



Especialização em *Business Intelligence*
Unidade Curricular de Análise de Dados
Ano Letivo de 2016/17
2º Semestre

2015 Flight Delays and Cancellations

Beatriz Loureiro (A68876), Hugo Rodrigues (A73476), João Fontes (A71184),
Pedro Lino (A66823)
Abril, 2017



Data de Recepção	
Responsável	
Avaliação	
Observações	

2015 Flight Delays and Cancellations

Beatriz Loureiro (A68876), Hugo Rodrigues (A73476), João Fontes (A71184), Pedro Lino (A66823)

Abril, 2017

Resumo

Um dos desafios de Aprendizagem Máquina passa pela construção de programas de computador que se melhorem automaticamente através de dados recolhidos e com a experiência absorvida de vários casos ao longo do tempo. Para isso, foi criado uma série de técnicas que nos permitem extrair conhecimento através de dados armazenados estruturadamente em bases de dados, *datasets*, etc. O objetivo deste trabalho foi estudar as diferentes metodologias de extração de conhecimento e, no final, estudar um *dataset* recorrendo à ferramenta R como auxílio nessa tarefa. No final tiramos as nossas próprias conclusões e dificuldades na aplicação destas técnicas de extração de conhecimento.

Área de Aplicação: *Business Intelligence, Data Science, Extração de Conhecimento de Dados, Data Mining*

Palavras-Chave: Extração de Conhecimento, R, Classificação, Clustering, Regras de Associação

Índice

1. Introdução	1
1.1. Contextualização	1
1.2. Apresentação do Caso de Estudo	1
1.3. Motivação e Objectivos	2
1.4. Estrutura do Relatório	2
2. Descrição e Compreensão da Natureza dos Dados	4
2.1. Descrição dos atributos do dataset	4
2.2. Análise Exploratória	6
Bibliografia	8
Lista de Siglas e Acrónimos	9
 Anexos	
I. Análise da Dispersão dos Dados Temporais	11

Índice de Figuras

Figura 1 - Localização dos Aeroportos no Mapa dos EUA	7
Figura 2 - <i>Boxplot</i> de Dispersão dos Dados Temporais	11
Figura 3 - Histogramas dos Atributos Temporais	12

Índice de Tabelas

No table of figures entries found.

1. Introdução

1.1. Contextualização

Um dos desafios da Aprendizagem Máquina passa pela construção de programas de computador que melhorem automaticamente com a experiência. Para tal, algoritmos que especificam, passo a passo, como resolver um problema, não são capazes de, por exemplo, encontrar padrões num grande conjunto de dados, pelo simples fato de que, de momento, é impossível codificar tantas e complexas características e todos os padrões possíveis para que o computador os encontre.

A partir desta noção de reconhecimento de padrões, podemos introduzir o conceito de *Machine Learning* [1], ou, traduzindo, Aprendizagem Máquina. Este conceito surge quando se pretende responder a questões mais complexas sobre um conjunto de dados. Tendo como exemplo a informação de vendas de uma loja, um SGBD consegue dizer quantas pessoas compraram um produto, mas não consegue identificar conjuntos de produtos que são frequentemente comprados em conjunto [2]. Para isso, é necessário usar um conjunto de técnicas que consigam classificar, prever, agrupar e extrair padrões destes dados, poupando assim os seres humanos dessa tarefa cansativa e de elevada dificuldade.

1.2. Apresentação do Caso de Estudo

Tendo sido propostos pelo docente 3 *datasets*, foi-nos atribuído o *dataset* “2015 Flight Delays and Cancellations”, disponível em [3], que contém os dados dos voos comerciais realizados no ano de 2015 nos EUA. Este *dataset* está dividido em 3 ficheiros CSV:

- ***airlines.csv*** - contém os códigos IATA das companhias aéreas e os seus respetivos nomes;
- ***airports.csv*** - contém os códigos IATA dos aeroportos, bem como os seus respetivos nomes e localização geográfica (cidade, estado, país e coordenadas geográficas)
- ***flights.csv*** - contém os dados dos voos comerciais realizados nos EUA no ano de 2015, desde a data de realização, a companhia aérea que o realizou, aeroportos de origem e destino, tempos espectáveis e reais de partida, de voo, chegada, tempos de atraso, tempos em que as rodas do avião saem do aeroporto de origem e em que chegam ao aeroporto de destino e os tempos de embarque e desembarque.

