Part_II_Breast_Cancer_Slides

October 11, 2022

1 Which types of patients are most likely to survive breast cancer?

by Klaus BONOU SELEGBE

1.1 Summary of Findings

Exploration showed that in this data set, 85% of patients are alive while approximately 15% are dead. The distribution of some interesting variables such as tumor size, stage variables, grade were studied. We learn from this exploration that - Alive patients are about 32 to 69 years old, their tumor size is among the smallest and they survive for a long time at least about more than 50 or 60 months - Dead patients are of all ages, their tumor size is also among the smallest, but not as alive patients...looking at the distribution, we could easily add that a large portion of dead patients have relatively large tumor with a size around 50 and 70 mm. They don't necessarily survive long before they die. - Proportionally, patients with an advanced stage for the 6th_stage (IIIC, IIIB, IIIA) and for n_stage(N3) are less likely to survive than if their cancer was at a lower stage (N1 or N2) or (IIA or IIB). - Proportionally patients with poorly differentiated or undifferentiated cells or whose carcinogenic cells are at an advanced grade (grade 3 and 4) are more likely to die than if their cells were well differentiated or if the carcinogenic cells were at a low grade (1 or 2) - Regarding marital status, separated, widowed, single and divorced patients respectively have less chance of surviving than married ones. - Finally The tumor_size increases as the stage or grade is high.

1.2 Investigation Overview

In this investigation, I wanted to examine patient characteristics that could be used to predict their vital status: whether they are more likely to survive or not. Emphasis was placed on variables describing cancer advancement such as cancer cell grade (grade), cancer cell invasion level (n_stage), cancer stage (6th_stage), tumor size (tumor_size). Relations with other variables such as the number of survival months (survival_months) and age will be briefly described.

1.3 Dataset Overview

The data consists of information about 4024 breast cancer patients. It presents each patient's age, marital status, different stages of cancer advancement, tumor size, number of months he survived before recovering or dying, vital status and many more. The dataset can be found here

```
In [2]: # import all packages and set plots to be embedded inline
        import numpy as np
        import pandas as pd
```

```
import matplotlib.pyplot as plt
import seaborn as sb

%matplotlib inline

# suppress warnings from final output
import warnings
warnings.simplefilter("ignore")
```

In [3]:
```
# load in the dataset into a pandas dataframe
breast_cancer_df = pd.read_csv('breast_cancer.csv')
```

1.3.1 Preliminary Wrangling

In [4]:
```
#Drop useless columns
breast_cancer_df.drop(['Regional Node Examined', 'Reginol Node Positive', 'T Stage '], a

# Rename column
breast_cancer_df.rename(columns = lambda x : x.strip().lower().replace(" ", "_"), inplac

# Replace " differentiated by ""
breast_cancer_df['differentiate'] = breast_cancer_df['differentiate'].apply(lambda x : x

# Replace " anaplastic; Grade IV" by "4"
breast_cancer_df['grade'] = breast_cancer_df['grade'].apply(lambda x : x.replace(" anapl
```

In [5]:
```
'''
    Change the type of column to categorical one with ordered values or not

            Parameters:
                    cols_dict (dict): A dict with column name as key and their values as
                    ordered (boolean): A boolean to indicate if values have to be oreder

            Returns:
                    void
'''
def change_type(cols_dict, ordered= True):
    if ordered :
        for col in cols_dict :
            breast_cancer_df[col] =  breast_cancer_df[col].astype(pd.api.types.Categoric
    else :
        for col in cols_dict :
            breast_cancer_df[col] =  breast_cancer_df[col].astype('category')
```

