

Cumbuca Case

Matheus Pasche

29/05/2022

DNA

O Rbase contém uma função relativamente pouco utilizada, mas poderosa para traduções simples: `chartr`

```
#' nucleotides
#' Esta função traduz DNA para RNA
#' @param dna character com nucleotideos
#'
#' @return
#' @export
#'
#' @examples
#' nucleotides("CGAT")
nucleotides <- function(dna) {

  if(!is.character(dna)){stop("Please insert character nucleotides, such as ACGT")}

  old <- "GCTA"
  new <- "CGAU"

  translate <- chartr(
    old = old,
    new = new,
    x = dna
  )

  return(translate)
}

##
test_that("Translate dna", {
  expect_equal(nucleotides("GACATGG"), "CUGUACC")
  expect_equal(nucleotides("AAATTT"), "UUUAAA")
  expect_error(nucleotides(1234))
})

## Test passed
```

NycFlights

É possível identificar que o dataframe possui alguns voos sem informação sobre partida e chegada. Como representam menos de 3% dos registros, optarei por eliminar a fim de evitar problemas.

```

flights <- nycflights13::flights %>%
  mutate(total_delay = arr_delay + dep_delay, month_year = floor_date(time_hour, 'month'),
         time = as_date(time_hour),
         month_year = ym(paste0(year, month)),
         status = case_when(total_delay > 0 ~ 'delay',
                           total_delay < 0 ~ 'early',
                           total_delay == 0 ~ 'in time'))

flights %>%
  summarise(across(everything(), ~ sum(is.na(.)))) %>%
  pivot_longer(cols = everything(),
              names_to = 'Variable',
              values_to = 'NA Count') %>%
  knitr::kable(caption = 'Contagem de NAs')

```

Table 1: Contagem de NAs

Variable	NA Count
year	0
month	0
day	0
dep_time	8255
sched_dep_time	0
dep_delay	8255
arr_time	8713
sched_arr_time	0
arr_delay	9430
carrier	0
flight	0
tailnum	2512
origin	0
dest	0
air_time	9430
distance	0
hour	0
minute	0
time_hour	0
total_delay	9430
month_year	0
time	0
status	9430

```
flights <- flights %>% filter(!is.na(status))
```

Questão 1:

A quantidade total de voos, que inclui atrasados e não atrasados, não variou consideravelmente ao longo dos meses.

```

flights %>%
  janitor::tabyl(month, status) %>%
  janitor::adorn_totals(where = 'col') %>%
  kable(caption = "Quantidade de voos por mês e status",
        format.args = list(big.mark = ".", decimal.mark = ","))

```

Table 2: Quantidade de voos por mês e status

month	delay	early	in time	Total
1	10.977	15.043	378	26.398
2	10.010	13.310	291	23.611
3	11.311	16.274	317	27.902
4	12.240	14.987	337	27.564
5	10.698	17.137	293	28.128
6	12.900	13.886	289	27.075
7	13.832	14.161	300	28.293
8	12.081	16.330	345	28.756
9	7.163	19.585	262	27.010
10	9.638	18.599	381	28.618
11	9.426	17.185	360	26.971
12	14.783	11.904	333	27.020

Tendo em vista que estamos interessados em analisar o atraso, foram retirados os voos com saldo total (embarque + desembarque) negativo. Abaixo podemos identificar os primeiros registros do atraso médio, mediano e estatísticas percentuais para cada dia do ano. Evidentemente não é prático exibir todos os 365 dias.

```

stats <- function(.data, variable) {
  .data %>%
    summarise(
      count = n(),
      mean = mean({{variable}}),
      median = median({{variable}}),
      sd = sd({{variable}}),
      var = var({{variable}}),
      'p.01'=quantile({{variable}}, probs = .1),
      'p.025'=quantile({{variable}}, probs = .25),
      'p.09'=quantile({{variable}}, probs = .9),
      'p.99'=quantile({{variable}}, probs = .99)
    )
}

flights %>%
  filter(status == 'delay') %>%
  with_groups(time, ~stats(., variable = total_delay))%>%
  head() %>%
  knitr::kable(caption = 'Registros de estatísticas descriticas por dia', digits=1)

```

Table 3: Registros de estatísticas descritivas por dia

time	count	mean	median	sd	var	p.01	p.025	p.09	p.99
2013-01-01	447	57.1	23.0	117.3	13769.2	4	9.0	139.8	514.7
2013-01-02	540	56.8	24.5	88.6	7841.8	4	9.0	157.2	399.3
2013-01-03	455	51.4	24.0	74.5	5542.9	4	10.0	129.0	349.5
2013-01-04	318	56.0	30.5	71.7	5143.6	5	10.2	130.3	334.4
2013-01-05	255	41.8	17.0	72.2	5214.7	4	7.0	120.0	340.3
2013-01-06	371	44.1	23.0	56.3	3171.6	4	9.0	111.0	284.0

Para fins de análise do desempenho e tendo em vista que em um único dia há registro de vários voos, agregaremos os dados em meses. Inicialmente podemos verificar que os períodos de maior atraso mediano são junho-julho e dezembro.

```
flights %>%
  filter(!is.na(air_time), status == 'delay') %>%
  with_groups(month, ~stats(., variable = total_delay)) %>%
  knitr::kable(caption = 'Registros de estatísticas descritivas por mês', digits=1)
```

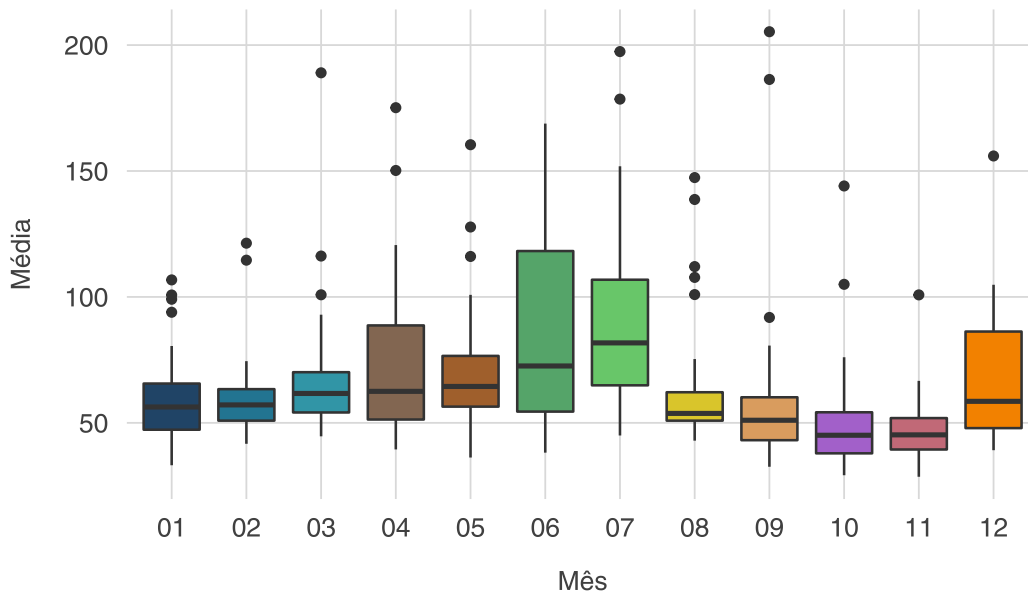
Table 4: Registros de estatísticas descritivas por mês

month	count	mean	median	sd	var	p.01	p.025	p.09	p.99
1	10977	63.8	27	97.3	9476.6	4.0	10	172.4	435.7
2	10010	63.8	28	93.8	8795.0	4.0	10	171.0	427.0
3	11311	76.2	35	105.2	11075.3	4.0	12	200.0	502.8
4	12240	79.5	37	110.2	12139.3	5.0	13	212.0	499.0
5	10698	78.5	40	104.7	10972.1	5.0	14	201.0	480.0
6	12900	98.6	51	126.6	16029.3	5.0	16	260.1	559.0
7	13832	98.4	50	125.9	15847.7	6.0	16	255.0	580.7
8	12081	71.8	33	97.1	9425.7	4.0	12	193.0	457.2
9	7163	72.6	29	115.1	13258.5	3.2	10	196.8	559.1
10	9638	55.5	24	83.1	6901.9	3.0	8	148.0	396.6
11	9426	50.7	23	77.6	6015.4	3.0	9	130.0	365.2
12	14783	70.9	35	100.0	9993.3	5.0	13	182.0	470.4

Tendo em vista que a quantidade total de voos no mês tende a se manter constante, o que poderia explicar que a quantidade de minutos de atraso médio em julho pode ser aproximadamente o dobro do observado em novembro? A explicação mais imediata com base em senso comum é a sazonalidade dos feriados e período de férias. Os meses de junho a agosto marcam as férias de verão dos colégios e universidades, enquanto o mês de dezembro é o período de reuniões familiares para as festas de fim de ano.

```
plot <- flights %>%
  filter(status == 'delay') %>%
  with_groups(time, ~stats(., variable = total_delay)) %>%
  mutate(month = format(time, '%m')) %>%
  ggplot(aes(x = month, y = mean, fill = month))+
  geom_boxplot(show.legend = F)+
  pilot::scale_fill_pilot()+
  pilot::theme_pilot()+
  labs(y = 'Média', x = 'Mês')
pilot::add_pilot_titles(plot, title = 'Média dos atrasos por mês')
```

Média dos atrasos por mês



A conexão exata entre os fatos de termos feriados e períodos festivos e a quantidade total de voos médios por mês não se elevar tanto nesses períodos pode se dar por algumas hipóteses, nem todas testáveis com os dados disponíveis: bagagens maiores nos períodos de feriados prolongados podem atrasar o embarque dos voos; ii) mudança da composição dos destinos: mais viagens para a Califórnia e Florida no verão; iii) questões de clima: a neve, os fortes ventos e a visibilidade podem atrapalhar a decolagem, sobretudo no inverno; iv) uma interseção em maior ou menor grau de todos os aspectos anteriores.

Como um dos critérios de avaliação é a análise direta ao ponto, destacarei um fato que corrobora a hipótese ii: se coletarmos os destinos ao longo dos meses, veremos que de fato há sazonalidade. A tabela abaixo mostra os 5 principais destinos dos voos partindo de Nova Iorque por mês. Há regularidade na relação com Atlanta e Chicago, provavelmente por relações financeiras, mas observe que os aeroportos de Los Angeles e São Francisco aparecem mais frequentemente no top 5 em períodos de férias.

```
flights %>%
  with_groups(c(dest, month), ~summarise(., count = n())) %>%
  with_groups(c(month), ~slice_max(., count, n = 5) %>%
    mutate(Top5 = row_number())) %>%
  left_join(nycflights13::airports, by = c("dest" = "faa")) %>%
  select(name, Top5, month) %>%
  pivot_wider(names_from = month, values_from = Top5) %>%
  kable(caption = 'Principais aeroportos de destino por mês')
```

Table 5: Principais aeroportos de destino por mês

name	1	2	3	4	5	6	7	8	9	10	11	12
Hartsfield Jackson Atlanta Intl	1	1	1	1	2	3	3	3	3	2	1	1
Chicago Ohare Intl	2	2	2	3	1	1	2	1	1	1	2	
General Edward Lawrence Logan Intl	3	3	3	4	4	4	4	4	4	4	4	

name	1	2	3	4	5	6	7	8	9	10	11	12
Orlando Intl	4	4	4	5			5	5				3
Fort Lauderdale Hollywood Intl	5	5	5									
Los Angeles Intl				2	3	2	1	2	2	3	3	2
San Francisco Intl					5	5						4
Charlotte Douglas Intl									5	5	5	5

Questão 2: Empresas aéreas e atrasos

Inicialmente, observamos que três empresas concentram mais de 10% cada dos voos atrasados: ExpressJet Airlines, UnitedAir Lines e JetBlue Airways.

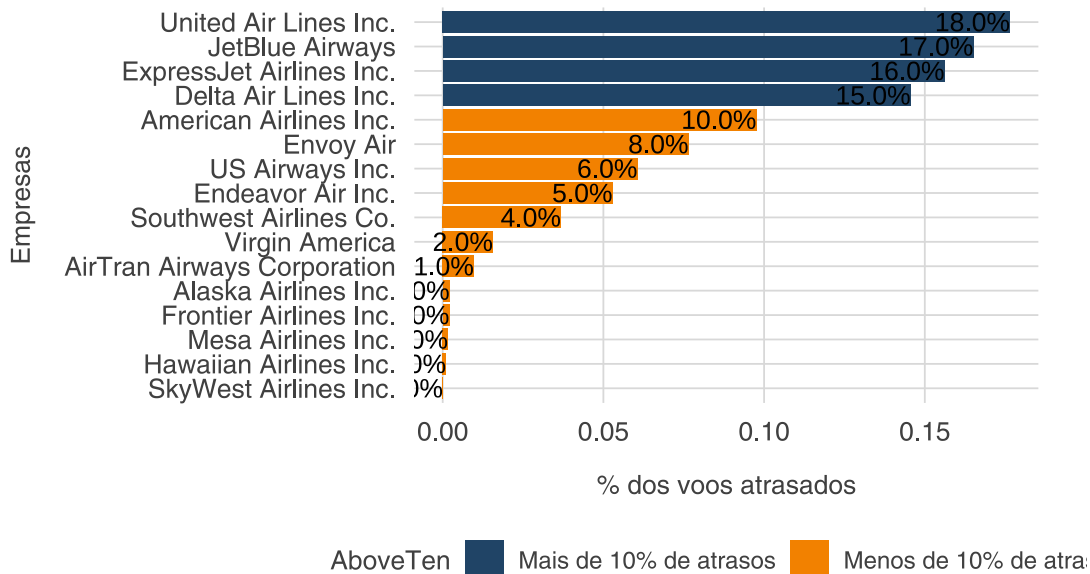
```
Companies = flights %>%
  filter(!is.na(air_time)) %>%
  with_groups(c(carrier ,status), ~summarise(.,count = n())) %>%
  left_join(nycflights13::airlines) %>%
  with_groups(name, ~mutate(., propDelay = count/sum(count), totalFlights = sum(count))) %>%
  select(name, status, count, propDelay, totalFlights)

plot = Companies %>%
  filter(status == 'delay') %>%
  mutate(Razao = totalFlights/sum(totalFlights),
         AboveTen = case_when( Razao >= .10 ~ 'Mais de 10% de atrasos',
                                TRUE~ 'Menos de 10% de atrasos')
  ) %>%

ggplot(aes(
  x = Razao,
  y = reorder(name, Razao),
  fill = AboveTen
))+
geom_bar(stat = 'identity')+
geom_text(aes(label = scales::percent(Razao %>%round(2))), hjust = 1)+
pilot::scale_fill_pilot()+
pilot::theme_pilot()+
labs(
  x= '% dos voos atrasados',
  y= 'Empresas')+
theme(legend.position="bottom")

pilot::add_pilot_titles(plot, title = 'Porcentagem dos voos atrasados')
```

Porcentagem dos voos atrasados



No entanto, não surpreendentemente as empresas também detêm os quatro maiores números totais de voos. Em uma operação em larga escala em um produto há décadas realizado de forma constante, regulada e previsível, é de certa forma esperado que a proporção de voos com atraso siga um nível de estabilidade. No entanto, a pergunta original é: elas concentram a maior parte do atraso?

```
OrdenacoesCompanies = Companies %>%
  filter(status == 'delay') %>%
  select(-status) %>%
  rename(`Count Delay` = count)

Porte <- OrdenacoesCompanies%>%
  mutate(
    Porte = case_when (totalFlights > 40000 ~ 'GrandePorte',
                      between(totalFlights, 12000, 40000) ~ 'MedioPorte',
                      TRUE ~ 'PequenoPorte')
  )

Porte %>%
  arrange(-totalFlights) %>%
  kable(digits = 3, format.args = list(big.mark = ".", decimal.mark = ","))
```

name	Count Delay	propDelay	totalFlights	Porte
United Air Lines Inc.	24.514	0,424	57.782	GrandePorte
JetBlue Airways	23.839	0,441	54.049	GrandePorte
ExpressJet Airlines Inc.	24.566	0,481	51.108	GrandePorte

name	Count Delay	propDelay	totalFlights	Porte
Delta Air Lines Inc.	16.616	0,349	47.658	GrandePorte
American Airlines Inc.	10.668	0,334	31.947	MedioPorte
Envoy Air	10.818	0,432	25.037	MedioPorte
US Airways Inc.	6.597	0,333	19.831	MedioPorte
Endeavor Air Inc.	6.892	0,399	17.294	MedioPorte
Southwest Airlines Co.	5.852	0,486	12.044	MedioPorte
Virgin America	1.888	0,369	5.116	PequenoPorte
AirTran Airways Corporation	1.854	0,584	3.175	PequenoPorte
Alaska Airlines Inc.	191	0,269	709	PequenoPorte
Frontier Airlines Inc.	412	0,605	681	PequenoPorte
Mesa Airlines Inc.	257	0,472	544	PequenoPorte
Hawaiian Airlines Inc.	87	0,254	342	PequenoPorte
SkyWest Airlines Inc.	8	0,276	29	PequenoPorte

Na tabela acima observamos que existe o quarteto já mencionado de grandes empresas com grandes quantidades de voos, seguido por um quinteto de empresas com quantidades de voos médias (American Airlines, Envoy Air, US Airways Inc., Endeavor Air Inc., SouthWest Airlines Co.) e as demais com poucos voos. Esta será a divisão por porte da empresa.

Podemos também dividir pela proporção de atrasos: as muito eficientes, com menos de 35% dos voos, as medianas que estão entre 35,1% até 45% de voos com atraso e as demais, que tem valores muito próximos ou maiores que a metade das operações em atraso.

```

Eficiencia <- OrdenacoesCompanies%>%
  mutate(
    Eficiencia = case_when (
      propDelay <.35 ~ 'MuitoEficientes',
      between(propDelay, .35, .45) ~ 'Eficientes',
      TRUE ~ 'PoucoEficientes'
    )
  )

Eficiencia %>%
  arrange(propDelay) %>%
  kable(digits = 3,format.args = list(big.mark = ".",decimal.mark = ","))

```

name	Count Delay	propDelay	totalFlights	Eficiencia
Hawaiian Airlines Inc.	87	0,254	342	MuitoEficientes
Alaska Airlines Inc.	191	0,269	709	MuitoEficientes
SkyWest Airlines Inc.	8	0,276	29	MuitoEficientes
US Airways Inc.	6.597	0,333	19.831	MuitoEficientes
American Airlines Inc.	10.668	0,334	31.947	MuitoEficientes
Delta Air Lines Inc.	16.616	0,349	47.658	MuitoEficientes
Virgin America	1.888	0,369	5.116	Eficientes
Endeavor Air Inc.	6.892	0,399	17.294	Eficientes
United Air Lines Inc.	24.514	0,424	57.782	Eficientes
Envoy Air	10.818	0,432	25.037	Eficientes
JetBlue Airways	23.839	0,441	54.049	Eficientes
Mesa Airlines Inc.	257	0,472	544	PoucoEficientes
ExpressJet Airlines Inc.	24.566	0,481	51.108	PoucoEficientes

name	Count Delay	propDelay	totalFlights	Eficiencia
Southwest Airlines Co.	5.852	0,486	12.044	PoucoEficientes
AirTran Airways Corporation	1.854	0,584	3.175	PoucoEficientes
Frontier Airlines Inc.	412	0,605	681	PoucoEficientes

Em um trabalho mais aprofundado seria adequado fazer uso de algoritmos de aproximação de vizinhanças a fim de definir não tão arbitrariamente os pontos de corte. No entanto, uma vez definido o corte, o problema vira uma questão de probabilidade condicional:

$$P(GrandePorte|Atraso) = \frac{P(GrandePorte \cap Atraso)}{P(Atraso)} \quad (1)$$

e equivalentemente:

$$P(PoucoEficiente|Atraso) = \frac{P(PoucoEficiente \cap Atraso)}{P(Atraso)} \quad (2)$$

A lógica é bastante simples: dado que um voo atrasou, é mais provável que ele pertença ao grupo das empresas com maior quantidade de voos ou das menos eficientes? Se um voo atrasou, há uma probabilidade de 66,3% desse voo pertencer a uma das quatro maiores, enquanto apenas 24,4% de probabilidade de pertencer às empresas com menor grau de eficiência. Portanto, podemos dizer que as empresas ExpressJet Airlines, UnitedAir Lines e JetBlue Airways concentram os atrasos.

Questão 3

Em primeiro lugar é preciso ressaltar que cerca de 14,7% dos voos não possuem registros de fabricante. Os dados evidentemente serão limpos, mas não representarão a totalidade dos voos.

Para o primeiro cenário em que estamos interessados em observar toda a base e a discriminação do fabricante, podemos recorrer à tabela da questão anterior.

```
flightsManufacturer <- flights %>%
  left_join(nycflights13::planes %>%
    select(manufacturer, tailnum)
  ) %>%
  filter(!is.na(manufacturer))
```

```
flightsManufacturer %>%
  with_groups(manufacturer, ~stats(., variable = distance))
```

```
## # A tibble: 35 x 10
##   manufacturer    count  mean median    sd   var  p.01 p.025  p.09  p.99
##   <chr>          <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 AGUSTA SPA      31 1765.  1089  757.  5.74e5 1089 1089  2586 2586
## 2 AIRBUS        46890 1431.  1069  787.  6.19e5  541  944  2475 2586
## 3 AIRBUS INDUSTRIE 40642  980.   950  620.  3.84e5  214  529  2133 2565
## 4 AMERICAN AIRCRAFT INC   38 1000.   733  312.  9.74e4  733  733  1389 1389
## 5 AVIAT AIRCRAFT INC    17 2049.  2586  732.  5.36e5 1089 1089  2586 2586
## 6 AVIONS MARCEL DASSAU~    4  544    544    0    0      544  544   544  544
## 7 BARKER JACK L      250 1345.  1069  600.  3.60e5  937  962.  2454 2569
## 8 BEECH           46 1622.  1089  682.  4.66e5 1085 1089  2586 2586
```

```
## 9 BELL 63 1461. 1372 728. 5.29e5 733 733 2586 2586
## 10 BOEING 82283 1567. 1416 794. 6.30e5 719 762 2565 2586
## # ... with 25 more rows
```

O primeiro dado a ser coletado é a grande dispersão entre os dados. Boeing e Airbus, as duas mais importantes produtoras de aviões, têm seus equipamentos em uso em diversos tipos de rotas, o que justifica a alta variância dos dados. O caso é uma descrição clara de um problema que pode ser resolvido com o uso da estatística de Fisher, com o uso da ANOVA.

```
flightsManufacturer %>%
  aov(formula = distance ~ manufacturer) %>%
  broom::tidy(summary()) %>%
  kable()
```

term	df	sumsq	meansq	statistic	p.value
manufacturer	34	55596839781	1635201170.0	4259.643	0
Residuals	278982	107096233639	383882.2		

O teste apresentou p-valor $< 5\%$, o que nos permite rejeitar a hipótese nula, indicando relação entre o fabricante e a distância.

Prosseguindo para a incorporação da empresa aérea na análise, observamos que o resultado se mantém.

```
flightsManufacturer %>%
  aov(formula = distance ~ manufacturer + carrier) %>%
  broom::tidy(summary()) %>%
  kable()
```

term	df	sumsq	meansq	statistic	p.value
manufacturer	34	55596839781	1635201170.0	5598.789	0
carrier	15	25620193870	1708012924.7	5848.090	0
Residuals	278967	81476039768	292063.4		

Apesar da significância estatística dos dados, o conceito de p-valor por si só é bastante problemático e utilizá-lo como ferramenta de negação de hipóteses sem maior exploração dos dados e compreensão de outras variáveis que possam influenciar na relação pode levar a uma conclusão errônea. Em especial, temos um dataset extremamente desbalanceado, com duas empresas responsáveis pela maioria das aeronaves, o que certamente influencia os resultados.

Pensamento estatístico

A inadimplência em queda linear à medida que o limite aumenta para um determinado público é algo esperado se a concessão estiver ajustada. Em boa parte das empresas que trabalham com cartão de crédito, o limite elevado é uma função da estimativa de renda, do comportamento dos gastos e do desempenho/comportamento (behavior) da relação com o cliente. Portanto, um cliente que possui limites elevados certamente possui estimativa de renda elevada nos bureaus de crédito e desempenho compatível no mercado. A relação entre alto limite e baixa inadimplência só é mantida se fatores externos, como taxa de juros e PIB estão constantes.

Um cenário de choque pode ser bastante didático. Imagine que ao invés de uma certa empresa de crédito por erro de sua análise de crédito acaba por conceder limites muito maiores que os compatíveis com a renda dos clientes para uma parte da carteira. Seguirá que o comportamento dos agentes no mercado possivelmente mudará. Se antes o cliente precisaria economizar e liberar limite para a viagem dos sonhos, com a mudança de política o comportamento no mercado será de comprar parceladamente, imaginando que terá como pagar no futuro. Ocorre que o país subitamente enfrenta uma crise econômica em que uma parte da população perde o emprego. Os clientes contemplados com limites elevados podem ter desempenhos iguais ou até piores que outros mais pobres.

Em cenários de crédito, a regra básica é realizar experimentos sempre que possível, e sempre de forma cautelosa. Se o stakeholder deseja realizar um aumento de limites, pode gerar um número aleatório para cada cliente de sua base. Se, por exemplo, o cliente retirar o número 1, receberá um aumento e caso contrário, não. Como o cartão de crédito não possui tempo de maturação e o comportamento dos agentes pode ser analisado em sequência, em pouquíssimo tempo haverá massa estatisticamente suficiente para realizar comparações por meio de testes estatísticos e avaliar se o trade-off entre inadimplência e receita compensa em comparação com o cenário habitual. Em caso afirmativo, o dígito aleatório pode ser expandido para a 1 a 5, por exemplo, até que se torne a política campeã.