Fundação Getulio Vargas
Escola de Matemática Aplicada

Relatório da Avaliação 2 de Linguagens de Programação

Michelli Almeida

Carlos Fonseca

André Costa

8 de dezembro de 2020

1 Introdução

Este trabalho tem por objetivo responder alguma perguntas de négocios, formuladas pelo próprio grupo, cujas respostas demandam a análise e a modelagem de dois bancos de dados, os quais serão apresentados em duas partes desse documento. A primeira parte tratará do banco de dados que compara alguns fatores com o valor médio das residências em subúrbios de Boston (EUA), alguns desses fatores serão usados para modelar a correlação que eles possuem entre si e , talvez, traçar uma expectiva de preço dos imovéis. A segunda parte tratará de tentar prever o valor da contratação de alguns

jogadores pela base de dados do jogo FIFA.

Para essa análise, utilizamos o banco de dados disponibilizado pelo professor da disciplina,o *Real States Values (RSV)* e o *FIFA19*. Dividimos as tarefas de cada aluno conforme nos foi repassado também pelo professor: Carlos Fonseca assumiu o papel 1, de Cientista de Dados; André Costa foi o aluno resposável pelo papel 2, o de Engenheiro de Dados e Michelli Almeida foi a Especialista em Visualização de Dados.

2 Banco de Dados: Real States Values

Na primeira parte do trabalho, após nos conectarmos à base de dados dos subúrbios de Boston,- realizada por meio do arquivo *SQLConnection.ipynb* - foi necessáio de tratar e limpar essa base - arquivo *LimpandoBD.ipynb* - pois havia algumas inconsistências nas colunas, por exemplo: remoção de entradas com dados inexistentes e alteração na coluna "CHAS"de T/F para 1/0. Após isso, obtivemos os dados para as respostas de 3 perguntas que surgiram.

Siglas e significados dos dados coletados da base:

- CRIM = Taxa de criminalidade
- ZN = Proporção de terrenos residenciais zoneados para lotes com mais de 25.000 pés quadrados
- INDUS = Proporção de industrias
- CHAS = Se há(T) ou não(F) rio limitando
- NOX = Concentração de Oxido Nitrico(PPM)
- RM = Média de quartos por residencias
- AGE = Proporção de unidades ocupadas pelo proprietário construídas antes de 1940
- DIS = Distância até o emprego
- RAD = Índice de acessibilidade a rodovias radiais

- TAX = Taxa de imposto sobre a propriedade
- PTRATIO = $\frac{aluno}{professor}$
- B = Proporção de pessoas negras
- LSTAT = Status inferior da população
- MEDV = Valor médio das casas

2.1 Existe índicios de relação entre os índices de Óxido Nítrico(NO) e a proporção de indústrias em um mesmo subúrbio?

Para responder a essa pergunta, usaremos o modelo de Regressão Linear, na qual a concentração de NO será a variável dependente, ou seja, o eixo y e a proporção das empresas será a variável independente, o eixo x.

```
x= pd.DataFrame(data = df["INDUS"])
y = df["NOX"]

x_treino1, x_teste1, y_treino1, y_teste1 = model_selection.
    train_test_split(x,y, test_size=.4,random_state=1)
```

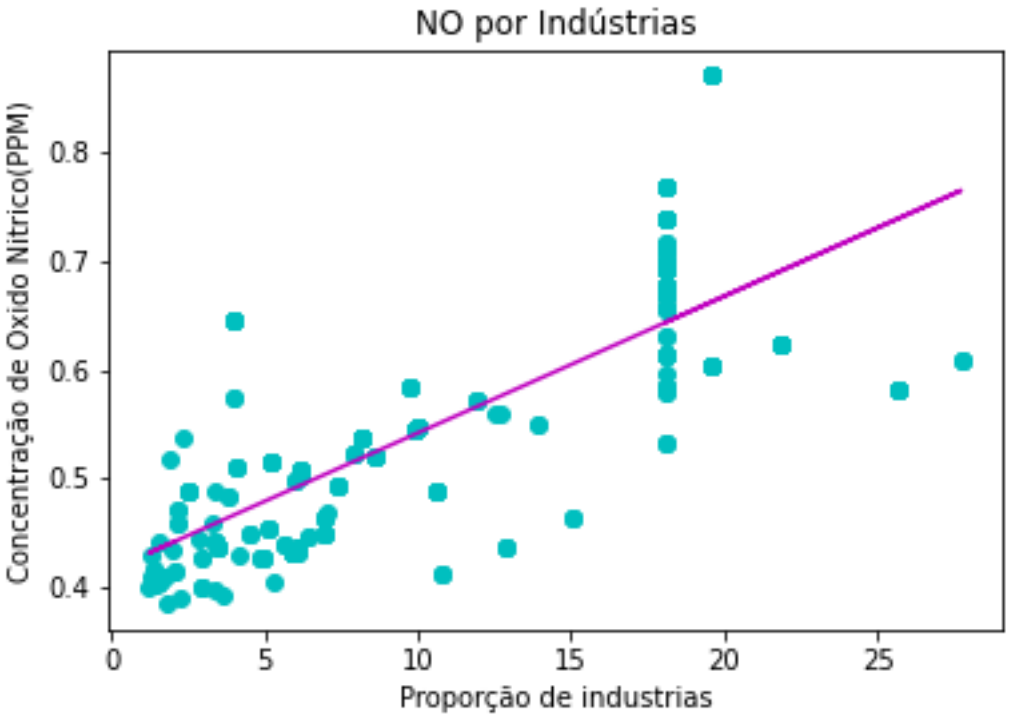
```
model1 = linear_model.LinearRegression()
model1.fit(x_treino1, y_treino1)
```

