Data Process second assignment
<First Part>

CAMPILLO MORALES Salvador
HUC-LHUILLERY Alexia
GONG Seong-Min
TRIBOUT Antoine

INDEX

1. INTRODUCTION

One of the most quoted and well-known principles in data science states the following: "Garbage in, garbage out". This implies that nothing that is done in a data analysis process will produce validable results allowing the extraction of useful knowledge if the data from which it starts are wrong. This can happen either because the sample taken is not representative of the population under analysis or because contamination has been introduced during the data extraction phase (too often not automated).

In addition, data analysis processes (statistical, machine learning based) require the analyst to have a deep understanding of the target data. This implies not only knowing the variables available, their meaning and the values they can take, but also getting additional insights about, among others, missing values, univariate statistical distributions or multivariate patterns on the data. Only in this way easily interpretable models can be produced, from which actual added value can be extracted.

The above two points highlight the importance of preliminary analysis and data cleaning before using the data for the intended purpose. Cleaning the data usually requires making decisions based on logical inferences and on the statistical, mathematical and computational resources available to the analyst, so it is not a linear process that can be executed following a set of fixed guidelines.

The aim of this practical application is to perform a complete set of cleansing and preprocessing tasks over a raw set composed of four datasets, producing a single output cleansed database containing all the information, ready to apply some further analysis on it. In fact, exploratory data analysis (both univariate and bivariate) will be applied over the cleansed results in order to get some useful insights about what's going on within the data.

These datasets belong to the clinical domain and contain information regarding cases of breast cancer. Two different kinds of datasets are present: one containing demographic and clinical information about patients and other containing information about tumor classification. As the assignment statement contains a complete list of the different variables and their meanings, a stop on this will be skipped, in order to avoid redundancies. The next sections go deeper into the characterization of the provided data, the problems found during the process and the logical, argued decisions taken on that.

The whole process of cleansing/preprocessing /exploring data has been performed using Pandas, a Python package providing a wide set of resources intended for that purpose, as the main frame of the project. As will be seen later, other Python packages (numpy and scikit-learn) will be punctually used.

2. DATA CLEANING AND PREPROCESSING

The following is a detailed explanation of the various cleaning and preprocessing tasks that have been undertaken to transform the raw data into a single dataset, now clean and available for further analysis.

2.1. Preliminary data loading / inspection / merging.

The EHR is an unique identifier associated with a patient and is present both on the individual patient data side and in tumor classification databases. It is the primary key that allows the instances of both types of databases (the ones regarding patient data and the others regarding tumor classification) to be merged into one. To do this, an inner join has been used, discarding those possible cases of patients for which there is only information in one of the two databases. Previously to that, the two pairs of identical datasets regarding its structure ('breast_cancer_data' and 'breast_cancer_data_2' on the one hand, /'breast_cancer_data_tnm' and /'breast_cancer_data_tnm_2' on the other) have been concatenated among them.

The main assumption done behind the merging process is that the basic unit in the final dataset will be a patient, so there is going to be only one row per each of them. Instances belonging to the /'breast_cancer_data_tnm' and /'breast_cancer_data_tnm_2 concatenated dataset having ehr's that appear more than one time have been deleted. This option has been chosen instead of other options (for instance, adding duplicated sets of columns for tumor classification or trying to use data structures as lists or dictionaries gathering all tumor classifications) because of two issues that could lead to confounding.

- Time sequential order of tumors isn't known when there are more than one in a patient.
- Only one diagnosis date is provided per patient, so that we can't assure which of the recorded tumors corresponds to that diagnosis date, or whether these tumors were diagnosed at the same time or not.

Moreover, that problem applies to an extremely low portion of cases (n=2), so a removal strategy is feasible and assumptions with reasonable confidence can be made regarding that hypothetical results are not going to be worsened due to this decision.

In addition, duplicates have been checked and dropped (n=3) from the 'breast_cancer_data' and 'breast_cancer_data_2' concatenated file.

Once done this first step, we have then checked the consistency of the merged output and looked at some useful general descriptors about the dataset, such as some statistical descriptors on the data, the proportion of missing values of each column and the type of each column and the number of rows with null values. A more precise insight about this can be achieved by executing the provided code, where these statistics are provided.

At that point, preliminary data analysis had led us to drop two columns, due to the following reasons:

- 'Unnamed: 0' which is an alphabetic incremental coding for instances, completely useless for data exploration/analysis and redundant with the fact that there is also an instance identifier.
- Side is a textual variable that is almost completely empty. This lack of filled values (n=229, ~94% of the total), added to the heterogeneity of filled ones, makes it impossible to perform some inference algorithm to recover loss of information with a reasonable confidence level, and prevents getting some useful information to it.

2.2. Cleaning and preprocessing tasks.

2.2.1. EHR and dates

Anonymization is a main issue when working with health records, especially if data is going to be shared with others. Malicious agents can get access to highly sensitive information about patients using variables as EHR or some dates. So, it is a good practice to delete or to mask sensitive data in order to prevent that possibility.

We have assumed that EHR's values correspond to actual patient EHR's, so we have used md5(), a hash function, over those values. The main property of hash functions relies on the fact that it is not possible to recover inputs using the outputs, which makes this kind of resource ideal for a purpose like that.

Regarding birth, diagnosis and death dates, a useful way to prevent this kind of problem is expressing that as ages. Doing that, information that is useful for analysis and knowledge retrieval is gathered in a way that ML algorithms can handle, and the sensitive part of that data is masked.

Ages at diagnosis, death and relapse have been taken with two decimals of precision, as well as the current age. For already dead patients, the current age field is not applicable so a blank value will be provided. For the age at relapse, given that only the year is provided, we have used 06-01 as date of reference, except for cases in which diagnosis_date and death_date are separated by less than year. In those cases, the age at relapse has been at the death date, in order to avoid setting a relapse date before diagnosis or after death. For cases in which death date came before relapse year, the relapse year value has been dropped.

