Análise dos acidentes na cidade de Porto Alegre no ano de 2016

Douglas Soares, Jônatas de Oliveira, Valdemiro Vieira

Recife, Brazil

Abstract

Este é o relatório final da disciplina de Introdução à Ciência dos Dados da Universidade Federal de Pernambuco. Este trabalho é uma análise dos acidentes de trânsito de Porto Alegre ocorridos em 2016, relacionando com os datasets com informações sobre semáforos, lombadas, pardais e ciclovias na mesma cidade.

Keywords: data science, EPTC, geodata, trânsito, acidentes

1. Introdução

Segundo estudo feito pelo Observatório Nacional de Segurança Viária (de São Paulo & Lajolo (2017)), cerca de 400 mil pessoas por ano, são afetadas por acidentes em trânsitos no Brasil, dentre as quais 47 mil morrem. Sendo assim, acidentes de trânsito se mostra um tópico essencial a ser analisado e explorado.

Os dados utilizados neste projeto foram obtidos do portal de dados abertos da cidade de Porto Alegre (EPTC (2016)) e contém informações dos acidentes de trânsito que ocorreram no ano de 2016, localizações de semáforos, pardais, lombadas eletrônicas e ciclovias implantadas na cidade. O dataset de acidentes contém um registro de mais de 12 mil acidentes e nele, pode-se encontrar data, hora, tipo do acidente, vítimas, veículos envolvidos, entre outras informações.

Foi utilizado neste projeto a linguagem Python e as bibliotecas Pandas, Matplotlib, StatsModel para o pré-processamento dos dados, análise exploratória e aprendizagem. E também foi utilizado a linguagem JavaScript e as bibliotecas Leaflet e D3 para geração do mapa com os acidentes, equipamentos de monitoramento e ciclovias, para uma melhor visualização dos dados e

a partir disso, criar novas hipóteses.

A partir do pré-processamento dos dados, análise exploratória e aprendizagem de todos os conjuntos de dados, serão avaliados algumas correlações entre eles, para inferir padrões e verificar hipóteses levantadas pelo grupo. Tais hipóteses são:

- Número de acidentes reduzem no final de semana.
- O centro da cidade é a região que ocorre mais acidentes.
- Existe uma relação entre a natureza do acidente com a quantidade de vítimas.
- Locais com equipamentos de fiscalização possuem menor índice de acidentes.
- Motos contribuem para o aumento de feridos.
- Ruas com mais equipamentos possuem menor quantidade de acidentes.

2. Pré-processamento

Como foi decidido fazer uma visualização geográfica dos dados, era preciso ter latitude e longitude de todos os pontos dos datasets escolhidos. No dataset de acidentes, todos os acidentes tinham latitude e longitude. E o de ciclovias tem um geojson descrevendo elas. Porém, os datasets de semáforos, lombadas e pardais não tinham essas informações. Tinham apenas um endereço descrito por extenso mesmo.

Inicialmente a ideia para obter esses endereços usando geopy e/ou com a api do Google Maps. Porém, ambas tem limite de requisições e não teve como obter todos as latitude e longitudes que era necessário. Aí foi necessário buscar uma alternativa. No fim foi usado Geocode Cells no Google Sheets. Esse complemento também usa a API do Google, porém lá se tem um limite maior de requisições e deu para obter todas as latitudes e longitudes dos semáforos, pardais e lombadas.

Algumas modificações nos datasets foram necessárias para começarem feitos os processamentos dos dados e geração dos gráficos para análise dos dados. Os meses no dataset estavam representados por números, então foi feito uma transformação para o formato literal. A hora dos acidentes estavam no formato hh:mm, porém para as análises, só era necessário a hora, com isso,

os minutos foram removidos. Os endereços dos semáforos, pardais e lombadas foram padronizados, pois estavam descritos com informações a mais. Por fim, alguns tipos de dados foram modificados para ajudar na plotagem dos gráficos.

