Análise das Solicitações 156 da Cidade do Recife

Lucas Alves Rufino (lar) e Rodrigo de Lima Oliveira (rlo)
Universidade Federal de Pernambuco

Abstract

Este projeto tem o intuito de realizar uma análise sobre os dados das solicitações 156 da cidade do Recife sobre serviços cotidianos fornecidos pela prefeitura, associando informações como dados de localização, temporal e socioeconômicos, a fim de promover a identificação de características que descrevem os dados, assim como a retirada de *insights*.
Keywords: Solicitações 156, Recife, Emlurb, Dados Públicos, localização, socioeconômico, temporal, pluviométrico.

1. Introdução

Com o objetivo de realizar uma análise sobre os dados das solicitações 156 da cidade do Recife, sobre os serviços cotidianos fornecidos pela prefeitura. Este relatório descreve as etapas de coleta, integração de dados de diversas fontes, preprocessamento das bases, aplicação de análises exploratória sobre hipoteses, assim como detecção de padrões nos dados.

Os dados que estão sendo estudados, traz uma informação que é utilizada para detectar necessidades da população, e fornecer melhores serviços a todos os cidadãos. O resultado da análise sobre os dados das solicitações 156 do Recife, é uma informação de interesse dos cidadãos, assim como do estado.

2. Metodologia

Os dados que foram analisados tratam sobre registros das solicitações e serviços do dia atual, em tempo real. Informações sobre demandas de serviços de arborização, drenagem, iluminação, limpeza, pavimentação, tapa buracos, entre outros serviços de reparação da cidade, solicitados pelos cidadãos via atendimento online ou pelo telefone 156 (links na última seção).

Os dados do *dataset* são atualizados periodicamente, foram contemplados dados até dia 11 de novembro de 2018. Foram consideradas informações a partir de 2012 quando o conteúdo começaram a ser coletado.

O *dataset* possui os seguintes atributos: código do grupo de serviço, descriçăo do grupo de serviço, código do serviço, descriçăo do serviço, descriçăo da situaçăo atual, endereço (logradouro, número, bairro e RPA) e datas (data de cadastro e data da última movimentaçăo). Além dos dados fornecidos no arquivo disponibilizado, o conjunto foi enriquecido com informações locais (latitude, longitude, áreas urbanas), temporais (períodos sazonais, eventos relevantes) e sociais (dados socioeconômicos por bairro do Recife).

O projeto foi realizado com auxilio de ferramentas como sistema de versão, jupyter notebook, python 3 e outras tecnologia. Bibliotecas de python como Request, APIs de localização, Pandas, BeautifulSoup, Holoviews, Matplotlib, Scipy, Scikit-Learn, Seaborn, Numpy e outras foram amplamente utilizadas durante o projeto. O projeto foi pensado e estruturado segundo o padrão Cookiecutter (1). A ferramenta de vizualização Tableau também foi utilizada para plotagens e *dashboard* de apresentação.

3. Bases de dados e coletas

A coleta de dados se deu inicialmente pelo site de dados abertos da prefeitura do Recife (2), as informações ja foram apresentadas em formato CSV, estruturados por ano, com seus respectivos dicionários de descrição.

O dados socioeconômicos foram obtidos através de *crawler* sobre o site da prefeitura do Recife sobre os perfis dos diferentes bairros da cidade (3). O site apresenta grande irregularidade de obtenção de informação, sendo por partes complexa a retirada de dados, contudo a mesma se mostrou possível. O *crawler* foi necessário devido a não disponibilidade de informações no site de origem dos dados, no caso, o CENSO do IBGE.

Os dados de localização foram obtidos através do *plug-in* para a plataforma de tabelas da Google chamada Geocode Cells (4). Para retirada de dados, basta selecionar a célula com o endereço e fazer a requisição, o resultado é a latitude e longitude na celula vizinha ao endereço, o dado por si só ja estava estruturado em planilha. Outras APIs como geopy e google maps foram utilizadas mas não davam liberdade para a obtenção de dados em laga escala. Cerca de 60.000 endereços foram crawleados repectivos ao ano de 2018.

Sobre as informações temporais, foram obtidas através de dois mecanismos, API da plataforma calendário (5). Os dados foram obtidos em formato JSON, sendo facilmente integrável. Sobre os dados de monitoramento pluviométrico, os mesmo foram obtidos no site da APAC (6). As informações de dados pluviometricos estavam registrados em um planilha HTML facilmente lida pela biblioteca Pandas.

4. Pré-processamento dos dados

Todas as bases trabalhandas possuiam um padrão bem regular para a execução do projeto, em sua maioria, o trabalho com relação ao pré-processamento ficou essencialmente sobre a tipagem dos dados, isso é, discriminar informações categóricas, numéricas, datas e afins. Outra modificação ficou a cargo da padronização de nomes utilizados diferentemente ao longo do ano em que as informações eram coletadas, como o nome dos grupos de serviço que passavam a ter nome diferente ou apresentavam um representação com carcteres especiais que precisava ser substituida. Quanto as dados pluviométricos, um importante pré-processamento que foi realizado, é a modificação dos grupos de coleta em cada endereço para a média de todos os grupos em um dia, isso porque seria muito complexo lidar com o dataset pluviometrico a nível de onde a coleta foi realizada e não sobre um valor médio diario de nível pluviométrico. Os dados processados foram salvos em formato .h5 (hdf) de forma a preservar a forma apos o pré-processamento, garantindo a tipagem das colunas.

