A3Data <> Pedro Blöss

Teste Técnico Cientista de Dados

Conteúdo

1. Introdução
    a. Problema e dados
    b. Método
2. Questões e hipóteses
3. Análise de inconsistências
4. Classificação dos incidentes
5. Tipo de incidente
6. Análise temporal
    a. Evolução de ocorrências e fatalidades
    b. Períodos (meses, horários, etc) de maior incidência
7. Locais e rotas
8. Análise sobre características gerais de aeronaves
    a. motor, fabricante, etc
9. Principais cenários de acidentes
10. Principais fatores
11. Conclusões

Introdução: Dados

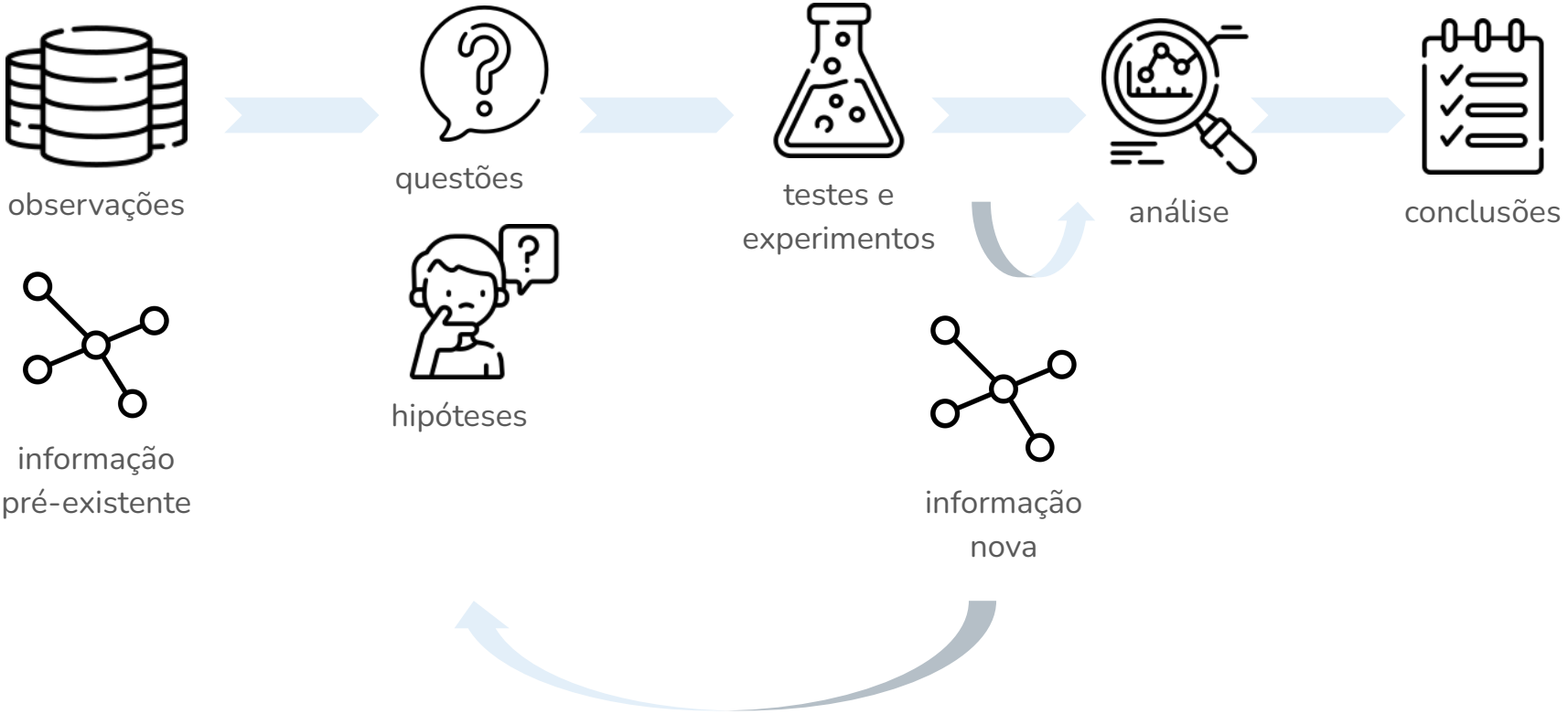
Neste teste, analisamos a base de dados "**Ocorrências Aeronáuticas na Aviação Civil Brasileira**".

