Desafio Técnico - A3Data

Cientista de Dados

Versão	Data	Nome	Modificação
1.0	11/04/2022	Isac Carvalho	Estudo/ Análise Exploratória

Visão Geral

O desafio técnico para a vaga de Cientista de Dados da A3Data consiste na exploração da base de dados "Ocorrências Aeronáuticas na Aviação Civil Brasileira" dos dados abertos do governo.

Fonte de Dados:

https://dados.gov.br/dataset/ocorrencias-aeronauticas-da-aviacao-civil-brasileira

Neste desafio deverá conter:

1. Apresentação em formato PDF com:
    a. Apresentação do desafio
    b. Explicação do processo utilizado
    c. Hipóteses levantadas
    d. Análise exploratória
    e. Conclusões e insights gerados
2. Constar o código no GitHub

Cenário Atual

A base de dados de ocorrências aeronáuticas é gerenciada pelo Centro de Investigação e Prevenção de Acidentes Aeronáuticos (CENIPA). Constam nesta base de dados as ocorrências aeronáuticas notificadas ao CENIPA nos anos de 2012 a 2021 que ocorreram em solo brasileiro.

Dentre as informações disponíveis estão os dados sobre as aeronaves envolvidas, fatalidades, local, data, horário dos eventos e informações taxonômicas típicas das investigações de acidentes (AIG). São resguardadas a privacidade de pessoas físicas/jurídicas envolvidas conforme previsto pela Lei de Acesso à Informação (Lei n° 12.527, de 18 de novembro de 2011).

O conjunto de dados

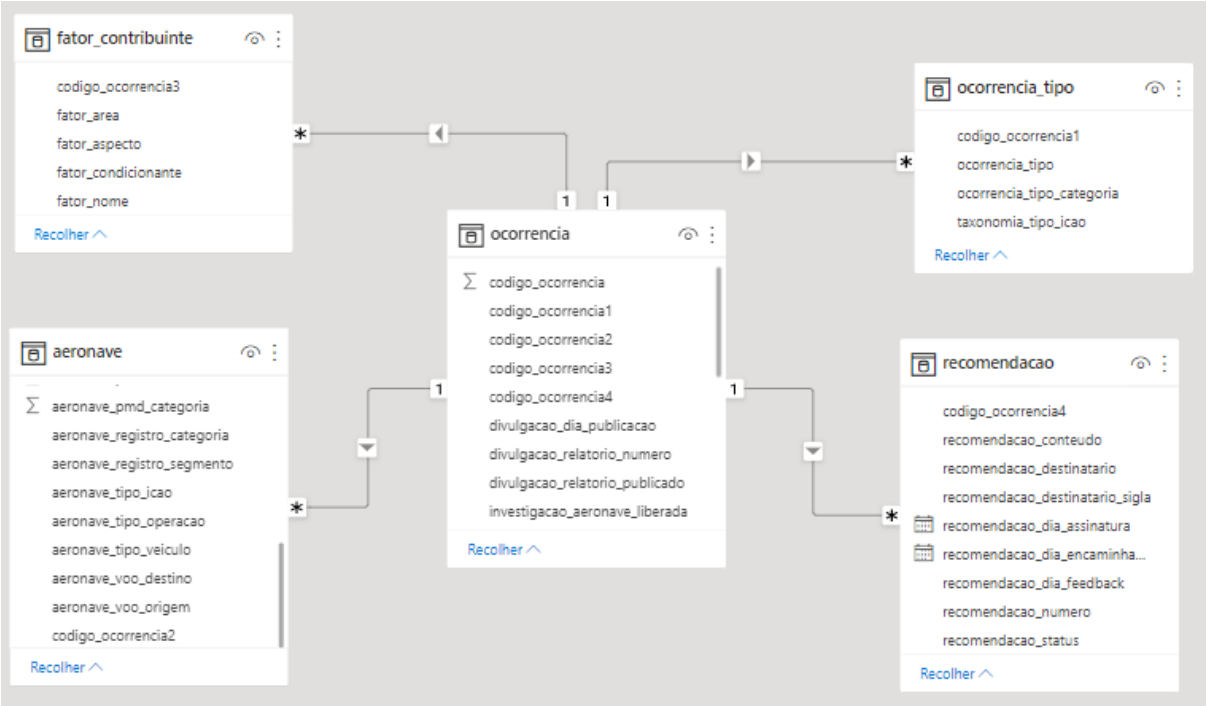
Visão Geral

O conjunto de dados consiste em 5 tabelas, sendo elas:

1. Ocorrência - Que traz dados sobre as ocorrências.
2. Tipo de ocorrência - Que traz dados sobre o tipo de ocorrência.
3. aeronave - Que traz dados sobre as aeronaves envolvidas nas ocorrências.
4. Fator Contribuinte - Que traz dados sobre os fatores contribuintes das ocorrências que tiveram investigações finalizadas.
5. Recomendação - Informações sobre as recomendações de segurança geradas nas ocorrências.

As 5 tabelas do conjunto de dados propostas para serem analisadas pelo desafio possuem pelo menos uma coluna nomeada como ‘codigo_ocorrencia’, porém cada tabela possui uma coluna ‘codigo_ocorrencia’ diferente, numeradas de 0 a 4, somente a tabela ‘ocorrencia’ possui todas as 5 colunas. Com isso, a tabela ‘ocorrencia’ vai ser essencial para relacionar com as demais tabelas.

Para facilitar a visualização podemos considerar a tabela ‘ocorrencia’ como tabela fato e as demais como tabelas dimensão.



Como exemplo, carreguei as 5 tabelas no Power BI. Na visão Schema é possível ver como as 5 tabelas se relacionam e que a tabela ‘ocorrencia’ se relaciona com todas as outras.

Tabela ‘ocorrencia’

São os dados que apresentam o tipo de ocorrência, a localização, a data do acontecimento, o total de recomendações de segurança e o total de aeronaves envolvidas.

A tabela ‘ocorrencia’ possui 5167 linhas e 22 colunas. As colunas são:

1	codigo_ocorrencia	12	ocorrencia_aerodromo
2	codigo_ocorrencia1	13	ocorrencia_dia
3	codigo_ocorrencia2	14	ocorrencia_hora
4	codigo_ocorrencia3	15	investigacao_aeronave_liberada
5	codigo_ocorrencia4	16	investigacao_status
6	ocorrencia_classificacao	17	divulgacao_relatorio_numero
7	ocorrencia_latitude	18	divulgacao_relatorio_publicado
8	ocorrencia_longitude	19	divulgacao_dia_publicacao
9	ocorrencia_cidade	20	total_recomendacoes
10	ocorrencia_uf	21	total_aeronaves_envolvidas
11	ocorrencia_pais	22	ocorrencia_saida_pista

Tabela ‘ocorrencia_tipo’

São os dados que apresentam o tipo de ocorrência como em Ocorrências, porém com uma visão mais detalhada por categorias.

A tabela ‘ocorrencia_tipo’ possui 5347 linhas e 4 colunas. As colunas são:

1	codigo_ocorrencia1
2	ocorrencia_tipo
3	ocorrencia_tipo_categoria
4	taxonomia_tipo_icao

Tabela ‘aeronave’

São os dados que apresentam todas as características das aeronaves, o nível de dano e a quantidade de fatalidades. Uma outra informação é o campo ‘aeronave_pmd’, onde PMD é a sigla para Peso Máximo de Decolagem.

A tabela ‘aeronave’ possui 5235 linhas e 23 colunas. As colunas são:

1	codigo_ocorrencia2	13	aeronave_ano_fabricacao
2	aeronave_matricula	14	aeronave_pais_fabricante
3	aeronave_operador_categoria	15	aeronave_pais_registro
4	aeronave_tipo_veiculo	16	aeronave_registro_categoria
5	aeronave_fabricante	17	aeronave_registro_segmento
6	aeronave_modelo	18	aeronave_voo_origem
7	aeronave_tipo_icao	19	aeronave_voo_destino
8	aeronave_motor_tipo	20	aeronave_fase_operacao
9	aeronave_motor_quantidade	21	aeronave_tipo_operacao
10	aeronave_pmd	22	aeronave_nivel_dano
11	aeronave_pmd_categoria	23	aeronave_fatalidades_total
12	aeronave_assentos		

Tabela ‘fator_contribuinte’

São os dados que apresentam todos os fatores e aspectos que contribuíram com as ocorrências.

