Introdução ao Tech Challenge

O Tech Challenge disponibilizou um dataset chamado 'flights.csv', uma planilha de 5819079 linhas e 31 colunas, contendo todos os voos dos EUA do ano de 2015, totalizando 578 Megabytes de dados. No Tech Challenge é pedido que o aluno gere insights sobre o sistema aeroviário e construa um modelo supervisionado e um modelo não supervisionado, buscando tentar prever se um voo irá atrasar ou não.

Na planilha possui-se as seguintes informações:

Coluna	Descrição	Tipo / Unidade
YEAR	Ano do voo (ex.: 2015)	Inteiro
MONTH	Mês do voo (1 a 12)	Inteiro
DAY	Dia do mês do voo (1 a 31)	Inteiro
DAY_OF_WEEK	Dia da semana (1 = Segunda, 7 = Domingo)	Inteiro
AIRLINE	Código da companhia aérea (ex.: AA = American Airlines)	Categórica
FLIGHT_NUMBER	Número do voo	Inteiro
TAIL_NUMBER	Número de registro da aeronave	Texto
ORIGIN_AIRPORT	Código IATA do aeroporto de origem (ex.: ATL)	Categórica
DESTINATION_AIRPORT	Código IATA do aeroporto de destino	Categórica
SCHEDULED_DEPARTURE	Horário de partida programado (HHMM)	Inteiro
DEPARTURE_TIME	Horário real de partida (HHMM)	Inteiro
DEPARTURE_DELAY	Atraso na partida (em minutos)	Numérico
TAXI_OUT	Tempo gasto taxiando até a decolagem (em minutos)	Numérico
WHEELS_OFF	Horário em que o avião decolou (HHMM)	Inteiro
SCHEDULED_TIME	Tempo total programado de voo (em minutos)	Numérico
ELAPSED_TIME	Tempo total real de voo (em minutos)	Numérico
AIR_TIME	Tempo no ar (em minutos)	Numérico
DISTANCE	Distância entre origem e destino (em milhas)	Numérico
WHEELS_ON	Horário em que as rodas tocaram o solo (HHMM)	Inteiro
TAXI_IN	Tempo taxiando até o portão de desembarque (em minutos)	Numérico
SCHEDULED_ARRIVAL	Horário de chegada programado (HHMM)	Inteiro
ARRIVAL_TIME	Horário de chegada real (HHMM)	Inteiro
ARRIVAL_DELAY	Atraso na chegada (em minutos)	Numérico
DIVERTED	Indica se o voo foi desviado (1 = sim, 0 = não)	Binária
CANCELLED	Indica se o voo foi cancelado (1 = sim, 0 = não)	Binária
CANCELLATION_REASON	Motivo do cancelamento (A = Airline, B = Weather, C = NAS, D = Security)	Categórica
AIR_SYSTEM_DELAY	Atraso causado por controle de tráfego aéreo	Numérico
SECURITY_DELAY	Atraso causado por problemas de segurança	Numérico
AIRLINE_DELAY	Atraso causado pela companhia aérea	Numérico
LATE_AIRCRAFT_DELAY	Atraso causado por chegada tardia da aeronave	Numérico
WEATHER_DELAY	Atraso causado por condições meteorológicas	Numérico

Modificando o DataFrame

Primeiramente, usando a biblioteca Pandas, renomeou-se as colunas, obtendo-se assim as seguintes colunas:

'Ano', 'Mês', 'Dia', 'Dia da Semana', 'Companhia Aérea', 'Número do Voo', 'Número de Registro da Aeronave', 'Aeroporto de Origem', 'Aeroporto de Destino', 'Horário Programado para Partida (HHMM)', 'Horário Real de Partida (HHMM)', 'Atraso na Partida', 'Tempo Gasto até Decolagem', 'Horário em que o Avião Decolou (HHMM)', 'Tempo de Voo Planejado', 'Tempo Real de Voo', 'Tempo no Ar', 'Distância', 'Momento do Pouso (HHMM)', 'Tempo do Pouso até Desembarque', 'Horário de Chegada Programado (HHMM)', 'Horário de Chegada Real (HHMM)', 'Atraso de Chegada', 'Desvio de Voo', 'Status de Cancelamento', 'Motivo de Cancelamento', 'Atraso por Controle de Espaço Aéreo', 'Atraso por Segurança', 'Atraso da Companhia', 'Atraso por Aeronave Anterior', 'Atraso por Condições Meteorológicas'

Fazendo análise dos dados faltantes para cada coluna, obteve-se:

```
flights.csv
shape: (5819079, 31)
Ano                                        0
Mês                                        0
Dia                                        0
Dia da Semana                              0
Companhia Aérea                            0
Número do Voo                              0
Número de Registro da Aeronave         14721
Aeroporto de Origem                        0
Aeroporto de Destino                       0
Horário Programado para Partida (HHMM)     0
Horário Real de Partida (HHMM)         86153
Atraso na Partida                      86153
Tempo Gasto até Decolagem              89047
Horário em que o Avião Decolou (HHMM)  89047
Tempo de Voo Planejado                     6
Tempo Real de Voo                     105071
Tempo no Ar                           105071
Distância                                  0
Momento do Pouso (HHMM)                92513
Tempo do Pouso até Desembarque         92513
Horário de Chegada Programado (HHMM)       0
Horário de Chegada Real (HHMM)         92513
Atraso de Chegada                     105071
Desvio de Voo                              0
Status de Cancelamento                     0
Motivo de Cancelamento                5729195
Atraso por Controle de Espaço Aéreo   4755640
Atraso por Segurança                  4755640
Atraso da Companhia                   4755640
Atraso por Aeronave Anterior          4755640
Atraso por Condições Meteorológicas   4755640
```

Primeiramente, decidiu-se eliminar do dataset algumas colunas que parecem ser irrelevantes, por exemplo:

'Ano' → Pois todas as linhas tem o mesmo valor.

‘Número do Voo’ → Pois tem função apenas de identificação.

‘Número de Registro da Aeronave’ → Pois a princípio não parece existir uma correlação entre aeronave e atraso. Só existiria uma correlação baixa caso.

‘Horário Real de Partida’ → Pois já existe um horário programado para partida e uma coluna de atraso na partida.

‘Horário em que o Avião Decolou’ → Pois ele parece já estar correlacionado com o Horário programado de partida.

‘Tempo Real de Voo’ → Pois já existe uma coluna ‘Tempo de Voo Planejado’

‘Tempo no Ar’ → Pois já existe uma coluna ‘Tempo de Voo Planejado’

‘Momento do Pouso’ → Pois já parece estar correlacionado com o tempo programado de voo e com o tempo de voo planejado.

‘Tempo do Pouso até o Desembarque’ → Pois parece já estar correlacionado com o Horário Previsto para Chegada

‘Horário de Chegada Real’ → Pois parece já estar correlacionado com o Horário Previsto para Chegada

*Apesar de o ‘Desvio do Voo’, o ‘Status de Cancelamento’ e o ‘Motivo de Cancelamento’ não serem interessantes para a construção de um modelo que prevê o atraso, eles foram mantidos no dataset para um posterior tratamento, onde se removeria todos os dados dessas linhas quem indiquem que um voo foi cancelado ou desviado, pois isso pode acrescentar um fator de imprevisibilidade dos modelos.

Tratando o DataSet

Após renomear as colunas, selecionou-se apenas os dados em que ‘Status de Cancelamento’ e ‘Desvio do Voo’ fossem iguais a 0 (transforando nosso dataset apenas em voos que não foram cancelados nem desviados).

