Exploratory Data Analysis

Chindu

5/26/2019

This is a Expploratory Data Analysis report carried out on a sample CRM dataset

Loading csv file into R studio

```
data<-read.csv("ReadyforModelling.csv")
```

Checking if R studio has identified the right structure for each variable

```
str(data)
```

```
## 'data.frame':    908 obs. of  31 variables:
##  $ X                     : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Allocated.Time        : Factor w/ 3 levels "Extremely Fast",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Web.or.Phone          : Factor w/ 2 levels "PHONE","WEB": 2 1 1 1 1 1 1 1 1 1 ...
##  $ Answered.by.specialist: Factor w/ 2 levels "","Yes": 2 2 1 2 1 1 2 2 2 2 ...
##  $ Holiday.Type          : Factor w/ 4 levels "A","B","E","OTHER": 2 1 1 2 1 1 1 3 1 1 ...
##  $ Accom.type            : Factor w/ 4 levels "grade1","grade2",..: 2 1 2 2 1 1 2 4 2 2 ...
##  $ Dep.Airport           : Factor w/ 8 levels "Any Airport",..: 8 7 8 4 7 5 4 4 4 6 ...
##  $ Lead.Time             : int  48 26 27 47 27 62 56 14 85 44 ...
##  $ Destination           : Factor w/ 15 levels " AA Resort"," AB",..: 4 8 5 15 2 2 1 15 1 3 ...
##  $ Duration              : int  14 14 14 17 14 14 14 14 14 10 ...
##  $ Adults                : int  2 2 2 2 2 3 2 3 1 4 ...
##  $ Children              : int  0 0 2 2 2 2 3 0 1 0 ...
##  $ Transport.Type        : Factor w/ 3 levels "A","B","None Required": 2 1 2 1 1 2 2 1 2 3 ...
##  $ Answered.Q            : Factor w/ 2 levels "NO","YES": 1 2 2 2 1 2 1 2 2 1 ...
##  $ Notes.Completed       : Factor w/ 2 levels "NO","YES": 1 1 1 1 1 2 1 2 1 1 ...
##  $ Title                 : Factor w/ 5 levels "Dr","Miss","Mr",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ Enquiry.Comments      : Factor w/ 2 levels "NO","YES": 1 2 1 2 1 1 1 1 1 1 ...
##  $ Booked.Status         : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ EnquiryYear           : int  2017 2018 2019 2018 2017 2018 2018 2017 2019 2017 ...
##  $ EnquiryMonth          : int  5 11 1 9 9 1 4 10 1 5 ...
##  $ EnquiryDay            : int  29 4 2 21 18 15 22 29 1 21 ...
##  $ EnquiryWeekday        : Factor w/ 7 levels "Friday","Monday",..: 2 4 7 1 2 2 4 4 6 4 ...
##  $ DepYear               : int  2018 2019 2019 2019 2018 2019 2019 2018 2020 2018 ...
##  $ DepMonth              : int  4 5 7 8 3 3 5 2 8 3 ...
##  $ DepDay                : int  29 5 10 14 30 26 23 9 16 28 ...
##  $ DepWeekday            : Factor w/ 7 levels "Friday","Monday",..: 4 4 7 7 1 6 5 1 4 7 ...
##  $ Enquiry.Timecat       : Factor w/ 2 levels "Business_Hour",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Enquiry.Time_class    : Factor w/ 3 levels "afternoon","morning",..: 2 1 2 1 3 2 1 1 1 3 ...
##  $ DepartureSeason       : Factor w/ 4 levels "fall","spring",..: 2 2 3 3 2 2 2 4 3 2 ...
##  $ Hotkey                : Factor w/ 2 levels "","Yes": 2 2 1 2 1 1 2 2 2 2 ...
##  $ Gender                : Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 1 1 ...
```

Changing structure of wrongly assigned variables and remove variables unrealated to the analysis

```
data$Answered.by.specialist<- factor(data$Answered.by.specialist)
data$Booked.Status<- factor(data$Booked.Status)
data$EnquiryYear<-factor(data$EnquiryYear)
data$DepYear<-factor(data$DepYear)
data$X<-NULL
```

