2015 Flight Delays and Cancellations

Beatriz Loureiro (a68876)
João Fontes (a71184)
Hugo Rodrigues (a73476)
Pedro Lino (a66823)

Índice

- Questões Colocadas
- Descrição dos Dados
- Preparação dos Dados
- Modelação e Implementação
- Resultados

Questões Colocadas

- ✓Descobrir a altura do ano mais propícia a haver menos atrasos nos voos
- ✓Descobrir que companhias conseguem atingir maior velocidade no tratamento de um voo
- ✓Criar um modelo que possa prever o atraso de um qualquer voo
- ✓Agrupar os aeroportos mediante os atrasos presentes nos voos que servem
- ✓Procurar padrões nas relações entre aeroportos e companhias aéreas

Descrição dos Dados

- ✓**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 e realizou o voo (1 a 28/30/31);
- ✓**DAY_OF_WEEK** - dia da semana em que se realizou o voo (1 - Domingo, …, 7 - Sábado);

Descrição dos Dados

- ✓**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);

Descrição dos Dados

- ✓**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 kilómetros;

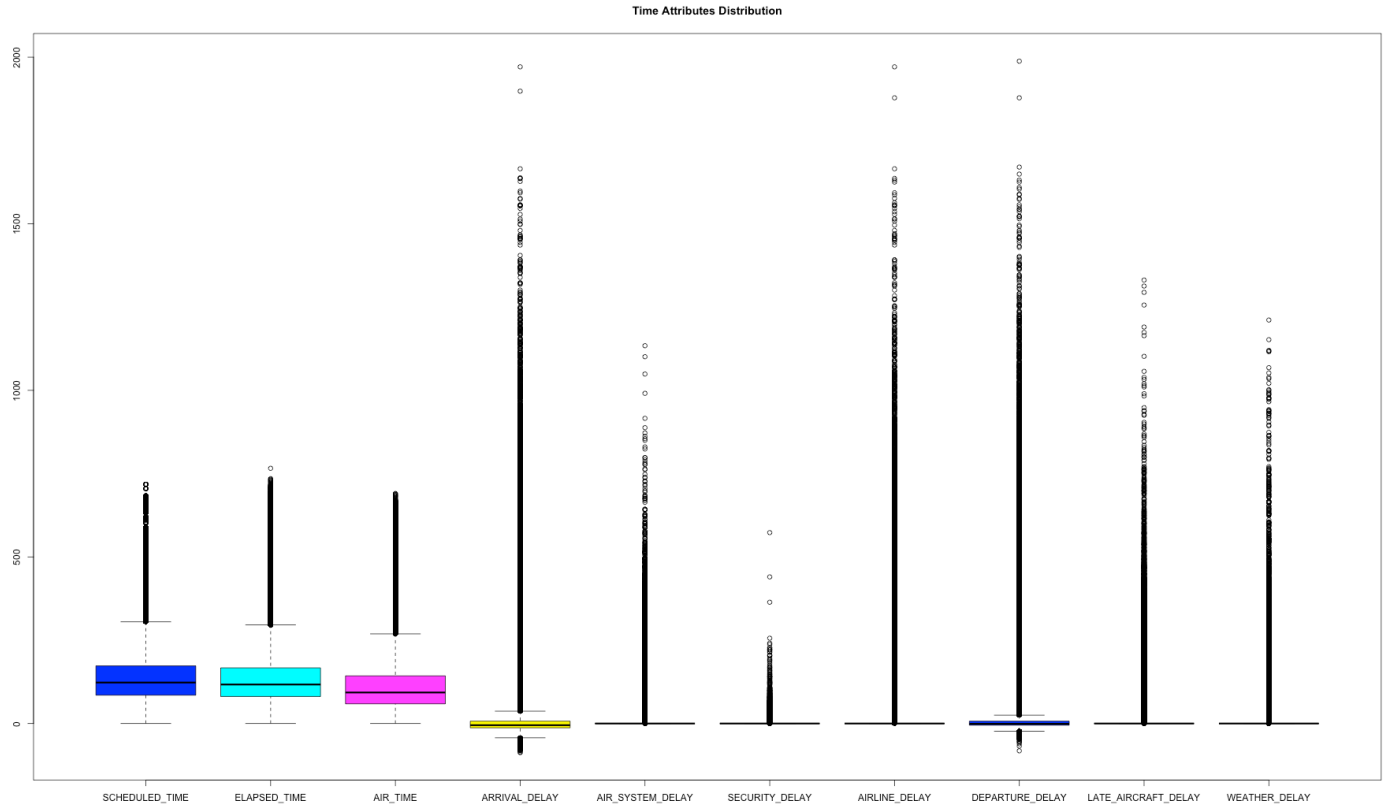
Descrição dos Dados

- ✓**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 climatéricas.