Como as colunas ‘Atraso na Partida’, ‘Atraso de Chegada’, ‘Atraso por Segurança’, ‘Atraso por Controle de Espaço Aéreo’, 'Atraso por Segurança', 'Atraso da Companhia', 'Atraso por Aeronave Anterior' e 'Atraso por Condições Meteorológicas' possuíam muitos dados ausentes, preencheu-se esses dados faltantes com as medianas. Além disso cortou-se 0.5% dos dados com os maiores e menores atrasos, visando remover outliers.

A partir do dataset, escolheu-se como variável principal a variável ‘Atraso de Chegada’, pois ela diz respeito ao atraso mais relevante.

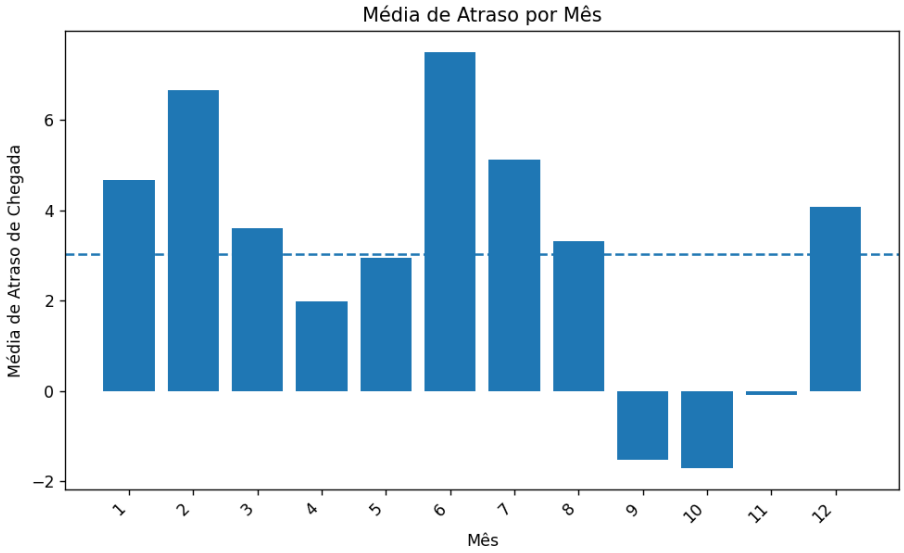
Gerando Insights Iniciais

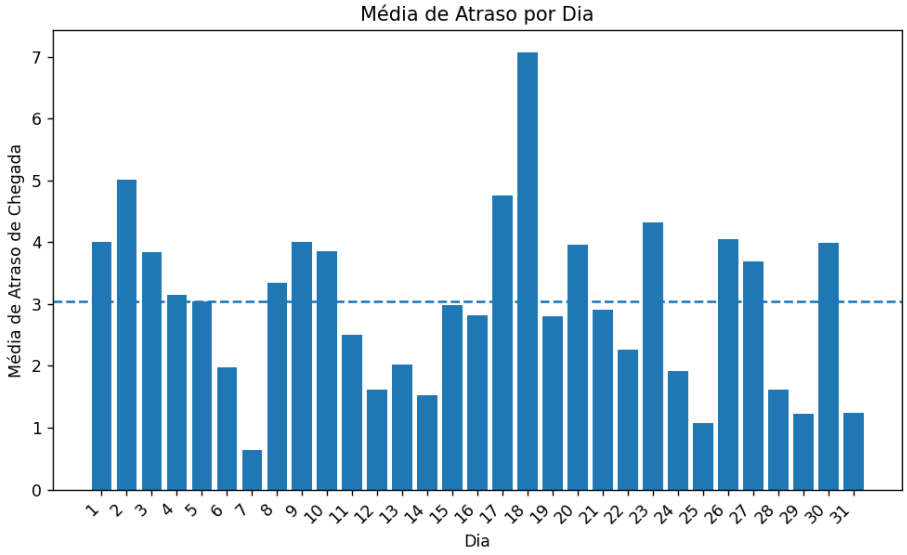
Primeiramente, calculou-se uma média global para o atraso e depois plotar para cada diferente tipo de dado em cada coluna, assim foi possível ter insights de fatores / eventos que poderiam influenciar no atraso.

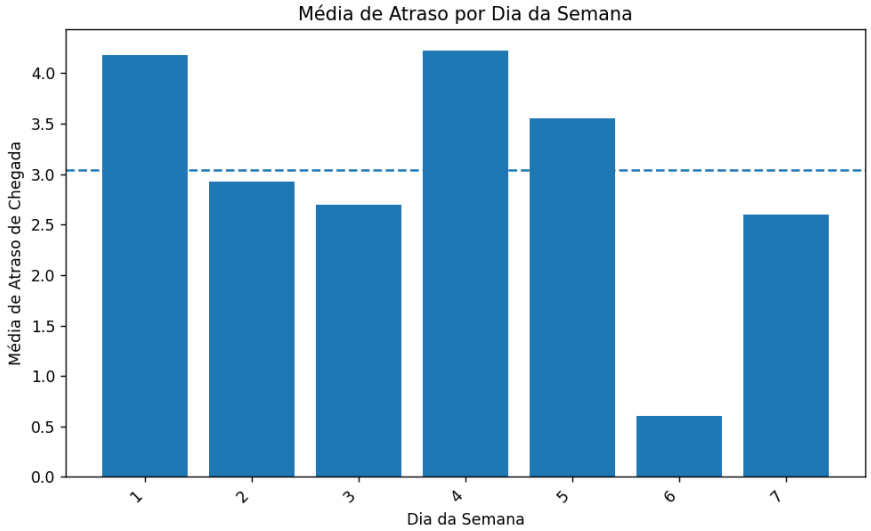
Principais Insights:

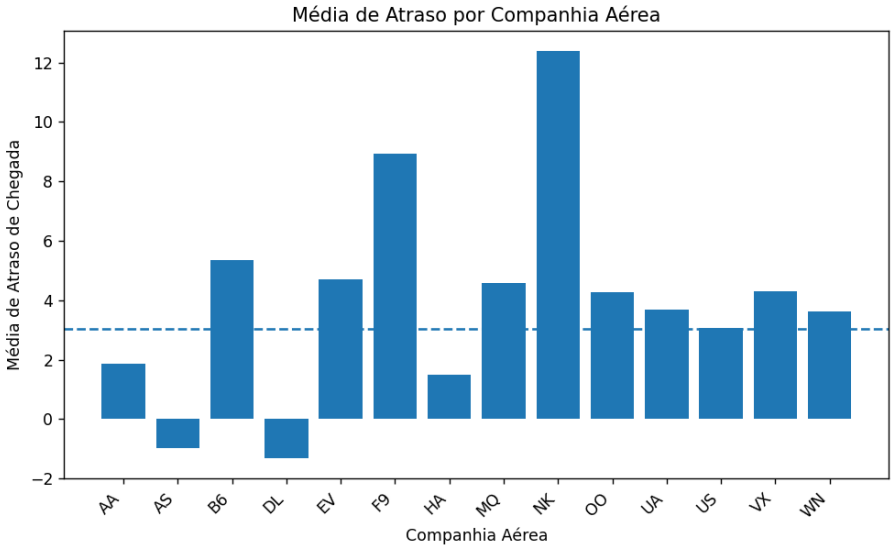
- Os meses de Janeiro, Fevereiro, Junho, Julho, Setembro e Dezembro possuem os maiores atrasos, possivelmente devido a sobrecarga do sistema aeroviário devido as férias.
- Os dias 2, 17 e 18 são os dias com maior tendência de atraso
- Os voos de segunda e quinta feira tem maior tendência de atraso enquanto os voos de sábado têm a menor tendência de atraso.
- As companhias F9 e NK são as que mais tem tendência de atraso
- Quanto maior a tendência de atraso na partida, maior a tendência de atraso na chegada
- Tempos de voos maiores levam a maior imprevisibilidade em relação a atraso, podendo chegar muito mais adiantado ou muito mais atrasado
- Quanto maior o atraso por controle do espaço aéreo, o atraso por segurança, o atraso da aeronave anterior, o atraso por condições meteorológicas ou o atraso da companhia maior a tendência de atraso para chegada.
- O horário de partida por volta das 17:00 até 20:00 tem a maior tendência de atraso, provavelmente devido ao trânsito típico desse horário.

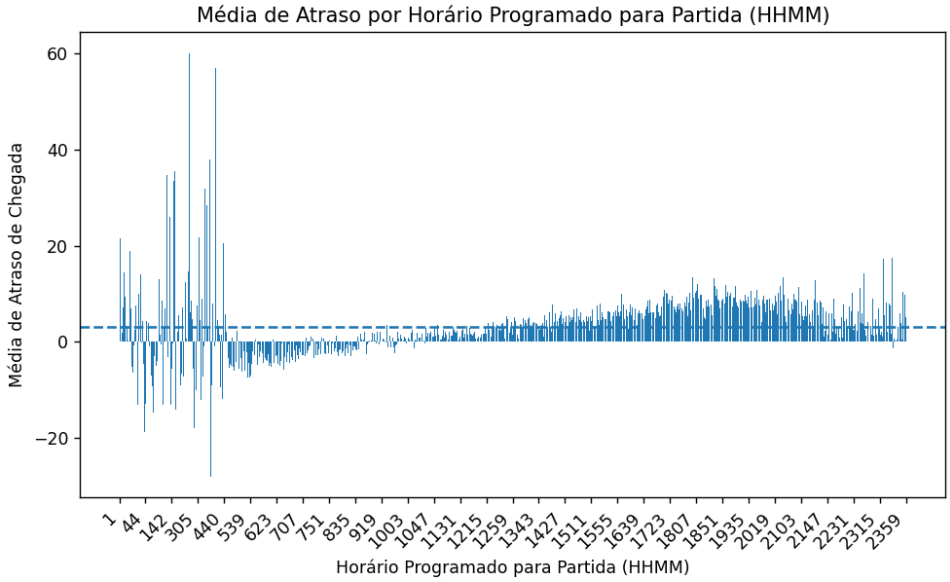
Os dados que levaram a esses insights são exibidos a seguir:

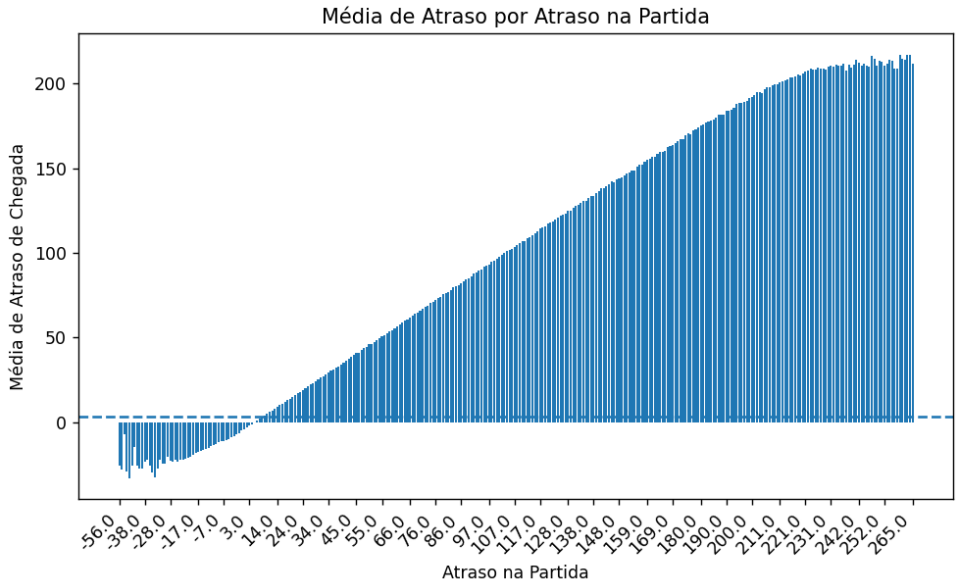


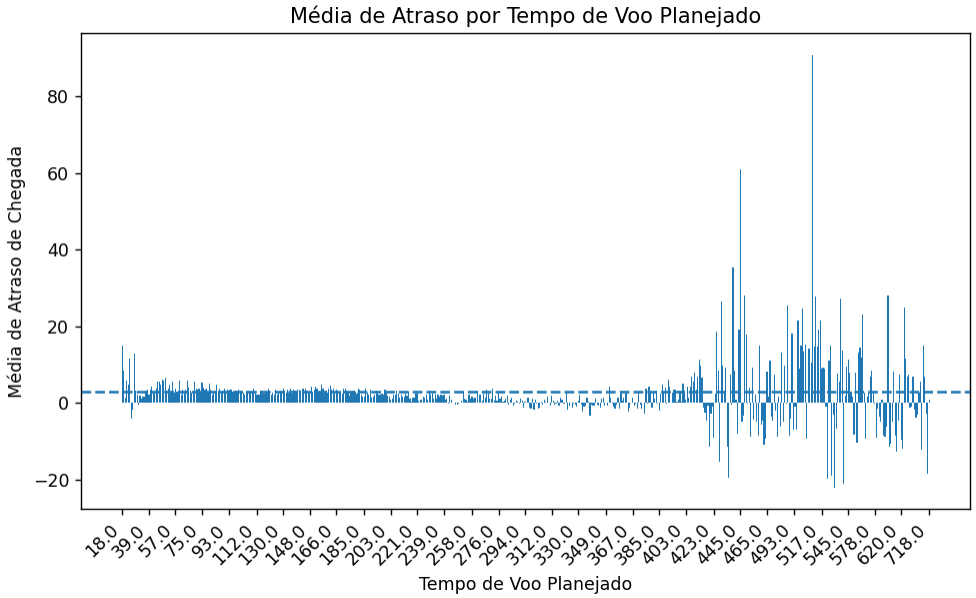


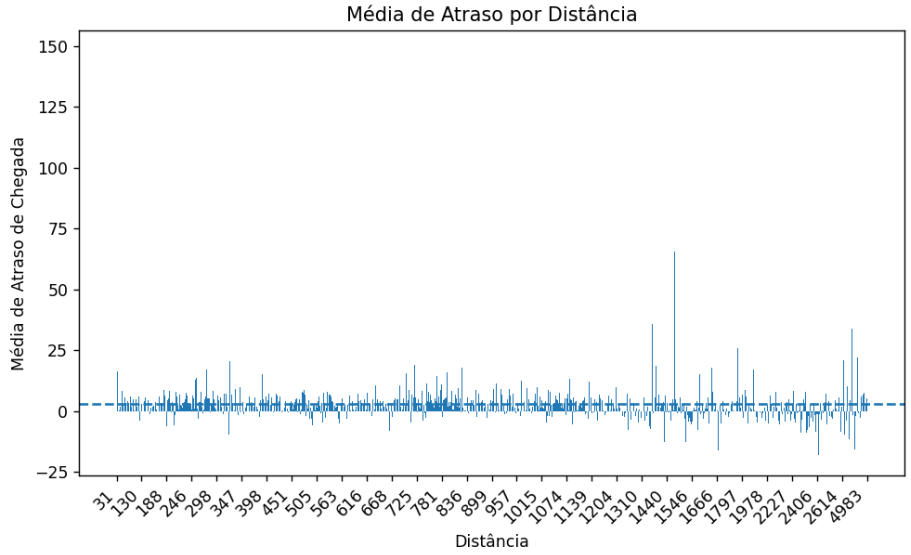

Média de Atraso por Dia da Semana
Média de Atraso de Chegada
Dia da Semana

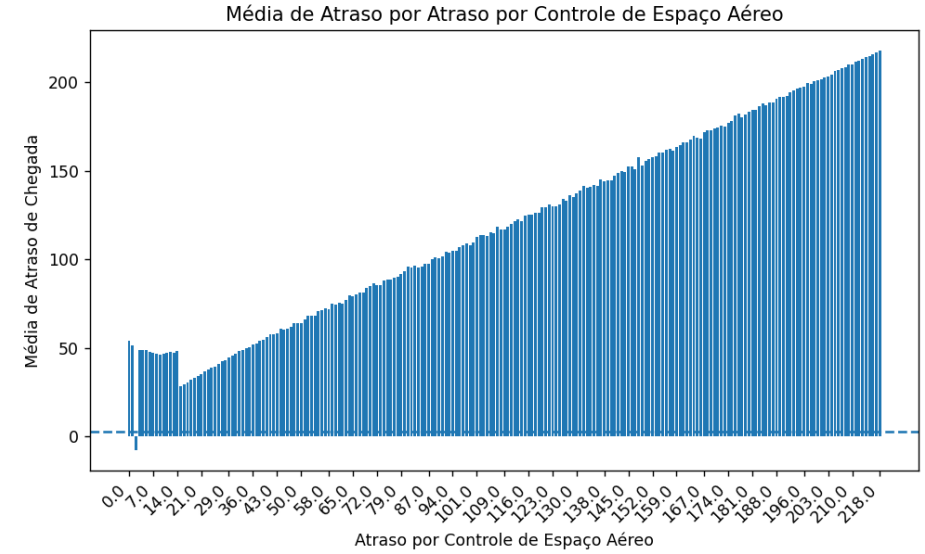

Média de Atraso por Companhia Aérea
Média de Atraso de Chegada
AA
AS
B6
DL
EV
F9
HA
MQ
NK
OO
UA
US
VX
WN
Companhia Aérea

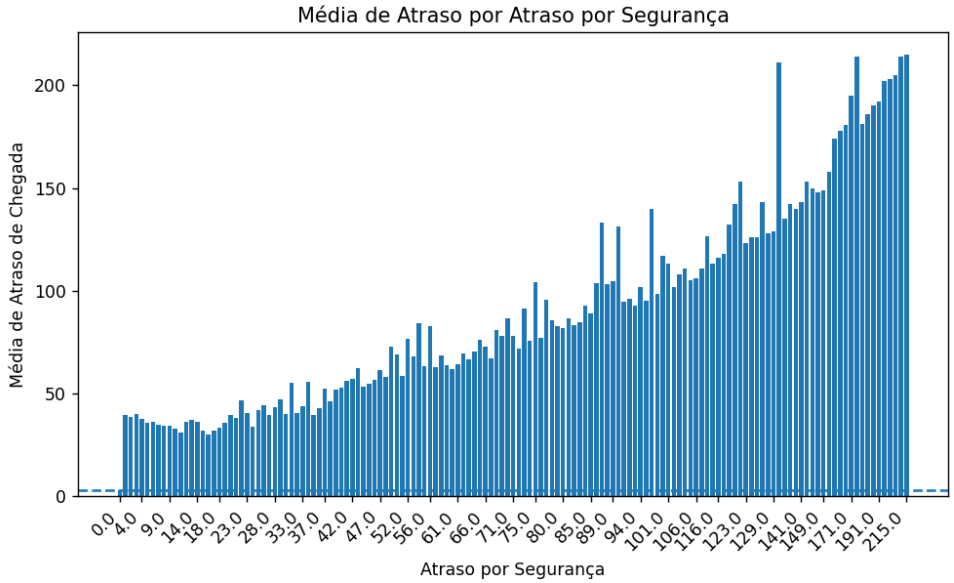

Média de Atraso por Horário Programado para Partida (HHMM)
Média de Atraso de Chegada
Horário Programado para Partida (HHMM)

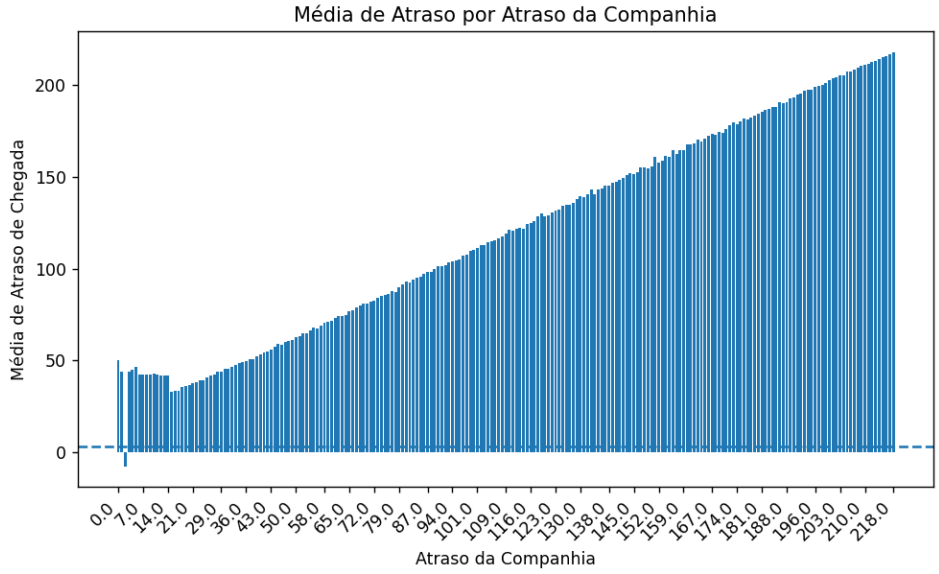

Média de Atraso por Atraso na Partida
Média de Atraso de Chegada
200
150
100
50
0
-56.0
-38.0
-28.0
-17.0
-7.0
3.0
14.0
24.0
34.0
45.0
55.0
66.0
76.0
86.0
97.0
107.0
117.0
128.0
138.0
148.0
159.0
169.0
180.0
190.0
200.0
211.0
221.0
231.0
242.0
252.0
265.0
Atraso na Partida

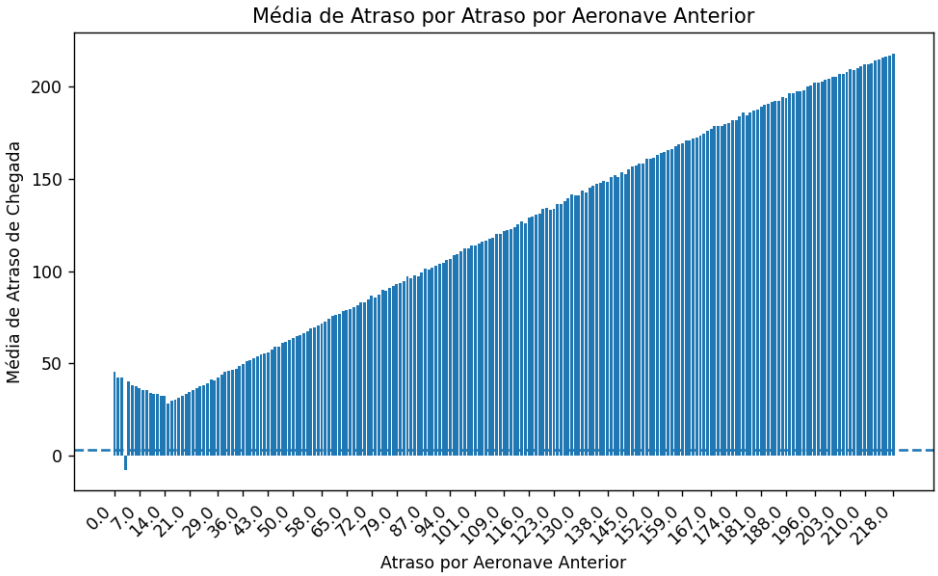

Média de Atraso por Tempo de Voo Planejado
Média de Atraso de Chegada
80
60
40
20
0
−20
18.0
39.0
57.0
75.0
93.0
112.0
130.0
148.0
166.0
185.0
203.0
221.0
239.0
258.0
276.0
294.0
312.0
330.0
349.0
367.0
385.0
403.0
423.0
445.0
465.0
493.0
517.0
545.0
578.0
620.0
718.0
Tempo de Voo Planejado

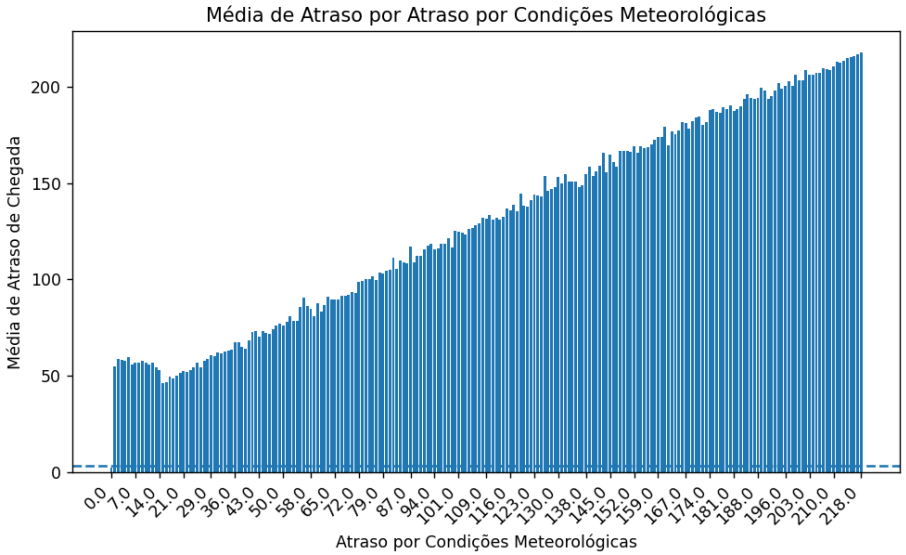

Média de Atraso por Distância
Média de Atraso de Chegada
150
125
100
75
50
25
0
−25
31
130
188
246
298
347
398
451
505
563
616
668
725
781
836
899
957
1015
1074
1139
1204
1310
1440
1546
1666
1797
1978
2227
2406
2614
4983
Distância


Média de Atraso por Atraso por Controle de Espaço Aéreo
Média de Atraso de Chegada
Atraso por Controle de Espaço Aéreo

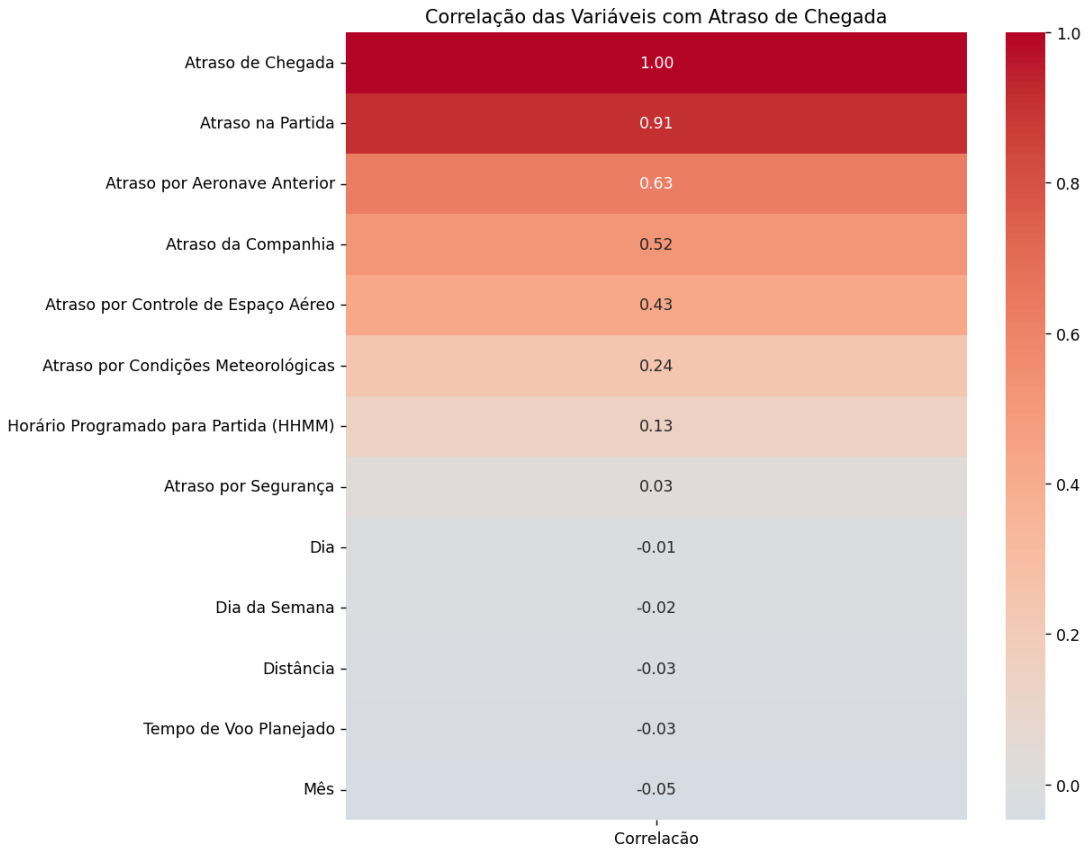

Média de Atraso por Atraso por Segurança
Média de Atraso de Chegada
Atraso por Segurança


Média de Atraso por Atraso da Companhia
Média de Atraso de Chegada
Atraso da Companhia



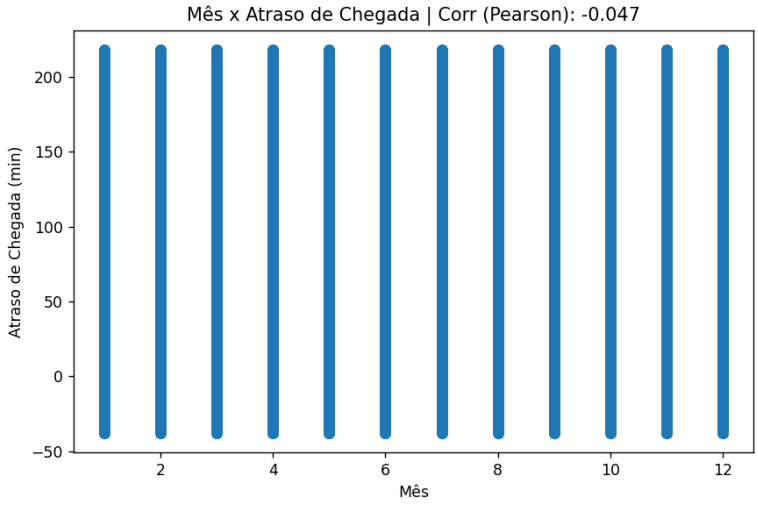


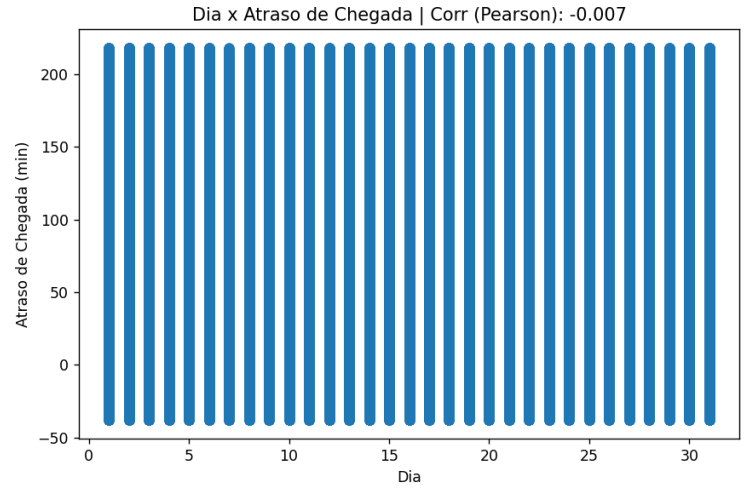
Após isso, buscou-se plotar a relação de cada coluna numérica com a coluna alvo, que no caso é a coluna '**Atraso de Chegada**', obtendo o seguinte heatmap:

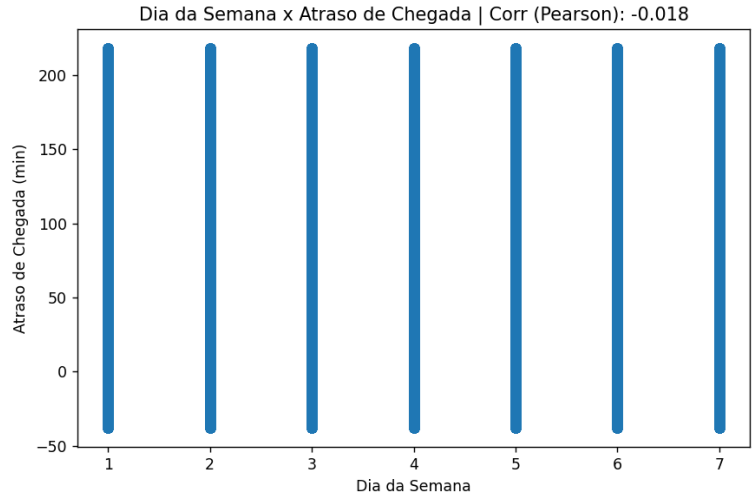


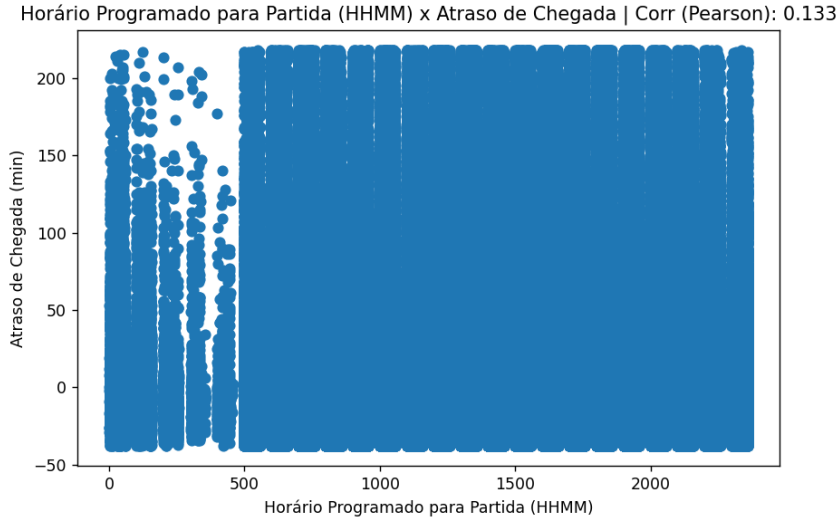
Selecionou-se as correlações que fossem maiores que 0.1 para a construção dos modelos. Isto é: 'Atraso na Partida', 'Atraso por Aeronave Anterior', 'Atraso da Companhia', 'Atraso por Controle e Espaço Aéreo', 'Atraso por Condições Meteorológicas' e 'Horário Programado para Partida (HHMM)'. Além disso também se incluiu a 'Companhia Aérea', pois pode ser um fator importante.

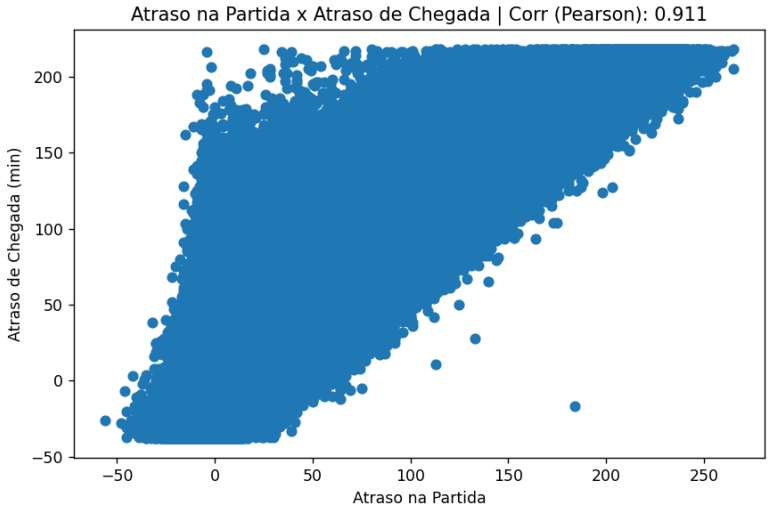
Os dados que levaram a esses insights são exibidos a seguir:

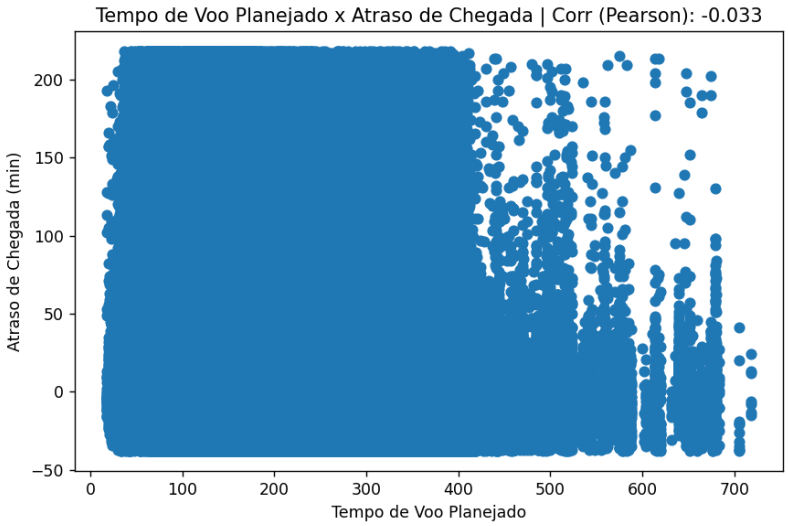


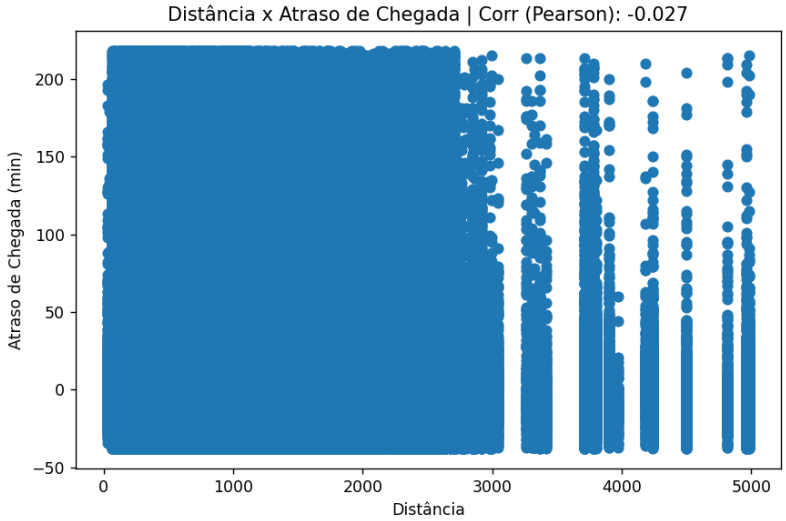

Dia x Atraso de Chegada | Corr (Pearson): -0.007
Atraso de Chegada (min)
Dia

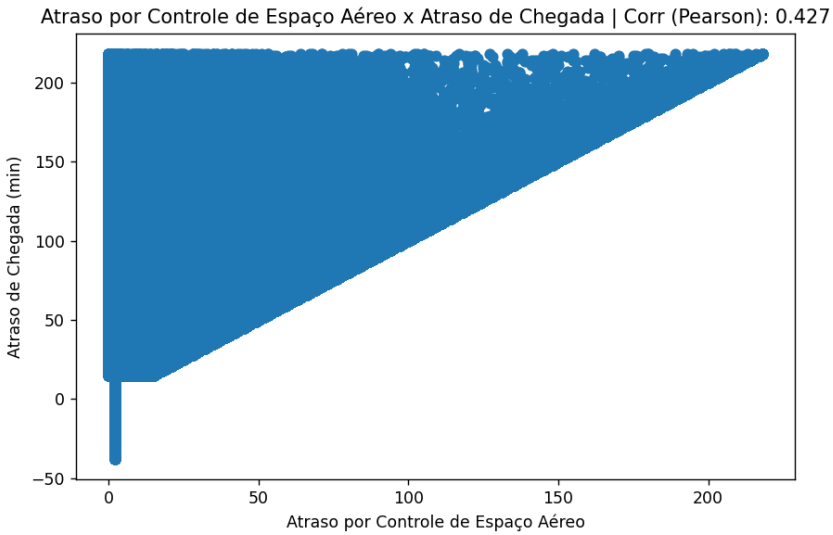

Dia da Semana x Atraso de Chegada | Corr (Pearson): -0.018
Atraso de Chegada (min)
Dia da Semana

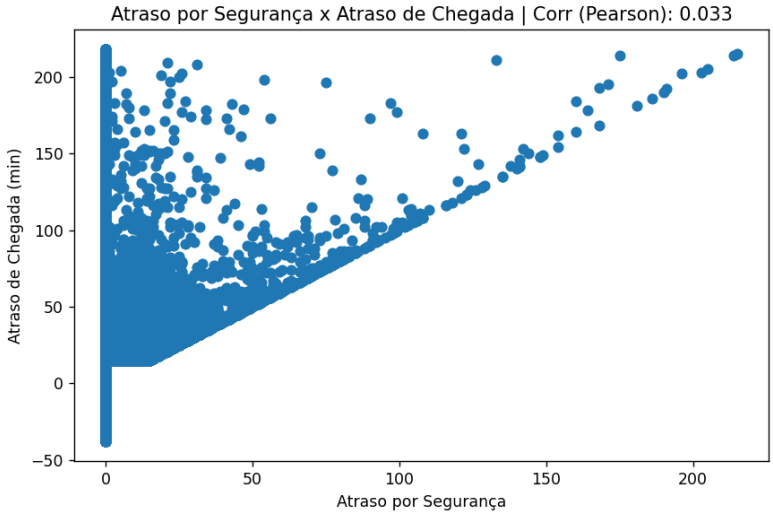

Horário Programado para Partida (HHMM) x Atraso de Chegada | Corr (Pearson): 0.133
Atraso de Chegada (min)
Horário Programado para Partida (HHMM)

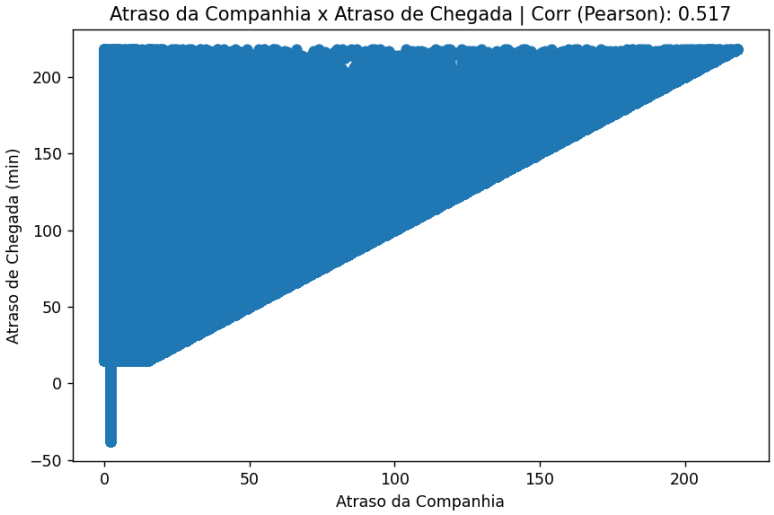

Atraso na Partida x Atraso de Chegada | Corr (Pearson): 0.911
Atraso de Chegada (min)
Atraso na Partida
200
150
100
50
0
−50
−50
0
50
100
150
200
250

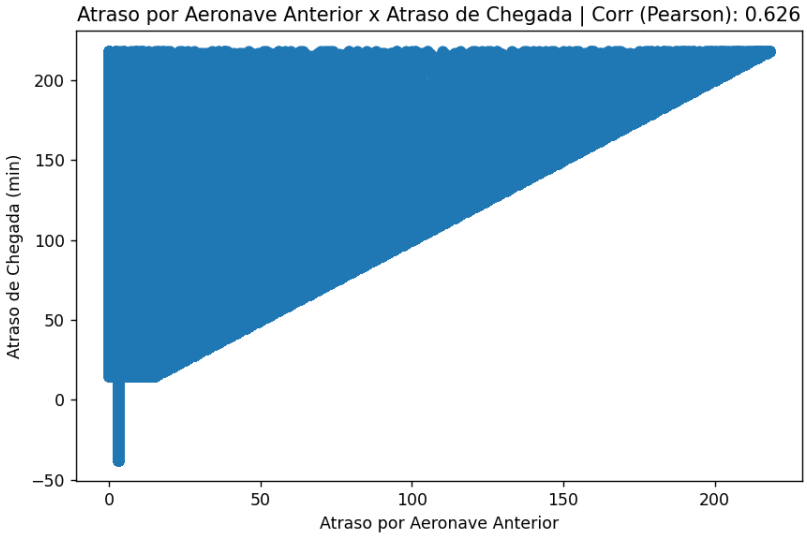

Tempo de Voo Planejado x Atraso de Chegada | Corr (Pearson): -0.033
Atraso de Chegada (min)
Tempo de Voo Planejado
200
150
100
50
0
−50
0
100
200
300
400
500
600
700

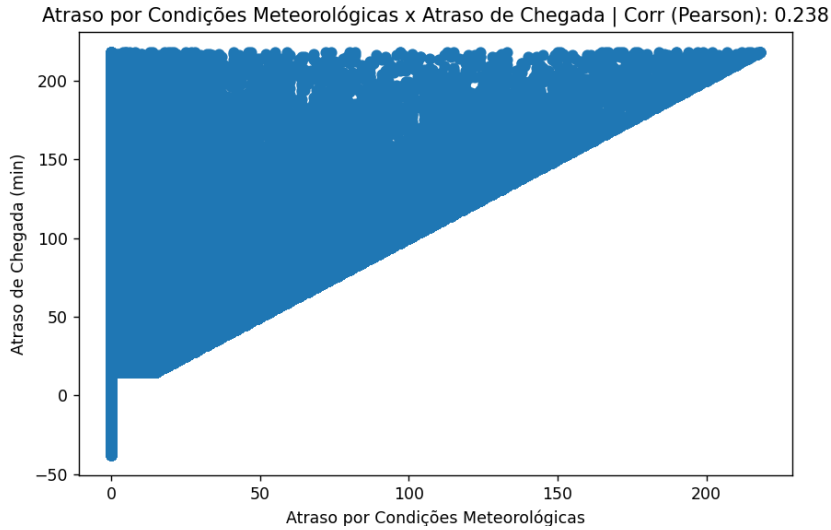

Distância x Atraso de Chegada | Corr (Pearson): -0.027
Atraso de Chegada (min)
Distância
200
150
100
50
0
−50
0
1000
2000
3000
4000
5000


Atraso por Controle de Espaço Aéreo x Atraso de Chegada | Corr (Pearson): 0.427
Atraso de Chegada (min)
Atraso por Controle de Espaço Aéreo
200
150
100
50
0
−50
0
50
100
150
200


Atraso por Segurança x Atraso de Chegada | Corr (Pearson): 0.033
Atraso de Chegada (min)
Atraso por Segurança
200
150
100
50
0
−50
0
50
100
150
200


Atraso da Companhia x Atraso de Chegada | Corr (Pearson): 0.517
Atraso de Chegada (min)
Atraso da Companhia
200
150
100
50
0
−50
0
50
100
150
200





Construção de Modelo Supervisionado de Classificação de ML

Assim, selecionou-se as colunas: 'Atraso na Partida', 'Atraso por Aeronave Anterior', 'Atraso da Companhia', 'Atraso por Controle e Espaço Aéreo', 'Atraso por Condições Meteorológicas', 'Horário Programado para Partida (HHMM)' e 'Companhia Aérea' para a construção de modelos de ML.

Assim, buscou-se criar um modelo de classificação que tentaria prever se um voo atrasaria ou não.

Para isso, usou-se os métodos: **Logistic Regression**, **Random Forest Classifier**, **Decision Tree Classifier** e **Extra Trees Classifier**. Desse dataset 2058515 amostras correspondem a classe atrasou (1) e 3599742 amostras correspondem a classe não atrasou (0), indicando dados **balanceados**.

Inserindo a Companhia:

	Logistic Regression	Random Forest	Decision Tree	Extra Trees Classifier
F1	0.762	0.756	0.751	0.753
Acurácia	0.849	0.843	0.842	0.843
Precisão	0.896	0.870	0.883	0.882
AUC	0.809	0.805	0.802	0.803

Não inserindo a Companhia

	Logistic Regression	Random Forest	Decision Tree	Extra Trees Classifier
F1	**0.759**	0.760	0.758	0.758
Acurácia	**0.850**	0.848	0.848	0.848
Precisão	**0.920**	0.898	0.902	0.902
AUC	**0.807**	0.807	0.806	0.806

Todos os modelos construídos apresentaram todos os parâmetros maiores que 0.75, indicando que todos são bons modelos.

A partir das métricas, decidiu-se que o melhor modelo é a regressão logística não utilizando a coluna da companhia aérea. Possuindo uma acurácia de 85% e uma precisão de 92%. Significando que o modelo acertou 85% das atribuições dos dados de teste.

O modelo de classificação foi exportado como 'classificação_atraso_voos.joblib'

Construção de Modelo Supervisionado de Regressão de ML

Também buscou-se trabalhar com modelo de regressão, obtendo assim o seguinte:

Inserindo a Companhia:

	Linear Regression	**Random Forest Regressor**	Decision Tree Regressor	Extra Trees Regressor
R^2	0.90	**0.92**	0.91	0.92
MAE	7.63	**6.00**	6.12	6.04
RMSE	9.56	**8.43**	8.62	8.53

Não inserindo a Companhia

	Linear Regression	Random Forest Regressor	Decision Tree	Extra Trees
R^2	0.90	0.92	0.92	0.92
MAE	7.64	6.04	6.08	6.03
RMSE	9.57	8.44	8.50	8.44

Assim, concluiu-se que o modelo Random Forest Regressor foi o melhor, devido ao maior R^2 e menor MAE / RMSE.

Assim, é possível a partir das variáveis: 'Atraso na Partida', 'Atraso por Aeronave Anterior', 'Atraso da Companhia', 'Atraso por Controle de Espaço Aéreo', 'Atraso por Condições Meteorológicas', 'Horário Programado para Partida (HHMM)' e 'Companhia Aérea' prever o atraso na chegada de um voo, o que pode ser utilizado para os sistemas aéreos em suas previsões para chegada.

Construção de Modelo Não Supervisionado de ML

Buscou-se trabalhar também com o modelo não supervisionado de PCA, para isso é necessário remover todas as colunas categóricas, removendo assim a coluna 'Companhia Aérea'.

Primeiramente, usou-se o StandartScaler() para padronizar os dados, então fez-se o cálculo de KMO, chegando nos seguintes valores:

Variável	KMO
Atraso na Partida	0.662
Atraso por Aeronave Anterior	0.313
Atraso da Companhia	0.237
Atraso por Controle de Espaço Aéreo	0.145
Atraso por Condições Meteorológicas	0.104
Horário Programado para Partida (HHMM)	0.652
Atraso de Chegada	0.614

O KMO calcula o quanto uma variável se correlaciona com as outras, as variáveis mais interessantes são aquelas com maior valor de KMO.

Então, selecionou-se as variáveis que tinham KMO maior que 0.50, pois significa que são variáveis se correlacionam bem com as outras. Assim, escolheu-se 'Atraso na Partida', 'Horário Programado para Partida (HHMM)' e 'Atraso de Chegada', sendo um conjunto de 3 variáveis.

A partir dessas 3 variáveis, criou-se uma análise de PCA com base em 2 variáveis, pois essa análise a partir de 2 variáveis explica 96% da variância:

Variável	Variância Explicada
PC1	0.65
PC2	0.31

A correlação entre PC1 e PC2 com as variáveis são:

Variável	PC1	PC2
Atraso na Partida	0.693	-0.135
Horário Programado para Partida (HHMM)	0.205	0.978
Atraso de Chegada	0.691	-0.154

Assim, PC1 está correlacionado com atraso na partida e atraso na chegada (seria um índice relacionado ao atraso), enquanto PC2 está mais correlacionado com o horário programado para partida.

Assim, foi possível reduzir a dimensionalidade e plotou-se o seguinte gráfico: