EDA (Exploratory Data Analysis) Report

➔ Code plan

1. First part - EDA
    a) Information
    b) Correlation
    c) Difference
    d) Outliers
    e) Missing values
2. Second part - Data cleaning
    a) Outliers
    b) Missing values
    c) the correlation is bigger than 0.7
    d) download file
3. Third part - EDA circular
4. Fourth part - Adding data
    a) download file
5. Fifth part - Feature selection

➔ Introduction

The aim of this study is to identify factors that may influence the likelihood of death during 5 years of follow-up, as well as to analyze their relationships with various medical and social characteristics of patients.

➔ Data overview

age	גיל בשנים
sex	מגדר (1=ז/2=נ)
marital_status	מצב משפחתי (M=נשוי/ U=לא נשוי)
ses	מצב סוציאו-כללי (1=נמוך / 2=בינוני/ 3=גבוע)
residence_cd	אזור מגורים (ראה הבא)
residence	אזור מגורים (urban=עיר/ rural=כפר)
weight	משקל (בקילוגרם)
height	גובה (בסנטימטר)
BMI	Body Mass Index – מדד מסה של הגוף
bp_sys	לחץ דם סיסטולי
bp_dias	לחץ דם דיאסטולי
bp_cat	קבוצת לחץ דם
smoking	עישון (ראה הבא)
smoking_status	מצב עישון (1=לא מעשון / 2=מעשן לשעבר / 3=מעשן)
HbA1c	המוגלובין מסוכרר (מדד של סוכר בדם ממוצע בשלושה החודשים האחרונים
glucose	סוכר בדם
creatinin	קראטינין – מדד תפקוד כליות
albumin	חלבון בדם
alb24h	איסוף חלבון בשתן במשך 24 שעות
ACR	יחס בין אלבומין לקראטיניו
cholesterol_total	סה"כ שומנים בדם

LDL	שומן רע בדם
HDL	שומן טוב בדם
triglycerides	שומן מורקב הדם
TSH	הורמון הקשור לפעילות בלוטת התריס
gravity_u	בדיקת משקל סגולי של השתן (Specific Gravity)
nitrites_u	ניטריטים בשתן
leuko_u	כדוריות לבנות בשתן
proteinuria	חלבון בשתן
WBC	כדוריות לבנות בדם
RBC	כדוריות אדומות בדם
platelets	תעשיות בדם (חשובים בקרישת הדם)
MCV	נפח קורפוסקולרית ממוצע בדם
MPV	מדידת נפח כדוריות אדומות דחוסות
Charlson	מדד סיכון מחושב
framingham_cvd	מדד סיכון מחושב
antidiabetics	תרופות לטיפול בסוכרת
ERD	מחלת כלייתית סופנית
CVD	מחלת לב
HTN	יתר לחץ דם
cancer	סרטן

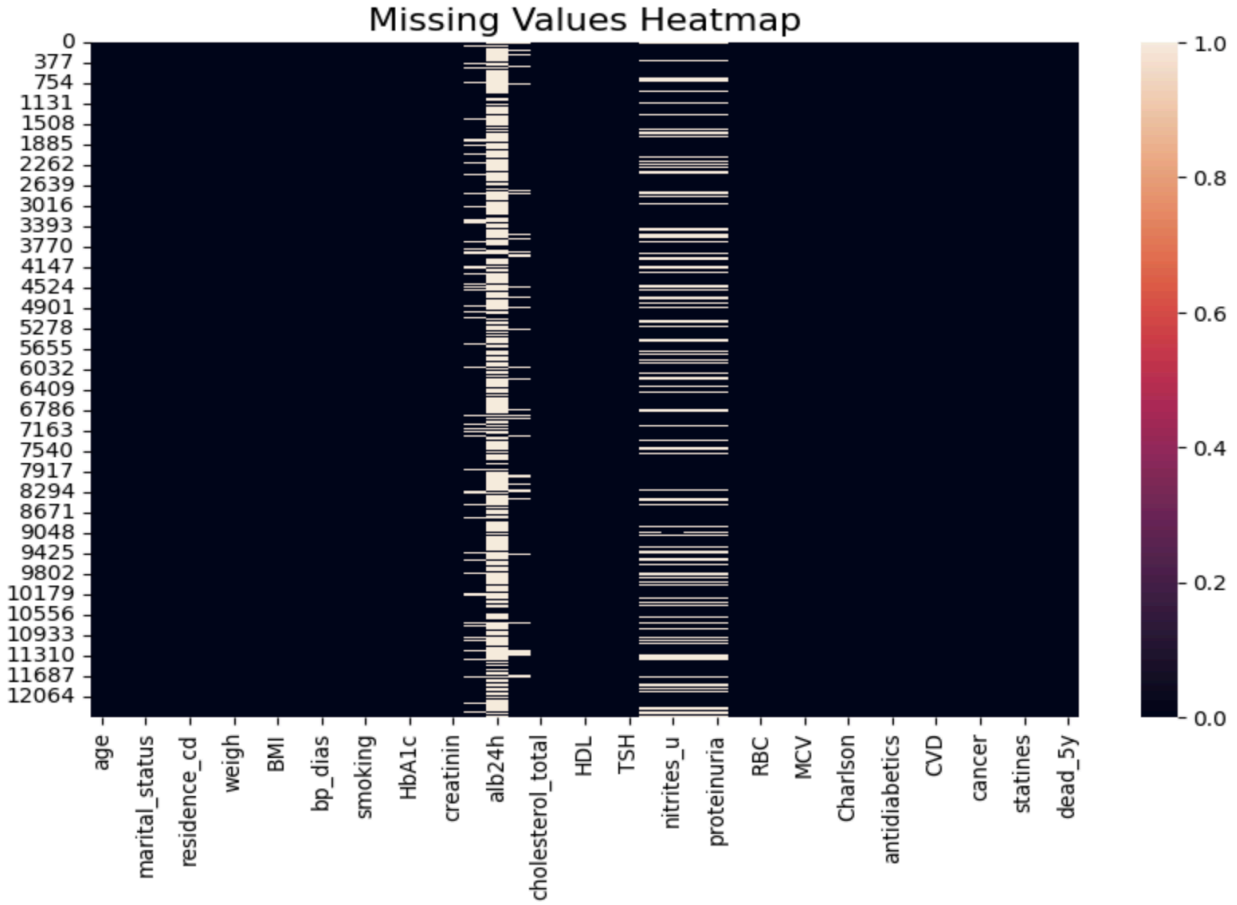
cardiovascular_meds	תרופות לטיפול במחלות לב
statines	תרופות להורדת השומנים בדם
immigrant	עולה חדש
dead_5y	*** מוות לאחר 5 שנים של מעקב (משתנה מטרה)

The data represent 12,439 patients with 45 variables, 26 of which are numeric and 19 are categorical.

Initial data quality check:

- No duplicates found

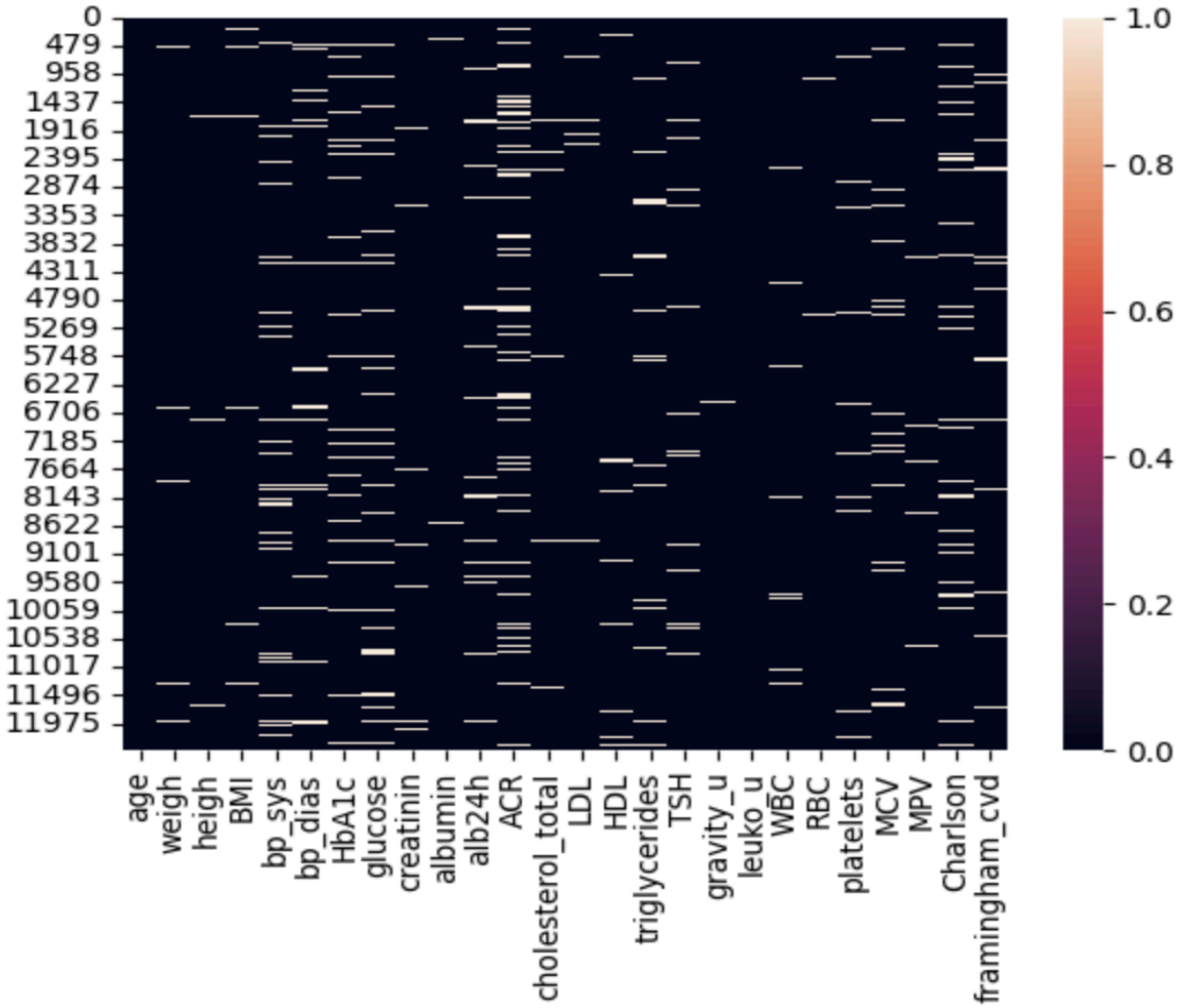
- Missing values detected



- Outliers detected

  *age and leuko_u is not considered as variables with extreme values

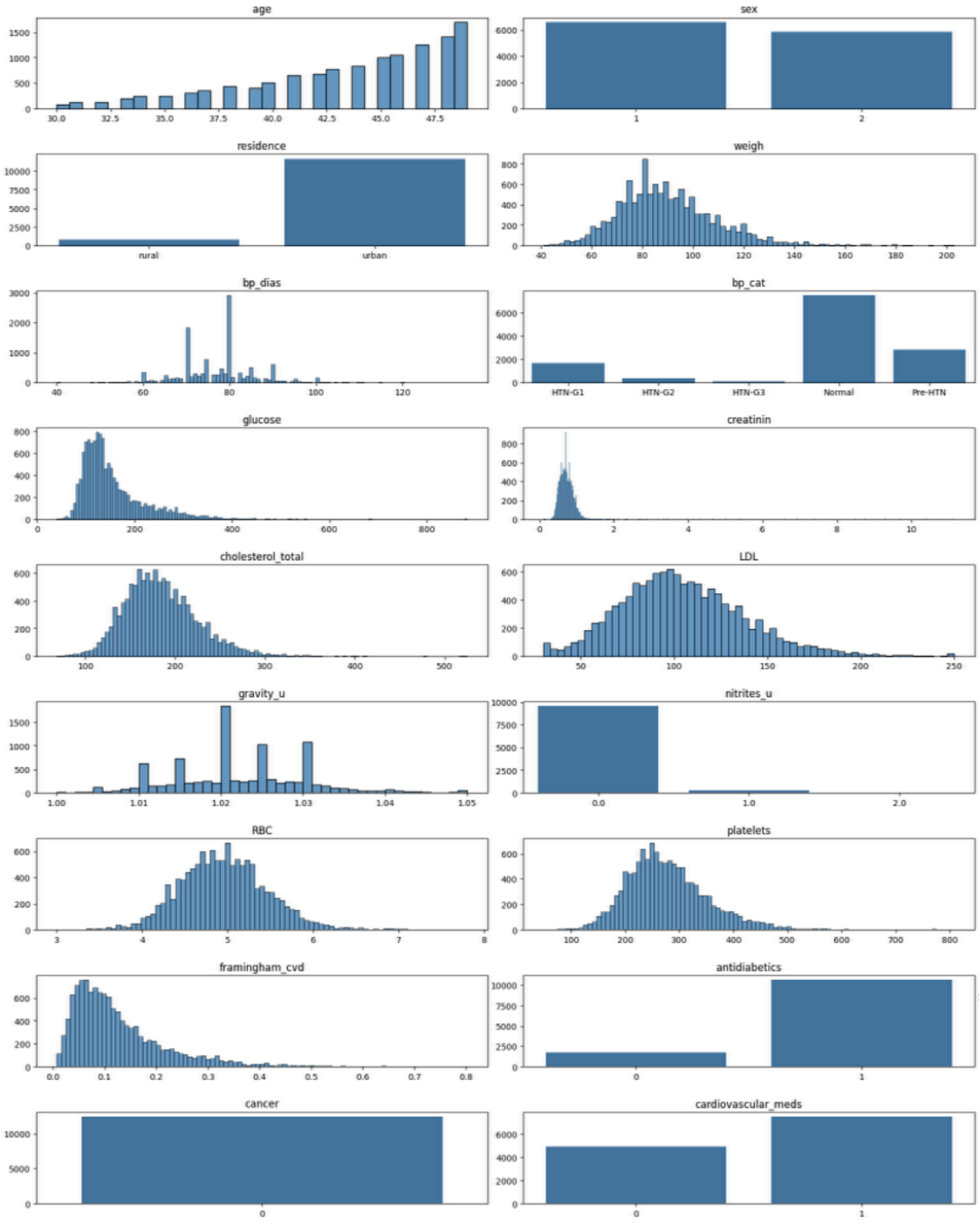
  age - the study involves people of all ages
  leuko_u - normal value for Leukocytes is 0-5 , and most of the outliers are 0.



Graphs are presented that show the amount of data for each variable :

- Histogram and KDE for numeric data
- Barplot for categorical data

For example 18 random variables:



➔ Relationships between variables

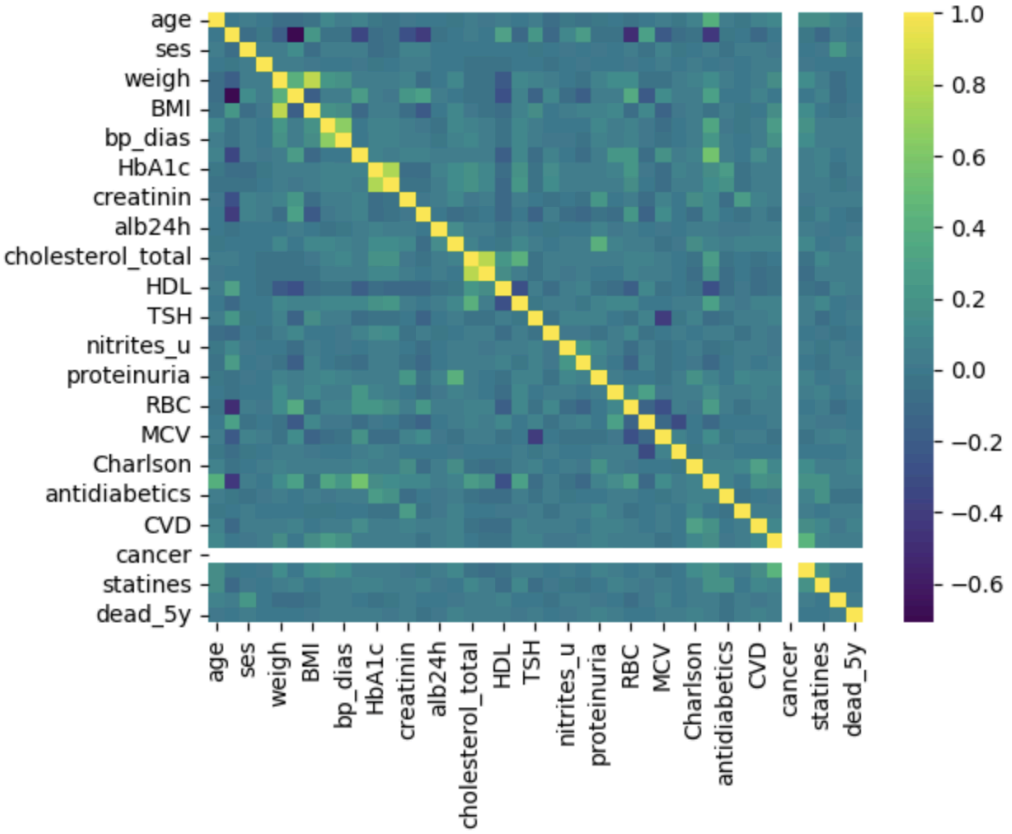
1. Numeric variables

- To calculate the correlation between numeric variables is used Pearson Correlation Coefficient. This approach allows us to identify which variables exhibit strong relationships with one another and to detect those that convey nearly identical information, particularly when the correlation coefficient exceeds 0.7.

correlation coefficient > 0.7 :

weigh and BMI: Corr = 0.80, p-value = 0.0000
HbA1c and glucose: Corr = 0.74, p-value = 0.0000
cholesterol_total and LDL: Corr = 0.82, p-value = 0.0000

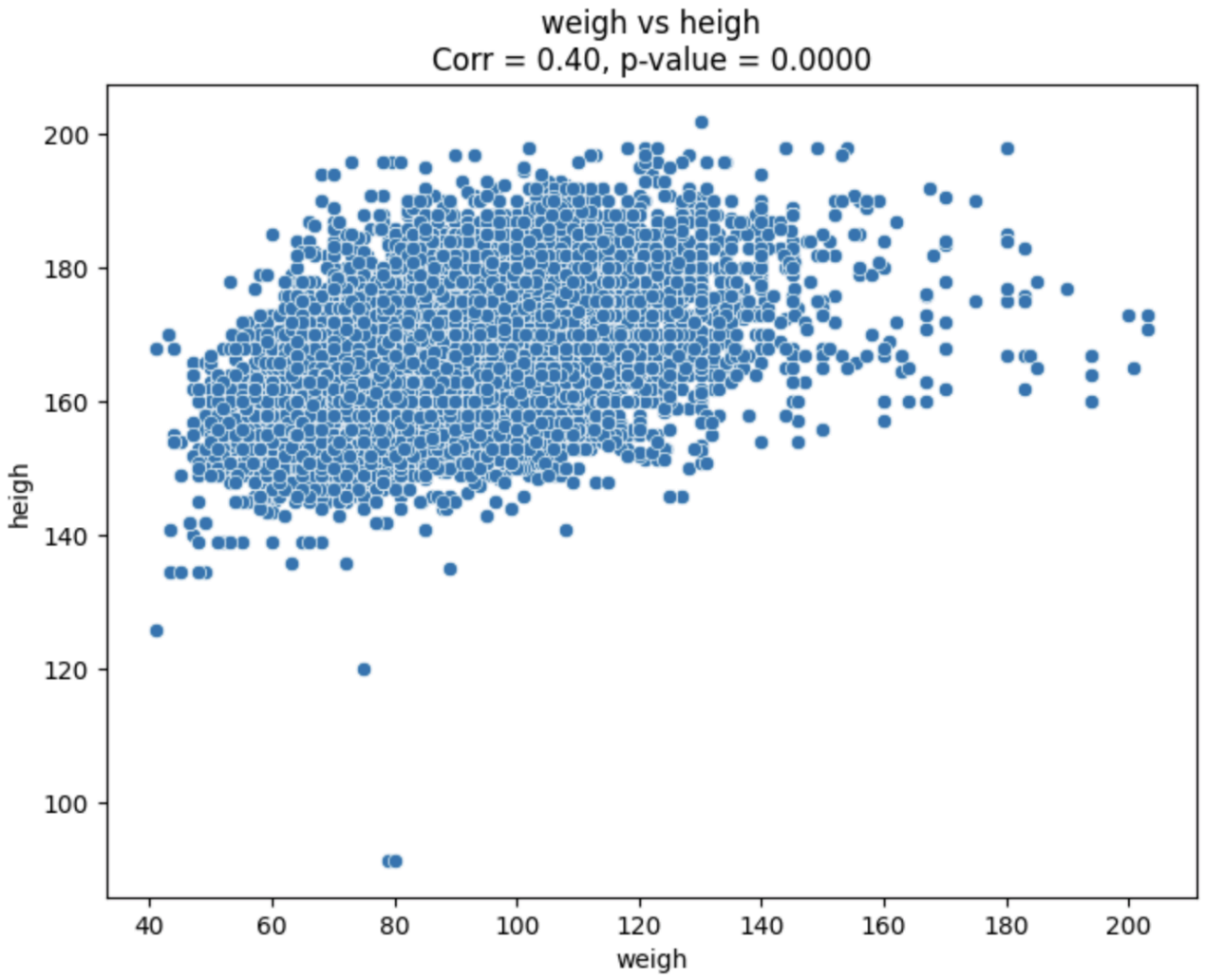
In the next part one of the variables in the pair will be removed.



*the variable cancer is painted white because the entire variable consists of 0

- Also shown are scatter plots for each pair of numerical variables with a p-value < 0.05. When the p-value is less than 0.05, it indicates a statistically significant relationship between the variables. In other words, the probability that the observed relationship occurred by chance is less than 5%. This allows us to confidently state that a genuine relationship exists between the variables, rather than it being a mere coincidence. In the context of data analysis, this suggests that one variable may influence the other, or that there is a meaningful pattern between them, which is important for further analysis and conclusion.

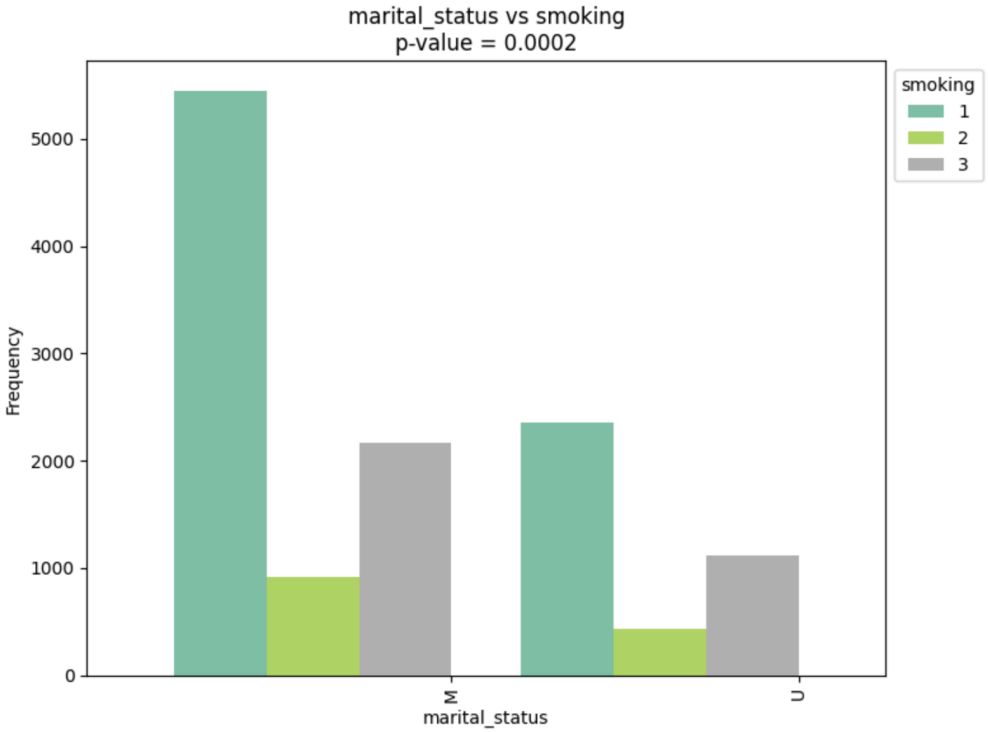
For example:



2. Categorical variables

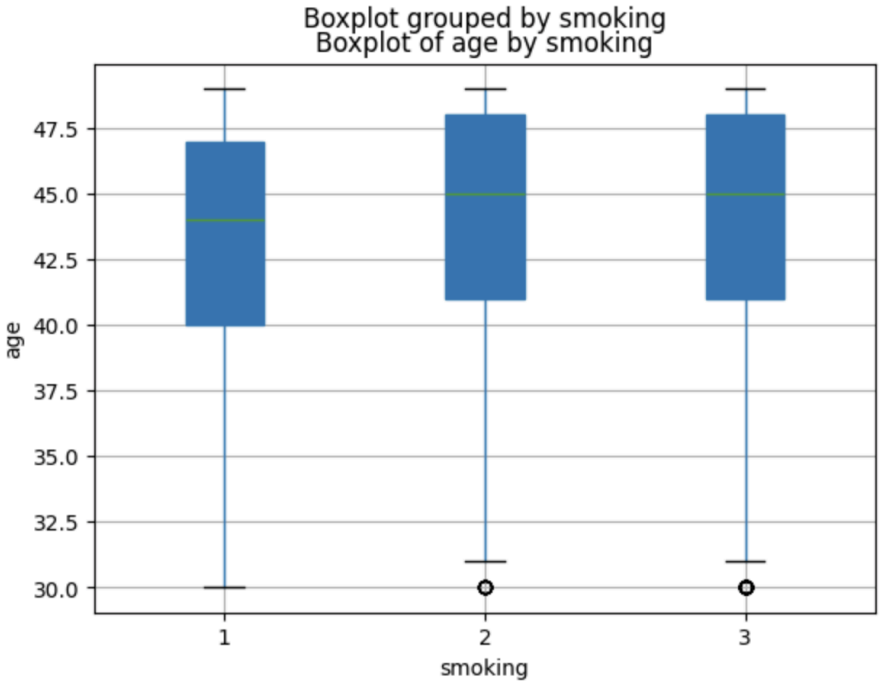
- Graphs were constructed to analyze the relationship between two categorical variables using the Chi-Square Test. This test evaluates whether there is a significant association between the variables. If the p-value < 0.05, the differences are considered statistically significant, indicating that the observed relationship is unlikely to have occurred by chance. In such cases, the null hypothesis, which assumes no association between the variables, can be rejected.

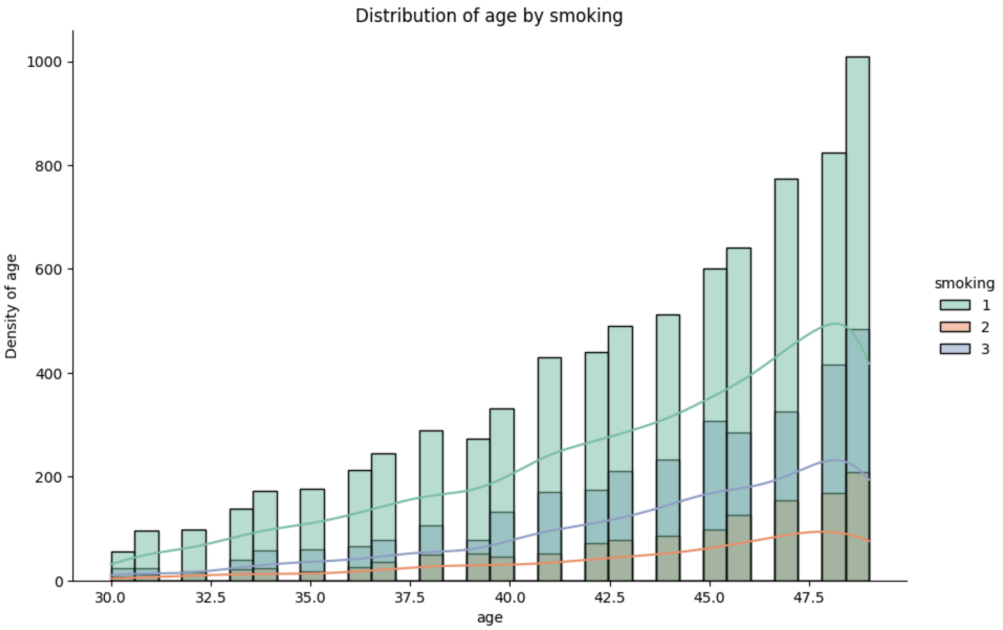
For example:



*The graph says that, for example: in marital status - M patients with smoking status - 1, more than 5000.
* This graph helps to understand the relationship and “proportion” between categorical data.

- Graphs were constructed to analyze the relationship between categorical and numerical variables using the ANOVA (Analysis of Variance) test. For cases where the p-value < 0.05, visualizations were created, including boxplots and distribution plots.





➔ Dealing with outliers and missing data

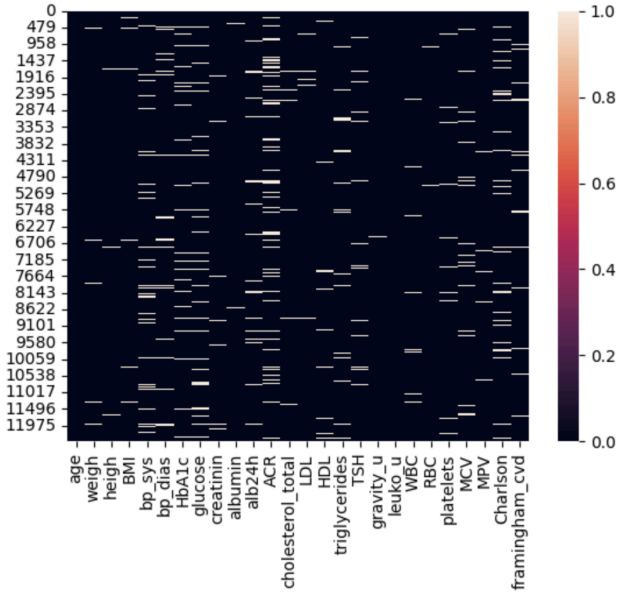
- Outliers

The principle of outlier analysis used allows us to assess their impact on the correlation between variables and the distribution of data. Removal of outliers is tested using two key metrics:

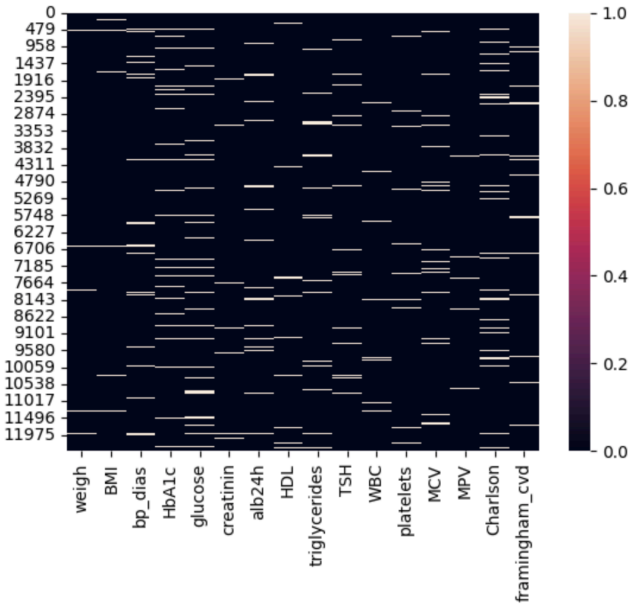
Correlation: If the p-value is less than 0.05 after removing outliers, this indicates significant changes in the relationship between variables. This is important because outliers can distort the strength and direction of the correlation, leading to incorrect conclusions.
Distribution: The Kolmogorov-Smirnov test helps determine whether the distribution of data changes after removing outliers. If the p-value is less than 0.05, then outliers significantly affect the shape of the variable distribution.

The XOR operator compares changes in correlation and distribution: if one has changed but the other has not, the variable is marked for deletion. If no change has been recorded or both aspects have changed, the variable remains in the data.

before cleaning:



after cleaning:



* after cleaning heatmap shows variables in which outliers were not cleared

- # Missing values

- First, remove variables that are more than 70% missing values

Variable: alb24h - 75%

- Then , 40% - 70% missing values. For numeric variables, the missing values are filled by splitting into 4 categories, and for categorical variables, a category 0 is added for missing values. After this, the missing values are replaced with 0, and the data type is updated to categorical.

Variable: there are no such variables

- Last one is up to 40% missing values. Checking if the missing values are MCAR (Missing Completely at Random) or MAR (Missing at Random) is crucial for understanding their nature. Significant p-values ($p < 0.05$) indicate that the missing data may depend on other variables, suggesting a pattern that should be considered in the analysis. Based on these findings, numerical variables are imputed with the mean or median, while categorical variables are filled with the most frequent value or a new class, depending on the context, to minimize data distortion.

Variables: ACR, gravity_u, nitrites_u, proteinuria, leuko_u, albumin, bp_sys, Charlson, glucose, triglycerides, framingham_cvd, TSH, MCV, HbA1c, bp_dias, antidiabetics, ERD, CVD, HTN, cancer, cardiovascular_meds, statines, immigrant, dead_5y

- we also remove one of the variables with a correlation greater than 0.7 for numerical values and p-value == 0 for categorical ones.

columns to drop: weigh, LDL, HbA1c, residence, smoking_status

* left variables that have a higher correlation with the main variable, as well as those that are shown as numbers (categorical)

➔ Conclusions

In the last section, we identified the variables with a significant impact on the probability of dying within 5 years. The key predictors include age, marital status, socioeconomic status, BMI, blood pressure (systolic and diastolic), smoking status, glucose level, creatinine level, albumin level, cholesterol level, TSH, urine specific gravity test, urine nitrites, urine protein, white blood cells, mean corpuscular volume, compressed red blood cell level, estimated risk index, end-stage renal disease, heart problems, hypertension, and immigrant status.

Additionally, two new variables created during the analysis — medication_count and blood_pressure_risk — were also found to have significant impacts. The medication_count variable summarizes the medications taken by the patient, where its value is derived from the first letter of the medication name. And the second additional variable is the risk of blood pressure.
Among all these predictors, variables such as age showed the strongest associations with mortality.

These findings can be applied in clinical settings to prioritize high-risk patients for preventive care. Further research could focus on refining the derived variables or exploring the causal relationships between the identified predictors and mortality.