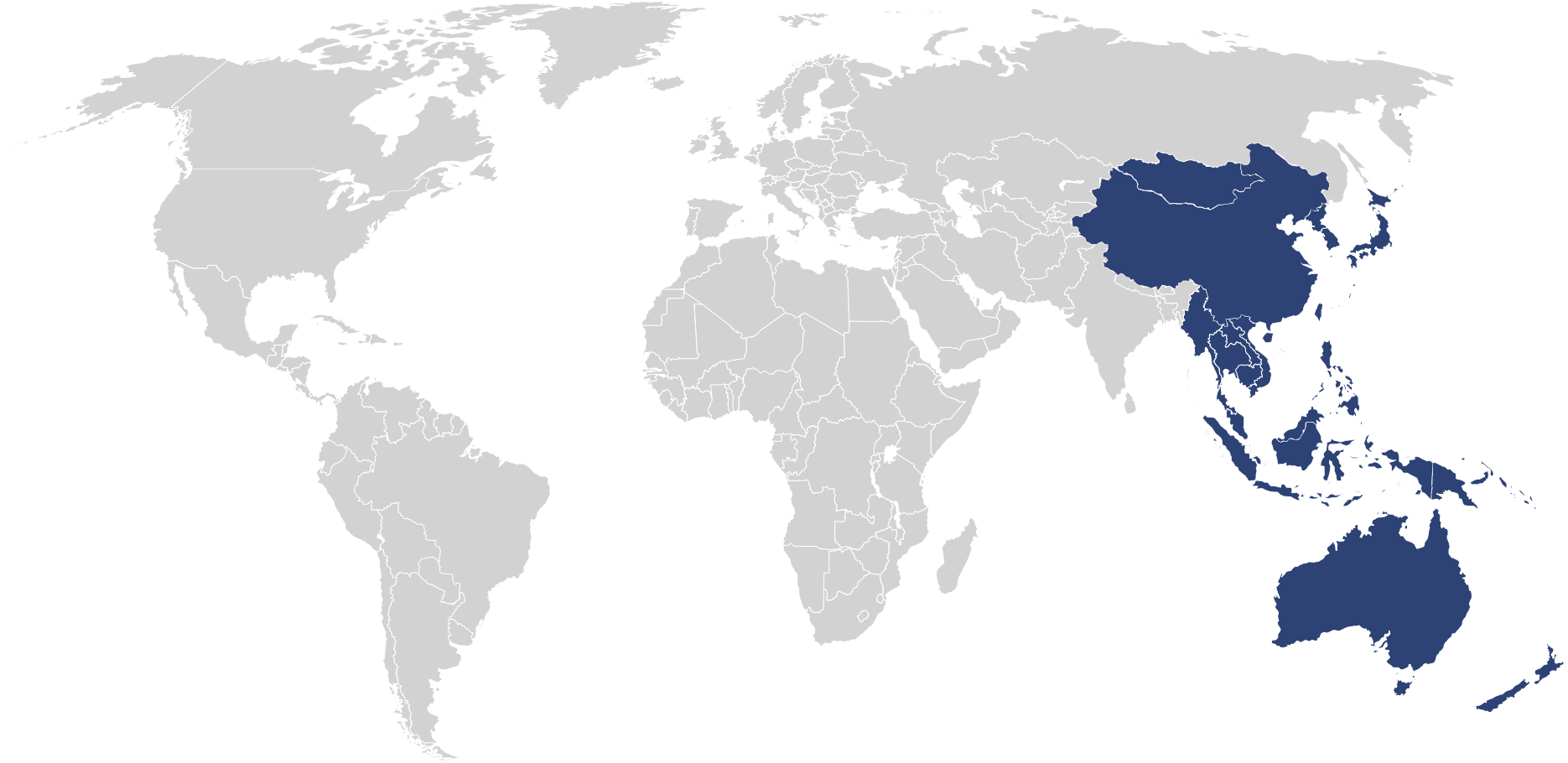
STA 592

Survival Analysis

Using Survival Analysis to Predict Customer Churn



What I will be taking about

1. Introduction
2. Methods and Data Diagnostics
3. Main Results
4. Conclusion

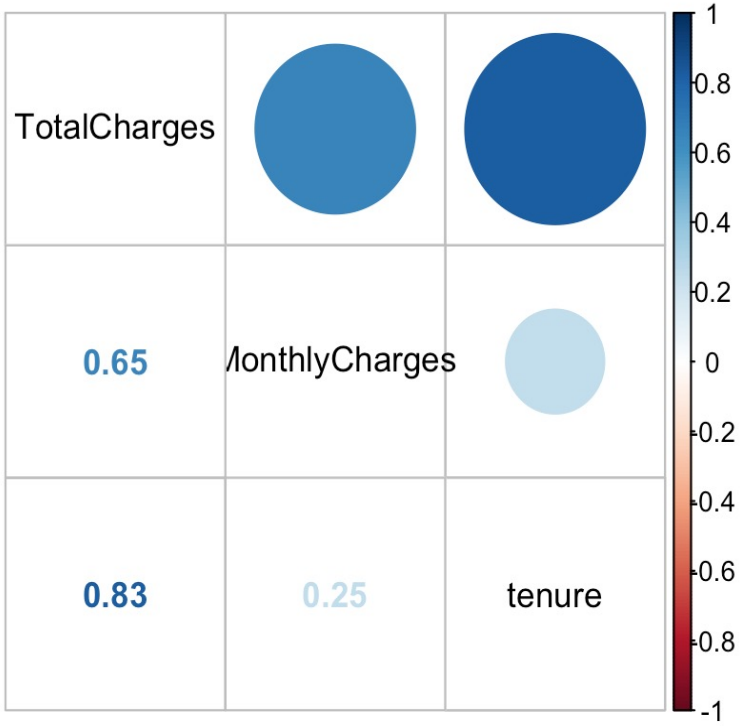
Introduction and Background

1. Motivation For The Analysis
2. Data Processing
3. Descriptive Statistics & Data Visuallization
4. Survival Model (Kaplan-Meier, Cox Proportional-Hazard, Log Rank Test)
5. Evaluation

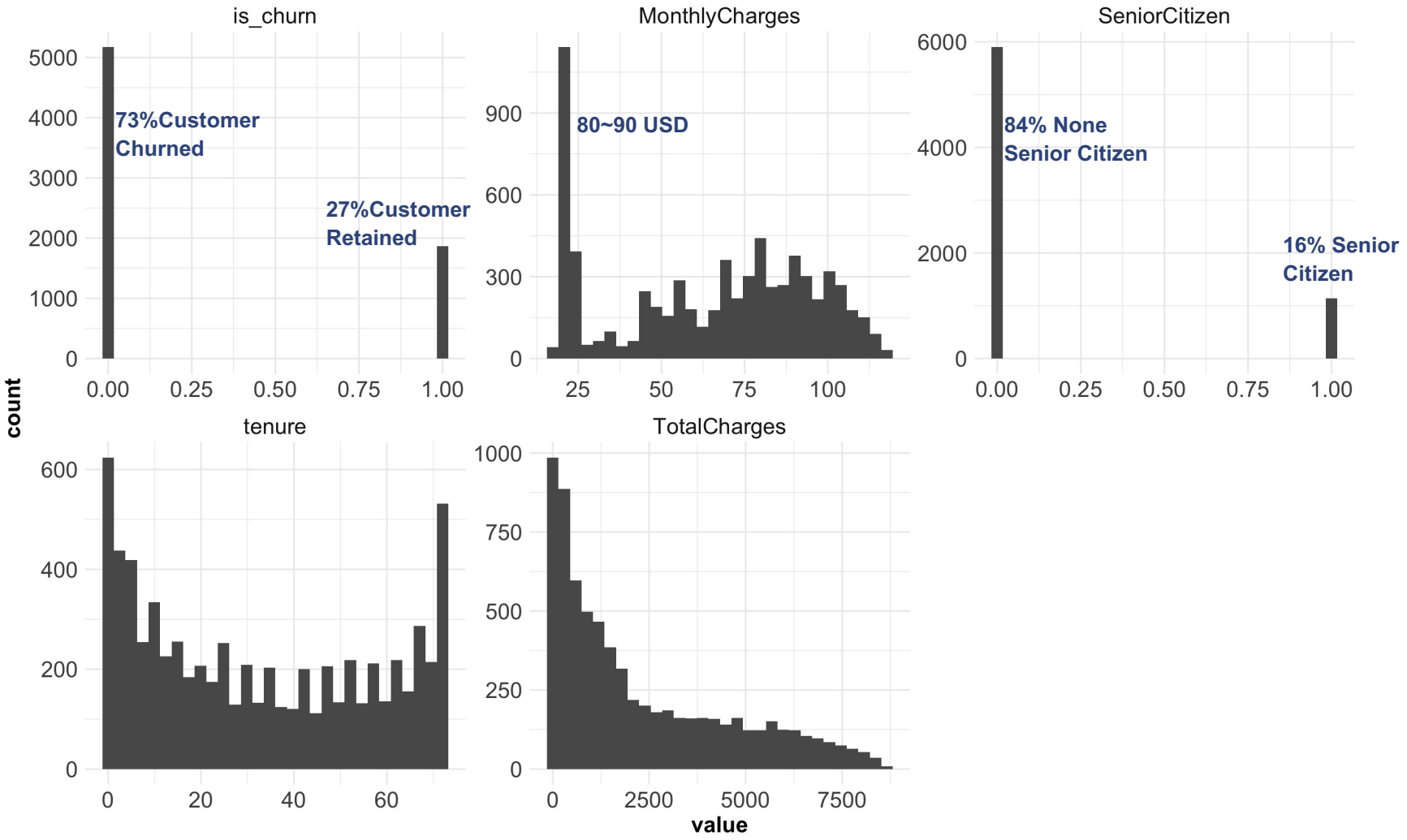
Data Overview

The raw data contains **7043** rows (customers) and **21** columns (features). The "Churn" column is the target.

"Total Charges" has 0.156% missing value in the dataset.

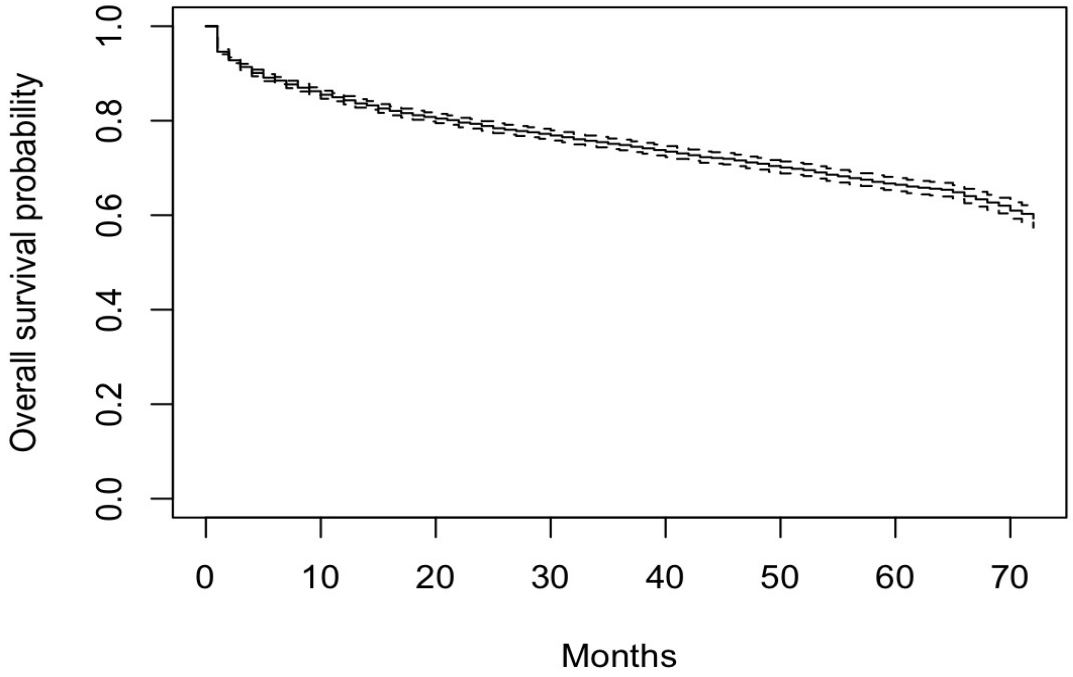


Distribution Of Numeric Feature(s)



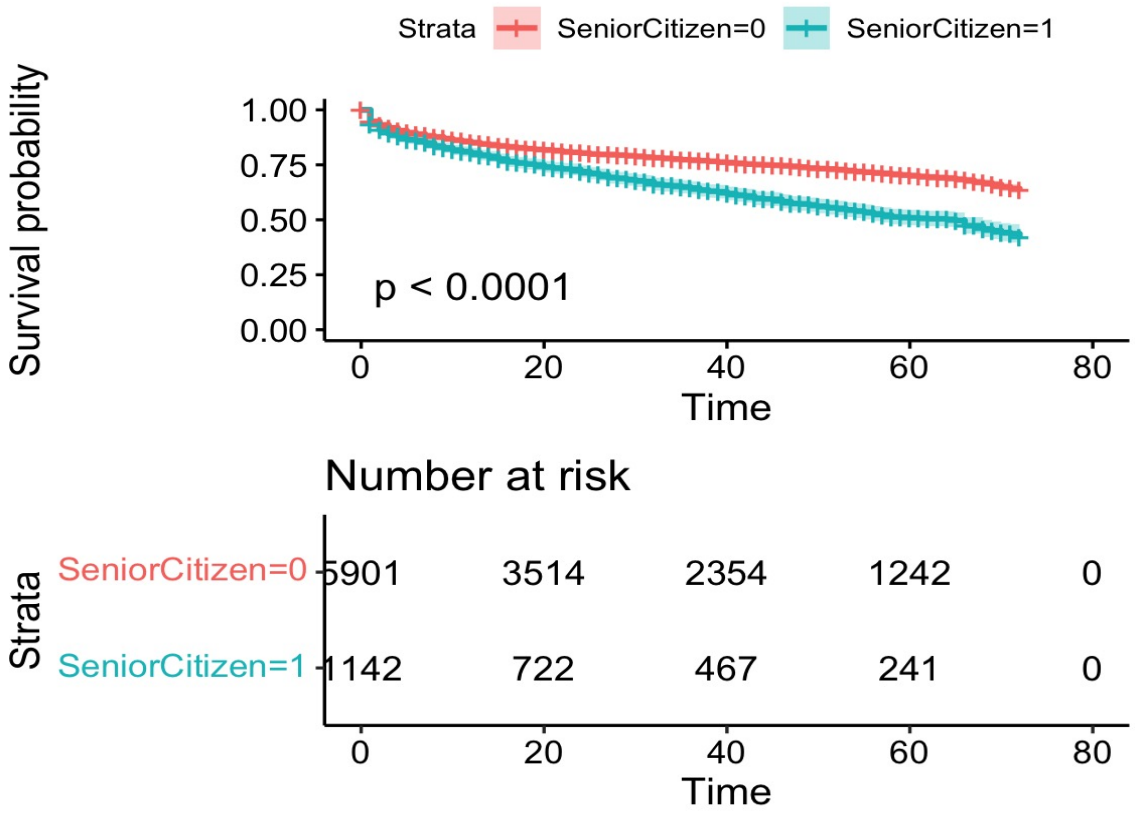
By Estimating Survival Curves With Kaplan-Meier Method

We Find 60 Months Probability Of Survival Is 66.4%



95%CI [65%-68%]

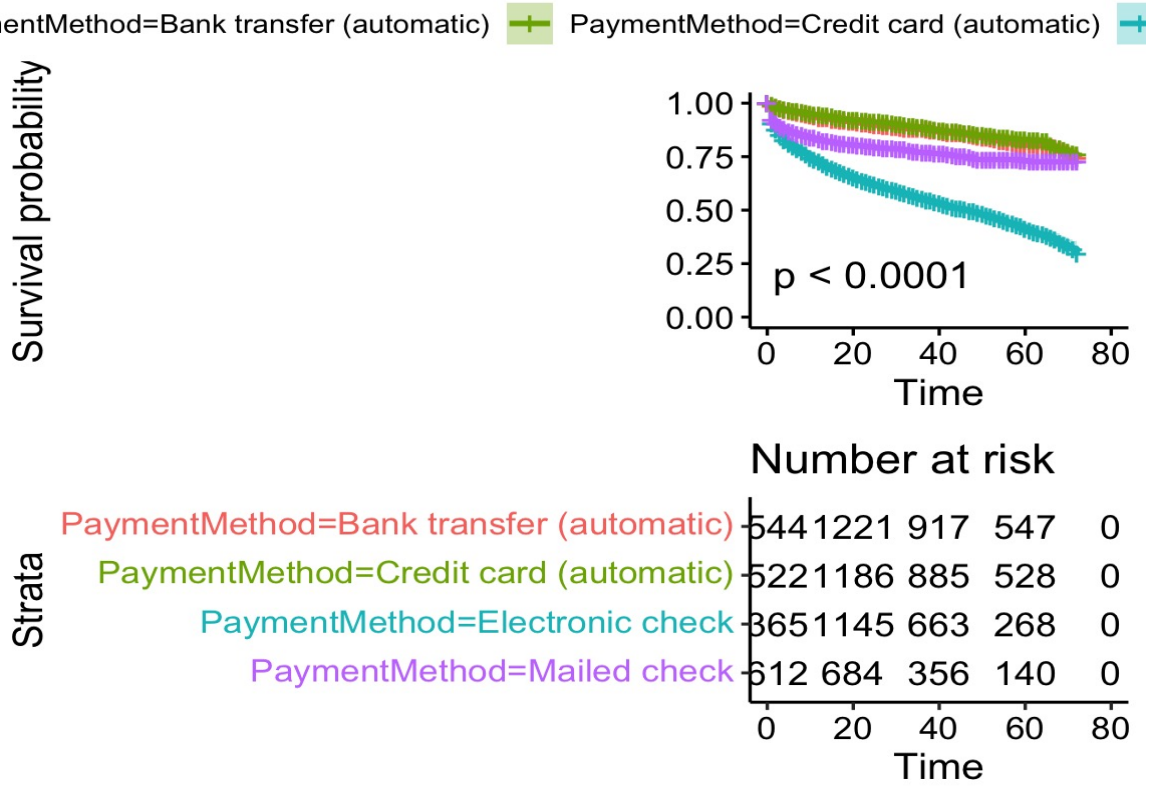
Kaplan-Meier Curve Indicate Senior Citizen Has Higher Churn Probability



Ho: Lifetime distributions under Senior Citizen and Not Senior Citizen are the same.

Ha: Lifetime distributions under Senior Citizen and Not Senior Citizen are not the same.

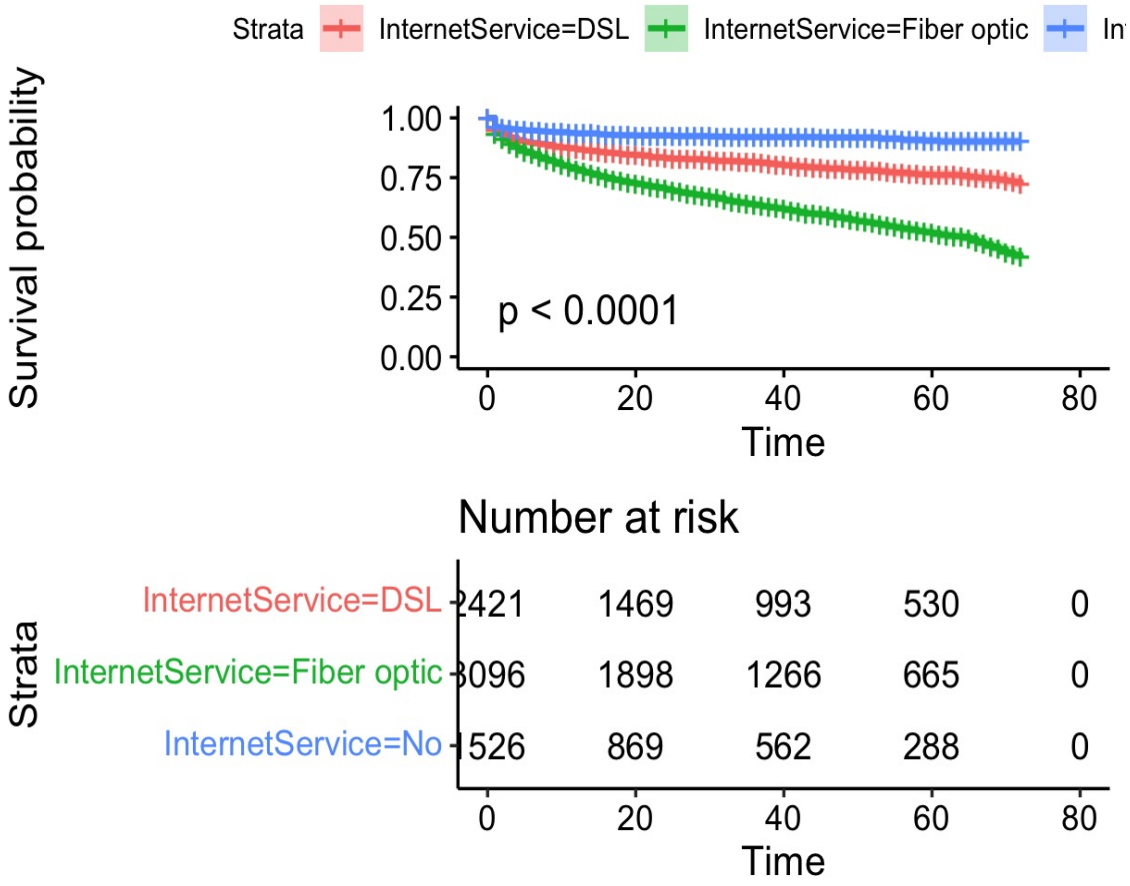
Kaplan-Meier Curve Indicate Payment Method With Electronic Check Has Higher Churn Probability



Ho: Lifetime distributions under Payment Method are the same.

Ha: Lifetime distributions Payment Method are not the same.

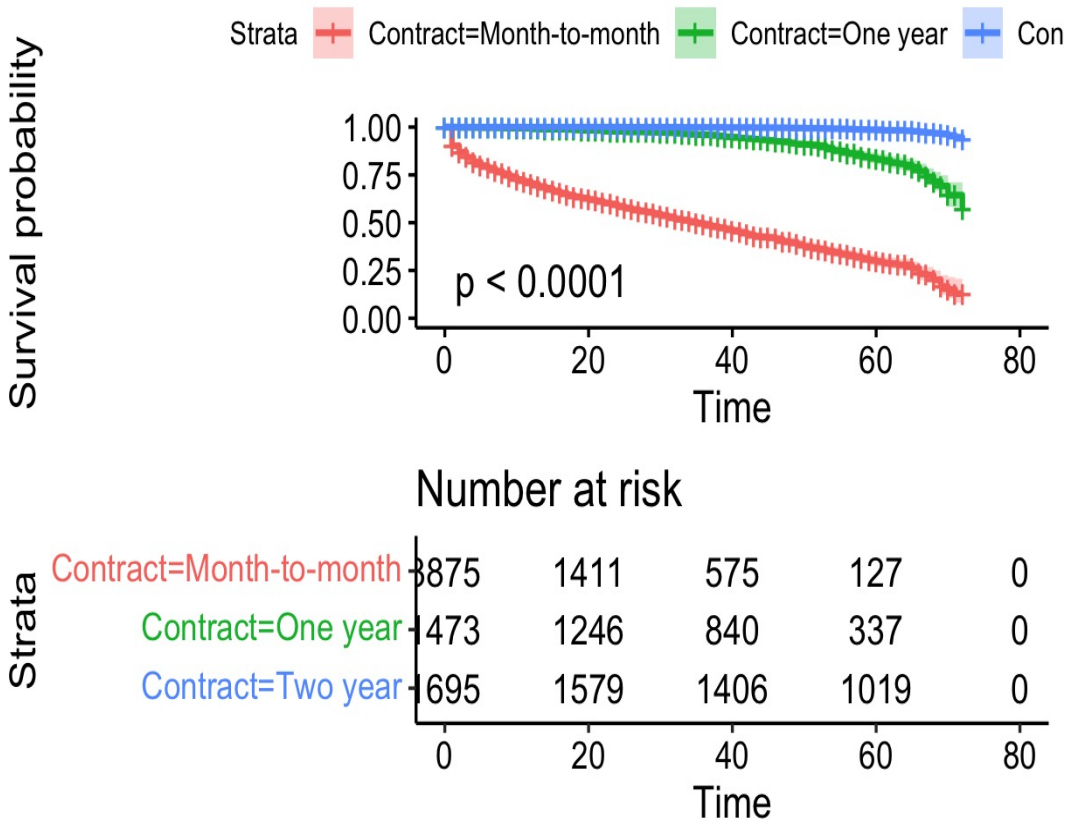
Kaplan-Meier Curve Indicate Internet Service For Fiber Has Higher Churn Probability



Ho: Lifetime distributions under **Internet Service** are the same.

Ha: Lifetime distributions **Internet Service** are not the same.

Kaplan-Meier Curve Indicate Month to Month Contract Has Higher Churn Probability

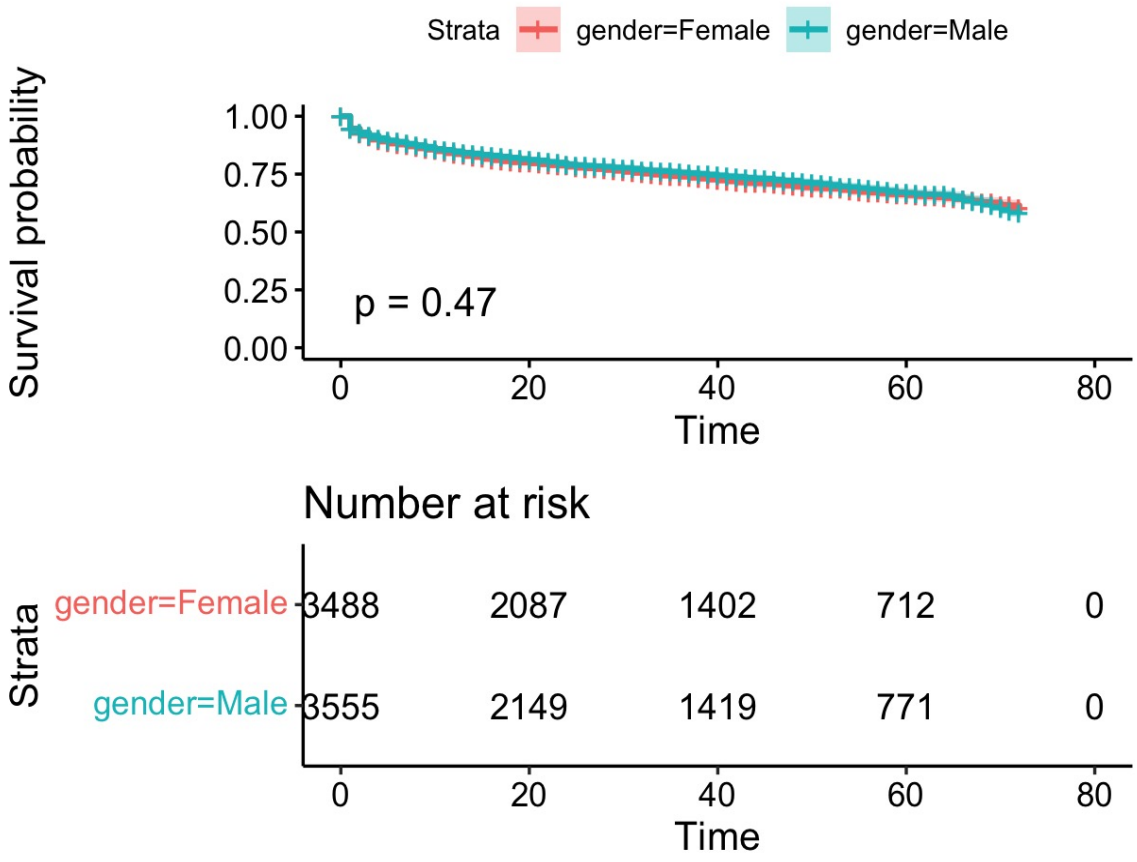


Ho: Lifetime distributions under **Contract Type** are the same.

Ha: Lifetime distributions **Contract Type** are not the same.

Kaplan-Meier Curve Indicate Gender Is Not Significant

There Is No Difference Between Gender



Ho: Lifetime distributions under **Gender** are the same.

Ha: Lifetime distributions **Gender** are not the same.

Log-Rank Test Comparing Survival Times Between Groups

```
survdiff(formula = surv_object ~ StreamingTV, data = telco_df)

                                  N Observed Expected (O-E)^2/E (O-E)^2/V
StreamingTV=No                 2810      942      624    162.47    252.07
StreamingTV=No internet service 1526      113      388    195.36    251.20
StreamingTV=Yes                2707      814      857      2.14      4.09

 Chisq= 368  on 2 degrees of freedom, p= <2e-16
Browse[1]> survdiff(surv_object ~ Contract, data=telco_df)
Call:
survdiff(formula = surv_object ~ Contract, data = telco_df)

                           N Observed Expected (O-E)^2/E (O-E)^2/V
Contract=Month-to-month 3875     1655      708      1265      2304
Contract=One year       1473      166      471       197       270
Contract=Two year       1695       48      690       597      1061

 Chisq= 2353  on 2 degrees of freedom, p= <2e-16
```

Cox Proportional-Hazard (PH): Likelihood ratio test indicate variables are significant

```
coxph(formula = Surv(tenure, is_churn) ~ Partner + PhoneService +
    InternetService * StreamingMovies + Contract + PaymentMethod,
    data = telco_df)

  n= 7043, number of events= 1869
```

	coef
PartnerYes	-0.62748
PhoneServiceYes	-0.16445
InternetServiceFiber optic	0.41625
InternetServiceNo	-0.35843
StreamingMoviesNo internet service	NA
StreamingMoviesYes	-0.25047
ContractOne year	-1.89399
ContractTwo year	-3.67663
PaymentMethodCredit card (automatic)	-0.06357
PaymentMethodElectronic check	0.66942
PaymentMethodMailed check	0.64931
InternetServiceFiber optic:StreamingMoviesNo internet service	NA
InternetServiceNo:StreamingMoviesNo internet service	NA
InternetServiceFiber optic:StreamingMoviesYes	0.01901
InternetServiceNo:StreamingMoviesYes	NA

Interaction Effect Appears For Steaming Movie, Steaming TV And Internet Service

```
Analysis of Deviance Table
 Cox model: response is  Surv(tenure, is_churn)
 Model 1: ~ Partner + PhoneService + InternetService + StreamingTV + StreamingMovies + TotalCharges
 Model 2: ~ Partner + PhoneService + InternetService + StreamingTV * StreamingMovies + TotalCharges
  loglik  Chisq Df P(>|Chi|)
1 -13002
2 -12994 15.407  1 8.666e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Analysis of Deviance Table
 Cox model: response is  Surv(tenure, is_churn)
 Model 1: ~ Partner + PhoneService + InternetService + StreamingTV + StreamingMovies + TotalCharges
 Model 2: ~ Partner + PhoneService + InternetService * StreamingTV + StreamingMovies + TotalCharges
  loglik  Chisq Df P(>|Chi|)
1 -13002
2 -12995 13.954  1 0.0001873 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

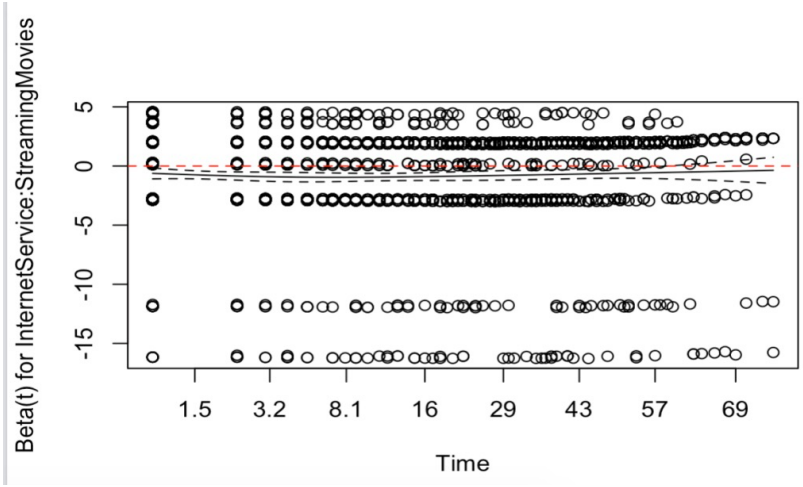
Proportional Hazards and Schoenfeld Residual Indicated Violation of Variables

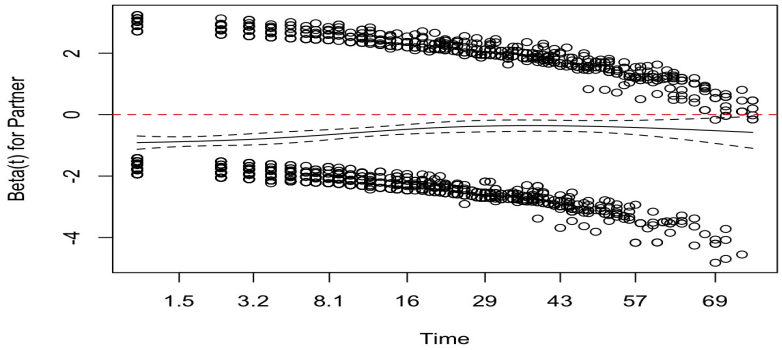
	chisq	df	p
Partner	22.18	1	2.5e-06
PhoneService	2.79	1	0.095
InternetService	27.62	2	1.0e-06
StreamingMovies	72.89	1	< 2e-16
Contract	95.57	2	< 2e-16
PaymentMethod	47.46	3	2.8e-10
GLOBAL	225.48	10	< 2e-16

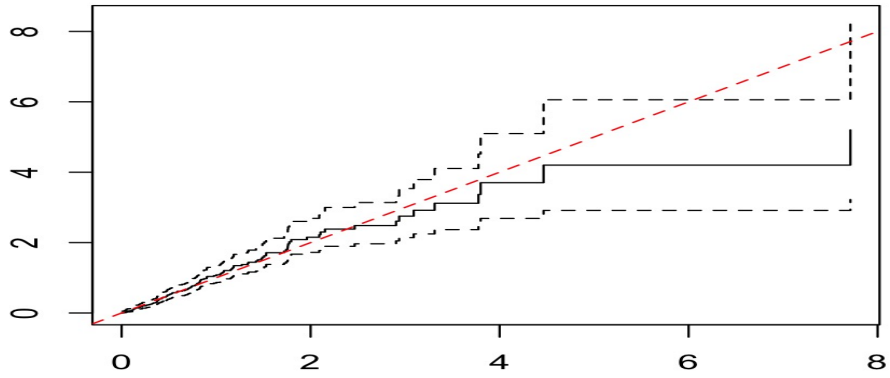
H_0: the effect of the j-th explanatory variable is constant over time (i.e. proportional hazards)

H_a: the effect is not constant over time (i.e. non-proportional hazard)

Overall Fit, Proportional Hazards and Schoenfeld Residuals







Executive Summary Model Solution

Churn is indeed high

- Average user churn risk of Internet service fiber users is 1.52 times that of non-fiber users
- The average churn risk of telephone service users is 1.95 times that of telephone service users
- Staying with the firm for 60 weeks is ~75% for non-senior citizens vs. ~50% for senior citizens.

Main driver

- Partner
- Internet Service
- Payment Method
- Senior Citizen

Suggestion

- Increase retention by tying long-term contracts with discount rates for customers who choose to use Internet fiber services and those who choose to Electronic Check .

THANK YOU