5. Integração dos dados

Os dados foram integrados a base de dados das solicitações através das principais columas de indexação de cada *dataset* externo, as principais colunas utilizadas são:

- dataset socioeconômico: merge pelo bairro
- dataset de localização: merge pelo bairro, logradouro e número
- dataset de datas e pluviométrico: merge pela data de demanda.

6. Avaliação de hipóteses

Sobre a análise foram formulados seis hipoteses iniciais a fim de compreender os dados que estavam sendo estudados, de forma a ter um caminho estruturado de exploração, com relação a proposta original existiu a modificação das hipoteses assim que eram retirados dos dados informações relevantes que indicavam um melhor caminho para exploração ou a demostração de uma análise mais relevante.

6.1. Hipótese 1: Análise das solicitações mais freguentes

Para responder essa questão foram utilizados uma serie de procediemntos de contagem a fim de evidênciar serviços que eram utilizados com maior freguência da base de dados, tanto quando ao macrogrupo de serviços fornecidos, quanto ao serviço fornecido especificamente.

Na figura 1, tem-se a visualização dos top 5 grupos de serviços mais frequentes no *dataset*, observe que o número de requisições relacionada a iluminação são bem maiores com relação aos grupos de limpeza urbana e de conservação. A escala logaritmica evidência esse grau de maior freguência.

Se analisar mais a fundo os serviços sobre iluminação pública (Figura 2), pode-se observar que o principal serviço entre eles é o serviço de manutenção de lâmpadas apagadas, seguida dos pedidos para religarem a rede elétrica e apagar determinada lâmpada de uma região. Esses dados podem ser vistos no gráfico acima.

Os dados também foram estudados de forma a analisar o serviço mais frequente por bairro, no qual ficou claro que iluminação pública esta presente em quase todos eles como o mais frequente. Caso o serviço de iluminação

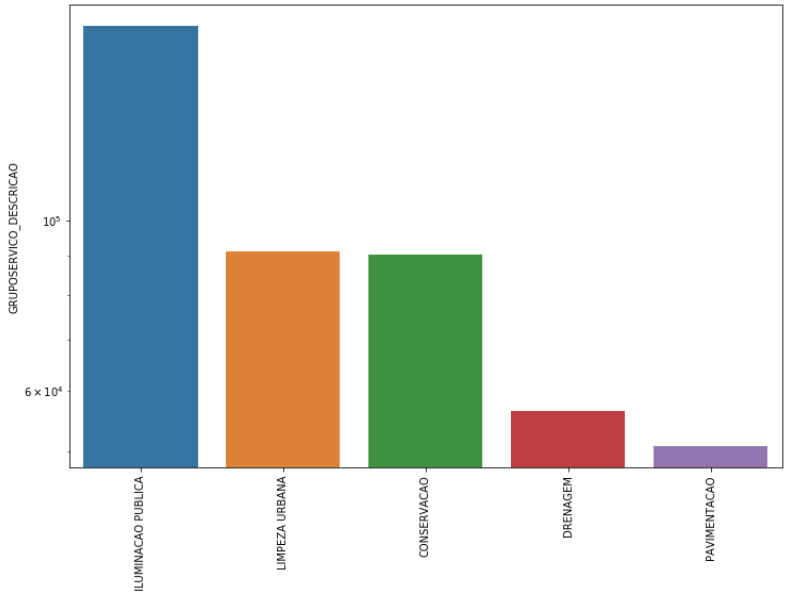

Figure 1: Grupos de serviços mais solicitados a prefeitura do Recife pelo 156

não seja o mais pedido, o de limpeza pública ganha na grande maioria dos casos.

6.2. Hipótese 2: Análise das solicitações sobre diposição gráfica

A fim de compreender como é o comportamento da disposição gráfica dos dados de solicitação na cidade do Recife, foram obtidos dados sobre latitude e longitude gráfica de forma a evidênciar essa carateristica, foram obtidos 60.000 dados de localização respectiva a base de 2018, não foram obtidos de outros anos devido o elevado numero de requisiões necessárias para a obtenção.

Através da plataforma tableau foi possível visualição a disposição gráfica de pontos no espaço, no caso da visualização na figura 3, tem-se a disposição de pontos de solicitação na cidade com a segmentação por bairros observe

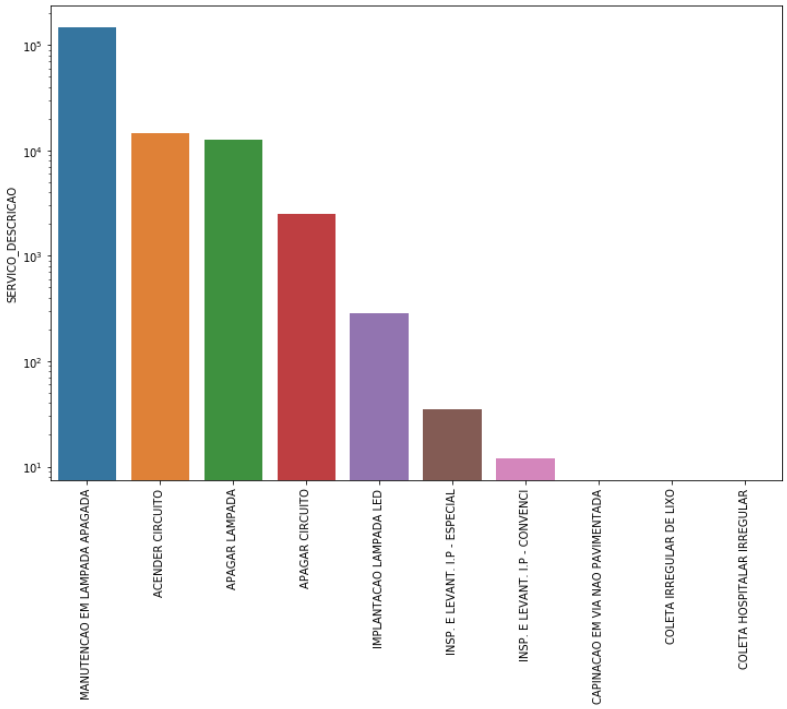

Figure 2: Grupos de serviços mais solicitados a prefeitura do Recife pelo 156

que os pontos parecem apresentar comportamente quase uniforme quanto ao número de serviços em toda a cidade.

