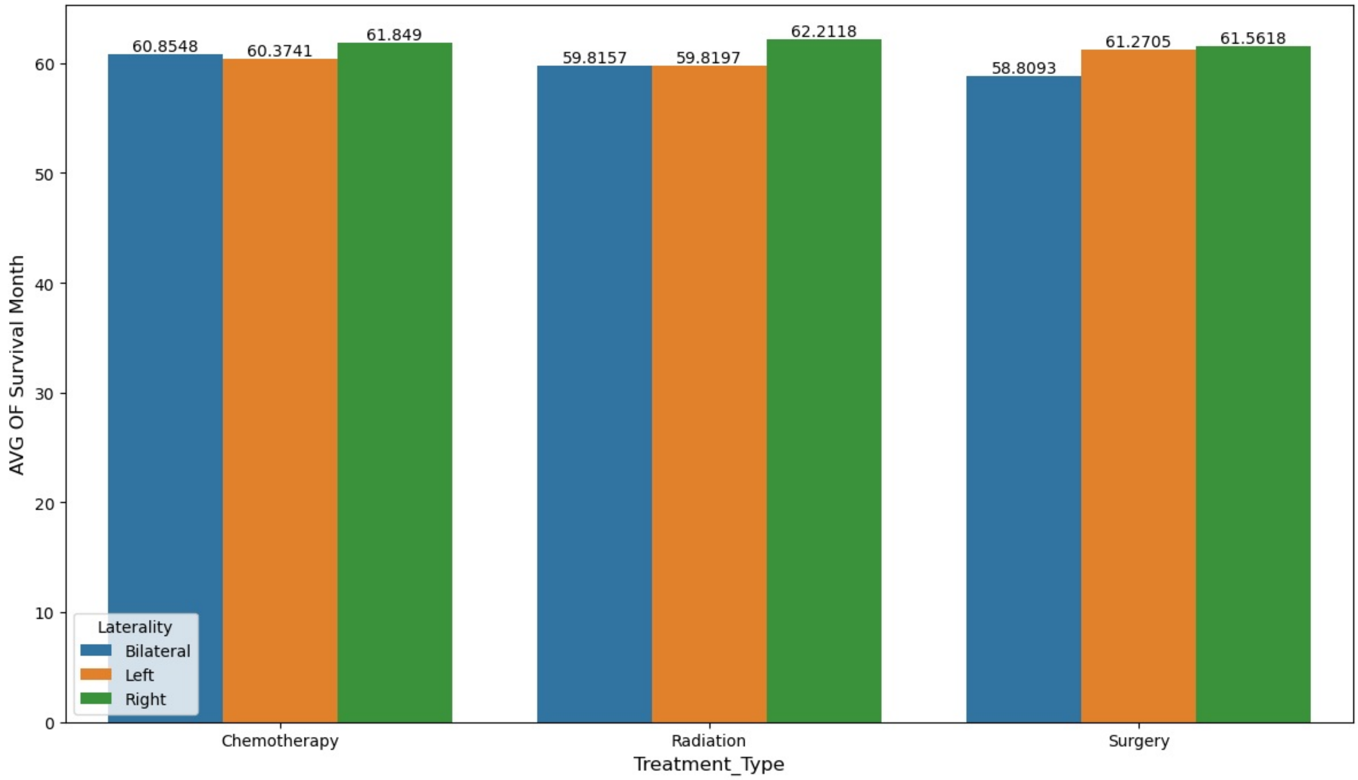



```
Conclusion--
  Patients diagnosed at early stages have shorter average survival times compared to
```

```
some later stages.
        This suggests that survival time may depend on multiple factors beyond stage at diagnosis, such as treatment type and
        patient health.
```

5.How does Laterality(left/right/bilateral) impact treatment or survival?

In [8]:
```python
Laterality_impact=pd.read_sql_query('''
    with t as(select Treatment_Type, Laterality, avg(Survival_Time_Months)as avg_month
    from  eye_cancer_analyes
    group by Treatment_Type ,Laterality)
    select Laterality, Treatment_Type, avg_month,
    AVG(avg_month) OVER (PARTITION BY Laterality order by avg_month DESC) as avg_laterality from t
    ''',conn)


# impact
plt.figure(figsize=(12,7))
x=sns.barplot(data=Laterality_impact,x=Laterality_impact['Treatment_Type'],y=Laterality_impact['avg_month'],
              hue=Laterality_impact['Laterality'])
for i in x.containers:
    x.bar_label(i)
# plt.title('b ', fontsize=16)
plt.xlabel('Treatment_Type', fontsize=12)
plt.ylabel('AVG OF Survival Month', fontsize=12)
plt.tight_layout()
plt.show()
```



