UNIVERSIDADE FEDERAL DE VIÇOSA
CAMPUS FLORESTAL
CIÊNCIA DA COMPUTAÇÃO

ANA CLÁUDIA, 1802
GUSTAVO GRAF, 1283
KAIO IGOR, 1760

ANÁLISE DE DADOS
RECEITAS DE CERVEJA DO AMIGO CERVEJEIRO

FLORESTAL
2018

Sumário

1 Introdução

Ao utilizar um conjunto de dados reais, há diversos impasses que dificultam sua visualização e extração de conhecimentos, pois, em grande maioria, há ruídos, erros e campos vazios. Este transtorno é notável, principalmente, quando as informações esperadas não são as mesmas descritas nos resultados. Desta forma, é necessário que seja feito uma série de procedimentos que facilite a representação desses dados, deixando-os de fácil visibilidade aos usuários comuns.

Na primeira parte do trabalho, foi o escolhido um conjunto de dados que será utilizado nas etapas seguintes, neste caso, o tema selecionado chama-se: Receitas de cerveja do amigo cervejeiro (em inglês, *Brewer's Friend Beer Recipes*). Na fase seguinte, a preparação dos dados, tem como objetivos : entender os atributos e objetos; tipos de atributos; domínio dos atributos; identificar ruídos ou informações ausentes; criar novos atributos; formatar valores; entre outros.

A terceira etapa, ocorre a análise exploratória e extração de conhecimento, assim, será gerado estatísticas descritivas, gráficos e tabelas para conhecer os dados. Portanto, extrairá correlações entre atributos e objetos. E, por fim, executaremos um algoritmo de aprendizagem de máquina para classificar ou agrupas os dados, para analisar se há algum acontecimento desconhecido, realizando uma análise preditiva.

2 Preparação dos dados

O conjunto de dados *Brewer's Friend Beer Recipes* apresenta 75 mil cervejas produzidas em casa de 176 estilos diferentes. Os registros são feitos por alguns usuários e classificados de acordo com os estilos definidos.

2.1 Atributos

A seguir, está descrito o nome das colunas do conjunto de dados, em sua confecção original, junto a sua descrição, respectivamente (KAGGLE, 2018):

BeerID	ID da cerveja.
Name	Nome da cerveja.
URL	Página que possui a receita.
Style	Tipo de fermentação.
Style ID	ID do tipo de fermentação.
Size(L)	Quantidade fabricada por receita.
OG	Quantidade antes da fermentação.
FG	Quantidade depois da fermentação.
ABV	Álcool por volume.
IBU	Unidade internacional de amargor.
Color	Cor da cerveja.
BoilSize	Fluido no início da fervura.
BoilTime	Tempo de fervimento.
BoilGravity	Quantidade antes da fervura.
Efficiency	Extração de açucares e grãos.
MashThick	Quantidade de água por quilo de grãos.
SugarScale	Quantidade de sólidos dissolvidos.
BrewMethod	Técnicas para sua fabricação.
PitchRate	Levedura adicionada ao fermentador.
PrimaryTemp	Temperatura durante fermentação.
PrimingMethod	Método para adição de açucares.
PrimingAmount	Quantidade de açucar usado.
UserId	ID do usuário.

Com base nessas informações, utilizou-se o Jupyter Notebook para implementação do código em Python no ambiente Anaconda para importação de suas bibliotecas. Nesta fase inicial, foram necessário importar três bibliotecas, sendo:

- **NumPy:** Pacote fundamental para manipulação de estruturas bidimensionais.
- **Pandas:** Fornece ferramentas de análise de dados.
- **Missingno:** Permite alguns módulos de visualização de dados.

- **Matplotlib.pyplot:** Oferece estrutura de plotagem de gráficos semelhante ao MatLab.

Em seguida, o conjunto de dados escolhido foi lido a partir da biblioteca Pandas e atribuido ao *dataframe* nomeado df_beer. Com isto, utilizando o Missingno, criou-se um gráfico que permite uma representação gráfica de valores nulos de cada coluna, como mostra a Figura 1.

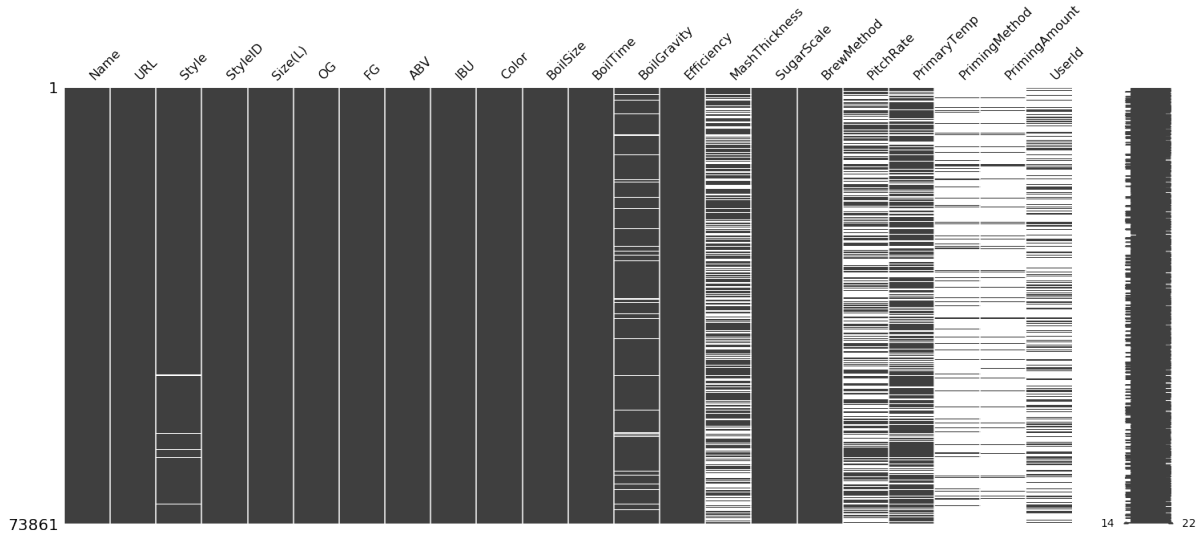

Figura 1 – Representação gráfica de valores nulos de cada coluna do conjunto de dados.

Como podemos ver na figura 1, as colunas *PrimingMethod* e *PrimingAmount* possuem muitos valores nulos, mais de 90% dos casos. Isso se dá pois essas etapas na fabricação da cerveja trazem um diferencial para a mesmo. Portanto, alguns cervejeiros não forneceram esses campos, já que ambos não são obrigatórios. Assim resolvemos apaga-las, utilizando a função *drop*.

2.2 Tipos dos atributos

Para visualizar os tipos de cada atributo do *Data Frame*, utilizamos a função *dtypes* que mostra todos os atributos e seus respectivos tipos como mostra a figura 2.

```
In [7]: df_beer.dtypes

Out[7]: Name             object
        URL              object
        Style            object
        StyleID           int64
        Size(L)         float64
        OG              float64
        FG              float64
        ABV             float64
        IBU             float64
        Color           float64
        BoilSize        float64
        BoilTime          int64
        BoilGravity     float64
        Efficiency      float64
        MashThickness   float64
        SugarScale       object
        BrewMethod       object
        PitchRate       float64
        PrimaryTemp     float64
        PrimingMethod    object
        PrimingAmount    object
        UserId          float64
        dtype: object
```

Figura 2 – Função *dtypes* para mostra dos tipos dos atributos.

2.3 Ruídos ou informações ausentes

Como citado anteriormente as colunas que possuíam mais valores nulos foram retiradas do *Data frame*. Além disso, na Figura 1, observamos que o atributo *Style*, referente ao *StyleID* igual a 111, possuía valor nulo. Dado que nossas análises terão grande dependência do atributo *Style*, optamos por excluí-los utilizando a função *dropna*. Assim como o *Style*, o atributo *Name* é de suma importância. Percebemos que 5 objetos do *Data frame*, apresentavam o valor *"???"* no campo *Name*. Por isso, resolvemos deletá-los.

2.4 Inclusão de Atributos

Analisando os atributos do *Data Frame*, percebemos que a quantidade de líquido antes do processo de fermentação da cerveja (OG) é menor que a quantidade resultante (FG). Pensando nesse aspecto, criamos um novo atributo, chamado *BrewLost*, que resume

a diferença entre *OG* e *FG*. Com isso poderemos relacionar qual *Style* tem uma maior perda nesse processo. A seguir a Figura 3 ilustra esse novo atributo no *Data Frame*.

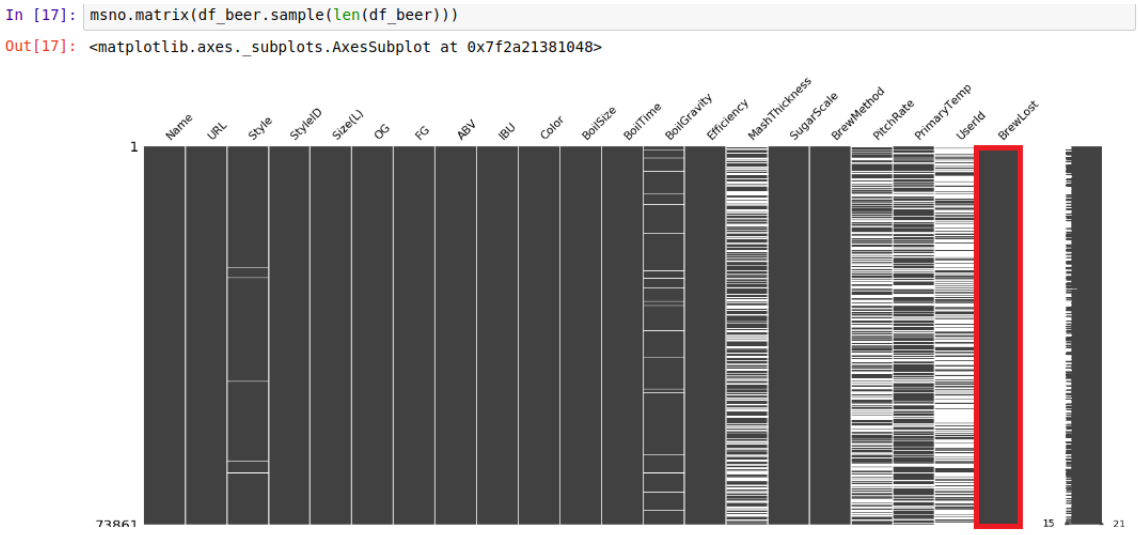

Figura 3 – A partir da *Missingno* podemos visualisar o novo atributo.

Referências

KAGGLE. **Brewer's Friend Beer Recipes**. 2018. Disponível em: <https://www.kaggle.com/jtrofe/beer-recipes>.