LISTA 3 – Pré-processamento: tratar os dados

Exercício 1 – Café da manhã

Disponibilizei um arquivo com compras que as pessoas costumam fazer em lanchonetes de café da manhã:

- Date: data da compra;
- Time: hora da compra;
- Transaction: identificador da compra (vários itens podem compor uma compra);
- Item: item que foi comprado (está dentro de uma transaction).

Nosso objetivo é verificar se há correlações interessantes entre os produtos, ou seja, produtos que frequentemente são comprados juntos.

Para tanto, precisaremos seguir os passos abaixo:

1. Importe os dados do arquivo para um data frame. No df, crie um campo novo chamado "Qtd" com a constante 1 (esse campo será usando posteriormente...);
2. Crie um df pivotado a partir do df original, através do método pandas pivot_table();
    a. A constante "Qtd" = 1 que criamos no passo anterior será usada aqui (values = 'Qtd');
    b. Nesse etapa, aproveite o parâmetro fill_value do método para preencher os espaços vazios com zeros;
3. Feito isso, crie uma nova matriz, agora com o valores resultantes da aplicação da equação de correlação de Pearson (correlação deve ser aplicada sobre a matriz pivotada do passo anterior).
4. Plote um gráfico de calor sobre a matriz de correlação.
5. Encontrou alguma correlação interessante entre os produtos? Quais?

Exercício 2 – NBA

Disponibilizei dois arquivos:

- **JOGADORES_NBA**: contém informações sobre os jogadores da NBA. Lá temos informações como peso, altura, ano e local de nascimento de cada atleta.
- **METRICAS_NBA**: contém o histórico de estatísticas dos jogadores/times em que jogaram. Lá temos diversas métricas de desempenho, abaixo destaco algumas delas:
    - Year: ano da estatística
    - Pos: posição do jogador
    - Age: idade
    - Tm: time
    - G: quantidade de jogos
    - MP: quantidade de minutos jogados
    - TS%: taxa de lançamentos com ponto
    - FTr: taxa de lançamento livre
    - PTS: pontos

Fonte dos dados: kaggle.com/drgilermo/nba-players-stats

Observe que há algumas dificuldades nos dados:

- Algumas métricas passaram a ser coletadas ao longo da série, portanto, muitas métricas estão nulas nos anos iniciais.
- Há valores ausentes mesmo em métricas que aparentemente estão preenchidas corretamente
- Há valores extremos também, poucos, mas eles estão lá.

Antes de aplicar as transformações, experimente calcular as médias de idade e peso, por exemplo. Tome nota dos valores que resultam.

Pré-processamento dos dados

1. Importe os arquivos de dados;
2. Junção: faça o merge dos arquivos (métricas/jogadores), use como chave o nome do jogador. Nesse contexto, é preciso incluir as informações de jogadores no layout das métricas (que possui mais registros);
3. Valores extremos:
    a. Há valores extremos nos campos: height, weight, TS% e FTr. Identifique-os e substitua-os pela média dos respectivos campos da série. Não esqueça que a média está distorcida pelos valores extremos, portanto, é preciso usar a média dos não extremos.
4. Valores ausentes:
    a. Inclui propositalmente valores ausentes nos campos, height, weight, TS% e FTr. Para esses, substitua a ausência pela média dos respectivos campos da série. Como os outliers já foram tratados anteriormente, esse passo será mais fácil;
    b. Há valores ausentes também nos campos birth_city, birth_state e collage. Trate-os incluindo uma categoria padrão, como por exemplo, "Não identificado".
5. Discretização:
    a. Discretize o campo Age em três categorias: Júnior, médio e sênior. Para isso, avalie a distribuição e aplique a abordagem de divisão por largura igual, usando intervalos de percentis iguais: (0.333).

Análise de dados:

1. Qual a média de idade/peso dos jogadores?
2. Qual a média de idade/peso dos jogadores na década de 50? Compare com a década de 90.
3. Qual o jogador que mais marcou pontos na série histórica (desconsiderando ano/time)? Achou o resultado? Se sim, joga o nome no Google 😉...
4. Qual foi o jogador que mais marcou pontos em uma única temporada? Em que ano isso aconteceu?
5. Qual jogador que jogou por mais tempo na NBA?
6. Qual jogador permaneceu mais tempo em um mesmo time? Qual é este time? Por quanto tempo ele ficou neste time?
7. Quais são os cinco atletas mais altos e mais baixos que já jogaram na NBA?
8. Quais são os cinco atletas mais pesados e mais leves que já jogaram na NBA?

Exercício 3 – Ramen Ratings

Disponibilizei um arquivo chamado RAMEN_RATINGS.xlsx (fonte: https://www.kaggle.com/residentmario/ramen-ratings) . Trata-se de um conjunto de avaliações sobre Ramen, um prato típico do Japão.

Lista de Exercícios 3

Análise de dados:

1. Qual o país tem a melhor avaliação média? E mediana? Os diferentes resultados sugerem erro nos dados (outliers)?
2. A marca "Nissin" é melhor avaliada em qual país? Use a mediana.
3. No geral, qual tipo de embalagem é melhor avaliada? Use a mediana aqui também

Pré-processamento dos dados:

Supondo que queremos prever a variável "Stars" e pretendemos fazer isso treinando um modelo de *machine learning*. Porém, o algoritmo que pretendemos usar aceita apenas variáveis numéricas.

1. Remova as variáveis "Review #" e "Variety".
2. Aplique codificação one-hot sobre as demais variáveis (exceto "Stars").