UNIVERSIDADE FEDERAL DA PARAÍBA - UFPB
CENTRO DE CIÊNCIAS SOCIAIS APLICADAS - CCSA
PROGRAMA DE PÓS GRADUAÇÃO EM ECONOMIA - PPGE

JOÃO ANDERSON DA SILVA FELIX

Coleta de preços *on-line* de forma automática e estimação do custo da cesta básica: uma aplicação para o município de João Pessoa

JOÃO PESSOA – PB
2023

JOÃO ANDERSON DA SILVA FELIX

Coleta de preços *on-line* de forma automática e estimação do custo da cesta básica: uma aplicação para o município de João Pessoa

Dissertação apresentada ao Programa de Pós-Graduação em Economia da Universidade Federal da Paraíba - UFPB, em cumprimento às exigências de conclusão do Curso de Mestrado em Economia.

Orientador: Prof. Dr. Cássio da Nóbrega Besarria
Coorientador: Prof. Me. Diénert de Alencar Vieira

JOÃO PESSOA – PB
2023

Catalogação na publicação
Seção de Catalogação e Classificação

F316c Felix, João Anderson da Silva.
Coleta de preços on-line de forma automática e estimação do custo da cesta básica : uma aplicação para o município de João Pessoa / João Anderson da Silva Felix. - João Pessoa, 2023.
38 f. : il.

Orientação: Cássio da Nóbrega Besarria.
Coorientação: Diénert de Alencar Vieira.
Dissertação (Mestrado) - UFPB/CCSA.

1. Previsão de preço. 2. Cesta básica. 3. Web-scraping. 4. Séries-temporais. I. Besarria, Cássio da Nóbrega. II. Vieira, Diénert de Alencar. III. Título.

UFPB/BC CDU 658.8.03-047.72(043)

Elaborado por WALQUELINE DA SILVA ARAUJO - CRB-15/514

JOÃO ANDERSON DA SILVA FELIX

COLETA DE PREÇOS ON-LINE DE FORMA AUTOMÁTICA E ESTIMAÇÃO DO CUSTO DA CESTA BÁSICA: UMA APLICAÇÃO PARA O MUNICÍPIO DE JOÃO PESSOA

Dissertação apresentada ao programa de Pós-Graduação em Economia da Universidade Federal da paraíba – UFPB, em cumprimento às exigências de conclusão do Curso de Mestrado em Economia.

Submetido à apreciação da banca examinadora, sendo aprovado em 25/02/2022.

gov.br Documento assinado digitalmente
Cassio da Nobrega Besarria
Data: 21/06/2022 15:52:27-0300
Verifique em https://verificador.iti.br

Prof. Dr. Cássio da Nóbrega Besarria
Orientados

Prof. Me. Diénert de Alencar Vieira
Coorientador

Prof. Dr. Jevuks Matheus de Araújo
Avaliador Interno

Lucas Lúcio Godeiro
Assinado de forma digital por Lucas Lúcio Godeiro
Dados: 2022.06.03 17:14:45 -03'00'

Prof. Dr. Lucas Lúcio Godeiro
Avaliador Externo

JOÃO PESSOA – PB
2022

Sumário

Lista de ilustrações

Lista de tabelas

Resumo

A coleta de informações de forma automática por meio de dados disponíveis na *internet* vem crescendo continuamente e rapidamente para uma ampla gama de produtos e serviços, incluindo a coleta de preços de produtos alimentícios. Esta disponibilidade de preços sobre produtos *on-line* abre a possibilidade para que possamos acompanhar a variação dos preços destes produtos e verificar o comportamento da inflação de uma determinada região de forma automática, por meio de tecnologias como do tipo *web-scraping*, gerando uma redução nos custos da pesquisa e obtendo os dados de forma mais rápida. Desta forma, para aproveitar da maior rapidez e economia de recursos que a coleta de dados *on-line* pode proporcionar, foi criado um algoritmo que coleta dados sobre os preços relativos a um conjunto de itens de uma determinada cesta básica para a região metropolitana de João Pessoa. Além disso, foi aplicado um conjunto de quatro algoritmos que fizeram a previsão do preço desta cesta de produtos. Os resultados indicam que o custo da cesta básica estimado com preços coletados de forma *on-line*, acompanharam a tendência dos dados gerados pelo órgão de estatística oficial de referência. Assim, as descobertas mostram que os dados coletados de forma *on-line* podem ser uma alternativa confiável para o acompanhamento de preços de produtos e podem ser usados para o acompanhamento da inflação de forma mais barata e rápida, se comparada aos métodos de coletas tradicionais.
Palavras-chave: web-scraping. cesta-básica. séries-temporais. previsão.

Abstract

The collection of information automatically through data available on the internet has been continuously and rapidly growing for a wide range of products and services, including the collection of prices of food products. This availability of prices on on-line products opens the possibility to monitor the price variation of these products and automatically check the inflation behavior of a given region, through technologies such as web-scraping, generating a reduction in research costs and obtaining more fast de data. Thus, to take advantage of the greater speed and economy of resources that on-line data collection can provide, an algorithm was created, which automatically collects data on prices related to a set of items from a certain basic food parcel for the metropolitan region of João Pessoa. In addition, a set of four algorithms was applied to predict the price of this basic food parcel. The results indicate that the cost of the basic food parcel, estimated with prices collected in an on-line way, satisfactorily followed the estimates made by the official statistical agency of reference. Thus, the findings show that data collected online can be a reliable alternative for tracking product prices, and can be used to track inflation in a cheaper and faster way, compared to traditional collection methods.
keywords: web-scraping. basic-food-parcel. time-series. forecasting.

1 Introdução

A coleta de informações de forma automática por meio de dados disponíveis na *internet* vem crescendo continuamente e rapidamente para uma ampla gama de produtos e serviços, incluindo a coleta de preços de produtos alimentícios. Junto a esta maior disponibilidade de preços sobre produtos em aplicativos ou *web pages*, também há novas formas de coleta de dados por meio de tecnologias como *web scraping*, criando-se, desta forma, possibilidades para que se possa construir novos índices de preços ou seguir metodologias de órgãos estatísticos oficiais com dados disponíveis *on-line*. Os métodos tradicionais de coleta de dados, como a pesquisa presencial, podem demandar elevados níveis de recurso e tempo. Devido a isso, muitas vezes temos um conjunto de dados com baixa frequência, *delay* na apresentação da informação e problemas com eventuais *outliers*, sujeitos a dias ou horários de coleta. Alguns estudos buscam contornar essas eventuais limitações utilizando novas metodologias para a coleta de dados, tais como: *web scraping* (ou raspagem de dados).

Kienle, German e Muller (2004) definem a tecnologia *web scraping* como uma técnica de raspagem que combina extração e análise para obter conteúdo de páginas formatadas em *HTML* (*HyperText Markup Language*). Ao empregar esta tecnologia para a coleta de dados, duas vantagens são destacadas na sua utilização: menor custo e maior rapidez na obtenção dos dados.

As vantagens da coleta de dados em sua forma automática é documentada em Hillen (2019). O menor custo na utilização dessa técnica pode ser encontrado na não utilização de entrevistadores (pesquisa presencial), ou na não necessidade de se pagar por uma API (Application Programming Interface) de banco de dados para se obter informações desejadas. Outra vantagem é obtida quando a rotina de extração de dados é construída, logo, fica a critério do pesquisador qual a frequência que os dados serão coletados (diário, semanal, mensal, trimestral etc), de forma automática, diminuindo assim o tempo necessário para a coleta das informações.

Contudo, há algumas desvantagens, como a falta de dados históricos, portanto, caso seja preciso fazer uma análise de algum problema de interesse, o pesquisador terá que iniciar por conta própria a coleta dos dados, caso estas informações ainda não existam. Outra desvantagem encontrada na coleta sobre preços de produtos, é a falta de informação de quantas vezes determinado produto foi visualizado e/ou adquirido, ver Hillen (2019) e Cavallo (2018). Quando não há essa informação, por exemplo, não há a possibilidade de ser criado um peso que cada produto pode ter ao se estimar o custo de uma cesta de bens.

Vale ressaltar que, apesar de abrir-se a possibilidade da obtenção dos dados a um menor custo, maior agilidade e de uma ampla utilização em diversos campos de pesquisa,

a utilização de *web scrapers* tem suas limitações legais. Os robots.txt, por exemplo, são arquivos disponibilizados por donos de *webpages*, que fornecem instruções a bots (aplicações que simulam a ação humana, também conhecidos como testes automatizados) quais ações são ou não permitidas em seus *sites*. Estes arquivos são um exemplo dos limites legais das coletas de dados automáticas na *internet*. Logo, a violação dos limites impostos por esses arquivos, podem desrespeitar regras, gerando assim punições legais.

Dadas as vantagens e desvantagens do emprego dessa tecnologia para a obtenção de dados e tendo uma noção de suas limitações legais, esta ferramenta pode ser aplicado em diferentes áreas de pesquisa. Atualmente, *web scraping* é utilizada para a obtenção de dados para diversos fins, como pesquisa de dados econômicos Edelman (2012), mercado imobiliário de aluguéis Boeing e Waddell (2017), eficiência energética Im et al. (2017), dados meteorológicos Kunang, Purnamasari et al. (2018) entre outros.

Para o acompanhamento de preços, Alvarez e Lein (2020) propõem um índice de inflação diário baseado por categorias de produtos, com dados obtidos por meio de *web scraping* para a Suíça. É encontrado pelos autores que o índice elaborado reflete de maneira satisfatória o índice de inflação oficial mensal. Para a Indonésia, Manik et al. (2015) constroem um índice de inflação diário por meio de dados obtidos por *web scraping*. O autores mostraram que os dados coletados diariamente são relevantes para o acompanhamento das mudanças de curto prazo dos preços.

À nível nacional, Cavallo (2013) faz o acompanhamento das variações dos preços *on-line* por meio da técnica de *web scraping* para países como: Argentina, Brasil, Chile, Colômbia e Venezuela. Para o Brasil, por exemplo, é mostrado que o comportamento dos preços coletados *on-line* acompanham a tendência da inflação oficial.

Diferentemente dos estudos apresentados anteriormente, esse trabalho tem o propósito de acompanhar os custos de uma cesta de produtos alimentícios para o município de João Pessoa, diariamente e de forma automática. Inicialmente, o acompanhamento das alterações nos preços desta cesta de bens será limitado a região metropolitana da capital do Estado da Paraíba[1], podendo ter sua análise ampliada em um momento futuro.

Os bens escolhidos para o acompanhamento e estimação do custo da cesta básica são os indicados pelo decreto lei n° 399/1938 que, entre outras coisas, estabelece quais itens serão considerados para a composição da cesta em cada região do país e a quantidade mínima que cada adulto deve consumir por mês. Ressalta-se que os itens e as quantidades dos produtos estão contidos na tabela 2.

O acompanhamento dos preços de cada item que compõem essa cesta básica foi feito por meio de um algoritmo que coleta dados sobre os preços dos produtos alimentícios, definidos pelo decreto lei n° 399/1938, em supermercados que ofertam produtos *on-line* na

[1] Compreende a região metropolitana de João pessoa: Bayeux, Cabedelo, Conde, Cruz do Espírito Santo, João Pessoa, Lucena, Mamanguape, Rio Tinto, Santa Rita, Alhandra, Caaporã e Pitimbu. - Lei complementar nº. 93 de 11 de dezembro de 2009

região metropolitana de João Pessoa.

Torna-se importante acompanhar o custo da cesta básica de forma diária, pois, a coleta de dados de alta frequência permite acompanhar o movimento instantâneo dos preços, facilitando uma melhor alocação de recursos pelos agentes econômicos como as famílias, firmas e governos.

Um exemplo disso, verificado com dados obtidos neste projeto, foi que a coleta de dados diária nos permitiu verificar em qual turno o custo da cesta obteve a maior média, assim como qual o dia da semana este conjunto de bens estava sendo estimado a um maior preço. No ano de 2021, o turno com o maior custo estimado foi à noite, com o preço de R$ 496,05, já o dia com a cesta estimada com o maior custo foi a sexta-feira com R$ 492,17.

Além do mais, para poder prover mais informações para uma melhor alocação de recursos, torna-se importante divulgar estes dados (cesta estimada *on-line*) e aplicar modelos preditivos para o conjunto de dados em mãos. Para isto, foram, construídas previsões por meio de técnicas de séries temporais, por meio do modelo ARIMA, e modelos de aprendizado de máquina, modelos Ridge, Lasso e Elastic Net, e a elaboração de um aplicativo web interativo para a divulgação desses dados.

2 Motivação

A inflação, definida usualmente como um aumento constante e sustentado dos preços[1], é uma das variáveis macroeconômicas mais importantes, impactando as decisões de diversos agentes na economia como famílias, firmas e governos.

Uma inflação alta diminui o poder de compra das famílias. Um aumento dos preços de forma constante fará com que a quantidade de bens e serviços que podiam ser adquiridos em um período anterior será menor em um período posterior, devido ao aumento dos preços, considerando uma variação nula na renda, assim diminuindo a qualidade de vida da população.

Além da diminuição do poder de compra, há evidências de que uma inflação alta é um fator que aumenta a desigualdade econômica de uma determinada economia Albanesi (2007). Isto porque, pessoas com uma menor renda são mais expostas a diminuição do poder de compra devido ao aumento dos preços. Esta maior exposição pode ser explicada por meio do acesso desigual ao mercado de crédito entre ricos e pobres, isto acontece pela maior possibilidade que pessoas com um maior nível de renda tendem a ter em possuir uma maior cesta de produtos financeiros que lhes protejam de uma alta inflacionária.

Um exemplo de como a inflação impacta de forma mais pesada as famílias de menor renda é demonstrada na tabela 1. Ela mostra a proporção que uma cesta básica, com preço médio de janeiro de 2022 (estimado por esse projeto), com valor de R$ 547,33, representa em relação ao rendimento mensal familiar.

Nesta tabela é mostrada a relação dos 10 bairros de maior renda e os 10 bairros de menor renda familiar média na cidade de João Pessoa [2]. Os valores dos rendimentos são referentes ao valor do rendimento nominal médio mensal dos domicílios particulares permanentes [3], estimado pelo Instituto Brasileiro de Geografia e Estatística (IBGE) no ano de 2010. Estes rendimentos foram atualizados pelos valores do Índice Nacional de Preços ao Consumidor Amplo (IPCA), com base no período de 2011-2020.

Tabela 1 – Proporção custo da cesta básica/renda

Ranking	**Bairro**	**Rend. (R$)**	**Prop. (%)**
1	Cabo Branco	15054,39	3,64%
2	Tambaú	13684,97	4,00%

continua na próxima página

[1] Definição utilizada por Friedman (1963).
[2] A relação dos 63 bairros da cidade de João Pessoa estará em anexo
[3] Tabela 3345

Tabela 1 – continuação da página anterior

Ranking	Bairro	Rend. (R$)	Prop. (%)
3	Manaíra	13259,02	4,13%
4	Jardim Oceania	12287,73	4,45%
5	Brisamar	11956,64	4,58%
6	Estados	11785,97	4,64%
7	Ponta Do Seixas	11521,10	4,75%
8	Miramar	11475,72	4,77%
9	Tambauzinho	11329,01	4,83%
10	Pedro Gondim	9887,87	5,54%
54	Mucumago	1795,74	30,48%
55	Padre Zé	1749,35	31,29%
56	Gramame	1720,14	31,82%
57	Grotão	1670,32	32,77%
58	Mumbaba	1636,38	33,45%
59	Jardim Veneza	1600,59	34,20%
60	Distrito Industrial	1569,16	34,88%
61	Alto Do Céu	1528,36	35,81%
62	Ilha do Bispo	1270,11	43,09%
63	São José	1149,09	47,63%

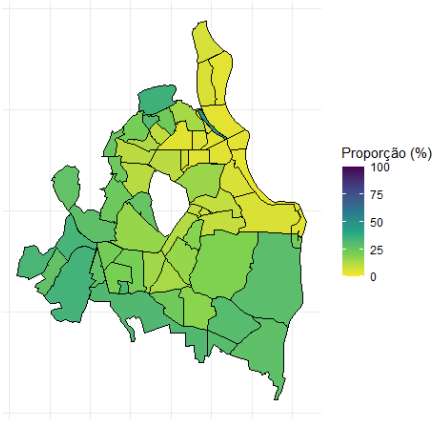
Fonte: Elaboração própria com dados do IBGE

O bairro de maior renda domiciliar, Cabo Branco, uma cesta básica representa apenas 3,64% do rendimento total, por outro lado, o bairro de menor renda, São José, uma cesta básica representa 47,63% do rendimento total familiar. Com esses dados, é possível observar que o bairro de maior renda possui um rendimento familiar médio 13,1 vezes maior do que o bairro de menor renda na cidade de João Pessoa.

Essa desigualdade fica um pouco mais evidente quando é plotado a proporção da cesta básica em relação a renda média familiar para todos os bairros da cidade, como é mostrado na figura 1.

Observa-se que os bairros de maior renda estão concentrados no litoral norte da cidade, enquanto que os bairros mais afastados desta área litorânea possuem uma proporção da cesta básica em relação a renda bem maior, consumindo uma fatia relativamente grande do orçamento e, consequentemente, fazendo com que gastos em outros serviços essenciais sejam diminuídos.

Figura 1 – Mapa proporção cesta básica/renda



Fonte: Elaboração com dados próprios e do IBGE

Observada a importância de acompanhar a evolução do aumento dos preços, este trabalho teve como objetivo principal mensurar o custo da cesta básica, diariamente e de forma automática, utilizando preços disponíveis *on-line* no município de João Pessoa.

Destaca-se a frequência da coleta das informações, dados coletados diariamente permitem acompanhar o movimento instantâneo dos preços, possibilitando uma melhor alocação de recursos dos agentes econômicos, tendo em vista que é possível verificar os turnos/dias em que ocorrem os picos de preços do produtos e consequentemente o preço da cesta básica.

Isto torna-se um beneficio principalmente para as famílias de menor renda, pois com o acesso a mais informação sobre a variação dos preços como em que dia da semana a cesta básica ou os itens atingem o maior preço, abri-se a possibilidade para que estes agentes aloquem da melhor forma possível a sua renda.

3 Metodologia

3.1 Web Scraping

Para a coleta dos preços dos produtos da cesta básica definida, foi construído um algoritmo que de forma automática e diária extrai estas informações em supermercados que ofertam seus produtos na *internet* na região metropolitana de João Pessoa.

A metodologia que permite com que se faça esta coleta de dados de forma *on-line* é conhecida como *web-scraping*. Mitchell (2018) define *web-scraping* como a prática de coletar dados de forma automática, normalmente, por meio de um programa automatizado sem a necessidade desta coleta ser feita por um ser humano.

Cavallo (2018), sintetiza como esta tecnologia pode ser utilizada para a coleta destes dados. Primeiro, um algoritmo seleciona uma lista de páginas *web* onde os dados de interesse são apresentados. Segundo, o código analisado percorre as páginas selecionadas e localiza as partes importantes do código, onde as variáveis de interesse estão inseridas, como preço e descrição dos produtos. Por fim, os dados são coletados e armazenados, contendo um registro por produto, na frequência desejada.

Estes algoritmos ou programas automatizados que percorrem várias páginas *web* são conhecidos como *web crawlers*, eles percorrem as páginas selecionadas e extraem as informações de interesse localizando em qual parte do código estão estas informações relevantes, como preços e descrição dos produtos. Abaixo é mostrado uma estrutura básica de uma página *web* e onde estes algoritmos conseguem localizar e extrair os dados de interesse.

```
<!DOCTYPE html>
     <html>
          <head>
              <title> Título </title>
          </head>
          <body>
              <a class="descrição do produto"
              <a class="preço produto"
          </body>
      </html>
```

A estrutura acima é um exemplo simples de como uma página *web* pode ser construída. Ela exemplifica como a maioria dos sites são criados, por meio da linguagem de marcação estruturada, conhecida como *HyperText Markup Language* (HTML).

Os documentos HTML são constituídos por <*tags*>, que estabelecem a estrutura desses sites por meio de elementos HTML como cabeçalho, parágrafos e links. Já a aparência ou o estilo das páginas são definidas por linguagens de programação como do tipo Cascading Style Sheets (CSS). Um dos principais conceitos da linguagem CSS é o de seletores (*CSS Selectors*), é por meio de um seletor que se aplicam regras CSS, uma regra CSS determina qual o estilo será utilizado em um elemento HTML.

Definida a construção da página *web*, estrutura e estilo, é possível localizar informações como descrição e preços dos produtos. Assim, algorítimos de raspagem de dados coletam dados de interesse de produtos utilizando-se de seletores CSS, ou por meio da linguagem XPath[1], para localizar elementos HTML. Definido os campos de interesse, onde estão localizados os preços e as descrições, basta aplicar a rotina que simula a ação humana e os dados serão coletados na frequência desejada e armazenados em um conjunto de dados.

Desta forma, utilizando essa tecnologia, foi construído um algoritmo por meio de uma linguagem de programação, que diariamente, e de forma automática, coleta dados de produtos em mercados na região metropolitana de João Pessoa, conforme decreto Lei n° 399/1938.

3.2 Analise preditiva para o índice de inflação

Tendo definido a região onde seria aplicado a coleta automática e, definido o conjunto de bens de interesse, também foi feito aplicações de modelos preditivos para a série temporal em questão (custo total da cesta básica). Como o objetivo principal do projeto foi a estimação do custo total da cesta básica utilizando dados coletados de forma automática, a descrição dos modelos preditivos mencionados a seguir não foi feita de forma completa, logo, apenas uma apresentação simples dos modelos escolhidos foi feita nesta seção.

3.2.1 ARIMA

O modelo Auto-Regressivo Integrado de Médias Móveis (ARIMA, sigla em inglês) é um dos modelos preditivos univariados mais aplicados na previsão de séries temporais e, é a junção de um modelo auto-regressivo (AR), uma parte integrada (I) e um modelo de média móvel (MA).

Na parte auto-regressiva, como explica Hyndman e Athanasopoulos (2018), a previsão da variável de interesse é feita realizando uma combinação linear de valores

[1] XPath é uma linguagem de programação que também pode ser utilizada para selecionar os elementos de um arquivo HTML por parte dos algorítimos de buscas.

passados da própria variável, ou seja, a parte auto-regressiva indica que é feita uma regressão da variável consigo mesma.

Desta forma, um modelo auto-regressivo de ordem p pode ser escrito da seguinte forma:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \tag{3.1}$$

Onde y_t é a variável de interesse, c é uma constante, ϕ é um parâmetro, p é a ordem de defasagem e ε_t é o ruído branco. A equação 3.1 é denominada de modelo auto-regressivo de ordem p.

Já o modelo de média móvel usa os erros de previsões passadas para realizar uma regressão. O modelo de média móvel pode ser escrito da seguinte forma:

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \tag{3.2}$$

Onde y_t é variável de interesse, c é uma constante, θ é um parâmetro, q define a ordem do modelo de média móvel e ε_t é o ruído branco. A equação 3.2 é denominado de modelo de média móvel de ordem q.

Por fim, a ordem de integração nos mostra qual a ordem de diferenciação é necessária para tornar a série estacionária.

Assim, como afirma Hyndman e Athanasopoulos (2018), se combinarmos a ordem de integração com um modelo auto-regressivo e um modelo de médias móveis é obtido um modelo ARIMA não sasonal. O modelo completo pode ser escrito da seguinte forma:

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \tag{3.3}$$

Onde y'_t é a série diferenciada referente a variável de interesse (podendo ou não ser diferenciado). A equação 3.3 é denominado uma modelo ARIMA(p,d,q).

Uma vez que o modelo tem a sua identificação realizada, valores de (p,d,q), é preciso estimar os valores dos parâmetros c, ϕ_1 (modelo auto-regressivo) e θ_1 (modelo de média móvel). Uma forma de identificar esta ordem é por meio de critérios de informação, um dos critérios de informações mais conhecidos é o critério de informação de Akaike Akaike (1974), este critério de informação pode ser escrito da seguinte forma:

$$\text{AIC} = -2\log(L) + 2(p + q + k + 1) \tag{3.4}$$

Onde L é a probabilidade dos dados, k = 1 se c $\neq$ 0 e k = 0 se c = 0. Os melhores modelos são aqueles que minimizam os valores de AIC. Contudo, vale ressaltar, que o critério de informação de Akaike pode ser uma boa alternativa para a identificação de (p e q) e tende a não ser uma boa alternativa para a seleção da ordem de integração (d).

3.2.2 Regressões Ridge, Lasso e Elastic Net

Os modelos Ridge, Lasso e Elastic Net podem ser uma alternativa não só ao método dos Mínimos Quadrados Ordinários (MQO), mas também, para a previsão de séries temporais. Estes três modelos, possuem grande semelhança com o MQO, contudo, diferenciam-se pelo benefício da inclusão de viés, para gerar previsões com um menor Erro Quadrático Médio (EQM). Isto porque, sob certas suposições, considera-se que o MQO é livre de viés e é o método de menor variância entre os modelos lineares, porém, isto não garante que o MQO fornecerá as melhores previsões.

Por isso, uma combinação entre variância e viés pode gerar predições com um menor erro quadrático médio, permitindo que os parâmetros sejam tendenciosos. Kuhn, Johnson et al. (2013), argumentam que pequenos aumentos de viés podem diminuir a variância, e consequentemente, gerar um menor EQM, se comparado ao MQO.

Uma forma de criar modelos com um maior balanceamento entre viés e variância, para tentar melhorar a acurácia do modelo, é adicionar uma penalidade a Soma dos Quadrados dos Erros (SQE).

$$SQE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3.5}$$

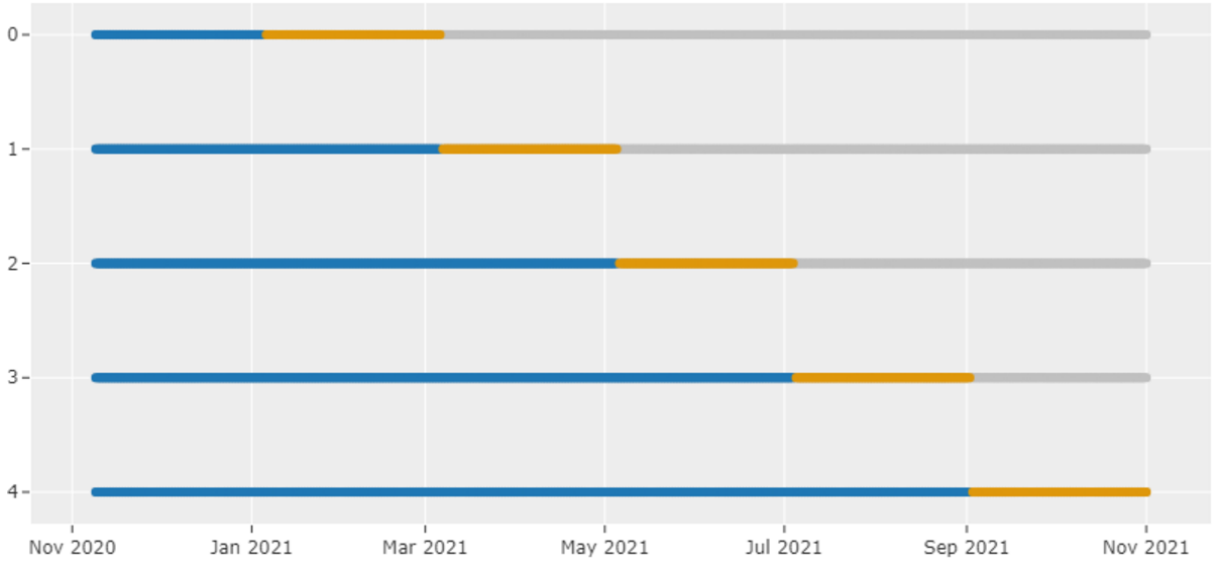
Como estes modelos tentam evitar um sobre ajustamento do modelo aos dados, há uma tentativa de se controlar o tamanho dos parâmetros estimados, por meio de uma diminuição do SQE. O controle, ou regularização, dos parâmetros estimados, pode ser obtido adicionando uma penalização ao SQE se os parâmetros estimados se tornarem grande demais.

Em suma, ao adicionar as penalidades, busca-se um *trade-off* entre viés e variância. Com isto, ao permitir um pouco mais de viés, tenta-se diminuir a variância, ao ponto de torna o EQM menor e, consequentemente, gerar melhores previsões.

Uma das formas de escolher o parâmetro que penaliza estes modelos é por meio da validação cruzada, ao utilizar esta técnica os dados do conjunto de dados são separados em k partes de mesmo tamanho, por exemplo, como foi feito neste trabalho, separar os dados em cinco partes, após isto, k-1 partes são utilizadas para o treinamento do modelo e a parte restante é designada para o teste do modelo ajustado.

A figura 2 mostra de forma mais intuitiva como os parâmetros dos modelos foram estimados. Nesta figura, a parte inicial 0, selecionou o período em azul para o ajuste do modelo e o período em amarelo para o teste do modelo ajustado. Após isto, passou-se para a segunda parte do processo, parte 1, e realizou o mesmo procedimento feito inicialmente utilizando um outro período, este processo foi feito de forma análoga para as partes restantes e, após finalizado o processo para todas as k partes, foi estimado os valores dos parâmetros que minimizavam a erro quadrático médio.

Figura 2 – Validação cruzada



Fonte: Elaboração com dados próprios

Após o processo de validação cruzada e definido os parâmetros que minimizam o erro quadrático médio, foi selecionado o período de treino e o período de teste (últimos 60 dias de 2021). A figura 3 ilustra os períodos escolhidos para os dados de treino (período em azul) e os dados de teste (período amarelo). No eixo y encontra-se o valor da cesta básica e no eixo x encontra-se o período de ajuste do modelo.

Figura 3 – Dados de treino e teste

Fonte: Elaboração com dados próprios

Os dados de treino representam o conjunto de dados onde os parâmetros dos modelos Lasso, Ridge e Elastic Net são calibrados, enquanto que os dados de teste, representam o conjunto de dados onde esses modelos são avaliados. Neste projeto, os dados de janeiro de 2022 foram utilizados para a previsão para fora da amostra.

3.2.2.1 Regressão Ridge

A regressão ridge Hoerl e Kennard (1970), penaliza o quadrado dos parâmetros estimados da regressão, como explica James et al. (2013), esse tipo de regressão é muito parecido com modelo de mínimos quadrados, a exceção é a forma de como os parâmetros são estimados. A regressão Ridge estima os parâmetros, $\hat{\beta^R}$, de forma a minimizar:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \text{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2 \tag{3.6}$$

Onde $\lambda \geq 0$ é um parâmetro de ajuste (*tuning parameter*), e é determinado separadamente, por meio da validação cruzada. A regressão Ridge busca estimar os coeficientes que se ajustem aos dados da melhor forma, fazendo com que a soma dos quadrados dos resíduos diminua.

Na equação 3.6, $\lambda\sum_j \beta_j^2$, é chamado de penalidade de encolhimento. Essa penalidade será menor quando $\beta_1, \cdots, \beta_p$ são próximos de zero, assim, encolherão as estimativas de β_j em direção a zero. Quando $\lambda = 0$, a penalidade não fará nenhum efeito e a regressão Ridge terá as mesmas estimativas que o modelo de mínimos quadrados ordinários, por outro lado, quando $\lambda \to \infty$, o impacto desta penalidade aumentará, e os coeficientes estimados pela regressão Ridge irá se aproximar de zero.

3.2.2.2 Regressão Lasso

A regressão Lasso (*least absolute shrinkage and selection operator*) Tibshirani (1996), é uma alternativa ao modelo Ridge, entretanto, neste caso, a penalização é efetuada no módulo dos parâmetros. A regressão Lasso estima os parâmetros, $\hat{\beta_\lambda^L}$, de forma a minimizar:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p} \mid \beta_j \mid = \text{RSS} + \lambda\sum_{j=1}^{p} \mid \beta_j \mid \tag{3.7}$$

Ao penalizar o módulo dos parâmetros, alguns destes parâmetros serão iguais a zero para determinados valores de *lambda*. Consequentemente o modelo lasso além de fazer a regularização, também fará a auto-seleção dos parâmetros.

3.2.2.3 Regressão Elastic Net

Por fim, a regressão Elastic Net Zou e Hastie (2005) é uma generalização da regressão Lasso, em suma, esta regressão combina os dois tipos de penalidades presentes nos modelos de regressão Ridge e Lasso.

Como explica Géron (2019), a regressão Elastic Net é um meio termo entre a regressão Ridge e a regressão Lasso. Este modelo combina os dois tipos de penalidade, como mostra a equação 3.8

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda_1 \sum_{j=1}^{p} \beta_j^2 + \lambda_2 \sum_{j=1}^{p} \mid \beta_j \mid \tag{3.8}$$

A vantagem de utilizarmos a Regressão Elastic Net, é que neste caso, podemos aproveitar a regularização da regressão Ridge com a seleção de variáveis da regressão Lasso.

3.3 Dados

Neste trabalho os dados foram coletados em supermercados varejistas que ofertam seus produtos de forma *on-line* na região metropolitana de João Pessoa, no mês de janeiro de 2022, o número de supermercados pesquisados foi de 16, de início, o número de supermercados foi de 11. O número de mercados foi aumentando e diminuindo à medida que alguns supermercados inciavam ou paravam com a oferta de produtos *on-line*.

Os produtos pesquisados foram os indicados pelo Decreto Lei n° 399 para a região 2, na qual a Paraíba faz parte. A localização dos mercados que atualmente fazem parte da coleta estão localizados nos mapas 11 e 12 no anexo B.

Os dados sobre o custo da cesta básica, com a frequência de coleta diária teve início no dia 09 de novembro de 2020 e vem sendo coletado de forma ininterrupta. A tabela 2 mostra as estatísticas descritivas mais recentes, mês de janeiro de 2022.

Tabela 2 – Estatística descritivas mês de janeiro de 2022

Produtos	**Peso**	**Peso coleta**	**Itens**	**Média (R$)**	**DP**	**Var**
Carne	4,5 KG	1,0 KG	22	176,380	3,160	10,010
Leite	6,0 L	1,0 L	72	32,820	1,850	3,420
Feijão	4,5 KG	1,0 KG	37	34,290	1,180	1,390
Arroz	3,6 KG	1,0 KG	81	17,020	0,150	0,020
Farinha	3,0 KG	1,0 KG	43	14,510	0,470	0,220
Legumes (Tomate)	12,0 KG	1,0 KG	11	89,820	6,450	41,600
Pão Francês	6,0 KG	1,0 KG	6	71,780	2,930	8,580
Café em Pó	300 GR	250 GR	235	9,280	0,200	0,040
Frutas (Banana)	90 unid. [2]	1,0 KG	13	42,830	2,060	4,260

continua na próxima página

[2] São considerados 11,25 quilogramas de banana

Tabela 2 – continuação da página anterior

Produtos	Peso	Peso coleta	Itens	Média (R$)	DP	Var
Açúcar	3,0 KG	1,0 kg	18	12,300	0,070	0,005
Óleo	750 ML	900 ML	41	8,240	0,030	0,001
Manteiga	750 GR	200 GR	109	38,000	0,920	0,850

Fonte: Decreto Lei n° 399 e elaboração própria

A tabela 2 mostra as quantidades mínimas indicadas pelo decreto lei n° 399, na coluna "Peso". A coluna "Peso coleta" mostra qual a medida de peso e volume que é considerada para o calculo do valor médio de cada item, por exemplo, para o produto Carne, após a coleta dos dados, apenas os itens que tenha como medida de peso igual a 1 KG (quilograma) são considerados, após isto, calcula-se a média de todos os itens com 1,0 KG de carne e multiplica-se por 4,5 kg, que dará o valor médio do produto ponderado pelo peso, como mostra a tabela 2 o valor médio para o produto Carne é de R$ 176,38 para o período considerado. Para o item manteiga, multiplica-se o valor médio de cada item com 200 gramas e multiplica-se por 3,75 [3]. Já para Óleo, multiplica-se o valor médio de cada item com 900 ml por $\frac{750}{900}$ (aproximadamente 0,8333) e, encontra-se o valor médio. O valor total da cesta é a soma das médias dos 12 produtos indicados, para o período de referência, o valor médio da cesta foi de R$ 547,33. Na mesma tabela, as colunas DP e Var representam, respectivamente, desvio padrão e variância em reais.

A coluna "Itens" mostra, em média, quantos itens são coletados por dia para cada produto, ou seja, para o produto Arroz, são coletados, em média, 81 itens. Vale afirmar que, se um produto, com as mesmas especificações for coletado em mais de um supermercado, será considerado na amostra quantas vezes este produto aparecer, ou seja, se o produto Leite de 1 L (litro) da marca A aparecer em 3 supermercados diferentes, serão computados 3 itens para a amostra.

As especifidades dos produtos considerados são as seguintes: Para a Carne coxão mole (chã de dentro), coxão duro (chã de fora) e patinho, o leite considerado é o do tipo integral, o tipo de feijão coletado é o carioca tipo 1, para frutas são considerados banana prata e nanica, o arroz considerado é o do tipo parboilizado tipo 1, e o óleo considerado é o de soja.

[3] 750 gramas dividido por 200 gramas

4 Resultados

Nesta seção serão observados como evoluiu o custo da cesta básica estimado por esse projeto, assim como os bens que compõem este conjunto de bens para a região metropolitana da cidade de João Pessoa. Como objetivo secundário, foram aplicados modelos de previsão de séries temporais para o custo da cesta, a frequência utilizada para as predições foram diárias.

A frequência diária foi escolhida para a predição devido a quantidade observações. Caso fosse escolhido a previsão mensal, encontraríamos um conjunto de dados com poucas observações, um total de 15 observações (quinze meses compreendendo novembro de 2020 a janeiro de 2022).

4.1 Cesta Básica

A figura 4 mostra a comparação da evolução do custo da cesta básica estimado com dados coletados *on-line* (linha tracejada preta), de periodicidade diária, e a coleta feita pelo Departamento Intersindical de Estatística e Estudos Socioeconômicos (DIEESE) (linha continua azul) feita de forma mensal.

Apesar de alguns períodos onde o custo da cesta básica estimado com preços coletados *on-line* variou em uma intensidade e/ou direção diferente da coleta presencial mensal, observa-se a capacidade da coleta feita de forma automática em acompanhar a tendência da variação do custo da cesta básica, que no período de acompanhamento destes preços foi de alta.

Figura 4 – Comparação do custo da cesta básica

Fonte: Elaboração com dados próprios e do DIEESE

A tabela 3 mostra as médias mensais da cesta básica coletada de forma automática

e a cesta estimada de forma mensal pelo DIEESE, além de suas variações em relação aos meses anteriores. A primeira coluna mostra qual o mês e o ano, a segunda e a terceira coluna mostram o custo estimado da cesta básica com preços coletados *on-line* e o custo da cesta estimado pelo DIEESE, respectivamente, por fim, a quarta e quinta coluna mostram a variação do custo da cesta básica em relação ao mês anterior, tanto da coleta *on-line*, quanto da coleta presencial.

Tabela 3 – Variação mensal

Produtos	**Coleta on-line (R$)**	**DIEESE (R$)**	**Coleta on-line**	**DIEESE**
Novembro 2020	441,42	454,85	—	—
Dezembro 2020	444,85	475,19	0,78%	4,47%
Janeiro 2021	451,91	471,87	1,59%	-0,70%
Fevereiro 2021	455,73	484,54	0,85%	2,69%
Março 2021	459,94	478,52	0,92%	-1,24%
Abril 2021	477,56	490,04	3,83%	2,41%
Maio 2021	503,79	491,63	5,49%	0,32%
Junho 2021	502,27	495,76	-0,30%	0,84%
Julho 2021	506,42	492,3	0,83%	-0,70%
Agosto 2021	493,22	490,93	-2,61%	-0,28%
Setembro 2021	484,87	476,63	-1,69%	-2,91%
Outubro 2021	511,76	491,12	5,55%	3,04%
Novembro 2021	524,9	508,91	2,57%	3,62%
Dezembro 2021	520,32	510,82	-0,87%	0,38%

Fonte: Elaboração com dados próprios e do DIEESE

Pela tabela 3 observamos que de início o custo da cesta básica estimado com preços obtidos de forma *on-line* foi inferior ao custo estimado pelo DIEESE, esse período compreendeu o intervalo de novembro de 2020 a abril de 2021, em média a cesta básica estimada neste intervalo, com preços captados automaticamente foi R$ 20,60 menor, se comparado a coleta feita de forma presencial. Por outro lado, a partir do mês de maio de 2021 a cesta *on-line* foi constantemente maior do que a cesta estimada presencialmente, neste intervalo a cesta estimada com preços coletados de forma automática foi em média R$ 11,18 maior, se comparado a cesta estimada com preços coletados presencialmente.

Ainda pela tabela 3, podemos observar qual foi a inflação acumulada para o período em análise, com os preços coletados de forma *on-line* a inflação acumulada foi de 17,87%, enquanto que na coleta presencial, a inflação acumulada foi de 12,31%.

Esse aumento da inflação, entre 2020 e 2021, pode ser explicado, em parte, pelo impacto da pandemia (COVID-19) que impactou os preços internacionais e causou a desvalorização do real, como explica Feijó, Araújo e Bresser-Pereira (2022).

Ao estimarmos o custo da cesta básica utilizando dados coletados diariamente, além das vantagens já mencionadas anteriormente, maior rapidez e menor custo, também foi possível observar em qual turno ou em qual dia da semana os preços dos itens estavam sendo coletados a um maior preço. Por turno, constatou-se que o turno da noite obteve a maior média de preços com um valor de R$ 496,05, seguido do turno da tarde com um valor de R$ 495,45 e, por fim o turno da manhã com um custo estimado de R$ 484,79.

Tabela 4 – Estatísticas descritivas por turno e dia - Valores em R$

Dia	**Geral**	**Manhã**	**Tarde**	**Noite**
Segunda-Feira	490,77	471,59	505,99	498,05
Terça-Feira	491,80	476,58	503,77	499,11
Quarta-Feira	490,22	486,53	493,59	489,59
Quinta-Feira	491,70	485,90	498,86	491,76
Sexta-Feira	492,17	495,15	496,23	475,44
Sábado	490,63	495,90	485,57	496,36
Domingo	491,62	482,92	492,04	516,00

Fonte: Elaboração com dados próprios

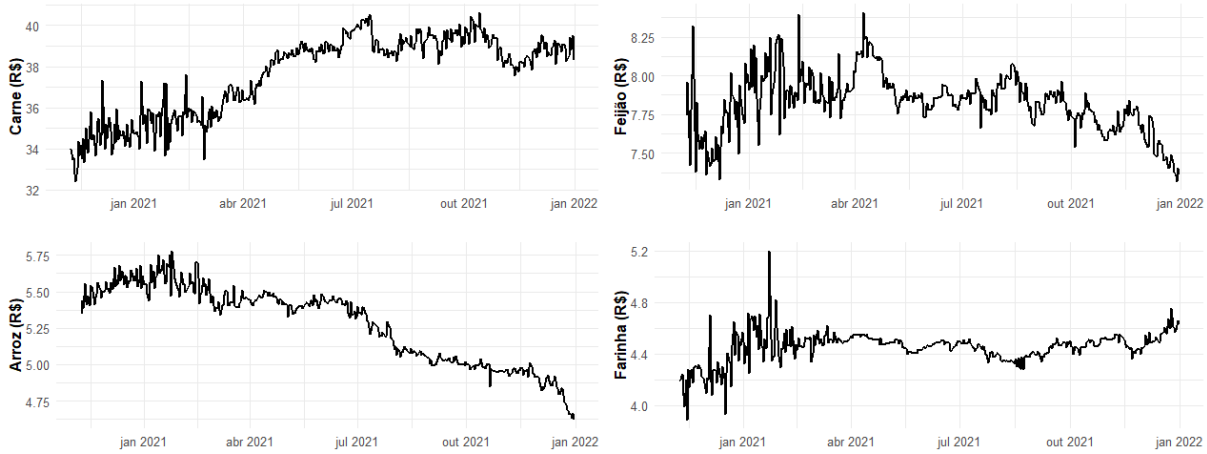
A tabela 4 mostra qual a média da cesta básica para cada dia da semana e, os dias da semana separados por turno de coleta para o ano de 2021[1]. A segunda coluna "Geral" mostra as médias das cestas básicas estimadas por dia da semana, como é visto, no ano de 2021, o dia com a cesta estimada com um maior preço foi a sexta-feira, com um valor de R$ 492,17. As colunas "Manhã", "Tarde" e "Noite" mostram os valores das cestas básicas por dia da semana e por turno. Por exemplo, quando a cesta foi estimada com preços coletados na segunda pela manhã, o valor foi estimado em R$ 471,59, quando estimado com preços coletados à tarde o valor foi de R$ 505,90 e, quando estimado com preços coletados à noite, o valor foi de R$ 498,05.

[1] Durante o ano de 2021 foram estimadas 151 cestas com dados coletados pela manhã, 144 com dados capturados à tarde e 70 cestas estimadas com dados coletados à noite

4.2 Produtos

Nesta subseção é analisado como os preços dos itens que compõe a cesta básica evoluíram no intervalo de coleta. A figura 5 mostra como o custo da carne, feijão, arroz e farinha evoluíram de 09 de novembro de 2020 a 31 de dezembro de 2021.

Figura 5 – Evolução produtos: Grupo 1



Fonte: Elaboração com dados próprios

Pela figura 5, é visto que o preço da carne teve uma tendência de alta do inicio da coleta até por volta do mês de julho de 2021, após este período é visto que o preço apresentou uma estabilização do valor observado e, por fim, voltou a apresentar uma tendência de alta.

Já o preço do feijão apresentou no início do período de coleta uma tendência de alta, esta alta de preços foi vista até o início de abril, após isto, o preço deste produto apresentou um pequeno período de estabilização e, por fim, a partir de agosto de 2021 o feijão apresentou uma tendência de baixa.

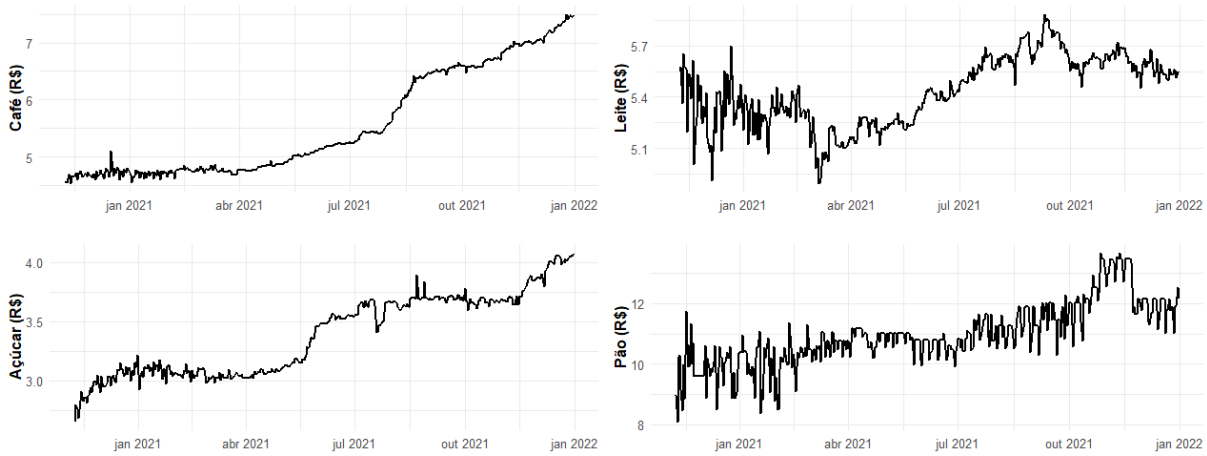
O arroz, desde o início da coleta foi um dos poucos produtos que não apresentaram uma tendência de alta em seu preço. Do início da pesquisa de preços até meados de janeiro de 2021, o preço deste produto permaneceu relativamente estável, como mostra a figura 5. Entretanto, a partir da segunda metade de janeiro de 2021 o valor do quilograma do arroz começou a apresentar uma tendência de queda, tendo esta tendência intensificado por volta do mês de junho de 2021.

O último item analisado da figura 5, a farinha, ao contrário dos demais itens, foi um dos produtos que mais apresentaram estabilidade na pesquisa de preço, à exceção do início do período. No princípio da pesquisa, foi visto um movimento de alta dos preços, após isto, o que foi constatado foi uma relativa estabilização nestes preços, ainda assim, o mês de dezembro foi de uma pequena elevação no preço deste produto.

A figura 6, mostra como evoluíram os preços de café, leite, açúcar e do pão. Assim como a figura 5, de início, é observado uma maior variabilidade dos preços pesquisados,

essa característica da maior variabilidade no início da coleta foi comum a todos os itens, isto pode ser explicado pelo maior alcance que a coleta de preços foi tendo ao longo da pesquisa [2].

Figura 6 – Evolução produtos: Grupo 2



Fonte: Elaboração com dados próprios

O café, ao contrário da maioria dos outros produtos, apresentou uma menor variabilidade dos preços desde o início da coleta. Isto é explicado pela quantidade de produtos que são coletados por dia, desde o principio o café foi o produto que obteve a maior média de itens coletados. Em relação ao preço deste produto, é observado que nos primeiros meses houve uma relativa estabilidade dos valores, essa estabilidade permaneceu até o início de abril de 2021, após este intervalo, houve um movimento de alta dos preços até a metade de 2021, após este segundo intervalo, é visto que há uma intensidade no aumento do preço do café para a região metropolitana de João Pessoa, tendência que é observada até o período final da coleta.

O leite iniciou o período de pesquisa com um custo estimado por volta de R$ 5,50 (Litro), após este período inicial foi observado uma queda no preço coletado, atingindo a quantia miníma de R$ 4,89 no início de março, após este pequeno período de queda, foi notado um aumento constante dos preços chegando a atingir o valor máximo no início de setembro, quando alcançou o valor de R$ 5,88, após atingir este preço, o custo do leite apresentou uma pequena queda e estabilizou-se em R$ 5,55.

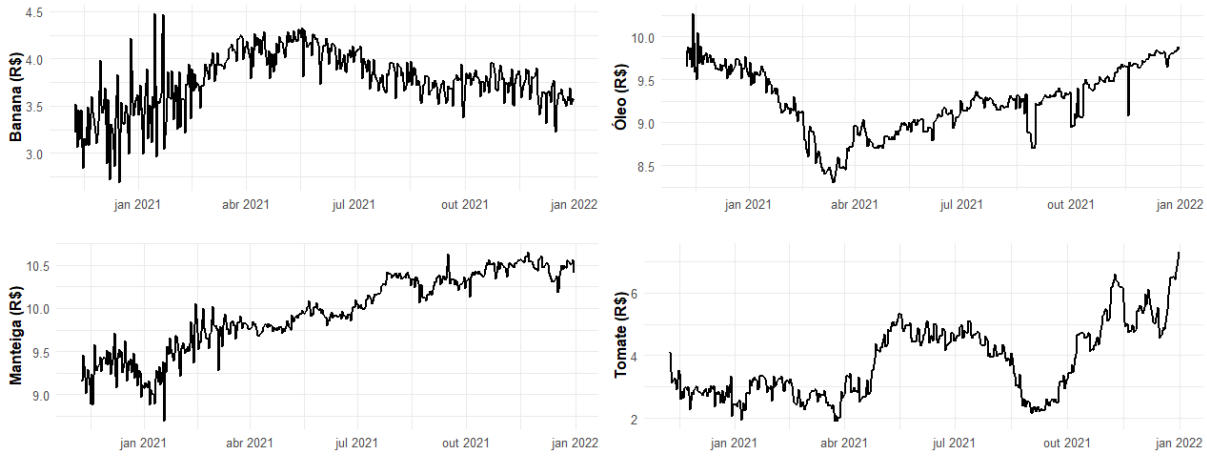
No começo da coleta de preços, o açúcar apresentou uma estabilidade nos valores observados, em média, o valor captado foi em torno de R$ 3,10 por quilograma. Após este período, o preço do açúcar apresentou o primeiro aumento mais intenso, esta elevação mais intensa foi vista em meados do mês de maio, ao fim desta maior intensificação o valor ficou estável por volta de R$ 3,68. Por fim, em meados de novembro observou-se a

[2] Para efeito de comparação, no primeiro mês de coleta do projeto, em média, eram considerados 415 produtos para a estimação da cesta. No mês de janeiro de 2022 este número foi de 694, um aumento de 67%.

segunda onda de elevação de preço deste produto, sendo notado uma constante elevação do custo até o fim do intervalo analisado, o preço do quilograma do açúcar fechou o ano de 2021 em R$ 4,08.

Como mostra a figura 6, o pão apresentou uma alta variabilidade nos preços, como já foi explicado, isto pode estar relacionado a quantidade de itens coletados por dia, como mostra a tabela 2, o pão é um dos produtos que conta com o menor número de itens coletados por dia, em média são coletado 7 produtos. Em relação aos preços captados, é visto que desde o início da coleta sempre houve uma tendência de alta, mesmo que em um menor nível, se comparado aos outros produtos. No período de observação, foi constatado que o menor valor observado foi R$ 8,09, nos primeiros dias da coleta, por outro lado, o maior preço visto foi de R$ 13,67 no fim de outubro de 2021.

Figura 7 – Evolução produtos: Grupo 3



Fonte: Elaboração com dados próprios

Por fim, é analisado o comportamento dos preços da banana, óleo, manteiga e tomate. Ao observar a figura 7, é visto que um dos produtos com maior aumento relativo (tomate) tem seu preço observado, como mostra a tabela 5.

O valor do quilograma da banana apresentou uma tendência inicial de alta, período de alta que durou de novembro de 2020 a junho de 2021, neste intervalo o custo da banana apresentou o seu menor e maior preço observado, R$ 2,68 em dezembro de 2020 e R$ 4,48 em janeiro de 2021. Após essa elevação dos preços, foi notado uma diminuição dos valores, tendo o custo do quilograma estimado por volta de R$ 3,70 ao fim da coleta.

O movimento do preço do óleo apresentou duas tendências de valores mais intensas, a primeira delas foi a de baixa, em novembro de 2020 o preço do óleo teve o valor médio de R$ 9,77, ao fim deste período de queda o preço do óleo chegou a atingir o valor de R$ 8,30 em março de 2021, após este preço mínimo o valor deste produto iniciou um período de alta voltando a atingir o valor observado ao fim de 2020, em novembro de 2021 o valor médio foi de R$ 9,60, valor R$ 0,17 menor se comparado ao mês de novembro de 2020.

A manteiga foi um dos produtos com maior aumento em seus preços, na maior parte do período de captura dos preços, houve um amento do valor deste produto, se for comparado o preço médio da manteiga em novembro de 2020 e novembro de 2021, é visto que o aumento foi de R$ 1,12, relativamente este aumento foi de 12,00%.

Por fim, é visualizado a evolução do custo tomate, o produto com o maior aumento percentual no período de coleta. O preço do tomate permaneceu relativamente estável até a metade de abril de 2021, quando foi observado um aumento no nível de preços sendo estimado por volta de R$ 4,50, após este aumento houve um diminuição dos preços em setembro de 2021, porém, voltando a ter outro período de aumento fechando com um valor de R$ 7,30 no último dia de dezembro de 2021.

Por último, verificamos qual foi a mudança percentual de cada produto, o período de comparação escolhido foram os meses de dezembro de 2020 e dezembro de 2021.

Tabela 5 – Variação dos preços dos produtos Dez 2020 - Dez 2021

Produto	**Dezembro 2020 (R$)**	**Dezembro 2021 (R$)**	**Nominal (R$)**	**Percentual (%)**
Carne	34,83	38,87	4,04	11,60
Leite	5,31	5,56	0,25	4,71
Feijão	7,72	7,50	-0,22	-2,85
Arroz	5,58	4,80	-0,78	-13,98
Farinha	4,31	4,55	0,24	5,57
Tomate	2,79	5,65	2,86	102,51
Pão	9,82	11,87	2,05	20,88
Café	4,73	7,25	2,52	53,28
Banana	3,37	3,56	0,19	5,64
Açúcar	3,05	3,98	0,93	30,49
Óleo	9,65	9,79	0,14	1,45
Manteiga	9,33	10,45	1,12	12,00

Fonte: Elaboração com dados próprios

A tabela 5 mostra que nominalmente o produto com maior aumento foi a carne, no mês de dezembro de 2021 o quilograma da carne foi R$ 4,04 maior se comparado ao mês de dezembro de 2020. Por outro lado, o arroz foi o produto que nominalmente teve a maior queda, uma diminuição de R$ 0,78. Percentualmente, o tomate foi o produto com o maior aumento, 102,5% e o arroz foi o produto com a maior queda percentual, uma diminuição de -13,98%.

4.3 Predições

Nesta seção serão aplicados os modelos preditivos escolhidos para a previsão do custo da cesta básica. O objetivo desta seção não consiste em fazer a comparação da acurácia dos modelo preditivos, pratica comum na literatura quando se é proposto a previsão de índices de inflação como Ülke, Sahin e Subasi (2018) ou variáveis macroeconômicas como Richardson, Mulder e Vehbi (2021), mas sim, fornecer diferentes alternativas na previsão do custo da cesta básica.

4.3.1 ARIMA

O modelo ARIMA é um dos modelos preditivos univariados mais aplicados na previsão de séries temporais Hyndman e Athanasopoulos (2018), e foi escolhido por ser um modelo de fácil aplicação. Na figura 8 observamos a evolução do custo da cesta básica.

Figura 8 – Evolução do custo da cesta básica

Fonte: Elaboração com dados próprios

para a previsão da série temporal utilizando o modelo ARIMA, é necessário que a série seja estacionária, para verificar a existência dessa condição na série que compreende o custo da cesta básica foi utilizado o teste elaborado por Kwiatkowski et al. (1992), ao realizar o teste foi verificado um valor de estatística de teste de 5,7866, valor este bem superior a qualquer valor crítico associado aos níveis de significância, como mostra a tabela 6.

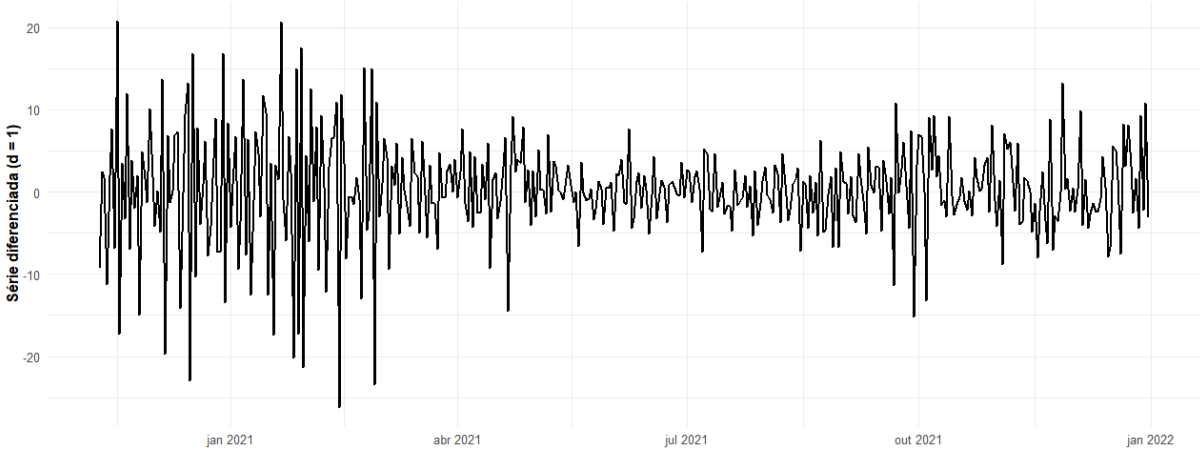
Tabela 6 – Teste de raíz unitária - KPSS

Nível de significância:	10%	5%	2,50%	1%
Valores críticos	0,347	0,463	0,574	0,739

Apesar da série original não ser estacionária, podemos verificar se a série diferenciada é estacionária, para isso foi verificado se a série diferenciada com uma ordem 1 de defasagem

é estacionária, a figura 9 mostra o custo da cesta básica com uma defasagem de ordem 1. Ao realizar o teste foi verificado um valor de estatística de teste de 0,0448, este valor é bem menor que qualquer valor crítico associado aos níveis de significância mostrados na tabela 6.

Figura 9 – Série diferenciada



Fonte: Elaboração com dados próprios

Após a execução do teste de raiz unitária, foi realizado a identificação da ordem do modelo, esta identificação foi realizado de forma automática pelo *software* escolhido. O modelo identificado que minimiza o critério de Akaike (AIC) foi o modelo ARIMA(2,1,2), com um valor AIC de 2622,9.

Finalizado as etapas de verificação de estacionariedade e identificação do modelo, foi feito a previsão da série em questão. Por ser uma série temporal de frequência diária, o horizonte escolhido para a previsão foi de tamanho 31. As previsões referentes a esses 31 dias estão contidos na tabela 7 no anexo A, e teve como uma raiz quadrada do erro-médio (RQEM) um valor de 9,23, foi o menor valor dentre os modelos utilizados.

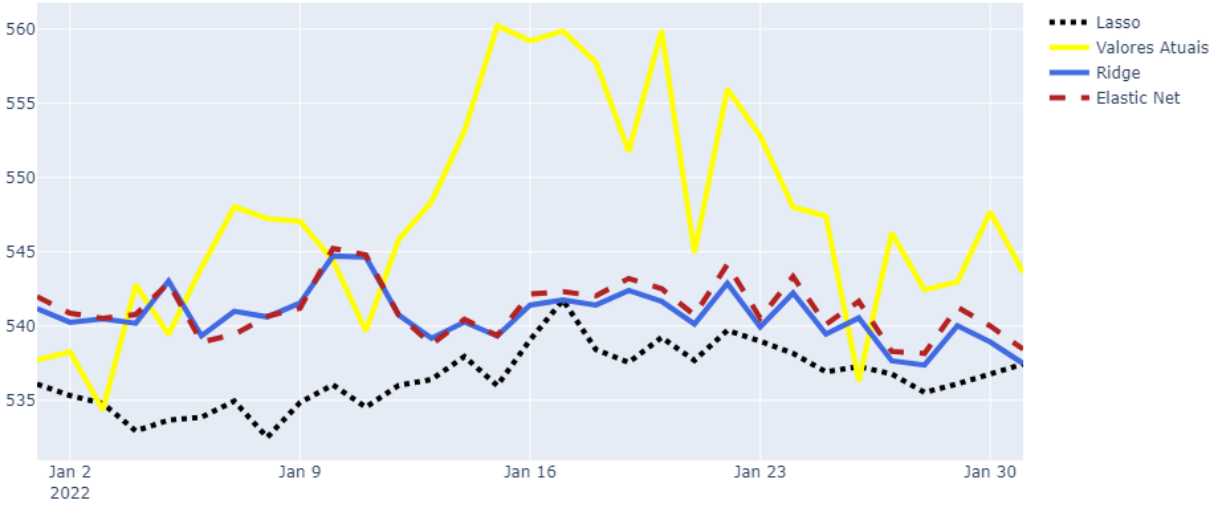
4.3.2 Lasso, Ridge e Elastic Net

Apesar dos modelos Lasso, Ridge e Elastic Net serem normalmente usados para a classificação de variáveis, estes algorítimos também podem ser usados na previsão de séries temporais univariadas.

Como mostrado na seção de metodologia, uma das formas de tentar melhorar a acurácia destes modelos é por meio da seleção do parâmetro λ, para os modelos Lasso e Ridge e λ e 1 - λ para o modelo Elastic Net. Após ter feito o processo de *tuning* dos modelos, foi estimado que os parâmetros que otimizavam os modelos foram os seguintes: 3,40895 para o λ do modelo Lasso; 5.7255 para o λ do modelo Ridge; 0,04359 λ e 0.39017 para 1 - λ para o modelo Elastic Net.

Após a otimização dos parâmetros foi feita a previsão para o custo da cesta básica, utilizando cada um dos três modelos, a tabela 7 no anexo A mostra todas as previsões feitas por cada um dos modelos utilizados para todos os dias do mês de janeiro de 2022.

Figura 10 – Previsões Ridge, Lasso e Elastic Net



Fonte: Elaboração com dados próprios

A figura 10 mostra as previsões dos modelos Lasso, Ridge e Elastic Net para o mês de janeiro de 2022 e os valores estimados do custo da cesta básica (linha amarela). É visto que os modelos Ridge e Elastic Net forneceram previsões similares, além disso, foram constantemente maiores do que o modelo Lasso. Entretanto, no último dia da previsão, os valores foram quase que idênticos para os três modelos

Também verifica-se que apesar dos modelos fornecerem previsões relativamente próximas dos valores atuais da cesta básica, o valor deste conjunto de bens, foi na maioria dos dias maior do que as previsões feitas. Nota-se que a variação nos preços desta cesta foi acompanhado por esses três modelos de previsão, porém, em uma intensidade maior. Em relação ao (RQEM), o modelo Lasso obteve um valor de 12,35, Ridge foi de 9,81 e Elastic Net um valor de 9,50.

5 Conclusões

Neste projeto foi construído um algoritmo que de forma automática e diaria coleta dados sobre os preços dos itens que compõem uma cesta de produtos para a região metropolitana de João Pessoa, além disso, com o objetivo de fazer o acompanhamento destes preços, foram aplicados algoritmos de previsão sobre os dados em mãos, os algoritmos escolhidos foram os modelos ARIMA, Lasso, Ridge e Elastic Net.

Com os dados coletados de forma *on-line* e, de frequência diária, foi possível verificar a capacidade dessas obervações coletadas de forma automática em acompanhar a tendência na mudança do custo da cesta básica, que no período de acompanhamento foi de alta, assim como foi observado pelo DIEESE. Em relação as previsões, foi visto que dos algoritmos de previsão escolhidos, o que teve melhor desempenho foi o modelo ARIMA, com o RQEM de 9,23.

Por fim, é possível notar que os dados coletados de forma automática em supermercados que ofertam seus produtos *on-line* são capazes de acompanhar as variações de preços, porém, de forma mais barata e rápida, se comparado aos métodos tradicionais de coleta, como a pesquisa de dados presencial. Sendo assim, um avanço em relação a coleta tradicional. No que se refere as previsões foi notado que dos modelos escolhidos, o modelo ARIMA, um dos modelos univariados para a previsão de séries temporais mais tradicionais foi o que teve melhor desempenho, se comparado aos modelos Lasso, Ridge e Elastic Net.

A Tabelas

Tabela 7 – Previsões custo da cesta básica

Data	Cesta	ARIMA	Lasso	Ridge	Elastic Net
01/01/2022	537,71	538,22	536,09	541,17	541,98
02/01/2022	538,25	537,65	535,32	540,25	540,87
03/01/2022	534,35	538,64	534,79	540,49	540,52
04/01/2022	542,77	538,14	532,94	540,19	540,77
05/01/2022	539,42	539,06	533,66	543,04	542,91
06/01/2022	543,93	538,62	533,86	539,35	538,93
07/01/2022	548,03	539,49	534,94	540,99	539,43
08/01/2022	547,24	539,09	532,50	540,62	540,65
09/01/2022	547,05	539,92	534,85	541,54	541,21
10/01/2022	544,42	539,57	536,04	544,72	545,23
11/01/2022	539,68	540,35	534,53	544,64	544,79
12/01/2022	545,82	540,04	536,01	540,76	540,74
13/01/2022	548,38	540,79	536,40	539,17	538,75
14/01/2022	553,14	540,52	537,95	540,27	540,46
15/01/2022	560,22	541,22	535,99	539,33	539,36
16/01/2022	559,20	540,99	539,07	541,41	542,15
17/01/2022	559,87	541,66	541,75	541,76	542,31
18/01/2022	557,77	541,46	538,42	541,41	542,03
19/01/2022	551,75	542,10	537,55	542,39	543,21
20/01/2022	559,91	541,93	539,23	541,66	542,52
21/01/2022	544,94	542,54	537,69	540,13	540,74
22/01/2022	555,96	542,39	539,71	542,88	544,14
23/01/2022	552,80	542,97	538,98	539,93	540,54
24/01/2022	548,01	542,86	538,16	542,23	543,33
25/01/2022	547,40	543,42	536,92	539,47	540,12
26/01/2022	536,28	543,33	537,27	540,54	541,67
27/01/2022	546,27	543,86	536,77	537,66	538,29
28/01/2022	542,43	543,79	535,53	537,37	538,16
29/01/2022	542,98	544,30	536,10	540,02	541,27
30/01/2022	547,68	544,25	536,78	538,92	540,01
31/01/2022	543,62	544,74	537,44	537,48	538,44

Fonte: Elaboração própria

Tabela 8 – Proporção custo da cesta básica/renda completo

Ranking	**Bairro**	**Rend. (R$)**	**Prop. (%)**
1	Cabo Branco	15054,39	3,64
2	Tambaú	13684,97	4,00
3	Manaíra	13259,02	4,13
4	Jardim Oceania	12287,73	4,45
5	Brisamar	11956,64	4,58
6	Estados	11785,97	4,64
7	Ponta Do Seixas	11521,10	4,75
8	Miramar	11475,72	4,77
9	Tambauzinho	11329,01	4,83
10	Pedro Gondim	9887,87	5,54
11	João Agripino	9574,24	5,72
12	Aeroclube	9429,06	5,80
13	Portal Do Sol	9167,56	5,97
14	Altiplano Cabo Branco	8755,62	6,25
15	Bessa	8234,95	6,65
16	Expedicionários	7133,04	7,67
17	Anatólia	6952,72	7,87
18	Jardim Cidade Universitária	5879,05	9,31
19	Bancários	5783,13	9,46
20	Jardim São Paulo	5524,00	9,91
21	Treze De Maio	5435,89	10,07
22	Torre	5005,89	10,93
23	Centro	4991,90	10,96
24	Jaguaribe	4849,07	11,29
25	Água Fria	4569,81	11,98
26	Tambiá	4319,19	12,67
27	Cuiá	3977,84	13,76
28	Ipês	3956,96	13,83
29	Ernesto Geisel	3721,67	14,71
30	Castelo Branco	3509,47	15,60
31	José Américo	3362,20	16,28
32	Cristo Redentor	3351,24	16,33
33	Valentina	3067,67	17,84
34	Cidade Dos Colibris	2946,66	18,57
35	Mangabeira	2801,39	19,54

continua na próxima página

Tabela 8 – continuação da página anterior

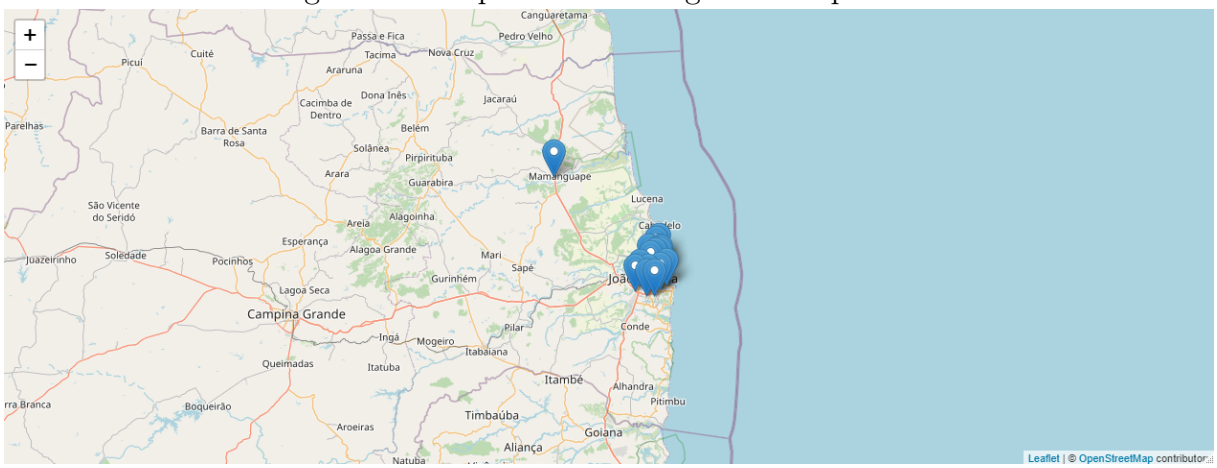
Ranking	Bairro	Rend. (R$)	Prop. (%)
36	Funcionários	2664,84	20,54
37	Roger	2536,26	21,58
38	João Paulo Ii	2530,82	21,63
39	Ernani Sátiro	2447,53	22,36
40	Penha	2437,97	22,45
41	Planalto Da Boa Esperança	2353,28	23,26
42	Mandacarú	2322,62	23,57
43	Trincheiras	2264,05	24,17
44	Cruz Das Armas	2236,06	24,48
45	Varadouro	2145,51	25,51
46	Varjão	2073,13	26,40
47	Costa E Silva	2004,20	27,31
48	Alto Do Mateus	1991,02	27,49
49	Indústrias	1982,06	27,61
50	Barra De Gramame	1932,16	28,33
51	Oitizeiro	1920,78	28,50
52	Costa Do Sol	1861,87	29,40
53	Paratibe	1804,99	30,32
54	Mucumago	1795,74	30,48
55	Padre Zé	1749,35	31,29
56	Gramame	1720,14	31,82
57	Grotão	1670,32	32,77
58	Mumbaba	1636,38	33,45
59	Jardim Veneza	1600,59	34,20
60	Distrito Industrial	1569,16	34,88
61	Alto Do Céu	1528,36	35,81
62	Ilha do Bispo	1270,11	43,09
63	São José	1149,09	47,63

Fonte: Elaboração própria com dados do IBGE

B Mapas

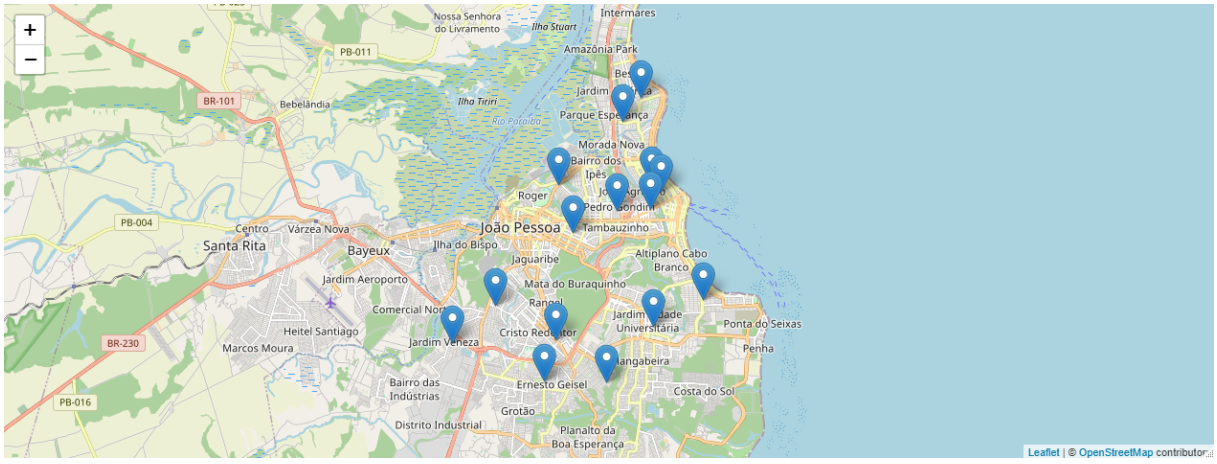
As figuras 11 e 12 mostram as localizações dos supermercados onde os preços dos produtos são coletados.

Figura 11 – Supermercados região metropolitana



A figura 11 mostra todos os supermercados, incluindo o supermercado localizado em Mamanguape, localizado na região metropolitana de João Pessoa, e que fica mais afastado dos demais supermercados. Já a figura 12 mostra os supermercado localizados na cidade de João Pessoa.

Figura 12 – Supermercados em João Pessoa



Referências

AKAIKE, H. A new look at the statistical model identification. *IEEE transactions on automatic control*, Ieee, v. 19, n. 6, p. 716–723, 1974.

ALBANESI, S. Inflation and inequality. *Journal of Monetary Economics*, Elsevier, v. 54, n. 4, p. 1088–1114, 2007.

ALVAREZ, S. E.; LEIN, S. M. Tracking inflation on a daily basis. *Swiss Journal of Economics and Statistics*, Springer, v. 156, n. 1, p. 1–13, 2020.

BOEING, G.; WADDELL, P. New insights into rental housing markets across the united states: Web scraping and analyzing craigslist rental listings. *Journal of Planning Education and Research*, SAGE Publications Sage CA: Los Angeles, CA, v. 37, n. 4, p. 457–476, 2017.

CAVALLO, A. Online and official price indexes: Measuring argentina's inflation. *Journal of Monetary Economics*, Elsevier, v. 60, n. 2, p. 152–165, 2013.

CAVALLO, A. Scraped data and sticky prices. *Review of Economics and Statistics*, MIT Press, v. 100, n. 1, p. 105–119, 2018.

EDELMAN, B. Using internet data for economic research. *Journal of Economic Perspectives*, v. 26, n. 2, p. 189–206, 2012.

FEIJÓ, C.; ARAÚJO, E. C.; BRESSER-PEREIRA, L. C. Política monetária no brasil em tempos de pandemia. *Brazilian Journal of Political Economy*, SciELO Brasil, v. 42, p. 150–171, 2022.

FRIEDMAN, M. *Inflation: Causes and consequences*. [S.l.]: Asia Publishing House, 1963.

GÉRON, A. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. [S.l.]: O'Reilly Media, 2019.

HILLEN, J. Web scraping for food price research. *British Food Journal*, Emerald Publishing Limited, 2019.

HOERL, A. E.; KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, Taylor & Francis, v. 12, n. 1, p. 55–67, 1970.

HYNDMAN, R. J.; ATHANASOPOULOS, G. *Forecasting: principles and practice*. [S.l.]: OTexts, 2018.

IM, J. et al. Energy efficiency in us residential rental housing: Adoption rates and impact on rent. *Applied Energy*, Elsevier, v. 205, p. 1021–1033, 2017.

JAMES, G. et al. *An introduction to statistical learning*. [S.l.]: Springer, 2013. v. 112.

KIENLE, H. M.; GERMAN, D.; MULLER, H. Legal concerns of web site reverse engineering. In: IEEE. *Proceedings. Sixth IEEE International Workshop on Web Site Evolution*. [S.l.], 2004. p. 41–50.

KUHN, M.; JOHNSON, K. et al. *Applied predictive modeling*. [S.l.]: Springer, 2013. v. 26.

KUNANG, Y. N.; PURNAMASARI, S. D. et al. Web scraping techniques to collect weather data in south sumatera. In: IEEE. *2018 International Conference on Electrical Engineering and Computer Science (ICECOS)*. [S.l.], 2018. p. 385–390.

KWIATKOWSKI, D. et al. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of econometrics*, Elsevier, v. 54, n. 1-3, p. 159–178, 1992.

MANIK, D. P. et al. A strategy to create daily consumer price index by using big data in statistics indonesia. In: IEEE. *2015 International Conference on Information Technology Systems and Innovation (ICITSI)*. [S.l.], 2015. p. 1–5.

MITCHELL, R. *Web scraping with Python: Collecting more data from the modern web*. [S.l.]: "O'Reilly Media, Inc.", 2018.

RICHARDSON, A.; MULDER, T. van F.; VEHBI, T. Nowcasting gdp using machine-learning algorithms: A real-time assessment. *International Journal of Forecasting*, Elsevier, v. 37, n. 2, p. 941–948, 2021.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 58, n. 1, p. 267–288, 1996.

ÜLKE, V.; SAHIN, A.; SUBASI, A. A comparison of time series and machine learning models for inflation forecasting: empirical evidence from the usa. *Neural Computing and Applications*, Springer, v. 30, n. 5, p. 1519–1527, 2018.

ZOU, H.; HASTIE, T. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, Wiley Online Library, v. 67, n. 2, p. 301–320, 2005.