Exploratory Data Analysis on a Cancer dataset

This dataset compiles cancer-related patient data collected from various hospital regions. It includes a wealth of information: demographic details, lifestyle factors, cancer diagnostics, treatment information, and outcomes for 17,686 patients. The data is meticulously organized to enable analysis of patterns in cancer diagnosis, treatment efficacy, and survival outcomes.

OBJECTIVES

- Discovering the datasets and understand it.
- Cleaning missing values and null values
- Creating new metrics and find relationships
- Validate the findings and write a summary

Import Libraries

Here imported some most popular and related libraries

In [98]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
```

In [8]:
```
# Lets read our csv files using panda
data = pd.read_csv('csv/cancer/cancer_issue.csv')
# lets make a copy of original datasets
df = data.copy()
```

Exploratory Data Analysis

In [17]:
```
# head(): This function displays the first five rows of the DataFrame by def
df.head()
```

Out[17]:

	PatientID	Age	Gender	Race/Ethnicity	BMI	SmokingStatus	FamilyHistory	Cance
0	1	80	Female	Other	23.3	Smoker	Yes	
1	2	76	Male	Caucasian	22.4	Former Smoker	Yes	
2	3	69	Male	Asian	21.5	Smoker	Yes	
3	4	77	Male	Asian	30.4	Former Smoker	Yes	P
4	5	89	Male	Caucasian	20.9	Smoker	Yes	

In [21]:
```
#info() will show us no of rows and columns and data types and memory usage
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17686 entries, 0 to 17685
Data columns (total 16 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   PatientID          17686 non-null  int64  
 1   Age                17686 non-null  int64  
 2   Gender             17686 non-null  object 
 3   Race/Ethnicity     17686 non-null  object 
 4   BMI                17686 non-null  float64
 5   SmokingStatus      17686 non-null  object 
 6   FamilyHistory      17686 non-null  object 
 7   CancerType         17686 non-null  object 
 8   Stage              17686 non-null  object 
 9   TumorSize          17686 non-null  float64
 10  TreatmentType      17686 non-null  object 
 11  TreatmentResponse  17686 non-null  object 
 12  SurvivalMonths     17686 non-null  int64  
 13  Recurrence         17686 non-null  object 
 14  GeneticMarker      13360 non-null  object 
 15  HospitalRegion     17686 non-null  object 
dtypes: float64(2), int64(3), object(11)
memory usage: 2.2+ MB
```

In [22]:
```
# lets check for the null values if exist
df.isnull().sum()
```

Out[22]:
```
PatientID              0
Age                    0
Gender                 0
Race/Ethnicity         0
BMI                    0
SmokingStatus          0
FamilyHistory          0
CancerType             0
Stage                  0
TumorSize              0
TreatmentType          0
TreatmentResponse      0
SurvivalMonths         0
Recurrence             0
GeneticMarker       4326
HospitalRegion         0
dtype: int64
```

Summary

Here we have 4326 null value in 'GeneticMarker' from the total entries of 17686.

In [33]:
```
# as geneticmaker has many null values we will perform a grouped query for t
df[df['GeneticMarker'].isnull()].groupby(['CancerType']).size()
```

Out[33]:
```
CancerType
Breast      702
Colon       749
Leukemia    688
Lung        750
Prostate    721
Skin        716
dtype: int64
```

In [34]:
```
# we will use mode of genetic makers based on the cancertype to replace the
# otherwise if no mode is suitable we will replace with 'unkown'

df['GeneticMarker'] = df.groupby(['CancerType'])['GeneticMarker'].transform(
    lambda x: x.fillna(x.mode()[0] if not x.mode().empty else 'Unknown')
)
```

In [42]:
```
hasnull = df['GeneticMarker'].isnull().sum()
print("GeneticMarker has",hasnull,"null values")
```

```
GeneticMarker has 0 null values
```

In [46]:
```
# lets see if we have null values
df.isnull().sum()
```

Out[46]:
```
PatientID           0
Age                 0
Gender              0
Race/Ethnicity      0
BMI                 0
SmokingStatus       0
FamilyHistory       0
CancerType          0
Stage               0
TumorSize           0
TreatmentType       0
TreatmentResponse   0
SurvivalMonths      0
Recurrence          0
GeneticMarker       0
HospitalRegion      0
dtype: int64
```

In [49]:
```
# Select all columns except 'GeneticMarker' and then call describe()
summary = df.loc[:, df.columns != 'PatientID'].describe()

