Chi-squared test

# A. Business Question

### Are paperless billing customers more or less prone to churn?

As market competition increases increases in the telecommunications industry with so many technological advances spurring the arrival of new players in streaming and information sharing, organizations must find ways to operate at finer margins and cut costs wherever possible.  
Two ways to achieve this are by retaining existing customers and to decrease the costs associated with traditional paper billing by postal mail. The costs to on-board a new customers are already identified as 10 times that of retention of an existing customer. With paperless billed customers, not only is the cost of sending paper bills eliminated, but paperless options also decrease the use of natural resources and therefore are much more eco friendly. It therefore would be valuable to an organization and its stakeholders to identify whether these highly values paperless customers are more prone to churn. Once this information is known, new strategic initiatives can be developed as needed to target the retention of paperless customers. The data set used for this analysis will be the churn\_clean.csv file with special focus on the Churn and PaperlessBilling variables

R will be used for this analysis. R is open source software that was specifically made for statistical analysis. Using R, we can ingest the raw data set, and leveraging an extensive library of data manipulation and visualization packages, clean and investigate the data. More information can be found on the R project website (<https://www.r-project.org/>).

# B. Chi Squared Test

Using a chi-squared test we can test the null hypothesis which in this case would state that churn results are normally or evenly distributed among customers with and without paperless billing.

df<-read.csv("c:/users/shua/documents/exploratory data analysis\_d207/churn\_clean.csv")  
tbl=table(df$Churn, df$PaperlessBilling)  
tbl

##   
## No Yes  
## No 3042 4308  
## Yes 1076 1574

chisq.test(tbl)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: tbl  
## X-squared = 0.4624, df = 1, p-value = 0.4965

The data showed the following frequencies:

* 3,042 occurrences of non-paperless billing resulting in no churn
* 1,076 occurrences of non-paperless billing resulting in churn
* 4,308 paperless billing resulting in no churn
* 1,574 paperless billing resulting in churn
* A chi square test was chosen as the variables are categorical, meaning they are “Yes”, “No” character values instead of numerical.
* The resulting p-value of the chi square test was .4965, meaning the null hypothesis that churn results are independent of paperless billing selection cannot be rejected. Practically speaking for the organization, it may be useful to target attraction of paperless billing customers for the material cost and environmental benefit, but not for reducing churn.

# C. Univariate Distribution

### Continuous Variables

* Tenure
* Income

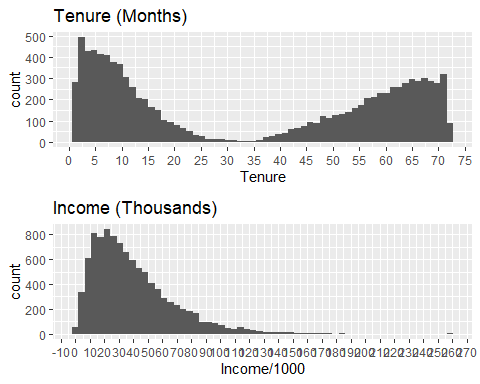
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.0.5

library(gridExtra)  
str(df)

## 'data.frame': 10000 obs. of 50 variables:  
## $ CaseOrder : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Customer\_id : chr "K409198" "S120509" "K191035" "D90850" ...  
## $ Interaction : chr "aa90260b-4141-4a24-8e36-b04ce1f4f77b" "fb76459f-c047-4a9d-8af9-e0f7d4ac2524" "344d114c-3736-4be5-98f7-c72c281e2d35" "abfa2b40-2d43-4994-b15a-989b8c79e311" ...  
## $ UID : chr "e885b299883d4f9fb18e39c75155d990" "f2de8bef964785f41a2959829830fb8a" "f1784cfa9f6d92ae816197eb175d3c71" "dc8a365077241bb5cd5ccd305136b05e" ...  
## $ City : chr "Point Baker" "West Branch" "Yamhill" "Del Mar" ...  
## $ State : chr "AK" "MI" "OR" "CA" ...  
## $ County : chr "Prince of Wales-Hyder" "Ogemaw" "Yamhill" "San Diego" ...  
## $ Zip : int 99927 48661 97148 92014 77461 31030 37847 73109 34771 45237 ...  
## $ Lat : num 56.3 44.3 45.4 33 29.4 ...  
## $ Lng : num -133.4 -84.2 -123.2 -117.2 -95.8 ...  
## $ Population : int 38 10446 3735 13863 11352 17701 2535 23144 17351 20193 ...  
## $ Area : chr "Urban" "Urban" "Urban" "Suburban" ...  
## $ TimeZone : chr "America/Sitka" "America/Detroit" "America/Los\_Angeles" "America/Los\_Angeles" ...  
## $ Job : chr "Environmental health practitioner" "Programmer, multimedia" "Chief Financial Officer" "Solicitor" ...  
## $ Children : int 0 1 4 1 0 3 0 2 2 1 ...  
## $ Age : int 68 27 50 48 83 83 79 30 49 86 ...  
## $ Income : num 28562 21705 9610 18925 40074 ...  
## $ Marital : chr "Widowed" "Married" "Widowed" "Married" ...  
## $ Gender : chr "Male" "Female" "Female" "Male" ...  
## $ Churn : chr "No" "Yes" "No" "No" ...  
## $ Outage\_sec\_perweek : num 7.98 11.7 10.75 14.91 8.15 ...  
## $ Email : int 10 12 9 15 16 15 10 16 20 18 ...  
## $ Contacts : int 0 0 0 2 2 3 0 0 2 1 ...  
## $ Yearly\_equip\_failure: int 1 1 1 0 1 1 1 0 3 0 ...  
## $ Techie : chr "No" "Yes" "Yes" "Yes" ...  
## $ Contract : chr "One year" "Month-to-month" "Two Year" "Two Year" ...  
## $ Port\_modem : chr "Yes" "No" "Yes" "No" ...  
## $ Tablet : chr "Yes" "Yes" "No" "No" ...  
## $ InternetService : chr "Fiber Optic" "Fiber Optic" "DSL" "DSL" ...  
## $ Phone : chr "Yes" "Yes" "Yes" "Yes" ...  
## $ Multiple : chr "No" "Yes" "Yes" "No" ...  
## $ OnlineSecurity : chr "Yes" "Yes" "No" "Yes" ...  
## $ OnlineBackup : chr "Yes" "No" "No" "No" ...  
## $ DeviceProtection : chr "No" "No" "No" "No" ...  
## $ TechSupport : chr "No" "No" "No" "No" ...  
## $ StreamingTV : chr "No" "Yes" "No" "Yes" ...  
## $ StreamingMovies : chr "Yes" "Yes" "Yes" "No" ...  
## $ PaperlessBilling : chr "Yes" "Yes" "Yes" "Yes" ...  
## $ PaymentMethod : chr "Credit Card (automatic)" "Bank Transfer(automatic)" "Credit Card (automatic)" "Mailed Check" ...  
## $ Tenure : num 6.8 1.16 15.75 17.09 1.67 ...  
## $ MonthlyCharge : num 172 243 160 120 150 ...  
## $ Bandwidth\_GB\_Year : num 905 801 2055 2165 271 ...  
## $ Item1 : int 5 3 4 4 4 3 6 2 5 2 ...  
## $ Item2 : int 5 4 4 4 4 3 5 2 4 2 ...  
## $ Item3 : int 5 3 2 4 4 3 6 2 4 2 ...  
## $ Item4 : int 3 3 4 2 3 2 4 5 3 2 ...  
## $ Item5 : int 4 4 4 5 4 4 1 2 4 5 ...  
## $ Item6 : int 4 3 3 4 4 3 5 3 3 2 ...  
## $ Item7 : int 3 4 3 3 4 3 5 4 4 3 ...  
## $ Item8 : int 4 4 3 3 5 3 5 5 4 3 ...

dist1<-ggplot(df, aes(x=Tenure))+  
 geom\_histogram(bins=60)+  
 scale\_x\_continuous(n.breaks=15)+  
 ggtitle("Tenure (Months)")  
dist2<-ggplot(df, aes(x=Income/1000))+  
 geom\_histogram(bins=60)+  
 scale\_x\_continuous(n.breaks=30)+  
 ggtitle("Income (Thousands)")  
grid.arrange(dist1, dist2)

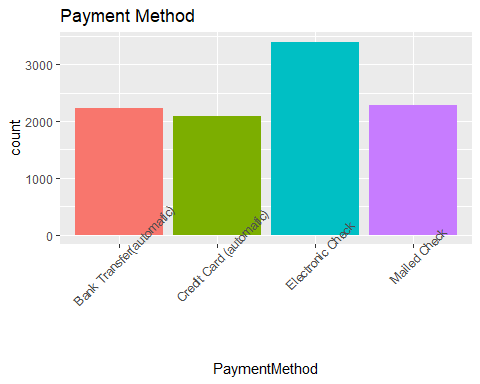


The library “ggplot2” is used which is useful for plotting data within R. More information on the library ggplot2 can be found on the tidyverse website (<https://ggplot2.tidyverse.org/>) When plotting Tenure on a histogram, interestingly the distribution is bimodial. The break between the modes falls between 30-35 months. When plotting Income (in thousands for readability) the distribution has a right skew with a long right tail, and the high point mode around 25 thousand.

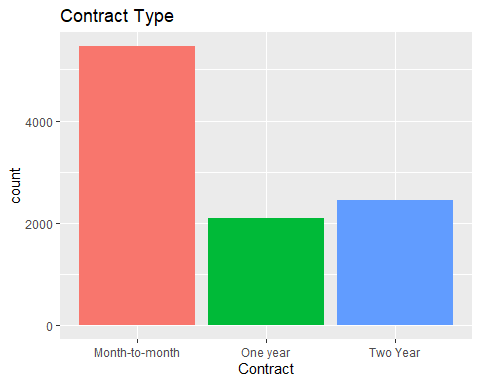
### Categorical Variables

* Payment Method
* Contract Type

bar1<-ggplot(df, aes(x=PaymentMethod, fill=PaymentMethod))+  
 geom\_bar()+  
 ggtitle("Payment Method")+  
 theme(axis.text.x = element\_text(angle=45))+  
 theme(legend.position = "none")  
bar2<-ggplot(df, aes(x=Contract, fill=Contract))+  
 geom\_bar()+  
 ggtitle("Contract Type")+  
 theme(legend.position = "none")  
bar1



bar2

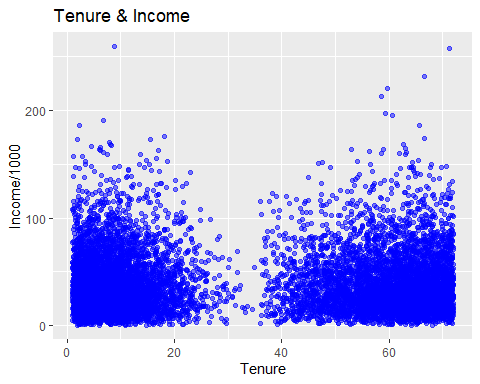


# D. Bivariate Distribution

### Continuous Variables

* Tenure
* Income

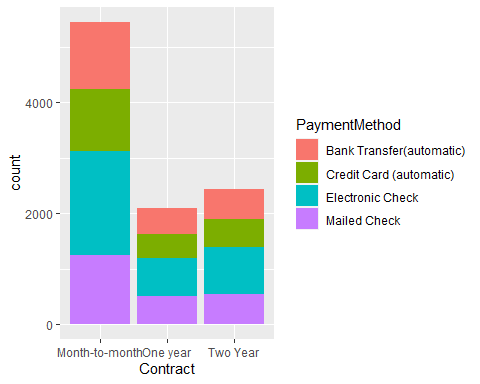
ggplot(df, aes(x=Tenure, y=Income/1000))+  
 geom\_point(position="jitter", color="blue", alpha=.5)+  
 ggtitle("Tenure & Income")



## Categorical Variables

* Payment Method
* Contract Type

ggplot(df, aes(x=Contract, fill=PaymentMethod))+  
 geom\_bar()



# E. Summary

Using a chi square test, we achieved a p-value result of .4965 which means that we cannot reject the null hypothesis. In this case, this means that we cannot state churn results are dependent on paperless billing selection. A p-value of .09 or below would have been needed to reject the null hypothesis with 90% confidence. This was not achieved using a chi square test.

This analysis was limited in that it did not have the ability to further explore the relationship between paperless billing and churn. There were several customer response questions in the data relating to their experience with customer service representatives. It would be helpful if the data set was expanded to include customer response questions which probed the customer’s comfort level with paperless billing in general such as prior use of paperless billing even with other company’s products as well as the importance of their providers commitment to eco friendly solutions.

If the company is going to be committed to eco friendly solutions as well as cutting billing costs, the next course of action should be to do additional analysis to uncover any relationships with paperless billing selections and other variables such as age, education level, or contract type. Once those relationships are known, that information can be further investigated to explore churn relationships. THe conclusion of these steps would be to identify highly valued customer relationships and to use this information to inform organizational strategic initiatives.