UHackthon

Solution

By 摆烂队

- 卢梦雨(Sending emails, Proofread)、
- 曾健洪(**NO contribution**)、
- 黄文超(大佬)

Repo: https://github.com/wenchaoh997/UHackthon2022-Q3

目录

CONTENT

- Introduction
- Exploratory Data Analysis
- Data Pre-processing
- Modeling
- Conclusion

介绍

Introduction

数据分析

Exploratory Data Analysis

数据预处理

Data Pre-processing

建模

Modeling

结论

Conclusion

Introduction 01

Introduction

- Predicting **the sales volume of the products** in the pre-market R&D stage for a period of time after they are launched is a problem with huge business value.
- **Higher prediction accuracy** can Guide the rational allocation of R&D resources in the R&D stage, or guide the rational production and stocking of the supply chain in the early stage of listing to reduce waste.

介绍

Introduction

数据探索

Exploratory Data Analysis

数据预处理

Data Pre-processing

建模

Modeling

结论

Conclusion

Exploratory Data Analysis 02

EDA - Original Data

info

size -> (60x, 8)

sales

size -> (162, 4)

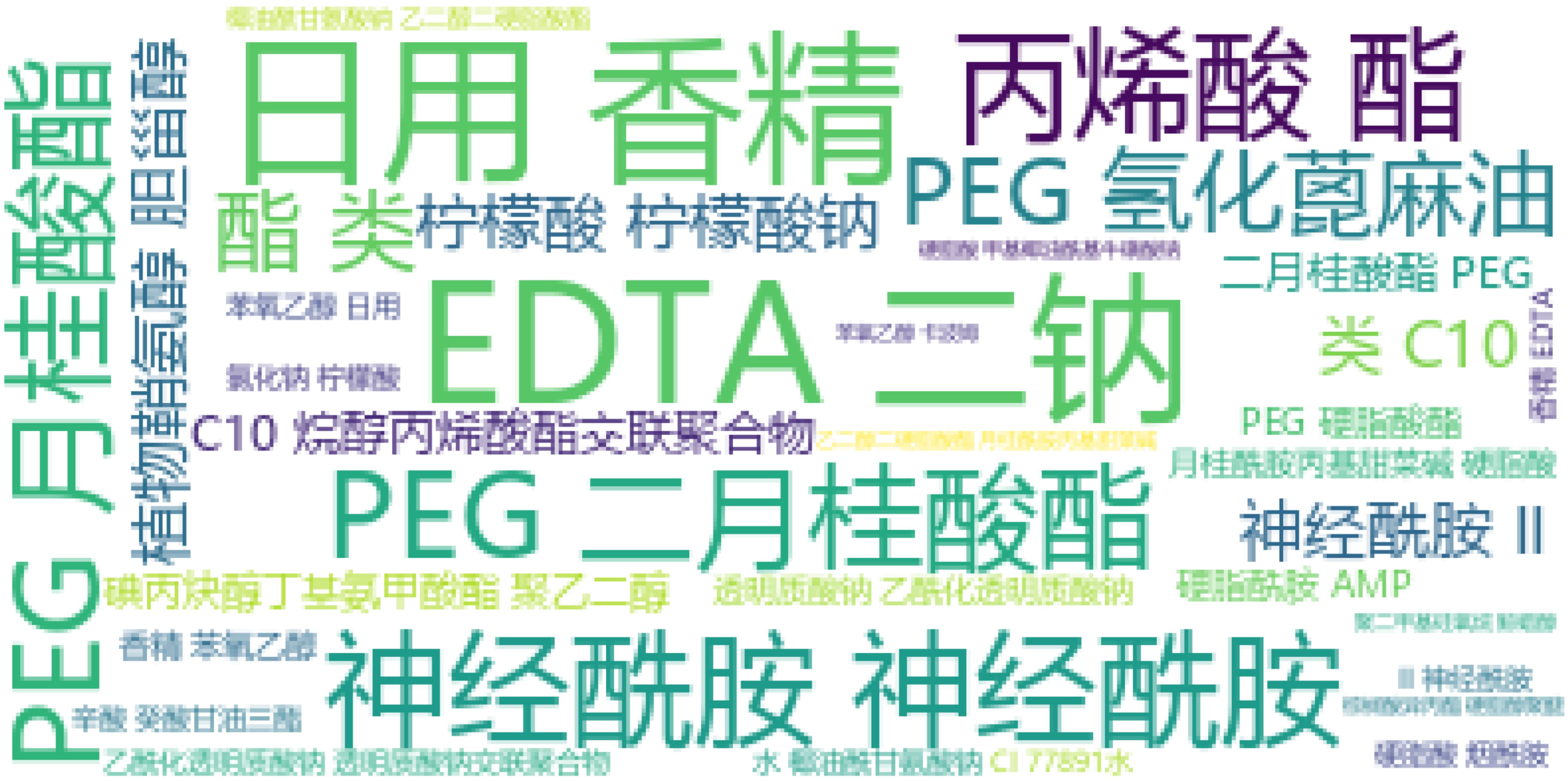
Types

int, string, float

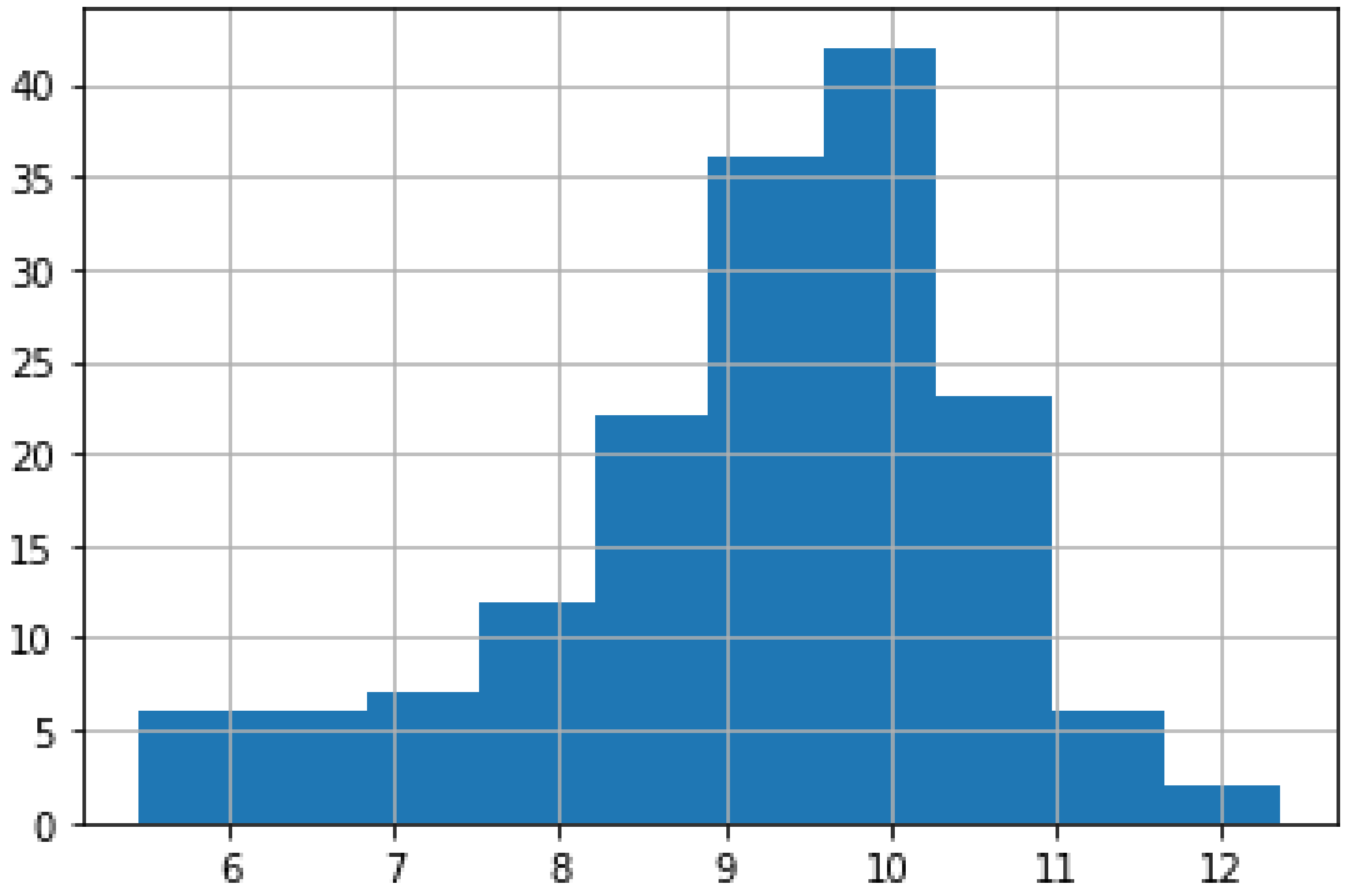
Duplicated IDs、

Insufficient Data

EDA - Word Cloud



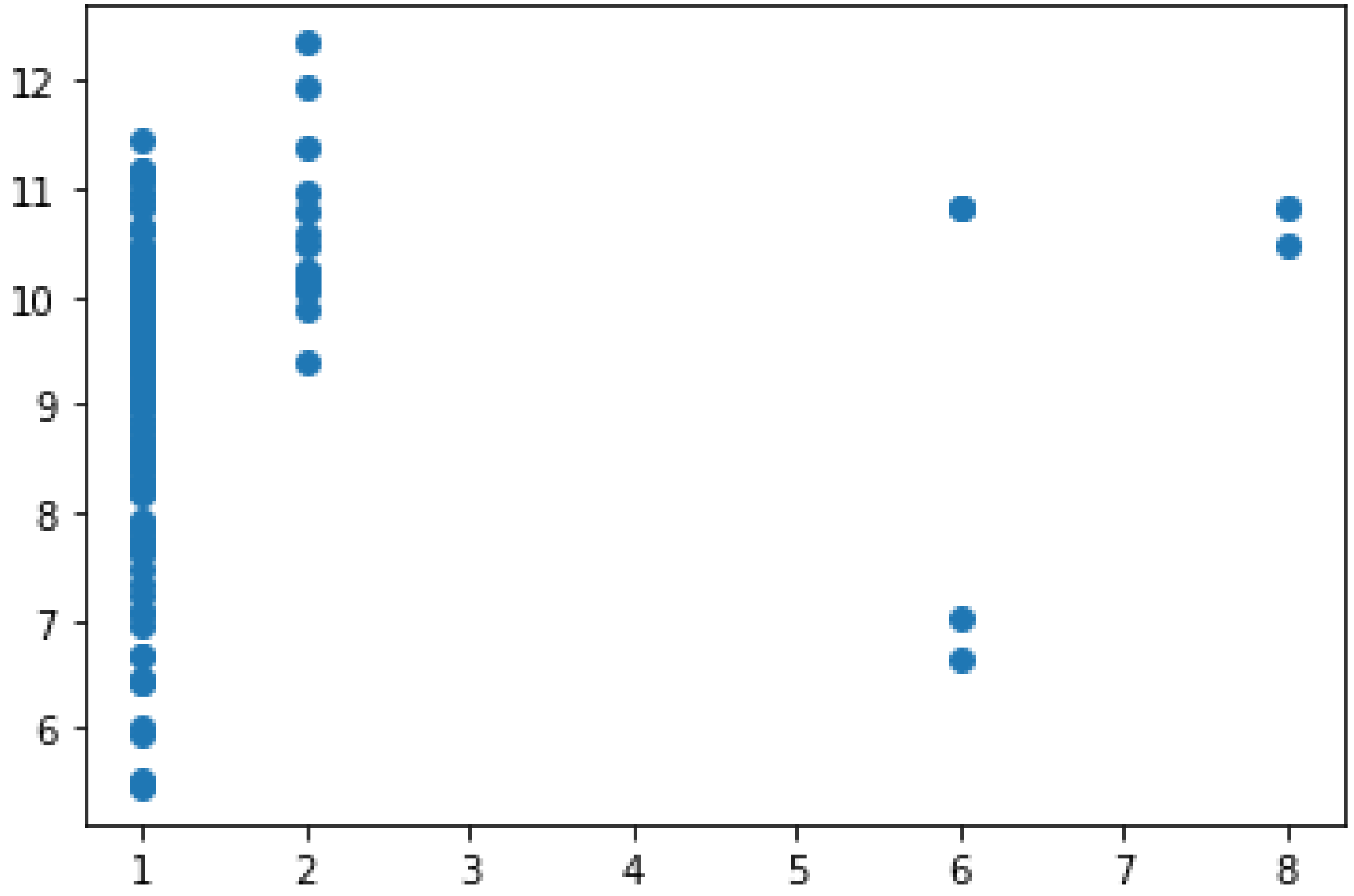
EDA - sales_value



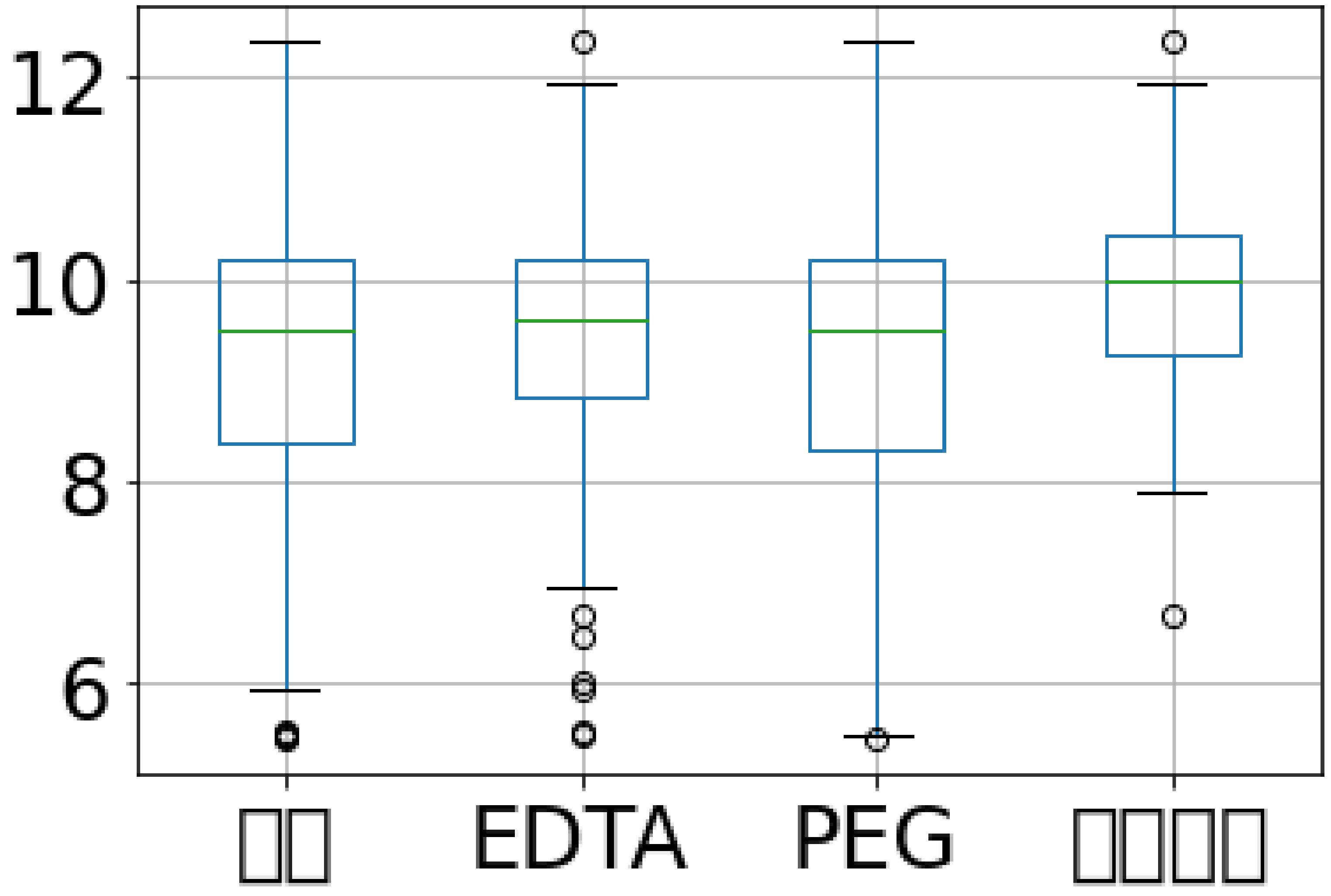
- Long tail

- Most of them are between 9 to 11

EDA - Others



counter VS sales_value



Boxplot on 香料、EDTA、PEG、神经酰胺

介绍

Introduction

数据探索

Exploratory Data Analysis

数据预处理

Data Pre-processing

建模

Modeling

结论