Para poder validar algumas hipóteses que as ciclovias ajudavam evitar acidentes envolvendo bicicletas e que equipamentos de fiscalização como pardais e lombadas ajudavam a diminuir os acidentes, foi necessário saber quantos acidentes aconteceram próximos a esses equipamentos e se os acidentes aumentaram à medida que se afastava desses equipamentos. Para isso, foi necessário construir funções que dizem a distância entre dois pontos e entre um ponto e uma linha, considerando latitude e longitude. Isso não é tão trivial pois isso não acontece num plano e sim na superfície da terra. Aí foram feitas pesquisas em diversos sites e foram de dúvidas até que foi encontrada uma que era uma aproximação excelente (Scripts (2017)).

Depois de se construir essas funções descritas, aí pode-se reconstruir os datasets acrescentando colunas. No dataset de acidentes, foi informado quantas ciclovias, semáforos, pardais e lombadas tinhas nas seguintes distâncias de raio: 10m, 20m, 50m, 100m, 200m, 500m e 1000m. Reforçando que apenas nos acidentes que envolviam bicicletas que foram acrescentadas as informações de ciclovias próximas, afinal esse era o objetivo. E nos datasets de lombadas, pardais e ciclovias foi acrescentado também quantos acidentes ocorreram nas seguintes distâncias de raio: 10m, 20m, 50m, 100m, 200m, 500m e 1000m. Reforçando novamente que no de ciclovias foram apenas os acidentes que envolveram alguma bicicleta.

Pode-se notar um problema depois que esses datasets foram construídos, a medida que os raios eram aumentados, outras ruas vizinhas eram também pegas o que prejudica na análise da eficácia dos equipamentos de fiscalização. Aí o ideal seria pegar apenas os acidentes que acontecem na mesma rua do pardal ou da lombada. Para isso, foi feito um novo pré-processamento para mapear as ruas dos datasets de pardais e lombadas para o de acidentes e vice-versa. Devido a irregularidade nos datasets e o fato dos datasets de lombadas e pardais serem pequenos, o mapeamento foi feito a mão mesmo. Depois disso, se conseguiu filtrar apenas os acidentes que aconteciam na mesma rua do pardal e da lombada.

No dataset de ciclovias, isso não se mostrou algo tão preocupante, afinal tinha como objetivo mostrar que nas ciclovias tinham nenhum ou pouquíssimos acidentes envolvendo bicicletas e em locais mais distantes de ciclovias tinham uma quantidade bem mais alta. Mostrando assim, consequentemente, que as

ciclovias ajudam a diminuir os acidentes envolvendo bicicletas. Já no dataset de semáforos não foi possível fazer pois era irregular a nomenclatura das ruas levando em consideração o de acidentes e era um pouco grande para fazer um mapeamento a mão. Sem levar em consideração que observando a visualização de acidentes versus semáforos no mapa, não mostra uma relação tão grande entre a quantidade de acidentes e a presença de semáforos.

Com as informações descritas acima, ainda foi construído outro dataset onde mostrava quantos acidentes tinha em determinada rua ou avenida e quantos pardais e lombadas tinham naquela determinada rua. Esse dataset foi feito para ser utilizado na parte de aprendizagem de máquina que será descrita mais à frente.

3. Análise Exploratória

Nesta etapa foi feito um estudo aprofundado com as informações dos datasets, correlacionando-as para entender melhor o motivo dos acidentes e extrair novos conhecimentos.

3.1. Mapas do Leaflet

Com o intuito de ter uma visão mais geográfica dos dados para que pudesse formular algumas hipóteses adicionais foi decidido fazer plotagens dos dados em mapas. Foi escolhido usar Leaflet de JavaScript por ela ser bem simples de ser utilizada.

