自动化生产线数据与影响因素分析

摘要：随着信息技术的快速进步，工业自动化技术日新月异，在显著提高制造效率的同时极大的降低了生产成本。为了提高生产效率，减少次品率，需要对生产线数据进行深入分析。

针对任务 1，我们借助 Python 的 Pandas 库和 Numpy 库对数据进行处理。通过统计产品的总产量和合格情况、提取每条故障信息后根据故障类型对故障信息进行分类整理，统计生产线有效生产时长并对其进行**相关性分析**。

针对任务 2，我们通过任务 1 中所统计的信息，借助 Python 的 Matplotlib 库和 Seaborn 库对多项数据进行了**可视化处理**，绘制了多种图表，直观展现了自动化生产线的时间，故障，产量等有效信息。

针对任务 3，我们尝试了多种影响因素分析方法：首先，我们通过可视化不同故障类型与合格率之间的关系，大致确定了分析方向；其次，在尝试传统主成分分析算法（PCA）失败后，我们选择**凸优化**对各种故障类型的影响权重进行了精确求解，并与可视化结果对比取得了一致的结论：A1 与 A4 对产品合格率的影响最大，其次是 A3，A2 对产品合格率的影响最少；最后，我们对产量与故障时间进行了线性假设，并进行了**显著性检验**，验证了我们的结果：产量与故障时间呈现显著负相关，在误差范围内相关性系数$r = -1$。

关键词：数据分析，数据可视化，相关性分析，凸优化，显著性检验

目录

任务 1 数据整理与统计

任务 1.1

借助 python 中的 pandas 库，我们分别统计了两条生产线每天的产品总数（包含不合格产品）、合格产品数、不合格产品数与合格率，并计算了各生产线全年的产品总数、合格产品数、不合格产品数与合格率，结果见表 1。

表 1. 各生产线全年的产品总数、合格产品数、不合格产品数与合格率

生产线	产品总数（件）	合格产品数（件）	不合格产品数（件）	合格率（%）
M101	1183016	1180910	2106	99.82
M102	1186838	1184461	2377	99.80

从表格中我们可以看到产品总数生产量足够大，能避免部分偶然情况的产生，同时两个生产线上的产品生产总量、不合格和合格产品的数量相差不大。

任务 1.2

按照月份、日期和开始时间升序，我们总结出列出两条生产线每次故障的相关信息，并根据以上信息我们总结出各生产线每种故障一年中第 25 次发生的相关信息，结果见表 2。

表 2. 各生产线每种故障一年中第 25 次发生的相关信息

生产线	故障类别	月份	日期	开始时间	持续时长
M101	A1	2	14	28553	247
M101	A2	1	4	19819	672
M101	A3	1	30	2717	865
M101	A4	1	13	22814	887
M102	A1	3	2	25982	581
M102	A2	1	5	7842	384
M102	A3	2	12	12307	737
M102	A4	1	14	28122	625

从图中可以看出每种故障类别在开始时间和日期上并没有表现出强烈的相关性，每次故障的持续时长也相差较大。

任务 1.3

根据任务 1.2 的结果，分别统计两条生产线各类故障每天发生的总次数和平均持续时长，按照生产线、月份、日期、故障类别升序排列，并汇总于表 3

表 3. 两条生产线各类故障发生的总次数、平均持续时长、故障发生频率

生产线 M101	A1	A2	A3	A4	汇总
总次数	240	2335	319	826	3720
平均持续时长（秒/次）	474.03	541.31	728.83	738.83	596.89
发生频率（次/天）	0.66	6.40	0.87	2.26	10.19

表 3 中可以清晰的看出生产线上各类故障的总次数、平均持续时长、发生频率均有明显的差别，具体的分析将在任务 2 中具体展开。

任务 1.4

有效工作时长等于当天开机时长减去因故障停机的时长。具体的，我们将‘故障类别’列中所有非空值设为 1，空值设为 0。用开机时长减去发生故障而停机的时长，得出有效工作时长，并计算出日平均有效工作时长，结果如表 4 所示。

