Project Proposals by Group 7

Instructions

1. Create a data folder under your working directory
2. Download the data to run the rmd file into the data directory.

- Dataset 1
- Dataset 2
- Dataset 3

Index

Proposal 1

Exploratory Data Analysis for dataset containing Lung Cancer associated SNPs from GWAS Catalog

Dataset can be found here

Dataset Link - Link

Appendix

Dataset Introduction

The dataset contains information about Single Nucleotide Polymorphisms (SNPs) associated with lung cancer sourced from the GWAS dataset. Each record within the dataset corresponds to a SNP reported in a published scientific literature and is reported to have a statistical link to lung cancer susceptibility or outcomes. The dataset includes key features such as SNP ID, the chromosome where the SNP is located at and its exact position in the chromosome, the significance of the SNP is captured through p-values, odds ratio which represents the probability of the said SNP to occur. Altogether, this dataset proves to be comprehensive resource for exploring genetic risk factors underlying lung cancer.

Dataset Justification

We chose this dataset as we acknowledge that lung cancer remains to be one of the leading cause of death in the world and understanding about the underlying risk factors that cause this fatal disease can help us understand the mechanism of the cancer and develop targeted therapies to prevent it. GWAS Catalog is a well-established knowledge resource for studying these risk factors that combines categorical columns such as SNP ID, Gene Name, ancestry as well as numerical variables such as p-values, odds ratio, and sample sizes. This makes the dataset a flexible option for numerical and categorical analysis. Its biomedical association, rich annotation, and potential to explore genetic risk factors make it a strong candidate for meaningful analysis.

Research Question

We plan to use the data from GWAS to perform eQTL analysis. eQTL or Expression Quantitative Trait Loci are genomic loci (positions) within the genome that influence gene expression levels. There are 2 types of eQTLs: cis-eQTLs are SNPs which occur within close proximity of a gene thereby controlling the gene expression by possibly altering the region where a transcription factor binds for gene expression. The second type is trans-eQTLs are SNPs which are located anywhere within the human genome, even on a different chromosome. These eQTLs are much harder to identify and often act together with other trans-eQTLs in coordinating gene expression. We would be using this dataset to identify potential eQTLs based on constraints such as p-value less than 5E-8, the proximity of the SNP to the gene etc. After identifying potential eQTLs for Lung Cancer, we would be comparing them against SNPs recorded in the established eQTL database for cancer PancanQTL

Data Pre-processing and cleanup

We found that the dataset contained a lot of NR values in the Risk Allele Frequency column which we plan to convert to NA values as they are not important because the column is actually numerical. We found dirty values within chromosome ID column which we plan to standardize to integers between 1 and 22, as well as X, Y, and MT. We found that certain numerical columns in the dataset that was misinterpreted to be string, which we would handling using as.integer or other similar functions for the other misinterpreted columns. We would also be checking for duplicate SNP IDs to maintain the uniqueness of each SNP.

Planned Statistical Methods

We could possibly implement classical model such as `linear regression` to understand the relationships between Expression which could be obtained from GEO (Gene Expression Omnibus) and Genotype which is the SNP that could be integrated using the existing R package `MatrixEQTL`. This will involve modeling gene expression as the dependent variable and SNP genotype (coded as 0, 1, or 2 based on minor allele count) as the independent variable, while controlling for potential confounders such as ancestry and study population. `Chi-square` tests will be employed to examine associations between categorical variables such as chromosome distribution and risk allele presence, while `correlation analysis` will assess relationships between continuous variables like p-values, odds ratios, and risk allele frequencies.

Limitations

The limitation of the dataset is that there are a lot of NA values in a lot of fields that would affect the downstream analysis. We plan to analyse the dataset using the proposed method using only a select fiew fields as some of the other fields contain descriptions/annotations that are not useful for analysis. We also found that the data type of certain fields are being misinterpreted so we would explore the field before changing the data type of the field. Additionally, the presence of 337 duplicate SNP records suggests inconsistent data curation practices, requiring careful selection criteria to retain the most reliable entries without introducing selection bias. These missing values could introduce bias if they are not missing at random, potentially skewing our understanding of allele frequencies and effect sizes.

Libraries

```r
library(ggplot2)
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v lubridate 1.9.4     v tibble    3.3.0
v purrr     1.1.0     v tidyr     1.3.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
library(stringr)
```

Exploratory Data Analysis

Loading our dataset

```r
data <- read.table('data/gwas-association-downloaded_2025-09-25-MONDO_0008903-withChildTraits
```

Summary of our dataset

```r
column_names <- colnames(data)
cat(paste0('Column names of the GWAS dataset:\t', column_names, '\n'))
```

```
Column names of the GWAS dataset:   DATE.ADDED.TO.CATALOG
 Column names of the GWAS dataset:  PUBMEDID
 Column names of the GWAS dataset:  FIRST.AUTHOR
 Column names of the GWAS dataset:  DATE
 Column names of the GWAS dataset:  JOURNAL
 Column names of the GWAS dataset:  LINK
 Column names of the GWAS dataset:  STUDY
 Column names of the GWAS dataset:  DISEASE.TRAIT
 Column names of the GWAS dataset:  INITIAL.SAMPLE.SIZE
 Column names of the GWAS dataset:  REPLICATION.SAMPLE.SIZE
 Column names of the GWAS dataset:  REGION
 Column names of the GWAS dataset:  CHR_ID
 Column names of the GWAS dataset:  CHR_POS
 Column names of the GWAS dataset:  REPORTED.GENE.S.
 Column names of the GWAS dataset:  MAPPED_GENE
 Column names of the GWAS dataset:  UPSTREAM_GENE_ID
 Column names of the GWAS dataset:  DOWNSTREAM_GENE_ID
 Column names of the GWAS dataset:  SNP_GENE_IDS
 Column names of the GWAS dataset:  UPSTREAM_GENE_DISTANCE
 Column names of the GWAS dataset:  DOWNSTREAM_GENE_DISTANCE
 Column names of the GWAS dataset:  STRONGEST.SNP.RISK.ALLELE
 Column names of the GWAS dataset:  SNPS
 Column names of the GWAS dataset:  MERGED
 Column names of the GWAS dataset:  SNP_ID_CURRENT
 Column names of the GWAS dataset:  CONTEXT
 Column names of the GWAS dataset:  INTERGENIC
 Column names of the GWAS dataset:  RISK.ALLELE.FREQUENCY
 Column names of the GWAS dataset:  P.VALUE
 Column names of the GWAS dataset:  PVALUE_MLOG
 Column names of the GWAS dataset:  P.VALUE..TEXT.
 Column names of the GWAS dataset:  OR.or.BETA
 Column names of the GWAS dataset:  X95..CI..TEXT.
 Column names of the GWAS dataset:  PLATFORM..SNPS.PASSING.QC.
 Column names of the GWAS dataset:  CNV
 Column names of the GWAS dataset:  MAPPED_TRAIT
 Column names of the GWAS dataset:  MAPPED_TRAIT_URI
 Column names of the GWAS dataset:  STUDY.ACCESSION
 Column names of the GWAS dataset:  GENOTYPING.TECHNOLOGY
```

```r
summary(data)
```

```
DATE.ADDED.TO.CATALOG    PUBMEDID        FIRST.AUTHOR           DATE
Length:1748           Min.   :18385676   Length:1748        Length:1748
Class :character      1st Qu.:28604730   Class :character   Class :character
Mode  :character      Median :31326317   Mode  :character   Mode  :character
                      Mean   :32965060
                      3rd Qu.:37689528
                      Max.   :40829600

  JOURNAL              LINK              STUDY          DISEASE.TRAIT
Length:1748        Length:1748        Length:1748        Length:1748
```

```
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character


INITIAL.SAMPLE.SIZE REPLICATION.SAMPLE.SIZE    REGION
Length:1748         Length:1748             Length:1748
Class :character    Class :character        Class :character
Mode  :character    Mode  :character        Mode  :character


   CHR_ID            CHR_POS          REPORTED.GENE.S.   MAPPED_GENE
Length:1748        Length:1748        Length:1748        Length:1748
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character


UPSTREAM_GENE_ID   DOWNSTREAM_GENE_ID SNP_GENE_IDS
Length:1748        Length:1748        Length:1748
Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character


UPSTREAM_GENE_DISTANCE DOWNSTREAM_GENE_DISTANCE STRONGEST.SNP.RISK.ALLELE
Min.   :      1        Min.   :     30          Length:1748
1st Qu.:  10198        1st Qu.:   8771          Class :character
Median :  29873        Median :  25382          Mode  :character
Mean   :  94663        Mean   :  95607
3rd Qu.:  96508        3rd Qu.:  86828
Max.   :3249767        Max.   :4177110
NA's   :1199           NA's   :1199
    SNPS               MERGED        SNP_ID_CURRENT        CONTEXT
Length:1748        Min.   :0.00000   Min.   :6.569e+03   Length:1748
Class :character   1st Qu.:0.00000   1st Qu.:4.400e+06   Class :character
Mode  :character   Median :0.00000   Median :1.234e+07   Mode  :character
                   Mean   :0.05492   Mean   :5.399e+07
                   3rd Qu.:0.00000   3rd Qu.:7.744e+07
                   Max.   :1.00000   Max.   :1.827e+09
                                     NA's   :114
  INTERGENIC     RISK.ALLELE.FREQUENCY    P.VALUE           PVALUE_MLOG
Min.   :0.0000   Length:1748          Min.   :0.000e+00   Min.   :  5.000
1st Qu.:0.0000   Class :character      1st Qu.:6.000e-10   1st Qu.:  5.398
Median :0.0000   Mode  :character      Median :4.000e-07   Median :  6.398
Mean   :0.3754                         Mean   :2.115e-06   Mean   :  9.308
3rd Qu.:1.0000                         3rd Qu.:4.000e-06   3rd Qu.:  9.222
Max.   :1.0000                         Max.   :1.000e-05   Max.   :178.097
NA's   :27
P.VALUE..TEXT.       OR.or.BETA      X95..CI..TEXT.
Length:1748        Min.   :  0.010   Length:1748
Class :character   1st Qu.:  1.070   Class :character
Mode  :character   Median :  1.176   Mode  :character
                   Mean   :  1.973
                   3rd Qu.:  1.645
                   Max.   :101.639
                   NA's   :121
PLATFORM..SNPS.PASSING.QC.     CNV            MAPPED_TRAIT
Length:1748                Length:1748        Length:1748
Class :character           Class :character   Class :character
Mode  :character           Mode  :character   Mode  :character
```

```
MAPPED_TRAIT_URI   STUDY.ACCESSION    GENOTYPING.TECHNOLOGY
Length:1748        Length:1748        Length:1748
Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character
```

We see that some variables such as Chromosome position and Pubmed ID which misinterpreted as different data types by R. So in the next section, we change the data type of the column to match their true quality.