Foram feitas quatro visualizações bem básicas para poder visualizar de maneira melhor a relação dos dados. A primeira foi relacionando os acidentes que tem alguma bicicleta envolvida com a rede de ciclovias da cidade, A segunda foi relacionado os acidentes com os semáforos. A terceira relacionando acidentes com pardais com acidentes. E a última relacionando lombadas com acidentes.

Com essas visualizações foram obtidos algumas visões que não dava para perceber apenas olhando os dados diretamente. Pôde-se perceber, por exemplo, que os acidentes envolvendo bicicletas estavam um pouco distantes de ciclovias, que semáforos não tinham um fator tão grande para evitar acidentes de trânsito pois eles continuavam acontecendo com frequência próximas a semáforos e que em vias que tinham pardais e/ou lombadas os acidentes aconteciam com menos frequência próximos a esses equipamentos de fiscalização.

Com essas observações, pode-se pensar em maneiras mais claras para validar essas hipóteses que extraímos dessas visualizações. Isso será relatado com mais detalhes adiante.

3.2. Visualizações

3.2.1. Análise das Regiões

Como geralmente os centros das cidades costumam ter uma maior movimentação de pessoas, por ser um local centrado no comércio, o grupo teve como hipótese que o centro da cidade de Porto Alegre é a região que possui mais acidentes. Para isso, ao analisar o mapa da cidade (Figura 1), nota-se que o tamanho das regiões é de uma diferença gritante.

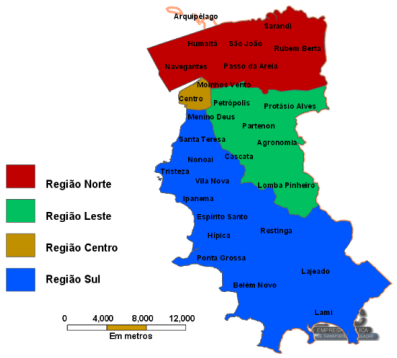

Figure 1: Regiões de Porto Alegre

Com isso, para pegar a proporção de acidentes em cada região, foi utilizado número total de acidentes que ocorreram naquela região pela quantidade de ruas de cada região. Na figura 2, pode-se notar que de fato, o centro é a região que mais ocorre mais acidentes, mesmo sendo a menor região.

Isso também pode ser visto no gráfico que mostra a quantidade de feridos, feridos gravemente e mortos por região (figura 3). No qual, o centro possui praticamente 1/3 de acidentes comparado às outras regiões.

3.2.2. Análise Temporal

Nesta etapa, foi tido como hipótese que o número de acidentes é reduzido no final de semana, pelo fato de não ter escola e o número de pessoas que trabalham reduz. Para validar a hipótese, foi feito o gráfico quantidade de acidentes por dias da semana (Figura 4), este gráfico confirma a hipótese e pode-se notar ainda que o número de acidentes no domingo cai para menos

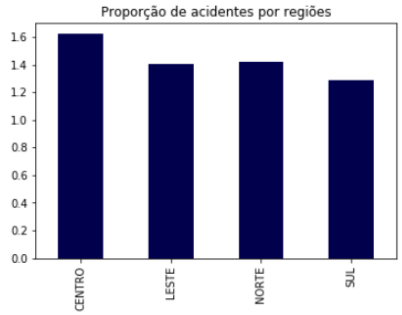

Figure 2: Proporção de Acidentes por Região

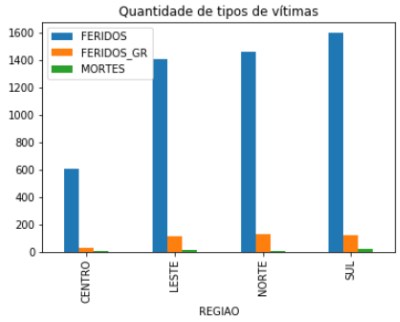

Figure 3: Quantidade de Vítimas por Tipo e Região

da metade e sábado cai para um pouco mais da metade comparado aos dias da semanas úteis.

