Survival Analysis in R

Ajinkya

2024-12-31

Step 1: Setting Up the Environment

```
# install.packages("flexsurv")
# installed.packages("survminer")
# install.packages("ggsurvplot")

# Load required libraries
library(survival)
```

```
## Warning: package 'survival' was built under R version 4.3.3
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.3
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
## Warning: package 'tidyr' was built under R version 4.3.3
```

```
## Warning: package 'dplyr' was built under R version 4.4.0
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to be
come errors
```

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.3.3
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(survminer)
```

```
## Warning: package 'survminer' was built under R version 4.3.3
```

```
## Loading required package: ggpubr
```

```
## Warning: package 'ggpubr' was built under R version 4.3.3
```

```
##
## Attaching package: 'survminer'
##
## The following object is masked from 'package:survival':
##
##     myeloma
```

```
library(flexsurv)
```

```
## Warning: package 'flexsurv' was built under R version 4.3.3
```

```
library(ggplot2)
library (ggcorrplot)
```

```
## Warning: package 'ggcorrplot' was built under R version 4.3.3
```

```
# Set the working directory (update this path as needed)
setwd("C:/Users/Ajinkyaa/OneDrive/Stata to R/New folder/Survival analysis")
```

Step 2: Load the dataset

```
dialysis <- read_csv("C:/Users/Ajinkyaa/OneDrive/Stata to R/New folder/Survival analysis/dial
ysis survival dataset.csv")
```

```
## Rows: 6805 Columns: 9
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## dbl (9): event, time, age, begin, center, disease_diabetes, disease_hypert, ...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Step 3: Data Inspection and Cleaning

```
# Preview the data
head(dialysis)
```

```
## # A tibble: 6 × 9
##   event  time   age begin center disease_diabetes disease_hypert disease_other
##   <dbl> <dbl> <dbl> <dbl>  <dbl>            <dbl>          <dbl>         <dbl>
## 1     0     1    59    35    120                0              1             0
## 2     0     3    49    38    120                0              1             0
## 3     0    18    49    22    120                0              0             1
## 4     0     2    52    21    120                0              1             0
## 5     1     1    89    41    120                0              1             0
## 6     1     3    72    33    120                0              0             0
## # ℹ 1 more variable: disease_renal <dbl>
```

```
str(dialysis)
```

```
## spc_tbl_ [6,805 × 9] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ event           : num [1:6805] 0 0 0 0 1 1 1 1 1 1 ...
##  $ time            : num [1:6805] 1 3 18 2 1 3 7 16 2 18 ...
##  $ age             : num [1:6805] 59 49 49 52 89 72 49 31 47 34 ...
##  $ begin           : num [1:6805] 35 38 22 21 41 33 24 10 37 22 ...
##  $ center          : num [1:6805] 120 120 120 120 120 120 120 120 120 120 ...
##  $ disease_diabetes: num [1:6805] 0 0 0 0 0 0 0 0 0 0 ...
##  $ disease_hypert  : num [1:6805] 1 1 0 1 1 0 0 0 0 1 ...
##  $ disease_other   : num [1:6805] 0 0 1 0 0 0 0 0 1 0 ...
##  $ disease_renal   : num [1:6805] 0 0 0 0 0 1 1 1 0 0 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   event = col_double(),
##   ..   time = col_double(),
##   ..   age = col_double(),
##   ..   begin = col_double(),
##   ..   center = col_double(),
##   ..   disease_diabetes = col_double(),
##   ..   disease_hypert = col_double(),
##   ..   disease_other = col_double(),
##   ..   disease_renal = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
summary(dialysis)
```

```
##      event             time            age           begin           center
##  Min.   :0.0000   Min.   : 1.00   Min.   : 0.0   Min.   : 1.00   Min.   : 120
##  1st Qu.:0.0000   1st Qu.: 3.00   1st Qu.:42.0   1st Qu.:12.00   1st Qu.:1039
##  Median :0.0000   Median :11.00   Median :53.0   Median :23.00   Median :2026
##  Mean   :0.2356   Mean   :14.16   Mean   :52.7   Mean   :22.78   Mean   :2553
##  3rd Qu.:0.0000   3rd Qu.:22.00   3rd Qu.:65.0   3rd Qu.:33.00   3rd Qu.:4163
##  Max.   :1.0000   Max.   :44.00   Max.   :97.0   Max.   :44.00   Max.   :5768
##  disease_diabetes disease_hypert   disease_other    disease_renal
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.1885   Mean   :0.4168   Mean   :0.1661   Mean   :0.2078
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
```

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

```
##            event             time              age            begin
##                0                0                0                0
##           center disease_diabetes   disease_hypert    disease_other
##                0                0                0                0
##    disease_renal
##                0
```

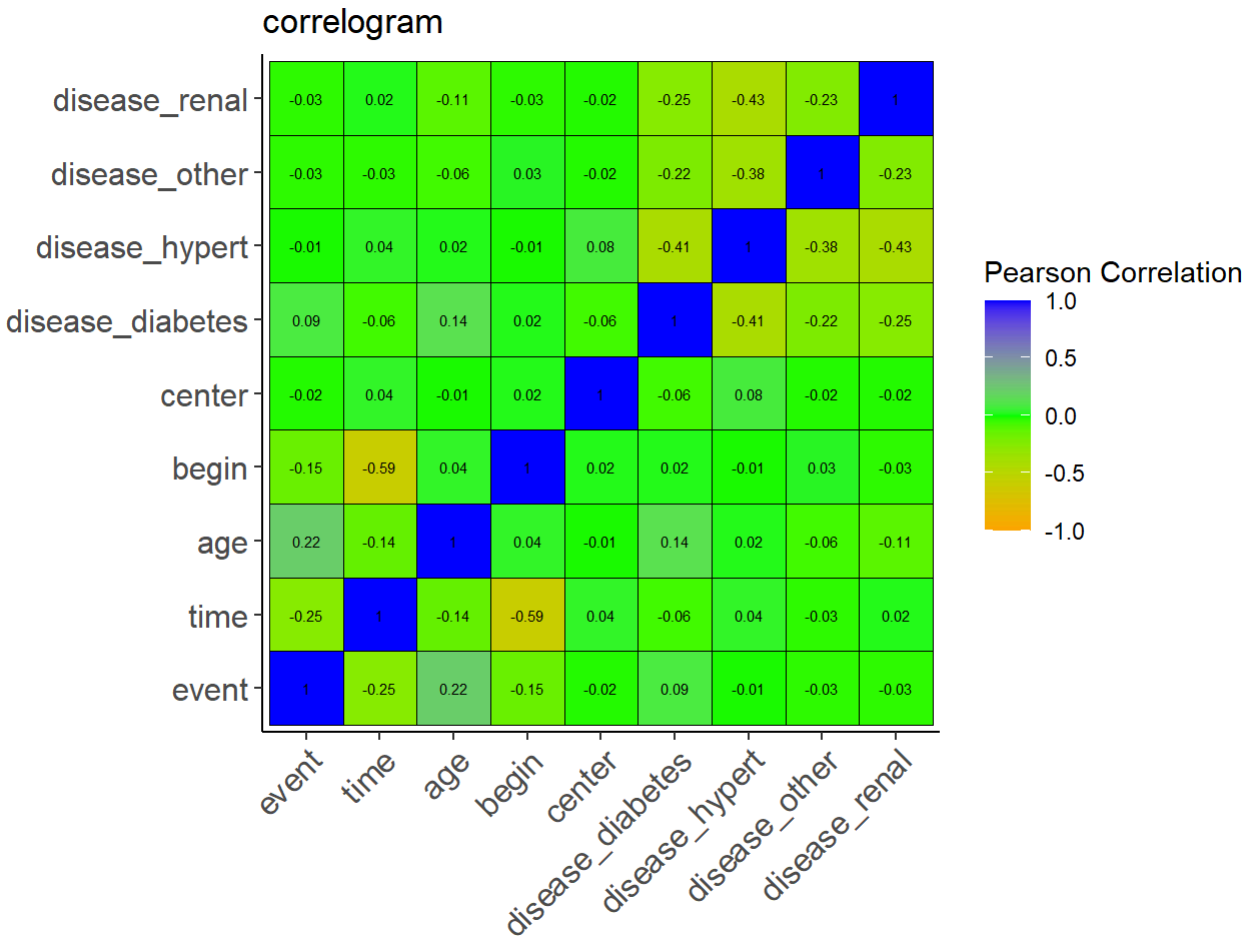
```
# Impute or remove missing values (example: removing rows with missing data)
dialysis <- dialysis %>% drop_na()
```

Step 4: (Exploratory Data Analysis (EDA)) Data Distribution Visualizations

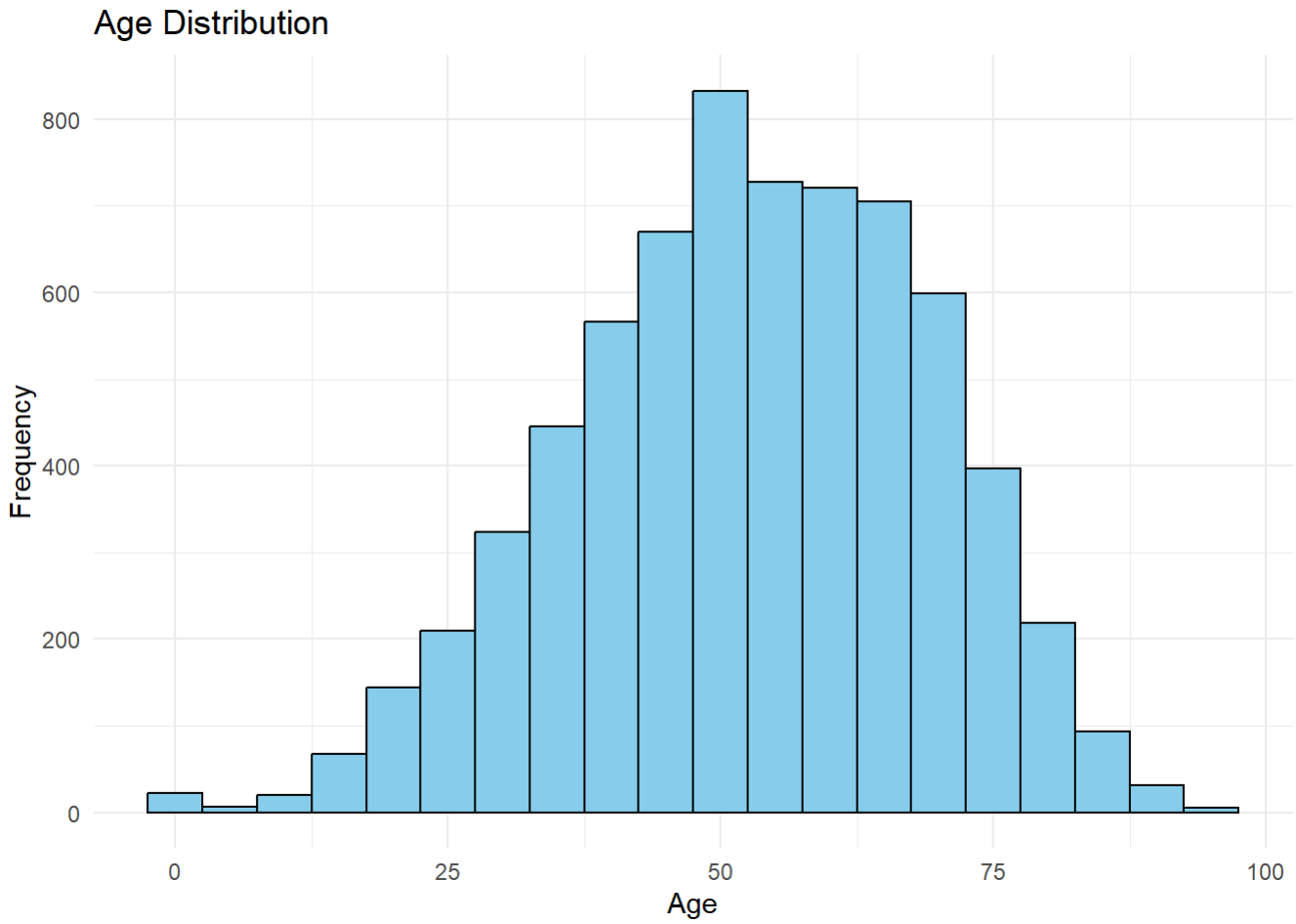
```
cr <- round(cor(dialysis), 2) #Store correlation matrix
cr
```

```
##                 event  time   age begin center disease_diabetes disease_hypert
## event            1.00 -0.25  0.22 -0.15  -0.02             0.09          -0.01
## time            -0.25  1.00 -0.14 -0.59   0.04            -0.06           0.04
## age              0.22 -0.14  1.00  0.04  -0.01             0.14           0.02
## begin           -0.15 -0.59  0.04  1.00   0.02             0.02          -0.01
## center          -0.02  0.04 -0.01  0.02   1.00            -0.06           0.08
## disease_diabetes 0.09 -0.06  0.14  0.02  -0.06             1.00          -0.41
## disease_hypert  -0.01  0.04  0.02 -0.01   0.08            -0.41           1.00
## disease_other   -0.03 -0.03 -0.06  0.03  -0.02            -0.22          -0.38
## disease_renal   -0.03  0.02 -0.11 -0.03  -0.02            -0.25          -0.43
##                 disease_other disease_renal
## event                   -0.03         -0.03
## time                    -0.03          0.02
## age                     -0.06         -0.11
## begin                    0.03         -0.03
## center                  -0.02         -0.02
## disease_diabetes        -0.22         -0.25
## disease_hypert          -0.38         -0.43
## disease_other            1.00         -0.23
## disease_renal           -0.23          1.00
```

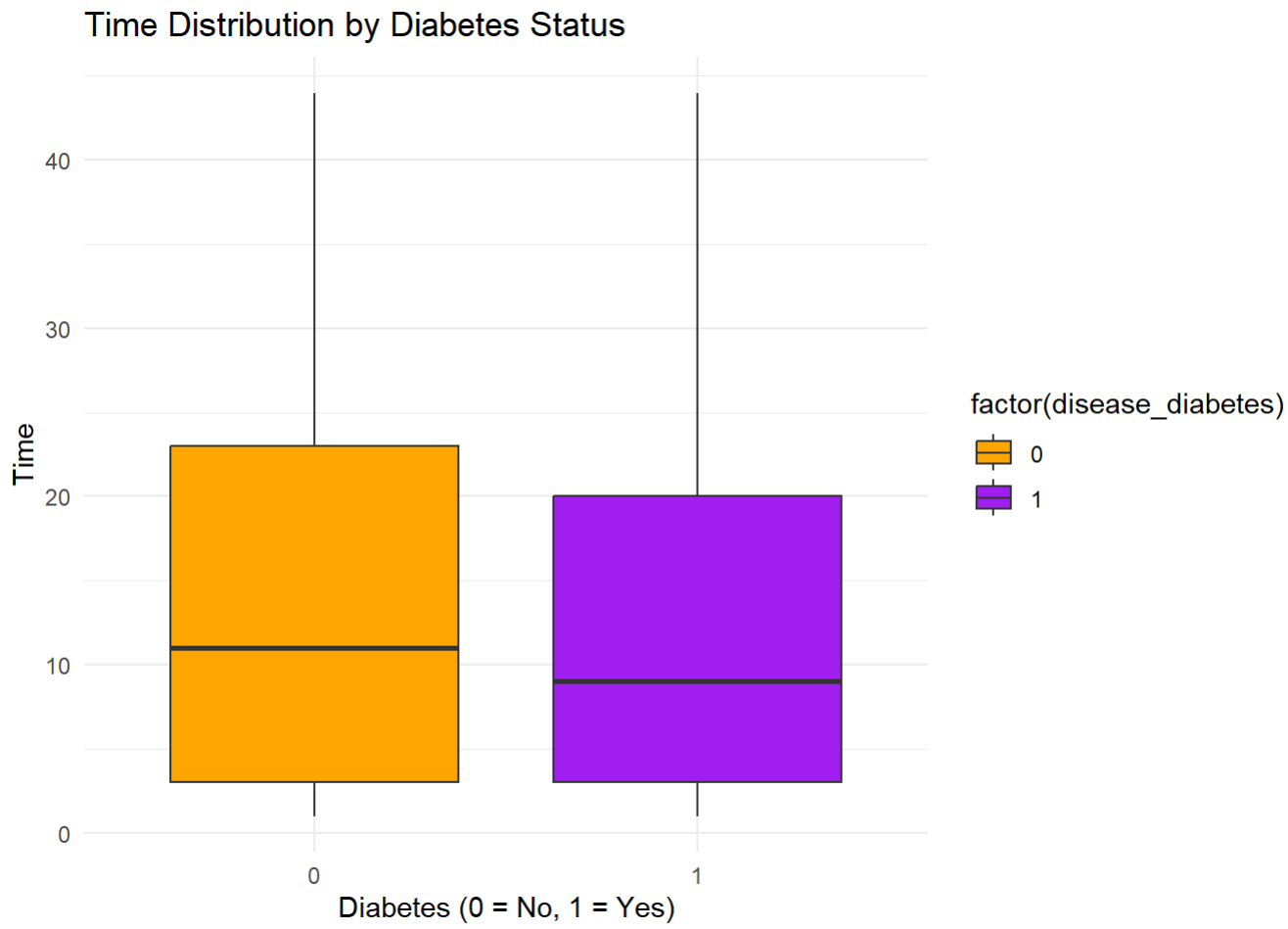
```
#Visualize your correlations (Creates a visual heatmap of the correlation matrix)
ggcorrplot(cr,title = "correlogram", lab_col = "black",
           lab = TRUE, legend.title = "Pearson Correlation",
           lab_size=2, ggtheme = theme_classic(),
           outline.color = "black",
           colors = c("orange", "green", "blue"))
```



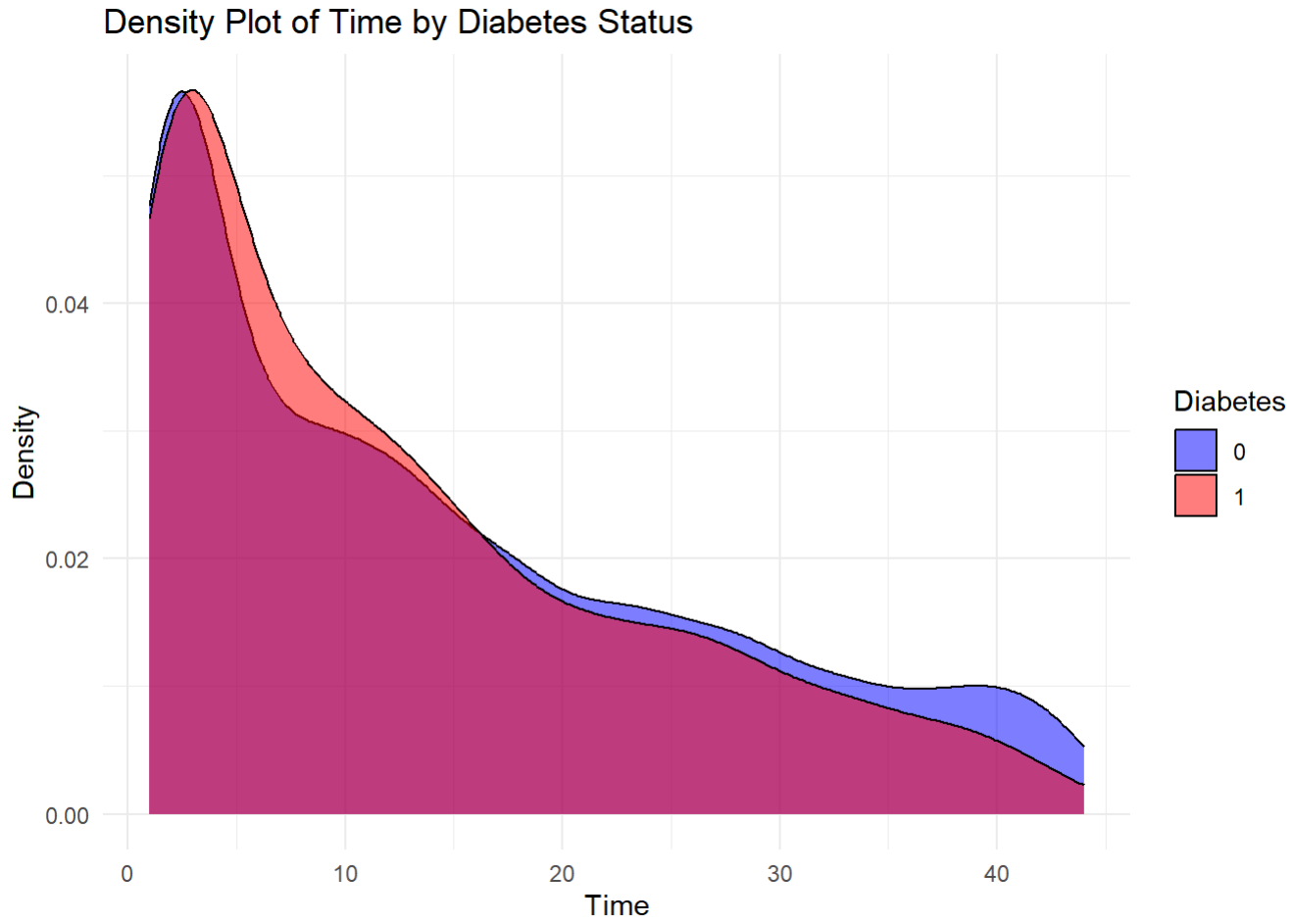
```
# 1. Histogram for Age
ggplot(dialysis, aes(x = age)) +
  geom_histogram(binwidth = 5, fill = "skyblue", color = "black") +
  labs(title = "Age Distribution", x = "Age", y = "Frequency") +
  theme_minimal()
```



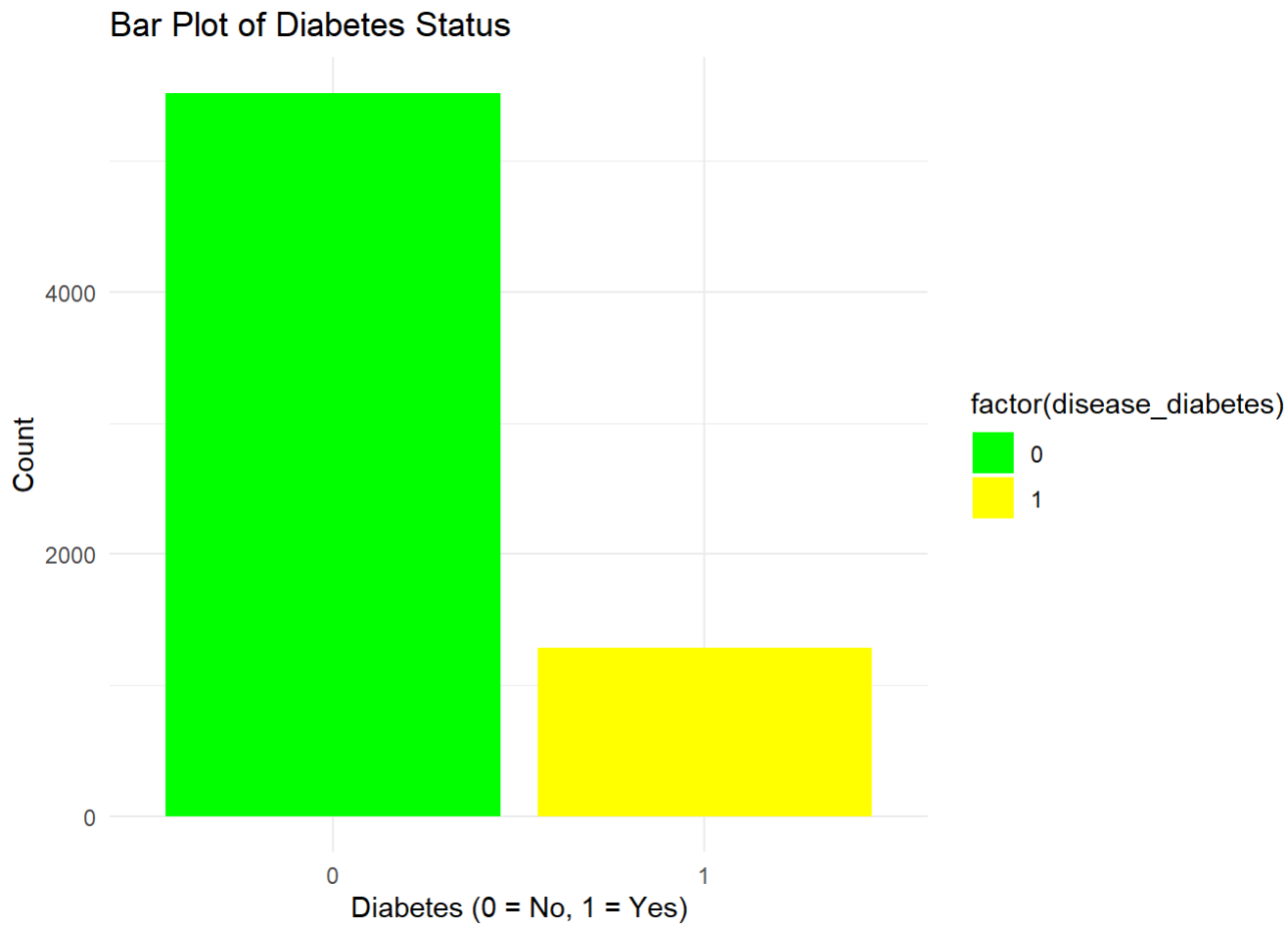

```
# 2. Boxplot for Time by Diabetes Status
ggplot(dialysis, aes(x = factor(disease_diabetes), y = time, fill = factor(disease_diabete
s))) +
  geom_boxplot() +
  labs(title = "Time Distribution by Diabetes Status", x = "Diabetes (0 = No, 1 = Yes)", y =
"Time") +
  scale_fill_manual(values = c("orange", "purple")) +
  theme_minimal()
```



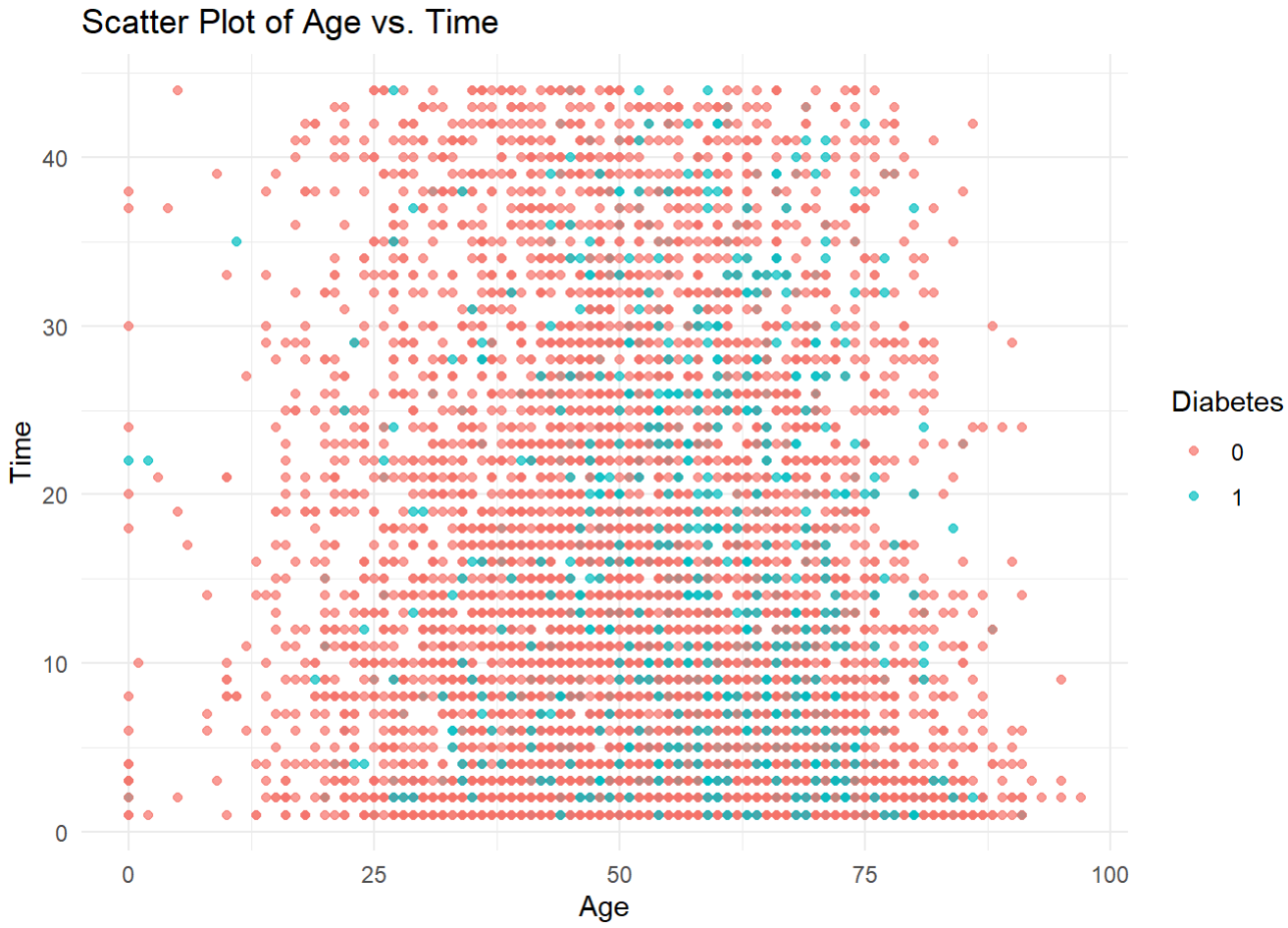

```
# 3. Density Plot for Time
ggplot(dialysis, aes(x = time, fill = factor(disease_diabetes))) +
  geom_density(alpha = 0.5) +
  labs(title = "Density Plot of Time by Diabetes Status", x = "Time", y = "Density", fill =
"Diabetes") +
  scale_fill_manual(values = c("blue", "red")) +
  theme_minimal()
```



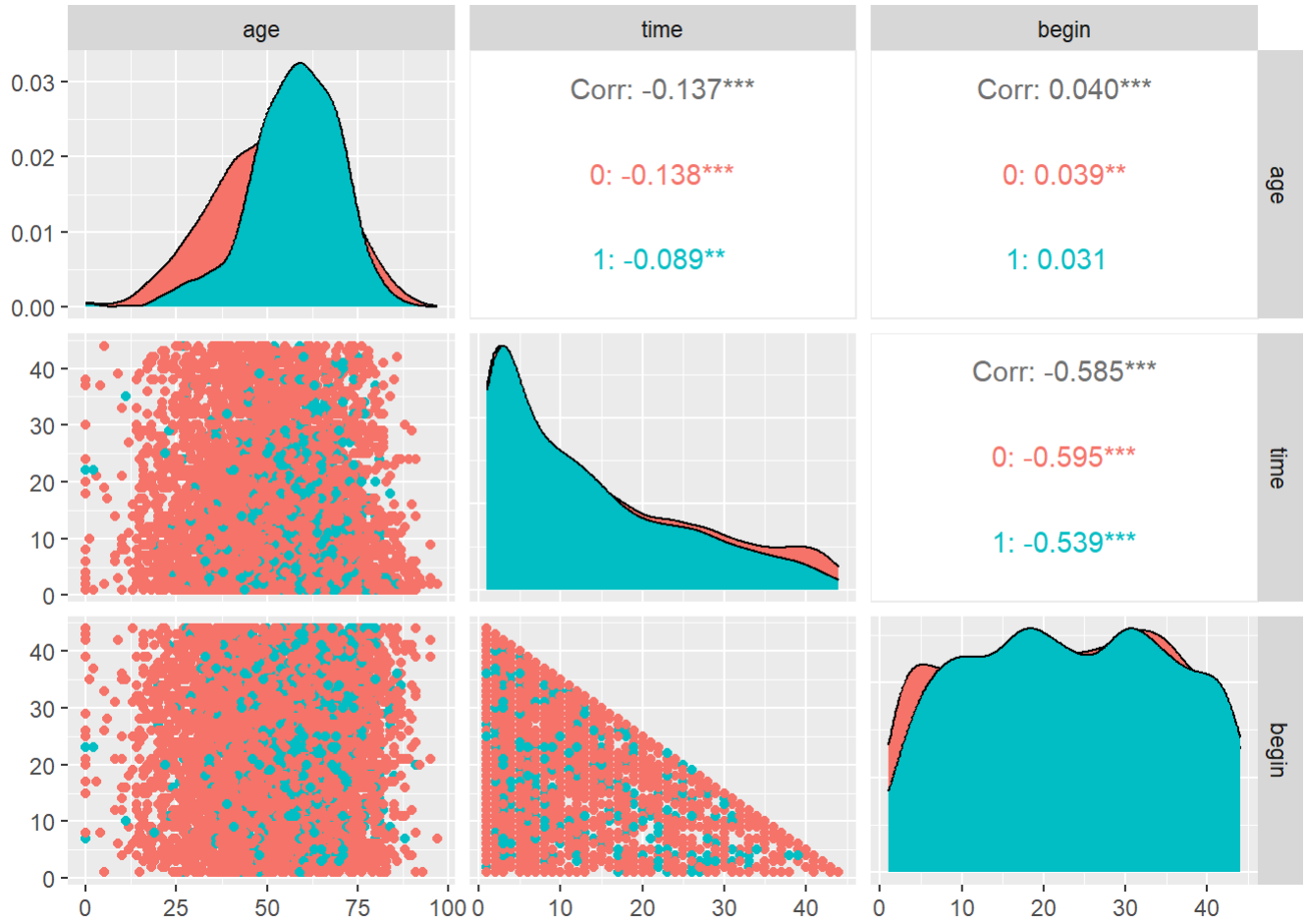

```
# 4. Bar Plot for Diabetes Status
ggplot(dialysis, aes(x = factor(disease_diabetes), fill = factor(disease_diabetes))) +
  geom_bar() +
  labs(title = "Bar Plot of Diabetes Status", x = "Diabetes (0 = No, 1 = Yes)", y = "Count")
+
  scale_fill_manual(values = c("green", "yellow")) +
  theme_minimal()
```



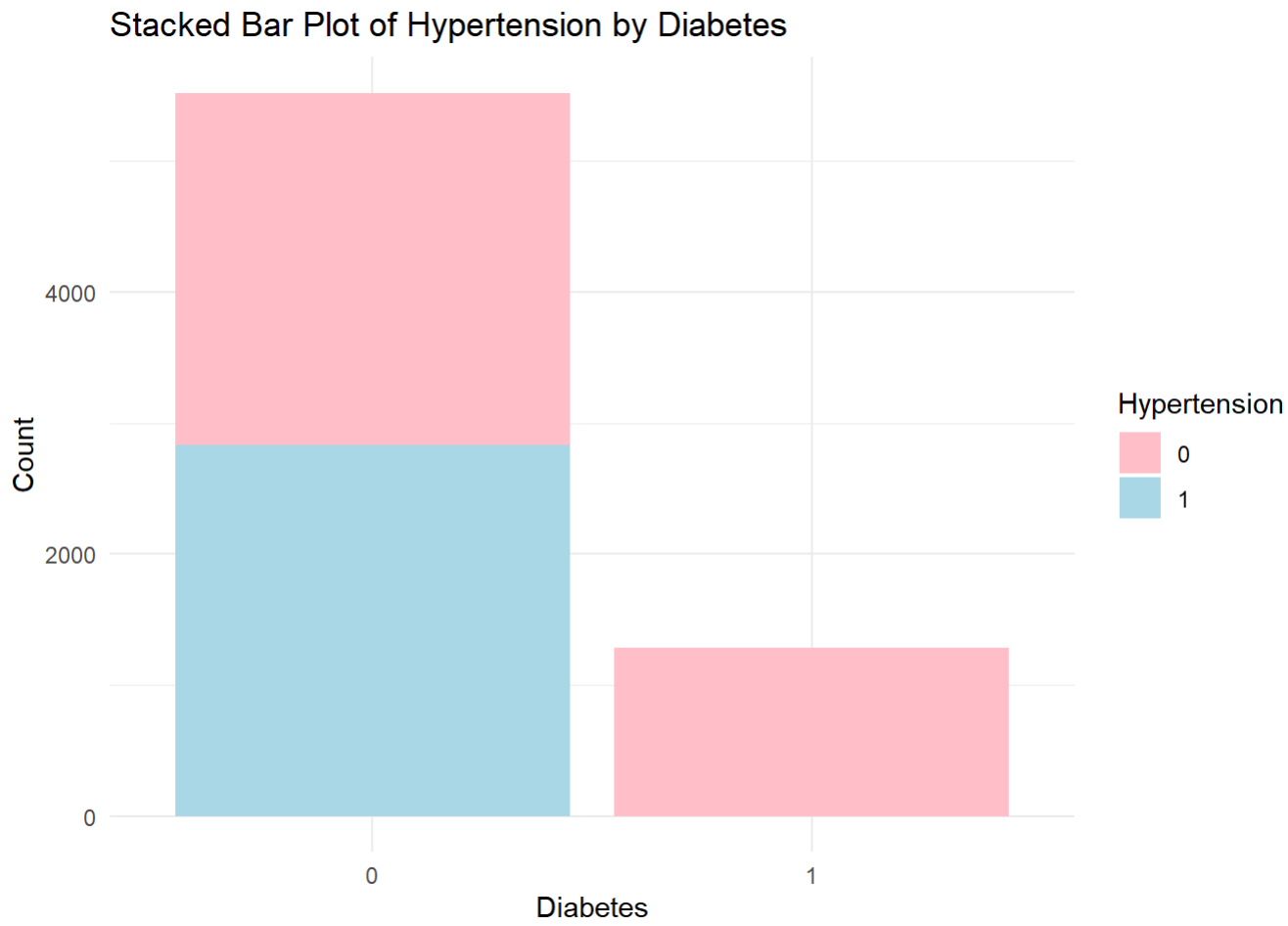

```
# 5. Scatter Plot for Age vs. Time
ggplot(dialysis, aes(x = age, y = time, color = factor(disease_diabetes))) +
  geom_point(alpha = 0.7) +
  labs(title = "Scatter Plot of Age vs. Time", x = "Age", y = "Time", color = "Diabetes") +
  theme_minimal()
```




```
# 6. Pairwise Plot for Selected Variables
library(GGally)
ggpairs(dialysis, columns = c("age", "time", "begin"), aes(color = factor(disease_diabetes)))
```




```
# 7. Stacked Bar Plot
ggplot(dialysis, aes(x = factor(disease_diabetes), fill = factor(disease_hypert))) +
  geom_bar(position = "stack") +
  labs(title = "Stacked Bar Plot of Hypertension by Diabetes", x = "Diabetes", y = "Count", f
ill = "Hypertension") +
  scale_fill_manual(values = c("pink", "lightblue")) +
  theme_minimal()
```



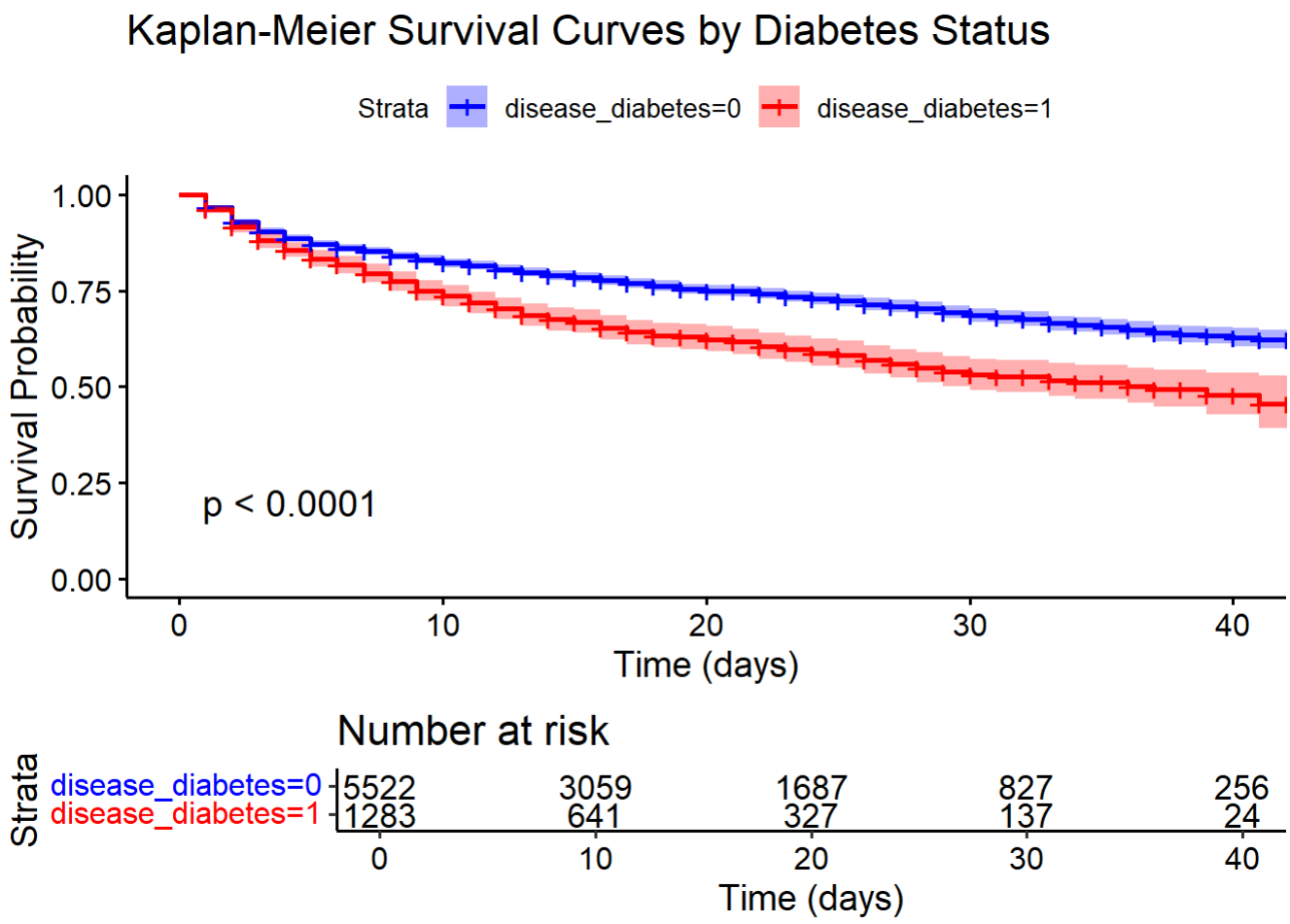
Step 5: Kaplan-Meier Model

```
# Fit Kaplan-Meier model
km_model <- survfit(Surv(time, event) ~ disease_diabetes, data = dialysis)

# Summary of the model
summary(km_model)
```

```
## Call: survfit(formula = Surv(time, event) ~ disease_diabetes, data = dialysis)
##
##                disease_diabetes=0
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1   5522     181    0.967 0.00240        0.963        0.972
##     2   5002     192    0.930 0.00349        0.923        0.937
##     3   4473     123    0.905 0.00409        0.897        0.913
##     4   4127      84    0.886 0.00447        0.877        0.895
##     5   3898      65    0.871 0.00476        0.862        0.881
##     6   3715      49    0.860 0.00497        0.850        0.870
##     7   3551      32    0.852 0.00511        0.842        0.862
##     8   3408      49    0.840 0.00533        0.829        0.850
##     9   3209      34    0.831 0.00549        0.820        0.842
##    10   3059      34    0.822 0.00565        0.811        0.833
##    11   2871      22    0.815 0.00576        0.804        0.827
##    12   2729      32    0.806 0.00594        0.794        0.818
##    13   2570      26    0.798 0.00609        0.786        0.810
##    14   2401      27    0.789 0.00626        0.777        0.801
##    15   2263      11    0.785 0.00634        0.773        0.797
##    16   2151      23    0.776 0.00651        0.764        0.789
##    17   2031      21    0.768 0.00667        0.756        0.782
##    18   1906      15    0.762 0.00680        0.749        0.776
##    19   1797      20    0.754 0.00698        0.740        0.768
##    20   1687       9    0.750 0.00707        0.736        0.764
##    21   1603       4    0.748 0.00712        0.734        0.762
##    22   1527      13    0.742 0.00727        0.728        0.756
##    23   1427      14    0.734 0.00746        0.720        0.749
##    24   1326      11    0.728 0.00762        0.714        0.743
##    25   1240      10    0.722 0.00778        0.707        0.738
##    26   1161      13    0.714 0.00801        0.699        0.730
##    27   1069       8    0.709 0.00817        0.693        0.725
##    28    996       7    0.704 0.00833        0.688        0.721
##    29    915      13    0.694 0.00866        0.677        0.711
##    30    827      11    0.685 0.00898        0.667        0.703
##    31    756       5    0.680 0.00914        0.663        0.698
##    32    718       4    0.676 0.00929        0.659        0.695
##    33    640      11    0.665 0.00977        0.646        0.684
##    34    574       3    0.661 0.00992        0.642        0.681
##    35    528       5    0.655 0.01021        0.635        0.675
##    36    476       5    0.648 0.01056        0.628        0.669
##    37    418       6    0.639 0.01107        0.618        0.661
##    38    373       2    0.635 0.01127        0.614        0.658
##    39    314       1    0.633 0.01142        0.611        0.656
##    40    256       2    0.629 0.01185        0.606        0.652
##    41    202       2    0.622 0.01253        0.598        0.647
##    43     90       1    0.615 0.01417        0.588        0.644
##
##                disease_diabetes=1
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1   1283      48    0.963 0.00530        0.952        0.973
##     2   1166      56    0.916 0.00786        0.901        0.932
##     3   1030      41    0.880 0.00939        0.862        0.898
##     4    935      25    0.856 0.01025        0.837        0.877
##     5    875      24    0.833 0.01103        0.812        0.855
##     6    819      15    0.818 0.01151        0.795        0.840
```

```
##     7    775      21    0.795 0.01217        0.772        0.820
##     8    735      20    0.774 0.01277        0.749        0.799
##     9    683      21    0.750 0.01339        0.724        0.777
##    10    641      12    0.736 0.01374        0.710        0.763
##    11    606      14    0.719 0.01415        0.692        0.747
##    12    563      13    0.702 0.01455        0.674        0.731
##    13    523      12    0.686 0.01495        0.658        0.716
##    14    484       8    0.675 0.01523        0.646        0.705
##    15    450       4    0.669 0.01538        0.639        0.700
##    16    426      10    0.653 0.01580        0.623        0.685
##    17    390       7    0.641 0.01613        0.611        0.674
##    18    366       5    0.633 0.01638        0.601        0.666
##    19    342       2    0.629 0.01649        0.598        0.662
##    20    327       3    0.623 0.01667        0.591        0.657
##    21    303       3    0.617 0.01688        0.585        0.651
##    22    282       6    0.604 0.01735        0.571        0.639
##    23    264       3    0.597 0.01760        0.564        0.633
##    24    244       4    0.587 0.01798        0.553        0.624
##    25    226       2    0.582 0.01820        0.548        0.619
##    26    210       5    0.568 0.01879        0.533        0.606
##    27    186       3    0.559 0.01922        0.523        0.598
##    28    168       3    0.549 0.01972        0.512        0.589
##    29    154       3    0.538 0.02028        0.500        0.580
##    30    137       2    0.531 0.02073        0.491        0.573
##    31    121       1    0.526 0.02102        0.487        0.569
##    33    101       2    0.516 0.02185        0.475        0.560
##    34     86       1    0.510 0.02241        0.468        0.556
##    36     63       1    0.502 0.02347        0.458        0.550
##    37     56       1    0.493 0.02470        0.447        0.544
##    39     36       1    0.479 0.02754        0.428        0.536
##    41     20       1    0.455 0.03507        0.391        0.529
```

```
# Kaplan-Meier survival plot
ggsurvplot(km_model,
           data = dialysis,
           conf.int = TRUE,
           risk.table = TRUE,
           pval = TRUE,
           title = "Kaplan-Meier Survival Curves by Diabetes Status",
           xlab = "Time (days)",
           ylab = "Survival Probability",
           palette = c("blue", "red"))
```



Step 6: Log-Rank Test

```
# Test survival differences between groups
log_rank <- survdiff(Surv(time, event) ~ disease_diabetes, data = dialysis)
log_rank
```

```
## Call:
## survdiff(formula = Surv(time, event) ~ disease_diabetes, data = dialysis)
##
##                         N Observed Expected (O-E)^2/E (O-E)^2/V
## disease_diabetes=0 5522     1200     1320      10.9      63.1
## disease_diabetes=1 1283      403      283      50.7      63.1
##
##  Chisq= 63.1  on 1 degrees of freedom, p= 2e-15
```

Step 7: Cox Proportional Hazards Model

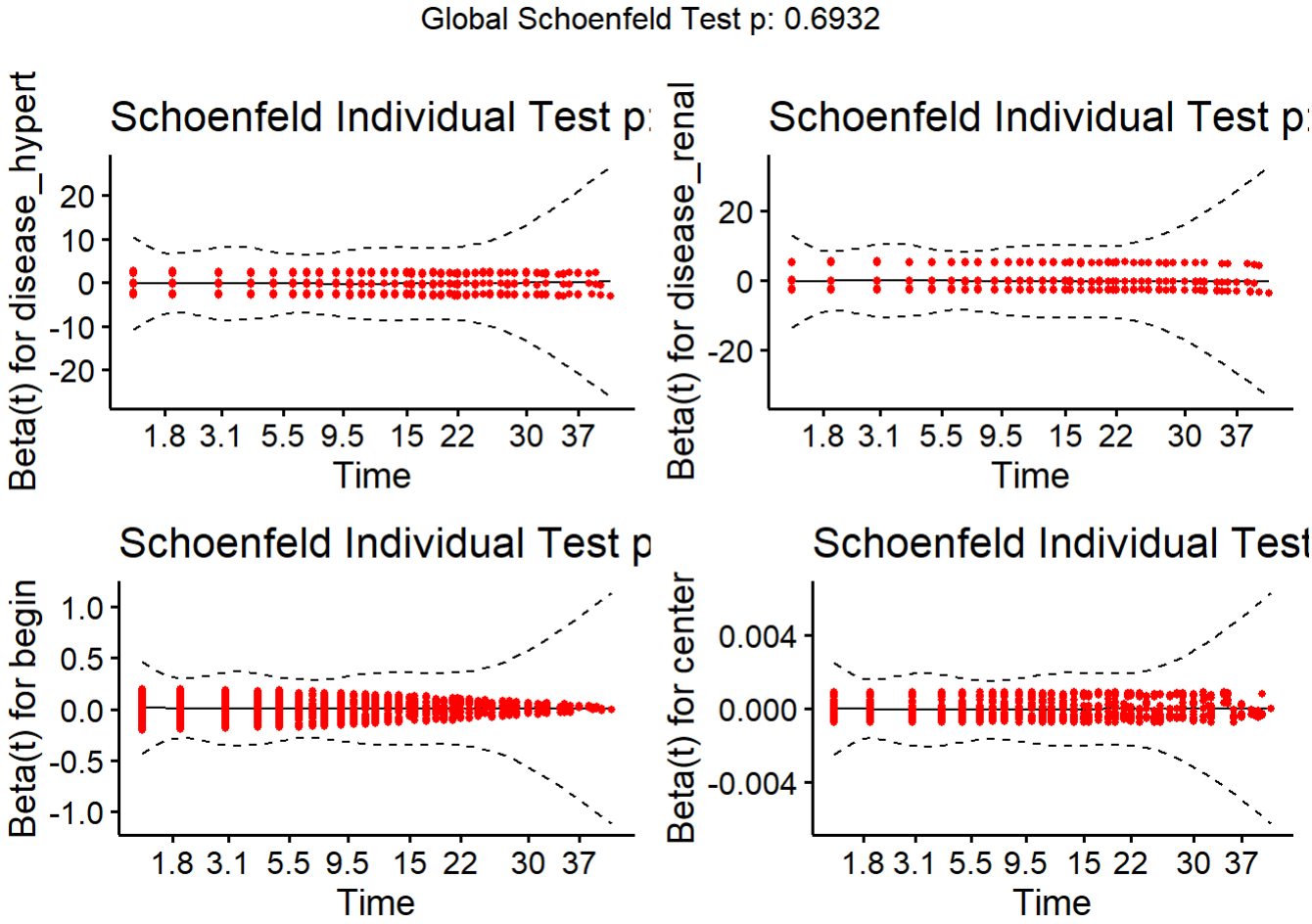
```
cox_model <- coxph(Surv(time, event) ~ disease_hypert + disease_renal + begin + center, data
= dialysis)

# Summary of the model
summary(cox_model)
```

```
## Call:
## coxph(formula = Surv(time, event) ~ disease_hypert + disease_renal +
##     begin + center, data = dialysis)
##
##   n= 6805, number of events= 1603
##
##                       coef  exp(coef)   se(coef)      z Pr(>|z|)
## disease_hypert -1.905e-01  8.265e-01  5.556e-02 -3.429 0.000606 ***
## disease_renal  -2.489e-01  7.796e-01  6.988e-02 -3.562 0.000368 ***
## begin           7.592e-03  1.008e+00  2.366e-03  3.209 0.001333 **
## center         -2.524e-05  1.000e+00  1.321e-05 -1.910 0.056130 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                exp(coef) exp(-coef) lower .95 upper .95
## disease_hypert    0.8265     1.2099    0.7412    0.9216
## disease_renal     0.7796     1.2826    0.6798    0.8941
## begin             1.0076     0.9924    1.0030    1.0123
## center            1.0000     1.0000    0.9999    1.0000
##
## Concordance= 0.544  (se = 0.008 )
## Likelihood ratio test= 32.98  on 4 df,   p=1e-06
## Wald test            = 33.2  on 4 df,   p=1e-06
## Score (logrank) test = 33.3  on 4 df,   p=1e-06
```

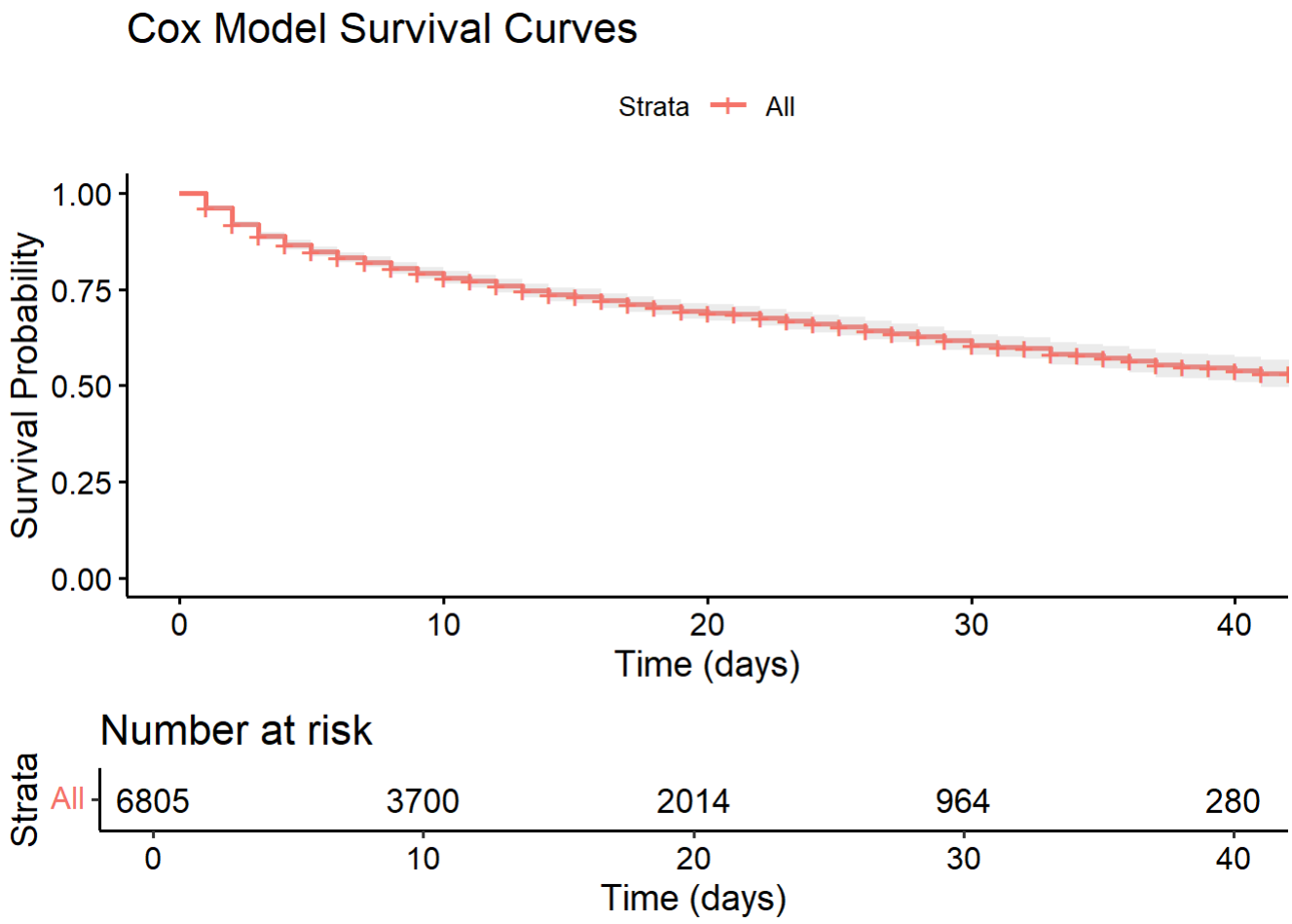
```
# Test proportional hazards assumption
ph_test <- cox.zph(cox_model)
ph_test
```

```
##                  chisq df    p
## disease_hypert 0.00675  1 0.93
## disease_renal  0.69199  1 0.41
## begin          1.31572  1 0.25
## center         0.05019  1 0.82
## GLOBAL         2.23212  4 0.69
```

```
# Visualize Schoenfeld residuals
ggcoxzph(ph_test)
```




```
# Plot survival curves based on Cox model
ggsurvplot(survfit(cox_model),
           data = dialysis,
           conf.int = TRUE,
           risk.table = TRUE,
           title = "Cox Model Survival Curves",
           xlab = "Time (days)",
           ylab = "Survival Probability")
```



Step 8: Parametric Survival Models

```
# 1. Exponential Model
exp_model <- flexsurvreg(Surv(time, event) ~ disease_hypert + disease_renal + begin + center,
                         data = dialysis, dist = "exponential")
summary(exp_model)
```

```
## disease_hypert=0.416752387950037,disease_renal=0.207788390889052,begin=22.7775165319618,ce
nter=2553.10639235856
##    time       est       lcl       ucl
## 1     1 0.9816002 0.9805447 0.9825909
## 2     2 0.9635390 0.9614679 0.9654848
## 3     3 0.9458101 0.9427622 0.9486766
## 4     4 0.9284075 0.9244205 0.9321610
## 5     5 0.9113250 0.9064356 0.9159329
## 6     6 0.8945568 0.8888006 0.8999873
## 7     7 0.8780972 0.8715087 0.8843193
## 8     8 0.8619404 0.8545532 0.8689241
## 9     9 0.8460809 0.8379276 0.8537969
## 10   10 0.8305132 0.8216255 0.8389330
## 11   11 0.8152320 0.8056405 0.8243279
## 12   12 0.8002319 0.7899665 0.8099771
## 13   13 0.7855078 0.7745974 0.7958761
## 14   14 0.7710547 0.7595274 0.7820206
## 15   15 0.7568675 0.7447506 0.7684063
## 16   16 0.7429413 0.7302612 0.7550291
## 17   17 0.7292713 0.7160537 0.7418847
## 18   18 0.7158529 0.7021227 0.7289691
## 19   19 0.7026814 0.6884627 0.7162784
## 20   20 0.6897522 0.6750684 0.7038086
## 21   21 0.6770610 0.6619347 0.6915560
## 22   22 0.6646032 0.6490566 0.6795166
## 23   23 0.6523747 0.6364290 0.6676868
## 24   24 0.6403711 0.6240471 0.6560630
## 25   25 0.6285884 0.6119060 0.6446415
## 26   26 0.6170226 0.6000012 0.6334188
## 27   27 0.6056695 0.5883280 0.6223916
## 28   28 0.5945253 0.5768819 0.6115563
## 29   29 0.5835862 0.5656585 0.6009096
## 30   30 0.5728484 0.5546534 0.5904483
## 31   31 0.5623081 0.5438624 0.5801691
## 32   32 0.5519617 0.5332814 0.5700689
## 33   33 0.5418058 0.5229063 0.5601445
## 34   34 0.5318367 0.5127330 0.5503929
## 35   35 0.5220510 0.5027576 0.5408110
## 36   36 0.5124454 0.4929763 0.5313960
## 37   37 0.5030165 0.4833853 0.5221449
## 38   38 0.4937611 0.4739808 0.5130548
## 39   39 0.4846761 0.4647594 0.5041229
## 40   40 0.4757581 0.4557174 0.4953466
## 41   41 0.4670043 0.4468512 0.4867231
## 42   42 0.4584115 0.4381576 0.4782496
## 43   43 0.4499769 0.4296331 0.4699237
## 44   44 0.4416974 0.4212745 0.4617428
```

```
# 2. Weibull Model
weibull_model <- flexsurvreg(Surv(time, event) ~ disease_hypert + disease_renal + begin + cen
ter,
                             data = dialysis, dist = "weibull")
summary(weibull_model)
```

```
## disease_hypert=0.416752387950037,disease_renal=0.207788390889052,begin=22.7775165319618,ce
nter=2553.10639235856
##    time       est       lcl       ucl
## 1     1 0.9719588 0.9685009 0.9749876
## 2     2 0.9502827 0.9455100 0.9546774
## 3     3 0.9307567 0.9249868 0.9360553
## 4     4 0.9126210 0.9063350 0.9186943
## 5     5 0.8955272 0.8886058 0.9021706
## 6     6 0.8792730 0.8718479 0.8864093
## 7     7 0.8637261 0.8558400 0.8712254
## 8     8 0.8487924 0.8405094 0.8567295
## 9     9 0.8344017 0.8258699 0.8428182
## 10   10 0.8204990 0.8113965 0.8293029
## 11   11 0.8070404 0.7974600 0.8161035
## 12   12 0.7939897 0.7840508 0.8035299
## 13   13 0.7813164 0.7710389 0.7911793
## 14   14 0.7689947 0.7582546 0.7792182
## 15   15 0.7570022 0.7456598 0.7678391
## 16   16 0.7453192 0.7335977 0.7564370
## 17   17 0.7339286 0.7217206 0.7455663
## 18   18 0.7228148 0.7102205 0.7351390
## 19   19 0.7119641 0.6988444 0.7248175
## 20   20 0.7013640 0.6878980 0.7148651
## 21   21 0.6910031 0.6771287 0.7050556
## 22   22 0.6808710 0.6665908 0.6954644
## 23   23 0.6709583 0.6561148 0.6858237
## 24   24 0.6612562 0.6460454 0.6762976
## 25   25 0.6517566 0.6361071 0.6669924
## 26   26 0.6424518 0.6266366 0.6580379
## 27   27 0.6333350 0.6170657 0.6493381
## 28   28 0.6243995 0.6076901 0.6408537
## 29   29 0.6156392 0.5985625 0.6325039
## 30   30 0.6070483 0.5895229 0.6243697
## 31   31 0.5986214 0.5805901 0.6162387
## 32   32 0.5903533 0.5719978 0.6083399
## 33   33 0.5822392 0.5634704 0.6007415
## 34   34 0.5742746 0.5548913 0.5931908
## 35   35 0.5664549 0.5464742 0.5857395
## 36   36 0.5587762 0.5382989 0.5783201
## 37   37 0.5512345 0.5303731 0.5711482
## 38   38 0.5438259 0.5224460 0.5641554
## 39   39 0.5365470 0.5148689 0.5572045
## 40   40 0.5293943 0.5076134 0.5503277
## 41   41 0.5223645 0.5001733 0.5435382
## 42   42 0.5154545 0.4928485 0.5368596
## 43   43 0.5086613 0.4856523 0.5302870
## 44   44 0.5019820 0.4785818 0.5238578
```

```
# 3. Compare Models
# Compare AIC values
model_comparison <- data.frame(
  Model = c("Exponential", "Weibull"),
  AIC = c(AIC(exp_model), AIC(weibull_model))
)
model_comparison
```

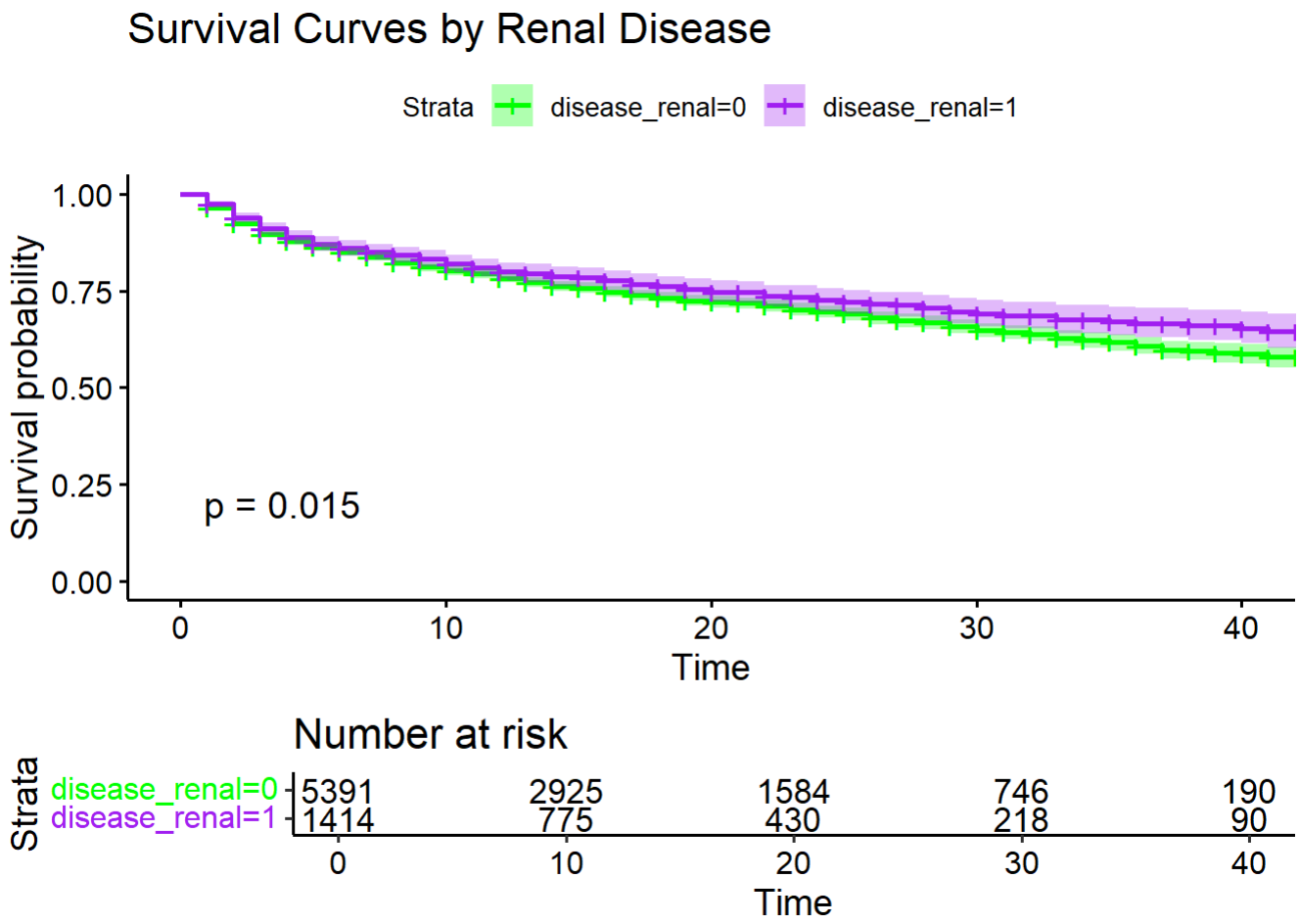
```
##         Model      AIC
## 1 Exponential 16209.46
## 2     Weibull 16148.71
```

Step 9: Advanced Analysis

```
# Predict survival probabilities at specific times using Weibull model
predict(weibull_model, newdata = dialysis, type = "survival", times = c(10, 20, 30))
```

```
## # A tibble: 6,805 × 1
##    .pred
##    <list>
##  1 <tibble [3 × 2]>
##  2 <tibble [3 × 2]>
##  3 <tibble [3 × 2]>
##  4 <tibble [3 × 2]>
##  5 <tibble [3 × 2]>
##  6 <tibble [3 × 2]>
##  7 <tibble [3 × 2]>
##  8 <tibble [3 × 2]>
##  9 <tibble [3 × 2]>
## 10 <tibble [3 × 2]>
## # ℹ 6,795 more rows
```

```
# Kaplan-Meier for subgroups analysis
km_model_subgroup <- survfit(Surv(time, event) ~ disease_renal, data = dialysis)
ggsurvplot(km_model_subgroup,
           data = dialysis,
           conf.int = TRUE,
           risk.table = TRUE,
           pval = TRUE,
           title = "Survival Curves by Renal Disease",
           palette = c("green", "purple"))
```



Step 10: Export Results

```
# Save plots and tables
ggsave("km_survival_plot.png")
```

```
## Saving 7 x 5 in image
```

```
write_csv(model_comparison, "model_comparison.csv")
```

Key Findings

1. Kaplan-Meier Analysis

Key Results

- **Survival Curves:**
  - Patients without diabetes have better survival probabilities than those with diabetes.
  - The curve for diabetic patients drops faster, meaning they are more likely to experience events (e.g., death) sooner.
- **Median Survival Time:**
  - Non-diabetic patients: Median survival time is not reached, indicating more than 50% of these patients survive throughout the study period.
  - Diabetic patients: Median survival time is around 37 days, meaning half of the diabetic patients die within this time.

Conclusion:
This analysis shows that diabetes significantly reduces survival during dialysis. Patients without diabetes generally live longer, and the gap between diabetic and non-diabetic groups widens over time.

2. Log-Rank Test

(The Log-Rank test checks if the survival differences between groups (e.g., diabetic vs. non-diabetic) are statistically significant)

Key Results

- **P-Value:**
  - The p-value is very small (2e-15), meaning there is a statistically significant difference in survival between the two groups.

Conclusion:
This test confirms that the difference in survival curves between diabetic and non-diabetic patients is not due to random chance. Diabetes has a real and measurable impact on survival outcomes.

3. Cox Proportional Hazards Model

(The Cox model estimates the effect of various factors (e.g., hypertension, renal disease) on the risk of death (hazard) while considering all factors together)

Key Results

- **Hypertension (`disease_hypert`):**
  - Patients with hypertension have a 17% lower risk of death than those without hypertension.
  - This suggests better survival for hypertensive patients, possibly due to effective treatment.
- **Renal Disease (`disease_renal`):**
  - Patients with renal disease have a 22% lower risk of death than those without renal disease.
  - This shows that renal disease patients, when managed well, can have better outcomes.
- **Start Time of Dialysis (`begin`):**
  - A later start in dialysis slightly increases the risk of death (by about 0.8% per unit of delay).
  - This indicates that starting dialysis earlier might improve survival.
- **Dialysis Center (`center`):**
  - The impact of the center is negligible, with minimal differences in outcomes across locations.

Conclusion:
- Hypertension and renal disease patients, when managed well, have better survival.
- Starting dialysis earlier can help improve survival chances.
- Where the treatment is provided (center) doesn’t significantly affect survival.

4. Parametric Survival Models

(These models (Exponential and Weibull) assume specific patterns for survival and hazard rates to provide more precise predictions)

Exponential Model Results:

- Assumes a constant hazard (risk of death) over time.
- Patients with hypertension and renal disease have lower risks of death.
- Starting dialysis later increases risk slightly.

Weibull Model Results:

- Accounts for time-varying hazards (risks change over time).

- Hazard decreases over time, meaning patients face higher risks earlier but stabilize later.
- Weibull model fits the data better (lower AIC value) than the Exponential model.

Conclusion:
The Weibull model shows that risks are not constant. Patients are at higher risk shortly after starting dialysis, but the risks reduce over time. This model is more realistic and accurate than the Exponential model.

5. Advanced Analysis

Predicting Survival Probabilities

- **Example Prediction:** Using the Weibull model, survival probabilities at 10, 20, and 30 days can be calculated.
- **Key Results:**
  - Survival probability decreases steadily over time:
    - At 10 days: ~83%
    - At 20 days: ~70%
    - At 30 days: ~57%

Conclusion:
This analysis shows how many patients are expected to survive beyond specific time points, helping doctors plan treatments accordingly.

Subgroup Analysis

- **Renal Disease Survival Curves:**
  - Patients with renal disease have slightly better survival compared to those without renal disease, as seen in their respective KM curves.

Conclusion:
Even within the group of dialysis patients, those with renal disease tend to fare better than others. This emphasizes the importance of managing such conditions effectively.

Summary of All Findings

1. **Kaplan-Meier & Log-Rank:** Diabetic patients face significantly worse survival outcomes than non-diabetic patients. The difference is statistically proven.
2. **Cox Model:**
   - Hypertension and renal disease are linked to better survival outcomes, likely due to effective management.
   - Starting dialysis earlier improves survival chances.
   - Dialysis center differences don’t matter much.
3. **Parametric Models:**
   - Weibull fits the data better, showing that risks decrease over time.
   - Early intervention is critical as risks are higher soon after starting dialysis.
4. **Predictions & Subgroups:** Survival probabilities at specific time points help in planning care. Renal disease patients show relatively better outcomes.