Changing the some data type of some columns as they were declared as str but were int

```r
data$CHR_POS <- as.integer(data$CHR_POS)
```

```
Warning: NAs introduced by coercion
```

```r
data$PUBMEDID <- as.character(data$PUBMEDID)
```

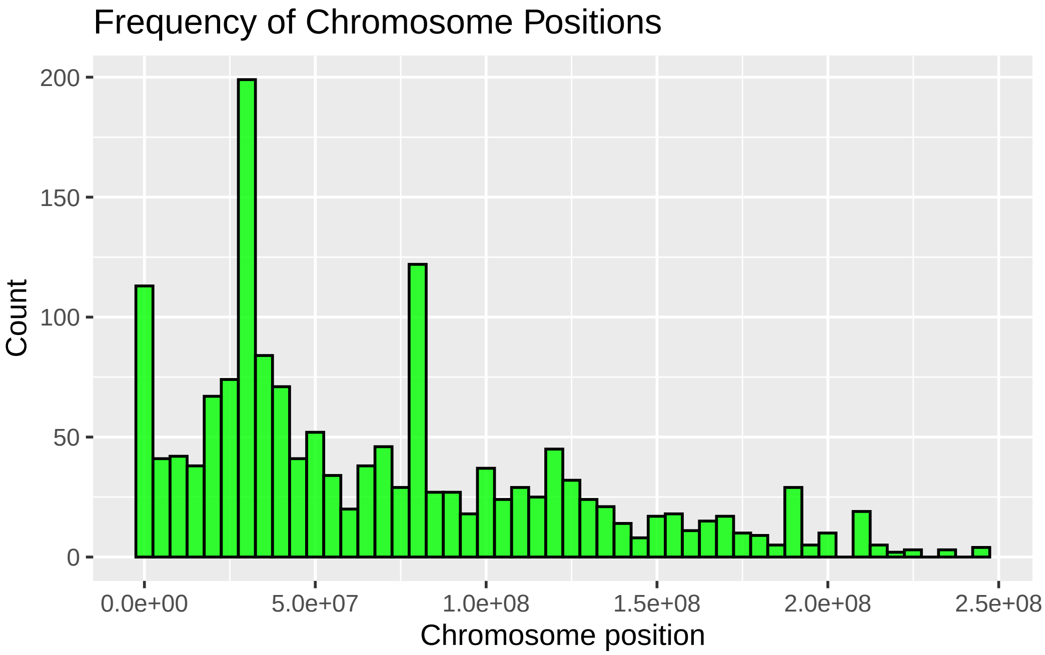
Variable Description

Histogram of the Positions of the SNPs

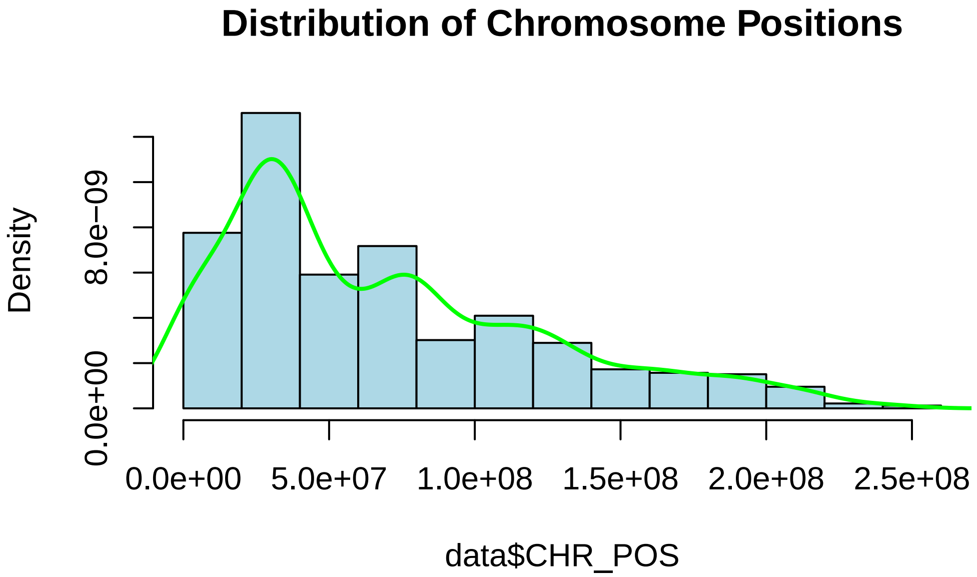
We are looking at the distribution of the positions where these SNPs occur within the genome to get an idea if most of the mutations occur upstream (towards the 5' end) or downstream (towards the 3' end) of the genome.

```r
ggplot(data, aes(CHR_POS, fill=CHR_ID)) +
    geom_histogram(bins = 50, fill = 'green', color = 'black', alpha = 0.8) +
    labs(
        title = 'Frequency of Chromosome Positions',
        x = 'Chromosome position',
        y = 'Count'
    )
```

```
Warning: Removed 124 rows containing non-finite outside the scale range
(`stat_bin()`).
```



```r
h <-hist(data$CHR_POS, probability =  TRUE, col = 'lightblue', main = 'Distribution of Chromo
lines(density(na.omit(data$CHR_POS)), col = 'green', lwd = 2)
```



```r
summary(data$CHR_POS)
```

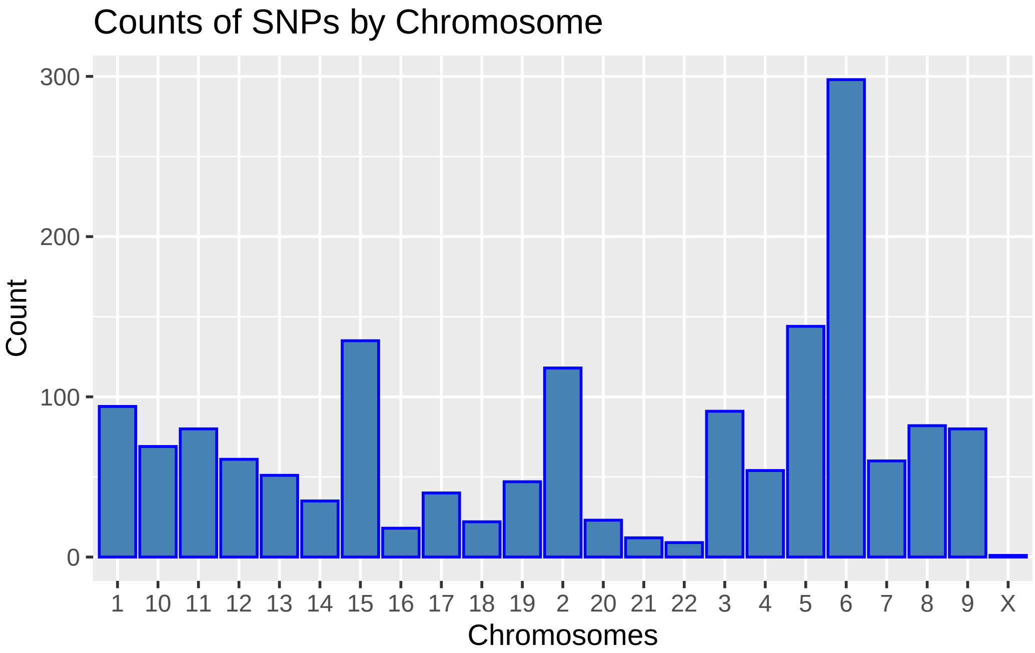
```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's 
   242581  28645826  51549004  68685302 101493781 245121489       124 
```

The SNP chromosome positions nearly span the entirety of the human genome with values ranging from 242kb (kilobases) to 245Mb (Megabases). The median SNP position is approximately 51.5Mb while the mean which is being skewed by larger values is 68.6Mb. SNPs are distributed across the genome, while clusters of SNPs occurring in certain regions may indicate the possibility of potential trans-eQTLs.

Barplot of Chromosomes

These chromosomes are condensed forms of the DNA where the SNPs occur and we are visualizng which of the chromosomes in the body seems to be associated with frequent mutations that leads to Lung cancer.

```r
valid_chr <- as.character(c(1:22, "X", "Y", "MT"))
data <- data %>%      # Cleaning the CHR_ID as it contained dirty mixed values
    mutate(
        CHR_ID_clean = str_trim(CHR_ID),
        CHR_ID_clean = str_extract(CHR_ID_clean, "^[0-9]+$|^X$|^Y$|^MT$"),
        CHR_ID_clean = ifelse(CHR_ID_clean %in% valid_chr, CHR_ID_clean, NA)
) %>%
    filter(!is.na(CHR_ID_clean))
ggplot(data, aes(x = CHR_ID_clean)) +
    geom_bar(fill = 'steelblue',color = 'blue', na.rm=TRUE) +
    labs(
        title = 'Counts of SNPs by Chromosome',
        x = 'Chromosomes',
        y = 'Count'
    )
```

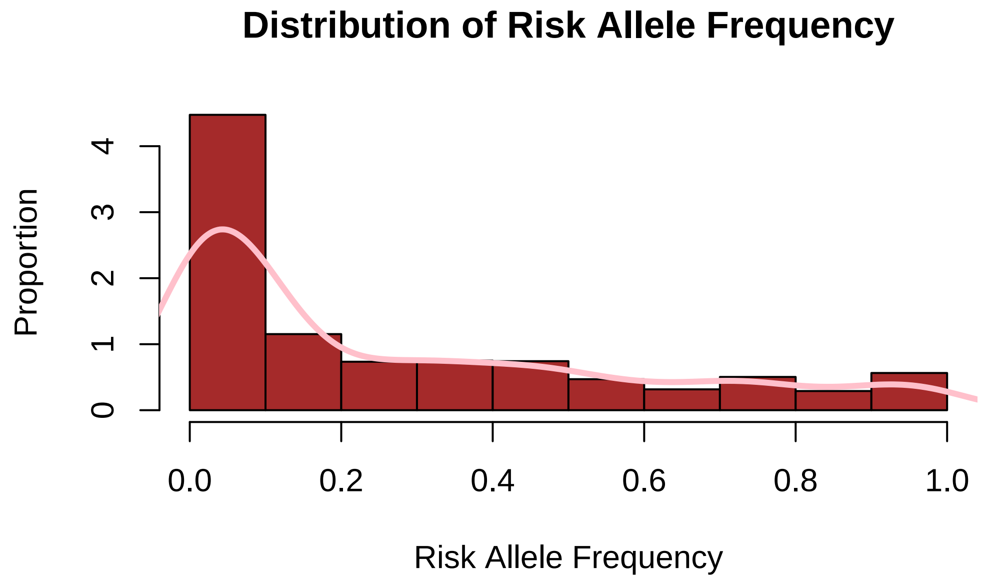


We find that most of the SNPs are found on Chromosome 6, followed by chromosomes 15 and 2. From this, we can infer that mutations associated with chromosome 6 may play a particularly important role in lung cancer susceptibility. At a glance, these findings indicate hotspots for targeted therapy that may decrease the risk of lung cancer susceptibility.

Histogram of Risk Allele frequency

Risk Allele Frequency is the proportion of chromosomes in a population that carry this risk allele as these alleles increase the likelihood of developing the disease.

```
data <- data %>%
    mutate(
        RAF = ifelse(RISK.ALLELE.FREQUENCY == "NR", NA, RISK.ALLELE.FREQUENCY),
        RAF = as.numeric(RAF)
    )

h <- hist(data$RAF, probability = TRUE,col = 'brown', main = 'Distribution of Risk Allele Fre
lines(density(na.omit(data$RAF)), col = 'pink', lwd = 3)
```



```r
summary(data$RAF)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
 0.0000  0.0240  0.1230  0.2728  0.4576  0.9998     453 
```

The Risk Allele Frequency (RAF) values range from 0 (rare alleles) to 0.9998 (nearly fixed alleles). The median RAF 0.1230, which is more useful in this feature of the dataset as it is right-skewed, indicating that half of the risk alleles occur in less than 12% of the population. The mean RAF 0.2728 though higher is affected by some common alleles. The IQR which lies between 0.0240 and 0.4576 captures majority of the moderately frequent risk alleles. Also 453 alleles lacked RAF values and were marked `NR` which were converted to `NA` values.