Ainda analisando a Figura 4, a sexta-feira é o dia que possui mais acidentes, o que é algo esperado pelo fato de ser o último dia útil da semana e as pessoas costumam sair para se divertir. Mas para entender melhor o porque, foi gerado alguns gráficos relacionando com as horas dos acidentes e notou-se que os horários dos acidentes assim como as quantidades nos dias das semanas eram semelhantes (Figura 5), tendo como o horário com mais acidentes o das 8 horas da manhã (Figura 6), o que faz sentido por ser um horário de pico no trânsito, e com um leve aumento na parte da noite da sexta-feira.

Nota-se na figura 5 que sábado e domingo possuem os maiores números de acidentes quando é comparado somente os horários da madrugada e pode ser visto mais claramente na figura 7, isso pode ser causado por pessoas que saem para se divertir e voltam, dirigindo alcoolizados ou aumentarem a velocidade

Figure 4: Quantidade de Acidentes por Dias da Semana

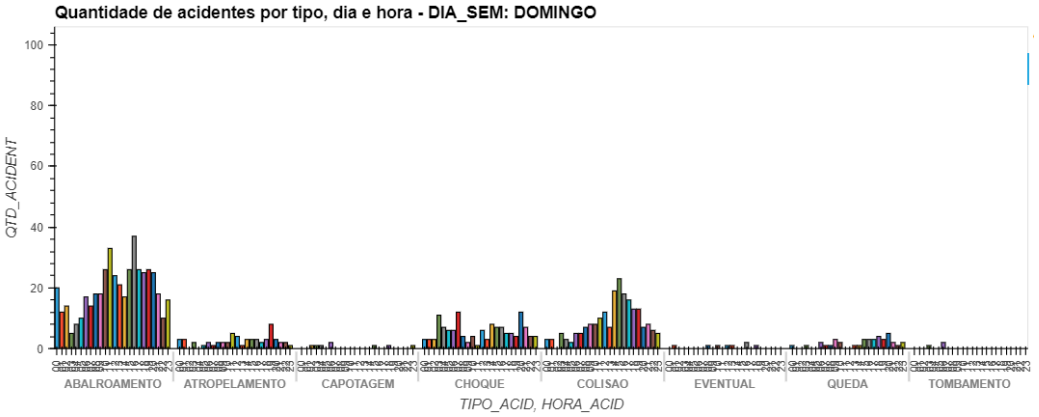

Figure 5: Quantidade de Acidentes por Dia e Hora

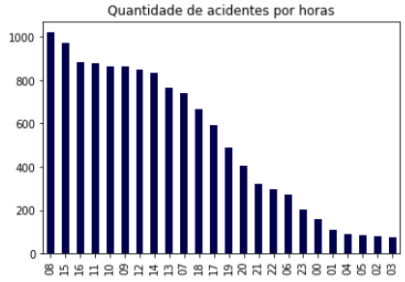

Figure 6: Quantidade de Acidentes por Horas

Figure 7: Quantidade de Acidentes da Madrugada por Dias da Semana

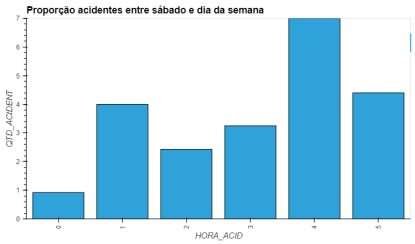

Figure 8: Proporção de Acidentes entre Sábado e Dia da Semana

por causa dos semáforos e lombadas estarem desligados, na madrugada da sexta pro sábado e do sábado para domingo.

Analisando a proporção entre os finais de semana com dia da semana (Figura 8), percebe-se que a quantidade de acidentes chega a ser 7 vezes maior.

Além disso, os tipos de acidentes que mais ocorrem nestes horários são os que envolvem batidas de carros (Figura 9), o que reforça o fato das pessoas dirigirem passando do limite máximo de velocidade ou alcoolizados. Analisando as horas separadamente, é possível ver que o número de acidentes de abalroamento às 4 horas da manhã chega a ser 16 vezes maior no sábado comparado a um dia útil, uma quantidade alarmante e que merece atenção a fim amenizar este problema.