Moreover, additional consistency checking have been performed.

- Following the GIGO principle, instances with impossible dates (in the future, taking 2022-12-01 as reference) have been considered as noisy and subsequently dropped out.
- The logical statements "date at diagnosis comes before age at death", "date at birth comes before age at diagnosis" and "date at diagnosis comes before age at relapse" have been checked. Given that we had some cases in which diagnosis and death dates were violating this rule, a flipping operation has been performed for those cases.

Finally, in the perspective of deleting all the dates to anonymize, we have added several two binary variables, potentially useful for exploration and analysis: "death_recorded" and "relapse recorded". For the cases in which death_recorded=0 or relapse_recorded=0, age_at_death and age_at_relapse respectively will be left as blank values, as filling operations are not applicable for those cases.

2.2.2. Processing data types

All different binary and categorical variables present in the dataset have been transformed to that, checking the absence of abnormal values and homogenizing some cases. For example, in columns regarding tumor grade and tumor classification, but also in some binary variables, several categorical levels are expressed both with doubles and integers. Prior to converting those variables to categorical ones, inconsistencies like that have been fixed. Moreover, neoadjuvant field values were expressed as "yes" and "no", so they have been converted to a 0/1 binary notation prior to converting to categorical, following the same homogenization purposes.

2.2.3. Dealing with missing values

Depending on the context of each of the different variables, several strategies have been applied over them in order to handle missing values, taking decisions based on logical inference. Next, a detailed relationship of that is presented.

2.2.3.1. Numeric variables

- Pregnancy, abort, birth, cesarean. Due to the high level of mutual information that these variables contain, looking overall is a desirable strategy in order to retrieve some missing values.

  There are cases in which a missing value regarding one variable can be recovered using the values for the other three variables. For instance, if pregnancy value is 0, then abort, birth and cesarean values are also 0. If we have non-missing data for pregnancies and births regarding a specific value, we can compute the number of abortions if it is not present. If the number of pregnancies is the same that the number of caesareans, the number both of abortions and births can be recovered if there is a missing value.

  Some additional assumptions have been taken in order to get rid of all missing values. For the cases in which pregnancy =NaN & abort=NaN & birth=NaN & cesarean=NaN, that four values have been assumed to be equal to 0. Finally, remaining cesarean missing values have been assumed to be equal to 0.

- Menarch age. It has been assumed that all patients have been fertile at some time. Missing values have been filled using the rest of filled values for menarche age variables. Median has been selected as the criterion to do that, with the aim of avoiding possible distortion effects caused by outliers, what could happen if mean were used.

- Menopause age. Unlike in menarche age, we can't assume that this variable is applicable to all patients. Patients can still be fertile, both at current time and at diagnosis date. What we have done is to compute the median for filled menopause age cases, assign that median value to patients older than that age and create an additional binary variable stating whether the patient had already reached menopause at diagnosis time or not.

  For patients younger than menopause median age that does not have a menopause age value recorded and for patients died at an age prior to menopause median age that does not have a menopause age value recorded, logical inference rules are not applicable, so that value remains missing, assuming that there is not menopause.

- Ki67. Again, a median strategy has been adopted in order to fill missing values. In that case, a regression strategy has been also tried, once filling missing values for the rest of both numeric and categorical variables. The key idea behind that is to see if it is possible to use the rest of available variables to predict with an acceptable accuracy ki67 values (improving then the expected results for a mean/median strategy).

  The dataset has been divided in two subsets: one with cases with filled values for ki67 and the other with the missing ones. For the first one, a linear regression model has been built using the scikit-learn machine learning package on Python.

  Algorithm accuracy has been evaluated using 10-fold cross validation and determination coefficient as key criterion. Due to the really poor fit, that strategy has been discarded.

Some folds validation presents even negative determination coefficients (what in this context means that even a horizontal line fits data better).

1st-fold	2nd-fold	3th-fold	4th-fold	5th-fold	6th-fold	7th-fodl	8th-fold	9th-fold	10th-fold
0.17	-3.41	-0.04	-0.01	0.51	-0.16	0.57	-0.34	-1.11	-0.23

Table 1. Results of 10-fold cross validation

2.2.3.2. Categorical and binary variables

- Binary variables (her2_positive, er_positive, pr_positive and neoadjuvant). The common denominator of these four cases is that there are variables with relatively few missing values to almost anyone. A replacement with the most prevalent value strategy has been adopted.

  For the case of the neoadjuvant variable, a checking task has been performed in all instances, in order to detect cases in which neoadjuvant=0 but there is some data recorded about tumor classification after neoadjuvant treatment. For those cases, neoadjuvant value has changed to neoadjuvant=1.

- Binary variables (invasive). What this variable has is that all filled values are 1's, it is a completely homogeneous variable if we don't take into account the missing values. In that context, three decisions could be taken about what to do: assume that missing values are 0s, delete instances with missing values or drop that variable.

  The homogeneity of that variable suggests that first assumption could be true, but it could be better if access to data gatherers was possible and that assumption could be validated, in order to avoid introducing some contamination to data. The second option does not seem to be a good idea, because a portion of cases would be lost just for getting a completely homogeneous variable, useless for exploratory analysis or forecasting. So, the selected approach has been the deletion of that column.

- Categorical variables (hist type and grade). In hist_type and grade, the number of missing values is much higher than in the previous three cases. We have imputed missing values for both variables to be 'unknown'. As we are preprocessing data for exploratory and descriptive purposes, this is a feasible strategy that can be adopted due to the circumstances. If this dataset were going to be used for classification/regression tasks, better approaches could be taken to fill those values (even dropping nan rows).

- Categorical variables (t,n,m). As there are few missing values, an analog strategy to the one chosen for binary variables with few missing values has been employed.

- Categorical variables (t_after_neoadj, n_after_neoadj, m_after_neoadj). Filling missing values proceeds only for cases in which neoadjuvant=1. As very little data is present about those variables, proceeding missing values have been filled with 'unknown' tags, following the same strategy and assumptions that the followed ones for hist type and grade.

2.2.4. Final adjustements.

Once completed the preprocessing, columns have been reordered in a fancier, more comfortable and logical way. Moreover, 'index' variable has been reset, ensuring it to be a one-step auto incremental variable for the final dataset.

3. Exploratory Analysis

Visualization is the most important way to explore data. In this subject, we used two libraries (Matplotlib, seaborn) to visualize and analyze refined data. In addition, the analysis was conducted by dividing it into univariate and bivariate analysis.

3.1. Univariate analysis

3.1.1. Histogram

In Univariate analysis, histograms were generated with age-related, pregnancy-related attributes, and KI67 respectively, and they expressed for frequency. Age-related variables were 'age_at_diagnosis', 'age_at_death', 'age_at_relapse', 'menarche_age', and 'menopause_age', respectively, and pregnancy-related variables were set to 'pregnancy', 'abort', 'birth', and 'cesarean'.

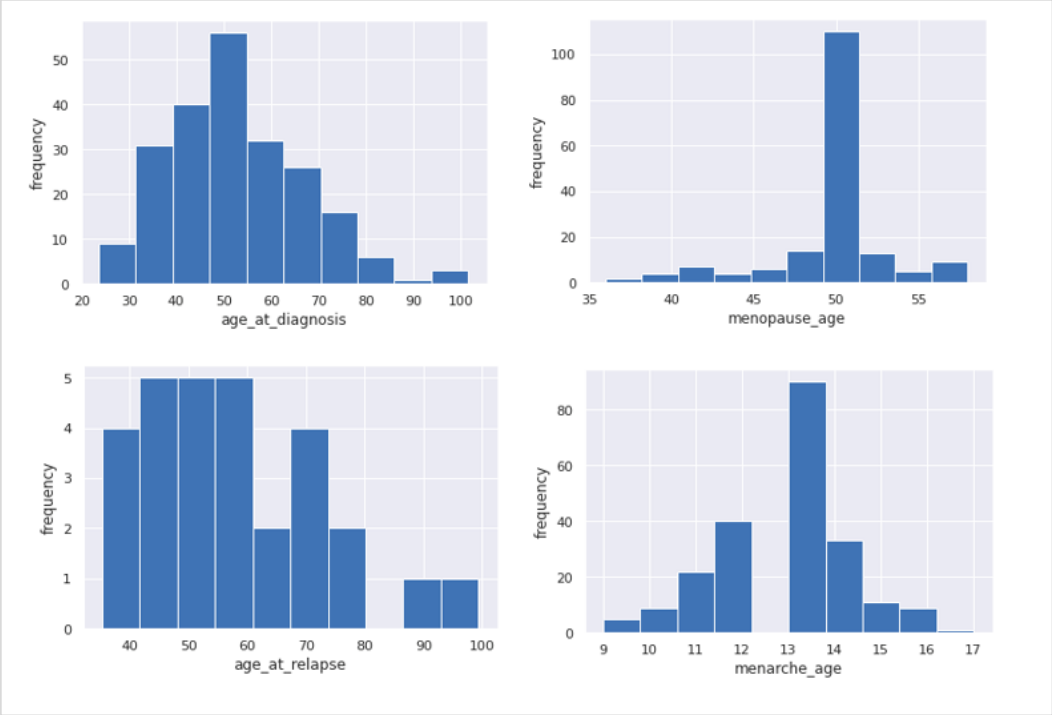

FIGURE 1. HISTOGRAM BY AGE-RELATED VARIABLES

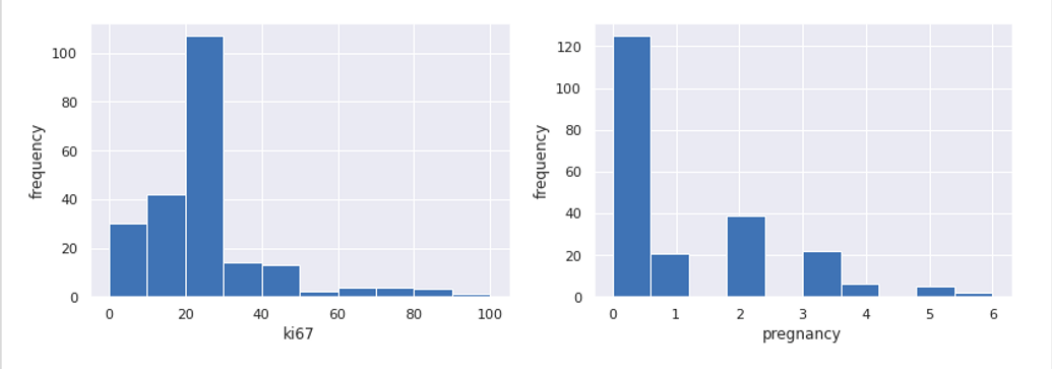

FIGURE 2. HISTOGRAM BY KI67 AND PREGNANCY

3.1.2. Correlation heatmap and Dendrogram

The correlation between each variable was expressed using a heatmap indicating the correlation between variables of the numerical type.