A base de dados contém 2 arquivos:

- **`ocorrencia.csv`:** informações e características sobre ocorrências de acidentes.
  - Ex: tipo, localidade, dia, horário, etc.
  - Dimensões: 2027 registros e 19 variáveis.
- **`aeronave.csv`**: informações e características sobre aeronaves.
  - Ex: qtd motores, peso máx., categoria, origem, destino, etc.
  - Dimensões: 2043 registros e 22 variáveis.

As bases possuem as variáveis "`codigo_ocorrencia`" e "`dia_extracao`" em comum.

Metodologia científica



Introdução: Objetivos

Responder às questões:

- Identificar padrões principais sobre ocorrências
- Quais são os **principais fatores contribuintes** para acidentes, e para acidentes graves?
- Quais são os principais insights que tiramos sobre os dados?

Questões e hipóteses

- Quais são as **principais causas de acidentes**? (e de acidentes mais graves?)
- Há **horários** e **locais mais propícios** a ocorrências? Imagina-se que existem rotas mais perigosas, por exemplo.
- Quais ocorrências são investigadas e possuem recomendações?
- Existem **tipos de naves** com **mais chances de acidentes**? (De um determinado fabricante, peso, etc)
- Supostamente, naves **mais antigas** são **menos seguras,** pois supostamente tecnologias mais modernas são mais robustas e têm maiores padrões de qualidade.
- Provavelmente, há menos **incidentes mais graves**. Quais são as **principais diferenças** com **acidentes não graves**?
- Será que a data da ocorrência influencia nos dados? Por exemplo, ocorrências antigas tem mais chances de já serem investigadas?

Análise de inconsistências

Possíveis inconsistências:

- Dados faltantes
  - 10 colunas possuem dados faltantes inicialmente.
  - Há colunas com alta taxa de faltantes
    - (ex: `saida_pista` com 87,46%)
- "Falsos não faltantes" (Ex: "n/a", "não atribuído", "-", etc)
  - Encontrados: "***" e "****".
  - Então, na realidade temos 22 colunas com faltantes (Fig. 3).
- Dados incongruentes

	coluna	Qtd faltantes	% faltantes
0	saida_pista	1787	87.469408
1	quantidade_fatalidades	1688	82.623593
2	dia_publicacao	1042	51.003426
3	relatorio_publicado	1042	51.003426
4	status_investigacao	209	10.230054
5	numero_relatorio	209	10.230054
6	quantidade_assentos	18	0.881057
7	quantidade_motores	9	0.440529
8	ano_fabricacao	4	0.195791
9	aerodromo	3	0.146843

Fig. 1: Valores faltantes

	uf	aerodromo	sera_investigada	c
0	RO	SJOG	***	
2	RO	****	SIM	
3	RR	****	SIM	
4	RS	****	SIM	
5	GO	****	***	
...	...	...	...	

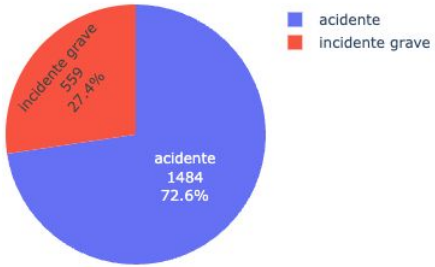
Fig. 2: Falsos não faltantes

	coluna	Qtd faltantes	% faltantes
0	saida_pista	1787	87.469408
1	quantidade_fatalidades	1688	82.623593
2	aerodromo	1229	60.156632
3	destino_voo	1204	58.932942
4	origem_voo	1110	54.331865
5	dia_publicacao	1042	51.003426
6	relatorio_publicado	1042	51.003426
7	status_investigacao	209	10.230054
8	numero_relatorio	209	10.230054
9	sera_investigada	206	10.083211
10	fabricante	110	5.384239
11	nivel_dano	69	3.377386
12	tipo_motor	28	1.370534
13	tipo_operacao	26	1.272638
14	categoria_aviacao	25	1.223691
15	quantidade_assentos	18	0.881057
16	modelo	15	0.734214
17	quantidade_motores	9	0.440529
18	categoria_registro	9	0.440529
19	equipamento	5	0.244738
20	ano_fabricacao	4	0.195791
21	uf	2	0.097895

Fig. 3: "Novos" valores faltantes

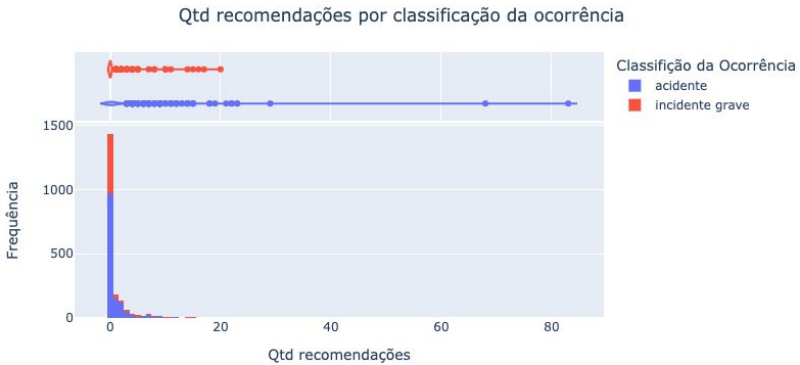
Classificação de incidente

Fig. 1



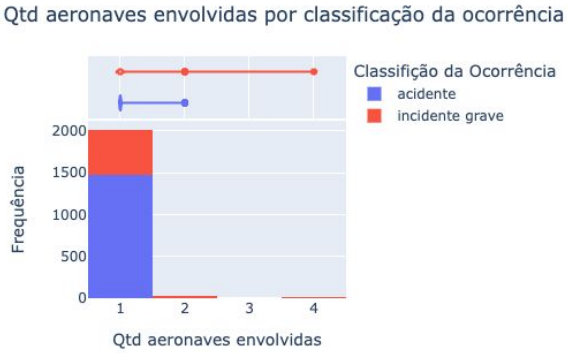
De fato, a **proporção de incidentes graves é pequena** (cerca de 1 a cada 4 ocorrências).

Fig. 2



Em média, a **quantidade de recomendações** é **menor** para **incidentes graves**.

Fig. 3



Casos com **muitas aeronaves** envolvidas geralmente ocorrem para **incidentes graves.**

Tipo de incidente

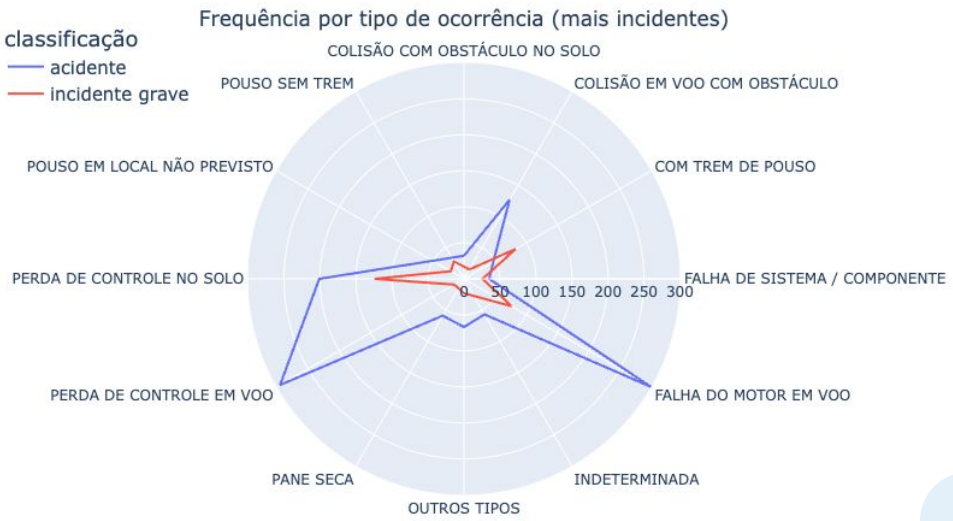


Fig. 1. Frequência de tipo de ocorrência, para cada classificação.

- **Acidentes não graves** tem na maioria incidentes por:
  - perda de controle
  - falha no motor
- Já **incidentes graves** têm como maiores tipos de acidentes:
  - perda de controle no solo
  - problemas com trem de pouso
  - falha no motor em voo

Entende-se que os seguintes **fatores** são **importantes** contribuintes para acidentes:

- **motor**
- **perda de controle em solo e em voo**
- **trem de pouso**

Período da ocorrência: Evolução das ocorrências e fatalidades

Fig. 1

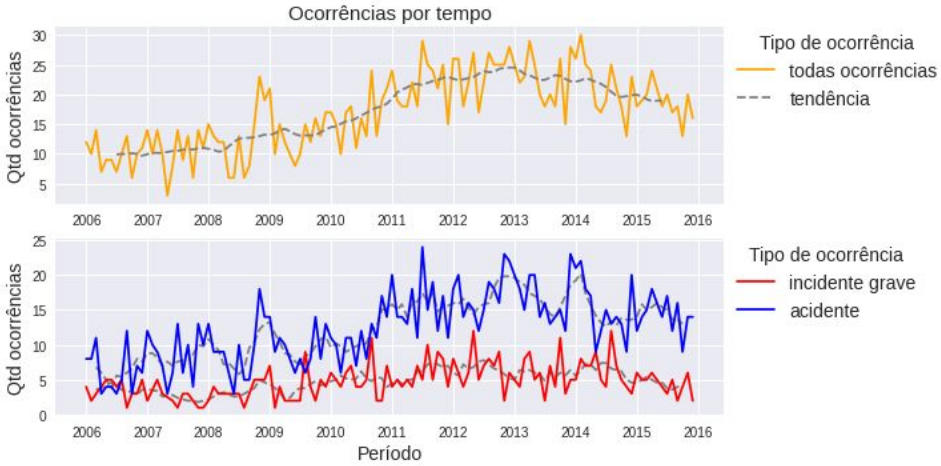
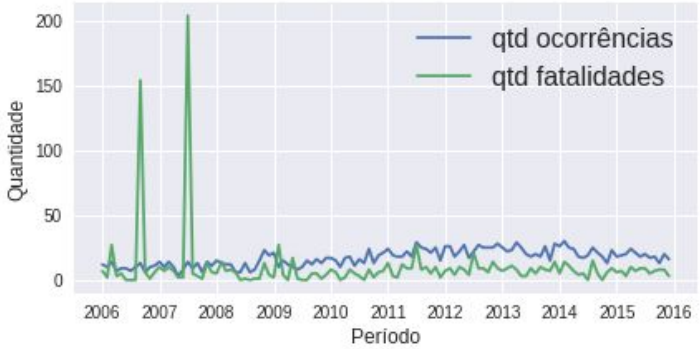


Fig. 2



A qtd de ocorrências teve crescimento até 2013, e depois leve decaimento.

Outliers: 2 eventos de taxa de fatalidade muito alta ocorreram, em 09/2006 e 07/2007.

Referências:

https://g1.globo.com/mato-grosso/noticia/2016/09/acidente-com-aviao-da-gol-que-matou-154-pessoas-completa-10-anos.html

https://www.cnnbrasil.com.br/nacional/acidente-da-tam-em-congonhas-completa-15-anos-veja-o-que-mudou-na-aviacao-brasileira/

Período da ocorrência: Evolução por mês do ano e por horário

Fig. 1

No geral, há menos ocorrências em Junho.

Incidentes graves ocorreram mais em **Agosto**.

Acidentes aconteceram mais no **final e início do ano.**

Fig. 2

Na madrugada (00 - 06) o movimento é menor, consequentemente há menos ocorrências.

Há **picos de ocorrências às 13h e 20h.**

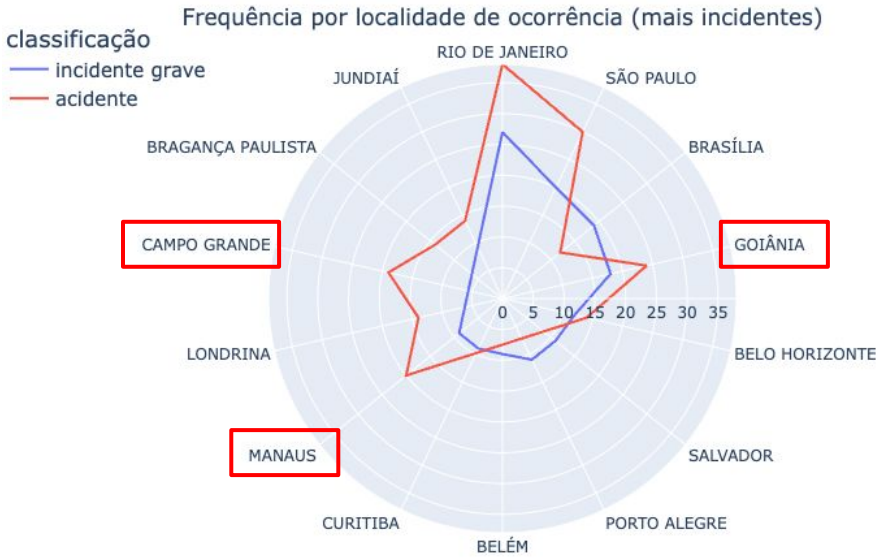
Podem ser horários tipicamente associados a períodos de refeição.

Locais e rotas

Lista de aeroportos do Brasil por movimento (2019)

Posição	Class	Aeroporto	Passageiros pagos (2019)[1]	Gestão	Cidade servida	Unidade federativa
1	–	Aeroporto Internacional de São Paulo-Guarulhos	42 248 207	GRU Airport	São Paulo	São Paulo
2	–	Aeroporto de São Paulo-Congonhas	22 281 896	Infraero	São Paulo	São Paulo
3	–	Aeroporto Internacional de Brasília	16 569 442	Inframérica	Brasília	Distrito Federal
4	–	Aeroporto Internacional Tom Jobim-Rio Galeão	13 518 783	Rio Galeão	Rio de Janeiro	Rio de Janeiro
5	–	Aeroporto Internacional de Belo Horizonte-Confins	10 734 359	CCR Aeroportos	Belo Horizonte	Minas Gerais
6	▲ 1	Aeroporto Internacional de Viracopos-Campinas	10 199 171	Viracopos	Complexo Metropolitano Expandido	São Paulo
7	▼ 1	Aeroporto do Rio de Janeiro-Santos Dumont	8 933 777	Infraero	Rio de Janeiro	Rio de Janeiro
8	–	Aeroporto Internacional do Recife-Guararapes	8 638 608	Aena Internacional	Recife	Pernambuco
9	–	Aeroporto Internacional de Porto Alegre-Salgado Filho	8 106 869	Fraport	Porto Alegre	Rio Grande do Sul
10	–	Aeroporto Internacional de Salvador-Dep. Luís Eduardo Magalhães	7 351 020	Vinci Airports	Salvador	Bahia

Ref.: https://pt.wikipedia.org/wiki/Lista_de_aeroportos_do_Brasil_por_movimento

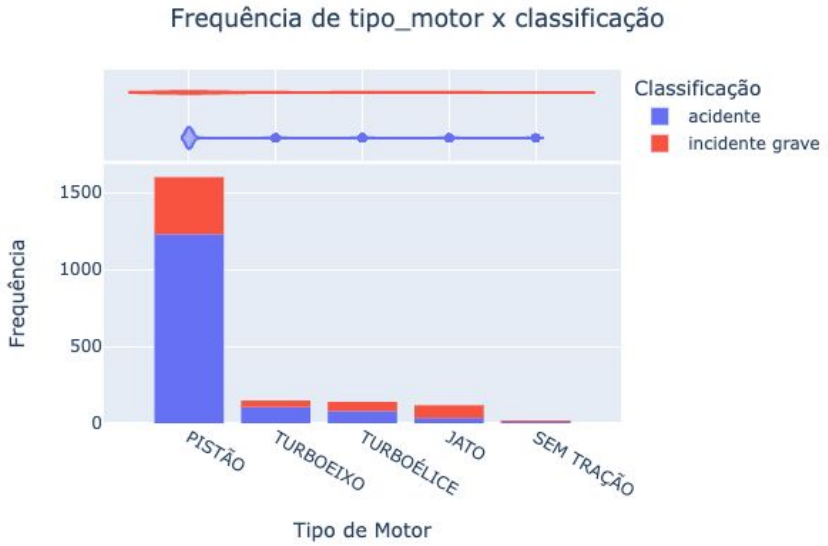


Naturalmente, **aeroportos mais movimentados** tem maior espaço amostral para acidentes.

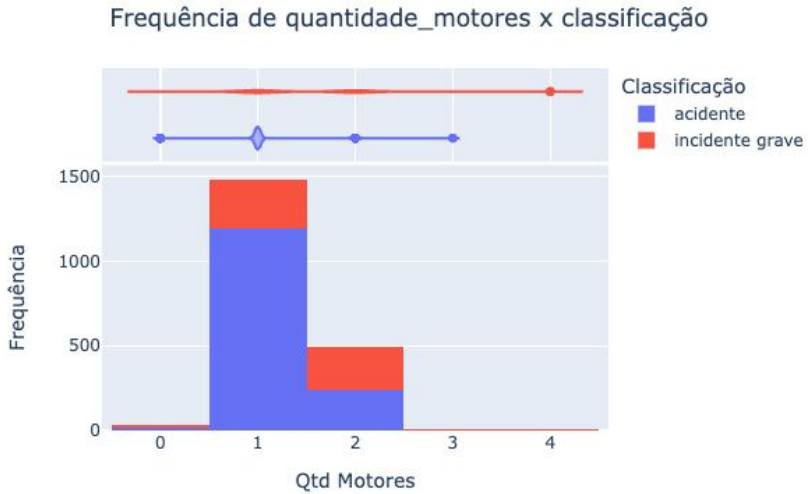
Então, espera-se que existam mais acidentes para localidades como:
São Paulo, Brasília, Rio de Janeiro, Recife, BH, Porto Alegre.

Localidades com aeroportos **não tão movimentados** tiveram **altas frequências** de ocorrências:
Manaus, Campo Grande, Goiânia.

Características de Motor

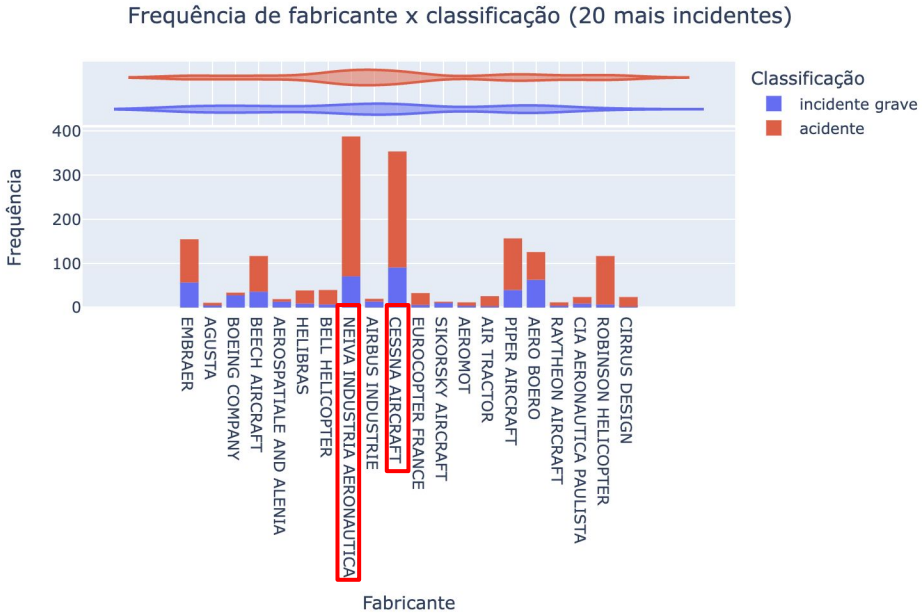


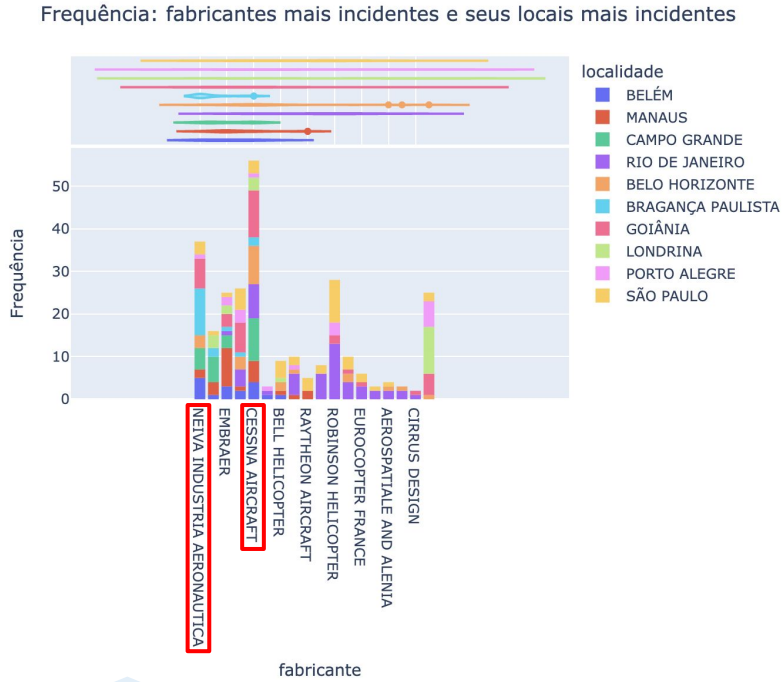
Não há diferenças estatisticamente significantes.
O tipo de motor "Pistão" é o mais frequente, e também é o que tem mais acidentes graves.



Para **acidentes não graves**, **geralmente há 1** motor.
Para **acidentes graves**, há uma **distribuição similar** de **1 e 2** (qtd motores).

Características de Fabricante





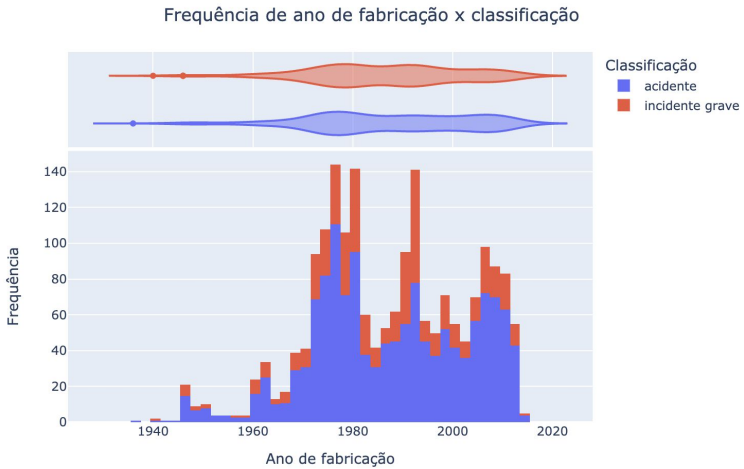
A distribuição de fabricantes por classificações de ocorrência é bastante similar.

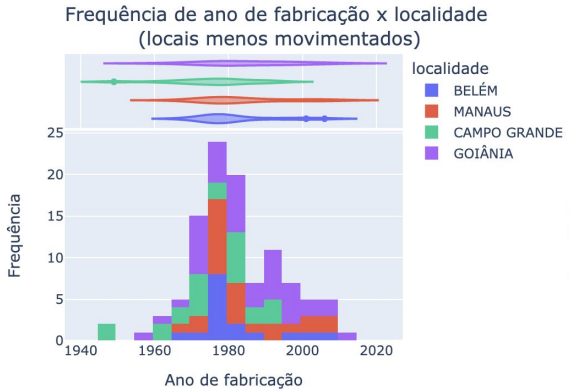
Há fabricantes com maior frequência de ocorrência, mas podem ser fabricantes mais comuns.

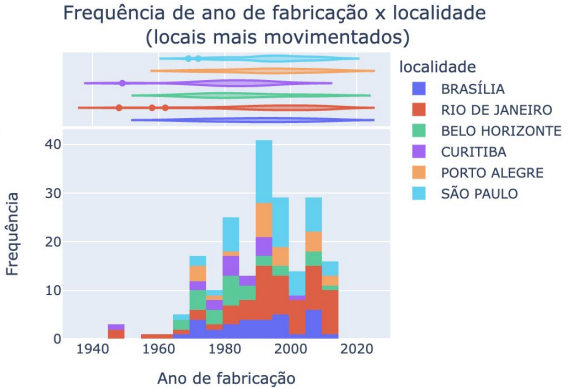
Para identificar se os fabricantes são mais comuns, vemos os locais.

Locais com alta taxa de acidentes e baixo tráfego aéreo relativo são frequentes nos fabricantes suspeitos: (Goiânia, Campo Grande, Manaus)

Ano de fabricação



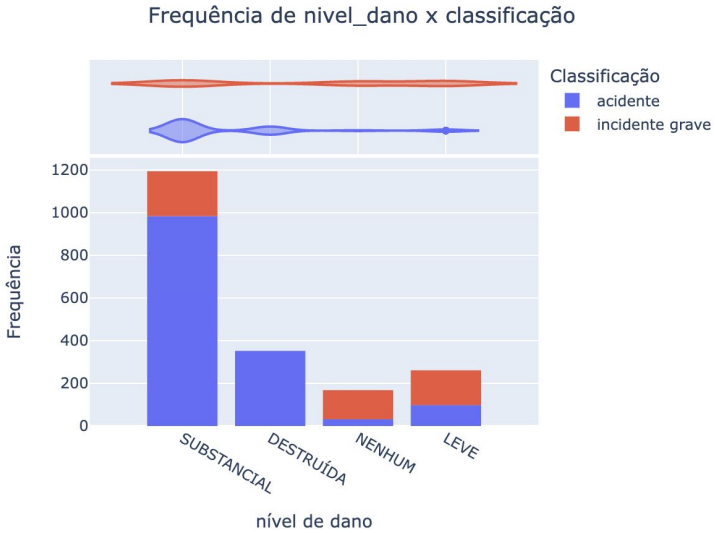


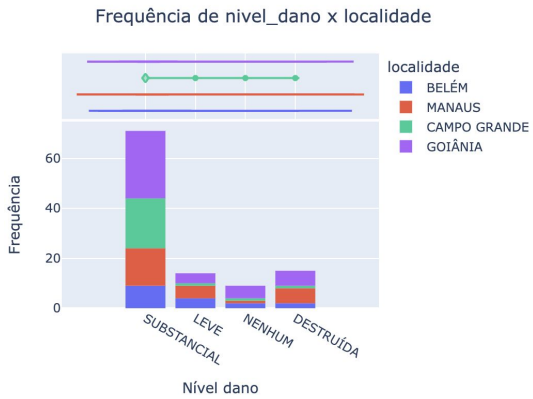


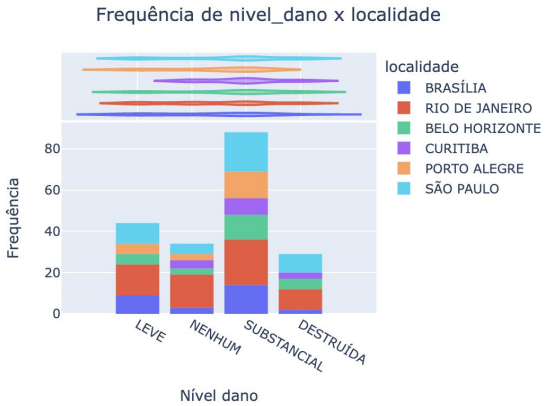
Em média, há **mais ocorrências** para **aeronaves com anos de fabricação mais antigos**, nos anos 80 e nos anos 90.

Locais **sem muito tráfego** e **alta taxa de ocorrências** possuem **aeronaves mais antigas**, comparando às idades de aeronaves de locais mais movimentados (SP, RJ, ...)

Nível de dano







Contraintuitivamente, **incidentes graves** possuem mais **níveis baixos de danificação**.

Locais **sem muito tráfego** e **alta taxa de ocorrências** possuem uma **proporção maior** de **dano "substancial" e "destruída"**, comparando às idades de aeronaves de locais mais movimentados (SP, RJ, ...)

Conclusões e comentários

Potenciais pontos de atenção:

- Problemas:
  - em motor
  - perda de controle (em solo e em voo)
  - trem de pouso
- Período:
  - meses de extremidades (Janeiro/Dezembro),
  - horários 13h e 20h.
- Locais e rotas: Goiânia, Manaus, Campo Grande
- Motor: quantidade 2
- Fabricantes suspeitos
- Ano de fabricação antigo

Voos com **características suspeitas** de alta **probabilidade de acidentes** podem ser **monitorados** com cautela.

Agradecimentos

Obrigado!

Pedro Blöss Braga

Referências:

- Magalhães, N. Marcos & De Limpa, Antonio C., "Noções de Probabilidade e Estatística", edusp.
- Bolfarine, Heleno & Sandoval, Monica C., "Introdução à Inferência Estatística", SBM.