

Relatório de Probabilidade e Estatística

Erico Meger¹, Igor Pinto², Victor Vechi³

¹Câmpus Pinhais – Instituto Federal do Paraná (IFPR)
Pinhais– PR – Brasil

ericomeger9@gmail.com, igorpinto103@gmail.com, victorvechi@gmail.com

Abstract. *This work analyzes data from the Bolsa Família program in Brazil between 2011 and 2016, focusing on the relationship between investment and the number of families served. Using a dataset available on Kaggle, descriptive statistical analyses and Pearson correlation were performed, resulting in a value of 0.9709, indicating a strong positive relationship between investment and beneficiaries. Additionally, a simple linear regression model was applied, showing that each increase of one million reais in investment leads to approximately 10 thousand new families served. The results revealed that, while the number of low-income families remained stable, the number of families served grew until 2014 and then declined. The Northeast region was identified as receiving the most investment. This work also provided practical understanding of data analysis and its applications in public policy.*

Resumo. *Este trabalho analisa os dados do programa Bolsa Família no Brasil entre 2011 e 2016, com foco na relação entre investimento e número de famílias atendidas. A partir de um conjunto de dados disponível no Kaggle, foram realizadas análises estatísticas descritivas e a correlação de Pearson, que resultou em um valor de 0,9709, indicando uma forte relação positiva entre investimento e beneficiários. Além disso, um modelo de regressão linear simples foi aplicado, mostrando que cada aumento de um milhão de reais no investimento resulta em aproximadamente 10 mil novas famílias atendidas. Os resultados revelaram que, embora o número de famílias de baixa renda tenha permanecido estável, o número de famílias atendidas cresceu até 2014 e depois apresentou um declínio. A região Nordeste foi identificada como a que recebeu mais investimento. O trabalho também proporcionou uma compreensão prática sobre análise de dados e suas aplicações em políticas públicas.*

Palavras-chave: Bolsa Família; Investimento; Análise de Dados; Regressão Linear; Políticas Públicas.

1. Conjunto de Dados:

Neste trabalho, utilizamos um conjunto de dados relacionado ao programa Bolsa Família, que é um programa de auxílio governamental destinado a famílias de baixa renda no Brasil. Os dados foram obtidos a partir de um dataset disponível no site Kaggle e cobrem o período de 2011 a 2017. No entanto, devido ao dataset estar incompleto para o ano de 2017, restringimos o período de análise aos anos de 2011 a 2016.

As variáveis do conjunto de dados podem ser classificadas em dois grupos. O primeiro grupo contém as variáveis quantitativas (numéricas), que incluem: número de

famílias atendidas pelo programa, valor do investimento por região, total de famílias de baixa renda por região e a porcentagem de famílias beneficiadas pelo programa. O segundo grupo contém a variável qualitativa (categórica), que corresponde à região geográfica, ano e mês.

A fonte dos dados é o Kaggle, uma plataforma amplamente utilizada para compartilhamento e análise de dados.

2. Resultado das Análises:

Foram calculadas as estatísticas descritivas (média, moda, mediana, variância e desvio padrão) das variáveis quantitativas, como o número de famílias contempladas, o investimento e o total de famílias de baixa renda, oferecendo uma visão geral das distribuições e dispersões dos dados.

media_familias_contempladas	2.738.747,53
media_investimento	407.652.531,68
media_familias_baixa_renda	4.018.991
mediana_familias_contempladas	763.827
mediana_investimento	105.159.204
varianca_familias_contempladas	5299498245758,453
varianca_investimento	1,3224028558002691e+17
varianca_familias_baixa_renda	8233900633548,702
desvio_padrao_familias_contempladas	2298864,37
desvio_padrao_investimento	363143155,59
desvio_padrao_familias_baixa_renda	2865489,25

Não foi observada uma moda explícita entre as variáveis categóricas, sendo elas amodais. Enquanto entre as variáveis contínuas, a única não amodal foi a quantidade de famílias de baixa renda. Os valores modais foram 1.246.826, 2.095.985, 8.380.881, 1.884.472 e 6.486.791, repetidos 72 vezes. Esses números correspondem ao total de famílias de baixa renda nas regiões: Centro-Oeste, Norte, Nordeste, Sul e Sudeste, respectivamente. Um ponto interessante, evidenciado também pelos gráficos, é que o número de famílias de baixa renda no Brasil permaneceu estável entre 2011 e 2016, sem alterações significativas ao longo do período.