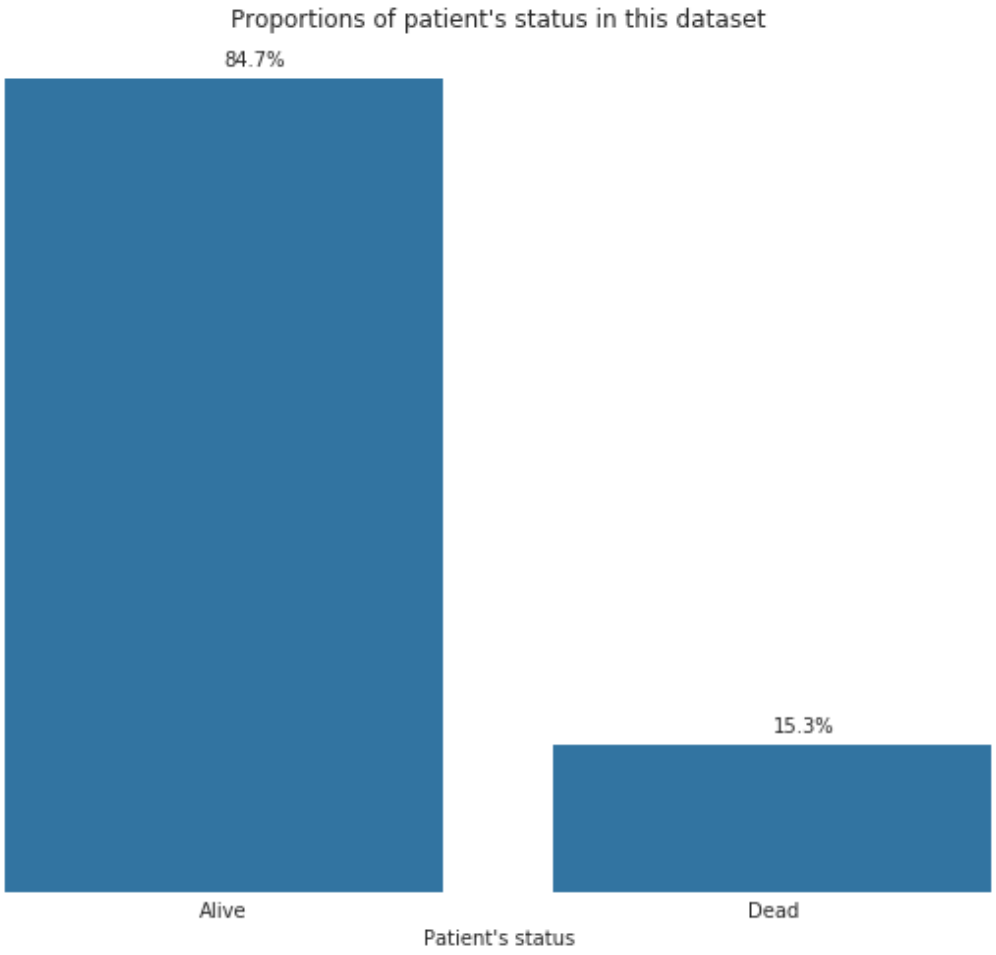
In [6]:
```
# Categorical columns with their ordered values
ordinal_col = {'n_stage': ['N1','N2','N3'],
               '6th_stage': ['IIA', 'IIB', 'IIIA', 'IIIB', 'IIIC'],
               'differentiate': ['Undifferentiated','Poorly', 'Moderately', 'Well'],
               'grade': ['1', '2', '3', '4']}
```

```
# Nominal columns with their values
nominal_col = {'race': ['White','Black','Other'],
               'marital_status': ['Single ', 'Married', 'Divorced', 'Widowed', 'Separat
               'a_stage': ['Regional', 'Distant'],
               'estrogen_status': ['Positive', 'Negative'],
               'progesterone_status': ['Positive', 'Negative'],
               'status': ['Alive', 'Dead']}

# The two lines below change the type of categorical columns to the rght types for both
change_type(ordinal_col)
change_type(nominal_col, ordered= False)
```

