# Week 5 lab BST 210

Daniel Herrera, Willow Duffell, Lauren Mock

10/9/2021

***Group 4 Members:*** Daniel Herrera
Willow Duffell
Lauren Mock

## Question 1

### General area/domain/subject area

The general subject area of this project is bone marrow transplant outcomes in children with hematologic diseases.

## Question 2

### Dataset and source

We will use data obtained from the UCI Machine Learning Repository: Center for Machine Learning and Intelligent Systems that was originally collected by researchers in the Faculty of Automatic Control, Electronics and Computer Science at Silesian University of Technology in Poland. This data set includes children with hematologic diseases, such as leukemia, who underwent bone marrow transplants. It includes information on the donor, the recipient, and the recipient's response to the treatment.

The source of this data set can be found here: https://archive.ics.uci.edu/ml/datasets/Bone+marrow+transplant%3A+children

And a downloadable version of the data set can be found here: https://www.kaggle.com/adamgudys/bone-marrow-transplant-children

## Question 3

### Primary questions

Our primary goals are to predict (before transplantation) whether or not a patient will survive, and to understand which factors are the most important in determining survival. We are particularly interested in determining the importance of the closeness of the match between the recipient and the donor for survival.

## Question 4

### Secondary questions

There is a lot of controversy around the source of hematopoietic stem cells used in transplantation (either peripheral blood (PBSC) or bone marrow (BM)). It would be interesting to see if there is an association between these two stem-cell sources and the likelihood of survival, and we will definitely need to be careful with confounding when addressing this question.

## Question 5

### Outcome(s)/endpoint(s)

Our primary outcome is survival status at the end of the follow-up period, which is a binary variable. We are interested in predicting a patient's probability of survival, as well as understanding the relative importance of factors that influence survival.

## Question 6

### Draft Statistical Analysis Plan

*I. Explain important variables (using "domain" knowledge)*

a. Identify potential confounders and effect measure modifiers (EMM)

b. Will require an understanding of how our variables are associated

*II. Exploratory Data Analysis (EDA using above)*

a. Visualizations (scatterplots - with jitter since survival 0-1, correlation matrices)

b. Visualize confounding and EMM Summary statistics

*III. Candidate Variables*

a. Choose best variables to use to predict survival

*IV. Model Selection*

a. Primary Question

Prediction of survival

1) Goal is to predict survival at end of follow-up (binary variable)
2) Does not need to account for confounding, because we only want the best possible prediction of survival
3) Perform logistic regression (binary outcome) with all variables that may be important
    - Trim model - remove unnecessary variables to avoid unnecessary complexity/overfitting
    - Backwards removal using adjusted R^2 (since we are more concerned with maximizing our predictive ability–the amount of variability in y explained by our covariates)
4) Explore K-means as a non-parametric method of predicting survival

Determining Risk Factors for Mortality (Association)

1) Goal is to determine the most important risk factors in determining whether or not a patient survives
    - This will only include data that is collected before transplantation
2) Must account for confounding, because we care about the strength of the relationship between each variable and survival

3) Perform logistic regression with all variables that may be important

- Trim model - remove unnecessary variables to avoid unnecessary complexity/overfitting
- Backwards removal using adjusted R^2 (since we are more concerned with maximizing our predictive ability–the amount of variability in y explained by our covariates)

b. Secondary Question–Source of hematopoietic stem cells

1) Goal is to determine association between the two stem-cell source methods (peripheral blood and bone marrow) and their outcome of survival, presumably using the relevant covariates discovered in our model above
2) Confounders will be important here, since we don't yet know which type of patients receive which treatment
3) Using Association model results from our primary question, add variable stem_cell_source and run an ANOVA test to determine which source has a better outcome of survival

*V. Check for multicollinearity using variance inflation factors*

a. We expect many of our variables to be highly correlated

*VI. Validate model using Cross-Validation, where appropriate*

a. Samples of the training/testing data

b. How well would the prediction model perform on other data?

*VII. Run sensitivity analysis using Ridge, LASSO, or Elastic Net*

## Question 7

**Biggest challenges foreseen**

One of the main challenges we foresee in answering our proposed questions is our lack of subject matter knowledge. In addressing which risk factors are most important in determining survival, we will need to have a comprehensive understanding of the potential confounders in our data set to ensure that the coefficients in our models are meaningful. In our secondary analysis, we will need to understand why a patient may be selected for a bone marrow transplant vs. a peripheral blood transplant, which will require a domain expert. We will also need to account for the many variable types included in our data, including numeric, categorical, and indicator.

## Question 8

**Domain expertise sought**

We will need to find a domain expert who has experience with hematologic diseases or bone marrow transplants in order to better understand our dataset and its variables. This will be especially important in identifying potential confounders, which cannot be identified through any statistical process alone. This will also give us some insight into possible effect measure modifiers. For now, we have discussed our dataset with Daniel's fiancee, who is a doctor, and Willow's mom, who is a nurse practitioner. Moving forward, we will be seeking additional expertise.

## Question 9

**What software package(s) will you use to complete this project?**

We will all be using R.

## Question 10

**Complete an initial round of exploratory analyses on your data that would be relevant to your plan and responses above, and include any plots, summaries, code and output. Please include exploratory analysis for outcome(s) of continuous form however/wherever possible even if your ultimate goals/questions involve a different form of outcome data such as binary, polytomous, etc. (You may consider this initial analysis as a potential sub-analysis later on.)**

### Explore Missing Values

Before we begin with any visualizations and summary statistics, we would like to explore missing-ness within our data set.

```
# lets look at a table of all missing values
colSums(is.na(bone))
```

```
##                       donor_age              donor_age_below_35
##                               0                               0
##                       donor_ABO                       donor_CMV
##                               0                               0
##                   recipient_age          recipient_age_below_10
##                               0                               0
##               recipient_age_int                 recipient_gender
##                               0                               0
##             recipient_body_mass                    recipient_ABO
##                               0                               0
##                    recipient_rh                    recipient_CMV
##                               0                               0
##                         disease                   disease_group
##                               0                               0
##                    gender_match                       ABO_match
##                               0                               0
##                      CMV_status           HLA_match..out.of.10.
##                               0                               0
##                   HLA_match_raw                    HLA_mismatch
##                               0                               0
##                         antigen                          allele
##                               0                               0
##                     HLA_group_1                     risk_group
##                               0                               0
##                stem_cell_source                  tx_post_relapse
##                               0                               0
## CD34_x1e6_per_kg...CD34kgx10d6                 CD3_x1e8_per_kg
##                               0                               0
##                CD3_to_CD34_ratio                   ANC_recovery
##                               0                               0
##                     PLT_recovery             acute_GvHD_II_III_IV
```

```
##                              0                              0
##              acute_GvHD_III_IV     time_to_acute_GvHD_III_IV
##                              0                              0
##         extensive_chronic_GvHD                        relapse
##                              0                              0
##                  survival_time                survival_status
##                              0                              0
```

Great, no missing data (spoiler: there is). Oh wait, there are "?" marks that we should examine. We see that there are actually a bit of "?" in our data set, so this may need to be addressed later, particularly for variables such as recipient CMV (14 ?'s), CMV_status (16) and extensive_chronic_GvHD (31 ?'s). For now, we will replace the "?"s in the data set.

```
# let us see the number of ?'s that appear
sort(colSums(bone == "?"), decreasing = TRUE) [1:10]
```

```
## extensive_chronic_GvHD             CMV_status          recipient_CMV
##                     31                     16                     14
##        CD3_x1e8_per_kg      CD3_to_CD34_ratio              donor_CMV
##                      5                      5                      2
##    recipient_body_mass           recipient_rh          recipient_ABO
##                      2                      2                      1
##              ABO_match
##                      1
```

```
# replace ?'s with true NA's
bone <- mutate_all(bone, ~replace(., . == "?", NA))

# Now we can see all of the ?'s have been replaced with NA values that are picked up
sort(colSums(is.na(bone)), decreasing = TRUE) [1:10]
```

```
## extensive_chronic_GvHD             CMV_status          recipient_CMV
##                     31                     16                     14
##        CD3_x1e8_per_kg      CD3_to_CD34_ratio              donor_CMV
##                      5                      5                      2
##    recipient_body_mass           recipient_rh          recipient_ABO
##                      2                      2                      1
##              ABO_match
##                      1
```

Well, it turns out there are some other problematic values. Some time-to-event variables have entries of 1,000,000, which will be challenging to handle in our EDA.

To illustrate, let us see the number of occurrences of the value 1000000. This value frequently occurs in the data amongst three variables. It seems as though these large values indicate that the event never occurred for that subject, but we will need to confirm this. If our assumption is correct, then we will need to seek out advice on how to proceed.

For now, we will leave these values as is. We don't want to replace them with NA, because they do not appear to be missing values.

```
# let us see the number of ?'s that appear
sort(colSums(bone == "1000000"), decreasing = TRUE) [1:3]
```

```
## time_to_acute_GvHD_III_IV              PLT_recovery              ANC_recovery
##                       145                        17                         5
```

```
# replace with NA
# bone <- mutate_all(bone, ~replace(., . == "1000000", NA))
```

Let's also make sure our true missing values do not exceed 5% of each column.

```
# look of missing values as % of column
missing_vals <- bone %>%
  is.na() %>%
  colSums() %>%
  `/`(nrow(bone)) %>%
  `*`(100)

missing_vals[missing_vals >= 5] %>% names()
```

```
## [1] "recipient_CMV"          "CMV_status"             "extensive_chronic_GvHD"
```

Only these three columns have $> 5\%$ missing values. We will need to investigate these missing values later if we choose to use these variables in our analysis.

### Basic Data Exploration

Now we can explore some more standard EDA. Let's familiarize ourselves with the data set by briefly viewing all the columns, column types, a brief amount of data. We can then explore summary statistics. Note: we will have to change some of our data from character data to numeric data.

**Numerical Variables:** donor_age, recipient_age, recipient_body_mass, CD34_x1e6_per_kg. . . CD34kgx10d6, CD3_x1e8_per_kg, CD3_to_CD34_ratio, ANC_recovery, PLT_recovery, time_to_acute_GvHD_III_IV, survival_time

**Categorical Variables:** donor_ABO, recipient_age_int, recipient_ABO, disease, CMV_status, HLA_match..out.of.10., antigen, allele, HLA_group_1

**Binary Variables:** donor_age_below_35 (yes/no), donor_CMV (present/absent), recipient_age_below_10 (yes/no), recipient_rh (plus/minus), recipient_CMV (present/absent), disease_group (malignant/nonmalignant), gender_match(female_to_male/other), ABO_match (matched/mismatched), HLA_mismatch (matched, mismatched), risk_group (high/low), stem_cell_source (peripheral_blood/bone_marrow),tx_post_relapse (yes/no), acute_GvHD_II_III_IV (yes/no), acute_GvHD_III_IV (yes/no), extensive_chronic_GvHD (yes/no), relapse (yes/no), survival_status (yes,no)

```
glimpse(bone)
```

```
## Rows: 187
## Columns: 38
## $ donor_age                          <dbl> 22.83014, 23.34247, 26.39452, 39.68493,~
## $ donor_age_below_35                 <chr> "yes", "yes", "yes", "no", "yes", "yes"~
## $ donor_ABO                          <chr> "A", "B", "B", "A", "A", "AB", "0", "0"~
## $ donor_CMV                          <chr> "present", "absent", "absent", "present~
## $ recipient_age                      <dbl> 9.6, 4.0, 6.6, 18.1, 1.3, 8.9, 14.4, 18~
## $ recipient_age_below_10             <chr> "yes", "yes", "yes", "no", "yes", "yes"~
## $ recipient_age_int                  <chr> "5_10", "0_5", "5_10", "10_20", "0_5", ~
## $ recipient_gender                   <chr> "male", "male", "male", "female", "fema~
## $ recipient_body_mass                <chr> "35", "20.6", "23.4", "50", "9", "40", ~
## $ recipient_ABO                      <chr> "A", "B", "B", "AB", "AB", "0", "A", "A~
## $ recipient_rh                       <chr> "plus", "plus", "plus", "plus", "minus"~
## $ recipient_CMV                      <chr> "present", "absent", "present", "absent~
## $ disease                            <chr> "ALL", "ALL", "ALL", "AML", "chronic", ~
## $ disease_group                      <chr> "malignant", "malignant", "malignant", ~
## $ gender_match                       <chr> "other", "other", "other", "other", "ot~
## $ ABO_match                          <chr> "matched", "matched", "matched", "misma~
## $ CMV_status                         <chr> "3", "0", "2", "1", "0", NA, NA, "1", "~
## $ HLA_match..out.of.10.              <dbl> 10, 10, 10, 10, 9, 10, 10, 7, 10, 9, 9,~
## $ HLA_match_raw                      <chr> "10-Oct", "10-Oct", "10-Oct", "10-Oct",~
## $ HLA_mismatch                       <chr> "matched", "matched", "matched", "match~
## $ antigen                            <chr> "0", "0", "0", "0", "2", "0", "0", "2",~
## $ allele                             <chr> "0", "0", "0", "0", "1", "0", "0", "3",~
## $ HLA_group_1                        <chr> "matched", "matched", "matched", "match~
## $ risk_group                         <chr> "high", "low", "low", "low", "high", "h~
## $ stem_cell_source                   <chr> "peripheral_blood", "bone_marrow", "bon~
## $ tx_post_relapse                    <chr> "no", "no", "no", "no", "no", "yes", "n~
## $ CD34_x1e6_per_kg...CD34kgx10d6     <dbl> 7.20, 4.50, 7.94, 4.25, 51.85, 3.27, 17~
## $ CD3_x1e8_per_kg                    <chr> "5.38", "0.41", "0.42", "0.14", "13.05"~
## $ CD3_to_CD34_ratio                  <chr> "1.33876", "11.078295", "19.01323", "29~
## $ ANC_recovery                       <int> 19, 16, 23, 23, 14, 16, 17, 22, 15, 16,~
## $ PLT_recovery                       <int> 51, 37, 20, 29, 14, 70, 29, 58, 14, 17,~
## $ acute_GvHD_II_III_IV               <chr> "yes", "yes", "yes", "yes", "no", "no",~
## $ acute_GvHD_III_IV                  <chr> "yes", "no", "no", "yes", "no", "no", "~
## $ time_to_acute_GvHD_III_IV          <int> 32, 1000000, 1000000, 19, 1000000, 1000~
## $ extensive_chronic_GvHD             <chr> "no", "no", "no", NA, "no", "no", NA, N~
## $ relapse                            <chr> "no", "yes", "yes", "no", "no", "no", "~
## $ survival_time                      <int> 999, 163, 435, 53, 2043, 2800, 41, 45, ~
## $ survival_status                    <int> 0, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, ~
```

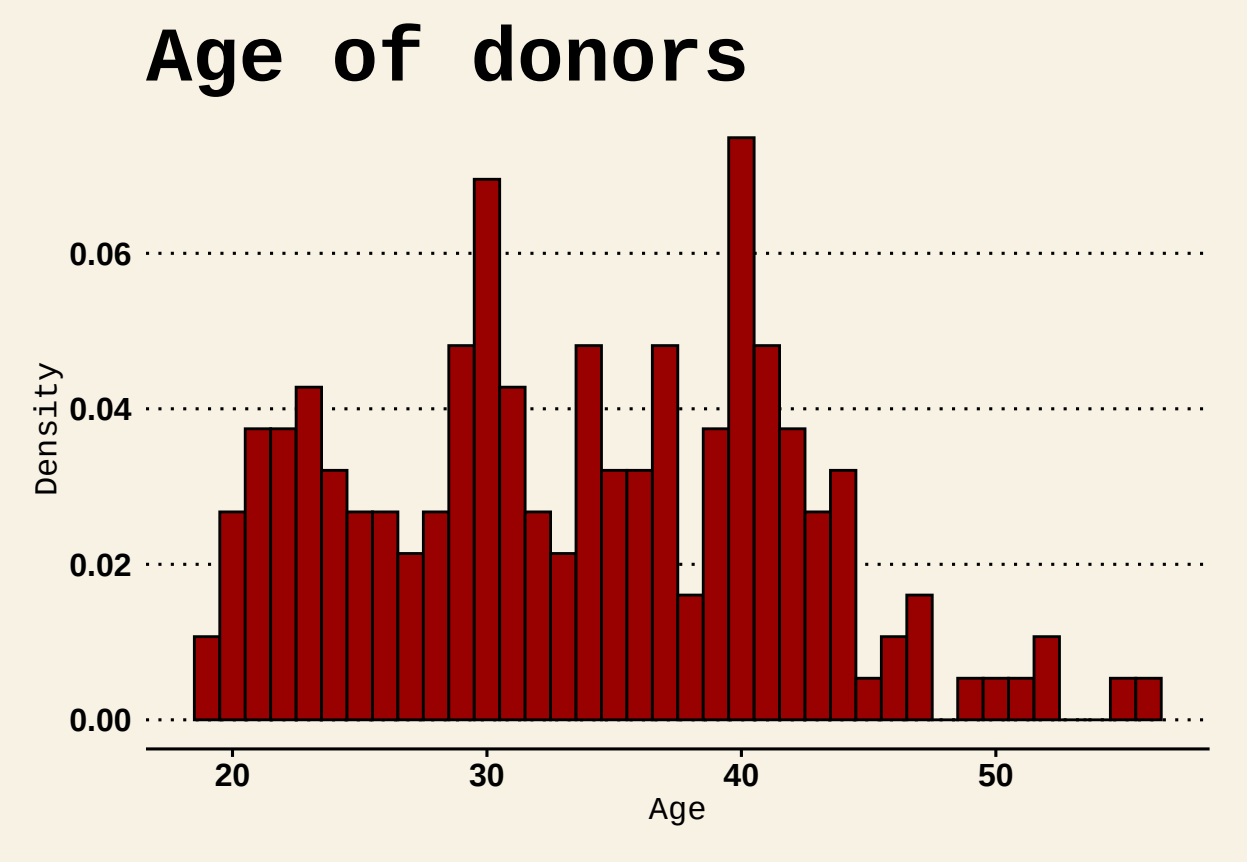
**Explore donor data**

```
summary(bone$donor_age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.65   27.04   33.55   33.47   40.12   55.55
```

```
# plot histogram of age distribution
ggplot(bone) +
  geom_histogram(aes(x = donor_age, y = ..density..), fill = "#990000", col = "black", binwidth = 1) +
  ggtitle("Age of donors", ) +
```

```
  xlab("Age") +
  ylab("Density") +
  theme_wsj()+ theme(axis.title=element_text(size=12))
```




```
# summary stats for donor blood type
bone %>%
  group_by(donor_ABO) %>%
  summarise(n = length(donor_ABO),
            proportion = length(donor_ABO)/nrow(bone))
```

```
## # A tibble: 4 x 3
##   donor_ABO     n proportion
##   <chr>     <int>      <dbl>
## 1 O            73     0.390
## 2 A            71     0.380
## 3 AB           15     0.0802
## 4 B            28     0.150
```

```
# summary stats for donor cytomegalovirus status
bone %>%
  group_by(donor_CMV) %>%
  summarise(n = length(donor_CMV),
            proportion = length(donor_CMV)/nrow(bone))
```

```
## # A tibble: 3 x 3
```

```
##   donor_CMV     n proportion
##   <chr>     <int>      <dbl>
## 1 absent      113     0.604
## 2 present      72     0.385
## 3 <NA>          2     0.0107
```

So far we can see the following:

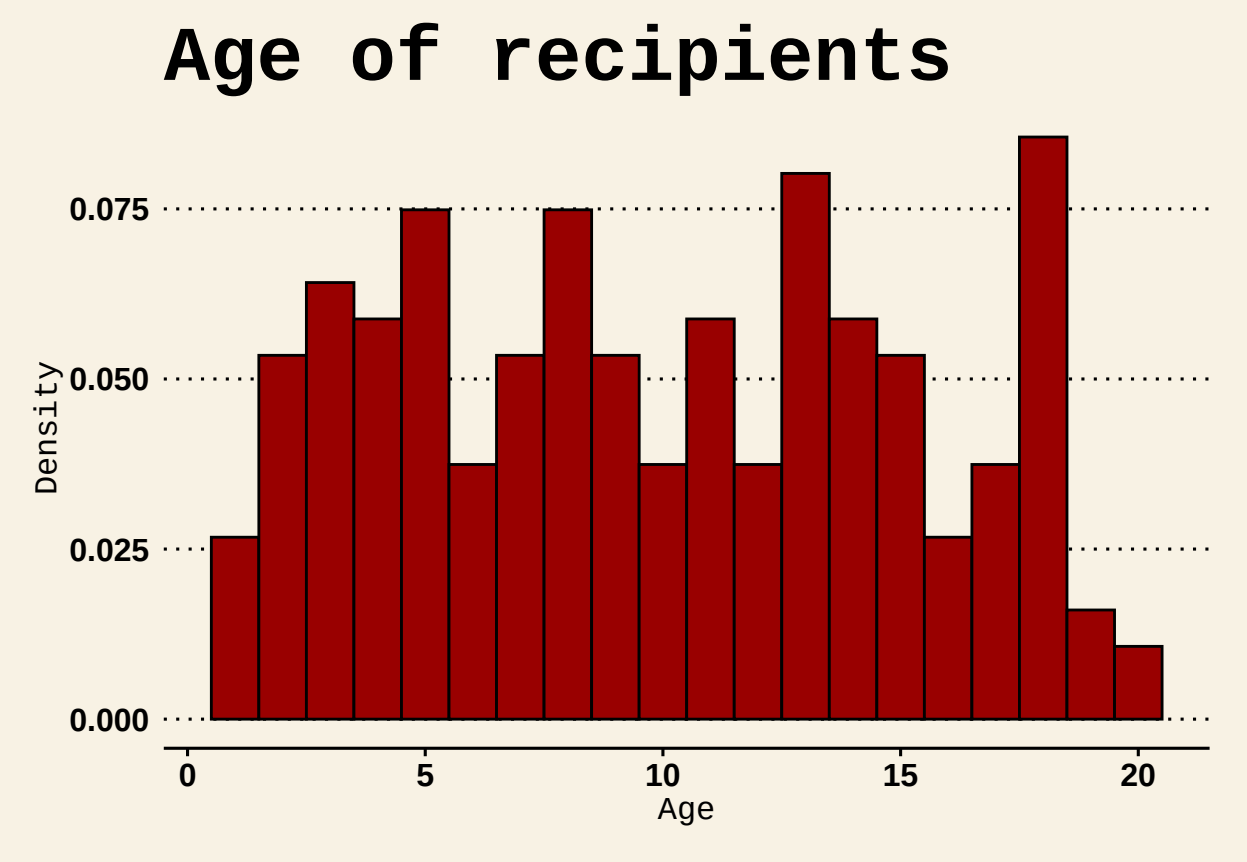
1) Most of our donors are between 20 and 50 years old
2) Donors' blood types are as follows: type O (39.0%), type A (38.0%), type AB (8.0%), B (15.0%)
3) Donors' cytomegalovirus status follows: absent (60.4%), present (38.5%), missing (0.1%)

