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$Children<-factor(data$Children)
data$Adults<-factor(data$Adults)
data$X<-NULL
```

Get a better understanding of numeric/interger variables

```
diagnose_numeric(data)
```

```
## # A tibble: 6 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 EnquiryMonth     1     3  5.60    5     8      12     0     0       0
## 4 EnquiryDay       1     8 15.8    15.5  23.2    31     0     0       0
## 5 DepMonth         1     5  7.15    8     9      12     0     0       0
## 6 DepDay           1     7 15.1    15    22      31     0     0       0
```

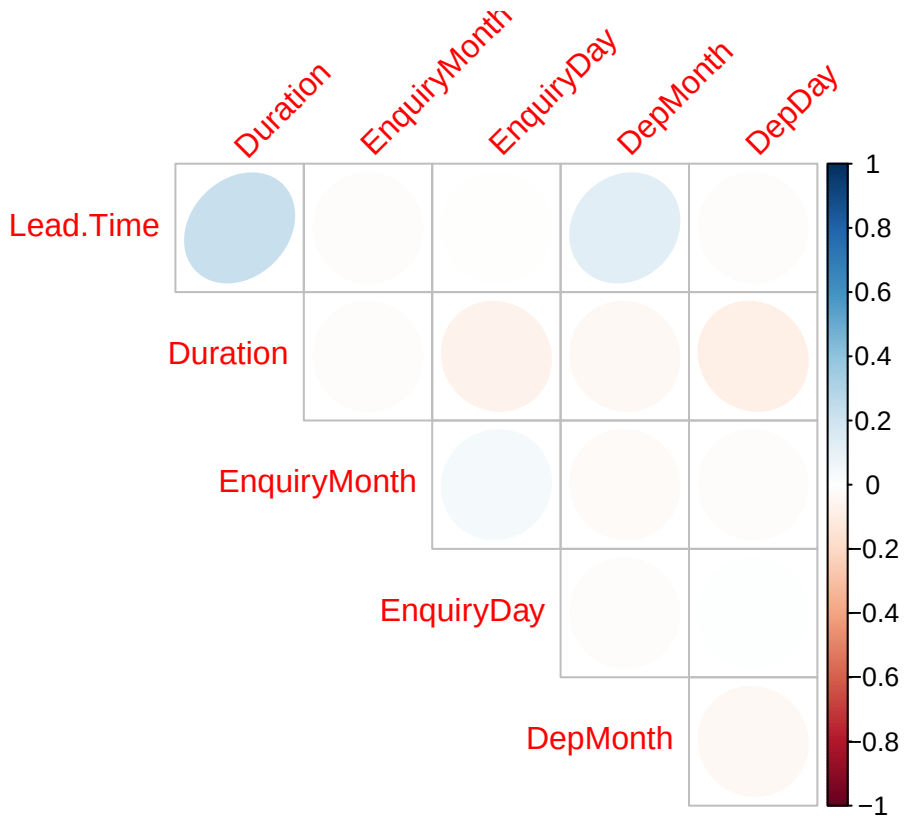
Get a better understanding of categorical variables

```
diagnose_category(data)
```

```
## # A tibble: 98 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 88 more rows
```

Checking correlation between numerical variables (fast plot)

```
plot_correlate(data)
```



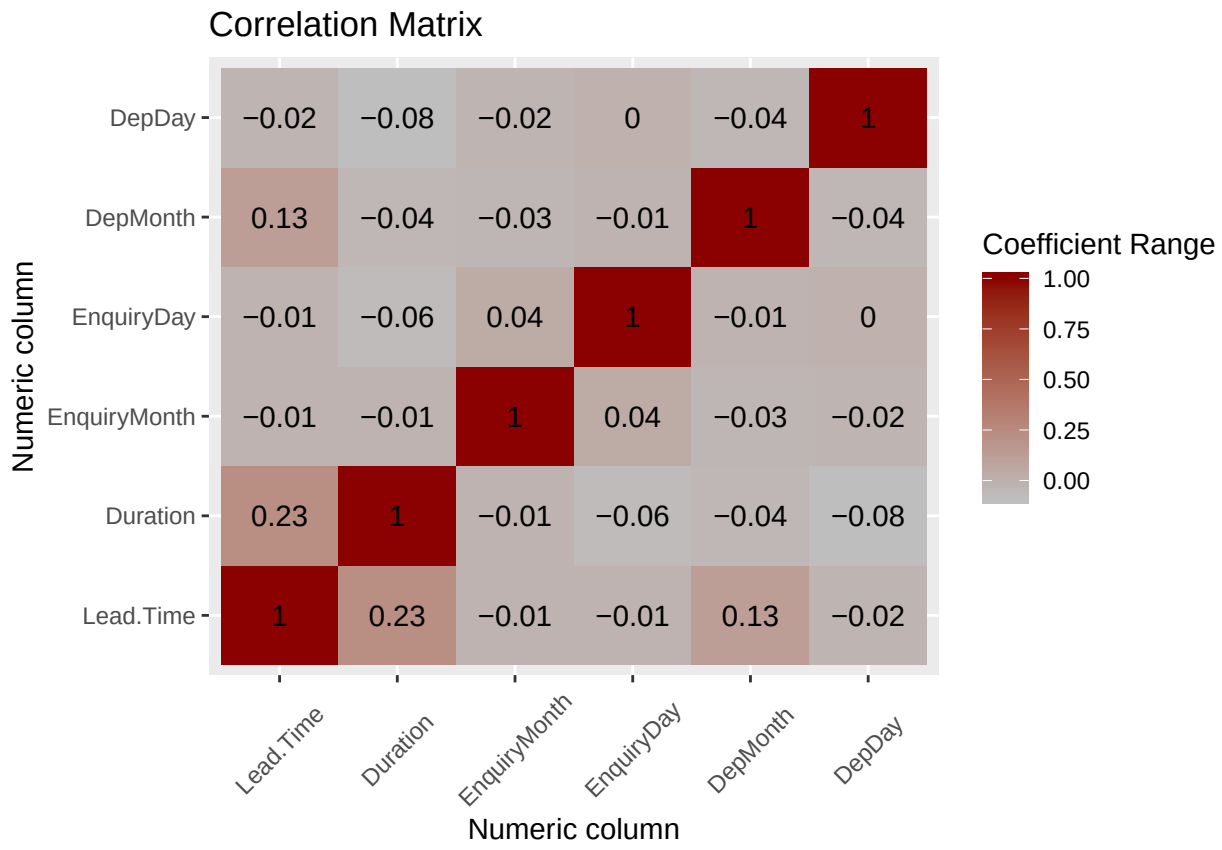
Detailed correlation plot

```
num.cols<-sapply(data,is.numeric)
data_numcols<-data[,num.cols]
cor(data_numcols)
```

```
##                 Lead.Time    Duration EnquiryMonth    EnquiryDay
## Lead.Time     1.000000000  0.22814624  -0.01184910 -0.0094899807
## Duration      0.228146245  1.00000000  -0.01011224 -0.0604599794
## EnquiryMonth -0.011849097 -0.01011224   1.00000000  0.0413274774
## EnquiryDay   -0.009489981 -0.06045998   0.04132748  1.0000000000
## DepMonth      0.129255970 -0.03771055  -0.02901795 -0.0132897607
## DepDay       -0.019142657 -0.08285208  -0.01886535  0.0004330459
##                 DepMonth        DepDay
## Lead.Time     0.12925597 -0.0191426570
## Duration     -0.03771055 -0.0828520793
## EnquiryMonth -0.02901795 -0.0188653477
## EnquiryDay   -0.01328976  0.0004330459
## DepMonth      1.00000000 -0.0385101707
## DepDay       -0.03851017  1.0000000000
```

```
melted_corr<-melt(cor(data_numcols))
ggplot(data=melted_corr,aes(x=Var1,y=Var2,fill=value))+
  geom_tile()+
  scale_fill_gradient(low="grey",high="darkred")+
  geom_text(aes(x=Var1,y=Var2,label=round(value,2)),size=4)+
```

```
  labs(title="Correlation Matrix",x="Numeric column",y="Numeric column",fill="Coefficient Range")+
  theme(axis.text.x=element_text(angle=45, vjust=0.5))
```



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

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

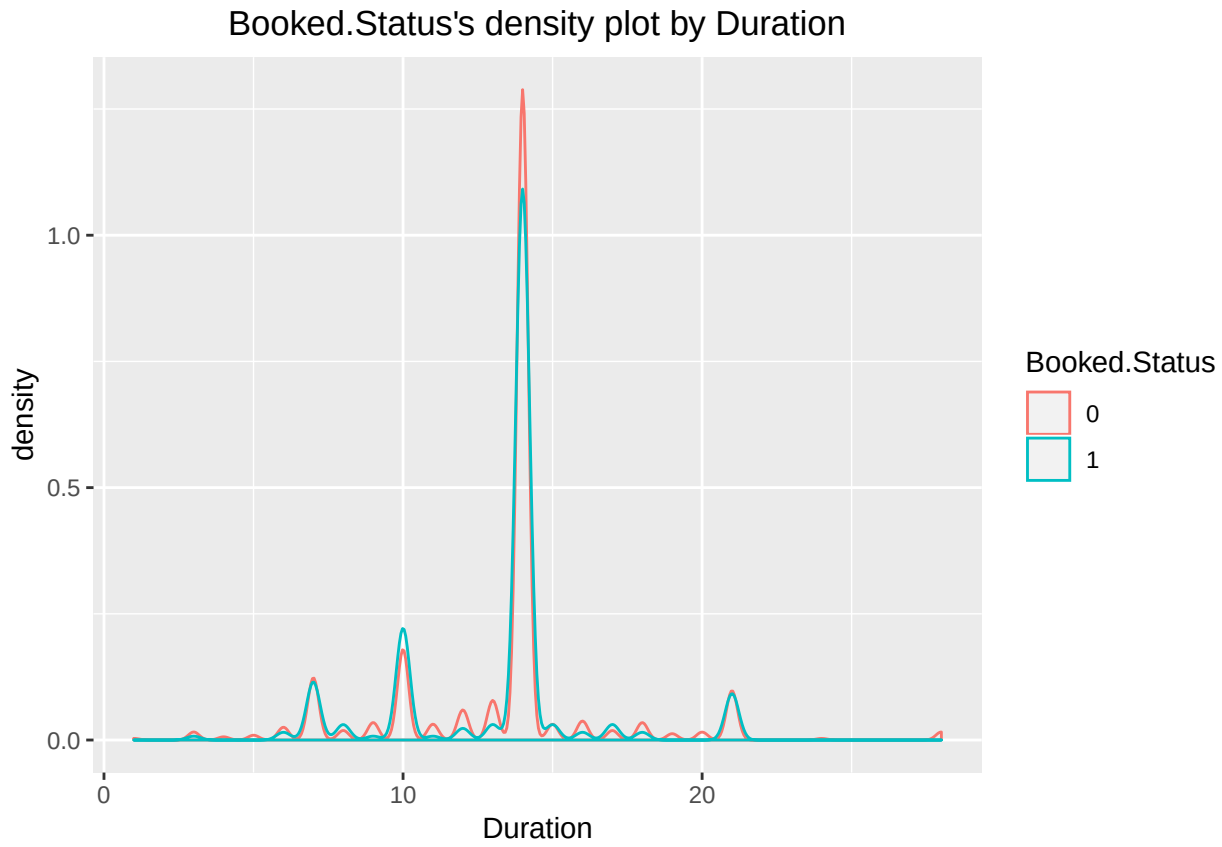
```
## # A tibble: 3 x 27
##   variable Booked.Status     n    na  mean    sd se_mean   IQR skewness
##   <chr>    <fct>         <dbl> <dbl> <dbl> <dbl>   <dbl> <dbl>    <dbl>
## 1 Duration 0               681     0  13.4  3.46   0.133     1    0.210
## 2 Duration 1               227     0  13.2  3.10   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
```

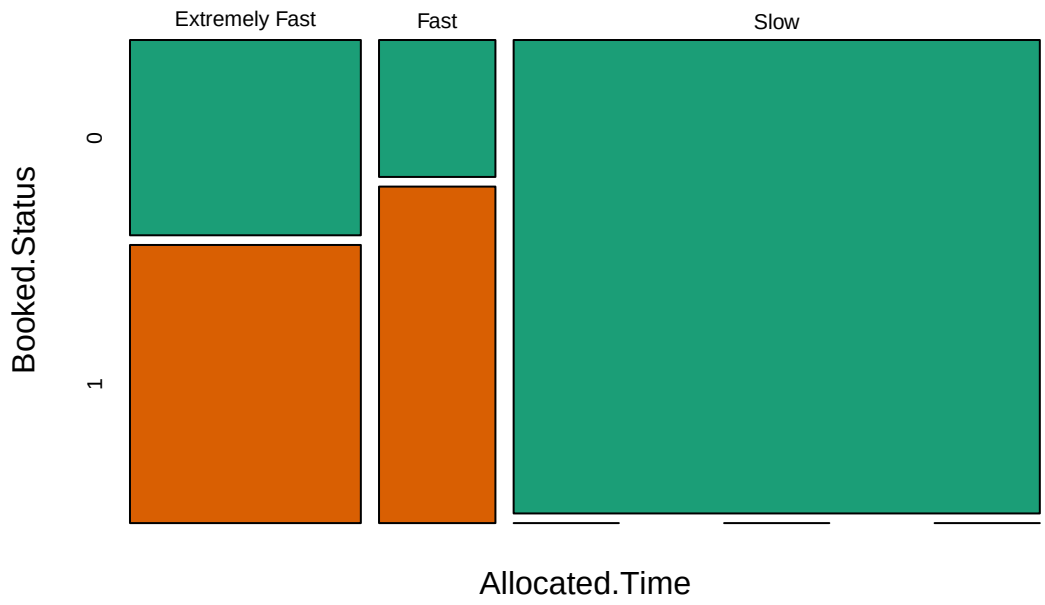


Exploring relation between target variable(BookedStatus) and a categorical variable

```
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
```



Checking for skewness in numeric variables (If skewness value lies above +1 or below -1, data is highly skewed. If it lies between +0.5 to -0.5, it is moderately skewed. If the value is 0, then the data is symmetric)

```
data %>%
  describe() %>%
  select(variable, skewness) %>%
  filter(!is.na(skewness)) %>%
  arrange(desc(abs(skewness)))
```

```
## # A tibble: 6 x 2
##   variable     skewness
##   <chr>           <dbl>
## 1 Lead.Time      0.415
## 2 DepMonth      -0.374
## 3 EnquiryMonth   0.170
## 4 Duration       0.152
## 5 DepDay         0.0809
## 6 EnquiryDay     0.0362
```

Lead.Time is highly skewed. To reduce the skewness and to achive a distribution that is close to a normal distribution, a sqrt transformation is used.

```
data$sqrt_lead.time<-sqrt(data$Lead.Time)

data %>%
```

```
  describe() %>%
  select(variable, skewness) %>%
  filter(!is.na(skewness)) %>%
  arrange(desc(abs(skewness)))
```

```
## # A tibble: 7 x 2
##   variable       skewness
##   <chr>             <dbl>
## 1 Lead.Time         0.415
## 2 DepMonth         -0.374
## 3 sqrt_lead.time   -0.284
## 4 EnquiryMonth      0.170
## 5 Duration          0.152
## 6 DepDay            0.0809
## 7 EnquiryDay        0.0362
```

The skewness is now reduced.

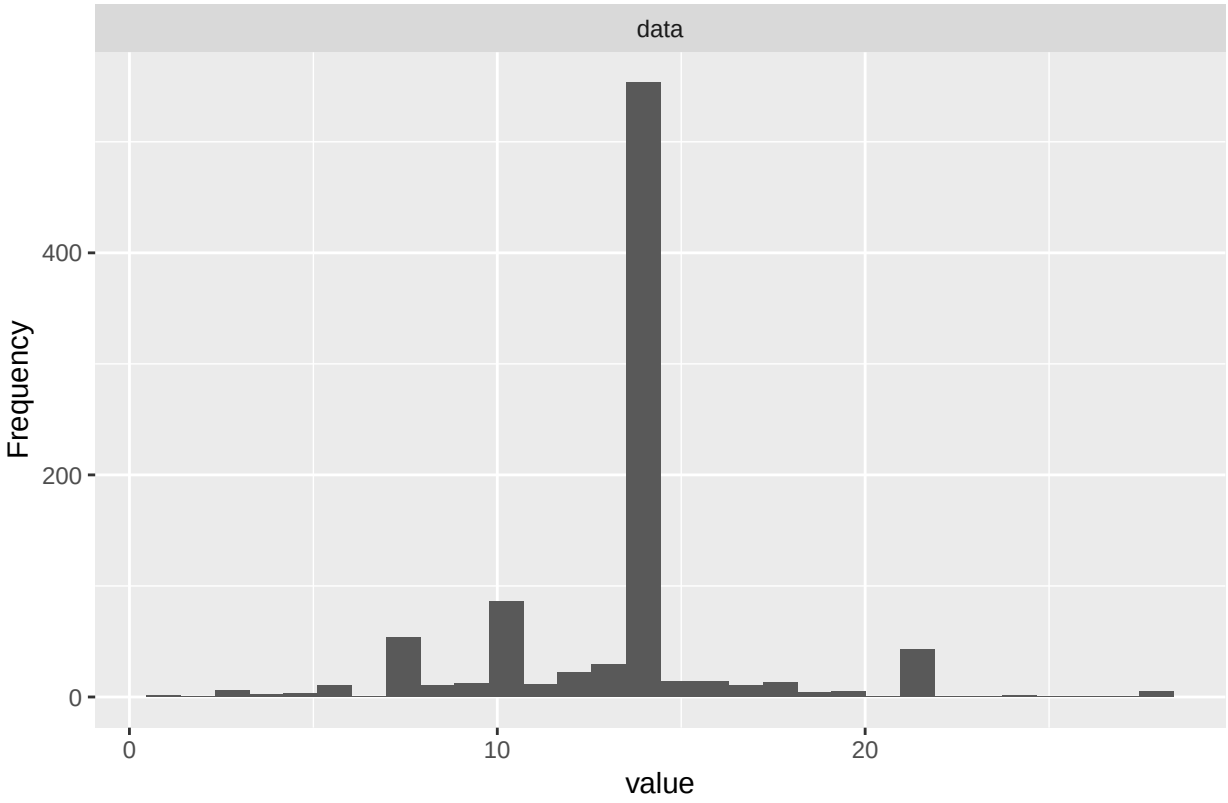
Diagnose anomalies of all numeric variables of data

```
diagnose_outlier(data)
```

```
##        variables outliers_cnt outliers_ratio outliers_mean with_mean
## 1      Lead.Time            4      0.4405286     120.25000 48.677313
## 2       Duration          290     31.9383260      12.13793 13.340308
## 3   EnquiryMonth            0      0.0000000           NaN  5.600220
## 4     EnquiryDay            0      0.0000000           NaN 15.759912
## 5       DepMonth            0      0.0000000           NaN  7.149780
## 6         DepDay            0      0.0000000           NaN 15.129956
## 7 sqrt_lead.time            2      0.2202643       1.00000  6.706558
##   without_mean
## 1    48.360619
## 2    13.904531
## 3     5.600220
## 4    15.759912
## 5     7.149780
## 6    15.129956
## 7     6.719156
```

The variable duration has approximately 32% observations identified as outliers

```
plot_histogram(data$Duration)
```



From the plot it is observed that the high skewness is due to majority of enquiries are for 7,10,14 or 21 days. tabulate values to get a better understanding.

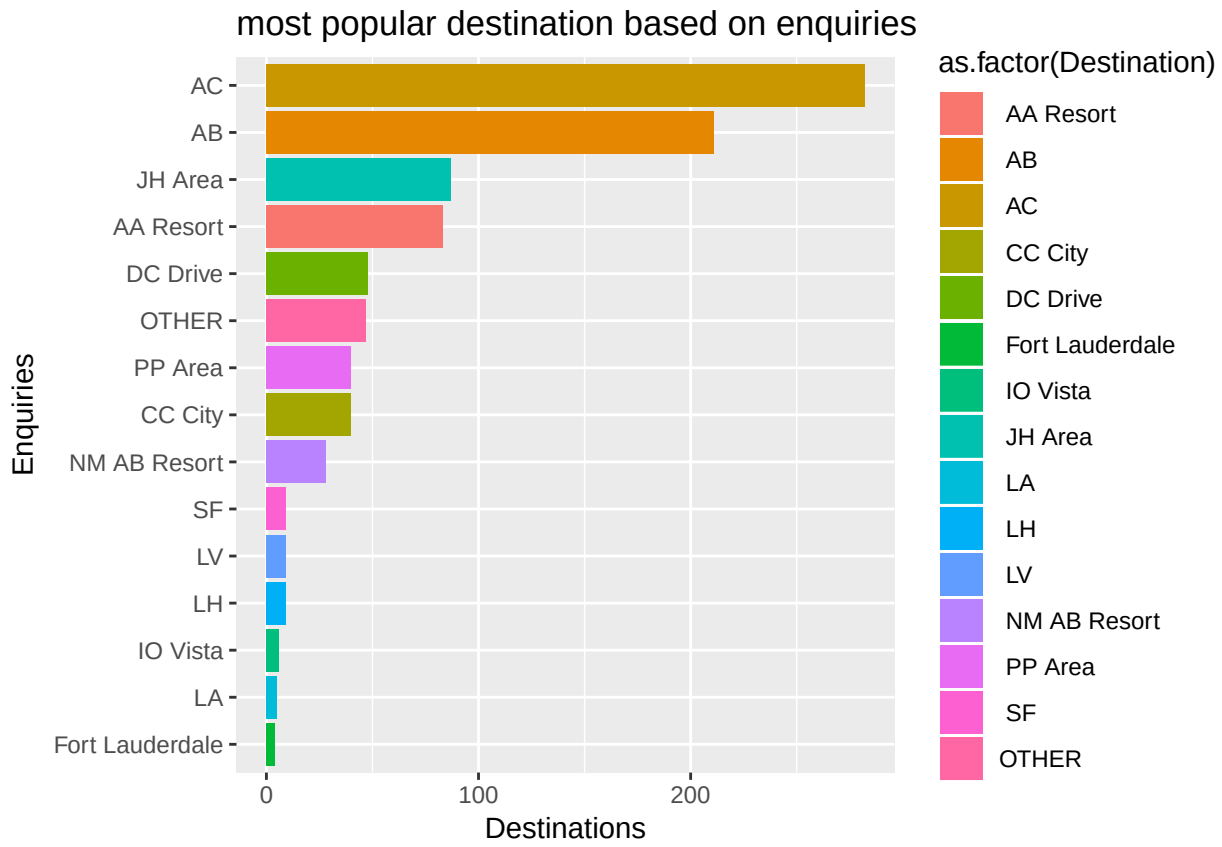
```
table(data$Duration)
```

```
##
##   1   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19
##   1   6   2   3  10  54  10  12  86  11  22  29 553  14  14  10  13   4
##  20  21  24  28
##   5  43   1   5
```

Answering questions using data visualisation techniques

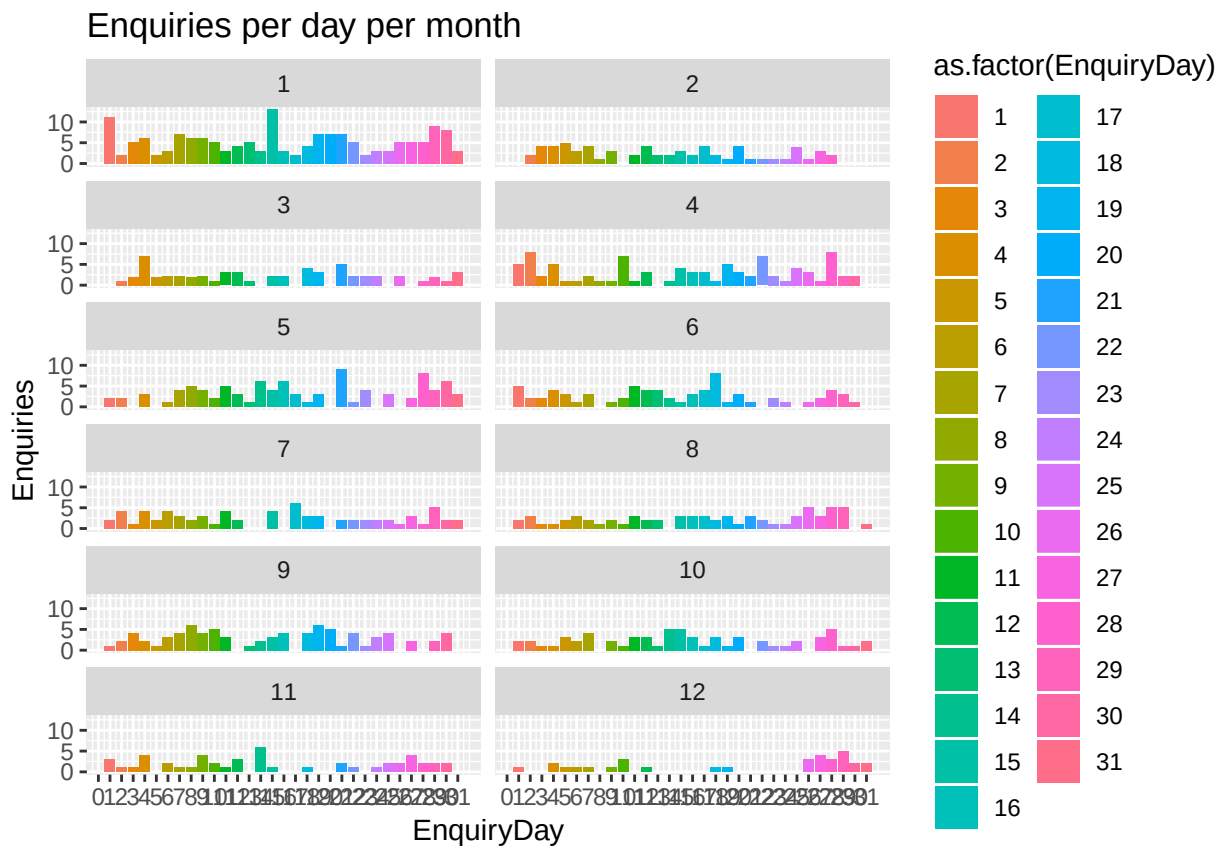
Desination by popularity and what is the total enquiries for each destination?

```
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")
```



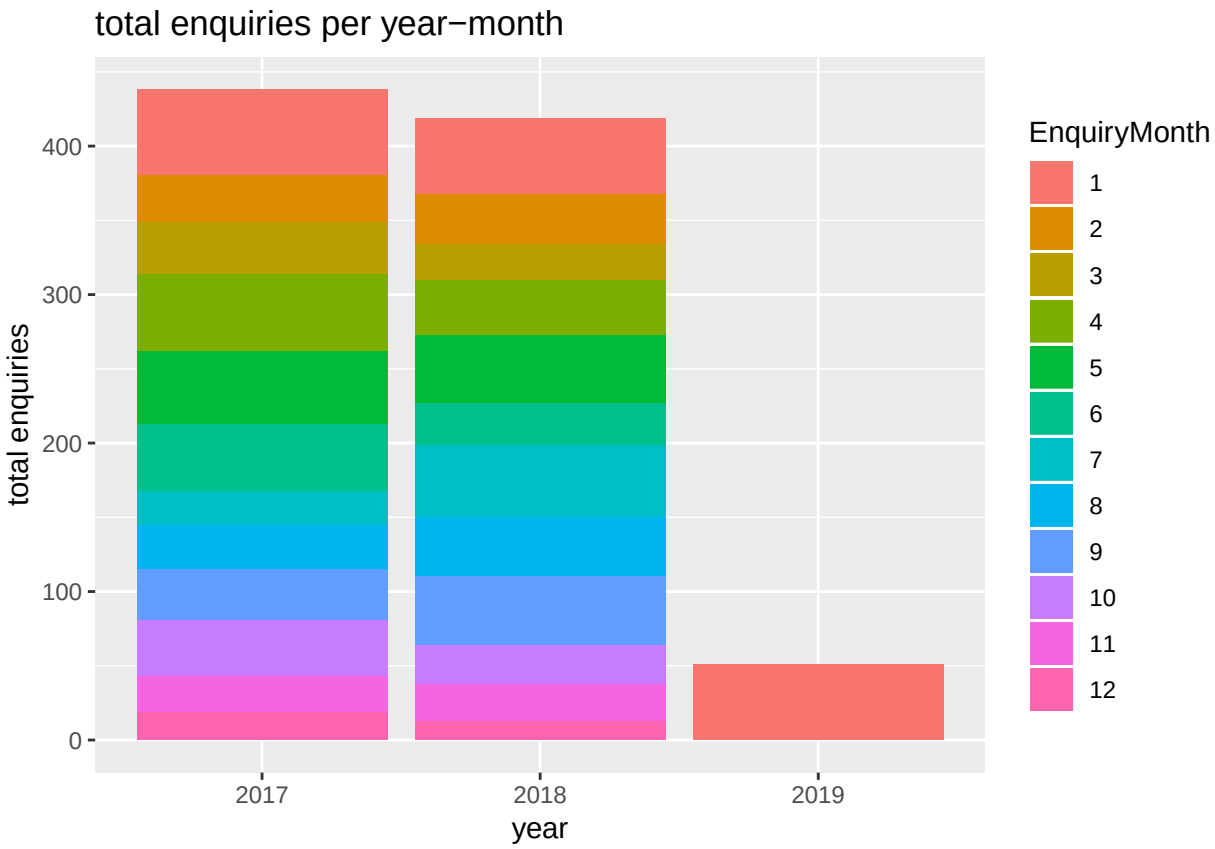
What are the day and month wise total enquiries?

```
day_month_sale<- data%>%group_by(EnquiryMonth,EnquiryDay)  %>% count(Destination)%>%arrange(EnquiryMonth

ggplot(data=day_month_sale, aes(x=EnquiryDay,y=n,fill=as.factor(EnquiryDay)))+geom_bar(stat="identity")+
labs(title= "Enquiries per day per month", x="EnquiryDay",y="Enquiries")
```



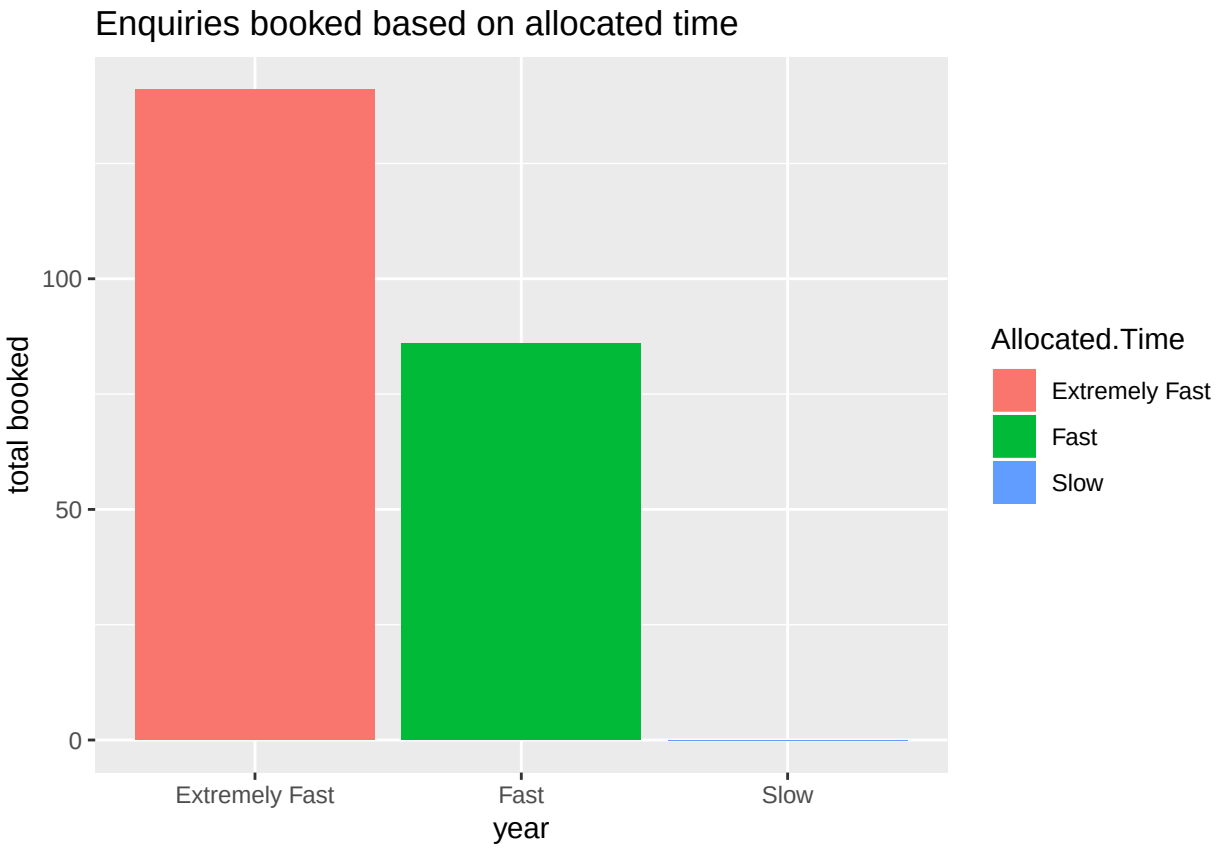
Total enquiries by year and month

```
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")
```



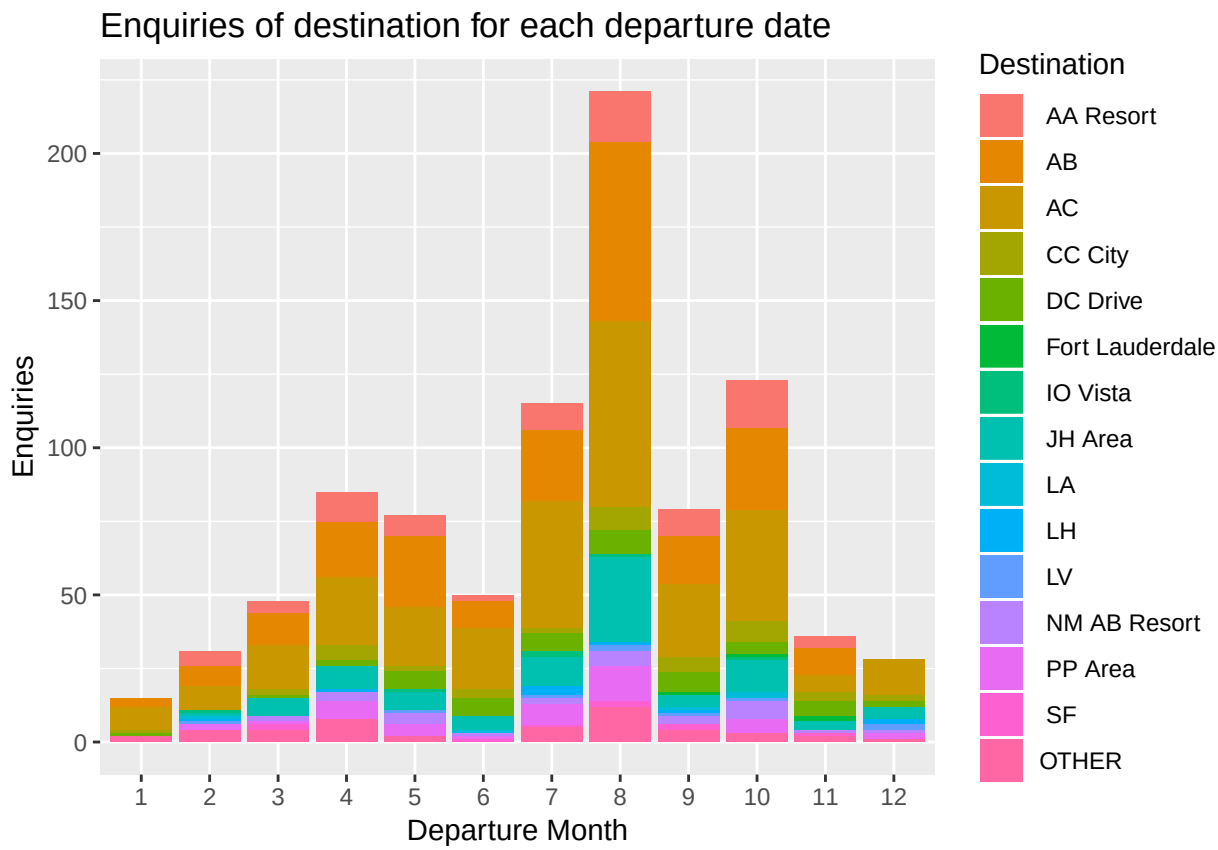
How many enquiries were booked based on Allocated.Time?

```
data$Booked.Status<-as.integer(data$Booked.Status)
data$Booked.Status<-ifelse(data$Booked.Status %in% 1,0,1)
booked_Allocated<-data%>%group_by(Allocated.Time)%>% summarise(booked=sum(Booked.Status)) %>%arrange(Al
ggplot(data=booked_Allocated,aes(x=Allocated.Time,y=booked,fill=as.factor(Allocated.Time)))+ geom_bar(s
  labs(title="Enquiries booked based on allocated time",x="year",y="total booked",fill="Allocated.Time"
```



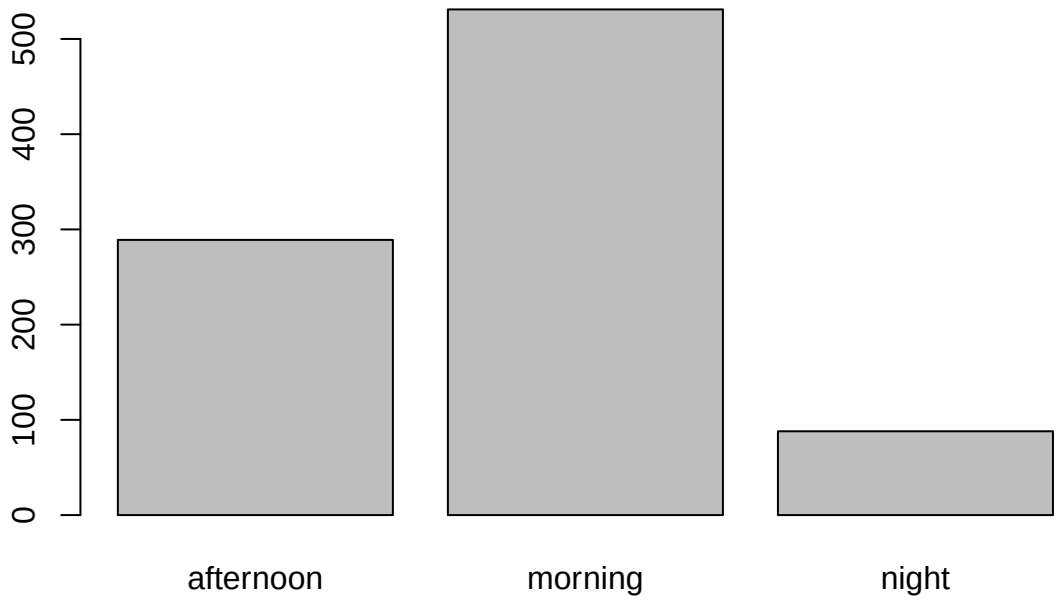
Which destinations are popular based on departure months?

```
ggplot(data,aes(x=factor(data$DepMonth),fill=Destination))+geom_bar()+
  labs(title="Enquiries of destination for each departure date",x="Departure Month",y="Enquiries",fill=
```

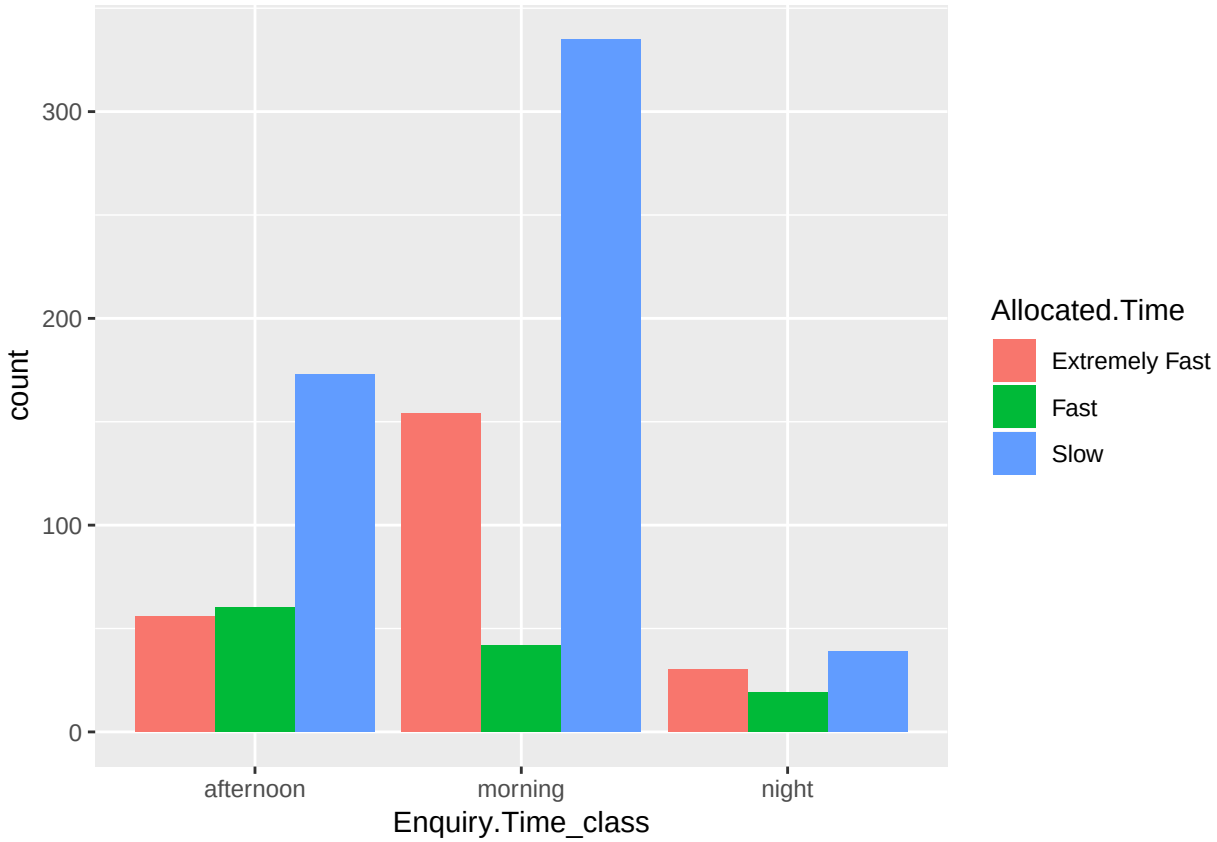


Which time of the day is the most enquiries coming in and Which period of the day is the Allocated.Time the worst?

```
plot(data$Enquiry.Time_class)
```




```
ggplot(data,aes(x=Enquiry.Time_class,fill=Allocated.Time)) + geom_bar(position="dodge")
```

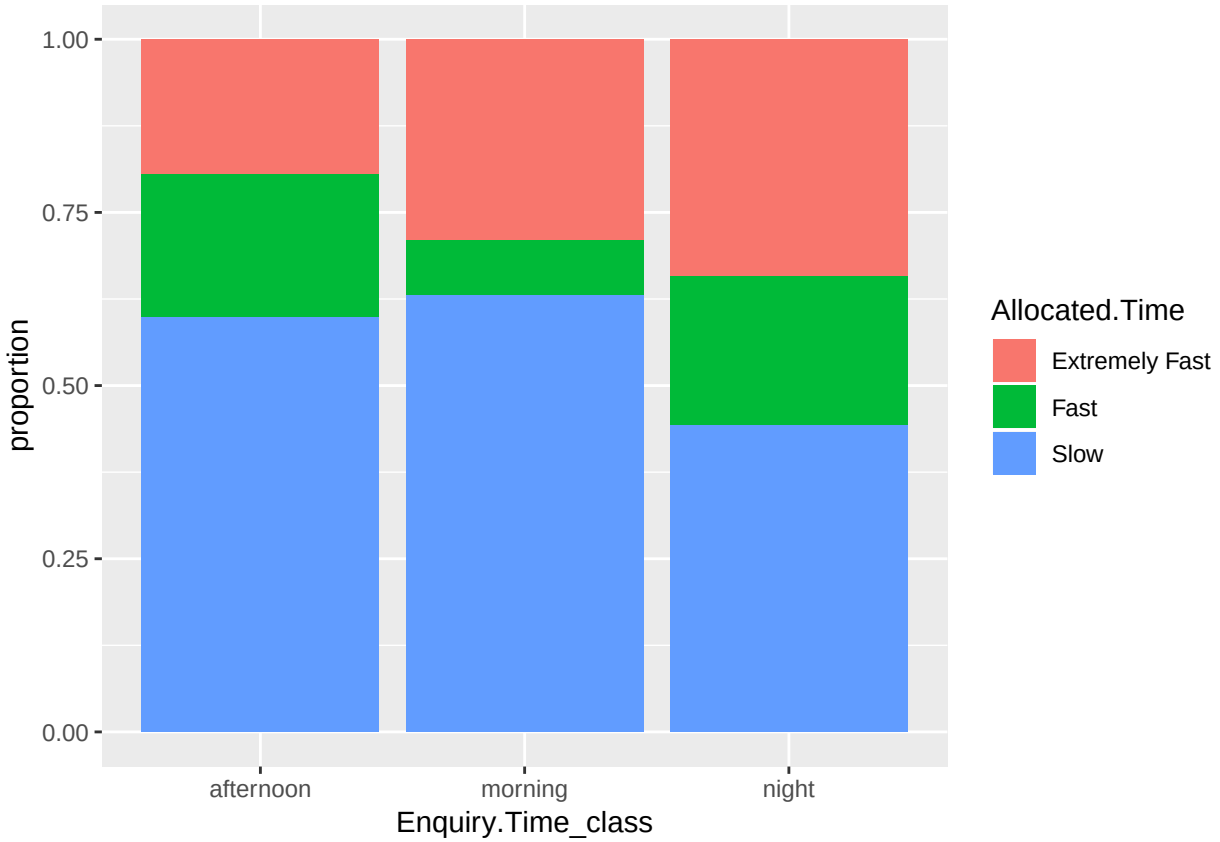


what is the Proportion of agent allocation speed for monring, afternoon and night?

```
tab_count<-table(data$Enquiry.Time_class,data$Allocated.Time)
prop.table(tab_count,1)
```

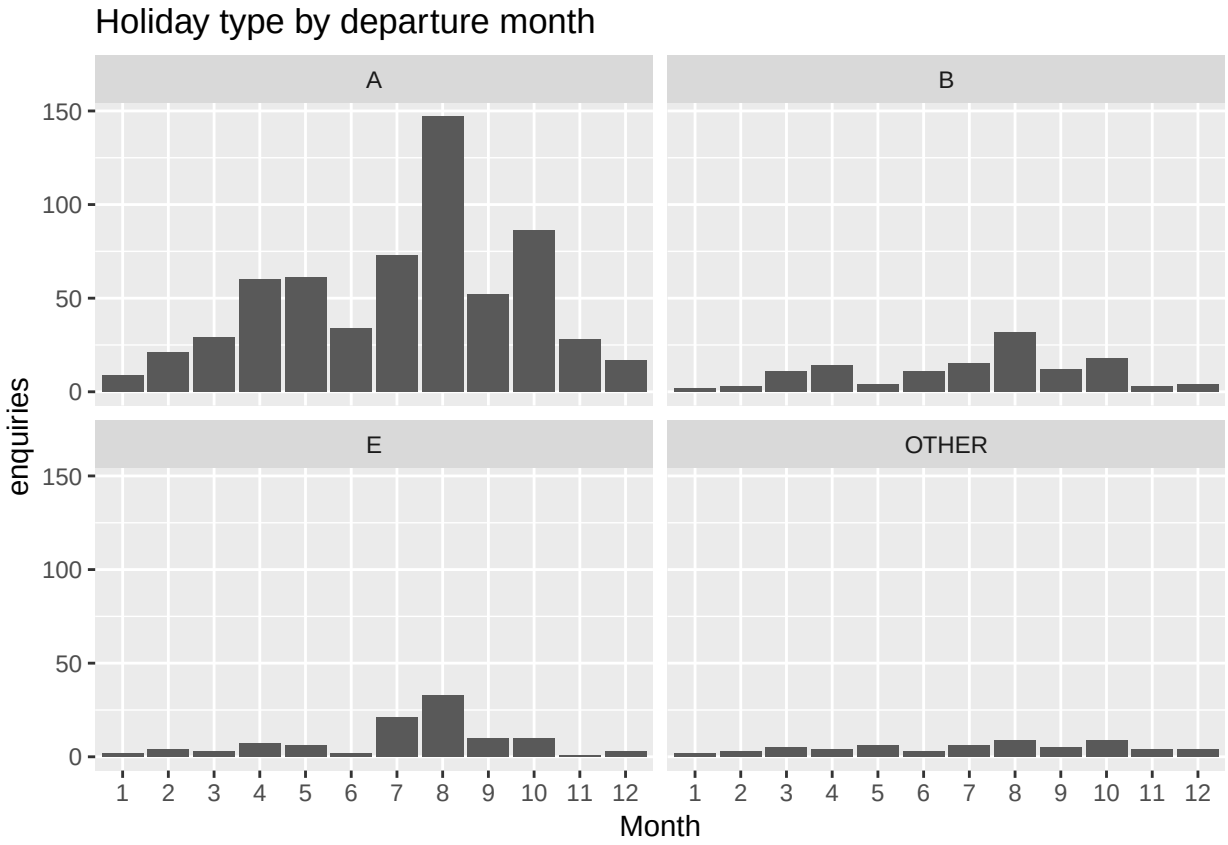
```
##
##             Extremely Fast       Fast       Slow
##   afternoon     0.19377163 0.20761246 0.59861592
##   morning       0.29001883 0.07909605 0.63088512
##   night         0.34090909 0.21590909 0.44318182
```

```
ggplot(data,aes(x=Enquiry.Time_class,fill=Allocated.Time)) + geom_bar(position="fill") + ylab("proporti
```



What are the most popular destinations each month?

```
ggplot(data,aes(x=factor(DepMonth))) + geom_bar() + facet_wrap(~Holiday.Type) +
  labs(title="Holiday type by departure month",x="Month",y="enquiries")
```

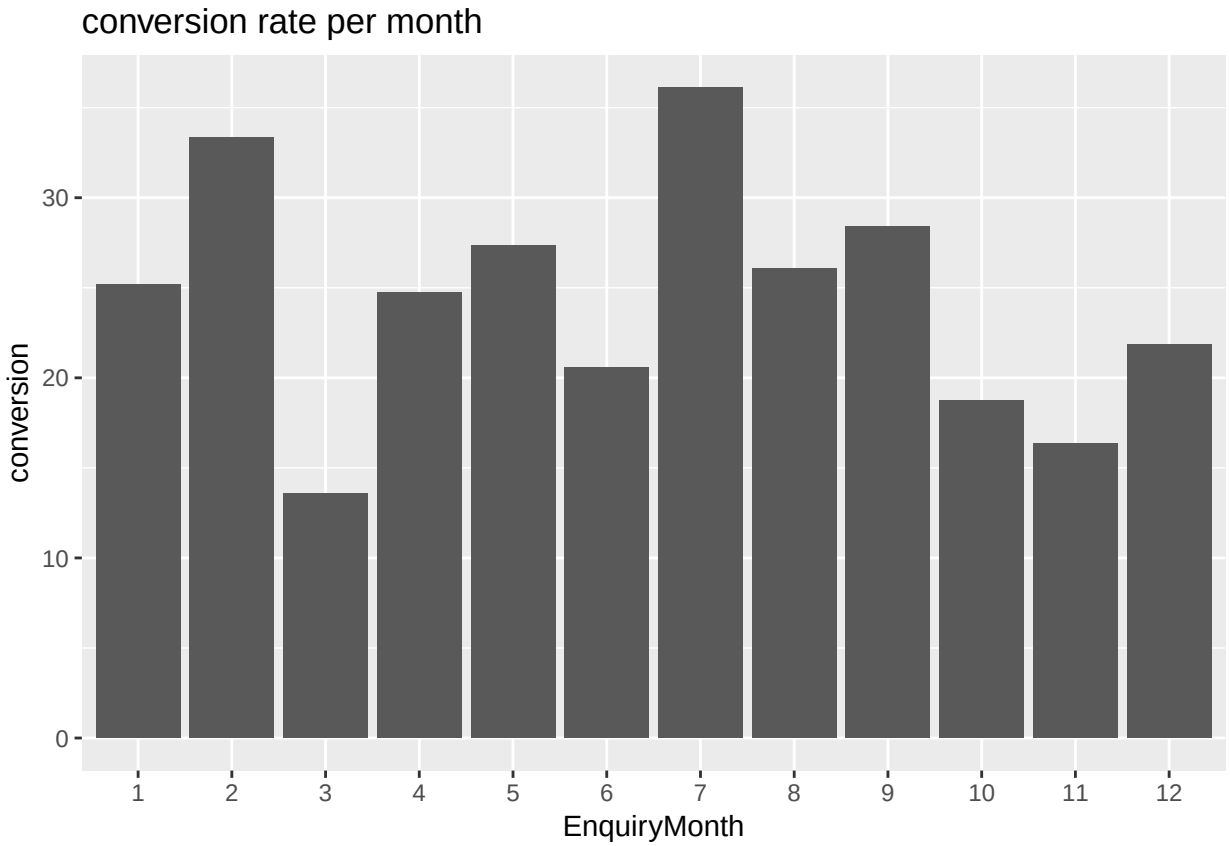


What is the conversion rate per month?

```
data$Booked<-as.integer(data$Booked.Status)
summarization <- sqldf("select EnquiryMonth, count(EnquiryMonth) as enquiries, sum(Booked) as totalbooke
summarization$totalbooked<- as.numeric(summarization$totalbooked)
summarization$enquiries<- as.numeric(summarization$enquiries)
conversionrate <- sqldf("select *, (totalbooked/enquiries)*100 as conversion from summarization")
data.frame(conversionrate)
```

```
##    EnquiryMonth enquiries totalbooked conversion
## 1             1       159          40   25.15723
## 2             2        66          22   33.33333
## 3             3        59           8   13.55932
## 4             4        89          22   24.71910
## 5             5        95          26   27.36842
## 6             6        73          15   20.54795
## 7             7        72          26   36.11111
## 8             8        69          18   26.08696
## 9             9        81          23   28.39506
## 10           10        64          12   18.75000
## 11           11        49           8   16.32653
## 12           12        32           7   21.87500
```

```
ggplot(conversionrate,aes(x=factor(EnquiryMonth),y=conversion))+geom_bar(stat="identity")+labs(title="c
```


conversion rate per month
conversion
30
20
10
0
1
2
3
4
5
6
7
8
9
10
11
12
EnquiryMonth