Desafio Técnico Cientista de Dados: Análise do Dataset Ocorrências Aeronáuticas Da Aviação Civil Brasileira

Fabrício Rodrigues de Souza

Descrição do desafio

Este desafio consiste na exploração, interpretação, análise e descrição do dataset **Ocorrências Aeronáuticas da Aviação Civil Brasileira**, bem como na sugestão de medidas a serem tomadas a partir da análise destes dados. O dataset consiste das tabelas citadas abaixo que contém dados de diferentes aspectos de incidentes registrados pela ANAC (Agência Nacional de Aviação Civil) durante os anos de 2010 a 2020. O código fonte, o arquivo de dependências necessárias para que ele execute corretamente e cópias das imagens utilizadas neste relatório estão disponíveis neste repositorio.

- ocorrencia: tabela principal do dataset, lista as ocorrências registradas e se conecta com outras quatro tabelas através de chaves estrangeiras;
- ocorrencia_tipo: tabela que se conecta à tabela principal e descreve o tipo da ocorrência;
- aeronave: tabela que se conecta à tabela principal e descreve as aeronaves de cada ocorrência;
- fator_contribuinte: tabela que se conecta à tabela principal e descreve os fatores que levaram à ocorrência;
- recomendacao: tabela que se conecta à tabela principal e descreve as recomendações dadas para cada ocorrência;
- reportes_2011_2020: tabela que descreve dados extras das ocorrências, não se conecta à nenhuma outra tabela.

Análises

O dataset é extremamente rico e complexo, apresentando diversos dados diferentes sobre os incidentes. O primeiro passo tomado foi a análise da modelagem do dataset, através da observação criteriosa da estrutura dos dados, disponível na Figura 1. Toda a análise foi feita utilizando a linguagem de programação Python no ambiente de desenvolvimento Jupyter Notebook. Para manipular as tabelas utilizamos a biblioteca Pandas. Podemos ver que as tabelas "ocorrencia" e "aeronave" são as que contêm mais dados sobre os incidentes. Por isso foi decidido começar a exploração por elas.

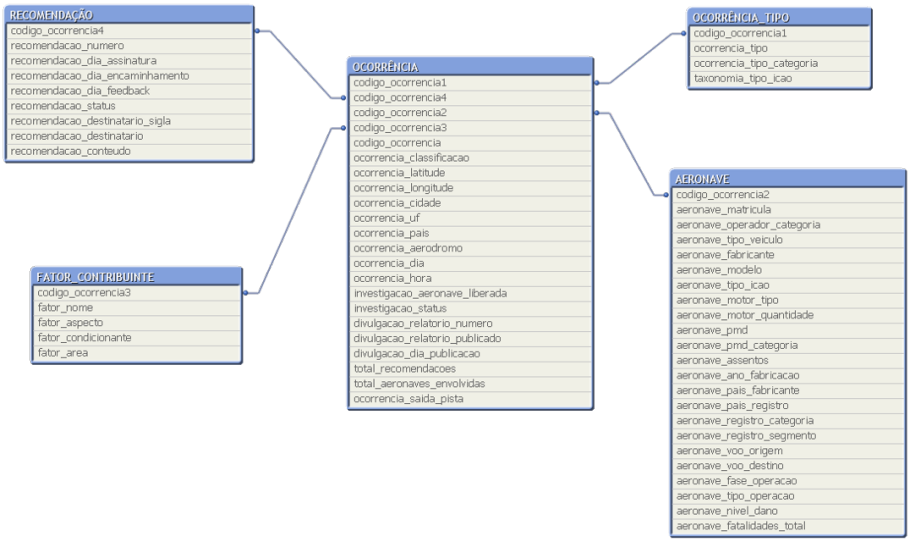


Figure 1. Estrutura dos dados.

Distribuição geográfica dos incidentes

Estão presentes para, aproximadamente, 72% dos incidentes os dados de latitude e longitude. Tendo em mente que a porcentagem de dados faltantes não é grande foram eliminados todos os dados que representavam valores faltantes. Os valores faltantes para estes dados são 'NaN' e as strings '***' e '******'.

O próximo desafio encontrado na análise dos dados geográficos foi a falta de padrão para a representação dos dados. Como os dados são adicionados na base por pessoas, é natural que algumas entradas estejam fora do padrão. As representações de latitude e longitude podem ser feitas em diversos formatos, o mais presente na base de dados é o formato decimal. Neste formato, latitudes podem variar de -90 a 90 e longitudes de -180 a 80. Através da aplicação de um filtro nas colunas de latitude e longitude foram removidos todos os dados que estavam fora deste padrão. Além disso foram removidos dados que não estavam nos intervalos de -40 a 10 para latitudes e -180 a -20 para longitudes. Fizemos este recorte para remover *outliers* que dificultavam as análises.

Com os dados limpos e formatados foi possível realizar diversos tipos de análises. Neste momento a algumas suposições foram feitas, "Os dados fornecidos são de uma empresa responsável por investigar todos os incidentes no local em que ele acontece", "Toda região representada pelas latitudes e longitudes é um espaço de terra contínuo, sem corpos d'água ou montanhas intrespassáveis", "Os times de investigação se deslocam por veículos terrestres", "Os times de investigação podem ser instalados em qualquer posição geográfica". Também foi suposto que a empresa em questão levantou o problema de que o tempo de chegada aos locais de incidentes tem sido muito longo, porque os times de investigação não estão bem posicionados.

Para resolver este problema utilizamos os dados de incidentes que já aconteceram para posicionarmos os times de investigação. Foi utilizada uma técnica não-supervisionada de agrupamento dos dados, O algoritmo *K-Means*. Este algoritmo agrega os dados baseado na média dos valores de cada ponto. É importante lembrar que só devemos utilizá-lo em cenários onde a média é um bom descritor dos pontos. Outro ponto que é necessário levantar é a medida de distância utilizada pelo algoritmo. Tradicionalmente ele utiliza a distância euclidiana, mas quando estamos utilizando pontos caracterizados por latitude e longitude é necessário que utilizemos a métrica de *haversine* para calcular esta distância, pois ela considera a curvatura da Terra em seus cálculos. Utilizamos este algoritmo porque ele consegue capturar bem o centro geográfico dos grupos e assim, podemos colocar os times nestes centros geográficos e minimizar a distância máxima percorrida pelos times de investigação. A Figura 2 apresenta a solução encontrada para o caso onde a empresa informou que possui 100 times de investigação.

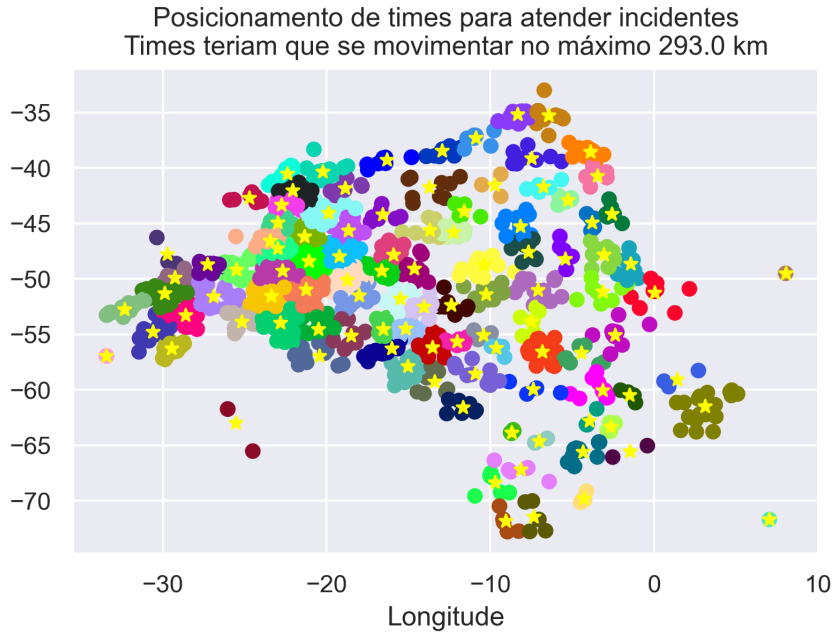


Figure 2. Distribuição geográfica dos times.

Cada grupo diferente está colorido com sua própria cor e os centros destes grupos são representados pelas estrelas amarelas. Nesta configuração os times teriam que se locomover no máximo 293 km para atender o incidente mais longe. Podemos ver na imagem que existem alguns pontos *outliers*. Para removê-los podemos entrar em contato com a empresa e questionar se aquele ponto realmente representa um acontecimento que é estatisticamente improvável de acontecer, por exemplo se é um ponto onde não existem rotas comerciais passando, ou se ele deve ser

considerado na análise. Se removermos estes *outliers* a distribuição dos times permitiria que eles se deslocassem ainda menos.

Tendo em vista este problema, exploramos outra face dele. Vamos considerar que a empresa deseja encontrar o número de times de investigação necessário para que a distância máxima percorrida fique abaixo de um limiar pré-definido. Para resolver este problema utilizamos novamente o algoritmo *K-Means*, mas nesta abordagem o número de grupos foi sendo aumentado à medida que a distância máxima de qualquer centro para todos os pontos que ele está responsável fosse maior que o limiar. A Figura 3 apresenta os resultados para esta análise. Neste caso definimos o limiar como 250 km. Podemos ver que o número necessário de times de investigação cresceu consideravelmente, de 100 times se deslocando no máximo 293 km para 143 times se deslocando no máximo 250 km. Este comportamento é esperado, porque uma quantidade menor de times deve cobrir um espaço maior.

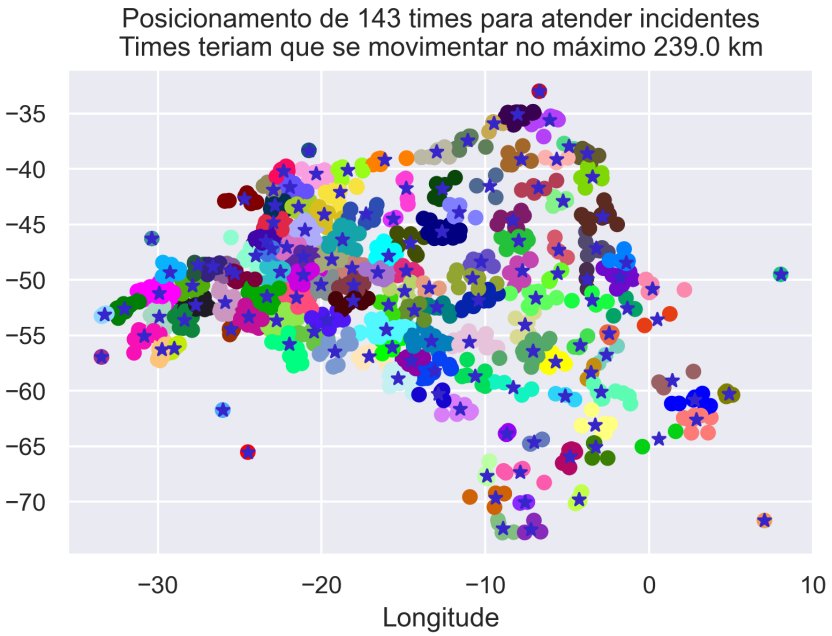
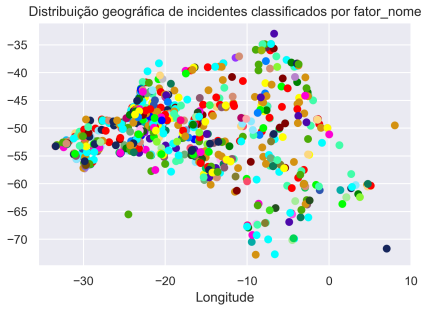
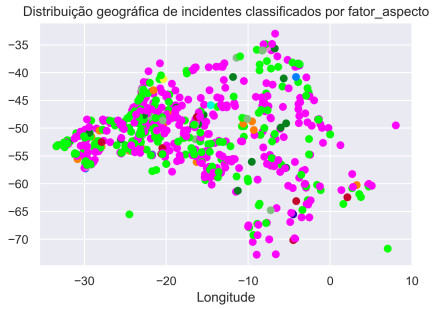


Figure 3. Distribuição geográfica dos times para percorrer no máximo 250 km.

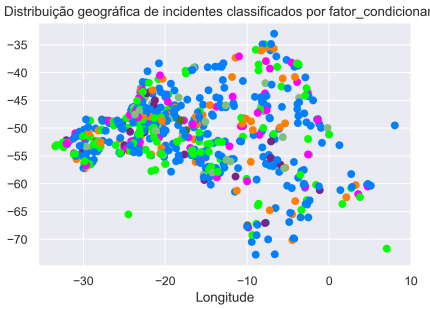
Levantamos a hipótese ”Quando aeronaves passam por esta área existe alguma interferência que leva à falha de instrumentos”. Para isso desenhamos todos os pontos e os colorimos de acordo com os quatro possíveis diferentes itens listados na tabela fator_contribuinte. A hipótese levantada se mostrou falsa, ao considerarmos todos os itens, fator_nome, fator_aspecto, fator_condicionante e fator_area, nenhum padrão foi demonstrado. As Figuras 4a, 4b, 4c, 4d apresentam como os pontos foram classificados. Se existisse algum padrão nas distribuições poderíamos inferir que algum evento se repetia naquela região e que a tripulação da aeronave deveria ser avisada ou treinada sobre esse possível evento.



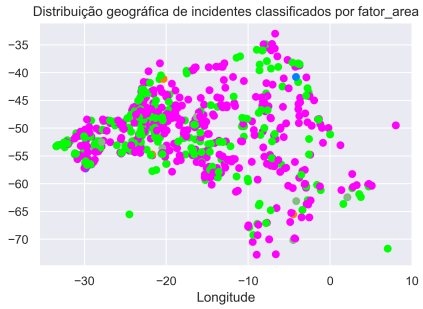
(a) Agrupamento por fator_nome.



(b) Agrupamento por fator_aspecto.



(c) Agrupamento por fator_condicionante



(d) Agrupamento por fator_area

Figure 4. Agrupamentos de fatores

Fatalidades

Fatalidades infelizmente acontecem durante operações de aeronaves, devemos fazer o máximo possível para que elas sejam evitadas, isso inclui investigar os motivos das fatalidades e tomar medidas para que minimizem sua frequência.

A tabela aeronave apresenta dados de todas as aeronaves envolvidas em incidentes. As aeronaves estão classificadas de acordo com três componentes, aeronave_tipo_veiculo, aeronave_motor_tipo e aeronave_motor_quantidade. Separamos cada incidente por cada valor de cada um dos três tipos citados. Foram removidos dados nulos e faltantes. Os dados nulos e faltantes neste caso são casos onde os campos possuem valores 'NaN' ou a string '***'. A Figura 5 apresenta a quantidade de fatalidades por combinação de tipo de aeronave. A Figura 6 apresenta a quantidade de incidentes por combinação de tipo de aeronave.

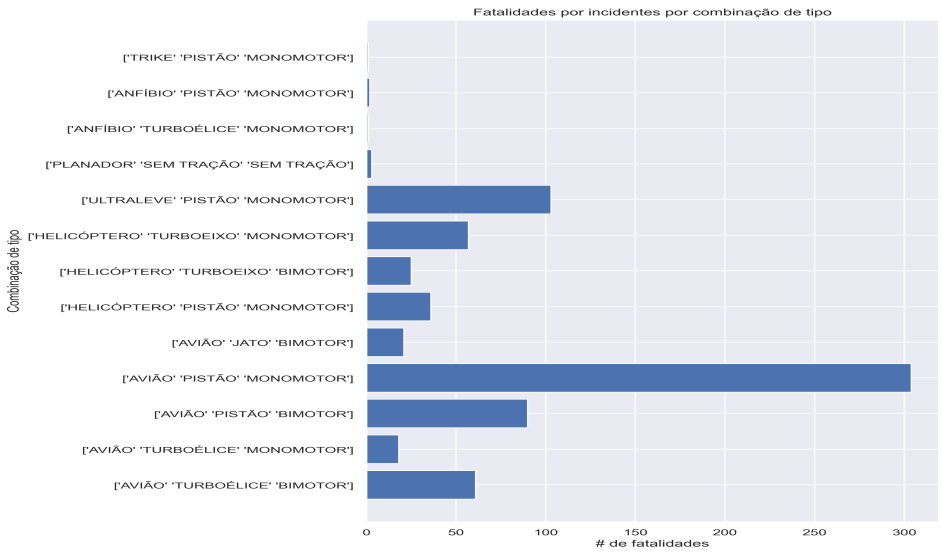


Figure 5. Fatalidades por combinação de tipo de aeronave.

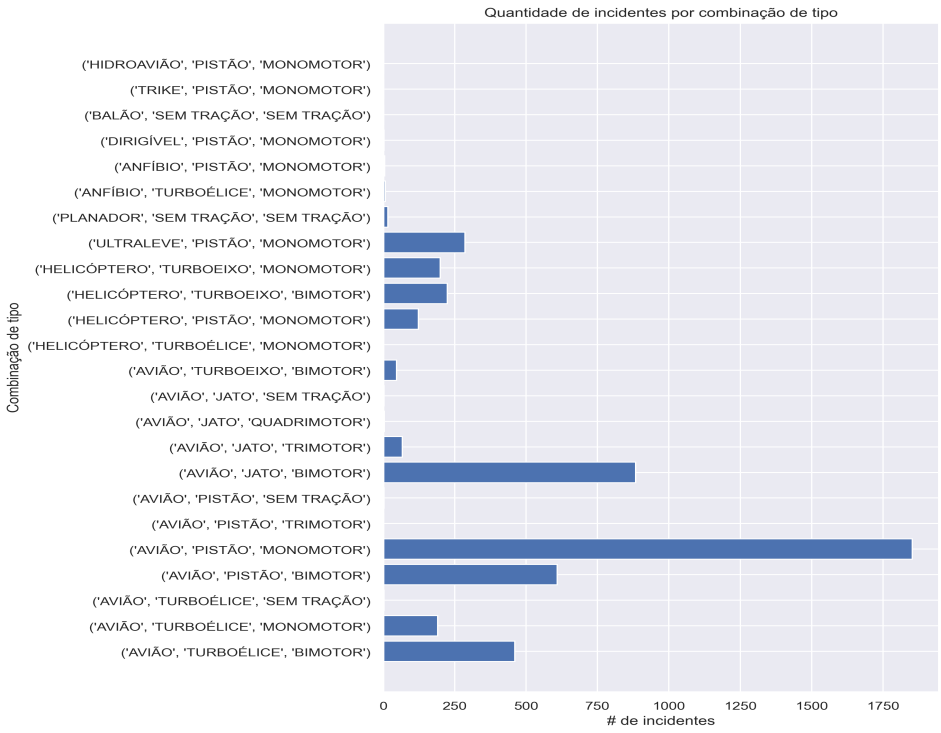


Figure 6. Quantidade de incidentes por combinação de tipo de aeronave.

Se partirmos do pressuposto que devemos fiscalizar ou criar oportunidades de ensino para tripulações e profissionais que atuam na combinação com maior quantidade de fatalidades, deveríamos direcionar nossos esforços para os aviões de pistão e monomotores. No entanto, é necessário que normalizemos os nossos dados. Faz sentido pensar que combinações de aeronaves que executam mais voos são mais propensas a incidentes com fatalidades. Ao normalizarmos o número de fatalidades pela quantidade de incidentes daquela combinação de aeronave obtemos os dados apresentados na Figura 7. Podemos ver que em 100% dos casos de incidentes com aeronaves do tipo trike de Pistão e Monomotor, temos uma fatalidade. O que significa que se a empresa em questão deseja evitar fatalidades seria interessante realizar uma campanha de conscientização e treinamento para os proprietários de veículos desta combinação. E é interessante sugerir que pessoas que lidam com todas as outras combinações de veículos também façam cursos recorrentes sobre segurança das suas aeronaves.

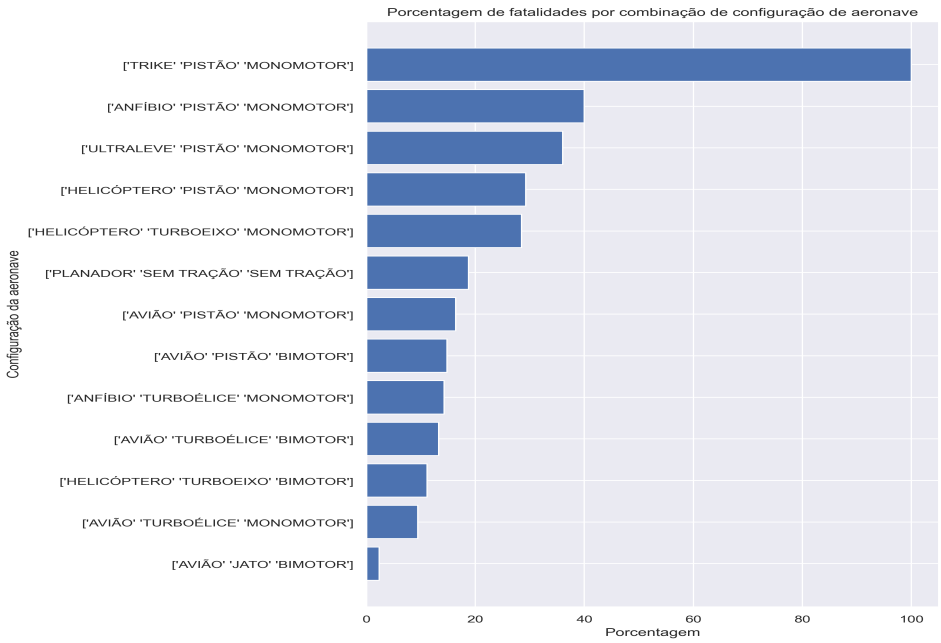


Figure 7. Porcentagem de fatalidades por combinação de tipo de aeronave.

Incidentes por fase de operação

O voo possui diversas fases de operação, que incluem a movimentação da aeronave do hangar, taxi, decolagem, cruzeiro, pouso e muitas outras. A hipótese levantada é de que existem fases onde incidentes ocorrem com mais frequência, logo devemos prestar atenção redobrada nessas etapas para que estes incidentes possam ser evitados. A Figura 8 apresenta os dados de incidentes por fase de operação. Podemos ver que a maioria dos incidentes acontecem nas fases de pouso, decolagem e cruzeiro, bem como em fases imediatamente adjacentes ao pouso e decolagem. Durante o cruzeiro é extremamente difícil prestar auxílio físico à aeronave, logo, quando consideramos esse tipo de auxílio que desejamos prestar, o foco da atenção deve ser maior durante o pouso e a decolagem.

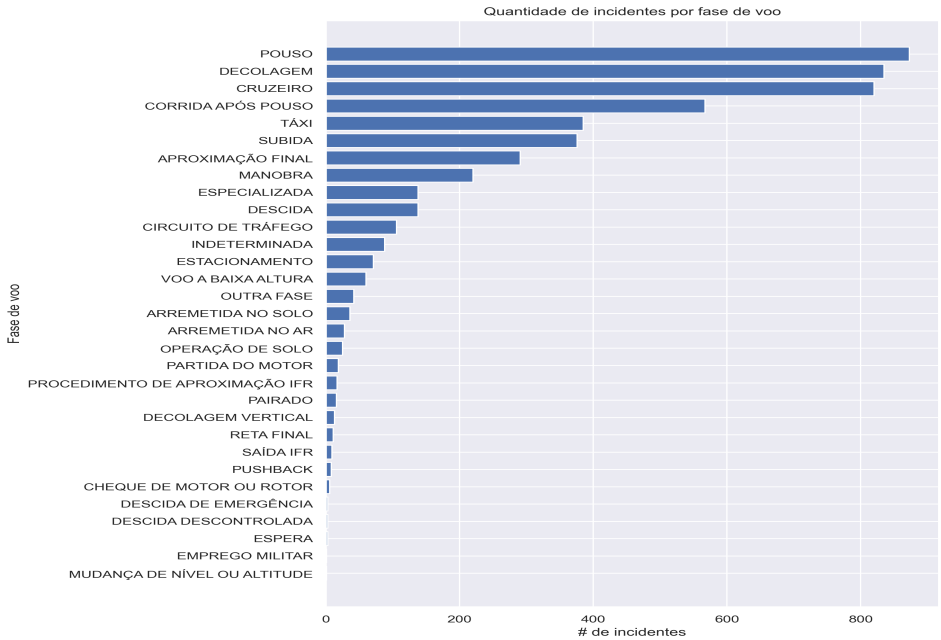


Figure 8. Quantidade de incidentes por fase de voo.

Incidentes por operadoras de voo

Uma informação de extrema importância é a quantidade de incidentes que cada operadora de voo está causando. O público geral merece estar ciente deste número, para que possa tomar decisões bem informadas quando desejar voar. A tabela reportes_2011_2020 apresenta os incidentes que aconteceram por operadora. Por questões de tempo e de visualização, alguns recortes da tabela foram feitos nesta análise. Apenas três operadoras foram consideradas, LATAM (TAM), Azul (AZU) e GOL (GLO), e apenas incidentes do ano de 2020 foram considerados. Apesar deste recorte é possível repetir a análise para todos os outros anos e operadoras. As operadoras foram escolhidas porque são três das maiores operadoras que atuam no Brasil e porque apresentavam seus nomes sem erros na base de dados. Outras operadoras apresentavam nomes inconstantes e com caracteres que geram erros durante a leitura.

Os dados presentes na tabela reportes_2011_2020 são apenas dos incidentes que aconteceram, não é correto comparar apenas estes dados absolutos entre as operadoras. Apesar de serem operadoras de porte parecido, é possível que em algum período uma das operadoras tenha realizado um número de voos consideravelmente menor, precisamos normalizar os dados em relação à quantidade de voos realizados pela operadora.

A ANAC disponibiliza em seu portal registros de todos os voos que aconteceram desde 2000 até 2022. Obtivemos os dados de todos os voos que aconteceram nos meses de 2020 e normalizámos a quantidade de incidentes por operadora. A Figura 9 apresenta estes dados. Podemos ver que no mês de maio de 2020 aconteceram uma quantidade significativamente maior que nos outros meses do ano. É interessante que a empresa em questão investigue o motivo desta diferença e se possível identificar maneiras de evitar que a quantidade de incidentes atinja este nível novamente.

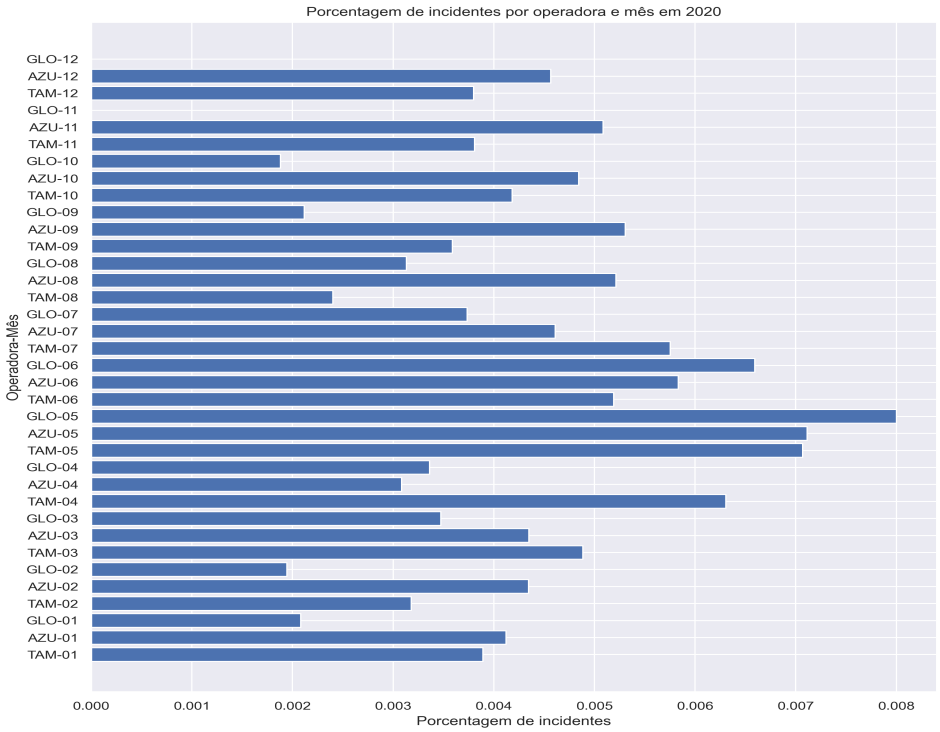


Figure 9. Porcentagem de incidentes por mês por operadora em 2020.

Conclusões

Ao longo deste documento foram descritas diversas análises e ações que podem ser feitas sobre os dados obtidos. A falta de padronização dos dados de latitude e longitude tomou um tempo considerável para contornar, além disso dados que lidam com nomes das operadoras precisam ser tratados profundamente para que possam ser utilizados, o que também levaria bastante tempo. O conjunto de dados ainda pode ser explorado. Podemos considerar os dados de recomendações feitas em relação aos incidentes e também podemos incluir os tipos de ocorrências nas análises. Devido à natureza completamente independente das ocorrências é difícil pensar em combinações de dados que podem ser utilizados para prever algo. Se possuíssemos dados históricos das aeronaves seria possível tentar prever momentos onde elas falhariam, mas não possuímos esses dados.