1.3.2 Distribution of Patient's status : The proportions

There is more patients alive than dead. Around 85% of patients of this dataset are alive whereas around 15% are dead.

```
In [9]: fig, axes = plt.subplots(figsize=[10,8])
        axes.get_yaxis().set_visible(False)

        base_color = sb.color_palette()[0]

        # Value counts for status variable
        status_count = breast_cancer_df['status'].value_counts()

        # Total number of values
        sum_count = breast_cancer_df.shape[0]

        #Plotting
        sb.countplot(data = breast_cancer_df, x = 'status', color = base_color);

        # This loop add the corresponding percent above each bar
        for i in range(status_count.shape[0]):
            count = status_count[i]
            percent_str = '{:.1f}%'.format(100*count/sum_count)
            plt.text(i, count+80, percent_str, va='center')

        # Enhance the visualization
        sb.set_style("whitegrid", {'axes.grid' : False, 'grid.linestyle': ''})
        axes.spines.clear()

        axes.set_title("Proportions of patient's status in this dataset");
        axes.set_xlabel("Patient's status");
```

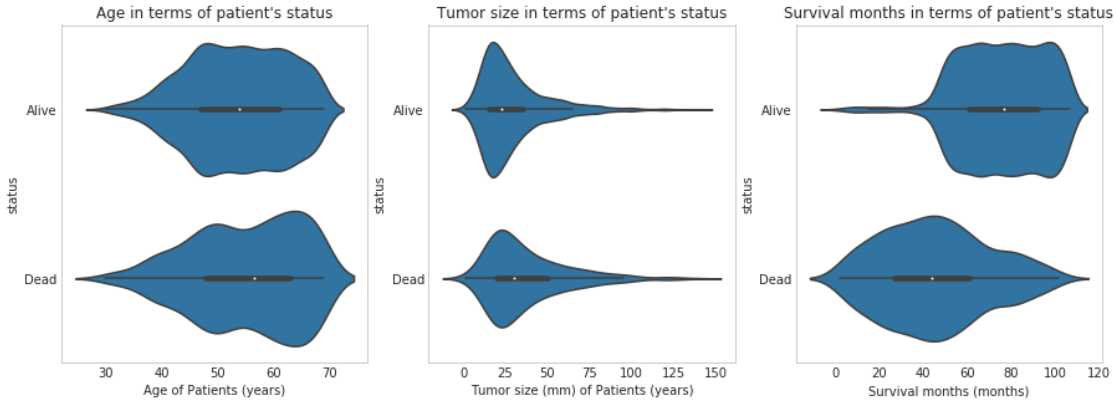


1.4 Patient's status vs their tumor size, age and survival months

- Alive patients are approximately 32 years old to 69, their tumor_size is among the smallest and they survive during a long time at least around 50 months
- Dead patients are of all ages, but a large number of dead patients are betweeen 60 and 70. Their tumor_size is among the smallest too but not as Alive patients...almost the same range, and they survive not necessarly a long time before dying.

```
In [8]: fig, axes = plt.subplots(ncols=3, figsize=[15,5])
        #axes.get_yaxis().set_visible(False)

        sb.violinplot(data=breast_cancer_df, x='age', y='status', ax=axes[0], color=base_color)
        sb.violinplot(data=breast_cancer_df, x='tumor_size', y='status', ax=axes[1], color=base_
        sb.violinplot(data=breast_cancer_df, x='survival_months', y='status', ax=axes[2], color=

        axes[0].set_title("Age in terms of patient's status")
        axes[0].set_xlabel('Age of Patients (years)');
```

```
axes[1].set_title("Tumor size in terms of patient's status")
axes[1].set_xlabel('Tumor size (mm) of Patients (years)');
axes[2].set_title("Survival months in terms of patient's status")
axes[2].set_xlabel('Survival months (months)');
```