Observando especificamente uma região, percebe-se que distribuição de tipos de serviço realizados na cidade também parece ter uma determinada uniformidade, além disso, uma característica importante é que a maioria dos dados apresentam uma rasterização sobre ruas, devido o caráter do endereço, sendo um comportamento já esperado.

6.3. Hipótese 3: Análise das solicitações sobre tempo de atendimento

Nas observações, foi visto que o tempo médio geral para realizar um atendimento é de 18 dias, serviços como: recuperação estrutural e relaciona-

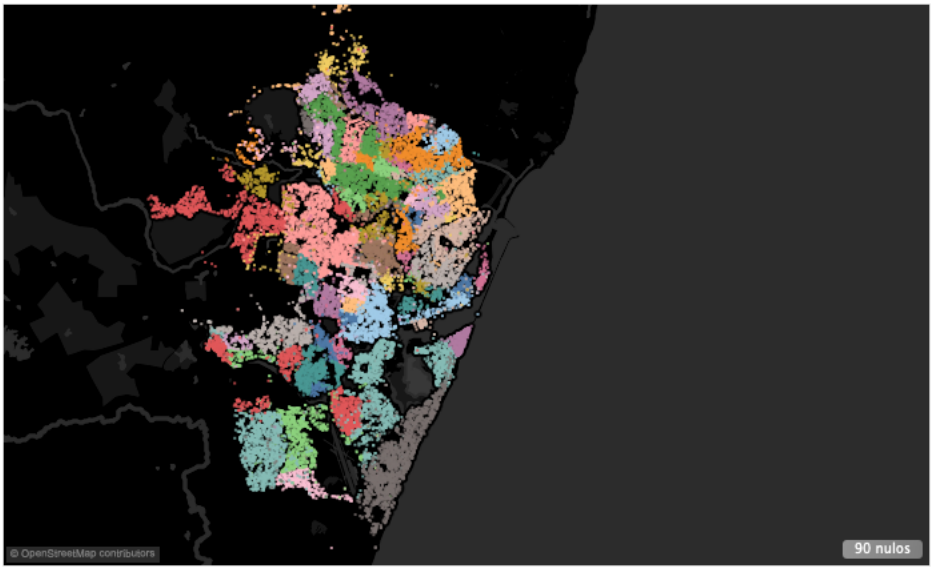

Figure 3: Disposição gráfica de pontos onde foram pedidos ações da solicitações 156 por bairro

Figure 4: Disposição gráfica de pontos onde foram pedidos ações da solicitações 156 no bairro de Santo Amaro, cores descrevem diferentes tipos de serviços.

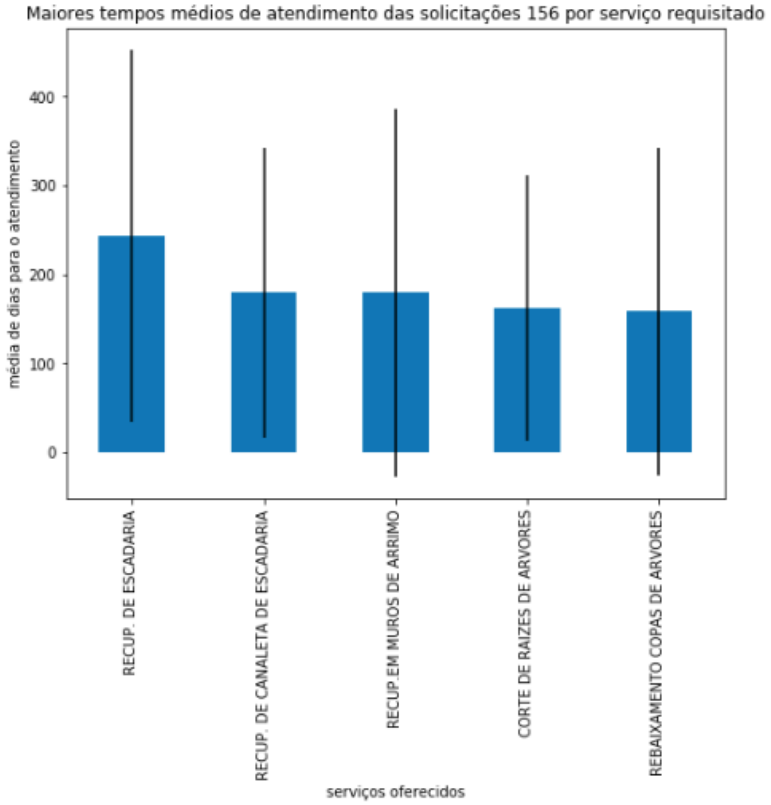

Figure 5: Maiores médias do tempo de atendimento dos serviços do 156

dos a iluminação publica, em média tomam mais tempo, cerca de 100 dias ou mais (figura 5). Já serviços como: remoção de animais mortos e de instalações provisorias tendem a ter atendimento rápido, cerca de 2 dias ou menos. Alguns serviços demotraram comportamente inusitado, sendo atendidos no mesmo dia em que foi requisitado na média, alguns são: cobranças de taxas irregulares, exumação, velório e levantamento topográfico (figura 6).