Partindo para analisar a relação dos acidentes com os meses (Figura 10), fevereiro é o mês com menos acidentes e março o que possui mais, o motivo deste acontecimento é por que como fevereiro é época de carnaval e férias de colégio e faculdade, de acordo com pesquisas (do Brasil (2016)), as pessoas costumam viajar durante o período de férias de meio de ano e fim de ano e, consequentemente, diminui o número de carros nas vias. No final de fevereiro faculdades e colégios costumam voltar às atividades, porém, excepcionalmente em 2016, houve paralisação dos professores e as aulas só

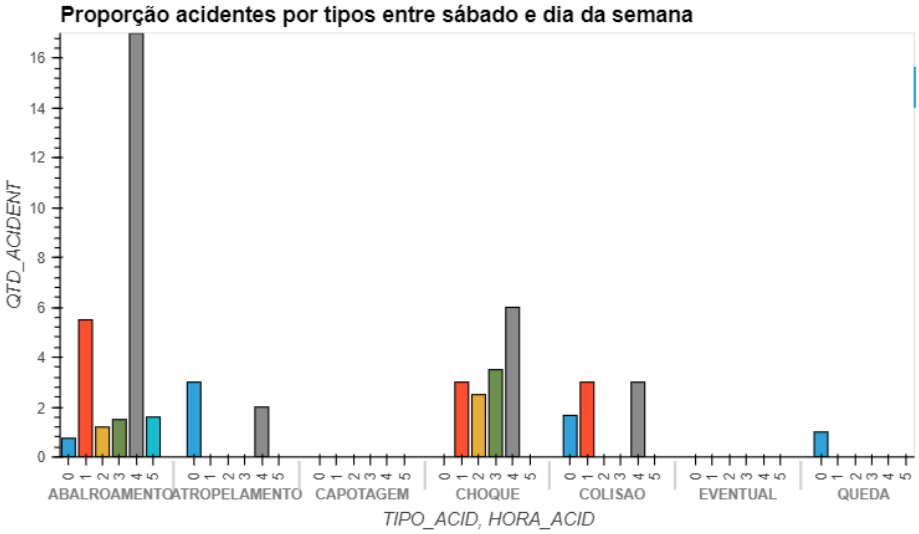

Figure 9: Proporção de Acidentes entre Sábado e Dia da Semana e Tipo de Acidente

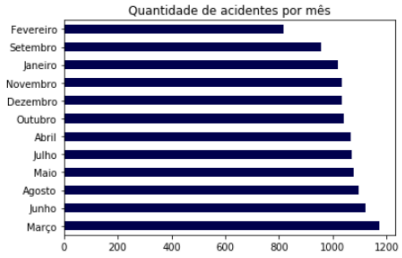

Figure 10: Quantidade de Acidentes por Mês

voltaram em março (Gauchazh (2016)), com esta notícia, as pessoas provavelmente voltaram de viagem no começo de março, sendo mais um ponto para aumentar o número de acidentes neste mês. Este fenômeno também pode ser visto nos meses de julho e agosto, mas de forma menos expressiva.

3.2.3. Análise de Vítimas

A fim de procurar alguma relação entre a natureza/tipo do acidente com a quantidade de vítimas, fez-se uma média das vítimas por tipo de acidente (Figura 11) que deu como resultado atropelamento e queda fatores que causam vítimas em quase todos os casos, o que é esperado, pois estes tipos de acidentes envolvem a pessoa diretamente.

