Machine Learning for Economists

Class 4: **Real** Financial Data

葛雷

中国人民大学经济院

2025 年 3 月 13 日

Real Financial Data

CSMAR & WRDS

Kaggle

Selenium

Homework3

Real Financial Data

CSMAR & WRDS

Kaggle

Selenium

Homework3

Real Financial Data

- What is the most important elements for Machine Learning? **Data**
- What makes the ML in finance unique? (we financial data)
- Why real data?

First look a the fake data

- sklearn.datasets is a good source for TOY data
- Good source for practice
- Only issue is that fake data is fake
- Lets check out why (Please follow to blank Ipynb)

CSMAR

CSMAR, short for China Stock Market & Accounting Research Database, is a comprehensive research-oriented database focusing on China Finance and Economy. CSMAR was developed by Shenzhen CSMAR Data Technology Co., Ltd based on academic research needs, meeting with the international professional standards while adapting to China’ s features.

CSMAR

- professional level financial data for stock & company study
- used by both financial companies and financial researchers

CSMAR: easy to use

- Easy to use especially for Python users
- We can use both UI and API (what is UA and API?)
- its check it with me step by step and login from lib

WRDS

USA counterpart of CSMAR[1]

[1] CSMAR followed WRDS's business model

Real Financial Data

CSMAR & WRDS

Kaggle

Selenium

Homework3

Kaggle

- Kaggle, a subsidiary of Google LLC
- Heavely platform for Quant Research (us)
- Codes, data, competition and more
- Let check it out! (Kaggle)

Kaggle

- Kaggle is most important data source for now
- You can search and find your interested research topics
- Let check it out! (Kaggle)

Real Financial Data

CSMAR & WRDS

Kaggle

Selenium

Homework3

Data from the internet

1. Internet has valuable data for the financial predictions
2. Internet data low quality? No
3. Selenium is a powerful and popular tool

But how to use?

- I will guide you to study this package
- but next time you should know how to learn any package by yourself

But how to use?

- Template + Documentation + CHATGPT + BING
- Template (from search bing and from CSDN, StackOverFlow, CHATGPT)
- Something unknown → search Bing + Documentation + ChatGPT

Real Financial Data

CSMAR & WRDS

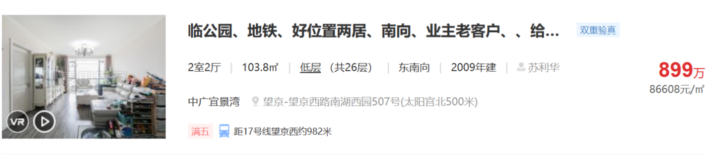
Kaggle

Selenium

Homework3

Homework3-1: Data Mining

- Housing Price Data from https://esf.fang.com/
- Housing Rent Data from https://zu.fang.com/
- Data needed: listed below



临公园、地铁、好位置两居、南向、业主老客户、、给... 双重验真

2室2厅 | 103.8㎡ | 低层 （共26层） | 东南向 | 2009年建 | 苏利华

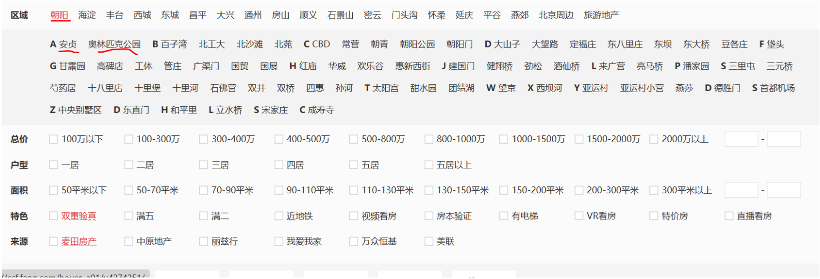
中广宜景湾 望京-望京西路南湖西园507号(太阳宫北500米)

满五 距17号线望京西约982米

899万

86608元/㎡

Homework3-2: Data Mining



区域 朝阳 海淀 丰台 西城 东城 昌平 大兴 通州 房山 顺义 石景山 密云 门头沟 怀柔 延庆 平谷 燕郊 北京周边 旅游地产

A 安贞 奥林匹克公园 **B** 百子湾 北工大 北沙滩 北苑 **C** CBD 常营 朝青 朝阳公园 朝阳门 **D** 大山子 大望路 定福庄 东八里庄 东坝 东大桥 豆各庄 **F** 垡头
G 甘露园 高碑店 工体 管庄 广渠门 国贸 国展 **H** 红庙 华威 欢乐谷 惠新西街 **J** 建国门 健翔桥 劲松 酒仙桥 **L** 来广营 亮马桥 **P** 潘家园 **S** 三里屯 三元桥
芍药居 十八里店 十里堡 十里河 石佛营 双井 双桥 四惠 孙河 **T** 太阳宫 甜水园 团结湖 **W** 望京 **X** 西坝河 **Y** 亚运村 亚运村小营 燕莎 **D** 德胜门 **S** 首都机场
Z 中央别墅区 **D** 东直门 **H** 和平里 **L** 立水桥 **S** 宋家庄 **C** 成寿寺

总价 □ 100万以下 □ 100-300万 □ 300-400万 □ 400-500万 □ 500-800万 □ 800-1000万 □ 1000-1500万 □ 1500-2000万 □ 2000万以上 □ - □

户型 □ 一居 □ 二居 □ 三居 □ 四居 □ 五居 □ 五居以上

面积 □ 50平米以下 □ 50-70平米 □ 70-90平米 □ 90-110平米 □ 110-130平米 □ 130-150平米 □ 150-200平米 □ 200-300平米 □ 300平米以上 □ - □

特色 □ 双重验真 □ 满五 □ 满二 □ 近地铁 □ 视频看房 □ 房本验证 □ 有电梯 □ VR看房 □ 特价房 □ 直播看房

来源 □ 麦田房产 □ 中原地产 □ 丽兹行 □ 我爱我家 □ 万众恒基 □ 美联

//esf.fang.com/house-a01/u4274251/

Homework3-2: Data Mining

- Team 1 北京-海淀 I：苏州桥、万柳、北太平庄、世纪城
- Team 2 北京-海淀 II：西三旗、清河、西二旗、上地
- Team 3 河北-廊坊 + 北京-通州：大厂、燕郊、马驹桥、亦庄
- Team 4 北京-昌平: 沙河、霍营、回龙观、天通苑
- Team 5 天津：中新生态城（滨海新区)、武清、劝业场（和平)、八里台（南开)
- Team 6 重庆-渝北：（Please choose blocks with both price and rental data）

- Each person only in charge of **one block** and only get first 20 pages if too many for you

Homework3-3: Data Research

- Collect Data from your teammates and merge the data (please feedback to TA if someone no response, so we can help both team and other student)

- Data description of your data and whether data has outliers

- Then get housing price per m2 and housing rent per m2 ($price/m2$ and $rent/m2$) for each block

- Calculate median price to rent ratio for each block

- Figure A: Bar Plot the median price to rent ratio for each block (The global fair value should around 200)

Homework3-4: Data Science

- Model 1
  $price/m2_i = \beta_0 m2_i + \beta_2 location_i + \beta_2 m2_i \times location_i + \epsilon_i$

- Model2
  $rent/m2_i = \beta_0 m2_i + \beta_2 location_i + \beta_3 m2_i \times location_i + \epsilon_i$

- Use model 1 and model 2 to predict price and rent for the $m2 = 50$, $m2 = 100$

- Figure B and C: Bar Plot the price to rent ratio for each block for the $m2 = 50$, $m2 = 100$

- Submission: only Ipynb codes to your personal folder (NO DATA PLEASE, Git is for codes not for data)