Contudo, o principal fator que determinará a resposta para esta pergunta é conhecido como R^2, o índice de correlação entre dois conjuntos de dados, no código, ele está expresso por:

```
print(model1.score(x_treino1, y_treino1))
```

$R^2 = 0.5782165624879906$

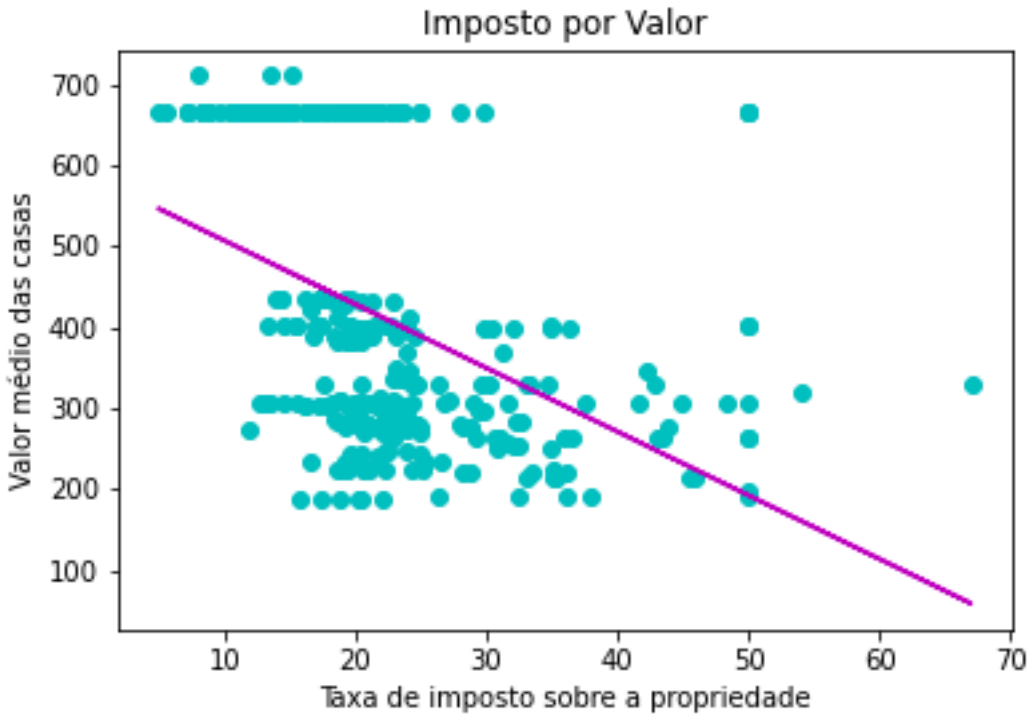
Ou seja, o grafico será esse:



Esse gráfico e o valor de R^2 demonstram haver uma correlação moderada entre os dados analisados, com 50%. Certamente, devem haver muitas outras variáveis que interferem na concentração de nitrogênio em um subúrbio e o alta taxa de gases produzidos por algumas indústrias deve ser um deles, contudo, matematicamente, não dá para afirmar que esses gases e as indústrias são os principais agentes emissores.

2.2 Há indicios de existir relação entre a taxa de imposto cobrada sobre o terreno e o valor dos terrenos?

Responderemos a essa pergunta com a ajuda do modelo de Regressão Linear, do mesmo modo que a pergunta anterior, na qual a taxa de impostos sobre a propriedade será a variável dependente, ou seja, o eixo y e a o valor médio dos imóveis será a variável independente, o eixo x.



$R^2 = 0.20342861999687656$

Esse gráfico e o valor de R^2 demonstram haver uma correlação bem fraca entre os dados analisados, apenas de aproximadamente 20%. O que podemos analisar a partir da imagem é que a relação entre eles é, aparentemente, inversa, além de que para $x = (10, 50)$ há grande concentração em $y = (200, 450)$. Isso significa que tanto as taxas cobradas de impostos quanto os valores de imóveis possuem uma certa moda, a maioria dos imovéis está em um valor médio acessível para os possiveis compradores e as taxas de impostos também, pois não podem ser abusivas.

2.3 É possivel prever o valor do imóvel dadas as outras variáveis?

Como foi mostrado no ínicio da primeira parte, o banco de dados *Real States Value* possui ao todo 14 conjuntos de dados a serem analisados. Será que a

soma de todos os outros 13 fatores pode auxiliar na previsão do valor médio dos imóveis em cada subúrbio?

Para isso, foram compiladas todas as outras 13 colunas de dados em uma só variável, que será a variável independente, e o preço estimado será a variável dependente. Porém, o que realmente será analisado é a previsão dos preços médios pela realidade deles. Ao plotar os dados, obtemos:

$R^2 = 0.6888198464951453$

Com um R^2 de 68 % pode-se dizer que é uma correlação moderada a forte, e isso pode ser explicado pela demanda dos imóveis por subúrbio. Isto é, quando uma família pensa em comprar uma casa, o primeiro passo é pesquisar bastante sobre um local ideal para a compra. Esse local ideal é determinado por vários fatores, dentre os mais importantes explicitados aqui, como: taxa de criminalidade, distância até o emprego, status inferior da população, e etc. Quando os fatores negativos são altos, a procura por imóveis naquele local diminui e, consequentemente, na tentativa de se tornarem mais

atraentes os preços dos imóveis diminuem também. Logo, há uma série de informações a se levar em consideração para determinar o valor de uma casa, e algumas delas estão correlacionadas pelo gráfico.

3 Banco de Dados: FIFA19

Na segunda parte do trabalho, após nos conectarmos à base de dados do jogo FIFA para console,- realizada por meio do arquivo *SQLConnection-FIFA.ipynb* - foi necessáio de tratar e limpar essa base - arquivo *LimpandoBDFIFA.ipynb* - pois havia algumas inconsistências nas colunas, por exemplo: entradas nulas em "Club", "Loaned From"e 'Release Clause"; remoção de aproximadamente 50 jogadores, os quais tinham muitos dados ausentes; remoção de colunas que não seriam aproveitáveis e possuiam muitos dados ausentes; remoção dos cifrões e unidades das colunas sobre dinheiro; criação de uma coluna indicando qual região do campo determinado jogador atua.

Vale ressaltar algumas observações relacionadas aos nomes das colunas de dados. 'Marking', que traduzindo fica 'Marcação',relaciona-se com o fato de que em um jogo de futebol qualquer jogador de defesa tem que ser melhor marcando que um jogador de ataque, afinal o jogador de defesa deve impedir o adversário de fazer um gol. Já o atacante joga para utrapassar essa defesa e marcar o gol. Então, o peso de cada variável para jogadores de diferentes posições será diferente também. Pensando no caso do goleiro, acredita-se que as variáveis que começam com GK são as importantes para a análise.

Após isso, obtivemos os dados para as respostas de 5 perguntas que surgiram.

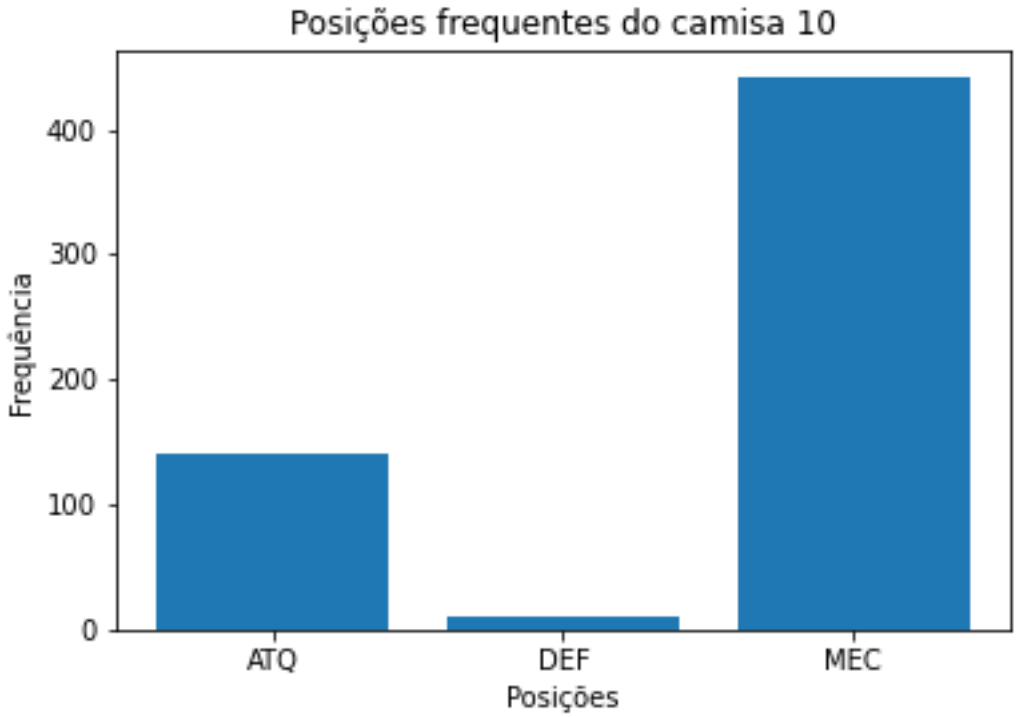
3.1 É familiar para um telespectador assíduo de jogos de futebol que o camisa 1 geralmente é o goleiro. Então, segundo os dados, qual seria posição em que os camisas 1 jogam?

Para responder, primeiro, vamos separar os camisas 1 de todos os outros números, e em seguida agrupá-los por posição, seguindo um histograma.

Como pudemos ver pelo histograma, a posição com mais ocorrências e única é, realmente, a de goleiro, que é ocupada por mais de 500 Número 1's.

3.2 É de consenso geral que frequentemente o melhor jogador do time é o camisa 10. Segundo os dados, qual é a posição mais frequente do camisa 10?

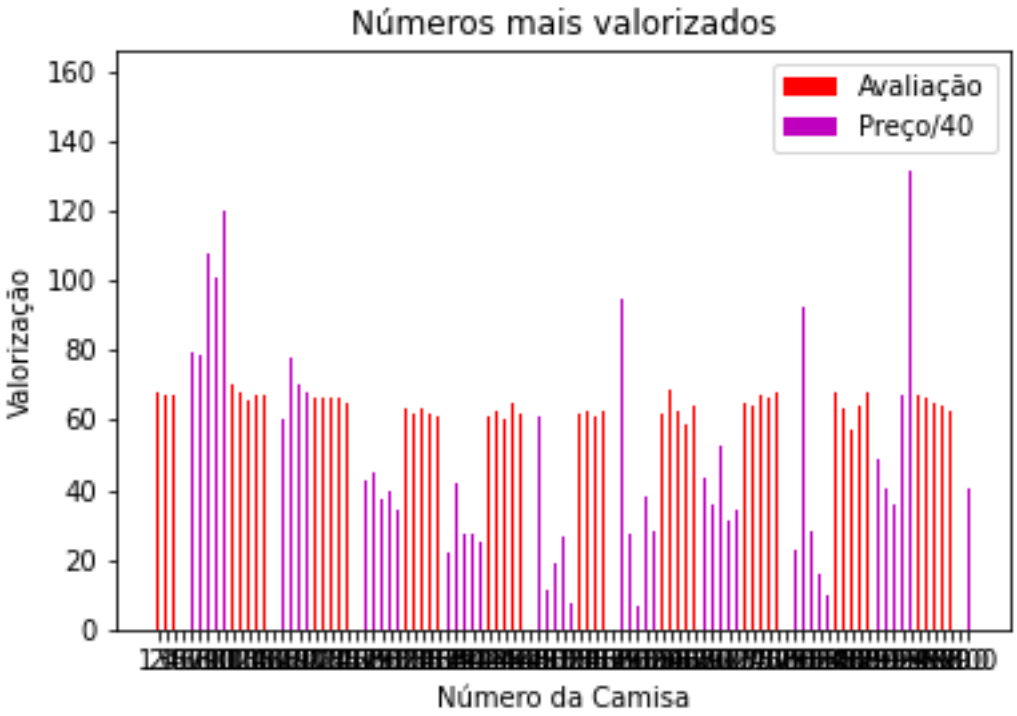
Para responder essa pergunta, teremos que fazer um processo similar ao feito na questão acima e plotar os dados em um histograma.



Pelo gráfico, é observável que ao contrário do Camisa 1, as posições do Camisa 10 variam, contudo, com uma larga diferença a maioria dos camisas 10 atuam como meio de campo. Deese modo, podemos concluir que o meio é lugar mais valorizado do futebol.

3.3 Qual é o número que tem melhor avaliação? E o maior preço de contratação?

Primeiramente, vamos pegar a tabela de dados e agrupá-los por preço e por número de camisa, após isso, iremos fazer a média da avaliação e do preço de cada número.

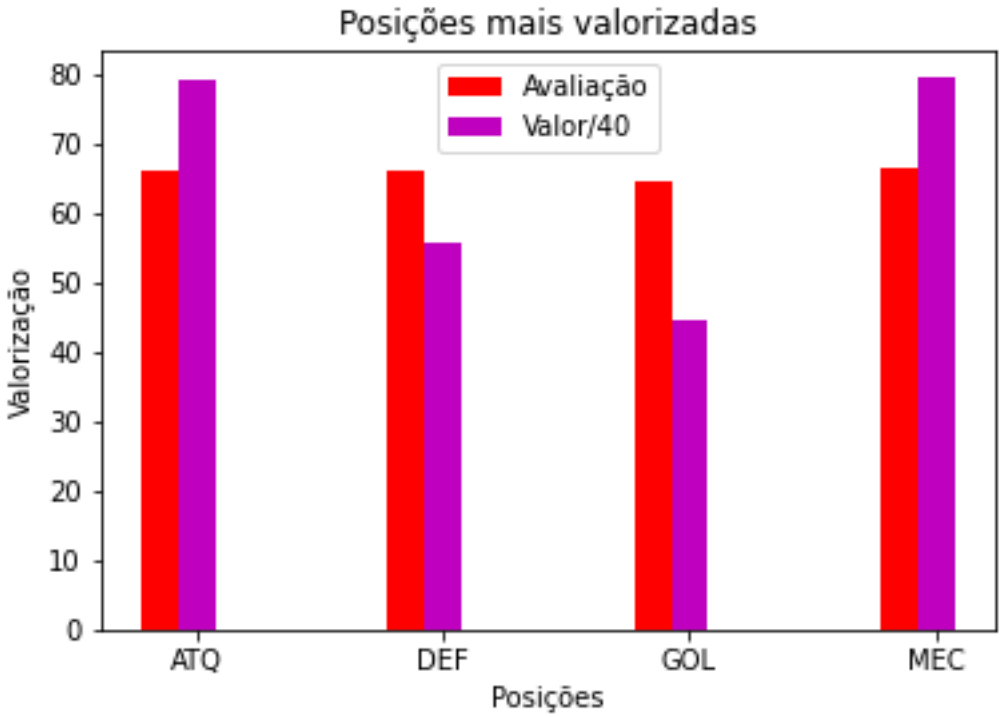


Pelo histograma não conseguimos analisar muitos dados e nenhuma resposta torna-se óbvia, entretando, ao calular por meio do *JupyterNotebook*, conseguimos obtê-la. Portanto, o número de camisa em que a média dos jogadores possuem a melhor avaliação é o número 79, todavia esse número é pouco usado, o que permite uma grande importância para a camisa 10, que é amplamente utilizada.

Já sobre os preços de contratação, os jogadores que usam as camisas 10 realmente levam vantagem, visto que, para telespectadores assíduos de futebol, o camisa 10 geralmente é o artilheiro do time.

3.4 Existe alguma posição que os jogadores são mais valorizados?

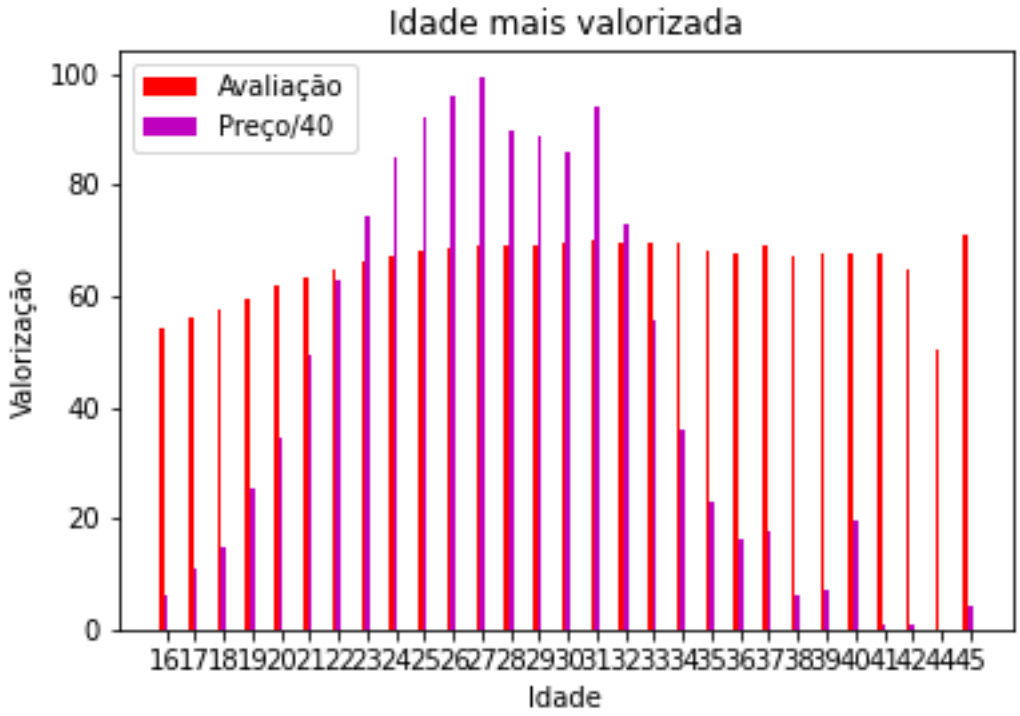
A solução para essa questão pode ser desenvolvida de modo análogo a questão anterior, contudo, em vez de utilizarmos o número que cada jogador usa, utilizaremos a região do campo em que ele joga.



Aparentemente, pelo gráfico de barras, as avaliações dos jogadores estão bem próximas, em um intervalo em aproximadamente $(65, 70)$, logo não há como concluir o melhor avaliado. Já sobre o preço da contratação, podemos concluir que a posição de goleiro é a mais desvalorizada entre as demais, enquanto que a posição de defensor está, aproximadamente, na média, e as posições de atacante e de meio de campo são as que possuem mais valor de contratação. Isso confirma as expectativas para essa resposta, dado o que obtivemos das questões 2 e 3, as quais afirmam que os camisas 10, geralmente, concentram-se no meio de campo e que os camisas 10 são os jogadores de maior preço de contratação.

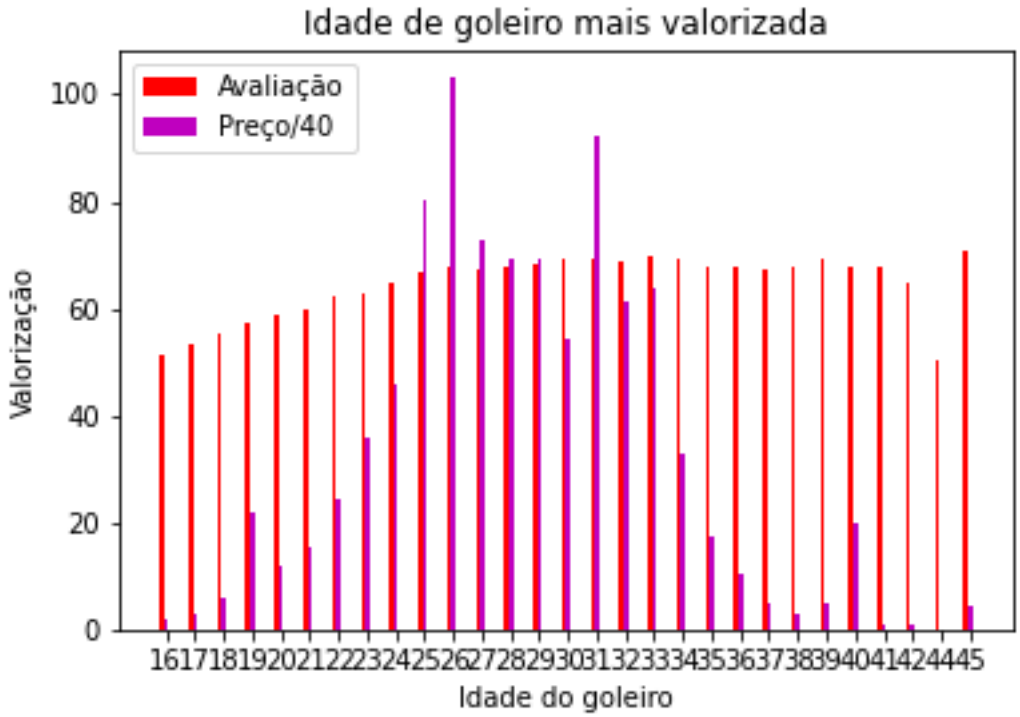
3.5 Em qual idade os jogadores são mais valorizados?

Da mesma maneira que as questões anteriores foram solucionadas, vamos agrupar os jogadores por idade e medir o nível de valorização pela avaliação de cada idade e pelo preço médio de contratação.



Vemos que os jogadores que estão nos extremos da carreira valem menos: os mais novos, por se supor a falta de experiência; e os mais velhos, por se supor a falta de disposição física. Nesse sentido, o topo da carreira acontece por volta dos 24 a 31 anos. A avaliação dos jogadores tambem segue, aparentemente, o mesmo criterio.

Vamos um pouquinho além, vamos observar o caso especifico de goleiros, que geralmente possuem as carreiras mais longas.



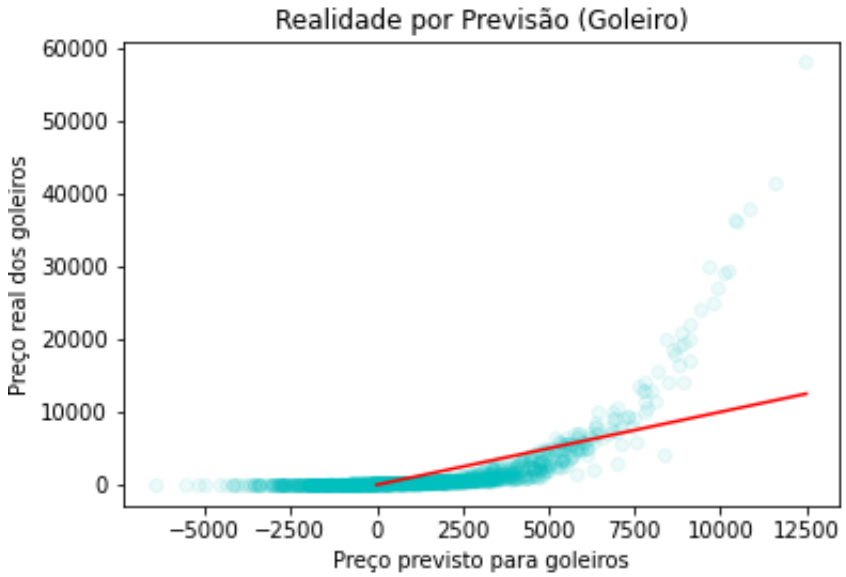
No gol, observamos que o jogador demora mais tempo para atingir o ápice que, na maioria das vezes, acontece aos 28 anos. Esse ápice mantem-se ao longo dos anos, atingindo seu fim por volta dos 35 anos. Claro, há jogadores exceções que podem continuar em alto nivel por mais tempo, um exemplo disso é que os goleiros com idade de 39 anos tem uma avaliação tão boa quanto os outros jogadores no ápice. Já analisando a faixa étaria de 45 anos, deve existir uma única exceção, o que explica os altos niveis de avaliação e os de valor.

3.6 Há alguma forma de prever matematicamente o preço da contratação de um jogador?

Como vimos nas questões acima, diferentes jogadores em diferentes regiões do campo tem diferentes valores, pois possuem diferentes paramêtros. Portanto, para almejar prever os valores dos jogadores, vamos dividi-los com base nas regiões que eles jogam.
Dadas as colunas dos fatores, a soma de todas elas, exceto a do preço de cada

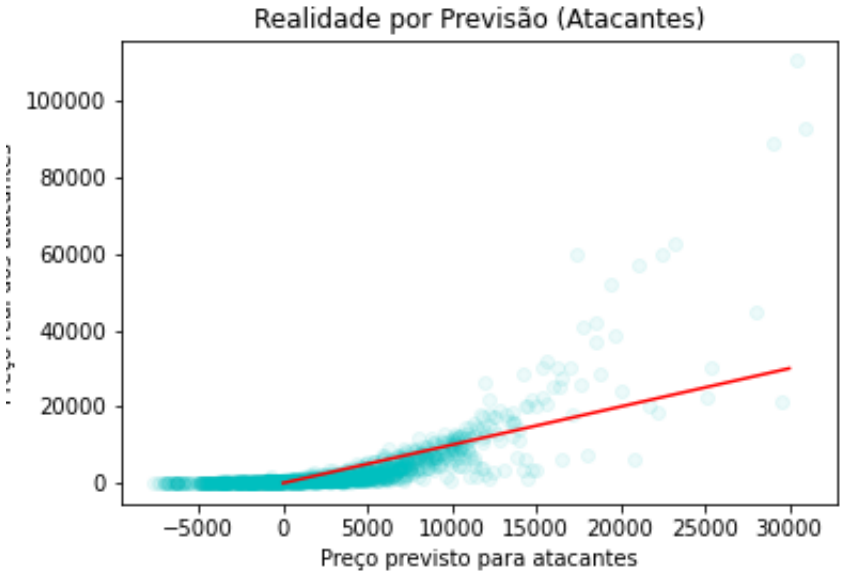
jogador, resultará em uma previsão de preço para cada jogador,que será o valor do x. O valor do y será o preço real de cada jogador.

- GOLEIROS



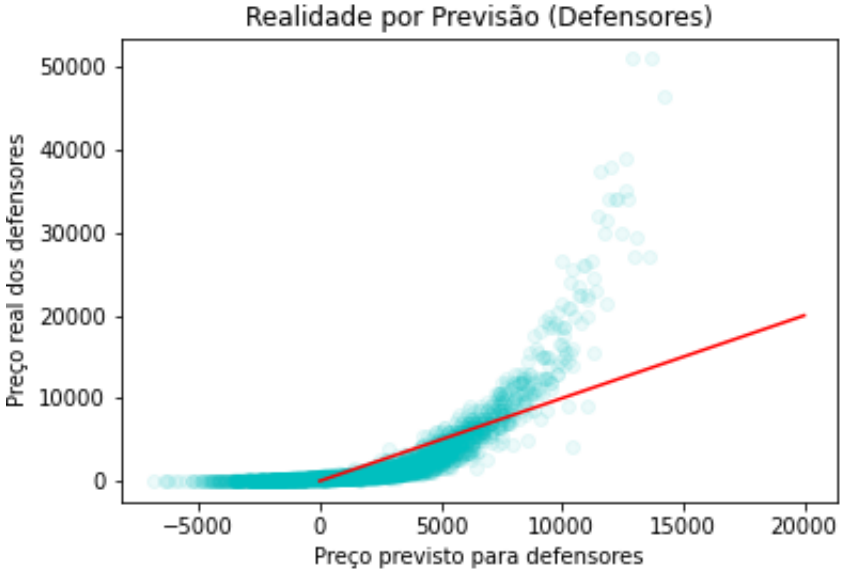
$R^2 = 0.43813365434363527$

- ATACANTES



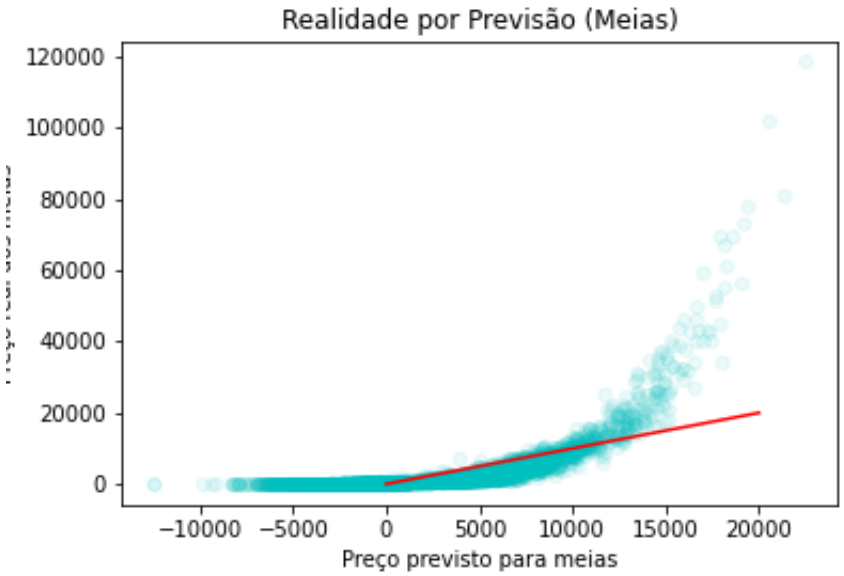
$R^2 = 0.5162631393108521$

- DEFENSORES



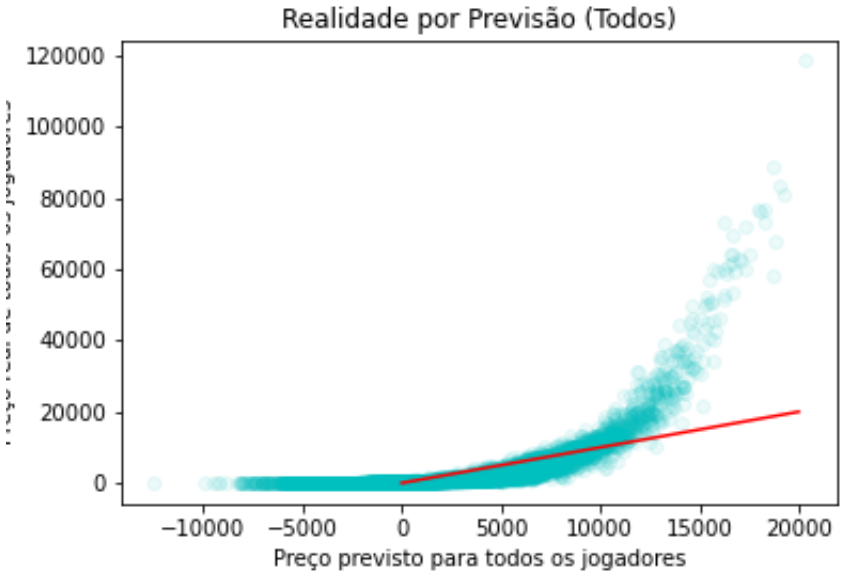
$R^2 = 0.5624281066954637$

- MEIAS



$R^2 = 0.4923064248584068$

- TODOS OS JOGADORES



$R^2 = 0.4758152997803411$

Analisando os gráficos e R^2 de cada categoria, podemos perceber o coeficiente de correlação no intervalo de $(0.43, 0.56)$, que é uma variação significativa. Entretanto, nesse intervalo, o R^2 é considerado fraco a moderado, ou seja, deve haver mais fatores internos e externos que determinam o preço de cada jogador que os analisados aqui. Matematicamente, a previsão mais correta é a da posição do defensor, enquanto que a pior é a da posição do goleiro - única cujo R^2 é menor que o R^2 do gráfico de todos os jogadores.