O tempo para execução de um atendimento pode variar em cerca de 61 dias para mais ou para menos. alguns serviços como: recuperação de praças,

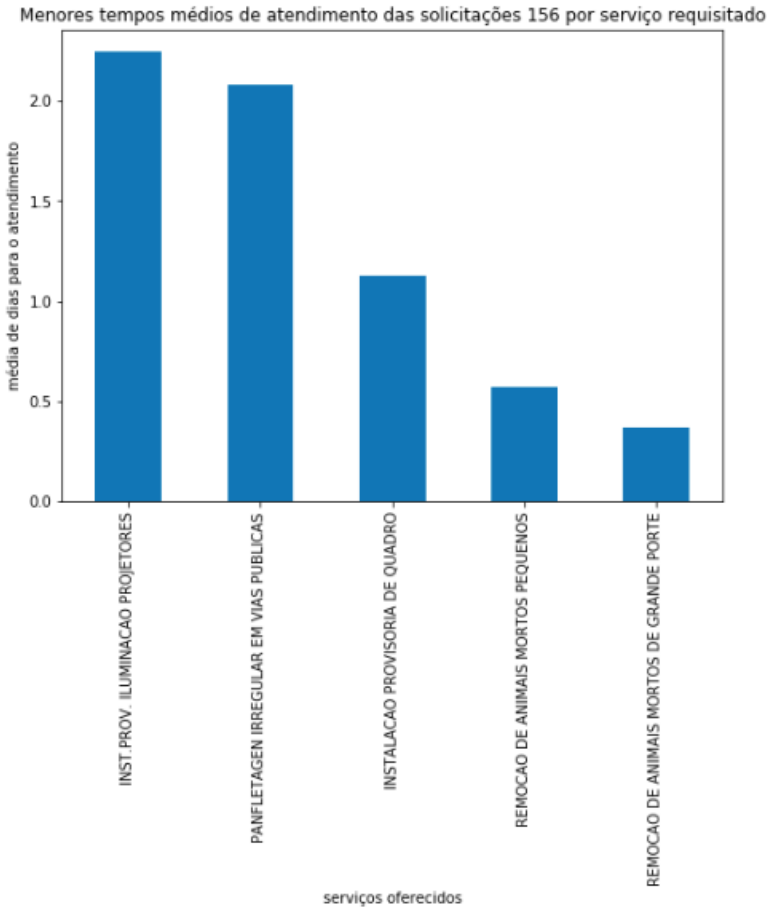

Figure 6: Menores médias do tempo de atendimento dos serviços do 156

brinquendo, instalação de luminarias e poda de copas de árvores, podem variar cerca de 150 dias para mais ou menos. Já os serviços mais rápidos, como: remoção de animais mortos e submissão de aceitação de projetos, possui variação baixa, cerca de 1 dia mais ou menos.

A distribuição dos intervalos indica que normalmente as solicitações são atendidas nos primeiros 90 dias desde sua requisição, isso porque normalmente as medias e medianas são proximos de 90 dias ou abaixo, com releção aos valores maximais de cada distribuição.

Já a correlação do tempo médio de cada tipo de serviço com relação ao ano, demostrou que a maioria dos serviços teve tempo médio, sendo reduzindo ao longo do ano, oferecendo um serviço mais rápido para os cidadões, é o caso de serviços funerários e de manutenção ou implantação de luminárias e quadras, isso é serviços estruturais. No entando, também notasse que alguns serviços pioraram, tendo um aumento no tempo médio de execução, como serviços educacionais, e de coleta que estão demorando mais tempo para serem execultados com o passar do ano (figura 7).

6.4. Hipótese 4: Análise das solicitações sobre dados socioeconômicos

O *dataset* de perfil socioeconômico é a primeira informação externa a ser utilizada no enriquecimentos dos dados. Vale reasaltar que alguns dados podem apresentar caráter único, que não possuem representatividade, ou seja, que não agrega informação útil, são os dados de RPA, microregião, e distancia para o marco zero.

Com o intuito de primeiramente perceber o comportamento dos dados socioeconômicos, foi realizada uma pequena análise exploratória através desses dados a fim de identificar suas caracteristicas. Análisando alguns dados sim-