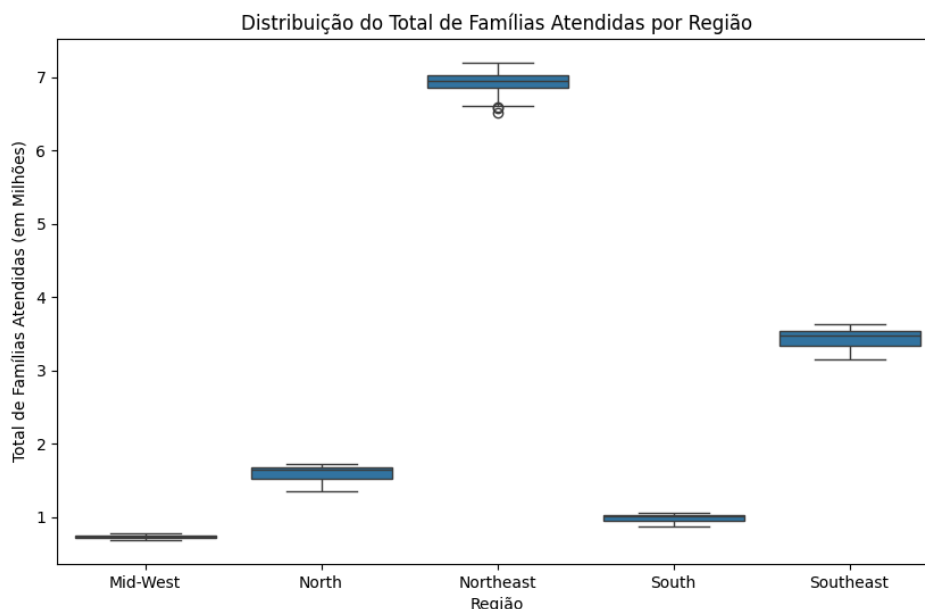
As estatísticas descritivas complementam essa análise, destacando que a média de famílias beneficiadas pelo Bolsa Família entre 2011 e 2016 foi de 2,7 milhões,

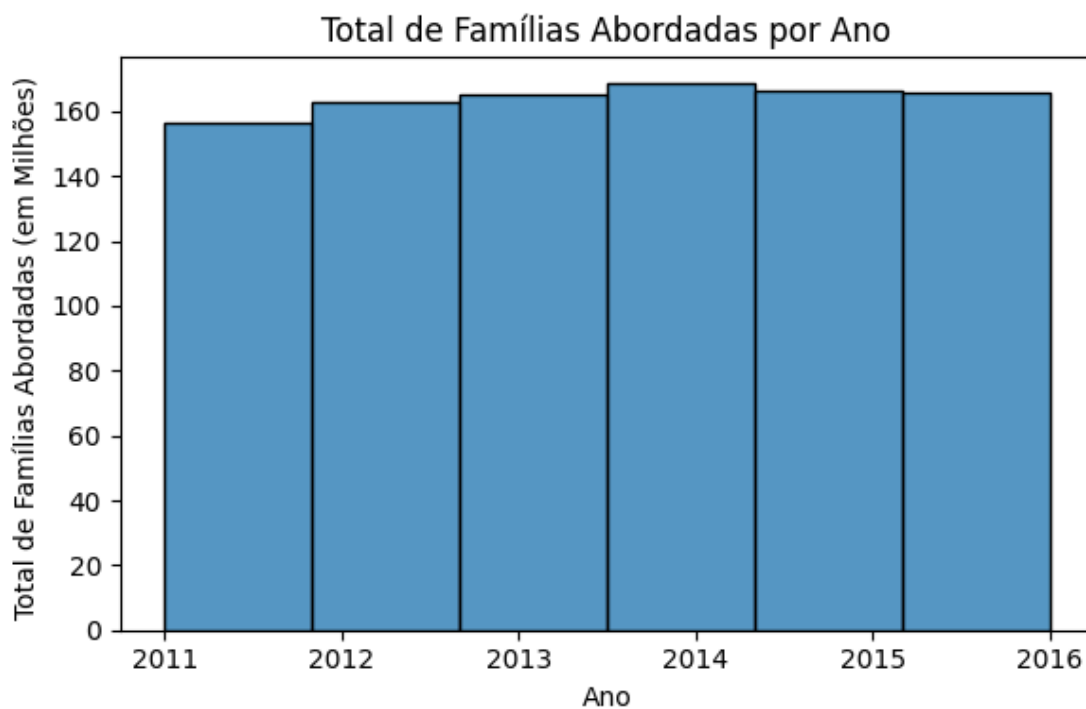
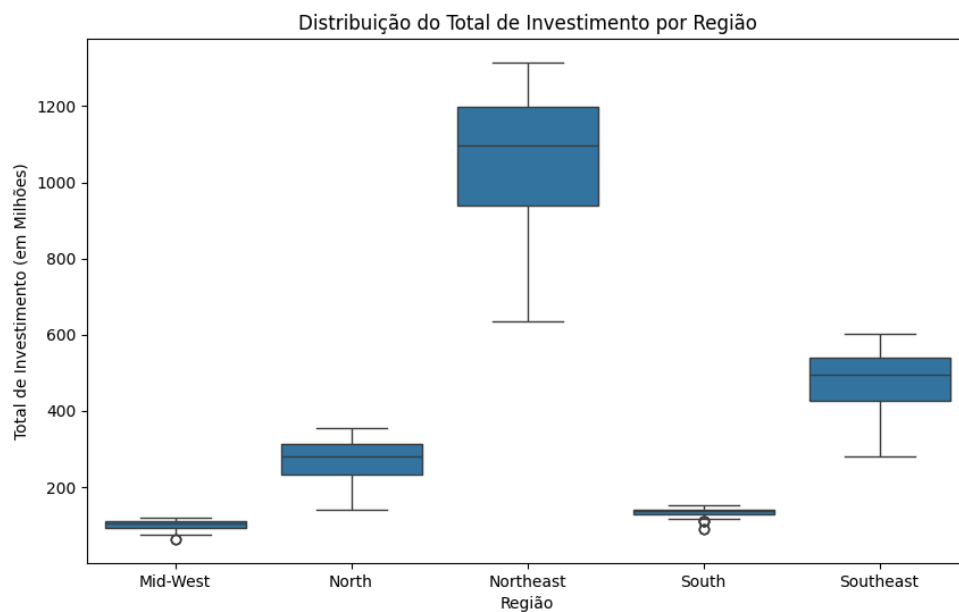
enquanto o investimento médio regional foi de R\$ 407,6 milhões. O número médio de famílias de baixa renda se manteve em 4 milhões, com pouca variação ao longo do período, reforçando a estabilidade observada nos gráficos.

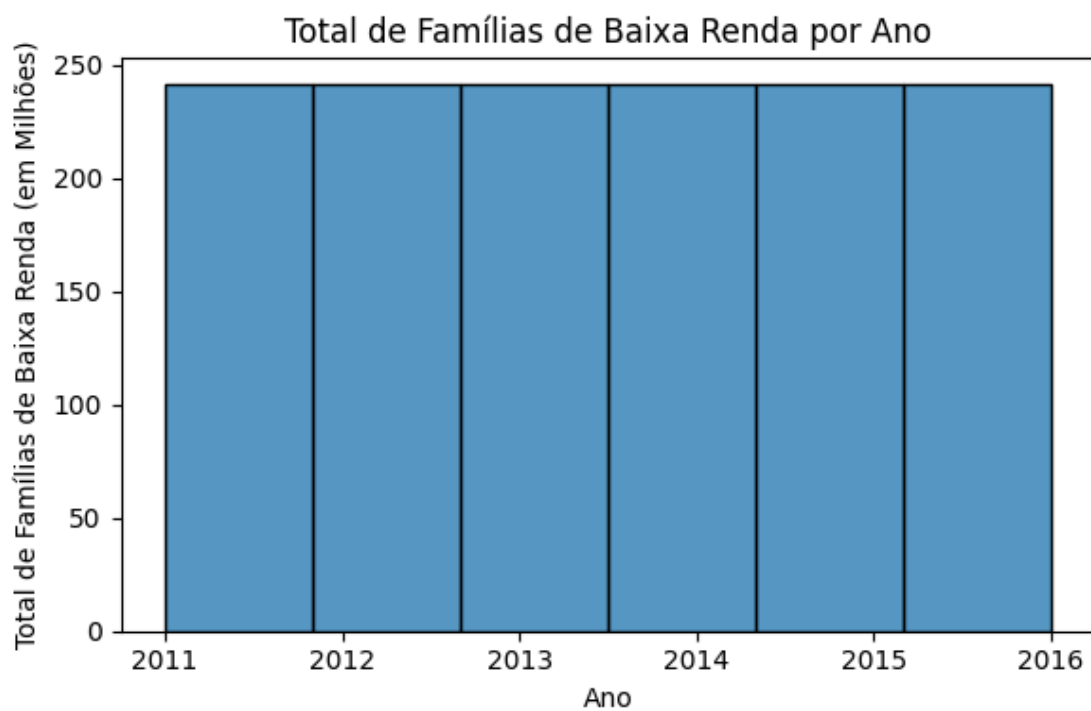
A mediana de 763.827 famílias beneficiadas, menor que a média, sugere uma distribuição assimétrica, influenciada por valores mais altos no Nordeste. O investimento também reflete essa assimetria, com uma mediana de R\$ 105,1 milhões, inferior à média, devido à concentração de investimentos em regiões específicas.

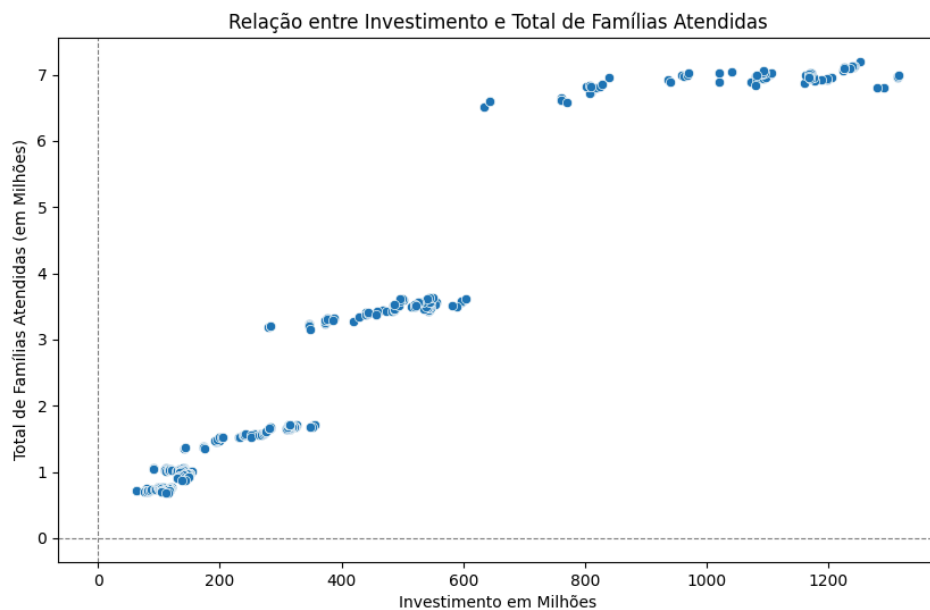
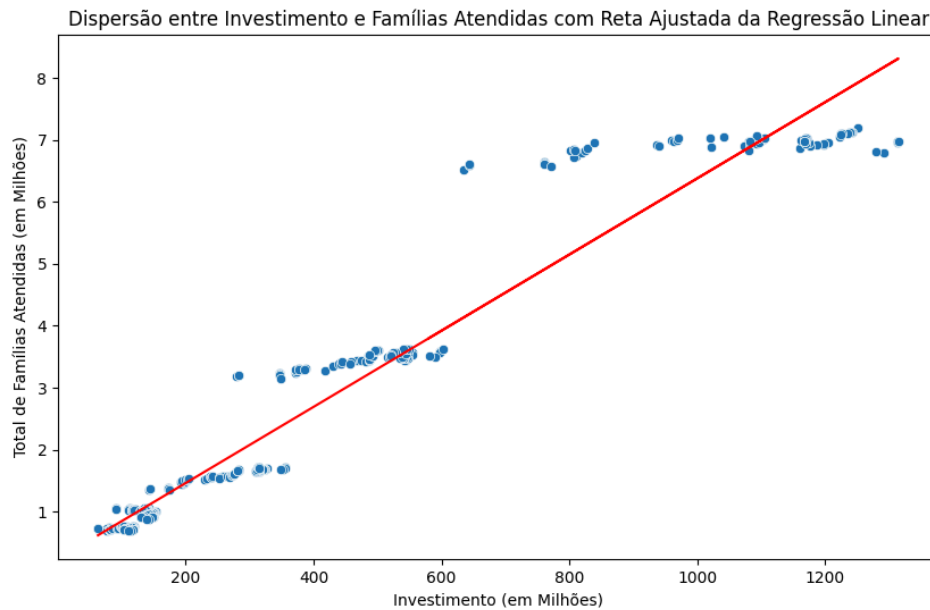
A alta variância no investimento ($1,32e+17$) e o desvio padrão de R\$ 363,1 milhões indicam uma grande disparidade entre as regiões, com o Nordeste recebendo significativamente mais recursos. Por outro lado, o desvio padrão das famílias beneficiadas, de 2,3 milhões, e das famílias de baixa renda, de 2,86 milhões, confirma uma consistência relativa nessas variáveis, apesar das variações regionais no investimento.

Foram gerados também gráficos para analisar as variáveis do conjunto de dados. O histograma do número de famílias contempladas e do valor de investimento por região foi criado para observar a distribuição desses dados. Também foi gerado um boxplot para identificar a dispersão das famílias de baixa renda nas diferentes regiões. Além disso, um gráfico de dispersão foi elaborado para visualizar a relação entre o número de famílias contempladas e o valor do investimento, complementado por uma reta de regressão linear que indica a tendência dessa relação.









As análises descritivas e os gráficos gerados revelam algumas tendências importantes nos dados. O histograma das famílias de baixa renda entre 2011 e 2016 indica uma relativa estabilidade no número total de famílias nessa condição, sem grandes variações ao longo dos anos. Por outro lado, o histograma das famílias contempladas pelo Bolsa Família mostra um crescimento contínuo no número de famílias atendidas até 2014, atingindo um pico nesse ano. A partir de 2015, observa-se um declínio gradual no número de famílias beneficiadas, refletindo possíveis mudanças no orçamento ou nas políticas do programa.

Os boxplots gerados também mostram dados valiosos sobre a distribuição dos investimentos e do número de famílias atendidas por região. A região Nordeste se destaca como a que recebe o maior volume de investimento, com uma significativa dispersão dos dados, indicando variações consideráveis ao longo do tempo. Da mesma forma, o Nordeste é a região que atende o maior número de famílias, refletindo uma relação direta entre o volume de recursos e o número de beneficiários. As regiões Norte e Sudeste também apresentam medianas elevadas tanto em investimento quanto em número de famílias atendidas, mas com menos variabilidade, sugerindo uma distribuição mais consistente ao longo do período analisado.

Por outro lado, as regiões Centro-Oeste e Sul possuem os menores volumes de investimento e o menor número de famílias atendidas, com uma distribuição mais concentrada dos dados. No caso da região Sul, foram observados outliers nos investimentos, indicando que, em determinados momentos, houve alocações significativamente menores em relação ao padrão geral. A presença de outliers no número de famílias atendidas na região Nordeste sugere que, embora a média tenha se mantido alta, houve períodos com uma redução atípica no número de beneficiários.

Esses resultados reforçam a correlação entre o montante de investimento e o número de famílias atendidas, destacando as diferenças regionais na alocação de recursos e na demanda do programa, com maior atenção voltada para as regiões mais vulneráveis, como o Nordeste.

3. Discussão da Correlação e Regressão:

A correlação de Pearson entre o investimento e o total de famílias atendidas pelo Bolsa Família foi de 0,9709. Esse valor indica uma correlação extremamente forte e positiva entre as variáveis, o que significa que, à medida que o investimento aumenta, o número de famílias beneficiadas também cresce proporcionalmente. Essa relação positiva sugere que o investimento é um fator chave na ampliação do alcance do programa.

Foi aplicado um modelo de regressão linear simples, com o investimento como variável independente (X) e o total de famílias atendidas como variável dependente (Y). O coeficiente angular (Beta) foi 0,01, enquanto a interseção (Alpha) foi 0,23, indicando que, para cada aumento de um milhão no investimento, espera-se um aumento de aproximadamente 10 mil no número de famílias atendidas.

O coeficiente de determinação (R^2) foi de 0,94, o que significa que 94% da variação no número de famílias atendidas pode ser explicada pelas variações no investimento. Esse alto valor de R^2 demonstra que o modelo ajustado tem um excelente poder preditivo, e o gráfico de dispersão com a reta ajustada da regressão confirma essa tendência positiva entre as variáveis analisadas.

4. Conclusão:

A análise realizada sobre o conjunto de dados do programa Bolsa Família entre os anos de 2011 e 2016 revelou resultados significativos sobre a relação entre investimento e número de famílias atendidas. A correlação de Pearson de 0,9709 indica uma associação extremamente forte e positiva entre essas variáveis, evidenciando que o aumento dos investimentos está diretamente relacionado à ampliação do número de beneficiários do programa.

Além disso, a aplicação do modelo de regressão linear simples demonstrou que, para cada incremento de um milhão de reais no investimento, espera-se um aumento de aproximadamente 10 mil famílias atendidas. O coeficiente de determinação (R^2) de 0,94 reforça a capacidade preditiva do modelo, sugerindo que a maioria da variação no número de famílias atendidas pode ser explicada pelas variações no investimento.

A análise estatística e gráfica também identificou padrões importantes: o número de famílias de baixa renda permaneceu relativamente estável durante o período analisado, enquanto o número de famílias contempladas pelo Bolsa Família mostrou um crescimento até 2014, seguido por um declínio a partir de 2015. A região Nordeste destacou-se como a que recebeu o maior volume de investimento e atendeu o maior número de famílias, refletindo uma política pública voltada para as áreas mais vulneráveis do país.

Este trabalho não apenas permitiu uma compreensão mais profunda dos dados relacionados ao Bolsa Família, mas também foi uma experiência enriquecedora do ponto de vista acadêmico. O processo de coleta e análise de dados, bem como a aplicação de técnicas estatísticas e a elaboração de gráficos, proporcionou um aprendizado valioso sobre como interpretar e apresentar informações de forma clara e precisa. Esses resultados são cruciais para a formulação de decisões futuras, pois evidenciam a necessidade de uma alocação estratégica dos recursos no programa Bolsa Família, orientando planejadores e gestores públicos na otimização dos investimentos para atender de forma mais eficaz às famílias em situação de vulnerabilidade social.