Conclusion

Data Pre-processing 03

Data Pre-processing - info

uid:

duplicated IDs, change into unique

bar_code:

all of them are 690... from China, drop

brand:

one-hot encoding

dTоMx:

distance to some "important" months

launch_date：

split into year, month and day

Ingredient：

union set, by uid

Counting：

how many times or versions are shown

Data Pre-processing - sales

- Merged by uid
- channel: EC -> 0 / DT -> 1
- sales_value: Min-Max Normalization

介绍

Introduction

数据探索

Exploratory Data Analysis

数据预处理

Data Pre-processing

建模

Modeling

结论

Conclusion

Modeling 04

Modeling - GAN

datacebo
CTGAN

- CTGAN, Conditional GAN for generating synthetic tabular data.
- Generating multiple tabular data based on our dataset.
  - with string attributes, overfit
  - w/o string attributes
- The available training set for validation.
- But is it reasonable to use such small dataset in this way?

Reference: CTGAN: https://github.com/sdv-dev/CTGAN

Modeling - LightGBM

- Simplify the model, Avoid overfitting
  - w/o string attributes
  - Less depth and leaves
  - Early stop
- Feature importance
- tweedie, for asymmetric distribution
- Additional noise
  - Inspired by DAE.
  - In our experiment, it can reduce the error.

Reference: LightGBM: https://github.com/Microsoft/LightGBM

Modeling - AutoGluon

- AutoML tech
- Additional noise

Reference: AutoGluon: https://auto.gluon.ai/stable/index.html

介绍

Introduction

数据探索

Exploratory Data Analysis

数据预处理

Data Pre-processing

建模

Modeling

结论

Conclusion

Conclusion 05

Conclusion - Discussion

Some interesting strategies i have not tried..

- Recursive training. Noise, load and train.
- DAE
- But I still have not idea on strings..

Conclusion - Discussion

- We used CTGAN to generate massive by the training set.
- The training set became validation set in our experiments.
- Tweddie loss function for regression issue.
- Simplify the model and avoid overfitting.
- Try LightGBM and AutoGluon.
- Submit our prediction.

Rules Changed? Whats wrong?

- **Deadline?**
- **Presentation rounds?**
- **Have not received the email?**
- **Scoring?**
- **…**

Thanks. And QA Time.

汇报人：张小可

卢梦雨(Sending emails)、曾健洪(**NO contribution**)、黄文超(大佬)