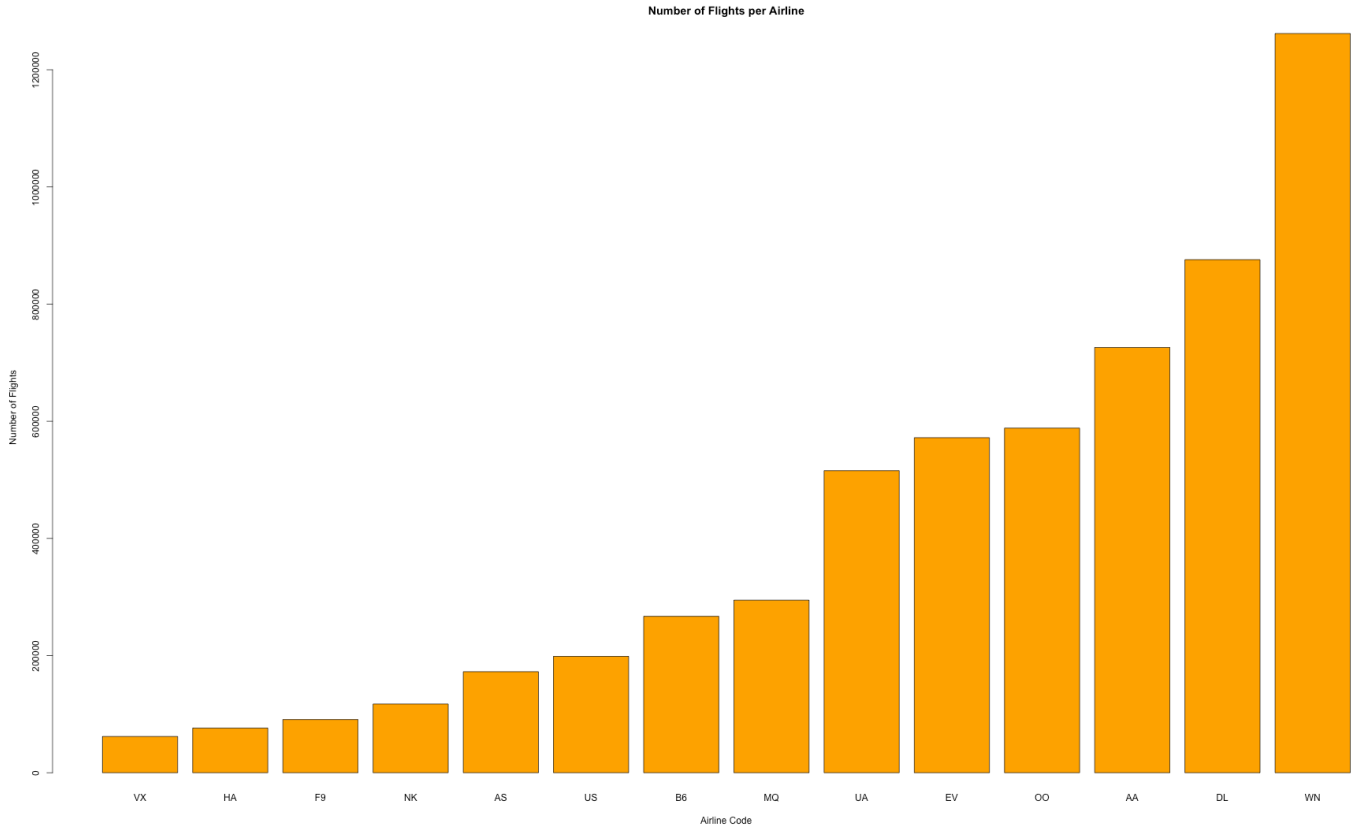
Descrição dos Dados

- ✓**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.

Descrição dos Dados



Descrição dos Dados



Preparação dos Dados

- **Códigos de Aeroportos errados (kernel Kaggle)**
  - ✓flights$ORIGIN_AIRPORT <-
    id.to.iata(flights$ORIGIN_AIRPORT)
  - ✓flights$DESTINATION_AIRPORT <-
    id.to.iata(flights$DESTINATION_AIRPORT)
- **Valores Nulos**
  - ✓flights[is.na(flights)] <- 0
- **Cálculos dos Atributos DELAY e DELAYED**
  - ✓flights[, "DELAY"] <- rowSums(flights[, delay.att])
  - ✓flights[, "DELAYED"] <- ifelse(flights$DELAY > 0, 0, 1)

Modelação e Implementação

- **Melhor altura do ano para viajar**
  - Usar capacidades de **Visualização de Dados** do R
  - **Dados**
    - atrasos médios por cada mês do ano civil
  - **Implementação**
    - Gráfico de barras → comando **barplot**

Modelação e Implementação

- **Melhor companhia aérea onde se viajar**
  - Usar capacidades de **Visualização de Dados** do R
  - **Dados**
    - tempos médios de tratamento de um voo (atrasos mais tempo de voo) para cada companhia aérea
    - valores de atraso médios por cada companhia aérea
  - **Implementação**
    - Gráfico de barras → comando **barplot**

Modelação e Implementação

- **Prever se um voo vai ser atrasado**
  - Usar **Classificação**
  - **Dados**
    - Amostra de 20% do *dataset* original
    - Atributos: *AIRLINE, DESTINATION_AIRPORT, ORIGIN_AIRPORT, MONTH, DAY, DAY_OF_WEEK, SCHEDULED_DEPARTURE, SCHEDULED_TIME, SCHEDULED_ARRIVAL, DELAYED*
  - **Implementação**
    - Regressão Linear → comando **lm**
    - Árvores de Decisão → comando **rpart**
    - *Naïve Bayes* → comando **naiveBayes**

Modelação e Implementação

- **Prever o atraso de um voo**
  - Usar **Regressão**
  - **Dados**
    - Amostra de 20% do *dataset* original
    - Atributos: *AIRLINE, DESTINATION_AIRPORT, ORIGIN_AIRPORT, MONTH, DAY, DAY_OF_WEEK, SCHEDULED_DEPARTURE, SCHEDULED_TIME, SCHEDULED_ARRIVAL, DELAY*
  - **Implementação**
    - Regressão Linear → comando **lm**

Modelação e Implementação

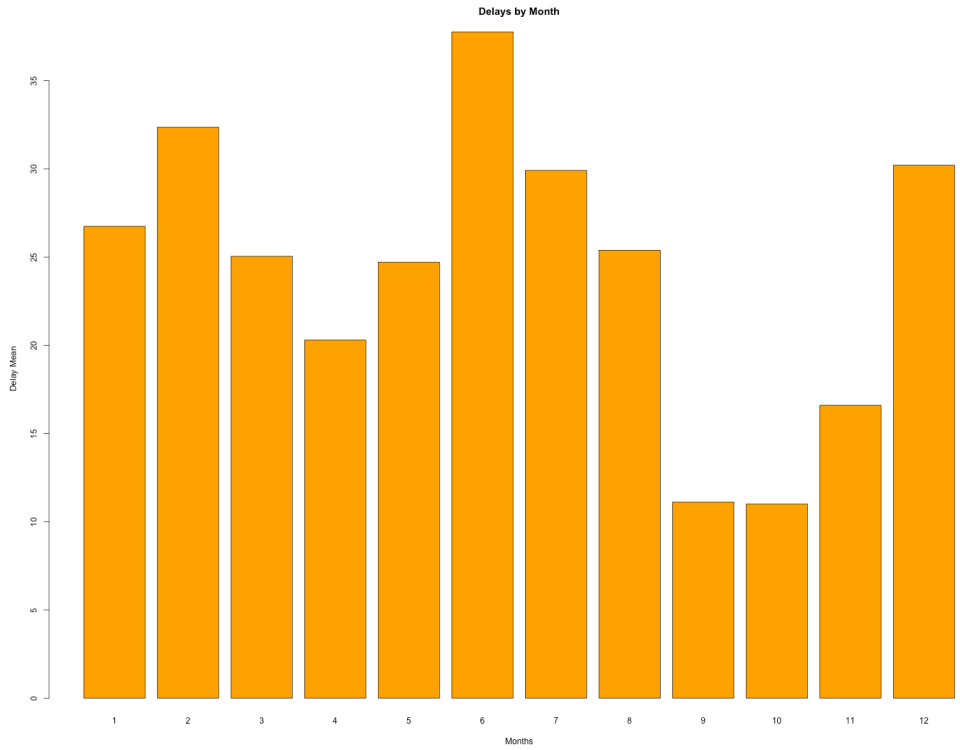
- **Agrupar aeroportos de acordo com os atrasos**
  - Usar ***Clustering***
  - Agrupamento por aeroportos de origem e destino
  - **Dados**
    - Tempos médios de atrasos por aeroporto
  - **Implementação**
    - *Clustering* Hierárquico → comando **hclust**
    - *Clustering K-Means* → comando **kmeans**

Modelação e Implementação

- **Procurar padrões entre aeroportos e companhias aéreas**
  - Usar **Regras de Associação**
  - **Dados**
    - Atributos: *AIRLINE, DESTINATION_AIRPORT, ORIGIN_AIRPORT, DELAY*
    - Atrasos discretizados em 5 intervalos de igual frequência
    - *Dataset* transformado em transações
  - **Implementação**
    - *Algoritmo Apriori*→ comando **apriori**

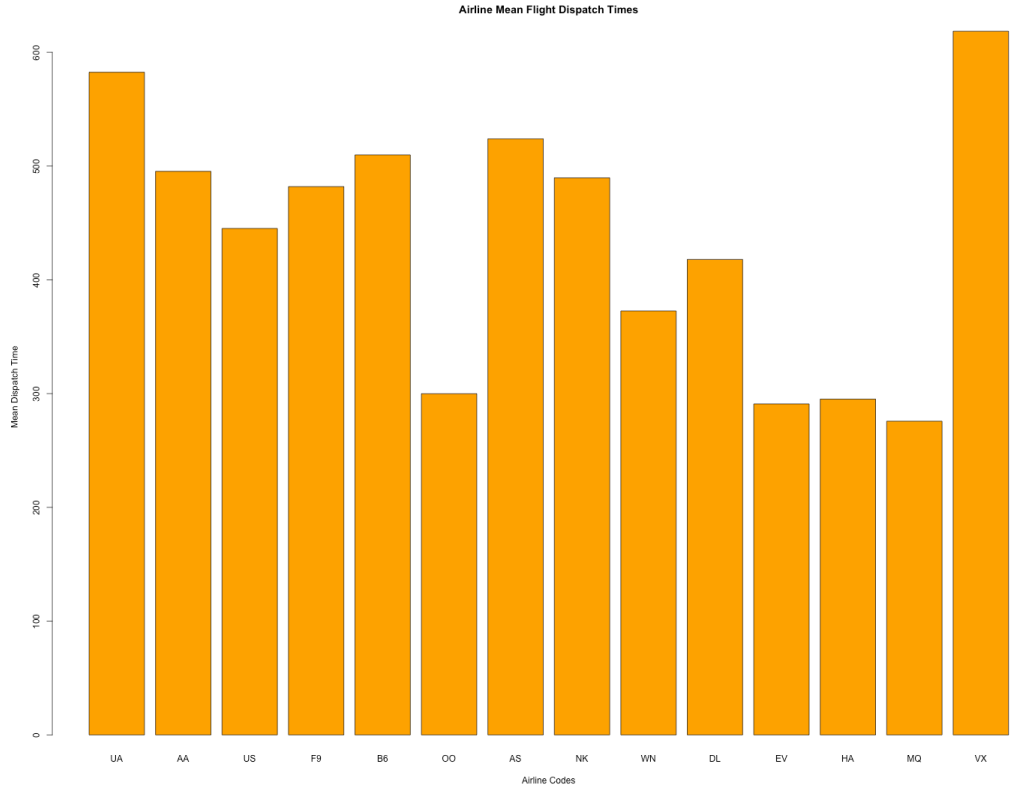
Resultados

- **Melhor altura do ano para viajar**



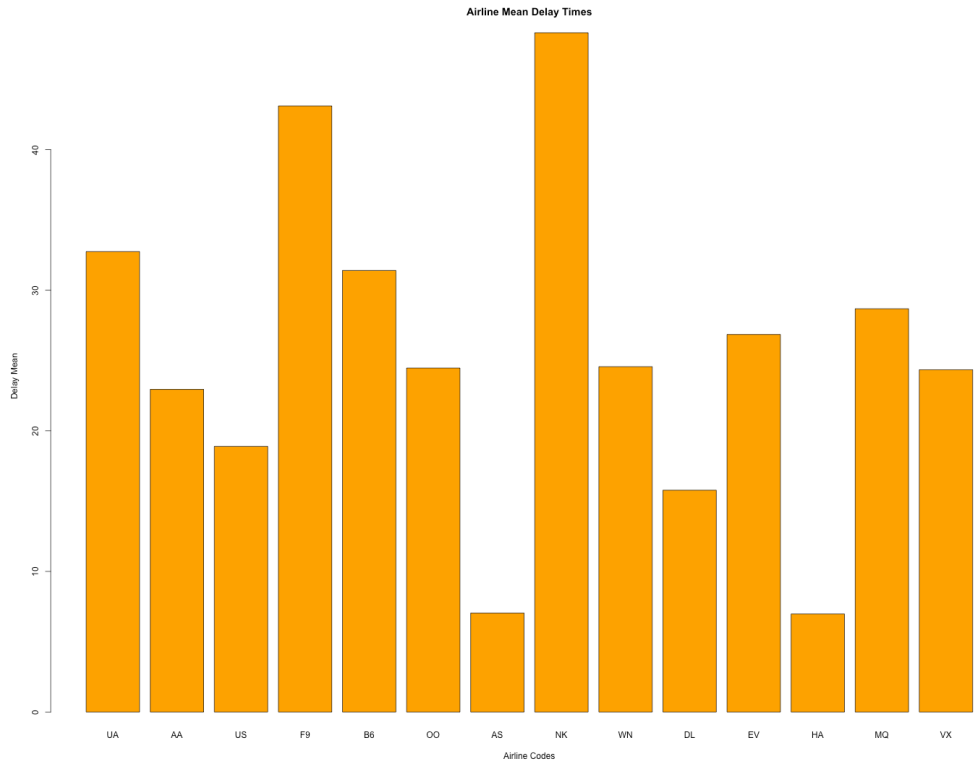
Resultados

- **Melhor altura do ano para viajar**



Resultados

- **Melhor companhia aérea onde se viajar**



Resultados

- **Prever se um voo vai ser atrasado**
  - Regressão Linear

```
Confusion Matrix and Statistics

          Reference
Prediction      0      1
         0  20890  17542
         1 122371 227135

               Accuracy : 0.6393
```

Resultados

- **Prever se um voo vai ser atrasado**
  - Árvores de Decisão

```
Confusion Matrix and Statistics

          Reference
Prediction      0      1
         0 108161 150578
         1  35100  94099

               Accuracy : 0.5214
```

Resultados

- **Prever se um voo vai ser atrasado**
  - *Naïve Bayes*

```
Confusion Matrix and Statistics

          Reference
Prediction     0      1
         0  43999  44514
         1  99262 200163

               Accuracy : 0.6294
```

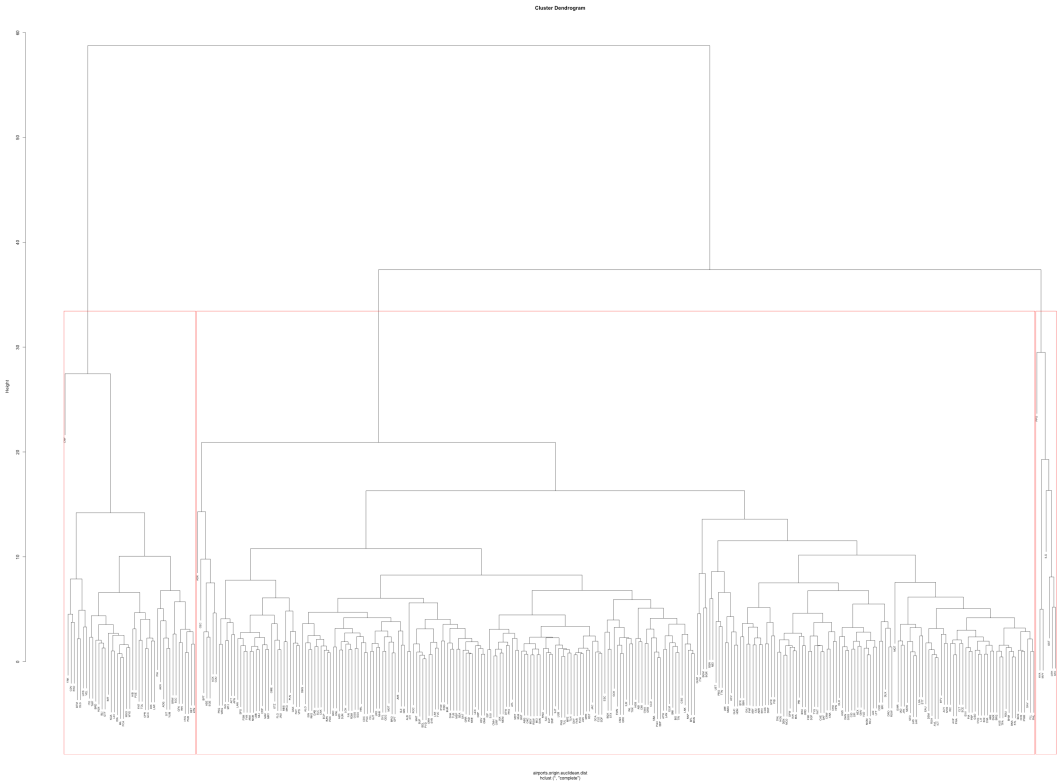
Resultados

- **Prever o atraso de um voo**
  - Regressão Linear

SSE	RMSE	MAE
4.479021e+09	1.074510e+02	5.316675e+01

Resultados

- Agrupar aeroportos de acordo com os atrasos

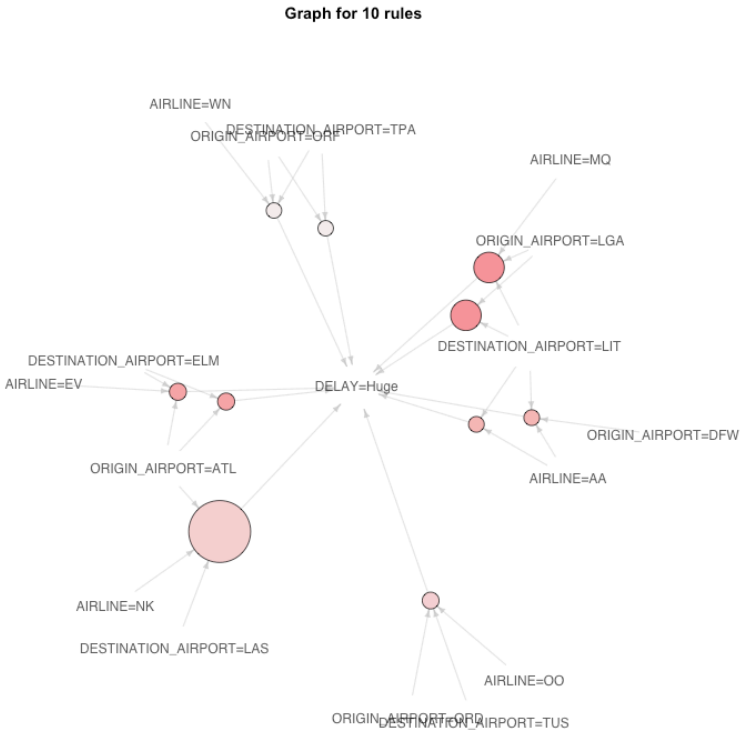


Resultados

- Agrupar aeroportos de acordo com os atrasos

Resultados

- **Procurar padrões entre aeroportos e companhias aéreas**



2015 Flight Delays and Cancellations

Beatriz Loureiro (a68876)
João Fontes (a71184)
Hugo Rodrigues (a73476)
Pedro Lino (a66823)