A tabela ‘fator_contribuinte’ possui 3464 linhas e 5 colunas. As colunas são:

1	codigo_ocorrencia3
2	fator_nome
3	fator_aspecto
4	fator_condicionante
5	fator_area

Tabela ‘recomendacao’

São os dados que apresentam todas as recomendações de segurança geradas nas ocorrências, o destinatário e o status desta recomendação.

A tabela ‘recomendacao’ possui 1197 linhas e 9 colunas. As colunas são:

1	codigo_ocorrencia4
2	recomendacao_numero
3	recomendacao_dia_assinatura
4	recomendacao_dia_encaminhamento
5	recomendacao_dia_feedback
6	recomendacao_conteudo
7	recomendacao_status
8	recomendacao_destinatario_sigla
9	recomendacao_destinatario

O conjunto de dados possui em sua maioria variáveis do tipo “character”, o que significa que terá necessidade de manipulação de classes.

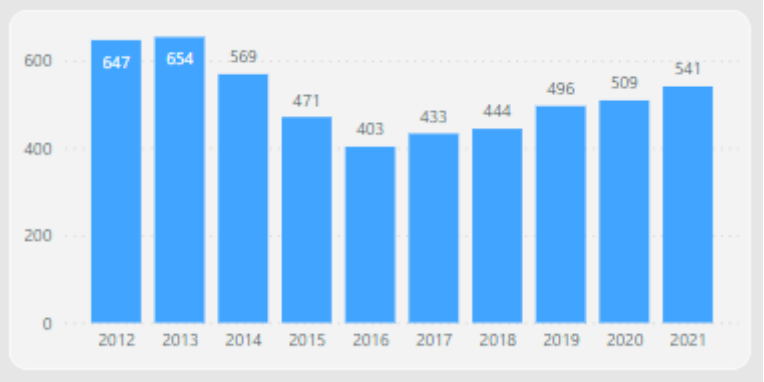
Etapa Data Mining

Nesta etapa foi a fase de explorar o conjunto de dados e entender o que precisa ser feito para deixar os dados mais compreensivos do ponto de vista de facilitar a próxima etapa que seria Data Analytics. Nesta etapa foi feita a modelagem dos dados utilizando a Linguagem R. Foi realizada uma padronização em relação aos dados vazios. Foi feita a manipulação de classes para que as variáveis numéricas e de data sejam exploradas e analisadas da maneira correta. Foi criado algumas colunas específicas na tabela 'ocorrência' relacionadas ao período da coluna 'ocorrencia_dia' (Mês, Dia-Mês, Semana do Mês, Dia da Semana e Hora). Explorando o conjunto de dados, foi detectado que existia colunas idênticas na tabela 'ocorrencia' e que não existia motivo para manter a variável 'codigo_ocorrencia' ser numerada de 0 a 4.

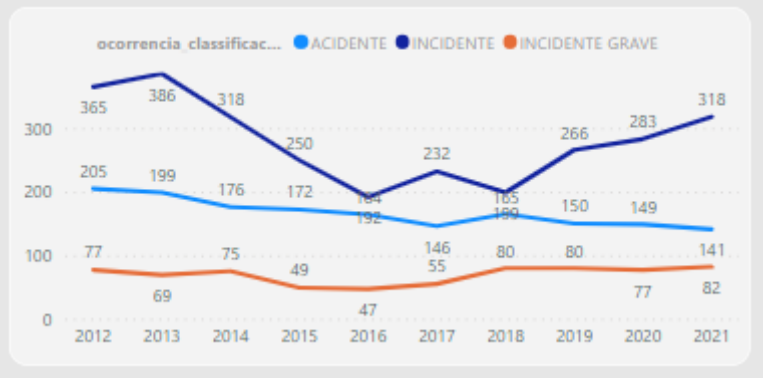
Etapa Data Analytics

Nesta etapa vou utilizar o Power BI para ajudar na análise de dados e na visualização de dados, nesta análise está sendo usado o conjunto de dados modelados com a Linguagem R, a escolha do uso do Power BI foi estritamente estratégica para se ganhar tempo, por se tratar de uma ferramenta versátil. Inicialmente vou analisar cada tabela individualmente, porém é possível relacionar os dados de Ocorrências com as demais tabelas para termos mais insights.

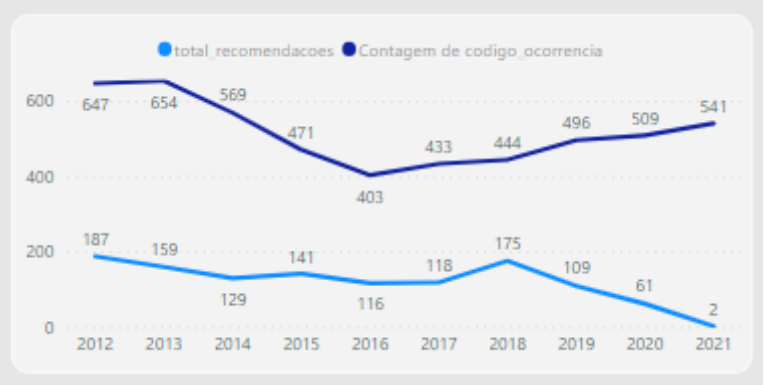
Ocorrências



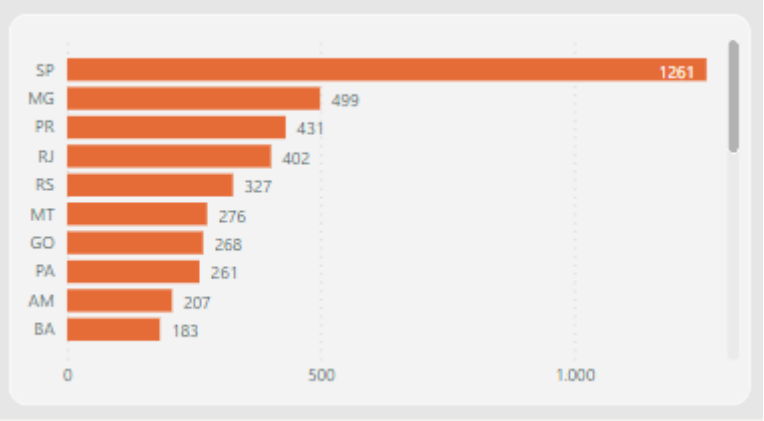
Neste primeiro gráfico que traz a Quantidade de Ocorrências por ano. Já conseguimos perceber que a base de dados de Ocorrências vai de 2012 a 2021, sendo que registrou um menor número de Ocorrências em 2016, mas que logo depois do mesmo, o número de ocorrências vem subindo ano após ano.



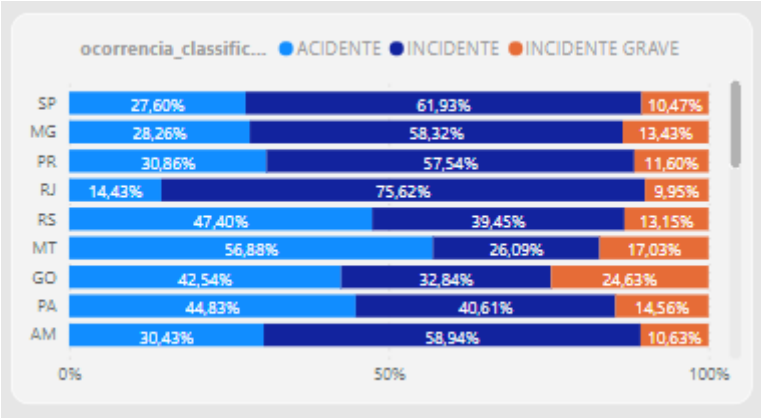
Se separarmos essas ocorrências por classificação, é perceptível que as ocorrências classificadas como Incidente são responsáveis por esse aumento de ocorrências desde do ano de 2016. Enquanto isso, as ocorrências classificadas como Incidente Grave vem caindo com o passar dos anos e as ocorrências classificadas como Acidente vem se mantendo em estabilidade.



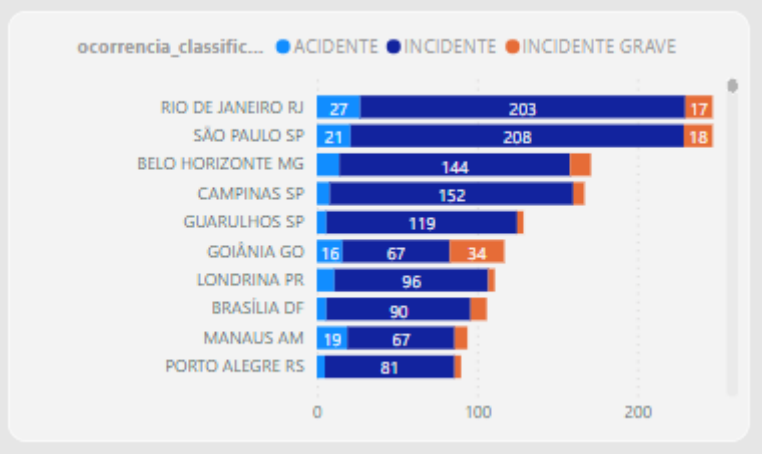
Um fator que pode estar acarretando nesse aumento de ocorrências é a diminuição das recomendações. É possível observar que ambas séries são inversamente proporcionais.



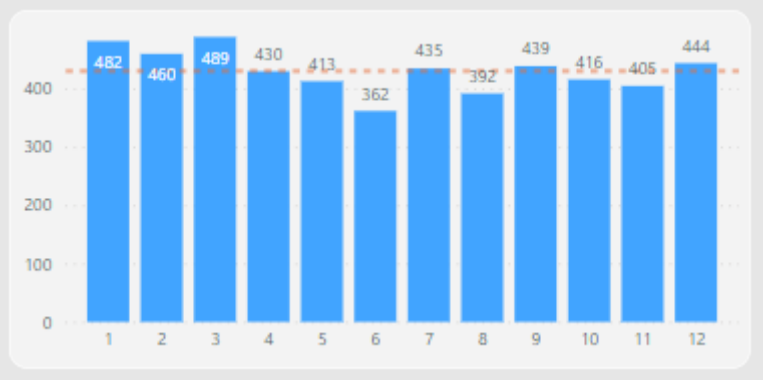
Olhando ocorrências separadas por estados, São Paulo tem a maioria das ocorrências, um pouco mais que 2 vezes e meia que o estado de Minas Gerais.



Olhando agora as ocorrências classificadas proporcionalmente e separadas por estado, é possível perceber que a categoria Acidente continua sendo predominante, principalmente nos 4 primeiros estados que possuem o maior número de Ocorrências.

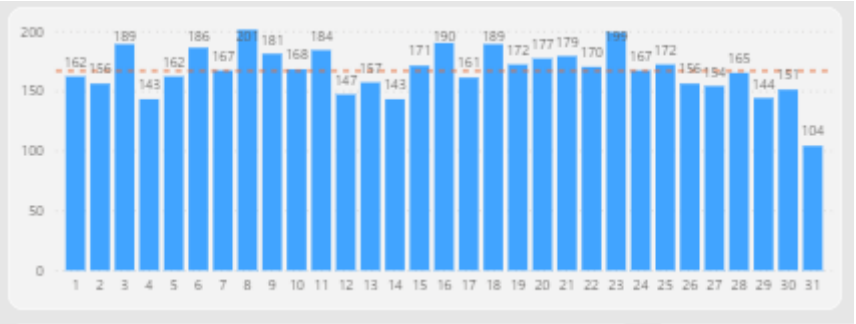


Se olharmos as ocorrências separadas por cidade, a categoria Acidente fica ainda mais evidente como a maioria. Destaque para São Paulo colocando 3 cidades com maiores ocorrências.

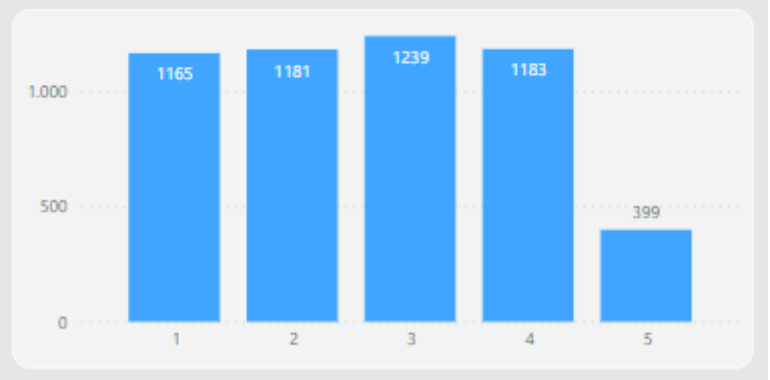


Podemos verificar também o comportamento das ocorrências em relação a sazonalidade. Olhando os dados desde 2012 a 2021 no acumulado dos meses, é possível ver que nos meses de Dezembro a Março tem um número maior de ocorrências em relação a média

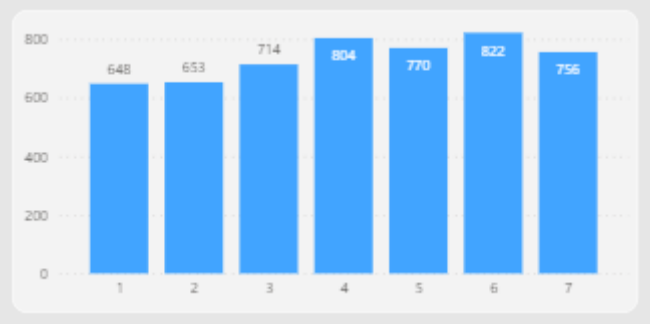
geral, podemos levar em consideração que nesses meses acontecem um número maior de viagens por conta dos períodos de férias e festas, o possível perceber esses mesmo fato acontecer no mês de julho, em comparação com o mês anterior e o mês posterior que possuem um menor volume de ocorrências no período, no mês de julho tem um aumento significativo por conta das férias de meio de ano.



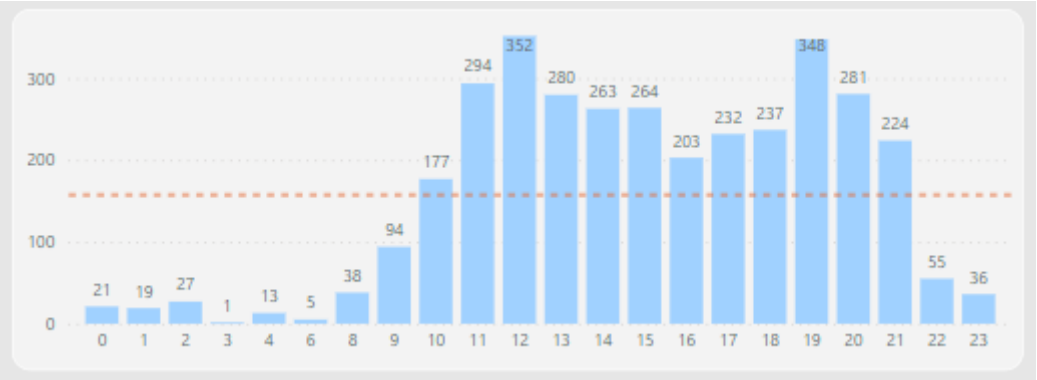
Olhando as ocorrências no acumulado nos dias do mês, não vemos nenhum comportamento atípico, somente no dia 31, mas como não são todos os meses que possuem 31 dias é aceitável ter um volume de ocorrências menor.



Olhando as ocorrências no acumulado das semanas do mês, é possível perceber que a terceira semana é o pico de ocorrências em todos os meses. A quinta semana é aceitável ter poucas ocorrências por ter menos dias.



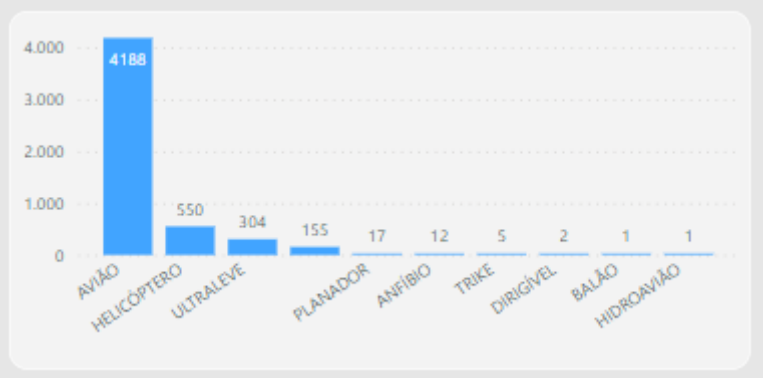
Olhando as ocorrências no acumulado dos dias da semana, onde a sequência de 1 a 7 se encaixa como segunda a domingo, é possível perceber que quinta-feira(4) e sábado(6) são os dias da semana com maiores ocorrências, sendo que segunda(1) e terça-feira(2) tem um menor volume.



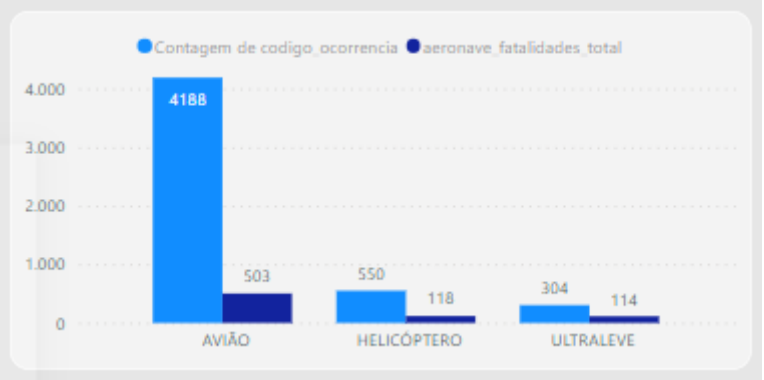
Olhando as ocorrências no acumulado das horas do dia, é possível perceber que há um volume maior de ocorrências a partir das 10 horas da manhã que vai até as 21 horas da noite. Os picos acontecem ao meio-dia e às 19 horas da noite.

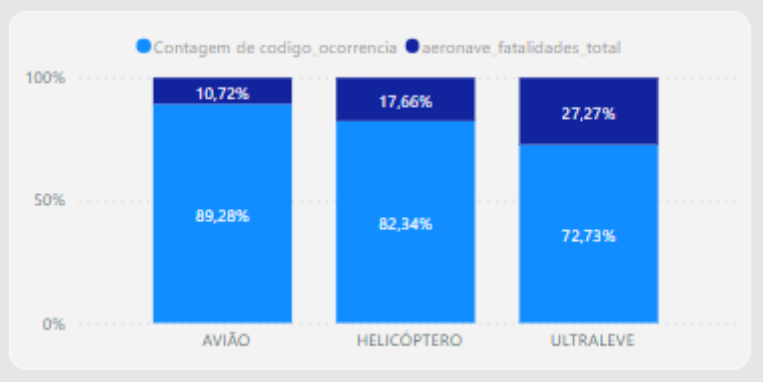
Levando em consideração a sazonalidade podemos concluir que os meses de Dezembro a Março, em toda terceira semana do mês, aos sábados das 10 às 21 horas, é um período crítico para as ocorrências aéreas acontecerem.

Aeronaves

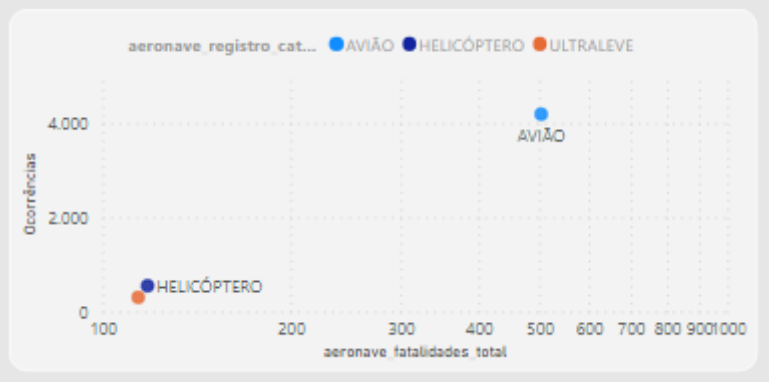


A grande maioria das ocorrências aéreas são de aeronaves da categoria avião.

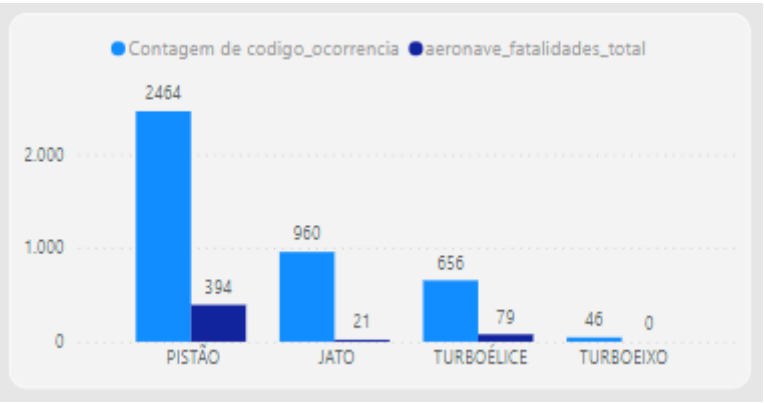


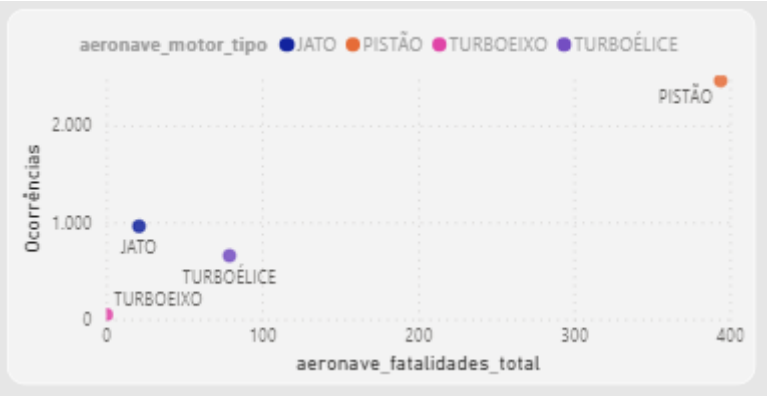


As três principais categorias de aeronaves que possuem um maior número de ocorrências, possuem uma fatia considerável de acidentes fatais. As aeronaves do tipo ultraleve proporcionalmente tem uma fatia maior de acidentes fatais, porém os aviões por ter uma quantidade de ocorrências muito maior consegue ser quase 4 vezes e meia maior em acidentes fatais em comparação aos ultraleves.



Podemos ver as três principais categorias de aeronaves neste gráfico de dispersão que reforça este ponto. Enquanto os ultraleves precisam de quase 3 ocorrências para ocasionar 1 acidente fatal, os aviões precisam de quase 8 ocorrências para ocasionar 1 acidente fatal, sendo assim podemos confirmar que os aviões são mais seguros que os ultraleves.





Agora neste gráfico de dispersão com apenas com os dados de aviões e separado por tipo do motor. Excluindo os do tipo eixo que não tiveram nenhum acidente fatal, os aviões do de pistão precisam de pouco mais de 6 ocorrências para ocasionar 1 acidente fatal, já os aviões turboélice precisam de pouco mais de 8 ocorrências para ocasionar 1 acidente fatal, já os aviões jato precisam de pouco mais de 45 ocorrências para ocasionar 1 acidente fatal, o que faz da categoria de aviões jato a mais seguro de forma disparada.