Save best for the Last?

Ravi Hela

2 December 2018

## Motivation

It is frequently observed that there is a surge of admissions during monthend day in JG courses. There is a race at month end to come at the top of the stack by showing most admissions. Other infuencer is pressure from Sales team at Head Office to bring admission. There is a concern that such factors possibly lead to compromises on quality of student. In this context we define quality of student as his disposition to pay his full fees. In our data set this is captured by feature AR% which we will refer to as Average Realisation. This AR% is a preprocessed variable inthiis dataset which is the ratio of collection over the booking value.

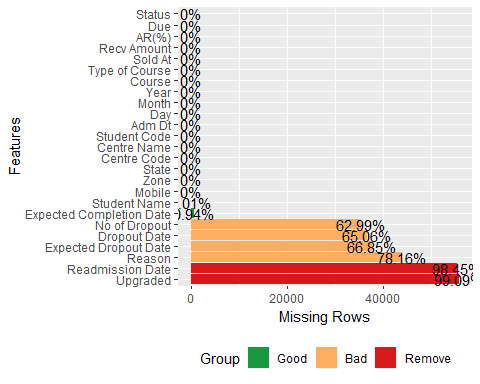
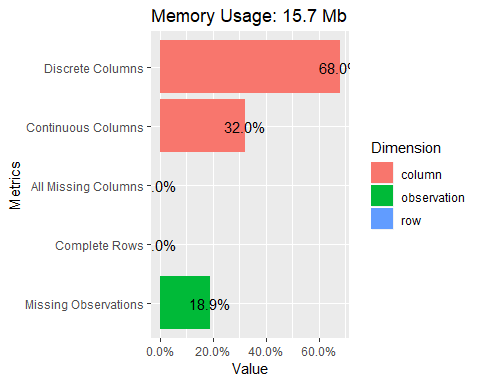
## Research Question:

### Do monthend admissions tend to have lower average realisation

#### Introduction to the Dataset

Lets get introduced to the dataset

## Observations: 56,376  
## Variables: 25  
## $ Zone <chr> "EAST", "EAST", "NORTH", "EAST", "E...  
## $ State <chr> "ASM", "ASM", "DLH", "ASM", "WB", "...  
## $ `Centre Code` <chr> "A034", "A081", "B031", "A109", "A0...  
## $ `Centre Name` <chr> "Silchar", "Chandmari", "North Camp...  
## $ `Student Name` <chr> "Rukiya parbin Barbhuiya", "Manobi ...  
## $ `Student Code` <chr> "A0341839", "A0811508", "B0311341",...  
## $ Mobile <dbl> 9954083617, 7576920310, 8860371725,...  
## $ `Adm Dt` <chr> "01-04-2015", "01-04-2015", "01-04-...  
## $ Day <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...  
## $ Month <int> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,...  
## $ Year <int> 2015, 2015, 2015, 2015, 2015, 2015,...  
## $ Course <chr> "Certified Industrial Accountant (C...  
## $ `Type of Course` <chr> "JG", "JG", "NJG", "NJG", "JG", "JG...  
## $ `Sold At` <int> 35600, 36100, 3500, 4000, 57000, 57...  
## $ `Recv Amount` <int> 32992, 22776, 3500, 4000, 45400, 50...  
## $ `AR(%)` <chr> "93%", "63%", "100%", "100%", "80%"...  
## $ Due <dbl> 2608.00, 13324.00, 0.00, 0.00, 1160...  
## $ Status <chr> "AUTODROPOUT", "AUTODROPOUT", "AUTO...  
## $ `Expected Completion Date` <chr> "Not Applicable", "Not Applicable",...  
## $ `Expected Dropout Date` <chr> NA, NA, NA, NA, "13-07-2018", "13-0...  
## $ `Readmission Date` <chr> NA, NA, NA, NA, "13-11-2017", "13-1...  
## $ `Dropout Date` <chr> "31-12-2016", "31-12-2016", "01-11-...  
## $ Upgraded <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA,...  
## $ Reason <chr> NA, NA, NA, "Not Interested", NA, N...  
## $ `No of Dropout` <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2,...



The data has 56376 rows and 25 columns. The dataset has admission date ranging from April 2015 to October 2018. We remove heavily missing variable. First 4 var are clearly factor var. Converting them to factor. We are only interested in JG courses hence we remove the NJG courses form the Dataset. Variables like Adm Dt, Course, AR(%), Status and Expected completion date have wrongly coded data type. We explore, confirm and fix these variables. We drop Mobile number as it is not relevant for this study.

rm\_var <- as.character((DataExplorer::profile\_missing(ar\_student) %>% filter(group %in% c("Remove", "Bad")) %>% select(feature))[, 1] %>% sapply(as.character))  
ar\_student <- ar\_student[ , -(which(names(ar\_student) %in% rm\_var))]  
  
ar\_student[, c(1:4)] <- lapply(ar\_student[, c(1:4)], factor)  
ar\_student <- ar\_student %>% filter(`Type of Course` == "JG")  
ar\_student$`Type of Course`<- NULL  
ar\_student$Mobile <- NULL  
  
head(ar\_student$`Adm Dt`)

## [1] "01-04-2015" "01-04-2015" "01-04-2015" "01-04-2015" "01-04-2015"  
## [6] "01-04-2015"

ar\_student$`Adm Dt` <- lubridate::dmy(ar\_student$`Adm Dt`)  
head(ar\_student$Course)

## [1] "Certified Industrial Accountant (CIA) 2014 - NM"   
## [2] "Certified Industrial Accountant (CIA) 2014 - NM"   
## [3] "Certified Industrial Accountant Plus (CIA+) 2014 - NM"  
## [4] "Certified Industrial Accountant Plus (CIA+) 2014 - NM"  
## [5] "Certified Industrial Accountant (CIA) 2014 - M"   
## [6] "Certified Industrial Accountant Plus (CIA+) 2014 - NM"

tail(ar\_student$Course)

## [1] "Certified Industrial Accountant Plus New"   
## [2] "Certified Industrial Accountant Express"   
## [3] "Certified Industrial Accountant Plus"   
## [4] "CIA Exemption"   
## [5] "Certified Industrial Accountant Plus (CIA+)"  
## [6] "CIA Exemption"

length(unique(ar\_student$Course))

## [1] 78

ar\_student$Course <- as.factor(ar\_student$Course)  
unique(ar\_student$Status)

## [1] "AUTODROPOUT" "ACTIVE" "DROPOUT" "DREQUEST" "TREQUEST"   
## [6] "Certified"

ar\_student$Status <- as.factor(ar\_student$Status)  
table(ar\_student$Status)

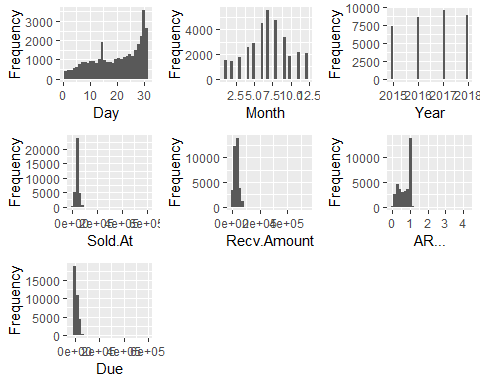
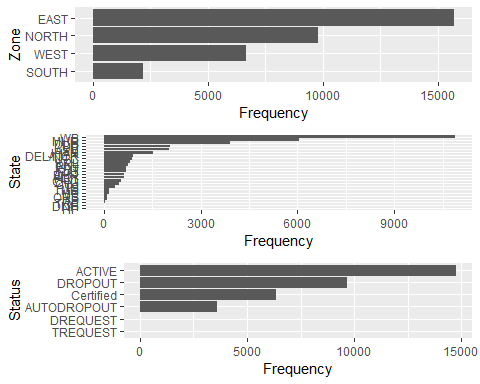
##   
## ACTIVE AUTODROPOUT Certified DREQUEST DROPOUT TREQUEST   
## 14764 3620 6330 5 9654 1

#### Exploratory Data Analysis

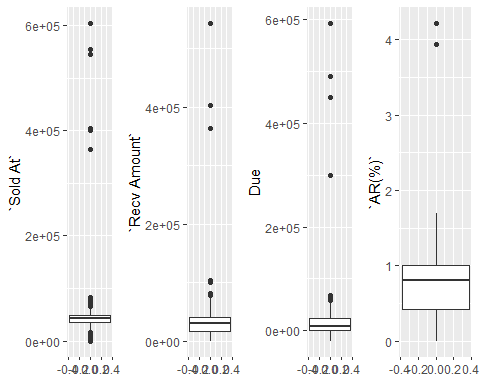
Lets have a look at the dataset once again

## Observations: 34,374  
## Variables: 17  
## $ Zone <fct> EAST, EAST, EAST, EAST, WEST, EAST,...  
## $ State <fct> ASM, ASM, WB, WB, MHR, WB, WB, WB, ...  
## $ `Centre Code` <fct> A034, A081, A044, A044, K000, A044,...  
## $ `Centre Name` <fct> Silchar, Chandmari, Dunlop, Dunlop,...  
## $ `Student Name` <chr> "Rukiya parbin Barbhuiya", "Manobi ...  
## $ `Student Code` <chr> "A0341839", "A0811508", "A0441375",...  
## $ `Adm Dt` <date> 2015-04-01, 2015-04-01, 2015-04-01...  
## $ Day <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...  
## $ Month <int> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,...  
## $ Year <int> 2015, 2015, 2015, 2015, 2015, 2015,...  
## $ Course <fct> Certified Industrial Accountant (CI...  
## $ `Sold At` <int> 35600, 36100, 57000, 57400, 32400, ...  
## $ `Recv Amount` <int> 32992, 22776, 45400, 50400, 32400, ...  
## $ `AR(%)` <dbl> 0.93, 0.63, 0.80, 0.88, 1.00, 0.80,...  
## $ Due <dbl> 2608.00, 13324.00, 11600.00, 7000.0...  
## $ Status <fct> AUTODROPOUT, AUTODROPOUT, ACTIVE, A...  
## $ `Expected Completion Date` <chr> "Not Applicable", "Not Applicable",...

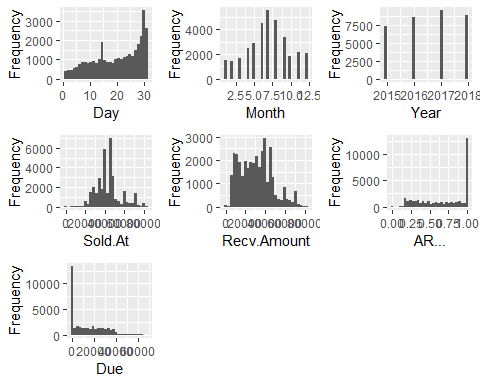
## 7 columns ignored with more than 50 categories.  
## Centre.Code: 160 categories  
## Centre.Name: 157 categories  
## Student.Name: 29426 categories  
## Student.Code: 34338 categories  
## Adm.Dt: 1234 categories  
## Course: 78 categories  
## Expected.Completion.Date: 1312 categories



There also seems to be some extreme values of variabale Sold at, Recv amount, Due and AR(%). lets look their distribution using boxplots.

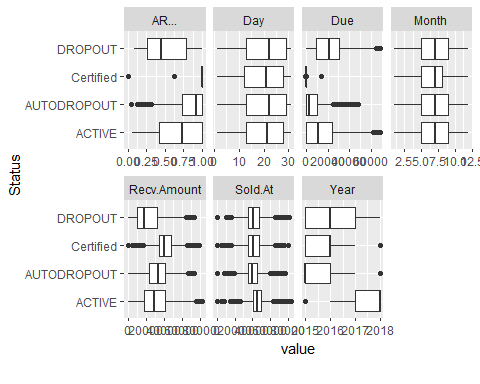


By bussiness knowledge we know the highest SP of a product shoud be less than a lakh. so remove such observation. AR(%) cannot be more than 100%. so we remove the outiers in AR(%). Lets clean up these above variables of their exceptional values.

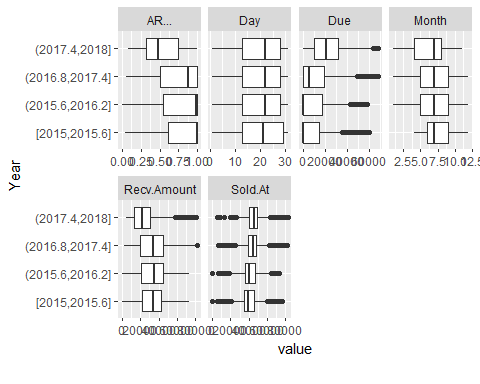
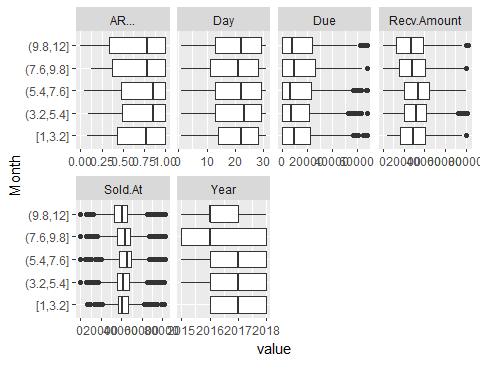
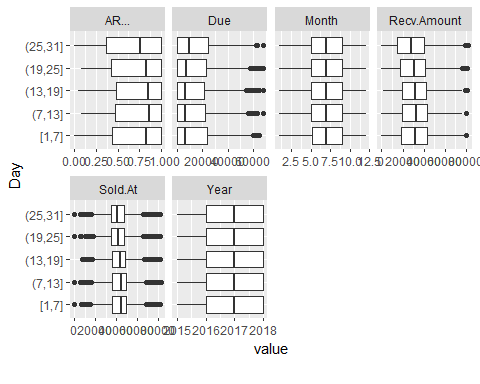
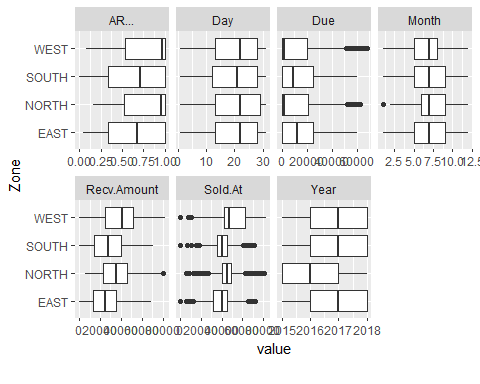
lets explore histograms once more for distribution after clean up 

1. We see from histograms that most admissions happen at Jun July and Aug.
2. We also see that Admissions peaks up during last days of the month.

Lets explore how our numeric variables are distributed across other categorical variable. We start with Status



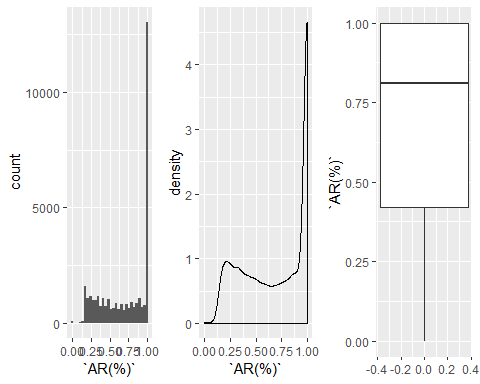
1. The median average realisation of dropouts are less than active student
2. Students dropping out has higer dues.
3. Bulk of the admissiom hapens in June-July-August
4. Received amount is highest for certified student and so reflected in median of Average realisation

Lets explore across other categorical variable 

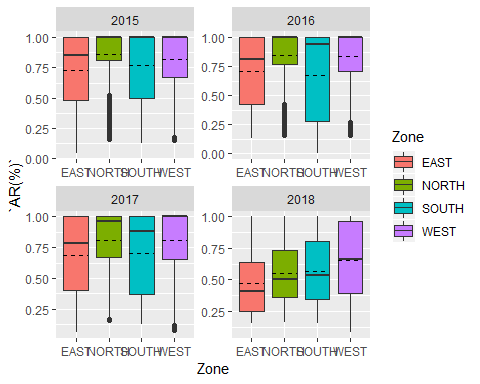
1. South and East have visibly lower averagae realisation than North and West.
2. Median of day of admission is above 22 indicates most Admission happens in end of 3rd to 4th week.
3. Median of Average realisation is lowest during last five days of the month.

Lets explore the distribution of Average Realisation

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

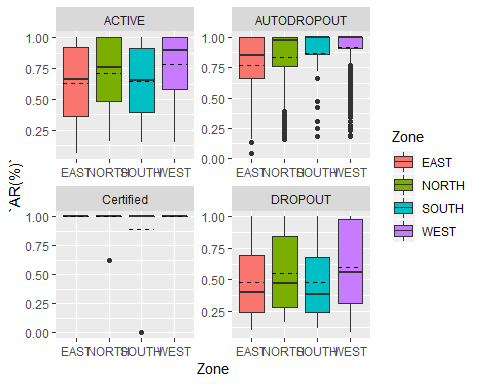


Looks like AR for most of the students is above 75%. We have see in the previous plot that Avg Realisation of East and South is very low caompared to North and West. Lets explore AR across Zone and facetted by Year to see if this observation is consisten YoY.

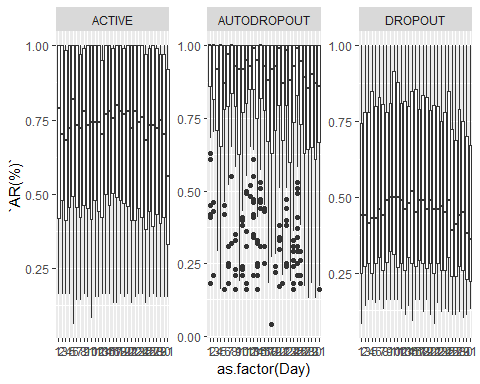


The trend of low recoevery is consistnet over last 3 years.

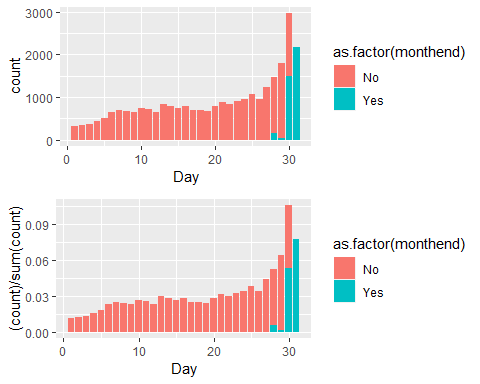
Lets see interaction of Average realisation across Zone and Student status.



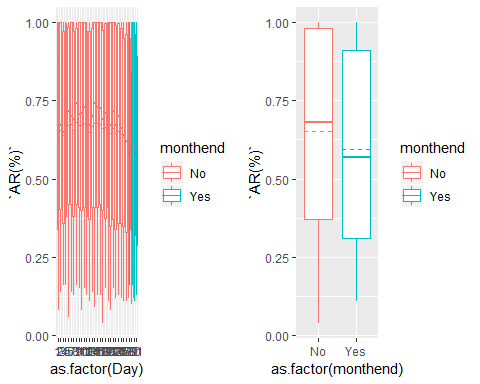
We observe South and East having Low median Avg realisation Compared to North aand West. Lets see how Average Realisation is distributed across days of the month.

Across Day 

We observe that the median value of Avg Realisation for Students admitted during end of days are lower in all three groups. Lets see what propotion of students get admitted at each day of the motnh. We want to see how it is different in month end. We first find out the month end dates. Then proceed to plot days Vs. number of admission. We have tagged monthend adissions as “Yes” and rest as “No”.



Plotting difference of Mean and Median across two groups and printing their summary statistics.



## # A tibble: 2 x 5  
## monthend no mean\_ar median\_ar prop  
## <chr> <int> <dbl> <dbl> <dbl>  
## 1 No 24122 0.652 0.68 0.862  
## 2 Yes 3864 0.593 0.570 0.138

We see the propotion of admission during month end is 14 %. This is 4.2 times more admssion than a normal day.

#### Formulate hypothesies

Lets formulate our Hypothesis about Average realisation among two groups of individual. A group who gets admitted in monthend and the group who are admitted in other days of the month. We test on both on their difference of mean and median. From the table and distribution lpots above we see that there is a difference of both Mean and Median in two groups. Is the difference statistically significant. We state the formal hypothesis as below.

H0 <- There is no significance difference between mean and median average realisation of monthend admitted student represented as group Y and rest of the group represented as N i.e Nbar - Ybar = 0

Ha <- Nbar - Ybar > 0

#### Compute test statistic

The observed differences of mean and median value of the two group are tabulate below:

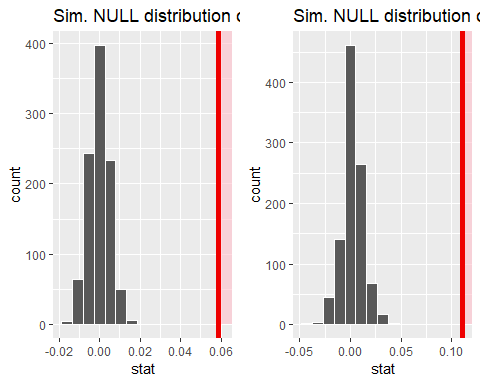
## # A tibble: 2 x 2  
## metric obs\_diff  
## <chr> <dbl>  
## 1 mean\_ar 0.0588  
## 2 median\_ar 0.11

#### Simulate NULL Distribution of mean and median.

We generate 1000 samples of the our data and shuffle the average realisation randomly to create a NULL distribution. This will generate 1000 random mean and median differences each. We plot the distribution and mark the observed actual difference. The observed difference lies very extreme end of the distribution in the plot below.

## # A tibble: 1 x 1  
## p\_value  
## <dbl>  
## 1 0

## # A tibble: 1 x 1  
## p\_value  
## <dbl>  
## 1 0



#### Test Conclusion

The pvalue for Mean 0 and pvalue of Median 0 is very extreme even if we define a strictest of confidence level of 99% as evident from the plot.

This gives a strong evidence that the differences is not by chance but there is some systematic influnce resulting in low realisation from monthend admissions.

Management should further investigate what could be the reason of such differences.