FUNDAÇÃO GETÚLIO VARGAS
CIÊNCIA DE DADOS

TRABALHO FINAL DA DISCIPLINA LINGUAGENS DE PROGRAMAÇÃO

ARI OLIVEIRA, JOÃO DONASOLO, LÍVIA MEINHARDT, LUIZ LUZ

DEZEMBRO DE 2020

Lista de Figuras

Figura 1 – Mudança temporal do percentual de base dos aeroportos 5
Figura 2 – Média do percentual da linha de base por mês 5
Figura 3 – Aeroportos dos USA - mais e menos afetados 7
Figura 4 – Aeroportos do Canadá - mais e menos afetados 8
Figura 5 – Mapa de correlação dos dados da FIFA 10
Figura 6 – Camisas mais usadas pelos jogadores com "Overall"acima de 85 . . 11
Figura 7 – Camisas mais usadas pelos jogadores com "Overall"acima de 90 . . 12
Figura 8 – Camisas mais usadas pelos melhores jogadores de cada clube . . . 13
Figura 9 – Reputação Internacional das nações 14
Figura 10 – Reputação Internacional dos clubes 14
Figura 11 – Distribuição das habilidades dos melhores atacantes 15
Figura 12 – Distribuição das habilidades dos piores atacantes 16

Sumário

1 – Introdução 1
1.1 Organização do trabalho 1

2 – Extração e Limpeza dos Dados 2
2.1 Extração de Dados 2
2.2 Limpeza de Dados 2
2.3 Diagrama de Solução 3

3 – Tráfico dos aeroportos em meio a pandemia da COVID-19 4
3.1 Analise exploratória 4
3.2 Perguntas formuladas 6
3.3 Visualizações 6
3.4 Modelo 8

4 – Jogadores da FIFA 9
4.1 Analise exploratória 9
4.2 Perguntas formuladas 9
4.3 Visualizações 10
4.4 Modelos 16

5 – Conclusão 18
5.1 Considerações Finais 18

1 Introdução

Esse relatório, elaborado para o trabalho final da disciplina, tem como objetivo apresentar as análises, soluções e conclusões do grupo acerca dos conjuntos de dados escolhidos: "COVID-19's Impact on Airport Traffic"e "FIFA 19 complete player dataset".

1.1 Organização do trabalho

Ari Oliveira: Cientista de Dados / Especialista de Negócio

João Donasolo: Engenheiro de Dados / Engenheiro de Software

Lívia Cereja: Especialista em Visualização de Dados

Luiz Luz: Especialista de Garantia da Qualidade

2 Extração e Limpeza dos Dados

2.1 Extração de Dados

O banco de dados ao qual a string de conexão fornecida faz referência é do tipo MS SQL server. Para fazer criar uma conexão com o servidor, foi utilizado o módulo python pyodbc e, para sua utilização, foi necessária a instalação do driver correspondente ODBC Driver 17 for SQL Server , disponível no site da Microsoft.

Após o download e carregamento dos scripts, foi feito um estudo preliminar do módulo através da documentação oficial disponível no github , seguindo-se da análise do banco de dados para sua compreensão.

Finalmente, foram escritos os códigos para extração dos dados do banco e carregamento destes em um arquivo csv para limpeza.
Alusão ao ELT: Extract $\rightarrow$Load $\rightarrow$Transform

2.2 Limpeza de Dados

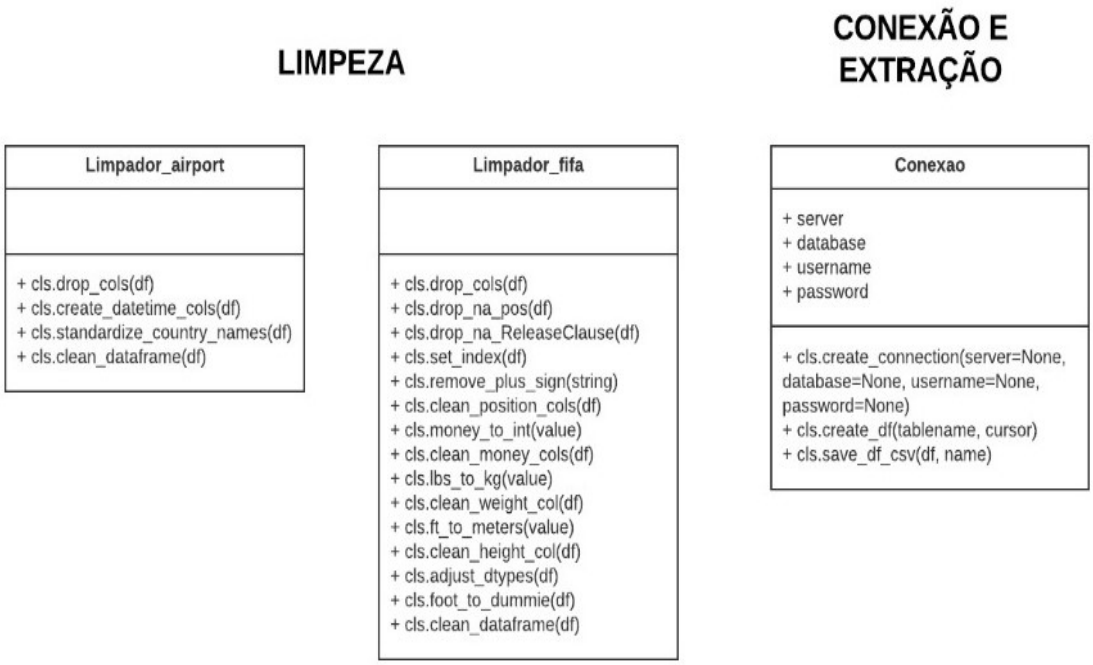
O processo de limpeza seguiu duas fases: Remoção de registros incompletos e transformação dos dados.

No dataset "FIFA 19 complete player dataset", foram removidas colunas desnecessárias (como as colunas com valores repetidos e as que continham links da web), além registros com informações faltantes. Em seguida, os dados foram transformados para que pudessem ser analisados. Por fim, valores como altura e peso foram transformados para seus valores correspondentes no Sistema Internacional.

Já no dataset "COVID-19's Impact on Airport Traffic", de forma análoga ao conjunto de dados da Fifa, foram removidas colunas desnecessárias. Além disso, foram padronizados os nomes dos países e separados dados distintos presentes na mesma coluna - data em dia, mês, ano.

2.3 Diagrama de Solução

Após os processos descritos, a solução foi ilustrada no diagrama da figura



3 Tráfico dos aeroportos em meio a pandemia da COVID-19

O dataset "COVID-19's Impact on Airport Traffic"contém dados do fluxo dos aeroportos desde o início da pandemia, de acordo com a porcentagem da linha de base do aeroporto registrado. A tabela 1 demostra algumas das suas colunas, utilizadas no desenvolvimento desse projeto.

