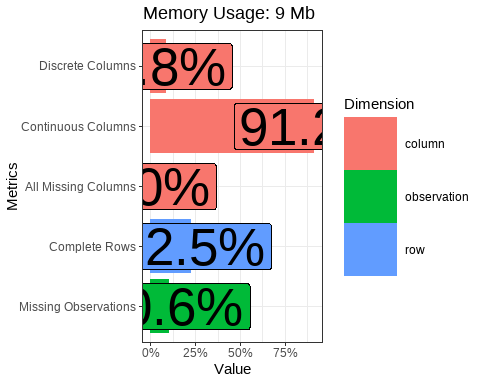
# Dataset Overview

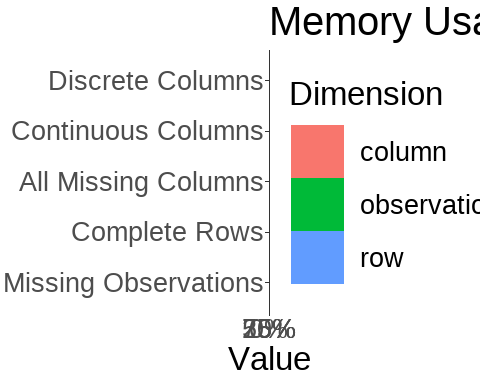
# libraries  
library(tidyverse)  
library(janitor)  
library(flextable)  
library(dlookr)  
library(feather)  
library(DataExplorer)  
# dataset  
datafile <- read\_feather("data/input/customers.feather") %>%   
 select(-New\_Status,-AMount\_Mobilised,-ProductID)

## Data Quality

### Dataset Dimensions.

plot\_intro(datafile,ggtheme = theme\_bw(),geom\_label\_args = list(size=14))+  
 theme(text = element\_text(size = 25))





introduce(datafile) %>%   
 t() %>%   
 as\_tibble(rownames = "Variable") %>%   
 rename('Values'="V1") %>%   
 filter(Variable != "memory\_usage")

## Warning: The `x` argument of `as\_tibble.matrix()` must have unique column names if `.name\_repair` is omitted as of tibble 2.0.0.  
## Using compatibility `.name\_repair`.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was generated.

## # A tibble: 8 x 2  
## Variable Values  
## <chr> <dbl>  
## 1 rows 12512  
## 2 columns 91  
## 3 discrete\_columns 8  
## 4 continuous\_columns 83  
## 5 all\_missing\_columns 0  
## 6 total\_missing\_values 120456  
## 7 complete\_rows 2820  
## 8 total\_observations 1138592

df\_dim = dim(datafile)

The dataset under study had 12512 records and 91 variables inclusive of target variable.

### Variable Explanations.

tibble(  
 ClientID = "Unique Client Identifier",  
 CreatedOn = "Date when the Account Was created",  
 No\_ofProducts = "Number of bank products the client is subscribed to",  
 Tenure = "The length of time a client has stayed with the institution since onboarding in days",  
 GenderID = "Clients Gender",  
 FirstTrxDate = "The first time the client made a transaction after onboarding",  
 LastTrxDate = "The last time the client made a transaction after onboarding",  
 TrxTenure = "Length of time in days between the first and the last transaction",  
 Lag\_Debit = "Difference between the current withdrawal and the preceeding withdrawal",  
 Lag\_Credit = "Difference between the current deposit and the preceeding deposit",  
 No\_Trx = "Number of transactions the client had done with the bank",  
 WeekendTrx = "Number of transactions done on weekends",  
 EndMonthTrx = "Number of transactions done at the end of the month",  
 MidMonthTrx = "Number of transactions done at the mod of the month",  
 TransactedProducts = "Number of products the client has actualy transacted after subscription",  
 Transactions\_per\_week = "Number of transactions a client made on a weeekly basis",  
 Transactions\_per\_month = "Number of transactions a client made on a monthly basis",  
 Transactions\_per\_year = "Number of transactions a client made on a yearly basis",  
 Transactions\_per\_day = "Number of transactions a client made on a daily basis",  
 DepositTrx = "Number of deposit transactions",  
 DepositTrx\_per\_day = "Number of deposit transactions made on a daily basis",  
 DepositTrx\_per\_week = "Number of deposit transactions made on a weekly basis",  
 DepositTrx\_per\_month = "Number of deposit transactions made on a monthly basis",  
 DepositTrx\_per\_year = "Number of deposit transactions made on a yearly basis",  
 Total\_Deposits = "Sum of deposit amount made",  
 Total\_Deposits\_per\_day = "Sum of deposit amount saved on a daily basis",  
 Total\_Deposits\_per\_week = "Sum of deposit amount saved on a weekly basis",  
 Total\_Deposits\_per\_month = "Sum of deposit amount saved on a monthly basis",  
 Total\_Deposits\_per\_year = "Sum of deposit amount saved on a yearly basis"  
) %>% t() %>% as\_tibble(rownames = "Variable") %>% rename("Explanation"="V1")

## # A tibble: 29 x 2  
## Variable Explanation   
## <chr> <chr>   
## 1 ClientID Unique Client Identifier   
## 2 CreatedOn Date when the Account Was created   
## 3 No\_ofProducts Number of bank products the client is subscribed to   
## 4 Tenure The length of time a client has stayed with the institution si~  
## 5 GenderID Clients Gender   
## 6 FirstTrxDate The first time the client made a transaction after onboarding   
## 7 LastTrxDate The last time the client made a transaction after onboarding   
## 8 TrxTenure Length of time in days between the first and the last transact~  
## 9 Lag\_Debit Difference between the current withdrawal and the preceeding w~  
## 10 Lag\_Credit Difference between the current deposit and the preceeding depo~  
## # ... with 19 more rows

### Quality

dlookr::diagnose(datafile) %>%  
 as.data.frame() %>% arrange(desc(missing\_count)) %>%   
 mutate\_at(.vars=names(select(.,-variables,-types)),.funs = function(x)round(x,2))

## variables types missing\_count missing\_percent  
## 1 HasActiveLoans character 9667 77.26  
## 2 Max\_LoansTaken numeric 9667 77.26  
## 3 Max\_LoanTaken numeric 9667 77.26  
## 4 Min\_LoanTaken numeric 9667 77.26  
## 5 Ave\_LoanTaken numeric 9667 77.26  
## 6 First\_LoanTaken numeric 9667 77.26  
## 7 Last\_LoanTaken numeric 9667 77.26  
## 8 sd\_LoanTaken numeric 9667 77.26  
## 9 var\_LoanTaken numeric 9667 77.26  
## 10 Lag\_Debit numeric 1671 13.36  
## 11 Lag\_Credit numeric 1671 13.36  
## 12 Var\_Deposits numeric 1671 13.36  
## 13 sd\_Deposits numeric 1671 13.36  
## 14 Var\_Withdrawals numeric 1671 13.36  
## 15 sd\_Withdrawals numeric 1671 13.36  
## 16 Var\_fixed\_dep numeric 1671 13.36  
## 17 sd\_fixed\_dep numeric 1671 13.36  
## 18 Var\_Deposits\_3 numeric 1671 13.36  
## 19 sd\_Deposits\_3 numeric 1671 13.36  
## 20 Var\_Withdrawals\_3 numeric 1671 13.36  
## 21 sd\_Withdrawals\_3 numeric 1671 13.36  
## 22 Var\_Deposits\_6 numeric 1671 13.36  
## 23 sd\_Deposits\_6 numeric 1671 13.36  
## 24 Var\_Withdrawals\_6 numeric 1671 13.36  
## 25 sd\_Withdrawals\_6 numeric 1671 13.36  
## 26 Var\_Deposits\_12 numeric 1671 13.36  
## 27 sd\_Deposits\_12 numeric 1671 13.36  
## 28 Var\_Withdrawals\_12 numeric 1671 13.36  
## 29 sd\_Withdrawals\_12 numeric 1671 13.36  
## 30 GenderID character 33 0.26  
## 31 ClientID character 0 0.00  
## 32 CreatedOn Date 0 0.00  
## 33 No\_ofProducts integer 0 0.00  
## 34 Tenure numeric 0 0.00  
## 35 FirstTrxDate Date 0 0.00  
## 36 LastTrxDate Date 0 0.00  
## 37 TrxTenure numeric 0 0.00  
## 38 No\_Trx integer 0 0.00  
## 39 WeekendTrx integer 0 0.00  
## 40 WeekdayTrx integer 0 0.00  
## 41 EndMonthTrxs integer 0 0.00  
## 42 MidMonthTrxs integer 0 0.00  
## 43 TransactedProducts integer 0 0.00  
## 44 Transactions\_per\_day numeric 0 0.00  
## 45 Transactions\_per\_week numeric 0 0.00  
## 46 Transactions\_per\_month numeric 0 0.00  
## 47 Transactions\_per\_year numeric 0 0.00  
## 48 DepositTrx numeric 0 0.00  
## 49 DepositTrx\_per\_day numeric 0 0.00  
## 50 DepositTrx\_per\_week numeric 0 0.00  
## 51 DepositTrx\_per\_month numeric 0 0.00  
## 52 DepositTrx\_per\_year numeric 0 0.00  
## 53 Total\_Deposits numeric 0 0.00  
## 54 Total\_Deposits\_per\_day numeric 0 0.00  
## 55 Total\_Deposits\_per\_week numeric 0 0.00  
## 56 Total\_Deposits\_per\_month numeric 0 0.00  
## 57 Total\_Deposits\_per\_year numeric 0 0.00  
## 58 Max\_Deposits numeric 0 0.00  
## 59 Av\_Deposits numeric 0 0.00  
## 60 Total\_Withdrawals numeric 0 0.00  
## 61 Max\_Withdrawals numeric 0 0.00  
## 62 Av\_Withdrawals numeric 0 0.00  
## 63 total\_fixed\_dep numeric 0 0.00  
## 64 max\_fixed\_dep numeric 0 0.00  
## 65 Av\_fixed\_dep numeric 0 0.00  
## 66 No\_Trx\_3 numeric 0 0.00  
## 67 No\_Trx\_6 numeric 0 0.00  
## 68 No\_Trx\_12 numeric 0 0.00  
## 69 Total\_Deposits\_3 numeric 0 0.00  
## 70 Max\_Deposits\_3 numeric 0 0.00  
## 71 Av\_Deposits\_3 numeric 0 0.00  
## 72 Total\_Withdrawals\_3 numeric 0 0.00  
## 73 Max\_Withdrawals\_3 numeric 0 0.00  
## 74 Av\_Withdrawals\_3 numeric 0 0.00  
## 75 Total\_Deposits\_6 numeric 0 0.00  
## 76 Max\_Deposits\_6 numeric 0 0.00  
## 77 Av\_Deposits\_6 numeric 0 0.00  
## 78 Total\_Withdrawals\_6 numeric 0 0.00  
## 79 Max\_Withdrawals\_6 numeric 0 0.00  
## 80 Av\_Withdrawals\_6 numeric 0 0.00  
## 81 Total\_Deposits\_12 numeric 0 0.00  
## 82 Max\_Deposits\_12 numeric 0 0.00  
## 83 Av\_Deposits\_12 numeric 0 0.00  
## 84 Total\_Withdrawals\_12 numeric 0 0.00  
## 85 Max\_Withdrawals\_12 numeric 0 0.00  
## 86 Av\_Withdrawals\_12 numeric 0 0.00  
## 87 days\_since\_last\_trx numeric 0 0.00  
## 88 HasFixedDeposit numeric 0 0.00  
## 89 LastDepositBalance numeric 0 0.00  
## 90 ChurnStatus character 0 0.00  
## 91 LoanHistory character 0 0.00  
## unique\_count unique\_rate  
## 1 3 0.00  
## 2 30 0.00  
## 3 280 0.02  
## 4 269 0.02  
## 5 1092 0.09  
## 6 245 0.02  
## 7 323 0.03  
## 8 966 0.08  
## 9 966 0.08  
## 10 1765 0.14  
## 11 979 0.08  
## 12 7306 0.58  
## 13 7306 0.58  
## 14 7091 0.57  
## 15 7091 0.57  
## 16 354 0.03  
## 17 354 0.03  
## 18 2910 0.23  
## 19 2910 0.23  
## 20 2651 0.21  
## 21 2651 0.21  
## 22 3877 0.31  
## 23 3876 0.31  
## 24 3690 0.29  
## 25 3690 0.29  
## 26 5140 0.41  
## 27 5140 0.41  
## 28 4925 0.39  
## 29 4925 0.39  
## 30 3 0.00  
## 31 12512 1.00  
## 32 731 0.06  
## 33 6 0.00  
## 34 731 0.06  
## 35 773 0.06  
## 36 761 0.06  
## 37 742 0.06  
## 38 262 0.02  
## 39 108 0.01  
## 40 215 0.02  
## 41 135 0.01  
## 42 212 0.02  
## 43 6 0.00  
## 44 4462 0.36  
## 45 4644 0.37  
## 46 4612 0.37  
## 47 4660 0.37  
## 48 95 0.01  
## 49 1961 0.16  
## 50 2034 0.16  
## 51 2011 0.16  
## 52 2042 0.16  
## 53 5555 0.44  
## 54 7515 0.60  
## 55 7609 0.61  
## 56 7574 0.61  
## 57 7723 0.62  
## 58 1167 0.09  
## 59 6303 0.50  
## 60 5011 0.40  
## 61 1257 0.10  
## 62 5898 0.47  
## 63 146 0.01  
## 64 115 0.01  
## 65 318 0.03  
## 66 93 0.01  
## 67 139 0.01  
## 68 204 0.02  
## 69 1670 0.13  
## 70 638 0.05  
## 71 2551 0.20  
## 72 1663 0.13  
## 73 625 0.05  
## 74 2373 0.19  
## 75 2544 0.20  
## 76 750 0.06  
## 77 3431 0.27  
## 78 2587 0.21  
## 79 824 0.07  
## 80 3279 0.26  
## 81 3691 0.29  
## 82 951 0.08  
## 83 4546 0.36  
## 84 3520 0.28  
## 85 985 0.08  
## 86 4219 0.34  
## 87 761 0.06  
## 88 2 0.00  
## 89 1090 0.09  
## 90 2 0.00  
## 91 2 0.00

**Observations and Inferences**

From the above ,it was noticed that a number of variables had missing values.The missing values occurrence where however systematic showing an indication that the client had never used the service/product before.These will later require imputation.

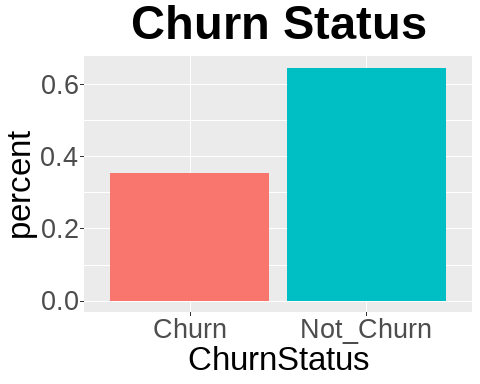
# Exploratory Data Analysis.

The main interest under exploratory analysis is to understand the underlying distrubutions on every variable and their interactions. ## Univariate Analysis. ### Target Variable

datafile %>% tabyl(ChurnStatus) %>% adorn\_totals("row")

## ChurnStatus n percent  
## Churn 4424 0.3535806  
## Not\_Churn 8088 0.6464194  
## Total 12512 1.0000000

plotcol = "purple"  
plots\_theme <- theme(text = element\_text(size=25),  
 plot.title = element\_text(size = 35,hjust = .5,face = "bold"))  
datafile %>% tabyl(ChurnStatus) %>%   
 ggplot(aes(ChurnStatus,percent)) +  
 geom\_col(aes(fill = factor(c("green","red"),labels = c("Churn","Not Churn"))))+  
 plots\_theme+theme(legend.position = 'none')+  
 labs(title = "Churn Status")



**Observations and inferences**

As we can observe from above,the institution had a 35% churn rate.

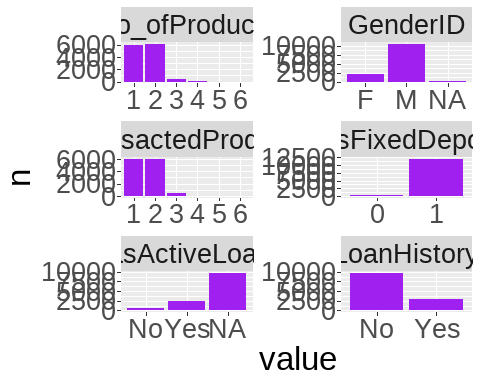
### Explanatory Variables

#### Categorical Variables

discretevars = datafile %>% select(-ChurnStatus) %>%   
 summarise\_all(.funs = function(x)length(unique(x))) %>%   
 t() %>%   
 as\_tibble(rownames = "Variable") %>%   
 filter(V1<=11) %>% pull(Variable)  
  
datafile %>% select(discretevars) %>%   
 mutate\_all(as.factor) %>%  
rowid\_to\_column() %>%   
 reshape2::melt(id.vars="rowid") %>%   
 group\_by(variable,value) %>%   
 tally() %>%   
 ggplot(aes(value,n))+  
 geom\_col(fill=plotcol)+  
 facet\_wrap(.~variable,scales = 'free',nrow = 3)+  
 plots\_theme

## Note: Using an external vector in selections is ambiguous.  
## i Use `all\_of(discretevars)` instead of `discretevars` to silence this message.  
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.  
## This message is displayed once per session.

## Warning: attributes are not identical across measure variables; they will be  
## dropped



**Observations and inference**

1. Majority of customers have subscribed for 2 products.
2. Most customers apear to be males interms of gender
3. Of the rigistered customers,majority have transacted in more than 2 products.
4. Majority of the customers have no fixed deposit with the bank.
5. Most clients who have transacted didnt have a running loan at the time pf the analysis
6. A bigger percentage of customers have never taken loans with the bank.

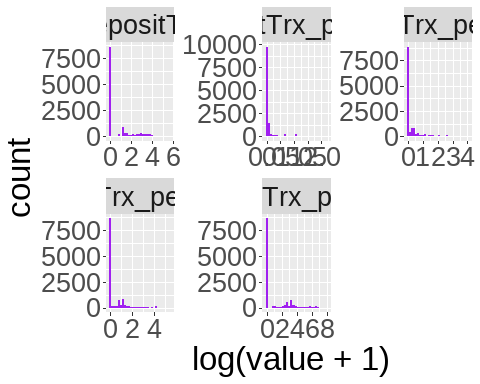
#### Numerical Variables

numvars <- datafile %>% select(-discretevars,-CreatedOn,-ClientID,-contains('Date'))

##### Deposits transactions count Related

datafile %>% select(contains('deposittrx')) %>%   
 rowid\_to\_column() %>%   
 reshape2::melt('rowid') %>%   
 ggplot(aes(log(value+1)))+  
 geom\_histogram(fill=plotcol)+facet\_wrap(.~variable,scales = 'free')+  
 plots\_theme

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

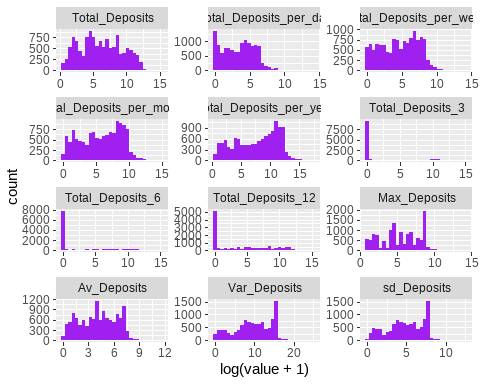


##### Deposit transaction amount Related

numvars %>%   
 select(contains('Total\_Deposits'),Max\_Deposits,Av\_Deposits,Var\_Deposits,sd\_Deposits) %>%   
 rowid\_to\_column() %>%   
 reshape2::melt('rowid') %>%   
 ggplot(aes(log(value+1)))+  
 geom\_histogram(fill=plotcol)+  
 facet\_wrap(.~variable,scales = 'free',ncol = 3,nrow=4)

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

## Warning: Removed 3342 rows containing non-finite values (stat\_bin).



# datafile

##### Transaction Count Related

###### Descriptive Statistics

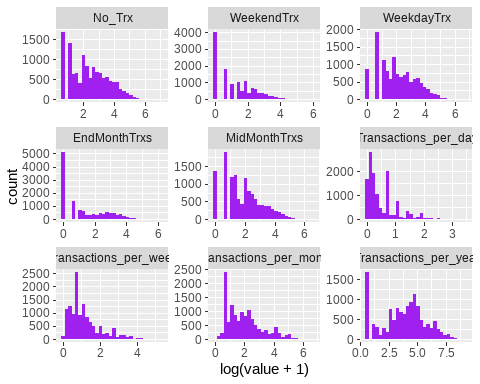
numvars %>%   
 select(No\_Trx,WeekendTrx,WeekdayTrx,EndMonthTrxs,MidMonthTrxs,  
 contains('Transactions\_per')) %>%   
 describe()

## # A tibble: 9 x 26  
## variable n na mean sd se\_mean IQR skewness kurtosis p00  
## <chr> <int> <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 No\_Trx 12512 0 22.4 41.2 0.368 22 7.64 125. 1   
## 2 WeekendT~ 12512 0 5.91 12.1 0.108 7 8.74 170. 0   
## 3 WeekdayT~ 12512 0 16.5 30.3 0.271 16 7.01 102. 0   
## 4 EndMonth~ 12512 0 8.53 16.8 0.150 10 5.84 90.9 0   
## 5 MidMonth~ 12512 0 13.9 29.3 0.262 12 7.38 103. 0   
## 6 Transact~ 12512 0 0.943 1.64 0.0146 0.889 4.87 42.6 0.00378  
## 7 Transact~ 12512 0 4.27 9.27 0.0829 2.45 6.89 83.8 0.0265   
## 8 Transact~ 12512 0 17.0 40.0 0.357 10.4 6.85 82.5 0.113   
## 9 Transact~ 12512 0 203. 488. 4.36 133. 6.81 81.7 1   
## # ... with 16 more variables: p01 <dbl>, p05 <dbl>, p10 <dbl>, p20 <dbl>,  
## # p25 <dbl>, p30 <dbl>, p40 <dbl>, p50 <dbl>, p60 <dbl>, p70 <dbl>,  
## # p75 <dbl>, p80 <dbl>, p90 <dbl>, p95 <dbl>, p99 <dbl>, p100 <dbl>

###### Visualization

numvars %>% select(No\_Trx,WeekendTrx,WeekdayTrx,EndMonthTrxs,MidMonthTrxs,contains('Transactions\_per')) %>%   
 rowid\_to\_column() %>%   
 reshape2::melt('rowid') %>%   
 ggplot(aes(log(value+1)))+  
 geom\_histogram(fill=plotcol)+facet\_wrap(.~variable,scales = 'free')

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



**Observations and Inferences**

* Number of transactions appear to be right skewed at Mean of 22.41 and sd = 41, transactions.Most clients have done 1-3 transactions only.
* Yearly transactions were observed to be evenly distributed across the year.
* Most customers make at most 3 transactions per month and less than 2 per week.

##### Withdrawals Related

###### Descriptive Statistics table

numvars %>% select(Total\_Withdrawals,Max\_Withdrawals,Av\_Withdrawals,Var\_Withdrawals,sd\_Withdrawals) %>% describe() %>%   
 mutate\_at(.vars = names(select(.,-variable)),.funs = function(x)round(x,2))

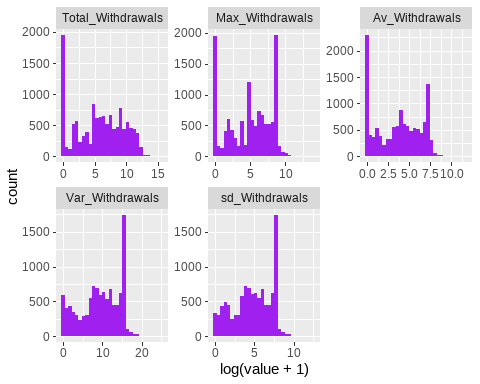
## # A tibble: 5 x 26  
## variable n na mean sd se\_mean IQR skewness kurtosis p00  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Total\_Withdr~ 12512 0 1.93e4 1.12e5 9.97e2 8.31e3 28.0 1113. 0  
## 2 Max\_Withdraw~ 12512 0 2.57e3 2.23e4 2.00e2 2.19e3 33.8 1366. 0  
## 3 Av\_Withdrawa~ 12512 0 4.27e2 1.65e3 1.47e1 4.43e2 40.7 2513. 0  
## 4 Var\_Withdraw~ 10841 1671 2.15e7 9.17e8 8.81e6 1.32e6 69.7 5153. 0  
## 5 sd\_Withdrawa~ 10841 1671 8.89e2 4.55e3 4.37e1 1.13e3 38.1 1925. 0  
## # ... with 16 more variables: p01 <dbl>, p05 <dbl>, p10 <dbl>, p20 <dbl>,  
## # p25 <dbl>, p30 <dbl>, p40 <dbl>, p50 <dbl>, p60 <dbl>, p70 <dbl>,  
## # p75 <dbl>, p80 <dbl>, p90 <dbl>, p95 <dbl>, p99 <dbl>, p100 <dbl>

###### Visualization

numvars %>% select(Total\_Withdrawals,Max\_Withdrawals,Av\_Withdrawals,Var\_Withdrawals,sd\_Withdrawals) %>%   
 rowid\_to\_column() %>%   
 reshape2::melt('rowid') %>%   
 ggplot(aes(log(value+1)))+  
 geom\_histogram(fill=plotcol)+facet\_wrap(.~variable,scales = 'free')

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

## Warning: Removed 3342 rows containing non-finite values (stat\_bin).



## Bivariate Analaysis

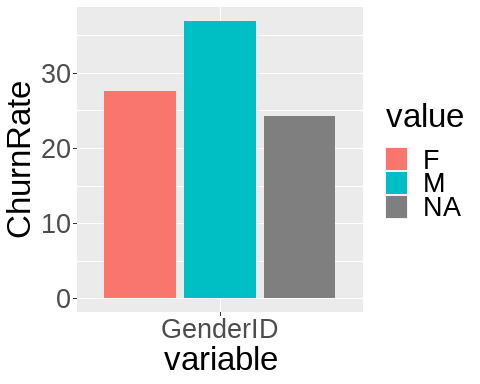
### Target Variable Vs Categorical Variables

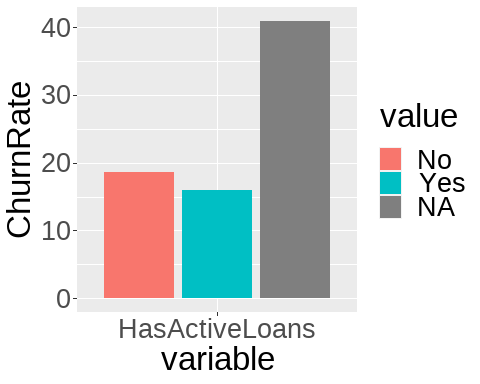
#### Descriptive Statistics

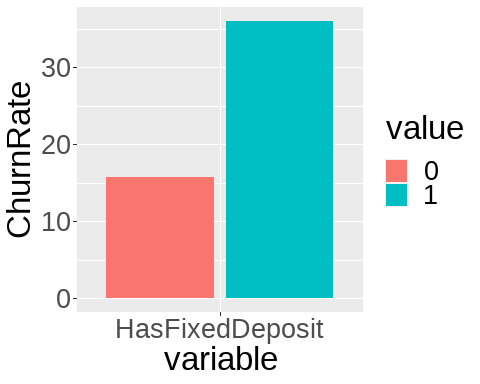
target\_cat\_sum <- datafile %>%   
 select(discretevars,ChurnStatus) %>%   
 rowid\_to\_column() %>%  
 reshape2::melt(id.vars=c('rowid',"ChurnStatus")) %>%   
 group\_by(ChurnStatus,variable,value) %>%   
 tally() %>%   
 reshape2::dcast(variable+value~ChurnStatus,value.var = "n",  
 fun.aggregate = sum,na.rm=T) %>%   
 split(.$variable) %>%   
 map\_df(~adorn\_totals(.,c('row','col')) %>%   
 mutate(ChurnRate = paste(round(Churn/Total \* 100,2) ,"%"))) %>%  
 as\_tibble()   
target\_cat\_sum

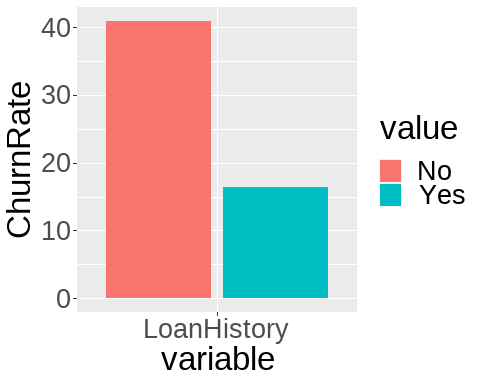
## # A tibble: 28 x 6  
## variable value Churn Not\_Churn Total ChurnRate  
## <chr> <chr> <int> <int> <dbl> <chr>   
## 1 No\_ofProducts 1 2999 2934 5933 50.55 %   
## 2 No\_ofProducts 2 1360 4689 6049 22.48 %   
## 3 No\_ofProducts 3 57 421 478 11.92 %   
## 4 No\_ofProducts 4 8 37 45 17.78 %   
## 5 No\_ofProducts 5 0 6 6 0 %   
## 6 No\_ofProducts 6 0 1 1 0 %   
## 7 Total - 4424 8088 12512 35.36 %   
## 8 GenderID F 565 1488 2053 27.52 %   
## 9 GenderID M 3851 6575 10426 36.94 %   
## 10 GenderID <NA> 8 25 33 24.24 %   
## # ... with 18 more rows

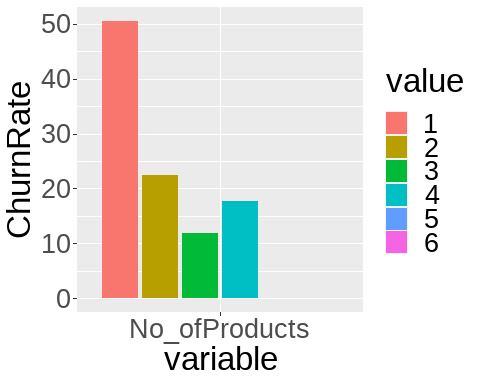
plots <- target\_cat\_sum %>%   
 mutate(ChurnRate = parse\_number(ChurnRate)) %>%   
 split(.$variable) %>%   
 map(~ggplot(.,aes(variable,ChurnRate,fill = value))+  
 geom\_col(position=position\_dodge(width=1))+  
 # labs(title = paste("Churn Rate Vs ",variable))+  
 plots\_theme)  
  
for (i in plots) {  
 print(i)  
}

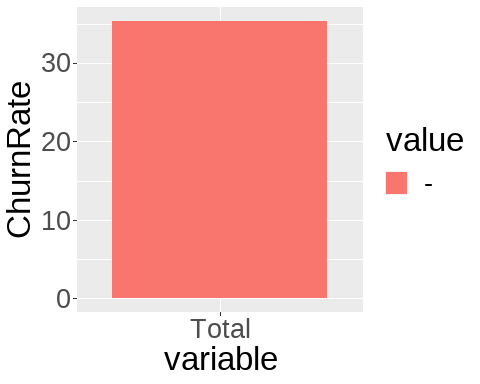


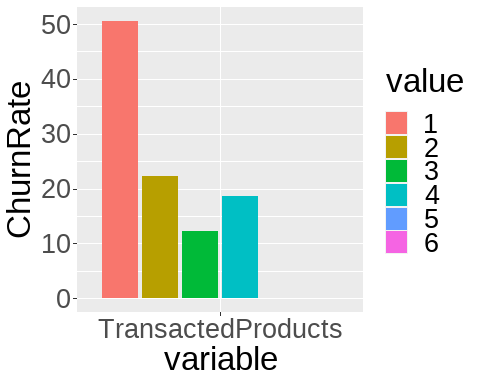












### Target Variable Vs Numerical Variables

##### Deposits transactions count Related

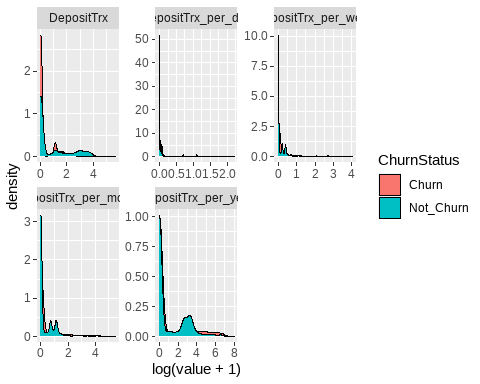
###### Descriptive Statistics

datafile %>%   
 select(contains('deposittrx'),ChurnStatus) %>%   
 group\_by(ChurnStatus) %>%   
 describe() %>%   
 mutate\_at(.vars = names(select(.,-variable,-ChurnStatus)),  
 .funs = function(x)round(x,2))

## # A tibble: 10 x 27  
## variable ChurnStatus n na mean sd se\_mean IQR skewness kurtosis  
## <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Deposit~ Churn 4424 0 0.94 3.13 0.05 0 9.44 150.   
## 2 Deposit~ Not\_Churn 8088 0 6.09 13.7 0.15 4 4.81 51.7  
## 3 Deposit~ Churn 4424 0 0.16 0.53 0.01 0 5.23 39.9  
## 4 Deposit~ Not\_Churn 8088 0 0.09 0.35 0 0.04 6.41 52.0  
## 5 Deposit~ Churn 4424 0 3.35 12.6 0.19 0 6.41 60.5  
## 6 Deposit~ Not\_Churn 8088 0 1.96 8.55 0.1 1.07 8.68 97.1  
## 7 Deposit~ Churn 4424 0 0.82 2.95 0.04 0 6.32 59.4  
## 8 Deposit~ Not\_Churn 8088 0 0.48 2 0.02 0.26 8.57 95.6  
## 9 Deposit~ Churn 4424 0 40.3 153. 2.3 0 6.41 60.4  
## 10 Deposit~ Not\_Churn 8088 0 23.6 104. 1.16 12.9 8.68 97.0  
## # ... with 17 more variables: p00 <dbl>, p01 <dbl>, p05 <dbl>, p10 <dbl>,  
## # p20 <dbl>, p25 <dbl>, p30 <dbl>, p40 <dbl>, p50 <dbl>, p60 <dbl>,  
## # p70 <dbl>, p75 <dbl>, p80 <dbl>, p90 <dbl>, p95 <dbl>, p99 <dbl>,  
## # p100 <dbl>

###### Visualization

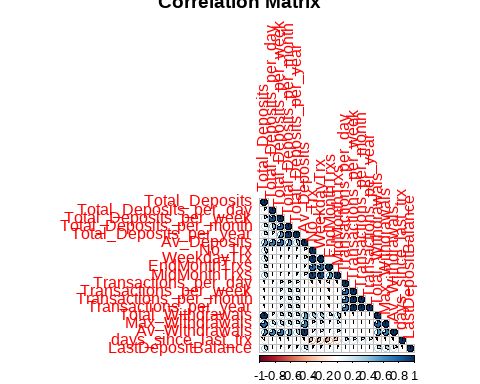
datafile %>%   
 select(contains('deposittrx'),ChurnStatus) %>%   
 rowid\_to\_column() %>%   
 reshape2::melt(c('rowid','ChurnStatus')) %>%   
 ggplot(aes(log(value+1),fill=ChurnStatus))+  
 geom\_density()+facet\_wrap(.~variable,scales = 'free')



### Correlation Analysis.

In order to plot correlation,features that exhibited some distributions were selected.

selected\_num\_vars <- datafile %>%   
 select(Total\_Deposits,Total\_Deposits\_per\_day,Total\_Deposits\_per\_week,  
 Total\_Deposits\_per\_month,Total\_Deposits\_per\_year,Av\_Deposits,  
 No\_Trx,WeekdayTrx,WeekdayTrx,EndMonthTrxs,MidMonthTrxs,  
 Transactions\_per\_day,Transactions\_per\_week,Transactions\_per\_month,  
 Transactions\_per\_year,Total\_Withdrawals,Max\_Withdrawals,Av\_Withdrawals,  
 days\_since\_last\_trx,LastDepositBalance)  
correlation <- cor(selected\_num\_vars)  
corrplot::corrplot(correlation,method = "pie",type = "lower",title = "Correlation Matrix",diag = T)



Observations and Inferences

* Some variables are highly correlated with others. eg Total withdrawals vs Total deposits.
* This correlation will be dealt with at feature engineering phase.

**Variable Selection for the next stage of analysis**

modelData <-   
datafile %>% select(discretevars,ChurnStatus) %>%   
 mutate(  
 GenderID = ifelse(is.na(GenderID),'Unknown',GenderID),  
 HasActiveLoans = ifelse(is.na(HasActiveLoans),"NO",HasActiveLoans)) %>%   
 mutate\_all(as.factor) %>%   
 bind\_cols(selected\_num\_vars)  
  
write\_feather(modelData,"data/output/model\_data.feather")