同济大学

TONGJI UNIVERSITY

《数据采集与集成技术》

实验报告

实验名称	数据集成
实验成员	2252813 肖颖
日　　期	2024 年 12 月 14 日

1、实验目的

1. 掌握基本的数据集成方法
2. 掌握简单的可视化操作

2、实验内容

本实验将整合来自多个数据源的信息，包括混凝土的成分配比、养护时间和强度测试结果。通过数据清洗、集成和可视化分析，展示集成的数据。

1 模式对齐：
将不同文件中表示相同含义但属性命名不同的某些数据列重命名，如可将将混凝土龄期信息中的属性 No.改为 number，为下面与原材料配比信息的数据融合作准备。

2 记录连接：
基于分类，比较记录对是否为统一实体的局部决策，这里具体可以举例为 element.xlsx 中的某一个行记录的是编号为 A 的混凝土原材料配比信息，age.xlsx 中有编号为 A 的混凝土的养护时间数据，这两条记录就是编号为 A 的混凝土实体的两个局部决策，数据项匹配。

3 数据融合
判断有无不同数据源同一实体的同一属性提供值时可能出现的冲突问题，融合后的数据有无缺失问题等，进行处理。

4 数据可视化
对集成后的数据作简单的数据分析，生成一些可视化结果展示。

3、实验步骤

1. 首先读取原始数据

```
element=pd.read_excel('element.xlsx')
age=pd.read_excel('age.xlsx')
strength=pd.read_excel('strength.xlsx')
```

2. 原来的表头名称太长了，对表头进行重命名，且相同含义的的数据列也需要改成相同的名字方便数据融合

```
strength.rename(columns={'serial_number': 'number'}, inplace=True)
strength.rename(columns={'Concrete compressive strength(MPa, megapascals) ': 'Concrete compressive strength'}, inplace=True)
age.rename(columns={'No.': 'number'}, inplace=True)
element.rename(columns={'Cement (component 1)(kg in a m^3 mixture)':'Cement'}, inplace=True)
element.rename(columns={'Blast Furnace Slag (component 2)(kg in a m^3 mixture)':'Blast Furnace Slag'}, inplace=True)
element.rename(columns={'Fly Ash (component 3)(kg in a m^3 mixture)':'Fly Ash'}, inplace=True)
element.rename(columns={'Water  (component 4)(kg in a m^3 mixture)':'Water'}, inplace=True)
element.rename(columns={'Superplasticizer (component 5)(kg in a m^3 mixture)':'Superplasticizer'}, inplace=True)
element.rename(columns={'Coarse Aggregate  (component 6)(kg in a m^3 mixture)':'Coarse Aggregate'}, inplace=True)
element.rename(columns={'Fine Aggregate (component 7)(kg in a m^3 mixture)':'Fine Aggregate'}, inplace=True)
```

3. 对三个表的数据进行融合，采用内联法，只保留相同的行，将结果保存为文件

```
merge = pd.merge(element, age, on='number', how='inner')
merge = pd.merge(merge, strength, on='number', how='inner')
# 保存合并后的文件
merge.to_excel("merge.xlsx", index=False)
```

4. 绘制水泥含量和抗压强度的散点图

```
plt.figure(figsize=(8,6))
plt.scatter(merge['Cement'],merge['Concrete compressive strength'])
plt.title('Scatter Plot of Cement vs Strength', fontsize=16)
plt.xlabel('Cement', fontsize=14)
plt.ylabel('Concrete compressive strength', fontsize=14)
plt.show()
```

5. 绘制不同年龄的平均抗压强度的柱状图

```
# 计算每个年龄的 strength 均值
age_strength_mean = merge.groupby('Age (day)')['Concrete compressive strength'].mean()

# 绘制柱状图
plt.figure(figsize=(10, 6))
age_strength_mean.plot(kind='bar', color='skyblue', alpha=0.8)

# 添加标题和轴标签
plt.title('Average Strength by Age', fontsize=16)
plt.xlabel('Age', fontsize=14)
plt.ylabel('Average Strength', fontsize=14)

# 显示图表
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

6. 计算属性相关性，绘制热力图

```
columns_of_interest = ['number', 'Cement', 'Blast Furnace Slag',
                       'Fly Ash', 'Water','Superplasticizer',
                       'Coarse Aggregate','Fine Aggregate',
                       'Concrete compressive strength','Age (day)']
correlation_matrix = merge[columns_of_interest].corr()

plt.figure(figsize=(8, 6))
sns.heatmap(correlation_matrix, annot=True, cmap="coolwarm", fmt=".2f", linewidths=0.5)

# 添加标题
plt.title('Correlation Heatmap of Attributes', fontsize=16)
plt.show()
```

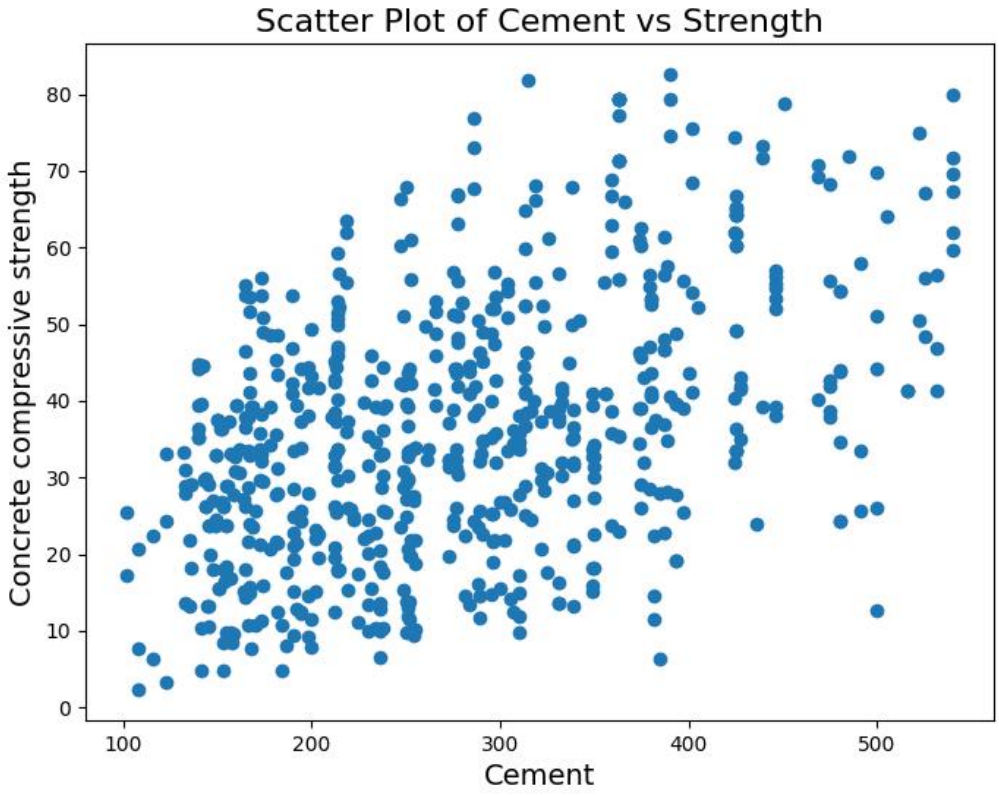
4、实验结果

融合后的表格内容如下

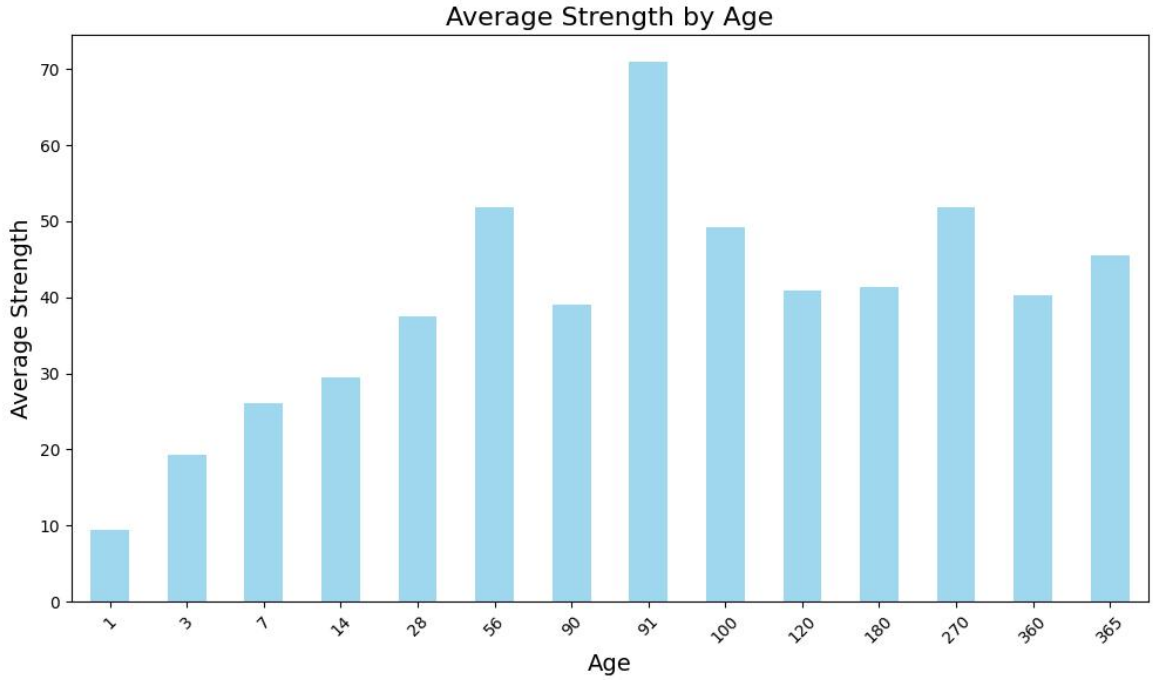
number	Cement	Furnace	Fly Ash	Water	rplastic	se Aggre	e Aggreg	Age (day)	ompressive
180000	540	0	0	162	2.5	1040	676	28	79.99
180291	182	45.2	122	170.2	8.2	1059.4	780.7	28	31.27
180535	393	0	0	192	0	940.6	785.6	90	48.85
180262	212.6	0	100.4	159.4	10.4	1003.8	903.8	56	44.4
180108	379.5	151.2	0	153.9	15.9	1134.3	605	7	47.1
180426	173.8	93.4	159.9	172.3	9.7	1007.2	746.6	14	29.55
180238	213.8	98.1	24.5	181.7	6.7	1066	785.5	100	49.97
180159	389.9	189	0	145.9	22	944.7	755.8	56	79.4
180852	298	0	107	186	6	879	815	28	42.64
180824	302	0	0	203	0	974	817	28	21.75
180297	168.9	42.2	124.3	158.3	10.8	1080.8	796.2	56	39.15
180896	313	161	0	178	10	917	759	28	52.44
180032	198.6	132.4	0	192	0	978.4	825.5	180	41.72
180533	289	0	0	192	0	913.2	895.3	3	11.65
180733	349	0	0	192	0	1056	809	28	33.61
180917	148	175	0	171	2	1000	828	28	26.92
180363	218.2	54.6	123.8	140.8	11.9	1075.7	792.7	100	63.53
180399	234	156	0	189	5.9	981	760	28	39.3
180690	288	192	0	192	0	932	717.8	7	23.52
180266	212	0	124.8	159	7.8	1085.4	799.5	28	38.5
180248	238.1	0	94.1	186.7	7	949.9	847	100	44.3
180818	525	0	0	189	0	1125	613	28	55.94
180752	540	0	0	173	0	1125	613	14	59.76
180757	350	0	0	203	0	974	775	7	18.13
180693	153	102	0	192	0	888	943.1	28	17.96
180425	167	75.4	167	164	7.9	1007.3	770.1	14	32.9
180219	166.1	0	163.3	176.5	4.5	1058.6	780.1	3	10.76
180747	500	0	0	200	0	1125	613	3	26.06
180097	375	93.8	0	126.6	23.4	852.1	992.6	7	45.7
180300	290.4	0	96.2	168.1	9.4	961.2	865	14	34.67
180771	331	0	0	192	0	978	825	3	13.52
180621	307	0	0	193	0	968	812	180	34.49

装 订 线

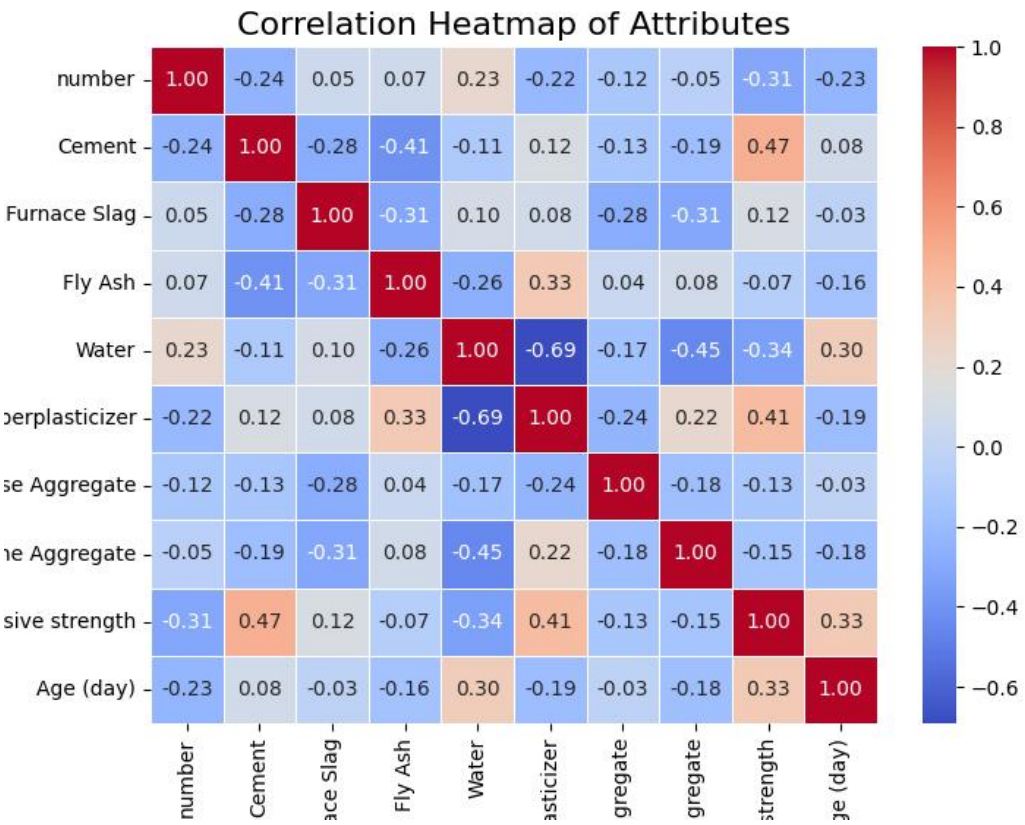
绘制散点图如下，可以看出水泥含量和抗压强度是呈正相关的



绘制柱状图如下：可以看出年龄最低的水泥平均强度最差，年龄居中的水泥平均强度最好



绘制属性相关性热力图如下：越接近 1（颜色越红）表示正相关越强，越接近-1（颜色越蓝）表示负相关越强。



5、实验总结

问题：最开始我在读取 excel 文件时遇到了编码问题的报错，后来发现我使用的函数是 read_csv 而不是 read_excel，read_csv 应该是适用于读取 csv 文件的，换成 read_excel 就没问题了。

心得体会：通过这个实验我掌握了基本的数据集成方法，主要就是模式对齐、记录连接和数据融合，并且学会了使用 python 来处理，并将处理后的数据进行可视化，分析不同属性之间的相关性。通过数据集成可以将不同源的数据聚集在一起进行处理和分析。