summary.T
```

Out[49]:

	count	mean	std	min	25%	50%	75%	max
Age	17686.0	53.758396	21.079473	18.0	35.0	54.0	72.0	90.0
BMI	17686.0	29.253805	6.203575	18.5	23.9	29.2	34.6	40.0
TumorSize	17686.0	5.499751	2.603107	1.0	3.3	5.5	7.7	10.0
SurvivalMonths	17686.0	60.387821	34.794859	1.0	30.0	60.0	91.0	120.0

Understand data with visualization

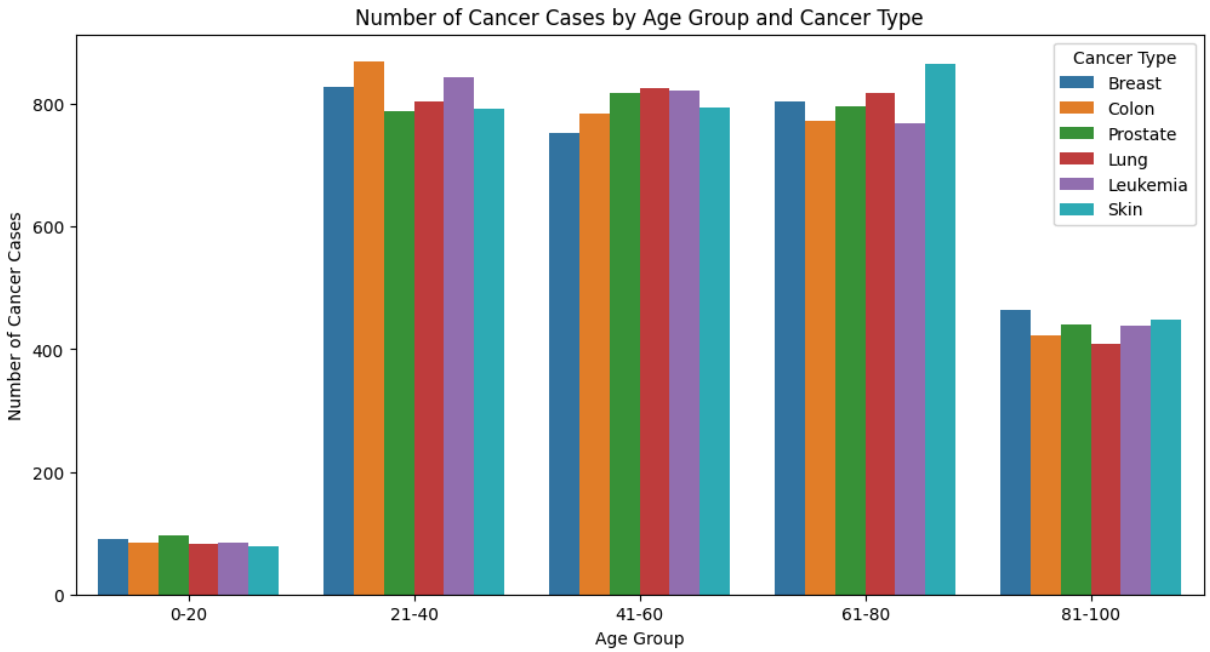
In [57]:
```
df.head()
```

Out[57]:

	PatientID	Age	Gender	Race/Ethnicity	BMI	SmokingStatus	FamilyHistory	Cance
0	1	80	Female	Other	23.3	Smoker	Yes	
1	2	76	Male	Caucasian	22.4	Former Smoker	Yes	
2	3	69	Male	Asian	21.5	Smoker	Yes	
3	4	77	Male	Asian	30.4	Former Smoker	Yes	P
4	5	89	Male	Caucasian	20.9	Smoker	Yes	

In [75]:
```
# Plots to find the relation between the Age and Cancer Occcurences