Checking for duplicated SNPs

```r
sum(duplicated(data$SNPS))
```

```
[1] 337
```

We found that there are around 337 duplicate SNPs that need to be filtered out. These duplicated records may arise from multiple studies reporting on the same SNP under different experimental conditions, sample populations, and statistical models. We plan to do the filtering based on biologically meaningful properties such as Risk Allele Frequency, p-value, odds ratio. This ensures that for each unique SNP, only the most reliable and informative record is being retained.

Citations

1. Sollis E, Mosaku A, Abid A, Buniello A, Cerezo M, Gil L, Groza T, Güneş O, Hall P, Hayhurst J, Ibrahim A, Ji Y, John S, Lewis E, MacArthur JAL, McMahon A, Osumi-Sutherland D, Panoutsopoulou K, Pendlington Z, Ramachandran S, Stefancsik R, Stewart J, Whetzel P, Wilson R, Hindorff L, Cunningham F, Lambert SA, Inouye M, Parkinson H, Harris LW. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. Nucleic Acids Res. 2023 Jan 6;51(D1):D977-D985. doi: 10.1093/nar/gkac1010. PMID: 36350656; PMCID: PMC9825413.

2. Chen, C., Liu, Y., Luo, M., Yang, J., Chen, Y., Wang, R., Zhou, J., Zang, Y., Diao, L., & Han, L. (2024). PancanQTLv2.0: a comprehensive resource for expression quantitative trait loci across human cancers. Nucleic acids research, 52(D1), D1400–D1406. https://doi.org/10.1093/nar/gkad916

Proposal 2

Appendix

1. Link to the Dataset:

Dataset Link - Link

Goad, Nathan, 2018, “Diabetic Ketoacidosis and Hyperchloremia Full Dataset,” https://doi.org/10.7910/DVN/P
Harvard Dataverse, V1.

2. Introduction to the Dataset:

This dataset contains clinical and laboratory data from a retrospective cohort study conducted at Wake Forest Baptist Medical Center. It includes information on adult patients admitted with diabetic ketoacidosis (DKA) and compares those who developed hyperchloremia during management to those who maintained normal chloride levels. The dataset captures baseline demographic, clinical, and laboratory measurements, as well as treatment details and a range of clinical outcomes. The goal is to explore the clinical impact of hyperchloremia, a common electrolyte imbalance often associated with DKA treatment, on patient outcomes.

3. Dataset Justification:

We selected the Diabetic Ketoacidosis and Hyperchloremia dataset for its strong relevance to current clinical practice and its robust structure for statistical analysis. Understanding the factors that influence outcomes in DKA patients is a critical public health concern, and this dataset provides a rich, real-world context for our research. It contains a diverse set of variables, including both continuous (e.g., blood glucose, length of stay) and categorical (e.g., AKI status, gender) data, making it an ideal platform for applying a wide range of descriptive, inferential, and predictive statistical methods taught in this course. The dataset is publicly available, ethically sourced, de-identified, and well-documented, ensuring that our analysis is transparent, reproducible, and compliant with all project guidelines.

4. RESEARCH QUESTION

A. Impact of IV Fluid Type on Chloride Levels This question investigates whether the type of intravenous fluid given during DKA treatment affects peak chloride levels and the likelihood of developing hyperchloremia. Since chloride-rich fluids like 0.9% NaCl may increase the risk of electrolyte imbalances compared to balanced solutions like Lactated Ringer’s or Plasma-Lyte, this analysis aims to see if outcomes differ based on fluid choice. Adjusting for factors such as total IV fluid volume and initial DKA severity ensures that the observed effects reflect fluid type rather than overall treatment intensity.

B. Predictors of Acute Kidney Injury (AKI) This question aims to identify which patient characteristics at admission (e.g., age, APACHE II score, baseline creatinine, admission glucose and pH) and treatment-related factors (e.g., volume of 0.9% NaCl, peak chloride, sepsis, hypotension, bicarbonate use) are most strongly linked to the development of AKI during hospitalization. Predictive modeling will help determine which variables contribute most to AKI risk, allowing clinicians to identify high-risk patients early and adjust management strategies accordingly.

5. VARIABLE DESCRIPTION

```r
library(readxl)
library(tidyverse)
library(knitr)

#load dataset into R

data <- read_excel("data/Hyperchloremia and DKA Dataset .xlsx")
head(data)
```

```
# A tibble: 6 x 50
  `Age (years)` `Weight (kg)` `BMI (kg/m2)` Diabetes Type (type 1 = 0, type 2 ~1
          <dbl>         <dbl>         <dbl>                                <dbl>
1            60          96            29.7                                    1
2            50         104.           25.5                                    1
3            25          85.5          23.6                                    0
4            47         130.           43.8                                    0
5            61          46.5          20.2                                    0
6            20         100.           30                                      0
# i abbreviated name: 1: `Diabetes Type (type 1 = 0, type 2 = 1)`
# i 46 more variables: `Gender (0 = male, 1 = female)` <dbl>,
#   `Unit (0 = Medical ICU, 1 = Intermediate Care)` <dbl>,
#   `APACHE II Score` <dbl>, `Admission Chloride (mEq/L)` <dbl>,
#   `Peak Chloride (mEq/L)` <dbl>,
#   `Peak Chloride Category (0 = normochloremia, 1 = hyperchloremia)` <dbl>,
#   `Admission Bicarbonate (mEq/L)` <dbl>, ...
```

```r
na_counts <- sapply(data, function(x) sum(is.na(x)))
na_counts <- na_counts[na_counts > 0]
print("Variables with Missing Values:")
```

```
[1] "Variables with Missing Values:"
```

```r
print(na_counts)
```

```
Hospital AKI, Time of Onset from inition of DKA treatment (hours) 
                                                               84 
                Duration of Admission Acute Kidney Injury (hours) 
                                                               14 
         Subcutaneous Insulin Overlap by 1-2 hr (0 = no, 1 = yes) 
                                                                2 
```

This output shows the number of missing values for each variable. Most variables have no missing values, indicating a nearly complete dataset, but a few variables like Hospital AKI, Time of Onset from initiation of DKA treatment (84 missing), Duration of Admission Acute Kidney Injury (14 missing), and Subcutaneous Insulin Overlap by 1-2 hr (2 missing) have some gaps that may need careful handling during analysis. Overall, the dataset is largely complete and ready for exploratory analysis.

The dataset contains 50 variables, of which 23 are categorical and 27 are numerical. Categorical variables describe groups or categories, while numerical variables represent measurable quantities.

Key Variables Selection for Analysis

Research Question 1: Impact of IV Fluid Type on Chloride Levels.The main outcome of interest is the Peak Chloride Category (0 = normochloremia, 1 = hyperchloremia), which indicates whether patients developed hyperchloremia during treatment. Key numerical variables include Peak Chloride (mEq/L) as the continuous measure of serum chloride, Total IV Fluid Volume (mL) to account for the overall amount of fluid administered, and Admission Bicarbonate (mEq/L) to reflect baseline metabolic status. These variables collectively allow assessment of how the type and quantity of intravenous fluids may influence chloride levels in patients with DKA.

Research Question 2: Predictors of Acute Kidney Injury (AKI).The primary outcome is the occurrence of Acute Kidney Injury in Hospital (0 = no, 1 = yes). Important categorical predictors include Sepsis (0 = no, 1 = yes), Hypotension (0 = no, 1 = yes), and Peak Chloride Category, to evaluate whether hyperchloremia contributes independently to AKI risk. Numerical predictors include Age (years), APACHE II Score as a measure of illness severity, Admission Serum Creatinine (mg/dL) to capture baseline kidney function, and Volume of 0.9% NaCl (mL) to explore potential fluid-related effects. Together, these variables provide a clinically meaningful framework for identifying factors associated with AKI in critically ill DKA patients.

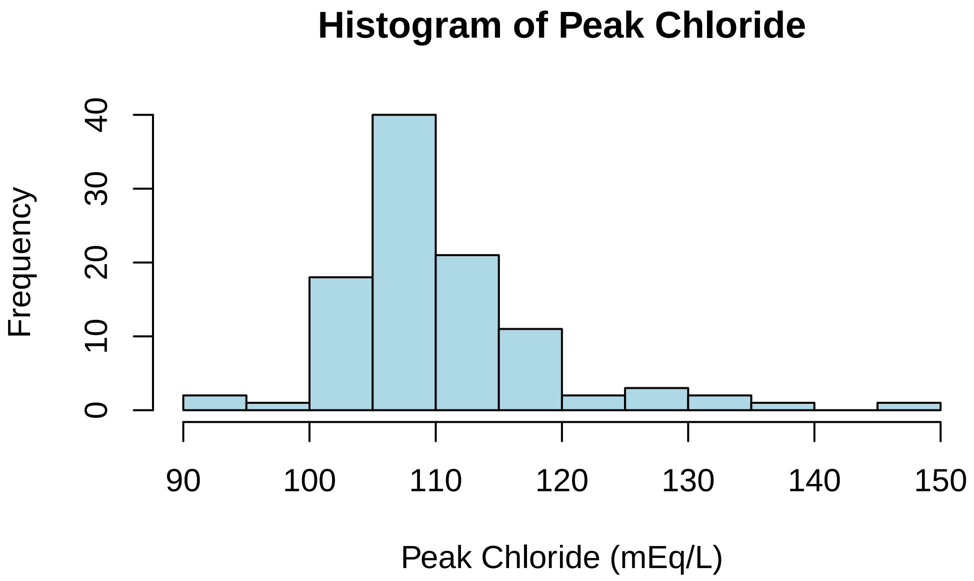
```
#key numerical variables list
numerical_vars <- c(
  "Peak Chloride (mEq/L)",
  "Total IV Fluid Volume (mL)",
  "Admission Bicarbonate (mEq/L)",
  "Age (years)",
  "APACHE II Score",
  "Admission Serum Creatinine (mg/dL)",
  "Volume of 0.9% NaCl (mL)"
)
# key categorical variables list
categorical_vars <- c(
  "Acute Kidney Injury In Hospital (0 = no, 1 yes)",
  "Sepsis (0 = no, 1 = yes)",
  "Hypotension (0 = no, 1 = yes)",
  "Peak Chloride Category (0 = normochloremia, 1 = hyperchloremia)"
)
# Converting categorical variables to factors with meaningful labels
data$`Acute Kidney Injury In Hospital (0 = no, 1 yes)` <-
  factor(data$`Acute Kidney Injury In Hospital (0 = no, 1 yes)`,
         levels = c(0,1),
         labels = c("No","Yes"))
data$`Sepsis (0 = no, 1 = yes)` <-
  factor(data$`Sepsis (0 = no, 1 = yes)`,
         levels = c(0, 1),
         labels = c("No", "Yes"))
data$`Hypotension (0 = no, 1 = yes)` <-
  factor(data$`Hypotension (0 = no, 1 = yes)`,
         levels = c(0, 1),
         labels = c("No", "Yes"))
data$`Peak Chloride Category (0 = normochloremia, 1 = hyperchloremia)` <-
  factor(data$`Peak Chloride Category (0 = normochloremia, 1 = hyperchloremia)`,
         levels = c(0, 1),
         labels = c("Normochloremia", "Hyperchloremia"))
# Summary of numerical variables
summary(data[numerical_vars])
```

```
 Peak Chloride (mEq/L) Total IV Fluid Volume (mL) Admission Bicarbonate (mEq/L)
 Min.   : 91.0         Min.   : 1200              Min.   : 2.00
 1st Qu.:106.0         1st Qu.: 3356              1st Qu.: 7.00
 Median :110.0         Median : 4531              Median :11.00
 Mean   :111.1         Mean   : 5424              Mean   :11.48
 3rd Qu.:114.0         3rd Qu.: 6482              3rd Qu.:15.00  
 Max.   :147.0         Max.   :17750              Max.   :22.00  
  Age (years)    APACHE II Score Admission Serum Creatinine (mg/dL)
 Min.   :18.00   Min.   : 6.00   Min.   :0.640                     
 1st Qu.:28.00   1st Qu.:10.00   1st Qu.:1.020                     
 Median :39.50   Median :15.00   Median :1.300                     
 Mean   :41.98   Mean   :16.23   Mean   :1.517                     
 3rd Qu.:54.00   3rd Qu.:21.00   3rd Qu.:1.732                     
 Max.   :83.00   Max.   :36.00   Max.   :4.200                     
 Volume of 0.9% NaCl (mL)
 Min.   : 1000           
 1st Qu.: 2125           
 Median : 3340           
 Mean   : 3513           
 3rd Qu.: 4417           
 Max.   :11325           
```

