Report for the Second Assignment (1st Part)

Data Processes

Academic year: 2022/23

Víctor Morcuende Castell

Guillermo Nájera Lavid

Javier Rocamora García

Antonio Ruiz García

TABLE OF CONTENTS

1. Dataset overview

As it is said in the project, the datasets we will be working on contain information about **patients of breast cancer** and **characteristics associated to their tumors**, as well as information about the **TNM classification of the tumor**.

To begin with and having already a quick view of the first dataset (*breast_cancer_data.csv*) after downloading it, we will proceed with a little analysis of the variables to know and comprehend in a better way what the data is telling us and their possible values:

- **ehr**: patient number => *identifier*
- **side**: breast side where the tumor => *right/left/nan*[1]
- **neoadjuvant**: whether a patient received neoadjuvant chemotherapy or not => *yes/no*
- **grade**: grade of the tumor (how quickly the malign cells are replicated) => *1/2/3*
    - ***grade '1':*** means the cancer is "well-differentiated" and resembles the tissue of origin
    - ***grade '2':*** intermediate grade. Most of the patients had a moderately differentiated tumor
    - ***grade '3':*** Cancer cells and tissue look very abnormal. These cancers are considered poorly differentiated since they no longer have an architectural structure or pattern. Grade 3 tumors are considered high grade.
- **er_positive**: whether the tumor has **estrogen receptors** => *1/0*
- **pr_positive**: whether the tumor has **progesterone receptors** => *1/0*
- **her2_positive**: whether the tumor has **HER2 overexpression** => *1/0*
- **ki67**: percentage (cellular marker) for **proliferation** => *%*
    - ⇨ ki67 is an excellent marker to determine the growth fraction of a given cell population and thus is strongly associated with tumor cell proliferation and growth
- **birth_date**: date of birth of the patient => *yyyy/mm/dd*
- **diagnosis_date**: date in which the **cancer was firstly diagnosed** => *yyyy/mm/dd*
- **death_date**: date of death of the patient => *yyyy/mm/dd*
- **recurrence year**: year where a previous cancer patient free of cancer get it again => *yyyy*
- **menarche_age:** age of the patient first menstruation
- **menopause_age:** age when the patient had first menstruation => *33-58 / nan*[1]
- **pregnancy:** patient number of pregnancies => *0-6 / nan*[1]
- **abort:** patient number of abortions => *0-3 / nan*[1]
- **birth:** patient number of natural births => *(-1)-6 / nan*[1]
- **caesarean:** patient number of cesareans => *0-2 / nan*[1]
- **hist_type:** histological type of the cancer => *ductal/lobular/unknown*

[1] *nan = unknown*

The second dataset used (*breast_cabcer_data_tnm.csv*) will give us the next information:

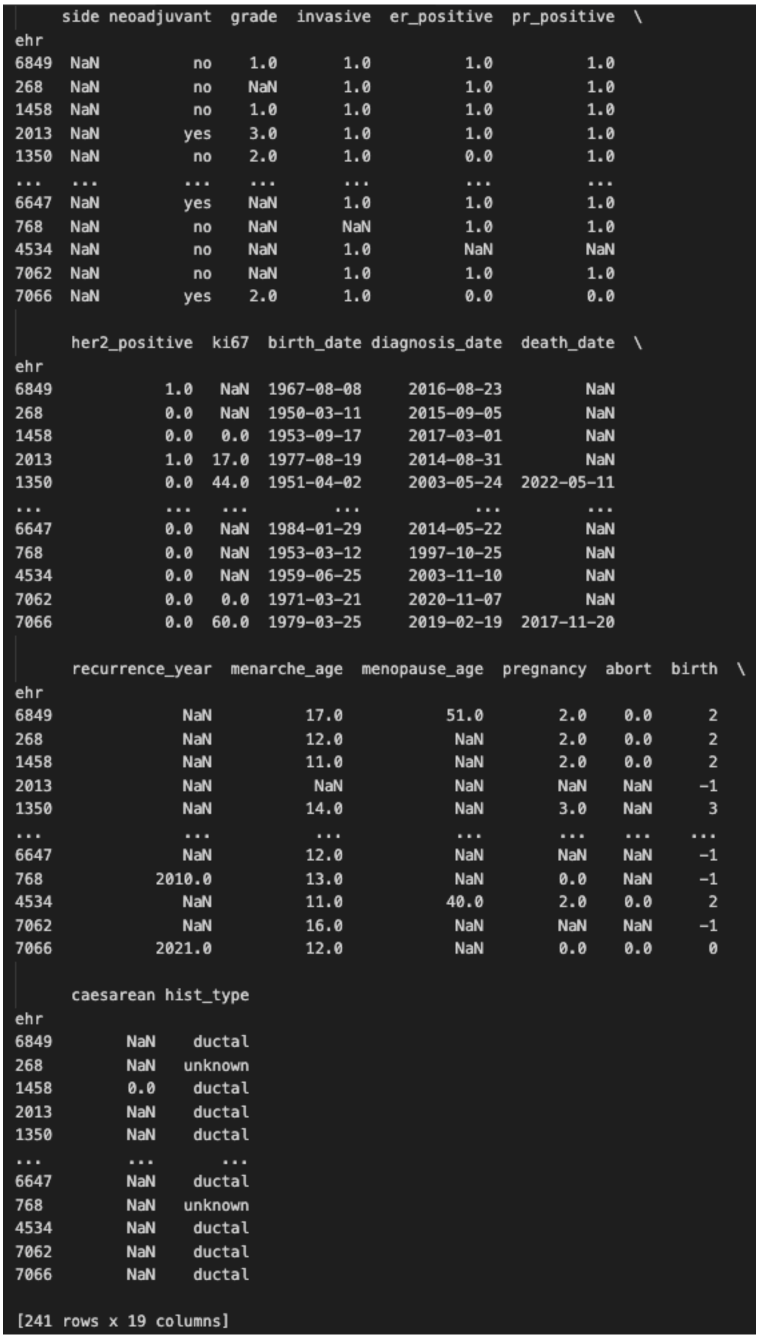
- **ehr**: patient number => *identifier*
- **n_tumor:** specific breast tumor identifier (increases for each ehr) => *1-3*
- **t:** T classification value of the breast tumor at diagnosis => *0-4/IS*
    - ⇨ this variable indicates where the tumor outbreak
        - ○ ***0***: there is no evidence about the primary tumor (couldn't be located)
        - ○ ***IS** (in-situ or precancer)*: means that cancer cells are reproducing in the topmost layer of tissue where the tumor originated, without invading deeper tissues.
        - ○ ***1 to 4***: describe the size of the tumor and how much it has spread to nearby structures in the body. The greater the T, the larger the tumor size and the greater the spread to nearby tissues.
- **n:** N classification value of the breast tumor at diagnosis => *0-3*
    - ⇨ this variable indicates whether the cancer has spread to nearby lymph nodes.
        - ○ ***0***: indicates that nearby lymph nodes do not contain cancer.
        - ○ ***1 to 3***: describe the size, location, and number of nearby lymph nodes affected by cancer. The higher the N number, the more cancer spreads to nearby lymph nodes.
- **m:** M classification value of the breast tumor at diagnosis => *0-1*
    - ⇨ this variable indicates whether the cancer has spread (if it has metastasized) to other parts of the body
        - ○ ***0***: indicates no distant spread of cancer.
        - ○ ***1***: indicates that there is spread to tissues or organs that are distant from where the cancer occurred.
- **t_after_neoadj:** T classification value of the breast tumor **after chemotherapy treatment** (neoadjuvant) => *0-4/IS*
- **n_after_neoadj:** N classification value of the breast tumor after chemotherapy treatment (neoadjuvant) => *0-3*
- **m_after_neoadjv:** M classification value of the breast tumor after chemotherapy treatment (neoadjuvant) => *0-1*

2. Preprocessing

Regarding the preparation, cleaning and preprocessing of the data, the first thing that we performed was a thorough analysis of the datasets' files (both excel and csv) in order to have a clear understanding of the whole background of the data as well as the variables contained in them, which represent different features or characteristics. Once having a clear understanding of the data, we had to work with, we decided to start the process of cleaning and preprocessing it thoroughly.

2.1. Preprocessing of Breast Cancer datasets

Firstly, we decided to start working on the data regarding the information about the patients and characteristics associated to their tumors, which is the data presented in the files 'breast_cancer_data.xlsx' as well as 'breast_cancer_data_2.xlsx'. In order to do so in a proper way, we decided that the main thing to do before going further was to merge both datasets in one. Indeed, as both datasets contain the same kind of information, we could perform the merge directly without further manipulation. Moreover, apart from deleting duplicated patients, we believed that the first column presented in the dataset (called "Unnamed: 0") was useless in this scenario, so we decided to delete it, since the dataset would have much more sense without it. The resultant dataset, which we will be working on, would be like this:



```
     side neoadjuvant  grade  invasive  er_positive  pr_positive  \
ehr
6849  NaN          no    1.0       1.0          1.0          1.0
268   NaN          no    NaN       1.0          1.0          1.0
1458  NaN          no    1.0       1.0          1.0          1.0
2013  NaN         yes    3.0       1.0          1.0          1.0
1350  NaN          no    2.0       1.0          0.0          1.0
...   ...         ...    ...       ...          ...          ...
6647  NaN         yes    NaN       1.0          1.0          1.0
768   NaN          no    NaN       NaN          1.0          1.0
4534  NaN          no    NaN       1.0          NaN          NaN
7062  NaN          no    NaN       1.0          1.0          1.0
7066  NaN         yes    2.0       1.0          0.0          0.0

      her2_positive  ki67  birth_date diagnosis_date  death_date  \
ehr
6849            1.0   NaN  1967-08-08     2016-08-23         NaN
268             0.0   NaN  1950-03-11     2015-09-05         NaN
1458            0.0   0.0  1953-09-17     2017-03-01         NaN
2013            1.0  17.0  1977-08-19     2014-08-31         NaN
1350            0.0  44.0  1951-04-02     2003-05-24  2022-05-11
...             ...   ...         ...            ...         ...
6647            0.0   NaN  1984-01-29     2014-05-22         NaN
768             0.0   NaN  1953-03-12     1997-10-25         NaN
4534            0.0   NaN  1959-06-25     2003-11-10         NaN
7062            0.0   0.0  1971-03-21     2020-11-07         NaN
7066            0.0  60.0  1979-03-25     2019-02-19  2017-11-20

      recurrence_year  menarche_age  menopause_age  pregnancy  abort  birth  \
ehr
6849              NaN          17.0           51.0        2.0    0.0      2
268               NaN          12.0            NaN        2.0    0.0      2
1458              NaN          11.0            NaN        2.0    0.0      2
2013              NaN           NaN            NaN        NaN    NaN     -1
1350              NaN          14.0            NaN        3.0    NaN      3
...               ...           ...            ...        ...    ...    ...
6647              NaN          12.0            NaN        NaN    NaN     -1
768            2010.0          13.0            NaN        0.0    NaN     -1
4534              NaN          11.0           40.0        2.0    0.0      2
7062              NaN          16.0            NaN        NaN    NaN     -1
7066           2021.0          12.0            NaN        0.0    0.0      0

      caesarean hist_type
ehr
6849        NaN    ductal
268         NaN   unknown
1458        0.0    ductal
2013        NaN    ductal
1350        NaN    ductal
...         ...       ...
6647        NaN    ductal
768         NaN   unknown
4534        NaN    ductal
7062        NaN    ductal
7066        NaN    ductal

[241 rows x 19 columns]
```