bins = [0, 20, 40, 60, 80, 100]  # define your age bins
labels = ['0-20', '21-40', '41-60', '61-80', '81-100'] # define age labels
df['age_group'] = pd.cut(df['Age'], bins=bins, labels=labels, right=False)
custom_palette = ['#1f77b4', '#ff7f0e', '#2ca02c', '#d62728', '#9467bd', '#1
```

In [113…
```
plt.figure(figsize=(12, 6))
sns.countplot(data=df, x='age_group', hue='CancerType', palette=custom_palet
plt.title('Number of Cancer Cases by Age Group and Cancer Type')
plt.xlabel('Age Group')
plt.ylabel('Number of Cancer Cases')
plt.legend(title='Cancer Type')
plt.show()
```



Key Insights

- There are less cancer occurences to younger people with age group 1-20 and elder one 81 plus and above
- The age group of 20-40 has more colon cancers compared to other
- The age group of 61 - 80 are more venerable to skin diseases and cancers

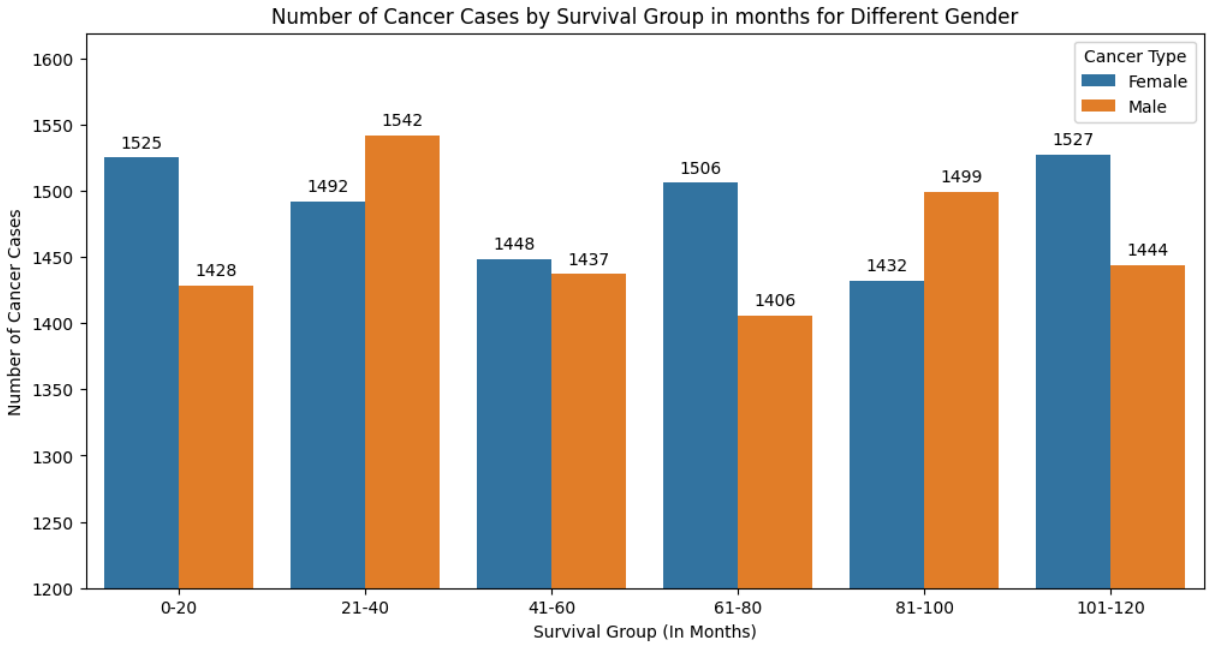
Survival Rates

In [93]:
```
#Lets check the Survival max and mins here
sm_max = df['SurvivalMonths'].max()
sm_min = df['SurvivalMonths'].min()

print(f"The Max Survival time in  Month is: {sm_max} ")
print(f"The Min Survival time in  Month is: {sm_min} ")
```

```
The Max Survival time in  Month is: 120
The Min Survival time in  Month is: 1
```

In [127…
```
# Lets define bins and labels as below. The bin values are in months
bins = [0, 20, 40, 60, 80, 100, 120]
labels = ["0-20", "21-40", "41-60", "61-80", "81-100", "101-120"]
df['survival_group'] = pd.cut(df['SurvivalMonths'], bins=bins, labels=labels
```

In [133…
```
custom_palette = ['#1f77b4', '#ff7f0e'] # custom color

plt.figure(figsize=(12, 6))
ax = sns.countplot(data=df, x='survival_group', hue='Gender', palette=custom
# ----
# Add count labels on top of each bar
for p in ax.patches:
    ax.annotate(format(p.get_height(), '.0f'),
                (p.get_x() + p.get_width() / 2.,
                 p.get_height()),
                 ha = 'center', va = 'center', xytext = (0, 9),
                 textcoords = 'offset points')
# ends here