1.3. Motivação e Objectivos

Com o forte crescimento da tecnologia, hoje em dia, todos os serviços procuram modernizar as suas infraestruturas tecnológicas por forma a satisfazer melhor tanto os seus clientes como a poder ter o melhor encaixe financeiro possível com o menor esforço. Uma análise do volume de negócios e da BD de clientes permite a uma empresa atingir estes objetivos e obter previsões sobre o seu futuro operacional.

Com isto em mente, podemos verificar que este *dataset* é propício para análise, uma vez que apresenta uma grande quantidade de dados (aproximadamente 6 milhões de registos de voos), o que leva a que as previsões e conclusões tiradas a partir deles possam apresentar uma maior exatidão.

Tendo como objetivo principal descobrir conhecimento específico sobre os vôs nos EUA no ano de 2015, podemos definir alguns objetivos de análise para este *dataset* com vista a atingir esse fim como sendo:

- Descobrir qual a altura do ano mais propícia a haver menos atrasos nos voos, usando as capacidades de visualização de gráficos da ferramenta R;
- Descobrir que companhias conseguem atingir maior velocidade no tratamento de um voo (tempo de voo + atrasos), usando também as técnicas de visualização de gráficos da ferramenta R;
- Criar um modelo que possa prever o atraso de um qualquer voo, usando técnicas de Previsão e Classificação de dados, testando diversos modelos e verificando a sua exatidão;
- Agrupar os aeroportos mediante os atrasos presentes nos voos que servem, usando técnicas de Clustering;
- Procurar padrões nas relações entre aeroportos e companhias aéreas, para procurar possíveis causas para os atrasos verificados, usando para isso a análise de Regras de Associação.

1.4. Estrutura do Relatório

A estrutura do relatório segue uma sequência lógica baseada na metodologia de DM largamente usada na indústria, denominada CRISP-DM, sendo que os capítulos do relatório retratam os diversos passos desta metodologia.

No primeiro capítulo será possível encontrar a primeira fase de análise do *dataset* e das suas particularidades, fazendo-se assim uma análise exploratória dos dados a usar.

O segundo capítulo retrata o processo de preparação dos dados para os processos de análise.

O terceiro capítulo apresenta as fases de desenvolvimento dos modelos para responder às questões apresentadas com objetivo de análise.

O quarto capítulo apresenta a avaliação dos modelos obtidos, usando diversas métricas de erro que nos permitam avaliar a exatidão e assertividade dos modelos criados.

O quinto capítulo irá retratar como foi feita a implementação dos modelos de resposta às questões apresentadas como objetivo na ferramenta R.

Por fim, irá ser feita uma síntese de todo o trabalho realizado no âmbito deste trabalho, bem como algumas linhas de orientação para a realização do trabalho futuro.

2. Descrição e Compreensão da Natureza dos Dados

Neste capítulo vamos analisar o o conjunto de dados proposto, “2015 Flight Delays and Cancellations”, disponibilizado através da plataforma Kaggle em [3]. Este *dataset* apresenta-nos um conjunto de dados relativo aos voos no ano de 2015 nos EUA, disponibilizados pelo *Department of Transportation*.

2.1. Descrição dos atributos do dataset

Os dados são disponibilizados em 3 ficheiros CSV, pelo que, de maneira a podermos analisar os seus atributos, começamos por carrega-los para a plataforma R.

```
airlines <- read.csv ("flight-delays/airlines.csv")
airports <- read.csv ("flight-delays/airports.csv")
flights <- read.csv ("flight-delays/flights.csv")
```

Após carregar os ficheiros, podemos assumir que estes ficheiros resultaram de exportações de tabelas de uma BD Relacional, podemos uni-los usando a função *merge*, equivalente a um *join* numa BD Relacional. É necessário alterar o nome da coluna que representa o nome da companhia aérea no *dataset airlines*, uma vez que iria causar um conflito de nomes e esse nome não iria aparecer no *dataset flights* após a execução da função *merge*.

```
colnames(airlines) <- c("IATA_CODE", "AIRLINE_NAME")
flights <- merge(flights, airports, by.x = "ORIGIN_AIRPORT", by.y =
"IATA_CODE")
flights <- merge(flights, airports, by.x = "DESTINATION_AIRPORT", by.y =
"IATA_CODE")
flights <- merge(flights, airlines, by.x = "AIRLINE", by.y = "IATA_CODE")
```

Assim, temos um só *data frame* para analisar, simplificando as operações que teremos de realizar.

De seguida, listamos os atributos presentes no *dataset*, apresentando os valores que podem tomar e uma breve explicação dos mesmos [4].

- **AIRLINE** - código IATA da companhia aérea que efetuou o voo;
- **DESTINATION_AIRPORT; ORIGIN_AIRPORT** - códigos IATA dos aeroportos de destino e origem;
- **YEAR** - ano em que se realizou o voo (sempre 2015);
- **MONTH** - mês em que se realizou o voo (1 a 12);
- **DAY** - dia em que se realizou o voo (1 a 28/30/31);
- **DAY_OF_WEEK** - dia da semana em que se realizou o voo (1 - Domingo, ..., 7 - Sábado);
- **FLIGHT_NUMBER** - identificador numérico que identifica cada voo;
- **TAIL_NUMBER** - identificador numérico que identifica a cauda do voo;
- **SCHEDULED_DEPARTURE; DEPARTURE_TIME; DEPARTURE_DELAY** - hora espectável e real de partida do voo e respetivo atraso (todas as horas são representadas por HHMM (representa a hora HH:MM), todos os atrasos são representados em minutos);
- **TAXI_OUT; TAXI_IN** - tempos em minutos entre o começo do embarque e a saída das rodas do avião do aeroporto de origem e a chegada das rodas do avião ao aeroporto de destino e o final do desembarque;
- **WHEELS_OFF; WHEELS_ON** - horas em que as rodas do avião saem do avião de origem e chegam ao aeroporto de destino;
- **SCHEDULED_TIME; ELAPSED_TIME; AIR_TIME** - tempos espectáveis, reais e de voo, em minutos, do avião;
- **DISTANCE** - distância percorrida, em quilómetros;
- **SCHEDULED_ARRIVAL; ARRIVAL_TIME; ARRIVAL_DELAY** - hora espectável e real de chegada do voo e respetivo atraso
- **DIVERTED; CANCELLED, CANCELLATION_REASON** - valores booleanos que indicam se o voo foi desviado, cancelado e, caso cancelado, a razão para o seu cancelamento;
- **AIR_SYSTEM_DELAY; SECURITY_DELAY; AIRLINE_DELAY; LATE_AIRCRAFT_DELAY WEATHER_DELAY** - tempos de atraso nos diversos estágios do voo, desde tempos de atraso no *check-in*, na segurança, atrasos da companhia aérea, na chegada atrasada do avião ou por causa das condições climáticas.

Os atributos descritos até este ponto são os atributos nativos do *dataset flights*. Os atributos seguintes foram adicionados a este *dataset* após a execução da função *merge*.

- **AIRPORT.{x,y}** - nomes dos aeroportos de origem e destino;
- **CITY.{x,y}; STATE.{x,y}; COUNTRY.{x,y}** - localização dos aeroportos de origem e destino (cidade, estado e país (sempre EUA));

- **LATITUDE.{x,y}; LONGITUDE.{x,y}** - coordenadas geográficas dos aeroportos de origem e destino;
- **AIRLINE_NAME** - nome da companhia aérea que efetuou o voo.

2.2. Análise Exploratória

Começamos por analisar como estão distribuídos os valores que representam tempos e atrasos, usando para isso a função *summary* do R.

```
time.att <- c("SCHEDULED_TIME", "ELAPSED_TIME", "AIR_TIME", "ARRIVAL_DELAY",
             "AIR_SYSTEM_DELAY", "SECURITY_DELAY", "AIRLINE_DELAY",
             "LATE_AIRCRAFT_DELAY", "WEATHER_DELAY")
summary(flights[ ,time.att])
```

Com isto, é-nos possível auferir a distribuição de valores que estes dados apresentam. O *output* da função é o seguinte:

SCHEDULED_TIME	ELAPSED_TIME	AIR_TIME	ARRIVAL_DELAY	AIR_SYSTEM_DELAY	SECURITY_DELAY
Min. : 18.0	Min. : 14.0	Min. : 7.0	Min. : -87.00	Min. : 0	Min. : 0
1st Qu.: 85.0	1st Qu.: 82.0	1st Qu.: 60.0	1st Qu.: -13.00	1st Qu.: 0	1st Qu.: 0
Median :123.0	Median :119.0	Median : 94.0	Median : -5.00	Median : 2	Median : 0
Mean :141.8	Mean :137.2	Mean :113.7	Mean : 4.89	Mean : 13	Mean : 0
3rd Qu.:174.0	3rd Qu.:169.0	3rd Qu.:144.0	3rd Qu.: 8.00	3rd Qu.: 18	3rd Qu.: 0
Max. :718.0	Max. :766.0	Max. :690.0	Max. :1971.00	Max. :1134	Max. :573
NA's :6	NA's :101784	NA's :101784	NA's :101784	NA's :4329554	NA's :4329554

AIRLINE_DELAY	LATE_AIRCRAFT_DELAY	WEATHER_DELAY
Min. : 0	Min. : 0	Min. : 0
1st Qu.: 0	1st Qu.: 0	1st Qu.: 0
Median : 2	Median : 4	Median : 0
Mean : 19	Mean : 24	Mean : 3
3rd Qu.: 19	3rd Qu.: 30	3rd Qu.: 0
Max. :1971	Max. :1331	Max. :1211
NA's :4329554	NA's :4329554	NA's :4329554

Podemos verificar que os dados apresentam muitos valores nulos, algo que irá dificultar a extração de conhecimento mais à frente e e que terão de ser tratados.

Podemos verificar ainda que os valores não nulos são muito díspares, pois apresentam grandes diferenças entre os valores máximos e os valores representativos do 3º quartil, o que vai gerar uma grande quantidade de *outliers*. Podemos explicar esta disparidade de valores pelo pouco ou nenhum atraso dos voos, o que leva a que os atrasos tenham um valor próximo de 0, levando assim a uma concentração dos dados aí.

Para comprovar-mos as afirmações feitas anteriormente, podemos usar um *boxplot* ou histogramas para auferir melhor qual é a distribuição dos dados. Estes gráficos podem ser encontrados no anexo I. Através da análise destes gráficos podemos comprovar a existência de uma grande dispersão dos dados e de um grande número de *outliers*, como foi referido anteriormente.

Após alguma pesquisa nos *kernels* da plataforma *Kaggle*, reparamos que alguns dos códigos dos aeroportos descritos nos atributos `ORIGIN_AIRPORT` e `DESTINATION_AIRPORT` não estão coerentes com os presentes no *dataset airports*. Estes códigos terão de ser tratados também na fase de processamento de dados.

A título de curiosidade, usamos o pacote *maps* da plataforma R para podermos visualizar a localização dos aeroportos presentes no *dataset airports* no mapa dos EUA.

```
map("usa")  
title("Airports")  
points(airports$LONGITUDE, airports$LATITUDE, col="red", cex=0.75)
```



Figura 1 - Localização dos Aeroportos no Mapa dos EUA

Bibliografia

- [1] João Gama, André Ponce De Leon Carvalho, Katti Faceli, Ana Carolina Lorena, and Márcia Oliveira, 2015. *Extração de Conhecimento de Dados - Data Mining*. Lisboa: Edições Sílabo.
- [2] Jiawei Han, Micheline Kamber, and Jian Pei, 2011. *Data mining: Concepts and techniques (the Morgan Kaufmann series in data management systems)*. Amsterdam: Morgan Kaufmann Publishers In.
- [3] Kaggle. (2017). *2015 Flight Delays and Cancellations* / Kaggle. [Online] Disponível em: <https://www.kaggle.com/usdot/flight-delays> [Acedido em 10 Abril 2017].
- [4] Transtats.bts.gov. (2017). *RITA / BTS / Transtats*. [Online] Disponível em: https://www.transtats.bts.gov/Fields.asp?Table_ID=236 [Acedido em 10 Abril 2017].

Lista de Siglas e Acrónimos

BD	Base de Dados
CRISP-DM	<i>CRoss Industry Standard Process for Data Mining</i>
CSV	<i>Comma-Separated Values</i>
DM	<i>Data Mining</i>
EUA	Estados Unidos da América
IATA	<i>International Air Transport Association</i>
SGBD	Sistema de Gestão de Bases de Dados

Anexos

I. Análise da Dispersão dos Dados Temporais

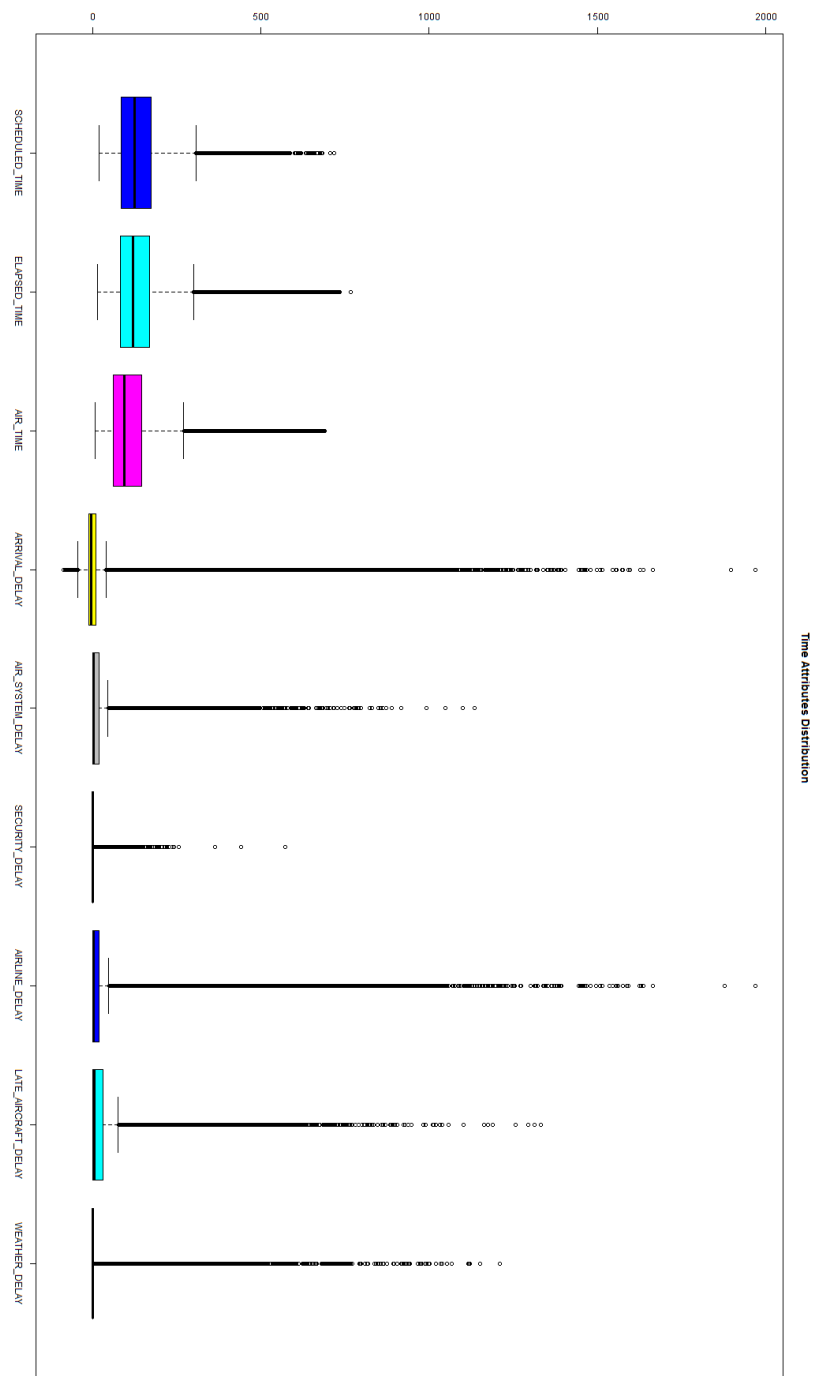


Figura 2 - Boxplot de Dispersão dos Dados Temporais

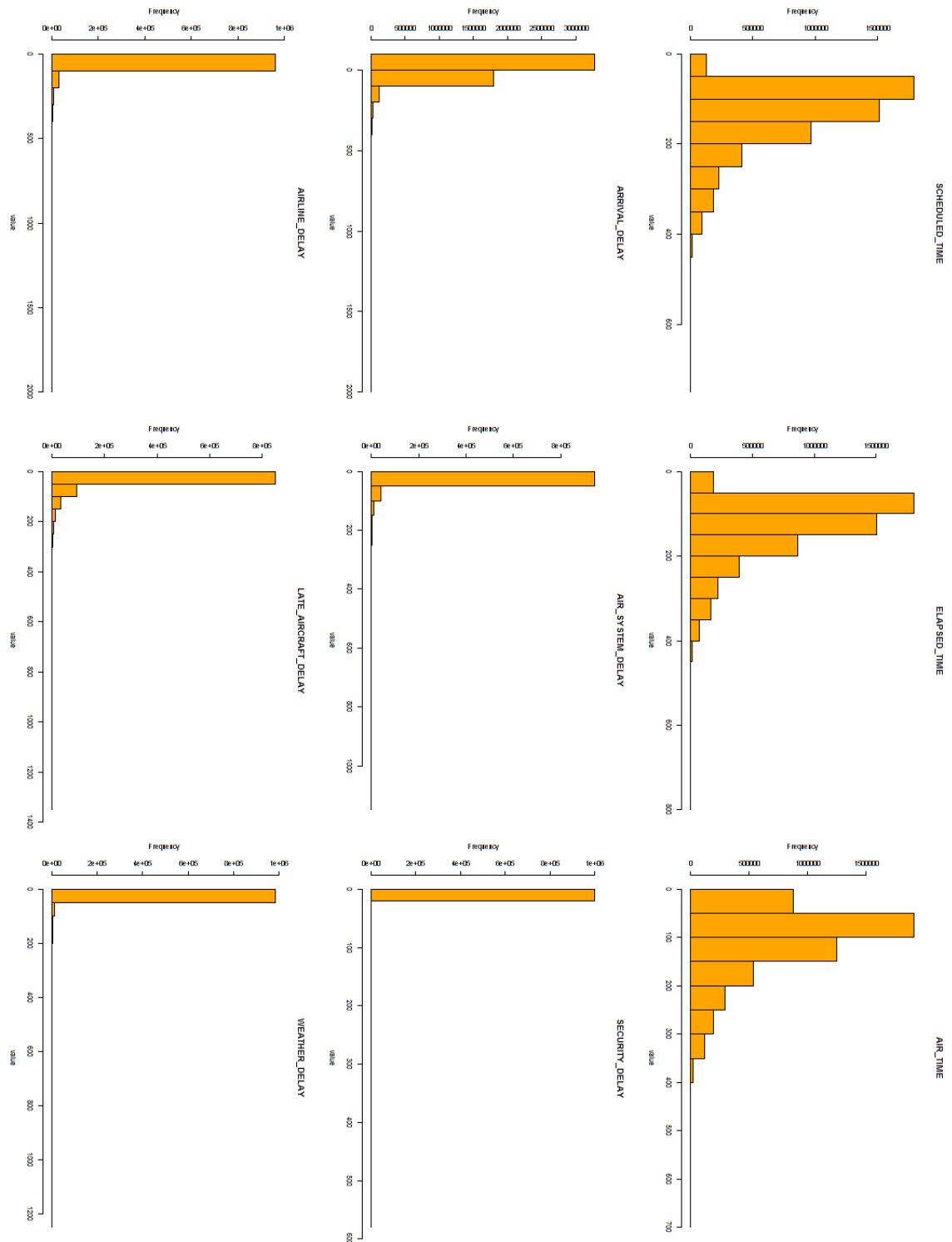


Figura 3 - Histogramas dos Atributos Temporais