Survival Analysis Project:

Methodologies and Decisions 1. Objective and Data Preparation Objective: The goal of this project is to analyze customer churn using survival analysis techniques. We aim to:

Estimate survival functions for different customer groups. Compare survival functions across multiple groups. Perform a Cox proportional hazards regression to explore the impact of different factors on survival. Data Preparation:

Data Source: Customer churn data is loaded from a CSV file. Data Cleaning: Columns are standardized using clean_names() from the janitor package to ensure consistent naming. The churn column is converted to binary format (1 for "Yes" and 0 for "No") for analysis. Character columns are converted to factors to facilitate categorical analysis. 2. Nonparametric Estimation of Survival Kaplan-Meier Estimator: The Kaplan-Meier estimator is used to estimate the survival function from lifetime data. It is a nonparametric method that provides an estimate of the probability of surviving past certain time points, considering censoring.

Implementation:

```
sfit <- survfit(Surv(tenure, churn) ~ contract, data = customer_churn_tbl)
```

Surv(tenure, churn): Creates a survival object where tenure represents the time to event and c

contract: Stratifies survival curves by the contract variable to estimate survival functions for each contract type. Visualization:

Kaplan-Meier Plot with Tidyquant Theme:

```
g1 <- ggsurvplot(
    sfit,
    conf.int = TRUE,
    data = customer_churn_tbl,
    palette = c("#E69F00", "#56B4E9", "#009E73"),
    ggtheme = theme_tq(),  # Apply tidyquant theme
    title = "Customer Churn Survival Plot"
)
print(g1)
```

This plot shows survival curves for different contract types with confidence intervals. The tidyquant theme is applied for a clean, professional look. It Provides a clear picture of survival probabilities across different contract types, indicating varying levels of customer retention.

3. Nonparametric Comparison of Groups Adding a Risk Table: A risk table provides additional insight by showing the number of subjects at risk at different time points.

Implementation:

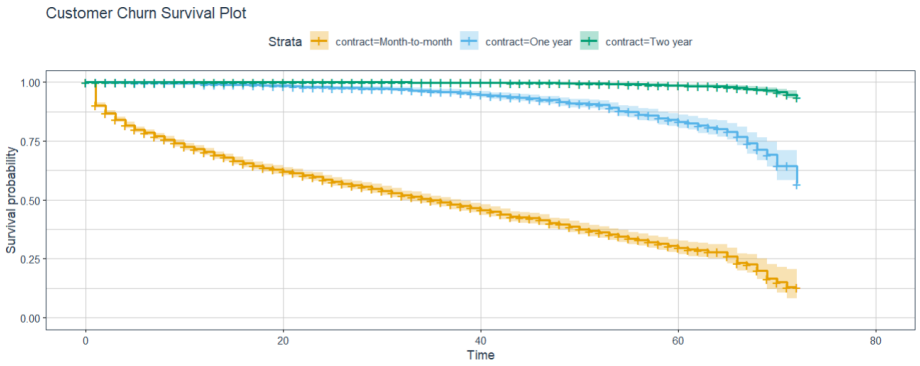

Figure 1: image

```
g2 <- ggsurvplot(
    sfit,
    conf.int = TRUE,
    data = customer_churn_tbl,
    risk.table = TRUE,    # Add risk table
    ggtheme = theme_tq(), # Apply tidyquant theme
    palette = c("#E69F00", "#56B4E9", "#009E73"),
    title = "Customer Churn Survival Plot with Risk Table"
)
```

Figure 2: image

risk.table = TRUE: Adds a table below the survival plot showing the number of individuals at risk at different time points. The table enhances the Kaplan-Meier plot by showing the number of customers at risk, adding valuable context to the survival analysis.

Customization and Combination:

```
g2_plot <- g2$plot +
    labs(title = "Customer Churn Survival Plot")

g2_table <- g2$table +
    theme_tq() +
    theme(panel.grid = element_blank())

combined_plot <- g2_plot / g2_table + plot_layout(heights = c(2, 1))
print(combined_plot)
```

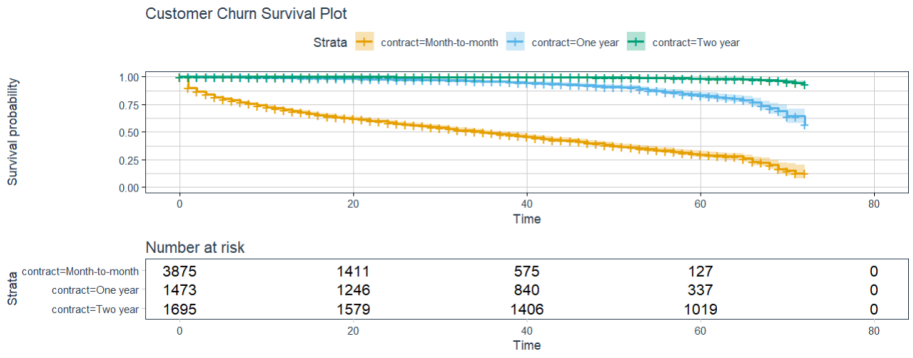

Figure 3: image

The plot and risk table are customized and combined using the patchwork package for comprehensive visualization.

4. Faceting by Groups Faceting: Faceting allows for the comparison of survival curves across different subgroups, such as gender, by creating separate panels.

Implementation:

```
unique_gender_levels <- unique(customer_churn_tbl$gender)
num_gender_levels <- length(unique_gender_levels)
palette_colors <- RColorBrewer::brewer.pal(n = num_gender_levels, name = "Set1")

g3 <- ggsurvplot_facet(
    sfit,
    conf.int = TRUE,
    data = customer_churn_tbl,
    facet.by = "gender",
    nrow = 1,  # Arrange facets in one row
    ggtheme = theme_tq(), # Apply tidyquant theme
    palette = palette_colors,
    title = "Survival Plot by Customer Gender"
)
print(g3)
```

facet.by = “gender”: Creates separate panels for each gender, facilitating a comparison of survival functions by gender. It allows for comparison of survival functions between genders, revealing potential differences in churn rates based on gender.

5. Semi-Parametric Cox Regression Cox Proportional Hazards Model: The Cox model is a semi-parametric method used to explore the relationship between survival time and one or more predictor variables. It provides

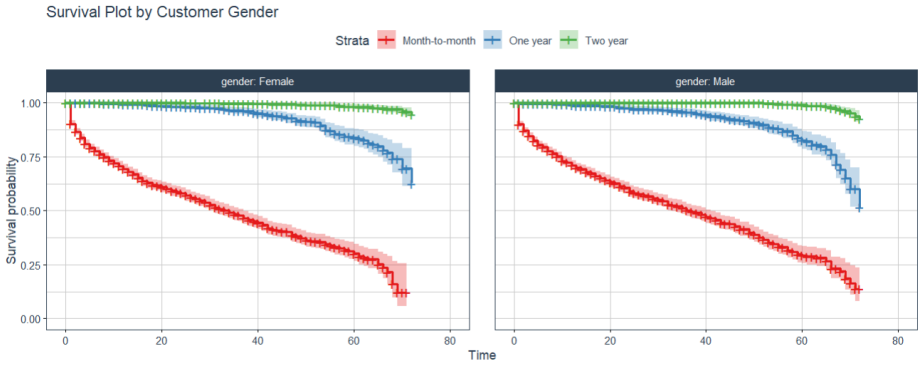

Figure 4: image

estimates of hazard ratios, which indicate the effect of covariates on the hazard or risk of the event occurring.

Implementation:

```
cox_model <- coxph(Surv(tenure, churn) ~ contract + gender, data = customer_churn_tbl)
summary(cox_model)

Call:
coxph(formula = Surv(tenure, churn) ~ contract + gender, data = customer_churn_tbl)

  n= 7043, number of events= 1869

                    coef exp(coef) se(coef)       z
contractOne year -2.19184   0.11171  0.08346 -26.261
contractTwo year -4.22809   0.01458  0.15637 -27.039
genderMale       -0.04787   0.95325  0.04632  -1.034
                 Pr(>|z|)
contractOne year   <2e-16 ***
contractTwo year   <2e-16 ***
genderMale          0.301
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                 exp(coef) exp(-coef) lower .95
contractOne year   0.11171      8.952   0.09485
contractTwo year   0.01458     68.586   0.01073
genderMale         0.95325      1.049   0.87052
                 upper .95
contractOne year   0.13156
contractTwo year   0.01981
```

```
genderMale         1.04385

Concordance= 0.775  (se = 0.004 )
Likelihood ratio test= 2620  on 3 df,   p=<2e-16
Wald test            = 1281  on 3 df,   p=<2e-16
Score (logrank) test = 2347  on 3 df,   p=<2e-16
```

coxph(Surv(tenure, churn) ~ contract + gender): Models the effect of contract and gender on the risk of churn.

Representation Graph:

```
g4 <- ggsurvplot(
    survfit(cox_model),
    conf.int = TRUE,
    data = customer_churn_tbl,
    ggtheme = theme_tq(), # Apply tidyquant theme
    palette = c("#E69F00", "#56B4E9"),
    title = "Survival Curves from Cox Model"
)
print(g4)
```

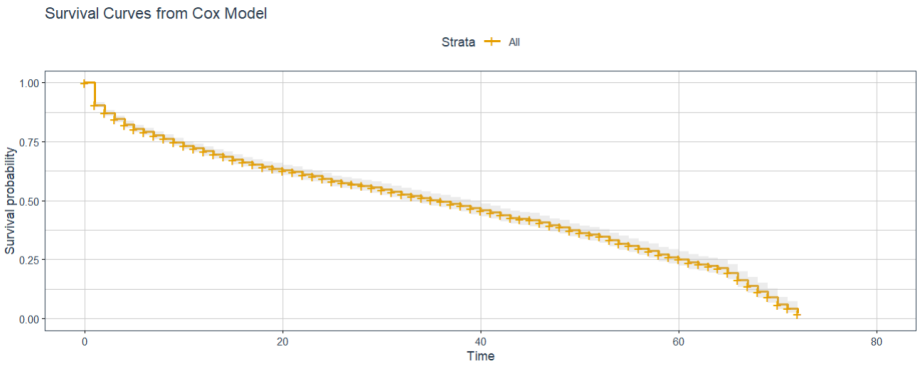

Figure 5: image

The survival curves derived from the Cox model provide insight into the adjusted survival probabilities considering the covariates. It quantifies the impact of contract type and gender on churn risk, offering a more detailed understanding of factors influencing customer retention.

Conclusion

This project utilizes survival analysis techniques to understand customer churn. Nonparametric methods, such as the Kaplan-Meier estimator, provide insights into survival functions and group comparisons. The Cox proportional hazards model offers a deeper understanding of how various factors affect survival. The

visualizations are customized for clarity and insight, using modern themes and plotting techniques to present the findings effectively.