Figure 1: Resultant dataset

Then, we started to divide the variables into 2 types, categorical and numerical variables, for the purpose of creating a homogeneous file and to begin our data wrangling process. Therefore, variables were divided as follows:

- Categorical variables:

- **Side**: as there was a lot of empty values here, we decided to add a new category called “unknown” if neither “left” or “right” appeared. The reason behind using this approach instead of the Simple Imputer class with a “most_frequent” strategy was because, as there were not enough values, it would not have made sense to use the “most_frequent” strategy since it would have changed the proportions of this variable and would have given the dataset a completely different and unreal representation of this variable.

- **Neoadjuvant**: in this case, we chose to apply the Simple Imputer class with a “most_frequent” strategy since almost all the cells had values, so thanks to this technique we could impute the null values. After that, we concluded to convert the “yes” or “no” values into numerical 1s or 0s, since we thought that, although the original neoadjuvant variable was categorical, we believed that it was better to transform it to numerical, changing these values into 1s and 0s, therefore avoiding to create a new column when the One Hot Encoder was applied.

- **Grade**: as with neoadjuvant, we performed the Simple Imputer class with a “most_frequent” strategy as there was a great variety of the 3 types of grades (1, 2 or 3).

- **Invasive**: regarding this variable, the only value presented here were 1s, so we decided to impute the null values by adding a 0 in the null values, as we understood invasive as whether the tumor is invasive (1) or not (0).

- **Er_positive**: here we chose to apply the Simple Imputer class with a “most_frequent” strategy because almost all the cells had values, so thanks to this technique we could impute the null values.

- **Pr_positive**: the same procedure was performed as with the “er_positive” variable.

- **Her2_positive**: the same procedure was performed as with the “er_positive” and “pr_positive” variables.

- **Hist_type**: regarding this variable, we only had to apply the One Hot Encoder technique (explained below).

Then, for the “side”, “grade” and “hist_type” variables we decided to apply the One Hot Encoder transformation in order to represent with numerical values each of the categories of the 3 variables.

To achieve this, we previously decided to create a duplicate of the DataFrame “df_cat” (which contained these variables, the categorical ones) called “df_aux”, in order to pop the variables which were already modified for the purpose of not applying the One Hot Encoder to them. These variables were: “neoadjuvant”, “invasive”, “er_positive”, “pr_positive” and “her2_positive”.

Finally, we merge the DataFrame “df_aux” with the one which possessed the categorical values with the One Hot Encoder technique applied (“df_cat_ohe”), resulting in a DataFrame that contained all the categorical variables preprocessed (“df_cat_def”).

After this process, we decided to analyze and treat some variables before starting to work with the numerical ones. In particular, we decided to swap the variables “birth_date”, “diagnosis_date” and “death_date” by the age at which the patient was diagnosed (“age_diagnosed”), the time in years they survived since they were diagnosed (“survival_time”) and the time in years that took for them to fall back again since they were diagnosed (“recurrence_time”). We chose to make these changes since we thought that they would be of way more use to us when trying to use this date for any predictions. After that, we just deleted from the DataFrame of numerical variables (“df_num”) the unused variables and merge the new ones.

Moreover, at this point we decided to create a new variable called “survived” (which had the value 1 if the patient was alive or 0 if not), that would become our target variable, since we believe that the most important use of this data would be trying to predict, based on all these features, if a patient with breast cancer survived or not.

Then, continuing with our analysis, we began to work with the other types of variables, the numerical ones, which are explained below.

- Numerical variables:

- **Ki67**: in this case, as there were plenty of values in this column and it represents a percentage, we decided to apply the Simple Imputer class with a “mean” strategy, for the purpose of imputing the nulls with the mean of all the values.

- **Menarche_age**: for this variable, we chose to also apply the Simple Imputer class with a “mean” strategy for those few patients that did not have values. However, as a mean is applied, after this process we rounded up the resulting number, since it does not make much sense to have a 13,48 age value.

- **Menopause_age**: the same procedure was performed as with the “menarche_age” variable.

- **Pregnancy**: regarding this column, we decided to add the 0 value to that patients which did not have data, since we believed that it would not make sense in this case to apply the mean or the most frequent.

- **Abort**: the same procedure was performed as with the “pregnancy” variable.

- **Caesarean**: the same procedure was performed as with the “pregnancy” and “abort” variables.

- **Birth**: as this variable had all the column with values, the only change we performed was modifying the negatives values (such as -1s) and transforming them to 0, as it would not make sense to have negative births.

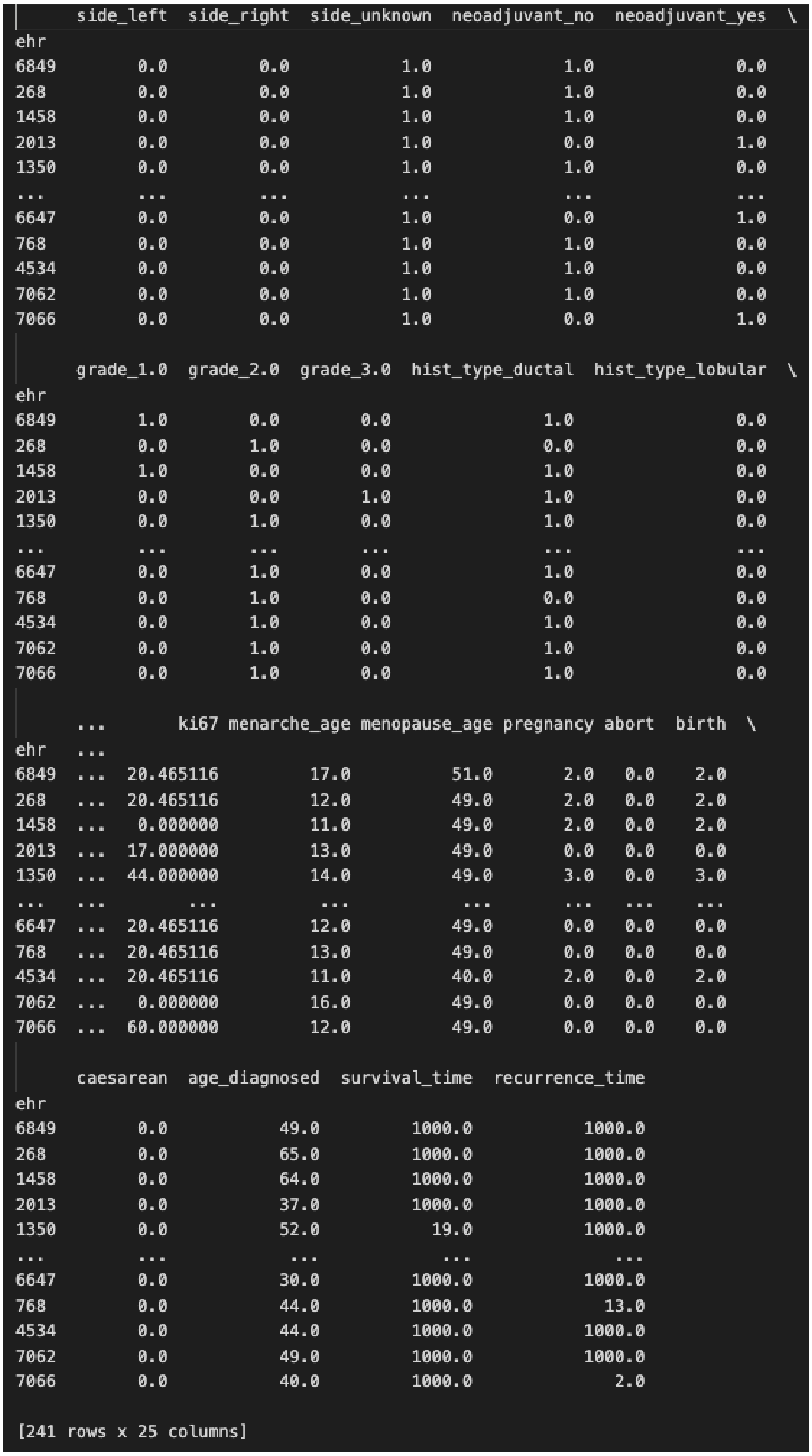
- **Age_diagnosed**: as explained above, for this variable we calculated through an apply function the age at which the patient was diagnosed with breast cancer.

- **Survival_time**: as we said, in this column we introduced the number of years which the patient lasted since they were diagnosed till their death. However, for the values in which the patient did not die, we decided to insert a huge value, for the purpose of clarifying that she is alive.

- **Recurrence_time**: as said before, this variable tells the time in years that took for the patient to fall back again in the cancer since the year they were diagnosed. As well as with “survival_time”, we chose to add a huge value in the cases where the patient did not fall back.