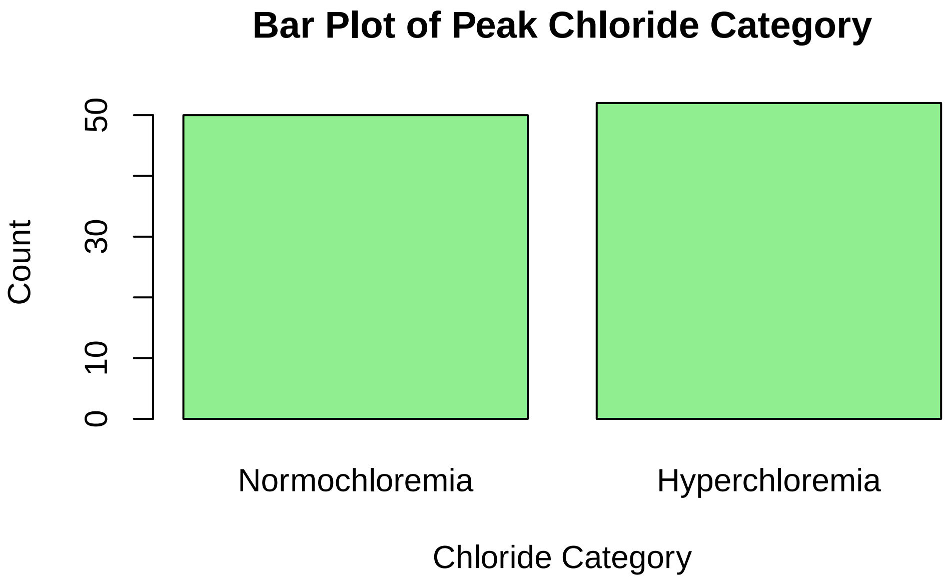
```r
# Summary of categorical variables
sapply(data[, categorical_vars], function(x) table(x))
```

```
    Acute Kidney Injury In Hospital (0 = no, 1 yes) Sepsis (0 = no, 1 = yes)
No                                               84                       96
Yes                                              18                        6
    Hypotension (0 = no, 1 = yes)
No                             82
Yes                            20
    Peak Chloride Category (0 = normochloremia, 1 = hyperchloremia)
No                                                               50
Yes                                                              52
```

```r
# Histogram for one numerical variable (e.g., Peak Chloride)
hist(data$`Peak Chloride (mEq/L)`,
     main = "Histogram of Peak Chloride",
     xlab = "Peak Chloride (mEq/L)",
     col = "lightblue",
     border = "black")
```



```
# Bar plot for one categorical variable (e.g., Peak Chloride Category)
barplot(table(data$`Peak Chloride Category (0 = normochloremia, 1 = hyperchloremia)`),
        main = "Bar Plot of Peak Chloride Category",
        xlab = "Chloride Category",
        ylab = "Count",
        col = "lightgreen")
```



Interpretation: The dataset shows a balanced distribution between the two categories of Peak Chloride, with nearly equal numbers of patients in the normochloremia and hyperchloremia groups, which is ideal for comparative analyses. The histogram of Peak Chloride reveals a right-skewed distribution, where most patients have chloride levels within the normal to mildly elevated range (approximately 100–115 mEq/L), but a small subset exhibits much higher values, creating a long tail. This indicates that while the majority of patients maintain typical chloride levels, there is a distinct group experiencing markedly elevated chloride, highlighting variability that is important for understanding the impact of IV fluids and identifying potential risk factors.

6.DATA PREPROCESSING AND CLEANUP

The analysis will begin with loading the dataset into R and reviewing its structure to understand the types and distributions of variables. Planned steps include checking for missing values and inconsistencies in key variables, as well as identifying and handling any duplicate records to prevent bias. Variable types will be verified and, if necessary, converted—for example, numeric data stored as characters will be recoded to numeric format. Variable names will be standardized to ensure they are clear and R-friendly. Key categorical variables, such as Sepsis (0 = no, 1 = yes) and Acute Kidney Injury In Hospital (0 = no, 1 = yes), may be converted into factors with meaningful labels to improve interpretability and facilitate plotting. Additionally, derived variables, such as a categorical variable representing the dominant IV fluid type, may be created to address specific research questions. Variables with a high proportion of missing data will be noted and potentially excluded from primary analyses to maintain data quality. These preparatory steps are designed to ensure the dataset is reliable, consistent, and ready for subsequent statistical and graphical analyses.

7.Descriptive statistics and visualizations

Descriptive Statistics

```
#summary table for numerical variables
num_summary <- data.frame(
  Variable = numerical_vars,
  Mean = as.numeric(sapply(numerical_vars, function(var) mean(data[[var]], na.rm = TRUE))),
  Median = as.numeric(sapply(numerical_vars, function(var) median(data[[var]], na.rm = TRUE)
  SD = as.numeric(sapply(numerical_vars, function(var) sd(data[[var]], na.rm = TRUE))),
  Min = as.numeric(sapply(numerical_vars, function(var) min(data[[var]], na.rm = TRUE))),
  Max = as.numeric(sapply(numerical_vars, function(var) max(data[[var]], na.rm = TRUE)))
)
print(num_summary)
```

```
                            Variable        Mean  Median           SD     Min
1              Peak Chloride (mEq/L)  111.147059  110.00    8.1159872   91.00
2         Total IV Fluid Volume (mL) 5424.138235 4531.25 3089.9950424 1200.00
3      Admission Bicarbonate (mEq/L)   11.480392   11.00    5.0497140    2.00
4                        Age (years)   41.980392   39.50   16.7698341   18.00
5                    APACHE II Score   16.225490   15.00    7.4007049    6.00
6 Admission Serum Creatinine (mg/dL)    1.517059    1.30    0.7511873    0.64
7           Volume of 0.9% NaCl (mL) 3512.768627 3340.00 1887.0604440 1000.00
      Max
1   147.0
2 17750.0
3    22.0
4    83.0
5    36.0
6     4.2
7 11325.0
```

```
#Summary table for categorical variables
cat_summary <- lapply(categorical_vars, function(var) {
  tbl <- table(data[[var]])
  prop <- prop.table(tbl) * 100
  list(Counts = tbl, Proportions = round(prop, 1))
})
names(cat_summary) <- categorical_vars
print("Categorical Variables Summary:")
```

```
[1] "Categorical Variables Summary:"
```

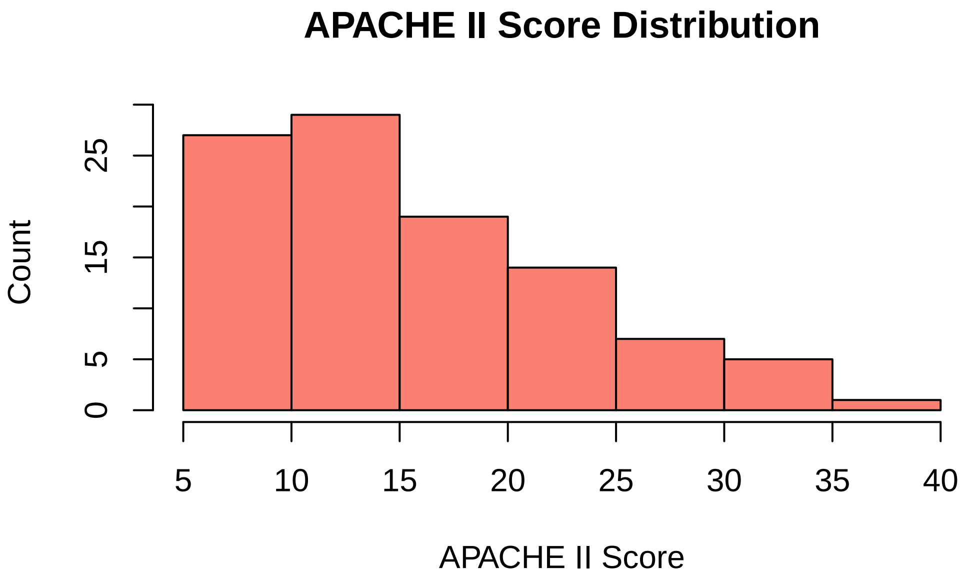
```
print(cat_summary)
```

```
$`Acute Kidney Injury In Hospital (0 = no, 1 yes)`
$`Acute Kidney Injury In Hospital (0 = no, 1 yes)`$Counts

 No Yes
 84  18

$`Acute Kidney Injury In Hospital (0 = no, 1 yes)`$Proportions

  No  Yes
82.4 17.6


$`Sepsis (0 = no, 1 = yes)`
$`Sepsis (0 = no, 1 = yes)`$Counts

 No Yes
 96   6

$`Sepsis (0 = no, 1 = yes)`$Proportions

  No  Yes
94.1  5.9


$`Hypotension (0 = no, 1 = yes)`
$`Hypotension (0 = no, 1 = yes)`$Counts

 No Yes
 82  20

$`Hypotension (0 = no, 1 = yes)`$Proportions
```

```
  No  Yes 
80.4 19.6 


$`Peak Chloride Category (0 = normochloremia, 1 = hyperchloremia)`
$`Peak Chloride Category (0 = normochloremia, 1 = hyperchloremia)`$Counts

Normochloremia Hyperchloremia 
            50             52 

$`Peak Chloride Category (0 = normochloremia, 1 = hyperchloremia)`$Proportions

Normochloremia Hyperchloremia 
            49             51 
```

