Data Preprocessing:

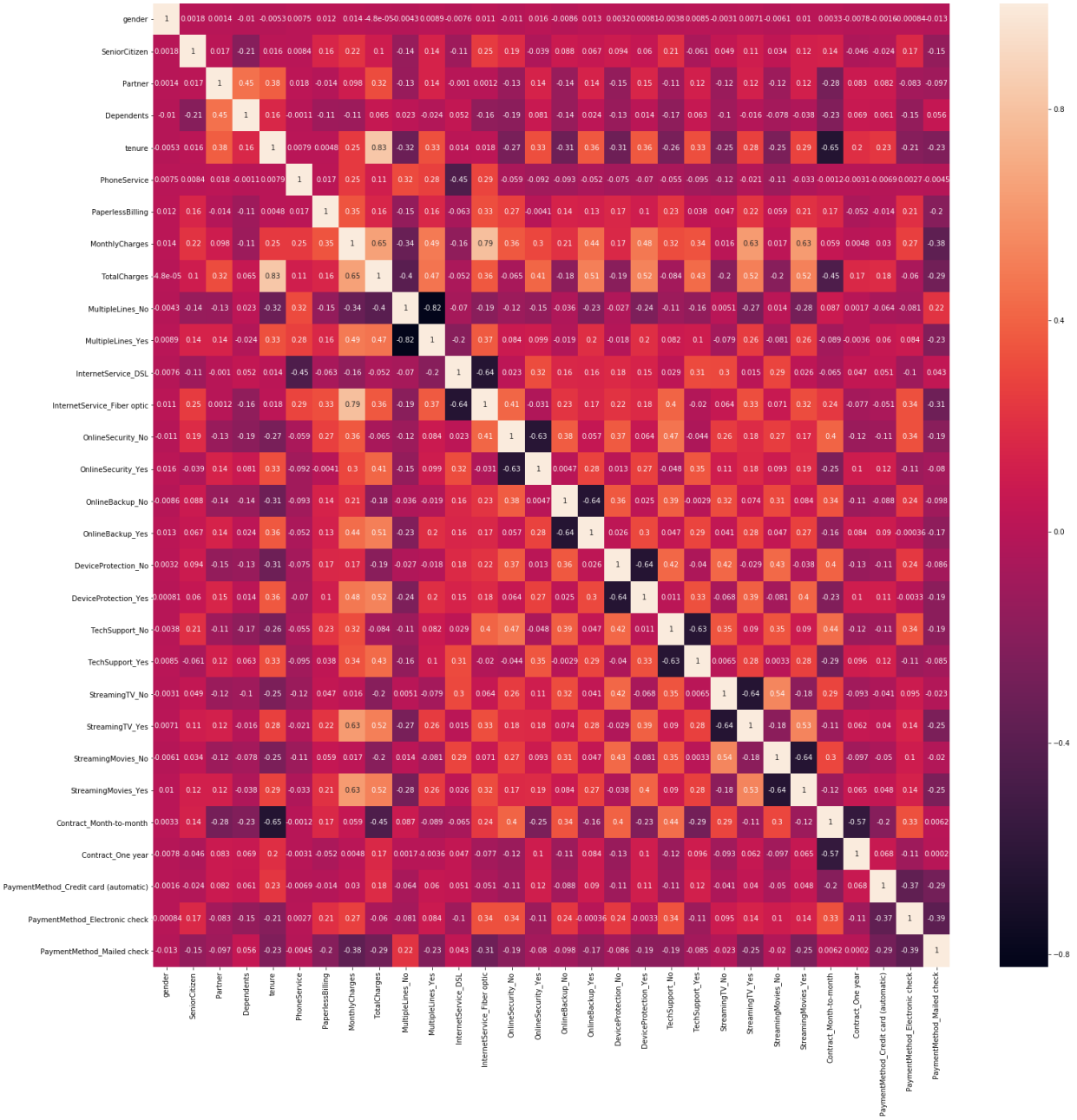
- Created Dummy variables (n-1) for all the categorical variables, standardized all the variables and computed percentage differences in mean for all variables across churn=Yes and churn=No
- Sorted those means in descending order and checked if there are any strong correlations
- There were 11 null values in the data, we have dropped those records

Percentage difference in means of features across churners and non-churners	
Feature	% Difference
Contract_One year	184.801495
OnlineSecurity_Yes	111.063729
tenure	109.409482
TechSupport_Yes	102.018706
PaymentMethod_Credit card (automatic)	101.127803
Dependents	96.878509
TotalCharges	66.820124
PaymentMethod_Electronic check	56.262699
InternetService_DSL	54.342424
PaymentMethod_Mailed check	52.321589
Contract_Month-to-month	51.441848
InternetService_Fiber optic	49.789059
OnlineSecurity_No	49.553071
SeniorCitizen	49.350583
TechSupport_No	49.280107
Partner	47.396836
OnlineBackup_No	45.568058
DeviceProtection_No	43.712319
OnlineBackup_Yes	31.648526
StreamingMovies_No	28.873789
PaperlessBilling	28.427658
StreamingTV_No	28.253521
DeviceProtection_Yes	24.408041
MonthlyCharges	17.643322
StreamingTV_Yes	15.993144
StreamingMovies_Yes	15.341837
MultipleLines_Yes	9.840995
MultipleLines_No	8.13063
gender	1.924915

PhoneService	0.860471

- On comparing the percentage difference between the means of all the features for churners and non-churners, finalized below 10 columns for our analysis.
    1. Contract
    2. Online Security
    3. Tenure
    4. Tech Support
    5. Payment
    6. Total Charges
    7. Internet service
    8. Dependents
    9. Senior Citizen
    10. Monthly Charges
- We have picked all the numeric columns which are Tenure, Monthly Charges and Total Charges and we have picked 7 categorical features that had the highest percentage difference in mean across churners and non-churners.
- Priority is given to numeric columns beacuse of the simple reason that dummy variables have less predictive power

Checking for correlation among selected features:



- We have using “No internet service” as equivalent “No” for columns selected after feature selection “Online Security” and “Tech Support”. We are using No Internet Service and No as the base for dummy variables as both mean almost one and the same thing, and having two instead of 1 dummy variables for each of these categorical features would introduce linear dependency among dummy variables which will in turn lead to multicollinearity

Answer-1

#Model-1

The SAS System

The LOGISTIC Procedure

Model Information	
Data Set	WORK.A4
Response Variable	dchurn
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	7032
Number of Observations Used	7032

Response Profile		
Ordered Value	dchurn	Total Frequency
1	0	5163
2	1	1869

Probability modeled is dchurn=1.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	8145.355	5937.260
SC	8152.213	6040.134
-2 Log L	8143.355	5907.260

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2236.0949	14	<.0001
Score	1940.6775	14	<.0001
Wald	1234.4663	14	<.0001

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept	1	-2.5779	0.2385	116.8462	<.0001		0.076
ddependents	1	-0.1699	0.0808	4.4197	0.0355	-0.0429	0.844
dinternet_dsl	1	0.9890	0.1582	39.0690	<.0001	0.2590	2.689
dinternet_fo	1	1.5634	0.2454	40.5819	<.0001	0.4279	4.775
donlinesecurity_Y	1	-0.4818	0.0849	32.1840	<.0001	-0.1201	0.618
dtech_Y	1	-0.4003	0.0877	20.8511	<.0001	-0.1002	0.670
dcontract_M	1	1.3846	0.1756	62.1909	<.0001	0.3797	3.993
dcontract_O	1	0.7005	0.1761	15.8315	<.0001	0.1571	2.015
dpayment_e	1	0.3649	0.0933	15.2925	<.0001	0.0951	1.440
dpayment_m	1	-0.0955	0.1136	0.7067	0.4005	-0.0221	0.909
dpayment_cc	1	-0.0701	0.1130	0.3844	0.5353	-0.0159	0.932
SeniorCitizen	1	0.2648	0.0830	10.1847	0.0014	0.0538	1.303
tenure	1	-0.0588	0.00625	88.5576	<.0001	-0.7954	0.943
MonthlyCharges	1	0.00346	0.00371	0.8703	0.3509	0.0574	1.003
TotalCharges	1	0.000342	0.000071	23.4615	<.0001	0.4278	1.000

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	84.3	Somers' D	0.686
Percent Discordant	15.7	Gamma	0.686
Percent Tied	0.0	Tau-a	0.268
Pairs	9649647	c	0.843

#Model-2

- Removing insignificant variables from above model, we are dropping Monthly Charges from the model.
- Also removing payment_method_creditcard and payment_method_mailedCheck, which means now the base for payment method dummy variable becomes credit card, mailed check and Automatic bank transfer.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	8145.355	5933.009
SC	8152.213	6015.308
-2 Log L	8143.355	5909.009

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2234.3461	11	<.0001
Score	1919.2987	11	<.0001
Wald	1227.7216	11	<.0001

Number of Observations Read	7032
Number of Observations Used	7032

Response Profile		
Ordered Value	dchurn	Total Frequency
1	0	5163
2	1	1869

Probability modeled is dchurn=1.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept	1	-2.5726	0.1995	166.3220	<.0001		0.076
ddependents	1	-0.1678	0.0807	4.3198	0.0377	-0.0423	0.846
dinternet_dsl	1	1.0902	0.1259	74.9826	<.0001	0.2855	2.975
dinternet_fo	1	1.7769	0.1330	178.5425	<.0001	0.4864	5.911
donlinesecurity_Y	1	-0.4685	0.0837	31.3096	<.0001	-0.1168	0.626
dtech_Y	1	-0.3795	0.0847	20.0967	<.0001	-0.0950	0.684
dcontract_M	1	1.3798	0.1756	61.7322	<.0001	0.3784	3.974
dcontract_O	1	0.7040	0.1761	15.9885	<.0001	0.1579	2.022
dpayment_e	1	0.4237	0.0680	38.7730	<.0001	0.1104	1.528
SeniorCitizen	1	0.2663	0.0829	10.3229	0.0013	0.0542	1.305
tenure	1	-0.0601	0.00586	105.1705	<.0001	-0.8136	0.942
TotalCharges	1	0.000369	0.000063	34.5315	<.0001	0.4615	1.000

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	84.3	Somers' D	0.686
Percent Discordant	15.7	Gamma	0.686
Percent Tied	0.0	Tau-a	0.268
Pairs	9649647	c	0.843

AIC and BIC have reduced on dropping insignificant variables, which means our model has improved on dropping few variables. So, Model# 2 would be our logistic regression model to predict which customers are likely to churn

Answer-2

Variable	Parameter	Estimate	Standardized Estimate	Pr > ChiSq	Odds Ratio
	Intercept	-2.5726		<.0001	0.076
Dependents	ddependents	-0.1678	-0.0423	0.0377	0.846
InternetService	dinternet_dsl	1.0902	0.2855	<.0001	2.975
	dinternet_fo	1.7769	0.4864	<.0001	5.911
Online Security	donlinesecurity_Y	-0.4685	-0.1168	<.0001	0.626
Tech Support	dtech_Y	-0.3795	-0.095	<.0001	0.684
Contract	dcontract_M	1.3798	0.3784	<.0001	3.974
	dcontract_O	0.704	0.1579	<.0001	2.022
Payment Method	dpayment_e	0.4237	0.1104	<.0001	1.528
Senior Citizen	SeniorCitizen	0.2663	0.0542	0.0013	1.305
Tenure	tenure	-0.0601	-0.8136	<.0001	0.942
Total Charges	TotalCharges	0.00037	0.4615	<.0001	1

McFadden R-sq:

27.43%

AIC Interpretation:

AIC (Null Model) : 8145.355

AIC (With covariates): 5933.009

AIC for the null model that is when there are no explanatory variables in the model is 8145 and AIC with explanatory variables added, comes down to 5933. AIC has lowered which means the model has improved. The lower the better.

BIC Interpretation:

BIC (Null Model) : 8152.213

BIC (With covariates): 6015.308

Likewise, BIC has also come down from 8152 to 6015 after the addition of covariates into the model, again a significant improvement in the model.

Significance of Variables:

All the variables in the model are significant, Monthly Charges was insignificant in our previous choice of model, but we have dropped it.

Interpreting the Coefficients:

By looking at the coefficients of the logistic regression, we can only tell the direction, whether the odds or the probability of churn will increase or decrease, we cannot comment by how much.

1. Dependents: Having dependent would decrease the odds of churn when compared to not having any dependents, all else remaining same. Or we can say customers having dependents are less likely to churn than customers not having any dependents

2. Internet_dsl: Having DSL Internet service would increase the odds of churn when compared to no internet service, all else remaining same. Or we can say customers with internet DSL are more likely to churn than customers not having any internet connection.

3. Internet_fo: Having Fiber Optic Internet service would increase the odds of churn when compared to no internet service, all else remaining same.

4. Online_security_Yes: Having opted for Online Security would decrease the odds of churn when compared to not opting for online security or when a customer doesn't have internet service, all else remaining same.

5. Tech_Support : Customers having opted for Tech Support are less likely to churn than customers who have not opted for Tech Support service and the customers who do not have internet service, all else remaining same.

6. Contract_Month to Month : Customers with month to month contract are more likely to churn as compared to customers with Two-year contracts, all else remaining same, which means customers with long term contracts are more likely to stay with the provider.

7. Contract_One Year : Customers with one-year contract are more likely to churn as compared to customers with Two-year contracts, which means customers with long term contracts are more likely to stay with the provider.

8. Payment method: Customers with payment method as electronic check are more likely to churn as compared to customers with any other payment method, all else remaining same.

9. Senior Citizen: Senior citizen customers are more likely to churn than customers who are not senior citizen, all else remaining same.

10. Tenure: With 1 unit increase in tenure, odds of churn decrease, all else remaining same

11. Total Charges: With 1 unit increase in Total Charges, odds of churn increase, all else remaining same

Percentage Concordant

Percentage concordant for our model is 84.3% which means out of 9649647 pairs, for 84.3% of the pairs, predicted probability(churn=Yes) is greater than predicted probability (churn=No), i.e. as that pair formed by an *event* (churn=1 or Yes) with a PHAT (predicted probability) higher than that of the *no-event (Churn=0 or No). Higher the better.*

Interpreting the Odds Ratio:

- By looking at the coefficients we could only tell if the odds of a customer to churn would increase or decrease, but with odds ratio we can compute by how much those odds would increase or decrease with a simple formula (odds ratio-1)*100

1. Dependents: Odds of customers with dependents to churn are less than Odds of customers without dependents to churn by 15.4%.

2. Internet_dsl: Customers with DSL Internet service are more likely to churn than customers not having any internet service by 197.5%.

3. Internet_fo: Customers with Fiber Optic Internet service are more likely to churn than customers not having any internet service by 491%.

4. Online_security_Yes: Customers with Online Security service are less likely to churn than customers who don't have Online security service or an internet service by 37.4%

5. Tech_Support : Customers who opted for Tech Support service are less likely to churn than customers who have not opted for Tech Support service or do not have internet service by 31.59%

6. Contract_Month to Month : Customers with month to month contract are more likely to churn as compared to customers with Two-year contracts by 297.4%.

7. Contract_One Year : Customers with one-year contract are more likely to churn as compared to customers with Two-year contracts by 102.19%.

8. Payment method: Customers with payment method as electronic check are more likely to churn as compared to customers with any other payment method by 52.8%.

9. Senior Citizen: Senior citizen customers are more likely to churn than who are not senior citizens by 30.49% .

10. Tenure: With 1 unit increase in tenure, odds of churn decrease by 5.8%

11. Total Charges: With 1 unit increase in Total Charges, odds of churn increase by a percentage very close to zero.

Answer-3

Tenure, **Internet Services** and **Total Charges** are the top three factors that affect churn in our model. These are highly significant in the model and have highest standardized estimates.

Answer-4

- **Tariff/Offers provided by competitors:** It is important to know what the other network providers are offering, it can be a good predictor of why a customer is churning
- **Types of complaints by customers and their frequency:** Can be a good predictor, based on complaints we can tell how happy/unhappy the customers are.
- **Occupation of customer :** Sometimes in a job, company offers free of cost phone services to the employee, these people could churn on their current network. Or else there are job that require frequent international travel, these customers would only prefer networks that provide international roaming on low prices, so occupation of a customer also has a significant role to play we believe.
- **OTT:** Over the top services can also be a good variable to include. A network provider giving free1 year Netflix subscription to users can tempt a user to churn from current network, so taking OTT services in to account would be a good idea.

- **Network Strength in location:** Network strength could also lead a customer to churn. On moving to a new city, if a user's current network has poor connectivity in that location, user would most likely switch to another network with better network strength.
- **Total Internet Usage :** It may be possible that customers could churn based on the internet provided by the other network providers. Given all other services provided are same, a person could churn depending on the total internet usage provided by the network provider based on his/her personal usage needs.
- **Talktime:** Like Internet usage, talk time offered could also help the model

Above listed are some other variables (if collected), that would help to improve the fit of the model.

Answer-5

Analysis Variable : hit				
N	Mean	Std Dev	Minimum	Maximum
7032	0.7979238	0.4015773	0	1.0000000

- We have got a hit-score of 80%, i.e. 80% of the predictions made by the model are correct.
- Out of 7032 observations that we had in the data, 5163 were for churn=No, which means our naïve prediction of all Non-churners would be 73.42% accurate.
- We have got accuracy higher than the naïve prediction accuracy.