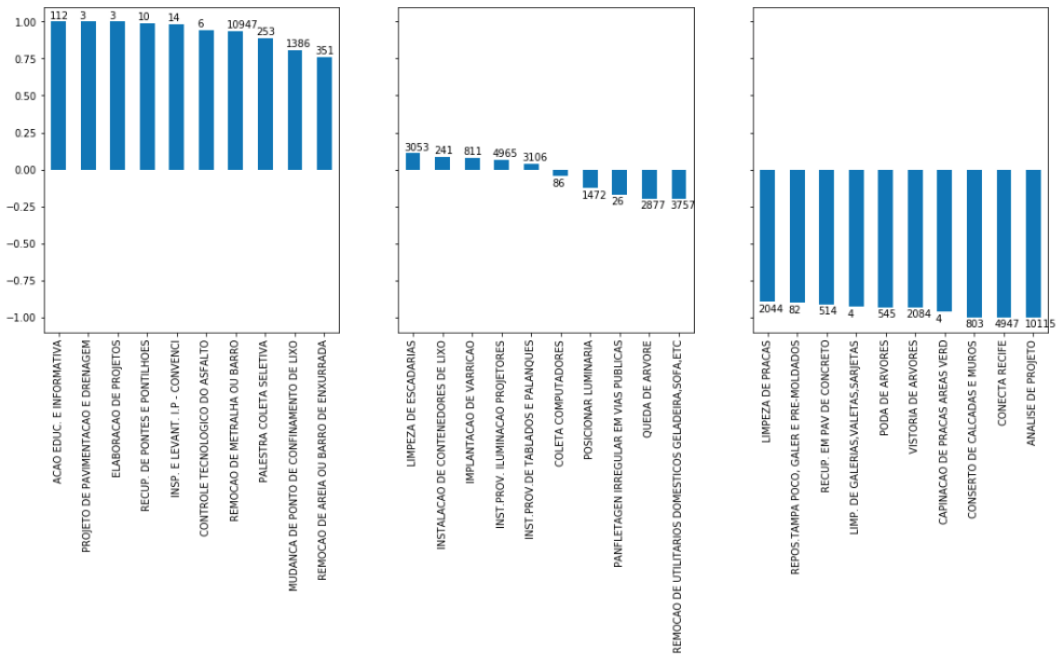

Figure 7: Correlação da média de valores por tipo de serviço entre os anos

ples, como a amplitude de renda, ou seja, a diferença entre os que ganham mais e ganham menos, bem como a média e o desvio padrão, pode-se observar que existe um elevado grau de desigualdade social entre os bairros, a amplitude obtida foi de R$ 10.772,79, que é um valor substanciamente elevado. A média dos valores está em R$ 2.912,72, com desvio padrão de R$ 2.659,81, observe que o desvio padrão é tão elevado quanto a média, evidenciando ainda mais essa noção de disparidade social.

Na obtenção de informações correlacionadas entra os dados socioeconômicos, pode-se citar os seguintes comentários: Uma relação forte, mais proxima de um, entre renda e a porcentagem de alfabetizados é uma relação que faz sentido e que foi evidênciadas nos dados, visto que níveis mais elevados e educação propiciam melhores ganhos, consequentimente aumento de renda. Existe uma relação mais fraca entre renda e crecimento anual da população. Essa relação pode ser mais fraca devido ao fato de que o crescimento populacional não impacta a curto prazo a renda media do local. Ao analisar a renda com a porcentagem de homem e mulheres, observousse um impacto das mulheres na renda um tanto significativo. Já para o porcentagem por faixa de idade, temos um fato curioso, a renda e a taxa de alfabetizados, tem relação apenas para a faixa entre 25 e 60 anos. O que indica que melhores condições de vida, como bons ganhos e uma educação básica, podem influenciar na longevidade de uma população.

Quando se fala em termos de correlação, considere que uma boa taxa de aceitação, e aquela em que apresentam valores acima de 0.5 ou -0.5 para indicar correlações positivas e negativas respectivamente, pode-se notar que a quantidade de solicitações tem relação com todas as colunas que dizem

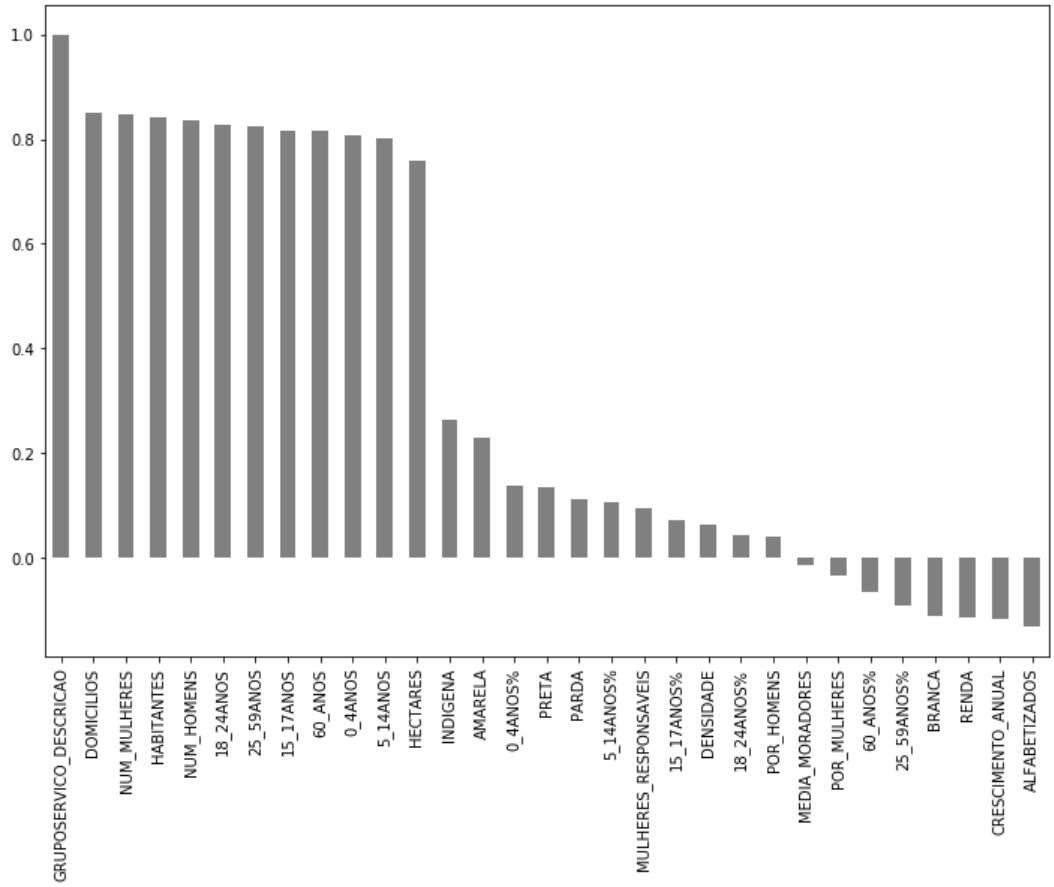

Figure 8: Correlação de número de solicitações por bairro com dados socioeconômicos

respeito, ou se relacionam com a quantidade de habitantes do bairro. Um fato importante a ser falado é que o número de solicitações não possui uma relação com a renda do (figura 8).

Como citado anteriormente, não é possível verificar um relação forte entre a renda do bairro, e questões sociais com o tipo de solicitação feita. Esse fato pode ser observado na figura 9 que mostra os valores máximos e minimos das correlações de solicitações feitas por tipos de serviços. Podemos notar,

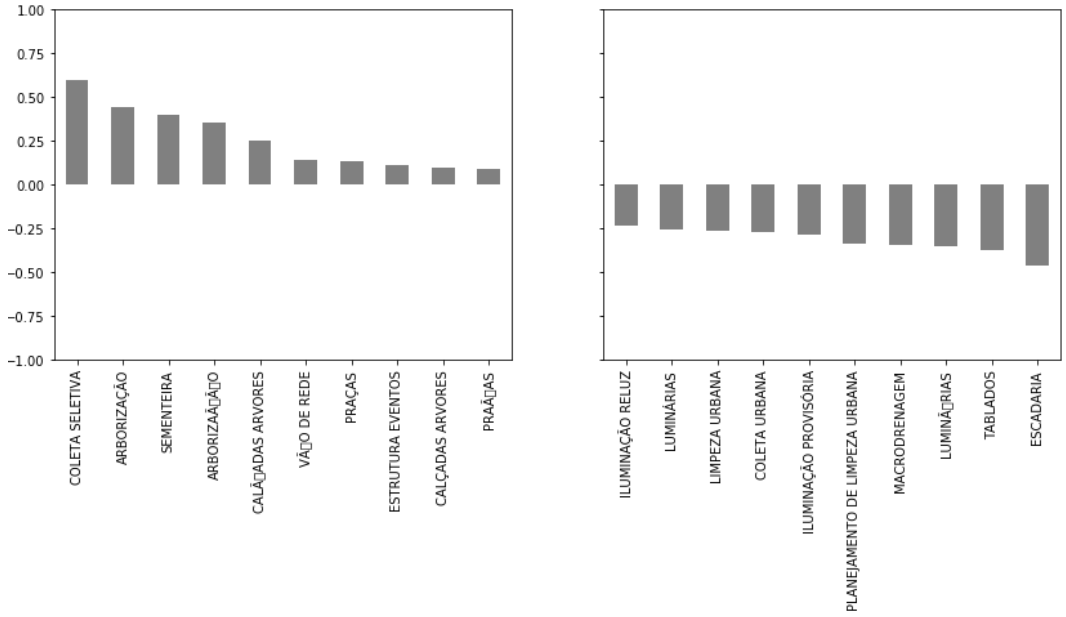

Figure 9: máximos e minimos das correlações de solicitações feitas por tipos de serviços

que os tipos de solicitação estão bem dispostos entre os diferentes bairros do Recife. Não havendo uma aglomeração especifica que permita identificar grupos agregados.

6.5. Hipótese 5: Análise das solicitações sobre dados pluviométricos

Sobre os dados pluviométricos, pode-se concluir que em média, Recife possui um volume ponderado de chuvas, cerca de 50mm, e que esse valor tende a se preservar durante os anos, permanecendo na faixa de 40 a 60mm por ano (figura 10). Observe também que com relação aos meses, Recife parece ter um período chuvoso entre e março e agosto, e um periodo com menor índice de chuvas entre setembro e fevereiro (figura 11). Pode-se resaltar também que a dispersão é bem elevada com relação a média, apresentando diferenças significativas durante ao ano (desvio de 115mm). Observe também o histograma,

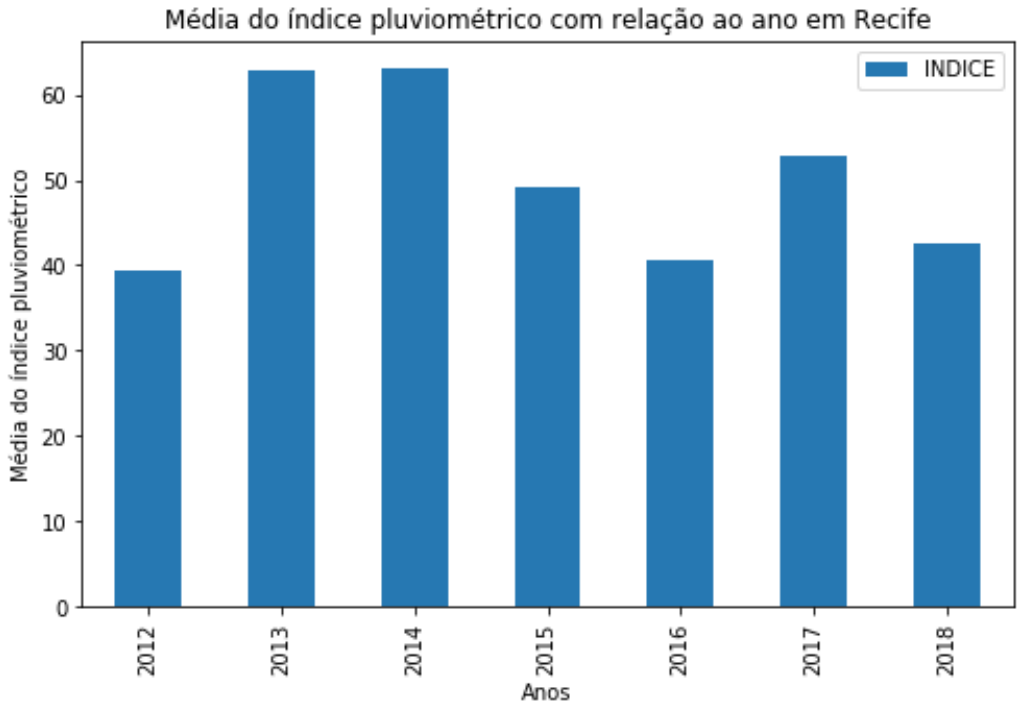

Figure 10: Média de índice pluviométrico durante ao ano

pode-se observar uma distribuição exponencial, onde temos um número elevado de chuvas leves, muito maior com relação a intemperes moderadas e elevadas (figura 12).

Quanto a aspectos que conectam os índices pluviométricos com os diferentes grupos de serviços e tipos específicos de serviços, pode-se dizer que os resultados foram pouco expressivos, aparentemente a correlação de Pearson e Spearman demostraram uma correlação fraca entre o índice e a arborização e drenagem, que foram os maiores valores ligados a pluviometria, ficando respectivamente 0.14 e 0.10, vale resaltar que os valores de p-value se mostraram expressivos nessa situação, evidenciando que realmente a cor-

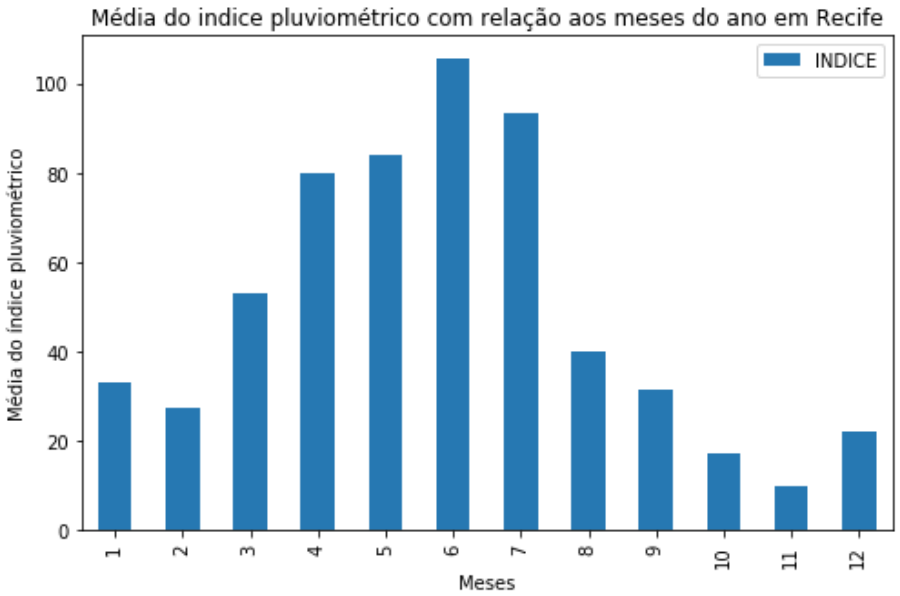

Figure 11: Média de índice pluviométrico ao mês durante os anos de 2012 a 2018

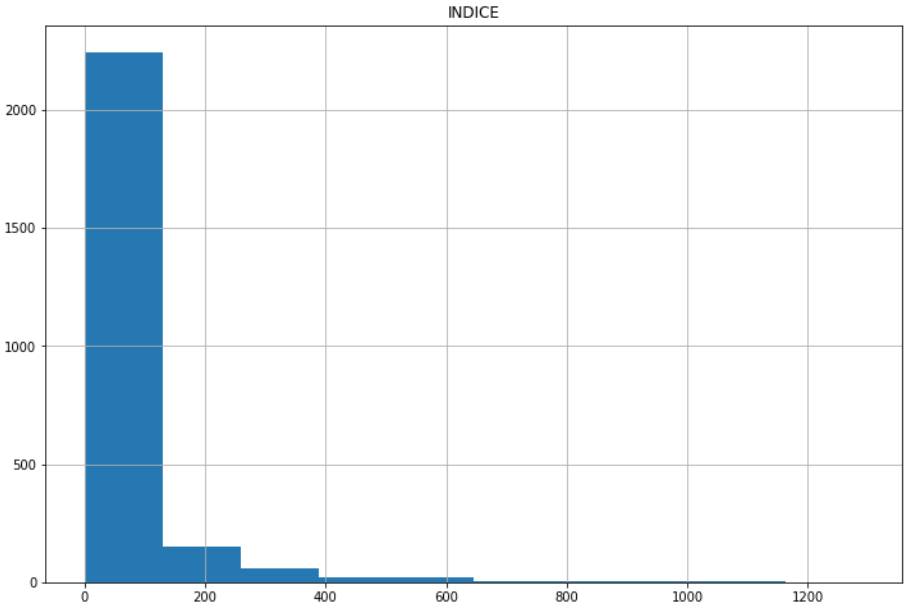

Figure 12: Histograma dos índices pluviométricos em Recife

relação é fraca, mas existente.

Quanto a tipos específicos de serviços oferecidos, observa-se que os que possuem maior correlação são serviços como: queda de árvore, recuperação de pavimento, limpeza de galeria, erradicação de árvores e vistorias de árvores, que apesar das correlações igualmente pequenas, entre 0.3 e 0.1, apresentam p-value significativamente baixo para ser considerado.

6.6. Hipótese 6: Análise de regressões sobre solicitações e dados pluviométricos

Além da análise de correlações entre os dados pluviométricos e as solicitações, como parte de uma tentativa de aplicar uma inteligência sobre os dados, utilizamos regressão linear multipla sobre os tipos de solicitações em relação ao índice de chuvas. Os dados utilizados foram todos os tipos de serviço, visto que a quantidade desses serviços são altos, e analisar eles aos pares seria muito trabalhoso e fora do tempo viável de projeto.

Segue na figura 13 a tabela com os principais resultados da regressão levando em conta o p-values mais significativos, vale ressaltar que foram poucos os tipos de serviço que apresentaram um valor de coef/p-value que refutam a hipotese nula. Contudo, aquelas que apresentam valores interessantes, tem valores de p-value extremamente próximas de zero. Esses serviços, sem uma análise númerica ou científica, são bem lógicas de serem observadas, como queda de árvores, erradicação de árvores e limpeza de galerias. Mesmo que pareça óbvio, essa inteligência pode ajudar a estruturar melhor os serviços e a preparação para atender esses serviços, melhorando alocação e diminuição no tempo de atendimento.

Vale também resaltar que quando que na obtenção da regressão multipla foi aplicada, o valor do coeficiente de determinação obtido R-Squared foi

	coef	p-value
Intercept	5,19e+01	1,31e-30
QUEDA_DE_ARVORE	4,85e+00	7,03e-18
ERRADICACAO_DE_ARVO..	5,28e+00	3,16e-12
LIMP_DE_GALERIAS_VAL..	1,74e+00	3,79e-08
LIMPEZA_DE_CANALETAS	4,69e+00	2,85e-07
MANUT_EM_LAMPADA_A..	-5,70e-01	1,53e-05
VISTORIA_DE_ARVORES	1,42e+00	5,84e-04
MANUTENCAO_EM_LAM..	-2,99e-01	3,38e-03
TUMULO_DE_FAMILIA	1,23e+02	4,42e-03
RECUP_DE_PAV_PARALE..	5,60e+00	4,82e-03
INSTALACAO_DE_CONTE..	-1,32e+01	7,72e-03

Figure 13: Tabela de regressão multipla de índice pluviométrico com relação aos tipos de serviços oferecidos

significativamente baixo, fazendo com que o modelo explicasse pouquissimo sobre o índice, cerca de 0,173

6.7. Hipótese 7: Análise de regressão não-linear sobre solicitações e dados pluviométricos

Também foi realizado a análise de regressão sobre algoritmos não-lineares como é o caso do kNN, assim foi possível avaliar se a existência de uma modelagem não-linear explicaria melhor o problema.

Para isso foi estruturado um *dataset* com datas vetorizadas de forma a evidênciar periodos de tempo e com a informação de indice pluviométrico normalizado, o objetivo era prever o número de solicitações 156 que seriam feitos para a indicação de queda de árvore, como resposta obtida, ficou evidência que apenas com esses dois criterios não era possível descrever com certeza o número de solicitações a ser realizada. O que válida isso é o coeficiente de determinação R-Squared onde foi obtido valor 0 ou negativo, e valores de erro médio absoluto de cerca de 2 solicitações para mais ou para menos, o

que é expressivamente alto.

7. Conclusão

Após as análises das hipóteses, e o estudos dos dados, notamos a importância da existência de dados abertos a respeito do cotidiano de uma cidade. Tais dados podem oferecer a possibilidade de que cientistas e empresas utilizem essas informações para ajudar a melhoras os serviços, prever e melhorar a distribuição de recursos, desafogando o estado da tarefa de lidar amplamente com esses dados.

Importantes premissas foram avalidadas no contexto desse projeto, umas das que mais aparentou surpresa foi a ausênca da relação do tipo de serviço com a renda do bairro. A integração de mais dados, bem como uma análise mais detalhada, com dados socioeconômicos mais atualizados, e mais granulares podem ajudar a identificar problemas específicos de cada faixa de renda, o que não conseguimos agora.

Estudos de dados abertos de cidades, e o fornecimento de *insigths* e informações é uma tendência. A transformação de cidades em cidades inteligentes passa pelo bom uso e pela disponibilidade dos dados referentes ao cotidiano da mesma. O que deixa em aberto possibilidades na área.

Referências

[1] Tech. Rep., accessed: 2018-11-29. [Online]. Available: https://drivendata.github.io/cookiecutter-data-science/

[2] Tech. Rep., accessed: 2018-11-29. [Online]. Available: http://dados.recife.pe.gov.br/dataset/central-de-atendimento-de-servicos-da-emlurb-156

[3] Tech. Rep., accessed: 2018-11-29. [Online]. Available: http://www2.recife.pe.gov.br/servico/perfil-dos-bairros

[4] Tech. Rep., accessed: 2018-11-29. [Online]. Available: https://chrome.google.com/webstore/detail/geocode-cells/pkocmaboheckpkcbnnlghnfccjjikmfc

[5] Tech. Rep., accessed: 2018-11-29. [Online]. Available: http://www.calendario.com.br/api_feriados_municipais_estaduais_nacionais.php

[6] Tech. Rep., accessed: 2018-11-29. [Online]. Available: http://www.apac.pe.gov.br/meteorologia/monitoramento-pluvio.php