Relatório Final - ATP de Preparação e Análise Exploratória d

Curso de Tecnologia em Big Data e Inteligência Analítica

Pontifícia Universidade Católica do Paraná (PUCPR)

`Coloque seu nome aqui - e seu email aqui`

Introdução

Nesta atividade, vamos trabalhar com um conjunto de dados da empresa Enron. A Enron foi um Estados Unidos. No seu auge, chegou a valer 65 bilhões de dólares e ela faliu em 24 dias. A Enr fusão da Houston Natural Gas e da InterNorth. Anos depois, quando Jeffrey Skilling tomou pos que criaram entidades fictícias, relatórios financeiros fracos e contabilidades erradas, que soma fracassados dos investidores. O esquema foi tão complexo que nem mesmo auditorias profiss problemas, e a empresa foi investigada pelo governo norte-americano, incluindo a sua suprema investigação incluía aproximadamente 500 mil e-mails trocados por empregados da Enron. Est regulatória de energia dos Estados Unidos durante sua investigação. A base de dados que vam destes e-mails, mas também de salário e ações da bolsa de valores dos envolvidos. Além disso interesse (Person of Interest, ou POI), que participaram da fraude; e funcionários que não partic

Dicas

Nesta análise e preparação de dados, nós precisamos ir além de explorações simples. Garanta acompanhada de uma análise crítica.

Importando as bibliotecas

Na célula abaixo, as principais bibliotecas para análise de dados são importadas. Sinta-se a vo com sua demanda e/ou preferência, contudo, garanta que todos os comandos de `import` seja

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline

# carregue os módulos de sua preferência aqui
```

Carregamento de dados

Na célula abaixo é realizado o carregamento da base de dados. Você **não** deve alterar o código

```
df = pd.read_csv('enron.csv')
```

Verificando se os dados carregaram corretamente:

```
df.head(5)
```

Análise descritiva de dados

Nesta etapa, você conduzirá uma análise descritiva da base de dados da Enron.

Para a base de dados como um todo, você deve reportar:

- O número de indivíduos (instâncias) na tabela;
- O número de variáveis descritivas (colunas ou atributos) destes indivíduos;
- O número de pessoas de interesse, isto é, fraudadores (POIs) e não-POIs;

Para cada uma das variáveis numéricas, você deve apresentar:

- Média
- Mediana
- Variância
- Desvio padrão
- Quartis

E para cada uma das variáveis categóricas:

- A moda
- Os valores únicos de cada variável

Além de apresentar estatísticas sobre cada variável da base de dados, a análise a ser conduzid
forma, você é convidado a extrair *insights* a partir destas estatísticas, verificando se os valores
realidade. **Lembre-se: todas estas estatísticas devem ser calculadas e apresentadas, de forma
ATP.**

```
# conduza sua análise descritiva aqui
# use quantas células de código e texto forem necessárias para atender os
# requisitos apresentados acima e a rubrica do projeto
```

Análise univariada de dados

Nesta seção, você conduzirá uma análise univariada de dados. Esta análise deve contemplar **to**
exceto o **nome do indivíduo**. O objetivo desta etapa é identificar o comportamento das variáveis
extraindo insights sobre cada variável individualmente. Desta forma, para cada variável, trabalh

- Crie um cabeçalho no relatório com o nome da variável;
- Crie visualizações univariadas (histogramas, box-plots, etc), de acordo com o tipo de dad
- Apresente as principais conclusões que podemos obter a partir destas visualizações.

```
# conduza sua análise univariada de dados aqui
# use quantas células de código e texto forem necessárias para atender os
# requisitos apresentados acima e a rubrica do projeto
```

Análise multivariada de dados

Nesta seção, você deverá conduzir uma análise de dados multivariada. O objetivo desta etapa é
variáveis da base de dados da Enron. Apesar de mais flexível que a etapa anterior, esta é uma d
que esta base de dados possui inúmeros relacionamentos interessantes a serem descobertos
visualizações, contudo, sempre no seguinte formato:

- Apresentar uma pergunta/hipótese sobre os dados,
- Criar uma visualização que responda esta pergunta ou confirme a hipótese, e
- Análise dos dados a partir desta visualização, respondendo textualmente a pergunta/hipó
  apresentando ao leitor os principais insights obtidos a partir da visualização.

Esta etapa do projeto requer que ao menos 20 visualizações bivariadas sejam criadas. Sinta-se
plots, violin plots, ou demais gráficos discutidos durante o curso. Note, contudo, que a utilizaçã
tipo de cada gráfico deve ser aderente com os tipos de dados sendo apresentados.

```
# conduza sua análise multivariada aqui
# use quantas células de código e texto forem necessárias para atender os
# requisitos apresentados acima e a rubrica do projeto
```

Visualizações efetivas

Nesta seção, você deve **escolher** e **melhorar** 5 visualizações criadas anteriormente. A chave aq
objetivo de apresentá-las a uma audiência que não conheça a base de dados da Enron e/ou nã
**garanta que o tamanho, cores, texturas e outras componentes visuais sejam bem escolhidas e
informação seja passada de forma clara e correta para a audiência.**

Para cada um destas visualizações, garanta que as seguintes etapas foram seguidas:

1. Criação de visualização: Criar novamente a visualização, garantindo que ela possui título
2. Cores: Garantir que o uso de cores é correto, de acordo com o objetivo da visualização e t
daltônicos
3. Cores e tamanhos: Garantir que texturas e tamanhos são utilizados de forma correta
4. Chart junk: Garantir que a visualização possui um baixo fator de "chart junk"
5. Avaliação por pares: você deverá angariar feedback de três pessoas sobre sua visualizaçã
pessoa, assim como deve discutir, de forma textual, se este feedback acarretou em algum
evidências das mudanças realizadas, isto é, a visualização **antes** e **depois** do *feedback*.
6. Descrição: cada visualização deve ser acompanhada de uma descrição que inclua as prin
extraídos a partir dela.

```
# crie as visualizações finais aqui
# novamente, use quantas células forem necessárias
```

Conclusão

Reflexão

Nesta seção você deve apresentar uma reflexão sobre sua atuação neste projeto e sobre os co
apresentar (1) o que você fez bem e (2) o que você poderia ter feito diferente. Esta reflexão dev
espaços).

Escreva sua resposta aqui.

Referências

Adicione na célula abaixo todas as referências utilizadas durante a preparação deste relatório.

```
Adicione as referências aqui.
```

Trabalhos futuros

Na célula abaixo, apresente ao menos 3 (três) idéias diferentes que você gostaria de atuar em u dados. Elas podem incluir, por exemplo, o uso de aprendizagem de máquina para um objetivo e tivemos tempo de fazer durante a execução deste trabalho. Você deve fornecer **detalhes** sobre as técnicas e/ou metodologia a ser seguida. Esta seção deve possuir ao menos 2500 caractere

```
Escreva sua resposta aqui.
```

Último passsos

1. Salve este relatório como um jupyter notebook em formato `.ipynb`
2. Salve uma cópia deste relatório como um arquivo PDF, isto é, com extensão `.pdf`
3. Compacte ambos em um único arquivo com extensão ZIP no seguinte formato:(`analise-`
4. Envie o seu relatório para avaliação no ambiente virtual de aprendizagem