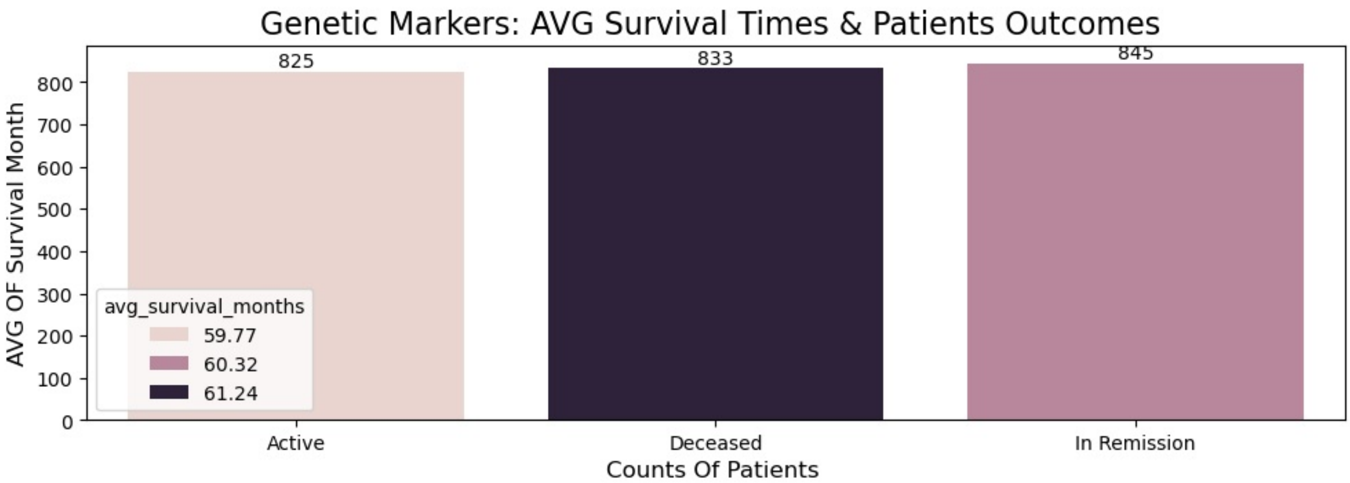

```
        *The right eye had the highest average survival time of 62.21 months.
        *Bilateral eye cases recorded the lowest survival time at 59.83 months.
        *Chemotherapy showed the best overall survival among all treatment types, with patients living the longest on average.
```

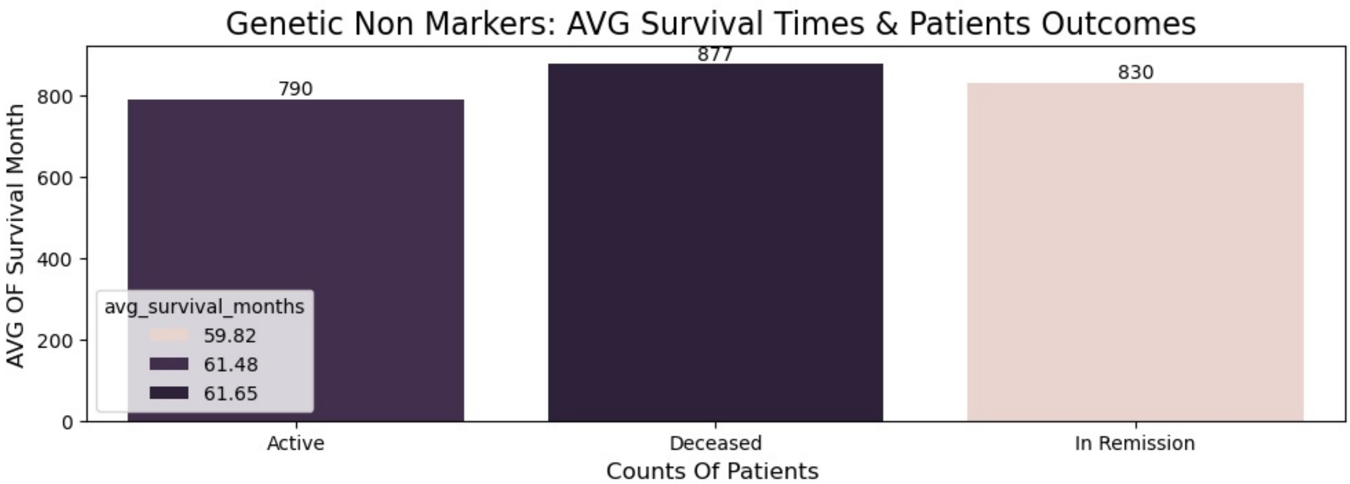
In [ ]:

6.Does the presence of genetic markers affect patient survival or outcoms?

