Midterm Project

Mengyun Li

November 29, 2017

Introduction

Data

```
#Remove the column with NA.
train <- read_csv("train.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   time = col_time(format = ""),
##   state = col_character(),
##   car_value = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
train <- na.omit(train)
head(train, n=9)
```

```
## # A tibble: 9 x 25
##   customer_ID shopping_pt record_type   day     time state location
##         <int>       <int>       <int> <int>   <time> <chr>    <int>
## 1    10000000           1           0     0 08:35:00    IN    10001
## 2    10000000           2           0     0 08:38:00    IN    10001
## 3    10000000           3           0     0 08:38:00    IN    10001
## 4    10000000           4           0     0 08:39:00    IN    10001
## 5    10000000           5           0     0 11:55:00    IN    10001
## 6    10000000           6           0     0 11:57:00    IN    10001
## 7    10000000           7           0     0 11:58:00    IN    10001
## 8    10000000           8           0     0 12:03:00    IN    10001
## 9    10000000           9           1     0 12:07:00    IN    10001
## # ... with 18 more variables: group_size <int>, homeowner <int>,
## #   car_age <int>, car_value <chr>, risk_factor <int>, age_oldest <int>,
## #   age_youngest <int>, married_couple <int>, C_previous <int>,
## #   duration_previous <int>, A <int>, B <int>, C <int>, D <int>, E <int>,
## #   F <int>, G <int>, cost <int>
```

As a customer shops an insurance policy, he/she will receive a number of quotes with different coverage options before purchasing a plan. From the data above, a customer with ID 10000000 recieved nine quote and purchased the last one.

```
#Select quotes which customer purchased
purchase <- train[train$record_type=="1", ]
head(purchase, n=3)
```

```
## # A tibble: 3 x 25
##   customer_ID shopping_pt record_type   day     time state location
```

```
##         <int>       <int>       <int> <int>   <time> <chr>    <int>
## 1    10000000           9           1     0 12:07:00    IN    10001
## 2    10000005           6           1     3 09:09:00    NY    10006
## 3    10000013           4           1     4 09:31:00    WV    10014
## # ... with 18 more variables: group_size <int>, homeowner <int>,
## #   car_age <int>, car_value <chr>, risk_factor <int>, age_oldest <int>,
## #   age_youngest <int>, married_couple <int>, C_previous <int>,
## #   duration_previous <int>, A <int>, B <int>, C <int>, D <int>, E <int>,
## #   F <int>, G <int>, cost <int>
```

```
ggplot(purchase, aes(x=shopping_pt)) +geom_bar()
```