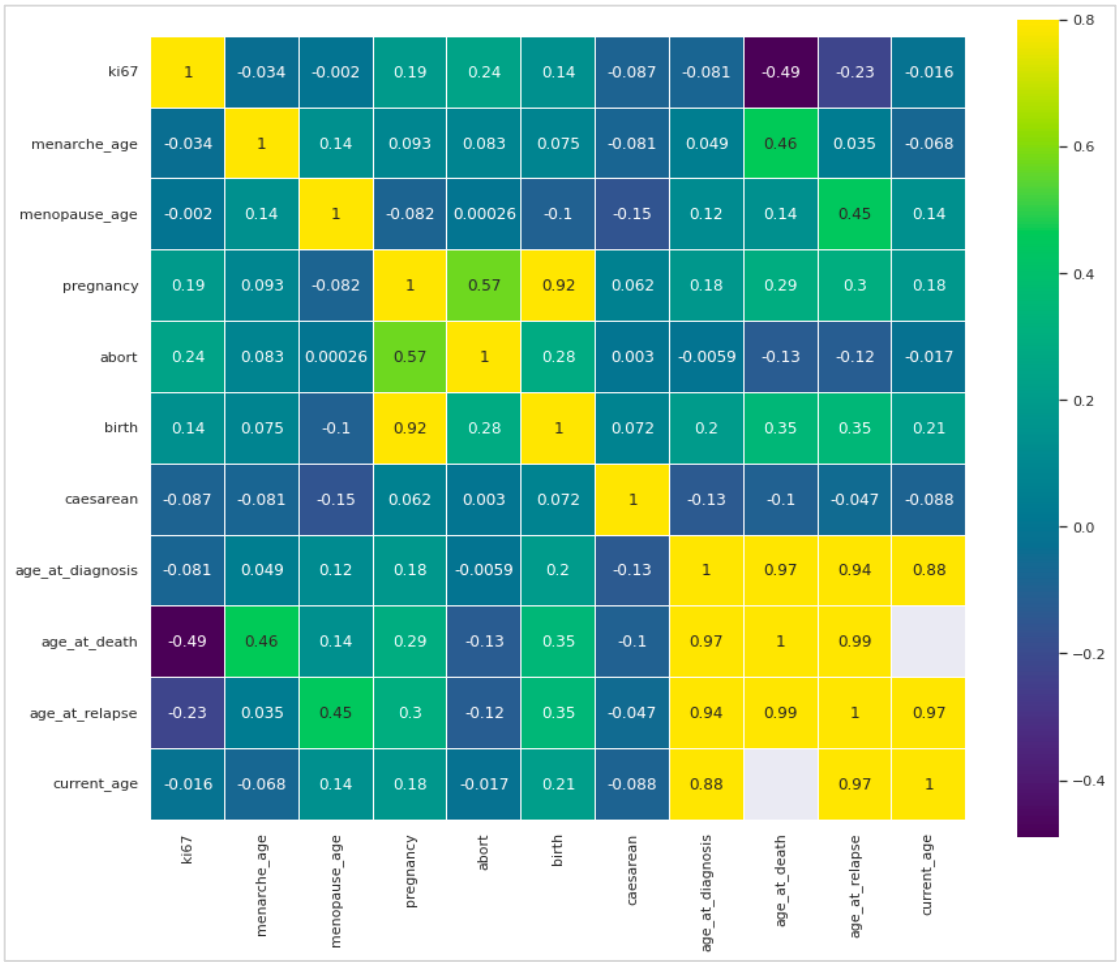

FIGURE 3. CORRELATION HEATMAP BETWEEN NUMERIC TYPE OF VARIABLES

The dendrogram is a tree structure that visualizes the correlation between variables. It allows more fully correlated variable completion, revealing trends deeper than the pairwise ones visible in the correlation heatmap.