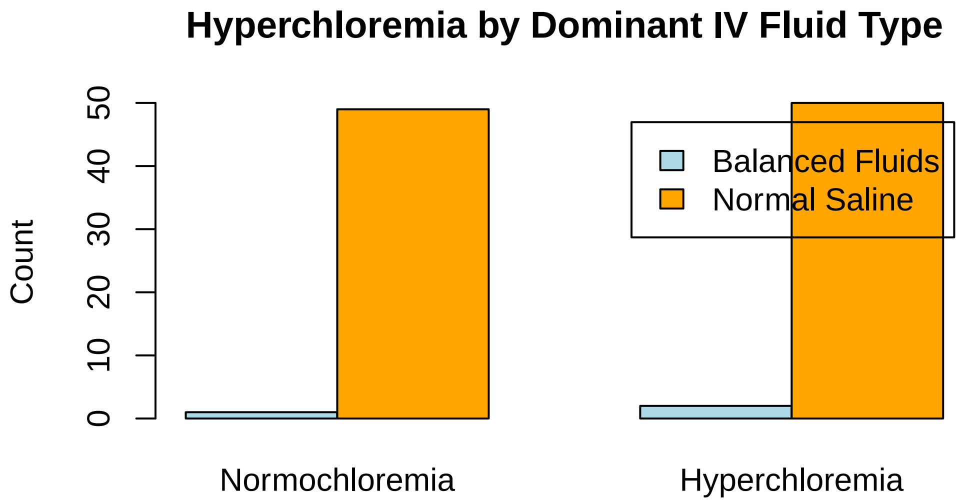
DATA VISUALIZATION

```r
# Histogram for APACHE II Score
hist(data$`APACHE II Score`,
     main = "APACHE II Score Distribution",
     xlab = "APACHE II Score",
     ylab = "Count",
     col = "salmon",
     border = "black")
```



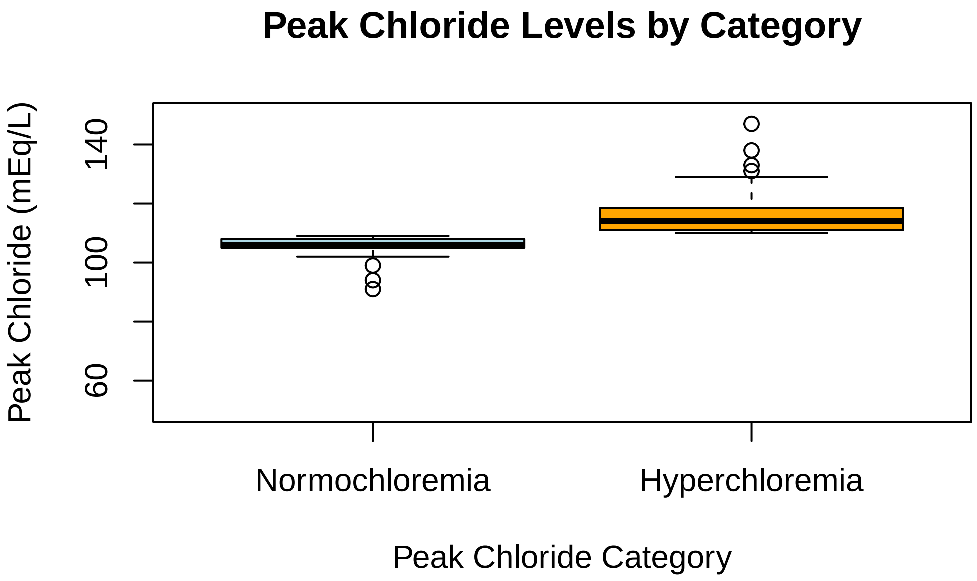
The histogram provide important insights into the baseline characteristics of the patient cohort that are directly relevant to predicting AKI. The APACHE II Score distribution is right-skewed, with most patients having lower scores but a long tail of critically ill patients, highlighting the wide range of illness severity that must be accounted for in the analysis.

```
data$Dominant_Fluid <- ifelse(data$`Volume of 0.9% NaCl (mL)` >
                               (data$`Volume of Lactated Ringer's (mL)` + data$`Volume of Pla
                               "Normal Saline", "Balanced Fluids")

#Barplot: Peak Chloride Category vs Dominant Fluid
chloride_fluid_table <- table(data$Dominant_Fluid, data$`Peak Chloride Category (0 = normochl
barplot(chloride_fluid_table,
        beside = TRUE,
        col = c("lightblue", "orange"),
        main = "Hyperchloremia by Dominant IV Fluid Type",
        ylab = "Count",
        legend.text = rownames(chloride_fluid_table))
```



The barplot compares the incidence of hyperchloremia between patients whose dominant IV fluid was normal saline versus balanced fluids. It visually demonstrates that nearly all patients received normal saline, reflecting real-world DKA management practices. Among these patients, there is a higher count of hyperchloremia compared to normochloremia, suggesting that greater exposure to saline may contribute to elevated chloride levels. In contrast, the small number of patients receiving balanced fluids makes it difficult to draw meaningful comparisons between fluid types. This plot supports the decision to shift the analysis from fluid type comparisons to examining the relationship between saline volume and hyperchloremia, as the variation in saline administration is sufficient to explore potential dose-dependent effects on chloride derangements.

```
# Boxplot of Peak Chloride by Peak Chloride Category
boxplot(data$`Peak Chloride (mEq/L)` ~ data$`Peak Chloride Category (0 = normochloremia, 1 =
        main = "Peak Chloride Levels by Category",
        xlab = "Peak Chloride Category",
        ylab = "Peak Chloride (mEq/L)",
        ylim=c(50,150),
        col = c("lightblue", "orange"))
```



The boxplot of Peak Chloride by Category confirms that patients classified as hyperchloremic have substantially higher median and overall chloride levels than those with normochloremia, validating the outcome variable. Similarly, patients who developed AKI show higher median APACHE II Scores and Admission Creatinine, indicating that greater illness severity and poorer baseline kidney function are important predictors of AKI in this cohort.

It reveals several important features and anomalies in the dataset. The APACHE II Score distribution is notably right-skewed, with most patients clustered at lower severity but a small subset showing very high scores, reflecting a minority of critically ill patients who may disproportionately influence outcomes. Similarly, Admission Creatinine and Peak Chloride also show right-skewed distributions, with a few patients exhibiting markedly elevated values, suggesting baseline kidney dysfunction or significant electrolyte derangements beyond the typical range. The barplot of fluid type vs hyperchloremia highlights a key imbalance: nearly all patients received normal saline, while very few received balanced fluids, limiting fluid-type comparisons and narrowing the analysis to saline exposure. Finally, the boxplot of Peak Chloride by Category confirms that hyperchloremic patients had substantially higher chloride levels, validating the outcome definition but also underscoring variability driven by outliers. Together, these anomalies—skewed distributions, outliers, and treatment imbalances—emphasize the need for careful statistical modeling and sensitivity analyses to avoid biased interpretations.

8.PLANNED STATISTICAL METHODS

In addition to the primary analyses, we plan to explore further relationships using multivariable models and subgroup analyses. Linear regression may be applied to examine continuous outcomes such as hospital length of stay, adjusting for key clinical variables like APACHE II Score, Admission Creatinine, and total IV fluid volume. Logistic regression will be used for binary outcomes such as the development of AKI, allowing assessment of predictors including sepsis, hypotension, peak chloride levels, and saline exposure. We may also perform correlation analyses and stratified comparisons to better understand dose-response relationships, such as the impact of varying saline volumes on hyperchloremia risk. Finally, sensitivity analyses could be conducted to account for potential confounding or skewed distributions, ensuring robust and reliable conclusions.

9.LIMITATIONS

Most patients received only normal saline, limiting comparisons with balanced fluids and narrowing the research focus to the impact of saline volume rather than fluid type. Small subgroup sizes (e.g., sepsis, hypotension) reduce statistical power. Skewed distributions of key variables may affect model assumptions, and unmeasured confounders could bias results. Findings may have limited generalizability to other populations or settings.

10.References

Goad, N. (2018). Diabetic ketoacidosis and hyperchloremia full dataset [Data set]. Harvard Dataverse. https://doi.org/10.7910/DVN/PX9K2R

11. Additional Data Visualizations and Interpretations

```r
# Barplot: AKI vs Sepsis
sepsis_aki_table <- table(data$`Sepsis (0 = no, 1 = yes)`,
                          data$`Acute Kidney Injury In Hospital (0 = no, 1 yes)`)

barplot(sepsis_aki_table,
        beside = TRUE,
        col = c("lightgreen", "salmon"),
        main = "AKI by Sepsis Status",
        ylab = "Count",
        ylim = c(0, 100),
        legend.text = rownames(sepsis_aki_table))
```

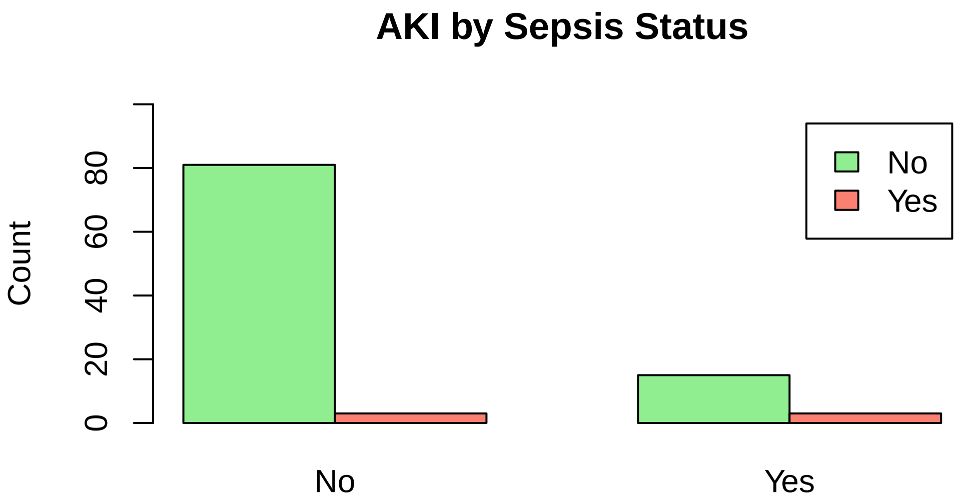


Figure A1.This bar plot shows the distribution of patients who did and did not develop AKI, categorized by whether they had sepsis on admission. The number of patients with sepsis is small, but a higher proportion of them appear to have developed AKI compared to the patients without sepsis. This suggests that sepsis is a strong contributing factor to AKI, which is consistent with clinical knowledge. The limited number of sepsis cases, however, may affect the statistical power of any formal analysis.

```
# Barplot: AKI vs Hypotension

hypotension_aki_table <- table(data$`Hypotension (0 = no, 1 = yes)`,
                               data$`Acute Kidney Injury In Hospital (0 = no, 1 yes)`)

barplot(hypotension_aki_table,
        beside = TRUE,
        col = c("lightyellow", "purple"),
        main = "AKI by Hypotension Status",
        ylab = "Count",
        ylim=c(0,100),
        legend.text = rownames(hypotension_aki_table))
```

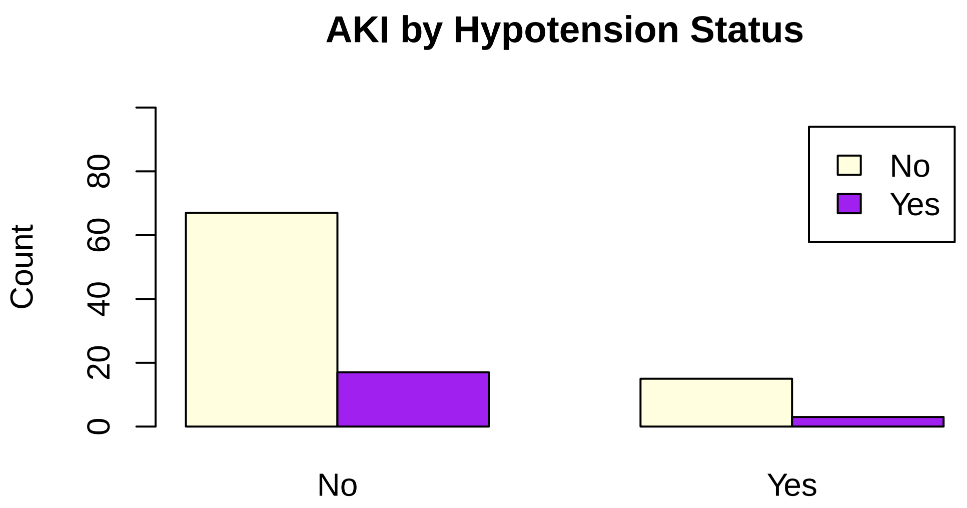


Figure A2.Similar to the sepsis plot, this bar plot compares AKI incidence between patients with and without hypotension. The data suggests that hypotension, like sepsis, is a significant risk factor for AKI, as a larger proportion of patients with hypotension developed kidney injury. This finding aligns with the well-established clinical relationship between poor blood flow (hypotension) and renal hypoperfusion.

```r
# Histogram for Admission Serum Creatinine
hist(data$`Admission Serum Creatinine (mg/dL)`,
     main = "Admission Creatinine Distribution",
     xlab = "Admission Serum Creatinine (mg/dL)",
     ylab = "Count",
     col = "lightgreen",
     border = "black")
```

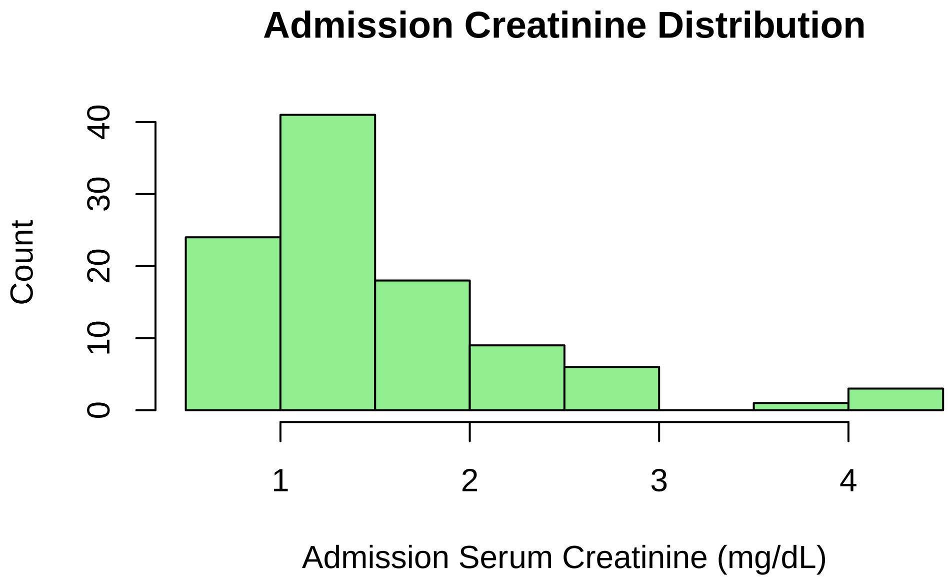


Figure A3:The histogram of admission creatinine levels shows a right-skewed distribution, indicating that most patients had creatinine values in the normal to low range (1.0-1.5 mg/dL). This is typical for diabetic ketoacidosis (DKA) patients. However, the presence of a long tail extending to high values (above 3.0 mg/dL) suggests that a small subset of patients had significantly elevated creatinine levels upon admission. This finding is critical because these patients likely had pre-existing kidney issues or severe illness, putting them at a higher risk for developing acute kidney injury (AKI) during their hospitalization. Therefore, admission creatinine is a vital predictor for AKI risk in this dataset.

```
boxplot(data$`Admission Serum Creatinine (mg/dL)` ~ data$`Acute Kidney Injury In Hospital (0
        main = "Admission Creatinine by AKI Status",
        xlab = "AKI Status",
        ylab = "Admission Serum Creatinine (mg/dL)",
        ylim=c(0,6),
        col = c("lightgreen", "salmon"))
```

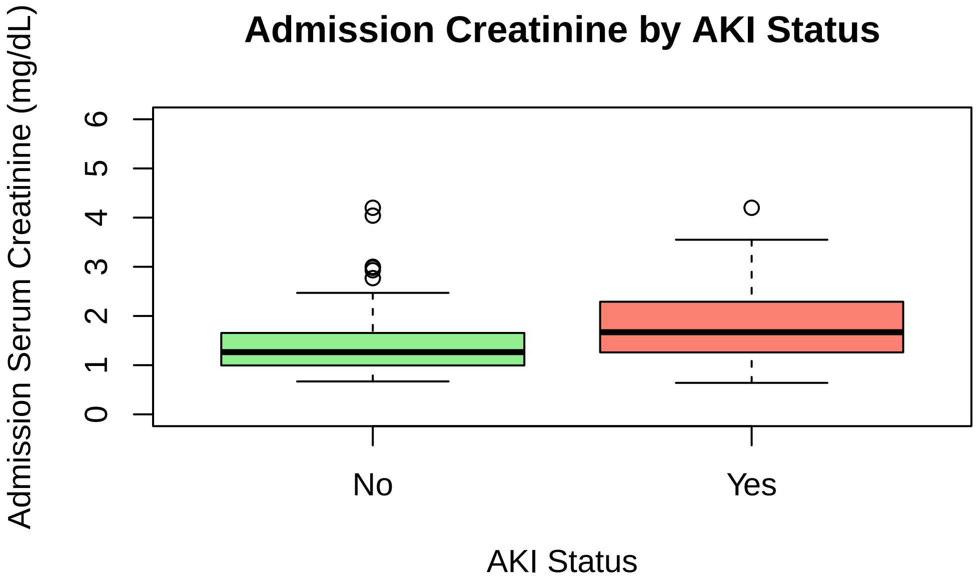
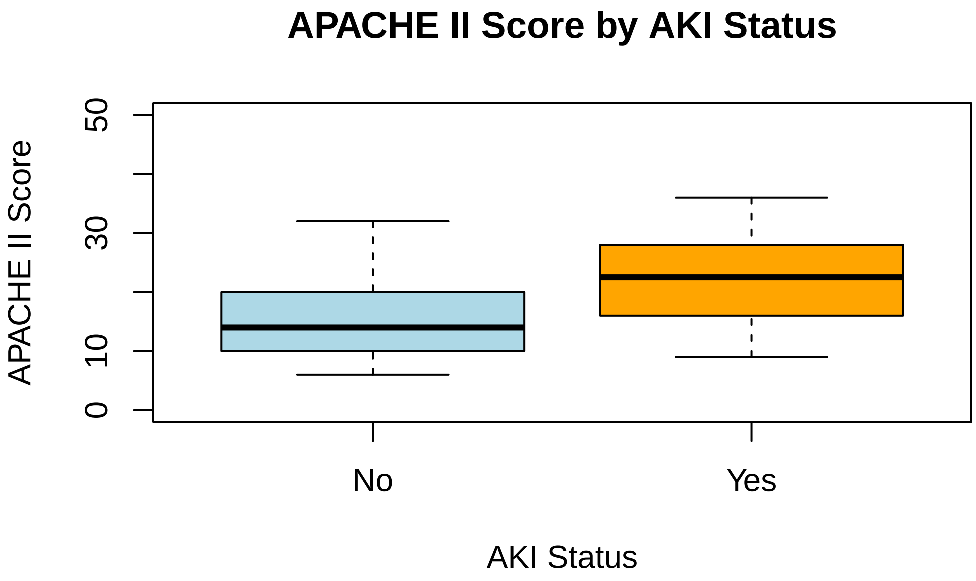


Figure A4.This boxplot visually confirms that patients who developed AKI during their hospitalization had a higher median admission serum creatinine than those who did not. The higher baseline creatinine level in the AKI group indicates that a degree of pre-existing kidney dysfunction or a higher degree of illness severity at the time of admission is a critical predictor for the development of AKI in DKA patients.

```
boxplot(data$`APACHE II Score` ~ data$`Acute Kidney Injury In Hospital (0 = no, 1 yes)`,
        main = "APACHE II Score by AKI Status",
        xlab = "AKI Status",
        ylab = "APACHE II Score",
        ylim=c(0,50),
        col = c("lightblue", "orange"))
```



FigureA5:This boxplot shows the distribution of APACHE II scores, a measure of illness severity, for both groups. Patients who developed AKI have a significantly higher median APACHE II score. This supports the hypothesis that a higher burden of illness on admission is a key predictor of AKI risk, as reflected by the score's components, which include physiological and demographic data. This plot validates the use of the APACHE II score as a crucial variable in the predictive models for AKI.

```
#Barplot: AKI vs Sepsis
sepsis_aki_table <- table(data$`Sepsis (0 = no, 1 = yes)`,
                          data$`Acute Kidney Injury In Hospital (0 = no, 1 yes)`)
barplot(sepsis_aki_table,
        beside = TRUE,
        col = c("lightgreen", "salmon"),
        main = "AKI by Sepsis Status",
        ylab = "Count",
        ylim = c(0, 100),
        legend.text = rownames(sepsis_aki_table))
```

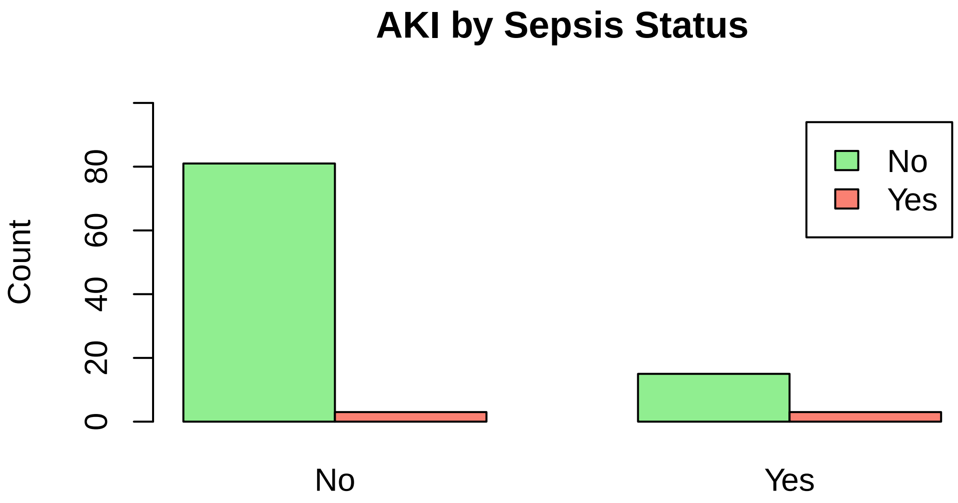


Figure A6.Patients with sepsis have a substantially higher proportion of AKI compared to non-septic patients, despite their smaller numbers. This indicates that sepsis is an important risk factor for AKI in this DKA cohort.

```
# Barplot: AKI vs Hypotension
hypotension_aki_table <- table(data$`Hypotension (0 = no, 1 = yes)`,
                               data$`Acute Kidney Injury In Hospital (0 = no, 1 yes)`)
barplot(hypotension_aki_table,
        beside = TRUE,
        col = c("lightyellow", "purple"),
        main = "AKI by Hypotension Status",
        ylab = "Count",
        ylim=c(0,100),
        legend.text = rownames(hypotension_aki_table))
```