Date	AirportName	PercentOfBaseline	Country
2020-08-04	John F. Kennedy International	54	United States
2020-08-05	Montreal Trudeau	83	Canada
2020-09-07	McCarran International	22	United States
2020-03-20	LaGuardia	67	United States
2020-07-22	Los Angeles International	76	United States

Tabela 1 – Representação de algumas colunas do dataset

3.1 Analise exploratória

A fase de analise exploratória desse dataset foi feita, principalmente, através de gráficos, essenciais para o entendimento dos dados. Além disso, a divisão em dados agrupados e uso de métodos como o *describe* foram bastante úteis. A figura 3.1 demostra visualmente as informações mais relevantes disponíveis e suas relações. A partir desse gráfico fica claro que os dados representam os registros diários dos percentual de linha de base de cada aeroporto, dividido, também, por país. Assim, através desses podemos observar que o fluxo não apresenta uma tendência clara, contanto com bastante variação ao longo de todo o período.

Para tentar identificar uma tendência foi projetado o gráfico 3.1, que destaca melhor as diferenças entre cada país presente no dataset pela média de cada mês. A analise isolada desse poderia deixar a entender que cada país tem um comportamento bastante distinto. Porém, a ausência de dados, principalmente, do Chile e da Austrália, que contam somente com registros de um único aeroporto cada, torna qualquer conclusão que envolve esses países inválida, uma vez que não podemos medir seus comportamentos somente com um registro.

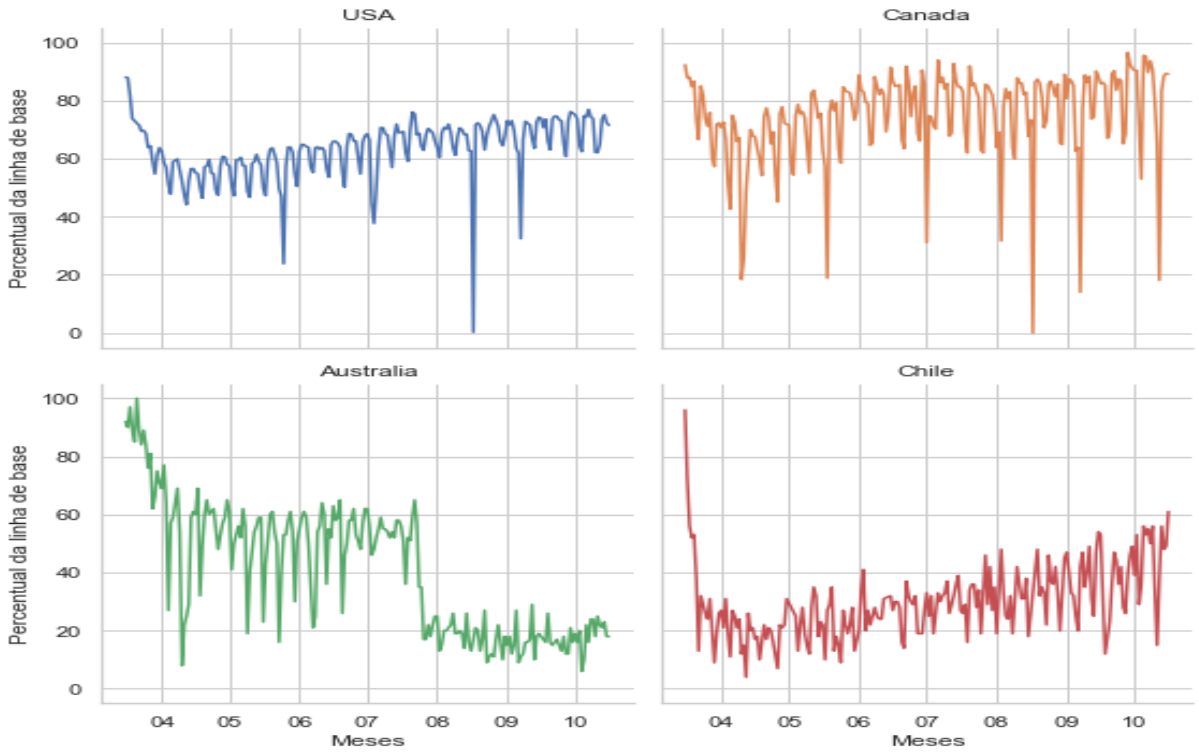


Figura 1 – Mudança temporal do percentual de base dos aeroportos

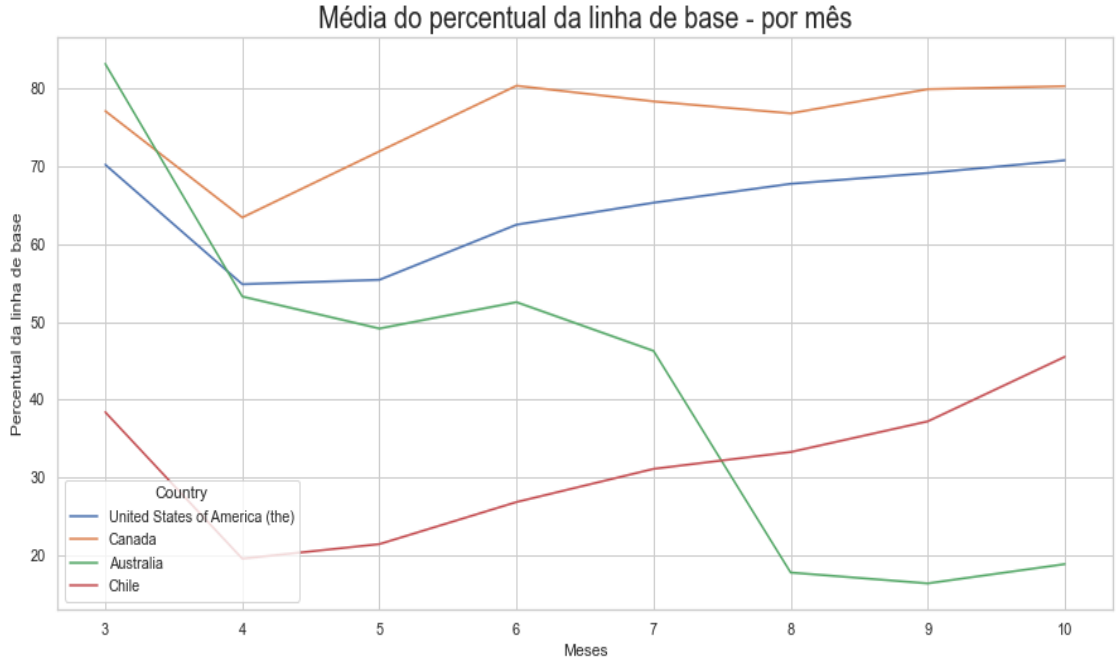


Figura 2 – Média do percentual da linha de base por mês

Constatada essa ausência, foi decidido focar nos dados dos Estados Unidos e Canadá, que apresentam maior volume e assim, podem ser devidamente explorados. A partir deles, foram procuradas outras possíveis relações a serem aprofundadas. A relação por dia do mês foi rapidamente destacada, mas foi percebido uma diferença considerável no comportamento dos aeroportos.

O arquivo *"analise inicial visualizações aeroporto.py"* contém todos as visualizações e tabelas utilizadas para essa análise, nesse relatório foram incluídos somente aqueles considerados mais relevantes.

3.2 Perguntas formuladas

A partir da análise inicial dos dados, foram formuladas as perguntas que gostaríamos de responder com visualizações ou modelos.

1. Quais aeroportos foram mais afetados, no geral?
2. Quais aeroportos foram menos afetados, no geral?
3. Conseguimos prever a porcentagem da linha de base em um determinado dia, dada as demais informações?

3.3 Visualizações

Para responder as perguntas formuladas a partir da análise inicial dos dados, foram construidas duas visualizações, no Tableau. A primeira para os dados dos Estados Unidos 3.3 e a segunda para os do Canadá 3.3. Em ambas, foi considerada a média - para cada aeroporto- por mês, do percentual da linha de base. Assim, podemos traçar aqueles que foram mais ou menos afetados em cada período.

No caso dos Estados Unidos é interessante observar que aqueles aeroportos mais/menos afetados durante o período inicial da pandemia, nesse caso o Daniel K.Inouye Internacional, que quase não sofreu alterações e o McCarran Internacional, que funcionou somente com 40% da sua base em março, permaneceram nessas categorias durante todo o período registrado. Além desses, os demais aeroportos que aparecem no gráfico, no geral, demostram uma tendência de permanecer, também, na mesma categoria, porém variam mais entre ela e o estado "mediano"(que seriam as porcentagens ausenteis na visualização - de 50% a 80% da base).

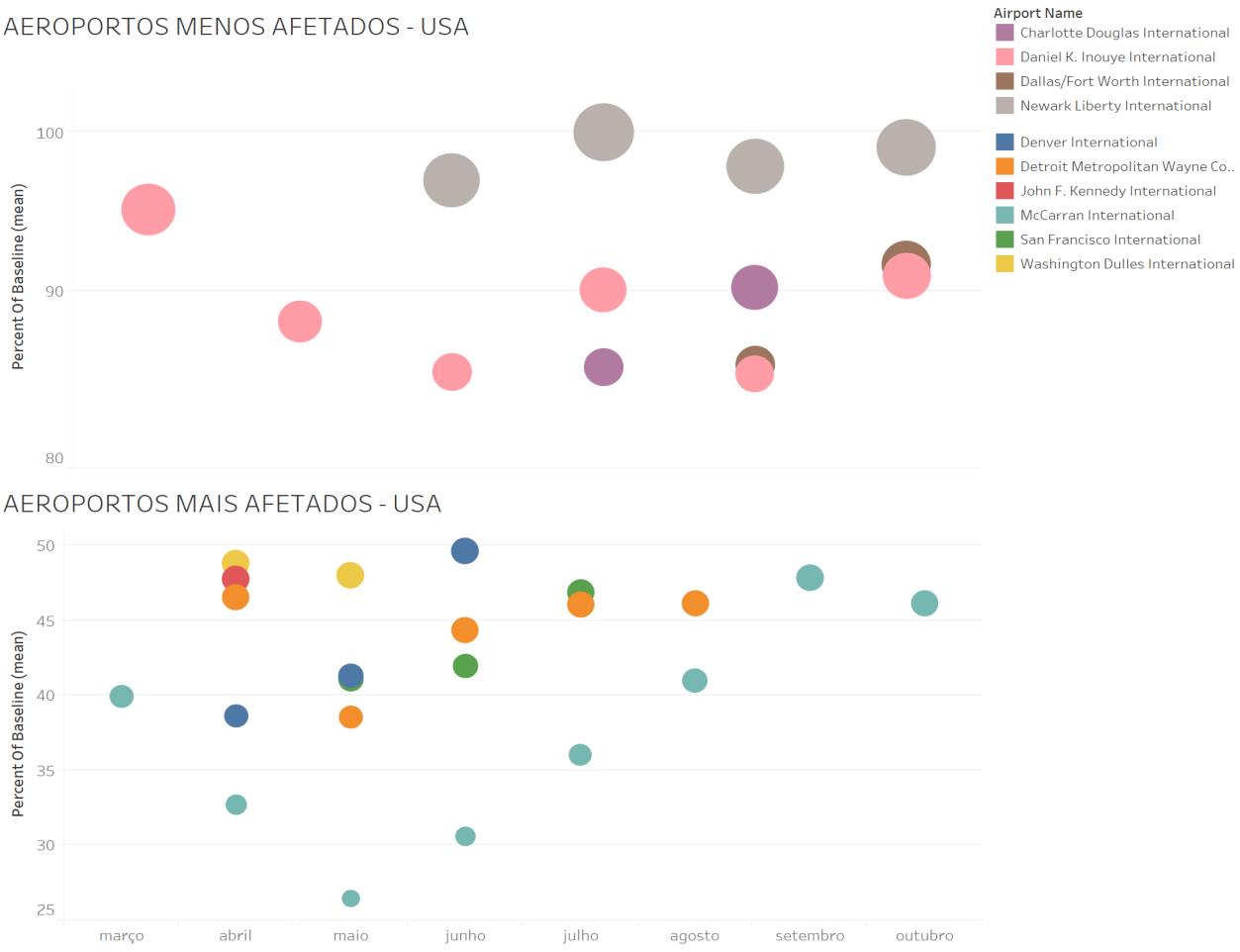


Figura 3 – Aeroportos dos USA - mais e menos afetados

Os aeroportos canadenses tem um comportamento, de certa forma, similar: aqueles que sofreram mais/menos logo no início tendem a permanecer na mesma categoria. Porém, nesse caso, a variação entre os estados mais extremos e o mediano parece ser maior. Também é interessante notar que, em ambos os países, um mesmo aeroporto nunca fez parte de cenários opostos, em períodos distintos. Esse fato mostra certa relação entre o aeroporto e o quão afetado ele foi (ou está sendo) pela pandemia.

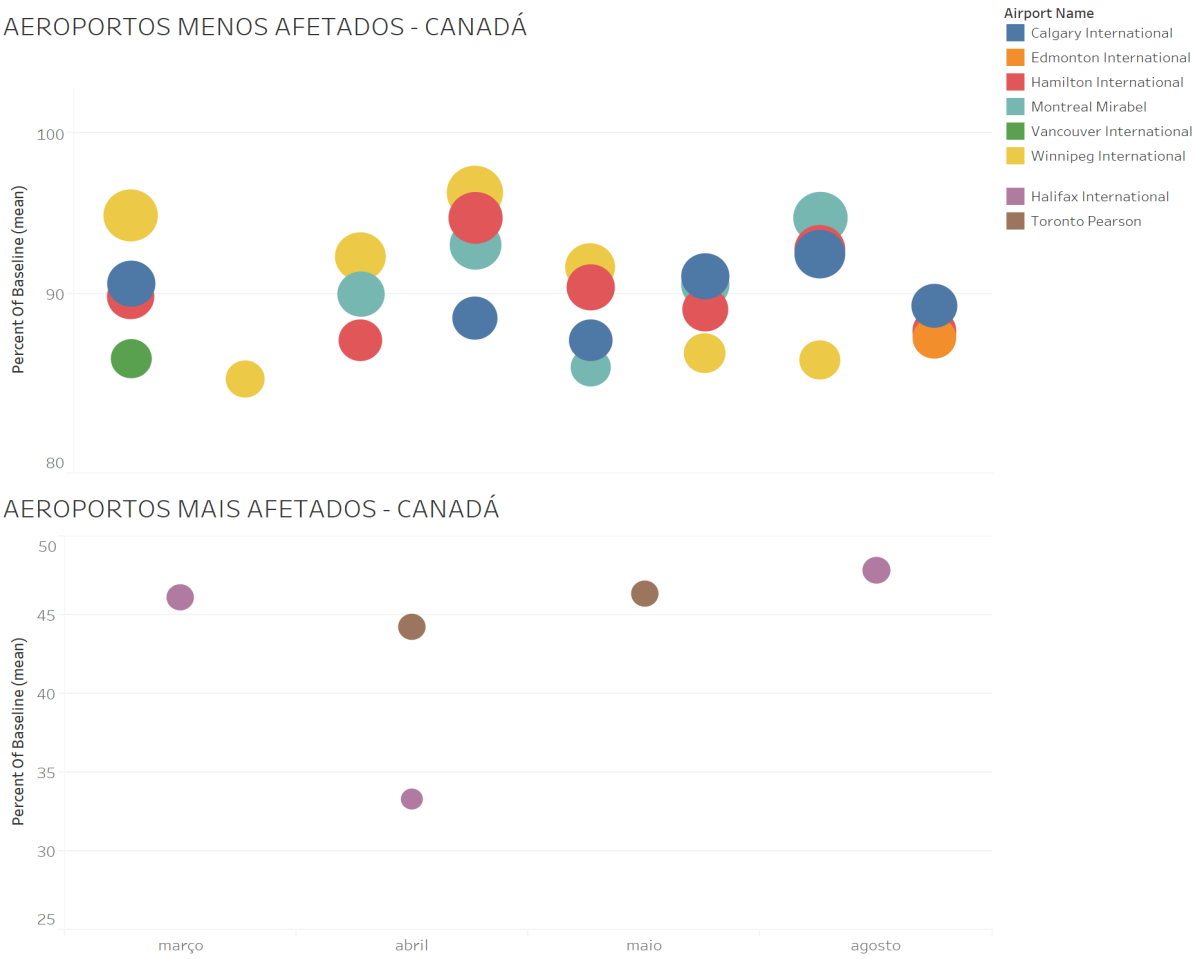


Figura 4 – Aeroportos do Canadá - mais e menos afetados

3.4 Modelo

Como comentado na seção 3.1, a série temporal representada pela figura 3.1 não obedece nenhuma tendência clara, logo um modelo de regressão Linear não parece tão indicado como abordagem. Todavia, a hipótese foi testada na prática no módulo *model airport.py* , aonde percebe-se que não há variáveis numéricas linearmente independentes suficientes para a construção de um modelo. Além disso, as variáveis categóricas também não podem ser usadas para o ajuste, assim não é possível fazer previsões a respeito do dataset utilizando regressões logísticas ou lineares simples.

Ademais, caso fosse possível construir os modelos a partir das variáveis disponíveis, a precisão do modelos, para dados externos, reais, provavelmente não seria confiável, uma vez que os dados utilizados não possuem registros suficientes, como comentado anteriormente (também na seção 3.1.

4 Jogadores da FIFA

O dataset "FIFA 19 complete player dataset"contém dados dos jogadores disponíveis no vídeo game FIFA 19, desde características gerais como idade, peso e altura até suas habilidades em campo. A tabela 2 exemplifica algumas das suas colunas (originalmente são 89), que foram utilizadas nesse projeto.

Overall	Club	Wage	Position	Jersey_Number	Finishing
69	Nîmes Olympique	8000.0	CDM	2.0	48.0
65	Al Raed	5000.0	RCB	66.0	23.0
74	Deportivo Cali	3000.0	CAM	8.0	50.0
65	Lincoln City	5000.0	RCM	4.0	56.0
69	Sagan Tosu	3000.0	RDM	36.0	49.0

Tabela 2 – Representação de algumas colunas do dataset

4.1 Analise exploratória

A análise inicial da fifa se deu por meio de diferentes abordagens, foram utilizados métodos como *describe*, além de gráficos e agrupamentos para entender melhor os dados contidos e suas relações. Como a base é bastante extensa, foi necessário escolher alguns focos, dentre os muitos possíveis nesse caso. Para auxiliar nesse processo, foram construídos mapas de calor da correlação das variáveis, como apresentado na figura 4.1.

4.2 Perguntas formuladas

A partir das análises feitas no módulo *analise inicial visualizações fifa.py* e discussões em grupo, foi possível definir as prioridades do projeto para esse dataset. Assim, as perguntas formuladas a serem respondidas pelas visualizações e modelos:

1. Qual o número mais usado pelos os melhores jogadores de cada clube?
2. Qual é a reputação das maiores nações?
3. Como é a distribuição das habilidades por posição?
4. O que mais influencia o "Overall"do jogador?
5. Conseguimos prever seu "Overall dadas suas características? Ou suas habilidades?

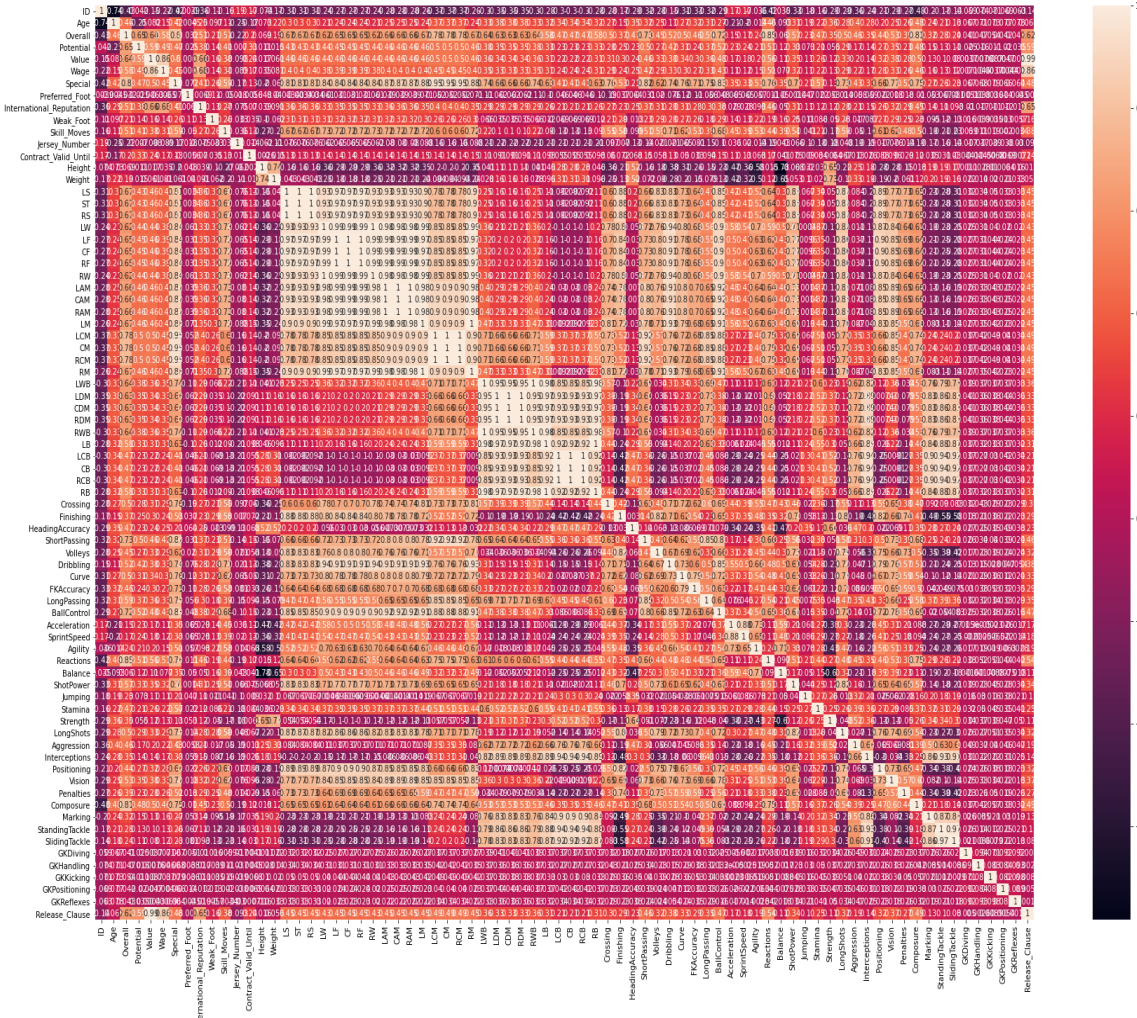


Figura 5 – Mapa de correlação dos dados da FIFA

4.3 Visualizações

Para descobrir a numeração mais utilizada pelos melhores jogadores foram feitas três visualizações, no Tableau. As duas primeiras consideram os melhores jogadores no geral, ou seja, entre todos os jogadores presentes no dataset. Já a última, considera o melhor de cada clube distinto, apresentando, assim, a distribuição da numeração entre os jogadores com maior "Overall"de cada clube.

O boxplot foi escolhido para o gráfico 4.3, pois existe uma número considerável de jogadores na condição estabelecida ("Overall"acima de 85) nos clubes presentes. Os clubes que foram descartados dessa visualização apresentaram baixa presença de jogadores nessa categoria.

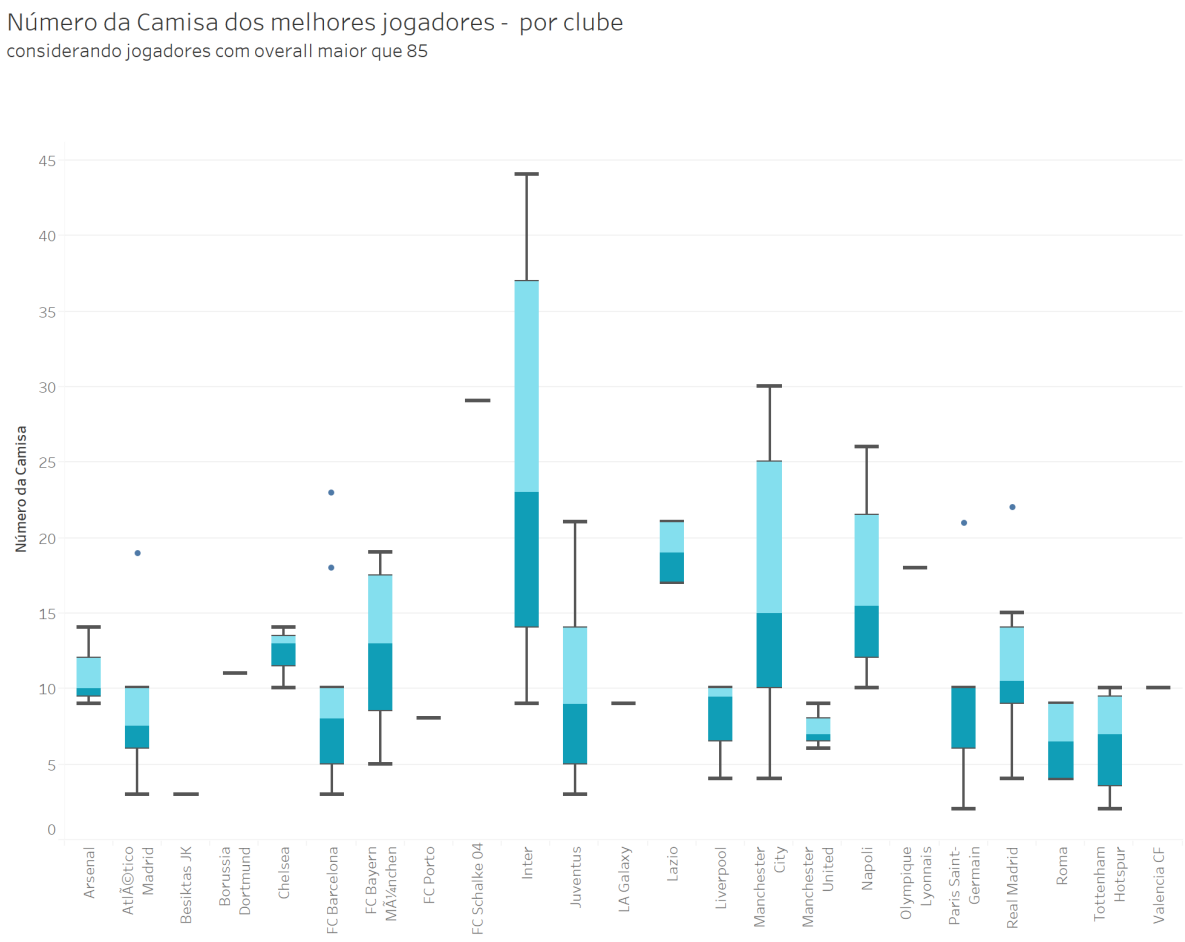


Figura 6 – Camisas mais usadas pelos jogadores com "Overall"acima de 85

O gráfico de "bolhas"foi escolhido no caso da figura 4.3, pois não há muitos jogadores nessa condição no dataset. Essa visualização é particularmente interessante, pois, de certa forma, acaba por filtrar os melhores clubes. Além disso, apresenta a escolha da numeração dos melhores jogadores de forma bastante clara e conforme o esperado - no geral utiliza camisa 10 ou 7.

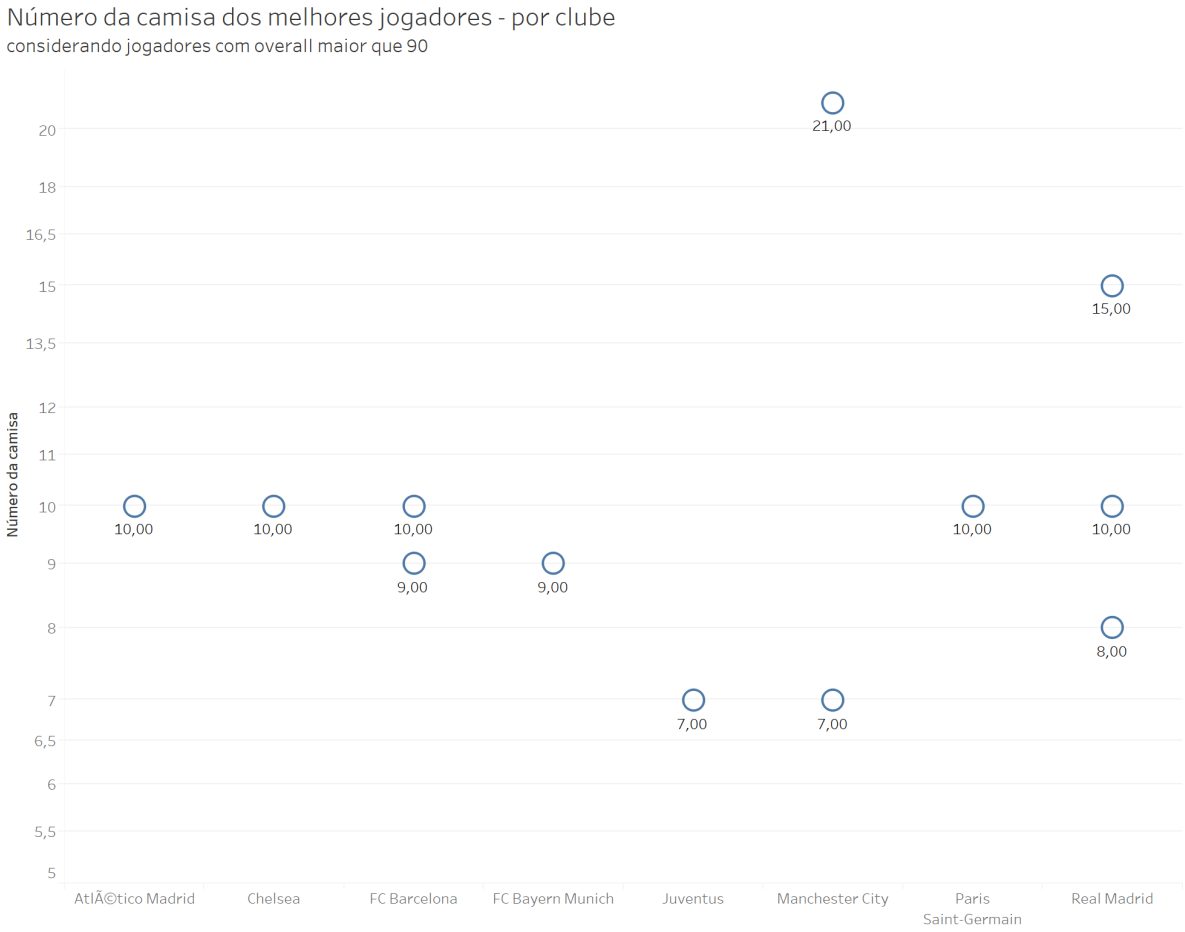


Figura 7 – Camisas mais usadas pelos jogadores com "Overall"acima de 90

Para finalizar essa pergunta, a visualização 4.3 apresenta os dados no formato de histograma, com a contagem no eixo y podemos claramente observar as camisas que mais se repetem. Além disso, foi adicionado um gráfico de "bolhas", cujo tamanho representa a proporção de uso daquela numeração, para os casos mais frequentes.

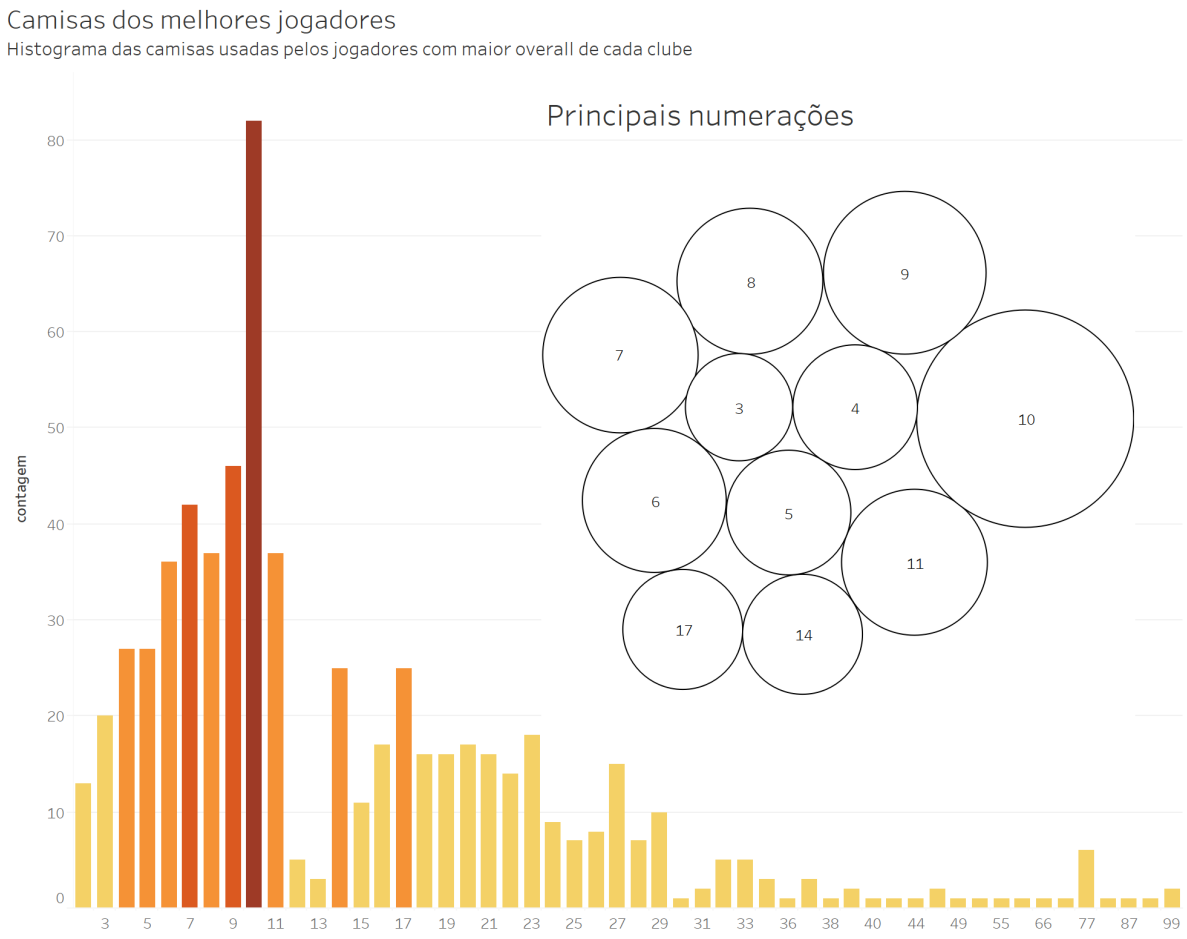


Figura 8 – Camisas mais usadas pelos melhores jogadores de cada clube

Agora, para analisar a relação da nacionalidade, figura 4.3 com a reputação internacional, foi utilizado o gráfico de violino. Como exitem muitas nações no dataset, essas foram filtradas pelo potencial acumulado de seus jogadores. Assim, os países presentes na visualização são aqueles com o maior potencial acumulado, respectivamente. Ao contrário do que se esperava inicialmente, a nacionalidade não tem nenhuma relação direta com a reputação do jogador. O que é facilmente justificado pelo simples fato de que a nacionalidade de alguém não define suas habilidades no futebol.

Como as nações não apresentam relação com a nacionalidade, foi adicionada a visualização que relaciona a reputação com os clubes, figura 4.3. Assim como anteriormente, os clubes foram filtrados e ordenados pelo potencial acumulado dos jogadores. O resultado foi bastante distinto. Nesse caso, o clube tem relação com as habilidades do jogador e podemos ver a distribuição da reputação internacional dos clubes apresentados.

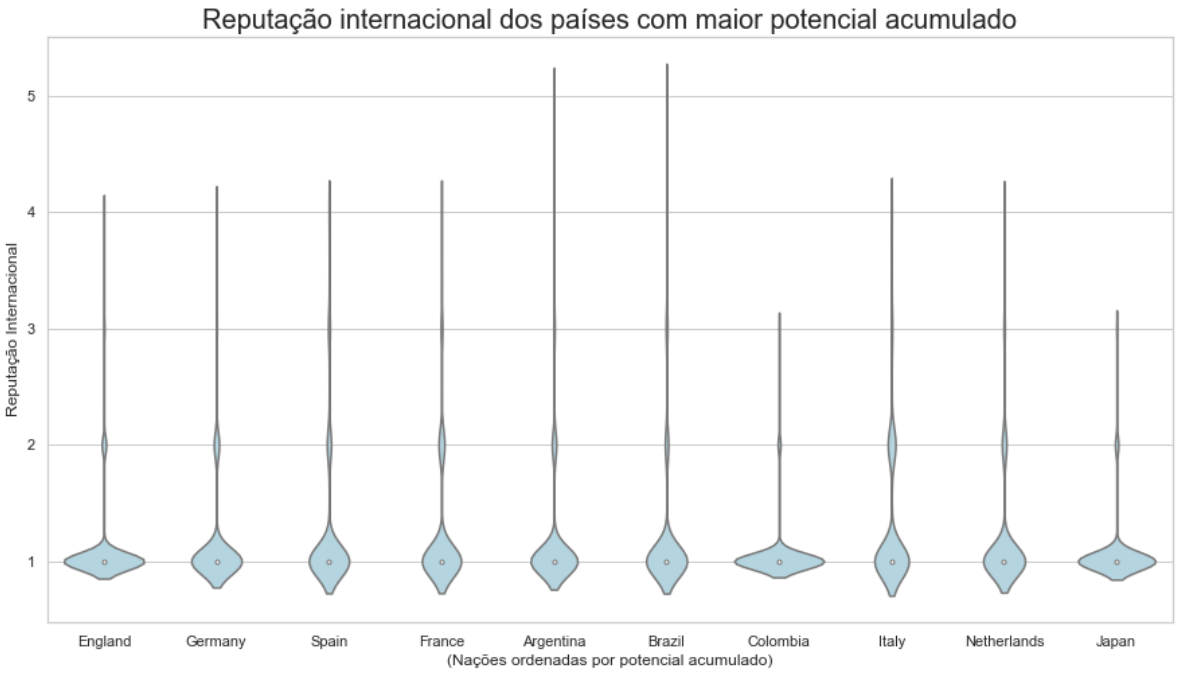


Figura 9 – Reputação Internacional das nações

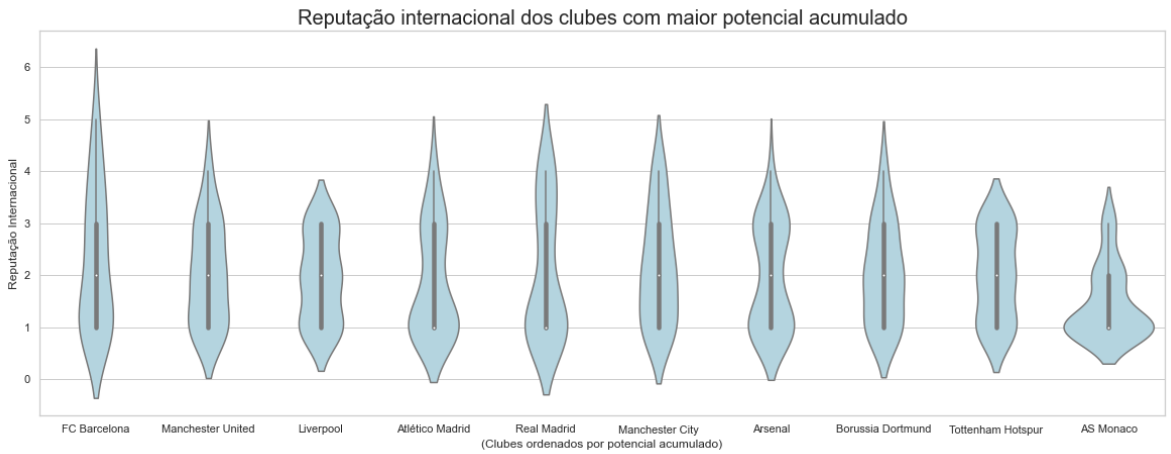


Figura 10 – Reputação Internacional dos clubes

Por fim, para relacionar as habilidades com as posições foram criados 6 visualizações semelhantes, em formato de "treemap". Estas apresentam quais habilidades são mais importantes para cada tipo de função do jogador - divididas entre ataque, defesa e meio de campo. Para tal, foi calculada a média das habilidades dos melhores e piores jogadores de cada categoria (novamente, de acordo com o "Overall", dessa

vez máximo e mínimo) e plotadas no formato escolhido, como exemplificado pelas figuras 4.3 e 4.3. As demais visualizações encontram-se na pasta correspondente.

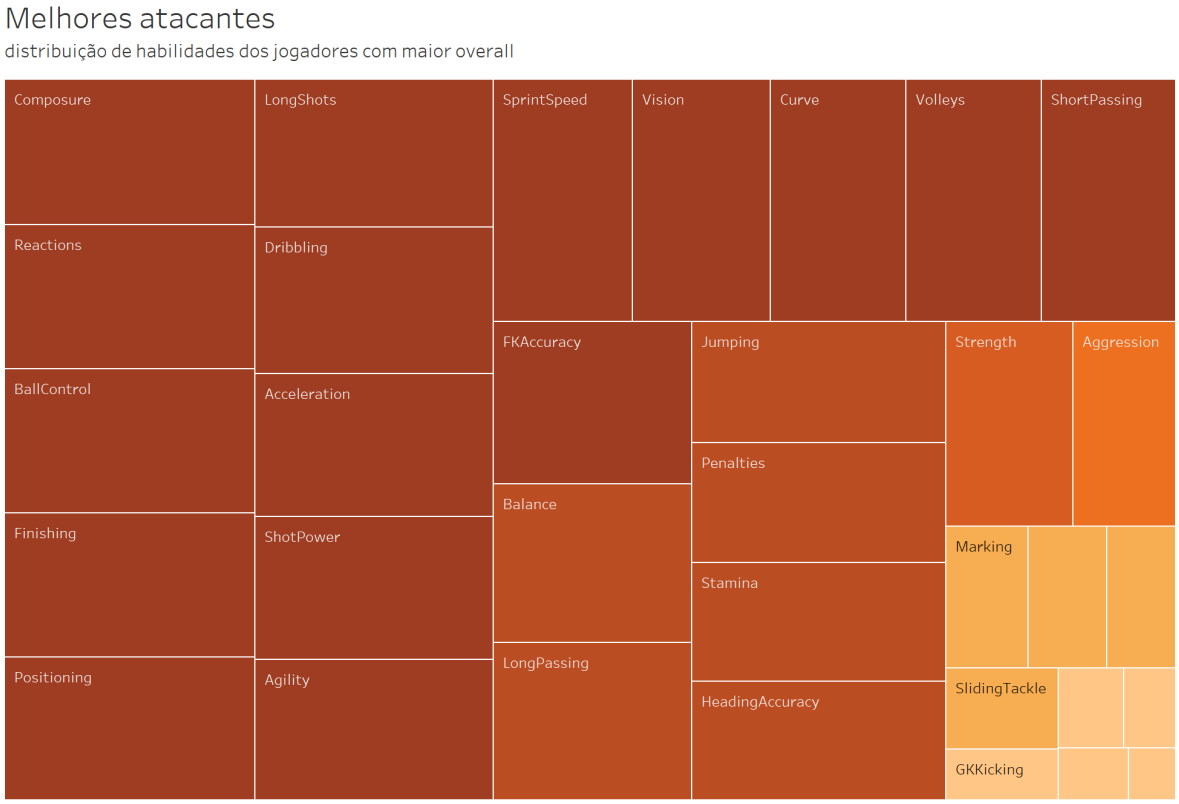


Figura 11 – Distribuição das habilidades dos melhores atacantes

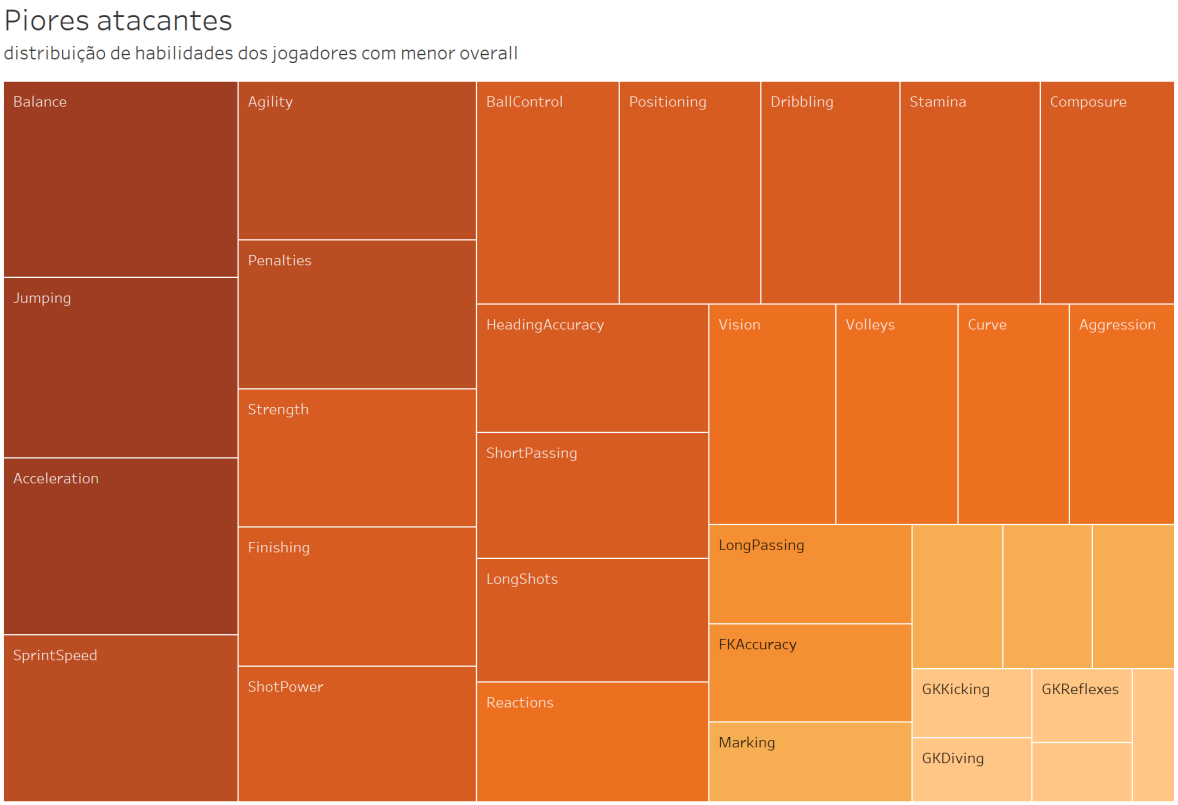


Figura 12 – Distribuição das habilidades dos piores atacantes

4.4 Modelos

A melhor escolha de modelo para previsões considerando várias features foi a regressão linear, uma vez que a regressão logística não atende o que foi pretendido resolver, de maneira precisa, já que o conjunto de dados é contínuo e o objetivo não é classifica-los, mas sim prever o valor do "Overall".

Para determinar as variáveis mais relevantes para o modelo, foi construído um gráfico que demostra a importância de cada uma delas. A partir desse, foram separados dois conjuntos: um com as características gerais do jogador (sua altura, peso, idade,etc) e outra com suas habilidades (aceleração, precisão de chute,etc)

Assim, foram construídos dois modelos distintos, com o mesmo objetivo: determinar o Overall do jogador. A avaliação de cada um foi feita segundo as métricas descritas:

- Coeficientes: São os pesos das features do modelo
- Intercept: O termo independente da regressão
- R2: Quanto mais próximo de 0, menos o modelo é preciso sobre a variabilidade dos dados, e quanto mais próximo de 1 melhor explica a variação entre as variáveis
- Erro Médio Absoluto: É a soma das distâncias dos pontos à reta(erros) dividido pela quantidade de pontos
- Erro Quadrático Médio: É o quadrado da soma das distâncias dos pontos à reta(erros) dividido pela quantidade de pontos

A avaliação, presente no módulo *fifa model.py*, permite concluir que a escolha entre os dois conjuntos de features é indiferente. Para qualquer uma delas o modelo é preciso e coerente para determinar o Overall, como desejado.

5 Conclusão

A partir do desenvolvimento desse trabalho, conseguimos entender melhor os desafios reais de analise de dados e construção de modelos baseados em um conjunto restrito de dados.

5.1 Considerações Finais

A divisão em tarefas determinada para esse trabalho trouxe novos embates para a realizazção do trabalho, bem como benefícios. A divisão proposta fez com que algumas partes dependessem de outras, assim, não poderiam começar a desenvolver sem que o outro já tenha terminado ou, ao menos, avançado em seu papel. Por outro lado, a organização do trabalho ficou mais clara e cada mebro da equipe pode focar na sua parte específica, no que diz respeito a escolha de ferramentas, estilo de solução e pesquisas.