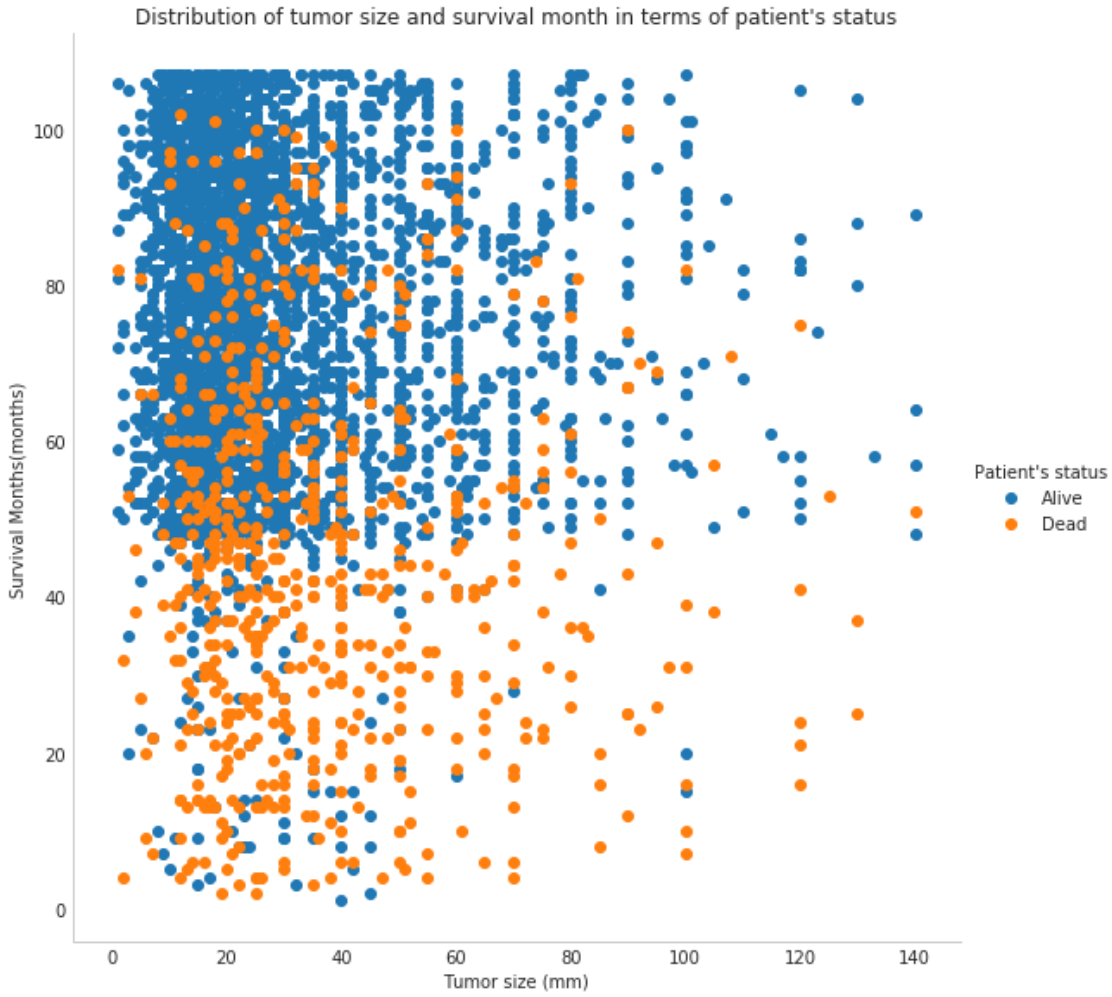


1.5 Patient's status vs both tumor size and survival months

Confirmation of the previous observations :

Alive patients in this dataset have mostly, a relative small tumor_size , and survive many months around over 50 or 60 Patients with tumor size great than 70mm are less likely to survive seing the the distribution of Dead patients in this dataset (15% of Dead patients) and the number of dead patients with tumor size great than 70mm.

```
In [11]: g = sb.FacetGrid(data=breast_cancer_df, hue='status', size=8)
         g.map(plt.scatter, 'tumor_size', 'survival_months');

         g.set(title= "Distribution of tumor size and survival month in terms of patient's statu
         g.add_legend(title="Patient's status");
```

