Prep and EDA\_v2

Ata, Catherine - Cortez, Angela - Foley, Amanda

5/11/2021

# PHASE 2: Data Preparation  
  
# Extract Data  
aug\_train\_complete <- read.csv('aug\_train.csv', na.strings=c("", "NA"))  
  
# Learning about the Data  
str(aug\_train\_complete)

## 'data.frame': 19158 obs. of 14 variables:  
## $ enrollee\_id : int 8949 29725 11561 33241 666 21651 28806 402 27107 699 ...  
## $ city : chr "city\_103" "city\_40" "city\_21" "city\_115" ...  
## $ city\_development\_index: num 0.92 0.776 0.624 0.789 0.767 0.764 0.92 0.762 0.92 0.92 ...  
## $ gender : chr "Male" "Male" NA NA ...  
## $ relevent\_experience : chr "Has relevent experience" "No relevent experience" "No relevent experience" "No relevent experience" ...  
## $ enrolled\_university : chr "no\_enrollment" "no\_enrollment" "Full time course" NA ...  
## $ education\_level : chr "Graduate" "Graduate" "Graduate" "Graduate" ...  
## $ major\_discipline : chr "STEM" "STEM" "STEM" "Business Degree" ...  
## $ experience : chr ">20" "15" "5" "<1" ...  
## $ company\_size : chr NA "50-99" NA NA ...  
## $ company\_type : chr NA "Pvt Ltd" NA "Pvt Ltd" ...  
## $ last\_new\_job : chr "1" ">4" "never" "never" ...  
## $ training\_hours : int 36 47 83 52 8 24 24 18 46 123 ...  
## $ target : int 1 0 0 1 0 1 0 1 1 0 ...

summary(aug\_train\_complete)

## enrollee\_id city city\_development\_index gender   
## Min. : 1 Length:19158 Min. :0.4480 Length:19158   
## 1st Qu.: 8554 Class :character 1st Qu.:0.7400 Class :character   
## Median :16983 Mode :character Median :0.9030 Mode :character   
## Mean :16875 Mean :0.8288   
## 3rd Qu.:25170 3rd Qu.:0.9200   
## Max. :33380 Max. :0.9490   
## relevent\_experience enrolled\_university education\_level major\_discipline   
## Length:19158 Length:19158 Length:19158 Length:19158   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## experience company\_size company\_type last\_new\_job   
## Length:19158 Length:19158 Length:19158 Length:19158   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## training\_hours target   
## Min. : 1.00 Min. :0.0000   
## 1st Qu.: 23.00 1st Qu.:0.0000   
## Median : 47.00 Median :0.0000   
## Mean : 65.37 Mean :0.2493   
## 3rd Qu.: 88.00 3rd Qu.:0.0000   
## Max. :336.00 Max. :1.0000

head(aug\_train\_complete)

## enrollee\_id city city\_development\_index gender relevent\_experience  
## 1 8949 city\_103 0.920 Male Has relevent experience  
## 2 29725 city\_40 0.776 Male No relevent experience  
## 3 11561 city\_21 0.624 <NA> No relevent experience  
## 4 33241 city\_115 0.789 <NA> No relevent experience  
## 5 666 city\_162 0.767 Male Has relevent experience  
## 6 21651 city\_176 0.764 <NA> Has relevent experience  
## enrolled\_university education\_level major\_discipline experience company\_size  
## 1 no\_enrollment Graduate STEM >20 <NA>  
## 2 no\_enrollment Graduate STEM 15 50-99  
## 3 Full time course Graduate STEM 5 <NA>  
## 4 <NA> Graduate Business Degree <1 <NA>  
## 5 no\_enrollment Masters STEM >20 50-99  
## 6 Part time course Graduate STEM 11 <NA>  
## company\_type last\_new\_job training\_hours target  
## 1 <NA> 1 36 1  
## 2 Pvt Ltd >4 47 0  
## 3 <NA> never 83 0  
## 4 Pvt Ltd never 52 1  
## 5 Funded Startup 4 8 0  
## 6 <NA> 1 24 1

# remove unnecessary variables: employee\_id and city   
aug\_train = subset(aug\_train\_complete, select = -c(1,2))  
  
head(aug\_train)

## city\_development\_index gender relevent\_experience enrolled\_university  
## 1 0.920 Male Has relevent experience no\_enrollment  
## 2 0.776 Male No relevent experience no\_enrollment  
## 3 0.624 <NA> No relevent experience Full time course  
## 4 0.789 <NA> No relevent experience <NA>  
## 5 0.767 Male Has relevent experience no\_enrollment  
## 6 0.764 <NA> Has relevent experience Part time course  
## education\_level major\_discipline experience company\_size company\_type  
## 1 Graduate STEM >20 <NA> <NA>  
## 2 Graduate STEM 15 50-99 Pvt Ltd  
## 3 Graduate STEM 5 <NA> <NA>  
## 4 Graduate Business Degree <1 <NA> Pvt Ltd  
## 5 Masters STEM >20 50-99 Funded Startup  
## 6 Graduate STEM 11 <NA> <NA>  
## last\_new\_job training\_hours target  
## 1 1 36 1  
## 2 >4 47 0  
## 3 never 83 0  
## 4 never 52 1  
## 5 4 8 0  
## 6 1 24 1

summary(aug\_train)

## city\_development\_index gender relevent\_experience  
## Min. :0.4480 Length:19158 Length:19158   
## 1st Qu.:0.7400 Class :character Class :character   
## Median :0.9030 Mode :character Mode :character   
## Mean :0.8288   
## 3rd Qu.:0.9200   
## Max. :0.9490   
## enrolled\_university education\_level major\_discipline experience   
## Length:19158 Length:19158 Length:19158 Length:19158   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## company\_size company\_type last\_new\_job training\_hours   
## Length:19158 Length:19158 Length:19158 Min. : 1.00   
## Class :character Class :character Class :character 1st Qu.: 23.00   
## Mode :character Mode :character Mode :character Median : 47.00   
## Mean : 65.37   
## 3rd Qu.: 88.00   
## Max. :336.00   
## target   
## Min. :0.0000   
## 1st Qu.:0.0000   
## Median :0.0000   
## Mean :0.2493   
## 3rd Qu.:0.0000   
## Max. :1.0000

str(aug\_train)

## 'data.frame': 19158 obs. of 12 variables:  
## $ city\_development\_index: num 0.92 0.776 0.624 0.789 0.767 0.764 0.92 0.762 0.92 0.92 ...  
## $ gender : chr "Male" "Male" NA NA ...  
## $ relevent\_experience : chr "Has relevent experience" "No relevent experience" "No relevent experience" "No relevent experience" ...  
## $ enrolled\_university : chr "no\_enrollment" "no\_enrollment" "Full time course" NA ...  
## $ education\_level : chr "Graduate" "Graduate" "Graduate" "Graduate" ...  
## $ major\_discipline : chr "STEM" "STEM" "STEM" "Business Degree" ...  
## $ experience : chr ">20" "15" "5" "<1" ...  
## $ company\_size : chr NA "50-99" NA NA ...  
## $ company\_type : chr NA "Pvt Ltd" NA "Pvt Ltd" ...  
## $ last\_new\_job : chr "1" ">4" "never" "never" ...  
## $ training\_hours : int 36 47 83 52 8 24 24 18 46 123 ...  
## $ target : int 1 0 0 1 0 1 0 1 1 0 ...

# Check continuous variables  
summary(aug\_train$city\_development\_index)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.4480 0.7400 0.9030 0.8288 0.9200 0.9490

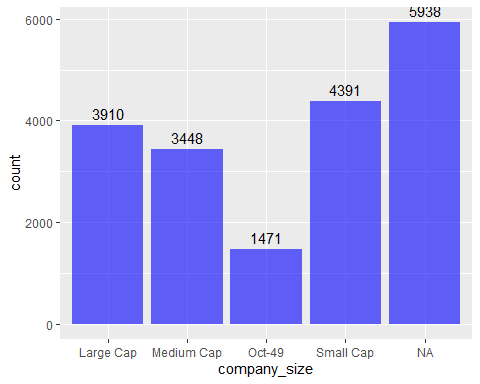
summary(aug\_train$training\_hours)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 23.00 47.00 65.37 88.00 336.00

# Data Conditioning  
  
# remove duplicates  
unique(aug\_train$company\_size)

## [1] NA "50-99" "<10" "10000+" "5000-9999" "1000-4999"  
## [7] "Oct-49" "100-500" "500-999"

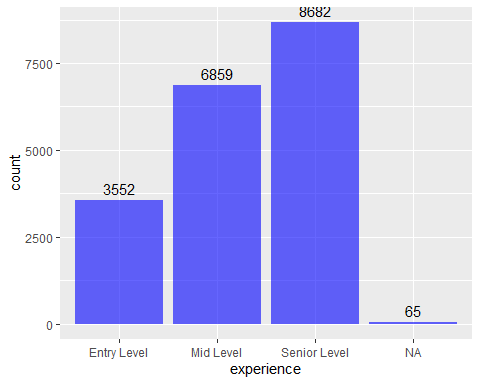
# Transform company\_size   
small\_cap = c('<10', '10/49','50-99')  
medium\_cap = c('100-500','500-999')  
large\_cap = c('1000-4999','5000-9999','10000+')  
all\_size = c(small\_cap,medium\_cap,large\_cap)  
  
# Assign elements in company\_size to correct cap  
for (i in all\_size){  
 if (i %in% small\_cap){  
 aug\_train$company\_size[aug\_train$company\_size == i] = "Small Cap"}  
 else if (i %in% medium\_cap){  
 aug\_train$company\_size[aug\_train$company\_size == i] = "Medium Cap"}  
 else if (i %in% large\_cap){  
 aug\_train$company\_size[aug\_train$company\_size == i] = "Large Cap"}}  
  
# Plot transformed variable  
ggplot(aug\_train, aes(company\_size)) + geom\_bar(fill = 'blue',alpha = 0.6) +  
 geom\_text( stat='count', aes(label=..count..), vjust=-0.5)



# remove duplicates  
unique(aug\_train$experience)

## [1] ">20" "15" "5" "<1" "11" "13" "7" "17" "2" "16" "1" "4"   
## [13] "10" "14" "18" "19" "12" "3" "6" "9" "8" "20" NA

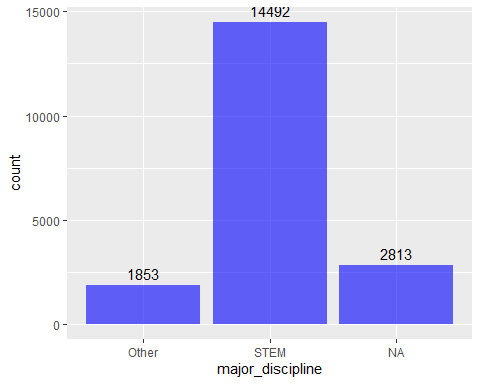
# Transform experience variable  
entry\_level = c('<1','1','2','3') # year of experience <1 - 3  
mid\_level = c('4','5','6','7','8','9') # year of experience 4-9  
senior\_level = c('10','11','12','13','14','15','16','17','18','19','20','>20') # year of experience 10 to >20  
all\_level = c(entry\_level,mid\_level,senior\_level) # combine all new variables   
# Assign elements in experience in correct level  
for (i in all\_level){  
 if (i %in% entry\_level){  
 aug\_train$experience[aug\_train$experience == i] = "Entry Level"}  
 else if (i %in% mid\_level){  
 aug\_train$experience[aug\_train$experience == i] = "Mid Level"}  
 else if (i %in% senior\_level){  
 aug\_train$experience[aug\_train$experience == i] = "Senior Level"}}  
# Plot transformed variable  
ggplot(aug\_train, aes(experience)) + geom\_bar(fill = 'blue',alpha = 0.6) +  
 geom\_text( stat='count', aes(label=..count..), vjust=-0.5)



# remove duplicates  
unique(aug\_train$major\_discipline)

## [1] "STEM" "Business Degree" NA "Arts"   
## [5] "Humanities" "No Major" "Other"

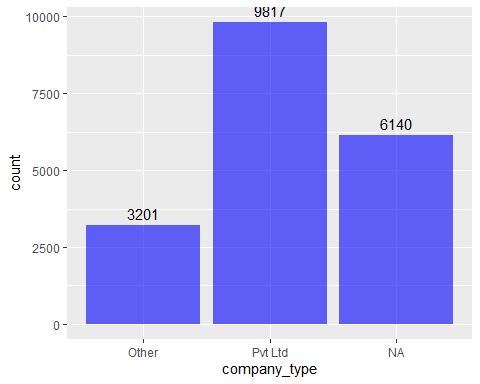
# Transform major\_discipline and Assign elements in major\_discipline to other\_major and label as "Other"  
other\_major = c('Arts','Business Degree','Humanities','No Major')  
for (i in other\_major){  
 if (i %in% other\_major){  
 aug\_train$major\_discipline[aug\_train$major\_discipline == i] = "Other"}}  
# Plot transformed variable  
ggplot(aug\_train, aes(major\_discipline)) + geom\_bar(fill = 'blue',alpha = 0.6) +  
 geom\_text( stat='count', aes(label=..count..), vjust=-0.5)



# Remove duplicates  
unique(aug\_train$company\_type)

## [1] NA "Pvt Ltd" "Funded Startup"   
## [4] "Early Stage Startup" "Other" "Public Sector"   
## [7] "NGO"

# Transform company\_type by Assigning elements to other\_company\_type then label as Other  
other\_company\_type = c("Early Stage Startup","Funded Startup","NGO","Public Sector")  
for (i in other\_company\_type){  
 aug\_train$company\_type[aug\_train$company\_type == i] = "Other"}  
# PLot transformed company\_type  
ggplot(aug\_train, aes(company\_type)) + geom\_bar(fill = 'blue',alpha = 0.6) +  
 geom\_text( stat='count', aes(label=..count..), vjust=-0.5)



# Remove duplicates and print out unique values  
unique(aug\_train$target)

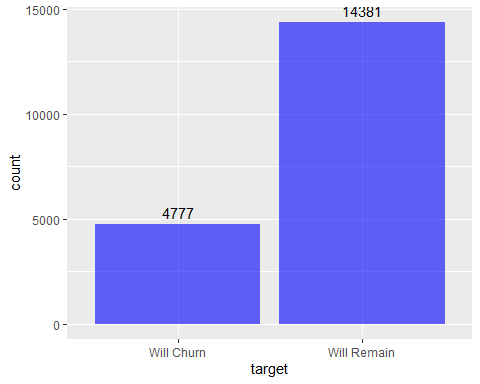
## [1] 1 0

# Provide appropriate name for target labels: 1 as positive label  
aug\_train$target[aug\_train$target == "1"] = "Will Churn"  
aug\_train$target[aug\_train$target == "0"] = "Will Remain"  
# Show new unique target labels: "Will Churn" as positive label  
unique(aug\_train$target)

## [1] "Will Churn" "Will Remain"

# Convert character to factor  
aug\_train$target = as.factor(aug\_train$target)

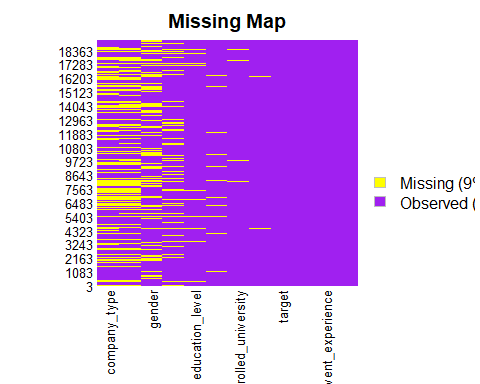
# PHASE 3: Model Planning  
# EDA and Data Visualization  
  
# plot distribution of target variable: Higher number of "Will Remain"  
ggplot(aug\_train, aes(target)) + geom\_bar(fill = 'blue',alpha = 0.6) +  
 geom\_text( stat='count', aes(label=..count..), vjust=-0.5)



# plot missing data: 9% are missing data  
missing = is.na(aug\_train)  
sum(missing)

## [1] 20733

missmap(aug\_train, main = 'Missing Map', col = c('yellow', 'purple'))



# Fill in missing data for gender with "Other"  
aug\_train$gender = as.character(aug\_train$gender)  
aug\_train$gender = replace\_na(aug\_train$gender,'Other')  
  
# Check for NAs  
sum(is.na(aug\_train$gender))

## [1] 0

unique(aug\_train$gender)

## [1] "Male" "Other" "Female"

# Fill in missing data for enrolled\_university with "Unknown"  
aug\_train$enrolled\_university = replace\_na(aug\_train$enrolled\_university,'Unknown')  
  
# Check for NAs  
sum(is.na(aug\_train$enrolled\_university))

## [1] 0

unique(aug\_train$enrolled\_university)

## [1] "no\_enrollment" "Full time course" "Unknown" "Part time course"

# Fill in missing data for education\_level with "Unknown"  
aug\_train$education\_level = replace\_na(aug\_train$education\_level,'Unknown')  
sum(is.na(aug\_train$education\_level))

## [1] 0

unique(aug\_train$education\_level)

## [1] "Graduate" "Masters" "High School" "Unknown"   
## [5] "Phd" "Primary School"

# Fill in missing data for company\_size with "Unknown"  
aug\_train$company\_size = replace\_na(aug\_train$company\_size,'Unknown')  
  
# Check for NAs  
sum(is.na(aug\_train$company\_size))

## [1] 0

unique(aug\_train$company\_size)

## [1] "Unknown" "Small Cap" "Large Cap" "Oct-49" "Medium Cap"

# Fill in missing data for major\_discipline with "Other"  
aug\_train$major\_discipline = replace\_na(aug\_train$major\_discipline ,'Other')  
  
# Check for NAs  
sum(is.na(aug\_train$major\_discipline ))

## [1] 0

unique(aug\_train$major\_discipline )

## [1] "STEM" "Other"

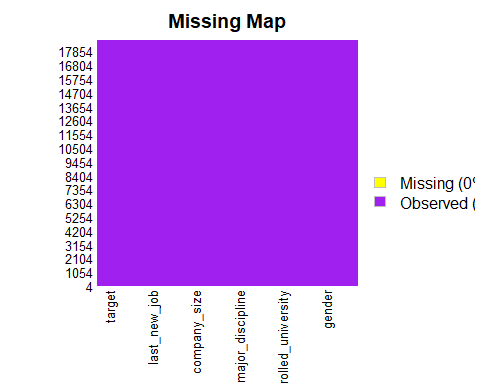
# Fill in missing data for company\_type with "Other"  
aug\_train$company\_type = replace\_na(aug\_train$company\_type ,'Other')  
  
# Check for NAs  
sum(is.na(aug\_train$company\_type ))

## [1] 0

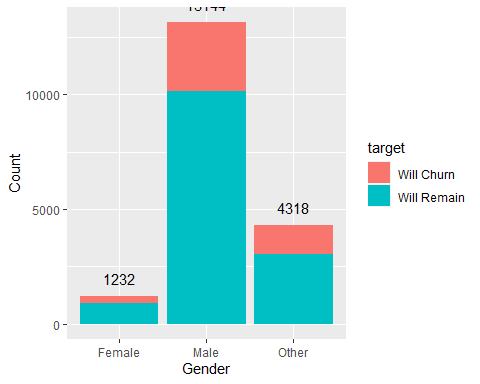
unique(aug\_train$company\_type )

## [1] "Other" "Pvt Ltd"

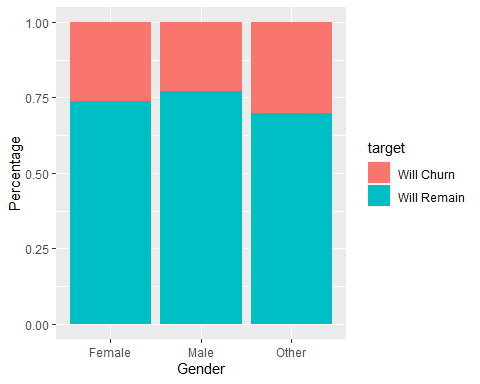
# Drop missing values in experience and last\_new\_job   
  
aug\_train = aug\_train[is.na(aug\_train$experience) != TRUE, ]  
aug\_train = aug\_train[is.na(aug\_train$last\_new\_job) != TRUE, ]  
  
# Check missing data again : 0 missing data  
missmap(aug\_train, main = 'Missing Map', col = c('yellow', 'purple'))



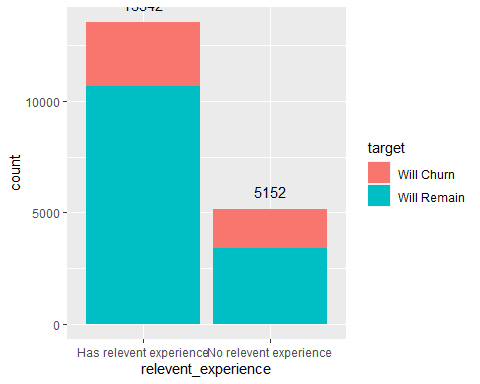
# Plot each variables  
# plot gender  
aug\_train$gender[aug\_train$target == "1"] = "Will Churn"  
# Shows Counts  
ggplot(aug\_train, aes(gender)) + geom\_bar(aes(fill = target)) +  
 labs(y= "Count", x= 'Gender')+   
 geom\_text( stat='count', aes(label=..count..), vjust=-1)



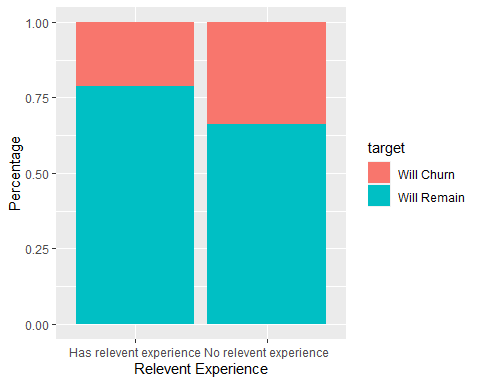
# Shows Percentage: "Other" is highest followed by the Female population  
ggplot(aug\_train, aes(gender)) + geom\_bar(aes(fill = target), position = 'fill') +  
 labs(y= "Percentage", x= 'Gender')



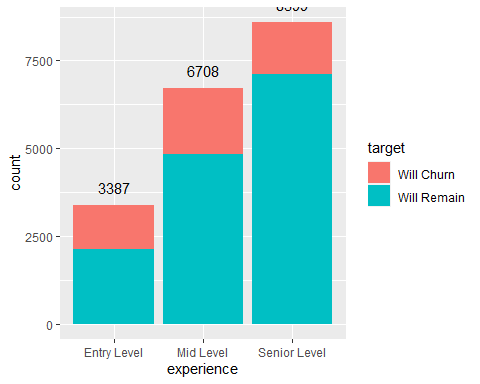
# plot relevant experience  
# Shows Counts  
ggplot(aug\_train, aes(relevent\_experience)) +   
 geom\_bar(aes(fill = target))+  
 geom\_text( stat='count', aes(label=..count..), vjust=-1)



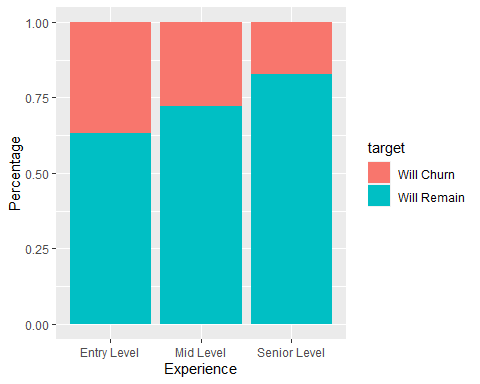
# SHows Percentage: Higher Churn percentage with population with No Relevant Experience  
ggplot(aug\_train, aes(relevent\_experience)) + geom\_bar(aes(fill = target), position = 'fill') +  
 labs(y= "Percentage", x= 'Relevent Experience')



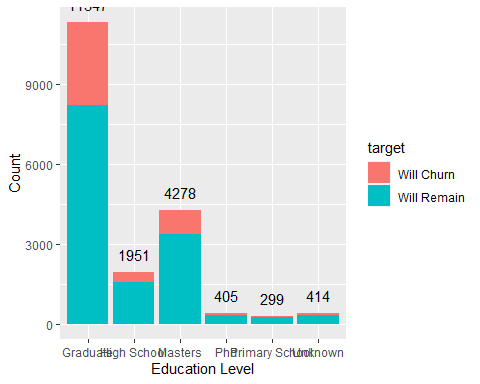
# plot experience  
# Shows COunt  
ggplot(aug\_train, aes(experience)) +   
 geom\_bar(aes(fill = target))+  
 geom\_text( stat='count', aes(label=..count..), vjust=-1)



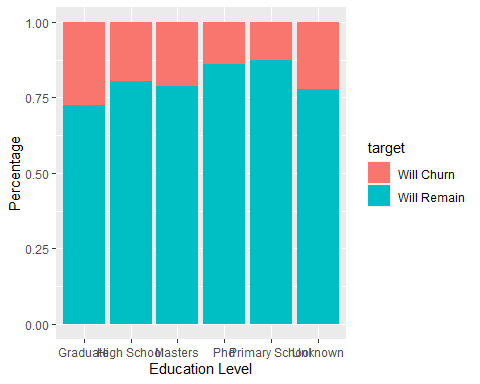
# Shows Percentage: Higher Churn in Entry Level   
ggplot(aug\_train, aes(experience)) + geom\_bar(aes(fill = target), position = 'fill') +  
 labs(y= "Percentage", x= 'Experience')



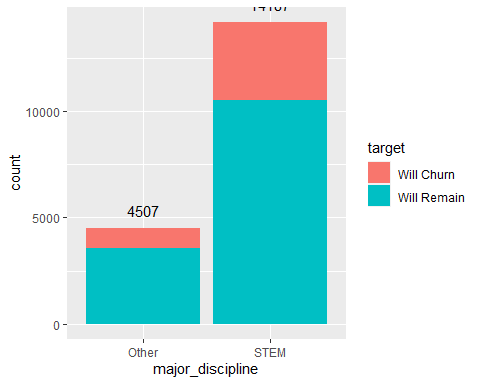
# plot education and percentage  
# Show Counts  
ggplot(aug\_train, aes(education\_level)) + geom\_bar(aes(fill = target)) +  
 labs(y= "Count", x= 'Education Level')+ geom\_text( stat='count', aes(label=..count..), vjust=-1)



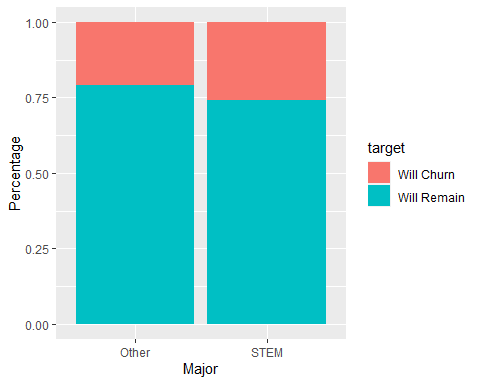
# Show percentage: Highest Churn among those with Graduate level  
ggplot(aug\_train, aes(education\_level)) + geom\_bar(aes(fill = target), position='fill') +  
 labs(y= "Percentage", x= 'Education Level')



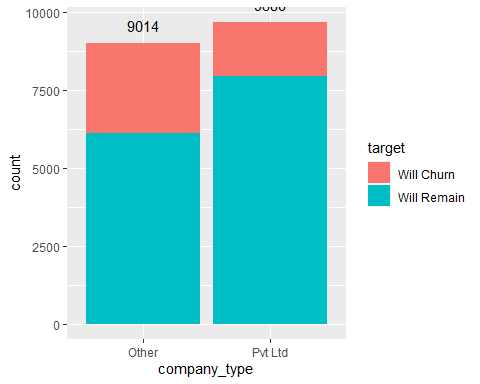
# plot major  
# Show Count  
ggplot(aug\_train, aes(major\_discipline)) +   
 geom\_bar(aes(fill = target))+  
 geom\_text( stat='count', aes(label=..count..), vjust=-1)



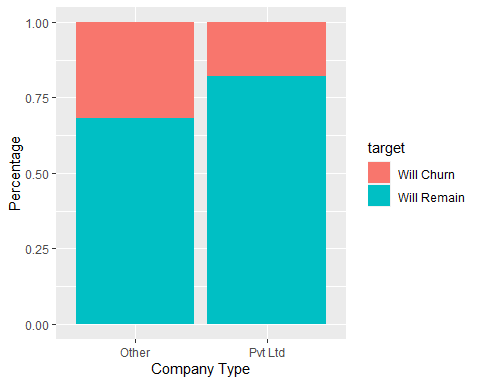
# Show Percentage: "STEM" has higher CHURN percentage  
ggplot(aug\_train, aes(major\_discipline)) + geom\_bar(aes(fill = target), position='fill') +  
 labs(y= "Percentage", x= 'Major')



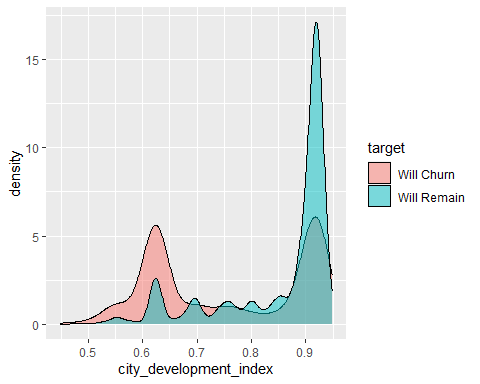
# plot company\_type  
# Show Count  
ggplot(aug\_train, aes(company\_type)) +   
 geom\_bar(aes(fill = target))+  
 geom\_text( stat='count', aes(label=..count..), vjust=-1)



# Show Percentage: "Other" has higher CHURN percentage  
ggplot(aug\_train, aes(company\_type)) + geom\_bar(aes(fill = target), position='fill') +  
 labs(y= "Percentage", x= 'Company Type')



# plot city\_development\_index: index around 0.9 and 0.62 have highest Churn while index at 0.9 will Remain  
# Not indicative of CHURN   
ggplot(aug\_train, aes(city\_development\_index, fill = target))+ geom\_density(alpha = 0.5)



# plot training\_hours: More or less the same ratio. Not very indicative of CHURN  
ggplot(aug\_train, aes(training\_hours, fill = target))+ geom\_density(alpha = 0.5)

