Taxa de ocupação em rotas rodoviárias

Marcos Ikino

Introdução

Este estudo consiste em uma análise de um conjunto de dados relacionado ao transporte de passageiros efetuado por empresas rodoviárias nas mais diversas rotas e períodos de tempo, ao longo dos meses de Julho e Agosto do ano de 2018.

O estudo está subdividido em duas partes. Na primeira parte o objetivo será realizar uma análise de negócio sobre o conjunto de dados, procurando responder questões envolvidas na análise, que abordam em sua base, as taxas de ocupação das viagens de ônibus, ao identificar padrões nas rotas ao longo de períodos do dia e do mês, gerando oportunidades para aperfeiçoar os itinerários das rotas, reduzindo custos e acrescentando receitas, ao possibilitar que gargalos operacionais sejam identificados e estudados. Com a identificação destes padrões será proposta mudança no sistema de precificação do negócio nas rotas que possuam potenciais de ganhos de receitas.

Na segunda parte do estudo será dado enfoque na construção de um modelo preditivo para tentar prever o número de assentos ocupados por viagem em função da empresa, rota, períodos de tempo, preço da passagem e número de assentos disponíveis, e a influência recebida do sistema de precificação pré-existente realizada na primeira parte na análise de negócio ao construir o modelo preditivo.

Parte 1 - Análise e proposta de negócio

Questões a serem desenvolvidas neste estudo:

1. Como a taxa de ocupação evolui ao longo das horas do dia e dos dias do mês?
2. Quais as viações com maiores taxas de ocupação e receita diárias?!
3. Qual é o potencial de receita que não é ganho nas 100 rotas com piores taxas de ocupação?
4. Rotas com muitos horários de partida por dia impactam na taxa de ocupação geral?
5. Há alguma rota de alguma viação em que é possível sugerir mudanças nos itinerários devido à taxa de ocupação observada?
6. Proposta de negócio: criar um conjunto de critérios de precificação dos assentos que potencialize a receita.

1. Variáveis dos dados

bus_company: Nome da viação de ônibus.
id_vehicle: ID do veículo. É único dentro da mesma viação.
id_route: ID da rota a ser feita pelo veículo.
departure_date: Data de partida, no formato AAAA-MM-DD
departure_time: Horário de partida, no formato HH:MM:SS
seat_price: Preço de cada assento.
occupied_seats: Assentos ocupados no veículo.
total_seats: Total de assentos no veículo.

2. Análise exploratória dos dados

2.1 Importando os dados

```
library(readxl)

data <- read.csv('tripbus.csv')
str(data)
```

```
## 'data.frame':    566176 obs. of  8 variables:
##  $ bus_company   : Factor w/ 130 levels "Company1","Company10",..: 1 43 1 54 54
65 54 54 43 65 ...
##  $ id_vehicle    : num  3.28e+18 1.13e+19 3.28e+18 4.16e+18 8.23e+18 ...
##  $ id_route      : int  23919 33045 13188 178967 66359 10775 178967 66359 33183
10775 ...
##  $ departure_date: Factor w/ 62 levels "2018-07-01","2018-07-02",..: 1 1 1 1 1 1
1 1 1 1 ...
##  $ departure_time: Factor w/ 839 levels "1899-12-31 02:00:00",..: 1 1 1 1 1 1 1
1 1 1 ...
##  $ seat_price    : num  16.4 50 44.6 25.7 76.8 ...
##  $ occupied_seats: int  25 25 34 12 24 26 7 10 42 4 ...
##  $ total_seats   : int  29 44 39 46 46 30 11 11 48 16 ...
```

```
summary(data)
```

```
##     bus_company       id_vehicle           id_route
##  Company40: 33396   Min.   :1.042e+14   Min.   :     2
##  Company1 : 23004   1st Qu.:4.514e+18   1st Qu.: 11033
##  Company43: 22917   Median :9.036e+18   Median : 34060
##  Company20: 18308   Mean   :9.167e+18   Mean   : 63536
##  Company4 : 17319   3rd Qu.:1.383e+19   3rd Qu.: 79712
##  Company44: 13665   Max.   :1.844e+19   Max.   :621028
##  (Other)  :437567
```

```
##     departure_date               departure_time     seat_price
##  2018-07-29: 10813   1899-12-31 19:00:00: 13901   Min.   :   1.65
##  2018-07-30: 10676   1899-12-31 09:00:00: 13257   1st Qu.:  29.90
##  2018-08-10: 10514   1899-12-31 13:00:00: 12663   Median :  52.01
##  2018-08-24: 10238   1899-12-31 12:00:00: 12238   Mean   :  70.33
##  2018-07-13: 10147   1899-12-31 14:00:00: 12176   3rd Qu.:  93.36
##  2018-08-05: 10135   1899-12-31 07:00:00: 12001   Max.   :1200.00
##  (Other)   :503653   (Other)            :489940
##  occupied_seats   total_seats
##  Min.   : 0.00   Min.   : 0.00
##  1st Qu.:10.00   1st Qu.:28.00
##  Median :20.00   Median :37.00
##  Mean   :21.85   Mean   :37.87
##  3rd Qu.:31.00   3rd Qu.:48.00
##  Max.   :86.00   Max.   :87.00
##
```

2.2 Transformações nas variáveis dos dados

```
# Transformação para variável categórica
categ_var <- c('id_vehicle', 'id_route')
data[,categ_var] <- lapply(data[,categ_var], factor)

# Transformação para formato de data
date_var <- c('departure_date', 'departure_time')
data[,date_var] <- lapply(data[,date_var], as.POSIXct)

# Tranformação para variável numérica
data$seat_price <- as.numeric(data$seat_price)

# Segregação dos dados das variáveis departure_date e departure_time
date_split <- data.frame(date = data$departure_date,
                 year = as.factor(format(data$departure_date, format = "%Y")),
                 month = as.factor(format(data$departure_date, format = "%m")),
                 day = as.factor(format(data$departure_date, format = "%d"))
                 )

time_split <- data.frame(time = data$departure_time,
                 hour = as.factor(format(data$departure_time, format = "%H")),
                 minute = as.factor(format(data$departure_time, format = "%M")),
                 second = as.factor(format(data$departure_time, format = "%S"))
                 )
```

```r
# Reconfigurando e reposicionando as novas variáveis
data <- cbind(data, date_split, time_split)
data[, c("date", "time", "second")] <- NULL
col_order <- c("bus_company", "id_vehicle", "id_route", "departure_date", "year",
"month", "day", "departure_time","hour", "minute", "seat_price", "occupied_seats",
"total_seats")
data <- data[, col_order]
```

```r
str(data)
```

```
## 'data.frame':    566176 obs. of  13 variables:
##  $ bus_company   : Factor w/ 130 levels "Company1","Company10",..: 1 43 1 54 54
65 54 54 43 65 ...
##  $ id_vehicle    : Factor w/ 21702 levels "1.04245e+14",..: 3975 13536 3975 5023
10056 20313 15974 21131 8145 16675 ...
##  $ id_route      : Factor w/ 1495 levels "2","4","5","6",..: 490 573 301 1293
831 202 1293 831 580 202 ...
##  $ departure_date: POSIXct, format: "2018-07-01" "2018-07-01" ...
##  $ year          : Factor w/ 1 level "2018": 1 1 1 1 1 1 1 1 1 1 ...
##  $ month         : Factor w/ 2 levels "07","08": 1 1 1 1 1 1 1 1 1 1 ...
##  $ day           : Factor w/ 31 levels "01","02","03",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ departure_time: POSIXct, format: "1899-12-31 02:00:00" "1899-12-31 02:00:00"
...
##  $ hour          : Factor w/ 22 levels "02","03","04",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ minute        : Factor w/ 60 levels "00","01","02",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ seat_price    : num  16.4 50 44.6 25.7 76.8 ...
##  $ occupied_seats: int  25 25 34 12 24 26 7 10 42 4 ...
##  $ total_seats   : int  29 44 39 46 46 30 11 11 48 16 ...
```

2.3 Criação da variável occupation_rate

Esta nova variável representa a taxa de ocupação dos assentos em cada viagem, sendo definida como a divisão entre as variáveis occupied_seats e total_seats.

Valores iguais a zero das variáveis total_seats e occupied_seats serão removidos por gerar incongruência nos cálculos.

```r
library(dplyr)

data <- data[!(data$total_seats == 0) & !(data$occupied_seats == 0),]
```

```r
# Criação da variável occupation_rate

data <- data %>%
  mutate(occupation_rate = occupied_seats/total_seats)

str(data)
```

```
## 'data.frame':    554155 obs. of  14 variables:
##  $ bus_company    : Factor w/ 130 levels "Company1","Company10",..: 1 43 1 54 54
65 54 54 43 65 ...
##  $ id_vehicle     : Factor w/ 21702 levels "1.04245e+14",..: 3975 13536 3975
5023 10056 20313 15974 21131 8145 16675 ...
##  $ id_route       : Factor w/ 1495 levels "2","4","5","6",..: 490 573 301 1293
831 202 1293 831 580 202 ...
##  $ departure_date : POSIXct, format: "2018-07-01" "2018-07-01" ...
##  $ year           : Factor w/ 1 level "2018": 1 1 1 1 1 1 1 1 1 1 ...
##  $ month          : Factor w/ 2 levels "07","08": 1 1 1 1 1 1 1 1 1 1 ...
##  $ day            : Factor w/ 31 levels "01","02","03",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ departure_time : POSIXct, format: "1899-12-31 02:00:00" "1899-12-31 02:00:00"
...
##  $ hour           : Factor w/ 22 levels "02","03","04",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ minute         : Factor w/ 60 levels "00","01","02",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ seat_price     : num  16.4 50 44.6 25.7 76.8 ...
##  $ occupied_seats : int  25 25 34 12 24 26 7 10 42 4 ...
##  $ total_seats    : int  29 44 39 46 46 30 11 11 48 16 ...
##  $ occupation_rate: num  0.862 0.568 0.872 0.261 0.522 ...
```

3. Questões para análise de negócio

As questões inicialmente propostas na análise de negócio serão desenvolvidas ao longo desta seção.

3.1 Como a taxa de ocupação evolui ao longo das horas do dia e dos dias do mês?

3.1.1 Cálculo das horas do dia

```r
library(ggplot2)

# Cálculo da taxa de ocupação média por hora do dia
data1 <- data %>%
  select(hour, occupied_seats, total_seats) %>%
  group_by(hour) %>%
  summarise(occup_rate_mean = round(mean(sum(occupied_seats)/sum(total_seats)),2))
data1
```

```
## # A tibble: 22 x 2
##    hour  occup_rate_mean
##    <fct>           <dbl>
##  1 02               0.68
##  2 03               0.63
##  3 04               0.55
##  4 05               0.45
##  5 06               0.48
##  6 07               0.53
##  7 08               0.56
##  8 09               0.56
##  9 10               0.56
## 10 11               0.55
## # ... with 12 more rows
```

```
data1 %>%
  ggplot(aes(hour, occup_rate_mean)) +
  geom_col(colour = "black", fill = "dodgerblue3") +
  theme(text = element_text(size = 11),
        axis.text.x = element_text(angle = 90, hjust = 0))
```

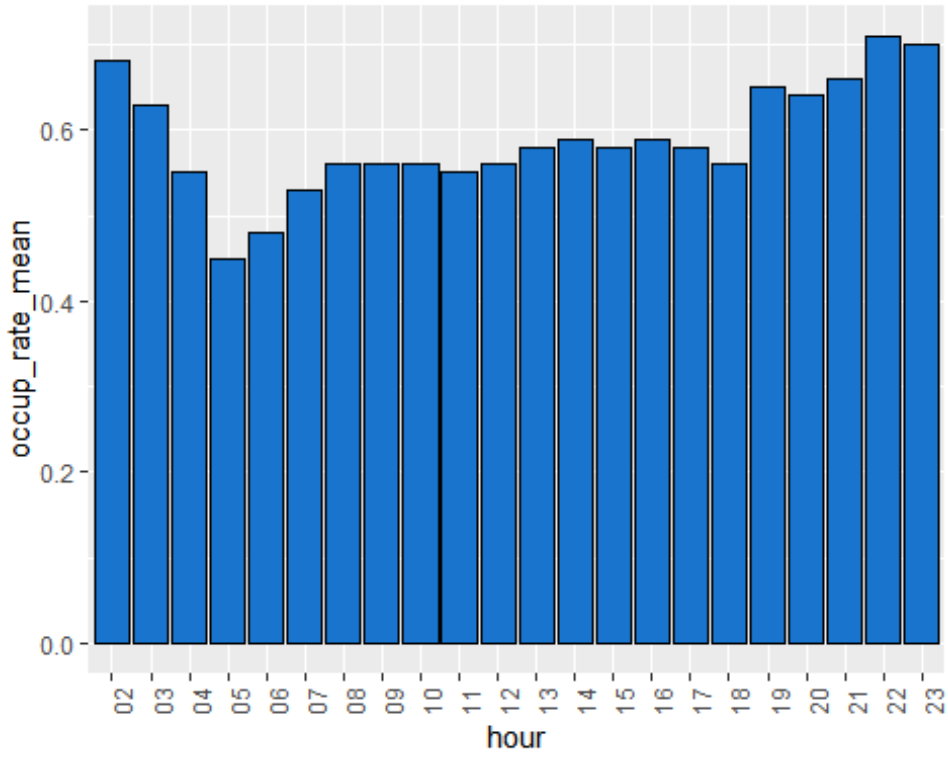

Gráfico 1: horas do dia x taxa de ocupação média

3.1.2 Cálculo dos dias do mês

```
# Cálculo da taxa de ocupação média por dia do mês

data2 <- data %>%
  select(day, occupied_seats, total_seats) %>%
  group_by(day) %>%
  summarise(occup_rate_mean = round(mean(sum(occupied_seats)/sum(total_seats)),2))
data2
```

```
## # A tibble: 31 x 2
##    day   occup_rate_mean
##    <fct>           <dbl>
##  1 01              0.55
##  2 02              0.52
##  3 03              0.56
##  4 04              0.54
##  5 05              0.62
##  6 06              0.61
##  7 07              0.56
##  8 08              0.580
##  9 09              0.6
## 10 10              0.64
## # ... with 21 more rows
```

```
data2 %>%
  ggplot(aes(day, occup_rate_mean)) +
  geom_col(fill = "darkviolet") +
  theme(text = element_text(size = 11),
        axis.text.x = element_text(angle = 90, hjust = 0))
```

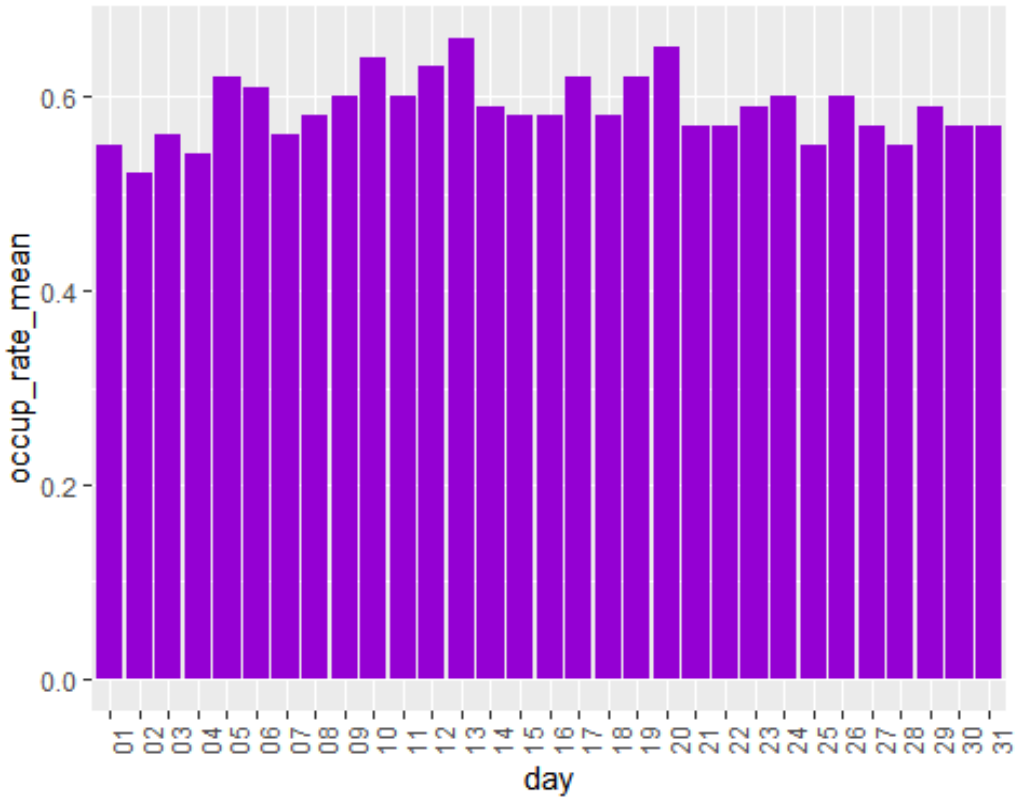

Gráfico 2: dias do mês x taxa de ocupação média

3.2 Quais as viações com maiores taxas de ocupação e receita diárias?

3.2.1 Criação da variável revenue, que representam receitas geradas pelas vendas das passagens.

```
data <- data %>%
  mutate(revenue = seat_price * occupied_seats)

str(data)
```

```
## 'data.frame':    554155 obs. of  15 variables:
##  $ bus_company    : Factor w/ 130 levels "Company1","Company10",..: 1 43 1 54 54
65 54 54 43 65 ...
##  $ id_vehicle     : Factor w/ 21702 levels "1.04245e+14",..: 3975 13536 3975
5023 10056 20313 15974 21131 8145 16675 ...
##  $ id_route       : Factor w/ 1495 levels "2","4","5","6",..: 490 573 301 1293
831 202 1293 831 580 202 ...
##  $ departure_date : POSIXct, format: "2018-07-01" "2018-07-01" ...
##  $ year           : Factor w/ 1 level "2018": 1 1 1 1 1 1 1 1 1 1 ...
##  $ month          : Factor w/ 2 levels "07","08": 1 1 1 1 1 1 1 1 1 1 ...
##  $ day            : Factor w/ 31 levels "01","02","03",..: 1 1 1 1 1 1 1 1 1 1
...
```

```
##  $ departure_time : POSIXct, format: "1899-12-31 02:00:00" "1899-12-31 02:00:00"
...
##  $ hour           : Factor w/ 22 levels "02","03","04",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ minute         : Factor w/ 60 levels "00","01","02",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ seat_price     : num  16.4 50 44.6 25.7 76.8 ...
##  $ occupied_seats : int  25 25 34 12 24 26 7 10 42 4 ...
##  $ total_seats    : int  29 44 39 46 46 30 11 11 48 16 ...
##  $ occupation_rate: num  0.862 0.568 0.872 0.261 0.522 ...
##  $ revenue        : num  409 1250 1516 308 1842 ...
```

3.2.2 Cálculo da taxa de ocupação média por empresa

```
data3 <- data %>%
  select(bus_company, occupation_rate) %>%
  group_by(bus_company) %>%
  summarise(occup_rate_mean = round(mean(occupation_rate),2)) %>%
  arrange(desc(occup_rate_mean))
data3
```

```
## # A tibble: 130 x 2
##    bus_company occup_rate_mean
##    <fct>                 <dbl>
##  1 Company23              0.93
##  2 Company100             0.86
##  3 Company89              0.83
##  4 Company124             0.82
##  5 Company42              0.81
##  6 Company20              0.79
##  7 Company37              0.79
##  8 Company110             0.78
##  9 Company121             0.78
## 10 Company48              0.78
## # ... with 120 more rows
```

3.2.3 Cálculo da taxa de receita média por empresa

Nos cálculos abaixo são apresentadas as empresas com as maiores receitas médias, o que não estão atreladas necessariamente com as maiores taxas de ocupação.

```
data4 <- data %>%
  group_by(bus_company) %>%
  summarise(revenue_mean = round(mean(revenue),0),
            occup_rate_mean = round(mean(occupation_rate),2)) %>%
```

```
  arrange(desc(revenue_mean))
data4
```

```
## # A tibble: 130 x 3
##    bus_company revenue_mean occup_rate_mean
##    <fct>              <dbl>           <dbl>
##  1 Company113          7673            0.65
##  2 Company27           6760            0.71
##  3 Company103          6722            0.69
##  4 Company23           5869            0.93
##  5 Company123          5835            0.63
##  6 Company115          5689            0.66
##  7 Company100          4201            0.86
##  8 Company98           3913            0.77
##  9 Company22           3911            0.72
## 10 Company130          3795            0.64
## # ... with 120 more rows
```

3.3 Qual é o potencial de receita que não é ganho nas 100 rotas com piores taxas de ocupação?

3.3.1 Cálculo do potencial de ganho (potential_revenue) por viagem e por rota, para as piores taxas de ocupação

Para o cálculo do potencial de receita que não é ganho nas piores rotas, utilizou-se das seguintes variáveis com suas respectivas funcionalidades:

worse_occup_rate_mean: piores taxas médias de ocupação.

general_occup_rate_mean: taxa média geral de ocupação, com valor de 0.59

seat_price_mean: preço médio do assento.

total_seats_mean: número total médio de assentos.

worse_revenue_gain: produtos das variáveis worse_occup_rate_mean, seat_price_mean e total_seats_mean.

optimized_revenue_gain: produtos das variáveis total_seats_mean, seat_price_mean e general_occup_rate_mean.

potential_revenue_gain: diferença entre a variável optimized_revenue_gain e worse_revenue_gain.

```
# Piores rotas com menores taxas médias de ocupação
data5 <- data %>%
  filter(occupation_rate > 0) %>%
  group_by(id_route) %>%
  summarise(worse_occup_rate_mean = round(mean(occupation_rate),4)) %>%
  arrange(worse_occup_rate_mean)

data5 <- data5[1:100, ]
```

```r
head(data5, 10)
```

```
## # A tibble: 10 x 2
##    id_route worse_occup_rate_mean
##    <fct>                   <dbl>
##  1 197739                 0.0463
##  2 86131                  0.0465
##  3 581226                 0.0466
##  4 79576                  0.0568
##  5 41563                  0.0656
##  6 1722                   0.066
##  7 1732                   0.0716
##  8 25127                  0.0721
##  9 1735                   0.0834
## 10 63341                  0.087
```

```r
# Criação de variáveis para cálculos de otimização de receitas

# Valor médio de taxa de ocupação geral
mean(data$occupation_rate)
```

```
## [1] 0.59
```

```r
data6 <- data %>%
  filter(occupation_rate > 0) %>%
  group_by(id_route) %>%
  summarise(worse_occup_rate_mean = round(mean(occupation_rate),4),
            general_occup_rate_mean = 0.59,
            seat_price_mean = round(mean(seat_price),2),
            total_seats_mean = as.integer(mean(total_seats)),
            worse_revenue_gain = round(worse_occup_rate_mean * seat_price_mean *
total_seats_mean,2),
            optimized_revenue_gain = round(total_seats_mean * seat_price_mean *
general_occup_rate_mean,2),
            potential_revenue_gain = optimized_revenue_gain - worse_revenue_gain
            ) %>%
  arrange(worse_occup_rate_mean)

data6 <- data6[1:100, ]
data6[, 1:4]
```

```
## # A tibble: 100 x 4
##    id_route worse_occup_rate_mean general_occup_rate_mean seat_price_mean
##    <fct>                   <dbl>                   <dbl>           <dbl>
##  1 197739                 0.0463                    0.59            5.25
##  2 86131                  0.0465                    0.59           11.3
##  3 581226                 0.0466                    0.59           21.8
##  4 79576                  0.0568                    0.59           30.6
```

```
##  5 41563                   0.0656                    0.59           32.3
##  6 1722                    0.066                     0.59            7.94
##  7 1732                    0.0716                    0.59            5.64
##  8 25127                   0.0721                    0.59           22.2
##  9 1735                    0.0834                    0.59            4.3
## 10 63341                   0.087                     0.59            9.79
## # ... with 90 more rows
```

```
data6[, 5:8]
```

```
## # A tibble: 100 x 4
##    total_seats_mean worse_revenue_ga~ optimized_revenue~ potential_revenue~
##               <int>             <dbl>              <dbl>              <dbl>
##  1               45              10.9               139.               128.
##  2               38              19.9               252.               233.
##  3               37              37.6               476.               439.
##  4               40              69.5               722.               652.
##  5               43              91.1               819.               728.
##  6               43              22.5               201.               179.
##  7               43              17.4               143.               126.
##  8               39              62.5               512.               449.
##  9               41              14.7               104.                89.3
## 10               37              31.5               214.               182.
## # ... with 90 more rows
```

3.4 Rotas com muitos horários de partida por dia impactam na taxa de ocupação geral?

Na tabela abaixo, a variável depart_day_mean representa o número médio de partidas totais por dia, juntamente com os seus valores de taxas de ocupação médios por rota, em ordem decrescente de número de partidas por dia. Os cálculos demonstram que o número de partidas não impacta na taxa de ocupação.

```
data7 <- data %>%
  select(id_route, departure_date, occupation_rate) %>%
  group_by(id_route, departure_date) %>%
  summarise(total_depart_day = n(), occup_rate_mean =
round(mean(occupation_rate),2))
data7
```

```
## # A tibble: 69,803 x 4
## # Groups:   id_route [1,495]
##    id_route departure_date      total_depart_day occup_rate_mean
##    <fct>    <dttm>                         <int>           <dbl>
##  1 2        2018-07-01 00:00:00              112           0.570
```

```
##  2 2        2018-07-02 00:00:00              105           0.5
##  3 2        2018-07-03 00:00:00              100           0.51
##  4 2        2018-07-04 00:00:00              101           0.45
##  5 2        2018-07-05 00:00:00               27           0.81
##  6 2        2018-07-06 00:00:00               29           0.87
##  7 2        2018-07-07 00:00:00               89           0.81
##  8 2        2018-07-08 00:00:00               94           0.580
##  9 2        2018-07-09 00:00:00               28           0.86
## 10 2        2018-07-10 00:00:00               82           0.74
## # ... with 69,793 more rows
```

```
data8 <- data7 %>%
  group_by(id_route) %>%
  summarise(depart_day_mean = round(mean(total_depart_day),0),
            occup_rate_mean = round(mean(occup_rate_mean),3)) %>%
  arrange(desc(depart_day_mean))
data8
```

```
## # A tibble: 1,495 x 3
##    id_route depart_day_mean occup_rate_mean
##    <fct>              <dbl>           <dbl>
##  1 1141                  90           0.654
##  2 2                     83           0.594
##  3 890                   76           0.554
##  4 79705                 72           0.522
##  5 47                    57           0.609
##  6 10739                 57           0.62
##  7 122464                49           0.556
##  8 6758                  48           0.572
##  9 24756                 46           0.603
## 10 10758                 44           0.587
## # ... with 1,485 more rows
```

3.5 Há alguma rota de alguma viação em que é possível sugerir mudanças nos itinerários devido à taxa de ocupação observada?

Foram selecionadas as rotas que exibiram valores menores que 59%, que representa a taxa de ocupação média geral. Nestas rotas, haveria a possiblidade de ter um estudo com possíveis mudanças nos itinerários.

```
data10 <- data6 %>%
  select(id_route, worse_occup_rate_mean) %>%
  group_by(id_route) %>%
  filter(worse_occup_rate_mean < 0.59) %>%
  arrange(worse_occup_rate_mean)
data10
```

```
## # A tibble: 100 x 2
## # Groups:   id_route [100]
##    id_route worse_occup_rate_mean
##    <fct>                    <dbl>
##  1 197739                  0.0463
##  2 86131                   0.0465
##  3 581226                  0.0466
##  4 79576                   0.0568
##  5 41563                   0.0656
##  6 1722                    0.066
##  7 1732                    0.0716
##  8 25127                   0.0721
##  9 1735                    0.0834
## 10 63341                   0.087
## # ... with 90 more rows
```

4. Proposta de negócio

Baseado na análise dos dados criou-se um conjunto de critérios de precificação dos assentos para potencializar a receita.

Como proposta de negócio para potencializar a receita, será criada uma métrica que calcula a receita gerada através da multiplicação das variáveis taxa de ocupação média (occup_rate_mean) e preço médio do assento (seat_price_mean). Este produto mede o quão eficiente está a geração de receita, o que permite identificar pontos a serem desenvolvidos ou otimizados. Os parâmetros de medição serão o dia da semana, hora do dia e dia do mês.

Criação da variável weekday para identificação do dia da semana

```
library(lubridate)

data$weekday <- as.factor(weekdays(data$departure_date))

col_order <- c("bus_company", "id_vehicle", "id_route", "departure_date", "year",
"month", "day", "weekday", "departure_time", "hour", "minute", "seat_price",
"occupied_seats", "total_seats", "occupation_rate", "revenue")

data <- data[, col_order]

str(data)
```

```
## 'data.frame':    554155 obs. of  16 variables:
##  $ bus_company    : Factor w/ 130 levels "Company1","Company10",..: 1 43 1 54 54
65 54 54 43 65 ...
##  $ id_vehicle     : Factor w/ 21702 levels "1.04245e+14",..: 3975 13536 3975
5023 10056 20313 15974 21131 8145 16675 ...
##  $ id_route       : Factor w/ 1495 levels "2","4","5","6",..: 490 573 301 1293
831 202 1293 831 580 202 ...
```

```
##  $ departure_date : POSIXct, format: "2018-07-01" "2018-07-01" ...
##  $ year           : Factor w/ 1 level "2018": 1 1 1 1 1 1 1 1 1 1 ...
##  $ month          : Factor w/ 2 levels "07","08": 1 1 1 1 1 1 1 1 1 1 ...
##  $ day            : Factor w/ 31 levels "01","02","03",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ weekday        : Factor w/ 7 levels "domingo","quarta-feira",..: 1 1 1 1 1 1
1 1 1 1 ...
##  $ departure_time : POSIXct, format: "1899-12-31 02:00:00" "1899-12-31 02:00:00"
...
##  $ hour           : Factor w/ 22 levels "02","03","04",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ minute         : Factor w/ 60 levels "00","01","02",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ seat_price     : num  16.4 50 44.6 25.7 76.8 ...
##  $ occupied_seats : int  25 25 34 12 24 26 7 10 42 4 ...
##  $ total_seats    : int  29 44 39 46 46 30 11 11 48 16 ...
##  $ occupation_rate: num  0.862 0.568 0.872 0.261 0.522 ...
##  $ revenue        : num  409 1250 1516 308 1842 ...
```

4.1 Cálculo para o dia da semana

A variável revenue_day na tabela abaixo mostra em ordem crescente, o padrão de geração de receitas por dia da semana.

```
data11 <- data %>%
  group_by(weekday) %>%
  summarise(occup_rate_mean = round(mean(occupation_rate),2),
            seat_price_mean = round(mean(seat_price),2)) %>%
  mutate(revenue_day = round(seat_price_mean * occup_rate_mean,2)) %>%
  arrange(revenue_day)
data11
```

```
## # A tibble: 7 x 4
##   weekday       occup_rate_mean seat_price_mean revenue_day
##   <fct>                   <dbl>           <dbl>       <dbl>
## 1 quarta-feira             0.53            69.7        37.0
## 2 terça-feira              0.54            70.7        38.2
## 3 quinta-feira             0.56            71.2        39.9
## 4 segunda-feira            0.61            70.4        43.0
## 5 sábado                   0.61            70.8        43.2
## 6 sexta-feira              0.65            71.6        46.5
## 7 domingo                  0.66            74.0        48.9
```

```
data11 %>%
  ggplot(aes(weekday, seat_price_mean)) +
  geom_col(fill = "darkorange") +
  theme(text = element_text(size = 11),
        axis.text.x = element_text(angle = 90, hjust = 0))
```

```
data11 %>%
  ggplot(aes(weekday, occup_rate_mean)) +
  geom_col(fill = "darkorange") +
  theme(text = element_text(size = 11),
        axis.text.x = element_text(angle = 90, hjust = 0))
```

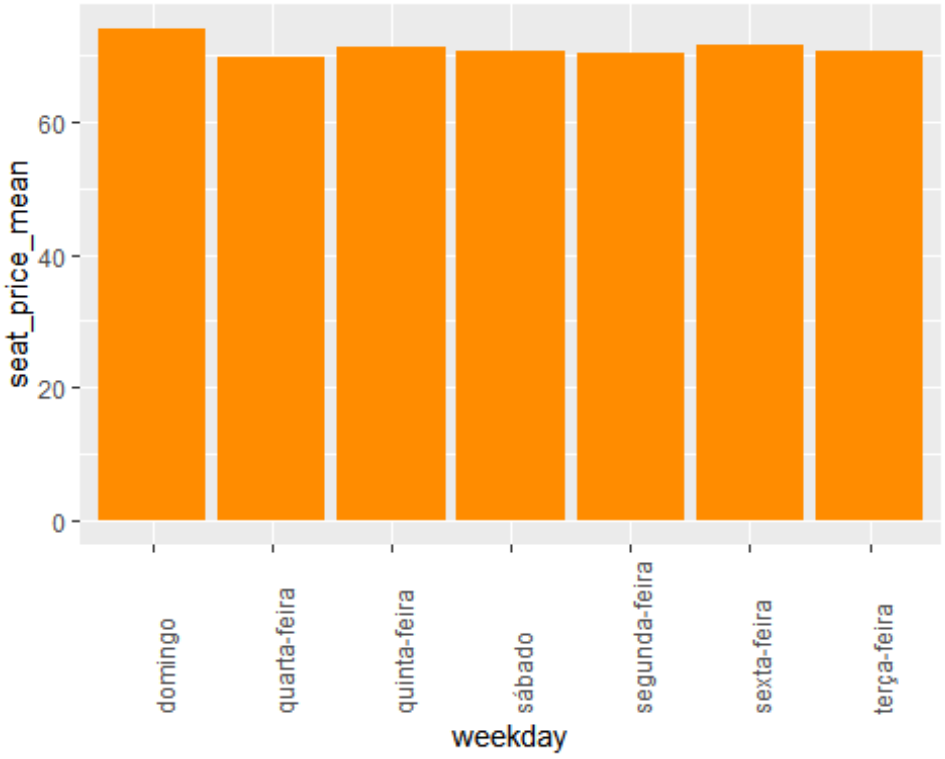

Gráfico 3: dia da semana x preço médio do assento

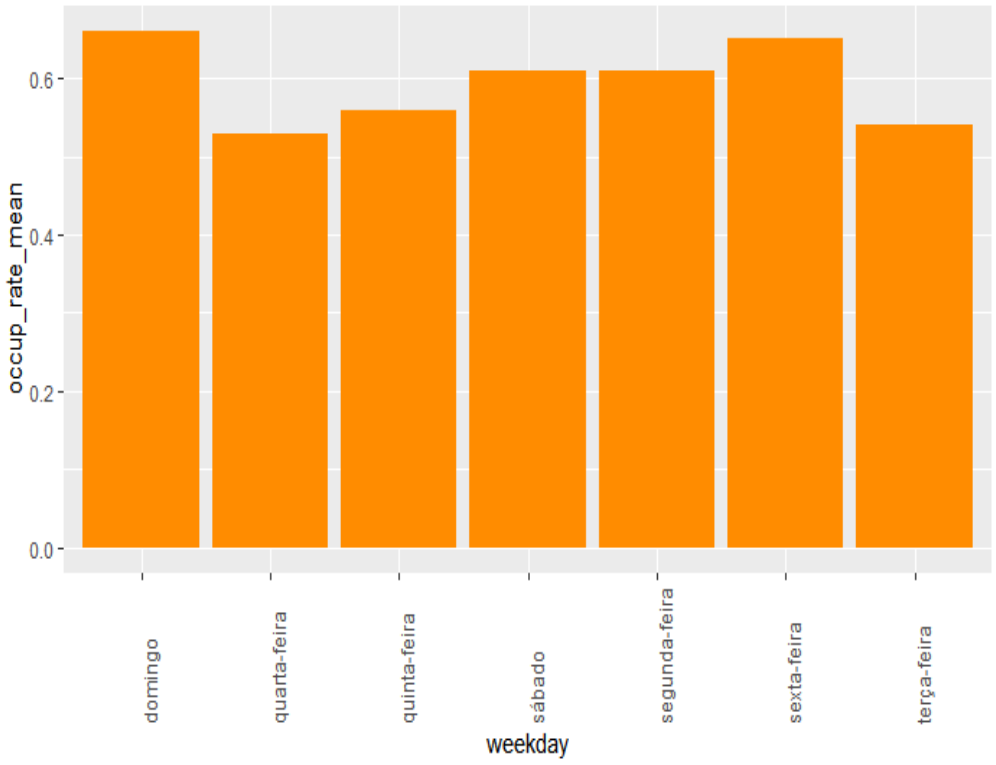

Gráfico 4: dia da semana x taxa de ocupação média

Pelos resultados notam-se claramente os dias em que se geram menos receitas, de terça a quinta-feira, tendo ainda os preços médios muito semelhantes aos outros dias da semana, aliadas às menores taxas de ocupação. O que poderia ser sugerido seria de haver uma diferenciação nos preços nestes dias, ao cobrar pela passagem um preço menor na tentativa de aumento na taxa de ocupação e nas receitas.

4.2 Cálculo pela hora do dia

A variável revenue_hour na tabela abaixo mostra em ordem crescente, o padrão de geração de receitas por hora do dia.

```
data12 <- data %>%
  group_by(hour) %>%
  summarise(occup_rate_mean = round(mean(occupation_rate),2),
            seat_price_mean = round(mean(seat_price),2)) %>%
  mutate(revenue_day = round(seat_price_mean * occup_rate_mean,2)) %>%
  arrange(revenue_day)
data12
```

```
## # A tibble: 22 x 4
##    hour  occup_rate_mean seat_price_mean revenue_day
##    <fct>           <dbl>           <dbl>       <dbl>
##  1 05              0.45             49.8        22.4
##  2 06              0.48             51.4        24.6
##  3 04              0.570            46.7        26.6
##  4 07              0.53             59.3        31.4
##  5 15              0.580            56.4        32.7
##  6 11              0.55             59.6        32.8
##  7 17              0.580            58.1        33.7
##  8 10              0.56             63.3        35.4
##  9 08              0.56             65.0        36.4
## 10 12              0.570            64.5        36.8
## # ... with 12 more rows
```

```
data12 %>%
  ggplot(aes(hour, seat_price_mean)) +
  geom_col(fill = "darkmagenta") +
  theme(text = element_text(size = 11),
        axis.text.x = element_text(angle = 90, hjust = 0))
```

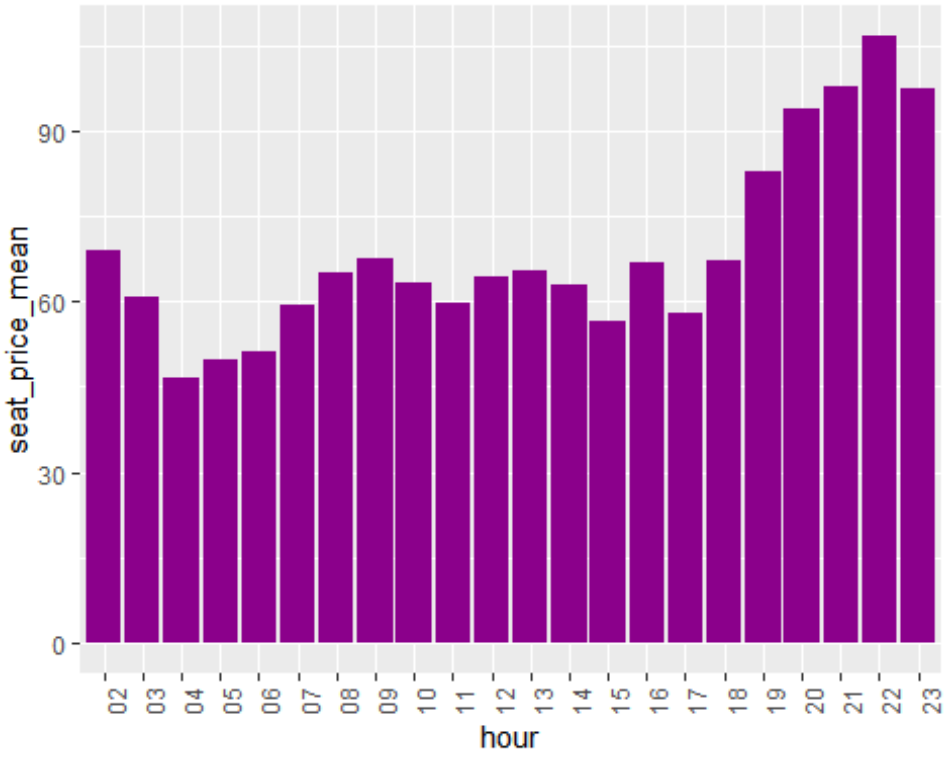

Gráfico 5: horas do dia x preço médio do assento

Observando os resultados, nas primeiras horas do dia, entre 4 a 6 da manhã, talvez não seja possível aumentar muito a demanda com uma eventual diminuição nos preços, uma vez que o valor médio cobrado nestes horários já está mais baixo comparado aos demais. Mas uma medida que poderia alavancar as receitas seria de diminuir os preços nos horários que apresentam taxa de ocupação entre 50 e 60%, vistos nos horários intermediários do dia, aproximando os valores das passagens aos horários de menor demanda.

4.3 Cálculo pelo dia do mês

A variável revenue_day na tabela abaixo mostra em ordem crescente, o padrão de geração de receitas por dia do mês.

```
data13 <- data %>%
  group_by(day) %>%
  summarise(occup_rate_mean = round(mean(occupation_rate),2),
            seat_price_mean = round(mean(seat_price),2)) %>%
  mutate(revenue_day = round(seat_price_mean * occup_rate_mean,2)) %>%
  arrange(revenue_day)
data13
```

```
## # A tibble: 31 x 4
##    day   occup_rate_mean seat_price_mean revenue_day
##    <fct>           <dbl>           <dbl>       <dbl>
##  1 02              0.53             69.8        37
##  2 04              0.54             68.6        37.0
##  3 01              0.55             70.0        38.5
##  4 03              0.56             69.0        38.7
##  5 25              0.56             69.7        39.0
##  6 28              0.55             71.2        39.2
##  7 07              0.56             70.5        39.5
##  8 27              0.580            68.7        39.9
##  9 31              0.570            70.8        40.4
## 10 18              0.580            69.8        40.5
## # ... with 21 more rows
```

```
data13 %>%
  ggplot(aes(day, seat_price_mean)) +
  geom_col(fill = "green3") +
  theme(text = element_text(size = 11),
        axis.text.x = element_text(angle = 90, hjust = 0))
```

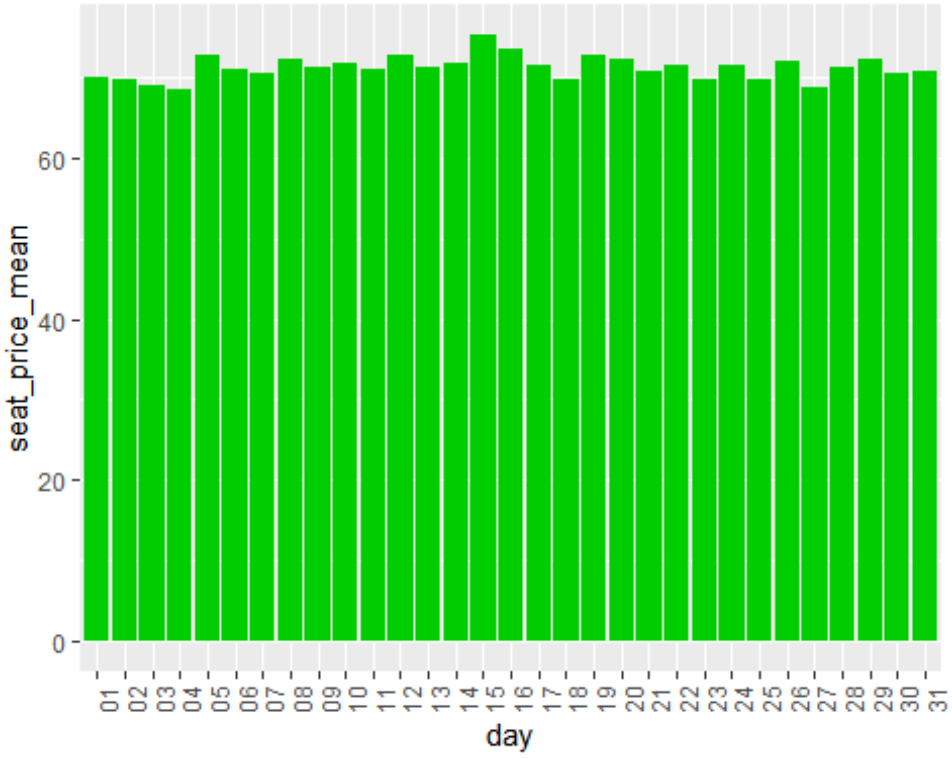

Gráfico 6: dias do mês x preço médio do assento

Conforme observado na tabela acima e no gráfico 2, o período consistente entre o final de cada mês e o início do mês subsequente exibe em média as menores receitas e, semelhantemente aos critérios adotados nos itens anteriores, a medida que também poderia ser adotada seria a de diminuição dos preços visando o incremento na receita, uma vez observado que os valores médios das passagens deste período são bastante semelhantes aos de maior demanda.

Parte 2 – Análise e construção do modelo preditivo

Nesta parte do estudo, com base em todas as análises já realizadas até aqui, questões relativas entre o **valor da passagem** e no **número de assentos ocupados** por viagem serão avaliadas com o objetivo de localizar um relacionamento entre eles. Para esta avaliação, um modelo preditivo será construído com a aplicação de regressão linear múltipla para prever o número de assentos ocupados (variável occupied_seats), em função do valor cobrado pela passagem e pelas demais variáveis envolvidas na construção do modelo preditivo.

Para maximizar os resultados para a construção do modelo, selecionaram-se as 100 melhores rotas em taxa de ocupação.

```
# Seleção das 100 melhores rotas em taxa de ocupação

data_best <- data %>%
  group_by(id_route) %>%
  summarise(best_occup_rate_mean = mean(occupation_rate)) %>%
  arrange(desc(best_occup_rate_mean))

list <- as.list(data_best[1:100,1])
data_best <- data[data$id_route %in% list[[1]],]
data_best <- droplevels(data_best)

str(data_best)
```

```
## 'data.frame':    17634 obs. of  16 variables:
##  $ bus_company    : Factor w/ 48 levels "Company1","Company100",..: 23 26 32 32
32 19 19 19 26 41 ...
##  $ id_vehicle     : Factor w/ 1329 levels "5.45487e+15",..: 174 1080 612 519 519
406 1132 406 2 1034 ...
##  $ id_route       : Factor w/ 100 levels "149","162","180",..: 4 41 84 48 84 58
33 62 41 65 ...
##  $ departure_date : POSIXct, format: "2018-07-01" "2018-07-01" ...
##  $ year           : Factor w/ 1 level "2018": 1 1 1 1 1 1 1 1 1 1 ...
##  $ month          : Factor w/ 2 levels "07","08": 1 1 1 1 1 1 1 1 1 1 ...
##  $ day            : Factor w/ 31 levels "01","02","03",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ weekday        : Factor w/ 7 levels "domingo","quarta-feira",..: 1 1 1 1 1 1
1 1 1 1 ...
##  $ departure_time : POSIXct, format: "1899-12-31 03:00:00" "1899-12-31 03:15:00"
...
##  $ hour           : Factor w/ 22 levels "02","03","04",..: 2 2 3 4 4 4 4 4 4 5
...
##  $ minute         : Factor w/ 48 levels "00","01","02",..: 1 14 1 1 1 25 25 25
37 1 ...
##  $ seat_price     : num  240 58.4 29.4 56.2 21.9 ...
##  $ occupied_seats : int  36 16 25 51 25 27 36 35 29 14 ...
```

```
##  $ total_seats    : int  37 27 26 52 26 28 37 37 36 26 ...
##  $ occupation_rate: num  0.973 0.593 0.962 0.981 0.962 ...
##  $ revenue        : num  8640 934 734 2864 548 ...
```

Construção do modelo preditivo

```
library(caret)

trainIndex <- createDataPartition(data_best$occupied_seats,
                                  p = 0.7,
                                  list = FALSE,
                                  times = 1)

train_data <- data_best[trainIndex, ]
test_data <- data_best[-trainIndex, ]

control <- trainControl(savePredictions = TRUE)

model_lm <- train(occupied_seats ~ bus_company + id_route + month + day + weekday +
hour + seat_price + total_seats,
                  data = train_data,
                  method = 'lm',
                  trControl = control,
                  preProcess = c('center', 'scale', 'nzv'))
summary(model_lm)

##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.280  -1.673   1.986   4.088  10.413
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              34.56031    0.05988 577.170  < 2e-16 ***
## bus_companyCompany20      0.18304    0.11512   1.590  0.11186
## bus_companyCompany37     -0.26004    0.07480  -3.476  0.00051 ***
## bus_companyCompany58     -0.03977    0.07221  -0.551  0.58179
## bus_companyCompany72     -0.33126    0.07135  -4.643 3.47e-06 ***
## id_route16270            -0.11078    0.08004  -1.384  0.16637
## id_route49805             0.16304    0.08132   2.005  0.04498 *
## id_route57433            -0.31774    0.07831  -4.057 5.00e-05 ***
## month08                  -1.38401    0.06046 -22.891  < 2e-16 ***
## `weekdayquarta-feira`    -0.62932    0.07704  -8.169 3.41e-16 ***
## `weekdayquinta-feira`    -0.56094    0.07801  -7.190 6.83e-13 ***
## weekdaysábado             0.06487    0.07735   0.839  0.40172
## `weekdaysegunda-feira`   -0.14449    0.07894  -1.830  0.06722 .
```

```
## `weekdaysexta-feira`    0.32355    0.07843   4.125 3.72e-05 ***
## `weekdayterça-feira`   -0.58032    0.07811  -7.430 1.16e-13 ***
## hour06                 -0.03777    0.06376  -0.592  0.55357
## hour07                 -0.11982    0.06355  -1.885  0.05940 .
## hour08                 -0.29714    0.06244  -4.759 1.97e-06 ***
## hour14                  0.18316    0.06264   2.924  0.00346 **
## hour15                 -0.31077    0.06315  -4.921 8.72e-07 ***
## hour19                  0.06548    0.06458   1.014  0.31065
## hour20                 -0.29732    0.06418  -4.633 3.64e-06 ***
## hour21                 -0.09929    0.06349  -1.564  0.11792
## hour22                 -0.07380    0.06493  -1.136  0.25578
## seat_price             -0.07896    0.08335  -0.947  0.34346
## total_seats            10.88472    0.06217 175.079  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.653 on 12319 degrees of freedom
## Multiple R-squared:  0.7337, Adjusted R-squared:  0.7331
## F-statistic:  1357 on 25 and 12319 DF,  p-value: < 2.2e-16


# Criação da variável occupied_seats_pred indicando a previsão de assentos ocupados
pelo modelo preditivo

test_data$occupied_seats_pred <- as.integer(predict(model_lm, test_data))

cols_select <- c('bus_company', 'id_route', 'month', 'day', 'weekday', 'hour',
'seat_price', 'occupied_seats','total_seats', 'occupied_seats_pred')

view <- test_data[, cols_select]

head(view[, 1:5])

##     bus_company id_route month day weekday
## 66    Company27      203    07  01 domingo
## 113   Company37   144449    07  01 domingo
## 386   Company20    47969    07  01 domingo
## 404   Company20    57433    07  01 domingo
## 409   Company20    43034    07  01 domingo
## 544   Company37    43034    07  01 domingo

head(view[, 6:10])

##     hour seat_price occupied_seats total_seats occupied_seats_pred
## 66    03     240.00             36          37                  34
## 113   04      29.37             25          26                  23
## 386   06      28.20             26          28                  26
## 404   06      18.90             26          28                  25
## 409   06      38.80             36          37                  34
## 544   06      46.80             43          43                  39
```

Interpretação dos resultados

Para a previsão do número de assentos ocupados em função do preço cobrado por assento em cada viagem, fixaram-se as condições de rota, empresa, hora, dia da semana e número total de assentos, variando apenas o preço da passagem no ensaio de simular o número de assentos ocupados resultante da elaboração do modelo preditivo.

Porém na prática a simulação não apresentou o resultado desejado de determinar a quantidade de assentos ocupados em função do preço da passagem, ou seja, aumentando e diminuindo o preço, era de se esperar respectivamente que diminuísse e aumentasse a taxa de ocupação. Este comportamento impreciso da simulação pode ser creditado ao seguinte motivo: há uma fraca relação entre a taxa de ocupação e os preços médios por passagem, claramente observada no resumo estatístico do modelo preditivo, onde a variável seat_price apresentou-se ser pouco significativa com o p-value igual a 0.34.

Mas por que houve este comportamento? Checando conjuntamente os gráficos 2 e 6 da variável dia do mês, com os gráficos 3 e 4 da variável dia da semana, nota-se que ao se comparar os dois conjuntos de gráficos, o relacionamento entre a taxa de ocupação média e o preço médio não está acoplado, ou seja, a queda ou subida na taxa de ocupação não é acompanhado pelo preço da passagem, exibindo este um padrão mais plano em seus gráficos, diferentemente do gráfico de taxa de ocupação. Este descompasso verificado na falha de relação entre estas variáveis é o provável causador na imprecisão de previsão do número de assentos ocupados.

A exceção se limitaria à variável hora do dia (gráficos 1 e 5), tanto a taxa de ocupação média quanto o preço médio apresentaram ter boa sincronia entre eles, e tal aproximação de ajuste é o que poderia ser adotado tanto para o dia do mês quanto para o dia da semana, com a lei de oferta e procura podendo estar mais presente, possibilitando que novos dados sejam alimentados e futuros modelos preditivos construídos com maior precisão na previsão dos assentos ocupados.

Conclusão

Este estudo teve por objetivo buscar respostas às questões da área de negócios através da análise dos dados, ao agrupá-los na identificação de padrões e de gargalos operacionais neles, destacando boas possibilidades de equacionar os preços das passagens num modelo preditivo, ao ajusta-los mais adequadamente à demanda para realizar a previsão do número de assentos ocupados, auxiliando às tomadas de decisões com a finalidade de tornar os processos mais claros e eficientes.