**Explore recipient data**

```
summary(bone$recipient_age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.600   5.050   9.600   9.932  14.050  20.200
```

```
# plot histogram of age distribution
bone %>%
ggplot() +
  geom_histogram(aes(x = recipient_age, y=..density..), fill = "#990000", col = "black", binwidth = 1) +
  ggtitle("Age of recipients" ) +
  xlab("Age") +
  ylab("Density") +
  theme_wsj()+ theme(axis.title=element_text(size=12))
```




```
# summary stats for blood type
bone %>%
  group_by(recipient_ABO) %>%
  summarise(n = length(recipient_ABO),
            proportion = length(recipient_ABO)/nrow(bone))
```

```
## # A tibble: 5 x 3
##   recipient_ABO     n proportion
##   <chr>         <int>      <dbl>
## 1 0                48    0.257
## 2 A                75    0.401
## 3 AB               13    0.0695
## 4 B                50    0.267
## 5 <NA>              1    0.00535
```

```
# summary stats for recipient cytomegalovirus status
bone %>%
  group_by(recipient_CMV) %>%
  summarise(n = length(recipient_CMV),
            proportion = length(recipient_CMV)/nrow(bone))
```

```
## # A tibble: 3 x 3
##   recipient_CMV     n proportion
##   <chr>         <int>      <dbl>
## 1 absent           73     0.390
```

```
## 2 present          100     0.535
## 3 <NA>              14     0.0749
```

So far we can see the following:

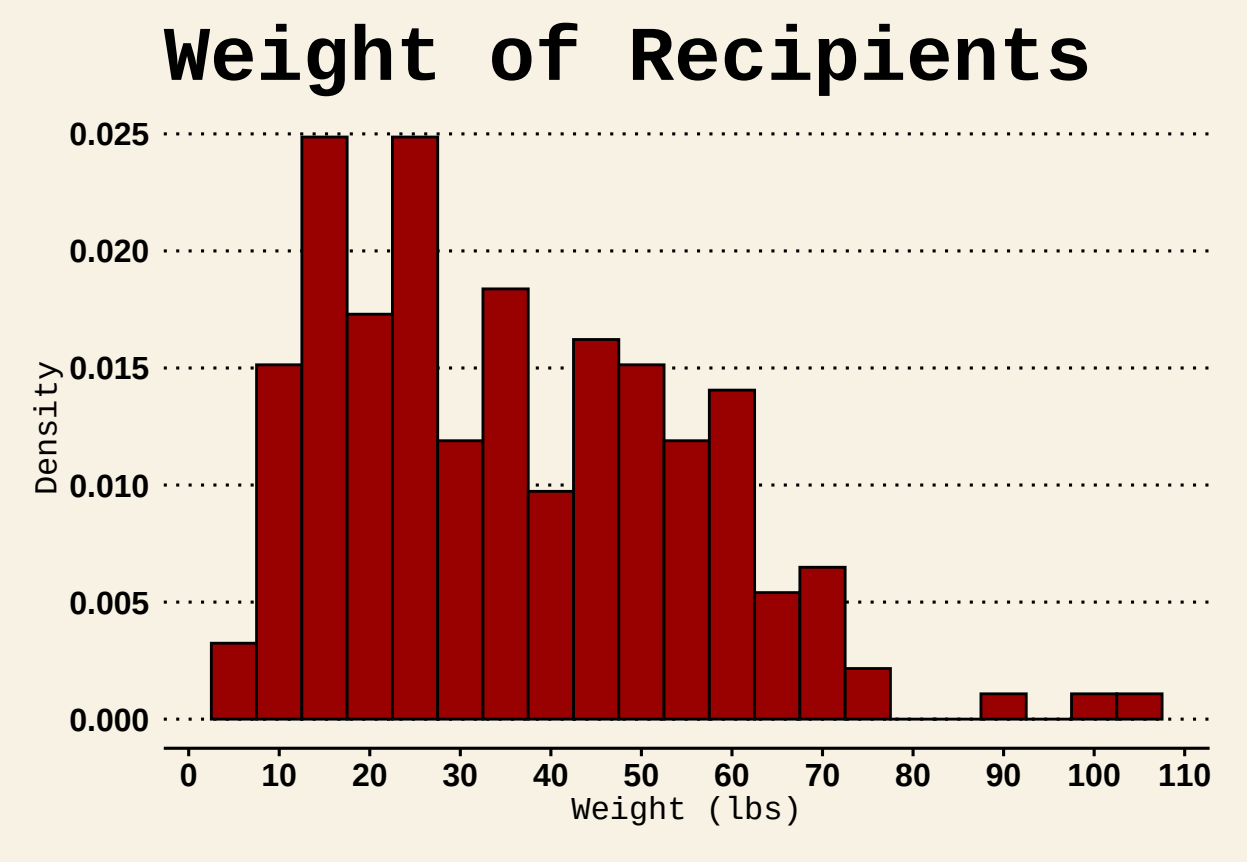
1) Our recipients are mostly between 5 and 15 years old (min age is 0.6 years old and max age is 20 years old)
2) Recipients' blood types are as follows: type O (25.7%), type A (40.1%), type AB (7.0%), B (26.7%), NA (0.5%)
3) Recipients' cytomegalovirus status follows: absent (39.0%), present (53.5%), missing (7.5%)

Now we can try to have a look at recipients' weight, rh factor (+ or -), and risk group.

Note: Changed data type of recipient body mass from character to numeric (Assuming weight has units of lbs)

```
# plot histogram of body mass distribution
bone %>%
ggplot() +
  geom_histogram(aes(x = as.numeric(recipient_body_mass), y = ..density..),
                 fill = "#990000", col = "black", binwidth = 5) +
  ggtitle("Weight of Recipients", ) +
  scale_x_continuous(breaks = seq(0, 110, by =10)) +
  xlab("Weight (lbs)") +
  ylab("Density") +
  theme_wsj()+ theme(axis.title=element_text(size=12))
```

```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```




```
# summary stats for recipient rh factor
bone %>%
  group_by(recipient_rh) %>%
  summarise(n = length(recipient_rh),
            proportion = length(recipient_rh)/nrow(bone))
```

```
## # A tibble: 3 x 3
##   recipient_rh     n proportion
##   <chr>        <int>      <dbl>
## 1 minus           27     0.144
## 2 plus           158     0.845
## 3 <NA>             2     0.0107
```

```
# summary stats for recipient risk group
bone %>%
  group_by(risk_group) %>%
  summarise(n = length(risk_group),
            proportion = length(risk_group)/nrow(bone))
```

```
## # A tibble: 2 x 3
##   risk_group     n proportion
##   <chr>      <int>      <dbl>
## 1 high          69      0.369
## 2 low          118      0.631
```

So far we can see the following about our recipients:

1) Our recipients weighed between 15lbs to 60lbs (min weight is 6 lbs and max weight is 103.4 lbs)
2) Recipients' rh factor (rhesus) are as follows: plus (84.5%), minus (14.4%), NA (1.1%)
3) Risk status follows: high (36.9%), low (63.1%)

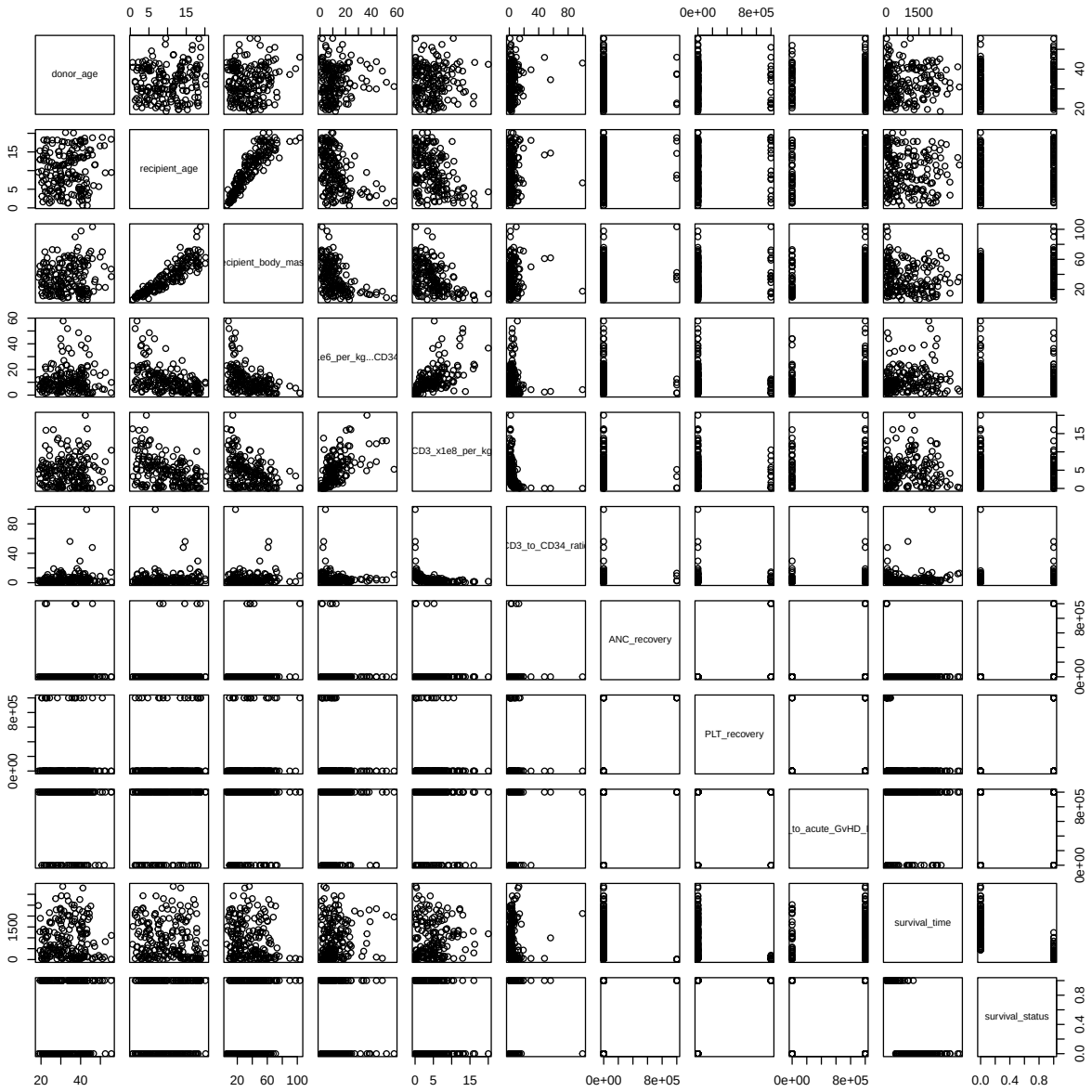
### Scatterplots

Looking only at the numerical variables next:

```
bone_num <- bone %>% select(c(donor_age, recipient_age, recipient_body_mass, CD34_x1e6_per_kg...CD34kgx
```

```
bone_num <- sapply(bone_num, as.numeric)
```

```
pairs(bone_num)
```




We see at first glance based off of the correlation scatter plots that recipient age and body mass are very highly correlated (as we would expect). We also see many plots with binary variables, and several plots with points that are very far from the main cloud of data.

We can also use a correlation matrix to see the relationships between our variables, where applicable. The data is quite large with many binary variables, so perhaps our visualizations will aid us more effectively.

```
# correlation matrix
cor(bone_num, use= "complete.obs") %>% as.data.frame() %>% pander()
```

Table 1: Table continues below

| | donor_age | recipient_age |
|---|---|---|
| **donor_age** | 1 | 0.108 |
| **recipient_age** | 0.108 | 1 |
| **recipient_body_mass** | 0.1172 | 0.8992 |
| **CD34_x1e6_per_kg... CD34kgx10d6** | 0.07269 | -0.4491 |
| **CD3_x1e8_per_kg** | 0.02355 | -0.4307 |
| **CD3_to_CD34_ratio** | 0.1449 | 0.05391 |
| **ANC_recovery** | 0.04305 | 0.1001 |
| **PLT_recovery** | -0.0366 | 0.09522 |
| **time_to_acute_GvHD_III_IV** | 0.004617 | 0.1516 |
| **survival_time** | -0.007969 | -0.1256 |
| **survival_status** | 0.07668 | 0.1925 |

Table 2: Table continues below

| | recipient_body_mass |
|---|---|
| **donor_age** | 0.1172 |
| **recipient_age** | 0.8992 |
| **recipient_body_mass** | 1 |
| **CD34_x1e6_per_kg... CD34kgx10d6** | -0.4657 |
| **CD3_x1e8_per_kg** | -0.4536 |
| **CD3_to_CD34_ratio** | 0.06139 |
| **ANC_recovery** | 0.1389 |
| **PLT_recovery** | 0.1245 |
| **time_to_acute_GvHD_III_IV** | 0.1447 |
| **survival_time** | -0.1408 |
| **survival_status** | 0.2337 |

Table 3: Table continues below

| | CD34_x1e6_per_kg... CD34kgx10d6 |
|---|---|
| **donor_age** | 0.07269 |
| **recipient_age** | -0.4491 |
| **recipient_body_mass** | -0.4657 |
| **CD34_x1e6_per_kg... CD34kgx10d6** | 1 |
| **CD3_x1e8_per_kg** | 0.5833 |
| **CD3_to_CD34_ratio** | -0.1306 |
| **ANC_recovery** | -0.0835 |
| **PLT_recovery** | -0.1811 |
| **time_to_acute_GvHD_III_IV** | 0.003353 |
| **survival_time** | 0.1583 |
| **survival_status** | -0.1633 |

Table 4: Table continues below

| | CD3_x1e8_per_kg | CD3_to_CD34_ratio |
|---|---|---|
| **donor_age** | 0.02355 | 0.1449 |
| **recipient_age** | -0.4307 | 0.05391 |
| **recipient_body_mass** | -0.4536 | 0.06139 |
| **CD34_x1e6_per_kg...CD34kgx10d6** | 0.5833 | -0.1306 |
| **CD3_x1e8_per_kg** | 1 | -0.3709 |
| **CD3_to_CD34_ratio** | -0.3709 | 1 |
| **ANC_recovery** | -0.09975 | 0.02466 |
| **PLT_recovery** | -0.1132 | -0.01089 |
| **time_to_acute_GvHD_III_IV** | -0.03535 | 0.03024 |
| **survival_time** | 0.06099 | 0.04402 |
| **survival_status** | -0.2323 | 0.08987 |

Table 5: Table continues below

| | ANC_recovery | PLT_recovery |
|---|---|---|
| **donor_age** | 0.04305 | -0.0366 |
| **recipient_age** | 0.1001 | 0.09522 |
| **recipient_body_mass** | 0.1389 | 0.1245 |
| **CD34_x1e6_per_kg...CD34kgx10d6** | -0.0835 | -0.1811 |
| **CD3_x1e8_per_kg** | -0.09975 | -0.1132 |
| **CD3_to_CD34_ratio** | 0.02466 | -0.01089 |
| **ANC_recovery** | 1 | 0.4829 |
| **PLT_recovery** | 0.4829 | 1 |
| **time_to_acute_GvHD_III_IV** | 0.08084 | -0.06482 |
| **survival_time** | -0.1671 | -0.3247 |
| **survival_status** | 0.1674 | 0.3467 |

Table 6: Table continues below

| | time_to_acute_GvHD_III_IV | survival_time |
|---|---|---|
| **donor_age** | 0.004617 | -0.007969 |
| **recipient_age** | 0.1516 | -0.1256 |
| **recipient_body_mass** | 0.1447 | -0.1408 |
| **CD34_x1e6_per_kg...CD34kgx10d6** | 0.003353 | 0.1583 |
| **CD3_x1e8_per_kg** | -0.03535 | 0.06099 |
| **CD3_to_CD34_ratio** | 0.03024 | 0.04402 |
| **ANC_recovery** | 0.08084 | -0.1671 |
| **PLT_recovery** | -0.06482 | -0.3247 |
| **time_to_acute_GvHD_III_IV** | 1 | 0.127 |
| **survival_time** | 0.127 | 1 |
| **survival_status** | -0.09932 | -0.7567 |

| | survival_status |
|---|---|
| **donor_age** | 0.07668 |
| **recipient_age** | 0.1925 |

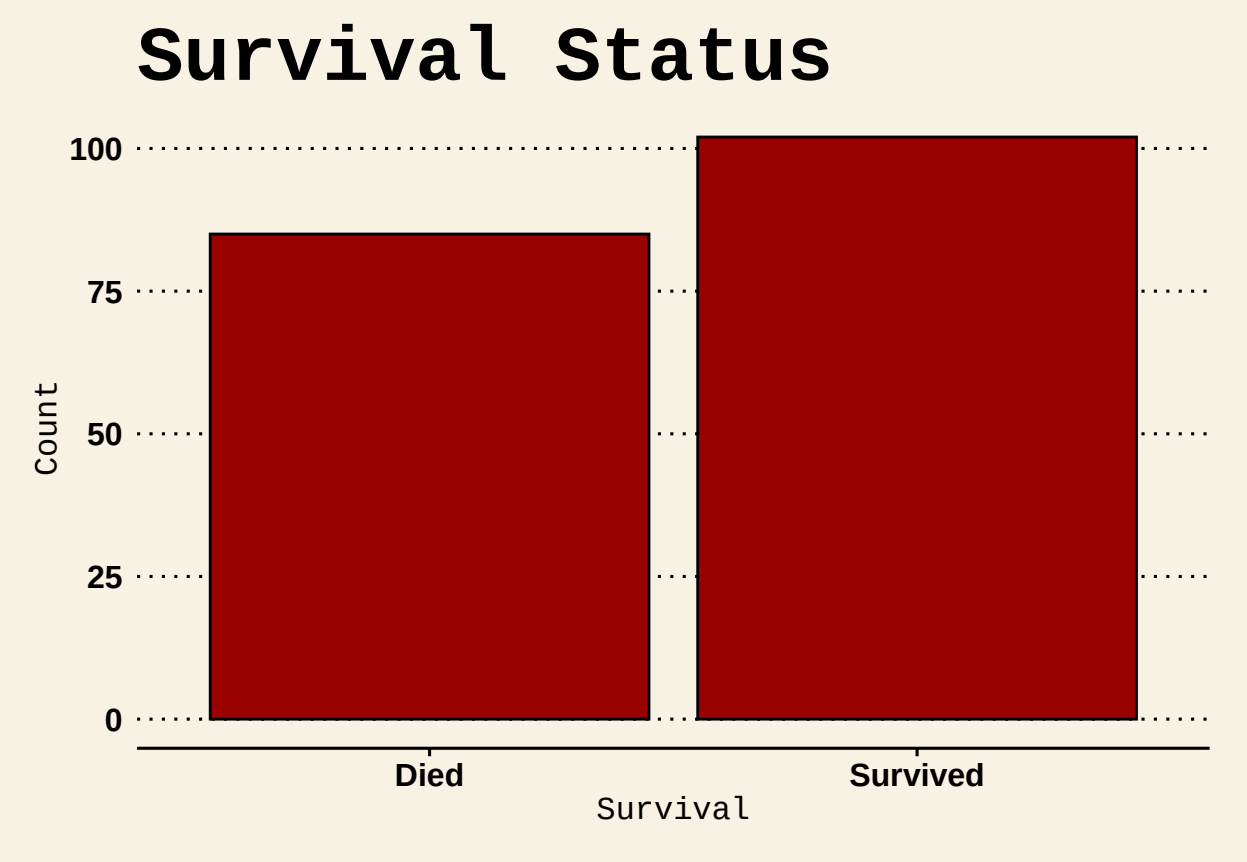
| | survival_status |
|---|---|
| **recipient_body_mass** | 0.2337 |
| **CD34_x1e6_per_kg...CD34kgx10d6** | -0.1633 |
| **CD3_x1e8_per_kg** | -0.2323 |
| **CD3_to_CD34_ratio** | 0.08987 |
| **ANC_recovery** | 0.1674 |
| **PLT_recovery** | 0.3467 |
| **time_to_acute_GvHD_III_IV** | -0.09932 |
| **survival_time** | -0.7567 |
| **survival_status** | 1 |

**Compare outcome variable (survival_status) and numeric variables**

1. Explore our binary outcome variable

In order to explore our outcome variable of interest, survival status after bone marrow transplant, we must first change it to numeric. This graph shows us that 45.5% of our bone marrow transplant patients unfortunately did not survive, while 54.5% did survive.

```
bone$survival_status <- ifelse(bone$survival_status == 1, "Died", "Survived") %>% as.factor()

bone %>%
ggplot(aes(x = survival_status)) +
  geom_histogram(fill = "#990000", col = "black", binwidth = 5, stat = "count") +
  ggtitle("Survival Status", ) +
  xlab("Survival") +
  ylab("Count") +
  theme_wsj()+ theme(axis.title=element_text(size=12))
```




```
bone %>%
  group_by(survival_status) %>%
  summarise(n = length(survival_status),
            proportion = length(survival_status)/nrow(bone))
```

```
## # A tibble: 2 x 3
##   survival_status     n proportion
##   <fct>           <int>      <dbl>
## 1 Died               85      0.455
## 2 Survived          102      0.545
```

2. Look at binary outcome with continuous predictors

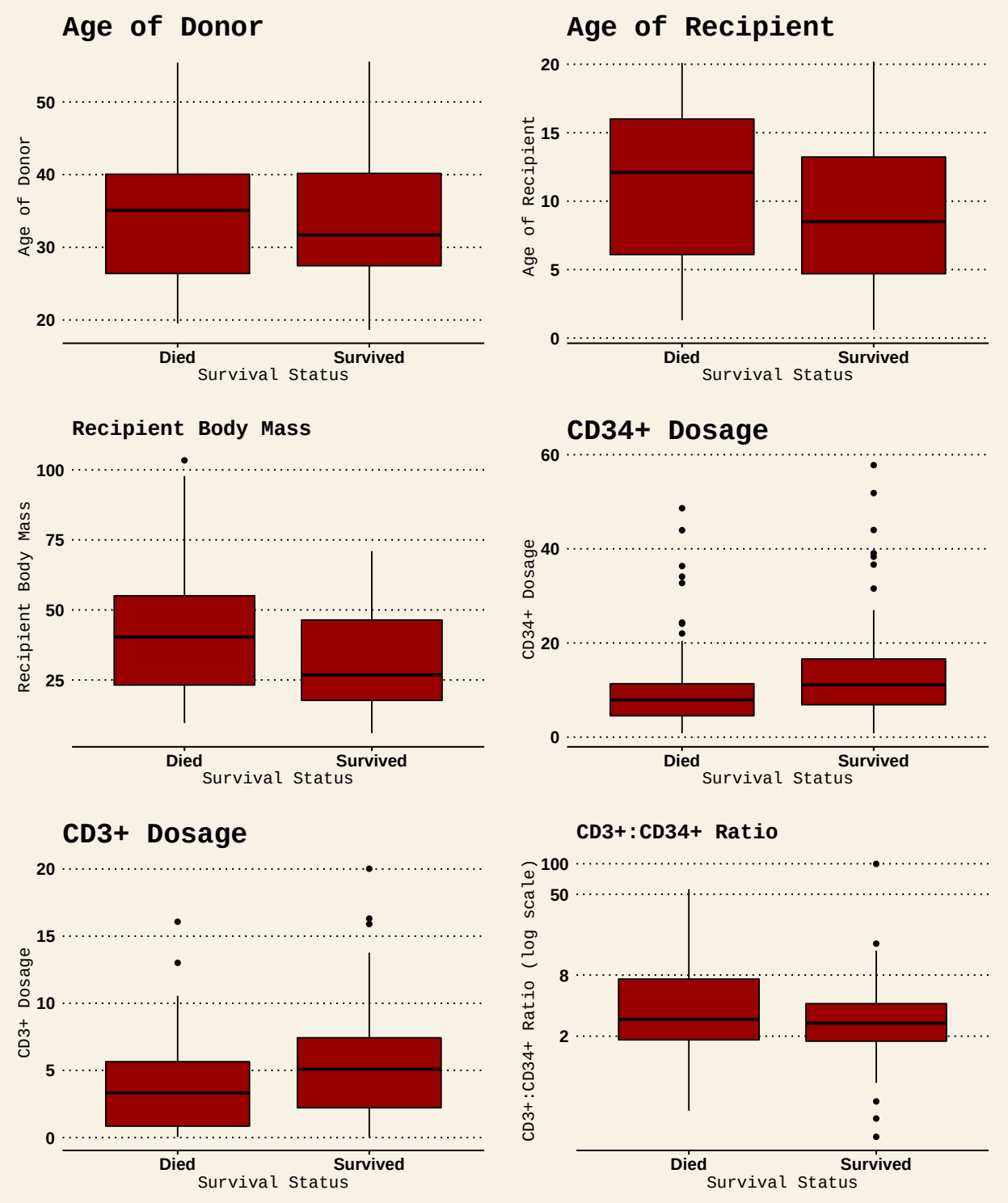
Now let's view our outcome variable, survival status, with our continuous predictor variables (listed below). We should pay close attention to any notable changes in distributions between those who survived compared to those who did not.

donor_age, recipient_age, recipient_body_mass, CD34_x1e6_per_kg. . . CD34kgx10d6, CD3_x1e8_per_kg, CD3_to_CD34_ratio, ANC_recovery, PLT_recovery, time_to_acute_GvHD_III_IV, survival_time

```
# construct plots of continuous predictors
c1 <- bone %>%
  ggplot(aes(x = survival_status, y = donor_age)) +
  geom_boxplot(fill = "#990000", col = "black") +
  ggtitle("Age of Donor") +
```

```
  xlab("Survival Status") +
  ylab("Age of Donor") +
  theme_wsj()+ theme(axis.title=element_text(size=12),
                     plot.title = element_text(size = 20))


c2 <- bone %>%
  ggplot(aes(x = survival_status, y = recipient_age)) +
  geom_boxplot(fill = "#990000", col = "black") +
  ggtitle("Age of Recipient") +
  xlab("Survival Status") +
  ylab("Age of Recipient") +
  theme_wsj()+ theme(axis.title=element_text(size=12),
                     plot.title = element_text(size = 20))


c3 <- bone %>%
  ggplot(aes(x = survival_status, y = as.numeric(recipient_body_mass))) +
  geom_boxplot(fill = "#990000", col = "black") +
  ggtitle("Recipient Body Mass") +
  xlab("Survival Status") +
  ylab("Recipient Body Mass") +
  theme_wsj()+ theme(axis.title=element_text(size=12),
                     plot.title = element_text(size = 15))


c4 <- bone %>%
  ggplot(aes(x = survival_status, y = as.numeric(CD34_x1e6_per_kg...CD34kgx10d6))) +
  geom_boxplot(fill = "#990000", col = "black") +
  ggtitle("CD34+ Dosage") +
  xlab("Survival Status") +
  ylab("CD34+ Dosage") +
  theme_wsj()+ theme(axis.title=element_text(size=12),
                     plot.title = element_text(size = 20))


c5 <- bone %>%
  ggplot(aes(x = survival_status, y = as.numeric(CD3_x1e8_per_kg))) +
  geom_boxplot(fill = "#990000", col = "black") +
  ggtitle("CD3+ Dosage") +
  xlab("Survival Status") +
  ylab("CD3+ Dosage") +
  theme_wsj()+ theme(axis.title=element_text(size=12),
                     plot.title = element_text(size = 20))


c6 <- bone %>%
  ggplot(aes(x = survival_status, y = as.numeric(CD3_to_CD34_ratio))) +
  geom_boxplot(fill = "#990000", col = "black") +
  scale_y_continuous(trans = "log", breaks = c(2,8,50,100)) +
  ggtitle("CD3+:CD34+ Ratio") +
  xlab("Survival Status") +
  ylab("CD3+:CD34+ Ratio (log scale)") +
```

```
  theme_wsj()+ theme(axis.title=element_text(size=12),
                     plot.title = element_text(size = 15))

# plot together
grid.arrange(c1,c2,c3,c4,c5,c6, ncol = 2, nrow = 3)
```



We can observe above the various relationships our variables have with the survival status outcome variable. I can not tell regression outcomes just from graphs, but we are particularly interested in variables such as CD34+ Dosage and CD3+ Dosage which tell us the amount of bone marrow transplanted.

### Comparison of Outcome (survival_status) and categorical variables

First, we will look at the comparison between our outcome variable and the different categorical variables which include:

**Categorical Variables:** donor_ABO, recipient_age_int, recipient_ABO, disease, CMV_status, HLA_match..out.of.10., antigen, allele and HLA_group_1

```
par(mfrow=c(3,3))

#Donor ABO
#2x2 table with margins
tab1 <- table(bone$survival_status,bone$donor_ABO)
tab1 <-prop.table(tab1,2)
addmargins(tab1) %>% pander(caption = "Donor Blood Type")
```

Table 8: Donor Blood Type

| | 0 | A | AB | B | Sum |
|---|---|---|---|---|---|
| **Died** | 0.4521 | 0.507 | 0.2667 | 0.4286 | 1.654 |
| **Survived** | 0.5479 | 0.493 | 0.7333 | 0.5714 | 2.346 |
| **Sum** | 1 | 1 | 1 | 1 | 4 |

```
tab1 <- tab1 %>% as.data.frame()

p1 <- ggplot(aes(Var2, Freq, fill = Var1), data = tab1) +
  geom_col(position = "dodge") +
  xlab("Donor Blood Type") +
  ylab ("Density (%)") +
  labs(fill='status') +
  scale_fill_manual(values=c("#990000", "burlywood2")) +
  theme(legend.position="top")

#Recipient age
bone$recipient_age_int <- factor(bone$recipient_age_int , levels=c("0_5", "5_10", "10_20"))
#2x2 table with margins
tab2 <- table(bone$survival_status,bone$recipient_age_int)
tab2 <-prop.table(tab2,2)
addmargins(tab2) %>% pander(caption = "Recipient Age")
```

Table 9: Recipient Age

| | 0_5 | 5_10 | 10_20 | Sum |
|---|---|---|---|---|
| **Died** | 0.3617 | 0.4118 | 0.5281 | 1.302 |
| **Survived** | 0.6383 | 0.5882 | 0.4719 | 1.698 |

| | 0_5 | 5_10 | 10_20 | Sum |
|---|---|---|---|---|
| **Sum** | 1 | 1 | 1 | 3 |

```
tab2 <- tab2 %>% as.data.frame()

p2 <- ggplot(aes(Var2, Freq, fill = Var1), data = tab2) +
  geom_col(position = "dodge") +
  xlab("Recipient Age Range") +
  ylab ("Density (%)") +
  labs(fill='status') +
  scale_fill_manual(values=c("#990000", "burlywood2")) +
  theme(legend.position="top")

#recipient_ABO
#2x2 table with margins
tab3 <- table(bone$survival_status,bone$recipient_ABO)
tab3 <-prop.table(tab3,2)
addmargins(tab3) %>% pander(caption = "Recipient Blood Type")
```

Table 10: Recipient Blood Type

| | 0 | A | AB | B | Sum |
|---|---|---|---|---|---|
| **Died** | 0.4375 | 0.48 | 0.3846 | 0.44 | 1.742 |
| **Survived** | 0.5625 | 0.52 | 0.6154 | 0.56 | 2.258 |
| **Sum** | 1 | 1 | 1 | 1 | 4 |

```
tab3 <- tab3 %>% as.data.frame()

p3 <- ggplot(aes(Var2, Freq, fill = Var1), data = tab3) +
  geom_col(position = "dodge") +
  xlab("Recipient Blood Type") +
  ylab ("Density (%)") +
  labs(fill='status') +
  scale_fill_manual(values=c("#990000", "burlywood2")) +
  theme(legend.position="top")


#Disease
#2x2 table with margins
# disease by survival
tab4 <- table(bone$survival_status,bone$disease)
tab4 <-prop.table(tab4,2)
addmargins(tab4) %>% pander(caption = "Disease")
```

Table 11: Disease

| | ALL | AML | chronic | lymphoma | nonmalignant | Sum |
|---|---|---|---|---|---|---|
| **Died** | 0.4412 | 0.4545 | 0.4222 | 1 | 0.375 | 2.693 |
| **Survived** | 0.5588 | 0.5455 | 0.5778 | 0 | 0.625 | 2.307 |

| | ALL | AML | chronic | lymphoma | nonmalignant | Sum |
|---|---|---|---|---|---|---|
| **Sum** | 1 | 1 | 1 | 1 | 1 | 5 |

```
tab4 <- tab4 %>% as.data.frame()

p4 <- ggplot(aes(Var2, Freq, fill = Var1), data = tab4) +
  geom_col(position = "dodge") +
  xlab("Disease Type") +
  ylab ("Density (%)") +
  labs(fill='status') +
  scale_fill_manual(values=c("#990000", "burlywood2")) +
  theme(legend.position="top")

#CMV_status

#2x2 table with margins
# CMV status by survival
tab5 <- table(bone$survival_status,bone$CMV_status)
tab5 <-prop.table(tab5,2)
addmargins(tab5) %>% pander(caption = "CMV Status")
```

Table 12: CMV Status

| | 0 | 1 | 2 | 3 | Sum |
|---|---|---|---|---|---|
| **Died** | 0.4583 | 0.3333 | 0.4561 | 0.4615 | 1.709 |
| **Survived** | 0.5417 | 0.6667 | 0.5439 | 0.5385 | 2.291 |
| **Sum** | 1 | 1 | 1 | 1 | 4 |

```
tab5 <- tab5 %>% as.data.frame()

p5 <- ggplot(aes(Var2, Freq, fill = Var1), data = tab5) +
  geom_col(position = "dodge") +
  xlab("CMV Status") +
  ylab ("Density (%)") +
  labs(fill='status') +
  scale_fill_manual(values=c("#990000", "burlywood2")) +
  theme(legend.position="top")


#HLA Match

#2x2 table with margins
# HLA by survival
tab6 <- table(bone$survival_status,bone$HLA_match..out.of.10.)
tab6 <-prop.table(tab6,2)
addmargins(tab6) %>% pander(caption = "HLA Match out of 10")
```

Table 13: HLA Match out of 10

| | 7 | 8 | 9 | 10 | Sum |
|---|---|---|---|---|---|
| **Died** | 0.6 | 0.4348 | 0.4769 | 0.4362 | 1.948 |
| **Survived** | 0.4 | 0.5652 | 0.5231 | 0.5638 | 2.052 |
| **Sum** | 1 | 1 | 1 | 1 | 4 |

```
tab6 <- tab6 %>% as.data.frame()

p6 <- ggplot(aes(Var2, Freq, fill = Var1), data = tab6) +
  geom_col(position = "dodge") +
  xlab("HLA Match (out of 10)") +
  ylab ("Density (%)") +
  labs(fill='status') +
  scale_fill_manual(values=c("#990000", "burlywood2")) +
  theme(legend.position="top")


#antigen

#2x2 table with margins
# Antigen by survival
tab7 <- table(bone$survival_status,bone$antigen)
tab7 <- prop.table(tab7,2)
addmargins(tab7)  %>% pander(caption = "Antigen")
```

Table 14: Antigen

| | 0 | 1 | 2 | 3 | Sum |
|---|---|---|---|---|---|
| **Died** | 0.4409 | 0.5238 | 0.4462 | 0.5714 | 1.982 |
| **Survived** | 0.5591 | 0.4762 | 0.5538 | 0.4286 | 2.018 |
| **Sum** | 1 | 1 | 1 | 1 | 4 |

```
tab7 <- tab7 %>% as.data.frame()

p7 <- ggplot(aes(Var2, Freq, fill = Var1), data = tab7) +
  geom_col(position = "dodge") +
  xlab("Antigen Difference") +
  ylab ("Density (%)") +
  labs(fill='status') +
  scale_fill_manual(values=c("#990000", "burlywood2")) +
  theme(legend.position="top")

#allele
#2x2 table with margins
# Allele by survival
tab8 <- table(bone$survival_status,bone$allele)
tab8 <- prop.table(tab8,2)
addmargins(tab8)  %>% pander(caption = "Allele")
```

Table 15: Allele

| | 0 | 1 | 2 | 3 | 4 | Sum |
|---|---|---|---|---|---|---|
| **Died** | 0.4409 | 0.4815 | 0.4375 | 0.5 | 1 | 2.86 |
| **Survived** | 0.5591 | 0.5185 | 0.5625 | 0.5 | 0 | 2.14 |
| **Sum** | 1 | 1 | 1 | 1 | 1 | 5 |

```
tab8 <- tab8 %>% as.data.frame()

p8 <- ggplot(aes(Var2, Freq, fill = Var1), data = tab8) +
  geom_col(position = "dodge") +
  xlab("Allele Difference") +
  ylab ("Density (%)") +
  labs(fill='status') +
  scale_fill_manual(values=c("#990000", "burlywood2")) +
  theme(legend.position="top")

#HLA_group_1

#2x2 table with margins
# HLA gorup 1 by survival
tab9 <- table(bone$survival_status,bone$HLA_group_1)
tab9 <- prop.table(tab9,2)
addmargins(tab9)  %>% pander(caption = "HLA Group 1")
```

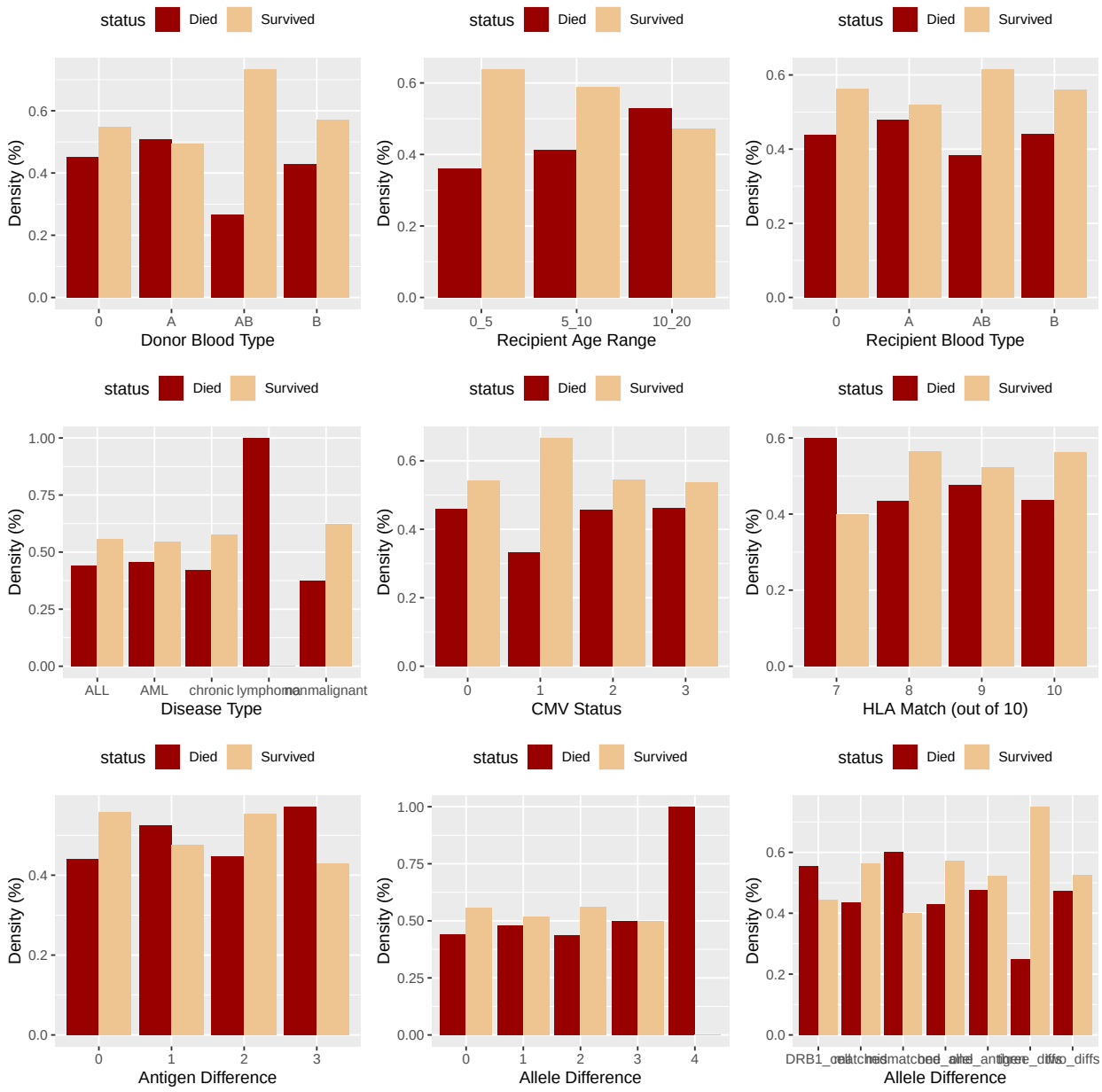
Table 16: HLA Group 1 (continued below)

| | DRB1_cell | matched | mismatched | one_allel | one_antigen |
|---|---|---|---|---|---|
| **Died** | 0.5556 | 0.4362 | 0.6 | 0.4286 | 0.4762 |
| **Survived** | 0.4444 | 0.5638 | 0.4 | 0.5714 | 0.5238 |
| **Sum** | 1 | 1 | 1 | 1 | 1 |

| | three_diffs | two_diffs | Sum |
|---|---|---|---|
| **Died** | 0.25 | 0.4737 | 3.22 |
| **Survived** | 0.75 | 0.5263 | 3.78 |
| **Sum** | 1 | 1 | 7 |

```
# Increase margin size
par(mar=c(8,1,1,1))

tab9 <- tab9 %>% as.data.frame()

p9 <- ggplot(aes(Var2, Freq, fill = Var1), data = tab9) +
  geom_col(position = "dodge") +
  xlab("Allele Difference") +
  ylab ("Density (%)") +
  labs(fill='status') +
  scale_fill_manual(values=c("#990000", "burlywood2")) +
  theme(legend.position="top")
```

```
grid.arrange(p1,p2,p3,p4,p5,p6,p7,p8,p9,
             nrow = 3, ncol = 3)
```




Looking at these resulting graphs and tables we see some interesting relationships that we will want to explore further. Starting with our first comparison:

(1) donor_ABO - the trend that sticks out the most with this comparison is the drastically better chances of survival if the donor has blood type AB compared to the other blood types

(2) recipient_age_int - There's not a trend that sticks out too much, it does look like though that recipients under 10 have an increased chance of survival

(3) recipient_ABO - Again here, there isn't a huge trend, again we see that a recipient having blood type AB has the best chance of survival among the other blood types. So we will definitely need to look into this specific trend.

(4) disease - All of these disease categories show the same probability of survival except leukemia, which shows a 0% chance of survival which is definitely something we need to consider and explore

(5) CMV_status - This variable shows serological compatibility of the donor and the recipient of hematopoietic stem cells according to cytomegalovirus infection prior to transplantation. So the higher the value the lower the compatibility. Here we see that lower values like 0 and 1 show a higher chance of survival which is what we expect.

(6) HLA_match..out.of.10. - This variable shows compatibility of antigens of the main histocompatibility complex of the donor and the recipient of hematopoietic stem cells (10/10, 9/10, 8/10, 7/10). As expected, a compatibility of 7/10 shows the least likely chances of survival.

(7) antigen - This variable shows the difference in antigens between the donor and the recipient (0-3). This graph shows varied results across the board with no noticeable trend.

(8) allele - This variable shows the allele difference between the donor and the recipient (0-4). The graph here pretty much shows the same chances of survival except when the allele difference was 4, then there was a 0% chance of survival, which definitely a factor we need to look into and consider in our modeling.

(9) HLA_group_1 - This variable shows the difference type between the donor and the recipient (HLA matched, one antigen, one allele, DRB1 cell, two allele or allel+antigen, two antigenes+allel, mismatched). This is one of our variables we need to get more information on from a domain expert in order to understand the matching process. However, we can still look at its exploratory graph. Here we see that there are significant differences in death and survival with couple of the categories. If the HLA group is mismatched there an increased risk of death and 3 differences increased survival significantly, which, logically, we would not have thought. So domain expertise will definitely be needed for this variable.

**Binary Variables:** And now, we will take a look at the comparison of our outcome variable and the extensive list of binary variables:

donor_age_below_35 (yes/no), donor_CMV (present/absent), recipient_age_below_10 (yes/no), recipient_rh (plus/minus), recipient_CMV (present/absent), disease_group (malignant/nonmalignant), gender_match(female_to_male/other), ABO_match (matched/mismatched), HLA_mismatch (matched, mismatched), risk_group (high/low), stem_cell_source (peripheral_blood/bone_marrow), tx_post_relapse (yes/no), acute_GvHD_II_III_IV (yes/no), acute_GvHD_III_IV (yes/no), extensive_chronic_GvHD (yes/no), relapse (yes/no), survival_status (yes,no)

```
#par(mfrow = c(4,4))

#Donor Age Group
#2x2 table with margins
tab10 <- table(bone$survival_status,bone$donor_age_below_35)
tab10 <- prop.table(tab10,2)
addmargins(tab10) %>% pander(caption = "Donor Age below 35")
```

Table 18: Donor Age below 35

| | no | yes | Sum |
|---|---|---|---|
| **Died** | 0.5181 | 0.4038 | 0.9219 |
| **Survived** | 0.4819 | 0.5962 | 1.078 |
| **Sum** | 1 | 1 | 2 |

```
tab10 <- tab10 %>% data.frame()

p10 <- ggplot(aes(Var2, Freq, fill = Var1), data = tab10) +
  geom_col(position = "dodge") +
  xlab("Donor Age Under 35") +
  ylab ("Density (%)") +
  labs(fill='status') +
  scale_fill_manual(values=c("#990000", "burlywood2")) +
  theme(legend.position="top")



#Donor CMV
#2x2 table with margins
tab11 <- table(bone$survival_status,bone$donor_CMV)
tab11 <- prop.table(tab11,2)
addmargins(tab11)  %>% pander(caption = "Donor CMV")
```

Table 19: Donor CMV

| | absent | present | Sum |
|---|---|---|---|
| **Died** | 0.4779 | 0.4167 | 0.8945 |
| **Survived** | 0.5221 | 0.5833 | 1.105 |
| **Sum** | 1 | 1 | 2 |

```
tab11 <- tab11 %>% data.frame()

p11 <- ggplot(aes(Var2, Freq, fill = Var1), data = tab11) +
  geom_col(position = "dodge") +
  xlab("Donor CMV Status") +
  ylab ("Density (%)") +
  labs(fill='status') +
  scale_fill_manual(values=c("#990000", "burlywood2"))+
  theme(legend.position="top")

#recipient_age_below_10
#2x2 table with margins
tab12 <- table(bone$survival_status,bone$recipient_age_below_10)
tab12 <- prop.table(tab12,2)
addmargins(tab12)  %>% pander(caption = "Recipient Age Below 10 yrs old")
```

Table 20: Recipient Age Below 10 yrs old

| | no | yes | Sum |
|---|---|---|---|
| **Died** | 0.5341 | 0.3838 | 0.9179 |
| **Survived** | 0.4659 | 0.6162 | 1.082 |
| **Sum** | 1 | 1 | 2 |

```
tab12 <- tab12 %>% data.frame()

p12 <- ggplot(aes(Var2, Freq, fill = Var1), data = tab12) +
  geom_col(position = "dodge") +
  xlab("Recipient Age Below 10") +
  ylab ("Density (%)") +
  labs(fill='status') +
  scale_fill_manual(values=c("#990000", "burlywood2"))+
  theme(legend.position="top")

#recipient_rh
#2x2 table with margins
tab13 <- table(bone$survival_status,bone$recipient_rh)
tab13 <- prop.table(tab13,2)
addmargins(tab13)  %>% pander(caption = "Recipient RH Status")
```

Table 21: Recipient RH Status

| | minus | plus | Sum |
|---|---|---|---|
| **Died** | 0.2963 | 0.4747 | 0.771 |
| **Survived** | 0.7037 | 0.5253 | 1.229 |
| **Sum** | 1 | 1 | 2 |

```
tab13 <- tab13 %>% data.frame()

p13 <- ggplot(aes(Var2, Freq, fill = Var1), data = tab13) +
  geom_col(position = "dodge") +
  xlab("Recipient RH Status") +
  ylab ("Density (%)") +
  labs(fill='status') +
  scale_fill_manual(values=c("#990000", "burlywood2"))+
  theme(legend.position="top")


#recipient_CMV
#2x2 table with margins
tab14 <- table(bone$survival_status,bone$recipient_CMV)
tab14 <- prop.table(tab14,2)
addmargins(tab14) %>% pander(caption = "Recipient CMV Stautus")
```

Table 22: Recipient CMV Stautus

| | absent | present | Sum |
|---|---|---|---|
| **Died** | 0.411 | 0.45 | 0.861 |
| **Survived** | 0.589 | 0.55 | 1.139 |
| **Sum** | 1 | 1 | 2 |

```
tab14 <- tab14 %>% data.frame()
```

```
p14 <- ggplot(aes(Var2, Freq, fill = Var1), data = tab14) +
  geom_col(position = "dodge") +
  xlab("Recipient CMV Presence") +
  ylab ("Density (%)") +
  labs(fill='status') +
  scale_fill_manual(values=c("#990000", "burlywood2"))+
  theme(legend.position="top")

#disease_group
#2x2 table with margins
tab15 <- table(bone$survival_status,bone$disease_group)
tab15 <- prop.table(tab15,2)
addmargins(tab15)  %>% pander(caption = "Disease Group")
```

Table 23: Disease Group

| | malignant | nonmalignant | Sum |
|---|---|---|---|
| **Died** | 0.471 | 0.375 | 0.846 |
| **Survived** | 0.529 | 0.625 | 1.154 |
| **Sum** | 1 | 1 | 2 |

```
tab15 <- tab15 %>% data.frame()

p15 <- ggplot(aes(Var2, Freq, fill = Var1), data = tab15) +
  geom_col(position = "dodge") +
  xlab("Disease Group") +
  ylab ("Density (%)") +
  labs(fill='status') +
  scale_fill_manual(values=c("#990000", "burlywood2"))+
  theme(legend.position="top")

#gender_match
#2x2 table with margins
tab16 <- table(bone$survival_status,bone$gender_match)
tab16 <- prop.table(tab16,2)
addmargins(tab16)  %>% pander(caption = "Gender Match")
```

Table 24: Gender Match

| | female_to_male | other | Sum |
|---|---|---|---|
| **Died** | 0.4688 | 0.4516 | 0.9204 |
| **Survived** | 0.5312 | 0.5484 | 1.08 |
| **Sum** | 1 | 1 | 2 |

```
tab16 <- tab16 %>% data.frame()

p16 <- ggplot(aes(Var2, Freq, fill = Var1), data = tab16) +
  geom_col(position = "dodge") +
  xlab("Gender Match") +
  ylab ("Density (%)") +
```

```
  labs(fill='status') +
  scale_fill_manual(values=c("#990000", "burlywood2"))+
  theme(legend.position="top")

#ABO_match
#2x2 table with margins
tab17 <- table(bone$survival_status,bone$ABO_match)
tab17 <- prop.table(tab17,2)
addmargins(tab17)  %>% pander(caption = "Blood Type Match")
```

Table 25: Blood Type Match

| | matched | mismatched | Sum |
|---|---|---|---|
| **Died** | 0.5192 | 0.4254 | 0.9446 |
| **Survived** | 0.4808 | 0.5746 | 1.055 |
| **Sum** | 1 | 1 | 2 |

```
tab17 <- tab17 %>% data.frame()

p17 <- ggplot(aes(Var2, Freq, fill = Var1), data = tab17) +
  geom_col(position = "dodge") +
  xlab("Match on Blood Group") +
  ylab ("Density (%)") +
  labs(fill='status') +
  scale_fill_manual(values=c("#990000", "burlywood2"))+
  theme(legend.position="top")

#HLA_match
#2x2 table with margins
tab18 <- table(bone$survival_status,bone$HLA_mismatch)
tab18 <- prop.table(tab18,2)
addmargins(tab18)  %>% pander(caption = "HLA Mismatch")
```

Table 26: HLA Mismatch

| | matched | mismatched | Sum |
|---|---|---|---|
| **Died** | 0.4528 | 0.4643 | 0.9171 |
| **Survived** | 0.5472 | 0.5357 | 1.083 |
| **Sum** | 1 | 1 | 2 |

```
tab18 <- tab18 %>% data.frame()

p18 <- ggplot(aes(Var2, Freq, fill = Var1), data = tab18) +
  geom_col(position = "dodge") +
  xlab("Match on HLA") +
  ylab ("Density (%)") +
  labs(fill='status') +
  scale_fill_manual(values=c("#990000", "burlywood2"))+
  theme(legend.position="top")
```

```
#risk_group
#2x2 table with margins
tab19 <- table(bone$survival_status,bone$risk_group)
tab19 <- prop.table(tab19,2)
addmargins(tab19)  %>% pander(caption = "Recipient Risk Group")
```

Table 27: Recipient Risk Group

| | high | low | Sum |
|---|---|---|---|
| **Died** | 0.5507 | 0.3983 | 0.949 |
| **Survived** | 0.4493 | 0.6017 | 1.051 |
| **Sum** | 1 | 1 | 2 |

```
tab19 <- tab19 %>% data.frame()

p19 <- ggplot(aes(Var2, Freq, fill = Var1), data = tab19) +
  geom_col(position = "dodge") +
  xlab("Risk Group") +
  ylab ("Density (%)") +
  labs(fill='status') +
  scale_fill_manual(values=c("#990000", "burlywood2"))+
  theme(legend.position="top")

#stem_cell_source
#2x2 table with margins
tab20 <- table(bone$survival_status,bone$stem_cell_source)
tab20 <- prop.table(tab20,2)
addmargins(tab20)  %>% pander(caption = "Stem Cell Source")
```

Table 28: Stem Cell Source

| | bone_marrow | peripheral_blood | Sum |
|---|---|---|---|
| **Died** | 0.5714 | 0.4207 | 0.9921 |
| **Survived** | 0.4286 | 0.5793 | 1.008 |
| **Sum** | 1 | 1 | 2 |

```
tab20 <- tab20 %>% data.frame()

p20 <- ggplot(aes(Var2, Freq, fill = Var1), data = tab20) +
  geom_col(position = "dodge") +
  xlab("Stem Cell Suurce") +
  ylab ("Density (%)") +
  labs(fill='status') +
  scale_fill_manual(values=c("#990000", "burlywood2"))+
  theme(legend.position="top")


#tx_post_relapse
#2x2 table with margins
tab21 <- table(bone$survival_status,bone$tx_post_relapse)
```

```
tab21 <- prop.table(tab21,2)
addmargins(tab21)  %>% pander(caption = "Treatment after 1st relapse")
```

Table 29: Treatment after 1st relapse

| | no | yes | Sum |
|---|---|---|---|
| **Died** | 0.4329 | 0.6087 | 1.042 |
| **Survived** | 0.5671 | 0.3913 | 0.9584 |
| **Sum** | 1 | 1 | 2 |

```
tab21 <- tab21 %>% data.frame()

p21 <- ggplot(aes(Var2, Freq, fill = Var1), data = tab21) +
  geom_col(position = "dodge") +
  xlab("Treatment After Relapse") +
  ylab ("Density (%)") +
  labs(fill='status') +
  scale_fill_manual(values=c("#990000", "burlywood2"))+
  theme(legend.position="top")



#acute_GvHD_II_III_IV
#2x2 table with margins
tab22 <- table(bone$survival_status,bone$acute_GvHD_II_III_IV)
tab22 <- prop.table(tab22,2)
addmargins(tab22)  %>% pander(caption = "Acute GVHD Stage")
```

Table 30: Acute GVHD Stage

| | no | yes | Sum |
|---|---|---|---|
| **Died** | 0.4133 | 0.4821 | 0.8955 |
| **Survived** | 0.5867 | 0.5179 | 1.105 |
| **Sum** | 1 | 1 | 2 |

```
tab22 <- tab22 %>% data.frame()

p22 <- ggplot(aes(Var2, Freq, fill = Var1), data = tab22) +
  geom_col(position = "dodge") +
  xlab("Acute GVHD stage II, III, IV") +
  ylab ("Density (%)") +
  labs(fill='status') +
  scale_fill_manual(values=c("#990000", "burlywood2"))+
  theme(legend.position="top")

#acute_GvHD_III_IV
#2x2 table with margins
tab23 <- table(bone$survival_status,bone$acute_GvHD_III_IV)
tab23 <- prop.table(tab23,2)
addmargins(tab23)  %>% pander(caption = "ACUTE GcVD stage III or IV")
```

Table 31: ACUTE GcVD stage III or IV

| | no | yes | Sum |
|---|---|---|---|
| **Died** | 0.4218 | 0.575 | 0.9968 |
| **Survived** | 0.5782 | 0.425 | 1.003 |
| **Sum** | 1 | 1 | 2 |

```
tab23 <- tab23 %>% data.frame()

p23 <- ggplot(aes(Var2, Freq, fill = Var1), data = tab23) +
  geom_col(position = "dodge") +
  xlab("Acute GVHD stage III or IV ") +
  ylab ("Density (%)") +
  labs(fill='status') +
  scale_fill_manual(values=c("#990000", "burlywood2"))+
  theme(legend.position="top")


#extensive_chronic_GvHD
#2x2 table with margins
tab24 <- table(bone$survival_status,bone$extensive_chronic_GvHD)
tab24 <- prop.table(tab24,2)
addmargins(tab24)  %>% pander(caption = "Chronic GvHD")
```

Table 32: Chronic GvHD

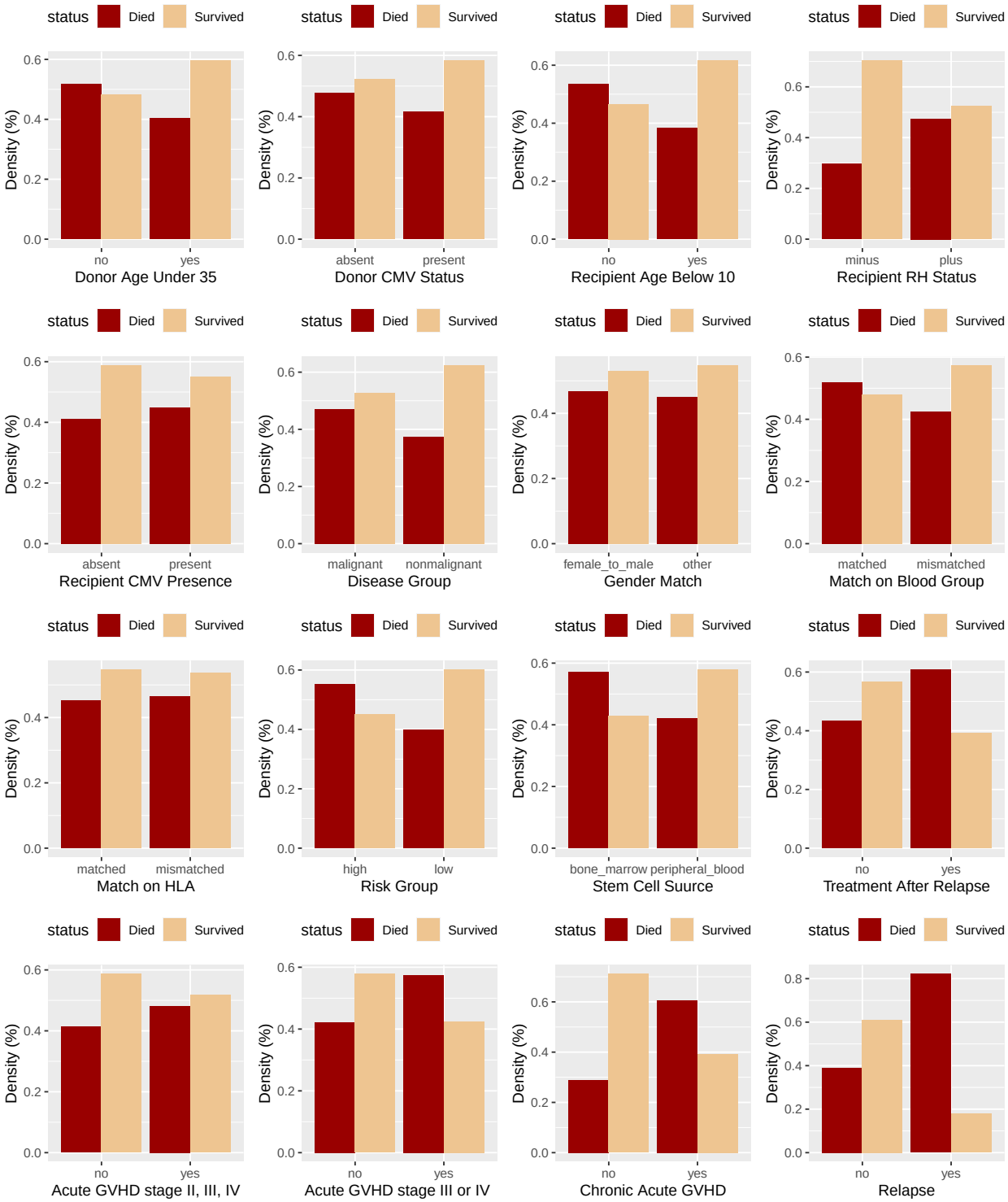
| | no | yes | Sum |
|---|---|---|---|
| **Died** | 0.2891 | 0.6071 | 0.8962 |
| **Survived** | 0.7109 | 0.3929 | 1.104 |
| **Sum** | 1 | 1 | 2 |

```
tab24 <- tab24 %>% data.frame()

p24 <- ggplot(aes(Var2, Freq, fill = Var1), data = tab24) +
  geom_col(position = "dodge") +
  xlab("Chronic Acute GVHD") +
  ylab ("Density (%)") +
  labs(fill='status') +
  scale_fill_manual(values=c("#990000", "burlywood2"))+
  theme(legend.position="top")

#relapse
#2x2 table with margins
tab25 <- table(bone$survival_status,bone$relapse)
tab25 <- prop.table(tab25,2)
addmargins(tab25) %>% pander(caption = "Relapse")
```

Table 33: Relapse

| | no | yes | Sum |
|---|---|---|---|
| **Died** | 0.3899 | 0.8214 | 1.211 |
| **Survived** | 0.6101 | 0.1786 | 0.7886 |
| **Sum** | 1 | 1 | 2 |

```
tab25 <- tab25 %>% data.frame()

p25 <- ggplot(aes(Var2, Freq, fill = Var1), data = tab25) +
  geom_col(position = "dodge") +
  xlab("Relapse") +
  ylab ("Density (%)") +
  labs(fill='status') +
  scale_fill_manual(values=c("#990000", "burlywood2"))+
  theme(legend.position="top")

grid.arrange(p10,p11,p12,p13,p14,p15,p16,p17,p18,p19,p20,p21,p22,p23,p24,p25,
             nrow = 4, ncol = 4)
```




Looking at these resulting graphs and tables we see some interesting relationships that we will want to explore further. Starting with our first comparison:

(1) donor_age_below_35 (yes/no) - There was a slightly better chance of survival if the donor was below 35 years old compared to if the donor was above 35 years old

(2)donor_CMV (present/absent) - Not a huge association of survival status with the presence versus the

absence of CMV in the donor

(3) recipient_age_below_10 (yes/no) - We saw here that there was a slightly increased chance of survival of the recipient was under 10 years old compared to the recipient being under 10 years old

(4) recipient_rh (plus/minus) - We found a large association with positive survival chances if the recipient had a minus rh presence compared to a plus rh presence (ex: blood type O- or AB- vs. O+)

(5)recipient_CMV (present/absent) - Not a huge association of survival status with the presence or absence of CMV in the recipient

(6)disease_group (malignant/nonmalignant) - In these plots we see having a nonmalignant disease had a better chance of survival than a malignant disease

(7) gender_match (female_to_male/other) - This variable showed no difference in survival status whether donor and recipient were of the same gender or different genders

(8) ABO_match (matched/mismatched) - Having a mismatched blood group had a better chance of survival than a matched blood type, which is very interesting since one would think having a matched blood type would increase survival probability. So, we will definitely look into this variable more

(9) HLA_mismatch (matched, mismatched) - Based of HLA match or mismatch, survival outcome was not significantly different.

(10) risk_group (high/low) - As expected, the lower risk group has a better chance of survival than the higher risk group

(11) stem_cell_source (peripheral_blood/bone_marrow) - We observed that a stem cell source from peripheral blood showed a much better chance for survival compared to a stem cell source from bone marrow. This is one of our secondary questions in this study, so we will come back to this association later on

(12) tx_post_relapse (yes/no) - As expected, there was an increased risk of death among those who were treated a second time after relapse compared to those who did not get treated after their first relapse.

(13) acute_GvHD_II_III_IV (yes/no) - There was not a significant difference in survival chances if acute GVHD was present in stage II, III or IV among the recipients

(14) acute_GvHD_III_IV (yes/no) - The same holds for this variable as above

(15) extensive_chronic_GvHD (yes/no) - However, for this variable, there was a much higher chance for survival if there was no development of GvHD compared to those who developed chronic GvHD.

(16) relapse (yes/no) - And lastly, as expected, those who underwent relapse had a much higher probability of death compared to those who did not undergo relapse.

## Question 11

Project Attestation: No member of this group is using these data or same/similar questions in any other course or course project, at HSPH.