Estatística Aplicada à Computação/Telemática

Projeto II

O projeto da segunda unidade da disciplina será composto por duas partes. Na primeira, vocês irão trabalhar com um projeto guiado, em que cada atividade, do pré-processamento dos dados à será indicada *a priori*, e você deverá, unicamente, executar. 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. Na segunda, a parte guiada será só referente ao pré-processamento dos dados. Uma vez que todas as ações referentes a essa etapa estejam concluídas, vocês deverao responder a um conjunto de perguntas e formular outras, usando os métodos de sua escolha, e, obviamente, justificando a escolha de cada método para responder as questões. Essa abordagem serve justamente para construir a independência e a criatividade de vocês em explorar os conjuntos de dados a disposição. Cada uma das partes será detalhada a seguir.

Parte I: Análise de Atividades Policiais

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.

Esse projeto, como comentei acima, é guiado, e será executado na plataforma Datacamp, no curso *Analyzing Police Activity with pandas*, que tem quatro partes (capítulos):

- **Preparing the data for analysis:** em que 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;
- **Exploring the relationship between gender and policing:** em que você explorará a questão do impacto do gênero de um motorista no comportamento policial durante a abordagem, usando, para isso, ferramentas de filtragem de dados, agrupamento, manipulação de strings e análise de distribuições de frequências;
- **Visual exploratory data analysis:** em que você responderá questões sobre horário de maior probabilidade de ser preso e se há aumento nas blitzes relacionadas a busca por drogas na região a partir do ferramental de geração e interpretação de gráficos;

- **Analyzing the effect of weather on policing:** em que você usará um segundo conjunto de dados para explorar o impacto das condições climáticas no comportamento policial durante as abordagens em blitzes, trabalhando com fusão de dados dos dois conjuntos, manipulação de dados categóricos, e outras habilidades mais avançadas.

Parte II: Análise de Emais Pessoais

Nesse projeto, vocês irão analisar dados de seus próprios emails pessoais, gerados a partir de uma ferramenta de *backup* do Google. Toda a etapa de pré-processamento *sugerida* (você é livro para realizar outras, caso ache necessário) está disponibilizada no notebook `PreProcessam_emails.ipynb`, e que foi explicado na aula gravada sobre o tema, ambos disponibilizados para vocês na plataforma Moodle.

Após o processamento inicial, vocês terão de responder as seguintes perguntas:

1. Quantos emails foram mandados por semana, por mês e por ano, considerando a janela de tempo dos dados baixados?

2. Há uma variação significativa na quantidade de emails enviados por cada período considerado na questão anterior?

3. Existe algum período, dos considerados nas duas questões anteriores, em que o número de emails enviados possa ser considerado um *outlier*? Você enxerga alguma justificativa para esse período ter esse *outlier*, caso haja?

4. Qual é o número médio de e-mails por hora? Essa média varia ao longo da semana?

5. Com quem me comunico com mais frequência, por meio de emails? Considere tanto por envio, quanto por recepção de emails;

6. Quais os temas mais comumente tratados nos emails considerados? (para esse caso, considere usar uma nuvem de palavras como gráfico; para saber mais, veja esse tutorial da Sigmoidal, um tutorial do Datacamp e o Python Graph Gallery);

7. Crie e responda mais duas perguntas relacionadas aos dados obtidos. Seja criativo!

PRAZO DE ENTREGA DOS DOIS PROJETOS

turma A – 05 de maio de 2020

turma B – 03 de maio de 2020

OBSERVAÇÕES

- o trabalho é individual;
- somente serão considerados os resultados apresentados em notebooks jupyter;
- o projeto é o principal componente da nota (vale 60%), portanto, deixar de fazê-lo implica em prejuízo na avaliação;
- obviamente, é mandatório que todas os reaultados obtidos sejam discutidos, com explicações claras e diretas sobre como foi feito cada passo constante em seu trabalho. Projetos que contenham apenas o código serão considerados incompletos, assim como projetos com comentários iguais, gramatica ou semanticalmente, serão considerados PLÁGIO. E plágio é crime!