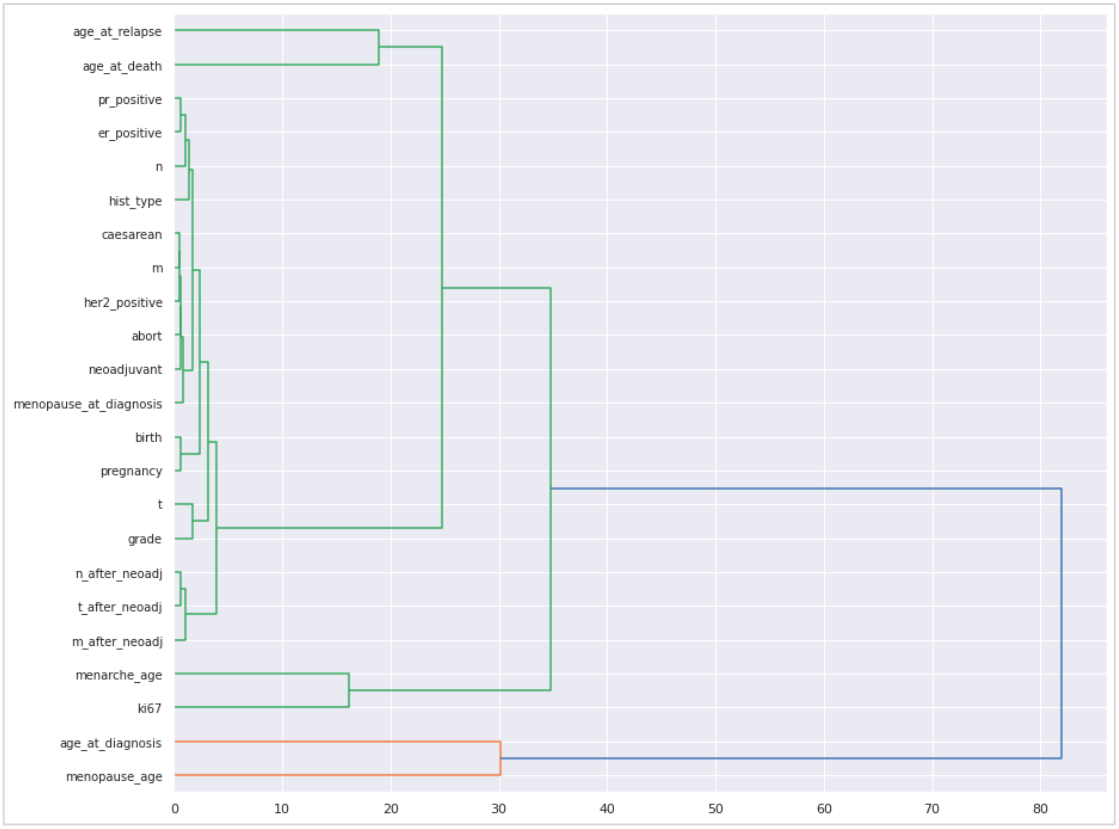

FIGURE 4. DENDOGRAM BETWEEN VARIABLES

3.1.3. Binary variables

To find out the proportion of the variables associated with the tumor divided into binary types, we expressed using a pie chart for those variables such as neoadjuvant, grade, er_positive, pr_positive, her2_positive and hist_type.

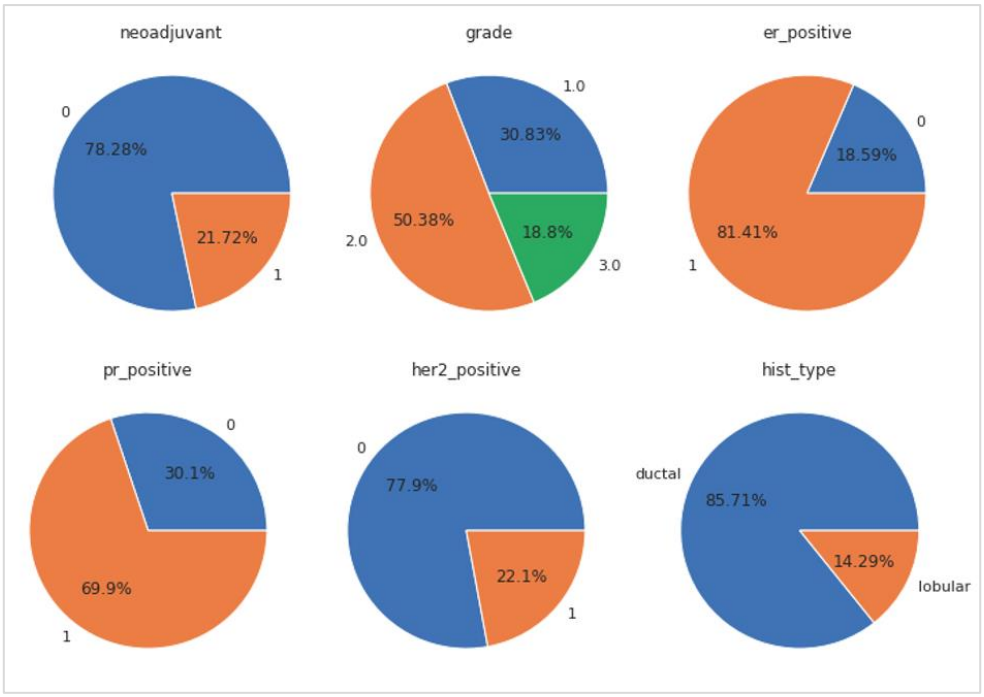

FIGURE 5. PIE CHART FOR BINARY VARIABLES

3.1.4. Bar chart (t,n,m classification)

T,n,m classification values of the tumor of the patient at diagnosis were generated by bar chart and those were expressed for frequency.

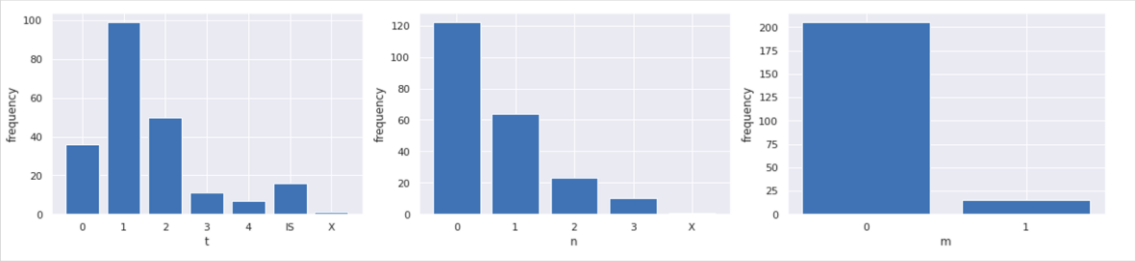

FIGURE 6. BAR CHART FOR T,N, M CLASSIFICATION

3.2. Bivariate analysis

In bivariate analysis, the association between each variable was investigated using the countplot of the seaborn library. First, 'age_at_diagnosis' was divided into five groups to express the count of grade, pregnancy, er_positive, pr_positive, her2_positive, and neoadjuvant, respectively. In addition, neoadjuvant compared the relationship with the recurrent age. Looking at the graph between age_at_diagnosis and pregnancy, It was found that patients with no pregnancy experience in all ages were significantly more confirmed to have breast cancer. For that reason, it was decided to graph the relationship between pregnancy and grade variables.