plt.title('Number of Cancer Cases by Survival Group in months for Different
plt.xlabel('Survival Group (In Months)')
plt.ylim(1200)
plt.ylabel('Number of Cancer Cases')
plt.legend(title='Cancer Type')
plt.show()
```



Key Insights

- There are more female who live shorter in terms of months from 1-20
- The survival time length is shorter for female while compared to male in if the survival length is in between 1-20 months
- The survial time length is longer for female during during the survival group of 21-40 months category

In [140…
```
df.head(2)
```

Out[140…

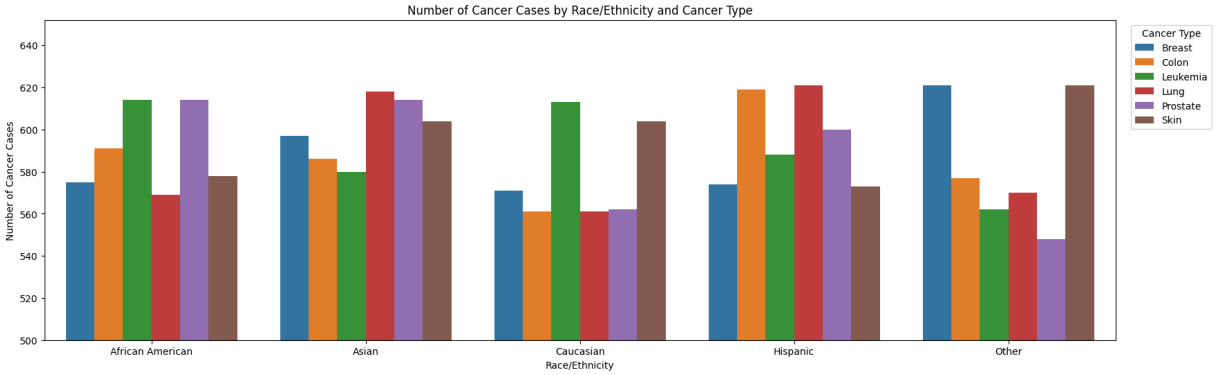
	PatientID	Age	Gender	Race/Ethnicity	BMI	SmokingStatus	FamilyHistory	Cance
0	1	80	Female	Other	23.3	Smoker	Yes	
1	2	76	Male	Caucasian	22.4	Former Smoker	Yes	

In [184…
```
# Create a new helper dataframe for plotting.
# df_by_ = df.groupby(['month','month_txt']).sum().sort_values('month', asce
# df_by_month

df_by_race_ctype = df.groupby(['Race/Ethnicity','CancerType']).count()
df_by_race_ctype =df_by_race_ctype.reset_index()

df_by_race_ctype.head()
```

Out[184…

	Race/Ethnicity	CancerType	PatientID	Age	Gender	BMI	SmokingStatus	FamilyH
0	African American	Breast	575	575	575	575	575	
1	African American	Colon	591	591	591	591	591	
2	African American	Leukemia	614	614	614	614	614	
3	African American	Lung	569	569	569	569	569	
4	African American	Prostate	614	614	614	614	614	

In [196…
```
plt.figure(figsize=(20, 6))
sns.barplot(data=df_by_race_ctype, x='Race/Ethnicity',y='PatientID', hue='Ca
plt.title('Number of Cancer Cases by Race/Ethnicity and Cancer Type')
plt.xlabel('Race/Ethnicity')
plt.ylabel('Number of Cancer Cases')
plt.ylim(500)
# plt.legend(title='Cancer Type')
plt.legend(title='Cancer Type', bbox_to_anchor=(1.01, 1), loc='upper left')
plt.show()
```



Key insights

- African American have higher number of Leukemia and Prostate cancer compared to others
- Asian have higher Lung and Prostate cancers compared to others
- Caucasian have more Leukemia and Skin cancers compared to others
- Hispanic have higher no of Colon and Lung cancers
- Other category have Breast and Skin cancers (need to verify where the other category falls)

Number of cancer patients categoried by Gender and Smoking Status

In [201…

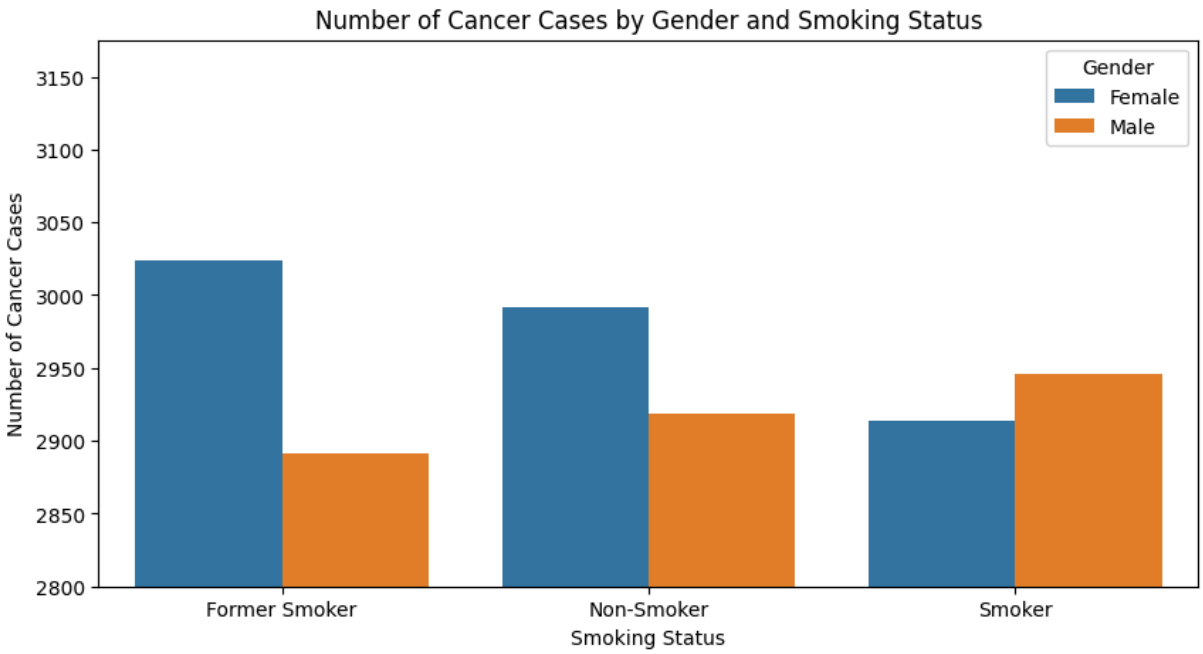
```
df_by_gender_smoke = df.groupby(['Gender','SmokingStatus']).count()
df_by_gender_smoke =df_by_gender_smoke.reset_index()
df_by_gender_smoke
```

Out[201…

	Gender	SmokingStatus	PatientID	Age	Race/Ethnicity	BMI	FamilyHistory	Can
0	Female	Former Smoker	3024	3024	3024	3024	3024	
1	Female	Non-Smoker	2992	2992	2992	2992	2992	
2	Female	Smoker	2914	2914	2914	2914	2914	
3	Male	Former Smoker	2891	2891	2891	2891	2891	
4	Male	Non-Smoker	2919	2919	2919	2919	2919	
5	Male	Smoker	2946	2946	2946	2946	2946	

In [213…