Get a better understanding of the entire dataset

```
summary(data)
```

```
##         Allocated.Time Web.or.Phone Answered.by.specialist Holiday.Type
##  Extremely Fast:240    PHONE:178       :439                A    :617
##  Fast           :121    WEB  :730    Yes:469                B    :129
##  Slow           :547                                       E    :102
##                                                            OTHER: 60
##
##
##
##   Accom.type       Dep.Airport    Lead.Time          Destination
##  grade1:326   MCH         :285   Min.   :  1.00    AC         :282
##  grade2:450   Lon All     :266   1st Qu.: 29.00    AB         :211
##  grade3: 50   Lon Gat     :129   Median : 47.00    JH Area    : 87
##  None  : 82   OTHER       : 59   Mean   : 48.68    AA Resort: 83
##               GG          : 58   3rd Qu.: 65.00    DC Drive : 48
##               Any Airport: 39   Max.   :121.00    OTHER      : 47
##               (Other)     : 72                     (Other)    :150
##     Duration         Adults         Children           Transport.Type
##  Min.   : 1.00   Min.   :1.000   Min.   :0.0000   A             :455
##  1st Qu.:13.00   1st Qu.:2.000   1st Qu.:0.0000   B             :237
##  Median :14.00   Median :3.000   Median :1.0000   None Required:216
##  Mean   :13.34   Mean   :3.305   Mean   :0.8744
##  3rd Qu.:14.00   3rd Qu.:4.000   3rd Qu.:2.0000
##  Max.   :28.00   Max.   :7.000   Max.   :5.0000
##
##  Answered.Q Notes.Completed  Title     Enquiry.Comments Booked.Status
##  NO :440    NO :652         Dr  :  4   NO :678          0:681
##  YES:468    YES:256         Miss:115   YES:230          1:227
##                             Mr  :367
##                             Mrs :382
##                             Ms  : 40
##
##
##  EnquiryYear  EnquiryMonth    EnquiryDay      EnquiryWeekday DepYear
##  2017:438     Min.   : 1.0   Min.   : 1.00   Friday   : 87    2017:118
##  2018:419     1st Qu.: 3.0   1st Qu.: 8.00   Monday   :137    2018:380
##  2019: 51     Median : 5.0   Median :15.50   Saturday :109    2019:355
##               Mean   : 5.6   Mean   :15.76   Sunday   :204    2020: 55
##               3rd Qu.: 8.0   3rd Qu.:23.25   Thursday :109
##               Max.   :12.0   Max.   :31.00   Tuesday  :129
##                                              Wednesday:133
##     DepMonth         DepDay          DepWeekday       Enquiry.Timecat
##  Min.   : 1.00   Min.   : 1.00   Friday   :121   Business_Hour:679
##  1st Qu.: 5.00   1st Qu.: 7.00   Monday   :141   Closed       :229
```

```
##  Median : 8.00   Median :15.00   Saturday :175
##  Mean   : 7.15   Mean   :15.13   Sunday   : 89
##  3rd Qu.: 9.00   3rd Qu.:22.00   Thursday :113
##  Max.   :12.00   Max.   :31.00   Tuesday  :126
##                                  Wednesday:143
##  Enquiry.Time_class DepartureSeason Hotkey    Gender
##  afternoon:289      fall  :238         :439   F:537
##  morning  :531      spring:210      Yes:469   M:371
##  night    : 88      summer:386
##                     winter: 74
##
##
##
```

Get a better understanding of numeric/interger variables

```
diagnose_numeric(data)
```

```
## # A tibble: 8 x 10
##   variables      min    Q1   mean median    Q3   max  zero minus outlier
##   <chr>        <int> <dbl>  <dbl>  <dbl> <dbl> <int> <int> <int>   <int>
## 1 Lead.Time        1    29 48.7     47    65     121     0     0       4
## 2 Duration         1    13 13.3     14    14      28     0     0     290
## 3 Adults           1     2  3.31     3     4       7     0     0       0
## 4 Children         0     0  0.874    1     2       5   437     0       0
## 5 EnquiryMonth     1     3  5.60     5     8      12     0     0       0
## 6 EnquiryDay       1     8 15.8     15.5  23.2    31     0     0       0
## 7 DepMonth         1     5  7.15     8     9      12     0     0       0
## 8 DepDay           1     7 15.1     15    22      31     0     0       0
```

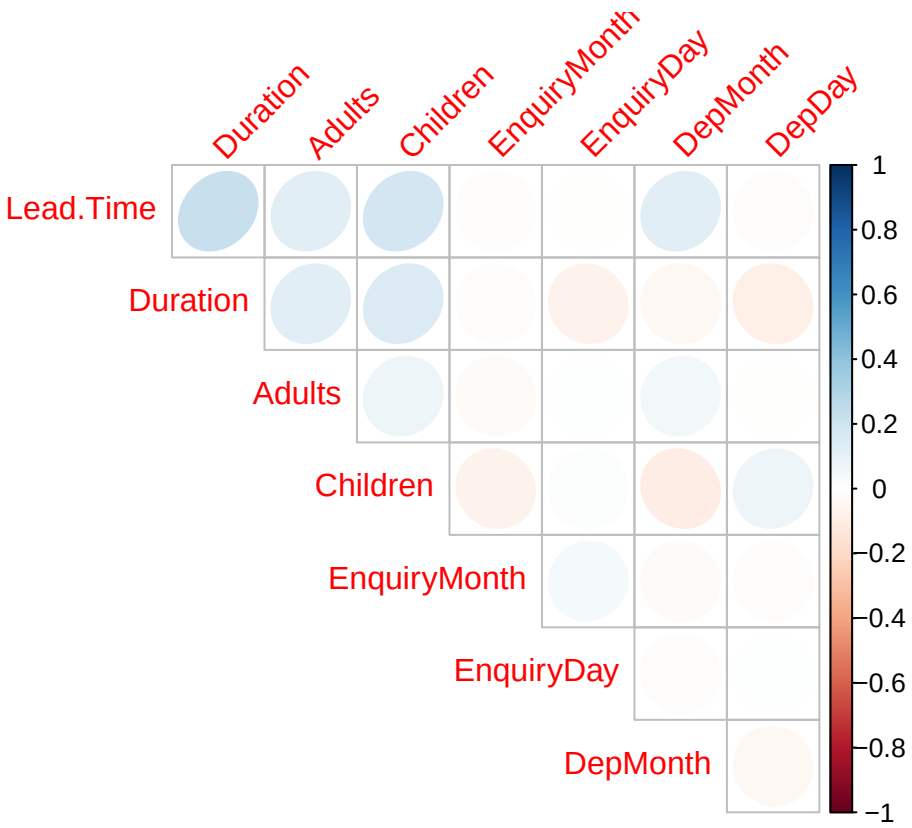
Get a better understanding of categorical variables

```
diagnose_category(data)
```

```
## # A tibble: 85 x 6
##    variables              levels             N  freq ratio  rank
##    <chr>                  <fct>          <int> <int> <dbl> <int>
##  1 Allocated.Time         Slow             908   547  60.2     1
##  2 Allocated.Time         Extremely Fast   908   240  26.4     2
##  3 Allocated.Time         Fast             908   121  13.3     3
##  4 Web.or.Phone           WEB              908   730  80.4     1
##  5 Web.or.Phone           PHONE            908   178  19.6     2
##  6 Answered.by.specialist Yes              908   469  51.7     1
##  7 Answered.by.specialist ""               908   439  48.3     2
##  8 Holiday.Type           A                908   617  68.0     1
##  9 Holiday.Type           B                908   129  14.2     2
## 10 Holiday.Type           E                908   102  11.2     3
## # ... with 75 more rows
```