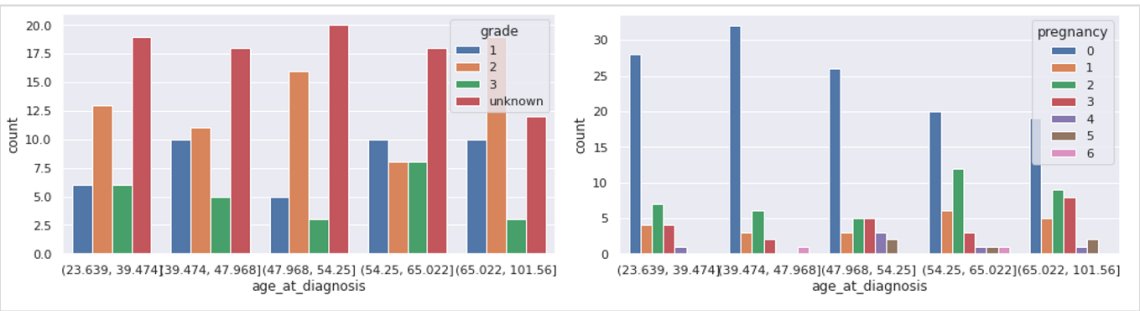

FIGURE 7. GRADE AND PREGNANCY BY AGE AT DIAGNOSIS

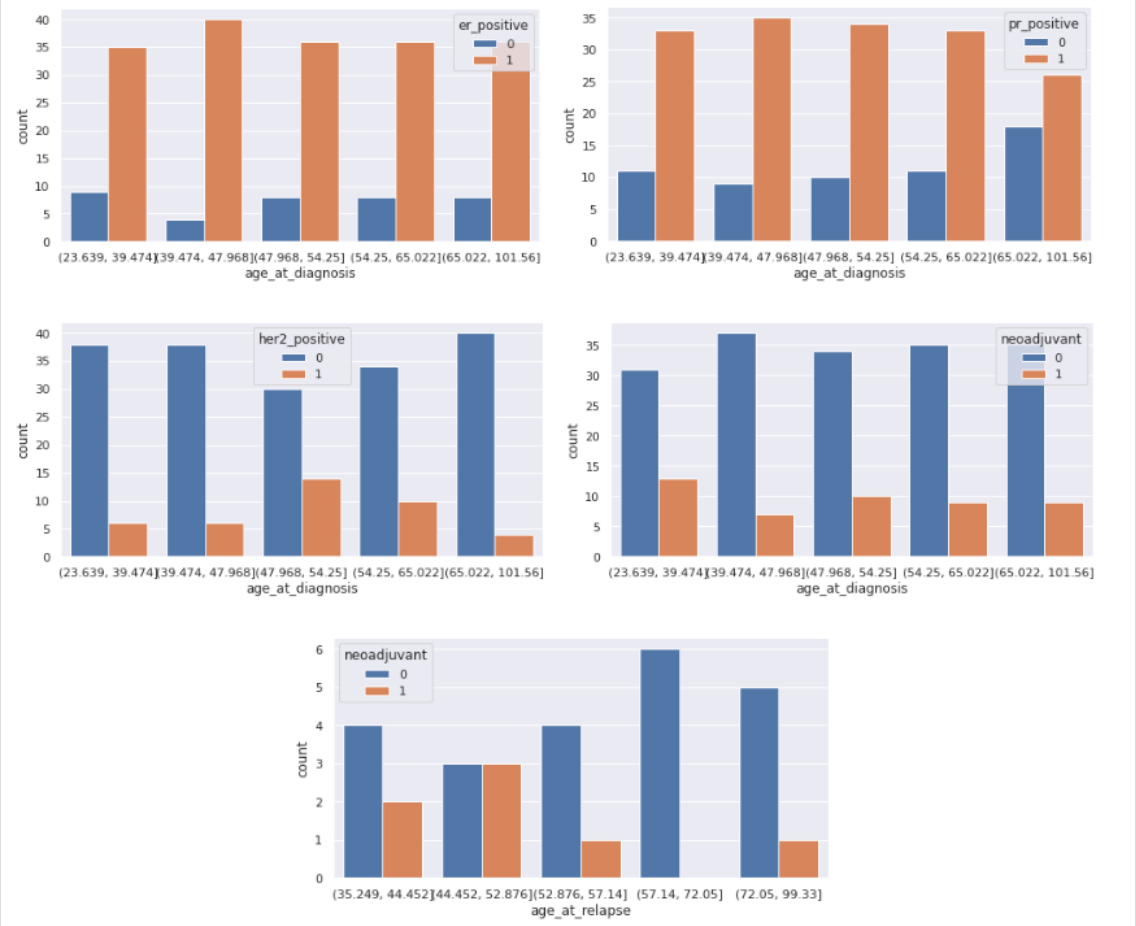

FIGURE 8. ER_POSITIVE, PR_POSITIVE, HER2_POSITIVE, NEOADJUVANT BY AGE_AT_DIAGNISES, NEOADJUVANT BY AGE_AT_RELAPSE

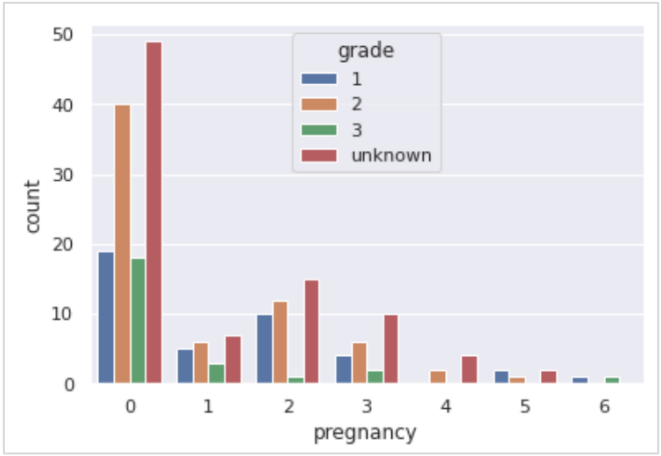

FIGURE 9. GRADE BY PREGNANCY

Subsequently, the hist_type variable was expressed by comparing it with the pregnancy, birth, abort, and grade variables.