```
plt.figure(figsize=(10, 5))
sns.barplot(data=df_by_gender_smoke, x='SmokingStatus',y='PatientID', hue='C
plt.title('Number of Cancer Cases by Gender and Smoking Status')
plt.xlabel('Smoking Status')
plt.ylabel('Number of Cancer Cases')
plt.ylim(2800)
# plt.legend(title='Cancer Type')
plt.legend(title='Gender',loc='upper right')
plt.show()
```



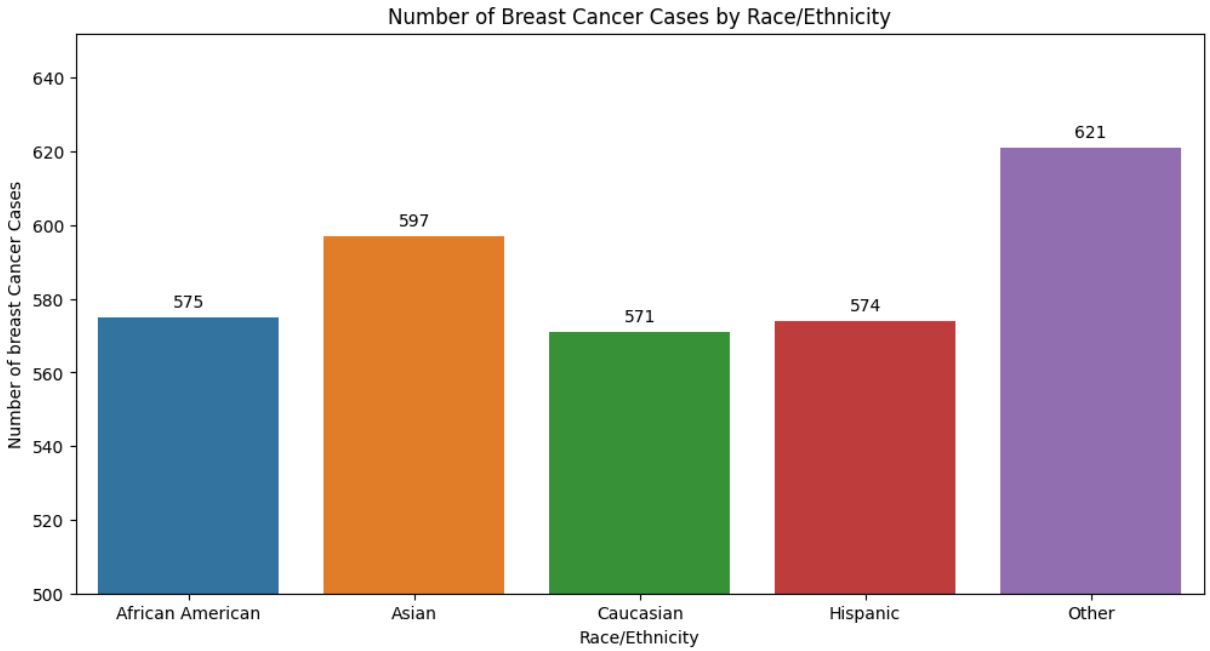
Key insights

- Former Female smoker has more cancer numbers compared to male
- Female non smoker has more cancers
- Male smoker has more cancers compared to females

In [225…
```
#Lets see which Race and Ethinicity has more breast cancer
df_breast_cancer = df[df['CancerType']=='Breast']
```

In [227…
```
# Lets group by Race and Ethinicity on df_breast_cancer df
df_by_race_breast_cancer = df_breast_cancer.groupby('Race/Ethnicity').count(
```

In [243…
```
#Lets plot a graph
plt.figure(figsize=(12, 6))
ax =sns.barplot(data=df_by_race_breast_cancer, x='Race/Ethnicity', y='Patier
# --------

# Add count labels on top of each bar
for p in ax.patches:
    ax.annotate(format(p.get_height(), '.0f'),
                (p.get_x() + p.get_width() / 2., p.get_height()),
                 ha = 'center', va = 'center', xytext = (0, 9),
                textcoords = 'offset points')

#---------

plt.title('Number of Breast Cancer Cases by Race/Ethnicity')
plt.xlabel('Race/Ethnicity')
plt.ylim(500)
plt.ylabel('Number of breast Cancer Cases')
plt.show()
```

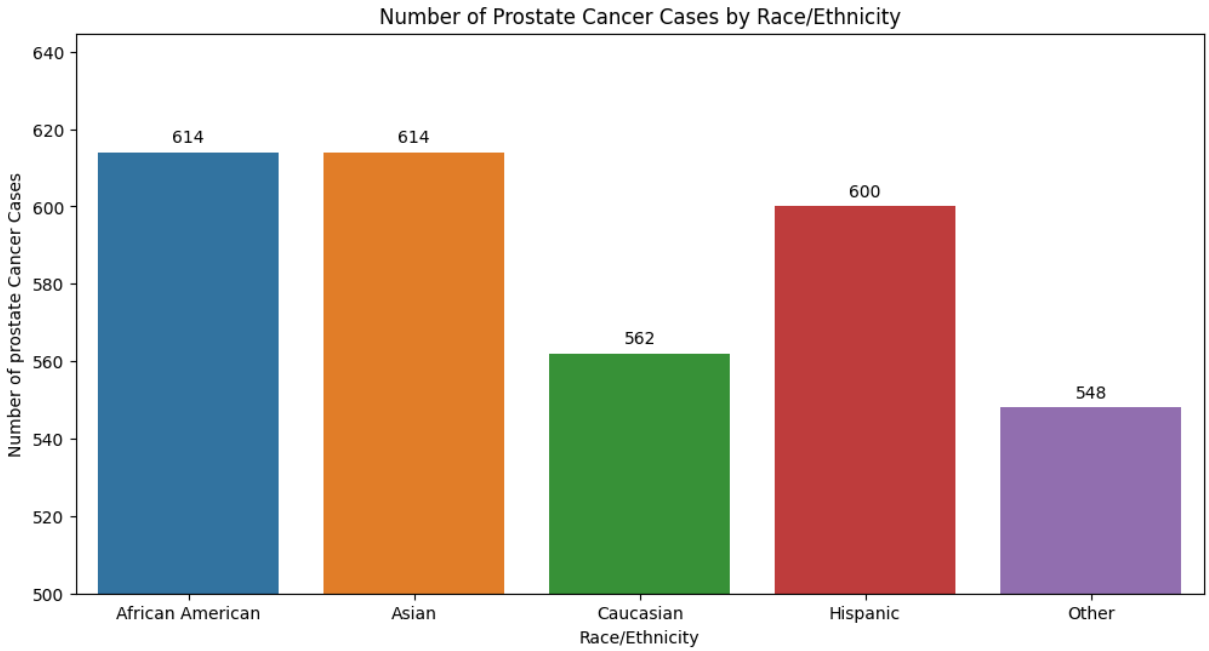


Key Findings

- Other category (Race / Ethnicity) has most breast cancers
- Asian people have more breast cancer compared to others.
- The number of breast cancer from highest to lowest are Other , Asian , African American, Hispanic and Caucasian respectively

In [238…
```
# Create a masking for Cancer Type of Prostate
df_prostate_cancer = df[df['CancerType']=='Prostate']
# Group by Race / Ethnicity
df_by_race_prostate_cancer = df_prostate_cancer.groupby('Race/Ethnicity').co
```

In [242…
```
# Lets plot the graphs

plt.figure(figsize=(12, 6))
ax =sns.barplot(data=df_by_race_prostate_cancer, x='Race/Ethnicity', y='Pati
# --------

# Add count labels on top of each bar
for p in ax.patches:
    ax.annotate(format(p.get_height(), '.0f'),
                (p.get_x() + p.get_width() / 2., p.get_height()),
                 ha = 'center', va = 'center', xytext = (0, 9),
                textcoords = 'offset points')

#---------

plt.title('Number of Prostate Cancer Cases by Race/Ethnicity')
plt.xlabel('Race/Ethnicity')
plt.ylim(500)
plt.ylabel('Number of prostate Cancer Cases')
plt.show()
```

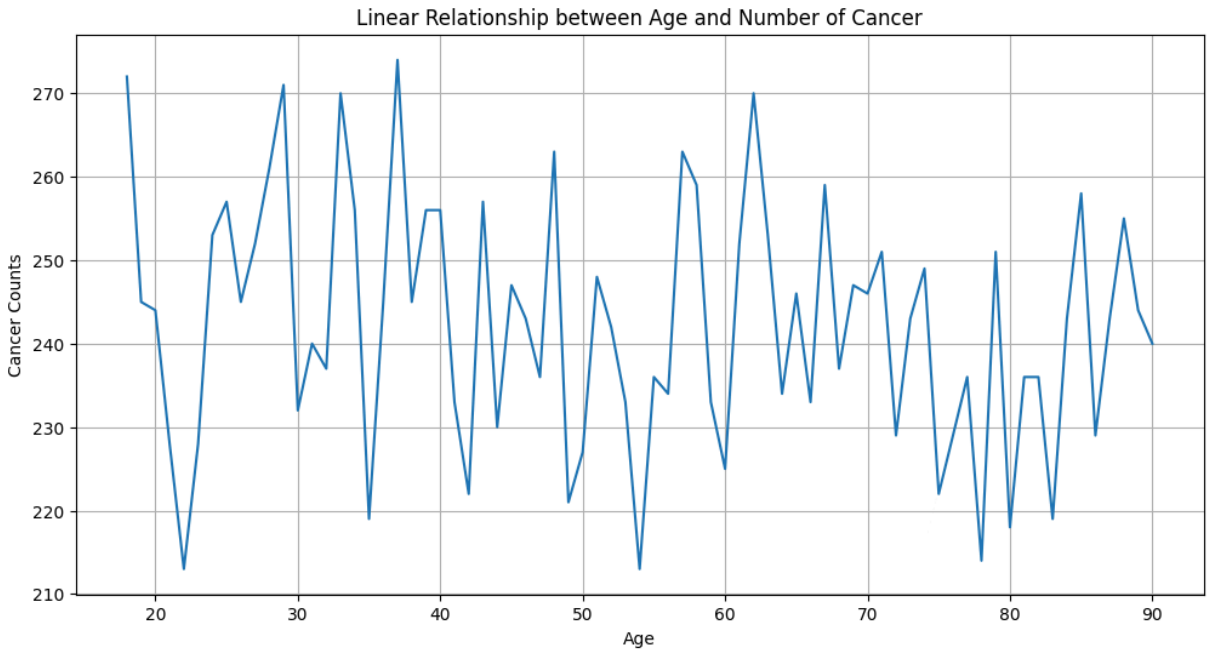


Key insights

- African american and Asian has most prostate cancer.
- Caucasian has least prostate cancer
- Need to verify the other category coz it has also large number of prostate cancers

In [268…

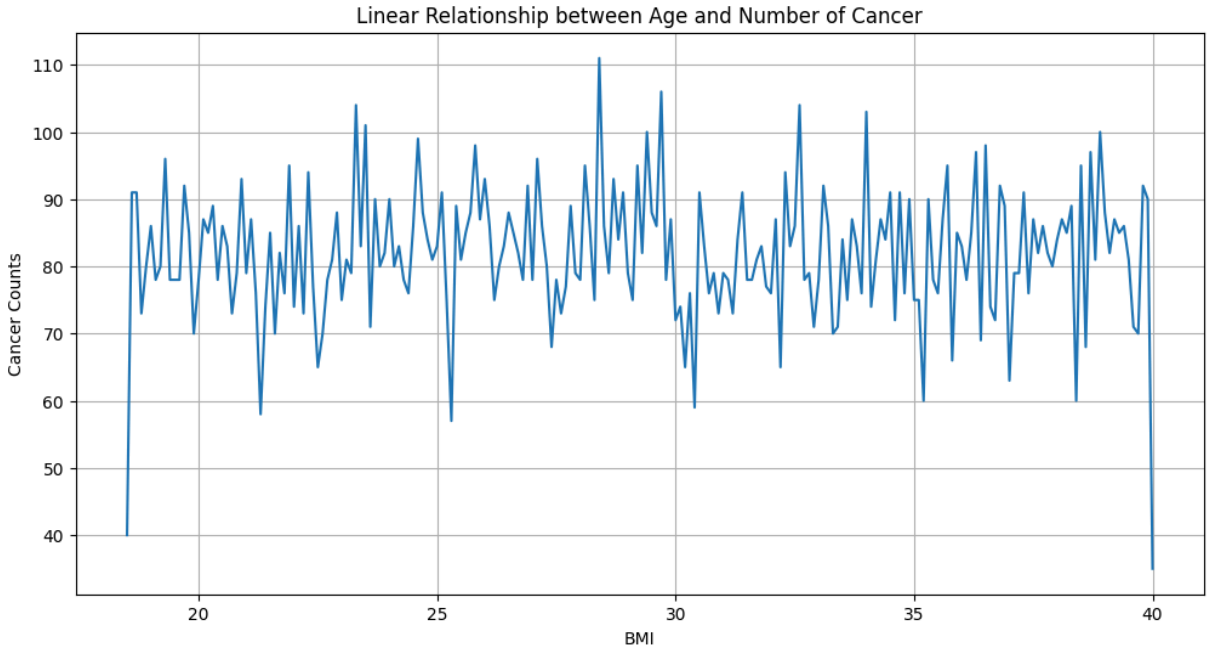
```
# Lets plot a line graph to check the correlation between the age and bmi.
df_by_age = df.groupby('Age')['CancerType'].count().reset_index()

plt.figure(figsize=(12, 6))
sns.lineplot(data=df_by_age, x='Age',y='CancerType')
plt.title('Linear Relationship between Age and Number of Cancer')
plt.xlabel('Age')
plt.ylabel('Cancer Counts')
plt.grid(True)
plt.show()
```



Key Insights

- There is no signicant relationship between Age and no of cancers

In [269…
```
# Lets plot a line graph to check the correlation between the BMI and no of
df_by_bmi = df.groupby('BMI')['CancerType'].count().reset_index()

plt.figure(figsize=(12, 6))
sns.lineplot(data=df_by_bmi, x='BMI',y='CancerType')
plt.title('Linear Relationship between Age and Number of Cancer')
plt.xlabel('BMI')
plt.ylabel('Cancer Counts')
plt.grid(True)
plt.show()
```



Key insights

- There is no high correlation between BMI and no of cancers

Summary

- The cancer dataset contains 17686 rows and 16 columns
- The genetic makers has so many null values, so i didnt use it for analysis
- There are less cancer occurences to younger people with age group 1-20 and elder one 81 plus and above
- The age group of 20-40 has more colon cancers compared to other
- The age group of 61 - 80 are more venerable to skin diseases and cancers
- There are more female who live shorter in terms of months from 1-20
- The survival time length is shorter for female while compared to male in if the survival length is in between 1-20 months
- The survial time length is longer for female during during the survival group of 21-40 months category

- African American have higher number of Leukemia and Prostate cancer compared to others
- Asian have higher Lung and Prostate cancers compared to others
- Caucasian have more Leukemia and Skin cancers compared to others
- Hispanic have higher no of Colon and Lung cancers
- Other category have Breast and Skin cancers (need to verify where the other category falls)
- Former Female smoker has more cancer numbers compared to male
- Female non smoker has more cancers
- Male smoker has more cancers compared to females
- There are no significant correlations between age and number of cancer cases
- There is no significant correlations between the bmi and number of cancer cases

Limitation

- The dataset is prefixed and conctrolled datasets.
- The data population sample is controlled one.
- Need to collect more samples and need to do more deeper analysis
- There is lots of null values in Genetic Makers and need to fill it out.
- "OTHER" many of the dataset has other categories, which takes large set of data sample. So need to pay attention to it.

Thanks -------- Cancer Issue Dataset - Licenced by Kaggle community

In [ ]: