

河南工业大学

课 程 设 计

课程设计名称: 数据分析与可视化综合实践

专 业 班 级: 数据科学与大数据技术 2101 班

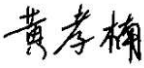

学 生 姓 名: 李俊

学 号: 211210100110

指 导 教 师: 黄孝楠

课程设计时间: 2023.12.18-2024.01.05

## 数据科学与大数据技术专业课程设计任务书

学生姓名	李俊	专业班级	大数据 2101	学 号	211210100110
题 目	地产图谱：多维数据解析中国房市——数据可视化				
课题性质	工程设计		课题来源		自拟课题
指导教师	黄孝楠		同组姓名		韩林欣
主要内容	<p>一、数据 近十年的房地产年度开发企业经营情况和负债率。全国以及各省的商品房年度平均价格，销售额，销售面积，近十年年度城市竣工面积和施工面积。</p> <p>二、基本功能</p> <ol style="list-style-type: none"> <li>1. 数据存储：将 csv 文件通过 Linux 系统传入 HDFS 端存储。</li> <li>2. 数仓分层：在 Hive 中处理数据并进行数据仓库分层。</li> <li>3. 数据迁移：在 MySQL 中建立相同格式的表，使用 DataX 将 Hive 的表同步到 MySQL 中。</li> <li>4. 后端开发：在后端编写 Servlet 建立后端数据访问接口。</li> <li>5. 数据可视化：前端通过 Ajax 请求访问后端接口获取数据并用 ECharts 进行可视化。</li> <li>6. 任务调度：使用 Azkaban 将业务进行调度。</li> <li>7. 数据挖掘：利用数据挖掘技术分析全国房地产风向变化</li> </ol>				
任务要求	<ol style="list-style-type: none"> <li>1. 根据房地产数据进行分析，分析近十年来的全国及各省房地产风向</li> <li>2. 对现有的数据进行清理提取等处理，并将处理后的数据通过 HDFS 建立 Hive 外部表并导入到 MySQL 数据库中，最后采用数据可视化工具对数据进行分析并得出数据分析图。</li> <li>3. 使用 Linux、Servlet、Hadoop、Hive、MySQL、Azkaban 等技术，并运用 ECharts、IntelliJ IDEA、MobaXterm、VMware、DataX 等操作工具。</li> <li>4. 在设计过程中撰写规范的设计报告，在设计完成后成功通过答辩</li> </ol>				
参考文献	<p>[1] 房地产大数据及其信息挖掘体系构建路径研究[J]. 隆林宁. 产业与科技论坛, 2023(04).</p> <p>[2] “互联网+”背景下众筹模式在房地产开发项目中的应用[J]. 郭庆军; 白思俊; 张华波; 朱海阔. 项目管理技术, 2016(01).</p> <p>[3] 用互联网思维和手段创新房地产业发展[J]. 刘志峰. 住宅产业, 2015(11).</p> <p>[4] 房地产行业“互联网+”模式研究与探讨[J]. 夏阳. 中国房地产, 2015(13).</p> <p>[5] 大数据时代: 房地产业的机遇与挑战[J]. 刘昱; 张玉娟. 河南商业高等专科学校学报, 2013.</p>				
审查意见	<p style="text-align: center;">同意。</p> <p style="text-align: center;">指导教师签字: </p> <p style="text-align: center;">教研室主任签字:  2023 年 12 月 4 日</p>				

# 目 录

1	需求分析 .....	1
1.1	需求背景 .....	1
1.2	数据分析需求 .....	1
1.3	技术需求分析 .....	2
2	概要设计 .....	2
2.1	数据准备 .....	2
2.2	数据预处理 .....	2
2.3	数据分层 .....	3
2.4	数据迁移 .....	4
2.5	任务调度 .....	4
2.6	数据挖掘 .....	4
2.7	数据可视化 .....	5
3	开发工具和编程语言 .....	6
3.1	开发工具 .....	6
3.2	编程语言 .....	6
4	详细设计及运行结果 .....	6
4.1	数据来源 .....	6
4.2	数据挖掘 .....	7
4.3	数据可视化 .....	14
5	调试分析 .....	25
6	总结 .....	27
7	参考文献 .....	28

# 1 需求分析

## 1.1 需求背景

随着中国经济的快速发展，房地产行业经历了巨大的变革。如今，这个行业不仅是国民经济的重要支柱，也是投资者、开发商、政府和普通民众关注的焦点。在这个庞大的市场中，数据的作用日益凸显。

然而，目前这些数据大多分散在各个机构、平台和部门中，缺乏一个统一、直观的展现方式。对于决策者来说，没有一个全局的视角，很难准确判断市场趋势和制定策略。对于普通民众来说，房地产市场的信息透明度不足，也增加了他们参与市场的难度。

地产图谱对近十年国家以及各省的房地产数据以及各个主要城市的施工竣工面积进行分析，为决策者提供我国房地产风向变化趋势。

## 1.2 数据分析需求

采集国家以及各个省份十年来的商品房年度平均价格，总均销售面积和总销售额，并采集出国家各个主要城市的近十年年度施工面积和竣工面积，并使用 Hive 对这些数据进行清洗，分层，汇总，最后达到数据挖掘和数据可视化的要求。

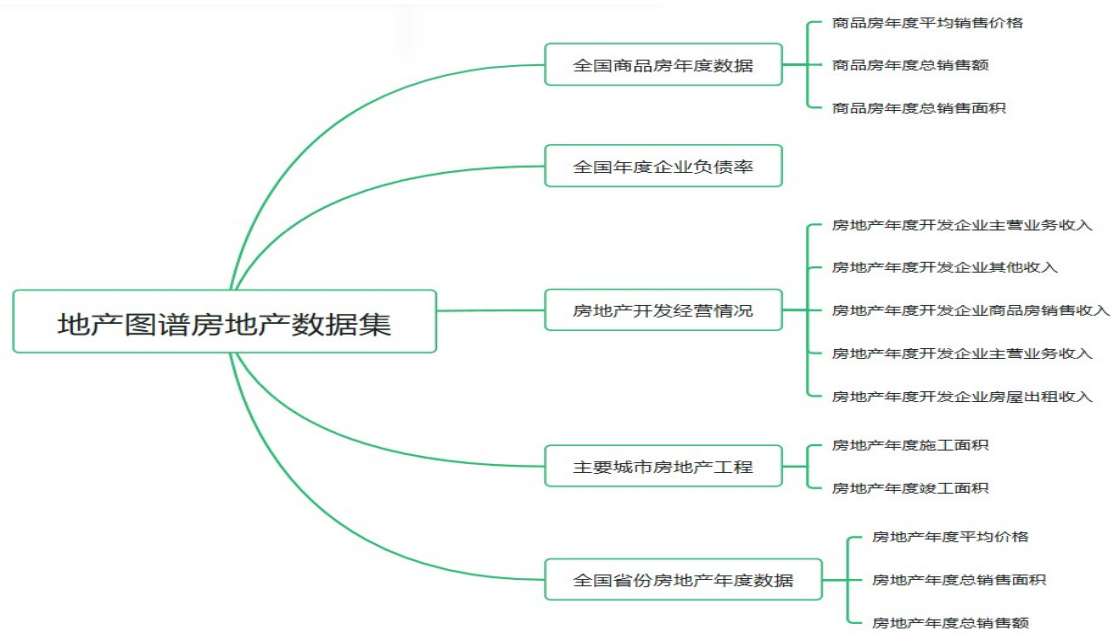


图 1-1 数据分析图

1.3 技术需求分析

- 1. 数据存储：将 csv 文件通过 Linux 系统传入 HDFS 端存储。
- 2. 数仓分层：在 Hive 中处理数据并进行数据仓库分层。
- 3. 数据迁移：在 MySQL 中建立相同格式的表，使用 Datanx 将 Hive 的表同步到 MySQL 中。
- 4. 后端开发：在后端编写 Servlet 建立后端数据访问接口。
- 5. 数据可视化：前端通过 Ajax 请求访问后端接口获取数据并用 ECharts 进行可视化。
- 6. 任务调度：使用 Azkaban 将业务转变为工作流。
- 7. 数据挖掘：利用数据挖掘技术分析全国房地产风向变化。

2 概要设计

2.1 数据准备

在国家统计局采集国家国家房地产年度数据

1. 房地产开发企业经营情况.csv	482	890	Microsoft Excel 逗号...	2023/12/15 10:40:50
2. 房地产开发商资产负债.csv	102	151	Microsoft Excel 逗号...	2023/12/15 10:24:59
3. 商品房年度平均价格V.csv	132	183	Microsoft Excel 逗号...	2023/12/15 10:00:15
4. 商品房年度总销售额V.csv	131	193	Microsoft Excel 逗号...	2023/12/15 9:59:32
5. 商品房年度总销售面积V.csv	135	198	Microsoft Excel 逗号...	2023/12/15 9:58:51
6. 商品房平均价格.csv	1,161	2,062	Microsoft Excel 逗号...	2023/12/15 9:57:28
7. 商品房销售额.csv	1,479	2,759	Microsoft Excel 逗号...	2023/12/15 9:56:48
8. 商品房销售面积.csv	1,512	2,810	Microsoft Excel 逗号...	2023/12/15 10:01:44
9. 主要城市竣工面积.csv	1,516	2,853	Microsoft Excel 逗号...	2023/12/15 9:54:37
10. 主要城市施工面积.csv	1,714	3,179	Microsoft Excel 逗号...	2023/12/15 9:54:22

图 2-1 国家房地产年度数据

2.2 数据预处理

数据预处理使用 Python 的 pandas 库进行数据预处理，并重新导出为 csv 文件。

名称	修改日期	类型	大小
<b>_completed_area_of_major_cities</b>	2023/12/26 9:07	文件夹	
annual_average_price	2023/12/26 8:49	文件夹	
annual_total_sales_area	2023/12/26 8:50	文件夹	
annual_total_sales_revenue	2023/12/26 8:56	文件夹	
assets_liabilities	2023/12/26 8:57	文件夹	
business_operations	2023/12/26 8:59	文件夹	
constructed_area_of_major_cities	2023/12/26 9:09	文件夹	
regional_housing_prices	2023/12/26 9:01	文件夹	
regional_housing_sales_area	2023/12/26 9:05	文件夹	
regional_housing_sales_revenue	2023/12/26 9:04	文件夹	

图 2-2 数据清洗结果

## 2.3 数据分层

数据在 Hive 中分层分为 ODS 原始数据层，DWD 清洗数据层，DWS 整合数据层和 ADS 应用数据层。

ads  
azkaban  
datax  
default  
dim  
dwd  
dws  
movie  
ods

图 2-3 数据分层

并通过以下方式进行维度表的设计。

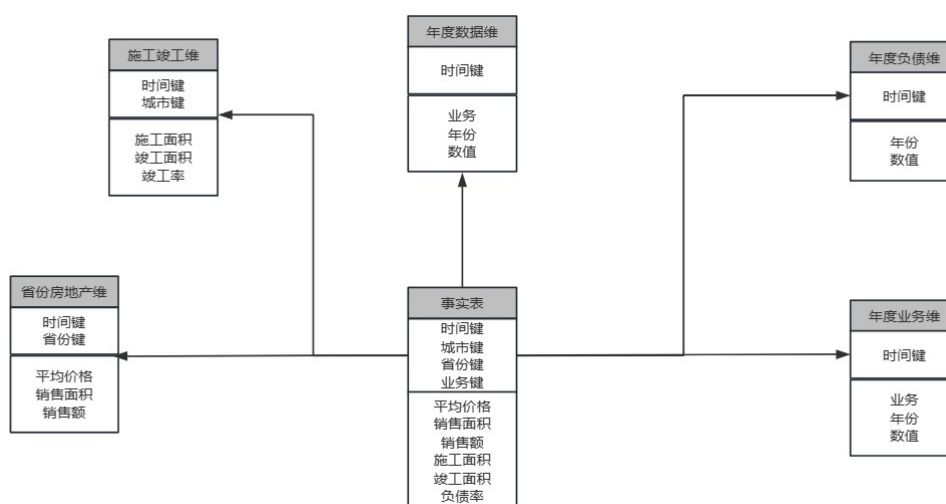


图 2-4 维度建模

## 2.4 数据迁移

数据迁移使用 Datax 完成，将 Hive 中的数据迁移到 MySQL 中。

目 > ETL				
名称	修改日期	类型	大小	
 annual_data.json	2023/12/21 11:37	JSON 源文件	3 KB	
 assets_liabilities.json	2023/12/21 11:24	JSON 源文件	2 KB	
 business_operations.json	2023/12/21 11:19	JSON 源文件	3 KB	
 ETL过程.docx	2023/12/21 12:46	Microsoft Word ...	12 KB	
 house_complete_rate.json	2023/12/21 11:11	JSON 源文件	3 KB	
 reader.json	2023/12/21 10:40	JSON 源文件	2 KB	
 regional_house.json	2023/12/21 11:00	JSON 源文件	3 KB	
 writer.json	2023/12/21 12:47	JSON 源文件	3 KB	

图 2-5 数据迁移文件

## 2.5 任务调度

任务调度使用 Azkaban 完成，将 Hive 中的 SQL 文件转变为 Azkaban 的 job 进行处理。

名称	修改日期	类型	大小
 ads层	2023/12/26 11:31	文件夹	
 dwd层	2023/12/26 9:34	文件夹	
 dws层	2023/12/26 10:04	文件夹	
 ods层	2023/12/25 19:36	文件夹	
 azkaban.project	2023/12/25 16:08	PROJECT 文件	1 KB
 basic.flow	2023/12/26 17:29	FLOW 文件	2 KB
 over.zip	2023/12/26 17:30	ZIP 压缩文件	1 KB

图 2-6 任务调度文件

## 2.6 数据挖掘

本次数据挖掘的目的是建立房价预测模型，数据清洗去除异常值，重复值。特征选择采取 Pearson 相关性系数和方差分析法，模型选择采用简单易懂，可解释性高的多元线性回归，模型训练使用 sklearn 机器学习库，模型预测所需的特征

指标准备采用多项式拟合得到未来指标，然后代入进行预测。模型评估采用多个角度评估，如 R 方，F-statistic（F 统计量）等。

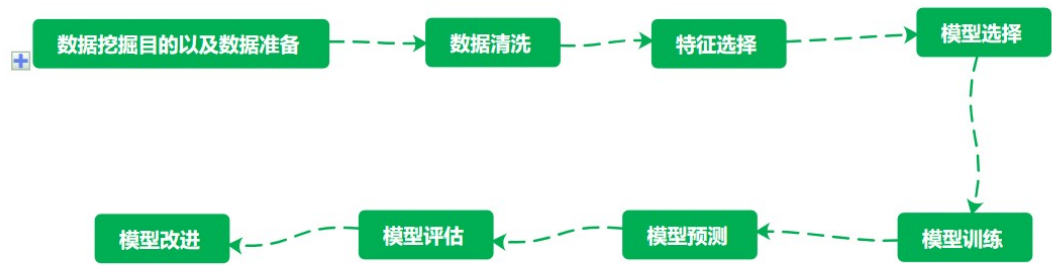


图 2-7 数据挖掘流程图

## 2.7 数据可视化

本次实践中可视化页面并没有像以前一样采用单一的可视化大屏。采用了模块化的设计，增加了交互性和可扩展性，为用户提供更详细的数据信息和好的使用体验。总体布局采用了 3 个主题域。通过使用 ECharts 完成数据可视化，通过 Ajax 获取后端接口的数据。具体可视化页面构造如下图所示。

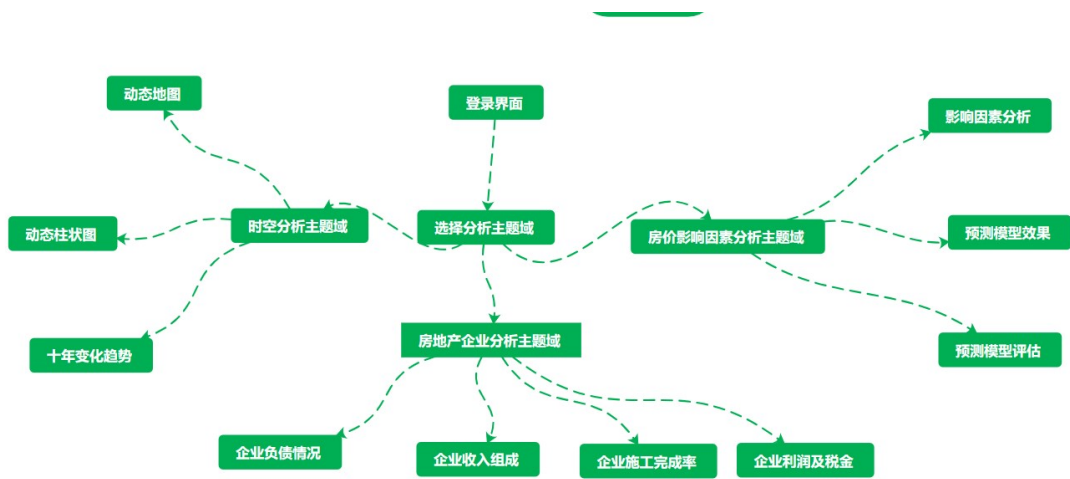


图 2-8 可视化页面模型图



## 3 开发工具和编程语言

### 3.1 开发工具

开发工具：

IDEA	--代码编辑器，
Datax	--数据迁移工具，
Linux	--后端大数据基础操作系统，
Hadoop	--提供 HDFS 和 Hive 运行环境，
Azkaban	--任务调度。

### 3.2 编程语言

编程语言：

Java，  
JavaScript，  
Python。

## 4 详细设计及运行结果

### 4.1 数据来源

房地产数据从国家统计局官网收集得到，在国家统计局官网搜寻年度数据，并以省份和主要城市为过滤条件进行收集。

采集数据包括房地产开发企业经营情况，房地产开发商资产负债，商品房年度平均价格，商品房年度总销售额，商品房年度总销售面积，商品房平均价格，商品房销售额，商品房销售面积，主要城市竣工面积，主要城市施工面积。

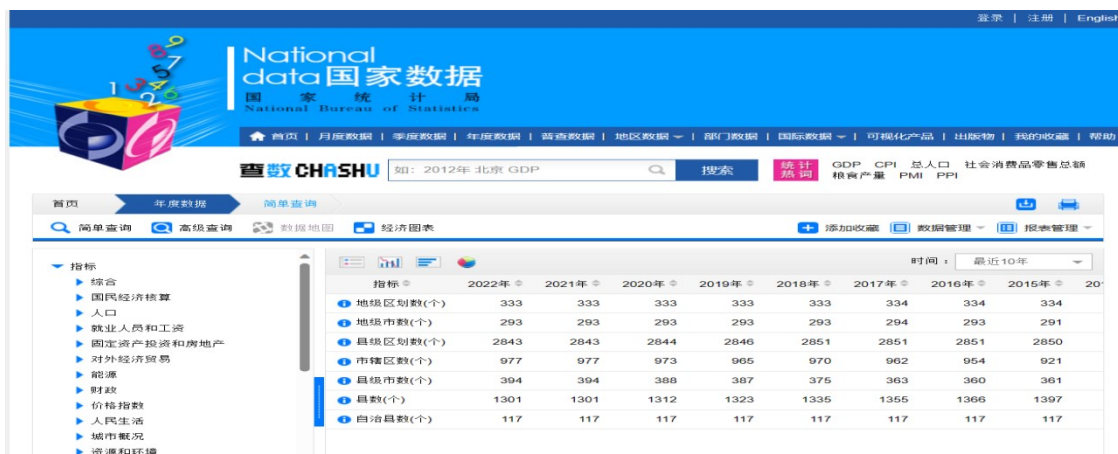


图 4-1 国家统计局官网

## 4.2 数据挖掘

### 4.2.1 数据挖掘目的以及介绍

房价预测有助于各方更好地理解 and 应对房地产市场的变化，从而更有效地进行投资、规划和决策。本次数据挖掘的目的是为了，根据现有的数据指标，建立预测模型，并且为决策者提供相应的参考。

本数据来源于国家统计局。数据选取了国家统计局近 20 年与房地产企业相关的数据。经过个人的筛选，本次数据训练选择的特征指标 11 个，指标如下：

表 4-1 特征选择表

特征	特征
企业平均从业人数(人)	企业个数(个)
企业待开发土地面积(万平方米)	企业计划总投资(亿元)
商品房销售面积(万平方米)	商品房销售额(亿元)
房地产开发企业资产负债率(%)	房地产开发企业资产总计(亿元)
房地产开发企业营业利润(亿元)	企业施工房屋面积(万平方米)
房地产开发企业竣工房屋面积(万平方米)	

选取的这 11 个指标都是和房价相关的因素，当然这些因素只是影响房价的一部分，并不能代表全部。

### 4.2.2 数据清洗

由于数据来源于国家统计局，相比于其他数据，国家统计局的数据完整度很

高，且基本没有脏数据，不过存在一些省份出现数据情况为 0 或者为空的情况，所以对于这种情况，所需要的仅为通过数据清洗，将采集的数据中的空数据用 0 填充，并保存为新的数据文件。

```
import pandas as pd
# 读取 CSV 文件
df = pd.read_csv('商品房销售额.csv')
# 检查每一行除去第一列是否有空值，并用 0 替代空值
df.iloc[:, 1:] = df.iloc[:, 1:].fillna(0)
# 将处理后的数据保存为新的 CSV 文件
df.to_csv('fenxi.csv', index=False)
```

#### 4.2.3 特征选择

但特征选择是一个重要的数据预处理过程，特征选择主要有两个功能。减少特征数量、降维，使模型泛化能力更强，减少过拟合增强对特征和特征值之间的理解。

好的特征选择能够提升模型的性能，更能帮助我们理解数据的特点、底层结构，这对进一步改善模型、算法都有着重要作用。

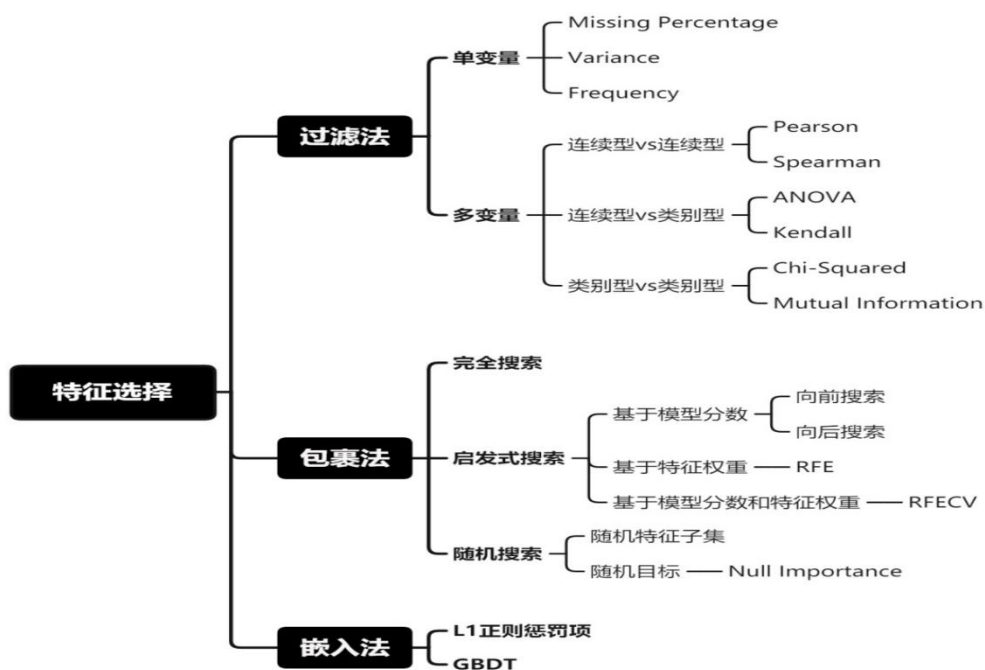


图 4-2 特征选择方法图

## ● 相关性分析

本次特征选择使用了过滤法，数据集中都是连续性变量。首先采用 Pearson 相关系数分析观察特征之间的关系。相关系数热力图如下图所示，0-11 分别代表上表格选取的特征。12 代表商品房房价。

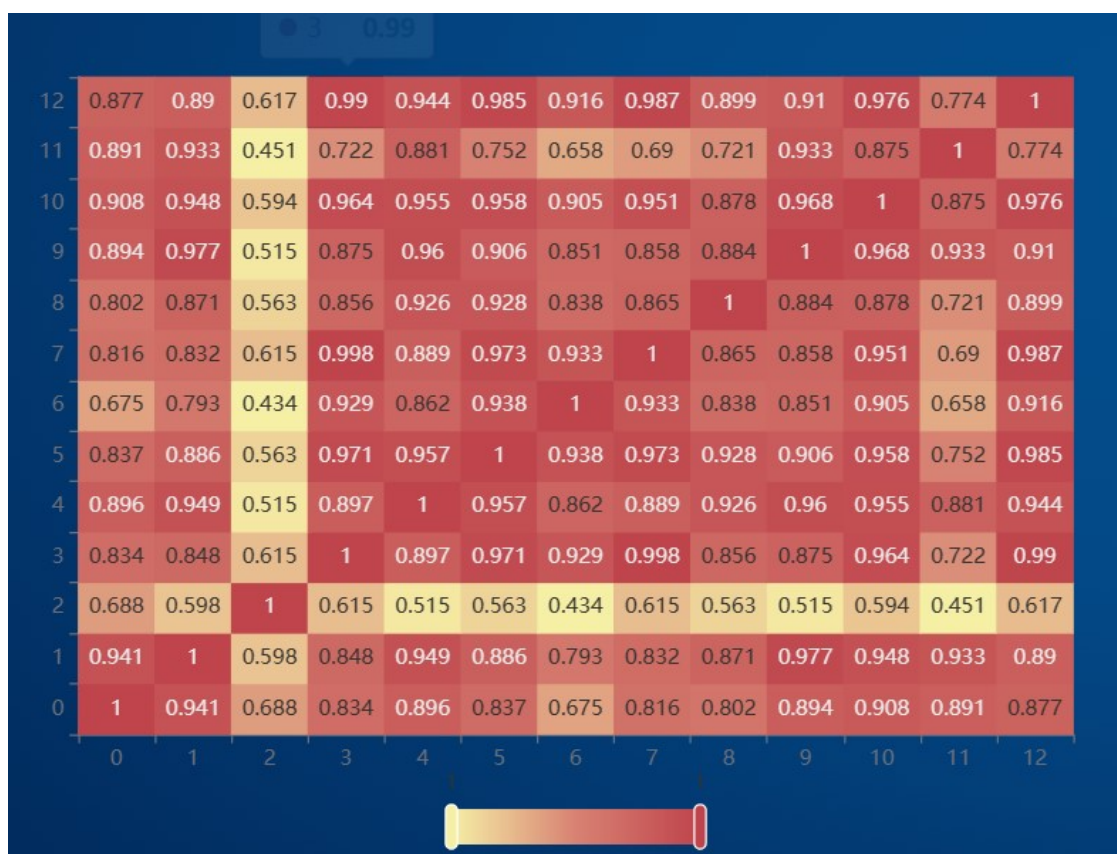


图 4-3 相关系数热力图

通过上图可以发现选取的样本具有多重共线性的问题，直接根据相关性选取特征不合适。因此选择使用方差法辅助特征选择。

## ● 方差分析

方差分析是通过计算一组特征的方差大小选择特征。如果方差太小说明该数据趋于稳定，信息含量较小，对模型的影响作用不大，检验剔除。如果方差较大，说明该数据波动大，含有较多信息量。

```
from sklearn.feature_selection import VarianceThreshold
# 初始化标准化对象
```

```
scaler = StandardScaler()
# 对数据集进行标准化
X_scaled = scaler.fit_transform(X)
# 设置方差阈值
threshold = 1
#初始化 VarianceThreshold 对象
selector = VarianceThreshold(threshold=threshold)
# 应用特征选择器到数据集
X_selected = selector.fit_transform(X_scaled)
# 获取被选取的特征的布尔掩码
selected_features_mask = selector.get_support()
# 使用掩码从原始数据中选择相应的列
selected_features = X.columns[selected_features_mask]
# 打印被选取的特征
print("Selected Features:")
print(selected_features)
```

经过计算选取的了如下的 3 个特征。这三个指标都是和房地产企业相关的指标。房地产开发企业平均从业人数(人)，房地产开发企业计划总投资(亿元)，房地产开发企业资产总计(亿元)。

#### 4.2.4 模型选择

本次数据挖掘选择的是多元线性回归模型，主要是根据以下几个原因选择：

- 线性关系假设:选择线性回归的前提是自变量和因变量之间存在线性关系。通过相关性分析可以发现变量之间的关系是线性的，那么线性回归可能是一个合适的选择。
- 简单性:线性回归是一种相对简单的模型，易于理解和解释。如果问题的复杂性可以通过一个线性模型来解释，而不需要引入更复杂的结构，那么线性回归可能是一个好的选择。
- 数据量:线性回归对于小到中等规模的数据集通常表现得很好，本数据集的数据量比较少,不适合选择太复杂的模型。

- 解释性要求:线性回归通常提供了清晰而直观的解释。系数表示了每个特征对目标变量的影响程度。

#### 4.2.5 模型训练

模型的构建采用 sklearn 机器学习库的线性回归模型，简单易懂，方便直接调用。

```
from sklearn.linear_model import LinearRegression
model = LinearRegression()
model.fit(X_selected, Y)
#打印模型的系数
print(model.coef_, model.intercept_)
```

#### 4.2.6 模型评估

模型的评估使用了 statsmodels 库，该库提供多个指标评价线性回归模型。具体使用代码如下。

```
#对模型进行评估
import statsmodels.api as sm
X2 = sm.add_constant(X_selected)
est = sm.OLS(Y,X2).fit()
print(est.summary())
```

运行后的结果如下图所示。

[ 527.48511223 550.26609592 1517.52601473] 6198.706499999999]

OLS Regression Results

Dep. Variable:	商品房平均销售价格(元/平方米)	R-squared:	0.990
Model:	OLS	Adj. R-squared:	0.988
Method:	Least Squares	F-statistic:	527.8
Date:	Mon, 01 Jan 2024	Prob (F-statistic):	3.33e-16
Time:	09:53:26	Log-Likelihood:	-139.11
No. Observations:	20	AIC:	286.2
Df Residuals:	16	BIC:	290.2
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
const	6198.7065	63.459	97.680	0.000	6064.179	6333.234
x1	527.4851	134.402	3.925	0.001	242.565	812.405
x2	550.2661	1250.929	0.440	0.666	-2101.586	3202.118
x3	1517.5260	1195.667	1.269	0.223	-1017.174	4052.226

Omnibus:	0.002	Durbin-Watson:	1.463
Prob(Omnibus):	0.999	Jarque-Bera (JB):	0.139
Skew:	0.002	Prob(JB):	0.933
Kurtosis:	2.591	Cond. No.	45.5

图 4-4 模型评估图

选取其中一些指标进行说明，Dep. Variable（因变量）：商品房平均销售价格(元/平方米) 是被预测的因变量。R-squared（R 平方）：衡量模型对观测数据的拟合程度，取值范围在 0 到 1 之间。0 表示模型无拟合能力，1 表示完美拟合。在这个例子中，R 平方为 0.9，说明模型很好地解释了因变量的变化。F-statistic（F 统计量）：用于检验模型整体的显著性。在这个例子中，F-statistic 为 527.8，对应的 Prob（F-statistic）为 3.33e-16，非常接近零，表示模型整体上是显著的。const 对应截距，x1、x2、x3 对应相应自变量的系数。’房地产开发企业平均从业人数(人)’，’房地产开发企业计划总投资(亿元)’，’房地产开发企业资产总计(亿元)’，其对应的参数是 527，550，1527。

#### 4.2.7 模型预测

进行房价模型预测，首先需要获取 3 个特征未来的指标，这三个特征采取多项式拟合的方法进行预测，多项式拟合模型是一种常用的机器学习方法，用于拟合数据集中的非线性关系。它通过在输入变量上构建多项式函数，并使用最小二乘法来拟合数据。这种模型的优点在于简单易用，并且可以适应各种数据集。尽量减少误差。

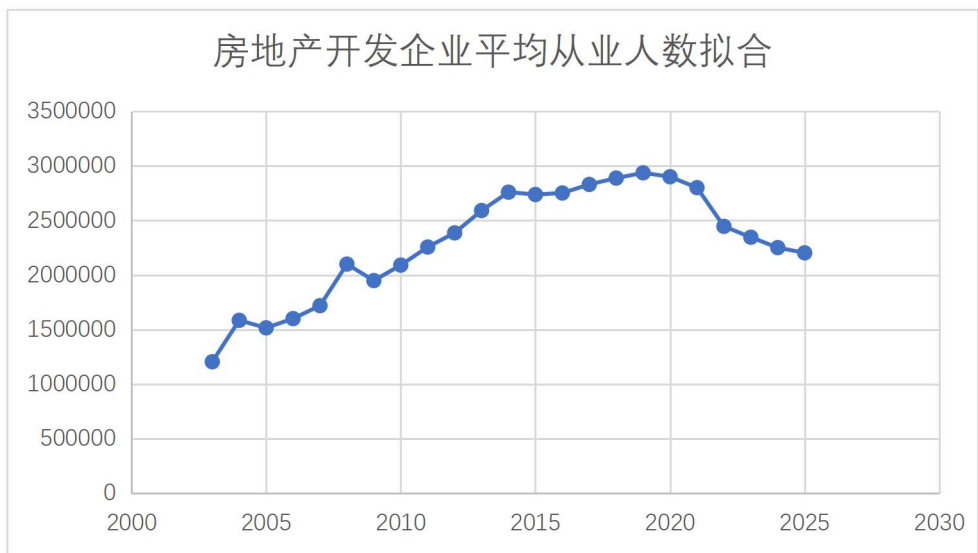


图 4-5 房地产开发企业从业人数拟合图

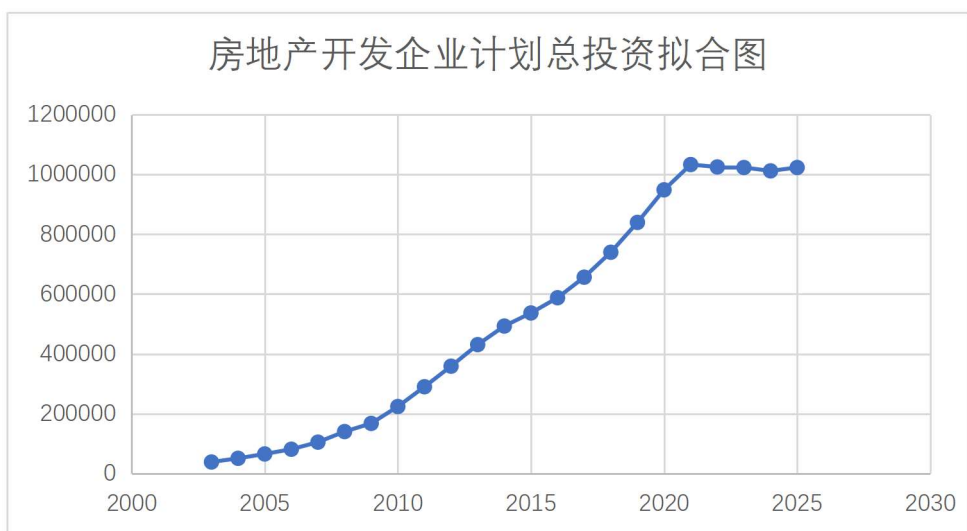


图 4-6 房地产开发企业计划投资拟合图

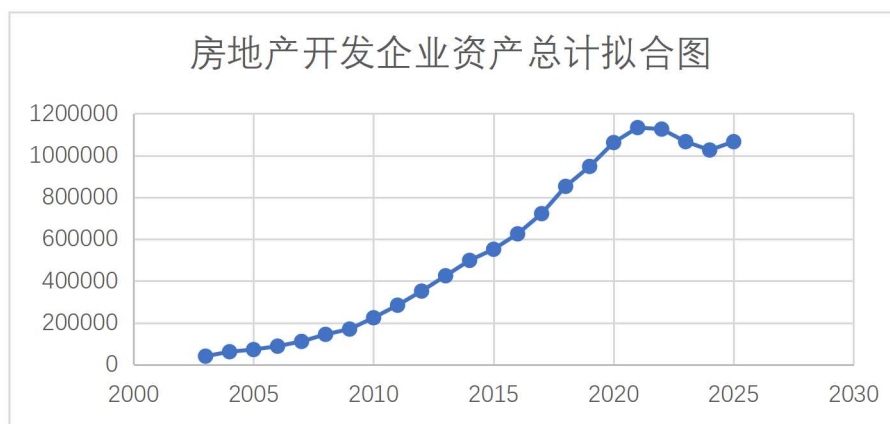


图 4-7 房地产开发企业资产总计拟合图



得到未来三年的指标后，带入模型预测即可，可以清楚的发现未来三年的房价还是会略微降低，但是基本保持平稳。商品房平均销售价格(元/平方米)，2023年，9869，2024年，9875，2025年，9658。

### 4.3 数据可视化

#### 4.3.1 可视化页面总体布局

本次实践中可视化页面并没有像以前一样采用单一的可视化大屏。采用了模块化的设计，增加了交互性和可扩展性，为用户提供更详细的数据信息和好的使用体验。总体布局采用了3个主题域。

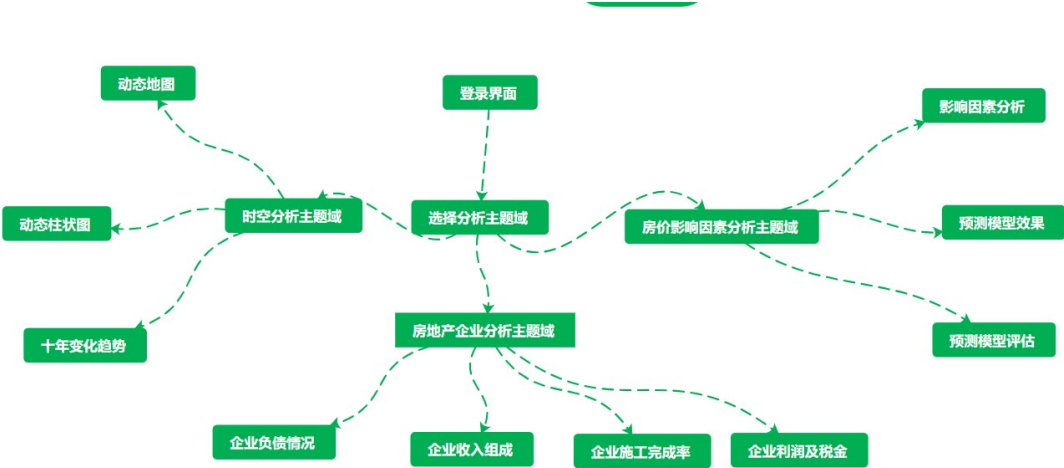


图 4-8 可视化页面模型图

#### 4.3.2 时空分析主题域

● 动态地图

该地图主要展示的是从2013年到2022年3个指标商品房价，商品房销售面积，商品房销售收入。地图颜色的深浅表面了当时当地的房价的情况，颜色越深，房价越高，地图随时间变化揭示各个省份的对比效果，可以通过左上角的点击切换指标，当然也可以通过点击地图具体的省市，查看该省市近10年房地产的变化情况.效果图如下：



图 4-9 动态中国地图

点击进入后的效果，该图展示的是河南省近 10 年的房价的变化趋势，房地产销售情况，房地产销售额。可以从图中看出 3 者都在 2019 年到达最大值，然后开始有下降的趋势。

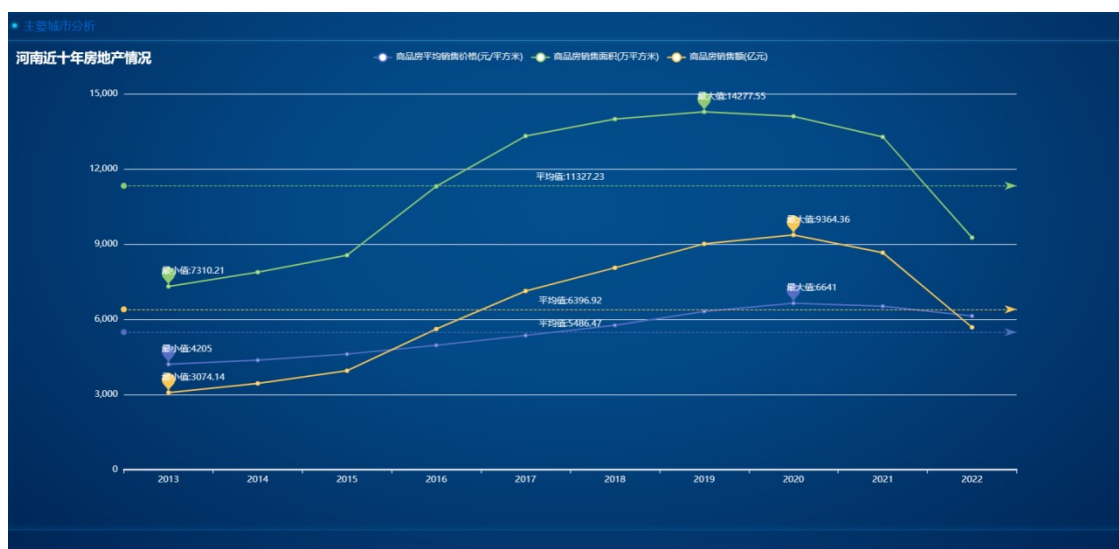


图 4-10 点击跳转图

#动态效果实现关键代码

a = 1

function run() {

if(a ==9){

```

        a=0
    }
    arr = response.filter(item => item.year === year[a]);
    const data1 = arr.map(item => ({
        name: item.city.slice(0, 2),
        value: item.price,
    }));
});
a = a+1;
myChart.setOption({
    series: [
        {
            data: newArrayWithModifiedValue1,
        }
    ],
    graphic: {
        elements: [
            {
                type: 'text',
                right: 70,
                bottom: 150,
                style: {
                    text: year[a]+'年',
                    font: 'bolder 80px monospace',
                    fill: 'white'
                },
            },
            {
                z: 100
            }
        ]
    }
});
setTimeout(function () {
    run();
}, 0);
setInterval(function () {
    run();
}, 3000);

```

这部分代码是实现跳转的关键代码

```
#实现跳转
myChart.on('click',function(params){
    var selectedRegion = params.name;
    window.location.href = 'route1.html'+'?region=' + selectedRegion;
})
```

