

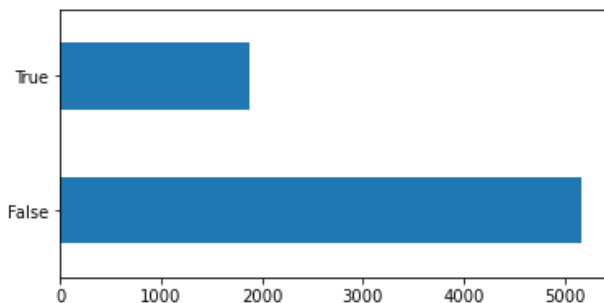


## Assignment #2

### Introduction:

The customer *churn rate* describes the rate at which customers leave a company or a service. For many organizations, this is a very important factor or behavior to model and understand. It is often more expensive to obtain a new customer base rather than to keep the existing one on board. Therefore it is worthwhile predicting which customers may want to leave and trying to keep them on board as customers.

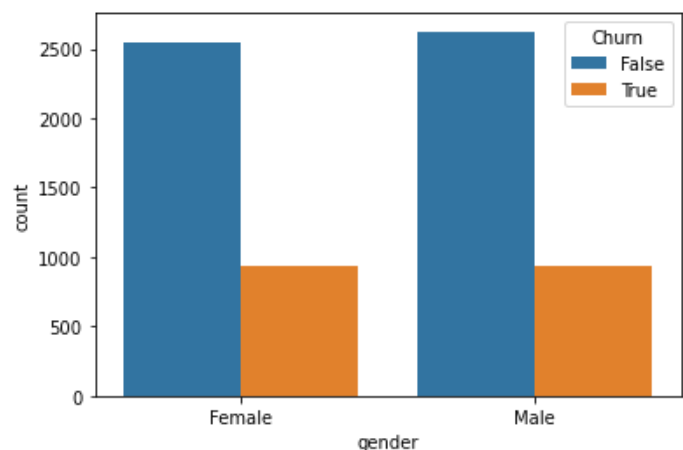
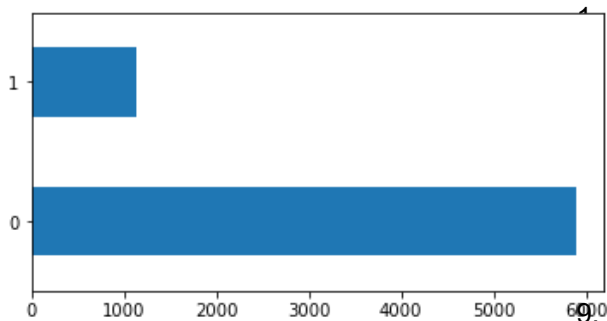
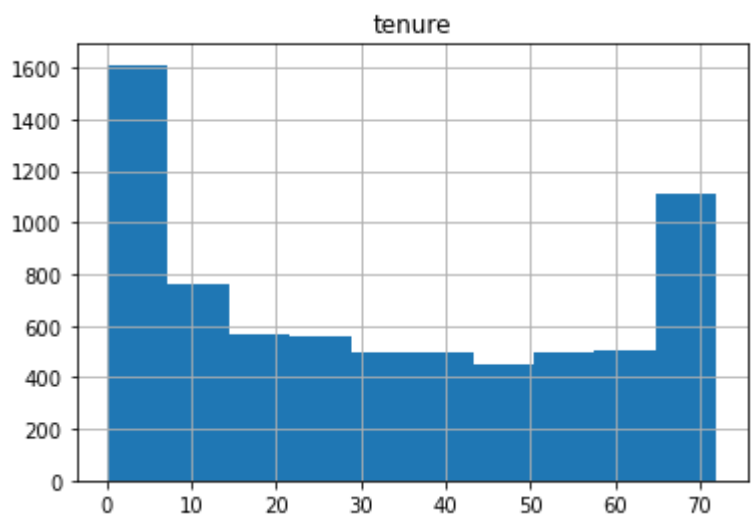
### Plots:



#### Churn:

Customers who do not churn: 5174: 73.46%

Customers who churn: 1869 : 26.54%

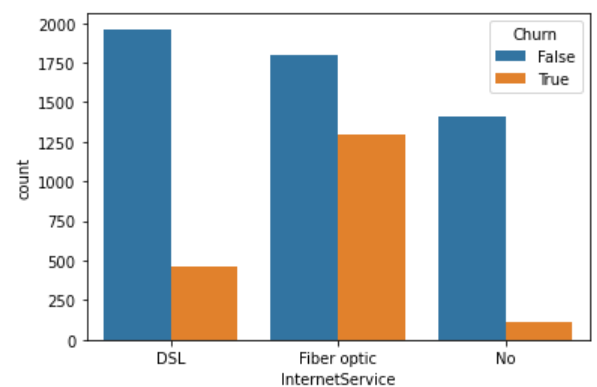
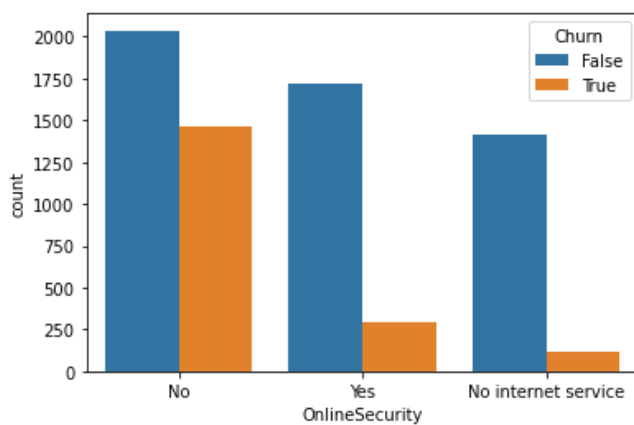
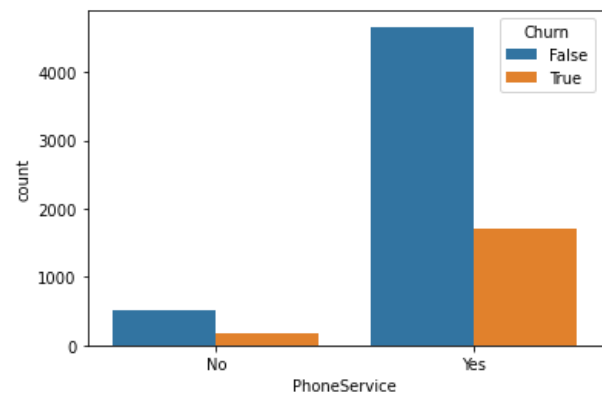
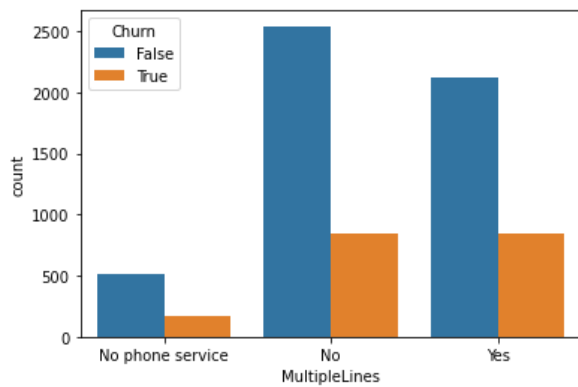
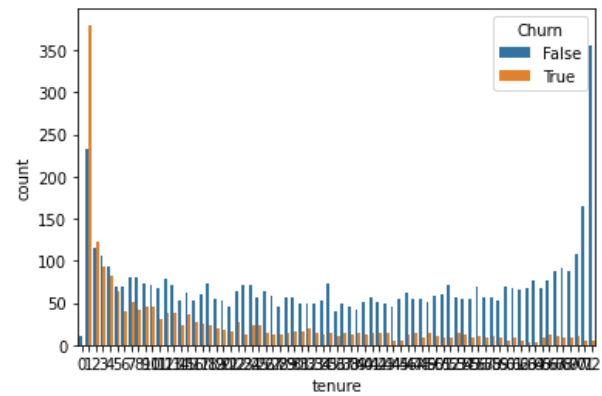
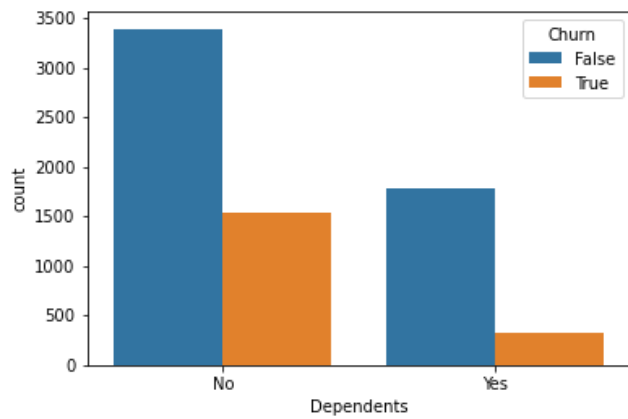
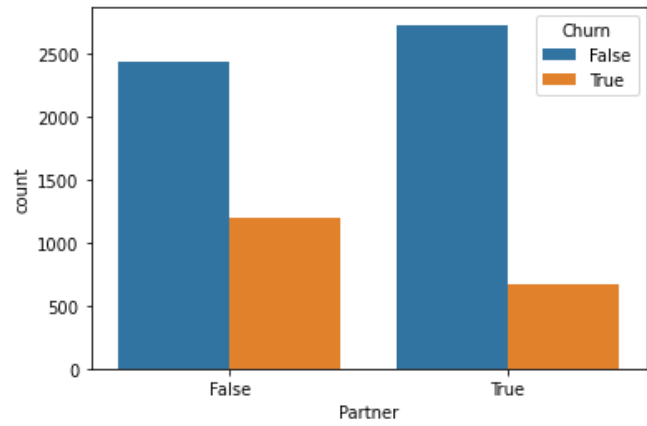
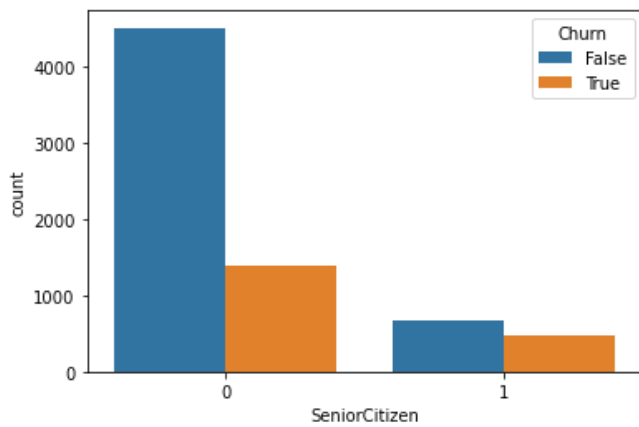


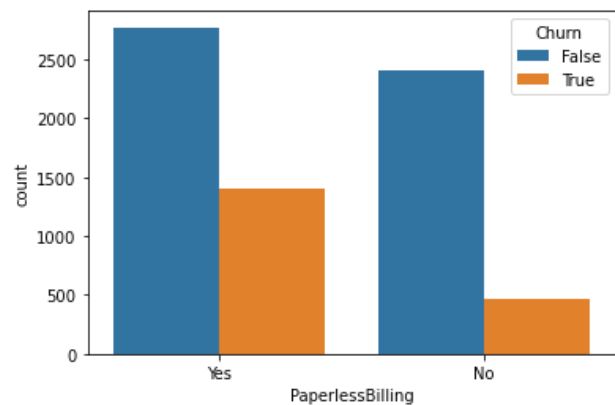
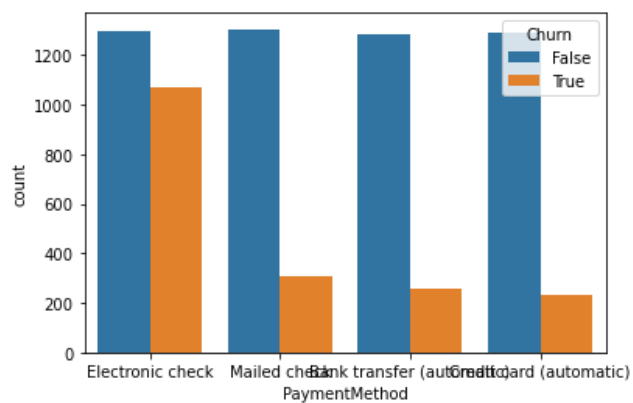
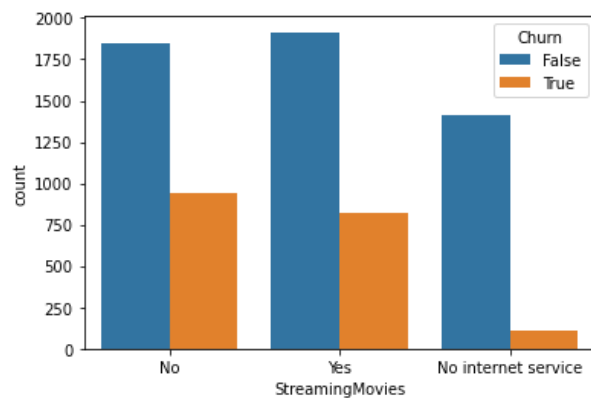
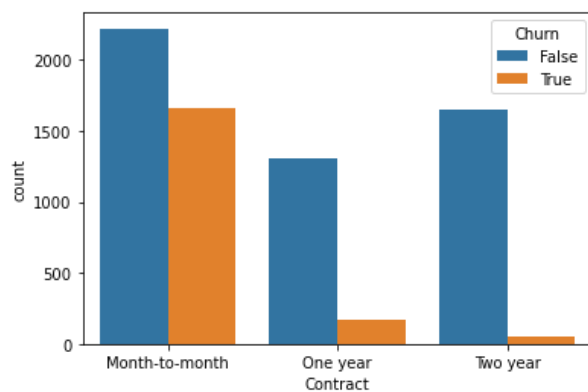
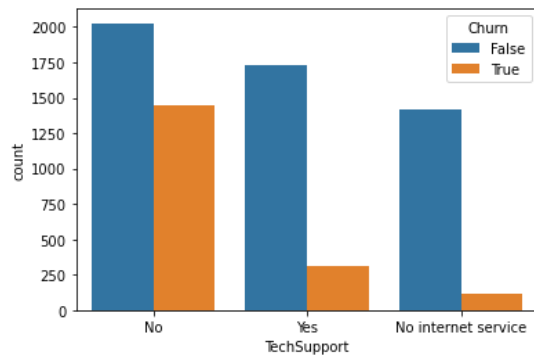
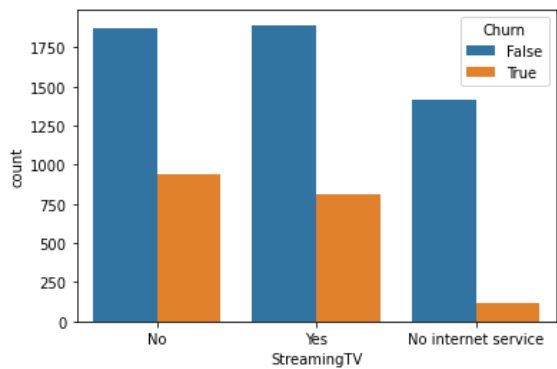
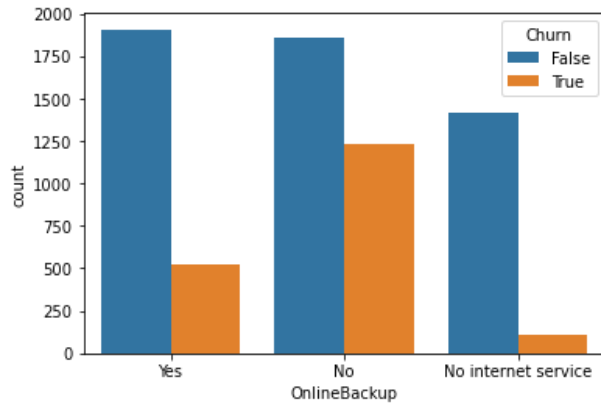
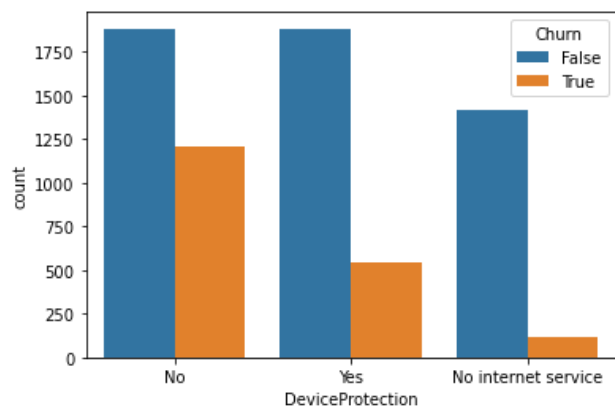
- Not senior citizen: 5901 = 83.785319%
  - Senior citizen: 1142 = 16.214681%
- Name: SeniorCitizen, dtype: int64

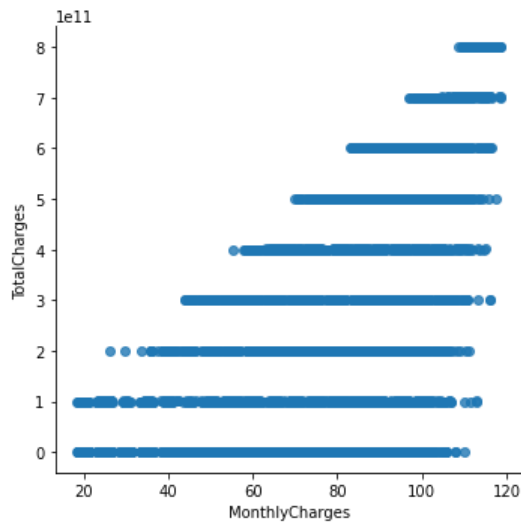
Gender wise churn rate:  
Approximately 50-50



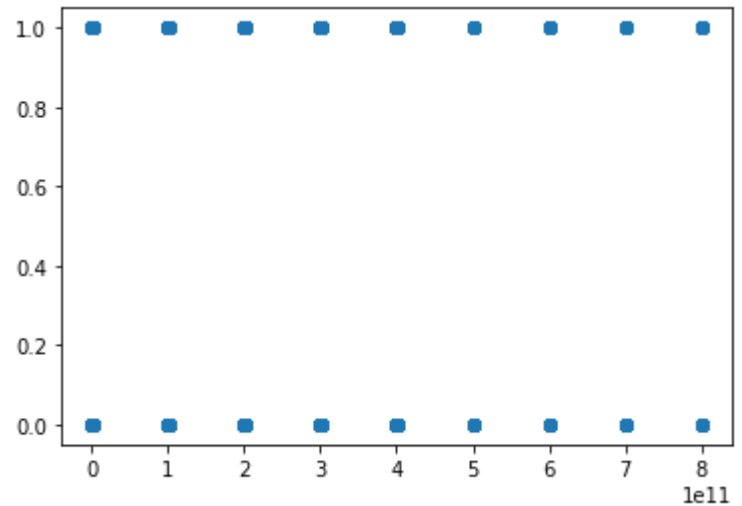
Other similar plots:



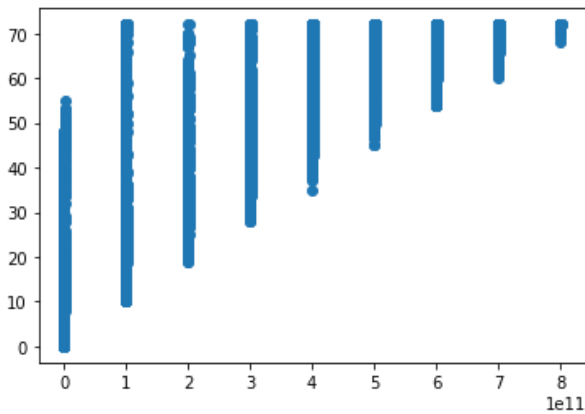




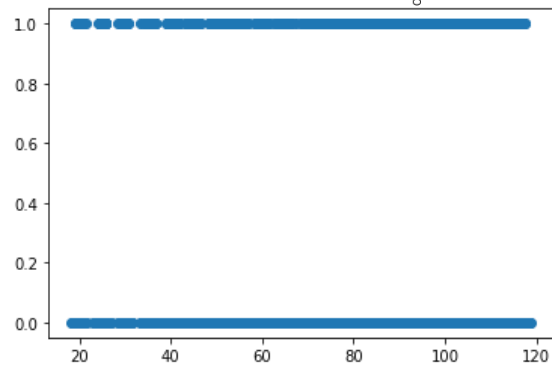
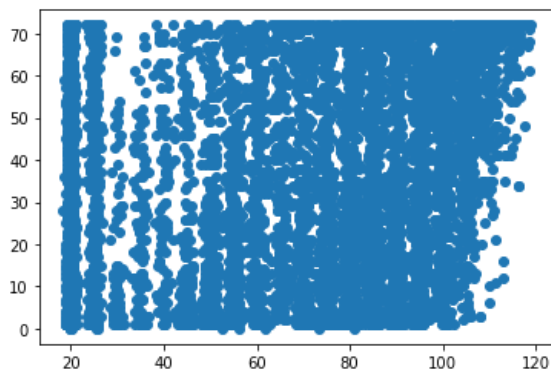
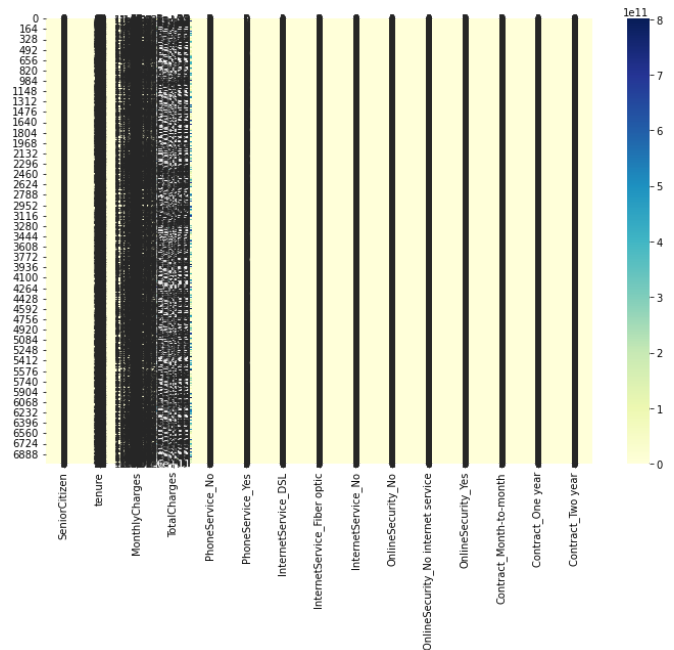
Monthly charges are increasing the total charges also increase, which shows a positive correlation.



TotalCharges vs. 'SeniorCitizen



TotalCharges vs. tenure





The figure is a density plot titled "Monthly charges by churn". The x-axis is labeled "Monthly Charges" and ranges from 0 to 140. The y-axis is labeled "Density" and ranges from 0.0000 to 0.0175. There are two data series: "No Churn" represented by a red line and a light red shaded area, and "Churn" represented by a blue line and a light blue shaded area. The "No Churn" distribution has a primary peak at approximately 20 with a density of about 0.0175, and a secondary, lower peak around 80. The "Churn" distribution has a primary peak at approximately 80 with a density of about 0.0185, and a secondary peak around 50. The two distributions overlap significantly between 40 and 100.

A density plot showing the distribution of 'Total Charges' for two groups: 'No Churn' (red line) and 'Churn' (blue line). The x-axis is labeled 'Total Charges' and ranges from 0 to 10,000. The y-axis is labeled 'Density' and ranges from 0.0000 to 0.0005. The 'Churn' distribution is highly peaked, with a maximum density of approximately 0.00048 at a total charge of about 500. The 'No Churn' distribution is broader and flatter, with a maximum density of approximately 0.00028 at a total charge of about 500. Both distributions are right-skewed, with the 'No Churn' group having a much longer tail extending towards 10,000.

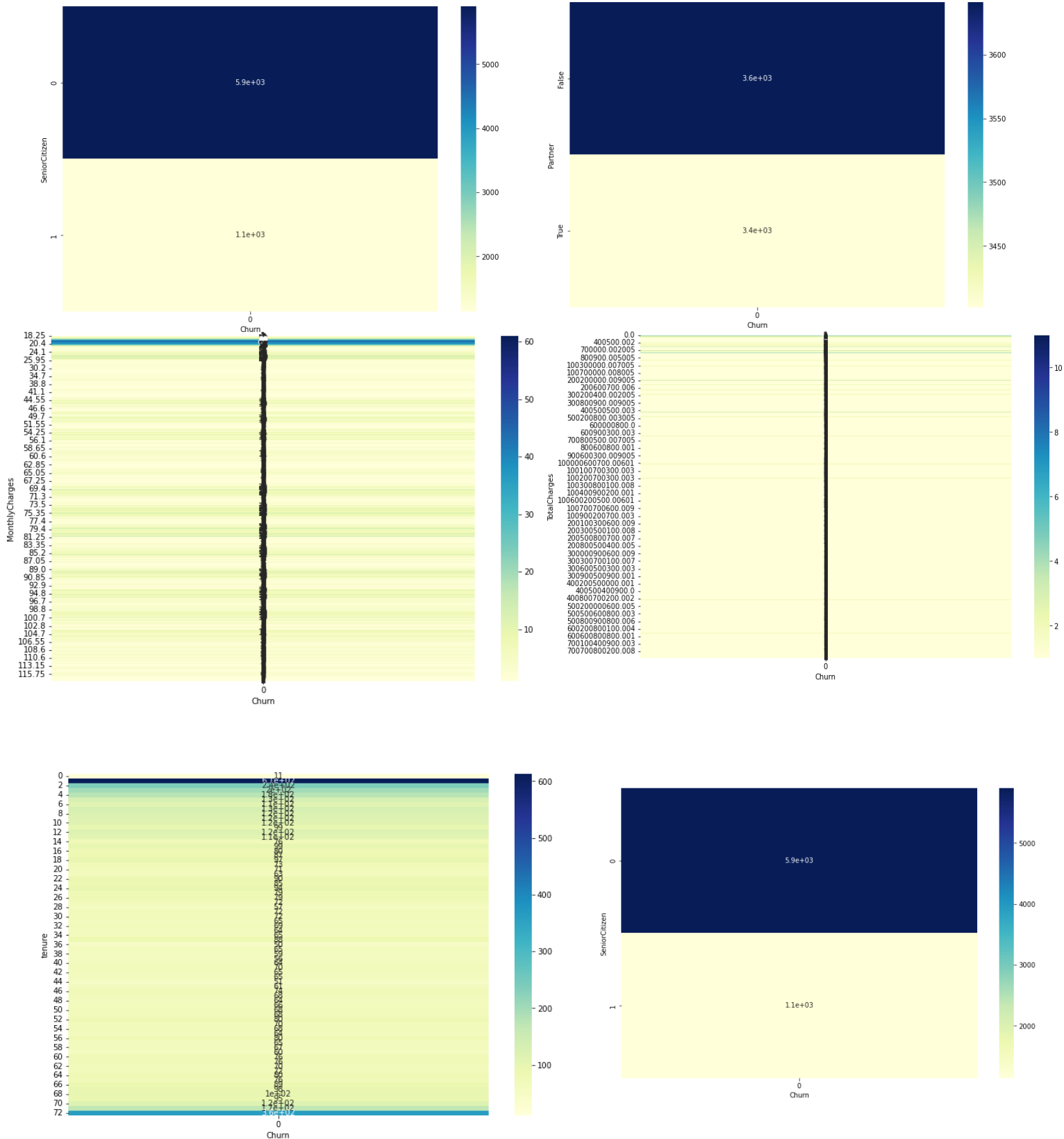
More churn is there with a lower value of total charge

[illegible]

The figure displays a 10x10 grid of heatmaps, each representing the pairwise correlation between two variables. The variables are listed on the left and top of the grid. The diagonal elements are all 1.0. The off-diagonal elements show correlation coefficients ranging from -0.75 to 0.14. The color scale ranges from -1.00 (dark blue) to 1.00 (dark red).

Variable 1 \ Variable 2	SeniorCitizen	income	MonthlyCharges	TotalCharges	Churn	gender_Female	gender_Male	PhoneService_No	PhoneService_Yes	InternetService_No	InternetService_Yes	OnlineSecurity_No	OnlineSecurity_Yes	StreamingTV_No	StreamingTV_Yes	No internet service	No internet service	Contract_Month-to-month	Contract_One year	Contract_Two year
SeniorCitizen	1	0.017	0.22	0.1	0.0086	0.0086	-0.11	0.26	-0.18	0.19	-0.18	-0.039	0.049	0.18	0.11	0.14	-0.046	-0.12		
income	0.017	1	0.25	0.8	-0.0084	-0.0084	0.013	0.02	-0.039	-0.26	-0.039	0.31	-0.65	0.2	0.56					
MonthlyCharges	0.22	0.25	1	0.64	-0.25	-0.16	0.79	-0.76	0.36	-0.76	0.3	0.017	-0.76	0.63	0.06	0.0049	-0.075			
TotalCharges	0.1	0.8	0.64	1	-0.11	0.11	-0.053	0.36	-0.37	-0.061	-0.37	0.41	-0.43	0.16	0.35					
Churn	0.0086	-0.0084	-0.25	-0.11	1	-1	0.45	-0.29	-0.17	0.058	-0.17	0.093	-0.12	0.032	0.0074	0.0028	-0.0035			
gender_Female	-0.11	0.26	-0.18	0.19	-0.18	1	-1	0.45	-0.29	-0.17	0.058	-0.17	0.093	-0.12	0.032	0.0074	0.0028	-0.0035		
gender_Male	0.26	-0.18	0.19	-0.18	0.19	-1	1	-0.45	0.29	0.17	-0.058	0.17	-0.093	0.12	-0.032	-0.0074	-0.0028	0.0035		
PhoneService_No	-0.0086	-0.0084	-0.25	-0.11	0.0086	0.0086	-0.013	0.02	-0.039	-0.26	-0.039	0.31	-0.65	0.2	0.56					
PhoneService_Yes	0.0086	0.0084	0.25	0.11	-0.0086	-0.0086	0.013	0.02	-0.039	-0.26	-0.039	0.31	-0.65	0.2	0.56					
InternetService_No	-0.18	-0.039	-0.76	-0.37	0.017	0.017	-0.38	-0.47	1	-0.52	1	-0.42	-0.22	0.038	0.22					
InternetService_Yes	0.18	0.039	0.76	0.37	-0.017	-0.017	0.38	0.47	-1	0.52	-1	0.42	-0.22	0.038	0.22					
OnlineSecurity_No	0.19	-0.26	0.36	0.061	0.058	-0.058	0.023	0.41	-0.52	1	-0.52	1	-0.63	0.4	-0.12	-0.35				
OnlineSecurity_Yes	-0.19	0.26	-0.36	-0.061	-0.058	0.058	-0.023	-0.41	0.52	-1	0.52	-1	0.63	-0.4	0.12	0.35				
StreamingTV_No	0.049	-0.25	0.017	-0.19	0.0084	-0.0084	0.023	0.013	0.02	-0.039	-0.26	-0.039	0.31	-0.65	0.2	0.56				
StreamingTV_Yes	-0.049	0.25	-0.017	0.19	-0.0084	0.0084	-0.023	-0.013	0.02	-0.039	-0.26	-0.039	0.31	-0.65	0.2	0.56				
No internet service	0.18	-0.039	-0.76	-0.37	0.017	0.017	-0.38	-0.47	1	-0.52	1	-0.42	-0.22	0.038	0.22					
No internet service	-0.18	0.039	0.76	0.37	-0.017	-0.017	0.38	0.47	-1	0.52	-1	0.42	-0.22	0.038	0.22					
Contract_Month-to-month	0.14	-0.05	0.06	-0.43	0.0074	-0.0074	0.006	0.28	0.22	0.4	0.22	-0.25	1	-0.57	-0.62					
Contract_One year	-0.046																			

## Chi-square tests:





## Observations based on plots and coded data:

- Gender distribution is relatively even.
- Roughly 2000 customers churned, and 5000 customers retained.
- Senior citizens are three times less likely to churn
- Partners are almost two times less likely to leave
- Customers without dependents are four times more likely to churn
- The dataset is imbalanced, with the majority of customers being active.
- There is multicollinearity between Monthly Charges and Total Charges.
- Most of the customers in the dataset are younger people.
- There are lots of customers with long and short tenures.
- Monthly charges have a lot of low charges and, aside from that, are reasonably normally distributed.
- The majority of customers that cancel their subscription have Phone Service-enabled.
- Customers with Internet Service as *Fiber-Optic* are more likely to cancel than those who have *DSL*
- Customers that do not have Online Security, Device Protection, Online Backup, and Tech Support services enabled are more likely to leave
- The strongest positive correlation with the target features is Monthly Charges and Age, while the negative correlation is with Partners, Dependents, and Tenure.
- There are a lot of new customers in the organization, followed by a loyal customer base above 70 months old.
- Customers with a month-to-month connection have a high probability of churning if they have subscribed to pay via electronic checks.