PROJETO 2: Analisando dados de voos nos EUA

Emanoeli Madalosso
Nanodegree Fundamentos de Data Science II

1. RESUMO

Este projeto utiliza um banco de dados fornecido pelo United States Department of Transportation [1], que fornece dados de voos nos EUA desde 1987, com informações como atrasos e cancelamentos de voos.

Os dados dos últimos 12 meses disponíveis (março de 2017 até fevereiro de 2018) foram selecionados para conduzir este trabalho. Eles podem ser baixados do repositório do Github deste projeto: https://github.com/emanoelim/fundamentos_data_science_2_proj_2, onde foram organizados por trimestre em arquivos .zip.

Durante uma análise exploratória foram levantadas duas perguntas sobre os dados:

1. Porque alguns meses tiveram um número muito maior de cancelamentos de voos?
2. Quais dias são melhores para viajar se eu não quiser sofrer com atrasos nos voos?

Para a primeira pergunta, a análise conduzida neste projeto mostrou que os eventos climáticos que aconteceram no último ano tiveram um grande impacto no número de cancelamento de voos.

Para a segunda pergunta foi possível descobrir que os melhores dias para viajar sem se preocupar com os atrasos são os sábados, exceto aqueles que procedem feriados importantes, entre outras descobertas.

Estas conclusões são apresentadas detalhadamente por meio de histórias no Tableau e estão acessíveis pelas seguintes urls:

História "Quais as causas dos picos de cancelamento de voos ao longo dos últimos 12 meses?":
Versão pré-feedback:
https://public.tableau.com/views/VoosEstadosUnidosltimoano-Verso1/Histria1?:embed=y&:display_count=yes
Versão pós-feedback:
https://public.tableau.com/views/VoosEstadosUnidosltimoano-Verso2/Histria1?:embed=y&:display_count=yes

História "Quais os dias mais tranquilos para viajar?":
Versão pré-feedback:
https://public.tableau.com/views/VoosEstadosUnidosltimoano-Verso1/Histria2?:embed=y&:display_count=yes
Versão pós-feedback:
https://public.tableau.com/views/VoosEstadosUnidosltimoano-Verso2/Histria2?:embed=y&:display_count=yes

2. DESIGN

Inicialmente o projeto tentou seguir as dicas aprendidas durante o curso para criar as visualizações. Após coletar o feedback de duas pessoas, algumas mudanças foram feitas para tornar a visualização mais compreensível e com informações mais relevantes.

2.1. Design inicial

Escolha dos tipos de gráficos:

Quando o objetivo era mostrar quantidades, por exemplo, a quantidade de voos cancelados para cada mês, foram usados gráficos de barras.

Para mostrar informações como a variação no número de voos ao longo do ano, por exemplo, foram usados gráficos de linhas.

Foi adotada a abordagem de usar tamanhos diferentes em gráficos onde era desejado ressaltar uma diferença quantitativa entre os dados com o objetivo de tornar a compreensão dessa diferença mais intuitiva.

Foram usados mapas sombreados para mostrar a variação de dados entre diferentes estados.

Também foram usados gráficos de dispersão quando havia necessidade de mostrar a correlação entre dois tipos de variáveis.

Escolha das cores:

Foi utilizada a paleta de cores chamada “Daltônico” do Tableau para as visualizações ficarem acessíveis para pessoas daltônicas. As cores usadas desta paleta foram o azul, o laranja e o cinza:

Estas cores são mais sóbrias, de modo a evitar que observador se canse visualmente e também tem um bom contraste entre elas. A maioria dos gráficos foi apresentado na cor azul e só foram usados gráficos com mais cores quando foi necessário ressaltar diferenças categóricas ou quantitativas.

Para montar as paletas categóricas foi utilizado o azul e laranja quando o objetivo era evidenciar dois tipos de grupos diferentes. Quando o objetivo era evidenciar um grupo em relação a outro, foi usado o azul e o cinza, sendo o azul para o grupo onde era desejado dar mais destaque e o cinza para o grupo de menos destaque.

Para paletas quantitativas, foram usadas apenas as do tipo sequenciais usando tons do azul acima, visto que não há gráficos com valores negativos e não foi necessário usar paletas de cores divergentes.

Escolha das linhas:

Foram utilizadas linhas finas para os eixos e para a grade. Nos gráficos de barra não foram usadas linhas para o contorno das barras, deixando os gráficos menos poluídos e chamando mais a atenção para os dados do que para outros detalhes.

Escolha dos textos e títulos:

Foi usada uma pequena quantidade de textos nos gráficos, como os textos referentes aos eixos e legendas, para que o observador compreenda o objetivo da visualização com rapidez.

Optou-se por colocar títulos em cada gráfico, pois alguns painéis tinham vários gráficos, assim ficaria mais fácil ver do que cada gráfico tratava.

Escolha dos tipos de interação:

Em vários gráficos foram utilizados filtros para que o observador possa interagir e comparar a variação nos dados conforme um mês de sua escolha, por exemplo.

2.2. Ajustes realizados após feedback

Os nomes das colunas foram traduzidos para o português e foram reescritos em formato de texto em vez de um nome de campo ou variável.

Os nomes dos eixos e demais textos dos gráficos também foram adequados para serem mais intuitivos e compreensíveis por pessoas de diferentes contextos.

Onde existia alguma sigla nas visualizações foi alterado para o significado da sigla para facilitar a contextualização. Siglas de estados, por exemplo, foram trocadas para o nome do estado por extenso. Também foram adicionadas as medidas de tempo nos gráficos representando os atrasos.

Algumas novas visualizações foram adicionadas para responder perguntas levantadas pelas pessoas consultadas no feedback, por exemplo: uma visualização mostrando quais companhias aéreas tinham mais cancelamentos ou atrasos. Da mesma forma que foram utilizados os nomes do estados em vez de siglas, nessa visualização foram usados os nomes das companhias aéreas em vez de seu código IATA (International Air Transport Association). Para isso foi criado um alias para cada companhia, pesquisando o significado do código no site Airline and airport code search [4].

Algumas questões referentes aos dados não foram respondidas, pois envolviam utilizar um intervalo de dados maior que o escopo inicial do projeto e até mesmo fontes adicionais de dados como, por exemplo, questões de impacto financeiro.

3. FEEDBACK

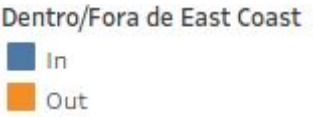
Foram realizados feedbacks com duas pessoas. Os feedbacks foram realizados pessoalmente. Foi feita uma breve explicação dos dados e das perguntas que queriam ser respondidas através dos gráficos. Foi solicitado que a pessoa fizesse comentários em voz alta durante sua exploração dos gráficos, sendo anotados esses comentários. Ao final da exploração foram feitas as seguintes perguntas:

1. Você tem alguma pergunta sobre os dados?
2. Qual você acha que é a principal conclusão desta visualização?
3. Há algo que você não entendeu?
4. Você tem alguma sugestão?

3.1. Feedback 1

- Dados da pessoa:
    1. Idade: 19
    2. Profissão: UX Designer
- Comentários anotados durante a exploração:
    1. Ao passar o mouse sobre pontos dos gráficos, as legendas não são muito intuitivas.
    2. Os títulos dos eixos as vezes estão com nome de “variáveis” em vez de terem um texto mais intuitivo.
    3. Achou interessante a possibilidade de poder selecionar visualizações de acordo com os meses do ano.

4. Nos gŕaficos com estados poderia mostrar o nome do estado ao passar o mouse sobre o mesmo em vez de mostrar só a sigla, pois como trata-se de dados de outro país teve dificuldade de lembrar o nome do estado a partir da sua sigla.
5. No gŕafico que mostra os cancelamentos da Costa Leste achou as legendas das cores confusas da maneira que estavam (figura abaixo), pensando que IN ou OUT indicava que eram voos entrando ou saindo do estado.



- Respostas para as perguntas pós exploração:
    1. Questionou se os dados usados eram de apenas uma empresa aérea. Também questionou o fato de o gráfico com o tempo médio de atraso x dia ter um crescimento linear em dezembro de 2017.
    2. Para a história “Quais as causas dos picos de cancelamento de voos ao longo dos últimos 12 meses?” conseguiu compreender que os cancelamentos decorriam dos eventos climáticos que aconteceram no país. Para a história “Quais os dias mais tranquilos para viajar?” apontou que caso morasse nos EUA as visualizações poderiam ajudar a decidir em que dias fazer uma viagem a passeio, visto que são o tipo de viagem onde há uma maior flexibilidade para escolha de datas. Apontou que optaria por viajar no sábado, que é o dia onde parece ter menos atrasos e que caso precisasse viajar em um dia com grande média de atrasos, como na sexta-feira, se programaria para a situação. Também disse que evitaria viajar em datas após feriados importantes devido a possibilidade de atrasos.
    3. Sentiu um pouco de dificuldade no gráfico de dispersão entre atrasos de saída x chegada.
    4. Sugeriu adicionar um gráfico mostrando quais companhias aéreas costumam atrasar mais, para ajudar decidir qual companhia escolher.

3.2. Feedback 2

- Dados da pessoa:
    1. Idade: 25
    2. Profissão: Desenvolvedor de software
- Comentários anotados durante a exploração:
    1. Os títulos dos eixos poderiam estar no formato de “texto” em vez de um formato de “variável”. Foi percebido isso em vários gráficos.
    2. Nos gráficos que mostram os atrasos não há informação sobre a medida de atraso (horas, minutos, etc).
- Respostas para as perguntas pós exploração:
    1. Teve questionamentos mais relacionados a outras questões que poderiam ser respondidas com o conjunto de dados utilizado, como por exemplo, a influência de atentados no número de voos ou a influência de eventos esportivos no número de voos, bem como o impacto econômico gerado.

2. Para a história “Quais as causas dos picos de cancelamento de voos ao longo dos últimos 12 meses?” constatou que o objetivo foi mostrar o impacto de eventos climáticos nos voos. Para a história “Quais os dias mais tranquilos para viajar?” constatou que o objetivo era mostrar quais dias tentar evitar caso queira realizar um voo sem incômodos por causa de atrasos.
3. No gráfico de dispersão que mostra a relação entre atrasos na saída e chegada existe um ponto do gráfico onde é possível ver vários voos que não tiveram atraso na saída mas tiveram atraso na chegada. O que pode ter provocado isso?
4. Sugeriu analisar um intervalo maior de dados, para ter visões de como o acesso a voos mudou ao longo dos últimos anos.

4. RECURSOS UTILIZADOS

Optou-se por utilizar os dados referentes aos voos dos Estados Unidos. Os dados foram baixados diretamente do site United States Department of Transportation [1], para obter dados mais condizentes com a atualidade e também pela possibilidade de escolher os campos desejados para montar o arquivo .csv. Foram selecionados dados no intervalo de um ano (de março de 2017 até fevereiro de 2018). Foram selecionados os campos abaixo:

Tipo do campo	**Nome do campo**	**Descrição**
Time period	Year	Ano
	Quarter	Trimestre
	Month	Mês
	DayofMonth	Dia
	DayofWeek	Dia da semana (1 - Monday, 7 - Sunday)
	FlightDate	Data do voo
Airline	UniqueCarrier	Código único da transportadora
	AirlineID	ID da companhia aérea
	Carrier	Transportadora
	TailNum	Número da cauda
	FlightNum	Número do voo
Origin	Origin	Aeroporto de origem
	OriginCityName	Cidade de origem
	OriginState	Estado de origem
Destination	Dest	Aeroporto de destino
	DestCityName	Cidade de destino

	DetState	Estado de destino
Departure performance	CRSDepTime	Horário agendado para partida
	DepTime	Horário real da partida
	DepDelay	Atraso na partida
Arrival performance	CRSArrTime	Hoarário agendado para chegada
	ArrTime	Horário real da chegada
	ArrDelay	Atraso na chegado
Cancellations and diversions	Cancelled	Indicador de voo cancelado (0 - não, 1 - sim)
	ActualElapsedTime	Tempo decorrido de voo
	AirTime	Duração do voo
	Distance	Distância
Cause of delay	CarrierDelay	Atraso pela transportadora
	WeatherDelay	Atraso pelo clima
	NASDelay	Atraso pelo National Airspace System
	SecurityDelay	Atraso por segurança
	LateAircraftDelay	Aeronave atrasada

Também foram consultadas algumas notícias sobre cancelamentos de voos nos EUA para tentar entender a causa dos picos de cancelamento em alguns meses. Elas estão listadas nas referências ao final do relatório.

5. REFERÊNCIAS

1. United States Department of Transportation. **"Airline On-Time Performance Data"**. 2018. Disponível em: https://www.transtats.bts.gov/DL_SelectFields.asp. Acesso em: maio de 2018.
2. Kottasová, I.; Disis, J.; Smith, A. **"Travel nightmare: Winter storm wipes out thousands of flights"**. 2018. Disponível em: http://money.cnn.com/2018/01/04/news/flight-cancellations-winter-storm/index.html. Acesso em: maio de 2018.
3. Mutzabaugh, B. **"Hurricane Irma: Florida flight cancellations spiking, now up to 4,200+"**. 2017. Disponível em: https://www.usatoday.com/story/travel/flights/todayinthesky/2017/09/08/hurricane-irma-2-000-florida-flights-already-canceled-more-likely/644768001. Acesso em: maio de 2018.
4. IATA. **"Airline and airport code search"**. 2018. Disponível em: http://www.iata.org/publications/Pages/code-search.aspx. Acesso em maio de 2018.

5. Wikipedia. **“List of U.S. state abbreviations”**. 2018. Disponível em: https://en.wikipedia.org/wiki/List_of_U.S._state_abbreviations. Acesso em: maio de 2018.