To end up with the numerical values, we need to emphasize that, since we saw some cases in which the number of pregnancies, aborts and births did not compute, meaning that there may have been less or more pregnancies than supposed due to contradictory data, we decided to increment the number of pregnancies for those cases that did not add up. We performed this through an auxiliary function which change the number of pregnancies if it was less or more than the sum of aborts and births.

To finish with this part, after we performed the above operations for both categorical and numerical variables, we merged the two DataFrames in order to have all the variables from the first 2 datasets together. The resulting dataset ("df_preprocessed") can be observed below:

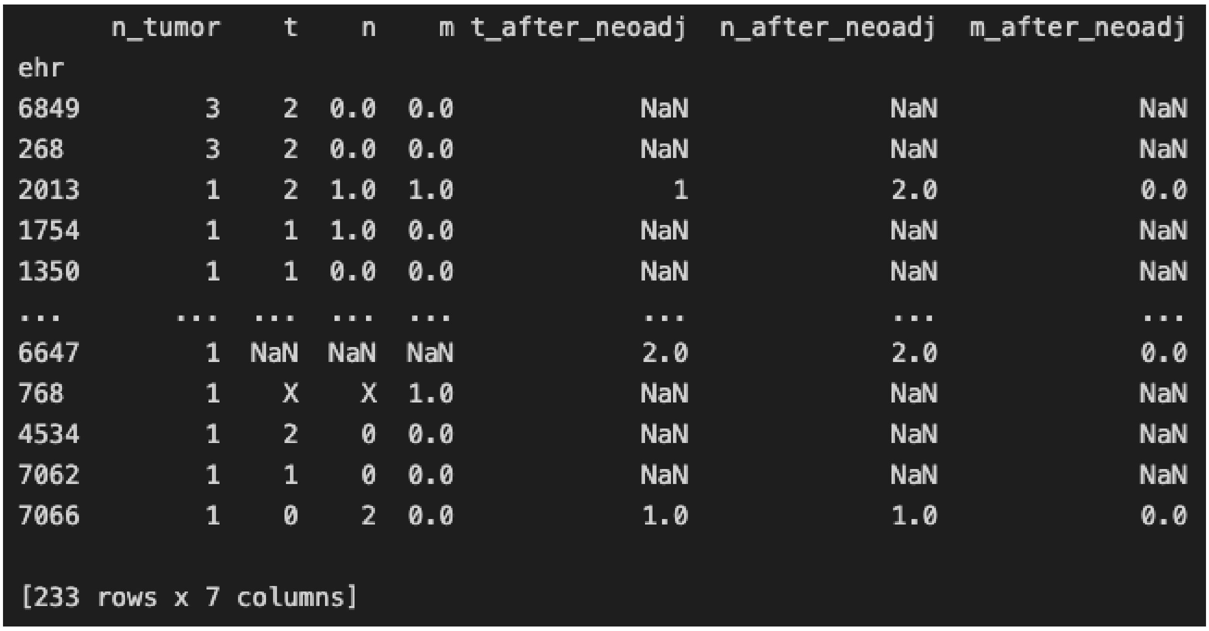


```
      side_left  side_right  side_unknown  neoadjuvant_no  neoadjuvant_yes  \
ehr
6849        0.0         0.0           1.0             1.0              0.0
268         0.0         0.0           1.0             1.0              0.0
1458        0.0         0.0           1.0             1.0              0.0
2013        0.0         0.0           1.0             0.0              1.0
1350        0.0         0.0           1.0             1.0              0.0
...         ...         ...           ...             ...              ...
6647        0.0         0.0           1.0             0.0              1.0
768         0.0         0.0           1.0             1.0              0.0
4534        0.0         0.0           1.0             1.0              0.0
7062        0.0         0.0           1.0             1.0              0.0
7066        0.0         0.0           1.0             0.0              1.0

      grade_1.0  grade_2.0  grade_3.0  hist_type_ductal  hist_type_lobular  \
ehr
6849        1.0        0.0        0.0               1.0                0.0
268         0.0        1.0        0.0               0.0                0.0
1458        1.0        0.0        0.0               1.0                0.0
2013        0.0        0.0        1.0               1.0                0.0
1350        0.0        1.0        0.0               1.0                0.0
...         ...        ...        ...               ...                ...
6647        0.0        1.0        0.0               1.0                0.0
768         0.0        1.0        0.0               0.0                0.0
4534        0.0        1.0        0.0               1.0                0.0
7062        0.0        1.0        0.0               1.0                0.0
7066        0.0        1.0        0.0               1.0                0.0

      ...       ki67 menarche_age menopause_age pregnancy abort  birth  \
ehr   ...
6849  ...  20.465116         17.0          51.0       2.0   0.0    2.0
268   ...  20.465116         12.0          49.0       2.0   0.0    2.0
1458  ...   0.000000         11.0          49.0       2.0   0.0    2.0
2013  ...  17.000000         13.0          49.0       0.0   0.0    0.0
1350  ...  44.000000         14.0          49.0       3.0   0.0    3.0
...   ...        ...          ...           ...       ...   ...    ...
6647  ...  20.465116         12.0          49.0       0.0   0.0    0.0
768   ...  20.465116         13.0          49.0       0.0   0.0    0.0
4534  ...  20.465116         11.0          40.0       2.0   0.0    2.0
7062  ...   0.000000         16.0          49.0       0.0   0.0    0.0
7066  ...  60.000000         12.0          49.0       0.0   0.0    0.0

      caesarean  age_diagnosed  survival_time  recurrence_time
ehr
6849        0.0           49.0         1000.0           1000.0
268         0.0           65.0         1000.0           1000.0
1458        0.0           64.0         1000.0           1000.0
2013        0.0           37.0         1000.0           1000.0
1350        0.0           52.0           19.0           1000.0
...         ...            ...            ...              ...
6647        0.0           30.0         1000.0           1000.0
768         0.0           44.0         1000.0             13.0
4534        0.0           44.0         1000.0           1000.0
7062        0.0           49.0         1000.0           1000.0
7066        0.0           40.0         1000.0              2.0

[241 rows x 25 columns]
```

Figure 2: Resultant dataset after several operations

2.2. Preprocessing of TNM datasets

In this part, the main objective of the datasets ‘breast_cancer_data_tnm.csv’ and ‘breast_cancer_data_tnm_2.csv’ is to give a detailed analysis of the characteristics of the tumors residing in the breast cancer patients. As with the previous set of datasets, we firstly combined them in one in order to work with it efficiently. After that, we erased the duplicated data, as we did before, keeping only the patients with the most severe tumors. The resultant dataset, which we will be working on, would be like this:



```
      n_tumor    t    n    m t_after_neoadj  n_after_neoadj  m_after_neoadj
ehr
6849        3    2  0.0  0.0            NaN             NaN             NaN
268         3    2  0.0  0.0            NaN             NaN             NaN
2013        1    2  1.0  1.0              1             2.0             0.0
1754        1    1  1.0  0.0            NaN             NaN             NaN
1350        1    1  0.0  0.0            NaN             NaN             NaN
...       ...  ...  ...  ...            ...             ...             ...
6647        1  NaN  NaN  NaN            2.0             2.0             0.0
768         1    X    X  1.0            NaN             NaN             NaN
4534        1    2    0  0.0            NaN             NaN             NaN
7062        1    1    0  0.0            NaN             NaN             NaN
7066        1    0    2  0.0            1.0             1.0             0.0

[233 rows x 7 columns]
```

Figure 3: Resultant dataset for tumors

In contrast with the previous dataset, in which there are great variety of variables, in this dataset we only have 7 variables, the number of tumors a patient has, the TNM classification of a tumor and the TNM classification of a tumor after neoadjuvant. Therefore, we decide to operate with all the variables together, without differentiating between numerical and categorical, since there is only one numerical (“n_tumor”).

First of all, we preprocessed the TNM and TNM after neoadjuvant variables. The way we performed this was by observing each variable and their dual after the treatment. In this case:

- If the first variable (“t”, “n” or “m”) was not null, it would keep that value
- If the first variable was null but the one after treatment was not, we inserted the value of the latter, meaning that the neoadjuvant treatment did not make any effect in that variable. However, if the variable after neoadjuvant was also null, we would insert a new value representing the category “unknown”, since we believe that it would not make sense to make up values for these variables.

For the values after neoadjuvant:

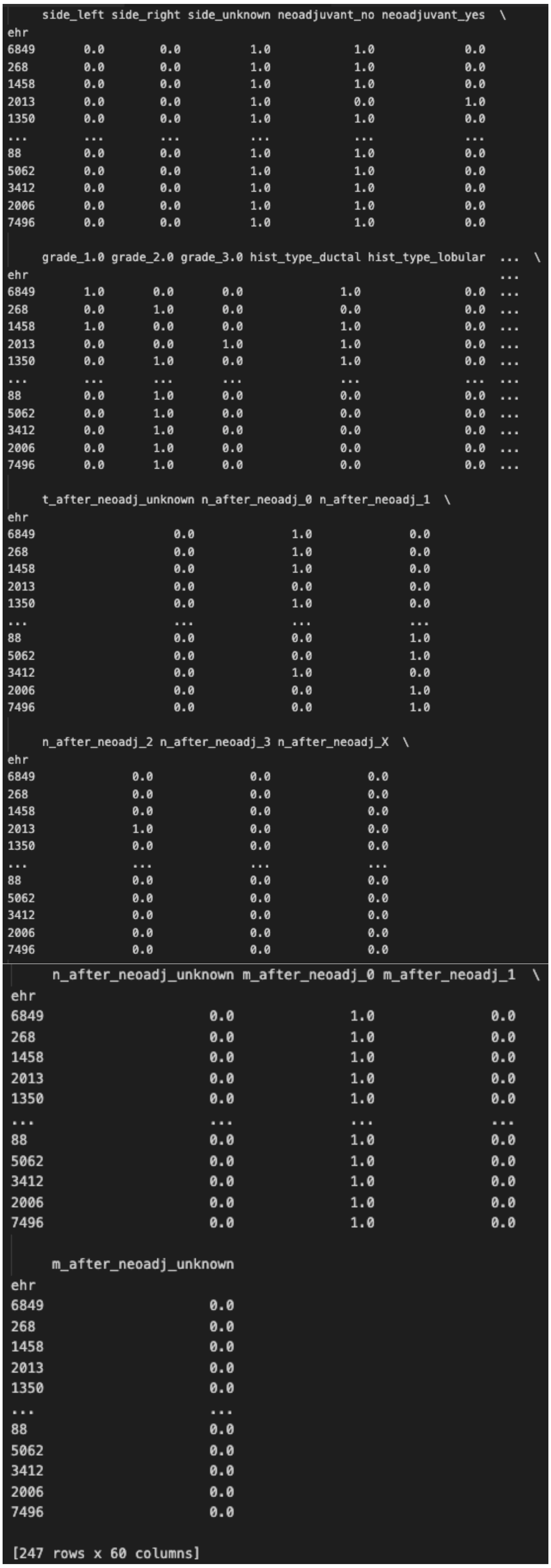
- If the variable (“t_after_neoadj”, “n_after_neoadj” or “m_after_neoadj”) was not null, it would keep that value

- If the variable was null but the original one (“t”, “n” or “m”) was not, we inserted the value of the original variable, meaning that the TNM values did not change after the neoadjuvant treatment.

After all the variables had imputed the nulls, we performed the One Hot Encoder for the categorical variables (all except “n_tumor”). In this way, as previously done, we duplicated the DataFrame “df_tnm” into a new DataFrame called “df_tnm_aux” in order to pop the “n_tumor” variable. Then, after the One Hot Encoder was applied to the rest of the variables, we merge the DataFrame “df_tnm_aux” with the One Hot Encoder one, resulting in a DataFrame that contained all the variables preprocessed (“df_tnm_def”).

Finally, once the 4 datasets were successfully cleaned, prepared and preprocessed, we decided to merge them, combining the “df_preprocessed” DataFrame (the one from the 2 first datasets), with the “df_tnm_def” DataFrame (the one from the 2 last ones). In order to use all the rows from the 4 datasets, that is, all the data, we chose to use the parameter “how” with “outer” value. The reason behind this decision lies in the fact that there are missing values in both datasets, meaning that there are patients in one dataset that do not correspond to any tumor in the other dataset and vice versa. By default, the merge method uses inner merge, instead of outer, which eliminates those entries that do not have matches on the other dataset. This result can also be achieved by using the join method, instead of merge.

Therefore, after we performed this merge, we decided to apply the Simple Imputer class with a “most_frequent” strategy in order to impute the nulls for these 14 variables. To end up, the final dataset can be appreciated below:

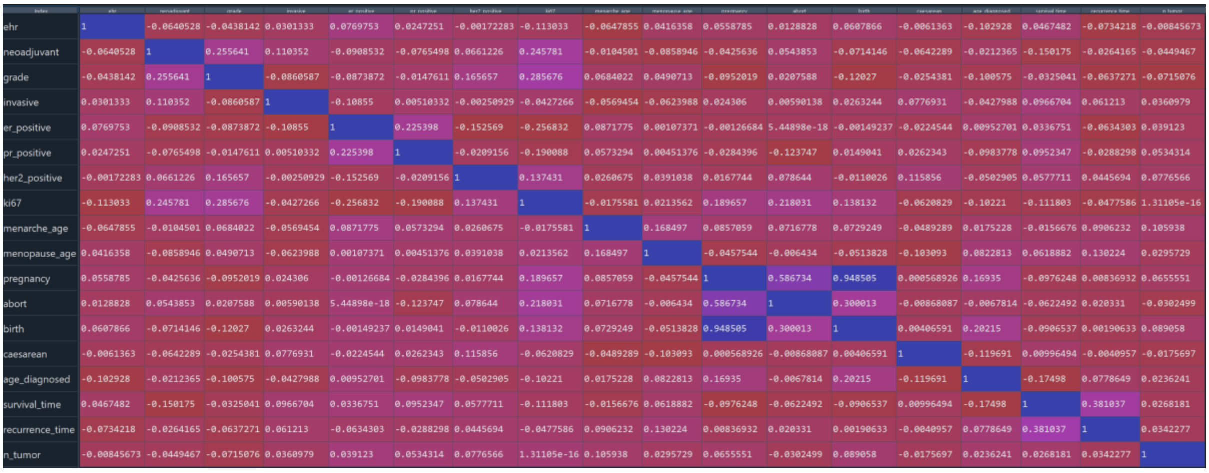


```
      side_left side_right side_unknown neoadjuvant_no neoadjuvant_yes  \
ehr
6849        0.0        0.0          1.0            1.0             0.0
268         0.0        0.0          1.0            1.0             0.0
1458        0.0        0.0          1.0            1.0             0.0
2013        0.0        0.0          1.0            0.0             1.0
1350        0.0        0.0          1.0            1.0             0.0
...         ...        ...          ...            ...             ...
88          0.0        0.0          1.0            1.0             0.0
5062        0.0        0.0          1.0            1.0             0.0
3412        0.0        0.0          1.0            1.0             0.0
2006        0.0        0.0          1.0            1.0             0.0
7496        0.0        0.0          1.0            1.0             0.0

      grade_1.0 grade_2.0 grade_3.0 hist_type_ductal hist_type_lobular  ...  \
ehr                                                                     ...
6849        1.0       0.0       0.0              1.0               0.0  ...
268         0.0       1.0       0.0              0.0               0.0  ...
1458        1.0       0.0       0.0              1.0               0.0  ...
2013        0.0       0.0       1.0              1.0               0.0  ...
1350        0.0       1.0       0.0              1.0               0.0  ...
...         ...       ...       ...              ...               ...  ...
88          0.0       1.0       0.0              0.0               0.0  ...
5062        0.0       1.0       0.0              0.0               0.0  ...
3412        0.0       1.0       0.0              0.0               0.0  ...
2006        0.0       1.0       0.0              0.0               0.0  ...
7496        0.0       1.0       0.0              0.0               0.0  ...

      t_after_neoadj_unknown n_after_neoadj_0 n_after_neoadj_1  \
ehr
6849                     0.0              1.0              0.0
268                      0.0              1.0              0.0
1458                     0.0              1.0              0.0
2013                     0.0              0.0              0.0
1350                     0.0              1.0              0.0
...                      ...              ...              ...
88                       0.0              0.0              1.0
5062                     0.0              0.0              1.0
3412                     0.0              1.0              0.0
2006                     0.0              0.0              1.0
7496                     0.0              0.0              1.0

      n_after_neoadj_2 n_after_neoadj_3 n_after_neoadj_X  \
ehr
6849               0.0              0.0              0.0
268                0.0              0.0              0.0
1458               0.0              0.0              0.0
2013               1.0              0.0              0.0
1350               0.0              0.0              0.0
...                ...              ...              ...
88                 0.0              0.0              0.0
5062               0.0              0.0              0.0
3412               0.0              0.0              0.0
2006               0.0              0.0              0.0
7496               0.0              0.0              0.0

      n_after_neoadj_unknown m_after_neoadj_0 m_after_neoadj_1  \
ehr
6849                     0.0              1.0              0.0
268                      0.0              1.0              0.0
1458                     0.0              1.0              0.0
2013                     0.0              1.0              0.0
1350                     0.0              1.0              0.0
...                      ...              ...              ...
88                       0.0              1.0              0.0
5062                     0.0              1.0              0.0
3412                     0.0              1.0              0.0
2006                     0.0              1.0              0.0
7496                     0.0              1.0              0.0

      m_after_neoadj_unknown
ehr
6849                     0.0
268                      0.0
1458                     0.0
2013                     0.0
1350                     0.0
...                      ...
88                       0.0
5062                     0.0
3412                     0.0
2006                     0.0
7496                     0.0

[247 rows x 60 columns]
```

Figure 4: Final dataset

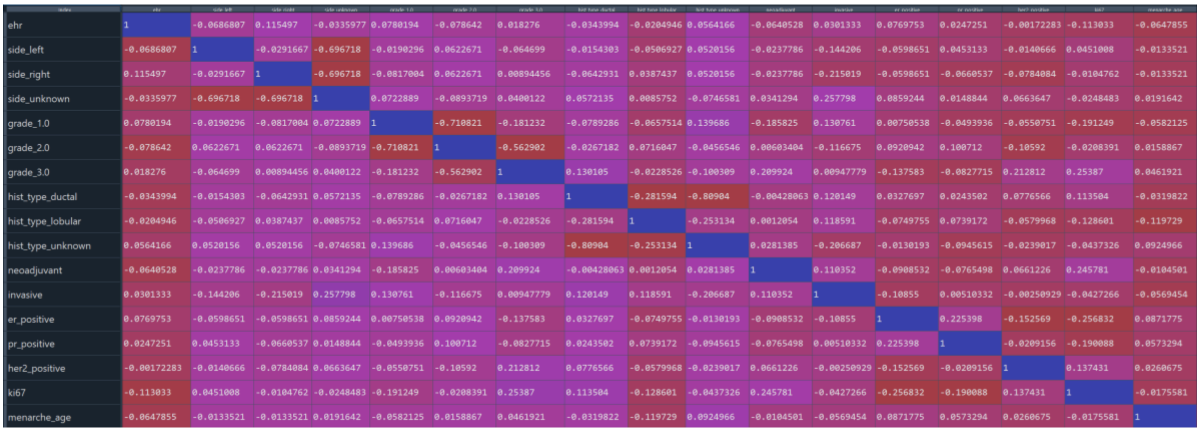
3. Analysis

Once having all the data cleared and pre-processed, now we can proceed with the task of generating a descriptive analysis of the latter. For the purpose of being able to visualize the data properly, we will consider the dataset generated before applying the One Hot Encoder, as each variable is contained in only one column. In this analysis we will be exploring each of the variables and commenting on those whose values throw interest insights or don't follow a common distribution. As there are almost 26 columns in the dataset (without considering the one hot encoder transformation which adds even more columns due to the categorical variables), plotting and analyzing all the possible correlations would be incredibly long and probably useless, since clearly not all the combinations will arise useful insights. In order to find and plot those correlations of interest, we will look at the correlation matrix from all the variables. On the graphic below, the correlation matrix between the numerical variables can be seen:



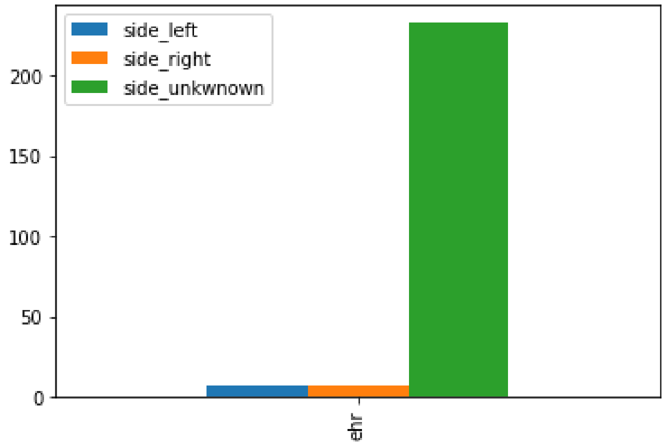
index	ehr	neoadjuvant	grade	invasive	er_positive	pr_positive	her2_positive	ki67	menarche_age	menopause_age	pregnancy	abort	birth	caesarean	age_diagnosed	survival_time	recurrence_time	n_tumor
ehr	1	-0.0640528	-0.0438142	0.0301333	0.0769753	0.0247251	-0.00172283	-0.113033	-0.0647855	0.0416358	0.0558785	0.0128828	0.0607866	-0.0061363	-0.102928	0.0467482	-0.0734218	-0.00845673
neoadjuvant	-0.0640528	1	0.255641	0.110352	-0.0908532	-0.0765498	0.0661226	0.245781	-0.0104501	-0.0858946	-0.0425636	0.0543853	-0.0714146	-0.0642289	-0.0212365	-0.150175	-0.0264165	-0.0449467
grade	-0.0438142	0.255641	1	-0.0860587	-0.0873872	-0.0147611	0.165657	0.285676	0.0684022	0.0490713	-0.0952019	0.0207588	-0.12027	-0.0254381	-0.100575	-0.0325041	-0.0637271	-0.0715076
invasive	0.0301333	0.110352	-0.0860587	1	-0.10855	0.00510332	-0.00250929	-0.0427266	-0.0569454	-0.0623988	0.024306	0.00590138	0.0263244	0.0776931	-0.0427988	0.0966704	0.061213	0.0360979
er_positive	0.0769753	-0.0908532	-0.0873872	-0.10855	1	0.225398	-0.152569	-0.256832	0.0871775	0.00107371	-0.00126684	5.44898e-18	-0.00149237	-0.0224544	0.00952701	0.0336751	-0.0634303	0.039123
pr_positive	0.0247251	-0.0765498	-0.0147611	0.00510332	0.225398	1	-0.0209156	-0.190088	0.0573294	0.00451376	-0.0284396	-0.123747	0.0149041	0.0262343	-0.0983778	0.0952347	-0.0288298	0.0534314
her2_positive	-0.00172283	0.0661226	0.165657	-0.00250929	-0.152569	-0.0209156	1	0.137431	0.0260675	0.0391038	0.0167744	0.078644	-0.0110026	0.115856	-0.0502905	0.0577711	0.0445694	0.0776566
ki67	-0.113033	0.245781	0.285676	-0.0427266	-0.256832	-0.190088	0.137431	1	-0.0175581	0.0213562	0.189657	0.218031	0.138132	-0.0620829	-0.10221	-0.111803	-0.0477586	1.31105e-16
menarche_age	-0.0647855	-0.0104501	0.0684022	-0.0569454	0.0871775	0.0573294	0.0260675	-0.0175581	1	0.168497	0.0857059	0.0716778	0.0729249	-0.0489289	0.0175228	-0.0156676	0.0906232	0.105938
menopause_age	0.0416358	-0.0858946	0.0490713	-0.0623988	0.00107371	0.00451376	0.0391038	0.0213562	0.168497	1	-0.0457544	-0.006434	-0.0513828	-0.103093	0.0822813	0.0618882	0.130224	0.0295729
pregnancy	0.0558785	-0.0425636	-0.0952019	0.024306	-0.00126684	-0.0284396	0.0167744	0.189657	0.0857059	-0.0457544	1	0.586734	0.948505	0.000568926	0.16935	-0.0976248	0.00836932	0.0655551
abort	0.0128828	0.0543853	0.0207588	0.00590138	5.44898e-18	-0.123747	0.078644	0.218031	0.0716778	-0.006434	0.586734	1	0.300013	-0.00868087	-0.0067814	-0.0622492	0.020331	-0.0302499
birth	0.0607866	-0.0714146	-0.12027	0.0263244	-0.00149237	0.0149041	-0.0110026	0.138132	0.0729249	-0.0513828	0.948505	0.300013	1	0.00406591	0.20215	-0.0906537	0.00190633	0.089058
caesarean	-0.0061363	-0.0642289	-0.0254381	0.0776931	-0.0224544	0.0262343	0.115856	-0.0620829	-0.0489289	-0.103093	0.000568926	-0.00868087	0.00406591	1	-0.119691	0.00996494	-0.0040957	-0.0175697
age_diagnosed	-0.102928	-0.0212365	-0.100575	-0.0427988	0.00952701	-0.0983778	-0.0502905	-0.10221	0.0175228	0.0822813	0.16935	-0.0067814	0.20215	-0.119691	1	-0.17498	0.0778649	0.0236241
survival_time	0.0467482	-0.150175	-0.0325041	0.0966704	0.0336751	0.0952347	0.0577711	-0.111803	-0.0156676	0.0618882	-0.0976248	-0.0622492	-0.0906537	0.00996494	-0.17498	1	0.381037	0.0268181
recurrence_time	-0.0734218	-0.0264165	-0.0637271	0.061213	-0.0634303	-0.0288298	0.0445694	-0.0477586	0.0906232	0.130224	0.00836932	0.020331	0.00190633	-0.0040957	0.0778649	0.381037	1	0.0342277
n_tumor	-0.00845673	-0.0449467	-0.0715076	0.0360979	0.039123	0.0534314	0.0776566	1.31105e-16	0.105938	0.0295729	0.0655551	-0.0302499	0.089058	-0.0175697	0.0236241	0.0268181	0.0342277	1

Moreover, on the graphic below, since we changed the categorical variables to numerical with the one hot encoder algorithm, now we are able to see the correlation matrix between all the possible numerical and categorical variables. It's a matrix with 60 rows and 60 columns, so we are showing here just a piece of it:

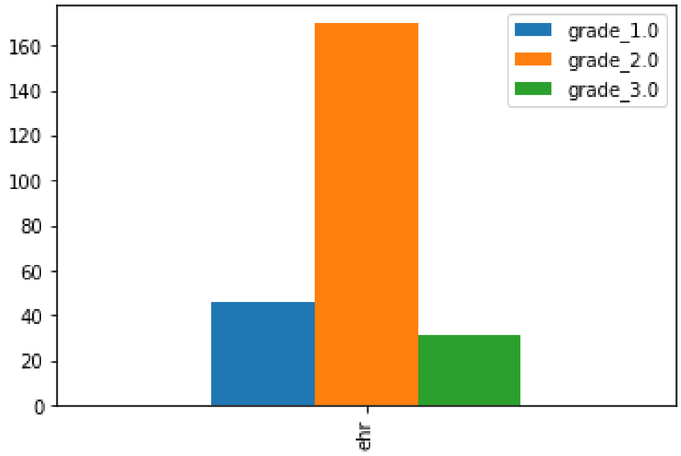


index	ehr	side_left	side_right	side_unknown	grade_1.0	grade_2.0	grade_3.0	hist_type_ductal	hist_type_lobular	hist_type_unknown	neoadjuvant	invasive	er_positive	pr_positive	her2_positive	ki67	menarche_age
ehr	1	-0.0686807	0.115497	-0.0335977	0.0780194	-0.078642	0.018276	-0.0343994	-0.0204946	0.0564166	-0.0640528	0.0301333	0.0769753	0.0247251	-0.00172283	-0.113033	-0.0647855
side_left	-0.0686807	1	-0.0291667	-0.696718	-0.0190296	0.0622671	-0.064699	-0.0154303	-0.0506927	0.0520156	-0.0237786	-0.144206	-0.0598651	0.0453133	-0.0140666	0.0451008	-0.0133521
side_right	0.115497	-0.0291667	1	-0.696718	-0.0817004	0.0622671	0.00894456	-0.0642931	0.0387437	0.0520156	-0.0237786	-0.215019	-0.0598651	-0.0660537	-0.0784084	-0.0104762	-0.0133521
side_unknown	-0.0335977	-0.696718	-0.696718	1	0.0722889	-0.0893719	0.0400122	0.0572135	0.0085752	-0.0746581	0.0341294	0.257798	0.0859244	0.0148844	0.0663647	-0.0248483	0.0191642
grade_1.0	0.0780194	-0.0190296	-0.0817004	0.0722889	1	-0.710821	-0.181232	-0.0789286	-0.0657514	0.139686	-0.185825	0.130761	0.00750538	-0.0493936	-0.0550751	-0.191249	-0.0582125
grade_2.0	-0.078642	0.0622671	0.0622671	-0.0893719	-0.710821	1	-0.562902	-0.0267182	0.0716047	-0.0456546	0.00603404	-0.116675	0.0920942	0.100712	-0.10592	-0.0208391	0.0158867
grade_3.0	0.018276	-0.064699	0.00894456	0.0400122	-0.181232	-0.562902	1	0.130105	-0.0228526	-0.100309	0.209924	0.00947779	-0.137583	-0.0827715	0.212812	0.25387	0.0461921
hist_type_ductal	-0.0343994	-0.0154303	-0.0642931	0.0572135	-0.0789286	-0.0267182	0.130105	1	-0.281594	-0.80904	-0.00428063	0.120149	0.0327697	0.0243502	0.0776566	0.113504	-0.0319822
hist_type_lobular	-0.0204946	-0.0506927	0.0387437	0.0085752	-0.0657514	0.0716047	-0.0228526	-0.281594	1	-0.253134	0.0012054	0.118591	-0.0749755	0.0739172	-0.0579968	-0.128601	-0.119729
hist_type_unknown	0.0564166	0.0520156	0.0520156	-0.0746581	0.139686	-0.0456546	-0.100309	-0.80904	-0.253134	1	0.0281385	-0.206687	-0.0130193	-0.0945615	-0.0239017	-0.0437326	0.0924966
neoadjuvant	-0.0640528	-0.0237786	-0.0237786	0.0341294	-0.185825	0.00603404	0.209924	-0.00428063	0.0012054	0.0281385	1	0.110352	-0.0908532	-0.0765498	0.0661226	0.245781	-0.0104501
invasive	0.0301333	-0.144206	-0.215019	0.257798	0.130761	-0.116675	0.00947779	0.120149	0.118591	-0.206687	0.110352	1	-0.10855	0.00510332	-0.00250929	-0.0427266	-0.0569454
er_positive	0.0769753	-0.0598651	-0.0598651	0.0859244	0.00750538	0.0920942	-0.137583	0.0327697	-0.0749755	-0.0130193	-0.0908532	-0.10855	1	0.225398	-0.152569	-0.256832	0.0871775
pr_positive	0.0247251	0.0453133	-0.0660537	0.0148844	-0.0493936	0.100712	-0.0827715	0.0243502	0.0739172	-0.0945615	-0.0765498	0.00510332	0.225398	1	-0.0209156	-0.190088	0.0573294
her2_positive	-0.00172283	-0.0140666	-0.0784084	0.0663647	-0.0550751	-0.10592	0.212812	0.0776566	-0.0579968	-0.0239017	0.0661226	-0.00250929	-0.152569	-0.0209156	1	0.137431	0.0260675
ki67	-0.113033	0.0451008	-0.0104762	-0.0248483	-0.191249	-0.0208391	0.25387	0.113504	-0.128601	-0.0437326	0.245781	-0.0427266	-0.256832	-0.190088	0.137431	1	-0.0175581
menarche_age	-0.0647855	-0.0133521	-0.0133521	0.0191642	-0.0582125	0.0158867	0.0461921	-0.0319822	-0.119729	0.0924966	-0.0104501	-0.0569454	0.0871775	0.0573294	0.0260675	-0.0175581	1