In [11]:
```python
genetic=pd.read_sql_query('''
                 SELECT Genetic_Markers, Outcome_Status, COUNT(*) AS patient_count,
                 ROUND(AVG(Survival_Time_Months), 2) AS avg_survival_months FROM eye_cancer_analyes
                 WHERE Genetic_Markers = 'BRAF Mutation'
                 GROUP BY Genetic_Markers, Outcome_Status
                 ORDER BY Outcome_Status''',conn)

non_genetic=pd.read_sql_query('''
                 SELECT  Genetic_Markers, Outcome_Status, COUNT(*) AS patient_count,
                 ROUND(AVG(Survival_Time_Months), 2) AS avg_survival_months
                 FROM eye_cancer_analyes WHERE Genetic_Markers = 'No'
```

```
              GROUP BY Genetic_Markers, Outcome_Status
             ORDER BY Outcome_Status''',conn)


plt.figure(figsize=(10,14))
plt.subplot(411)
x=sns.barplot(data=genetic,x=genetic['Outcome_Status'],y=genetic['patient_count'],
              hue=genetic['avg_survival_months'])
for i in x.containers:
    x.bar_label(i)
plt.title('Genetic Markers: AVG Survival Times & Patients Outcomes ', fontsize=16)
plt.xlabel('Counts Of Patients ', fontsize=12)
plt.ylabel('AVG OF Survival Month', fontsize=12)
plt.tight_layout()

plt.subplot(412)
y=sns.barplot(data=non_genetic,x=non_genetic['Outcome_Status'],y=non_genetic['patient_count'],
              hue=non_genetic['avg_survival_months'])
for i in y.containers:
    y.bar_label(i)
plt.title('Genetic Non Markers: AVG Survival Times & Patients Outcomes ', fontsize=16)
plt.xlabel('Counts Of Patients ', fontsize=12)
plt.ylabel('AVG OF Survival Month', fontsize=12)
plt.tight_layout()

plt.show()
```





* We observe that there is not much difference between Genetic Markers and Non-Genetic Markers.

7.What treatments(surgery, rediation, chemo) are most effective for different cancer types?