Checking correlation between numerical variables

```
plot_correlate(data)
```



Exploring relation between target varaible (Booked.Status) and a numeric variable

```
categ<-target_by(data,Booked.Status)
cat_num<-relate(categ,Duration)
cat_num
```

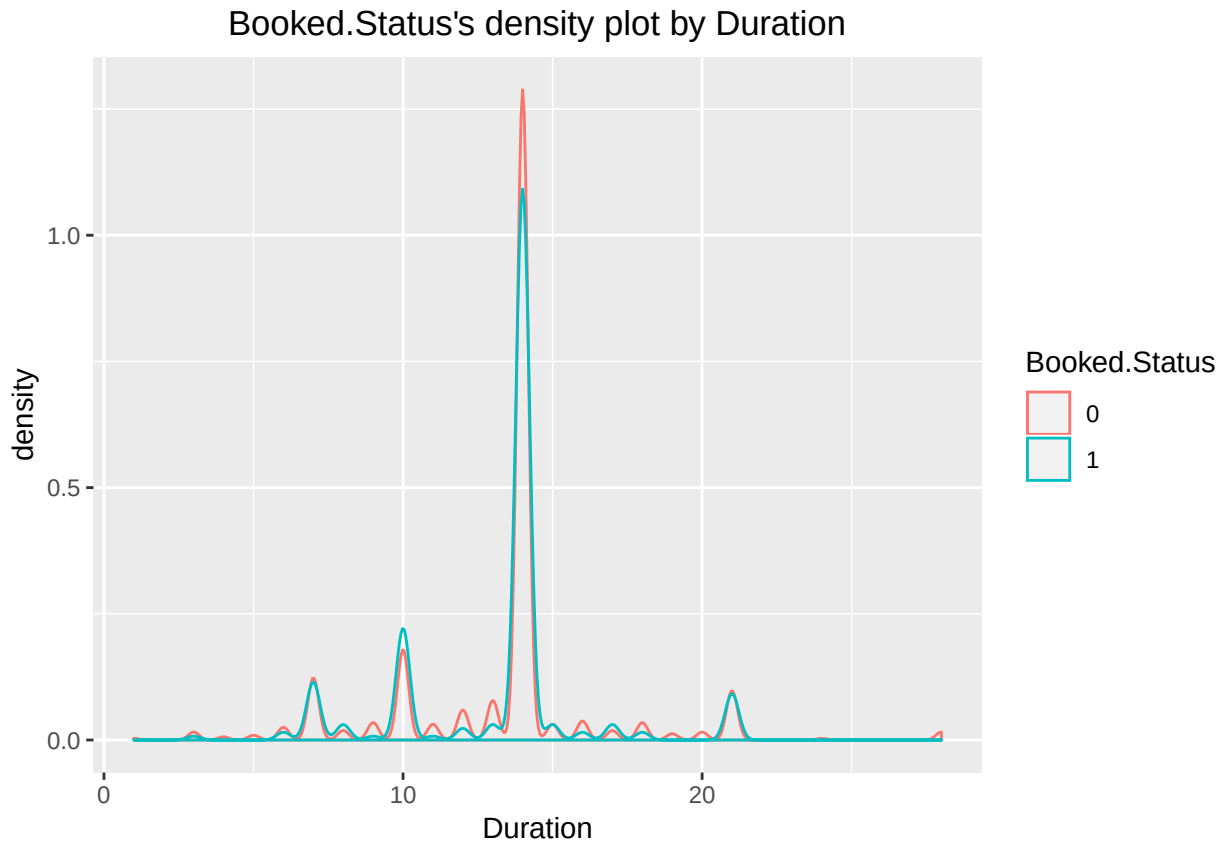
```
## # A tibble: 3 x 27
##   variable Booked.Status     n    na  mean    sd se_mean   IQR skewness
##   <chr>    <fct>         <dbl> <dbl> <dbl> <dbl>   <dbl> <dbl>    <dbl>
## 1 Duration 0               681     0  13.4  3.46   0.133     1    0.210
## 2 Duration 1               227     0  13.2  3.10   0.206     1   -0.122
## 3 Duration total           908     0  13.3  3.37   0.112     1    0.152
## # ... with 18 more variables: kurtosis <dbl>, 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>
```

```
summary(cat_num)
```

```
##    variable         Booked.Status       n               na
##  Length:3           0    :1       Min.   :227.0   Min.   :0
##  Class :character   1    :1       1st Qu.:454.0   1st Qu.:0
##  Mode  :character   total:1       Median :681.0   Median :0
```

```
##                                  Mean   :605.3   Mean   :0
##                                  3rd Qu.:794.5   3rd Qu.:0
##                                  Max.   :908.0   Max.   :0
##       mean             sd           se_mean            IQR
##  Min.   :13.22   Min.   :3.099   Min.   :0.1119   Min.   :1
##  1st Qu.:13.28   1st Qu.:3.236   1st Qu.:0.1223   1st Qu.:1
##  Median :13.34   Median :3.372   Median :0.1326   Median :1
##  Mean   :13.31   Mean   :3.310   Mean   :0.1501   Mean   :1
##  3rd Qu.:13.36   3rd Qu.:3.416   3rd Qu.:0.1691   3rd Qu.:1
##  Max.   :13.38   Max.   :3.460   Max.   :0.2057   Max.   :1
##     skewness           kurtosis          p00             p01
##  Min.   :-0.12192   Min.   :1.456   Min.   :1.000   Min.   :4.000
##  1st Qu.: 0.01503   1st Qu.:2.123   1st Qu.:1.000   1st Qu.:4.500
##  Median : 0.15197   Median :2.789   Median :1.000   Median :5.000
##  Mean   : 0.07988   Mean   :2.422   Mean   :1.667   Mean   :5.087
##  3rd Qu.: 0.18078   3rd Qu.:2.906   3rd Qu.:2.000   3rd Qu.:5.630
##  Max.   : 0.20958   Max.   :3.022   Max.   :3.000   Max.   :6.260
##       p05          p10           p20             p25          p30
##  Min.   :7   Min.   :9.0   Min.   :10.00   Min.   :13   Min.   :14
##  1st Qu.:7   1st Qu.:9.0   1st Qu.:10.00   1st Qu.:13   1st Qu.:14
##  Median :7   Median :9.0   Median :10.00   Median :13   Median :14
##  Mean   :7   Mean   :9.2   Mean   :10.33   Mean   :13   Mean   :14
##  3rd Qu.:7   3rd Qu.:9.3   3rd Qu.:10.50   3rd Qu.:13   3rd Qu.:14
##  Max.   :7   Max.   :9.6   Max.   :11.00   Max.   :13   Max.   :14
##       p40          p50          p60          p70          p75
##  Min.   :14   Min.   :14   Min.   :14   Min.   :14   Min.   :14
##  1st Qu.:14   1st Qu.:14   1st Qu.:14   1st Qu.:14   1st Qu.:14
##  Median :14   Median :14   Median :14   Median :14   Median :14
##  Mean   :14   Mean   :14   Mean   :14   Mean   :14   Mean   :14
##  3rd Qu.:14   3rd Qu.:14   3rd Qu.:14   3rd Qu.:14   3rd Qu.:14
##  Max.   :14   Max.   :14   Max.   :14   Max.   :14   Max.   :14
##       p80          p90             p95             p99          p100
##  Min.   :14   Min.   :15.00   Min.   :20.10   Min.   :21   Min.   :21.00
##  1st Qu.:14   1st Qu.:15.50   1st Qu.:20.55   1st Qu.:21   1st Qu.:24.50
##  Median :14   Median :16.00   Median :21.00   Median :21   Median :28.00
##  Mean   :14   Mean   :15.67   Mean   :20.70   Mean   :21   Mean   :25.67
##  3rd Qu.:14   3rd Qu.:16.00   3rd Qu.:21.00   3rd Qu.:21   3rd Qu.:28.00
##  Max.   :14   Max.   :16.00   Max.   :21.00   Max.   :21   Max.   :28.00
```

```
plot(cat_num) # relationship between booked.status and duration is represented using a desity plot
```

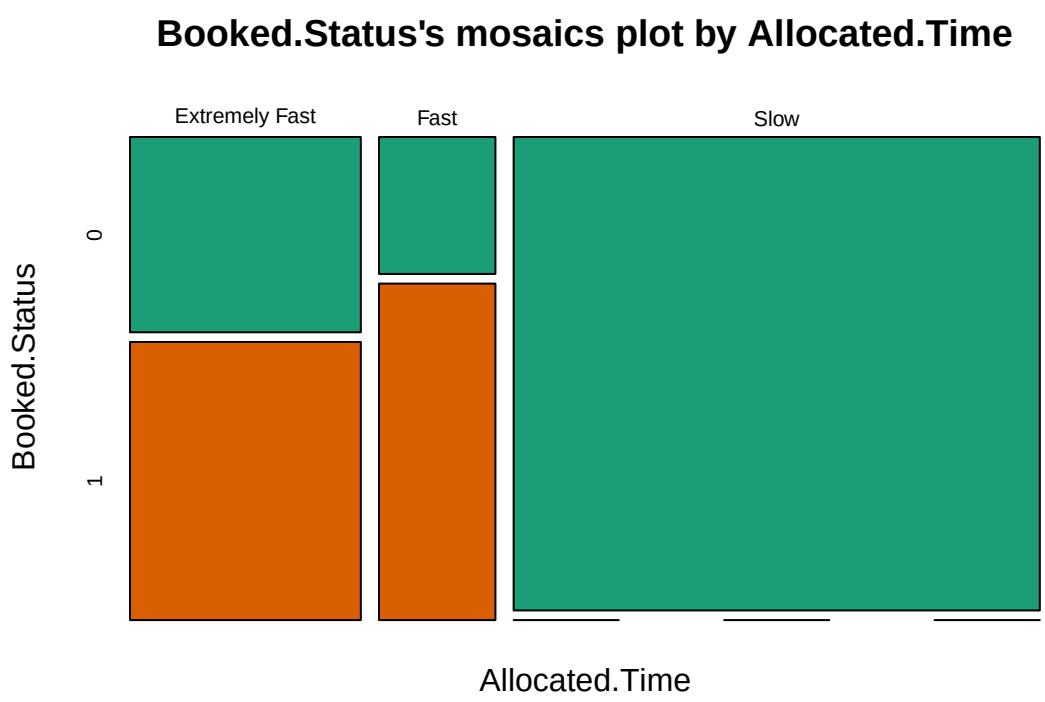


Exploring relation between target variable(BookedStatus) and a categorical variable

```
cat_cat<-relate(categ,Allocated.Time)
cat_cat
```

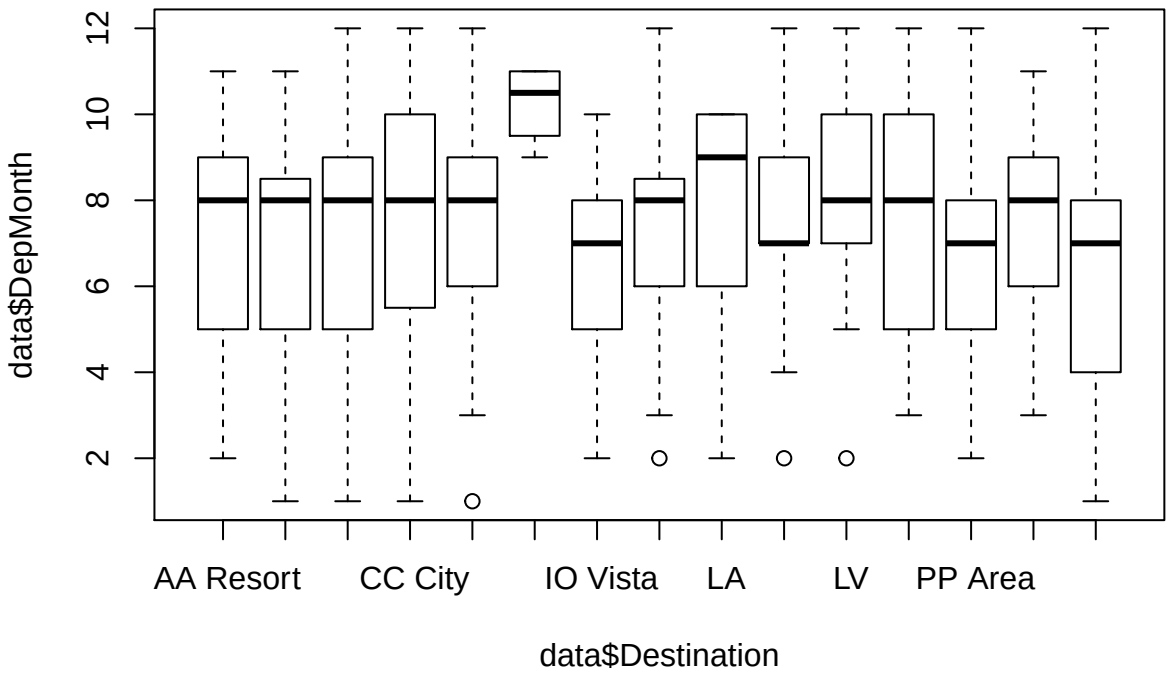
```
##               Allocated.Time
## Booked.Status Extremely Fast Fast Slow
##             0              99   35  547
##             1             141   86    0
```

```
plot(cat_cat) #mosaics plot
```



Understanding

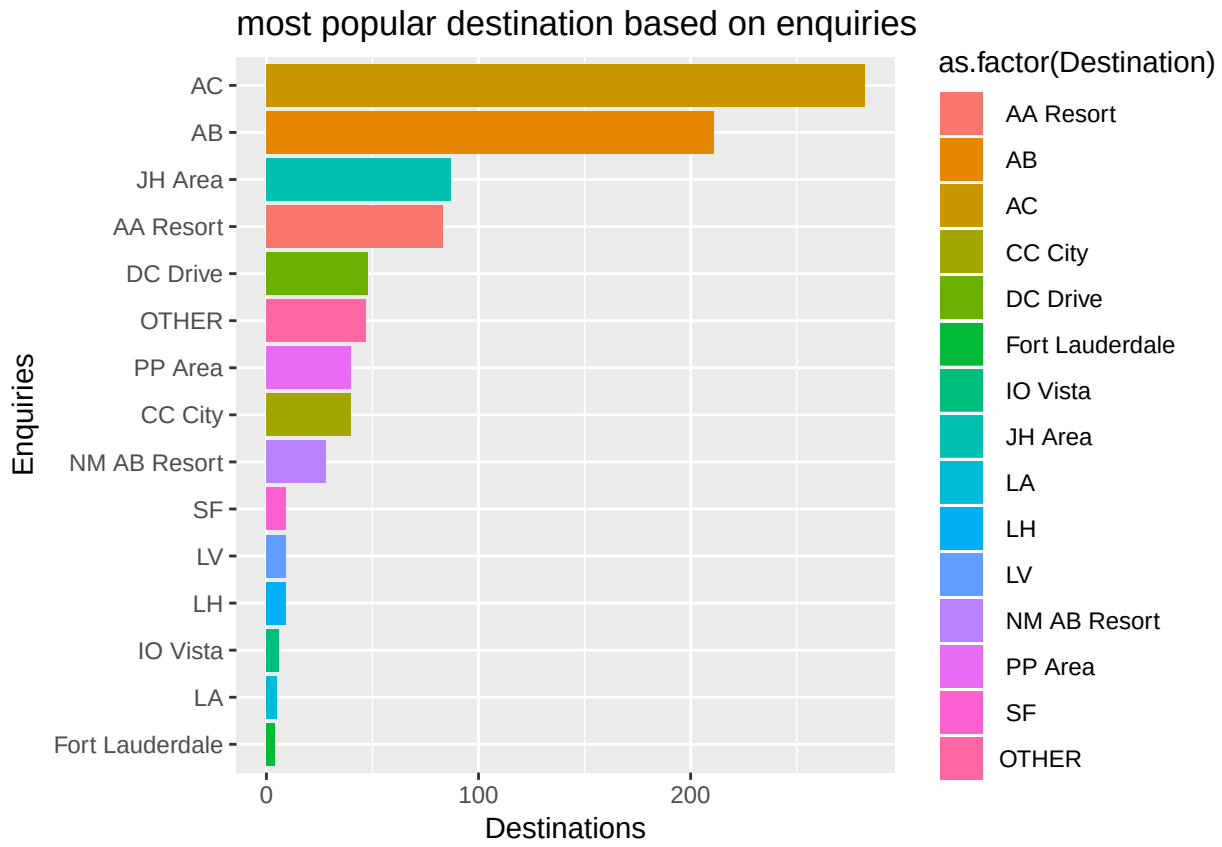
```
boxplot(data$DepMonth~data$Destination)
```



Desination by popularity and what is the total enquiries for each destination?

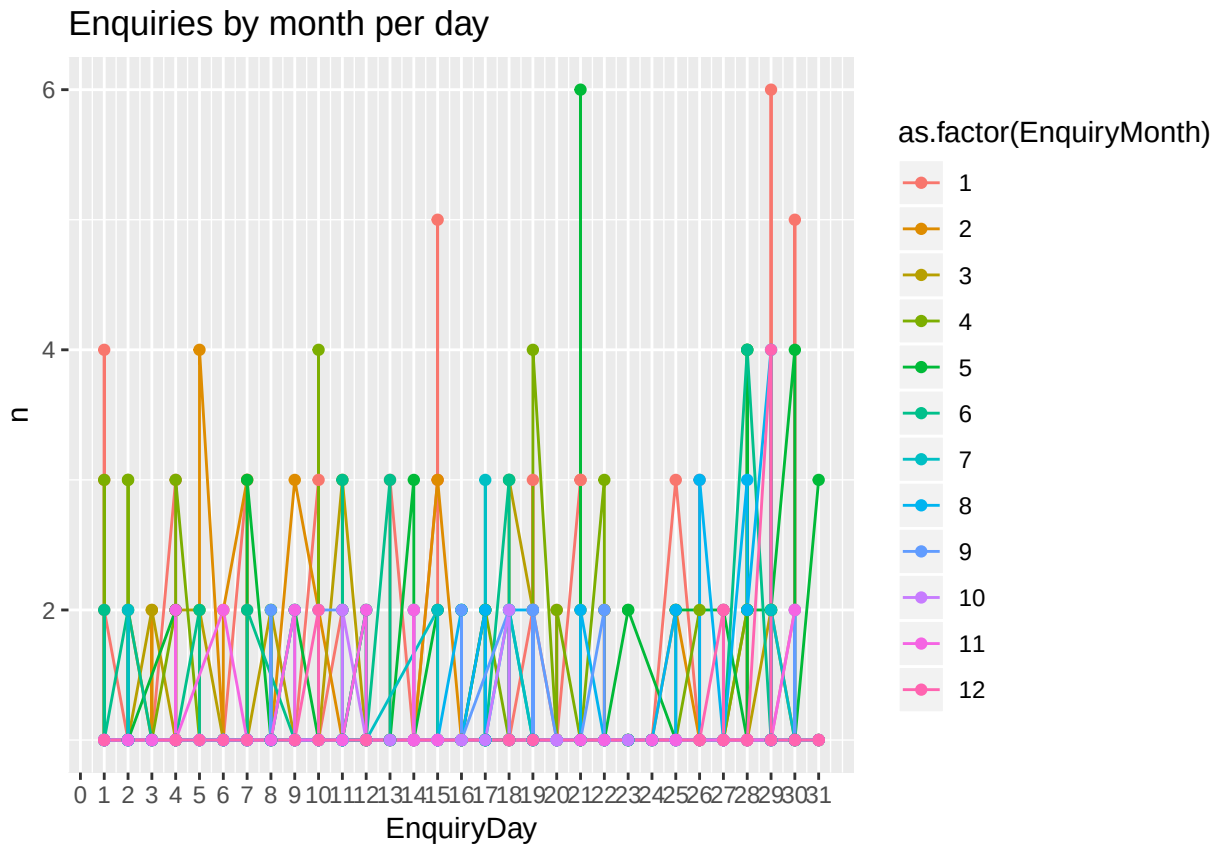
```
pop_destination<- data %>% group_by(Destination) %>% count(Destination) %>%ungroup()

ggplot(data=pop_destination,aes(x=reorder(as.factor(Destination),n),y=n,fill=as.factor(Destination)))+g
  labs(title= "most popular destination based on enquiries", x="Enquiries",y="Destinations")
```

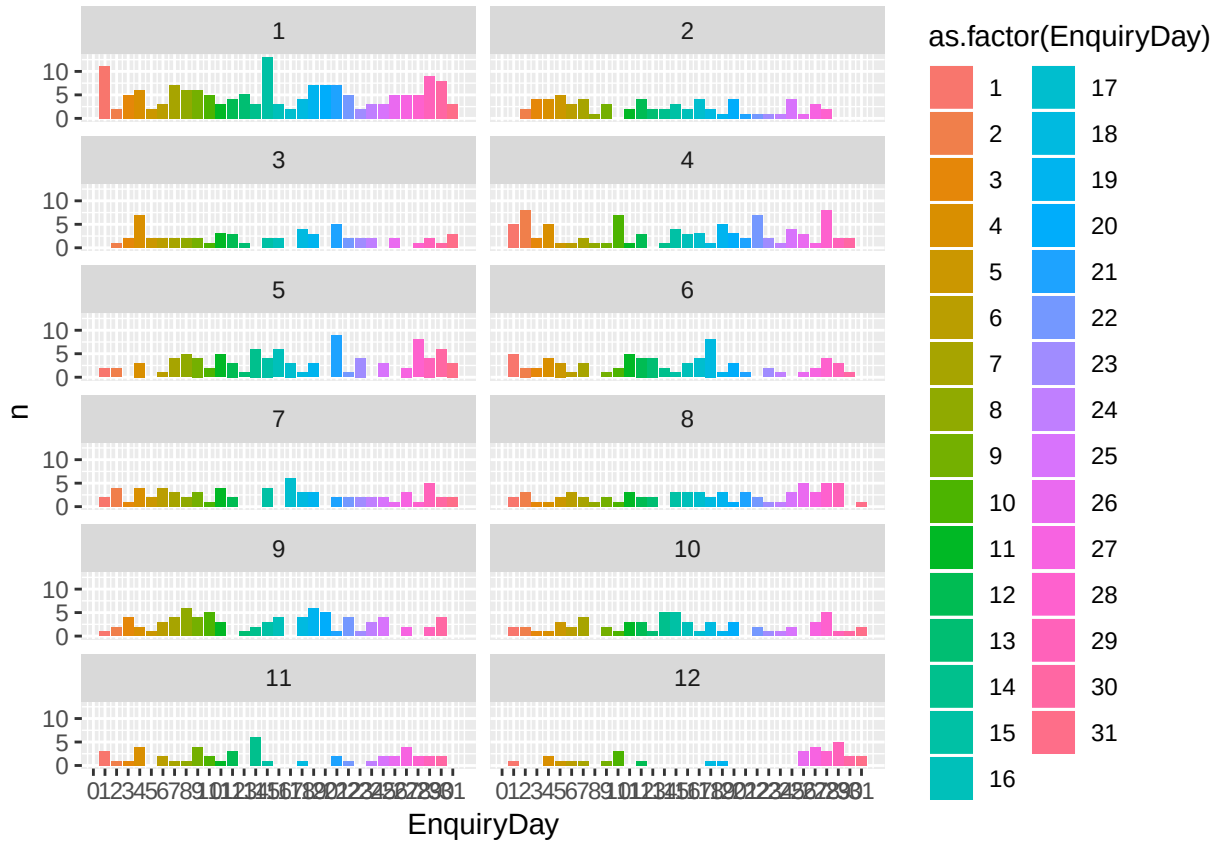


What are the day and month wise total enquiries?

```
day_month_sale<- data%>%group_by(EnquiryMonth,EnquiryDay)  %>% count(Destination)%>%arrange(EnquiryMonth
ggplot(data=day_month_sale,aes(x=EnquiryDay,y=n,group=EnquiryMonth,color=as.factor(EnquiryMonth)))+geom_
  labs(title="Enquiries by month per day")
```

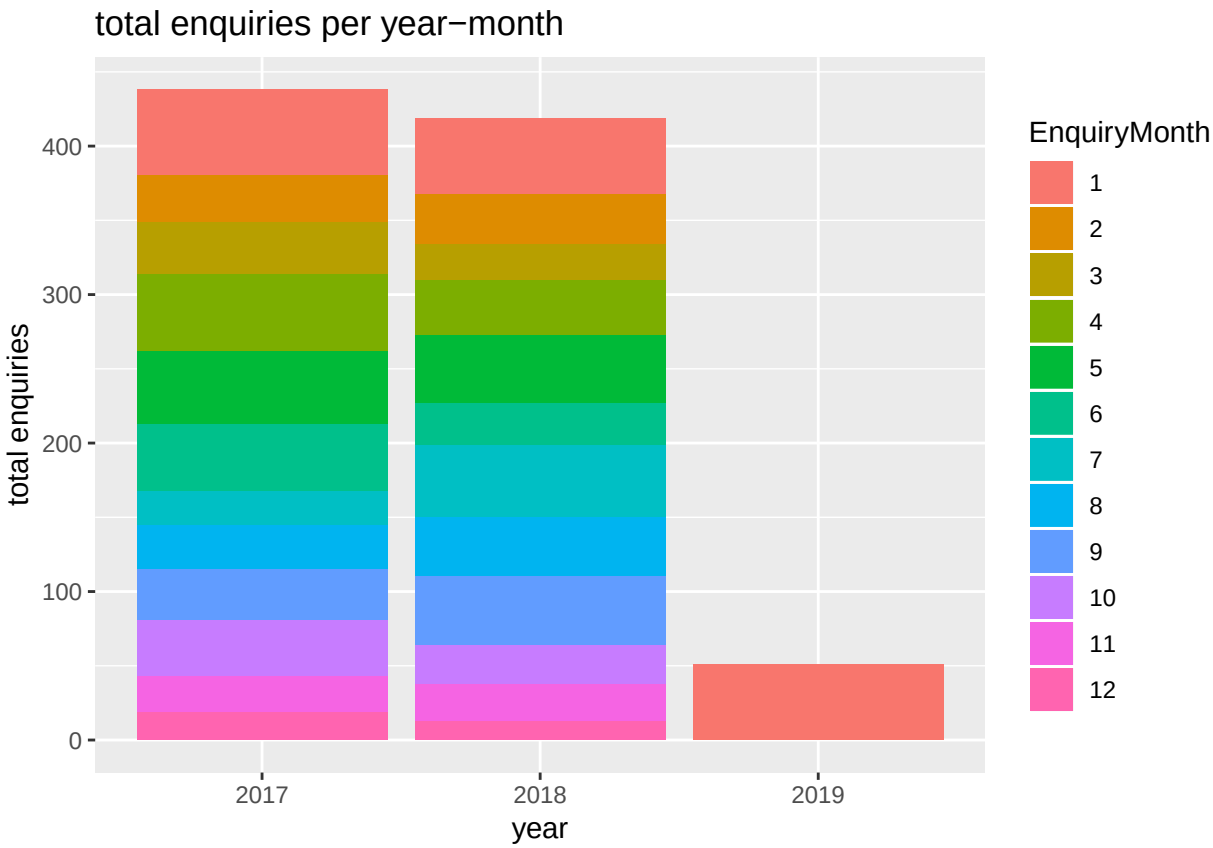



```
ggplot(data=day_month_sale, aes(x=EnquiryDay,y=n,fill=as.factor(EnquiryDay)))+geom_bar(stat="identity")+
```



Total enquiries by year and month

```
year_month<- data%>%group_by(EnquiryYear,EnquiryMonth) %>% count(Destination)%>%arrange(EnquiryYear)%>%
ggplot(data=year_month,aes(x=EnquiryYear,y=n,fill=as.factor(EnquiryMonth)))+ geom_bar(stat="identity")+
  labs(title="total enquiries per year-month",x="year",y="total enquiries",fill="EnquiryMonth")
```



How many enquiries were booked based on Allocated.Time

```
data$Booked.Status<-as.integer(data$Booked.Status)
booked_Allocated<-data%>%group_by(Allocated.Time)%>% summarise(booked=sum(Booked.Status)) %>%arrange(Al
```