表 4. 各生产线的日平均有效工作时长（小时/天）

生产线	日平均有效工作时长（小时/天）
M101	6.31
M102	6.33

生产线的日工作时长为 8 小时，而生产线上的日平均有效工作时长在 6.3 小时左右，说明优化故障的出现可以很大程度上增加日平均有效工作时长从而提高生产效率。

任务 1.5

为完成生产线 M101 的相关性分析，首先我们利用 python seaborn 库中的 corr 函数计算每天推出的电路板数量、抓取的元件数量和抓取的故障次数的 Pearson 相关系数。

注 1.（Pearson 相关系数） Pearson 相关系数是一种用于度量两个变量之间线性相关程度的统计量，其取值范围从-1 到+1：

- *+1 表示完全正相关，即一个变量增加，另一个变量也增加。*
- *-1 表示完全负相关，即一个变量增加，另一个变量减少。*
- *0 表示没有线性相关关系*

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

其中，x_i 和y_i是两个变量的各个数据点。$\bar{x}$和$\bar{y}$是两个变量的平均值

我们得出相关系数矩阵如下：

表 5. 相关系数矩阵

	推出累计数	抓取累计数	抓取故障次数
推出累计数	1.000000	0.999998	-0.637233
抓取累计数	0.999998	1.000000	-0.637400
抓取故障次数	-0.637233	-0.637400	1.000000

进一步，我们绘制热力图与散点图来更直观地展示这些变量之间的关系。

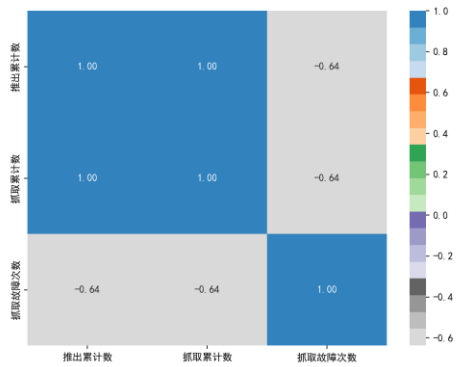

图 1. 相关系数矩阵热力图

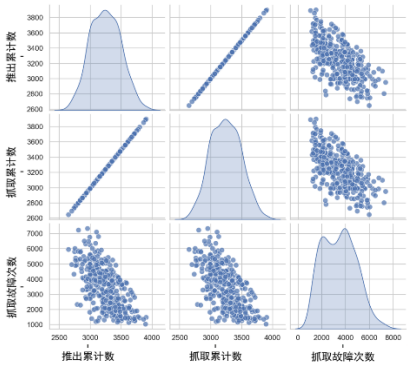

图 2. 多变量散点图

热力图和相关性矩阵形象地展示了‘推出累计数’、‘抓取累计数’与 ‘抓取故障次数’之间的相关系数。我们可以得出以下结论：

1）‘推出累计数’与‘抓取累计数’有明显的相关性，相关系数达到了 0.999998，近似接近 1，说明两者间有极强的正相关。抓取设备的运作频率与电路板推出数量呈正向关系是因为，设备的抓取需要电路板的正确推出为前提，另外，由**任务 1.1** 可知，整个自动化流程具有较高的合格率，因此，抓取的故障也发生次数较少，由此，可以认为‘推出累计数’与‘抓取累计数’数量大致相当，故呈现极高的相关性；

2）‘抓取故障次数’与‘推出累计数’或‘抓取累计数’均无正相关性，相反，它们间具有-0.637400 的负相关性。因为故障的发生可能存在于电路板传送、元件抓取、元件安装、电路板检测 4 道工序中的任一步骤，而推出与抓取计数的前提是它们没有发生故障，由此整个流程发生故障的概率降低，故‘抓取故障次数’与‘推出累计数’或‘抓取累计数’呈现负相关性。

任务 2 生产线运行情况的可视化分析

任务 2.1

根据各月的产品总数（包含不合格产品），以月份为横坐标，绘制两条生产线的堆叠柱状图。

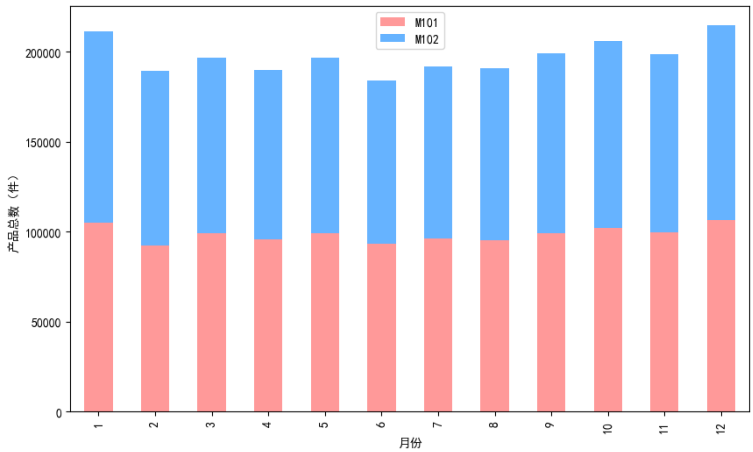

图 3. 两条生产线的堆叠柱状图

从图中我们可观察出，M101 和 M102 两条生产线每月生产的产品总数持平，即生产效率相当。

任务 2.2

根据生产线每天的不合格产品数（取值范围为[0,35]）和不合格率（取值范围为[0,0.9%]），分别绘制两条生产线的双 Y 轴折线图。

由于不合格产品数和不合格率整体趋势相当，为了避免不合格产品数与不合格率重叠影响观察，我们将不合格率的 y 轴上下限重新设定为[-0.3%，0.9%]，结果如图 4.1 与 4.2。

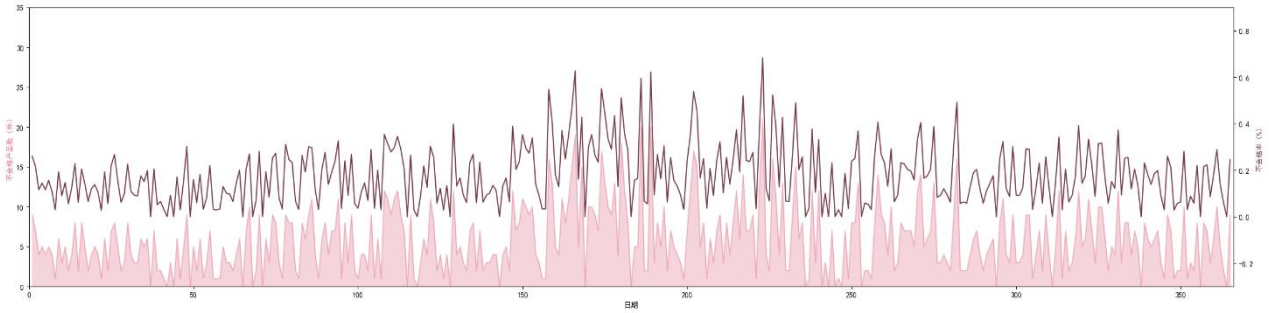

图 4.1. 生产线 M101 的双 Y 轴折线图

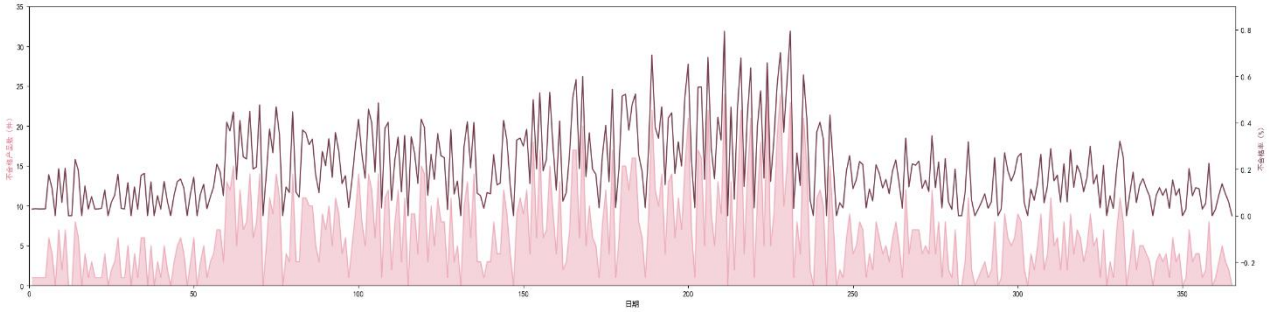

图 4.2. 生产线 M101 的双 Y 轴折线图

从图中，我们可以观察出，生产线 M101 与生产线 M102 在全年中不合格产品数与不合格率走势相当，即在年初和年末较低，在年中相对较高。具体而言，在 0 至 60 天（1 至 2 月）和 240 至 365 天（8 月至 12 月）不合格产品数维持在 8 至 16 个，不合格率维持在 0 至 0.4%；在 60 至 240 天（3 至 7 月）不合格产品数提升到 8 至 33 个，相应不合格率取值提升到 0 至 0.8%。

任务 2.3

根据不同故障类别的全年发生总次数，绘制两条生产线各故障类别的占比双层环形图（其中，外圈为 M101，内圈为 M102）。

从图中我们可以观察出，两条生产线各故障类比占比大致相同。对于生产线 M101 和 M102，占比最大的是故障 A2，分别占 62.00%与 62.77%，其次是故障 A4，分别占 23.34%与 22.20%，接着是 A3，分别占 8.71%与 8.58%，占比最低的是 A1，分别占 5.95%与 6.45%。

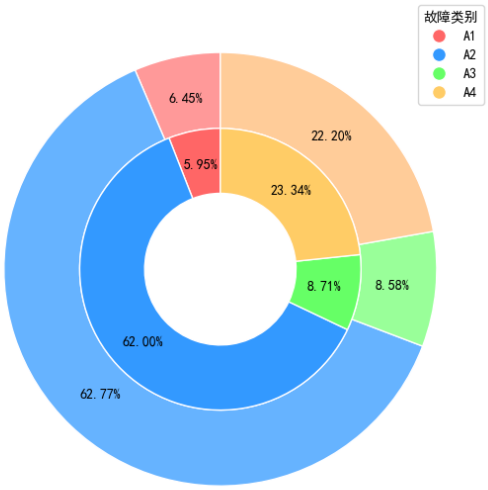

图 5. 两条生产线各故障类别的占比双层环形图

任务 2.4

根据不同故障类别，以持续时长（单位：秒）为横坐标，分别绘制两条生产线全年故障发生持续时长的叠加直方图。

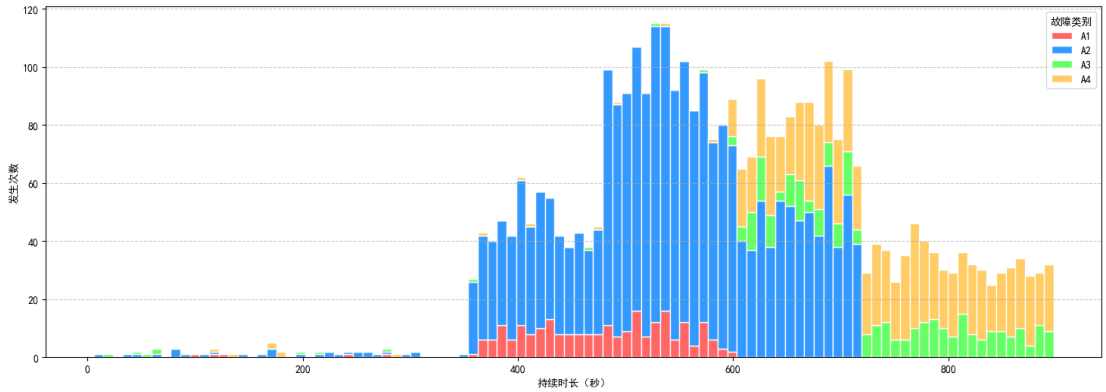

图 6.1. 生产线 M101 全年故障发生持续时长的叠加直方图

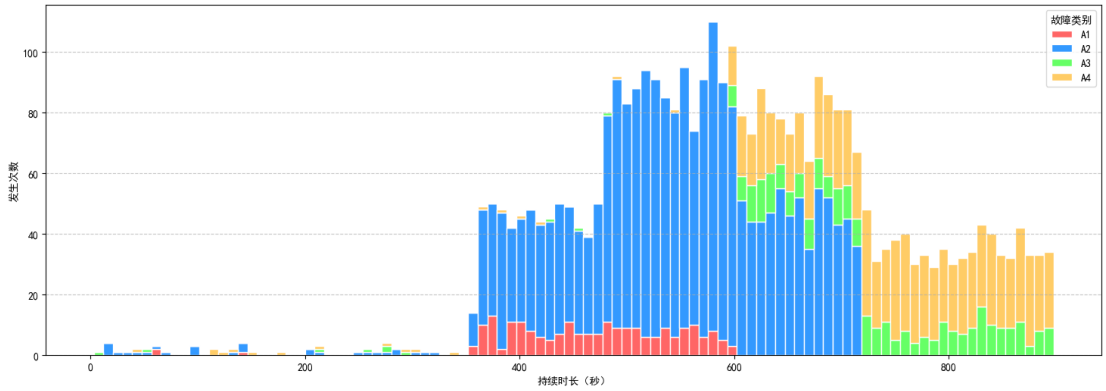

图 6.2. 生产线 M102 全年故障发生持续时长的叠加直方图

从两张图表可以看出，两个生产线全年四种故障发生持续时长在直方图上的表现几乎一致。除了极少量数据，其余四种故障发生的持续时长均有明显的

规律。例如 A1 故障均匀的分布在 400 秒和 600 秒之间，A2 故障分布于 350s 和 700s 左右，500s 到 600s 之间是 A2 故障发生的主要持续时长，且 A2 的发生次数遥遥领先。A3 和 A4 的故障发生持续时长较长，分布在 600 秒到 800 秒之间。A1、A2、A4 的故障发生没有明显的突出，在时间轴上分布的相对均匀。

任务 2.5

根据 4 月 26 日前 100 秒生产线数据，以时间（单位：秒）为横坐标，电路板推出次序为纵坐标，绘制生产线 M101 包含 4 个工序的甘特图。

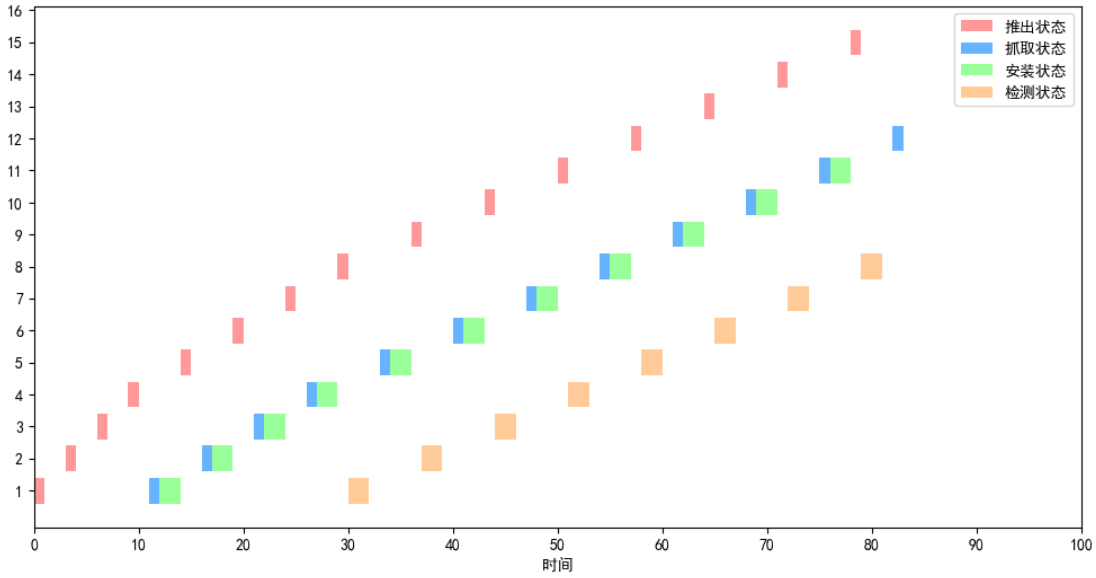

图 7. 生产线 M101 包含 4 个工序的甘特图

从甘特图上色块弯曲的程度可以看出，从一开始推出状态的快速运行后逐渐减速，最后达到 4 个工序速度一致的状态。其中抓取状态和安装状态是承接完成，其余工序之间均有等待的时间。

任务 3　生产线影响因素分析

3.1 可视化分析

首先，我们通过可视化分析 A1,A2,A3,A4 与合格率之间的关系。因为直观上来说，合格率与各种故障的发生次数呈现负相关性，并且不合格的产品数量远小于合格产品数量，故单从合格率的角度来看，各类故障的发生次数对合格率的影响不明显。因此我们采用直接分析各类故障与不合格的产品数量之间的关系来反映其与合格率之间的关系。

因为原始数据中的统计次数均是以天为单位进行计数，故在本任务中的分析仍然以天作为基本单位进行分析。通过控制变量数量，作任意三类故障与不合格产品数量的关系图如下。其中 x、y、z 轴分别表示三类故障的发生次数，不合格产品数（件）通过散点的大小与颜色表示，颜色越红、半径越大表述数量越多，具体的数值对应关系通过右侧的颜色条表示。

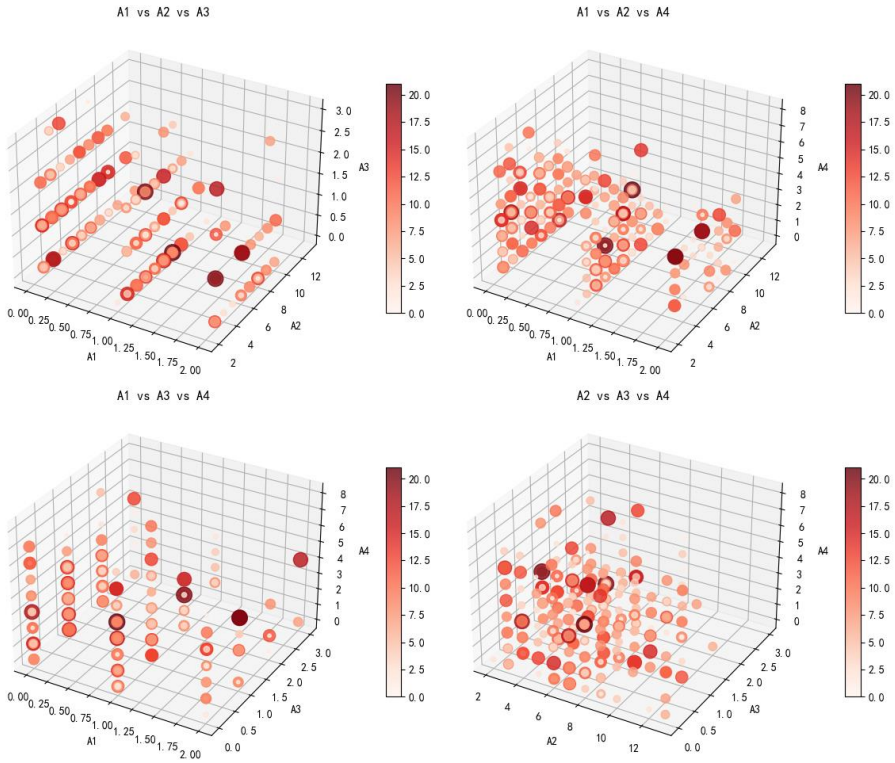

图 8. 任意三类故障的数量与不合格产品数（件）之间的关系散点图

从中可以初步观察到，A2 故障的发生次数与不合格产品数呈一定的负相关或者不相关性。为进一步确认其中的关系，我们单一故障的发生次数与不合格产品数的散点图如下。

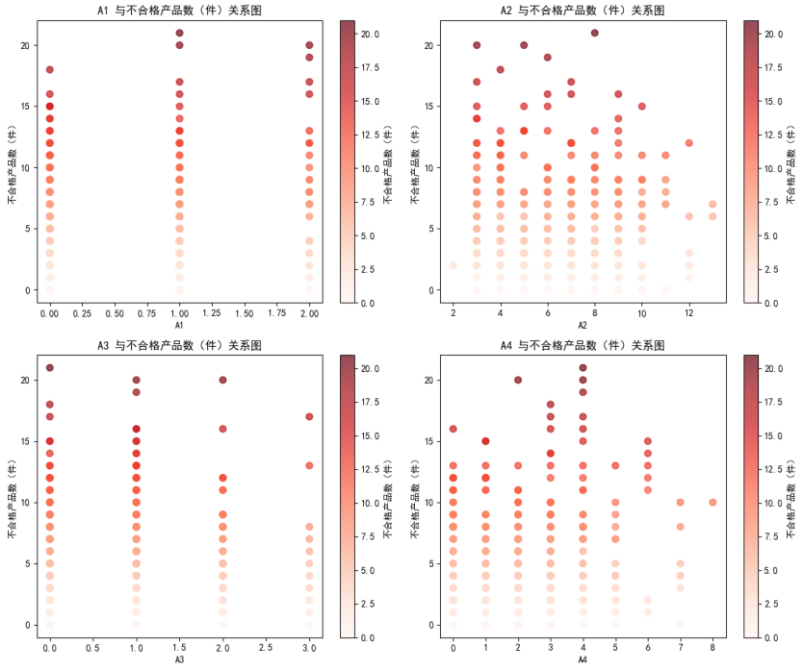

图 9. 各故障的发生次数与不合格产品数的散点图

通过上图可以进一步得知 A2 故障的发生次数越多，相对来说不合格产品数越少，尽管这违背了常识，但可能从侧面反映了 A2 故障可能并不对产品合格率

造成直接影响，而其余三类故障并无此关系。但可以观察到，不合格产品数>15时的概率集中在 A4 故障发生在 3-5 次之间。

在有以上初步结论的基础上，我们使用 PCA 主成分分析法和凸优化求解进一步分析 A1,A2,A3,A4 与合格率之间的关系。

3.2 PCA 的尝试

在前面的分析中我们得知，有一些故障类型如 A2、A4 可以观察到它们与不合格产品数的关系，而对于如 A1、A3 的故障来说，似乎这种关系从图中来看并不是很明显，故我们尝试用主成分分析（PCA）来进行特征降维。

注 2. 主成分分析（PCA）是一种用于降维的统计方法，旨在通过将数据投影到新的坐标轴（主成分）上，最大化数据的方差，从而减少数据的维度，同时尽可能保留数据中的主要信息。

其基本原理是，首先计算数据的协方差矩阵，该矩阵反映了各个变量之间的关系。接着，求解协方差矩阵的特征值和特征向量，特征向量确定了新的坐标轴方向（即主成分），特征值则表示各主成分的方差大小。然后，选择特征值的贡献率最大的 k 个主成分，这些主成分包含了数据中的主要信息。最后，将原始数据投影到这些主成分上，从而实现数据的降维。

通过代码进行计算，我们得到四个特征值的贡献率依次为：0.31869483，0.2728078，0.23418587，0.17431151。但根据主成分分析方法的一般原理，通常只能当累计贡献率>85%时才可成功进行降维操作，但根据实际计算得到的数据，我们无法有理由相信 PCA 在此是行得通的。然而，PCA 对任务三进行的尝试也间接说明了 A1,A2,A3,A4 与合格率之间的关系是存在的，且不能被忽视，因此我们还需进一步探寻其中的关联。