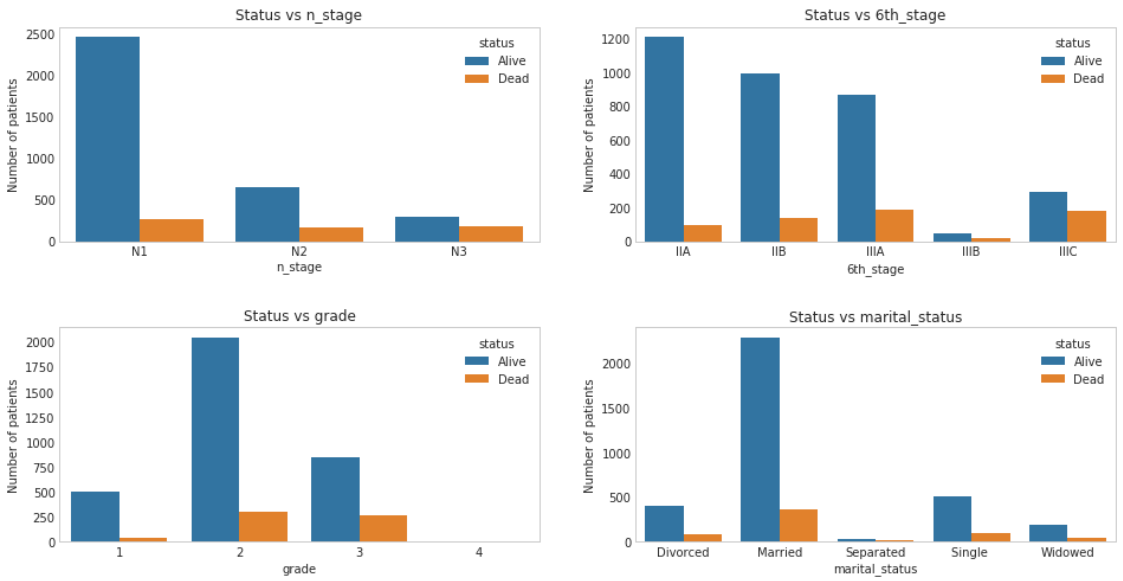


1.6 Influence of cancer stage on the patient's status

- Proportionally patients with advanced stage (N3), (it means with a high degree of invasion for carcinogenic cells) are less likely to survive than if their cancer was at a lower stage(N1 or N2).
- Patients with advanced stage for 6th_stage(IIIC, IIIB, IIIA), are less likely to survive than if their cancer's 6th stage was at a lower stage(IIA or IIB).
- Proportionally, patients die most when the carcinogenic cells is at advanced grade (grade 3 and 4)
- Concerning marital status, separated, widowed, single and divorced patients respectively are less likely to survive than married ones.

Patients with advanced stage are less likely to survive than if their cancer was at a lower stage.

In [13]:
```
'''
        Draw grouped grouped bar for precised variables

                Parameters:
                        axis_row (int): Number of rows of the axis in which we have to drau
                        axis_col (int): Number of cols of the axis in which we have to drau

                        x_vars (list): List of columns for which we have to draw grouped ba

                Returns:
                        grouped bar
        '''

        def draw_grouped_bar(axis_row, axis_col, x_vars):
            cpt = 0
            for i in range(axis_row):
                for j in range(axis_col):
                    sb.countplot(data = breast_cancer_df, x=x_vars[cpt], hue='status', ax=axes[
                    axes[i,j].set_title('Status vs '+x_vars[cpt])
                    axes[i,j].set_ylabel('Number of patients');
                    cpt +=1

        fig, axes = plt.subplots(2,2, figsize=[16,8])
        plt.subplots_adjust(wspace=0.2, hspace=0.4)

        draw_grouped_bar(2, 2, ['n_stage', '6th_stage', 'grade', 'marital_status'])
```