### ● 动态柱状图

该图以动态变化的效果生动的展示了我国主要城市这 10 年来房地产施工的面积排名，为决策者提供分析的建议效果图如下：

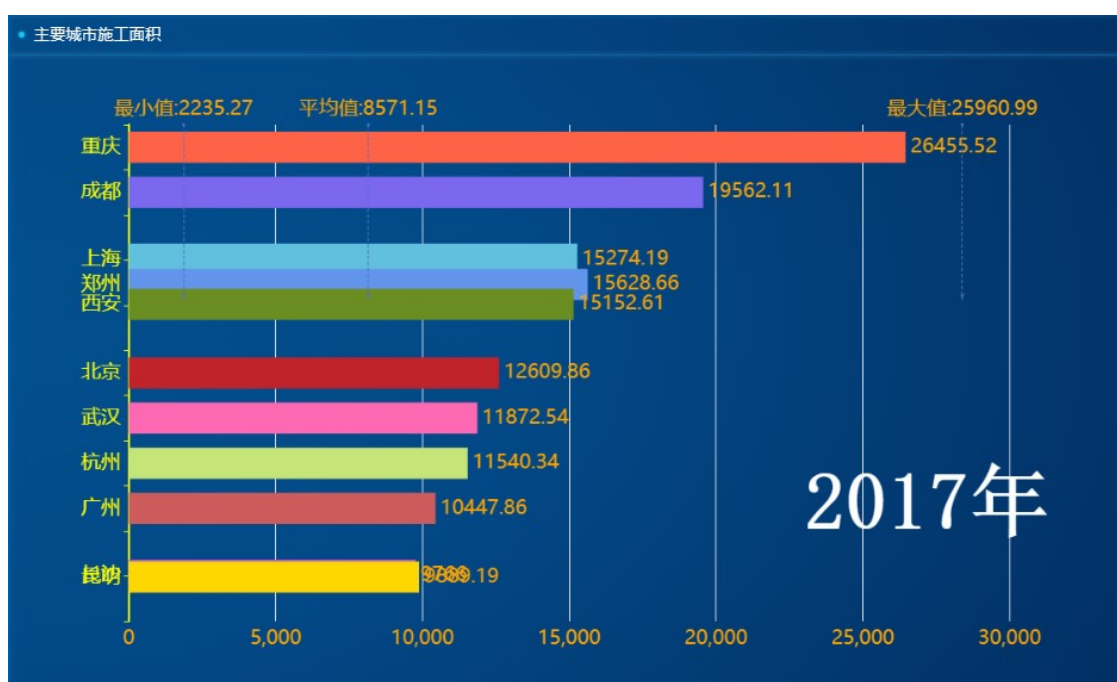


图 4-11 动态柱状图

实现关键代码：

```
myChart.setOption({
    series: [
        {
            type: 'bar',
            data: bianhua_arr3,
            id : 'TSSQuestion',

            label: {
                show: true,
                position: 'right',
```

```

        valueAnimation: true
    },
    progressive : 5000,
    animation: true,
    animationDurationUpdate: 1000,
    animationEasing: 'linear',
    animationEasingUpdate: 'linear',

```

### ● 10 年趋势变化

该图展示了从 2013 年到 2022 年 3 个指标商品房价，商品房销售面积，商品房销售收入，不过这不是某个省或城市的指标，而是全国性的数据。可以简单的看出这些指标都在 2022 年开始下降。

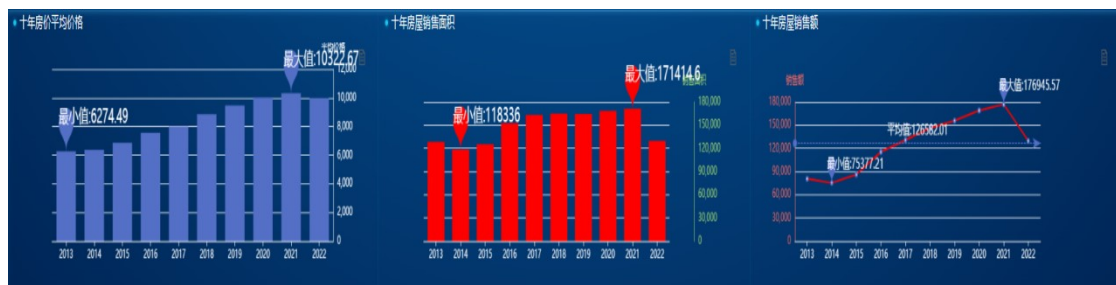


图 4-12 十年趋势变化图

关键代码。

```

#显示最大值，最小值
markPoint: {
    data: [
        { type: 'min', name: '最小值'},
        { type: 'max', name: '最大值'}],
    label: {
        position: 'start', // 文字位置
        formatter: '{b}:{c}',
        color: 'white', // 文字颜色
        fontSize: 20,
    }
    },
#显示平均线
markLine: {
    data: [{ type: 'average', name: '平均值' }],

```

```
label: {  
    position: 'middle', // 文字位置  
    formatter: '{b}:{c}',  
    color: 'white', // 文字颜色  
    fontSize: 15,  
},
```

#### 4.3.3 房地产企业分析主题域

##### ● 企业负债情况

企业负债率是企业负债总额占企业资产总额的百分比。这个指标反映了在企业的全部资产中由债权人提供的资产所占比重的大小，反映了债权人向企业提供信贷资金的风险程度，也反映了企业举债经营的能力。

但是对企业来说：一般认为，资产负债率的适宜水平是 40%~60%。

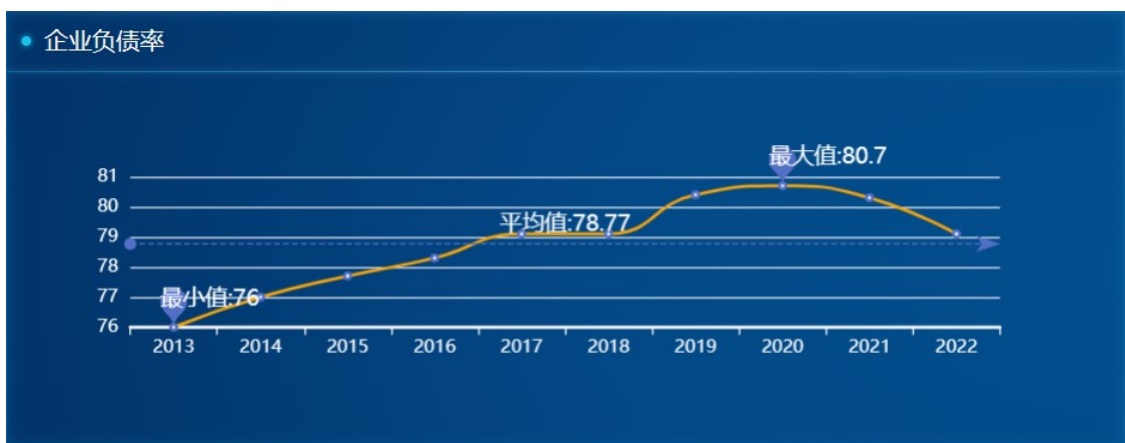


图 4-13 企业负债率折线图

##### ● 企业收入组成

企业收入组成分析对企业管理和决策具有重要作用，它能够提供深入的业务洞察，帮助企业更好地了解其盈利模式和经营状况可以通过图中看出房地产企业主要营收还是商品房销售收入，企业营收结构单一，容易受到不稳定因素的影响。需要增加其他收入的比例。



图 4-14 企业收入组成堆叠折线图

关键代码

#堆叠图关键部分

```
series: [{
  name: '企业土地转让收入(亿元)',
  type: 'line',
  stack: 'Total',
  smooth: true,
  lineStyle: {
    width: 0
  },
  showSymbol: false,
  areaStyle: {
    opacity: 0.8,
    color: new ECharts.graphic.LinearGradient(0, 0, 0, 1, [
      {
        offset: 0,
        color: 'rgb(0, 221, 255)',
      },
      {
        offset: 1,
        color: 'rgb(77, 119, 255)'
      }
    ])
  }
},]
```

## ● 企业施工完成率

房地产企业施工的完成率显然是人们非常关注的一点，最近 2 年来烂尾楼的数量越来越多。企业能否按时交房成为了重中之重。这个滚动列表，展示的就是

当年施工的完成率，虽然有一定的偏见，但是也有部分的参考价值。

地区	年份	施工面积	完成面积	当年完成率
青岛	2013	7072.55	957.34	0.14
郑州	2013	9721.23	1137.47	0.12
武汉	2013	8545.13	679.31	0.08

图 4-15 地区企业房产建设完成列表图

关键代码

```
// 使用循环遍历数据，并创建相应的<li>元素
for (var i = 0; i < scroll_arr2.length; i=i+5) {
    var ulElement = document.createElement('ul');
    for(var j = i;j< i+5;j++){
        var liElement = document.createElement('li');
        liElement.textContent = scroll_arr2[j];
        liElement.className = 'list-item';
        ulElement.appendChild(liElement);
        // 将<ul>添加到目标元素下
        scrollContainer.appendChild(ulElement);
    }
}
```

● 企业利润及税金

利润分析能够帮助企业评估其盈利能力，了解企业在特定时期内实现的净利润水平。这有助于评估经营的健康状况。税金分析有助于了解国家政策，以及企业的经营情况。通过这个展示的图表可以看出房地产企业的利润在 2018 年就开始下降了，至 2022 年以及降至和 2013 年前后相同的水平。说明企业的盈利能力开始降低，企业开始出现问题。





图 4-16 企业营业利润及税金折线图

#### 4.3.4 房地产影响因素分析主题域

该主题域主要是对数据挖掘进行可视化。

热力图展示各特征指标之间的相关性，可以看出特征间的相关性的比较强。

从中筛选出 3 个特征房地产开发企业平均从业人数(人)', '房地产开发企业计划总投资(亿元)', '房地产开发企业资产总计(亿元)', 可以看出这三个因素对房价影响的占比。房地产开发企业资产总计影响占比最大高达 0.58, 另外 2 个因素基本相当大约都有 0.2 的影响。

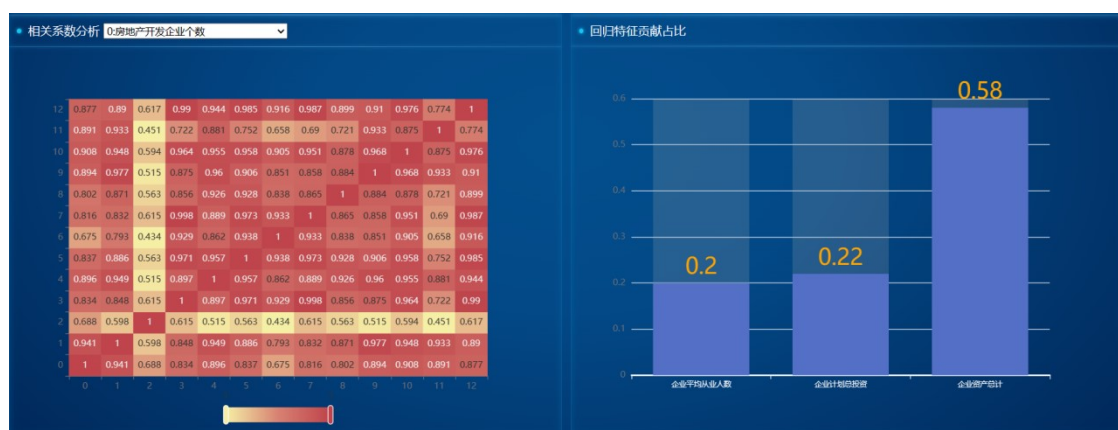


图 4-17 相关系数热力图及特征贡献图

模型评估是对预测模型进行的评价。预测展示是将实际值与预测值进行的对比演示通过这个预测值与真实值的结合图来看，未来房价还是可能会下降。但是下降的趋势并不会太大，房价会逐步归于平稳。并且结合最近几个月的政策，国家对房地产企业的帮助会越来越多，不会让房地产企业有严重的问题。

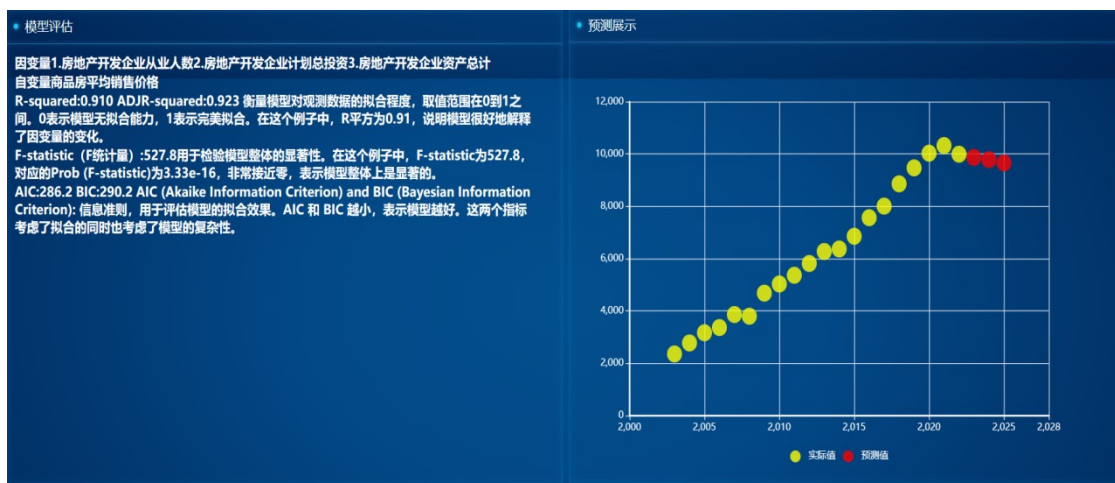


图 4-18 模型预测图

#### 4.3.5 三个主题域总体演示效果

##### ● 时空分析主题域

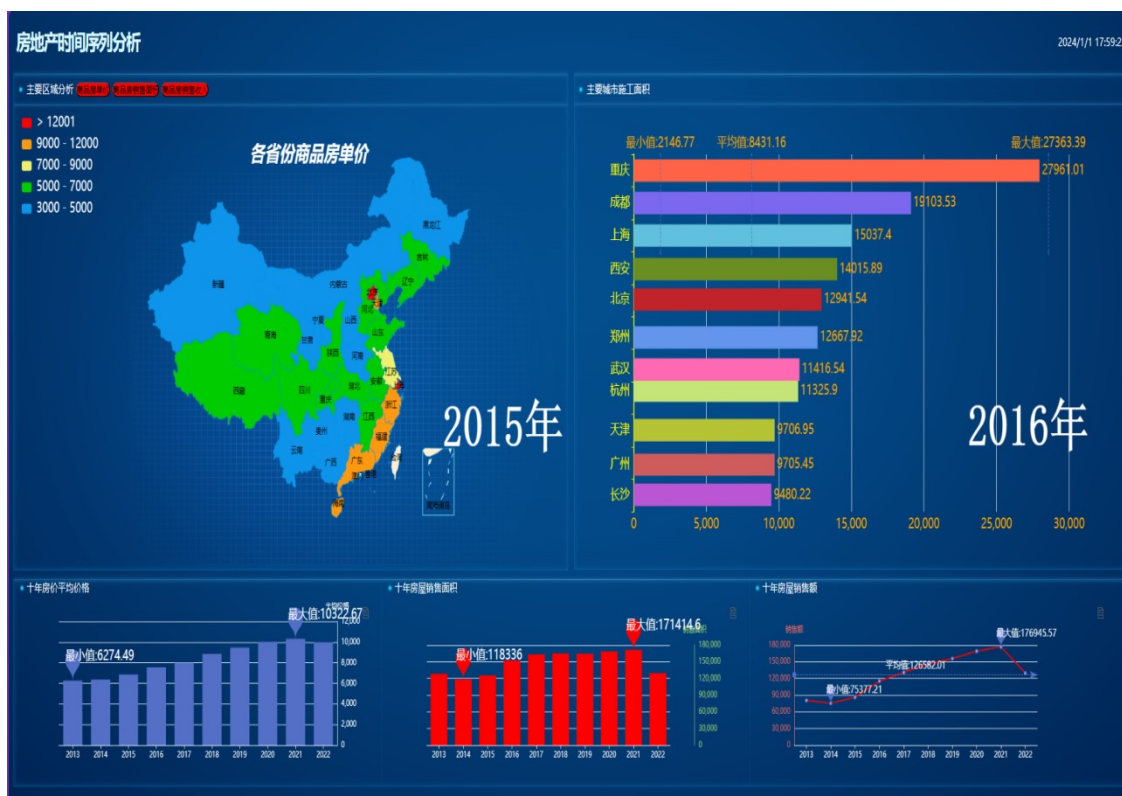
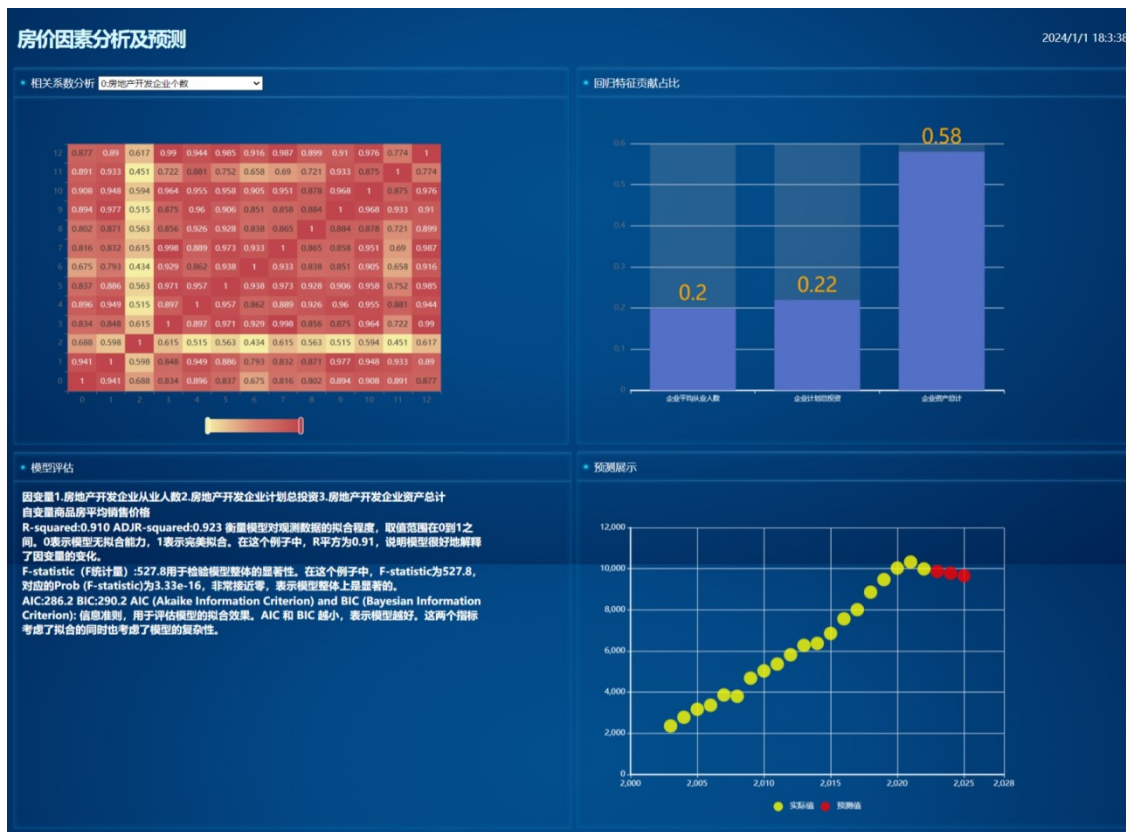


图 4-19 时空分析主题域大屏

● 房地产企业分析主题域



● 房地产影响因素分析主题域



## 5 调试分析

### 1. ECharts 动态效果无法正常开启。

问题现象：动态柱状图实现过程中出现，图发生变化是一瞬间，不是一个持续的过程，没有过渡的动画效果。

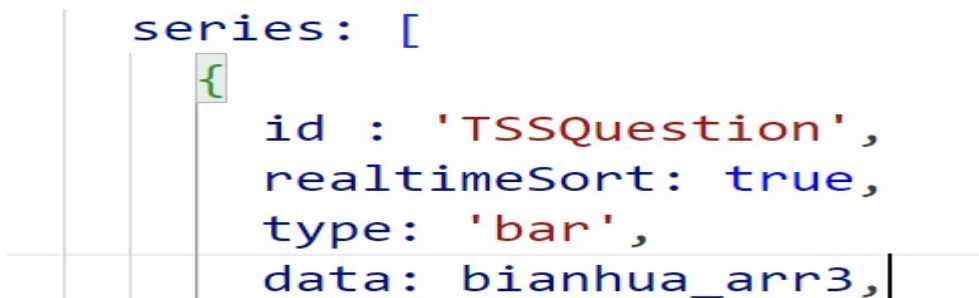
问题分析：在开始遇到这个问题时第一时间去查询配置项，认为是下面 2 个配置没有开启的原因，这 2 个配置是控制动画过渡时长所需要的。

`animationDuration: 300,`

`animationDurationUpdate: 300,`

但是在后续配置中发现还是存在这个问题。

解决方法：经过研究各种动态图的案例发现，图的变化继承效果是需要依靠 `id` 这个属性来绑定的，



```
series: [  
  {  
    id : 'TSSQuestion',  
    realtimeSort: true,  
    type: 'bar',  
    data: bianhua_arr3,  
  },  
]
```

图 5-1 配置图

ECharts 通过 `series` 中的 `id` 属性来检测图的前后继承关系。如果 `id` 属性没有指定，则会默认配置一个新的 `id`。如果图形变化后前后 `id` 不一致则会导致过渡的动画效果不能开启成功。

### 2. ECharts 动态变化图形具体实现方法。

问题现象：实现一个动态的图形最重要的在合适的时间替换掉这个图形中数据，这样就可以实现动态的效果了。当然，实现这个替换的方法有很多，这里简述一下笔者在实现过程中的方法。起初，准备使用 `js` 中的定时器来控制 `for` 循环，一步一步获取数据并且替换。但是经过实验才发现，虽然定时器可以在表面上控制 `for` 循环一步一步执行。但是实际上 `for` 循环在刚开始的那一刻已经是运行完成了，获取到的变量一直是最后一个。

原因分析：JavaScript 语言的一大特点就是单线程，也就是说，同一个时间只能做一件事。`for` 循环时一个原子语句，不会受到其他影响，要么全部执行，要么

不执行，不受定时器时间的影响。

解决方法：最后，思考了一些，发现定时器本身就是一个 for 循环，只需要在外部添加变量，在内部递增变量，就可以实现类似于循环的效果。

### 3. 实现 ECharts 地图点击跳转并且根据跳转参数绘图。

问题现象：需要实现点击地图，并且获取省市参数，绘制相应的图形。

原因分析：这是 2 个任务需要相互配合，点击地图后需要获取当前地图的名字，将其加入到跳转地址后的参数部分，接着获取 url 地址，解析其中的参数，使用参数对其数据进行过滤，绘制需要的图形

解决方法：编写函数解析 url 参数。

```
function getQuery() {  
    let href = window.location.href//获取 url  
    href = decodeURI(href);//使用 decodeURI() 对一个编码后的 URI 进行解码：  
    //或者使用 decodeURIComponent(href)  
    let query = href.substring(href.indexOf("?") + 1);//获取？后面的参数部分  
    let param = query.split("&");//将参数部分-用&分割，转换成数组  
    let obj = {}  
    for (var i = 0; i < param.length; i++) {  
        let per = param[i].split("=");//将数组中的每一项-用=分割，转换成数组  
        obj[per[0]] = per[1]//以键值对的形式储存  
    }  
    return obj;  
}
```

4. 报错遇到：Reshape your data either using `array.reshape(-1, 1)` if your data has a single feature or。

问题现象：在数据挖掘数据标准化这一步出现以上报错。

原因分析：这个错误通常发生在我们试图使用某些机器学习算法或库时，数据的形状不符合要求。具体来说，这个错误通常发生在我们试图处理只有一个特征的数据时。

解决方法：最后检测代码发现房价也被标准化了，这一步应该去掉，因变量不需要进行标准化。

## 6 总结

本次实际项目，我主要负责 2 个方面分别是数据挖掘和数据可视化下面是我对这 2 个方面的总结与展望。

首先，数据挖掘方面，按照正常的数据挖掘的流程，建立了对应的模型，预测出了相应的结果。在整个过程中学习到了很多东西，如处理数据的方法，数据的变换与选择，模型评估的指标等。本次练习加深我对数据挖掘的理解，数据挖掘不单单只是表面现象，不单单只是套代码。这是一个系统性的工程，每一步都会对最后的结果产生重要的影响。每一步都需要认真且耐心的做。当然，本次做的数据挖掘工作是较为简单的。因为数据收集的比较简单。如果要改进这次的工作，我认为应该先从数据采集方面进行，应该选项更多和房地产看似无关，实际上很有关系的变量，如 GDP，失业率等。通过更为广泛的数据集才能提供模型的灵敏性，鲁棒性。当然一个好的数据集也可以让我们选择其他更优秀复杂的模型和数据处理的方法。

接着是数据可视化方面，本次数据可视化做的略显粗糙，根本原因是，自己大量修改了别人的模板，但是自身的技术力有限，不能良好的驾驭整个布局。这次可视化选择图表也比较单一，容易产生审美疲劳。所以会导致最后的可视化结果有点唐突。个人认为，应该将可视化的交互性做的多，提高用户的自由度，用户不在是冷冰冰的对着图表，而是可以和图表互动，可以和图表动起来。一旦图表动起来，用户的思维也会动起来，这样就会碰撞出思维的火花，引出很多天马行空的想象。当然数据可视化最主要的就是直观，也需要保留其简单易读的特点。

最后，虽然本次实践还有很多方面需要改进，但是相对于上次的实践已经有了较大进步，希望在未来的实践可以改正这些缺点，更上一层楼。

## 7 参考文献

- [1] 房地产大数据及其信息挖掘体系构建路径研究[J]. 隆林宁.产业与科技论坛,2023(04)
- [2] 互联网+”背景下众筹模式在房地产开发项目中的应用[J]. 郭庆军;白思俊;张华波;朱海阔.项目管理技术,2016(01)
- [3] 用互联网思维和手段创新房地产业发展[J]. 刘志峰.住宅产业,2015(11)
- [4] 房地产行业“互联网+”模式研究与探讨[J]. 夏阳.中国房地产,2015(13)
- [5] 大数据时代:房地产业的机遇与挑战[J]. 刘昱;张玉娟.河南商业高等专科学校学报,2013