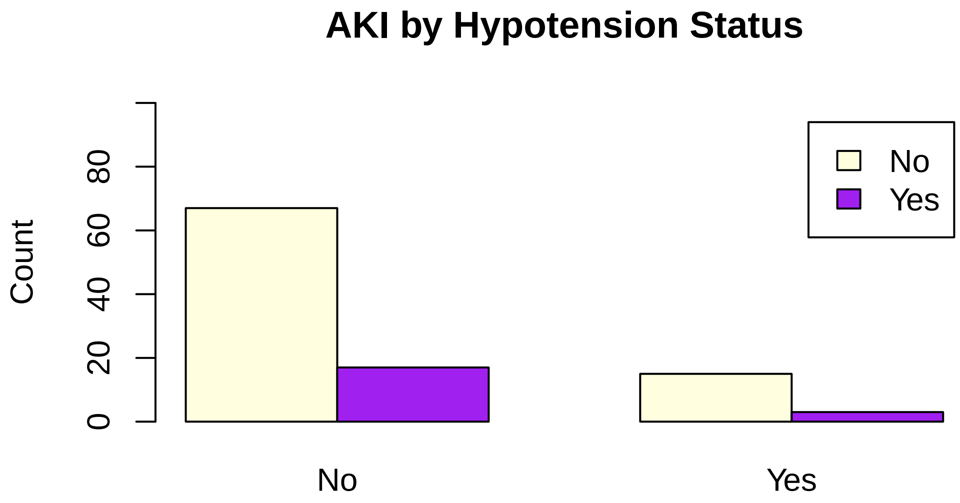


Figure A7. Patients who were hypotensive on admission show a markedly higher proportion of AKI than those without hypotension. This suggests that hypotension is a strong predictor of AKI development in this population.

Proposal 3

Link to dataset:

Dataset Link - Link

Download the data using the `original data` option that is available in the page. Shift the tar file to the data directory and unarchive the directory using the following command.

cd data tar xvf brca_tcga_pan_can_atlas_2018.tar.gz

Appendix

Introduction

The Breast Invasive Carcinoma dataset from TCGA PanCancer Atlas (2018) contains clinical information for 1,084 patients with breast cancer. The dataset includes demographic details such as age, sex, race, and ethnicity; clinical features including tumor stage, lymph node status, and subtype; and outcome data such as overall survival, disease-free survival, and progression-free survival. It offers a well-documented, de-identified resource that can be used to explore associations between clinical features and patient outcomes.

Dataset Justification

This dataset was selected because it meets the requirements for size, complexity, and biomedical relevance. With 38 variables and over 1,000 patients, it provides ample opportunities to apply descriptive statistics and visualization techniques. It includes both categorical and continuous variables, such as age, sex, subtype, and survival months. The dataset is ethically sourced and publicly available through cBioPortal, making it a strong candidate for meaningful analysis.

Research Questions

1. What is the age distribution of breast cancer patients, and does it differ by molecular subtype?
2. How do tumor stage and subtype relate to overall survival?
3. Are there differences in overall survival between patients who received radiation therapy and those who did not?

Data Wrangling and Cleanup

```r
# Load dataset
data <- read.delim("data/brca_tcga_pan_can_atlas_2018/data_clinical_patient.txt", comment.cha

# Inspect dataset
dim(data)
```

```
[1] 1084   38
```

```r
colnames(data)
```

```
 [1] "PATIENT_ID"
 [2] "SUBTYPE"
 [3] "CANCER_TYPE_ACRONYM"
 [4] "OTHER_PATIENT_ID"
 [5] "AGE"
 [6] "SEX"
 [7] "AJCC_PATHOLOGIC_TUMOR_STAGE"
 [8] "AJCC_STAGING_EDITION"
 [9] "DAYS_LAST_FOLLOWUP"
[10] "DAYS_TO_BIRTH"
[11] "DAYS_TO_INITIAL_PATHOLOGIC_DIAGNOSIS"
[12] "ETHNICITY"
[13] "FORM_COMPLETION_DATE"
[14] "HISTORY_NEOADJUVANT_TRTYN"
[15] "ICD_10"
[16] "ICD_O_3_HISTOLOGY"
[17] "ICD_O_3_SITE"
[18] "INFORMED_CONSENT_VERIFIED"
[19] "NEW_TUMOR_EVENT_AFTER_INITIAL_TREATMENT"
[20] "PATH_M_STAGE"
[21] "PATH_N_STAGE"
[22] "PATH_T_STAGE"
[23] "PERSON_NEOPLASM_CANCER_STATUS"
[24] "PRIMARY_LYMPH_NODE_PRESENTATION_ASSESSMENT"
[25] "PRIOR_DX"
[26] "RACE"
[27] "RADIATION_THERAPY"
[28] "WEIGHT"
```

```
[29] "IN_PANCANPATHWAYS_FREEZE"
[30] "OS_STATUS"
[31] "OS_MONTHS"
[32] "DSS_STATUS"
[33] "DSS_MONTHS"
[34] "DFS_STATUS"
[35] "DFS_MONTHS"
[36] "PFS_STATUS"
[37] "PFS_MONTHS"
[38] "GENETIC_ANCESTRY_LABEL"
```

```r
str(data)
```

```
'data.frame':   1084 obs. of  38 variables:
 $ PATIENT_ID                                : chr  "TCGA-3C-AAAU" "TCGA-3C-AALI" "TCGA-3C-A
 $ SUBTYPE                                   : chr  "BRCA_LumA" "BRCA_Her2" "BRCA_LumB" "BRC
 $ CANCER_TYPE_ACRONYM                       : chr  "BRCA" "BRCA" "BRCA" "BRCA" ...
 $ OTHER_PATIENT_ID                          : chr  "6E7D5EC6-A469-467C-B748-237353C23416" "
 $ AGE                                       : int  55 50 62 52 50 42 52 70 59 56 ...
 $ SEX                                       : chr  "Female" "Female" "Female" "Female" ...
 $ AJCC_PATHOLOGIC_TUMOR_STAGE               : chr  "STAGE X" "STAGE IIB" "STAGE IIB" "STAGE
 $ AJCC_STAGING_EDITION                      : chr  "6TH" "6TH" "7TH" "7TH" ...
 $ DAYS_LAST_FOLLOWUP                        : int  4047 4005 1474 1448 348 1477 303 259 437
 $ DAYS_TO_BIRTH                             : int  -20211 -18538 -22848 -19074 -18371 -1539
 $ DAYS_TO_INITIAL_PATHOLOGIC_DIAGNOSIS      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ ETHNICITY                                 : chr  "Not Hispanic Or Latino" "Not Hispanic O
 $ FORM_COMPLETION_DATE                      : chr  "1/13/14" "7/28/14" "7/28/14" "7/28/14"
 $ HISTORY_NEOADJUVANT_TRTYN                 : chr  "No" "No" "No" "No" ...
 $ ICD_10                                    : chr  "C50.9" "C50.9" "C50.9" "C50.9" ...
 $ ICD_O_3_HISTOLOGY                         : chr  "8520/3" "8500/3" "8500/3" "8500/3" ...
 $ ICD_O_3_SITE                              : chr  "C50.9" "C50.9" "C50.9" "C50.9" ...
 $ INFORMED_CONSENT_VERIFIED                 : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ NEW_TUMOR_EVENT_AFTER_INITIAL_TREATMENT   : chr  "No" "No" "No" "No" ...
 $ PATH_M_STAGE                              : chr  "MX" "M0" "M0" "M0" ...
 $ PATH_N_STAGE                              : chr  "NX" "N1A" "N1A" "N0 (I+)" ...
 $ PATH_T_STAGE                              : chr  "TX" "T2" "T2" "T1C" ...
 $ PERSON_NEOPLASM_CANCER_STATUS             : chr  "With Tumor" "Tumor Free" "Tumor Free" "
 $ PRIMARY_LYMPH_NODE_PRESENTATION_ASSESSMENT: chr  "Yes" "Yes" "Yes" "Yes" ...
 $ PRIOR_DX                                  : chr  "No" "No" "No" "No" ...
 $ RACE                                      : chr  "White" "Black or African American" "Bla
 $ RADIATION_THERAPY                         : chr  "No" "Yes" "No" "No" ...
 $ WEIGHT                                    : logi  NA NA NA NA NA NA ...
 $ IN_PANCANPATHWAYS_FREEZE                  : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ OS_STATUS                                 : chr  "0:LIVING" "0:LIVING" "0:LIVING" "0:LIVI
 $ OS_MONTHS                                 : num  133.1 131.7 48.5 47.6 11.4 ...
 $ DSS_STATUS                                : chr  "0:ALIVE OR DEAD TUMOR FREE" "0:ALIVE OR
 $ DSS_MONTHS                                : num  133.1 131.7 48.5 47.6 11.4 ...
 $ DFS_STATUS                                : chr  "1:Recurred/Progressed" "0:DiseaseFree"
 $ DFS_MONTHS                                : num  59.4 131.7 48.5 NA 11.4 ...
 $ PFS_STATUS                                : chr  "1:PROGRESSION" "0:CENSORED" "0:CENSORED"
 $ PFS_MONTHS                                : num  59.4 131.7 48.5 47.6 11.4 ...
 $ GENETIC_ANCESTRY_LABEL                    : chr  "EUR" "AFR" "AFR_ADMIX" "AFR" ...
```

```r
# Convert categorical variables into factors
data$SEX <- as.factor(data$SEX)
data$SUBTYPE <- as.factor(data$SUBTYPE)
data$AJCC_PATHOLOGIC_TUMOR_STAGE <- as.factor(data$AJCC_PATHOLOGIC_TUMOR_STAGE)
data$RADIATION_THERAPY <- as.factor(data$RADIATION_THERAPY)
data$OS_STATUS <- as.factor(data$OS_STATUS)

# Check for missing values
colSums(is.na(data))
```

```
                                PATIENT_ID
                                         0
                                   SUBTYPE
                                         0
                       CANCER_TYPE_ACRONYM
                                         0
                          OTHER_PATIENT_ID
                                         0
                                       AGE
                                         0
                                       SEX
                                         0
               AJCC_PATHOLOGIC_TUMOR_STAGE
                                         0
                      AJCC_STAGING_EDITION
                                         0
                        DAYS_LAST_FOLLOWUP
                                       104
                             DAYS_TO_BIRTH
                                        15
      DAYS_TO_INITIAL_PATHOLOGIC_DIAGNOSIS
```

```
                                         0
                                 ETHNICITY
                                         0
                      FORM_COMPLETION_DATE
                                         0
                 HISTORY_NEOADJUVANT_TRTYN
                                         0
                                    ICD_10
                                         0
                         ICD_O_3_HISTOLOGY
                                         0
                              ICD_O_3_SITE
                                         0
                 INFORMED_CONSENT_VERIFIED
                                         0
   NEW_TUMOR_EVENT_AFTER_INITIAL_TREATMENT
                                         0
                              PATH_M_STAGE
                                         0
                              PATH_N_STAGE
                                         0
                              PATH_T_STAGE
                                         0
             PERSON_NEOPLASM_CANCER_STATUS
                                         0
PRIMARY_LYMPH_NODE_PRESENTATION_ASSESSMENT
                                         0
                                  PRIOR_DX
                                         0
                                      RACE
                                         0
                         RADIATION_THERAPY
                                         0
                                    WEIGHT
                                      1084
                  IN_PANCANPATHWAYS_FREEZE
                                         0
                                 OS_STATUS
                                         0
                                 OS_MONTHS
                                         0
                                DSS_STATUS
                                         0
                                DSS_MONTHS
                                         2
                                DFS_STATUS
                                         0
                                DFS_MONTHS
                                       143
                                PFS_STATUS
                                         0
                                PFS_MONTHS
                                         2
                    GENETIC_ANCESTRY_LABEL
                                         0
```

```
# Summary of numeric variables
summary(data$AGE)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  26.00   49.00   58.00   58.42   67.00   90.00 
```

```
summary(data$OS_MONTHS)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00   14.72   27.01   40.83   55.00  282.90 
```

Cleanup summary

The dataset contains 1,084 patients and 38 variables. Some survival fields contain missing values (e.g., DFS months), but overall the dataset is well-structured. Categorical variables were converted to factors for analysis.