1.7 Influence of cancer stage (6th stage), tumor size on patient's status

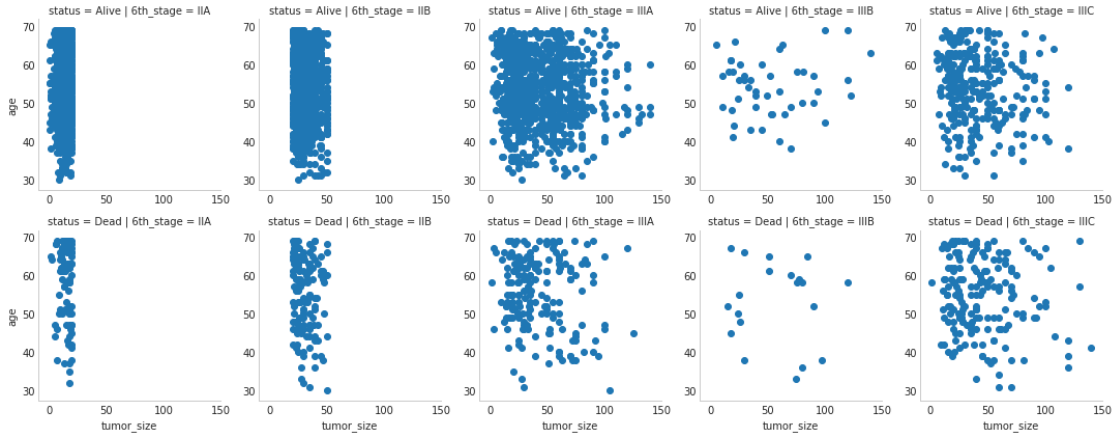
Let's try to understand more 6th stage values. ### Staging System for Breast Cancer

6th_stage	Tumor size	n_stage
IIA	T1 or T2	N1
IIB	T2	N1
IIIA	T1, T2, T3	N2
IIIA	T3	N1
IIIB	T4	N1 or N2
IIIC	AnyT	N3

Note : *T1 <=> [1mm - 20mm] T2 <=> [2mm - 50mm] T3 <=> More than 50mm T4 <=> Any size with direct extension to chest wall and/or to skin*

The most the stage is high, the less patients are likely to survive. The survival_months does not influence really here. However, the tumor_size increases as the stage is high

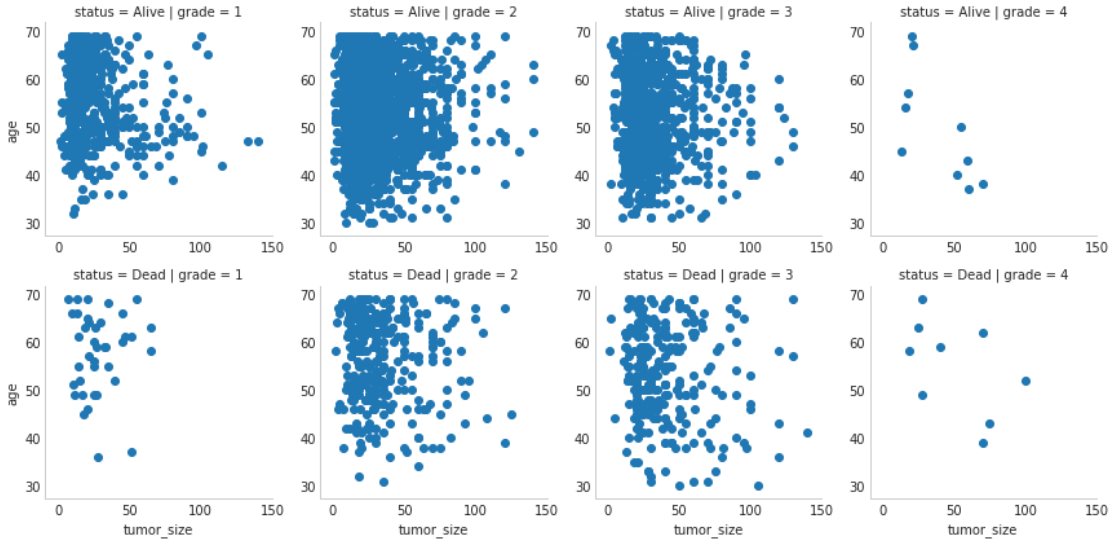
```
In [14]: '''
                    Draw scatter plots between 4 variables (2 categorical and 2 numerical)

                    Parameters:
                            cat_var (str) : column which represent the second categorical varia

                    Returns:
                            scatter plot
         '''

         def draw_multi_scatter_plot(cat_var):
             g = sb.FacetGrid(data=breast_cancer_df, row='status', col=cat_var, size=3)
             g.map(plt.scatter, 'tumor_size', 'age');
             g.add_legend();

         draw_multi_scatter_plot('6th_stage')
```



