Customer Churn Prediction:

I. Describe the process by which you cleaned, processed, and partitioned data as necessary.

Churn: It is a measure of how many customers stop using the product. In our case we are trying to find out how many subscribers of telephone and/or internet of a large telco will stop using the service. This is a classification problem (Logistic Regression) as we are not trying to find a continuous quantity but are trying to find a discrete label which when an input is given, the model would help us determine if the value is part of a pre-identified group (Is there a Churn? "Yes" or "No"). In simple words, when DV is categorical we use Logistic Regression.

Steps for data cleaning, pre-processing and partitioning are as follows:

1. First, we need to find if there are NA values in the data and remove those rows.
2. Then, we remove all the columns that would not make much sense as predictors.
3. Then, we need to factorize our DV and subset the data for the analysis into three groups as phone_service, internet_service and both_services.

II. What predictors do you think contributes to the churn of (i) only telephone customers, (ii) only Internet service customers, and (iii) customers who subscribe to both phone and Internet services? Explain the rationale for your answer.

Relevant Factors for [y <- Churn]				
Predictor	**Phone**	**Internet**	**Both**	**Rationale**
Gender	+	+	+	Usage of these services vary greatly between Males and Females. So, gender certainly contributes to the churn in all the services.
SeniorCitizen	+	+	+	Most of the older population has gotten online and uses one or both the services. Based on the recent studies the adoption and the churn rates are quite high.
Partner	+	+	+	There is a duality in most cases where one partner uses the phone a lot and the other uses the internet. There are also couples that use both regularly in the process of finding the best service and that could lead to churn.
Dependents	+	+	+	Based on the wording, this could correlate with partner, but the churn will be there as a lot more people in the family could be using these services based on personal choices and preferences.
Tenure	+	+	+	Tenure plays a vital role for all services as this metric could help us understand how long a customer has used the service.
PhoneService	+		+	Key predictor in our analysis for finding the churn.
MultipleLines	+		+	Subset of the phone service as having multiple lines shows that the customer trusts the service and relies on it on a frequent basis.
InternetService		+	+	Key predictor in our analysis for finding the churn.
TechSupport		+	+	For a service like internet, tech support is an important factor, where it shows the service providers ability to sort out operational issues, which is a major pain point for most of the customers.
StreamingMovies		+	+	Most of the service providers offer movie plans as part of their broadband bundles and this could help with gaining new customers.
MonthlyCharges	+	+	+	Charges play an important role in predicting customer churn and in most cases are dependent on the individual service itself based on the usage.
TotalCharges	+	+	+	Same as above but billed yearly and is very dependent on the service itself.

Irrelevant Factors		
Predictor	**Effect**	**Rationale**
CustomerID	No Effect	Used for indexing so has no effect on our DV.
OnlineSecurity/ OnlineBackup	No Effect	Most internet subscribers do not need an online backup and most operators offer basic security features on their broadband plans.
DeviceProtection	No Effect	Internet is a service and does not pose any physical threat to the user devices.
StreamingTV	No Effect	Most customers have a dedicated TV at home.
Contract	No Effect	Contractual agreements do not have any effect on the churn.
PaperlessBilling	No Effect	Paperless billing will not have any effect on the churn.
PaymentMethod	No Effect	Payment method will not have any effect on the churn.

III. Create training and test data sets with a 75:25 split using a random seed of 1024. Use the training data to train three logit models with the variables you identified in Question II. Combine the outputs of the three modes using stargazer.

```
#Stargazer
stargazer(phone_logit, internet_logit, both_logit, type='text', single.row = TRUE)
```

```
##
## ===============================================================================
##                                        Dependent variable:
##                    ---------------------------------------------------------
##                                             churn
##                             (1)                 (2)                (3)
## -------------------------------------------------------------------------------
## genderMale                -0.004 (0.077)     -0.040 (0.076)     -0.056 (0.081)
## seniorcitizen           0.405*** (0.099)   0.446*** (0.095)   0.339*** (0.100)
## partnerYes                0.038 (0.092)      0.027 (0.090)      -0.032 (0.096)
## dependentsYes           -0.372*** (0.105)  -0.329*** (0.105)  -0.162 (0.111)
## tenure                  -0.084*** (0.008)  -0.063*** (0.008) -0.074*** (0.010)
## multiplelinesYes         0.373*** (0.092)                     0.388*** (0.096)
## techsupportYes                             -0.834*** (0.092) -0.708*** (0.099)
## streamingmoviesYes                           0.062 (0.093)      -0.102 (0.105)
## monthlycharges           0.029*** (0.002)  0.020*** (0.003)  0.027*** (0.004)
## totalcharges            0.0003*** (0.0001) 0.0002** (0.0001) 0.0003** (0.0001)
## Constant                -1.655*** (0.163)  -0.758*** (0.209) -1.384*** (0.292)
## -------------------------------------------------------------------------------
## Observations                  4,764              4,134              3,624
## Log Likelihood             -2,078.735         -2,060.649         -1,835.049
## Akaike Inf. Crit.          4,175.470          4,141.299          3,692.099
## ===============================================================================
## Note:                                           *p<0.1; **p<0.05; ***p<0.01
```

IV. What are the top three predictors of churn of (i) only telephone customers, (ii) only Internet service customers, and (iii) customers who subscribe to both phone and Internet services. Explain using marginal effects how much each predictor contributes to churn occurrence.

Odds and Probability Percentages

Phone Service

	odd_phone	prob_phone
(Intercept)	0.1911033	0.1604423
genderMale	0.9962687	0.4990654
seniorcitizen	1.4995681	0.5999309
partnerYes	1.0392097	0.5096140
dependentsYes	0.6891147	0.4079739
tenure	0.9191004	0.4789225
multiplelinesYes	1.4520795	0.5921829
monthlycharges	1.0298312	0.5073482
totalcharges	1.0002755	0.5000689

Internet Service

	odd_internet	prob_internet
(Intercept)	0.4685039	0.3190349
genderMale	0.9611909	0.4901057
seniorcitizen	1.5615696	0.6096144
partnerYes	1.0272919	0.5067311
dependentsYes	0.7197955	0.4185355
tenure	0.9386142	0.4841676
techsupportYes	0.4342869	0.3027894
streamingmoviesYes	1.0639819	0.5154996
monthlycharges	1.0204057	0.5050499
totalcharges	1.0001879	0.5000470

Both Services

	odd_both	prob_both
(Intercept)	0.2506800	0.2004349
genderMale	0.9453833	0.4859625
seniorcitizen	1.4038160	0.5839948
partnerYes	0.9688884	0.4920992
dependentsYes	0.8504340	0.4595862
tenure	0.9287954	0.4815417
multiplelinesYes	1.4735288	0.5957193
techsupportYes	0.4923974	0.3299372
streamingmoviesYes	0.9030975	0.4745409
monthlycharges	1.0271267	0.5066909
totalcharges	1.0002556	0.5000639

Churn: Measure of customers who stop using the service **Interpretation in terms of [Odds -> Probability/(1 - Probability)]**	
Top 3 Predictors – Telephone Service	
1. Senior citizens have	**50% higher odds** of churn
2. Customers with multiple lines have	**45% higher odds** of churn compared to those that do not have multiple lines
3. Customers with partners have	**4% higher odds** of churn compared to those that do not have a partner
Top 3 Predictors – Internet Service	
1. Senior citizens have	**56% higher odds** of churn
2. Customers that stream movies have	**6.4% higher odds** of churn
3. Customers with partners have	**2.7% higher odds** of churn compared to those that do not have a partner
Top 3 Predictors - Both Services	
1. Customers with multiple lines have	**47% higher odds** of churn compared to those that do not have multiple lines
2. Senior citizens have	**40% higher odds** of churn
3. Customers that pay monthly have	**2.7% higher odds** of churn

Churn: Measure of customers who stop using the service **Interpretation in terms of [Probability -> Event/Total]**	
Top 3 Predictors – Telephone Service	
4. Senior citizens have	**60% likelihood** of churn
5. Customers with multiple lines have	**59% likelihood** of churn compared to those that do not have multiple lines
6. Customers with partners have	**51% likelihood** of churn compared to those that do not have a partner
Top 3 Predictors – Internet Service	
4. Senior citizens have	**61% likelihood** of churn
5. Customers that stream movies have	**51.5% likelihood** of churn
6. Customers with partners have	**50.6% likelihood** of churn compared to those that do not have a partner
Top 3 Predictors - Both Services	
4. Customers with multiple lines have	**59.5% likelihood** of churn compared to those that do not have multiple lines
5. Senior citizens have	**58% likelihood** of churn
6. Customers that pay monthly have	**50.6% likelihood** of churn

V. Fit your models using test data, and compute recall, precision, F1-score, and AUC values for each of your three models. Create a table with these values.

Confusion Matrix Statistics on Test Data

Phone Service

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1047  212
##          1  124  205
##
##                Accuracy : 0.7884
##                  95% CI : (0.7675, 0.8083)
##     No Information Rate : 0.7374
##     P-Value [Acc > NIR] : 1.353e-06
##
##                   Kappa : 0.4138
##
##  Mcnemar's Test P-Value : 2.072e-06
##
##             Sensitivity : 0.4916
##             Specificity : 0.8941
##          Pos Pred Value : 0.6231
##          Neg Pred Value : 0.8316
##               Precision : 0.6231
##                  Recall : 0.4916
##                      F1 : 0.5496
##              Prevalence : 0.2626
##          Detection Rate : 0.1291
##    Detection Prevalence : 0.2072
##       Balanced Accuracy : 0.6929
##
##        'Positive' Class : 1
```

Internet Service

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 816 198
##          1 141 223
##
##                Accuracy : 0.754
##                  95% CI : (0.7304, 0.7765)
##     No Information Rate : 0.6945
##     P-Value [Acc > NIR] : 5.874e-07
##
##                   Kappa : 0.3974
##
##  Mcnemar's Test P-Value : 0.002354
##
##             Sensitivity : 0.5297
##             Specificity : 0.8527
##          Pos Pred Value : 0.6126
##          Neg Pred Value : 0.8047
##               Precision : 0.6126
##                  Recall : 0.5297
##                      F1 : 0.5682
##              Prevalence : 0.3055
##          Detection Rate : 0.1618
##    Detection Prevalence : 0.2642
##       Balanced Accuracy : 0.6912
##
##        'Positive' Class : 1
```

Both Services

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 721 186
##          1 104 197
##
##                Accuracy : 0.7599
##                  95% CI : (0.7348, 0.7838)
##     No Information Rate : 0.6829
##     P-Value [Acc > NIR] : 2.305e-09
##
##                   Kappa : 0.4119
##
##  Mcnemar's Test P-Value : 1.970e-06
##
##             Sensitivity : 0.5144
##             Specificity : 0.8739
##          Pos Pred Value : 0.6545
##          Neg Pred Value : 0.7949
##               Precision : 0.6545
##                  Recall : 0.5144
##                      F1 : 0.5760
##              Prevalence : 0.3171
##          Detection Rate : 0.1631
##    Detection Prevalence : 0.2492
##       Balanced Accuracy : 0.6941
##
##        'Positive' Class : 1
```

Relevant Confusion Matrix Statistics

Metric	Phone Service	Internet Service	Both Services
1. Recall	**49%**	**53%**	**51.4%**
2. Precision	**62.3%**	**61%**	**65.4%**
3. F1-Score	**55%**	**57%**	**57.6%**
4. AUC	**69%**	**69%**	**69.4%**

Based on our analysis, Internet model fits better for the prediction. That's because we have an unbalanced sample, where we are trying to avoid the worst-case scenario. Consider the cases below:

Precision (Factor: False Positive): Customer was predicted to churn but did not (Effect: Customer is using the service - Good).
Recall (Factor: False Negative): Customer was not predicted to churn but did (Effect: Customer is not using the service - Bad).

Now, in any business case, Recall is costlier than Precision. So, we must select a model that has the highest Recall, which in this case is the Internet model with 53% .