Tool Selection

There are many options in the telecommunication industry for customers to choose from. When a customer leaves one company for another its known as churn. Companies can put time and resources into reducing churn by analyzing customers' data and then determining what factors contribute to the churn rate. The goal of this project is to apply Principal Component Analysis and Logistic Regression Analysis using SAS software to reduce the churn rate. Principal Component Analysis (PCA) is a statistical technique used to analyze the correlations between a high number of variables and to offer explanation of the variables in terms of a smaller number of variables, referred to as principal components, without losing information (Principal Component Analysis, n.d.). By completing PCA we can determine which areas a company should focus on to retain customers, therefore not wasting efforts on factors that do not reduce churn rate.

Logistic Regression is a type of regression analysis used to describe data and to explain relationships between one dependent binary variable and one or more independent variables (Thanda, 2020). Since a customer will either churn or they will not, there is only two possible outcomes, making it a binary outcome. Logistic regression is easier to train staff to use, and it is easier to implement compared to other methods (Thanda, 2020).

SAS is a programming language that is commonly used for analytical purposes (Advantages of SAS, 2019) . SAS makes the manipulation of data very straightforward and presents data in a wide variety of meaningful reports that can be saved in many different formats. With the capability of analyzing statistics with sophisticated models and a wide range of techniques, SAS works well to analyze data like the dataset for this project. SAS Studio also allows SAS users to integrate Java, R programming language or even C into their SAS programs. SAS users with the expertise in those programming languages can take full advantages of them to further enhance their SAS code. The SAS interface is easier to understand and visualize data than Python and R programming language (Advantages of SAS, 2019).

Data Exploration and Preparation

The goal of this project is to determine factors that can reduce churn. Churn is our target variable for this project. In the data set, churn will be either a “Yes” or a “No”, making it a binary variable. The variables that will be analyzed as they affect churn are the predictor variables. In this data set, the predictor variables have two different formats, Char and Num, as seen in the SAS output table below.

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
21	Churn	Char	3	$3.	$3.
16	Contract	Char	14	$14.	$14.
5	Dependents	Char	3	$3.	$3.
12	DeviceProtection	Char	19	$19.	$19.
9	InternetService	Char	11	$11.	$11.
19	MonthlyCharges	Num	8	BEST12.	BEST32.
8	MultipleLines	Char	16	$16.	$16.
11	OnlineBackup	Char	19	$19.	$19.
10	OnlineSecurity	Char	19	$19.	$19.
17	PaperlessBilling	Char	3	$3.	$3.
4	Partner	Char	3	$3.	$3.
18	PaymentMethod	Char	25	$25.	$25.
7	PhoneService	Char	3	$3.	$3.
3	SeniorCitizen	Num	8	BEST12.	BEST32.
15	StreamingMovies	Char	19	$19.	$19.
14	StreamingTV	Char	19	$19.	$19.
13	TechSupport	Char	19	$19.	$19.
20	TotalCharges	Num	8	BEST12.	BEST32.
1	customerID	Char	10	$10.	$10.
2	gender	Char	6	$6.	$6.
6	tenure	Num	8	BEST12.	BEST32.

The variables that are in Char format will be changed to Num format by coding and converting them to numbers. For example, the variable Partner has either a “Yes” or “No” answer. In the same way Churn was converted, Partner will also be converted where “No” equals 0 and “Yes” equals 1. Some of the variables MulitpleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, and StreamingMovies have a value of “No Internet Service”. These will be changed to “No”, and then “No” will be equal to 0, and “Yes” equal to 1 for further analysis. PaymentMethod will be changed so that “Electronic check” will be equal to 1, “Mailed Check” equal to 2, “Bank transfer (automatic)” equal to 3, and Credit card (automatic)” equal to 4. Since tenure covers a wide range of time, they were grouped into bins of one year and given a number one to six to represent years, respectively. The CustomerID variable will be dropped as it has no impact on the data analysis. By running PROC MEANS it is discovered there are 11 missing inputs for TotalCharges, as seen in the table below. These inputs account for less than one percent of the data, so they will be dropped.

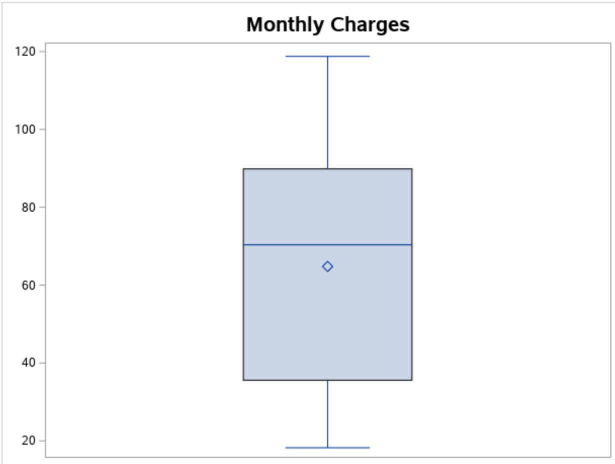
The MEANS Procedure

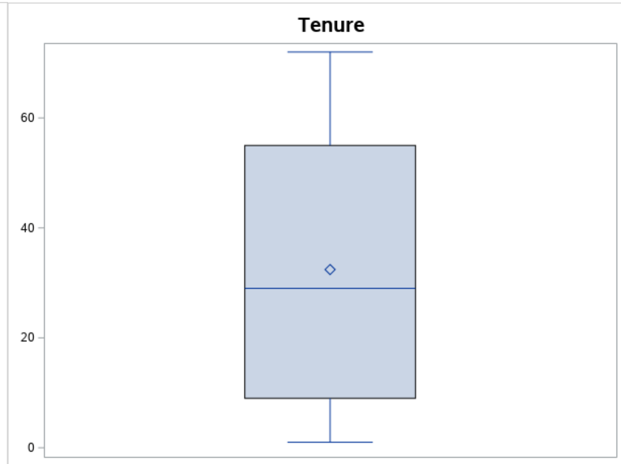
Variable	N Miss	N
SeniorCitizen	0	7043
tenure	0	7043
MonthlyCharges	0	7043
TotalCharges	11	7032

A correlation matrix was made with Monthly Charges and Total Charges as seen below. Since they have such a strong correlation, Total Charges will be removed.

	MonthlyCharges	TotalCharges
MonthlyCharges	1.00000 7043	0.65106 <.0001 7032
TotalCharges	0.65106 <.0001 7032	1.00000 7032

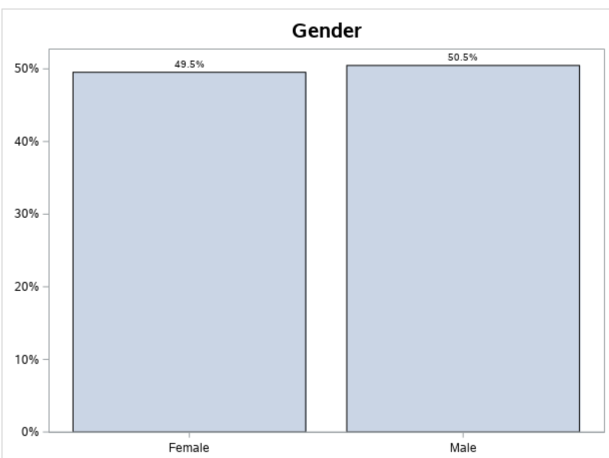
Monthly Charges and tenure were plotted using a box and whisker plot to see if any outliers were present as seen below. No outliers were present.

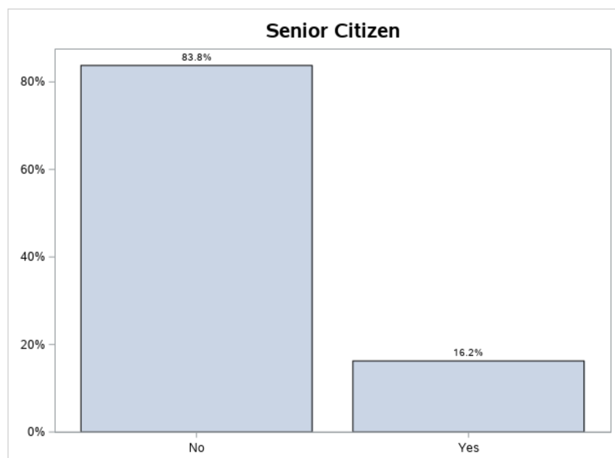




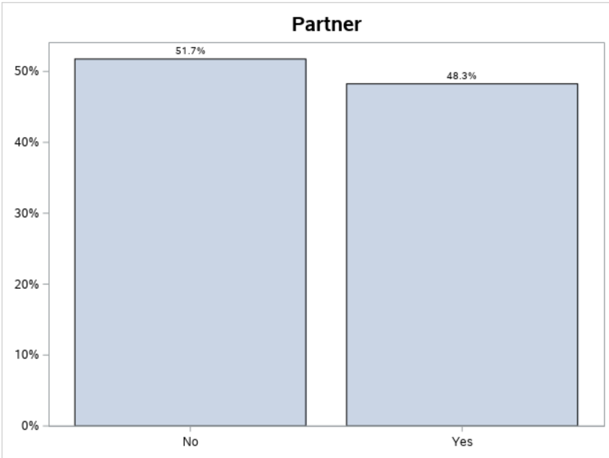
Data Analysis

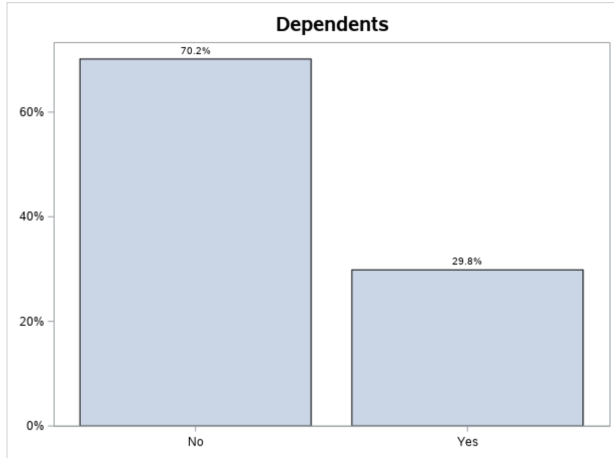
The graphs below show the distribution of the variables in the data set.



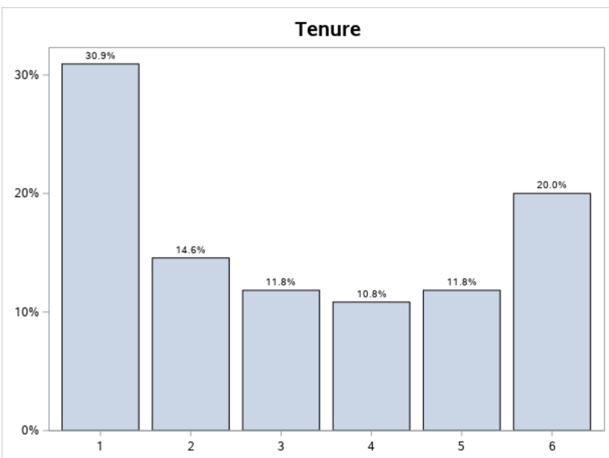


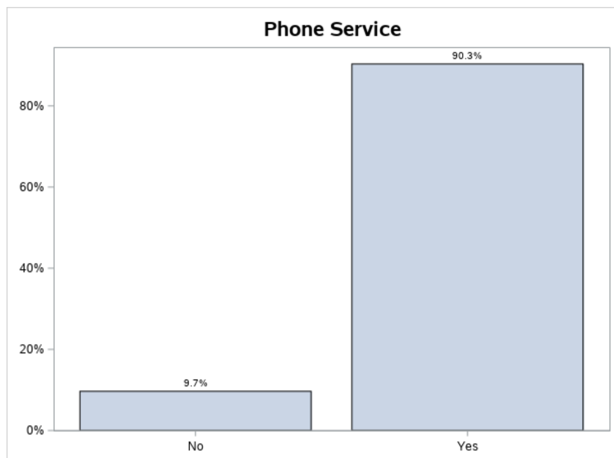
Gender is distributed evenly in the data set. Senior citizens were not well represented in the data.



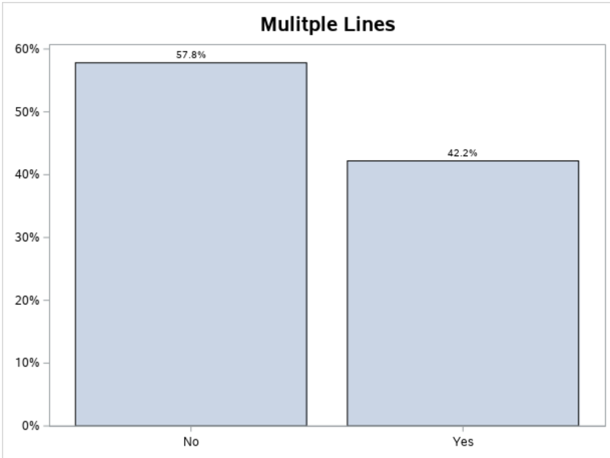


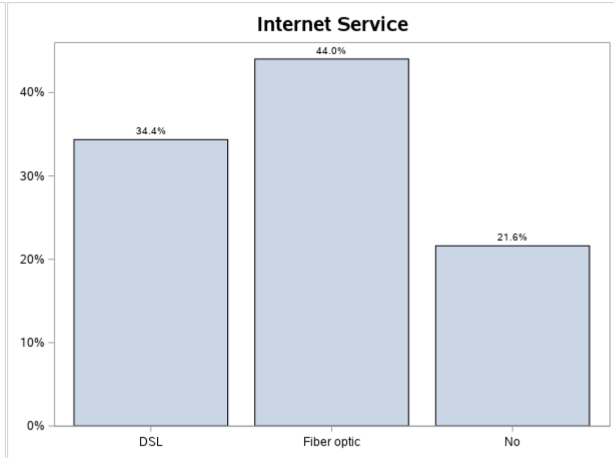
The number of customers who had a partner were close to equal, while the number of customers with dependents was not.



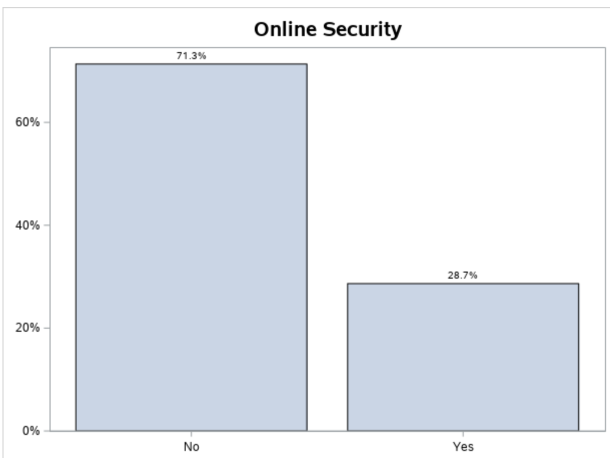


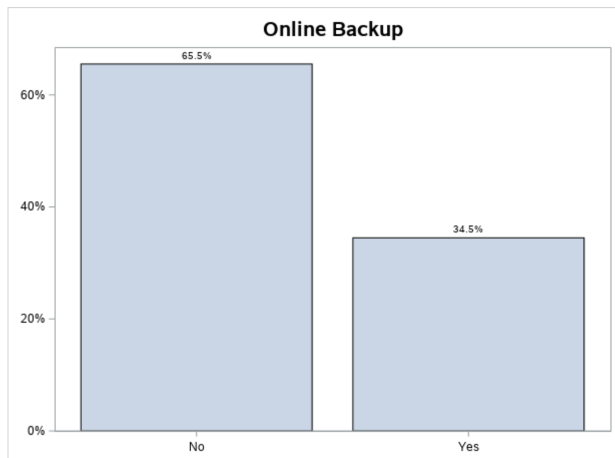
As mentioned before, tenure was placed into 1-year groups. Customers who were with the company up to one year were most represented. It makes sense that most customers had phone service.

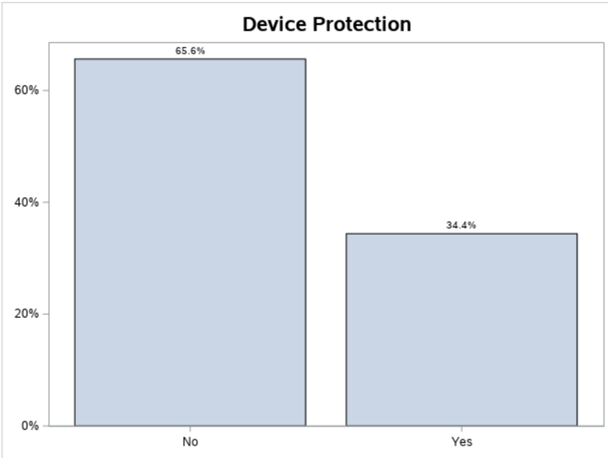


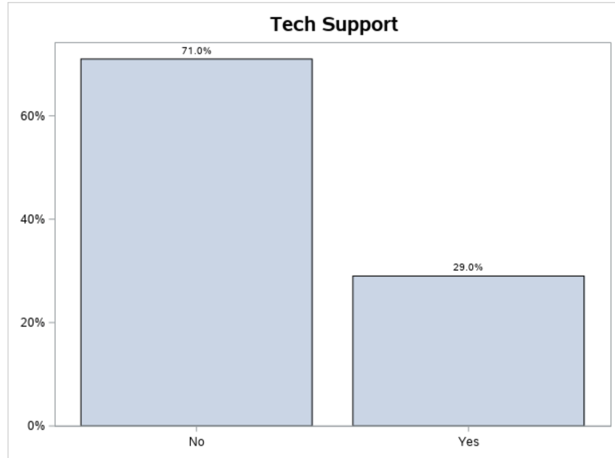


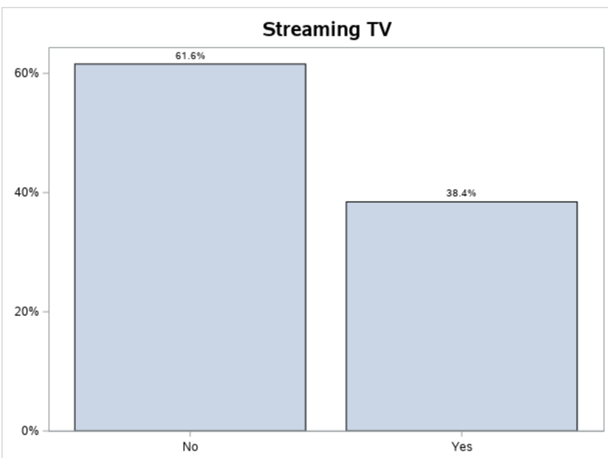
Fiber Optics was the preferred method of internet service.

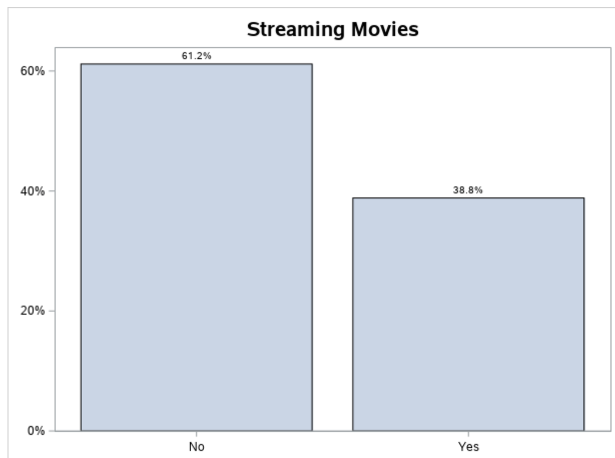


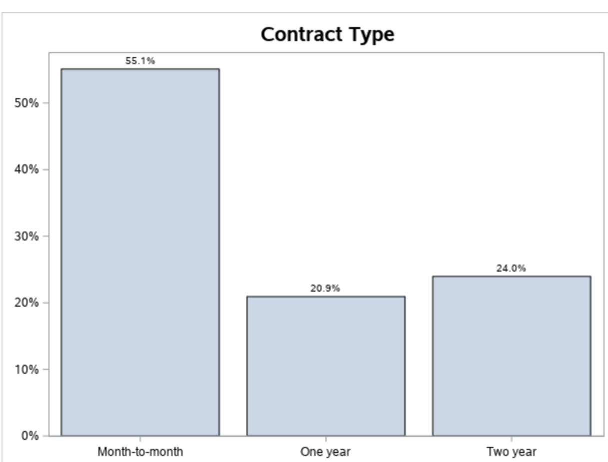


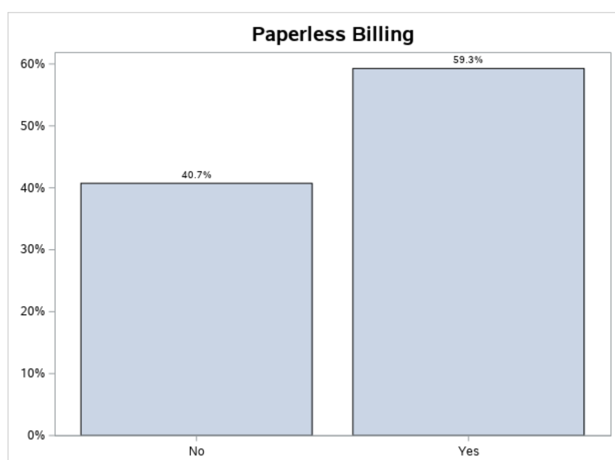

Device Protection
65.6%
34.4%
60%
40%
20%
0%
No
Yes

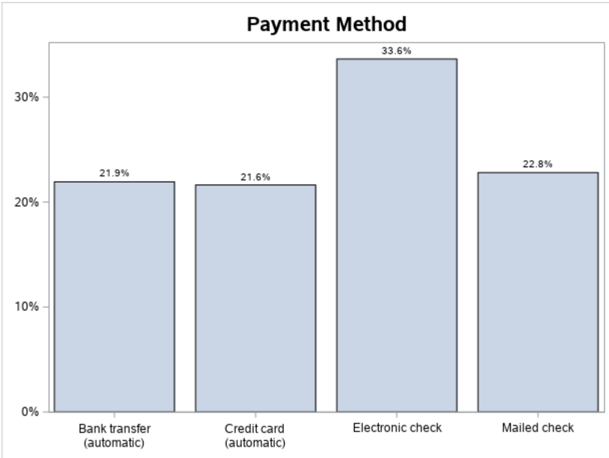

Tech Support
71.0%
29.0%
60%
40%
20%
0%
No
Yes

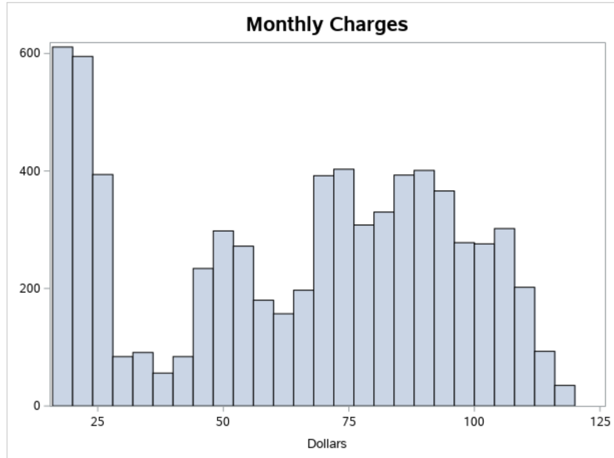

Streaming TV
61.6%
38.4%
60%
40%
20%
0%
No
Yes

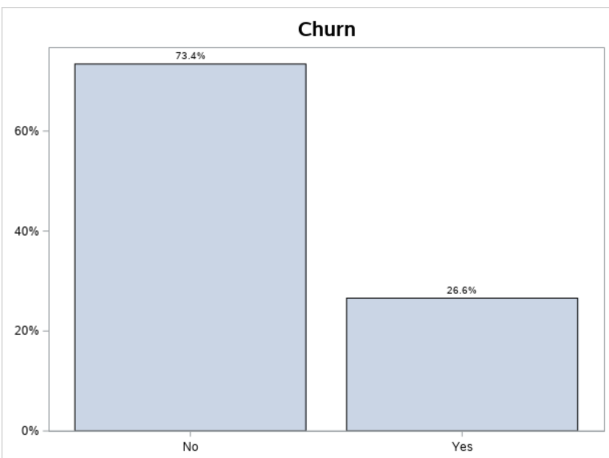

Streaming Movies
61.2%
38.8%
60%
40%
20%
0%
No
Yes


Contract Type
55.1%
20.9%
24.0%
50%
40%
30%
20%
10%
0%
Month-to-month
One year
Two year


Paperless Billing
40.7%
59.3%
60%
50%
40%
30%
20%
10%
0%
No
Yes

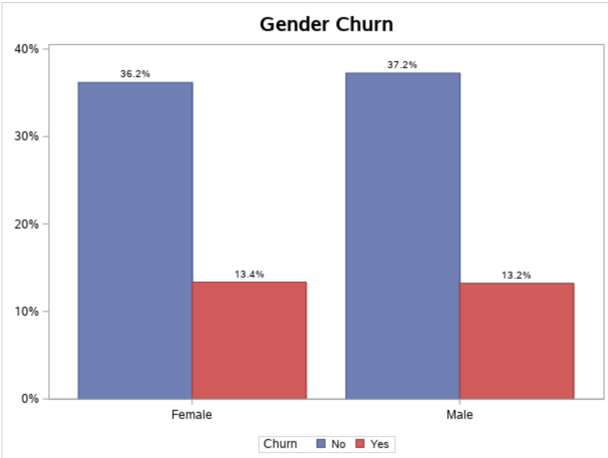


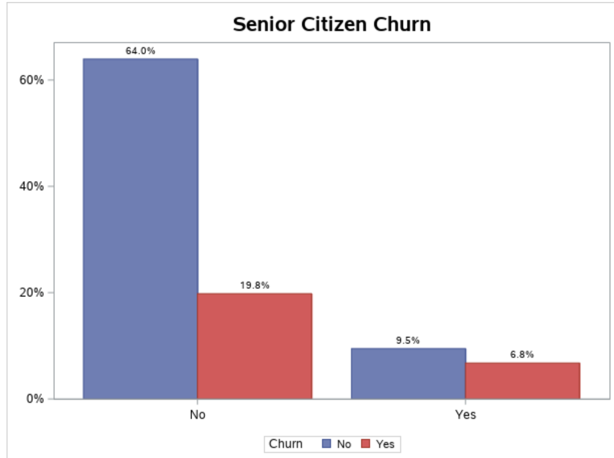


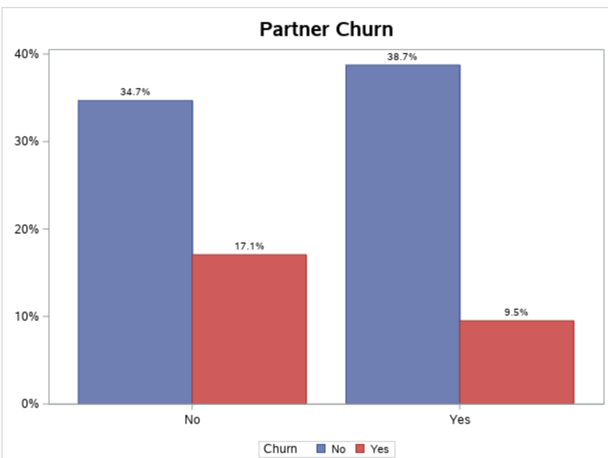


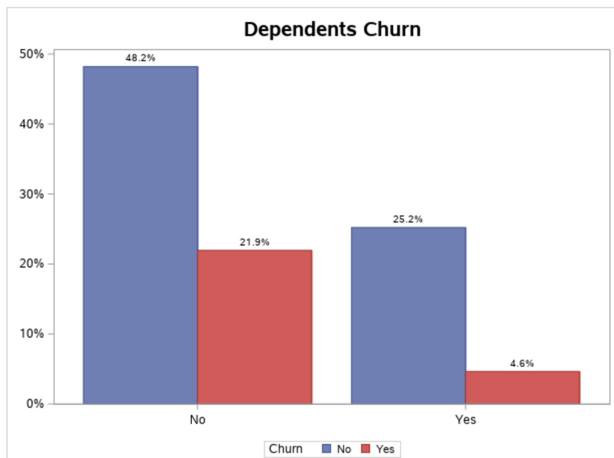
The churn graph shows the churn rate to be 26.6%.

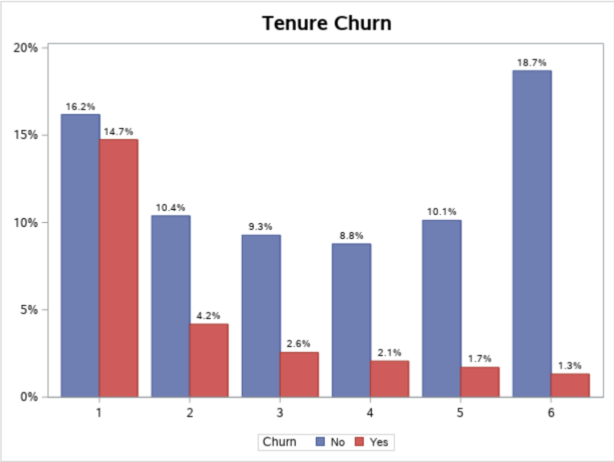
The bivariate distributions of the data are presented in the bar graphs below.

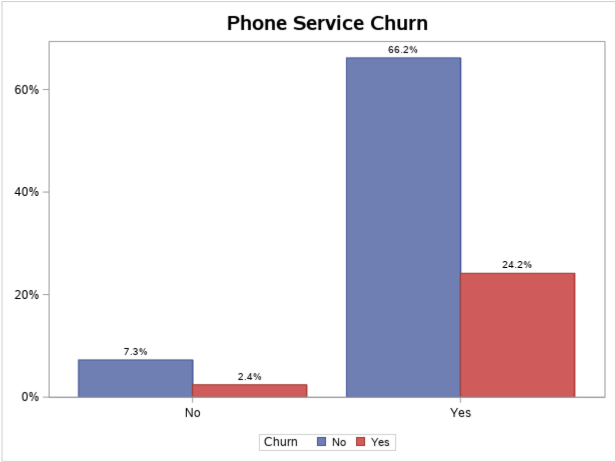

Gender Churn
40%
30%
20%
10%
0%
36.2%
13.4%
37.2%
13.2%
Female
Male
Churn
No
Yes


Senior Citizen Churn
60%
40%
20%
0%
64.0%
19.8%
9.5%
6.8%
No
Yes
Churn
No
Yes

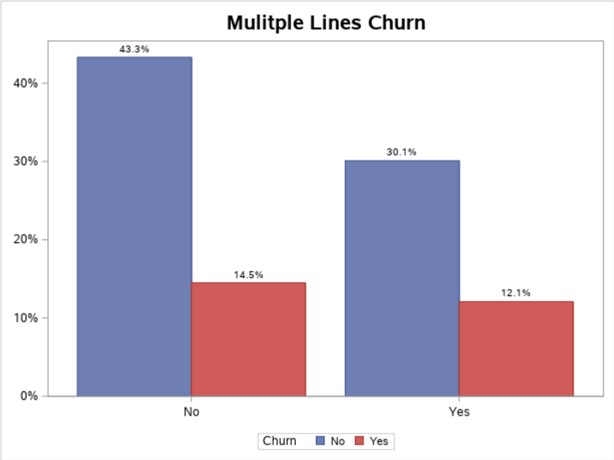

Partner Churn
40%
30%
20%
10%
0%
34.7%
17.1%
38.7%
9.5%
No
Yes
Churn
No
Yes

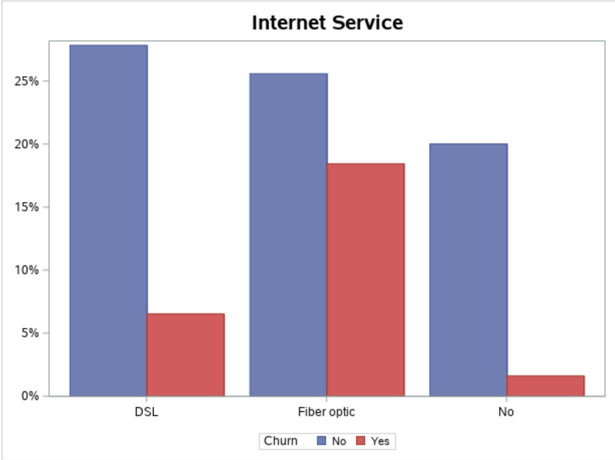

Dependents Churn
50%
40%
30%
20%
10%
0%
48.2%
21.9%
25.2%
4.6%
No
Yes
Churn
No
Yes



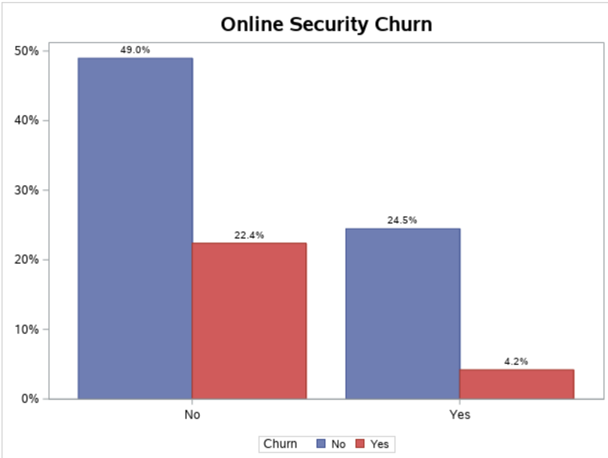


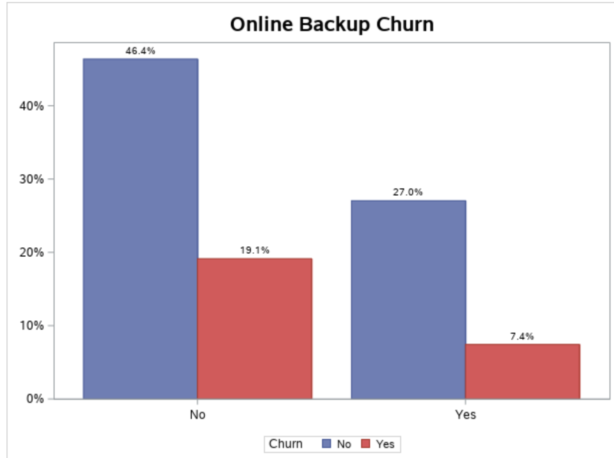
Customers who were with the company under a year had the highest churn rate, while customers who have been with the company the longest had the lowest churn rate.

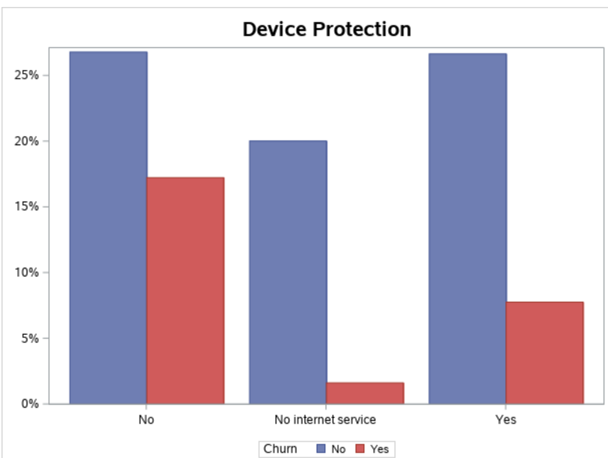


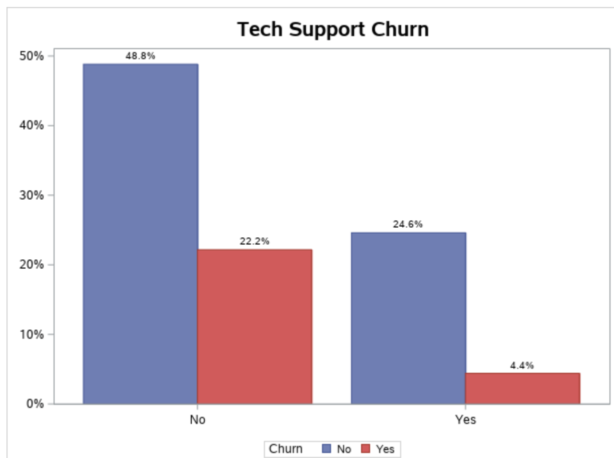


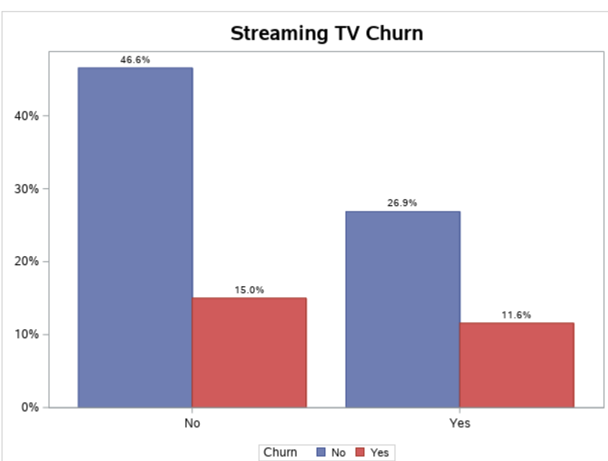
Customers with Fiber Optic internet service had a higher churn rate than did DSL customers.

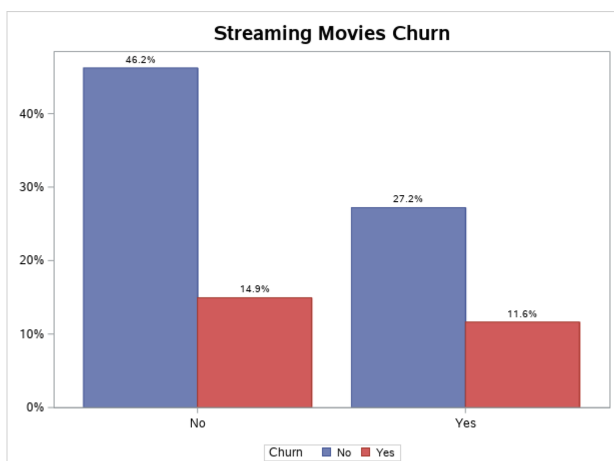

Online Security Churn
50%
40%
30%
20%
10%
0%
49.0%
22.4%
24.5%
4.2%
No
Yes
Churn No Yes

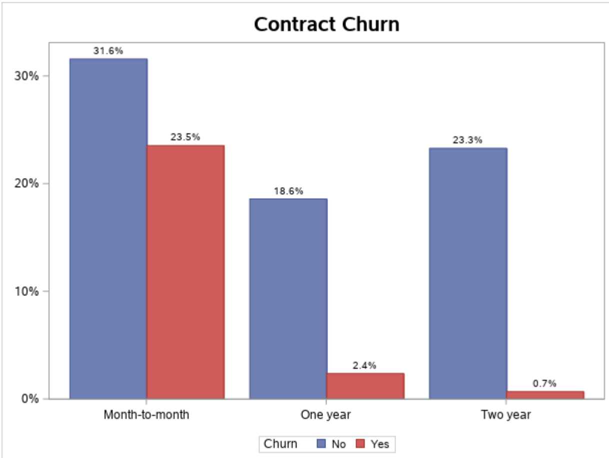

Online Backup Churn
40%
30%
20%
10%
0%
46.4%
19.1%
27.0%
7.4%
No
Yes
Churn No Yes

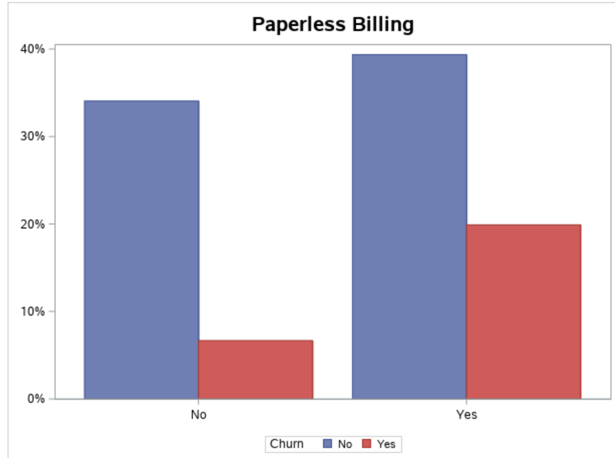

Device Protection
25%
20%
15%
10%
5%
0%
No
No internet service
Yes
Churn No Yes


Tech Support Churn
50%
40%
30%
20%
10%
0%
48.8%
22.2%
24.6%
4.4%
No
Yes
Churn No Yes

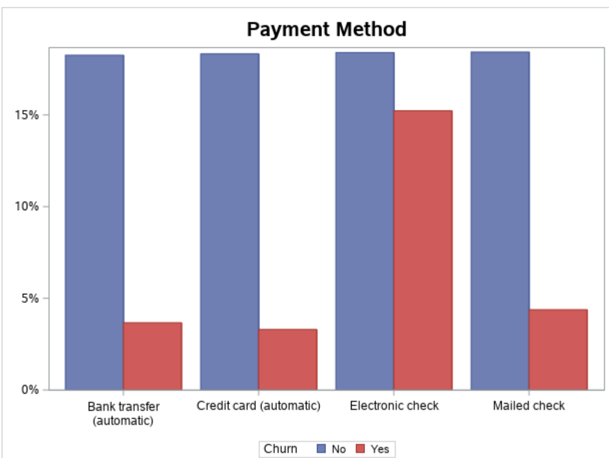

Streaming TV Churn
40%
30%
20%
10%
0%
46.6%
15.0%
26.9%
11.6%
No
Yes
Churn No Yes


Streaming Movies Churn
40%
30%
20%
10%
0%
46.2%
14.9%
27.2%
11.6%
No
Yes
Churn No Yes





Contracts on a month-to-month basis had the highest churn rate.



Payment by Electronic Check had the highest churn rate

As mentioned earlier, Principal Component Analysis (PCA) allows content in a data set to be summarized by smaller groups of data that can be more easily visualized and analyzed. SAS makes

Correlation Matrix

	SeniorCitizen	MonthlyCharges	gender_new	partner_new	dep_new	phonser_new	papbill_new	tenure_new	multi_new	online_new	backup_new	device_new	tech_new	tv_new	movie_new	internet_new	contract_new	payment_new
SeniorCitizen	1.0000	0.2199	-.0018	0.0170	-.2106	0.0084	0.1563	0.0160	0.1430	-.0386	0.0667	0.0595	-.0606	0.1054	0.1198	0.2590	-.1418	-.0937
MonthlyCharges	0.2199	1.0000	-.0138	0.0978	-.1123	0.2480	0.3519	0.2419	0.4909	0.2964	0.4415	0.4826	0.3383	0.6297	0.6272	0.9054	-.0727	-.0746
gender_new	-.0018	-.0138	1.0000	-.0014	0.0103	-.0075	-.0119	0.0060	-.0089	-.0163	-.0131	-.0008	-.0085	-.0071	-.0101	-.0096	0.0001	-.0049
partner_new	0.0170	0.0978	-.0014	1.0000	0.4523	0.0184	-.0140	0.3699	0.1426	0.1433	0.1418	0.1536	0.1202	0.1245	0.1181	0.0009	0.2941	0.1333
dep_new	-.2106	-.1123	0.0103	0.4523	1.0000	-.0011	-.1101	0.1575	-.0243	0.0808	0.0236	0.0139	0.0631	-.0165	-.0384	-.1778	0.2406	0.1240
phonser_new	0.0084	0.2480	-.0075	0.0184	-.0011	1.0000	0.0167	0.0075	0.2795	-.0917	-.0521	-.0701	-.0951	-.0214	-.0335	0.0942	0.0030	-.0031
papbill_new	0.1563	0.3519	-.0119	-.0140	-.1101	0.0167	1.0000	0.0037	0.1637	-.0041	0.1271	0.1041	0.0375	0.2242	0.2116	0.3776	-.1755	-.1018
tenure_new	0.0160	0.2419	0.0060	0.3699	0.1575	0.0075	0.0037	1.0000	0.3242	0.3181	0.3559	0.3520	0.3169	0.2732	0.2781	0.0322	0.6626	0.3324
multi_new	0.1430	0.4909	-.0089	0.1426	-.0243	0.2795	0.1637	0.3242	1.0000	0.0986	0.2022	0.2017	0.1004	0.2578	0.2592	0.3451	0.1075	0.0360
online_new	-.0386	0.2964	-.0163	0.1433	0.0808	-.0917	-.0041	0.3181	0.0986	1.0000	0.2833	0.2749	0.3545	0.1755	0.1874	0.1565	0.2457	0.1628
backup_new	0.0667	0.4415	-.0131	0.1418	0.0236	-.0521	0.1271	0.3559	0.2022	0.2833	1.0000	0.3031	0.2937	0.2816	0.2745	0.3072	0.1553	0.0962
device_new	0.0595	0.4826	-.0008	0.1536	0.0139	-.0701	0.1041	0.3520	0.2017	0.2749	0.3031	1.0000	0.3329	0.3899	0.4023	0.3134	0.2196	0.1110
tech_new	-.0606	0.3383	-.0085	0.1202	0.0631	-.0951	0.0375	0.3169	0.1004	0.3545	0.2937	0.3329	1.0000	0.2775	0.2802	0.1645	0.2940	0.1672
tv_new	0.1054	0.6297	-.0071	0.1245	-.0165	-.0214	0.2242	0.2732	0.2578	0.1755	0.2816	0.3899	0.2775	1.0000	0.5334	0.4296	0.1042	-.0142
movie_new	0.1198	0.6272	-.0101	0.1181	-.0384	-.0335	0.2116	0.2781	0.2592	0.1874	0.2745	0.4023	0.2802	0.5334	1.0000	0.4268	0.1091	-.0043
internet_new	0.2590	0.9054	-.0096	0.0009	-.1778	0.0942	0.3776	0.0322	0.3451	0.1565	0.3072	0.3134	0.1645	0.4296	0.4268	1.0000	-.2889	-.1787
contract_new	-.1418	-.0727	0.0001	0.2941	0.2406	0.0030	-.1755	0.6626	0.1075	0.2457	0.1553	0.2196	0.2940	0.1042	0.1091	-.2889	1.0000	0.3595
payment_new	-.0937	-.0746	-.0049	0.1333	0.1240	-.0031	-.1018	0.3324	0.0360	0.1628	0.0962	0.1110	0.1672	-.0142	-.0043	-.1787	0.3595	1.0000

running PCA easier than other analytics software because of the point and click capability. Clicking on the Principal Component Analysis option brings up prewritten code and gives a box to add the variables to be analyzed. After all the variables are added, a correlation matrix is produced, as seen below.

Before conducting PCA, it is important to check for correlations between variables. If any of the correlations are too high, above .9, you may need to remove one of the variables from the analysis. If the correlations are too low, below .1, they could be problematic as well. An alternative to removing the variable is to combine them for a new variable to be analyzed. From running the correlation analysis previously, we determined Total Charges and Monthly Charges were highly correlated and Total Charges was dropped. After confirming the results of the correlation matrix are acceptable, then next step is to look at the eigenvalues. The first component will always account for the most variance and will have the highest eigenvalue. Each successive component will account for less variance. Using the Kaiser criterion, (Kaiser,1960), any eigenvalue with a value greater than 1 will be kept. The output below shows the eigenvalues:

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	4.29879073	1.60022323	0.2388	0.2388
2	2.69856751	1.32902547	0.1499	0.3887
3	1.36954204	0.23230341	0.0761	0.4648
4	1.13723863	0.13062173	0.0632	0.5280
5	1.00661690	0.01384695	0.0559	0.5839
6	0.99276995	0.11882319	0.0552	0.6391
7	0.87394676	0.05359882	0.0486	0.6876
8	0.82034794	0.10380640	0.0456	0.7332
9	0.71654154	0.03015061	0.0398	0.7730
10	0.68639093	0.05349772	0.0381	0.8112
11	0.63289321	0.03115911	0.0352	0.8463
12	0.60173410	0.02053350	0.0334	0.8797
13	0.58120060	0.10161874	0.0323	0.9120
14	0.47958186	0.01538817	0.0266	0.9387
15	0.46419369	0.07594364	0.0258	0.9645
16	0.38825005	0.13760830	0.0216	0.9860
17	0.25064175	0.24988994	0.0139	1.0000
18	0.00075181		0.0000	1.0000

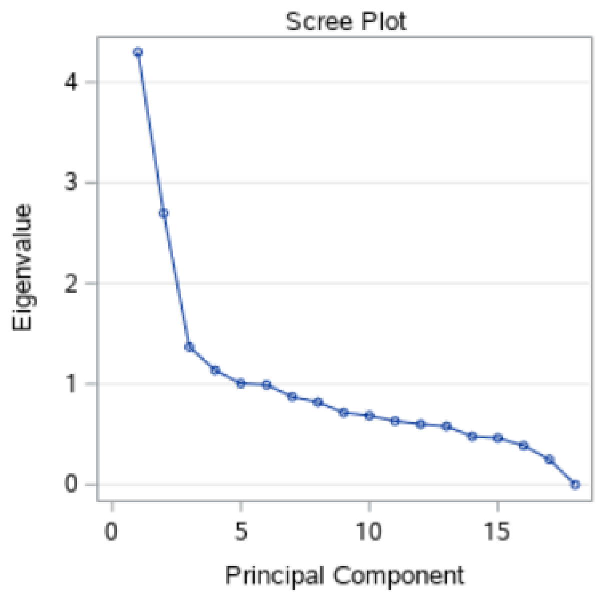
The eigenvectors are the correlation between the principal components and the original variables.

Using the table below, the data shows which variables are correlated for each principal component.

Eigenvectors					
	Prin1	Prin2	Prin3	Prin4	Prin5
gender_new	-.008714	0.005234	-.001069	-.017002	0.986972
SeniorCitizen	0.100936	-.228756	0.057846	0.675885	0.040033
dep_new	-.018593	0.272194	0.083457	-.544281	0.043295
tenure_new	0.254296	0.382267	0.187338	0.227949	0.037973
phonser_new	0.036997	-.096778	0.701667	-.217614	-.041590
multi_new	0.253456	-.048724	0.504325	0.050843	0.003446
internet_new	0.339315	-.335600	-.019225	-.101822	-.010907
contract_new	0.105293	0.499730	0.160094	0.140433	0.025420
papbill_new	0.168267	-.250330	-.030243	0.023931	-.004767
payment_new	0.043061	0.357581	0.132441	0.248016	-.036102
MonthlyCharges	0.438760	-.206877	0.078911	-.138894	-.009265
online_new	0.212089	0.218952	-.204334	-.053047	-.089441
backup_new	0.278263	0.090326	-.105104	0.058240	-.039475
device_new	0.309180	0.105169	-.155487	0.014133	0.036189
tech_new	0.245092	0.219858	-.241151	-.101030	-.040287
tv_new	0.340045	-.049386	-.099253	-.112587	0.050609
movie_new	0.341336	-.045448	-.110678	-.063198	0.044675

In the first principal component, Monthly Charges, type of Internet Service, Streaming Movies, Streaming TV, Device Protection, Online Backup, and Tenure are the most correlated.

In the second principal component, Partner status, Dependent status, Tenure, Contract type and Payment type are the most correlated. Using the table, the correlations for the third, fourth and fifth principal component can be seen. In addition to the eigenvectors table, a scree plot can show how many components should be kept.



This also shows that 5 principal components are above the eigenvalue of 1 and should be kept.

By using Principal Component Analysis, each component shows the variables that most correlated with each and therefore should be studied together for the most benefit.

Where PCA is a type of descriptive analysis, Logistic Regression is used for predictive analysis. Logistic Regression is appropriate when the dependent variable is binary, which in this case there is only two outcomes, either a customer will churn, or they will not. SAS does an excellent job and is more straightforward than other logistic analysis software because of the ease of use of language and the point-and-click capabilities (Advantages of SAS, 2019). Using SAS binary logistic regression option, the variables are entered as well as the dependent variable and the code is generated and run. SAS also can take categorical data and automatically generate dummy variables. The first table below shows that after running backward stepwise regression, there are 7 variables that can be taken out of the analysis and does not affect the results. The next table titled “Type 3 Analysis of Effects” shows the variables that are significantly significant at a p-value of 0.05.

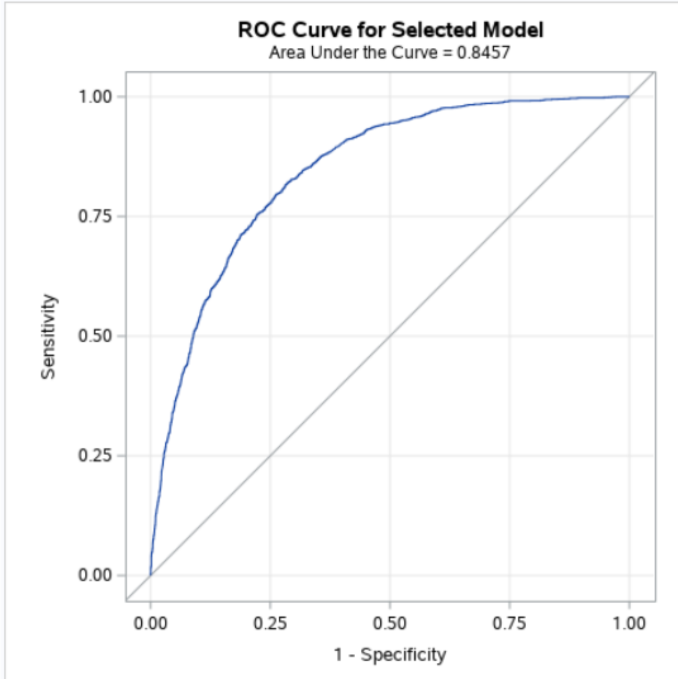
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	PhoneService	1	17	0.0090	0.9244
2	Partner	1	16	0.1339	0.7144
3	gender	1	15	0.1530	0.6957
4	OnlineBackup	1	14	0.2674	0.6051
5	DeviceProtection	1	13	1.5361	0.2152
6	Dependents	1	12	3.1709	0.0750

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
SeniorCitizen	1	8.6703	0.0032
InternetService	1	60.0848	<.0001
OnlineSecurity	1	9.3729	0.0022
TechSupport	1	5.1519	0.0232
StreamingTV	1	26.2568	<.0001
StreamingMovies	1	29.0889	<.0001
Contract	2	107.1111	<.0001
PaperlessBilling	1	19.0900	<.0001
PaymentMethod	3	25.7882	<.0001
tenure_new	5	241.2818	<.0001
multi_new	1	19.8116	<.0001
MonthlyCharges	1	17.9976	<.0001

The next output from the data shows the odd ratio estimates. The Point Estimate shows the relationship between the variable having a No value to a Yes value. The higher the Point Estimate the stronger the relationship.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
SeniorCitizen 0 vs 1	0.784	0.666	0.922
InternetService DSL vs No	7.121	3.730	13.594
InternetService Fiber optic vs No	31.368	12.001	81.992
OnlineSecurity No vs Yes	1.315	1.104	1.567
TechSupport No vs Yes	1.230	1.029	1.471
StreamingTV No vs Yes	0.608	0.502	0.735
StreamingMovies No vs Yes	0.596	0.494	0.719
Contract Month-to-month vs Two year	5.296	3.716	7.548
Contract One year vs Two year	2.438	1.712	3.470
PaperlessBilling No vs Yes	0.722	0.624	0.836
PaymentMethod Bank transfer (automatic) vs Mailed check	0.994	0.794	1.243
PaymentMethod Credit card (automatic) vs Mailed check	0.925	0.737	1.161
PaymentMethod Electronic check vs Mailed check	1.379	1.142	1.664
tenure_new 1 vs 6	5.914	4.292	8.150
tenure_new 2 vs 6	2.362	1.715	3.251
tenure_new 3 vs 6	1.603	1.158	2.219
tenure_new 4 vs 6	1.701	1.228	2.357
tenure_new 5 vs 6	1.291	0.934	1.786
multi_new 0 vs 1	0.676	0.569	0.803
MonthlyCharges	0.977	0.967	0.988

The next graph shows the Receiver Operator Characteristic (ROC) curve. The curve shows how effective this model is at predicting the likelihood of a customer churning. The curve also shows the trade-off between sensitivity and specificity. The Area Under the Curve score shows how good the model is at predicting churn. The data shows that the model is 85% effective at prediction.



The Hosmer-Lemeshow Test table shows the observed values and the expected values. These values also show this a good model for predicting churn.

Partition for the Hosmer and Lemeshow Test					
		churn_new = 1		churn_new = 0	
Group	Total	Observed	Expected	Observed	Expected
1	703	8	7.19	695	695.81
2	703	15	16.13	688	686.87
3	703	27	32.22	676	670.78
4	703	63	61.00	640	642.00
5	703	113	108.96	590	594.04
6	703	167	160.95	536	542.05
7	703	226	233.45	477	469.55
8	703	307	312.95	396	390.05
9	703	418	409.89	285	293.11
10	705	525	526.26	180	178.74

Logistic Regression is a quality tool to use for this data because its easier to implement, interpret, and very efficient to train. The outputs are easy to read and understand, and SAS makes running logistic regression and Principal Component Analysis very straightforward.

The tables and graphs are easy to read and interpret whereas outputs from other statistical analysis languages are not as well represented visually.

Data Summary

The data shows Fiber Optics service has having more influence on Churn than any other factor. Streaming services had the least impact. The lower the Tenure, the more influence it had on the Churn rate. If a customer paid online rather than mailing their payment, it seemed to help lower churn. The telecommunications company should focus on newer customers and customers who have Fiber Optics. If more effort could be put into these customers to keep them happier, the churn rate could be reduced.

References

Advantages of SAS: Disadvantages of SAS programming. (2019, May 01). Retrieved January 31, 2021, from https://data-flair.training/blogs/disadvantages-and-advantages-of-sas/#:~:text=%20Advantages%20of%20SAS%20%201%20Easy%20to,easy.%20We%20can%20understand%20and%20correct...%20More

Principal component analysis. (n.d.). Retrieved January 31, 2021, from https://www.real-statistics.com/multivariate-statistics/factor-analysis/principal-component-analysis/#:~:text=Principal%20component%20analysis%20is%20a%20statistical%20technique%20that,[xi]%20be%20any%20k%20%C3%97%201%20random%20vector.

Thanda, A. (2020, April 14). What is logistic regression? Retrieved January 31, 2021, from https://careerfoundry.com/en/blog/data-analytics/what-is-logistic-regression/#:~:text=Logistic%20regression%20is%20used%20for%20classification%20problems%20when,regression%20analysis,%20and%20different%20types%20of%20logistic%20regression.