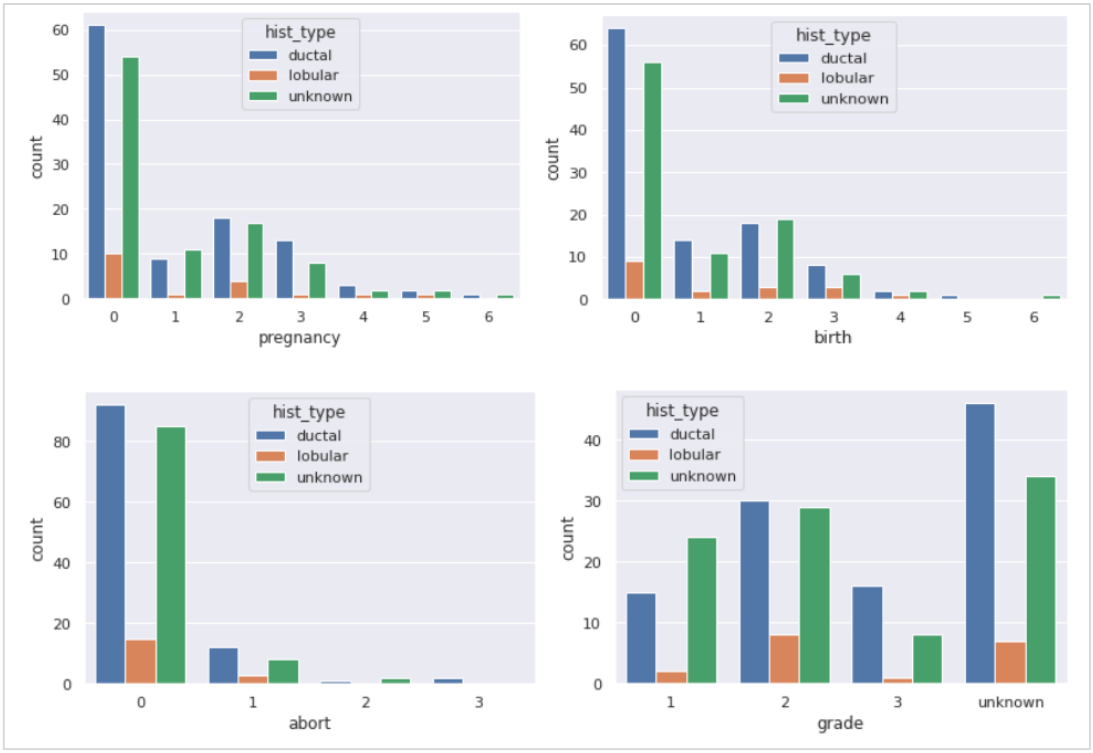

FIGURE 10. HIST TYPE BY PREGNANCY, BIRTH, ABORT AND GRADE

Thereafter, it was generated using countplot in order that ki67 by er_positive, neoadjuvant and relapse_recorded and hist_type by hist_type by pregnancy, birth, abort, and grade.

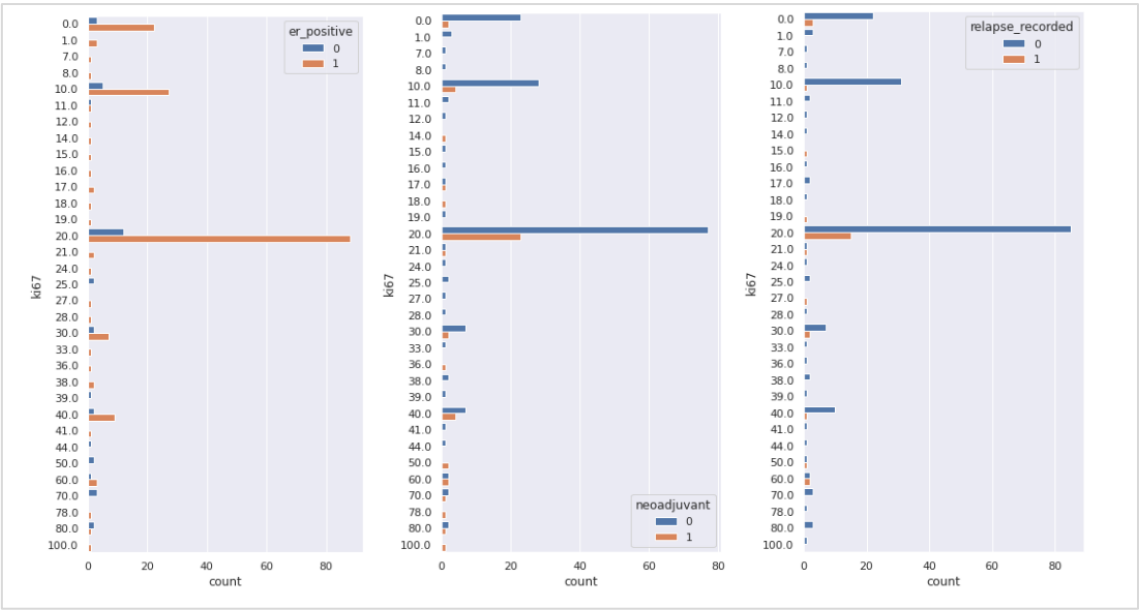

FIGURE 11. ER POSITIVE, NEOADJUVANT AND RELAPSE RECORDED BY KI67

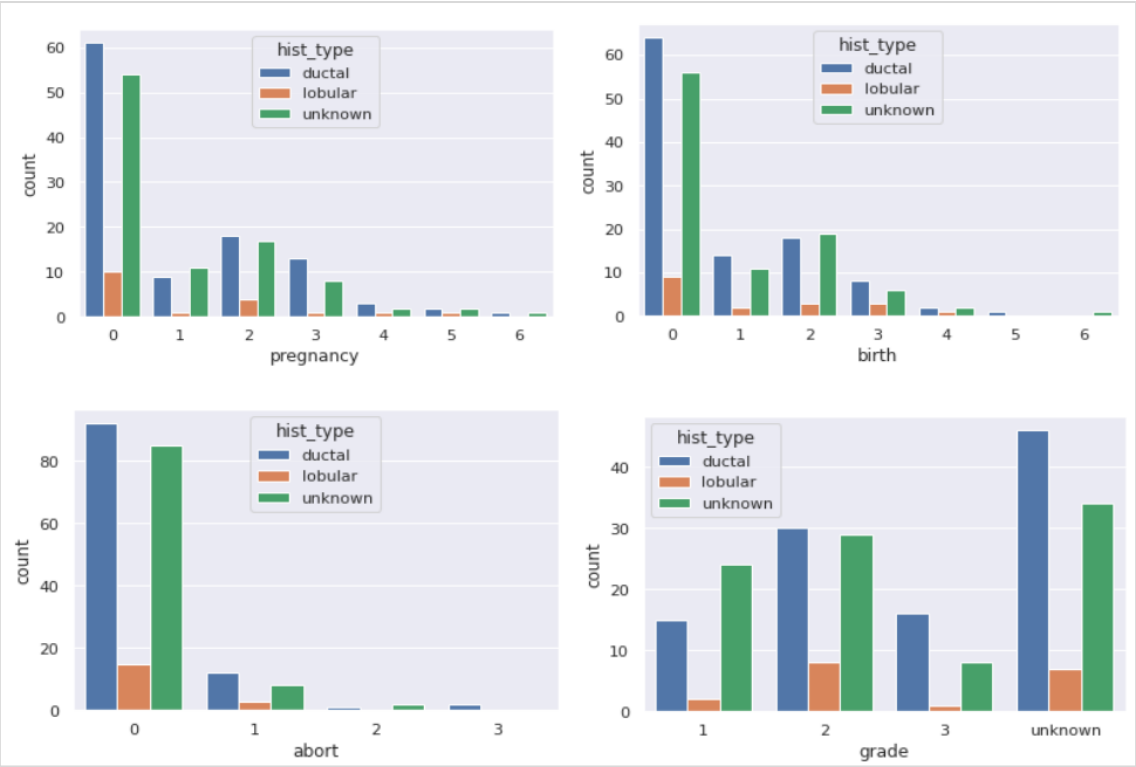

FIGURE 12. HIST TYPE BY PREGNANCY, BIRTH, ABORT AND GRADE

4. Results

In the beginning, looking at the histogram Figure 1 of univariate analysis, the age at which the diagnosis is distributed in various ways from 30s to 60s, but in particular, it can be seen that it increased noticeably between the ages of 45 and 65. However, menopause and menarcheage were found to be between the early 50s and 13-14 years of age, respectively, which is difficult to conclude that there is the same correlation as the average premenstrual and menopause of women.

Interestingly, noteworthy results were found in the Ki67 and pregnancy variables Figure 2, the percentage for proportion indicated by the column ki67 is distributed mainly between 0% and 30%. The mean is 21%, with a standard deviation of 19 which indicates that the distribution is not concentrated around the mean value. We can notice that more than 75% of the values are under a ki67 of 30%.

In addition, in the Pregnancy variable, it was found that about 60% of patients had no pregnancy experience. In dendrogram and heatmap Figure 3, Figure 4, which express the correlation between variables, Pregnancy abort and birth are related, also there was a strong positive correlation between the age of the recurred patient and the age of death(0.98), and a normal negative correlation between ki67 and the age of death was found(-0.49).

In the pie chart that analyzed the binary variable Figure 5, less than 1/4 of patient received neoadjuvant chemotherapy and half of the patients are in grade 2 and 1/3 in grade 1 also, 80% of the tumor has estrogen receptors, 70% of the tumor has progesterone receptors respectively. Finally, it was found that nearly 80% of the tumor doesn't have a HER2 protein overexpression. Lastly, in the bar chart Figure 6 that analyzed t,n, and m classification in univariate analysis, In the T classification, it was found that 1 is the most frequent value, then it is 2 and in the N classification, 0 and 1 values are the most frequent. In M classification, the value is mainly 0.

In the next section bivariate analysis, as mentioned above, most of the analysis was generated through the count plot of the seaborn library. First of all, the most notable outcome is between pregnancy and age_at_diagnosis, where people with no pregnancy experience were overwhelmingly confirmed at all ages diagnosed Figure 7. Also, A large majority of grade 2 for the oldest patients and the youngest quartile of ages contains more grade 3 than the other quartiles. We can observe more grade 1 for the ages 40-49 and 55-65. Looking at the figure Figure 8, We were able to find that between age and positive response of estrogen, progesterone, and negative test of HER2 protein were not correlated. Likewise, there was no noticeable correlation between neoadjuvant chemotherapy and age. However, among the patients diagnosed with recurrence, young patients, i.e., those under the age of 50, were found to have neoadjuvant chemotherapy.

The biggest characteristic common in univariate and bivariate variables can be seen that diagnosis occurs more in patients with no pregnancy or birth experience than in the group that do not. Figure 9,Figure 10,Figure 12 When the analysis is conducted based on the hist_type, it can be seen that the ductal type is clearly more than the lobular type in common. However, in the history_type, the classes classified as unknown had difficulty in making a vast and accurate analysis. Accurate classification tended to be difficult because it was a class that could not be classified as a median or an average value.

Finally, looking at the graph classified based on ki 67 Figure 11,The rate of proliferation tended to be found in 10 multiples, especially when it was 20%, which can be estimated to be associated with estrogen receptors. Likewise, neoadjuvant showed a graph with the same trend as estrogen receptors. It was also confirmed through the graph that there was no recurrence around 20%.