1.8 Influence of agressivness of the carcinogenic cells (grade), tumor size on patient's status

The most the grade is high (from garde 3), the less patients are likely to survive than the other grades. The tumor_size increases too as the grade is high

```
In [15]: draw_multi_scatter_plot('grade')
```



2 General Conclusion

2.1 Patients with small tumor size , a lower grade(the agressivness of the carcinogenic cells) and a lower cancer stage are more likely to survive.

2.1.1 Generate Slideshow

Once you're ready to generate your slideshow, use the `jupyter nbconvert` command to generate the HTML slide show.

```
In [ ]: # Use this command if you are running this file in local
        !jupyter nbconvert 'Part_II_Breast_Cancer_Slides.ipynb' --to slides --post serve --no-in
```

```
[NbConvertApp] Converting notebook Part_II_Breast_Cancer_Slides.ipynb to slides
[NbConvertApp] Writing 653901 bytes to Part_II_Breast_Cancer_Slides.slides.html
[NbConvertApp] Redirecting reveal.js requests to https://cdnjs.cloudflare.com/ajax/libs/reveal.j
Serving your slides at http://127.0.0.1:8000/Part_II_Breast_Cancer_Slides.slides.html
Use Control-C to stop this server
/usr/bin/xdg-open: 778: /usr/bin/xdg-open: x-www-browser: not found
/usr/bin/xdg-open: 778: /usr/bin/xdg-open: firefox: not found
/usr/bin/xdg-open: 778: /usr/bin/xdg-open: iceweasel: not found
/usr/bin/xdg-open: 778: /usr/bin/xdg-open: seamonkey: not found
```

```
/usr/bin/xdg-open: 778: /usr/bin/xdg-open: mozilla: not found
/usr/bin/xdg-open: 778: /usr/bin/xdg-open: epiphany: not found
/usr/bin/xdg-open: 778: /usr/bin/xdg-open: konqueror: not found
/usr/bin/xdg-open: 778: /usr/bin/xdg-open: chromium-browser: not found
/usr/bin/xdg-open: 778: /usr/bin/xdg-open: google-chrome: not found
/usr/bin/xdg-open: 778: /usr/bin/xdg-open: www-browser: not found
/usr/bin/xdg-open: 778: /usr/bin/xdg-open: links2: not found
/usr/bin/xdg-open: 778: /usr/bin/xdg-open: elinks: not found
/usr/bin/xdg-open: 778: /usr/bin/xdg-open: links: not found
/usr/bin/xdg-open: 778: /usr/bin/xdg-open: lynx: not found
/usr/bin/xdg-open: 778: /usr/bin/xdg-open: w3m: not found
xdg-open: no method available for opening 'http://127.0.0.1:8000/Part_II_Breast_Cancer_Slides.sl
```

2.1.2 Submission

If you are using classroom workspace, you can choose from the following two ways of submission:

1. **Submit from the workspace**. Make sure you have removed the example project from the /home/workspace directory. You must submit the following files:

- Part_I_notebook.ipynb
- Part_I_notebook.html or pdf
- Part_II_notebook.ipynb
- Part_I_slides.html
- README.md
- dataset (optional)

2. **Submit a zip file on the last page of this project lesson**. In this case, open the Jupyter terminal and run the command below to generate a ZIP file.

   ```
   zip -r my_project.zip .
   ```

   The command abobve will ZIP every file present in your /home/workspace directory. Next, you can download the zip to your local, and follow the instructions on the last page of this project lesson.

In [ ]: