Case técnico - Data Person - Cumbuca

Lucas Ketzer

Mai/2022

Transcrição DNA -> RNA

Considerando que para cada letra no DNA existe uma transcrição correspondente para RNA, o problema pode ser facilmente solucionado ao considerar a string como uma lista de caracteres. Tendo a lista, basta substituir cada elemento na lista por seu correspondente mapeado.

Em Python:

```
import unittest

def transcribe_dna_to_rna(dna: str) -> str:
    """
    Transcribes a DNA sequence into a RNA sequence.

    Takes a string that represents a DNA sequence, substituting
    each letter for it's corresponding match in a RNA sequence,
    returning the transcribed sequence as a string.

    Parameters
    ----------
    seq : str
        The DNA sequence to be transcribed.

    Returns
    -------
    str
        The transcribed RNA sequence.
    """

    nucleotide_switcher = {
        "G":"C",
        "C":"G",
        "T":"A",
        "A":"U"
    }

    # Get the corresponding RNA letter for each letter in DNA
    rna = [nucleotide_switcher.get(l) for l in dna]

    return "".join(rna)

class TestTranscription(unittest.TestCase):
    """
```

```
    Class used to test if DNA sequences
    are being correctly transcribed into
    RNA sequences.

    Methods
    -------
    test_one
        Tests if the first mentioned sequence in the
        challenge is being correctly transcribed.
    test_two
        Tests it the second mentioned sequence in the
        challenge ins being correctly transcribed.
    """

    def test_one(self):
        self.assertEqual(
            transcribe_dna_to_rna("GGCTA"),
            "CCGAU",
            "Sequence incorrectly transcribed, should be CCGAU"
        )

    def test_two(self):
        self.assertEqual(
            transcribe_dna_to_rna("ACTGATA"),
            "UGACUAU",
            "Sequence incorrectly transcribed, should be UGACUAU"
        )
if __name__ == '__main__':
  unittest.main(exit = False)
```

```
## <unittest.main.TestProgram object at 0x7f8ad8a7df70>
##
## ..
## ----------------------------------------------------------------------
## Ran 2 tests in 0.000s
##
## OK
```

Em R:

```
library(magrittr)
library(testthat)
```

```
##
## Attaching package: 'testthat'
```

```
## The following objects are masked from 'package:magrittr':
##
##     equals, is_less_than, not
```

```
#' transcribe_dna_to_rna(dna)
#'
#' @description
#' Takes a string that representes a DNA sequence,
#' substituting each letter for it's corresponding match
#' in a RNA sequence, returning the transcribed sequence as
```

```
#' a character.
#'
#' @param dna A DNA sequence, as a character, to be transcribed
#' @return The transcribed RNA sequence, as a character
transcribe_dna_to_rna <- function(dna) {
    nucleotide_switcher <- c(
        "G" = "C",
        "C" = "G",
        "T" = "A",
        "A" = "U"
    )

    rna <- dna %>%
            stringr::str_split(., "") %>%
            unlist(.) %>%
            dplyr::recode(., !!!nucleotide_switcher) %>%
            paste(., collapse = "")

    return(rna)
}

test_that("Check if sequences in the challenge are being correctly transcribed", {
    expect_equal(transcribe_dna_to_rna("GGCTA"), "CCGAU")
    expect_equal(transcribe_dna_to_rna("ACTGATA"), "UGACUAU")
})
```

```
## Test passed
```

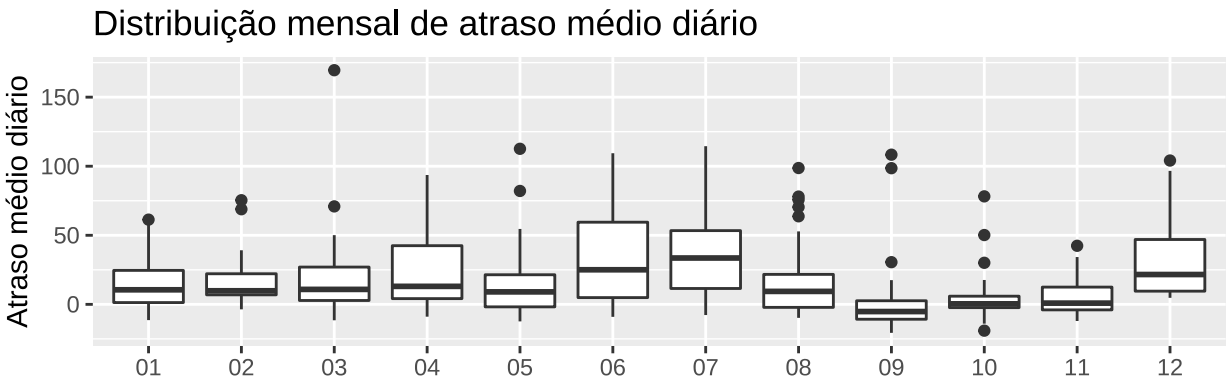
Processando e explorando dados em um banco relacional - nycflights13

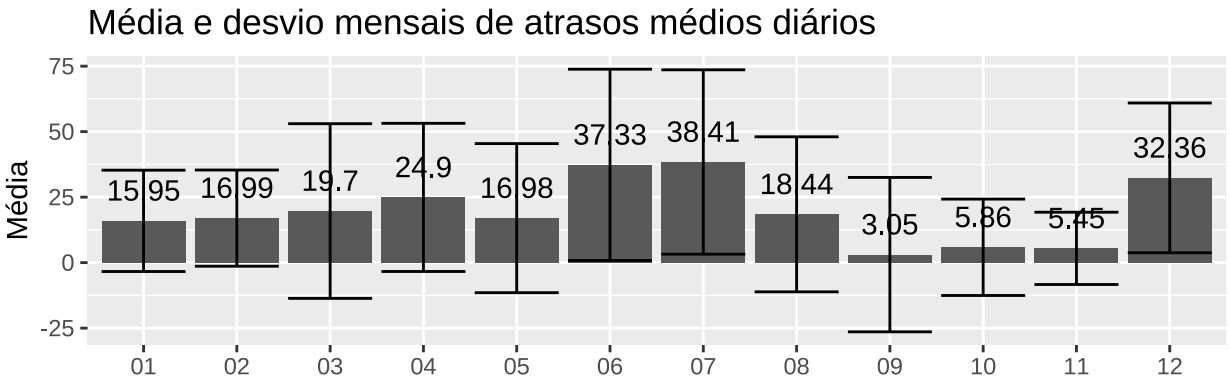
As seções seguintes serão focadas em efetivamente responder as perguntas levantadas em relação a base. O código completo, contendo alguns detalhes a mais do tratamento da base, estará na versão em markdown deste documento.

Tendências ao longo do ano

Como existem vários vôos que ocorrem em um mesmo dia, é necessário trabalhar com estatísticas descritivas para cada dia do ano - como por exemplo, o atraso médio diário (ou seja, a média de todos os atrasos que ocorrem em um dia) - todas as métricas aqui descritas são medidas descritivas de um único dia, agrupadas, principalmente, num recorte temporal mensal.

Verifica-se que, em média, os atrasos médios diários tendem a ser menores em setembro, outubro e novembro, enquanto tendem a ser maiores em junho, julho e dezembro:





É possível apontar algumas possibilidades para essas tendências: setembro, outubro e novembro são meses do outuno, além de apresentarem uma redução significativa após agosto (fim do verão, e, possivelmente, das férias de algumas pessoas.)

Um aumento na média mensal do atraso médio diário em junho e julho poderia ser explicado pelo início do período do verão e férias, enquanto em dezembro temos semanas com atrasos mais intensos que poderiam ser possivelmente explicados por festas de fim de ano.

De janeiro a maio, verificam-se distribuições relativamente similares, com uma variação máxima de 5 minutos a mais na média de março a abril. De todo modo, existe uma variação bastante grande na no atraso médio diário para todos os meses - repare na tabela que traz medidas resumo dos nonagésimo e nonagésimo nono percentil de todos os atrasos que ocorrem em um dia, mês a mês:

```
## # A tibble: 12 x 9
## # Groups:   month [12]
##    month mean_var sd_var variance_var   p01   p25   p50   p90   p99
##    <chr>    <dbl>  <dbl>        <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 01        79.9   56.6        3201.  28    34.8  65.3  169.  169.
##  2 10        55.3   56.3        3173.  23.4  25.3  37.1   88    88
##  3 11        50.0   39.8        1586.  16    23.4  35.5  113.  113.
##  4 12       115.    74.5        5548.  49    59.3  84.8  209   209
##  5 02        82.4   51.7        2668.  44.7  51.7  63.2  115.  115.
##  6 03        94.2   79.9        6384.  24.1  50    72.9  167   167
##  7 04       109.    83.2        6922.  28.7  48.7  73.7  227.  227.
##  8 05        96.1   81.9        6705.  28.6  43.3  70.1  191.  191.
##  9 06       146.   104.        10721.  38.8  54.6 116.   292.  292.
## 10 07       152.   102.        10372.  44.7  73.2 140.   316   316
## 11 08        87.3   76.5        5845.  30    39.2  61.4  222   222
## 12 09        57.5   92.7        8591.  10.1  14.2  27.9  106.  106.
```

```
## # A tibble: 12 x 9
## # Groups:   month [12]
##    month mean_var sd_var variance_var   p01   p25   p50   p90   p99
##    <chr>    <dbl>  <dbl>        <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 01        287.   99.6        9926.  182.  206.  268.  413.  518.
##  2 10        220.   74.7        5586.  147.  171.  194.  277.  449.
##  3 11        218.   78.8        6202.  158.  168.  206.  288.  469.
##  4 12        328.  120.        14408.  210.  239.  287   492.  658.
##  5 02        306.   91.5        8370.  207.  248.  308.  375.  592.
##  6 03        309.  112.        12568.  216.  238.  286.  371.  681.
##  7 04        338.  137.        18832.  201.  254.  291.  510.  683.
##  8 05        303.  115.        13226.  202.  239.  272.  480.  632.
##  9 06        384.  159.        25302.  224.  260.  349.  614.  714.
## 10 07        391.  143.        20507.  235.  281.  374.  625.  745.
## 11 08        289.  107.        11462.  199.  216.  252.  431.  577.
## 12 09        250.  144.        20641.  142.  167.  208.  378.  715.
```

Os percentis 90 e 99 representam o 90 e 99 maiores atrasos: ou seja, no caso do percentil, 90% dos atrasos são menores ou iguais a este valor. Repare, por exemplo, que no mês de setembro, com menor média de atrasos médios diários, pode-se afirmar que, em 75% dos dias do mês, 90% dos atrasos que ocorreram foram de até 54 minutos ou mais, e em 75% destes mesmos dias, 99% dos atrasos que ocorreram foram de até 260 minutos ou mais. Tem-se aqui um forte indício de alta variabilidade nos dados, que pode ser facilmente apontado ao se verificar a distribuição dos percentis 25% e 50% dos atrasos que ocorrem em um dia:

```
##     0%    10%    20%    30%    40%    50%    60%    70%    80%    90%   100%
## -37.75 -29.00 -25.00 -23.00 -21.00 -20.00 -18.00 -15.00 -13.00  -8.40  35.00
```

```
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##  -27  -17  -14  -12   -9   -7   -4    0    5   17  118
```

De 75% dos atrasos computados em um dia, ao menos 90% são menores ou iguais a -8 minutos: o mesmo se verifica para 50% dos atrasos computados em um dia, com 60% deles sendo menores ou iguais a -4 minutos. Se existe uma boa parcela de vôos adiantados, e, mesmo assim, verificam-se atrasos médios muito menores em alguns meses e muito altos em outros, uma boa explicação seria a existência de casos extremos em algumas semanas do ano:

```
## # A tibble: 53 x 10
## # Groups:   week [53]
##    week  month mean_var sd_var variance_var   p01   p25   p50   p90   p99
##    <chr> <chr>    <dbl>  <dbl>        <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 25    06         526.  157.        24798.  352.  385.  562.  695.  728.
##  2 27    07         507.  147.        21652.  381.  413.  459.  699.  762.
##  3 15    04         452.  139.        19281.  299.  352.  491.  580.  669.
##  4 20    05         394.  135.        18144.  297.  322.  346.  553.  651.
##  5 29    07         393.  147.        21618.  270.  295.  366.  555.  674.
##  6 30    07         387.  125.        15548.  261.  303.  374.  509.  614.
##  7 26    06         372.  151.        22767.  220.  228.  396.  552.  570.
##  8 03    01         364.   95.6        9144.  285.  308.  334   475.  540.
##  9 23    06         361.  157.        24576.  230.  247.  297.  547.  642.
## 10 11    03         355.  130.        16814.  254.  284.  328.  473.  615.
## # ... with 43 more rows
```

Junho possui uma semana onde, em média, 90% dos atrasos computados foram de 526 minutos ou menos, enquanto julho possui uma semana onde 90% dos atrasos computados foram de 507 minutos ou menos. Estes mesmos meses possuem um atraso médio diário elevado em comparação a outros meses - a existência de casos extremos muito possivelmente joga o atraso médio mensal para cima quando comparado com outros meses.

Pontuam-se algumas limitações: considerar outros fatores, como o clima do mês e semanas que possuem feriados poderiam explicar melhor tendências anuais, principalmente dentro do contexto de um modelo de regressão. Fatores externos aos à sazonalidade poderiam, também, explicar atrasos em maior riqueza de detalhes: marcar aeroportos que são internacionais ou não, considerar a potência do motor do aviãso e a fabricante poderiam ajudar a auxiliar fatores sazonais - para fins de explicar possíveis tendências ao longo do ano, no entanto, essas análises seriam muito exageradas.

Concentração de atrasos por empresas aéreas

Uma forma simples de verificar uma empresa com maior concentração de atrasos seria analisar o % de atrasos desta empresa em relação ao total - entretanto, se estivermos interessados em afirmar se há uma maneira de discriminar o percentual de atrasos entre uma empresa e outra, é necessário aprofundar a nossa análise.

Veja, por exemplo, a tabela abaixo, com o % de vôos atrasados e não atrasados entre vôos para uma mesma empresa:

```
## # A tibble: 16 x 3
##    name                           N     Y
##    <fct>                      <dbl> <dbl>
##  1 Frontier Airlines Inc.     0.395 0.605
##  2 AirTran Airways Corporation 0.416 0.584
##  3 Southwest Airlines Co.     0.514 0.486
##  4 ExpressJet Airlines Inc.   0.519 0.481
##  5 Mesa Airlines Inc.         0.528 0.472
##  6 JetBlue Airways            0.559 0.441
##  7 Envoy Air                  0.568 0.432
##  8 United Air Lines Inc.      0.576 0.424
##  9 Endeavor Air Inc.          0.601 0.399
## 10 Virgin America             0.631 0.369
## 11 Delta Air Lines Inc.       0.651 0.349
## 12 American Airlines Inc.     0.666 0.334
## 13 US Airways Inc.            0.667 0.333
## 14 SkyWest Airlines Inc.      0.724 0.276
## 15 Alaska Airlines Inc.       0.731 0.269
## 16 Hawaiian Airlines Inc.     0.746 0.254
```

Esses percentuais nos dão a probabilidade condicional de um vôo pertencer a esta companhia aérea e estar atrasado. Veja o caso da Frontier Airlines Inc.: a probabilidade condicional de um vôo pertencer a esta empresa e estar atrasado é de aproximadamente 60% - se não há relação entre um vôo estar atrasado ou não e este pertencer a uma companhia aérea específica, esperamos que estes percentuais estejam próximos ao % total de atrasos e vôos não atrasados na base:

```
## # A tibble: 1 x 2
##       N     Y
##   <dbl> <dbl>
## 1 0.587 0.413
```

Comparando ao total da base, destacam-se as empresas Frontier Airlines Inc. e AirTran Airways Corporation, concentrando um percentual de atrasos com 17 pontos ou mais acima do percentual total da base. Em relação aos menores atrasos, temos as empresas SkyWest Airlines Inc., Alaska Airlines Inc. e Hawaiian Airlines Inc., com um percentual de vôos não atrasados até 14 pontos maior do que o percentual total da base - o restante das companhias aéreas tem um percentual de 8% ou menos em relação ao referencial.

Podemos agrupar as companhias com base nestes percentuais - vamos criar 3 grupos: os com os maiores atrasos, menos atrasos e sem diferenças perceptíveis, respectivamente, grupos 1, 2 e 3.

```
## # A tibble: 3 x 3
##   group     N     Y
```

```
##   <fct> <dbl> <dbl>
## 1 1     0.412 0.588
## 2 3     0.589 0.411
## 3 2     0.735 0.265
```

Esta comparação percentual com o total da base pode dar uma boa noção se há associação entre o grupo de empresas aéreas e o fato do vôo estar atrasado ou não. Formalmente, é necessário entender se esta diferença é estatisticamente significante: isso pode ser feito através de um teste de hipóteses.

Ele pode ser entendido como uma forma de avaliar o quão "difícil" seria encontrar estes valores de atraso caso nossa hipótese fosse falsa: por exemplo, se realizarmos o teste a um nível de significância de 5%, e o teste passasse, isso nos diria que só encontraríamos um valor de teste tão extremo em 5% das vezes caso empresa aérea e atrasos não fossem relacionados - vamos ao teste:

```
##
##  Pearson's Chi-squared test
##
## data:  company_groups_delays_freq
## X-squared = 588.34, df = 2, p-value < 2.2e-16
```

O p-valor é menor do que 0.05. Existem evidências que apontam que há diferença na concentração de atrasos entre os grupos: considerando o maior percentual de atrasos no grupo 1, não seria inadequado afirmar que há evidências que ele concentre mais atrasos do que os grupos restantes.

Repetindo o teste para empresas aéreas como um todo:

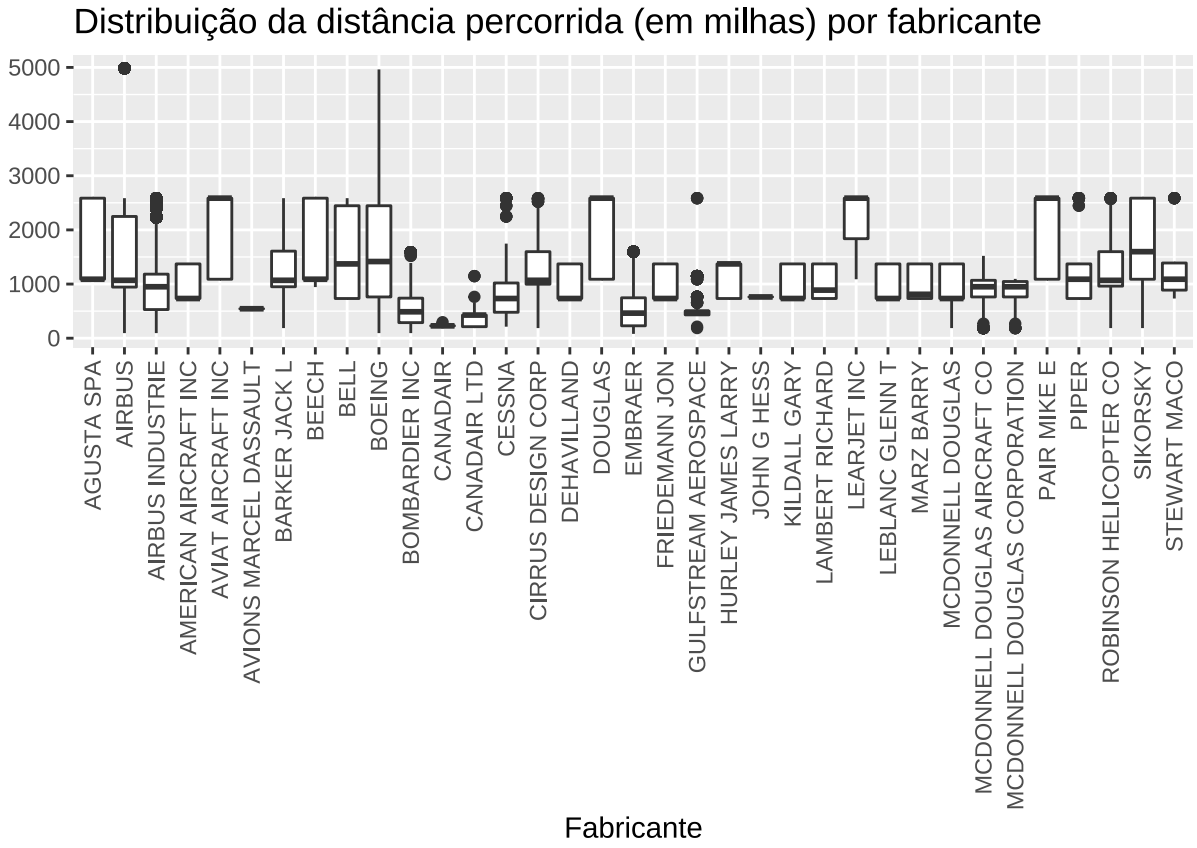
```
##
##  Pearson's Chi-squared test
##
## data:  company_delays_freq
## X-squared = 4287.5, df = 15, p-value < 2.2e-16
```

Também existem evidências que apontam para diferenças significativas entre o percentual de atrasos das empresas como um todo, com as empresas no grupo 1 concentrando mais atrasos em particular. Para aprofundar uma análise no tema, seria mais interessante, no entanto, prosseguir em uma análise com grupos, afim de simplificar análises.

Relação entre distância percorrida e fabricante/empresa

Embora estejamos lidando com uma situação similar à da análise anterior - verificar se uma variável está relacionada a diferentes grupos - essa análise requer técnicas diferentes. Considerando que temos vários grupos, uma análise tabular seria muito complexa, e a distância não é uma variável que pode ser classificada como "aconteceu" ou "não aconteceu".

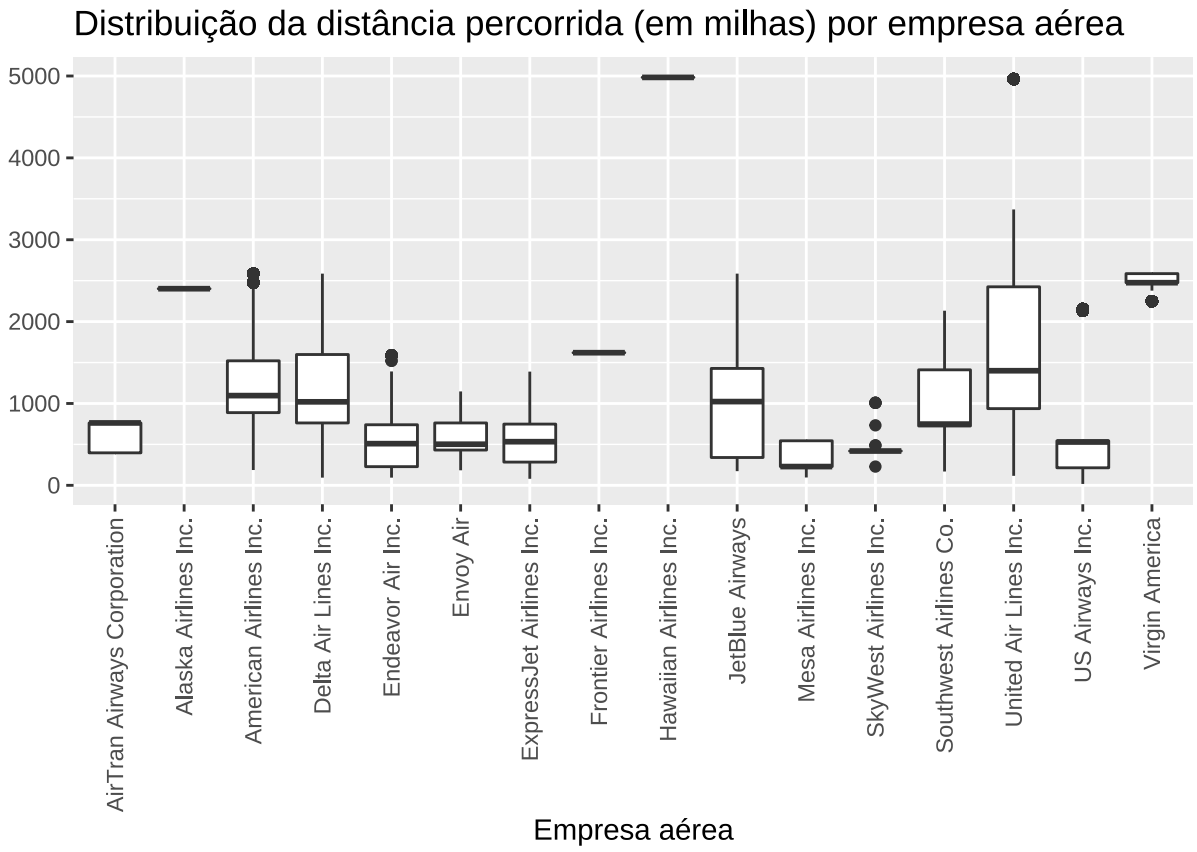
Uma técnica interessante é a ANOVA, que testa se uma média para diferentes grupos é diferente - sabendo que elas o são, podemos coletar evidências para afirmar que as duas variáveis estão relacionadas. Precisamos preparar nossa base e verificar algumas hipóteses: primeiramente, vamos verificar se os dados tendem a variar de maneira parecida, e se existem dados que são muito discrepantes dos demais.



Note o formato das caixas formadas pelo gráfico - elas tem comprimentos diferentes, indicando que os dados não tendem a variar de forma homogênea. Os pontos indicam valores muito discrepantes dos demais, ditos outliers - o gráfico não exibe uma quantidade tão grande de pontos de modo a exigir um tratamento de substituição de outliers: podemos prosseguir para o teste, assumindo variância desigual entre os grupos.

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  distance and manufacturer
## F = 19903, num df = 32.00, denom df = 637.45, p-value < 2.2e-16
```

Veja que o p-valor < 0.05, apontando evidências de que existe relação entre o fabricante e a distância percorrida pelo avião. Repetiremos o mesmo processo para empresas aéreas, apenas exibindo o gráfico e o resultado do teste para manter a brevidade:




```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  distance and name
## F = 146451, num df = 12.0, denom df = 1620.7, p-value < 2.2e-16
```

A conclusão se mantém, portanto, a nível de empresas aéreas.

Pensamento estatístico

- A conclusão se mantém desde todos os outros fatores sejam constantes (ex: a inadimplência cai a medida que o limite aumenta, desde que o número de negativações seja o mesmo)
- Através de exemplos: se, hipoteticamente, pegarmos alguns tomadores de crédito com limite alto, mas que tem um bom histórico de pagamentos de conta, podemos mostrar de maneira prática que existem outros fatores que podem afetar essa visualização. Nesse exemplo hipotético, o fato de que pessoas com limite alto tendem a pagar suas contas em dia.
- Uma técnica comumente aplicada para testar esse tipo de hipótese são testes *champion challenger*: nesse contexto, testar a concessão de limites atual x uma concessão de limites mais relaxada. Isso pode ser feito ao se selecionar, aleatoriamente, novos tomadores de crédito dentro de um mesmo grupo de risco para terem um limite maior, e outros para terem o limite que normalmente teriam, para garantir que os outros fatores foram devidamente controlados.
- Ainda sugeriria o mesmo desenho de experimento - seria possível orientar o risco que se deseja tomar ao limitar o percentual do portfolio que seria sujeitado ao experimento, reduzindo a amostra e o risco. Ainda seria possível diversificar o risco ao, por exemplo, sujeitar o experimento a grupos específicos dos quais se espera menor perda antes de realizar o experimento de política com os grupos restantes,

buscando uma implementação gradual.