02\_Data\_Visualization

Matt Dube

12/4/2018

Table of Contents

Load libraries and set theme

library(ggplot2)  
library(dplyr)  
library(ggthemes)  
library(here)  
library(readr)  
library(purrr)  
library(tidyr)  
library(vcd)  
library(corrplot)

theme\_set(theme\_light())

Load data

customer <- read\_csv(here("00\_Data/raw", "WA\_Fn-UseC\_-Telco-Customer-Churn.csv"))

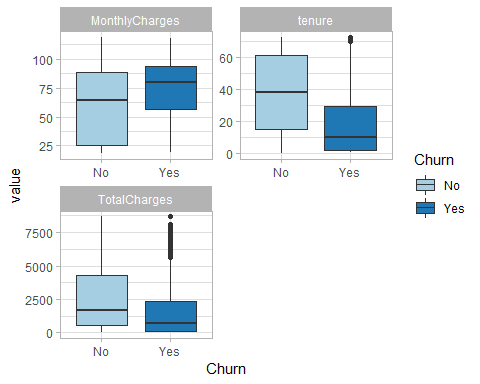
Functions to assist with plotting

plotNumData <- function(varList, inputData, numCols=2) {  
 gather(data=inputData, varList, key = "var", value = "value") %>%  
 ggplot(aes(x=Churn, y=value, fill=Churn)) +  
 geom\_boxplot() +  
 facet\_wrap(~ var, scales = "free", ncol = numCols) +  
 scale\_fill\_brewer(palette = "Paired")  
}  
  
plotCharData <- function(varList, inputData, numCols=2) {  
 gather(data=inputData, varList, key = "var", value = "value") %>%  
 ggplot(aes(x=value, fill=Churn)) +  
 geom\_bar() +  
 coord\_flip() +  
 facet\_wrap(~ var, scales="free", ncol = numCols) +  
 scale\_fill\_brewer(palette = "Paired")  
}  
  
plotHistograms <- function(varList, inputData, numCols=2) {  
 gather(data=inputData, varList, key = "var", value = "value") %>%  
 ggplot(aes(x=value, fill=Churn)) +  
 geom\_histogram() +  
 facet\_wrap(~ var, scales = "free", ncol = numCols) +  
 scale\_fill\_brewer(palette = "Paired")  
}

Boxplots of numeric features

numVarToPlot <- c("tenure", "MonthlyCharges", "TotalCharges")  
plotNumData(numVarToPlot, customer)

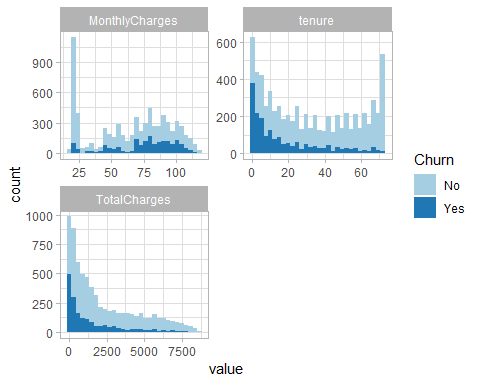
## Warning: Removed 11 rows containing non-finite values (stat\_boxplot).

 Histograms of numeric features

plotHistograms(numVarToPlot, customer)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 11 rows containing non-finite values (stat\_bin).

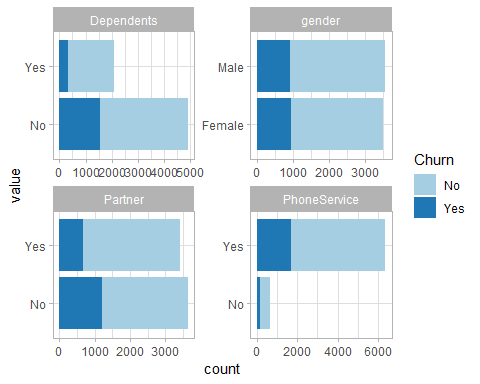


Notes on numeric features:

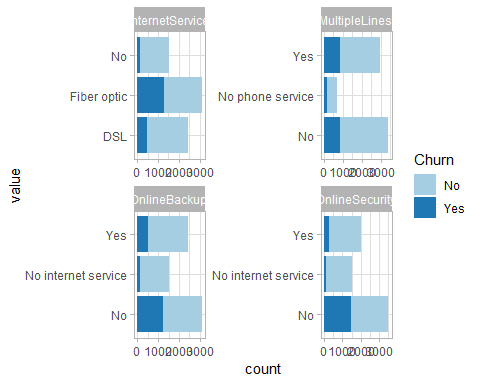
<<<<<<< HEAD \* customers with smaller tenures and higher monthly charges are more likely to churn. \* there are a few long tenured customers who have churned, and appear as outliers in the boxplot. \* total charges has a few potential outliers as well that need to be examined.

Bar plots of character features

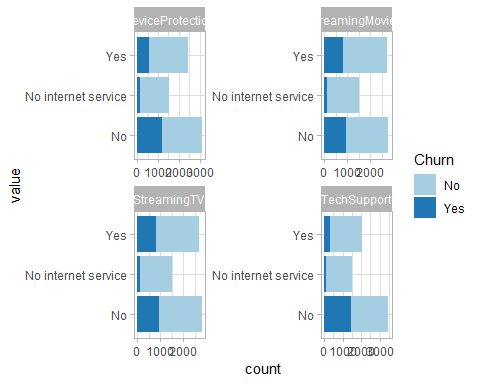
varToPlot <- c("gender", "Partner", "Dependents", "PhoneService")  
plotCharData(varToPlot, customer)



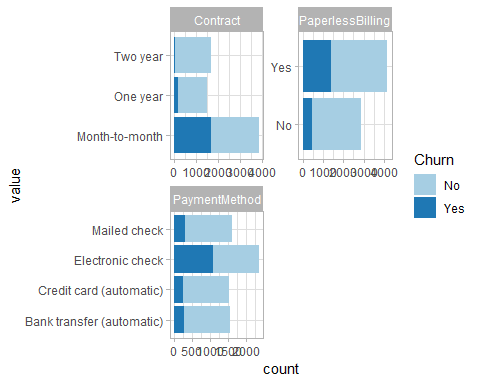
varToPlot <- c("MultipleLines", "InternetService", "OnlineSecurity", "OnlineBackup")  
plotCharData(varToPlot, customer)



varToPlot <- c("DeviceProtection", "TechSupport", "StreamingTV", "StreamingMovies")  
plotCharData(varToPlot, customer)



varToPlot <- c("Contract", "PaperlessBilling", "PaymentMethod")  
plotCharData(varToPlot, customer)



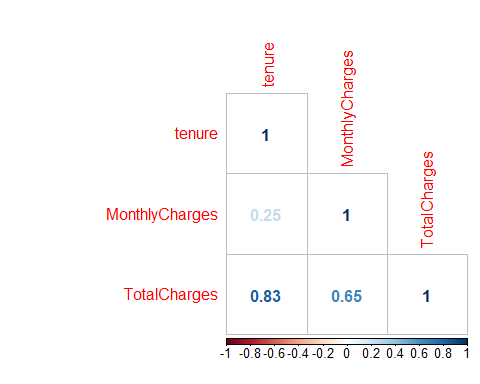
* customers will smaller tenures and higher monthly charges are more likely to churn.
* there are a few long tenured customers who have churned, and appear as outliers in the boxplot.
* total charges has a few potential outliers as well that need to be examined.

Review correlation between tenure, TotalCharges and MonthlyCharges.

cust\_numeric <-   
 customer %>%  
 select(-SeniorCitizen) %>%  
 keep(is.numeric)

M <- cor(cust\_numeric,use="complete.obs")

corrplot(M, method="number", type="lower")



Tenure and TotalCharges have a strong correlation. MonthlyCharges and TotalCharges are also strongly correlated. When we get to data cleaning TotalCharges should be removed - or at the minimum not used as a predictor in our models.