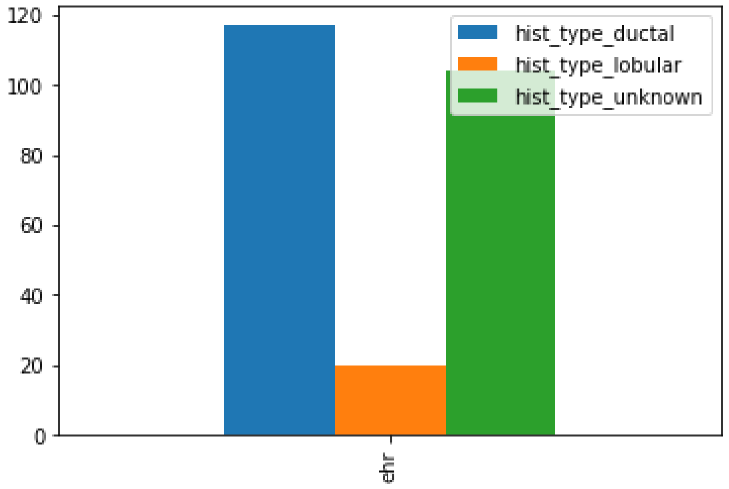
To begin with, as it can be seen in the next bar graphic plot, for most of the patients, the side of the tumor is unknown:



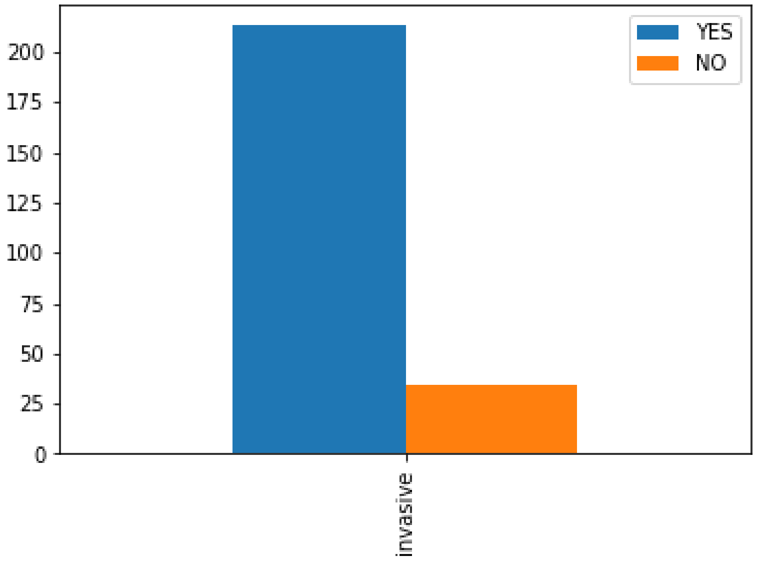
Regarding the grades of the tumors, as it can be appreciated in the next bar graphic plot, most of the patients in this dataset had a ‘grade 2’ tumor, indicating an intermediate grade, in other words, most of the patients had a moderately differentiated tumor:



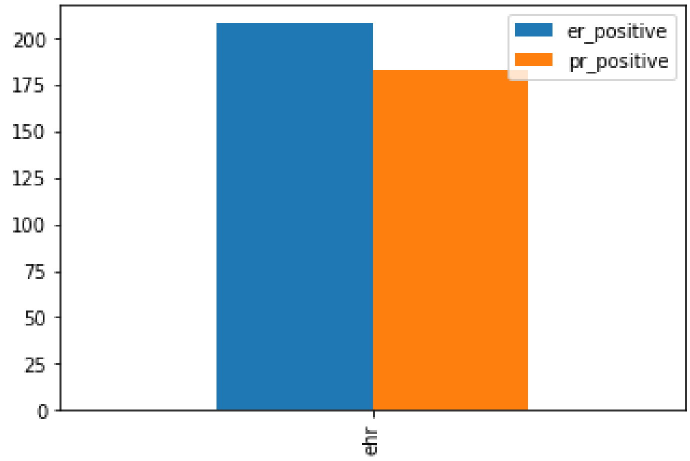
Paying attention to the histological type of the cancer of the patients, as it is shown in the next bar graphic plot, for the patients where the histological type was determined, a great majority had a ductal histological type of cancer (117 patients out of 247 in total). On the other hand, it was found that only 20 patients had a lobular histological type of cancer. The rest of values (104), which were unknown, still represent a great percentage with respect to the total, that’s the reason why it wasn’t worth to apply the Simple Imputer to them. Still, from the bar graphic plot one can realize that many of the unknown histological type could appear to be ductal instead of lobular.



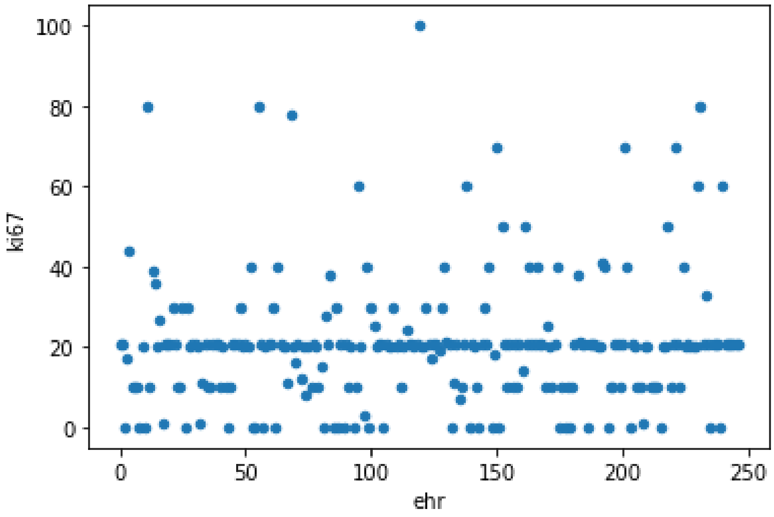
Furthermore, as it can be seen in the next bar graphic plot, it was found that most of the tumors from the patients were invasive, with only an amount of 34 tumors out of 247 not being invasive:



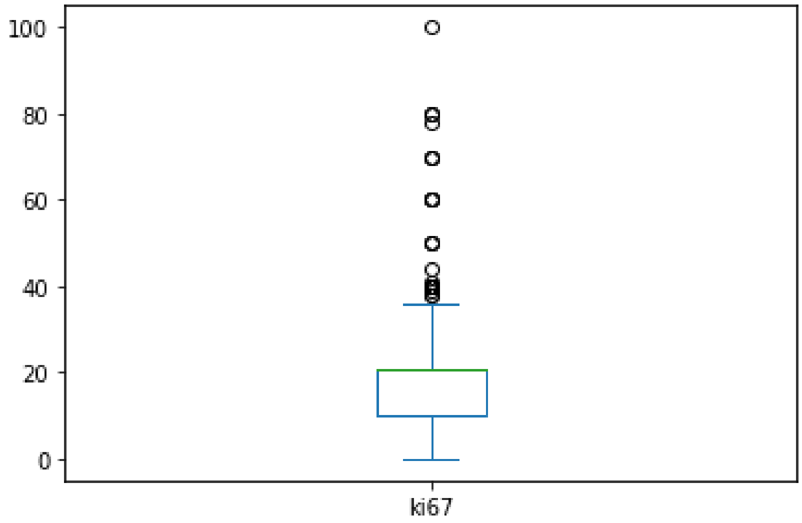
In relation to whether the tumor has any kind of receptors or not, as it can be appreciated in the next bar graphic plot, 208 out of 247 patients were detected estrogen receptors, while 183 patients were also detected progesterone receptors, meaning that both types of receptors were in most of the patients' tumors. In this case, since most of the patients have been detected both receptors, we arrive to the conclusion that both types of receptors are strongly correlated with the proliferation of tumors without further correlation graphics.



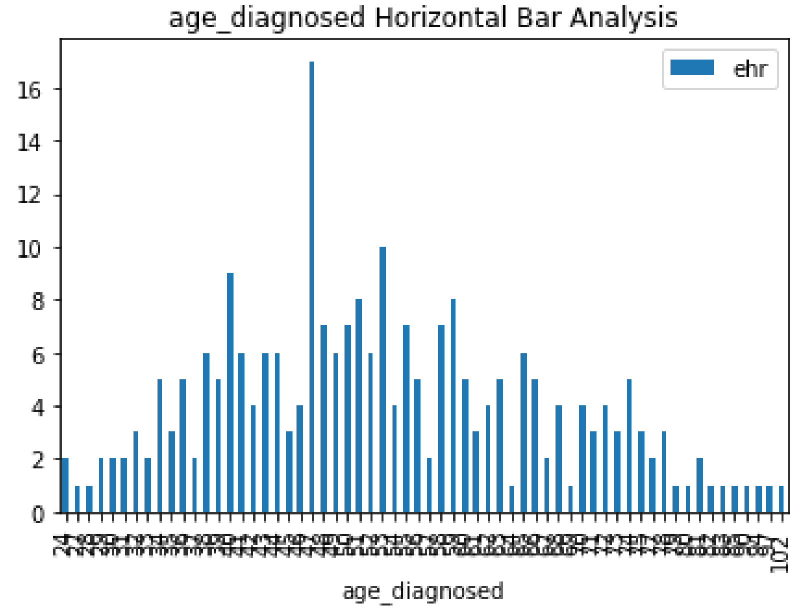
With respect to the percentage of proliferation of the tumor, as it is shown in the next scatter plot graphic, most of the patients have a 20% of proliferation. This means that in many patients, 2 in every 10 cells are dividing, which corresponds to a grade 2 neuroendocrine tumor (NET), meaning that the cells look more likely to grow and spread abnormally (they are also called moderately differentiated tumors).



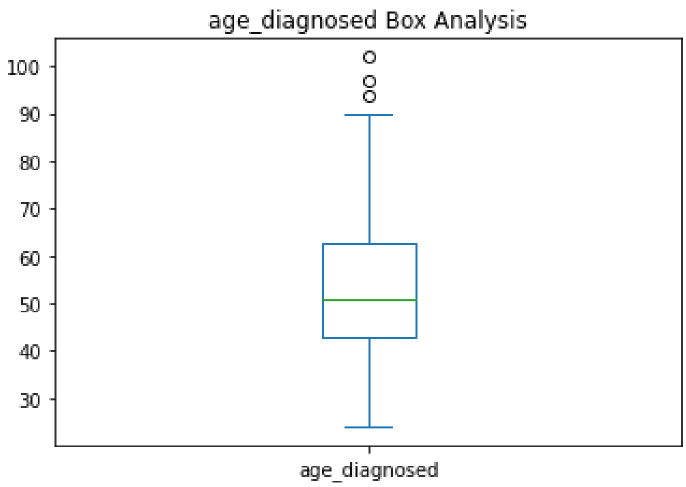
In the next box plot graphic, it can be seen more accurately how the values are distributed. As we were saying before, the mean is close to 20, as most of the values are there, and we can see that the percentiles between 0-40 and values over 40 are more difficult to be seen.



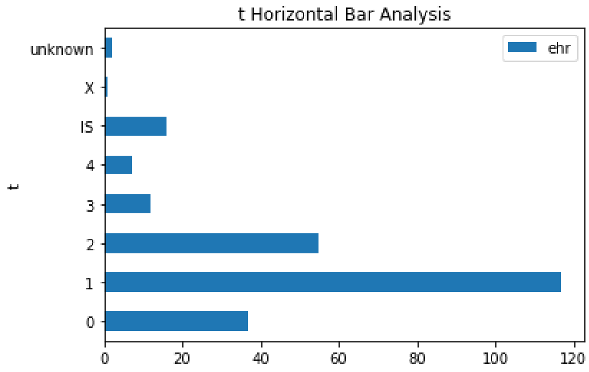
Regarding the age at which the tumor was diagnosed, as it can be observed in the next horizontal bar plot graphic, a great majority of patients (17 in total) were diagnosed at the age of 47. Also, from this graphic it can be deduced that the major part of detections take place at 40 years old and above.

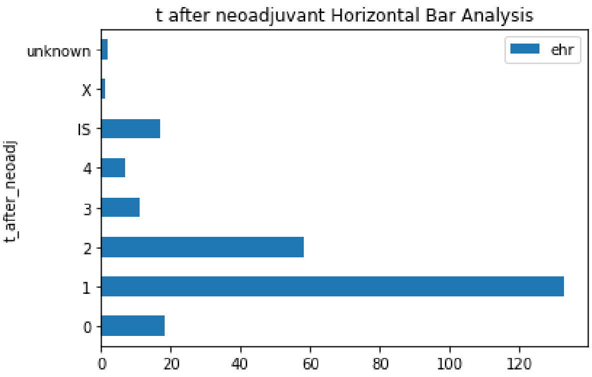


In the next box plot graphic, it can be appreciated more accurately how the values are distributed. As it can be seen, the inter-quartile range corresponds to the age group between 45-65 years old, and the middle quartile correspond approximately to the group of 47 years.

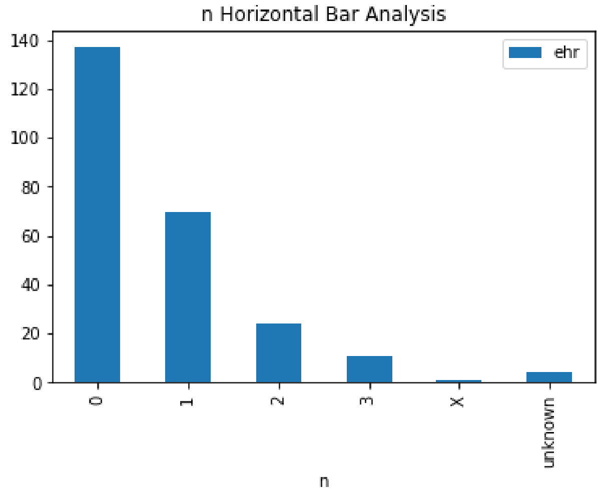


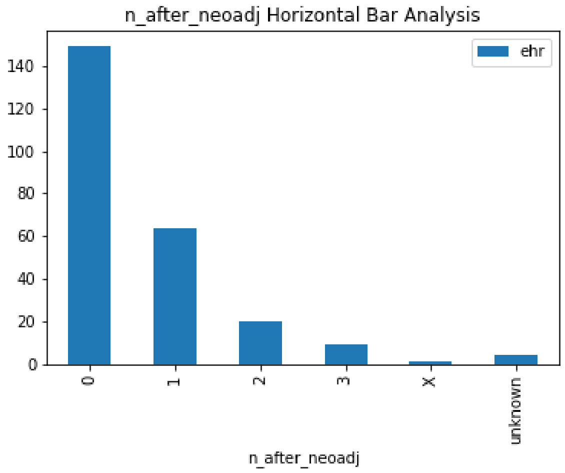
With respect to the T classification value of the tumor of the patient at diagnosis (t), as it can be seen in the horizontal bar plot graphic on the left, a great majority of patients (almost 120) were classified with a type 1 T classification value, which means that for most of the patients, the extent of the main tumor is not so large, at least at the time of diagnosis. The number of patients increases in the graphic of the right, which is the same but this time referring to the T classification value of the tumor of the patient after neoadjuvant chemotherapy treatment. That is expected, since it would mean that the treatment makes a difference for the patients.



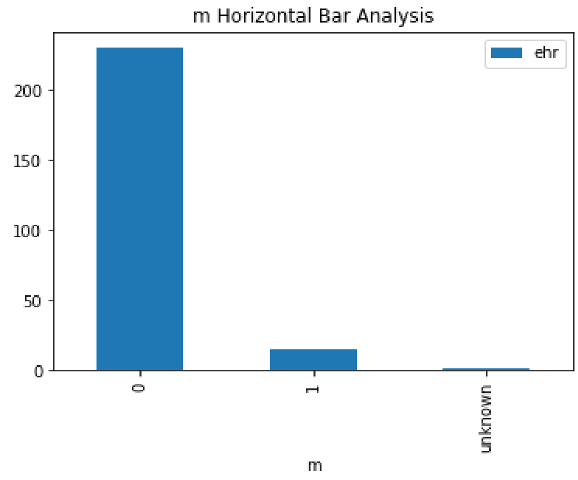


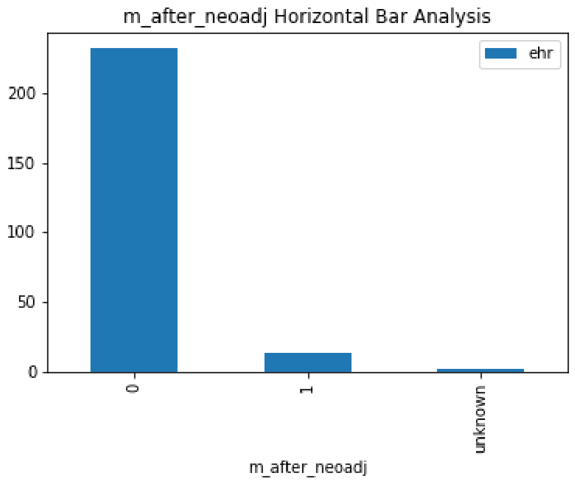
Paying attention to the N classification value of the tumor, as it can be observed in the horizontal bar plot graphic of the left, a great majority of patients (almost 140) were classified with a type 0 N classification value, which means that for most of the patients no lymph nodes have been affected at least at the moment of diagnosis. The number of patients increases in the graphic of the right, which is the same but this time referring to the N classification value of the tumor after neoadjuvant chemotherapy was applied. That makes sense since it would mean that the treatment indeed makes improvement on the patients.