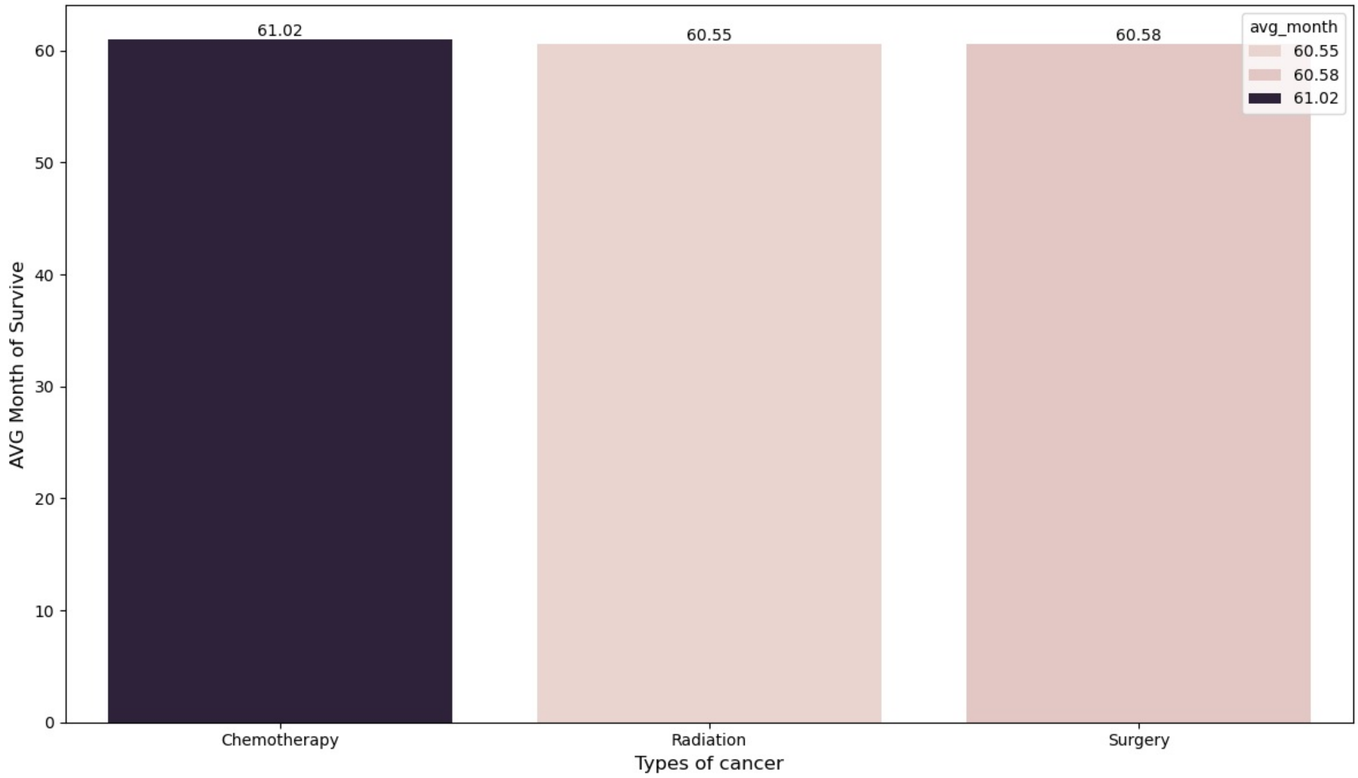
In [21]:
```
pd.read_sql_query('''
         WITH treatment_stats AS (
                SELECT Cancer_Type, Treatment_Type, Outcome_Status, COUNT(*) AS patient_count
                FROM eye_cancer_analyes GROUP BY Cancer_Type, Treatment_Type, Outcome_Status)
         SELECT Cancer_Type,Treatment_Type, Outcome_Status, patient_count, DENSE_RANK() OVER (
         PARTITION BY Cancer_Type, Treatment_Type ORDER BY patient_count DESC) AS outcome_rank
          FROM treatment_stats ORDER BY Cancer_Type, Treatment_Type, outcome_rank
          ''',conn)
```

Out[21]:

	Cancer_Type	Treatment_Type	Outcome_Status	patient_count	outcome_rank
0	Lymphoma	Chemotherapy	Deceased	182	1
1	Lymphoma	Chemotherapy	Active	178	2
2	Lymphoma	Chemotherapy	In Remission	174	3
3	Lymphoma	Radiation	In Remission	186	1
4	Lymphoma	Radiation	Deceased	183	2
5	Lymphoma	Radiation	Active	167	3
6	Lymphoma	Surgery	Deceased	192	1
7	Lymphoma	Surgery	In Remission	191	2
8	Lymphoma	Surgery	Active	184	3
9	Melanoma	Chemotherapy	Deceased	199	1
10	Melanoma	Chemotherapy	In Remission	178	2
11	Melanoma	Chemotherapy	Active	174	3
12	Melanoma	Radiation	Deceased	199	1
13	Melanoma	Radiation	In Remission	194	2
14	Melanoma	Radiation	Active	173	3
15	Melanoma	Surgery	Deceased	194	1
16	Melanoma	Surgery	In Remission	192	2
17	Melanoma	Surgery	Active	188	3
18	Retinoblastoma	Chemotherapy	Deceased	212	1
19	Retinoblastoma	Chemotherapy	Active	193	2
20	Retinoblastoma	Chemotherapy	In Remission	175	3
21	Retinoblastoma	Radiation	Active	193	1
22	Retinoblastoma	Radiation	In Remission	193	1
23	Retinoblastoma	Radiation	Deceased	168	2
24	Retinoblastoma	Surgery	In Remission	192	1
25	Retinoblastoma	Surgery	Deceased	181	2
26	Retinoblastoma	Surgery	Active	165	3

In [23]:

