Statistical Data Analysis

After we completed data wrangling and data visualization, it is time to use statistical inference tool to explore the relationship between independent variables and dependent variable. For this particular project, we are going to perform survival analysis to predict telecom customers churn. In this context, “tenure” and “churn” will be treated as time-event variables. During the process of data wrangling, we found this dataset has 16 categorical variables and 3 numerical variables that can be used to build model.

First of all, we divide all observations into two groups using ‘churn’, i.e. churned and non-churned. We are interested to know how the numerical variables’ mean differ between these two subgroups. How would we explain them statistically? The result of t-test indicates that statistically significance does exist since p-values are all extremely low.

```
the mean of MonthlyCharge for not churned:  61.2651236953999
the mean of MonthlyCharge for churned:  74.4413322632423
Ttest Result(statistic=-16.53673801593631, pvalue=2.706645606888261e-60)
```

```
the mean of TotalCharges for not churned:  2549.911441824514
the mean of TotalCharges for churned:  1531.7960941680035
Ttest Result(statistic=16.978779727124437, pvalue=2.127211613240394e-63)
```

```
the mean of tenure for not churned:  37.56996521066873
the mean of tenure for churned:  17.979133226324237
Ttest_indResult(statistic=31.57955051135377, pvalue=7.99905796059022e-205)
```

We can confidently reject the null hypothesis and conclude that all these numerical variables differ between two groups. Please also be noted that these three variables were considered individually. In survival analysis, we should treat variable 'tenure' as time factor that may bring censoring issue. This will be addressed in model building and model performance evaluation that follows.

Since the majority of the features are categorical, after we had an idea about cardinality, we are more interested in how each labeled group affect the performance of time-event in the context of survival analysis. Rather than using traditional chi-square contingence table, we use Log rank test together with

Kaplan-Meier curve to figure out how those groups differ from each other. The analysis indicates that out the 16 categorical variables, only "gender" , 'Phone Service" and ‘MultipleLines” have large value of p-values. We don't have sufficient evidence to reject null hypothesis for them. The statics of other variables is statistically significant. They are more important to predict the customer churn.