Exploratory Data Analysis

Descriptive Statistics

```
# Age distribution
summary(data$AGE)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  26.00   49.00   58.00   58.42   67.00   90.00
```

```r
# Sex distribution
table(data$SEX)
```

```

Female   Male
  1072     12
```

```r
# Subtype distribution
table(data$SUBTYPE)
```

```

             BRCA_Basal   BRCA_Her2   BRCA_LumA   BRCA_LumB BRCA_Normal
        103         171          78         499         197          36
```

```r
# Tumor stage distribution
table(data$AJCC_PATHOLOGIC_TUMOR_STAGE)
```

```

              STAGE I   STAGE IA   STAGE IB   STAGE II  STAGE IIA  STAGE IIB
         5         89         86          6          6        355        255
 STAGE III STAGE IIIA STAGE IIIB STAGE IIIC   STAGE IV    STAGE X
         2        155         28         64         19         14
```

```r
# Overall survival
summary(data$OS_MONTHS)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   14.72   27.01   40.83   55.00  282.90
```

```r
table(data$OS_STATUS)
```

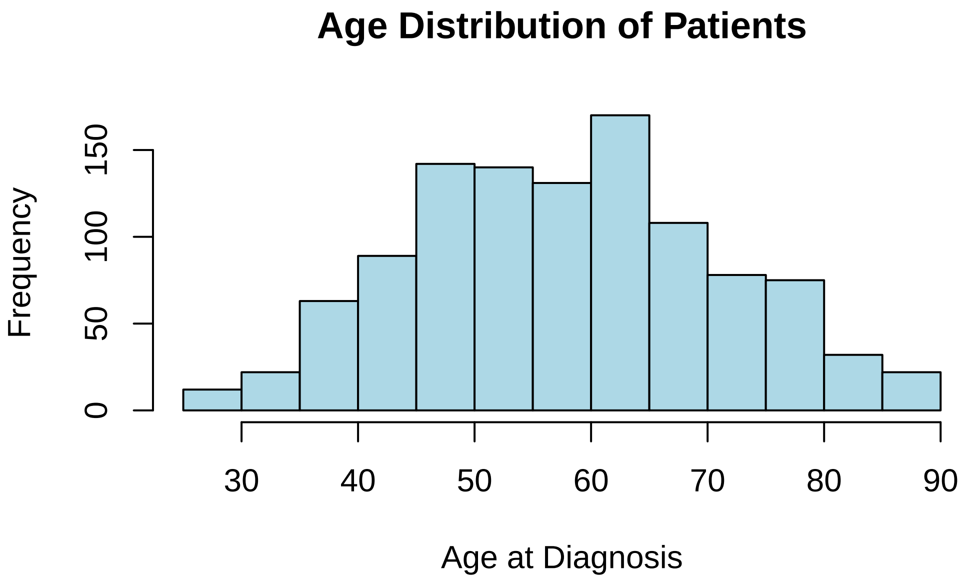
```

  0:LIVING 1:DECEASED
       933        151
```

Summary:

The median age at diagnosis was around 58 years, with patients ranging from 26 to 90 years. The majority of patients were female, with only a small number of males. The most common subtypes were Luminal A and Luminal B. Tumor stages varied, with many patients diagnosed at Stage II. Median overall survival was about 27 months, with survival times ranging up to nearly 283 months.

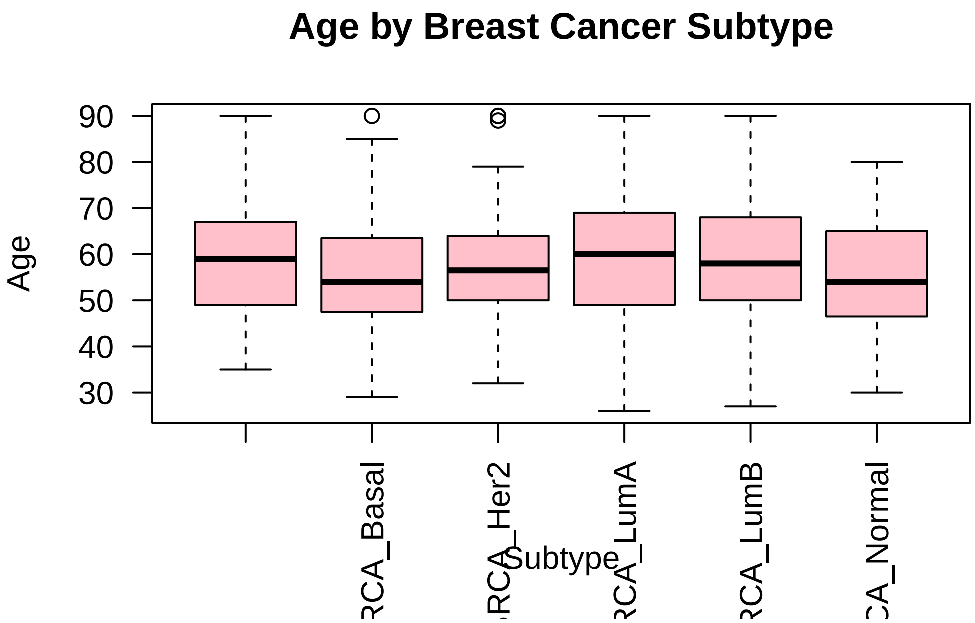
Visualizations

```r
hist(data$AGE,
     main="Age Distribution of Patients",
     xlab="Age at Diagnosis",
     col="lightblue", border="black")
```



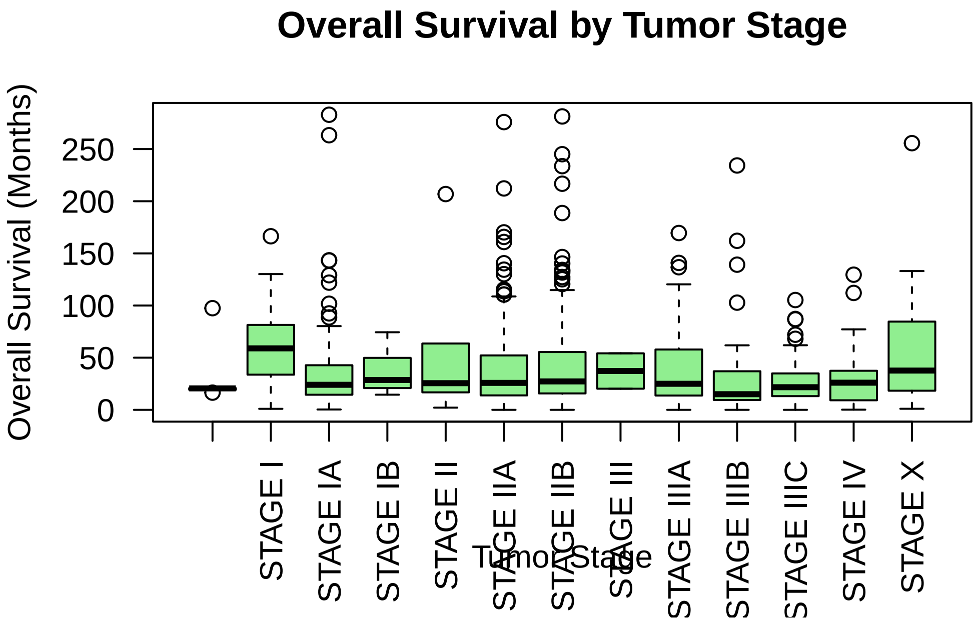
Age by Subtype

```
boxplot(AGE ~ SUBTYPE, data=data,
        main="Age by Breast Cancer Subtype",
        xlab="Subtype", ylab="Age",
        col="pink", las=2)
```



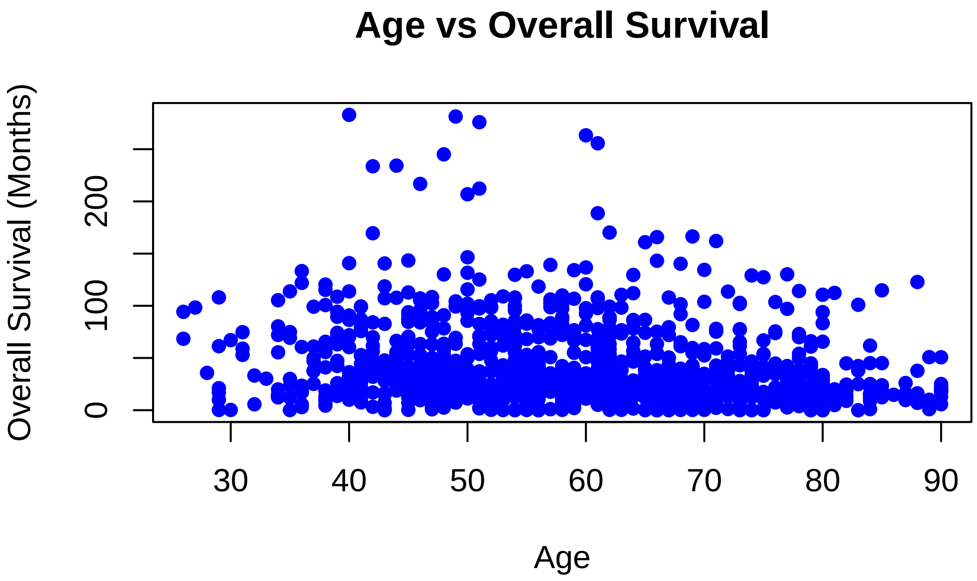
Survival by Tumor Stage

```
boxplot(OS_MONTHS ~ AJCC_PATHOLOGIC_TUMOR_STAGE, data=data,
        main="Overall Survival by Tumor Stage",
        xlab="Tumor Stage", ylab="Overall Survival (Months)",
        col="lightgreen", las=2)
```



Age vs Overall Survival

```
plot(data$AGE, data$OS_MONTHS,
     main="Age vs Overall Survival",
     xlab="Age", ylab="Overall Survival (Months)",
     pch=16, col="blue")
```



Planned Statistical Methods

Planned analyses include chi-square tests to examine associations between categorical variables such as subtype and tumor stage, and t-tests to compare survival outcomes between groups such as radiation vs non-radiation patients. Correlation and regression methods may also be used to evaluate relationships between age and survival outcomes.

Limitations

The dataset is cross-sectional and lacks detailed treatment history beyond radiation and neoadjuvant therapy. Some survival data are incomplete, particularly for disease-free and progression-free survival. The dataset is also biased toward female patients, with very few male cases available for comparison.

References

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A. P., Sander, C., & Schultz, N. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer discovery, 2(5), 401–404. https://doi.org/10.1158/2159-8290.CD-12-0095