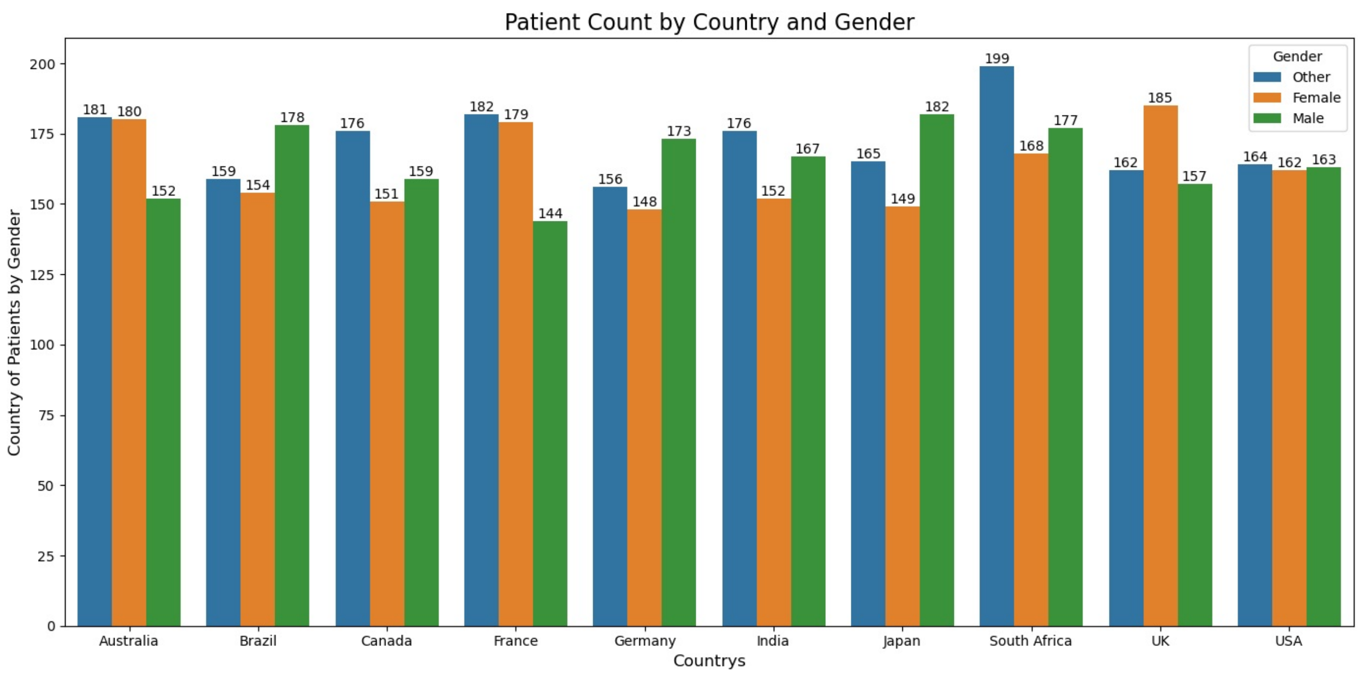
```
treatment=pd.read_sql_query('''
                 select Treatment_Type ,round(avg(Survival_Time_Months),2) as avg_month from eye_cancer_analyes
                 group by Treatment_Type
                 ''',conn)



plt.figure(figsize=(12,7))
x=sns.barplot(data=treatment,x=treatment['Treatment_Type'],y=treatment['avg_month'],
              hue=treatment['avg_month'])
for i in x.containers:
    x.bar_label(i)
# plt.title('b ', fontsize=16)
plt.xlabel('Types of cancer', fontsize=12)
plt.ylabel(' AVG Month of Survive ', fontsize=12)
plt.tight_layout()
plt.show()
```




```
* Chemotherapy shows slightly higher survival than radiation or surgery
```

9.Analysis of Eye Cancer Impact by Country and Gender.

In [32]:
```
country=pd.read_sql_query('''
                    SELECT Country, Gender, COUNT(Country) AS Gender_count,
                    COUNT(Country) OVER (PARTITION BY Country) AS country_of_patients
                    FROM eye_cancer_analyes GROUP BY Country, Gender
                    ORDER BY Country , Gender_count DESC
                    ''',conn)


plt.figure(figsize=(14,7))
x=sns.barplot(data=country,x=country['Country'],y=country['Gender_count'],
              hue=country['Gender'])
for i in x.containers:
    x.bar_label(i)
plt.title('Patient Count by Country and Gender', fontsize=16)
plt.xlabel('Countrys', fontsize=12)
plt.ylabel('Country of Patients by Gender', fontsize=12)
plt.tight_layout()
plt.show()
```