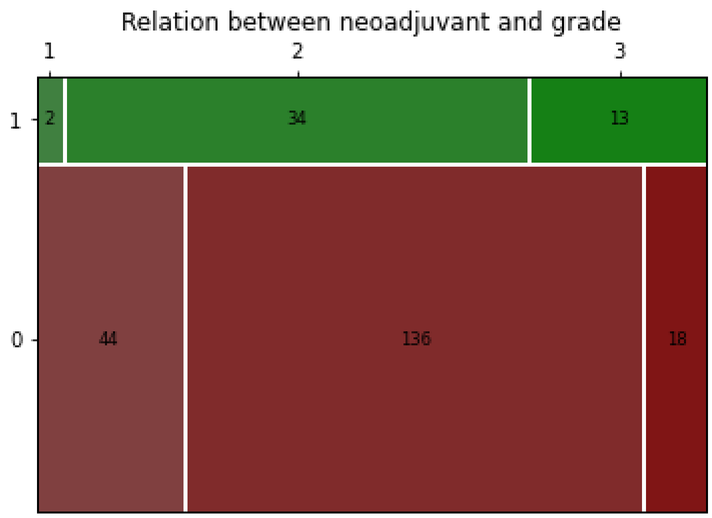


Regarding the M classification value of the tumor of the patient at diagnosis "m", as it can be seen in the horizontal bar plot graphic of the left, a great majority of patients (almost 230) were classified with a type 0 M classification value, which means that for most of the patients, the cancer has not metastasized at the moment of diagnosis. The number of patients increases in the graphic of the right, which is the same but this time referring to the M classification value of the tumor of the patient after neoadjuvant chemotherapy treatment. As with the previous T and N, this makes sense since it would mean that the treatment actually makes improvement on the patients.

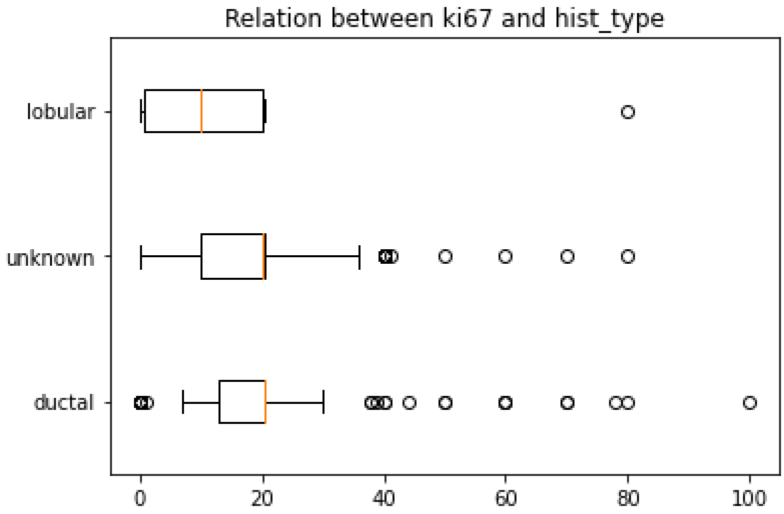




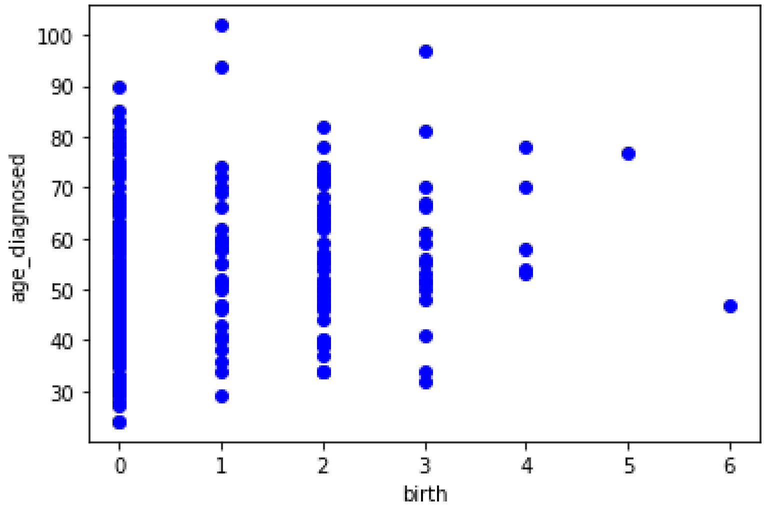
One of the interest correlations that we found was with the patients that received chemotherapy (neoadjuvant) and the grade of the tumors which indicate how quickly the malign cells are replicated. In this case, as it can be seen in the next mosaic graphic plot, the data shows how the patients who received neoadjuvant chemotherapy tent to have a higher grade of replication than the ones who didn't (those represent a 26% from the total against a 9% from the total in those patients who didn't receive chemotherapy).



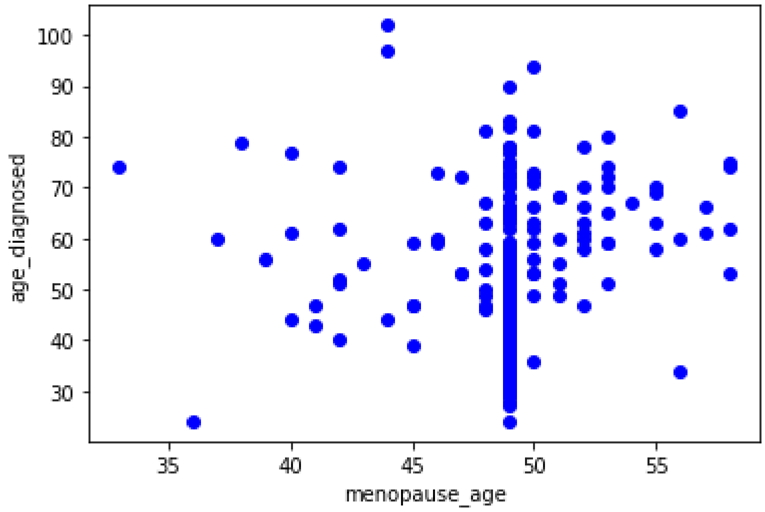
Another interesting correlation we found was between the cellular marker of proliferation (ki67) and the histological type of the cancer of the patient (hist_type). In this case, as it can be seen in the next box graphic plot, when the histological case of proliferation is lobular, then the cellular marker of proliferation tends to be lower than when the case is ductal or for other cases.



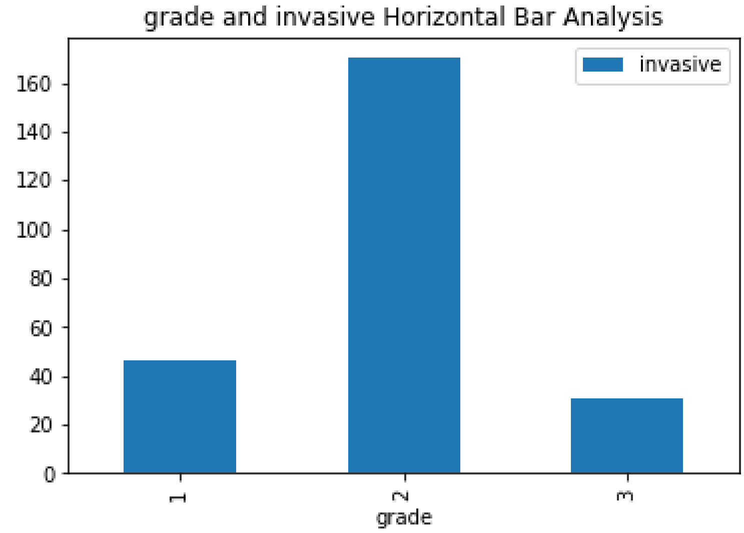
Also, another interesting correlation we found was between the age_diagnosed of the cancer (age_diagnosed) and the number of natural births of the patient (birth). In this case, as it can be seen in the next scatter graphic plot, it's not necessary to give birth to develop the cancer and indeed in this dataset most of the cases that were diagnosed didn't give birth. Between the birth values we found that in this dataset from those who gave birth, those who gave birth 2 times are the most propense to develop it.



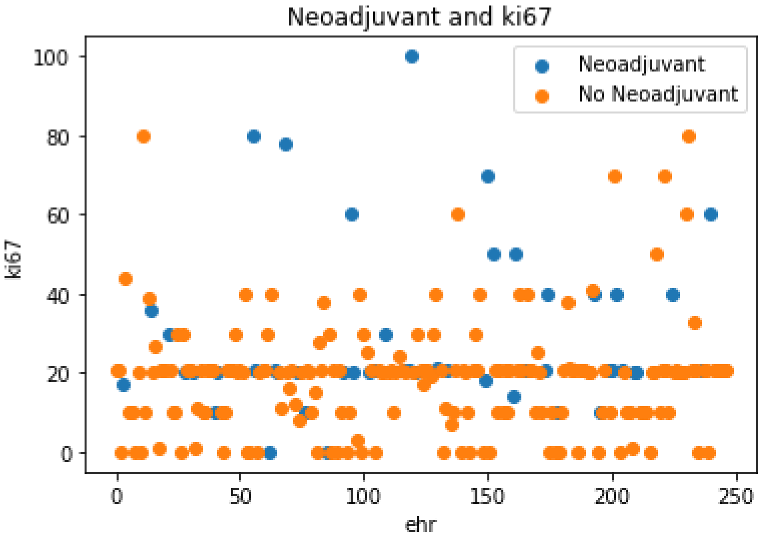
Another interesting correlation we found was between the age diagnosed of the cancer (age_diagnosed) and the age at which the patient had the menopause (menopauseage). In this case, as it can be seen in the next scatter graphic plot, most of the cases found are found in the median, which is 49, but still from those cases outside the normal, it's more usual in those who have a delayed menopause than an early one.



Another interesting correlation we found was between the grade of the tumor (grade) and whether if it was invasive or not (invasive). In this case, as it can be seen in the next bar graphic plot, from all the invasive cases, the grade of the tumor where most invasive cases were found was the tumor with grade 2.



In conclusion, we would like to include another interesting correlation that we found, which affects whether a patient has received neoadjuvant chemotherapy (neoadjuvant) and the percentage for proliferation (ki67). In this case, as it can be seen in the next scatter graphic plot, a number of patients that received neoadjuvant chemotherapy, can see their percentage of proliferation significantly reduced.



References

- Med-Surg Nursing: Tumor Classification & the 5 Types of Cancer - LevelUpRN
- Estadificación del cáncer (cancer.org)
- Neoadjuvant therapy - Wikipedia