CSML1000 Winter 2020, Group 8, Course Project

Steven Wang, Tarun Bagga, Paul Doucet, Jerry Khidaroo, Nikola Stevanovic

3/19/2020

Contents

Load Libraries

```
# Load packages
library(knitr)
library(arules)
library(arulesViz)
library(dplyr)
library(data.table)
library(shinydashboard)
library(shiny)
library(ggplot2)
library(lubridate)
```

```
#library(Hmisc)
#library(funModeling)
library(tidyverse)
library(survival)
library(broom)

library(survival)
library(ranger)
library(ggplot2)
library(dplyr)
library(ggfortify)
```

```
## Warning: package 'ggfortify' was built under R version 3.6.3
```

```
library(DataExplorer)
```

```
## Warning: package 'DataExplorer' was built under R version 3.6.3
```

1. Business Understanding

- Business Problem: The business problem we are trying to solve fir the final project is to facioptimizing travel agent bookings so that they can accurately gauge when the surge will occur for the consummers to fly towards sun destinations and make a profit while doing so. To do so we will build a model that will take historical travel data and use that to let travel agents know how far in advance they should prebook hotel rooms in popular travel locations.
- Project Plan:
  - Load and get an understanding of the dataset, its target variable and its features.
  - Make any modifications to the dataset needed to enable learning algorithms to be run on the data.
  - Identify the features of the dataset that are important in predicting the target variable (dianosis in this case).
  - Build and evaluate a few models from the dataset by appling various machine learning algoritms as appropiate and testing them.
  - Identify the best model to use for the project.
  - Build a shiny app that deploys the selected model with a user interface for end users to imput measurement values from a study and obtain a pridiction result.
  - Identify any ethical considerations that should be addressed at each stage of the process.
- Business Success Criteria: Success criteria for this business case would be that travel agents can use the application to streamline their booking process and also use the data provided to increase profitability by booking aggressively at hotels before surge takes place.
- Ethical Framework Questions:
  - How could your system negatively impact individuals? This is very important question to ask as we have to look at how our recommendations are not creating biases or undesired negative results for the customers.As this system is based on historical data, some years might have had outside factors that skew the data and while an effort has been made to remove outliers there is still a chance that a booking date will be shown that is not the best predicted date for a possible destination.

- Who is most vulnerable and why? The most vulnerable people are the travel agents who would use the application, a common use case shall be that the booking agents try to prebook rooms expecting a lot of sales to take place in the coming months, in case the model has not predicted an extreme event to take place such as pandemic or a weather related event then they might not reach the expected amount of sales and will either lose money or not profit as much as possible.
- How much error in predictions can your business accept for this use case? As this could potentially affect the business profit and bottom line, we need to try and keep error to a minimum.
- Will you need to explain which input factors had the greatest influence on outputs? Yes, the important input factors will need to be explained as the model is time to event and we need to imagine where the data is coming from and data.
- Do you need PII or can you provide group-level data? The analysis requires customers travel booking and booking costs data however any PII can be anonymised.

2. Data Understanding

- Ethical Framework Questions:
  - Have you de-identified your data and taken measures to reduce the probability of reidentification? Yes the data is de-identified
  - Will socially sensitive features like gender or ethnic background influence outputs? No, socially sensitive features will not influence outputs
  - Are seemingly harmless features like location hiding proxies for socially sensitive features? No, there are no socially sensitive features

2.1 Get Data File

- For this final project we wanted to look at a real use case which presented real issues with data with skewness, data imbalance and ambiguity we came across the vaccations Dataset used is obtained from: Air Canada Hackathon Data

```
# Load data
unzip("./input/acme-travel-vacation.zip", exdir = "input") # unzip file
raw = read.table("./input/acme-travel-vacation.csv", sep="\t", header=TRUE)

# Load in airport dataset
raw_airport = read.csv("./input/cities_IATA_long_lat.csv", header=TRUE)
dataset_airport <- raw_airport

# check data
str(raw)
```

2.2 Load and check data

```
## 'data.frame':    652446 obs. of  42 variables:
##  $ BKG_ID                      : int  1171217 1220260 1236226 1240215 1266467 1268100 1269930 12760
##  $ SEND_DATE                   : int  20191223 20191223 20191223 20191223 20191223 20191223 2019122
##  $ PACKAGE_ID                  : Factor w/ 1 level "NULL": 1 1 1 1 1 1 1 1 1 1 ...
##  $ DESTINATION                 : Factor w/ 38 levels "ANTIGUA","BRIDGETOWN",..: 22 22 28 22 22 3 28
```

```
##  $ PACKAGE_TYPE_INDICATOR       : Factor w/ 4 levels "A","F","H","P": 4 4 4 4 4 4 4 4 4 1 ...
##  $ PROPERTY_ID                  : Factor w/ 834 levels "ANU","ANUBLU",..: 412 412 576 469 412 205 569
##  $ ACCOM_TYPE_PAID              : Factor w/ 5616 levels "ANUBLUBEAC","ANUBLUCOVE",..: 2700 2701 3871
##  $ ACCOM_SIZE_PAID              : Factor w/ 1 level "NULL": 1 1 1 1 1 1 1 1 1 1 ...
##  $ ACCOM_LEVEL_PAID             : Factor w/ 1 level "NULL": 1 1 1 1 1 1 1 1 1 1 ...
##  $ PARTY_SIZE                   : int  27 16 35 31 7 86 29 38 26 75 ...
##  $ ACCOM_TYPE_RECEIVED          : Factor w/ 1 level "NULL": 1 1 1 1 1 1 1 1 1 1 ...
##  $ ACCOM_SIZE_RECEIVED          : Factor w/ 1 level "NULL": 1 1 1 1 1 1 1 1 1 1 ...
##  $ ACCOM_LEVEL_RECEIVED         : Factor w/ 1 level "NULL": 1 1 1 1 1 1 1 1 1 1 ...
##  $ TRUE_ORIGIN                  : Factor w/ 73 levels "ANY","CUN","CUR",..: 67 51 59 67 67 67 67 67 6
##  $ MAIN_FLIGHT_ORIGIN           : Factor w/ 21 levels "ANY","CUN","CUR",..: 20 20 20 20 20 20 20 20 2
##  $ MAIN_FLIGHT_DESTINATION      : Factor w/ 38 levels "ANU","ANY","AZS",..: 18 18 21 18 18 7 21 7 7 1
##  $ INBOUND_FLIGHT_NUMBER        : Factor w/ 144 levels "AC1230","AC1240",..: 80 80 67 80 80 82 67 82
##  $ INBOUND_FEEDER_FLIGHT_NUMBER : Factor w/ 1007 levels "-1","AC10","AC100",..: 434 365 914 435 1 845
##  $ INBOUND_TRAVEL_CLASS         : Factor w/ 24 levels "-1","A","C","D",..: 6 6 6 6 6 6 6 24 6 6 ...
##  $ OUTBOUND_FLIGHT_NUMBER       : Factor w/ 153 levels "-1","AC1231",..: 81 81 68 81 81 84 68 84 84 4
##  $ OUTBOUND_FEEDER_FLIGHT_NUMBER: Factor w/ 942 levels "-1","AC101","AC102",..: 413 343 856 413 1 565
##  $ OUTBOUND_TRAVEL_CLASS        : Factor w/ 25 levels "-1","A","B","C",..: 7 7 7 7 7 7 7 25 7 7 ...
##  $ BKG_TYPE                     : Factor w/ 2 levels "GRP","IND": 1 1 1 1 1 1 1 1 1 1 ...
##  $ STATUS                       : Factor w/ 1 level "BKD": 1 1 1 1 1 1 1 1 1 1 ...
##  $ RATE_HEADER_ID               : Factor w/ 1 level "NULL": 1 1 1 1 1 1 1 1 1 1 ...
##  $ RATE_CODE                    : Factor w/ 1 level "NULL": 1 1 1 1 1 1 1 1 1 1 ...
##  $ START_DATE                   : int  20170623 20170109 20170123 20170518 20170506 20170127 2017011
##  $ LENGTH_OF_STAY               : int  15 11 20 8 7 19 11 11 8 9 ...
##  $ BKG_DATE                     : int  20150615 20150819 20150909 20150914 20151014 20151016 2015101
##  $ REVENUE                      : num  55766 30289 62028 56597 15662 ...
##  $ SURCHARGES_AND_TAXES         : num  17673 10061 21187 20338 3333 ...
##  $ ANCILLARY_REVENUE            : Factor w/ 518 levels "-120","-200",..: 518 518 518 518 518 518 518
##  $ TOTAL_COST                   : num  46505 32750 58881 52881 8596 ...
##  $ LIST_PRICE                   : Factor w/ 1 level "NULL": 1 1 1 1 1 1 1 1 1 1 ...
##  $ MARKET                       : Factor w/ 1 level "NULL": 1 1 1 1 1 1 1 1 1 1 ...
##  $ SOURCE                       : Factor w/ 2 levels "B2B","B2C": 1 1 1 1 1 1 1 1 1 1 ...
##  $ RATE_GRP                     : Factor w/ 1 level "NULL": 1 1 1 1 1 1 1 1 1 1 ...
##  $ CXL_DATE                     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ MARGIN                       : num  7279 -2561 3146 3446 1516 ...
##  $ ACCOMMODATION_STAR_RATING    : Factor w/ 13 levels "","0","2.5*",..: 10 10 12 12 10 12 10 12 12 13
##  $ HOTEL_CJAIN_AFFILIATION      : Factor w/ 96 levels "ACCOR HOTELS",..: 8 8 5 8 8 64 49 64 5 57 ...
##  $ AGENCY                       : Factor w/ 3382 levels "1200002323","1508411717",..: 2459 2551 697 8
```

2.3 Initial Data Collection Report:

- There is two files provided as part of the given dataset for this semi supervised learning model:
  1. acme-travel-vacation.csv
  2. cities_IATA_long_lat.csv Each of the datset contains different type of times series data for historical bookeing oif sun destination hotels by various travel operators.

Let us explore and deep dive within the various datasets provided.

```
For this section there are 2 CSV files provided, namely acme-travel-vacation and cities_IATA_long_lat. T
```

Let us look the counts of records inside the files. Using the built-in R functions we can see that there are 652446 travel records by operators in the acme-travel-vacation file. The travel csv file contains

16 features: The Destination of the trip (DESTINATION) The property id destination (PROPERTY_ID) The party size of the customers (PARTY_SIZE) The flight destination 3 letter code (MAIN_FLIGHT_DESTINATION) The start date of the trip (START_DATE) The length of the trip (LENGHT_OF_STAY) The date that the trip was booked(BKG_DATE) The revenue made from the trip (REVENUE) The total cost of the trip(TOTAL_COST) The profit margin for the agent(MARGIN) The star rating for the accomodation(ACCOMMODATION_STAR_RATING) The chain that the hotel is affiliated with(HOTEL_CHAIN_AFFILIATION) The time to event(TTE) The price per night for the trip(Price_PerNight) The booking date as a weekday(Wday_BookingDate) The start date as a weekday(Wday_StartDate) Overall, there are 652446 rows in our data, there are also 16 variables in our datasett

The summary statistcs of the raw dataset is shown above.

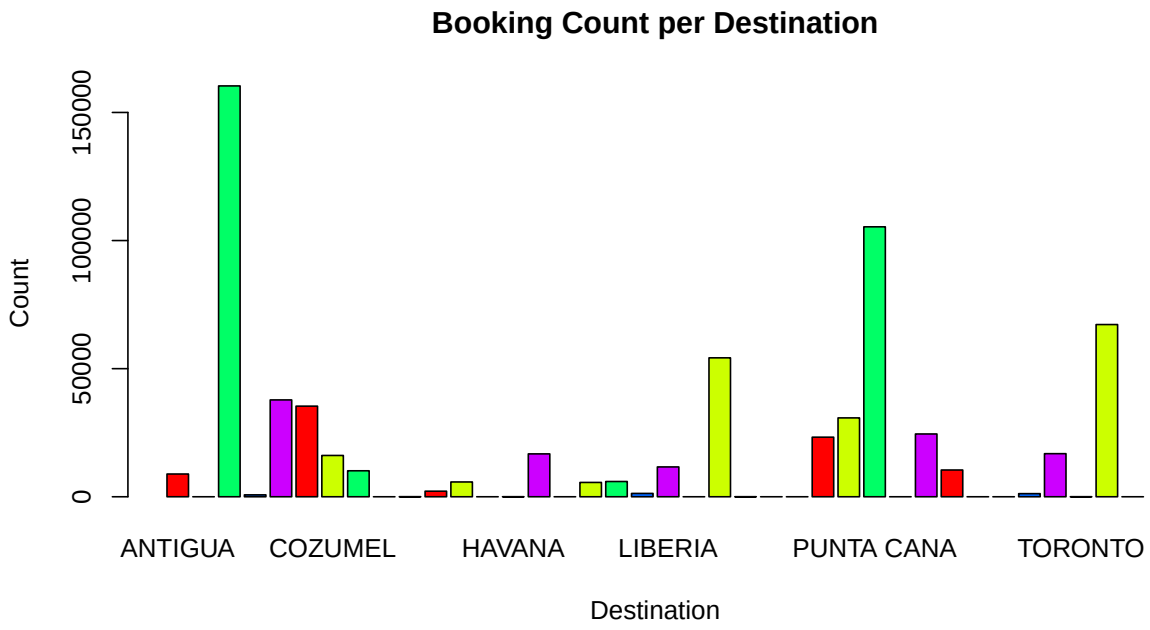
```
head(raw, 12)
```

```
##     BKG_ID SEND_DATE PACKAGE_ID DESTINATION PACKAGE_TYPE_INDICATOR PROPERTY_ID
## 1  1171217  20191223       NULL MONTEGO BAY                      P      MBJBAH
## 2  1220260  20191223       NULL MONTEGO BAY                      P      MBJBAH
## 3  1236226  20191223       NULL  PUNTA CANA                      P      PUJNLM
## 4  1240215  20191223       NULL MONTEGO BAY                      P      MBJRBA
## 5  1266467  20191223       NULL MONTEGO BAY                      P      MBJBAH
## 6  1268100  20191223       NULL      CANCUN                      P      CUNMOO
## 7  1269930  20191223       NULL  PUNTA CANA                      P      PUJMAJ
## 8  1276039  20191223       NULL      CANCUN                      P      CUNMOO
## 9  1284906  20191223       NULL      CANCUN                      P      CUNNSC
## 10 1291341  20191223       NULL   ST. KITTS                      A         N/A
## 11 1292325  20191223       NULL  PUNTA CANA                      P      PUJNLM
## 12 1298984  20191223       NULL  PUNTA CANA                      P      PUJMAJ
##    ACCOM_TYPE_PAID ACCOM_SIZE_PAID ACCOM_LEVEL_PAID PARTY_SIZE
## 1       MBJBAHJRST            NULL             NULL         27
## 2       MBJBAHJSST            NULL             NULL         16
## 3       PUJNLMDTRO            NULL             NULL         35
## 4       MBJRBAJRST            NULL             NULL         31
## 5       MBJBAHJRST            NULL             NULL          7
## 6       CUNMOOSDGV            NULL             NULL         86
## 7       PUJMAJJRSS            NULL             NULL         29
## 8       CUNMOOSDGV            NULL             NULL         38
## 9       CUNNSCDTRO            NULL             NULL         26
## 10             N/A            NULL             NULL         75
## 11      PUJNLMDTRO            NULL             NULL         32
## 12      PUJMAJJRSM            NULL             NULL         39
##    ACCOM_TYPE_RECEIVED ACCOM_SIZE_RECEIVED ACCOM_LEVEL_RECEIVED TRUE_ORIGIN
## 1                 NULL                NULL                 NULL         YYZ
## 2                 NULL                NULL                 NULL         YXE
## 3                 NULL                NULL                 NULL         YYC
## 4                 NULL                NULL                 NULL         YYZ
## 5                 NULL                NULL                 NULL         YYZ
## 6                 NULL                NULL                 NULL         YYZ
## 7                 NULL                NULL                 NULL         YYZ
## 8                 NULL                NULL                 NULL         YYZ
## 9                 NULL                NULL                 NULL         YYZ
## 10                NULL                NULL                 NULL         YYZ
```

```
## 11                NULL                NULL                NULL         YYZ
## 12                NULL                NULL                NULL         YYZ
##    MAIN_FLIGHT_ORIGIN MAIN_FLIGHT_DESTINATION INBOUND_FLIGHT_NUMBER
## 1                 YYZ                     MBJ                AC1804
## 2                 YYZ                     MBJ                AC1804
## 3                 YYZ                     PUJ                AC1772
## 4                 YYZ                     MBJ                AC1804
## 5                 YYZ                     MBJ                AC1804
## 6                 YYZ                     CUN                AC1810
## 7                 YYZ                     PUJ                AC1772
## 8                 YYZ                     CUN                AC1810
## 9                 YYC                     CUN                AC1564
## 10                YYZ                     SKB                AC1730
## 11                YYZ                     PUJ                AC1772
## 12                YYZ                     PUJ                AC1772
##    INBOUND_FEEDER_FLIGHT_NUMBER INBOUND_TRAVEL_CLASS OUTBOUND_FLIGHT_NUMBER
## 1                         AC689                    F                AC1805
## 2                         AC457                    F                AC1805
## 3                        AC8838                    F                AC1773
## 4                         AC691                    F                AC1805
## 5                            -1                    F                AC1805
## 6                        AC8695                    F                AC1811
## 7                            -1                    F                AC1773
## 8                            -1                    Z                AC1811
## 9                            -1                    F                AC1811
## 10                       AC8953                    F                AC1731
## 11                       AC8546                    F                AC1773
## 12                       AC8919                    F                AC1773
##    OUTBOUND_FEEDER_FLIGHT_NUMBER OUTBOUND_TRAVEL_CLASS BKG_TYPE STATUS
## 1                          AC698                     F      GRP    BKD
## 2                          AC466                     F      GRP    BKD
## 3                         AC8835                     F      GRP    BKD
## 4                          AC698                     F      GRP    BKD
## 5                             -1                     F      GRP    BKD
## 6                         AC8168                     F      GRP    BKD
## 7                             -1                     F      GRP    BKD
## 8                          AC152                     Z      GRP    BKD
## 9                             -1                     F      GRP    BKD
## 10                        AC8952                     F      GRP    BKD
## 11                        AC8547                     F      GRP    BKD
## 12                        AC8928                     F      GRP    BKD
##    RATE_HEADER_ID RATE_CODE START_DATE LENGTH_OF_STAY BKG_DATE   REVENUE
## 1            NULL      NULL   20170623             15 20150615  55766.26
## 2            NULL      NULL   20170109             11 20150819  30288.70
## 3            NULL      NULL   20170123             20 20150909  62027.60
## 4            NULL      NULL   20170518              8 20150914  56596.82
## 5            NULL      NULL   20170506              7 20151014  15662.48
## 6            NULL      NULL   20170127             19 20151016 197882.26
## 7            NULL      NULL   20170114             11 20151019  37825.36
## 8            NULL      NULL   20170116             11 20151026  59270.22
## 9            NULL      NULL   20170506              8 20151104  42104.03
## 10           NULL      NULL   20170114              9 20151112 199417.22
## 11           NULL      NULL   20170114             12 20151113  56507.13
## 12           NULL      NULL   20170113              9 20151120  69683.34
```

```
##    SURCHARGES_AND_TAXES ANCILLARY_REVENUE TOTAL_COST LIST_PRICE MARKET SOURCE
## 1              17672.57              NULL   46505.24       NULL   NULL    B2B
## 2              10060.72              NULL   32749.92       NULL   NULL    B2B
## 3              21186.67              NULL   58881.49       NULL   NULL    B2B
## 4              20337.95              NULL   52881.01       NULL   NULL    B2B
## 5               3333.40              NULL    8596.38       NULL   NULL    B2B
## 6              44545.51              NULL  164774.33       NULL   NULL    B2B
## 7              11395.69              NULL   33199.59       NULL   NULL    B2B
## 8              10680.44              NULL   55222.49       NULL   NULL    B2B
## 9              11643.91              NULL   38937.92       NULL   NULL    B2B
## 10             34398.07              NULL   50489.98       NULL   NULL    B2B
## 11             19031.71              NULL   55312.37       NULL   NULL    B2B
## 12             21838.48              NULL   65132.85       NULL   NULL    B2B
##    RATE_GRP CXL_DATE    MARGIN ACCOMMODATION_STAR_RATING
## 1      NULL        0   7279.22                      4.5*
## 2      NULL        0  -2561.22                      4.5*
## 3      NULL        0   3146.11                        5*
## 4      NULL        0   3445.81                        5*
## 5      NULL        0   1516.10                      4.5*
## 6      NULL        0  32657.93                        5*
## 7      NULL        0   4625.77                      4.5*
## 8      NULL        0   4047.73                        5*
## 9      NULL        0   2909.61                        5*
## 10     NULL        0 148657.24                       N/A
## 11     NULL        0   -641.48                        5*
## 12     NULL        0   4550.49                      4.5*
##           HOTEL_CJAIN_AFFILIATION     AGENCY
## 1  BAHIA PRINCIPE HOTEL AND RSRTS 7097226500
## 2  BAHIA PRINCIPE HOTEL AND RSRTS 7804383808
## 3                     AM RESORTS 4165324949
## 4  BAHIA PRINCIPE HOTEL AND RSRTS 2049873333
## 5  BAHIA PRINCIPE HOTEL AND RSRTS 4162407700
## 6                  PALACE RESORTS 9058952433
## 7     MAJESTIC HOTELS AND RESORTS 9024314932
## 8                  PALACE RESORTS 2503743393
## 9                     AM RESORTS 4032405350
## 10                         N/A 9052587778
## 11                    AM RESORTS 8076237449
## 12    MAJESTIC HOTELS AND RESORTS 5197524363
```

```r
a=table(raw$DESTINATION)
barplot(a,main="Booking Count per Destination",
        ylab="Count",
        xlab="Destination",
        col=rainbow(5),
        )
```



Feature Engineering

3. Data Preparation

a) Data Modification We modified the data quite heavily. Firstly we removed a bunch of columns that were completely NULL and then focused on identifying the data columns that would be useful for our business case. As we were doing our business case based on hotel booking we did not not need most of the data concerning the flight information or flight codes.

Column Removals Based on the business problem of trying to find optimal booking window for the agent we had removed most of teh columns visually as they were not providing sufficent analysis to the model we were trying to create. This can be seen on the previous sections where removed the features.

```
# Remove Lines for the ID and Null X final Columns
# travel_data_full <- travel_data_full[2:32]
```

Scale Data Set We did not sclae the dataset as our model did not require scaling of data. We decided to remove the outliers and that seemed sufficent enough to get a good enough survivi=al analysis model. Hence Scaling was not included in this report.

Split Data into Train and Test Sets Since we are using a semi supervised learning model, we are not able to use any kind of test and train split on the original datset and all the data shall be used for modelling purposes.

```
# prop.table(table(travel_data_train$diagnosis))*100
# prop.table(table(travel_data_test$diagnosis))*100
```

For feature enginnering we would be removing null features and ## Removing Null Values We did a summary on the main datset and traced all the features which had null values and removed it from the original dataset.

```
summary(raw)
```

```
##      BKG_ID            SEND_DATE        PACKAGE_ID              DESTINATION
##  Min.   : 1171217   Min.   :20191223   NULL:652446   CANCUN          :160357
##  1st Qu.:21253668   1st Qu.:20191223                 PUNTA CANA      :105358
##  Median :21471904   Median :20191223                 VARADERO        : 67211
##  Mean   :21077364   Mean   :20191223                 MONTEGO BAY     : 54226
##  3rd Qu.:21686243   3rd Qu.:20191223                 CAYO COCO       : 37801
##  Max.   :31023988   Max.   :20191223                 CAYO SANTA CLARA: 35364
##                                                      (Other)         :192129
##  PACKAGE_TYPE_INDICATOR  PROPERTY_ID       ACCOM_TYPE_PAID   ACCOM_SIZE_PAID
##  A: 22041                N/A    : 22205   N/A       : 22205   NULL:652446
##  F:  6440                CUNCOB :  9200   PUJBAHJRST:  8166
##  H:  4883                MBJBAH :  8861   CUNCOBJRST:  7188
##  P:619082                CUNSIA :  8729   PUJGREJRST:  6396
##                          MBJRBA :  8649   CCCMJASTRD:  6239
##                          PUJBAH :  8169   CUNSIAJRSS:  5701
##                          (Other):586633   (Other)   :596551
##  ACCOM_LEVEL_PAID   PARTY_SIZE      ACCOM_TYPE_RECEIVED ACCOM_SIZE_RECEIVED
##  NULL:652446      Min.   :  0.000   NULL:652446         NULL:652446
##                   1st Qu.:  2.000
##                   Median :  2.000
##                   Mean   :  2.979
##                   3rd Qu.:  3.000
##                   Max.   :450.000
##
##  ACCOM_LEVEL_RECEIVED  TRUE_ORIGIN     MAIN_FLIGHT_ORIGIN
##  NULL:652446          YYZ    :284550   YYZ    :340244
##                       YUL    :215534   YUL    :235568
##                       YOW    : 30020   YVR    : 22829
##                       YVR    : 22530   YOW    : 17526
##                       YYC    : 17147   YYC    : 13226
##                       YHZ    : 15317   YHZ    : 11908
##                       (Other): 67348   (Other): 11145
##  MAIN_FLIGHT_DESTINATION INBOUND_FLIGHT_NUMBER INBOUND_FEEDER_FLIGHT_NUMBER
##  CUN    :160357          AC1810 : 53107        -1     :552905
##  PUJ    :105358          AC1712 : 46070        AC8970 :  2409
##  VRA    : 67211          AC1986 : 43564        AC158  :  2193
##  MBJ    : 54226          AC1804 : 42741        AC441  :  1894
##  CCC    : 37801          AC1882 : 36841        AC401  :  1872
##  SNU    : 35364          AC1746 : 35338        AC481  :  1865
##  (Other):192129          (Other):394785        (Other): 89308
##  INBOUND_TRAVEL_CLASS OUTBOUND_FLIGHT_NUMBER OUTBOUND_FEEDER_FLIGHT_NUMBER
##  F      :605988       AC1811 : 49882         -1     :553159
##  P      : 15380       AC1713 : 46279         AC430  :  3102
##  C      : 15142       AC1987 : 43636         AC470  :  2902
##  -1     :  4915       AC1805 : 42956         AC428  :  2551
##  Z      :  3557       AC1883 : 36978         AC472  :  1488
##  A      :  2817       AC1747 : 35343         AC8761 :  1435
##  (Other):  4647       (Other):397372         (Other): 87809
##  OUTBOUND_TRAVEL_CLASS BKG_TYPE     STATUS       RATE_HEADER_ID RATE_CODE
```

```
##  F      :606319        GRP: 15288   BKD:652446   NULL:652446    NULL:652446
##  P      : 15743        IND:637158
##  C      : 15172
##  -1     :  4890
##  Z      :  3224
##  A      :  2749
##  (Other):  4349
##    START_DATE       LENGTH_OF_STAY       BKG_DATE           REVENUE
##  Min.   :20170101   Min.   :  0.000   Min.   :20150615   Min.   :-6159935
##  1st Qu.:20171125   1st Qu.:  7.000   1st Qu.:20170913   1st Qu.:    2272
##  Median :20180804   Median :  7.000   Median :20180605   Median :    3149
##  Mean   :20182499   Mean   :  7.354   Mean   :20179498   Mean   :    4758
##  3rd Qu.:20190408   3rd Qu.:  7.000   3rd Qu.:20190226   3rd Qu.:    4675
##  Max.   :20210827   Max.   :999.000   Max.   :20191223   Max.   :14757630
##
##  SURCHARGES_AND_TAXES ANCILLARY_REVENUE   TOTAL_COST      LIST_PRICE
##  Min.   :-14143.1     NULL   :527454    Min.   :  -3628   NULL:652446
##  1st Qu.:   228.6     80     : 24079    1st Qu.:   2103
##  Median :   328.3     98     : 15557    Median :   2816
##  Mean   :   604.3     0      : 13107    Mean   :   4041
##  3rd Qu.:   416.9     200    : 11313    3rd Qu.:   4025
##  Max.   :768372.7     118    :  5729    Max.   :6888711
##                       (Other): 55207
##   MARKET       SOURCE       RATE_GRP         CXL_DATE     MARGIN
##  NULL:652446   B2B:519517   NULL:652446   Min.   :0   Min.   :-6216023
##                B2C:132929                 1st Qu.:0   1st Qu.:     -67
##                                           Median :0   Median :     173
##                                           Mean   :0   Mean   :     637
##                                           3rd Qu.:0   3rd Qu.:     530
##                                           Max.   :0   Max.   : 6004311
##
##  ACCOMMODATION_STAR_RATING                   HOTEL_CJAIN_AFFILIATION
##  4.5*   :208187            BAHIA PRINCIPE HOTEL AND RSRTS:106553
##  4*     :182257            BLUE DIAMOND                  : 67346
##  5*     :168015            AM RESORTS                    : 46017
##  3.5*   : 36194            IBEROSTAR HOTELS AND RESORTS  : 39834
##  3*     : 26303            BARCELO HOTEL GROUPS          : 34895
##  N/A    : 22259            MELIA HOTELS INTL CUBA        : 31406
##  (Other):  9231            (Other)                       :326395
##         AGENCY
##  9052836020: 52652
##  9050002020: 39457
##  8663705911: 38362
##  5140002020: 27594
##  9055287800: 23734
##  4166792369: 20014
##  (Other)   :450633
```

This section shall remove the null columns that are not useful for the analysis.

```
dataset <- raw

# Remove the following columns
# Remove the following columns
```

```
dataset$PACKAGE_ID <- NULL
dataset$ACCOM_SIZE_PAID <- NULL
dataset$ACCOM_LEVEL_PAID <- NULL
dataset$ACCOM_TYPE_RECEIVED <- NULL
dataset$ACCOM_SIZE_RECEIVED <- NULL
dataset$ACCOM_LEVEL_RECEIVED <- NULL
dataset$STATUS <- NULL
dataset$RATE_HEADER_ID <- NULL
dataset$RATE_CODE <- NULL
dataset$LIST_PRICE <- NULL
dataset$MARKET <- NULL
dataset$RATE_GRP <- NULL
dataset$CXL_DATE <- NULL
dataset$SEND_DATE <- NULL
dataset$BKG_ID <- NULL
dataset$AGENCY <- NULL
dataset$PACKAGE_TYPE_INDICATOR <- NULL
dataset$ACCOM_TYPE_PAID <- NULL
dataset$TRUE_ORIGIN <- NULL
dataset$MAIN_FLIGHT_ORIGIN <- NULL
#dataset$MAIN_FLIGHT_DESTINATION <- NULL
#dataset$MAIN_FLIGHT_DESTINATION <- NULL
dataset$INBOUND_FLIGHT_NUMBER <- NULL
dataset$INBOUND_FEEDER_FLIGHT_NUMBER <- NULL
dataset$INBOUND_TRAVEL_CLASS <- NULL
dataset$OUTBOUND_FLIGHT_NUMBER <- NULL
dataset$OUTBOUND_FEEDER_FLIGHT_NUMBER <- NULL
dataset$OUTBOUND_TRAVEL_CLASS <- NULL
dataset$BKG_TYPE <- NULL
dataset$SURCHARGES_AND_TAXES <- NULL
dataset$ANCILLARY_REVENUE <- NULL
#dataset$TOTAL_COST <- NULL
dataset$SOURCE <- NULL
```

Data Transformation

We are performing transformation on all the date column by converting all 8 digit integer to y/m/d format. This is important step as it allows our model to deciopher dates properly in correct format.

```
# Make certain adjustments to the data in the following columns
dataset$START_DATE <- as.Date(as.character(dataset$START_DATE), format="%Y%m%d")
# BKG_DATE
dataset$BKG_DATE <- as.Date(as.character(dataset$BKG_DATE), format="%Y%m%d")
# ACCOMMODATION_STAR_RATING
dataset$ACCOMMODATION_STAR_RATING <- gsub("\\*", "", dataset$ACCOMMODATION_STAR_RATING)
dataset$ACCOMMODATION_STAR_RATING <- as.factor(dataset$ACCOMMODATION_STAR_RATING)

# TTE - Time between vacation start and booking date
dataset$TTE <- dataset$START_DATE - dataset$BKG_DATE
dataset$TTE <- as.numeric(dataset$TTE)
# Price
dataset$Price_PerNight <- dataset$REVENUE / dataset$PARTY_SIZE / dataset$LENGTH_OF_STAY
summary(dataset$Price_PerNight)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -29333.0    145.0    188.9      Inf    246.3      Inf
```

```
# Day of week of booking
dataset$Wday_BookingDate <- weekdays(as.Date(dataset$BKG_DATE))
dataset$Wday_BookingDate <- as.factor(dataset$Wday_BookingDate)
# Day of week of start of stay
dataset$Wday_StartDate <- weekdays(as.Date(dataset$START_DATE))
dataset$Wday_StartDate <- as.factor(dataset$Wday_StartDate)
```

Looking at the first few values of each column along with last few records

```
head(dataset,25)
```

```
##    DESTINATION            PROPERTY_ID PARTY_SIZE MAIN_FLIGHT_DESTINATION
## 1  MONTEGO BAY                 MBJBAH         27                      MBJ
## 2  MONTEGO BAY                 MBJBAH         16                      MBJ
## 3   PUNTA CANA                 PUJNLM         35                      PUJ
## 4  MONTEGO BAY                 MBJRBA         31                      MBJ
## 5  MONTEGO BAY                 MBJBAH          7                      MBJ
## 6       CANCUN                 CUNMOO         86                      CUN
## 7   PUNTA CANA                 PUJMAJ         29                      PUJ
## 8       CANCUN                 CUNMOO         38                      CUN
## 9       CANCUN                 CUNNSC         26                      CUN
## 10   ST. KITTS                    N/A         75                      SKB
## 11  PUNTA CANA                 PUJNLM         32                      PUJ
## 12  PUNTA CANA                 PUJMAJ         39                      PUJ
## 13 MONTEGO BAY                 MBJBAH         39                      MBJ
## 14      CANCUN                 CUNIBB         21                      CUN
## 15      CANCUN                 CUNCOB         27                      CUN
## 16     GRENADA                 GNDSLS         50                      GND
## 17      CANCUN                 CUNSPB         24                      CUN
## 18      CANCUN                 CUNMOO         14                      CUN
## 19  PUNTA CANA                 PUJHAR         25                      PUJ
## 20 GEORGE TOWN GGTHBSI_37947_10051798          2                      GGT
## 21 GEORGE TOWN                 GGTSEB          2                      GGT
## 22      CANCUN                 CUNMGD         76                      CUN
## 23   CAYO COCO                 CCCMJA         32                      CCC
## 24 MONTEGO BAY                 MBJBRV         87                      MBJ
## 25 MONTEGO BAY                 MBJBAH         35                      MBJ
##    START_DATE LENGTH_OF_STAY   BKG_DATE   REVENUE TOTAL_COST    MARGIN
## 1  2017-06-23             15 2015-06-15  55766.26   46505.24   7279.22
## 2  2017-01-09             11 2015-08-19  30288.70   32749.92  -2561.22
## 3  2017-01-23             20 2015-09-09  62027.60   58881.49   3146.11
## 4  2017-05-18              8 2015-09-14  56596.82   52881.01   3445.81
## 5  2017-05-06              7 2015-10-14  15662.48    8596.38   1516.10
## 6  2017-01-27             19 2015-10-16 197882.26  164774.33  32657.93
## 7  2017-01-14             11 2015-10-19  37825.36   33199.59   4625.77
## 8  2017-01-16             11 2015-10-26  59270.22   55222.49   4047.73
## 9  2017-05-06              8 2015-11-04  42104.03   38937.92   2909.61
## 10 2017-01-14              9 2015-11-12 199417.22   50489.98 148657.24
```

```
## 11 2017-01-14             12 2015-11-13  56507.13   55312.37   -641.48
## 12 2017-01-13              9 2015-11-20  69683.34   65132.85   4550.49
## 13 2017-05-06              8 2015-11-26  54909.10   49015.51   5788.59
## 14 2017-01-19              9 2015-12-11  28616.72   27235.58   1281.14
## 15 2017-02-04              7 2015-12-11  39982.53   35851.91   4030.62
## 16 2017-05-01             10 2015-12-21 146083.78  127953.28  14790.50
## 17 2017-01-22             17 2015-12-30  39766.03   36248.26   3517.77
## 18 2017-04-20              8 2016-01-04  25531.22   23132.79   1648.43
## 19 2017-04-04             10 2016-01-08  53749.02   51437.59   2311.43
## 20 2017-11-13            371 2016-01-10   9132.34    7160.45   1971.89
## 21 2017-11-13            371 2016-01-11   7618.00    5629.44   1455.30
## 22 2017-01-07             34 2016-01-12 121945.85  127842.60  -6721.84
## 23 2017-02-28              8 2016-01-20  37724.84   35738.65   1886.19
## 24 2017-04-22              7 2016-01-20 753664.58  228494.40 525583.02
## 25 2017-05-28             15 2016-01-20  50791.15   54564.12  -3772.97
##    ACCOMMODATION_STAR_RATING        HOTEL_CJAIN_AFFILIATION TTE Price_PerNight
## 1                        4.5 BAHIA PRINCIPE HOTEL AND RSRTS 739      137.69447
## 2                        4.5 BAHIA PRINCIPE HOTEL AND RSRTS 509      172.09489
## 3                          5                      AM RESORTS 502       88.61086
## 4                          5 BAHIA PRINCIPE HOTEL AND RSRTS 612      228.21298
## 5                        4.5 BAHIA PRINCIPE HOTEL AND RSRTS 570      319.64245
## 6                          5                PALACE RESORTS 469      121.10297
## 7                        4.5    MAJESTIC HOTELS AND RESORTS 453      118.57480
## 8                          5                PALACE RESORTS 448      141.79478
## 9                          5                    AM RESORTS 549      202.42322
## 10                       N/A                          N/A 429      295.43292
## 11                         5                      AM RESORTS 428      147.15398
## 12                       4.5    MAJESTIC HOTELS AND RESORTS 420      198.52803
## 13                       4.5 BAHIA PRINCIPE HOTEL AND RSRTS 527      175.99071
## 14                         4   IBEROSTAR HOTELS AND RESORTS 405      151.41122
## 15                       4.5 BAHIA PRINCIPE HOTEL AND RSRTS 421      211.54778
## 16                         5               SANDALS RESORTS 497      292.16756
## 17                         4      SANDOS HOTELS AND RESORTS 389       97.46576
## 18                         5                PALACE RESORTS 472      227.95732
## 19                         5               AIC HOTEL GROUP 452      214.99608
## 20                         5               SANDALS RESORTS 673       12.30774
## 21                         5               SANDALS RESORTS 672       10.26685
## 22                       4.5          BARCELO HOTEL GROUPS 361       47.19267
## 23                         4         MELIA HOTELS INTL CUBA 405      147.36266
## 24                         5     MELIA HOTELS INTERNATIONAL 458     1237.54447
## 25                       4.5 BAHIA PRINCIPE HOTEL AND RSRTS 494       96.74505
##    Wday_BookingDate Wday_StartDate
## 1            Monday         Friday
## 2         Wednesday         Monday
## 3         Wednesday         Monday
## 4            Monday       Thursday
## 5         Wednesday       Saturday
## 6            Friday         Friday
## 7            Monday       Saturday
## 8            Monday         Monday
## 9         Wednesday       Saturday
## 10         Thursday       Saturday
## 11           Friday       Saturday
## 12           Friday         Friday
```

```
## 13         Thursday       Saturday
## 14           Friday       Thursday
## 15           Friday       Saturday
## 16           Monday         Monday
## 17        Wednesday         Sunday
## 18           Monday       Thursday
## 19           Friday        Tuesday
## 20           Sunday         Monday
## 21           Monday         Monday
## 22          Tuesday       Saturday
## 23        Wednesday        Tuesday
## 24        Wednesday       Saturday
## 25        Wednesday         Sunday
```

```
names(dataset)
```

```
##  [1] "DESTINATION"               "PROPERTY_ID"
##  [3] "PARTY_SIZE"                "MAIN_FLIGHT_DESTINATION"
##  [5] "START_DATE"                "LENGTH_OF_STAY"
##  [7] "BKG_DATE"                  "REVENUE"
##  [9] "TOTAL_COST"                "MARGIN"
## [11] "ACCOMMODATION_STAR_RATING" "HOTEL_CJAIN_AFFILIATION"
## [13] "TTE"                       "Price_PerNight"
## [15] "Wday_BookingDate"          "Wday_StartDate"
```

```
str(dataset)
```

```
## 'data.frame':    652446 obs. of  16 variables:
##  $ DESTINATION              : Factor w/ 38 levels "ANTIGUA","BRIDGETOWN",..: 22 22 28 22 22 3 28 3 3
##  $ PROPERTY_ID              : Factor w/ 834 levels "ANU","ANUBLU",..: 412 412 576 469 412 205 569 205
##  $ PARTY_SIZE               : int  27 16 35 31 7 86 29 38 26 75 ...
##  $ MAIN_FLIGHT_DESTINATION  : Factor w/ 38 levels "ANU","ANY","AZS",..: 18 18 21 18 18 7 21 7 7 24 .
##  $ START_DATE               : Date, format: "2017-06-23" "2017-01-09" ...
##  $ LENGTH_OF_STAY           : int  15 11 20 8 7 19 11 11 8 9 ...
##  $ BKG_DATE                 : Date, format: "2015-06-15" "2015-08-19" ...
##  $ REVENUE                  : num  55766 30289 62028 56597 15662 ...
##  $ TOTAL_COST               : num  46505 32750 58881 52881 8596 ...
##  $ MARGIN                   : num  7279 -2561 3146 3446 1516 ...
##  $ ACCOMMODATION_STAR_RATING: Factor w/ 9 levels "","0","2.5","3",..: 7 7 8 8 7 8 7 8 8 9 ...
##  $ HOTEL_CJAIN_AFFILIATION  : Factor w/ 96 levels "ACCOR HOTELS",..: 8 8 5 8 8 64 49 64 5 57 ...
##  $ TTE                      : num  739 509 502 612 570 469 453 448 549 429 ...
##  $ Price_PerNight           : num  137.7 172.1 88.6 228.2 319.6 ...
##  $ Wday_BookingDate         : Factor w/ 7 levels "Friday","Monday",..: 2 7 7 2 7 1 2 2 7 5 ...
##  $ Wday_StartDate           : Factor w/ 7 levels "Friday","Monday",..: 1 2 2 5 3 1 3 2 3 3 ...
```

```
tail(dataset)
```

```
##        DESTINATION PROPERTY_ID PARTY_SIZE MAIN_FLIGHT_DESTINATION START_DATE
## 652441      CANCUN      CUNMOO         12                     CUN 2020-04-20
## 652442  PUNTA CANA      PUJCHC         30                     PUJ 2020-07-18
## 652443     COZUMEL      CZMALL         13                     CZM 2020-02-14
## 652444      CANCUN      CUNRIT         12                     CUN 2020-02-13
```

```
## 652445  PUNTA CANA      PUJGRE         30                    PUJ 2020-06-06
## 652446  PUNTA CANA      PUJELB         12                    PUJ 2020-03-18
##        LENGTH_OF_STAY   BKG_DATE  REVENUE TOTAL_COST MARGIN
## 652441              7 2019-12-23 22892.16  10319.279      0
## 652442              7 2019-12-23 64798.30  35570.480      0
## 652443              7 2019-12-23 29134.00  14618.397      0
## 652444              7 2019-12-23 15060.28   6042.715      0
## 652445              7 2019-12-23 51202.10  32228.573      0
## 652446              7 2019-12-23 25408.92  17697.720      0
##        ACCOMMODATION_STAR_RATING        HOTEL_CJAIN_AFFILIATION TTE
## 652441                         5                PALACE RESORTS 119
## 652442                         5                  BLUE DIAMOND 208
## 652443                         4           BARCELO HOTEL GROUPS  53
## 652444                       3.5       ALSOL HOTELS AND RESORTS  52
## 652445                         5 BAHIA PRINCIPE HOTEL AND RSRTS 166
## 652446                         5            OCEAN HOTELS BY H10  86
##        Price_PerNight Wday_BookingDate Wday_StartDate
## 652441       272.5257           Monday         Monday
## 652442       308.5633           Monday       Saturday
## 652443       320.1538           Monday         Friday
## 652444       179.2890           Monday       Thursday
## 652445       243.8195           Monday       Saturday
## 652446       302.4871           Monday      Wednesday
```

Data Split

The data is split into yearly format to visualize the data year on year. This allows us to see what is variance in the data for each year and guides us to create a better Survival analysis model.

```
##############################################################################################
##############################################################################################
##############################################################################################
##############################################################################################
All_2019 <- dataset[dataset$START_DATE < as.Date("2020-01-01") &
                        dataset$START_DATE > as.Date("2018-12-31") #&
                        # dataset$DESTINATION=="CANCUN" &
                        # dataset$LENGTH_OF_STAY >= 1 &
                        # dataset$LENGTH_OF_STAY <= 14 &
                        # dataset$MARGIN > 0
                        ,]

All_2018 <- dataset[dataset$START_DATE < as.Date("2019-01-01") &
                       dataset$START_DATE > as.Date("2017-12-31") #&
                       # #dataset$DESTINATION=="CANCUN" &
                       # dataset$LENGTH_OF_STAY >= 1 &
                       # dataset$LENGTH_OF_STAY <= 14 &
                       # dataset$MARGIN > 0
                      ,]

All_2017 <- dataset[dataset$START_DATE < as.Date("2018-01-01") &
                       dataset$START_DATE > as.Date("2016-12-31") #&
                       # dataset$DESTINATION=="CANCUN" &
                       # dataset$LENGTH_OF_STAY >= 1 &
```

```
                          # dataset$LENGTH_OF_STAY <= 14 &
                          # dataset$MARGIN > 0
                         ,]

All_2016 <- dataset[dataset$START_DATE < as.Date("2017-01-01") &
                          dataset$START_DATE > as.Date("2015-12-31") #&
                          # dataset$DESTINATION=="CANCUN" &
                          # dataset$LENGTH_OF_STAY >= 1 &
                          # dataset$LENGTH_OF_STAY <= 14 &
                          # dataset$MARGIN > 0
                         ,]

All_2015 <- dataset[dataset$START_DATE < as.Date("2016-01-01") &
                          dataset$START_DATE > as.Date("2014-12-31") #&
                          # dataset$DESTINATION=="CANCUN" &
                          # dataset$LENGTH_OF_STAY >= 1 &
                          # dataset$LENGTH_OF_STAY <= 14 &
                          # dataset$MARGIN > 0
                         ,]
```

We have created our model above and ran it on all the training data by saving into seperate excel format.

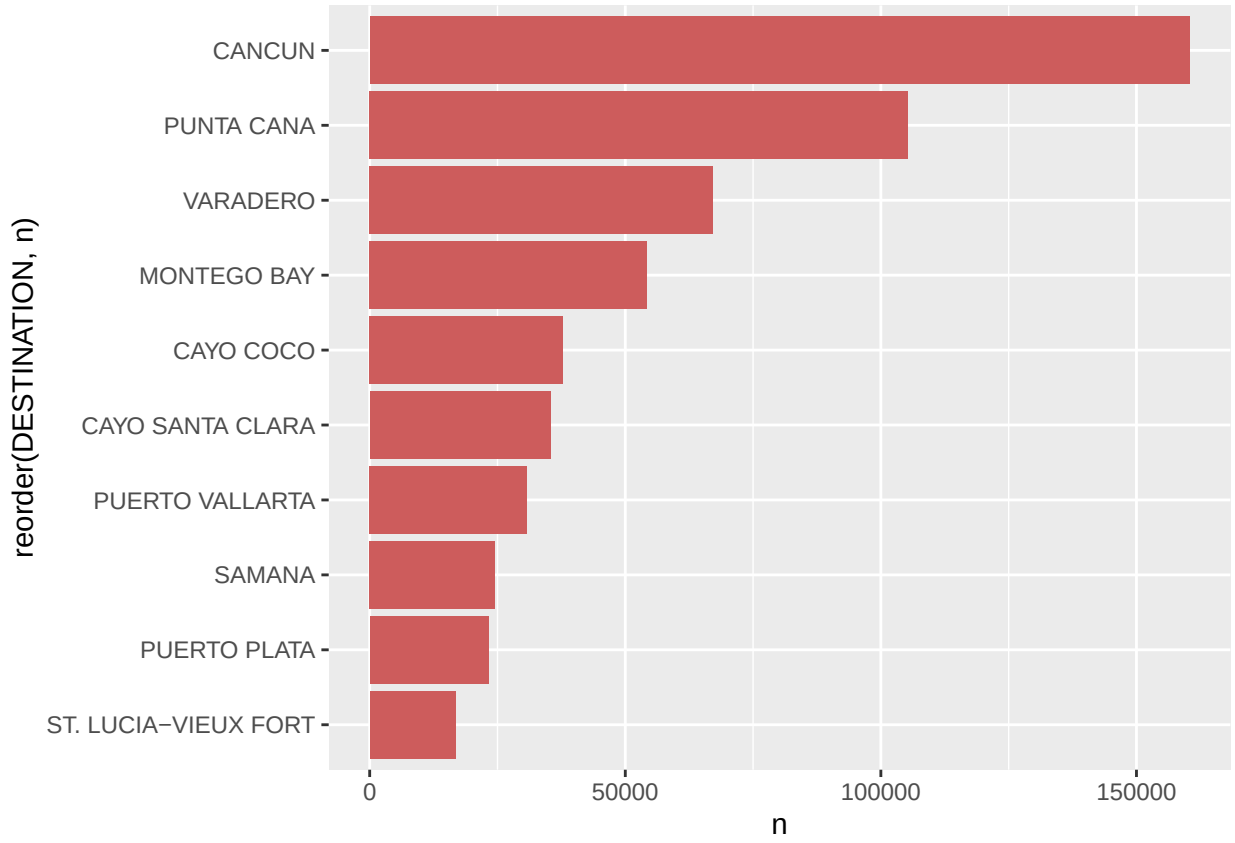
```
dataset_flt <- dataset
write.csv(dataset_flt,'dataset_flt.csv')
write.csv(All_2019,'Vacation_2019.csv')
write.csv(All_2018,'Vacation_2018.csv')
write.csv(All_2017,'Vacation_2017.csv')
write.csv(All_2016,'Vacation_2016.csv')
write.csv(All_2015,'Vacation_2015.csv')
```

Graph Histogram of Popular Destinations per Booking

```
tmp <- dataset %>%
  group_by(DESTINATION) %>%
  summarise(n=n()) %>%
  arrange(desc(n))
tmp <- head(tmp, n=10)
tmp
```

```
## # A tibble: 10 x 2
##    DESTINATION              n
##    <fct>                <int>
##  1 CANCUN              160357
##  2 PUNTA CANA          105358
##  3 VARADERO             67211
##  4 MONTEGO BAY          54226
##  5 CAYO COCO            37801
##  6 CAYO SANTA CLARA     35364
##  7 PUERTO VALLARTA      30780
##  8 SAMANA               24505
##  9 PUERTO PLATA         23255
## 10 ST. LUCIA-VIEUX FORT  16817
```

```
tmp %>%
  ggplot(aes(x=reorder(DESTINATION,n), y=n))+
  geom_bar(stat="identity",fill="indian red")+
  coord_flip()
```



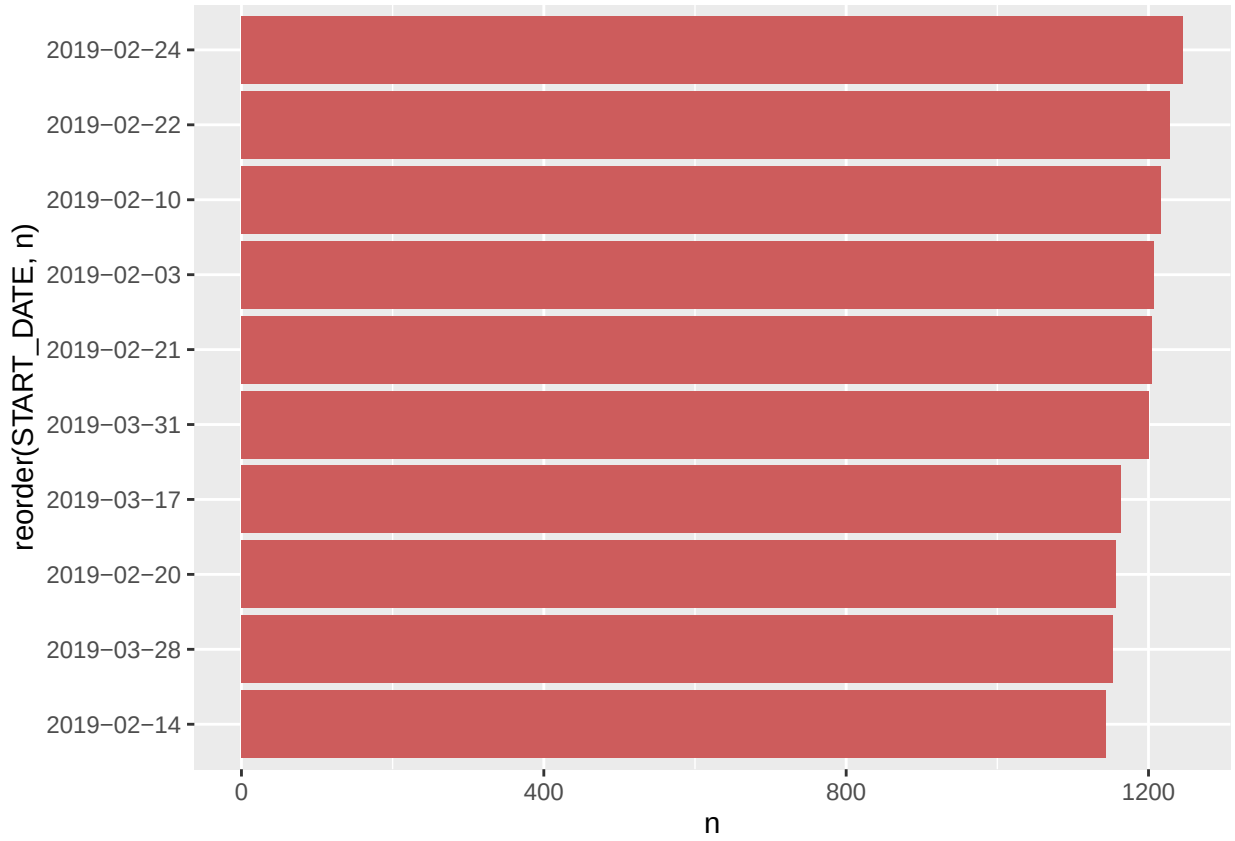
Graph Histogram of Popular Start Date of Bookings

```
tmp <- All_2019 %>%
  group_by(START_DATE) %>%
  summarise(n=n()) %>%
  arrange(desc(n))
tmp <- head(tmp, n=10)
tmp
```

```
## # A tibble: 10 x 2
##    START_DATE     n
##    <date>     <int>
##  1 2019-02-24  1246
##  2 2019-02-22  1229
##  3 2019-02-10  1216
##  4 2019-02-03  1207
##  5 2019-02-21  1204
##  6 2019-03-31  1201
##  7 2019-03-17  1164
##  8 2019-02-20  1157
##  9 2019-03-28  1153
```

```
## 10 2019-02-14  1144
```

```
tmp %>%
  ggplot(aes(x=reorder(START_DATE,n), y=n))+
  geom_bar(stat="identity",fill="indian red")+
  coord_flip()
```



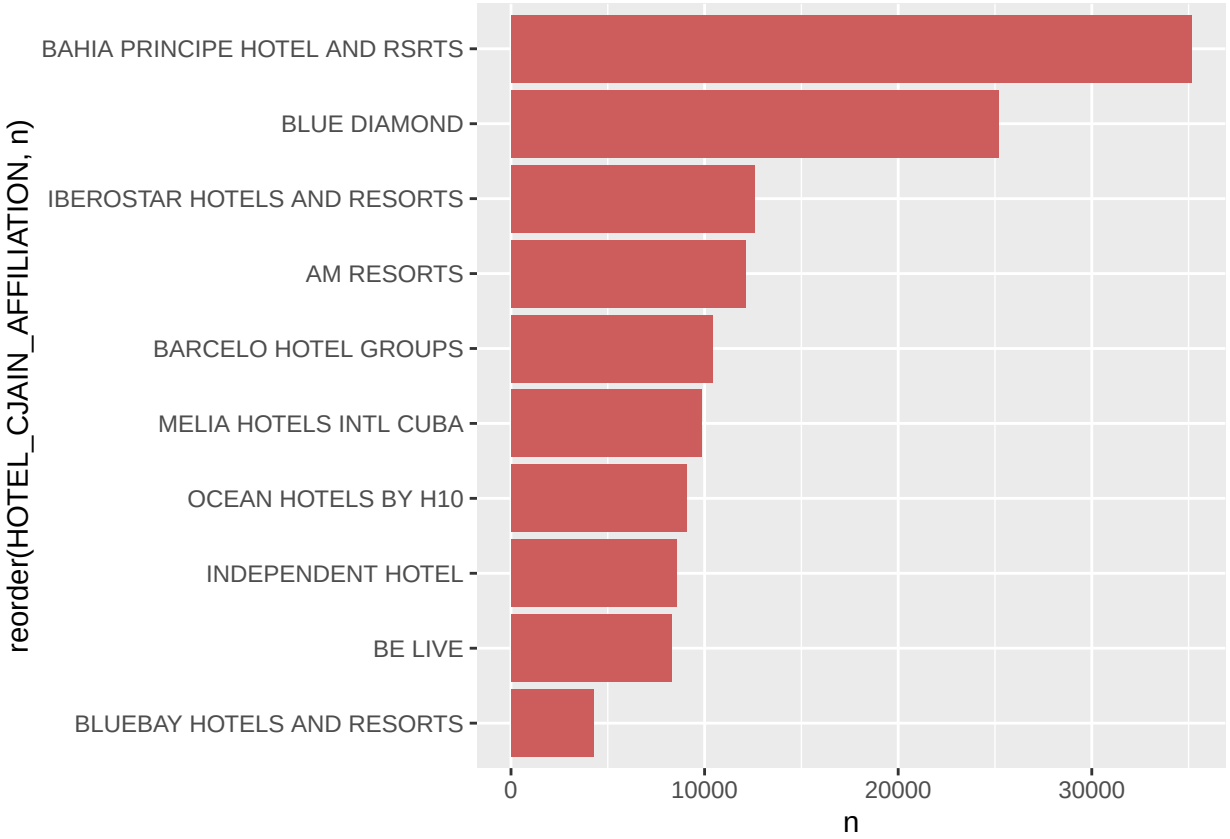
Graph Histogram of Popular Destinations per Booking

```
tmp <- All_2019 %>%
  group_by(HOTEL_CJAIN_AFFILIATION) %>%
  summarise(n=n()) %>%
  arrange(desc(n))
tmp <- head(tmp, n=10)
tmp
```

```
## # A tibble: 10 x 2
##    HOTEL_CJAIN_AFFILIATION            n
##    <fct>                          <int>
##  1 BAHIA PRINCIPE HOTEL AND RSRTS 35144
##  2 BLUE DIAMOND                   25206
##  3 IBEROSTAR HOTELS AND RESORTS   12585
##  4 AM RESORTS                     12133
##  5 BARCELO HOTEL GROUPS           10431
##  6 MELIA HOTELS INTL CUBA          9819
##  7 OCEAN HOTELS BY H10             9044
```

```
##  8 INDEPENDENT HOTEL               8539
##  9 BE LIVE                         8287
## 10 BLUEBAY HOTELS AND RESORTS      4279
```

```
tmp %>%
  ggplot(aes(x=reorder(HOTEL_CJAIN_AFFILIATION,n), y=n))+
  geom_bar(stat="identity",fill="indian red")+
  coord_flip()
```



4. Data Modeling

- Ethical Framework Questions:
  - Does your use case require a more interpretable algorithm? No, the algorithm that we are using is interpretable for the audience that we are targeting.
  - Should you be optimizing for a different outcome than accuracy to make your outcomes fairer? No, as this is something that affects potential profit for companies accuracy is very important.
  - Is it possible that a malicious actor has compromised training data and created misleading results? We thoroughly went through the data and made sure that any outliers or suspicious entries were dealt with accordingly.

4. A) Data Modeling Survival Analysis

Build a Time to Event Model based on all values to start.

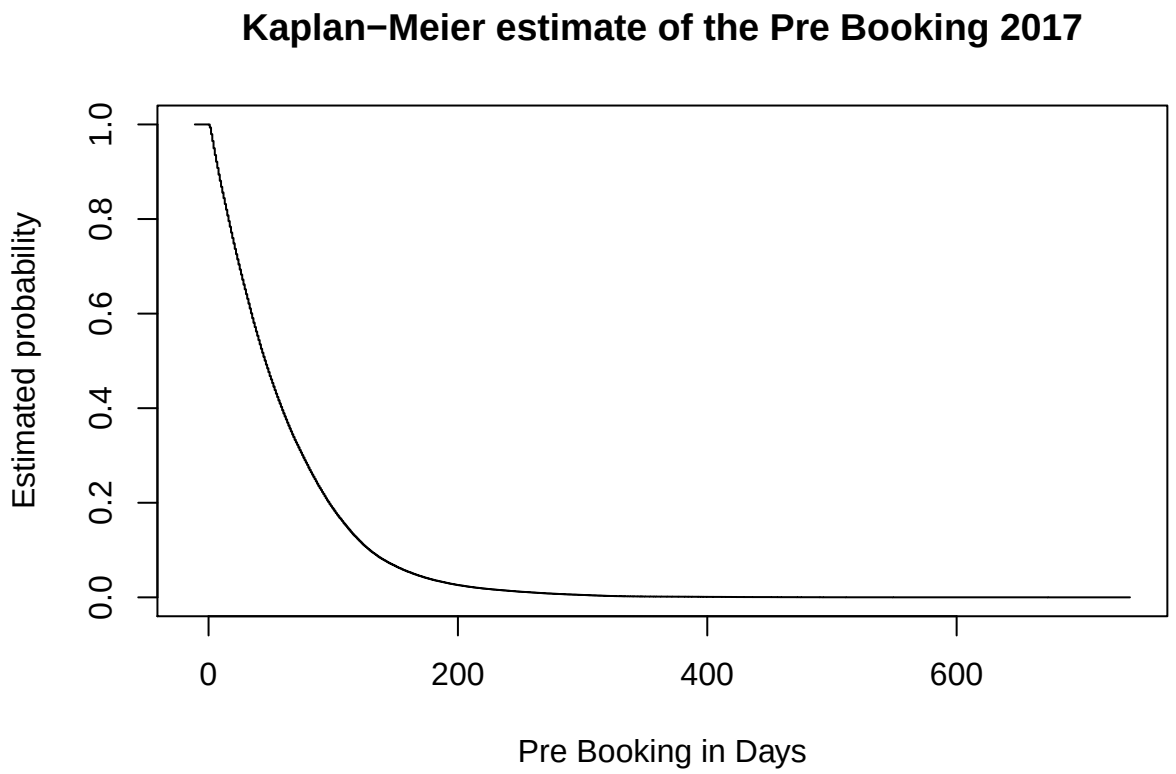
This gives us a model that is reponsible to giving us optimal time to book a hotel

Kaplan-Meier Model

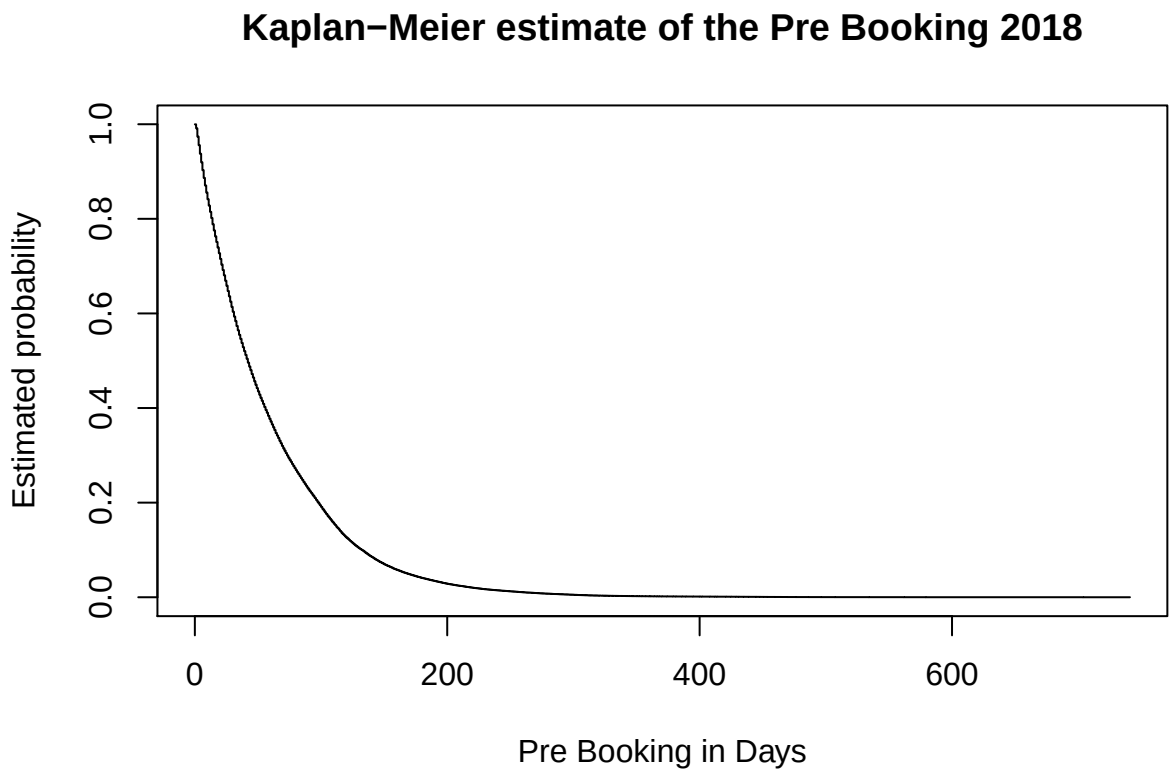
According to IBM The Kaplan-Meier procedure is a method of estimating time-to-event models in the presence of censored cases. The Kaplan-Meier model is based on estimating conditional probabilities at each time point when an event occurs and taking the product limit of those probabilities to estimate the survival rate at each point in time.

We perfomed Kaplan-Meier Survival Analysis to identofy the period where the agent shall

```
km_2017 <- survfit(Surv(All_2017$TTE)~1)
#summary(km_2017)
plot(km_2017,conf.int=FALSE, mark.time=TRUE,main="Kaplan-Meier estimate of the Pre Booking 2017", xlab=
```




```
km_2018 <- survfit(Surv(All_2018$TTE)~1)
#summary(km_2018)
plot(km_2018,conf.int=FALSE, mark.time=TRUE,main="Kaplan-Meier estimate of the Pre Booking 2018", xlab=
```




```
km_2019 <- survfit(Surv(All_2019$TTE)~1)
summary(km_2019)
```

```
## Call: survfit(formula = Surv(All_2019$TTE) ~ 1)
##
##  time n.risk n.event survival  std.err lower 95% CI upper 95% CI
##  -284 211361       1 1.00e+00 4.73e-06     1.00e+00     1.00e+00
##     0 211360      52 1.00e+00 3.44e-05     1.00e+00     1.00e+00
##     1 211308    1465 9.93e-01 1.84e-04     9.92e-01     9.93e-01
##     2 209843    3139 9.78e-01 3.19e-04     9.77e-01     9.79e-01
##     3 206704    3503 9.61e-01 4.19e-04     9.61e-01     9.62e-01
##     4 203201    3808 9.43e-01 5.03e-04     9.42e-01     9.44e-01
##     5 199393    3633 9.26e-01 5.69e-04     9.25e-01     9.27e-01
##     6 195760    3648 9.09e-01 6.26e-04     9.08e-01     9.10e-01
##     7 192112    3525 8.92e-01 6.74e-04     8.91e-01     8.94e-01
##     8 188587    3570 8.75e-01 7.18e-04     8.74e-01     8.77e-01
##     9 185017    3487 8.59e-01 7.57e-04     8.57e-01     8.60e-01
##    10 181530    3259 8.43e-01 7.90e-04     8.42e-01     8.45e-01
##    11 178271    3117 8.29e-01 8.20e-04     8.27e-01     8.30e-01
##    12 175154    3104 8.14e-01 8.46e-04     8.12e-01     8.16e-01
##    13 172050    3132 7.99e-01 8.71e-04     7.97e-01     8.01e-01
##    14 168918    3083 7.85e-01 8.94e-04     7.83e-01     7.86e-01
##    15 165835    2904 7.71e-01 9.14e-04     7.69e-01     7.73e-01
##    16 162931    2879 7.57e-01 9.33e-04     7.55e-01     7.59e-01
##    17 160052    2953 7.43e-01 9.50e-04     7.41e-01     7.45e-01
```

```
##    18 157099    2712 7.30e-01 9.65e-04     7.29e-01     7.32e-01
##    19 154387    2633 7.18e-01 9.79e-04     7.16e-01     7.20e-01
##    20 151754    2696 7.05e-01 9.92e-04     7.03e-01     7.07e-01
##    21 149058    2661 6.93e-01 1.00e-03     6.91e-01     6.95e-01
##    22 146397    2433 6.81e-01 1.01e-03     6.79e-01     6.83e-01
##    23 143964    2458 6.69e-01 1.02e-03     6.67e-01     6.72e-01
##    24 141506    2441 6.58e-01 1.03e-03     6.56e-01     6.60e-01
##    25 139065    2324 6.47e-01 1.04e-03     6.45e-01     6.49e-01
##    26 136741    2388 6.36e-01 1.05e-03     6.34e-01     6.38e-01
##    27 134353    2397 6.24e-01 1.05e-03     6.22e-01     6.26e-01
##    28 131956    2370 6.13e-01 1.06e-03     6.11e-01     6.15e-01
##    29 129586    2394 6.02e-01 1.06e-03     6.00e-01     6.04e-01
##    30 127192    2190 5.91e-01 1.07e-03     5.89e-01     5.94e-01
##    31 125002    2307 5.80e-01 1.07e-03     5.78e-01     5.83e-01
##    32 122695    2206 5.70e-01 1.08e-03     5.68e-01     5.72e-01
##    33 120489    2175 5.60e-01 1.08e-03     5.58e-01     5.62e-01
##    34 118314    2207 5.49e-01 1.08e-03     5.47e-01     5.51e-01
##    35 116107    2146 5.39e-01 1.08e-03     5.37e-01     5.41e-01
##    36 113961    1894 5.30e-01 1.09e-03     5.28e-01     5.32e-01
##    37 112067    2050 5.21e-01 1.09e-03     5.18e-01     5.23e-01
##    38 110017    2034 5.11e-01 1.09e-03     5.09e-01     5.13e-01
##    39 107983    1995 5.01e-01 1.09e-03     4.99e-01     5.04e-01
##    40 105988    2064 4.92e-01 1.09e-03     4.90e-01     4.94e-01
##    41 103924    1952 4.82e-01 1.09e-03     4.80e-01     4.85e-01
##    42 101972    1925 4.73e-01 1.09e-03     4.71e-01     4.75e-01
##    43 100047    1802 4.65e-01 1.08e-03     4.63e-01     4.67e-01
##    44  98245    1672 4.57e-01 1.08e-03     4.55e-01     4.59e-01
##    45  96573    1644 4.49e-01 1.08e-03     4.47e-01     4.51e-01
##    46  94929    1554 4.42e-01 1.08e-03     4.40e-01     4.44e-01
##    47  93375    1668 4.34e-01 1.08e-03     4.32e-01     4.36e-01
##    48  91707    1667 4.26e-01 1.08e-03     4.24e-01     4.28e-01
##    49  90040    1596 4.18e-01 1.07e-03     4.16e-01     4.21e-01
##    50  88444    1547 4.11e-01 1.07e-03     4.09e-01     4.13e-01
##    51  86897    1489 4.04e-01 1.07e-03     4.02e-01     4.06e-01
##    52  85408    1459 3.97e-01 1.06e-03     3.95e-01     3.99e-01
##    53  83949    1474 3.90e-01 1.06e-03     3.88e-01     3.92e-01
##    54  82475    1401 3.84e-01 1.06e-03     3.82e-01     3.86e-01
##    55  81074    1385 3.77e-01 1.05e-03     3.75e-01     3.79e-01
##    56  79689    1271 3.71e-01 1.05e-03     3.69e-01     3.73e-01
##    57  78418    1268 3.65e-01 1.05e-03     3.63e-01     3.67e-01
##    58  77150    1179 3.59e-01 1.04e-03     3.57e-01     3.61e-01
##    59  75971    1206 3.54e-01 1.04e-03     3.52e-01     3.56e-01
##    60  74765    1215 3.48e-01 1.04e-03     3.46e-01     3.50e-01
##    61  73550    1135 3.43e-01 1.03e-03     3.41e-01     3.45e-01
##    62  72415    1150 3.37e-01 1.03e-03     3.35e-01     3.39e-01
##    63  71265    1095 3.32e-01 1.02e-03     3.30e-01     3.34e-01
##    64  70170    1064 3.27e-01 1.02e-03     3.25e-01     3.29e-01
##    65  69106    1108 3.22e-01 1.02e-03     3.20e-01     3.24e-01
##    66  67998    1109 3.16e-01 1.01e-03     3.14e-01     3.18e-01
##    67  66889    1051 3.11e-01 1.01e-03     3.10e-01     3.13e-01
##    68  65838    1067 3.06e-01 1.00e-03     3.04e-01     3.08e-01
##    69  64771     954 3.02e-01 9.99e-04     3.00e-01     3.04e-01
##    70  63817    1010 2.97e-01 9.94e-04     2.95e-01     2.99e-01
##    71  62807     950 2.93e-01 9.90e-04     2.91e-01     2.95e-01
```

```
##    72  61857     945 2.88e-01 9.85e-04     2.86e-01     2.90e-01
##    73  60912     905 2.84e-01 9.81e-04     2.82e-01     2.86e-01
##    74  60007     999 2.79e-01 9.76e-04     2.77e-01     2.81e-01
##    75  59008     917 2.75e-01 9.71e-04     2.73e-01     2.77e-01
##    76  58091     880 2.71e-01 9.66e-04     2.69e-01     2.73e-01
##    77  57211     890 2.66e-01 9.62e-04     2.65e-01     2.68e-01
##    78  56321     793 2.63e-01 9.57e-04     2.61e-01     2.65e-01
##    79  55528     816 2.59e-01 9.53e-04     2.57e-01     2.61e-01
##    80  54712     811 2.55e-01 9.48e-04     2.53e-01     2.57e-01
##    81  53901     771 2.51e-01 9.44e-04     2.50e-01     2.53e-01
##    82  53130     715 2.48e-01 9.39e-04     2.46e-01     2.50e-01
##    83  52415     782 2.44e-01 9.35e-04     2.42e-01     2.46e-01
##    84  51633     794 2.41e-01 9.30e-04     2.39e-01     2.42e-01
##    85  50839     747 2.37e-01 9.25e-04     2.35e-01     2.39e-01
##    86  50092     717 2.34e-01 9.20e-04     2.32e-01     2.35e-01
##    87  49375     722 2.30e-01 9.16e-04     2.28e-01     2.32e-01
##    88  48653     705 2.27e-01 9.11e-04     2.25e-01     2.29e-01
##    89  47948     649 2.24e-01 9.07e-04     2.22e-01     2.26e-01
##    90  47299     655 2.21e-01 9.02e-04     2.19e-01     2.22e-01
##    91  46644     591 2.18e-01 8.98e-04     2.16e-01     2.20e-01
##    92  46053     619 2.15e-01 8.94e-04     2.13e-01     2.17e-01
##    93  45434     674 2.12e-01 8.89e-04     2.10e-01     2.14e-01
##    94  44760     643 2.09e-01 8.84e-04     2.07e-01     2.10e-01
##    95  44117     637 2.06e-01 8.79e-04     2.04e-01     2.07e-01
##    96  43480     552 2.03e-01 8.75e-04     2.01e-01     2.05e-01
##    97  42928     542 2.01e-01 8.71e-04     1.99e-01     2.02e-01
##    98  42386     565 1.98e-01 8.67e-04     1.96e-01     2.00e-01
##    99  41821     578 1.95e-01 8.62e-04     1.93e-01     1.97e-01
##   100  41243     641 1.92e-01 8.57e-04     1.90e-01     1.94e-01
##   101  40602     603 1.89e-01 8.52e-04     1.88e-01     1.91e-01
##   102  39999     605 1.86e-01 8.47e-04     1.85e-01     1.88e-01
##   103  39394     602 1.84e-01 8.42e-04     1.82e-01     1.85e-01
##   104  38792     560 1.81e-01 8.37e-04     1.79e-01     1.83e-01
##   105  38232     607 1.78e-01 8.32e-04     1.76e-01     1.80e-01
##   106  37625     599 1.75e-01 8.27e-04     1.74e-01     1.77e-01
##   107  37026     622 1.72e-01 8.21e-04     1.71e-01     1.74e-01
##   108  36404     631 1.69e-01 8.16e-04     1.68e-01     1.71e-01
##   109  35773     651 1.66e-01 8.10e-04     1.65e-01     1.68e-01
##   110  35122     536 1.64e-01 8.05e-04     1.62e-01     1.65e-01
##   111  34586     611 1.61e-01 7.99e-04     1.59e-01     1.62e-01
##   112  33975     633 1.58e-01 7.93e-04     1.56e-01     1.59e-01
##   113  33342     563 1.55e-01 7.87e-04     1.54e-01     1.57e-01
##   114  32779     611 1.52e-01 7.81e-04     1.51e-01     1.54e-01
##   115  32168     658 1.49e-01 7.75e-04     1.48e-01     1.51e-01
##   116  31510     614 1.46e-01 7.68e-04     1.45e-01     1.48e-01
##   117  30896     562 1.44e-01 7.63e-04     1.42e-01     1.45e-01
##   118  30334     565 1.41e-01 7.57e-04     1.39e-01     1.42e-01
##   119  29769     589 1.38e-01 7.50e-04     1.37e-01     1.40e-01
##   120  29180     611 1.35e-01 7.44e-04     1.34e-01     1.37e-01
##   121  28569     625 1.32e-01 7.37e-04     1.31e-01     1.34e-01
##   122  27944     511 1.30e-01 7.31e-04     1.28e-01     1.31e-01
##   123  27433     540 1.27e-01 7.25e-04     1.26e-01     1.29e-01
##   124  26893     568 1.25e-01 7.18e-04     1.23e-01     1.26e-01
##   125  26325     496 1.22e-01 7.12e-04     1.21e-01     1.24e-01
```

```
##   126  25829     536 1.20e-01 7.06e-04     1.18e-01     1.21e-01
##   127  25293     510 1.17e-01 7.00e-04     1.16e-01     1.19e-01
##   128  24783     531 1.15e-01 6.93e-04     1.13e-01     1.16e-01
##   129  24252     488 1.12e-01 6.87e-04     1.11e-01     1.14e-01
##   130  23764     451 1.10e-01 6.81e-04     1.09e-01     1.12e-01
##   131  23313     450 1.08e-01 6.76e-04     1.07e-01     1.10e-01
##   132  22863     386 1.06e-01 6.71e-04     1.05e-01     1.08e-01
##   133  22477     412 1.04e-01 6.65e-04     1.03e-01     1.06e-01
##   134  22065     417 1.02e-01 6.60e-04     1.01e-01     1.04e-01
##   135  21648     398 1.01e-01 6.54e-04     9.93e-02     1.02e-01
##   136  21250     408 9.86e-02 6.48e-04     9.73e-02     9.99e-02
##   137  20842     427 9.66e-02 6.43e-04     9.53e-02     9.79e-02
##   138  20415     351 9.49e-02 6.38e-04     9.37e-02     9.62e-02
##   139  20064     342 9.33e-02 6.33e-04     9.21e-02     9.46e-02
##   140  19722     333 9.17e-02 6.28e-04     9.05e-02     9.30e-02
##   141  19389     324 9.02e-02 6.23e-04     8.90e-02     9.14e-02
##   142  19065     347 8.86e-02 6.18e-04     8.74e-02     8.98e-02
##   143  18718     358 8.69e-02 6.13e-04     8.57e-02     8.81e-02
##   144  18360     329 8.53e-02 6.08e-04     8.41e-02     8.65e-02
##   145  18031     275 8.40e-02 6.03e-04     8.28e-02     8.52e-02
##   146  17756     299 8.26e-02 5.99e-04     8.14e-02     8.38e-02
##   147  17457     303 8.12e-02 5.94e-04     8.00e-02     8.23e-02
##   148  17154     309 7.97e-02 5.89e-04     7.86e-02     8.09e-02
##   149  16845     306 7.83e-02 5.84e-04     7.71e-02     7.94e-02
##   150  16539     314 7.68e-02 5.79e-04     7.56e-02     7.79e-02
##   151  16225     294 7.54e-02 5.74e-04     7.43e-02     7.65e-02
##   152  15931     240 7.42e-02 5.70e-04     7.31e-02     7.54e-02
##   153  15691     272 7.30e-02 5.66e-04     7.19e-02     7.41e-02
##   154  15419     242 7.18e-02 5.62e-04     7.07e-02     7.29e-02
##   155  15177     250 7.06e-02 5.57e-04     6.95e-02     7.17e-02
##   156  14927     249 6.94e-02 5.53e-04     6.84e-02     7.05e-02
##   157  14678     240 6.83e-02 5.49e-04     6.72e-02     6.94e-02
##   158  14438     222 6.73e-02 5.45e-04     6.62e-02     6.83e-02
##   159  14216     216 6.62e-02 5.41e-04     6.52e-02     6.73e-02
##   160  14000     228 6.52e-02 5.37e-04     6.41e-02     6.62e-02
##   161  13772     211 6.42e-02 5.33e-04     6.31e-02     6.52e-02
##   162  13561     215 6.31e-02 5.29e-04     6.21e-02     6.42e-02
##   163  13346     223 6.21e-02 5.25e-04     6.11e-02     6.31e-02
##   164  13123     218 6.11e-02 5.21e-04     6.00e-02     6.21e-02
##   165  12905     228 6.00e-02 5.16e-04     5.90e-02     6.10e-02
##   166  12677     199 5.90e-02 5.13e-04     5.80e-02     6.00e-02
##   167  12478     205 5.81e-02 5.09e-04     5.71e-02     5.91e-02
##   168  12273     192 5.72e-02 5.05e-04     5.62e-02     5.82e-02
##   169  12081     217 5.61e-02 5.01e-04     5.52e-02     5.71e-02
##   170  11864     193 5.52e-02 4.97e-04     5.43e-02     5.62e-02
##   171  11671     189 5.43e-02 4.93e-04     5.34e-02     5.53e-02
##   172  11482     174 5.35e-02 4.89e-04     5.26e-02     5.45e-02
##   173  11308     174 5.27e-02 4.86e-04     5.17e-02     5.36e-02
##   174  11134     179 5.18e-02 4.82e-04     5.09e-02     5.28e-02
##   175  10955     165 5.11e-02 4.79e-04     5.01e-02     5.20e-02
##   176  10790     189 5.02e-02 4.75e-04     4.92e-02     5.11e-02
##   177  10601     199 4.92e-02 4.71e-04     4.83e-02     5.01e-02
##   178  10402     179 4.84e-02 4.67e-04     4.75e-02     4.93e-02
##   179  10223     163 4.76e-02 4.63e-04     4.67e-02     4.85e-02
```

```
##   180  10060     175 4.68e-02 4.59e-04     4.59e-02     4.77e-02
##   181   9885     148 4.61e-02 4.56e-04     4.52e-02     4.70e-02
##   182   9737     181 4.52e-02 4.52e-04     4.43e-02     4.61e-02
##   183   9556     173 4.44e-02 4.48e-04     4.35e-02     4.53e-02
##   184   9383     164 4.36e-02 4.44e-04     4.28e-02     4.45e-02
##   185   9219     159 4.29e-02 4.41e-04     4.20e-02     4.37e-02
##   186   9060     166 4.21e-02 4.37e-04     4.12e-02     4.29e-02
##   187   8894     153 4.14e-02 4.33e-04     4.05e-02     4.22e-02
##   188   8741     131 4.07e-02 4.30e-04     3.99e-02     4.16e-02
##   189   8610     122 4.02e-02 4.27e-04     3.93e-02     4.10e-02
##   190   8488     150 3.94e-02 4.23e-04     3.86e-02     4.03e-02
##   191   8338     130 3.88e-02 4.20e-04     3.80e-02     3.97e-02
##   192   8208     162 3.81e-02 4.16e-04     3.73e-02     3.89e-02
##   193   8046     136 3.74e-02 4.13e-04     3.66e-02     3.82e-02
##   194   7910      86 3.70e-02 4.11e-04     3.62e-02     3.78e-02
##   195   7824     134 3.64e-02 4.07e-04     3.56e-02     3.72e-02
##   196   7690     129 3.58e-02 4.04e-04     3.50e-02     3.66e-02
##   197   7561     127 3.52e-02 4.01e-04     3.44e-02     3.60e-02
##   198   7434     133 3.45e-02 3.97e-04     3.38e-02     3.53e-02
##   199   7301     123 3.40e-02 3.94e-04     3.32e-02     3.47e-02
##   200   7178     102 3.35e-02 3.91e-04     3.27e-02     3.43e-02
##   201   7076     148 3.28e-02 3.87e-04     3.20e-02     3.35e-02
##   202   6928     106 3.23e-02 3.84e-04     3.15e-02     3.30e-02
##   203   6822     101 3.18e-02 3.82e-04     3.11e-02     3.26e-02
##   204   6721     135 3.12e-02 3.78e-04     3.04e-02     3.19e-02
##   205   6586     115 3.06e-02 3.75e-04     2.99e-02     3.14e-02
##   206   6471     130 3.00e-02 3.71e-04     2.93e-02     3.07e-02
##   207   6341     117 2.94e-02 3.68e-04     2.87e-02     3.02e-02
##   208   6224     126 2.89e-02 3.64e-04     2.81e-02     2.96e-02
##   209   6098     110 2.83e-02 3.61e-04     2.76e-02     2.90e-02
##   210   5988      86 2.79e-02 3.58e-04     2.72e-02     2.86e-02
##   211   5902      99 2.75e-02 3.55e-04     2.68e-02     2.82e-02
##   212   5803     106 2.70e-02 3.52e-04     2.63e-02     2.77e-02
##   213   5697      94 2.65e-02 3.49e-04     2.58e-02     2.72e-02
##   214   5603     120 2.59e-02 3.46e-04     2.53e-02     2.66e-02
##   215   5483      87 2.55e-02 3.43e-04     2.49e-02     2.62e-02
##   216   5396      80 2.52e-02 3.41e-04     2.45e-02     2.58e-02
##   217   5316      93 2.47e-02 3.38e-04     2.41e-02     2.54e-02
##   218   5223      99 2.42e-02 3.35e-04     2.36e-02     2.49e-02
##   219   5124     108 2.37e-02 3.31e-04     2.31e-02     2.44e-02
##   220   5016      91 2.33e-02 3.28e-04     2.27e-02     2.40e-02
##   221   4925      98 2.28e-02 3.25e-04     2.22e-02     2.35e-02
##   222   4827      79 2.25e-02 3.22e-04     2.18e-02     2.31e-02
##   223   4748      81 2.21e-02 3.20e-04     2.15e-02     2.27e-02
##   224   4667      64 2.18e-02 3.17e-04     2.12e-02     2.24e-02
##   225   4603      83 2.14e-02 3.15e-04     2.08e-02     2.20e-02
##   226   4520      67 2.11e-02 3.12e-04     2.05e-02     2.17e-02
##   227   4453      95 2.06e-02 3.09e-04     2.00e-02     2.12e-02
##   228   4358      67 2.03e-02 3.07e-04     1.97e-02     2.09e-02
##   229   4291      51 2.01e-02 3.05e-04     1.95e-02     2.07e-02
##   230   4240      54 1.98e-02 3.03e-04     1.92e-02     2.04e-02
##   231   4186      54 1.95e-02 3.01e-04     1.90e-02     2.01e-02
##   232   4132      64 1.92e-02 2.99e-04     1.87e-02     1.98e-02
##   233   4068      57 1.90e-02 2.97e-04     1.84e-02     1.96e-02
```

```
##   234   4011      64 1.87e-02 2.94e-04     1.81e-02     1.93e-02
##   235   3947      73 1.83e-02 2.92e-04     1.78e-02     1.89e-02
##   236   3874      78 1.80e-02 2.89e-04     1.74e-02     1.85e-02
##   237   3796      62 1.77e-02 2.87e-04     1.71e-02     1.82e-02
##   238   3734      52 1.74e-02 2.85e-04     1.69e-02     1.80e-02
##   239   3682      63 1.71e-02 2.82e-04     1.66e-02     1.77e-02
##   240   3619      59 1.68e-02 2.80e-04     1.63e-02     1.74e-02
##   241   3560      51 1.66e-02 2.78e-04     1.61e-02     1.72e-02
##   242   3509      57 1.63e-02 2.76e-04     1.58e-02     1.69e-02
##   243   3452      62 1.60e-02 2.73e-04     1.55e-02     1.66e-02
##   244   3390      54 1.58e-02 2.71e-04     1.53e-02     1.63e-02
##   245   3336      48 1.56e-02 2.69e-04     1.50e-02     1.61e-02
##   246   3288      61 1.53e-02 2.67e-04     1.48e-02     1.58e-02
##   247   3227      64 1.50e-02 2.64e-04     1.45e-02     1.55e-02
##   248   3163      68 1.46e-02 2.61e-04     1.41e-02     1.52e-02
##   249   3095      50 1.44e-02 2.59e-04     1.39e-02     1.49e-02
##   250   3045      46 1.42e-02 2.57e-04     1.37e-02     1.47e-02
##   251   2999      47 1.40e-02 2.55e-04     1.35e-02     1.45e-02
##   252   2952      49 1.37e-02 2.53e-04     1.32e-02     1.42e-02
##   253   2903      55 1.35e-02 2.51e-04     1.30e-02     1.40e-02
##   254   2848      59 1.32e-02 2.48e-04     1.27e-02     1.37e-02
##   255   2789      48 1.30e-02 2.46e-04     1.25e-02     1.35e-02
##   256   2741      45 1.28e-02 2.44e-04     1.23e-02     1.32e-02
##   257   2696      39 1.26e-02 2.42e-04     1.21e-02     1.31e-02
##   258   2657      40 1.24e-02 2.41e-04     1.19e-02     1.29e-02
##   259   2617      48 1.22e-02 2.38e-04     1.17e-02     1.26e-02
##   260   2569      29 1.20e-02 2.37e-04     1.16e-02     1.25e-02
##   261   2540      51 1.18e-02 2.35e-04     1.13e-02     1.22e-02
##   262   2489      49 1.15e-02 2.32e-04     1.11e-02     1.20e-02
##   263   2440      42 1.13e-02 2.30e-04     1.09e-02     1.18e-02
##   264   2398      32 1.12e-02 2.29e-04     1.08e-02     1.17e-02
##   265   2366      26 1.11e-02 2.28e-04     1.06e-02     1.15e-02
##   266   2340      33 1.09e-02 2.26e-04     1.05e-02     1.14e-02
##   267   2307      45 1.07e-02 2.24e-04     1.03e-02     1.11e-02
##   268   2262      36 1.05e-02 2.22e-04     1.01e-02     1.10e-02
##   269   2226      38 1.04e-02 2.20e-04     9.93e-03     1.08e-02
##   270   2188      35 1.02e-02 2.18e-04     9.77e-03     1.06e-02
##   271   2153      40 1.00e-02 2.16e-04     9.58e-03     1.04e-02
##   272   2113      32 9.85e-03 2.15e-04     9.43e-03     1.03e-02
##   273   2081      27 9.72e-03 2.13e-04     9.31e-03     1.01e-02
##   274   2054      51 9.48e-03 2.11e-04     9.07e-03     9.90e-03
##   275   2003      40 9.29e-03 2.09e-04     8.89e-03     9.71e-03
##   276   1963      21 9.19e-03 2.08e-04     8.79e-03     9.60e-03
##   277   1942      41 8.99e-03 2.05e-04     8.60e-03     9.41e-03
##   278   1901      35 8.83e-03 2.03e-04     8.44e-03     9.24e-03
##   279   1866      28 8.70e-03 2.02e-04     8.31e-03     9.10e-03
##   280   1838      36 8.53e-03 2.00e-04     8.14e-03     8.93e-03
##   281   1802      27 8.40e-03 1.98e-04     8.02e-03     8.80e-03
##   282   1775      33 8.24e-03 1.97e-04     7.87e-03     8.64e-03
##   283   1742      34 8.08e-03 1.95e-04     7.71e-03     8.47e-03
##   284   1708      35 7.92e-03 1.93e-04     7.55e-03     8.30e-03
##   285   1673      22 7.81e-03 1.91e-04     7.44e-03     8.20e-03
##   286   1651      22 7.71e-03 1.90e-04     7.34e-03     8.09e-03
##   287   1629      27 7.58e-03 1.89e-04     7.22e-03     7.96e-03
```

```
##   288   1602      31 7.43e-03 1.87e-04     7.08e-03     7.81e-03
##   289   1571      44 7.22e-03 1.84e-04     6.87e-03     7.59e-03
##   290   1527      22 7.12e-03 1.83e-04     6.77e-03     7.49e-03
##   291   1505      22 7.02e-03 1.82e-04     6.67e-03     7.38e-03
##   292   1483      33 6.86e-03 1.80e-04     6.52e-03     7.22e-03
##   293   1450      28 6.73e-03 1.78e-04     6.39e-03     7.09e-03
##   294   1422      26 6.60e-03 1.76e-04     6.27e-03     6.96e-03
##   295   1396      28 6.47e-03 1.74e-04     6.14e-03     6.82e-03
##   296   1368      20 6.38e-03 1.73e-04     6.05e-03     6.73e-03
##   297   1348      21 6.28e-03 1.72e-04     5.95e-03     6.62e-03
##   298   1327      39 6.09e-03 1.69e-04     5.77e-03     6.43e-03
##   299   1288      24 5.98e-03 1.68e-04     5.66e-03     6.32e-03
##   300   1264      22 5.88e-03 1.66e-04     5.56e-03     6.21e-03
##   301   1242      12 5.82e-03 1.65e-04     5.50e-03     6.15e-03
##   302   1230      28 5.69e-03 1.64e-04     5.38e-03     6.02e-03
##   303   1202      30 5.55e-03 1.62e-04     5.24e-03     5.87e-03
##   304   1172      24 5.43e-03 1.60e-04     5.13e-03     5.75e-03
##   305   1148      23 5.32e-03 1.58e-04     5.02e-03     5.64e-03
##   306   1125      24 5.21e-03 1.57e-04     4.91e-03     5.53e-03
##   307   1101      15 5.14e-03 1.56e-04     4.84e-03     5.45e-03
##   308   1086      26 5.02e-03 1.54e-04     4.72e-03     5.33e-03
##   309   1060      40 4.83e-03 1.51e-04     4.54e-03     5.13e-03
##   310   1020      18 4.74e-03 1.49e-04     4.46e-03     5.04e-03
##   311   1002      30 4.60e-03 1.47e-04     4.32e-03     4.90e-03
##   312    972      32 4.45e-03 1.45e-04     4.17e-03     4.74e-03
##   313    940      23 4.34e-03 1.43e-04     4.07e-03     4.63e-03
##   314    917      13 4.28e-03 1.42e-04     4.01e-03     4.56e-03
##   315    904      22 4.17e-03 1.40e-04     3.91e-03     4.46e-03
##   316    882      17 4.09e-03 1.39e-04     3.83e-03     4.37e-03
##   317    865      27 3.96e-03 1.37e-04     3.71e-03     4.24e-03
##   318    838      19 3.87e-03 1.35e-04     3.62e-03     4.15e-03
##   319    819      17 3.79e-03 1.34e-04     3.54e-03     4.07e-03
##   320    802      16 3.72e-03 1.32e-04     3.47e-03     3.99e-03
##   321    786      24 3.61e-03 1.30e-04     3.36e-03     3.87e-03
##   322    762      17 3.52e-03 1.29e-04     3.28e-03     3.79e-03
##   323    745      17 3.44e-03 1.27e-04     3.20e-03     3.70e-03
##   324    728      14 3.38e-03 1.26e-04     3.14e-03     3.63e-03
##   325    714      21 3.28e-03 1.24e-04     3.04e-03     3.53e-03
##   326    693      13 3.22e-03 1.23e-04     2.98e-03     3.47e-03
##   327    680      11 3.17e-03 1.22e-04     2.93e-03     3.41e-03
##   328    669      26 3.04e-03 1.20e-04     2.82e-03     3.29e-03
##   329    643      14 2.98e-03 1.18e-04     2.75e-03     3.22e-03
##   330    629       5 2.95e-03 1.18e-04     2.73e-03     3.19e-03
##   331    624       6 2.92e-03 1.17e-04     2.70e-03     3.16e-03
##   332    618      10 2.88e-03 1.16e-04     2.66e-03     3.11e-03
##   333    608      12 2.82e-03 1.15e-04     2.60e-03     3.06e-03
##   334    596       7 2.79e-03 1.15e-04     2.57e-03     3.02e-03
##   335    589       8 2.75e-03 1.14e-04     2.53e-03     2.98e-03
##   336    581       5 2.73e-03 1.13e-04     2.51e-03     2.96e-03
##   337    576      17 2.64e-03 1.12e-04     2.43e-03     2.87e-03
##   338    559      12 2.59e-03 1.11e-04     2.38e-03     2.81e-03
##   339    547       8 2.55e-03 1.10e-04     2.34e-03     2.77e-03
##   340    539       5 2.53e-03 1.09e-04     2.32e-03     2.75e-03
##   341    534       5 2.50e-03 1.09e-04     2.30e-03     2.73e-03
```

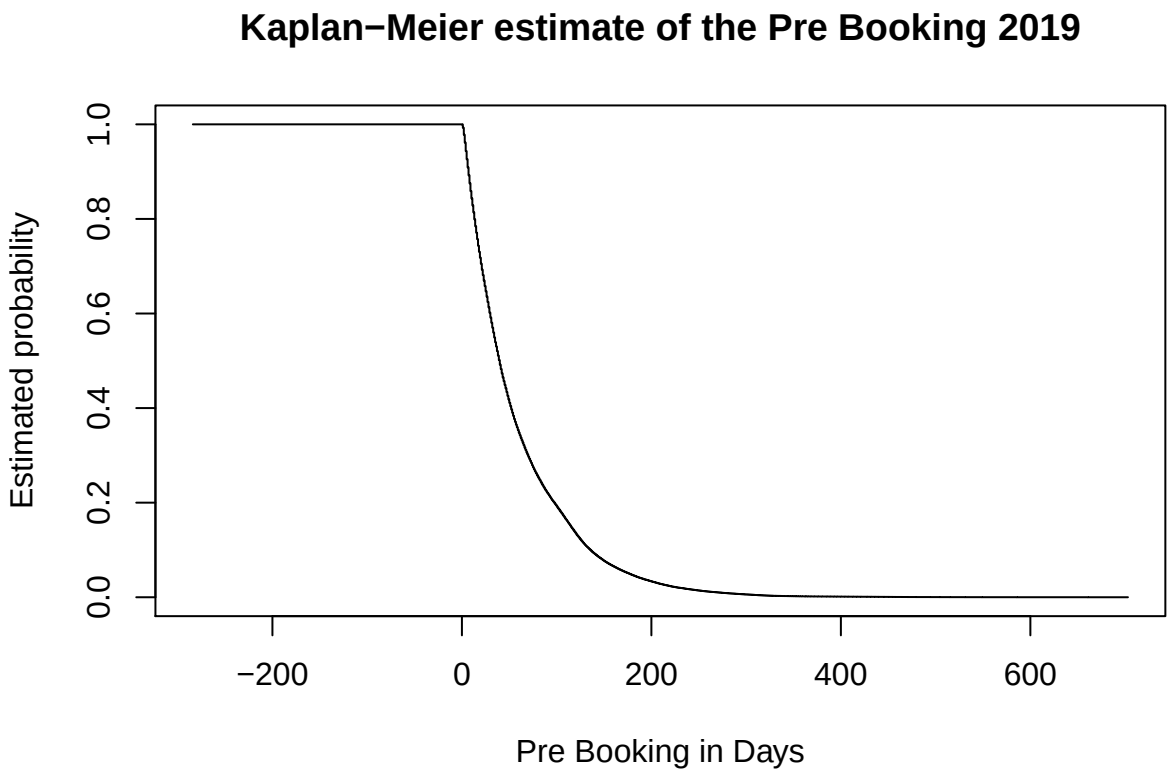
```
##   342    529       1 2.50e-03 1.09e-04     2.29e-03     2.72e-03
##   343    528       8 2.46e-03 1.08e-04     2.26e-03     2.68e-03
##   344    520      11 2.41e-03 1.07e-04     2.21e-03     2.63e-03
##   345    509       4 2.39e-03 1.06e-04     2.19e-03     2.61e-03
##   346    505       7 2.36e-03 1.05e-04     2.16e-03     2.57e-03
##   347    498       7 2.32e-03 1.05e-04     2.13e-03     2.54e-03
##   348    491       4 2.30e-03 1.04e-04     2.11e-03     2.52e-03
##   349    487       7 2.27e-03 1.04e-04     2.08e-03     2.48e-03
##   350    480       2 2.26e-03 1.03e-04     2.07e-03     2.47e-03
##   351    478       6 2.23e-03 1.03e-04     2.04e-03     2.44e-03
##   352    472       9 2.19e-03 1.02e-04     2.00e-03     2.40e-03
##   353    463      10 2.14e-03 1.01e-04     1.95e-03     2.35e-03
##   354    453       6 2.11e-03 9.99e-05     1.93e-03     2.32e-03
##   355    447       9 2.07e-03 9.89e-05     1.89e-03     2.28e-03
##   356    438       3 2.06e-03 9.86e-05     1.87e-03     2.26e-03
##   357    435       5 2.03e-03 9.80e-05     1.85e-03     2.24e-03
##   358    430       2 2.02e-03 9.78e-05     1.84e-03     2.23e-03
##   359    428       9 1.98e-03 9.68e-05     1.80e-03     2.18e-03
##   360    419       6 1.95e-03 9.61e-05     1.77e-03     2.15e-03
##   361    413       5 1.93e-03 9.55e-05     1.75e-03     2.13e-03
##   362    408       9 1.89e-03 9.44e-05     1.71e-03     2.08e-03
##   363    399       6 1.86e-03 9.37e-05     1.68e-03     2.05e-03
##   364    393       5 1.84e-03 9.31e-05     1.66e-03     2.03e-03
##   365    388       6 1.81e-03 9.24e-05     1.64e-03     2.00e-03
##   366    382       5 1.78e-03 9.18e-05     1.61e-03     1.97e-03
##   367    377       5 1.76e-03 9.12e-05     1.59e-03     1.95e-03
##   368    372       2 1.75e-03 9.09e-05     1.58e-03     1.94e-03
##   369    370       1 1.75e-03 9.08e-05     1.58e-03     1.93e-03
##   370    369       5 1.72e-03 9.02e-05     1.55e-03     1.91e-03
##   371    364       3 1.71e-03 8.98e-05     1.54e-03     1.89e-03
##   372    361       4 1.69e-03 8.93e-05     1.52e-03     1.87e-03
##   373    357       3 1.67e-03 8.89e-05     1.51e-03     1.86e-03
##   374    354       1 1.67e-03 8.88e-05     1.50e-03     1.85e-03
##   375    353       4 1.65e-03 8.83e-05     1.49e-03     1.83e-03
##   376    349       4 1.63e-03 8.78e-05     1.47e-03     1.81e-03
##   377    345       1 1.63e-03 8.77e-05     1.46e-03     1.81e-03
##   379    344       3 1.61e-03 8.73e-05     1.45e-03     1.79e-03
##   380    341       6 1.58e-03 8.65e-05     1.42e-03     1.76e-03
##   381    335       3 1.57e-03 8.61e-05     1.41e-03     1.75e-03
##   382    332       4 1.55e-03 8.56e-05     1.39e-03     1.73e-03
##   383    328       2 1.54e-03 8.54e-05     1.38e-03     1.72e-03
##   385    326       3 1.53e-03 8.50e-05     1.37e-03     1.70e-03
##   386    323       1 1.52e-03 8.48e-05     1.37e-03     1.70e-03
##   387    322       3 1.51e-03 8.44e-05     1.35e-03     1.68e-03
##   388    319       5 1.49e-03 8.38e-05     1.33e-03     1.66e-03
##   389    314       8 1.45e-03 8.27e-05     1.29e-03     1.62e-03
##   390    306       2 1.44e-03 8.24e-05     1.29e-03     1.61e-03
##   391    304       2 1.43e-03 8.22e-05     1.28e-03     1.60e-03
##   392    302       2 1.42e-03 8.19e-05     1.27e-03     1.59e-03
##   393    300       3 1.41e-03 8.15e-05     1.25e-03     1.57e-03
##   394    297       4 1.39e-03 8.09e-05     1.24e-03     1.55e-03
##   395    293       4 1.37e-03 8.04e-05     1.22e-03     1.53e-03
##   396    289       3 1.35e-03 8.00e-05     1.21e-03     1.52e-03
##   397    286       2 1.34e-03 7.97e-05     1.20e-03     1.51e-03
```

```
##   398    284       3 1.33e-03 7.93e-05     1.18e-03     1.49e-03
##   400    281       5 1.31e-03 7.85e-05     1.16e-03     1.47e-03
##   401    276       5 1.28e-03 7.78e-05     1.14e-03     1.44e-03
##   402    271       3 1.27e-03 7.74e-05     1.12e-03     1.43e-03
##   403    268       3 1.25e-03 7.70e-05     1.11e-03     1.41e-03
##   404    265       2 1.24e-03 7.67e-05     1.10e-03     1.40e-03
##   405    263       2 1.23e-03 7.64e-05     1.09e-03     1.39e-03
##   406    261       3 1.22e-03 7.59e-05     1.08e-03     1.38e-03
##   407    258       1 1.22e-03 7.58e-05     1.08e-03     1.37e-03
##   408    257       5 1.19e-03 7.51e-05     1.05e-03     1.35e-03
##   409    252       3 1.18e-03 7.46e-05     1.04e-03     1.33e-03
##   410    249       1 1.17e-03 7.45e-05     1.04e-03     1.33e-03
##   411    248       2 1.16e-03 7.42e-05     1.03e-03     1.32e-03
##   412    246       3 1.15e-03 7.37e-05     1.01e-03     1.30e-03
##   413    243       6 1.12e-03 7.28e-05     9.87e-04     1.27e-03
##   414    237       3 1.11e-03 7.23e-05     9.74e-04     1.26e-03
##   415    234       3 1.09e-03 7.19e-05     9.61e-04     1.24e-03
##   416    231       2 1.08e-03 7.16e-05     9.52e-04     1.23e-03
##   417    229       7 1.05e-03 7.05e-05     9.21e-04     1.20e-03
##   418    222       1 1.05e-03 7.03e-05     9.17e-04     1.19e-03
##   419    221       2 1.04e-03 7.00e-05     9.08e-04     1.18e-03
##   420    219       1 1.03e-03 6.98e-05     9.03e-04     1.18e-03
##   421    218       3 1.02e-03 6.93e-05     8.90e-04     1.16e-03
##   422    215       2 1.01e-03 6.90e-05     8.81e-04     1.15e-03
##   423    213       1 1.00e-03 6.89e-05     8.77e-04     1.15e-03
##   424    212       3 9.89e-04 6.84e-05     8.64e-04     1.13e-03
##   426    209       2 9.79e-04 6.80e-05     8.55e-04     1.12e-03
##   427    207       1 9.75e-04 6.79e-05     8.50e-04     1.12e-03
##   428    206       3 9.60e-04 6.74e-05     8.37e-04     1.10e-03
##   429    203       8 9.23e-04 6.60e-05     8.02e-04     1.06e-03
##   431    195       1 9.18e-04 6.59e-05     7.97e-04     1.06e-03
##   432    194       3 9.04e-04 6.54e-05     7.84e-04     1.04e-03
##   433    191       2 8.94e-04 6.50e-05     7.75e-04     1.03e-03
##   434    189       2 8.85e-04 6.47e-05     7.67e-04     1.02e-03
##   435    187       3 8.71e-04 6.41e-05     7.53e-04     1.01e-03
##   436    184       4 8.52e-04 6.34e-05     7.36e-04     9.86e-04
##   437    180       1 8.47e-04 6.33e-05     7.32e-04     9.80e-04
##   438    179       6 8.19e-04 6.22e-05     7.05e-04     9.50e-04
##   439    173       2 8.09e-04 6.18e-05     6.96e-04     9.40e-04
##   440    171       3 7.95e-04 6.13e-05     6.83e-04     9.25e-04
##   441    168       1 7.90e-04 6.11e-05     6.79e-04     9.19e-04
##   442    167       1 7.85e-04 6.09e-05     6.75e-04     9.14e-04
##   443    166       2 7.76e-04 6.06e-05     6.66e-04     9.04e-04
##   444    164       4 7.57e-04 5.98e-05     6.48e-04     8.84e-04
##   445    160       1 7.52e-04 5.96e-05     6.44e-04     8.79e-04
##   446    159       3 7.38e-04 5.91e-05     6.31e-04     8.63e-04
##   447    156       1 7.33e-04 5.89e-05     6.27e-04     8.58e-04
##   448    155       2 7.24e-04 5.85e-05     6.18e-04     8.48e-04
##   449    153       5 7.00e-04 5.75e-05     5.96e-04     8.23e-04
##   450    148       2 6.91e-04 5.71e-05     5.87e-04     8.12e-04
##   451    146       1 6.86e-04 5.70e-05     5.83e-04     8.07e-04
##   452    145       5 6.62e-04 5.60e-05     5.61e-04     7.82e-04
##   453    140       2 6.53e-04 5.56e-05     5.53e-04     7.71e-04
##   454    138       2 6.43e-04 5.52e-05     5.44e-04     7.61e-04
```

```
##   455    136       7 6.10e-04 5.37e-05     5.14e-04     7.25e-04
##   456    129       5 5.87e-04 5.27e-05     4.92e-04     7.00e-04
##   457    124       1 5.82e-04 5.25e-05     4.88e-04     6.94e-04
##   458    123       4 5.63e-04 5.16e-05     4.70e-04     6.74e-04
##   459    119       3 5.49e-04 5.09e-05     4.58e-04     6.58e-04
##   460    116       1 5.44e-04 5.07e-05     4.53e-04     6.53e-04
##   461    115       1 5.39e-04 5.05e-05     4.49e-04     6.48e-04
##   462    114       2 5.30e-04 5.01e-05     4.40e-04     6.38e-04
##   463    112       2 5.20e-04 4.96e-05     4.32e-04     6.27e-04
##   465    110       2 5.11e-04 4.92e-05     4.23e-04     6.17e-04
##   466    108       2 5.02e-04 4.87e-05     4.15e-04     6.07e-04
##   467    106       2 4.92e-04 4.82e-05     4.06e-04     5.96e-04
##   468    104       1 4.87e-04 4.80e-05     4.02e-04     5.91e-04
##   469    103       3 4.73e-04 4.73e-05     3.89e-04     5.76e-04
##   470    100       3 4.59e-04 4.66e-05     3.76e-04     5.60e-04
##   473     97       3 4.45e-04 4.59e-05     3.63e-04     5.44e-04
##   474     94       3 4.31e-04 4.51e-05     3.51e-04     5.29e-04
##   475     91       1 4.26e-04 4.49e-05     3.46e-04     5.24e-04
##   476     90       1 4.21e-04 4.46e-05     3.42e-04     5.18e-04
##   477     89       1 4.16e-04 4.44e-05     3.38e-04     5.13e-04
##   478     88       2 4.07e-04 4.39e-05     3.29e-04     5.03e-04
##   479     86       1 4.02e-04 4.36e-05     3.25e-04     4.97e-04
##   480     85       1 3.97e-04 4.34e-05     3.21e-04     4.92e-04
##   488     84       1 3.93e-04 4.31e-05     3.17e-04     4.87e-04
##   489     83       3 3.78e-04 4.23e-05     3.04e-04     4.71e-04
##   491     80       1 3.74e-04 4.20e-05     3.00e-04     4.66e-04
##   492     79       1 3.69e-04 4.18e-05     2.96e-04     4.61e-04
##   493     78       3 3.55e-04 4.10e-05     2.83e-04     4.45e-04
##   494     75       2 3.45e-04 4.04e-05     2.75e-04     4.34e-04
##   495     73       2 3.36e-04 3.99e-05     2.66e-04     4.24e-04
##   496     71       1 3.31e-04 3.96e-05     2.62e-04     4.19e-04
##   497     70       1 3.26e-04 3.93e-05     2.58e-04     4.13e-04
##   499     69       2 3.17e-04 3.87e-05     2.50e-04     4.03e-04
##   500     67       1 3.12e-04 3.84e-05     2.45e-04     3.97e-04
##   502     66       1 3.08e-04 3.81e-05     2.41e-04     3.92e-04
##   503     65       2 2.98e-04 3.75e-05     2.33e-04     3.82e-04
##   504     63       3 2.84e-04 3.66e-05     2.20e-04     3.66e-04
##   505     60       1 2.79e-04 3.63e-05     2.16e-04     3.60e-04
##   506     59       2 2.70e-04 3.57e-05     2.08e-04     3.50e-04
##   507     57       2 2.60e-04 3.51e-05     2.00e-04     3.39e-04
##   511     55       1 2.55e-04 3.48e-05     1.96e-04     3.34e-04
##   512     54       2 2.46e-04 3.41e-05     1.87e-04     3.23e-04
##   513     52       1 2.41e-04 3.38e-05     1.83e-04     3.17e-04
##   514     51       1 2.37e-04 3.35e-05     1.79e-04     3.12e-04
##   515     50       1 2.32e-04 3.31e-05     1.75e-04     3.07e-04
##   516     49       1 2.27e-04 3.28e-05     1.71e-04     3.01e-04
##   517     48       1 2.22e-04 3.24e-05     1.67e-04     2.96e-04
##   519     47       1 2.18e-04 3.21e-05     1.63e-04     2.91e-04
##   520     46       2 2.08e-04 3.14e-05     1.55e-04     2.80e-04
##   524     44       3 1.94e-04 3.03e-05     1.43e-04     2.63e-04
##   525     41       1 1.89e-04 2.99e-05     1.39e-04     2.58e-04
##   528     40       2 1.80e-04 2.92e-05     1.31e-04     2.47e-04
##   531     38       1 1.75e-04 2.88e-05     1.27e-04     2.42e-04
##   533     37       1 1.70e-04 2.84e-05     1.23e-04     2.36e-04
```

```
##   534     36       1 1.66e-04 2.80e-05     1.19e-04     2.31e-04
##   537     35       1 1.61e-04 2.76e-05     1.15e-04     2.25e-04
##   542     34       3 1.47e-04 2.63e-05     1.03e-04     2.09e-04
##   543     31       2 1.37e-04 2.55e-05     9.53e-05     1.97e-04
##   547     29       2 1.28e-04 2.46e-05     8.76e-05     1.86e-04
##   548     27       1 1.23e-04 2.41e-05     8.38e-05     1.81e-04
##   549     26       1 1.18e-04 2.37e-05     7.99e-05     1.75e-04
##   550     25       1 1.14e-04 2.32e-05     7.61e-05     1.69e-04
##   552     24       1 1.09e-04 2.27e-05     7.23e-05     1.64e-04
##   558     23       1 1.04e-04 2.22e-05     6.85e-05     1.58e-04
##   559     22       1 9.94e-05 2.17e-05     6.48e-05     1.52e-04
##   561     21       1 9.46e-05 2.12e-05     6.10e-05     1.47e-04
##   563     20       1 8.99e-05 2.06e-05     5.73e-05     1.41e-04
##   564     19       1 8.52e-05 2.01e-05     5.37e-05     1.35e-04
##   581     18       3 7.10e-05 1.83e-05     4.28e-05     1.18e-04
##   584     15       1 6.62e-05 1.77e-05     3.92e-05     1.12e-04
##   586     14       1 6.15e-05 1.71e-05     3.57e-05     1.06e-04
##   587     13       1 5.68e-05 1.64e-05     3.22e-05     1.00e-04
##   593     12       1 5.20e-05 1.57e-05     2.88e-05     9.40e-05
##   594     11       1 4.73e-05 1.50e-05     2.55e-05     8.79e-05
##   604     10       1 4.26e-05 1.42e-05     2.22e-05     8.18e-05
##   605      9       1 3.78e-05 1.34e-05     1.89e-05     7.57e-05
##   606      8       1 3.31e-05 1.25e-05     1.58e-05     6.95e-05
##   618      7       1 2.84e-05 1.16e-05     1.28e-05     6.32e-05
##   627      6       1 2.37e-05 1.06e-05     9.85e-06     5.68e-05
##   634      5       1 1.89e-05 9.46e-06     7.10e-06     5.04e-05
##   647      4       1 1.42e-05 8.19e-06     4.58e-06     4.40e-05
##   659      3       1 9.46e-06 6.69e-06     2.37e-06     3.78e-05
##   661      2       1 4.73e-06 4.73e-06     6.66e-07     3.36e-05
##   703      1       1 0.00e+00      NaN           NA           NA
```

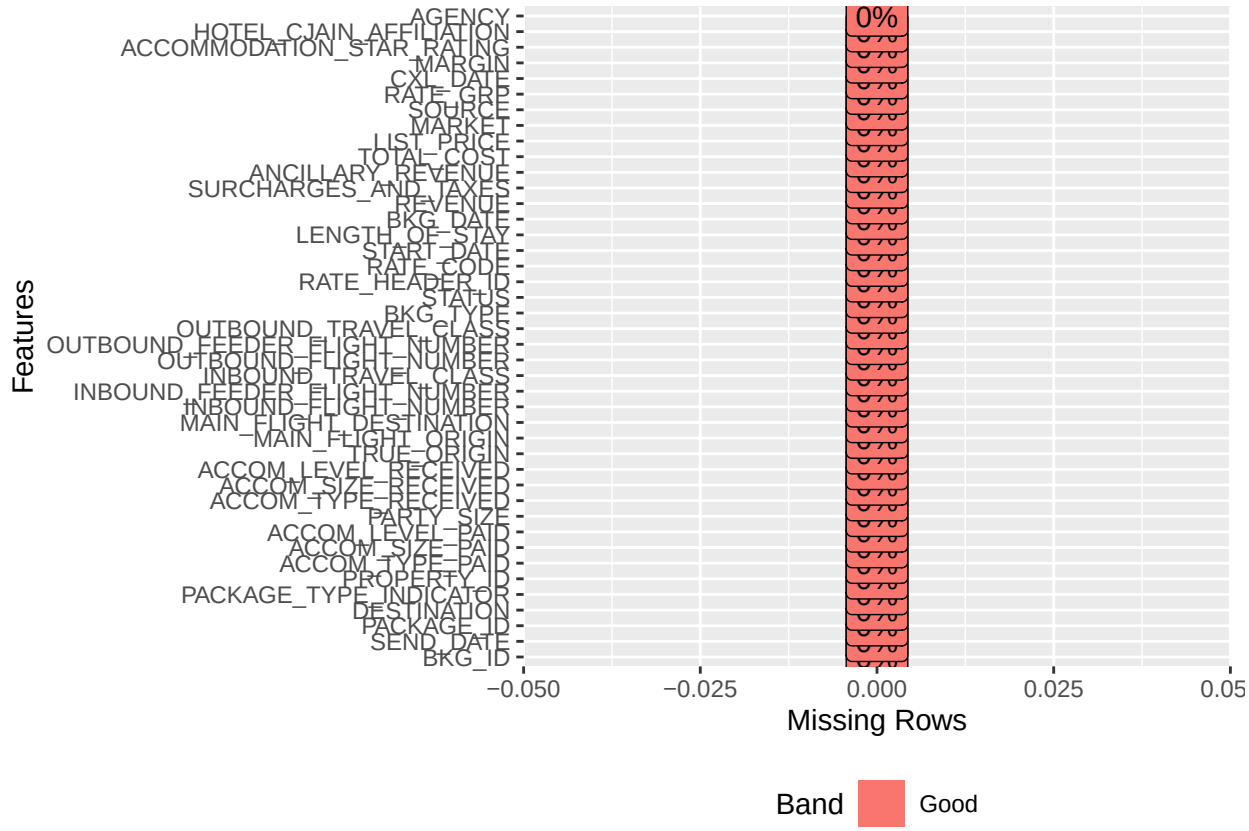
```
plot(km_2019,conf.int=FALSE, mark.time=TRUE,main="Kaplan-Meier estimate of the Pre Booking 2019", xlab=
```



Box Cox Model

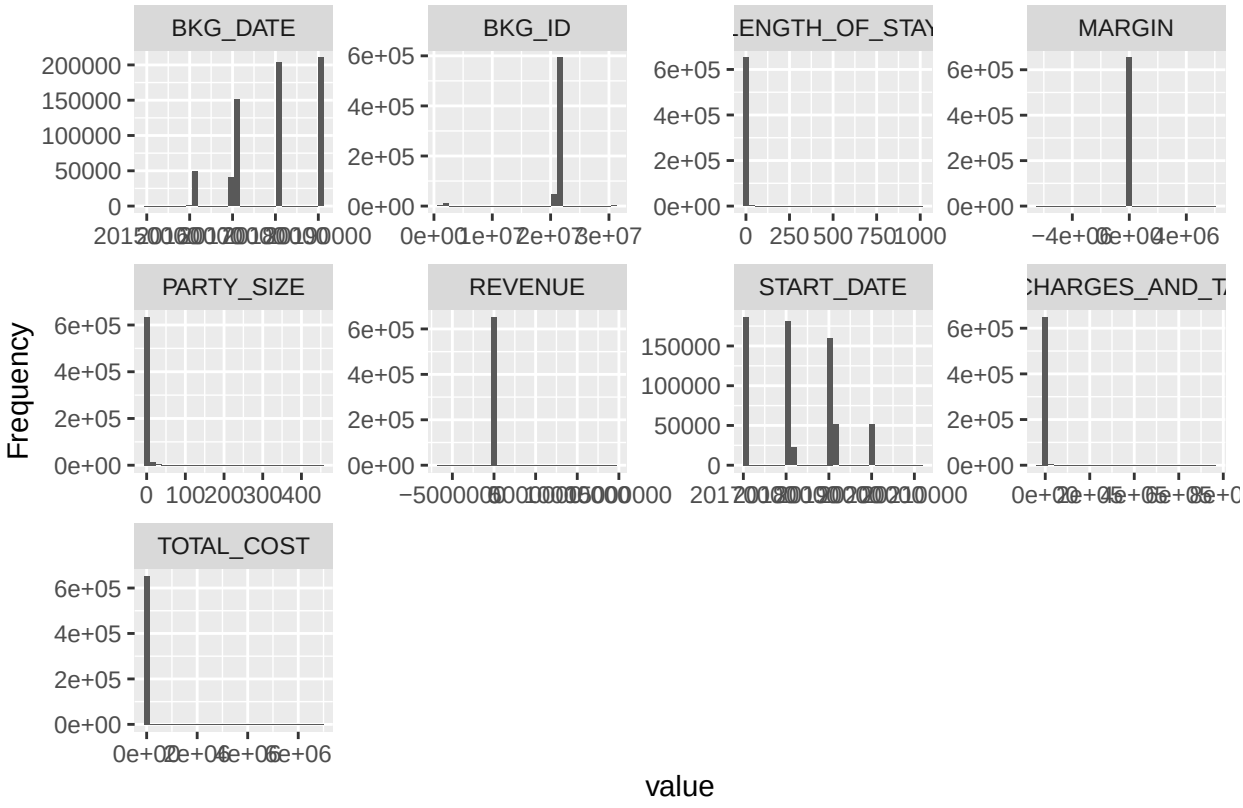
A Box Cox transformation is a way to transform non-normal dependent variables into a normal shape. Normality is an important assumption for many statistical techniques; if your data isn't normal, applying a Box-Cox means that you are able to run a broader number of tests.

```
plot_missing(raw)
```



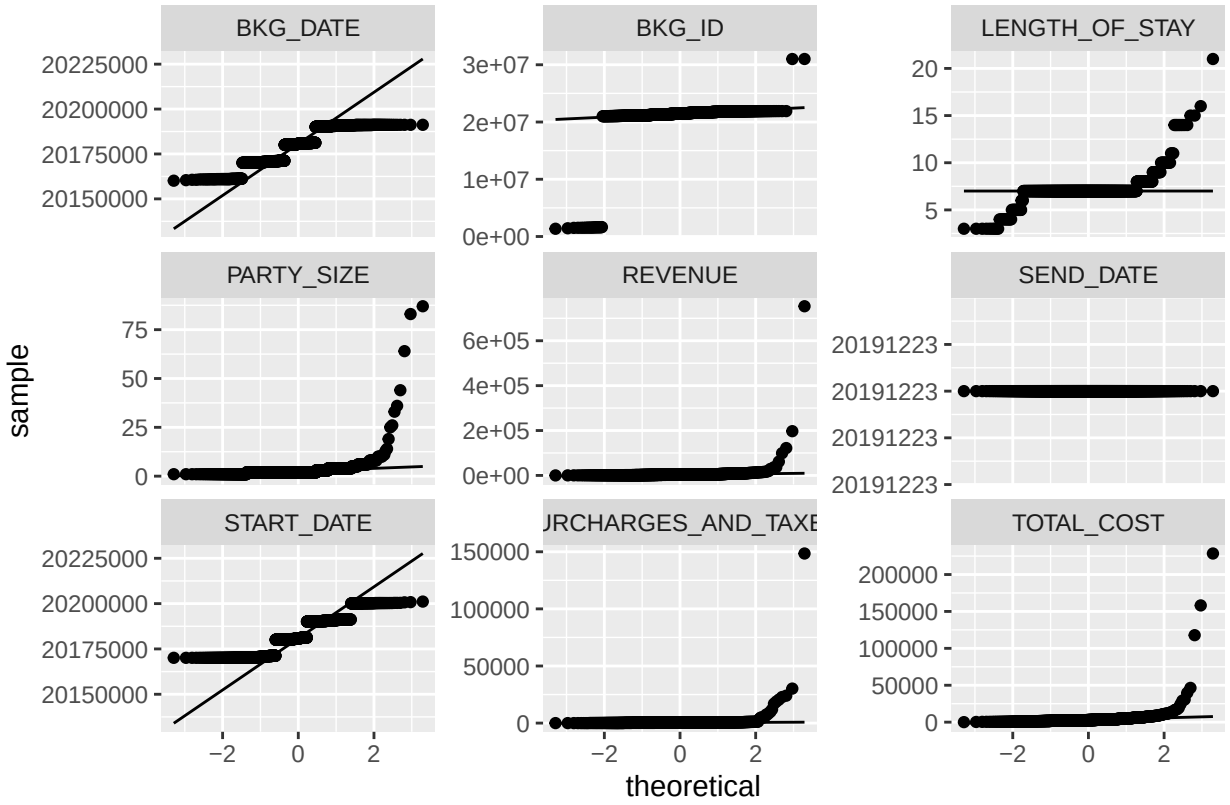
This graph shows summary statistics for all the features we have in our dataset with regards to frequency of occurances.

```
plot_histogram(raw)
```

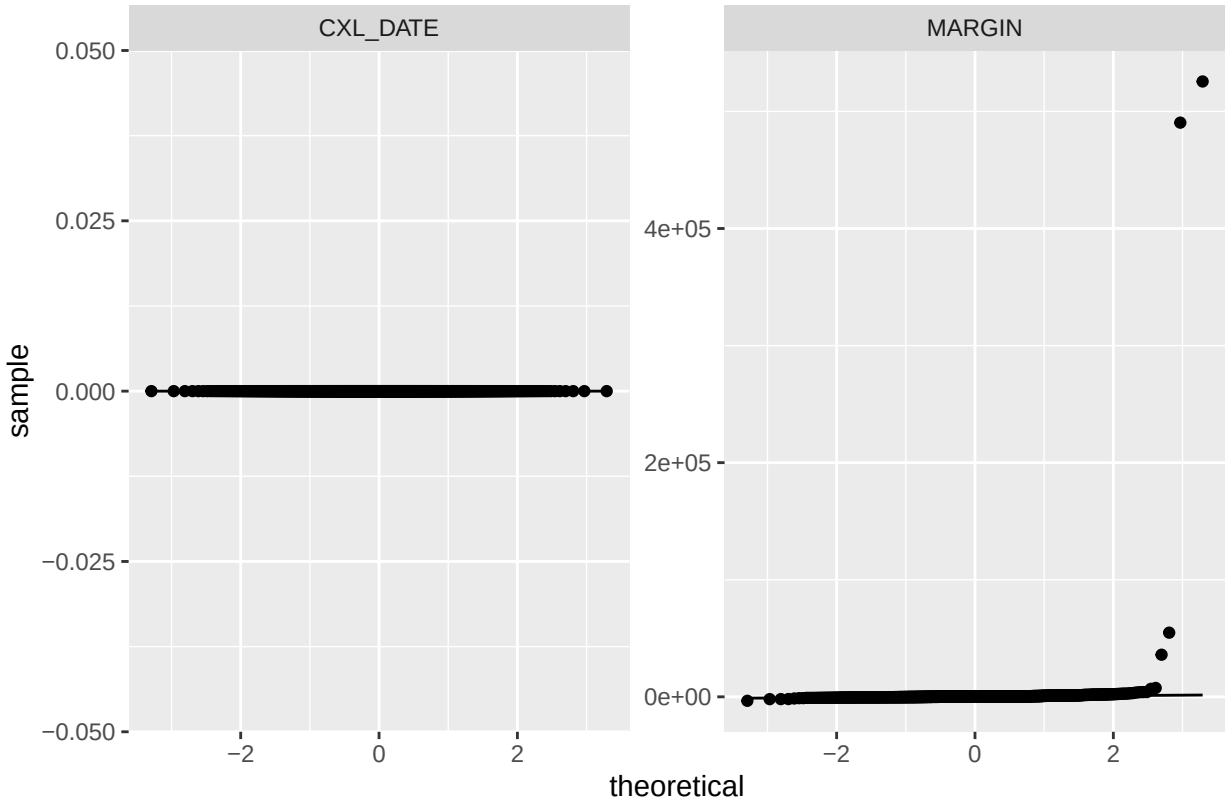


When Analyzing the dataset we see that few outliers show up

```
plot_qq(raw, sampled_rows = 1000L)
```


BKG_DATE
BKG_ID
LENGTH_OF_STAY
PARTY_SIZE
REVENUE
SEND_DATE
START_DATE
URCHARGES_AND_TAXE
TOTAL_COST
sample
theoretical
20225000
20200000
20175000
20150000
3e+07
2e+07
1e+07
0e+00
20
15
10
5
75
50
25
0
6e+05
4e+05
2e+05
0e+00
20191223
150000
100000
50000
200000
−2
0
2

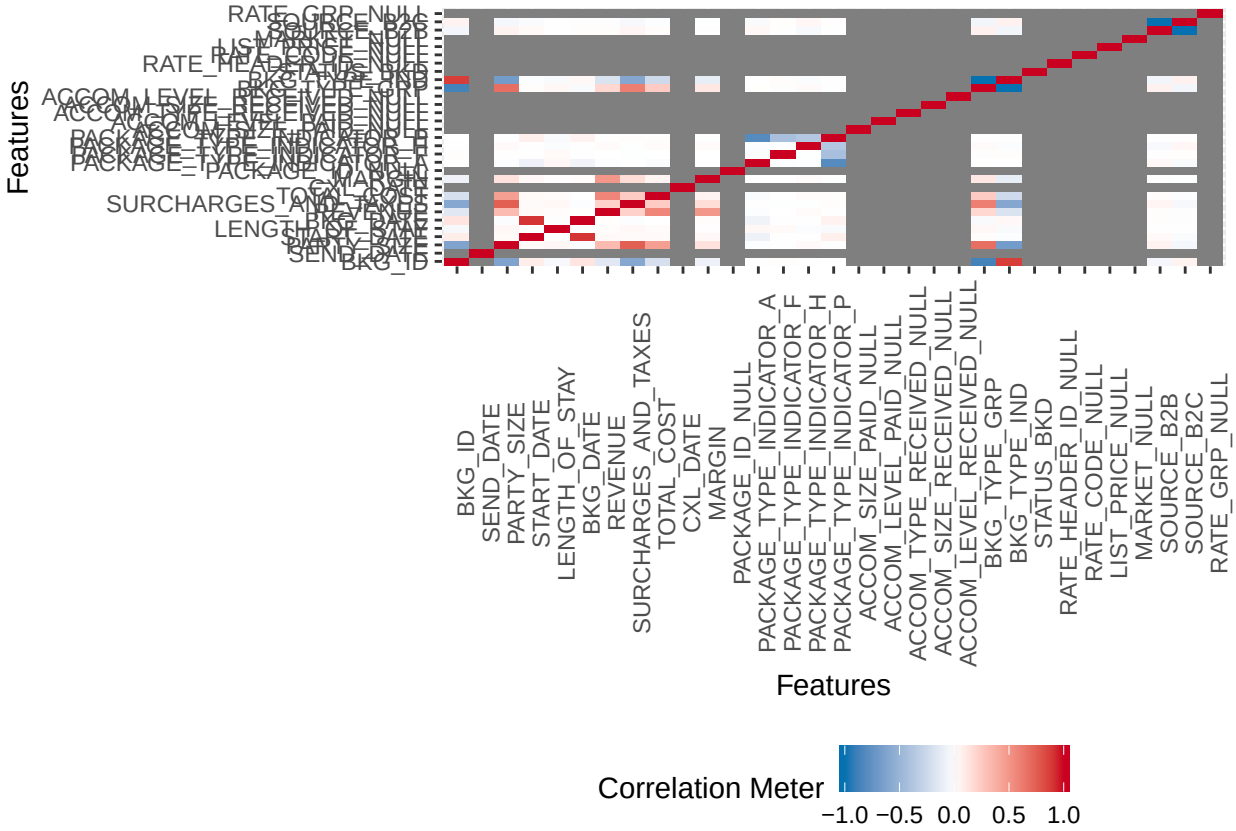
Page 1



Correlation graph to see any correlation between the data

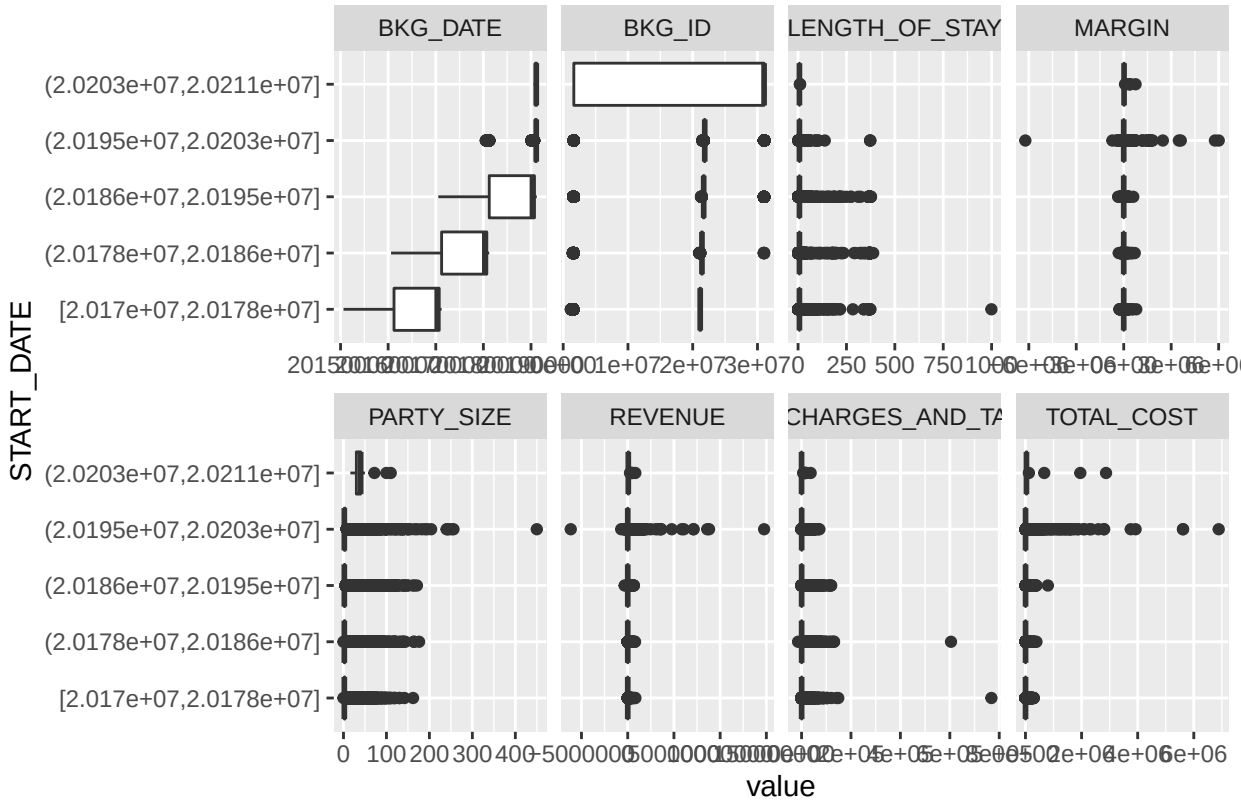
```
plot_correlation(na.omit(raw), maxcat = 10L)
```

```
## Warning in cor(x = structure(list(BKG_ID = c(1171217L, 1220260L, 1236226L, : the
## standard deviation is zero
```



The boxplot gives ability to deciphers the outliers from the dataset

```
plot_boxplot(raw, by = "START_DATE")
```




```
#############################################################################
# Get the list of all the DESTINATION from the Raw data set
Name_of_Dest <- levels(dataset$DESTINATION)
Name_of_Dest
```

```
##  [1] "ANTIGUA"              "BRIDGETOWN"           "CANCUN"
##  [4] "CARTAGENA"            "CAYO COCO"            "CAYO SANTA CLARA"
##  [7] "COZUMEL"              "CURACAO"              "EDMONTON"
## [10] "GENERIC"              "GEORGE TOWN"          "GRENADA"
## [13] "GUADELOUPE"           "HAVANA"               "HOLGUIN"
## [16] "HONOLULU"             "HUATULCO"             "IXTAPA/ZIHUATANEJO"
## [19] "LA ROMANA"            "LIBERIA"              "LONDON"
## [22] "MONTEGO BAY"          "MONTREAL"             "OTTAWA"
## [25] "PENTICTON, BC"        "PUERTO PLATA"         "PUERTO VALLARTA"
## [28] "PUNTA CANA"           "QUEBEC"               "SAMANA"
## [31] "SAN JOSE DEL CABO"    "SAULT ST MARIE"       "ST. JOHNS, NL"
## [34] "ST. KITTS"            "ST. LUCIA-VIEUX FORT" "TORONTO"
## [37] "VARADERO"             "WINNIPEG"
```

```
# Get the count of the DESTINATION from the Raw data set
Num_of_Dest <- length(Name_of_Dest)
Num_of_Dest
```

```
## [1] 38
```

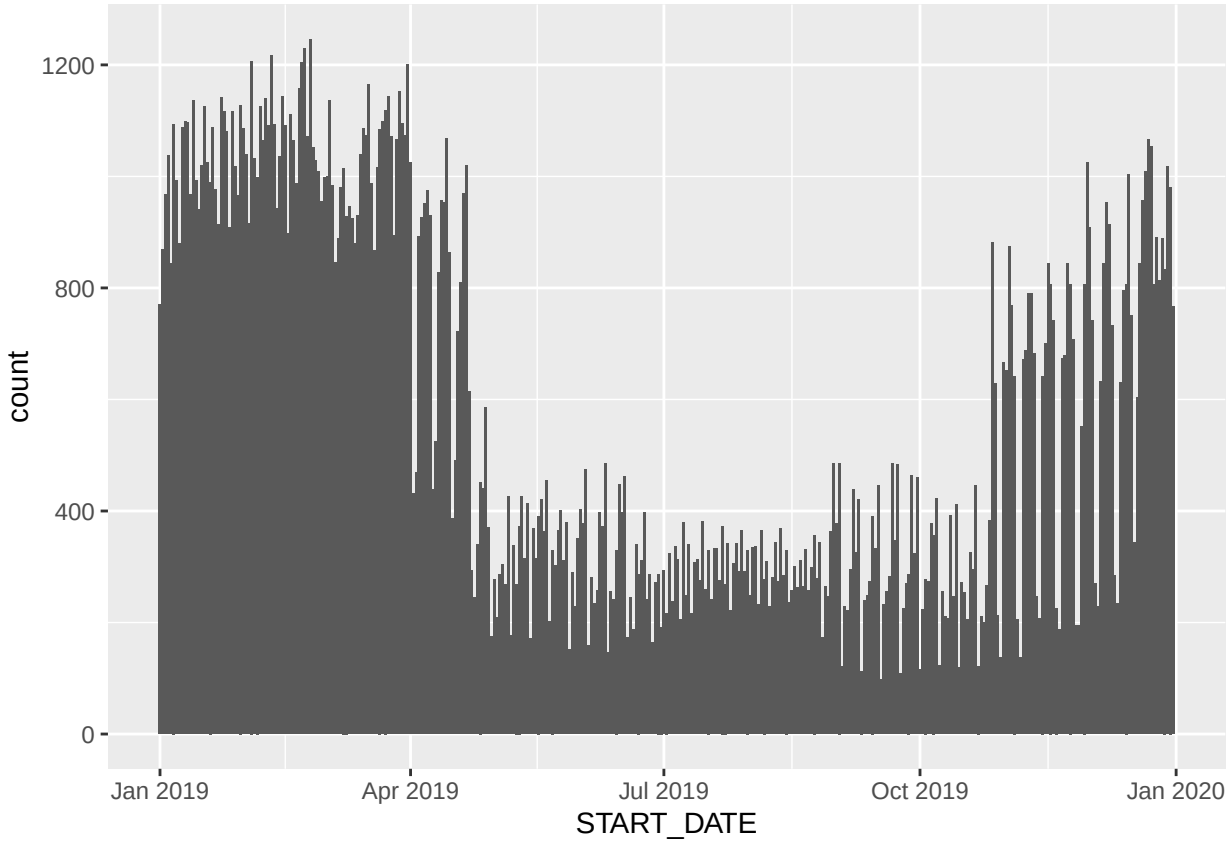
```r
# Values for the TTE Dataframe
df_TTE <- NULL      # DataFrame
TTE_v <- NULL       # TTE Value of interest
dest_dest_v <- NULL # Destination to match in the airports.csv for map data
dest_airport_v <- NULL# Main Destination airport to match with file as Destination names may not match

for(i in 1:Num_of_Dest) {
  # Get Destination Label
  dest_dest <- Name_of_Dest[i]
  dest_dest
  Destination_num <- dataset[dataset$DESTINATION==dest_dest,]
  dest_airport <- as.character(Destination_num$MAIN_FLIGHT_DESTINATION[1])
  dest_airport
  # Count the number of rows in the data frame
  cnt <- nrow(Destination_num)
  # Perform the Kaplan-Meier estimate of the Pre Booking
  km_curv <- survfit(Surv(Destination_num$TTE)~1)
  res <- summary(km_curv)
  # Create a data frame with only the Destination and TTE
  cols <- lapply(c(2:6), function(x) Destination_num[x])
  # Extract the columns you want
  cols <- lapply(c(2,6) , function(x) res[x])
  # Combine the columns into a data frame
  tbl <- do.call(data.frame, cols)
  # Extract values at % Required TODO: We could use this as a variable for more dynamic app?
  tbl_5pct <- tbl[tbl$surv < 0.05,]
  # Get the TTE as the first value in the subset
  TTE <- tbl_5pct$time[1]
  TTE_v <- c(TTE_v, TTE)
  dest_dest_v <- c(dest_dest_v, dest_dest)
  dest_airport_v <- c(dest_airport_v, dest_airport)
}
# dest_dest_v <- as.factor(dest_dest_v)
dest_airport_v <- as.character(dest_airport_v)
dest_dest_v <- as.character(dest_dest_v)
TTE_v <- as.character(TTE_v)
df_TTE <- data.frame("Destination" = dest_dest_v,"IATA" = dest_airport_v, "TTE" = TTE_v)
write.csv(df_TTE,'TTE_5pct.csv')
```

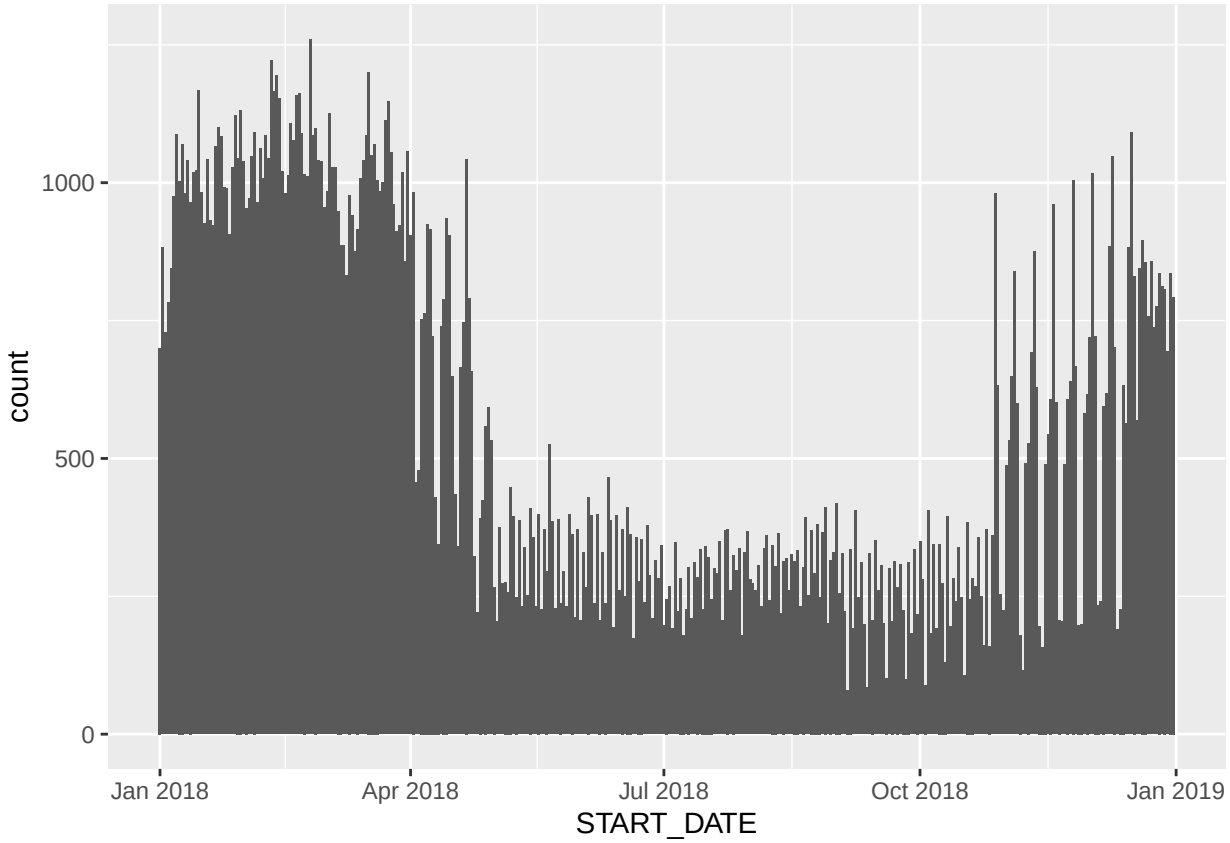
```r
# Merge airport data for the dataset
# Airport File Raw
raw_airport = read.csv("./input/cities_IATA_long_lat.csv", header=TRUE)
df_IATA <- raw_airport
# Airport Code and ETA
TTE_aiport = read.csv("./input/IATA_TTE_5pct.csv", header=TRUE)
df_IATA_TTE <- TTE_aiport
# Merge two datasets
mydata <- merge(df_IATA,df_IATA_TTE,by="IATA")
mydata$X <- NULL
mydata$Destination <- NULL
write.csv(mydata,'IATA_TTE_5pct.csv')
```

preprocess the data for modeling

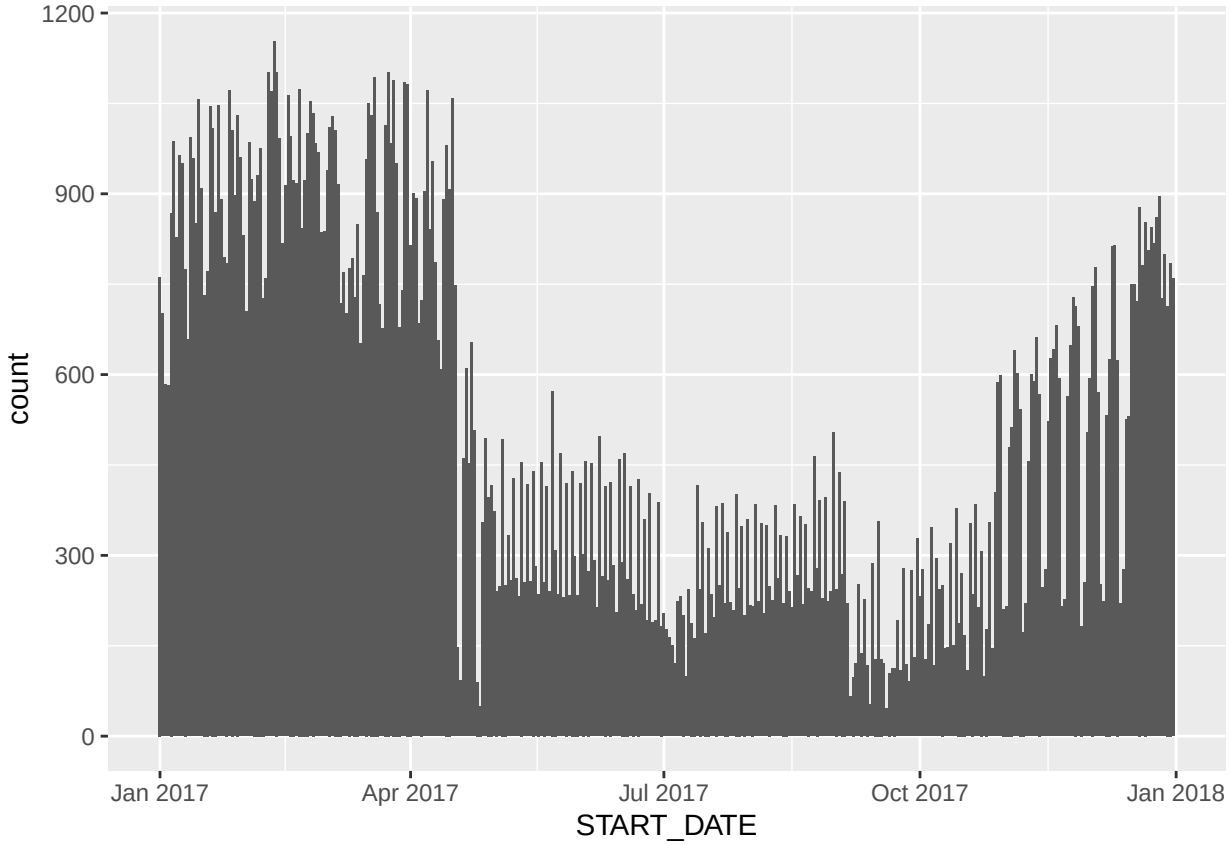
```
ggplot(data = All_2019) +
  geom_bar(mapping = aes(x=START_DATE))
```



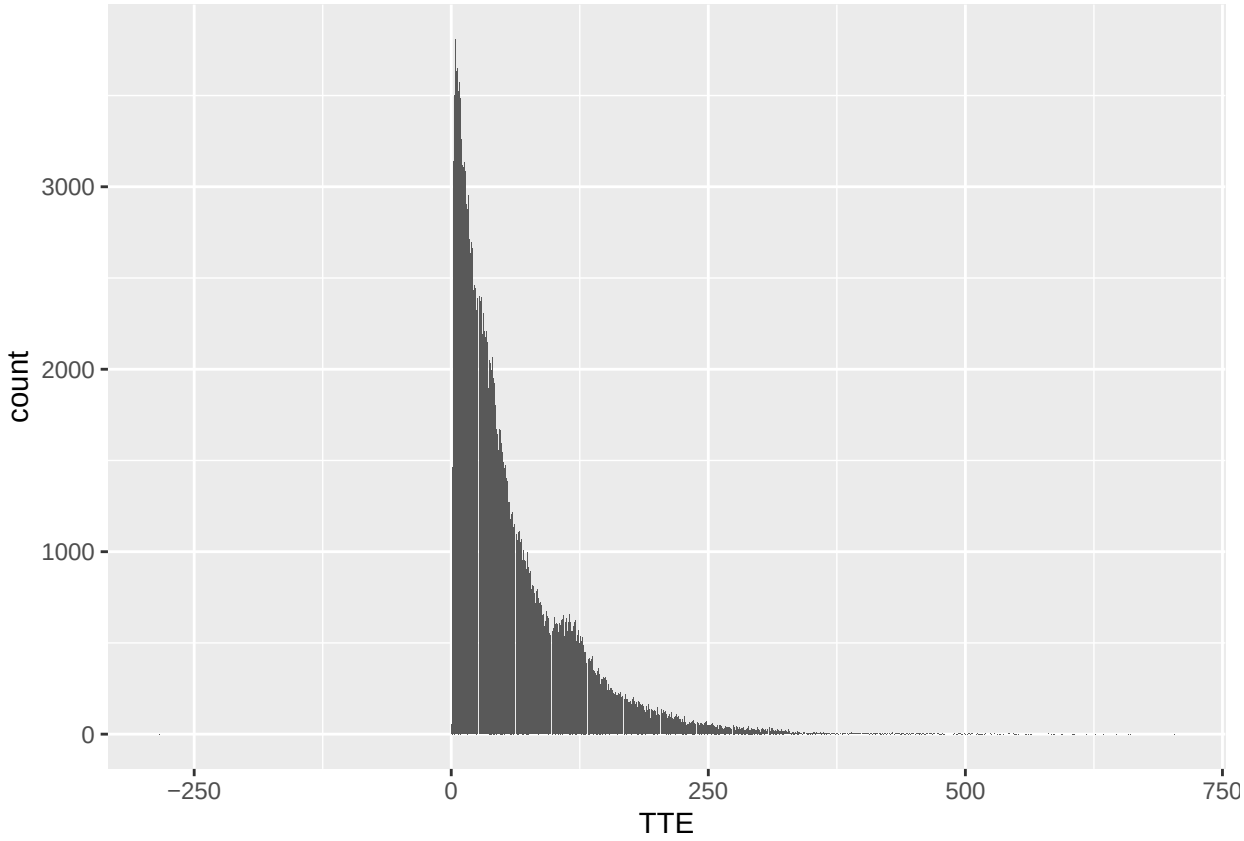

```
ggplot(data = All_2018) +
  geom_bar(mapping = aes(x=START_DATE))
```



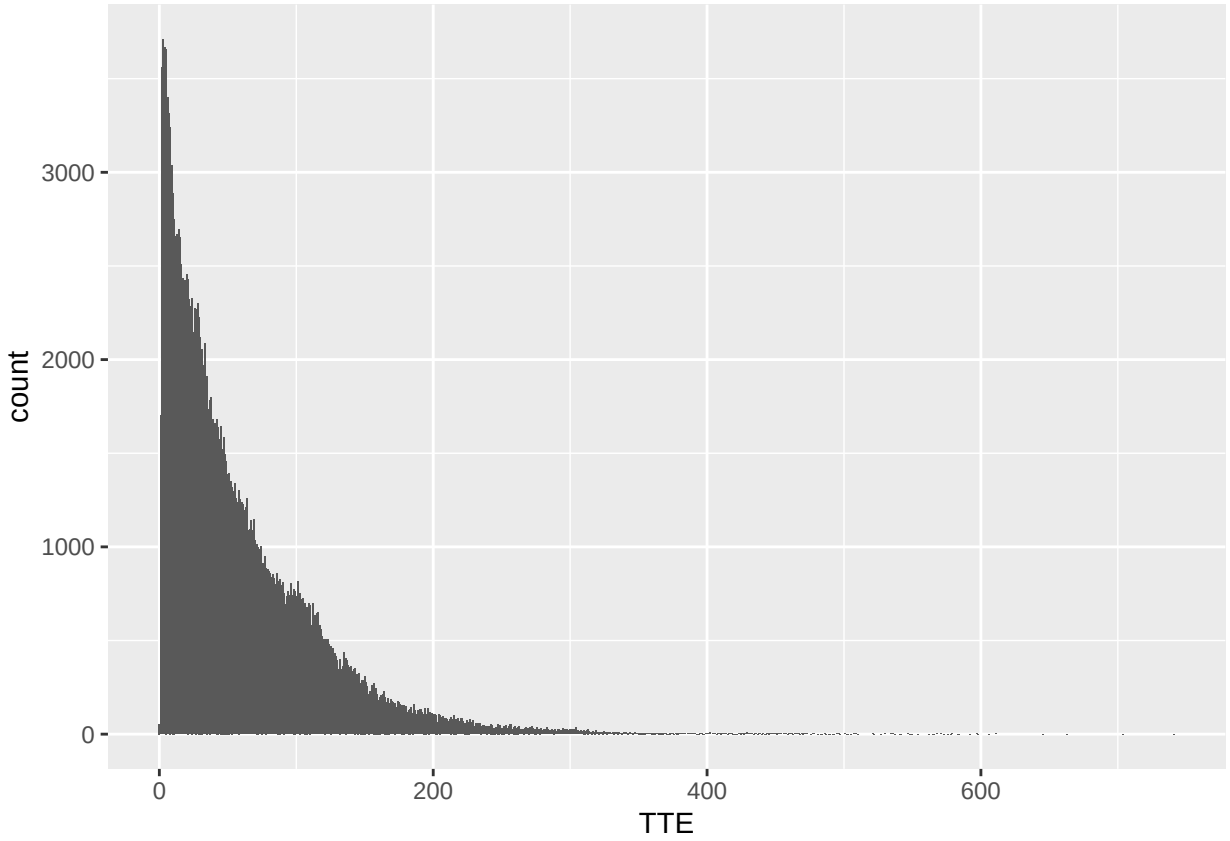

```
ggplot(data = All_2017) +
  geom_bar(mapping = aes(x=START_DATE))
```



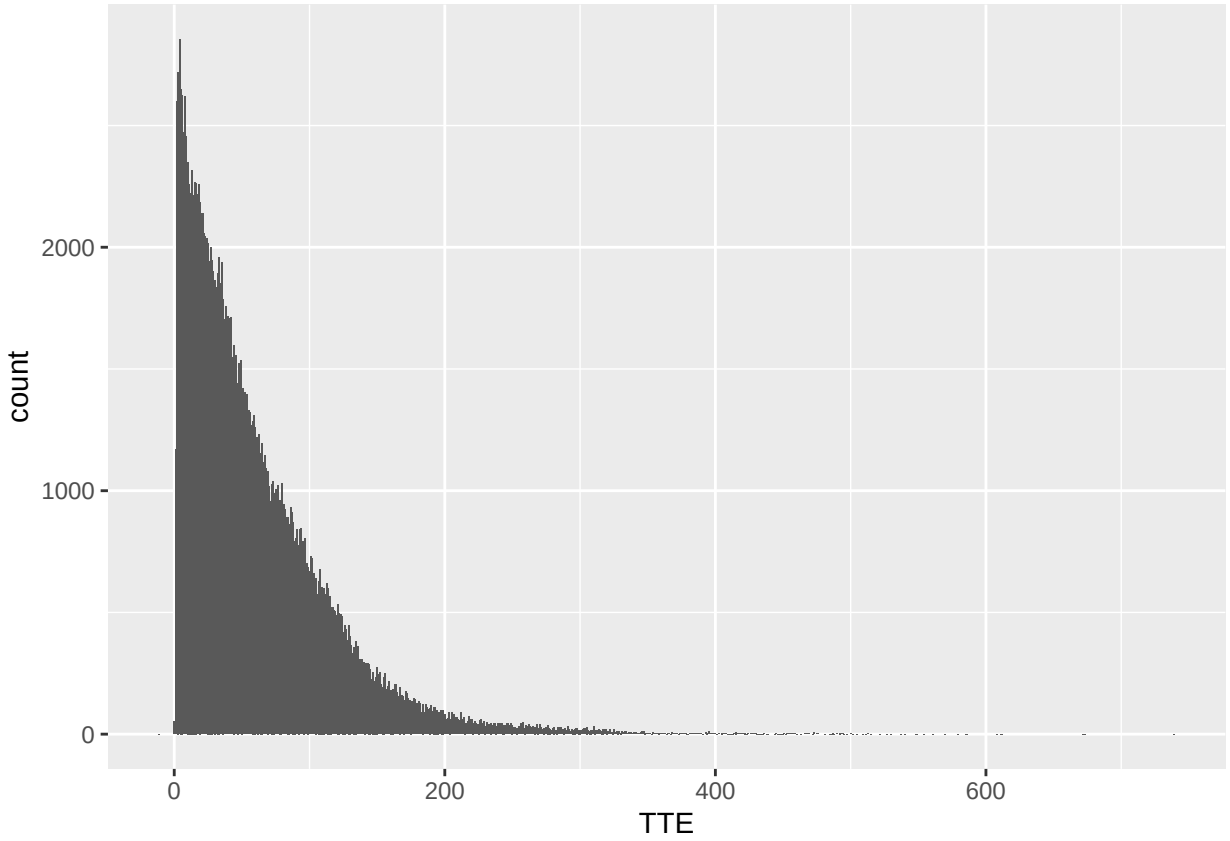

```
ggplot(data = All_2019) +
  geom_bar(mapping = aes(x=TTE))
```



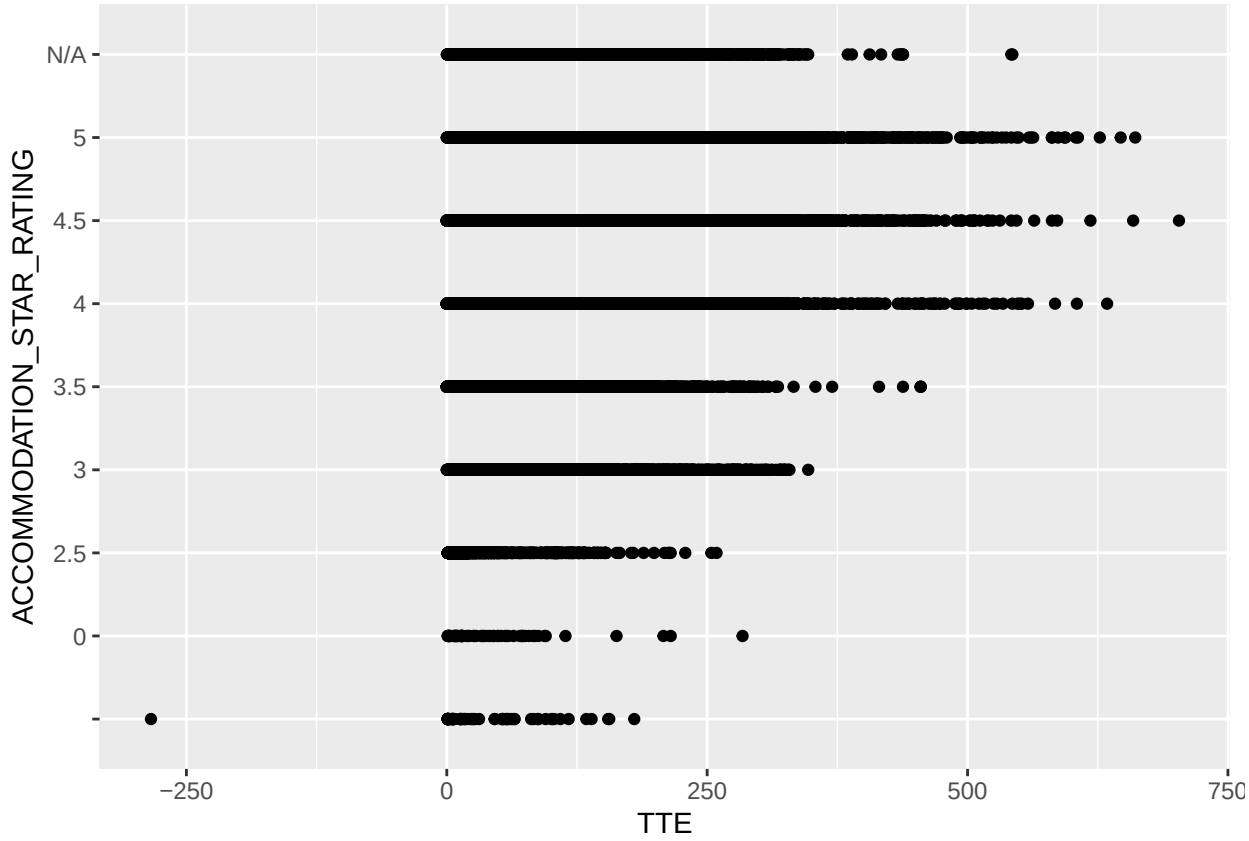

```
ggplot(data = All_2018) +
  geom_bar(mapping = aes(x=TTE))
```



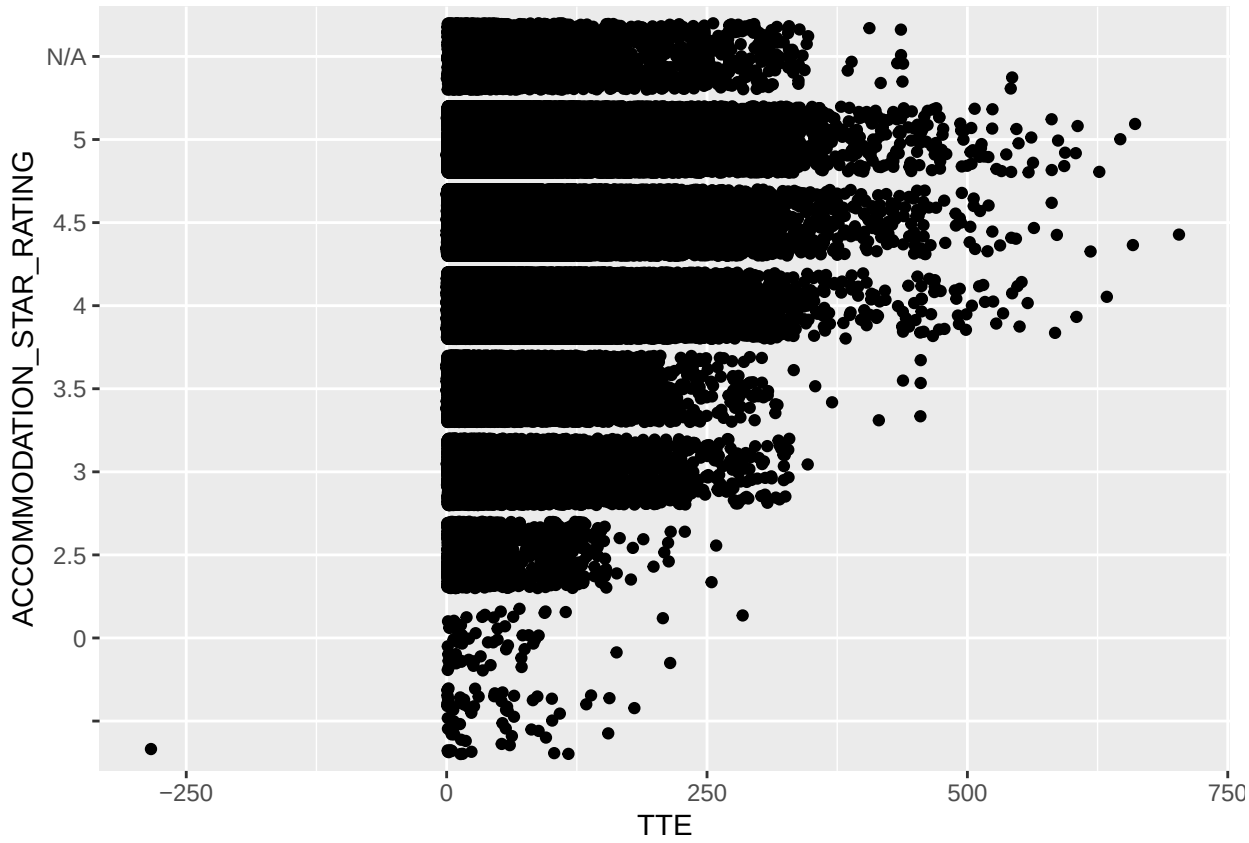

```
ggplot(data = All_2017) +
  geom_bar(mapping = aes(x=TTE))
```



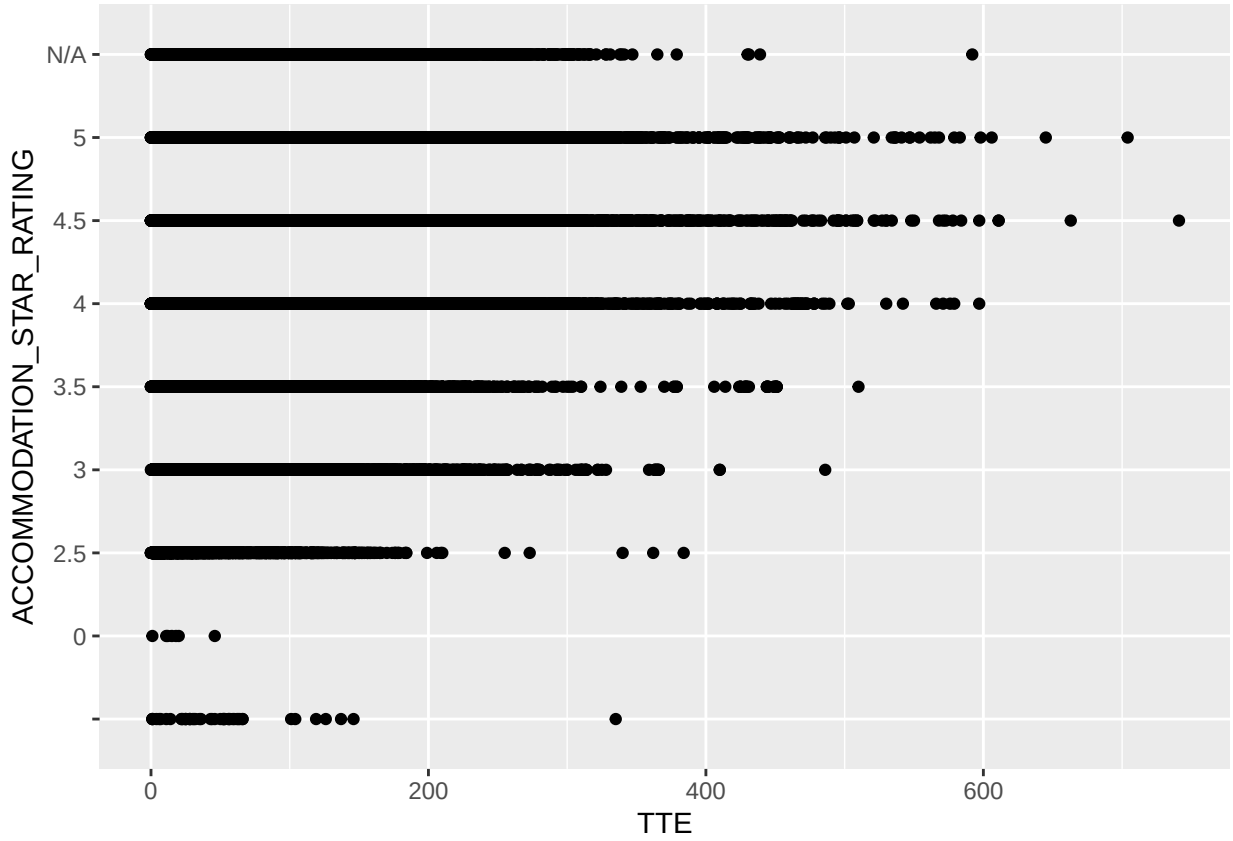

```
ggplot(data = All_2019) +
  geom_point(mapping = aes(x=TTE, y=ACCOMMODATION_STAR_RATING))
```



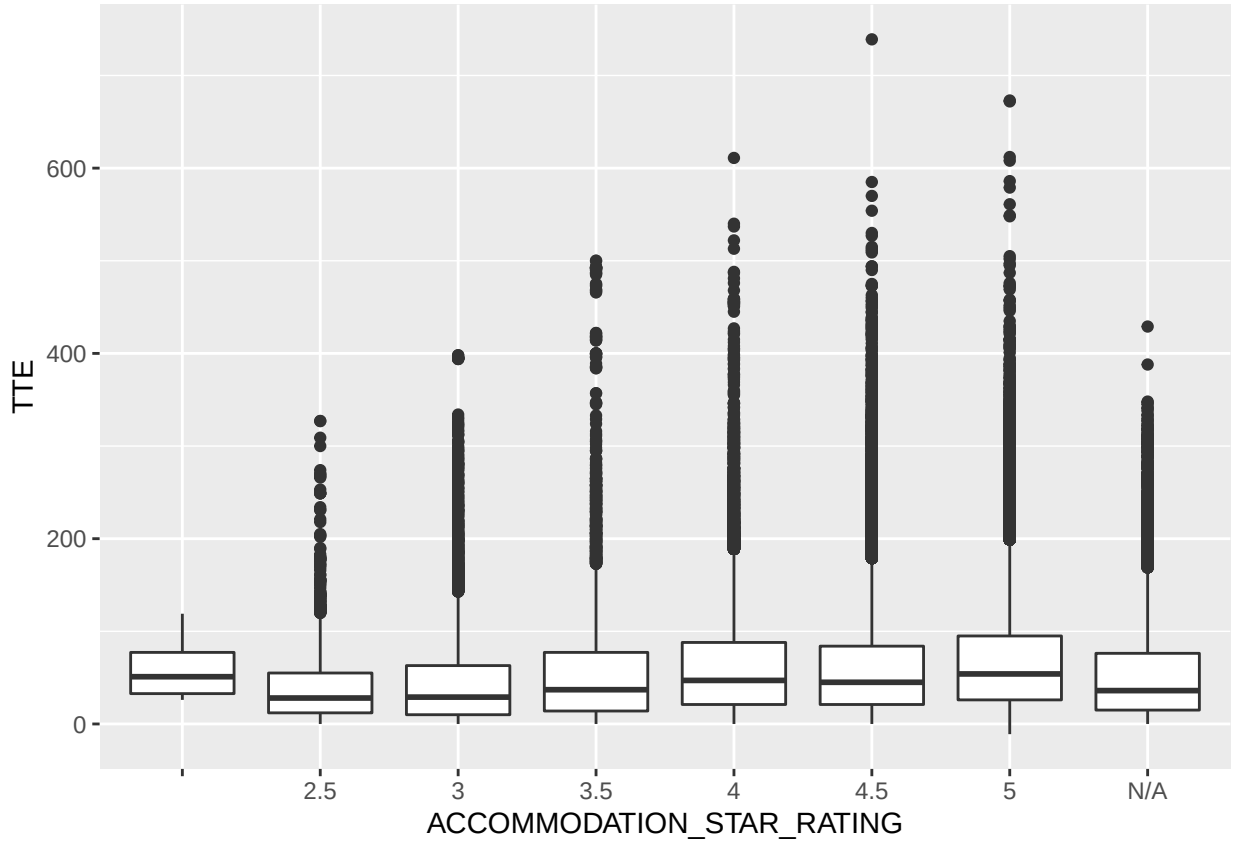

```
f <- ggplot(All_2019, aes(TTE, ACCOMMODATION_STAR_RATING))
f + geom_jitter()
```




```
ggplot(data = All_2018) +
  geom_point(mapping = aes(x=TTE, y=ACCOMMODATION_STAR_RATING))
```

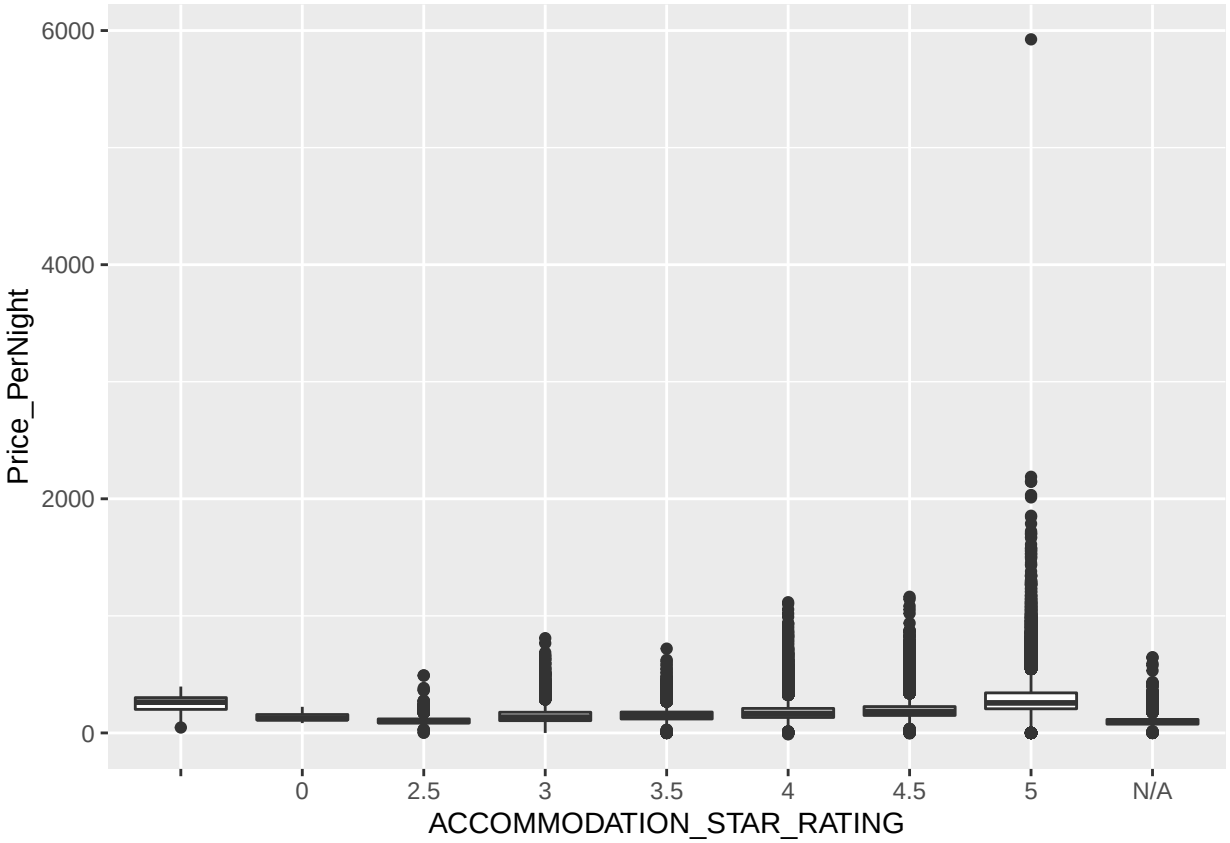



```
ggplot(data = All_2017, mapping = aes(x = ACCOMMODATION_STAR_RATING, y = TTE)) +
  geom_boxplot(mapping = aes(group = cut_width(ACCOMMODATION_STAR_RATING, 0.1)))
```

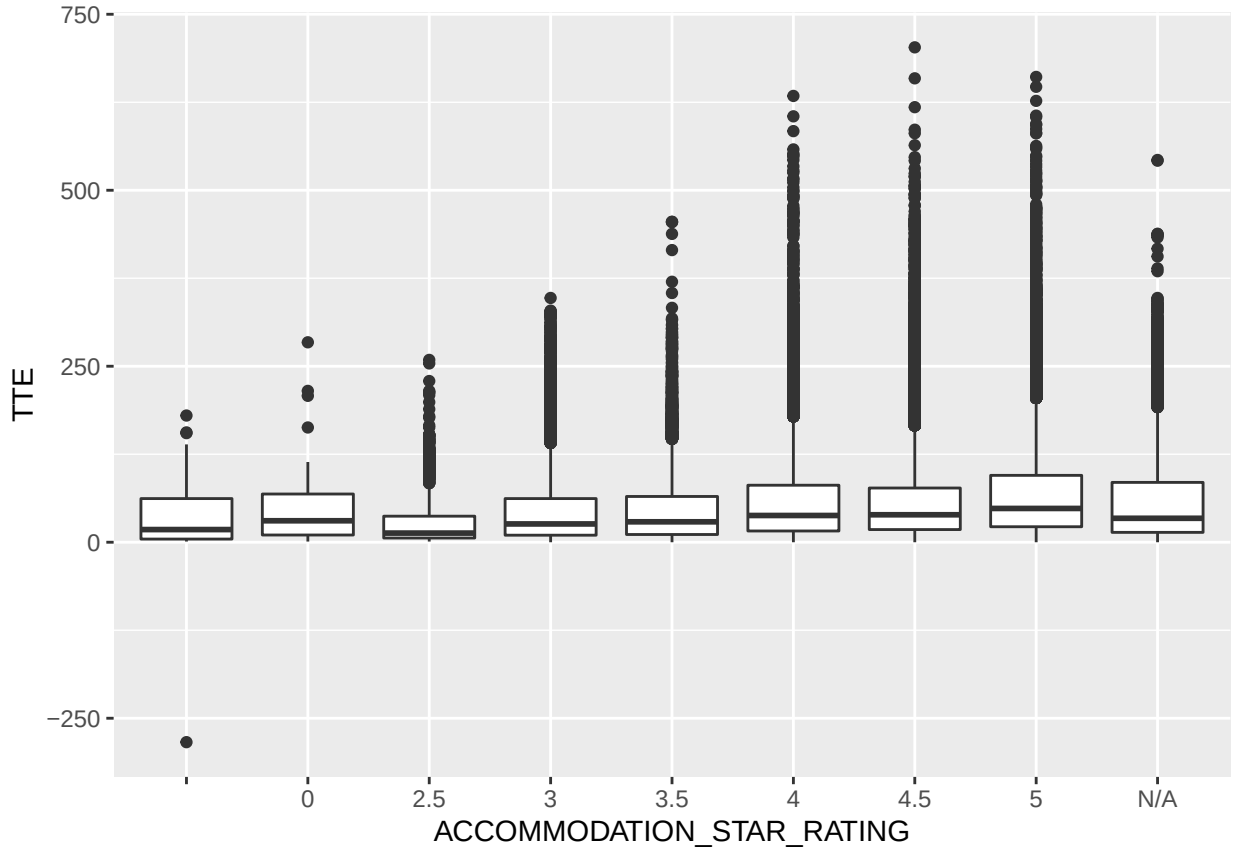



```
ggplot(data = All_2018, mapping = aes(x = ACCOMMODATION_STAR_RATING, y = Price_PerNight)) +
  geom_boxplot(mapping = aes(group = cut_width(ACCOMMODATION_STAR_RATING, 0.1)))
```

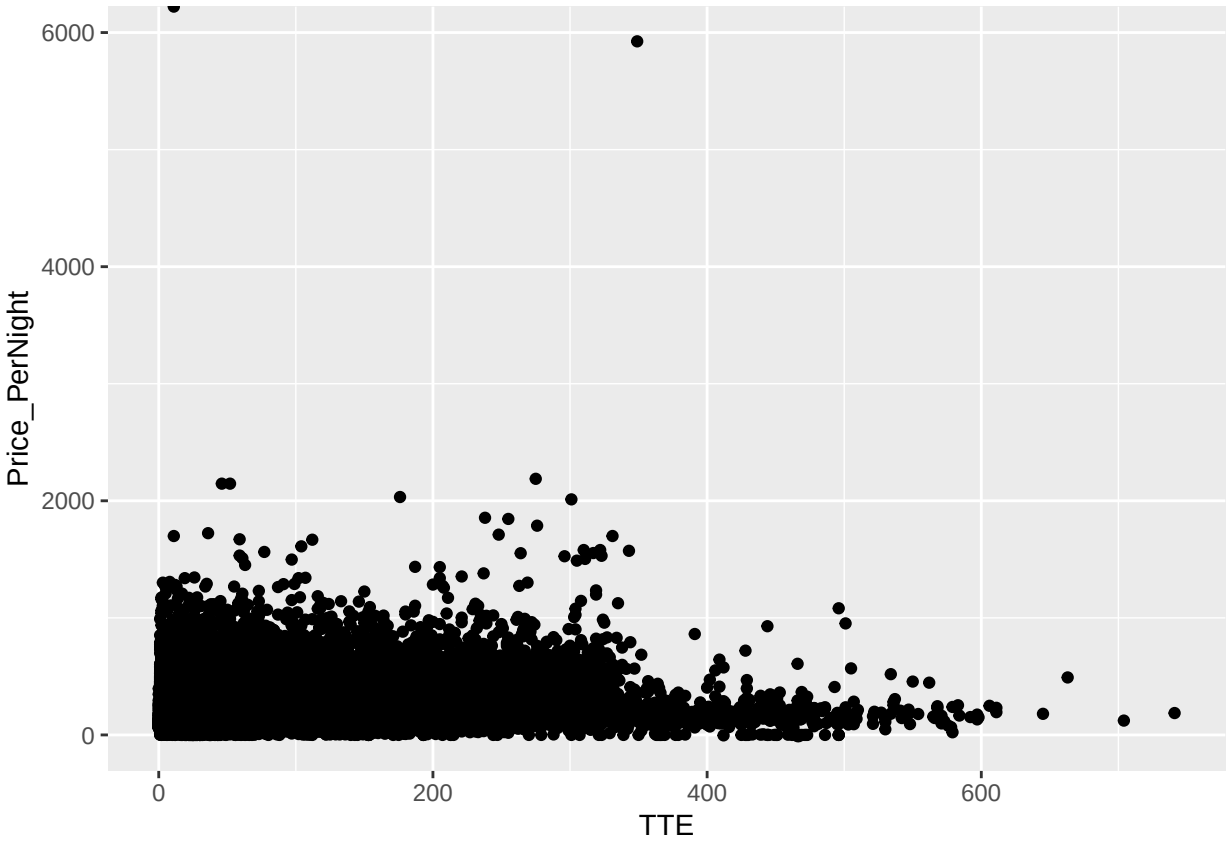
```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```




```
ggplot(data = All_2019, mapping = aes(x = ACCOMMODATION_STAR_RATING, y = TTE)) +
  geom_boxplot(mapping = aes(group = cut_width(ACCOMMODATION_STAR_RATING, 0.1)))
```




```
All_2018$ACCOMMODATION_STAR_RATING <- as.factor(All_2018$ACCOMMODATION_STAR_RATING)
ggplot(data = All_2018) +
  geom_point(mapping = aes(x= TTE, y = Price_PerNight))
```




```
str(All_2018)
```

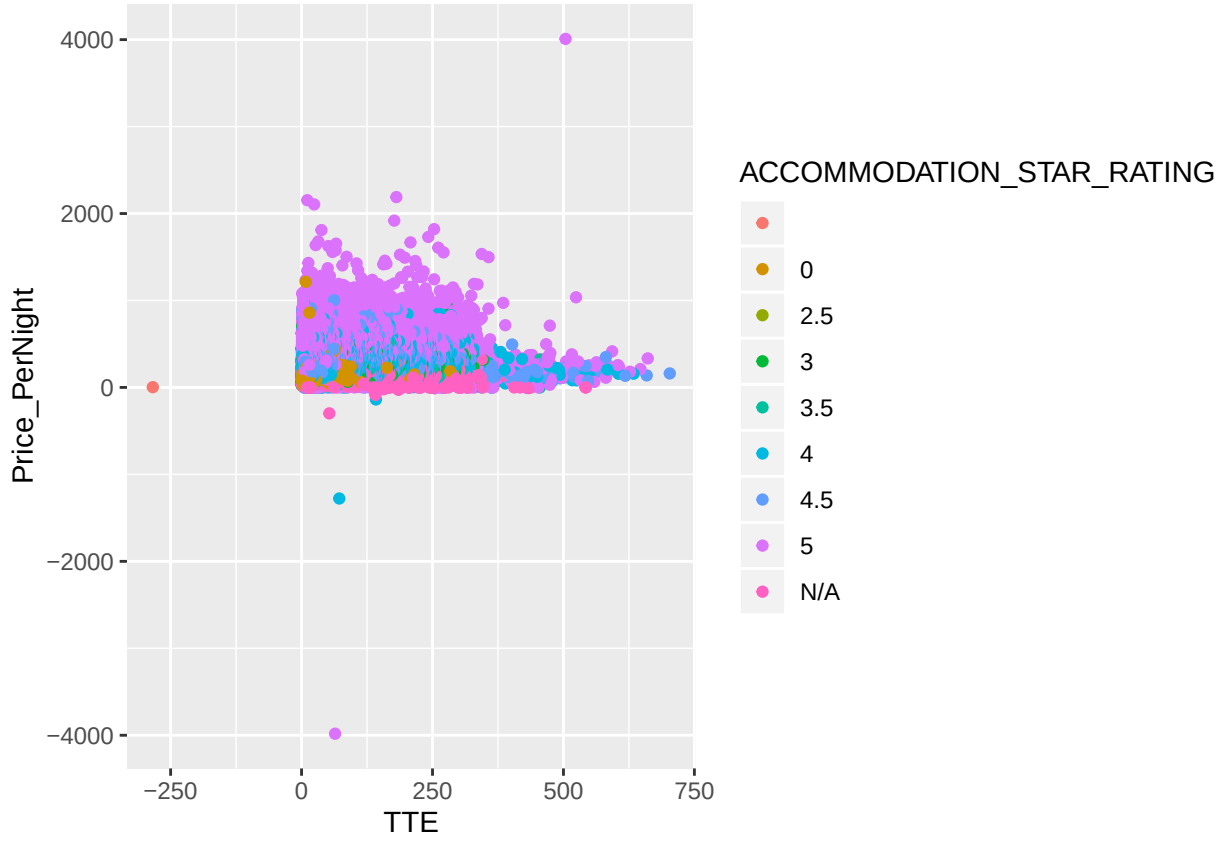
```
## 'data.frame':    203195 obs. of  16 variables:
##  $ DESTINATION              : Factor w/ 38 levels "ANTIGUA","BRIDGETOWN",..: 3 3 22 22 22 22 8 8 3 3
##  $ PROPERTY_ID              : Factor w/ 834 levels "ANU","ANUBLU",..: 143 172 469 469 412 469 288 288
##  $ PARTY_SIZE               : int  18 48 55 62 62 53 29 28 45 90 ...
##  $ MAIN_FLIGHT_DESTINATION  : Factor w/ 38 levels "ANU","ANY","AZS",..: 7 7 18 18 18 18 8 8 7 7 ...
##  $ START_DATE               : Date, format: "2018-01-20" "2018-01-13" ...
##  $ LENGTH_OF_STAY           : int  9 15 9 7 7 16 8 8 14 13 ...
##  $ BKG_DATE                 : Date, format: "2016-09-21" "2016-09-23" ...
##  $ REVENUE                  : num  38327 74204 535566 0 0 ...
##  $ TOTAL_COST               : num  37799 69152 111154 118328 95310 ...
##  $ MARGIN                   : num  -672 2019 423062 -118328 -95310 ...
##  $ ACCOMMODATION_STAR_RATING: Factor w/ 9 levels "","0","2.5","3",..: 8 7 8 8 7 8 6 6 8 8 ...
##  $ HOTEL_CJAIN_AFFILIATION  : Factor w/ 96 levels "ACCOR HOTELS",..: 5 82 8 8 8 8 5 5 64 64 ...
##  $ TTE                      : num  486 477 496 496 496 496 478 478 466 460 ...
##  $ Price_PerNight           : num  237 103 1082 0 0 ...
##  $ Wday_BookingDate         : Factor w/ 7 levels "Friday","Monday",..: 7 1 2 2 2 2 7 7 6 1 ...
##  $ Wday_StartDate           : Factor w/ 7 levels "Friday","Monday",..: 3 3 4 4 4 4 1 1 3 7 ...
```

```
str(All_2019)
```

```
## 'data.frame':    211361 obs. of  16 variables:
##  $ DESTINATION              : Factor w/ 38 levels "ANTIGUA","BRIDGETOWN",..: 3 11 12 20 28 20 37 22 1
##  $ PROPERTY_ID              : Factor w/ 834 levels "ANU","ANUBLU",..: 223 312 326 386 599 376 810 481
```

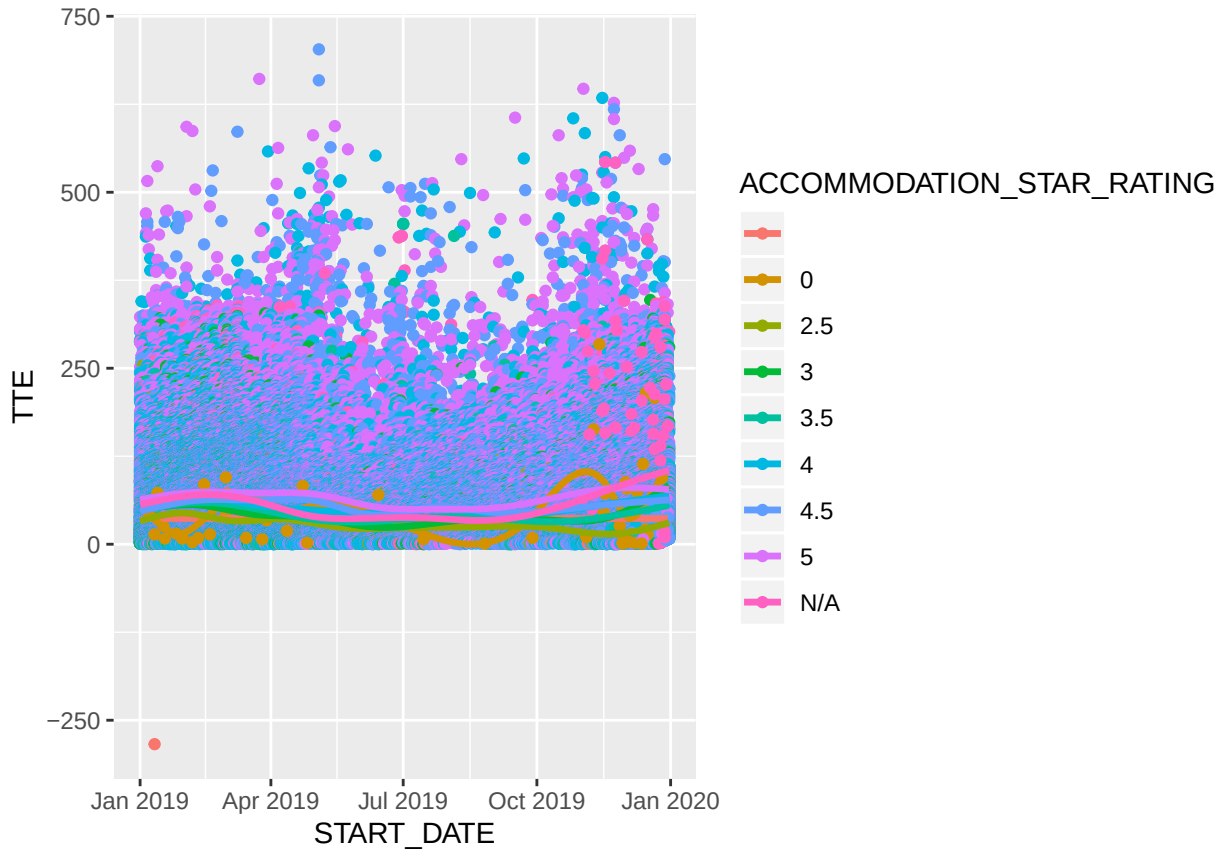
```
##  $ PARTY_SIZE                : int  48 56 43 84 33 171 7 14 35 13 ...
##  $ MAIN_FLIGHT_DESTINATION   : Factor w/ 38 levels "ANU","ANY","AZS",..: 7 10 11 16 21 16 27 18 18 18
##  $ START_DATE                : Date, format: "2019-05-04" "2019-03-24" ...
##  $ LENGTH_OF_STAY            : int  9 9 9 18 9 14 7 7 12 9 ...
##  $ BKG_DATE                  : Date, format: "2017-05-31" "2017-06-01" ...
##  $ REVENUE                   : num  69497 168387 160690 280387 41032 ...
##  $ TOTAL_COST                : num  57334 167931 143042 229295 35071 ...
##  $ MARGIN                    : num  12164 456 17649 51092 5961 ...
##  $ ACCOMMODATION_STAR_RATING: Factor w/ 9 levels "","0","2.5","3",..: 7 8 8 8 7 8 7 8 7 6 ...
##  $ HOTEL_CJAIN_AFFILIATION   : Factor w/ 96 levels "ACCOR HOTELS",..: 61 77 77 5 8 5 14 77 65 5 ...
##  $ TTE                       : num  703 661 593 587 659 537 586 516 531 558 ...
##  $ Price_PerNight            : num  161 334 415 185 138 ...
##  $ Wday_BookingDate          : Factor w/ 7 levels "Friday","Monday",..: 7 5 2 5 1 6 2 6 5 2 ...
##  $ Wday_StartDate            : Factor w/ 7 levels "Friday","Monday",..: 3 4 3 7 3 4 3 4 7 3 ...
```

```
ggplot(data = All_2019) +
  geom_point(mapping = aes(x= TTE, y = Price_PerNight, colour = ACCOMMODATION_STAR_RATING))
```




```
d <- ggplot(data=All_2019, aes(x=START_DATE, y=TTE,
                               colour=ACCOMMODATION_STAR_RATING))
d + geom_point() +
  geom_smooth(fill=NA, size=1.2)
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Final Data Preparation for Box Cox Model

```
df <- raw[c("DESTINATION", "PROPERTY_ID", "START_DATE", "PARTY_SIZE", "LENGTH_OF_STAY", "BKG_DATE", "MA
#df["PRE_BKG_DAYS"] <- df["START_DATE"] - df["BKG_DATE"]
# df["START_YEAR"] <- substr(df$START_DATE, 1, 4) # - as.Date(df$BKG_DATE,"%Y%m%d")
# df["START_MONTH"] <- substr(df$START_DATE, 5, 6)
# df["START_DAY_OF_MTH"] <- substr(df$START_DATE, 7, 8)
# df["START_MTH_AND_DAY"] <- substr(df$START_DATE, 5, 8)
str(df)
```

```
## 'data.frame':    652446 obs. of  7 variables:
##  $ DESTINATION   : Factor w/ 38 levels "ANTIGUA","BRIDGETOWN",..: 22 22 28 22 22 3 28 3 3 34 ...
##  $ PROPERTY_ID   : Factor w/ 834 levels "ANU","ANUBLU",..: 412 412 576 469 412 205 569 205 217 493 .
##  $ START_DATE    : int  20170623 20170109 20170123 20170518 20170506 20170127 20170114 20170116 2017
##  $ PARTY_SIZE    : int  27 16 35 31 7 86 29 38 26 75 ...
##  $ LENGTH_OF_STAY: int  15 11 20 8 7 19 11 11 8 9 ...
##  $ BKG_DATE      : int  20150615 20150819 20150909 20150914 20151014 20151016 20151019 20151026 2015
##  $ MARGIN        : num  7279 -2561 3146 3446 1516 ...
```

```
df_group <- aggregate(df[,4:7], df[,1:3], FUN = sum )
# df_group <- aggregate(df[-c("PROPERTY_ID","START_DATE")], df[c("PROPERTY_ID","START_DATE")], FUN = su
#df_group["MAX_OCCP"] <- df_group["PARTY_SIZE"]
#aggregate(df$MAX_OCCP, by = list(df$PROPERTY_ID), max)
```

```
#tapply(df_group$Value, df$PROPERTY_ID, max)
# df_group %>%
#   group_by(PROPERTY_ID) %>%
#   summarise(MAX_OCCP = max(PARTY_SIZE))
df_group["START_YEAR"] <- substr(df_group$START_DATE, 1, 4) # - as.Date(df$BKG_DATE,"%Y%m%d")
df_group["START_MONTH"] <- substr(df_group$START_DATE, 5, 6)
df_group["START_DAY_OF_MTH"] <- substr(df_group$START_DATE, 7, 8)
df_group["START_MTH_AND_DAY"] <- substr(df_group$START_DATE, 5, 8)
str(df_group)
```

```
## 'data.frame':    186610 obs. of  11 variables:
##  $ DESTINATION      : Factor w/ 38 levels "ANTIGUA","BRIDGETOWN",..: 1 1 1 1 1 1 30 30 30 30 ...
##  $ PROPERTY_ID      : Factor w/ 834 levels "ANU","ANUBLU",..: 7 8 13 18 19 20 22 23 24 25 ...
##  $ START_DATE       : int  20170101 20170101 20170101 20170101 20170101 20170101 20170101 20170101 20
##  $ PARTY_SIZE       : int  3 34 3 20 18 2 19 8 9 2 ...
##  $ LENGTH_OF_STAY   : int  8 64 14 65 35 7 42 7 28 7 ...
##  $ BKG_DATE         : int  20160707 181448516 40321951 181448728 100805200 20160731 120965043 201609
##  $ MARGIN           : num  716 -348 1158 8097 7931 ...
##  $ START_YEAR       : chr  "2017" "2017" "2017" "2017" ...
##  $ START_MONTH      : chr  "01" "01" "01" "01" ...
##  $ START_DAY_OF_MTH : chr  "01" "01" "01" "01" ...
##  $ START_MTH_AND_DAY: chr  "0101" "0101" "0101" "0101" ...
```

```
kable(head(df_group,20))
```

DESTINATION	PROPERTY_ID	START_DATE	PARTY_SIZE	LENGTH_OF_STAY	BKG_DATE	MARGI
ANTIGUA	ANUGAB	20170101	3	8	20160707	715.5
ANTIGUA	ANUHAL	20170101	34	64	181448516	-347.9
ANTIGUA	ANUJOL	20170101	3	14	40321951	1158.3
ANTIGUA	ANUSAN	20170101	20	65	181448728	8097.3
ANTIGUA	ANUSJA	20170101	18	35	100805200	7931.3
ANTIGUA	ANUSUG	20170101	2	7	20160731	779.5
SAMANA	AZSBCO	20170101	19	42	120965043	3867.8
SAMANA	AZSBSR	20170101	8	7	20160920	2244.8
SAMANA	AZSLCA	20170101	9	28	80644175	4294.2
SAMANA	AZSLUX	20170101	2	7	20161017	679.1
SAMANA	AZSPOR	20170101	23	42	120965624	7657.9
SAMANA	AZSVIV	20170101	16	57	161286762	4360.1
CAYO COCO	CCCDAI	20170101	5	14	40322044	1076.2
CAYO COCO	CCCMCA	20170101	22	42	120966147	4926.6
CAYO COCO	CCCMEL	20170101	2	8	20161223	1509.4
CAYO COCO	CCCMFL	20170101	24	21	60483073	4639.2
CAYO COCO	CCCMJA	20170101	16	28	60483260	4148.8
CAYO COCO	CCCMOJ	20170101	15	21	60483073	2990.1
CAYO COCO	CCCPCC	20170101	10	28	80644097	2571.1
CAYO COCO	CCCPLA	20170101	8	14	40322141	3326.7

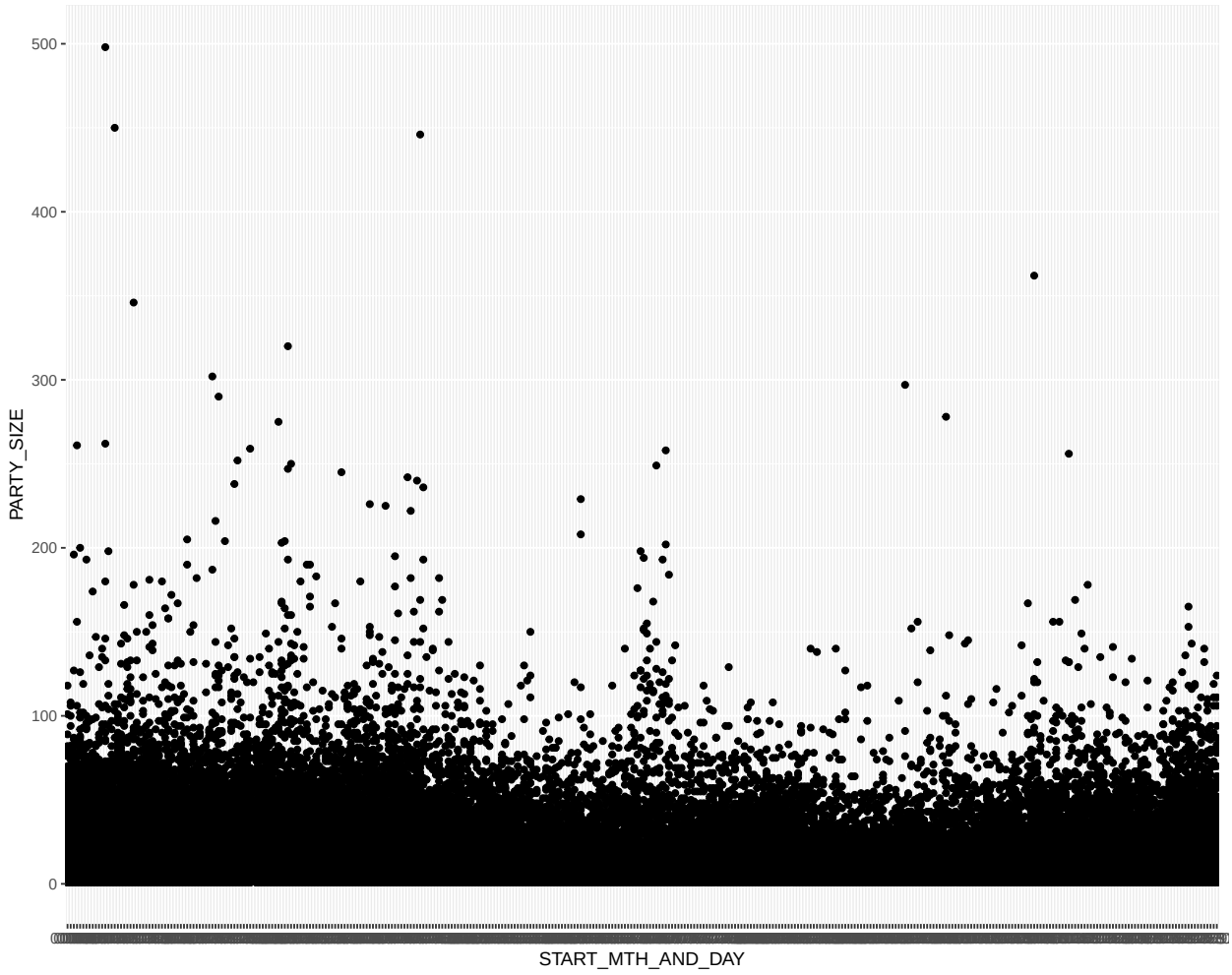
```
kable(tail(df_group,20))
```

	DESTINATION	PROPERTY_ID	START_DATE	PARTY_SIZE	LENGTH_OF_STAY	BKG_DATE
186591	CANCUN	CUNMOO	20210412	30	7	2019121
186592	CANCUN	CUNKAN	20210415	40	10	2019111
186593	MONTEGO BAY	MBJPAL	20210421	30	7	2019120
186594	PUNTA CANA	PUJPAL	20210424	50	7	2019112
186595	MONTEGO BAY	MBJLAD	20210426	40	7	2019111
186596	CANCUN	CUNPRI	20210429	16	7	2019072
186597	PUNTA CANA	PUJDPC	20210429	40	10	2019111
186598	CANCUN	CUNAKU	20210430	40	7	2019091
186599	CANCUN	CUNBBA	20210501	30	7	2019121
186600	CANCUN	CUNCAR	20210501	30	7	2019110
186601	PUNTA CANA	PUJDPC	20210502	40	7	2019111
186602	PUNTA CANA	PUJROY	20210502	100	7	2019091
186603	PUNTA CANA	PUJTUR	20210503	50	7	2019111
186604	MONTEGO BAY	MBJRBA	20210515	40	7	2019120
186605	CANCUN	CUNMGR	20210516	40	7	2019121
186606	CANCUN	CUNNSC	20210521	40	7	2019121
186607	CANCUN	CUNLBA	20210522	30	7	2019103
186608	MONTEGO BAY	MBJBAH	20210703	40	7	2019110
186609	CANCUN	CUNCAR	20210710	30	7	2019092
186610	GENERIC	CUNMGR	20210827	30	7	2019121

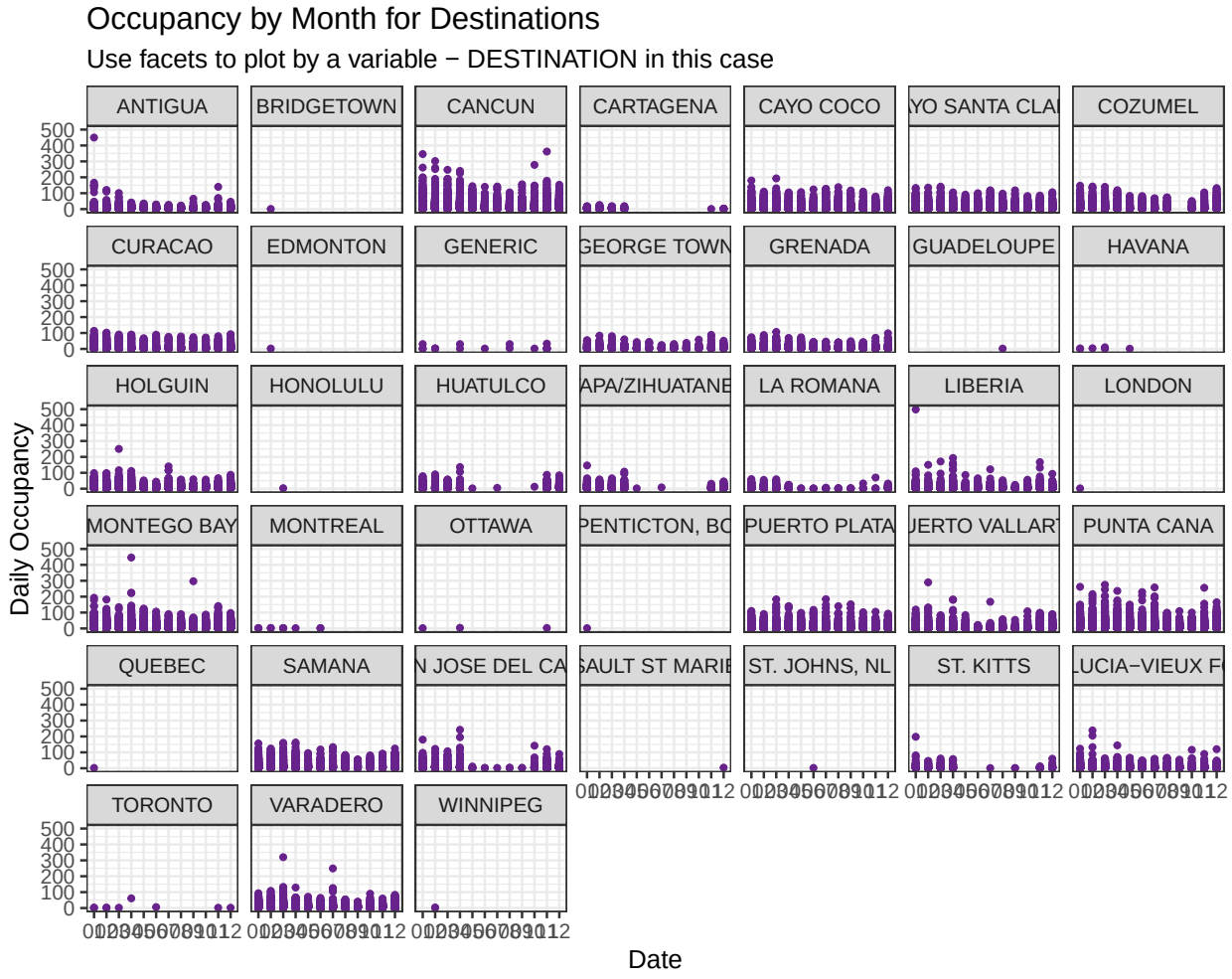
```
summary(df_group)
```

```
##           DESTINATION     PROPERTY_ID       START_DATE         PARTY_SIZE
##  CANCUN          :52384   N/A    :  5767   Min.   :20170101   Min.   :  1.00
##  PUNTA CANA      :31798   MBJBAH :  1215   1st Qu.:20171215   1st Qu.:  2.00
##  VARADERO        :20515   CUNCOB :  1174   Median :20181011   Median :  5.00
##  MONTEGO BAY     :20388   CUNSIA :  1166   Mean   :20183186   Mean   : 10.42
##  PUERTO VALLARTA:11124   MBJRBA :  1161   3rd Qu.:20190530   3rd Qu.: 12.00
##  CAYO COCO       : 6857   CUNBAH :  1058   Max.   :20210827   Max.   :498.00
##  (Other)         :43544   (Other):175069
##  LENGTH_OF_STAY      BKG_DATE             MARGIN          START_YEAR
##  Min.   :  1.00   Min.   :2.015e+07   Min.   :-6216023   Length:186610
##  1st Qu.:  7.00   1st Qu.:2.019e+07   1st Qu.:      10   Class :character
##  Median : 14.00   Median :4.036e+07   Median :     533   Mode  :character
##  Mean   : 25.71   Mean   :7.055e+07   Mean   :    2228
##  3rd Qu.: 29.00   3rd Qu.:8.075e+07   3rd Qu.:    1602
##  Max.   :999.00   Max.   :1.897e+09   Max.   : 6004311
##
##  START_MONTH        START_DAY_OF_MTH   START_MTH_AND_DAY
##  Length:186610      Length:186610      Length:186610
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
```

```
ggplot(aes(x=START_MTH_AND_DAY,y=PARTY_SIZE),data=df_group)+
  geom_point()
```




```
df_group %>%
  na.omit() %>%
ggplot(aes(x = START_MONTH, y = PARTY_SIZE)) +
      geom_point(color = "darkorchid4") +
      facet_wrap( ~ DESTINATION ) +
      labs(title = "Occupancy by Month for Destinations",
           subtitle = "Use facets to plot by a variable - DESTINATION in this case",
           y = "Daily Occupancy",
           x = "Date") + theme_bw(base_size = 15) #+
```




```
    # adjust the x axis breaks
    #scale_x_date(date_breaks = "5 years", date_labels = "%m-%Y")
```

```
# Make certain adjustments to the data in the following columns
df_group$START_DATE <- as.Date(as.character(df_group$START_DATE), format="%Y%m%d")
# BKG_DATE
df_group$BKG_DATE <- as.Date(as.character(df_group$BKG_DATE), format="%Y%m%d")

# TTE - Time between vacation start and booking date
df_group$TTE <- df_group$START_DATE - df_group$BKG_DATE
df_group$TTE <- as.numeric(df_group$TTE)

# Day of week of booking
df_group$Wday_BookingDate <- weekdays(as.Date(df_group$BKG_DATE))
df_group$Wday_BookingDate <- as.factor(df_group$Wday_BookingDate)
# Day of week of start of stay
df_group$Wday_StartDate <- weekdays(as.Date(df_group$START_DATE))
df_group$Wday_StartDate <- as.factor(df_group$Wday_StartDate)
```

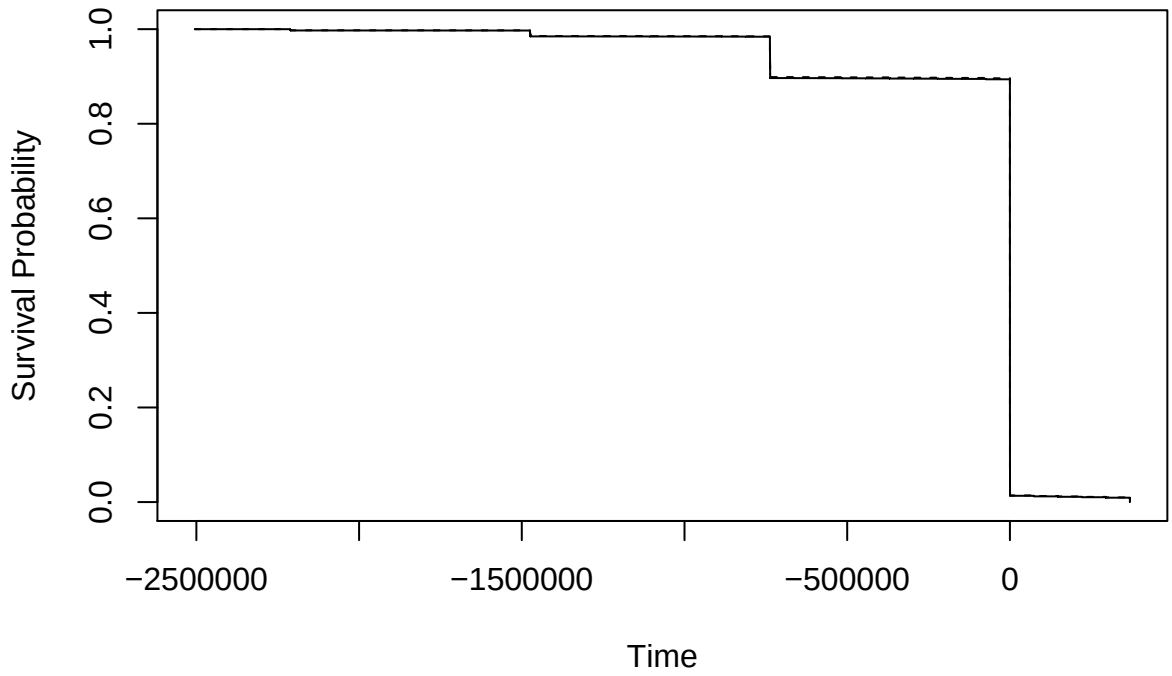
```
names(df_group)
```

```
##  [1] "DESTINATION"       "PROPERTY_ID"       "START_DATE"
##  [4] "PARTY_SIZE"        "LENGTH_OF_STAY"    "BKG_DATE"
##  [7] "MARGIN"            "START_YEAR"        "START_MONTH"
## [10] "START_DAY_OF_MTH"  "START_MTH_AND_DAY" "TTE"
## [13] "Wday_BookingDate"  "Wday_StartDate"
```

```
box_cox<- coxph(Surv(TTE) ~ 1 , data = df_group)
summary(box_cox)
```

```
## Call:  coxph(formula = Surv(TTE) ~ 1, data = df_group)
##
## Null model
##   log likelihood= -864014.2
##   n=83609 (103001 observations deleted due to missingness)
```

```
curve_cox <- survfit(box_cox)
N <- length(unique(lung$PROPERTY_ID))
group <- lung$PROPERTY_ID
plot(curve_cox,xlab="Time",ylab="Survival Probability", mark.time = F, col=1:N)
```




```
# legend(
#   "topright",
#   legend=unique(group),
#   col=1:N,
```

```
#   horiz=FALSE,
#   bty='n', lty=2:3)

#plot(curve_cox)
```

6. Final Model Analysis and Selection

Cost Analysis We cost for getting incorrect model would be general mistrust on the application by the agents and loss of revenue for the hotel and the travel agents in genrat. There is no other cost that can be assicated with this model.As this could potentially affect the business profit and bottom line, we need to try and keep error to a minimum.

Model Comparison We have used the following two models to do prediction for the instacart shopping use case.The following Machine Learning Algorithms were used in this analysis: Kaplan-Meier - According to IBM The Kaplan-Meier procedure is a method of estimating time-to-event models in the presence of censored cases. The Kaplan-Meier model is based on estimating conditional probabilities at each time point when an event occurs and taking the product limit of those probabilities to estimate the survival rate at each point in time. Box Cox - A Box Cox transformation is a way to transform non-normal dependent variables into a normal shape. Normality is an important assumption for many statistical techniques; if your data isn't normal, applying a Box-Cox means that you are able to run a broader number of tests.

KM estimates the survival curve for one group only. KM doesnt provide any kind of comparison between groups. As we are only looking at prebooking dates for our model KM is a better fit for us compared to box cox that works with multiple groups.

Selected Model: We selected Kaplier-Maier model as it gives us much comfortable results that we can take it to the app and explain it to the end client. ## 7. Deployment

Shiny App Url: https://csml1000-group8.shinyapps.io/CSML1000-Group8-CourseProject/ #### Summary Explanation

- Limitations of our analysis:
  - Due to processing and resource limitations we used a full datset but removed lot of features and split it into years.
  - The analysis is based on data provided by Vaccation Company and may inherit any biases that exists in their customer base relative to the general population.
- Further steps:
  - The analysis can be expanded to include all of the original data as well as any other similar sources that may be available. We also wanted to tackle the Ethical framework for the model by adding the percentages rules along with our time to event dates. The percentage would allow us to protect actors from creatiung huge nmistakes for themselves which would either be underiable, illegal or dangerous for the actors.
- Ethical Framework Questions:
  - Can a malicious actor infer information about individuals from your system? No. There is no PII present.
  - Are you able to identify anomalous activity on your system that might indicate a security breach? This would need to be considered for each specific deployment.
  - Do you have a plan to monitor for poor performance on individuals or subgroups? N/A since No demographic data is present.

- Do you have a plan to log and store historical predictions if a consumer requests access in the future? N/A since No demographic data is present.
- Have you documented model retraining cycles and can you confirm that a subject's data has been removed from models? N/A since No demographic data is present.

References

Yihui Xie, J. J. Allaire, Garrett Grolemund, 2019, R Markdown: The Definitive Guide https://bookdown.org/yihui/rmarkdown/markdown-syntax.html

Jonathan McPherson, 2016, R Notebooks https://blog.rstudio.com/2016/10/05/r-notebooks

Adam Kardash, Patricia Kosseim, 2018, Responsible AI in Consumer Enterprise, integrate.ai

J Marcus W. Beck, 2018, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6262849/