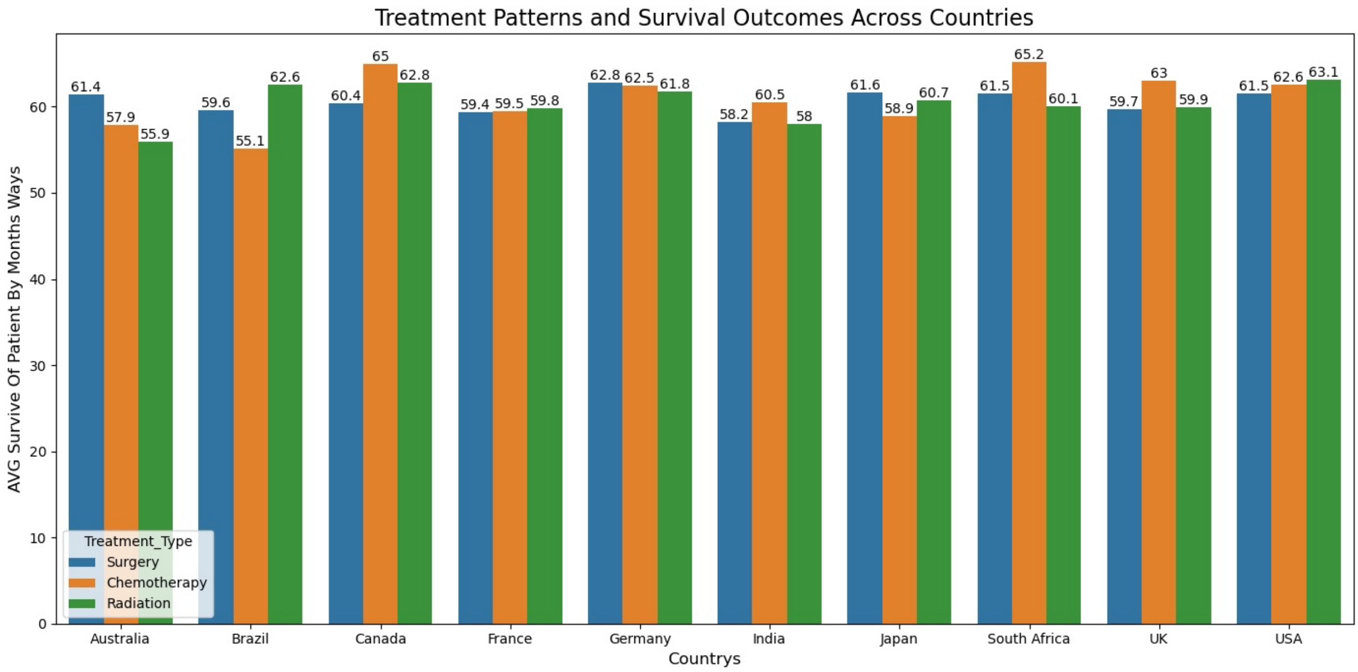


In [ ]:

```
In [ ]:
```

10.How do treatment patterns and Survival differ across Counties?

```
In [36]: country_survive=pd.read_sql_query('''
                              SELECT Country, Treatment_Type, COUNT(*) AS patient_count,
                              ROUND(AVG(Survival_Time_Months),1) AS avg_survival_months
                              FROM eye_cancer_analyes
                              GROUP BY Country, Treatment_Type
                              ORDER BY Country, avg_survival_months DESC
                              ''',conn)


plt.figure(figsize=(14,7))
x=sns.barplot(data=country_survive,x=country_survive['Country'],y=country_survive['avg_survival_months'],
              hue=country_survive['Treatment_Type'])
for i in x.containers:
    x.bar_label(i)
plt.title('Treatment Patterns and Survival Outcomes Across Countries', fontsize=16)
plt.xlabel('Countrys', fontsize=12)
plt.ylabel('AVG Survive Of Patient By Months Ways', fontsize=12)
plt.tight_layout()
plt.show()
```




```
     * We see that 4 to 5 countrys survivel time was logger to others countrys
     *Canada has the highest average survival, chemotherapy is the most widely favored and
top-performing treatment,
       and South Africa’s chemotherapy shows the best survival outcome globally.
```

11.

Hypothesis Testing --

independent t test:

- H0 (Null Hypothesis): There is no difference in mean survival time between patients with a BRAF mutation and those without.
- H1 (Alternative Hypothesis): There is a difference in mean survival time between patients with a BRAF mutation and those without.

```
In [21]: braf_group = data[data['Genetic_Markers'] == 'BRAF Mutation']['Survival_Time_Months'].dropna()
none_group = data[data['Genetic_Markers'] == 'No']['Survival_Time_Months'].dropna()

# Perform independent t test
t_stat, p_value =ttest_ind(braf_group, none_group)

print("T-statistic:", (t_stat))
print("p-value:", round(p_value, 4))

alpha=0.05

if p_value < alpha:
```

```
    print("Significant difference in survival times between BRAF Mutation and None groups.")
else:
    print("No significant difference in survival times between BRAF Mutation and None groups.")
```

```
T-statistic: -0.5518551943558958
p-value: 0.5811
No significant difference in survival times between BRAF Mutation and None groups.
```

In [ ]:

- Chi-Square Test--

In [58]:

```
from scipy.stats import chi2_contingency

#H0: The Treatment_Type and Outcome_Status are independent.
#H1: The Treatment_Type and Outcome_Status are dependent.

# Create a contingency table between 'Treatment_Type' and 'Outcome_Status'
contingency_table = pd.crosstab(data['Treatment_Type'], data['Outcome_Status'])

# Perform the chi-square test
chi2, p, dof, expected = chi2_contingency(contingency_table)

# Print the results
print("Chi-square statistic:", chi2)
print("P-value:", p)
print("Degrees of freedom:", dof)
print("Expected frequencies table:",expected)

alpha=0.05

if p< alpha:
    print("The Treatment_Type and Outcome_Status are dependent.")
else:
    print("The Treatment_Type and Outcome_Status are independent")
```

```
Chi-square statistic: 4.270855929444442
P-value: 0.37058876248501893
Degrees of freedom: 4
Expected frequencies table: [[537.795 569.43  557.775]
 [534.888 566.352 554.76 ]
 [542.317 574.218 562.465]]
The Treatment_Type and Outcome_Status are independent
```

- Since the p-value (0.371) is greater than the significance level (alpha = 0.05), we conclude that the Treatment_Type and Outcome_Status are independent.
  - Thats mean's a patient's outcome is not dependent on the type of treatment they received.

In [ ]:

In [ ]:

In [ ]:

- Patientd Trend(Year,Month,Days)

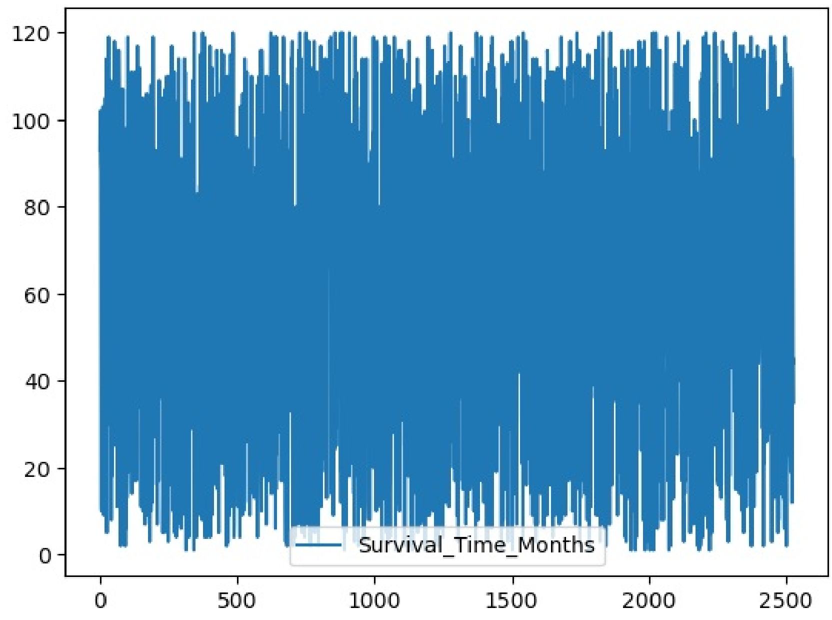
In [39]:

```
trend=pd.read_sql_query('''
               SELECT substr(Date_of_Diagnosis, 1, 10) AS Date, Survival_Time_Months FROM eye_cancer_analyes
               where Date between'2022-01-01' and '2024-12-30'
               ''',conn )
```

In [40]:

```
sns.lineplot(trend)
```

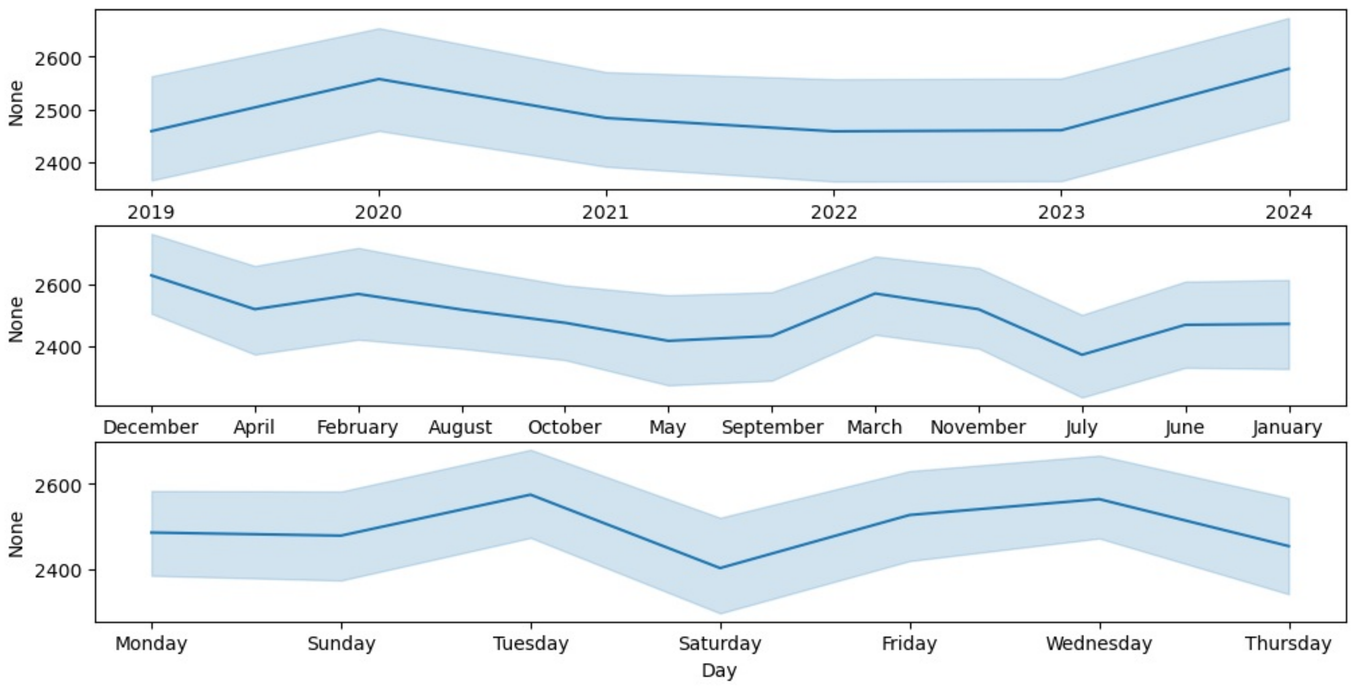
Out[40]: <Axes: >



In [ ]:

In [48]:
```
plt.figure(figsize=(12,8))
plt.subplot(411)
sns.lineplot(data=data,x=data['Year'],y=data.index)
plt.subplot(412)
sns.lineplot(data=data,x=data['Month'],y=data.index)
plt.subplot(413)
sns.lineplot(data=data,x=data['Day'],y=data.index)
```

Out[48]: <Axes: xlabel='Day', ylabel='None'>



In [ ]:

In [ ]:

In [ ]:

In [ ]:

- Conclusions of this projects::

  - Demographic Insights: The 61-70 age group shows the highest number of eye cancer cases, followed by the 31-40 and 11-20 age groups. Eye cancer is observed across all age groups, including children. Across various cancer types, males generally have a slightly higher case count compared to females and other genders.

  - Common Cancer Types: Retinoblastoma is identified as the most common type of eye cancer, with Melanoma and Lymphoma being the next most frequent types.

  - Treatment and Survival Outcomes:

    ```
    * Treatment Effectiveness: Chemotherapy appears to be the most effective treatment type, showing slightly higher average survival
      times than radiation or surgery.
    * Laterality: Patients with cancer in the right eye have the highest average survival time at 62.21 months. In contrast, patients
      with bilateral eye cancer have the lowest average survival time, at 59.83 months.
    * Stage at Diagnosis: There is no direct correlation between an earlier stage at diagnosis and longer survival time. This suggests
      that survival is influenced by multiple factors beyond just the stage, such as the treatment type and the patient's overall
      health.
    ```

  - Genetic Marker Analysis: A hypothesis test comparing patients with a BRAF mutation to those without showed no significant difference in mean survival times (p-value: 0.5811). This indicates that, based on this data, the presence of this specific genetic marker does not significantly affect survival time.

  - Geographical Trends: Canada recorded the highest average survival time among all countries. Additionally, South Africa's chemotherapy treatment showed the best survival outcome globally.

In [ ]: