Kailin Liu

QBIO 490

14 October 2022

Midterm Project

Influence of Pre and Post-Menopausal Status on Breast Cancer Survivability

Introduction

Breast cancer is the most common cancer for females in the United States and is the second leading cause of female cancer death (The Cancer Genome Atlas Network, 2012). The American Cancer Society estimates that in 2022, approximately 287,850 new cases of invasive breast cancer will be diagnosed while 43,250 women will die from breast cancer. Breast cancers have a variety of risk factors including controllable factors, such as physical activity and diet habits, as well as non-controllable ones such as increased age, genetic mutations, and reproductive history (Surakasula et.al, 2014). This study focuses on the relationship between the uncontrollable factor of menopause status and breast cancer survivability, specifically whether pre-menopausal or post-menopausal women have better breast cancer survival. Furthermore, utilizing data from The Cancer Genome Atlas (a publicly available multi-omic data set organized by National Cancer Institute and National Human Genome Research Institute), we analyzed what genes are more or less mutated for each group, how such mutations influence treatment, and the impact of such therapies on the survivability of the patients. Using the R programming language in R Studio, we identified the relationship between age and menopause status, the relationship between menopause status and survival rates, GATA Binding Protein 3 (GATA3) as one of the key differentially expressed genes between pre and post-menopausal groups, and the efficacy of GATA3 linked treatments on patient survival.

Methods

Setting Up

R Studio and the R programming language were used to analyze data from The Cancer Genome Atlas (TCGA). First, from outside R Studio through Terminal, a folder titled midsemester_project_lastname was created under the Desktop folder qbio_490_firstname. Next, moving into R Studio with a new R Script titled midterm_firstname, an outputs folder under the previously created midsemester_project_lastname folder was created to store future data and figures. The working directory was set to this outputs folder. Necessary packages were then loaded into R Studio. BiocManager, TCGAbiolinks, SummarizedExperiment, and maftools were installed using BiocManager (as BiocManager is necessary to install and manage packages from the Bioconductor project) while the packages ggplot2, survival, and survminer were installed normally as they were already implemented in the base installation of R.

To analyze patient data (such as barcode, menopause status, vital status, age, etc.) TCGA was accessed (and stored under the variable clinicial_query) using the R package GDCquery, the accession code “TCGA-BRCA”, the data category “Clinical”, and the file type “xml”. GDCdownload was then used to download the queried data under clinical_query. Two data frames were created with clinical and drug information and titled “clinical” and “clinical.drug” respectively. Clinical information was accessed using the R package GDCprepare_clinic with the query set to “clinical_query” and clinical.info set to “patient”. To access drug information, the same R package GDCprepare_clinical was used with query set to “clinical_query” and clinical.info set to “drug”.

Editing Data Frames

Since this study focuses on pre-menopause and post-menopause groups, we preprocessed our clinical data to select only these two kinds of patients. We removed any patients who were labeled “Indeterminate (neither Pre or Postmenopausal)”, “Peri (6-12 months since last menstrual period)”, or not labeled at all, as these patients’ menopause statuses were ambiguous and thus did not have the requisite data to pass our selection criteria. Furthermore, as the status names for the pre and post-menopause groups were long and would affect the dimensions of future graphs, the menopause_status column was overwritten and renamed to “Pre” and “Post” respectively using an ifelse loop. The edited clinical data frame was then written and saved as a file to the outputs folder as “analysis_clinical_data.csv”.

In preparation for plotting basic boxplots and survival analysis plots, we subset the clinical data frame for only the columns we will be using. Using the data.frame function we subset the bcr_patient_barcode, days_to_death, days_to_last_followup, vital_status, age_at_initial_pathologic_diagnosis, and menopause_status columns to a new data frame titled “analysis”. For clarity we then retitled the column names of the analysis data frame to "patient_barcode", "days_to_death", "days_to_last_followup","vital_status", "age_at_initial_pathologic_diagnosis", "menopause_status” for the corresponding columns.

Plotting

First, we created a boxplot (to visualize the relationship between pre or post-menopause status and age) using the boxplot S3 method for class 'formula', the formula “analysis$age_at_initial_pathologic_diagnosis ~ analysis$menopause_status”, data “clinical”, the xlab “Menopause Status”, the ylab “Age”, the main “Age vs Menopause Status”, and the cex.axis set to 0.5. The plot was saved to the outputs folder as a jpeg file titled “menopause_age_boxplot.jpg” for future reference.

Next, we created a Kaplan-Meier (KM) plot to compare the survival of pre and post-menopause patients. Before plotting, we needed information on follow-up time and a status indicator. For the follow-up time, we stored info from the days_to_death column (or days_to_last_followup column as some patients never had a registered death event and are labeled NA) in a new “survival_time” column under the analysis data frame. Next, for the status indicator, we stored boolean variables of TRUE or FALSE (based on whether the patient had a death event or no recorded event/null in the vital_status column) in a new “death_event” column under the analysis data frame. To graph the Kaplan Meier plot, we initialized a survival object using the Surv function where time utilized the survival_time analysis column and event utilized the death_event analysis column. Next, a fit object was created to categorize the previously created survival object by the two strata “Pre” and “Post” (from the menopause_status column of analysis). Formatting information about fit, pval, ggtheme, and legend was stored under survplot_menopause using the ggsurvplot function. Finally, the KM plot titled KM_plot_menopause was created and saved as a jpg in the outputs folder.

Setting up for Gene Analysis

To further understand the mechanism involved in the different survival trends of pre and post-menopause patients, we accessed mutation information from The Cancer Genome Atlas (TCGA) using the R package GDCquery, the accession code “TCGA-BRCA”, the data category "Simple Nucleotide Variation", the access “open”, the data.type “Masked Somatic Mutation”, and the workflow.type "Aliquot Ensemble Somatic Variant Merging and Masking". The accessed information was stored under the variable maf_query which was then downloaded via GDC download. A Mutation Annotation Format (MAF) file was then created using the package GDCprepare on maf_query. To ensure that the MAF package can read the clinical file, we then renamed the bcr_patient_barcode column of the clinical data frame to Tumor_Sample_Barcode. We then saved and overwrote the old clinical data frame with this new one. Finally, the clinical data from the MAF file was read in using read.maf , the maf "maf", the clinicalData "clinical", and isTCGA set to TRUE", and stored under the name maf_object.

Creating Co-Oncoplots

To study the point mutations specific to pre and post-menopause patients, we subset the maf_object's clinical.data into two smaller mafs (one for Pre and one for Post) using masking and subsetMaf. We then plotted a co-oncoplot to see the percentages of point mutations for each gene of each type of patient side by side. The co-oncoplot was then saved as a jpg the output folder.

Studying Efficacy of Drugs

Based on the results in the co-oncoplot, we selected GATA3 as one of the differentially expressed genes. To utilize the clinical.drug data frame, we selected for patients who underwent Hormonal Therapy as patients with the GATA3 biomarker are typically prescribed hormonal therapies. We also removed any duplicates from clinical.drug data frame (as many patients are duplicated for reasons including differing length of treatment time or just duplicates of completely same data altogether). Furthermore, the same drug type was spelled differently (ex. Tamoxifen as tamoxifen or TAMOXIFEN). As plotting the survival plot would have case-sensitive strata, we then standardized the names of each drug to the same spelling and capitalized the first letter(ex. Tamoxifen, Letrozole, Arimidex, etc.). Furthermore, in order to utilize the previously created follow-up time and a status indicator created in the Plotting section, we masked the analysis data frame to only include patients with barcodes similar to the ones in the clinical.drug data frame. As the analysis data frame also had duplicates, we removed those as well so that the subsetted clinical.drug and analysis data frames had the same row dimensions and could thus be plotted. Then, following the same steps as in the Plotting section, we created a KM plot utilizing drug names as the strata.

Results

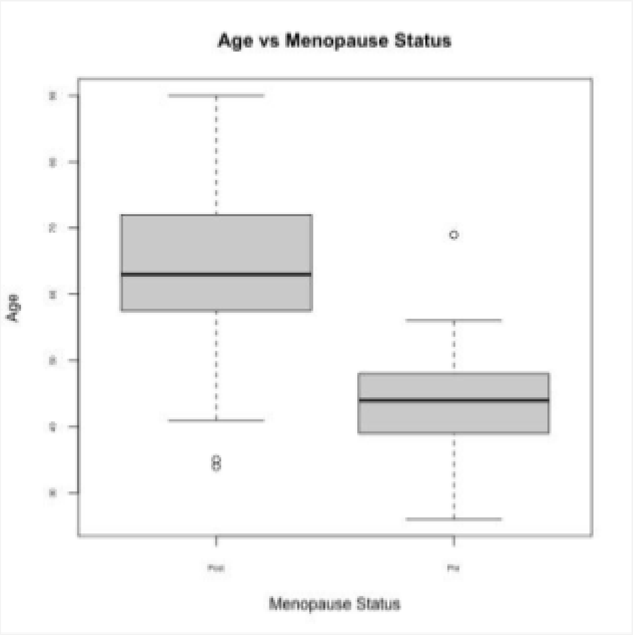

Figure 1: Boxplot showing the relationship between Post or Pre Menopause status and age. The median age for Post-Menopause is approximately 63 years old and the median age for Pre-Menopause is approximately 44 years old

Using data from the TCGA clinical samples, we first analyzed the relationship between age and menopause status to both visualize the age difference and confirm that the provided samples in the TCGA are within the normal range of pre and post-menopause ages. In Figure 1, the left boxplot of post-menopausal patient ages has approximately a lower range of 40 years old, a median of approximately 63, and an upper range of 90. Meanwhile, the right box plot of pre-menopausal patient ages has approximately a lower range of 25 years old, a median of 44, and an upper range of 55.

To analyze the difference in survivability between pre and post-menopausal patients, we created a Kaplan Meier plot with survival probaility as the y-axis and time in days as the y-axis.

Both pre and post-menopausal patients show an exponential decrease in survival probability with increased time. Pre-menopausal patients denoted by the blue line have a significantly shorter survival time (ending before the 4000 days mark) than post-menopausal patients (whose time in days extends up to the 6000 days mark).

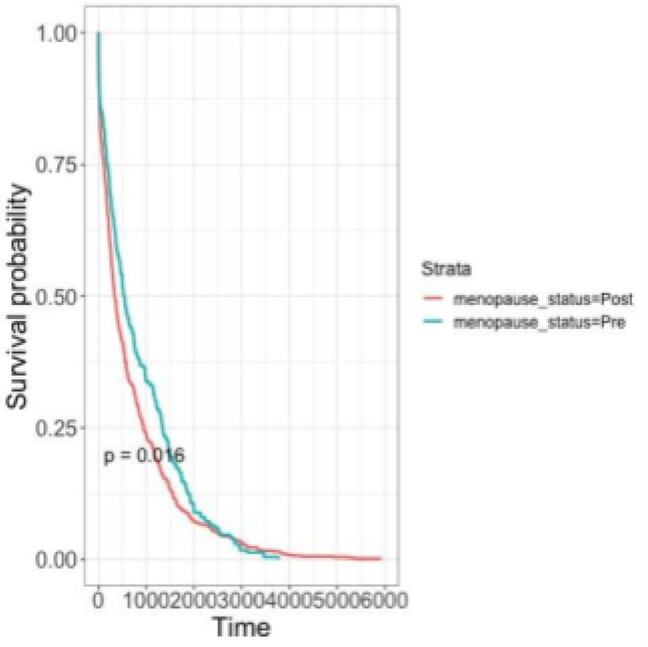

Figure 2: Kaplan-Meier plot showing the survival of Pre and Post-menopause patients. Pre-menopause patients have a shorter survival time than Post-menopause patients. P-value of 0.0084 denotes the statistical significance of the found results

To study the mechanism behind why pre-menopausal patients have a significantly shorter survival time than post-menopausal patients, a MAF co-oncoplot showing 6 genes (PIK3CA, TP53, TTN, CDH1, MUC16, and GATA3) and their point mutation percentages. The left oncoplot containing data from the pre-menopausal sample showed that PIK3CA was 32% mutated, TP53 was 33% mutated, TTN was 9% mutated, CDH1 was also 9% mutated, MUC16 was 10% mutated, and GATA3 was 21% mutated. The right oncoplot

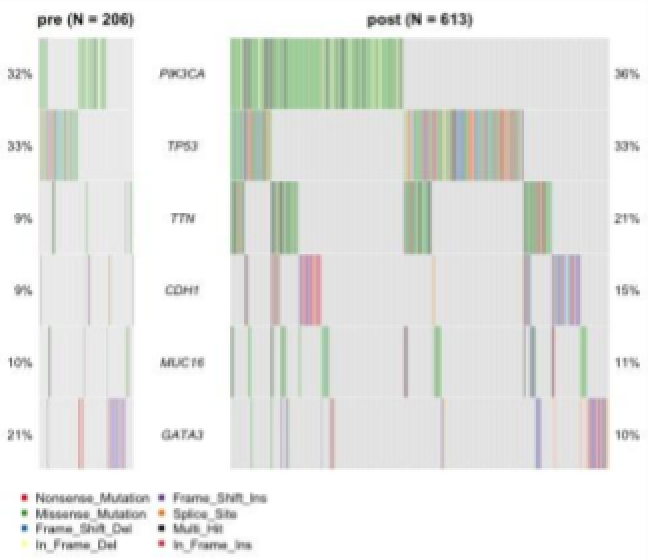

Figure 3: Co-oncoplot showing the point mutation percentages of PIK3CA, TP53, TTN, CDH1, MUC16, and GATA3 for pre and post-menopause patients side by side. Notably, TTN mutations are doubled in the post-menopause patient sample in comparison to the pre-menopause patients sample while GATA3 mutations are halved in the post-menopause patient sample.

containing data from the post-menopausal sample showed that PIK3CA was 36% mutated, TP53 was 33% mutated, TTN was 21% mutated, CDH1 was 15% mutated, MUC16 was 11% mutated, and GATA3 was 10% mutated. GATA3 was mutated twice as much in pre-menopausal patients than in post-menopausal patients.

To understand the therapeutic implications of the GATA3 mutation on pre and post-menopausal patients, we created a second Kaplan-Meier plot to visualize the survivability of patients with Hormone Therapy, a common therapy given to GATA3 mutated patients (Gustin et. al, 2017). For this survival plot, survival probability was the y-axis, and time in days was the x-axis. Seven different drug types correlated with hormone therapy were included. In order of shortest to longest survival time: Anastrole patients survived up to approximately 600 days, Fermara up to approx. 800-900 days, Aromasin up to approx. 1100 days, Arimidex up to approx. 1600 days, Letrozole up to approx. 2450 days, other/unspecified drugs up to 2950 days, and Tamoxifen past 3000 days.

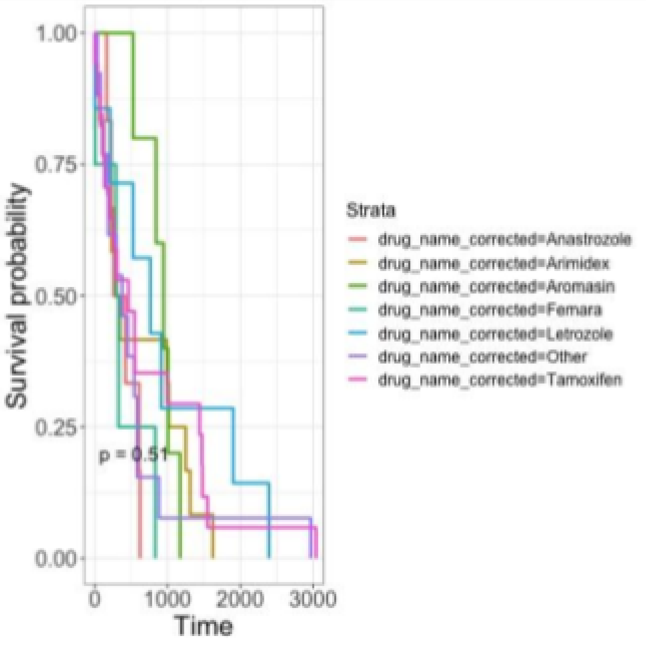

Figure 4: Kaplan-Meier plot showing the survival different drug types in patients treated with Hormonal Therapy. Anastrazole has the shortest survival time while Tamoxifen has the longest. P-value of 0.51 denotes the statistical significance of the found results

Discussion

In this study we attempted to determine the effect of menopause status on survival in breast cancer patients. Our study showed that pre-menopause breast cancer patients have a shorter survival time than post-menopause patients, as indicated in Figure 2. This is interesting as typically menopause begins between ages 40-50 (Mayo Clinic, 2020) and one of the uncontrollable factors that increase the risk of breast cancer is aging; women are more likely to be diagnosed with breast cancer at an older age (American Cancer Society, 2022). However, previous published literature indicates that younger breast cancer patients are more susceptible to more aggresive cancers than older patients (Chen et. al, 2016). In Figure 1, our boxplot visualization of the ages of our pre and post-menopause patients supported this finding as the median pre-menopausal age was 44 years while the median post-menopausal age was 63 years. Simultaneously in Figure 2, pre-menopause/younger patients survived up to 4000 days whereas post-menopause/older patients survived 2000 days more (up to 6000 days).

To further investigate why pre-menopause patients had a shorter survival time, we analyzed point mutation percentages in common breast cancer biomarkers such as Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha (PIK3CA), Tumor Protein p53 (TP53), Titin (TTN), E-cadherin (CDH1), Mucin16 (MUC16), and GATA Binding Protein 3 (GATA3). Specifically, PIK3CA, TP53, MUC16, and GATA3 were of greater interest as across both groups, they were mutated more than 10%. This was in line with previous findings where somatic mutations in PIK3CA, TP53, GATA3 were >10% across different breast cancer subtypes (The Cancer Genome Atlas Network, 2012). Using a co-oncoplot, Figure 3, we identified GATA3 mutations as a potential gene mutation involved in affecting survivability. Notably, GATA3 mutations were doubled in the pre-menopause sample (21% mutated) in comparison to the post-menopause sample (10% mutated). Further research in GATA3 revealed that the GATA3 transcription factor is one of the most frequently mutated genes in breast cancer and dysregulation of transcription due to GATA3 mutations may block terminal differentiation, thus promoting tumor growth (Gustin et. al, 2017). Thus, the increased mutation rate of GATA3 in pre-menopausal patients, leading to more aggressive tumor growth, could be one of the factors involved in decreased survivability.

To understand the clinical consequences of the GATA3 mutation, we studied both pre and post-menopausal patients who underwent hormonal therapy and found that Tamoxifen was one of the more efficient drugs for promoting survival time. We selected hormone therapy as GATA3 is a strong biomarker for Estrogen Receptor Positive Breast Cancer subtypes (ER Positive Breast Cancer) (McCleskey et. al, 2015) and ER Positive Breast Cancers are typically treated with hormone therapy (National Cancer Institute, 2022). Using a Kaplan-Meier survival plot, Figure 4, we found that patients treated with Anastrazole had the shortest survival time of around 600 days while patients treated with Tamoxifen had the longest survival time going past 3000 days.

While this study identified that pre-menopausal women have a shorter survival time than post-menopausal women and that GATA3 is one of the key differentially mutated genes between the two groups, there are limitations to our findings. First, the studied samples were from The Cancer Genome Atlas which is based on U.S. patients and thus the results cannot be projected on the international breast cancer population. Second, the mutations studied were simple nucleotide variations and somatic mutations, which is beneficial in removing the confounding factors of family history but also simplifies the identification of differentially mutated genes. Finally, while GATA3 was differentially mutated between our two groups, the mechanism behind how GATA3 decreases the survivability of pre-menopause patients was not analyzed in this study. There may possibly be additional mutations along its pathway that also contribute or are bigger contributors to the decrease in survivability. Nonetheless, while we believe that these limitations have not impacted the primary outcomes of this study, future work could seek to include non-U.S based samples, other mutation-type data, additional controls, and further analyze the pathway of GATA3 and related mutations, pathways, and risk factors that contribute to the decrease in survivability for pre-menopausal breast cancer patients.

References

"Breast Cancer Statistics: How Common Is Breast Cancer?" *American Cancer Society*, https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html#:~:text=The%20American%20Cancer%20Society%27s%20estimates,will%20die%20from%20breast%20cancer.

"Breast Cancer: Breast Cancer Information & Overview." *American Cancer Society*, https://www.cancer.org/cancer/breast-cancer.html.

"The Cancer Genome Atlas Program." *National Cancer Institute*, https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga.

Chen, Hai-long, et al. "Effect of Age on Breast Cancer Patient Prognoses: A Population-Based Study Using the Seer 18 Database." *PLOS ONE*, vol. 11, no. 10, 2016, https://doi.org/10.1371/journal.pone.0165409.

"Comprehensive Molecular Portraits of Human Breast Tumours." *Nature*, vol. 490, no. 7418, 2012, pp. 61–70., https://doi.org/10.1038/nature11412.

Gustin, John P., et al. "*gata3* Frameshift Mutation Promotes Tumor Growth in Human Luminal Breast Cancer Cells and Induces Transcriptional Changes Seen in Primary *gata3* Mutant Breast Cancers." *Oncotarget*, vol. 8, no. 61, 2017, pp. 103415–103427., https://doi.org/10.18632/oncotarget.21910.

"Hormone Therapy for Breast Cancer Fact Sheet." *National Cancer Institute*, https://www.cancer.gov/types/breast/breast-hormone-therapy-fact-sheet#:~:text=Hormone%20therapy%20is%20also%20a,to%20treat%20metastatic%20breast%20cancer.

McCleskey, Brandi C., et al. “GATA3 Expression in Advanced Breast Cancer: Prognostic Value and Organ-Specific Relapse.” *American Journal of Clinical Pathology*, vol. 144, no. 5, 2015, pp. 756–763., https://doi.org/10.1309/ajcp5mmr1fjvvtpk.

Surakasula, Aruna, et al. “A Comparative Study of Pre- and Post-Menopausal Breast Cancer: Risk Factors, Presentation, Characteristics and Management.” *Journal of Research in Pharmacy Practice*, vol. 3, no. 1, 2014, p. 12., https://doi.org/10.4103/2279-042x.132704.

Part 3:
Review Questions

General Concepts

1. **What is TCGA and why is it important?**

    a. TCGA stands for The Cancer Genome Atlas and is a joint cancer genomics program between the National Cancer Institute and the National Human Genome Research institute. It contains genomic, epigenomic, transcriptomic, and proteomic data from 20,000 different samples across 33 different cancer types.

2. **What are some strengths and weaknesses of TCGA?**

    a. Strengths: a wide array of data, pre-existing categorizations, built-in tools, filters, and visualizers

    b. Weaknesses: certain details have missing info that needs to be masked out, the humanistic correlation between data and patient is removed

Coding Skills

1. **What commands are used to save a file to your GitHub repository?**

    a. git add [name_of_file]
    b. git commit -m “[message]”
    c. git push

2. **What command(s) must be run in order to use a standard package in R?**

    a.
    ```
    if (!require(package)){
       install.packages("package")
       }
    library(package_name)
    ```

3. **What command(s) must be run in order to use a Bioconductor package in R?**

    a.
    ```
    if (!require("BiocManager", quietly = TRUE))
       install.packages("BiocManager")
       BiocManager::install(version = "3.15")
    library(BiocManager)

    if (!require("package_name", quietly = TRUE))
       BiocManager::install("package_name")
    library(package_name)
    ```

4. **What is boolean indexing? What are some applications of it?**

    a. Boolean indexing includes a boolean vector (a vector of TRUEs and FALSEs) and applying it to a data frame as a "mask." The steps of boolean masking include first defining what data you are looking for in a logical statement (ex. Ifelse or is.na) and assigning it to a variable (usually a descriptively named mask). Then the mask is applied to the data frame (ex. Data_frame[row_mask, col_mask]) and assigning this to a new data frame or rewriting the old data frame. The mask selects for TRUE values so all FALSE values generated by the logical statement will be removed in the new data frame. Boolean indexing is useful in subsetting data or removing data that could cause issues in plotting or running further analyses.

5. **Draw a mock up (just a few rows and columns) of a sample dataframe. Show an example of the following and explain what each line of code does.**

    a. **an ifelse() statement**

        i. mydf$size_category <- ifelse(mydf$values >= 10, "large", "small)

            1. mydf$size_category creates a new column that the output of the ifelse statement will go into

            2. mydf$values calls the values column of the mydf data frame

            3. mydf$values >= 10 sets up the parameters

            4. if the number in values column is greater than or equal to 10, then it will be assigned as "large"

            5. if it is anything else (ex. less than 10) then it will be assigned as "small"

    b. **boolean indexing**

        i. na_mask <- ifelse(is.na(mydf$value), T, F)

        ii. mydf[!na_mask,]

            1. The ifelse statement checks if there are NA values in the value column of mydf. If a value is NA, then the boolean TRUE will go into the na_mask vector. If there is not, then the boolean FALSE will go into the na_mask vector.

2. Then the mask is applied to the data frame on the row side. Thus it will get rid of any rows that have na values (instead of columns. There is an ! in front of na_mask because we want to keep the non-NA values (the mask selects for true values but by putting ! in front of it, we keep the false values instead).