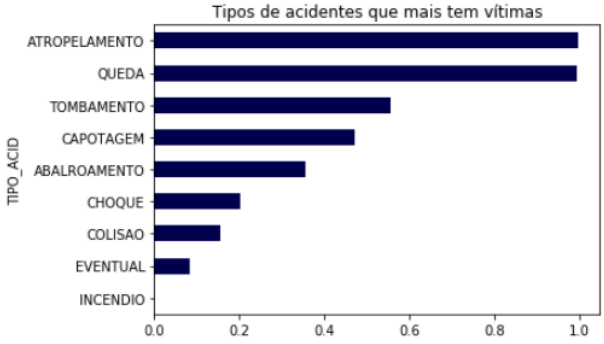

Figure 11: Tipos de Acidentes por Quantidade de Vítimas

Figure 12: Quantidade de Acidentes Relacionado a Distância de Pardais

3.2.4. Cruzamento de Datasets

Como dito anteriormente, foi feito um cruzamento dos datasets para entendermos o comportamento dos acidentes perto dos equipamentos de monitoramentos e ciclovias implantados na cidade.

Para entender tais comportamentos, foi gerado os gráficos de quantidade de acidentes por distância dos equipamentos e ciclovias (Figura 12, 13 e 14). O resultado era esperado, no qual tem um crescimento exponencial quanto mais longe, o que infere que tais implantações diminuem a quantidade de acidentes.

As próximas visualizações foi utilizada como distância a de até 200 metros dos equipamentos e ciclovias.

Para reforçar as hipóteses que locais com fiscalização (lombadas eletrônicas e pardais) possuem um menor índice de acidentes, observa-se na figura 15, que o número de acidentes perto dos equipamentos é bem menor comparado ao

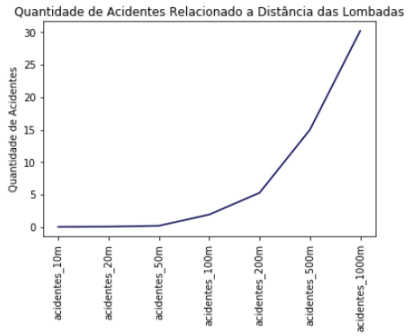

Figure 13: Quantidade de Acidentes Relacionado a Distância de Lombadas

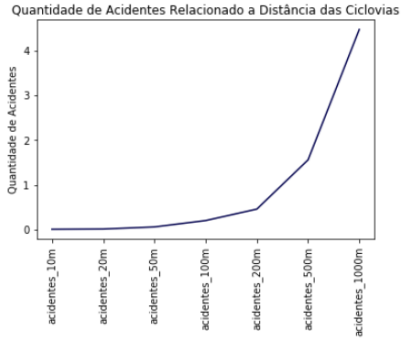

Figure 14: Quantidade de Acidentes Relacionado a Distância de Ciclovias

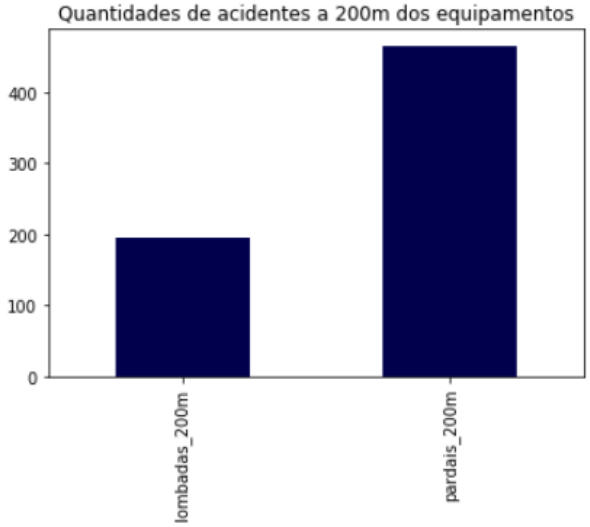

Figure 15: Quantidade de Acidentes à 200 Metros dos Equipamentos

total de acidentes na cidade, principalmente próximo a lombadas eletrônicas (200 acidentes de 12 mil), que são equipamentos mais visíveis quando comparado a pardais.

4. Aprendizagem

Com o objetivo de observar padrões e/ou extrair algum tipo de conhecimento dos dados, se fez necessário uso de técnicas de aprendizagem para tentar validar hipóteses. Mais especificamente, foi utilizado o algoritmo de Regressão Linear Simples e Múltipla. Esta técnica de aprendizado é essencialmente boa, pois é simples, no sentido de tentar representar os dados a partir de um hiperplano, predizendo um valor numérico a partir das variáveis. Apesar das limitações, principalmente por ser simples, esta técnica oferece bastante informações sobre a base estudada.

Através do senso comum, o grupo tomou como hipótese sobre a relação entre a quantidade de vítimas de um acidente (feridos, feridos gravemente e morte) com os transportes, que acidentes onde motos estão envolvidas tende a aumentar o número de vítimas. Para provar ou refutar esta hipótese, foi utilizado a regressão linear com o objetivo de relacionar os veículos envolvidos nos acidentes com a quantidade de vítimas. Primeiro, foi calculado a correlação entre a quantidade de cada veículo envolvido no acidente com a quantidade de afetados, para tentar perceber quais as variáveis que mais influenciam no número de vítimas. As correlações mais altas e positivas, foram indicadas por: ônibus, moto e bicicleta. Com isso, obteve-se um direcionamento para cálculo das regressões.

Calculada a regressão simples para cada variável, notou-se que realmente os veículos apontados pela correlação apresentam uma relação de crescimento de acidentes quando os mesmos veículos citados estão envolvidos. Porém, para ônibus o p-value se apresentou muito alto, garantindo menos de 99% de certeza, logo a regressão não representa tão bem os dados para ônibus. Já para moto e bicicleta, o p-value apresentou valores bastante pequenos, como apresentado nas figuras 16 e 17, confirmando que a regressão representa bem os dados para estes transportes, provando então a hipótese do grupo que moto causa mais feridos quando está envolvida em acidentes, como pode ser visto através dos coeficientes que fazem a quantidade de acidentes crescer quando o números destes mesmos veículos aumentam, e especificamente moto apresenta um coeficiente bem alto. Além disso, outra regressão simples foi feita para carro, e através da figura 18, percebe-se que, por o p-value ser

	coef	SE	t	p-value
Intercept	0.235474	0.005793	40.647792	0.0
MOTO	0.716620	0.011608	61.735648	0.0

Figure 16: Relação Entre Moto e Afetados em Acidentes

	coef	SE	t	p-value
Intercept	0.397993	0.005834	68.218444	0.000000e+00
BICICLETA	0.587014	0.053831	10.904802	1.454508e-27

Figure 17: Relação Entre Bicicleta e Afetados em Acidentes

muito baixo (tendendo a zero) e o coeficiente ser negativo, quando carros estão envolvidos em acidentes, a quantidade de feridos tende a diminuir.

Outra hipótese levantada pelo grupo, está relacionada com as figuras 12 e 13, e tem a ver com o número de equipamentos em cada rua. Como mostrado nas imagens, o número de acidentes aumenta exponencialmente em fator do aumento da distância dos equipamentos em questão (lombadas e pardais), logo supõe-se que a quantidade de acidentes nas ruas com equipamentos, era menor do que as que possuíam. Para provar ou refutar esta hipótese foi realizado uma regressão linear múltipla que tem como objetivo prever a quantidade de acidentes baseado na quantidade dos equipamentos citados, baseando-se nos dados gerados agrupados por rua. Porém, através da figura 19, pode-se concluir que os equipamentos eles influenciam para o aumento de acidentes, pois o p-value é alto (a regressão representa bem os dados) e o coeficiente é positivo, fazendo com que a função cresça junto com a quantidade de equipamentos. Sendo assim, haveria uma divergência entre os resultados das figuras 12 e 13 com a da regressão. Com uma pesquisa aprofundada do grupo sobre assunto, foi encontrado que os equipamentos de trânsito eles são aplicados em ruas que já possuem alto índice de acidentes para tentar atingir uma redução dessa quantidade, e que realmente reduzem esta quantidade, evitando cerca de 3 óbitos e 34 acidentes por ano (Perkons (2016)).

	coef	SE	t	p-value
Intercept	0.798580	0.011313	70.589547	0.0
AUTO	-0.281404	0.007068	-39.811800	0.0

Figure 18: Relação Entre Carro e Afetados em Acidentes

	coef	SE	t	p-value
Intercept	6.867063	0.732220	9.378415	2.781459e-20
N_PARDAIS	58.297061	1.971692	29.567018	4.120380e-148
N_LOMBADAS	14.222351	2.910391	4.886750	1.149620e-06

Figure 19: Relação do Número de Acidentes com os Equipamentos (Lombadas e Pardais)

5. Conclusão

Ao fazer as análises dos gráficos e tabelas pôde-se notar a necessidade de implantar mais equipamentos de monitoramento para reduzir a quantidades de acidentes, como visto nas figura 12 e 13, os números crescem de forma exponencial quando é aumentado a distância dos equipamentos. As lombadas eletrônicas demonstraram-se serem mais efetivas em relação aos pardais. E também existe uma necessidade de aumentar as fiscalizações no período da madrugada, no qual há um aumento considerável de acidentes nos finais de semana, podendo ser causado, por exemplo dos equipamentos estarem desligados.

Também pode-se concluir através de aprendizagem de máquina que motos e bicicletas influenciam mais na quantidade de feridos em acidentes do que outros tipos de veículos. O que faz sentido, afinal eles são veículos abertos e tendem a provocar muito mais danos do que veículos fechados como carros e ônibus.

Na figura 14, também pode-se observar que as ciclovias ajudam a prevenir bastante acidentes envolvendo bicicletas. O que também faz muito sentido pois ali terão locais mais seguros para bicicletas transitarem do que em outras ruas que elas terão que competir espaço com carros e outros veículos. Também foi feita uma aprendizagem para tentar mostrar que ruas que possuem aparelhos de monitoramento tem menos acidentes que as que não possuem. Porém, isso foi refutado. Aprofundando a pesquisa, já que não parecia ter sentido, pode-se perceber que os aparelhos de monitoramento reduzem sim a quantidade de acidentes e eles eram colocados em vias com um índice mais alto de acidentes. E como não tinha dados para fazer uma análise temporal disso, não deu para mostrar isso.

References

do Brasil, G. (2016). Cresce número de brasileiros que pretendem viajar nos próximos seis meses. URL: http://www.brasil.gov.br/noticias/turismo/2016/11/cresce-numero-de-brasileiros-q accessed: 2018-11-29.

EPTC (2016). Conjunto de dados abertos de porto alegre. URL: http://datapoa.com.br/dataset?organization=eptc accessed: 2018-11-29.

Gauchazh (2016). Ao vivo: acompanhe a movimentação de volta às aulas em porto alegre. URL: https://gauchazh.clicrbs.com.br/educacao-e-emprego/noticia/2016/02/ao-vivo-acomp accessed: 2018-11-29.

Perkons (2016). Ao vivo: acompanhe a movimentação de volta às aulas em porto alegre. URL: https://gauchazh.clicrbs.com.br/educacao-e-emprego/noticia/2016/02/ao-vivo-acomp accessed: 2018-11-29.

Scripts, M. T. (2017). Movable type scripts. URL: https://www.movable-type.co.uk/scripts/latlong.html accessed: 2018-11-29.

de São Paulo, F., & Lajolo, M. (2017). Trânsito no brasil mata 47 mil por ano e deixa 400 mil com alguma sequela. URL: https://www1.folha.uol.com.br/seminariosfolha/2017/05/1888812-transito-no-brasil accessed: 2018-11-29.