Estatística Aplicada à Computação/Telemática

Projeto II – Análise de Atividades Policiais em Rhode Island

Nesse projeto, você explorará o conjunto de dados do Stanford Open Policing Project e analisará o impacto do gênero no comportamento policial durante abordagens. Aqui você terá oportunidade de praticar limpeza de dados confusos, criar visualizações, combinar e remodelar conjuntos de dados e manipular dados de séries temporais.

Você irá trabalhar com um projeto guiado, em que cada atividade, do pré-processamento dos dados à análise desejada será indicada *a priori*, e você deverá executar e interpretar os resultados. Esse tipo de abordagem é interessante para que tenhamos uma ideia de que tipos de processamentos e de perguntas podemos realizar com um ou mais conjuntos de dados. Cada atividade será descrita a seguir, devendo ser executada em um único *notebook* com as devidas identificações. Para esse projeto, serão analisados os dados da região de Rhode Island, e os dados estão disponíveis no Moodle, em dois arquidos: *police.csv* e *weather.csv*. A descrição de cada uma das variáveis pode ser vista nesse link.

1. **Preparação dos dados para a análise**

   Antes de iniciar sua análise, é fundamental que você primeiro examine e limpe o conjunto de dados, para tornar o trabalho com ele um processo mais eficiente. Nesta parte, você praticará a correção de tipos de dados, manipulação de valores ausentes e eliminação de colunas e linhas enquanto aprende sobre o conjunto de dados do *Stanford Open Policing Project*.

   (a) Importe o dataset `police.csv`, e indique quantas variáveis estão disponíveis e quantos registros policiais estão catalogados;

   (b) Conte o número de dados faltosos de cada variável;

   (c) Como estamos trabalhando somente com dados de um único estado, não faz sentido mantermos informações de variáveis como `county_name` e `state`. Exclua essas variáveis do conjunto;

   (d) Durante as análises, a coluna `driver_gender` será crítica para muitas de suas análises. Como apenas uma pequena fração das linhas está faltando `driver_gender`, remova essas linhas do conjunto de dados;

   (e) Vá na descrição dos dados e verifique que tipo de variável são `search_conducted`, `is_arrested` e `district`; depois compare com os tipos de dados que estão armazena-

dos no conjunto de dados. Havendo necessidade, faça a modificação dos tipos de dados para essas variáveis;

(f) A data e a hora de cada parada de tráfego estão armazenadas em colunas separadas: `stop_date` e `stop_time`. Combine essas duas colunas em uma única coluna, nomeando-a como `stop_datetime` e converta no formato data e hora (tipo `datetime`). Isso habilitará atributos baseados em data convenientes que usaremos posteriormente no projeto;

(g) Por fim, transforme a coluna `stop_datetime` no índice do `dataframe`.

2. **Análise do gênero sobre as infrações**

O gênero do motorista influencia o comportamento da polícia durante uma parada no trânsito? Nesta parte, você explorará essa questão enquanto pratica filtragem, agrupamento e muito mais!

(a) Antes de comparar as infrações cometidas por cada gênero, você deve examinar as infrações cometidas por todos os motoristas para obter uma compreensão básica dos dados. Construa uma distribuição de frequências da variável `violation` e responda: qual a infração mais comum e a menos notificada?

(b) O interesse agora é responder a seguinte questão: *motoristas do sexo masculino e feminino tendem a cometer diferentes tipos de infrações de trânsito?* Para isso, crie uma tabela de contingência para frequência absoluta e outra para frequência relativa, contendo a distribuição conjunta das variáveis `driver_gender` e `violation`.

(c) Construa um gráfico de barras agrupadas para ilustrar os dados das tabelas de contingência construídas;

(d) Quando um motorista é parado por excesso de velocidade, muitas pessoas acreditam que o gênero influencia se o motorista receberá uma multa ou um aviso. *Você pode encontrar evidências disso no conjunto de dados?* Para tentar responder essa pergunta, crie uma tabela de contingência considerando as variáveis `driver_gender` e `stop_outcome` e então vai comparar a porcentagem de paradas resultados de uma "Citation"versus um "Warning"(veja o link que descreve as variáveis caso tenha dúvidas de compreensão);

(e) *O gênero afeta a escolha de veículos a serem revistados?* Para responder essa pergunta, primeiro, calcule a porcentagem de todas as paradas no DataFrame que resultam em uma revista de veículo;

(f) Em seguida, filtre o DataFrame por gênero e calcule a taxa de pesquisa para cada grupo separadamente. Dica: você executará o mesmo cálculo para ambos os gêneros ao mesmo tempo usando `groupby`;

(g) Considere agora a hipótese de que a taxa de revista varia de acordo com o tipo de infração, e a diferença na taxa de revista entre homens e mulheres é porque eles tendem a cometer infrações diferentes. Calcule a taxa de infração para cada combinação de gênero e infração. *Homens e mulheres são revistados com a mesma taxa para cada infração?*

3. **Análise exploratória visual dos dados**

   É mais provável que você seja preso em uma determinada hora do dia? As paradas relacionadas às drogas estão aumentando? Nesta parte, você responderá a essas e outras questões analisando o conjunto de dados visualmente, uma vez que os gráficos podem ajudá-lo a entender as tendências de uma forma que o exame dos dados brutos não pode.

   (a) Quando um policial para um motorista, uma pequena porcentagem dessas paradas termina em uma prisão. Isso é conhecido como taxa de prisão. Você descobrirá se a taxa de prisão varia de acordo com a hora do dia. Primeiro, você calculará a taxa de prisão em todas as paradas no DataFrame, calculando a média da coluna `is_arrested` ;

   (b) Em seguida, você calculará a taxa de prisão por hora usando o atributo de hora do índice. A hora varia de 0 a 23, considerando que *0* é meia noite e *12h* é meio dia. Para isso, agrupe (usando `groupby`) pelo atributo de hora do índice do DataFrame, calculando a média dos valores agrupados. No final crie uma nova variável `hourly_arrest_rate` com os valores encontrados da taxa de prisão por hora;

   (c) Agora crie um gráfico de linha mostrando a variável `hourly_arrest_rate`, colocando o rótulo *Horas* no eixo-x, e *Taxa de Prisões*, no eixo-y, e o título de *Taxa de Prisões por Hora do Dia*;

   (d) Em uma pequena parte das paradas de trânsito, drogas são encontradas no veículo durante uma busca. Agora, você avaliará se essas interrupções relacionadas à drogas estão se tornando mais comuns com o tempo. A coluna booleana `drug_related_stop` indica se drogas foram encontradas durante uma determinada parada. Você calculará a taxa anual de drogas reamostrando essa coluna e, em seguida, usará um gráfico de linha para visualizar como a taxa mudou ao longo do tempo;

   (e) Ainda falando sobre drogas, consideremos a hipótese de que, o aumento ou a diminuição das apreensões de drogas estão associadas ao aumento ou diminuição das abordagens políciais, ou seja, mais abordagens, geram mais apreensões e menos abordagens, menos apreensões de drogas. Podemos testar essa hipótese calculando a taxa de abordagens anual e, em seguida, comparando-a com a taxa anual de medicamentos. Se a hipótese for verdadeira, você verá que ambas as taxas aumentam com o tempo. Para isso, calcule a taxa de pesquisa anual reamostrando a coluna `search_conducted` e salve o resultado como `Annual_search_rate`. Concatene `Annual_drug_rate` e `Annual_search_rate` ao longo do eixo das colunas e gere gráficos de linha para os dados desse resultado da concatenação;

   (f) O estado de Rhode Island está dividido em seis distritos policiais, também conhecidos como zonas. *Como as zonas se comparam em termos de quais infrações são detectadas pela polícia?* Para isso, crie uma distribuição conjunta entre as variáveis `district` e `violation`, usando uma tabela de contingência. Depois, selecione as linhas das zonas `'Zona K1'` a `'Zona K3'`, gere um gráfico de barras agrupadas que ilustre os resultados obtido na tabela, e responda a questão colocada.

4. **Analisando o efeito do clima no policiamento**

   Nesta última parte, você usará um segundo conjunto de dados, `weather.csv`, para explorar o impacto das condições meteorológicas no comportamento da polícia durante as paradas de trânsito. Você vai praticar mesclar e remodelar conjuntos de dados, avaliando se uma fonte de dados é confiável, trabalhando com dados categóricos e outras habilidades avançadas.

   (a) Comece explorando as temperaturas apresentadas no conjunto de dados: carregue o conjunto, selecione as variáveis relativas à temperatura (`TMIN`, `TAVG`, `TMAX`), imprima as principais medidas resumo usando o comando `describe` e plote os três boxplots dessas variáveis em um mesmo gráfico. O que você poderia comentar sobre as temperaturas, com base nos resultados obtidos? PS.: ao decidir se os valores parecem razoáveis, lembre-se de que a temperatura é medida em graus Fahrenheit, não Celsius!

   (b) Para a variável `TDIFF`, que representa a diferença entre as temperaturas, apresente as medidas resumo e plote um histograma para essa variável. O que pode dizer sobre a distribuição de dados?

   (c) Você agora preparará os DataFrames das abordagens de trânsito e de classificação do clima (o dessa sessão) para que estejam prontos para serem mesclados. No DataFrame sobre abordagens no trânsito, você transformará o índice `stop_datetime` para uma coluna (`reset_index`), pois o índice será perdido durante a mesclagem. Com o DataFrame meteorológico, selecione as colunas `DATE` e `rating` e coloque em um novo dataframe;

   (d) Agora, mescle os dataframes gerados em um novo dataframe, unidos usando a coluna *stop_date* de do dataframe policial e a coluna `DATE` do novo dataframe gerado a partir dos dados meteorológicos. Assim que a mesclagem for concluída, defina `stop_datetime` novamente como o índice;

   (e) A partir desse novo dataframe criado, levante duas questões e as responda usando qualquer técnica que ache necessária.