3.3 凸优化求解

无监督学习的 PCA 并不能很好的反映数据内部结构，因此，我们引入监督学习的凸优化方法。

首先，需要明确的不同故障类型 A1,A2,A3,A4 是影响产品合格率的主要因素，而影响产量的主要因素则是故障时间。

为了更精确分析影响产品合格的可能因素，我们希望求出各个故障类型 A1,A2,A3,A4 对合格的影响幅度占比（w_3，w_3，w_3，w_4），即故障类型影响权重矩阵W,我们旨在最小化以下目标函数：

$$\min_{\boldsymbol{W}} \|\boldsymbol{W}\boldsymbol{X}-\boldsymbol{Y}\|_F^2$$

注 3. 其中，Frobenius 范数定义为：

$$\|A\|_F^2=\sum_{i,j}|a_{ij}|^2$$

对于权重矩阵W,它往往应具备元素和为 1 与非负性两个局部属性，但是，由前面 3.1，3.2 小节分析，我们可知，各故障类型可能与产品合格率并不简单呈正相关，因此，我们创新性地取消非负约束，构造出以下优化目标：

$$\min \|\boldsymbol{W}\boldsymbol{X} - \boldsymbol{Y}\|_F^2 \quad \text{subject to} \quad \boldsymbol{W}^\top \mathbf{1} = \mathbf{1},$$

展开 Frobenius 范数，我们有

$$\|\boldsymbol{W}\boldsymbol{X}^\top - \boldsymbol{Y}^\top\|_F^2 = (\boldsymbol{W}\boldsymbol{X}^\top - \boldsymbol{Y}^\top)^\top(\boldsymbol{W}\boldsymbol{X}^\top - \boldsymbol{Y}^\top)$$

$$= \boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{W}^\top - 2\boldsymbol{Y}^\top\boldsymbol{W}\boldsymbol{X} + \boldsymbol{Y}^\top\boldsymbol{Y}.$$

引入拉格朗日乘子λ,接着，我们构造相应的拉格朗日问题：

$$\mathcal{L}(\boldsymbol{W}, \lambda) = \boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{W}^\top - 2\boldsymbol{Y}^\top\boldsymbol{W}\boldsymbol{X} + \boldsymbol{Y}^\top\boldsymbol{Y} - \lambda(\boldsymbol{W}^\top\mathbf{1} - 1).$$

对权重矩阵W求导，我们得到：

$$\nabla_{\boldsymbol{W}}\mathcal{L} = 2\boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{X} - 2\boldsymbol{Y}^\top\boldsymbol{X} - \lambda\mathbf{1}.$$

令导数为 0，我们解得：

$$2\boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{X} = 2\boldsymbol{Y}^\top\boldsymbol{X} + \lambda\mathbf{1}.$$

对于拉格朗日乘子λ，我们得到：

$$\lambda = \frac{2\left(\boldsymbol{Y}^\top\boldsymbol{X}\mathbf{1} - \boldsymbol{W}\boldsymbol{X}^\top\boldsymbol{X}\mathbf{1}\right)}{\mathbf{1}^\top\mathbf{1}}.$$

为了简化计算，我们通过 python 中 scipy.optimize.minimize 迭代求解带约束的最优化问题，避免了手动推导拉格朗日乘数中的高额时间复杂度。

```
最优解 W: [0.36946779 0.00636694 0.2358107  0.38835457]
W * ones(4) 是否接近 1: True
```

图 10. scipy.optimize.minimize 迭代求解结果

我们得到故障流程 A1,A2,A3,和 A4 对影响产品合格率的权重系数分别为 0.37，0.01，0.24，和 0.39，也即，A2 对产品合格率的影响最少，接近于 0. 这符合我们 3.1 中可视化观察到的反常结果。同时，A1 与 A4 对产品合格率的影响最大，其次是 A3。

3.4 故障对生产线产量的影响

为探究影响生产线的产量的因素，我们从实际入手可以很明显的知道，生产线的生产时间与产量密切相关，通常生产时间越长，产量越高。故我们从生产时间入手可以得知，故障发生的持续时间对产量至关重要，我们以天为单位散点图观察。

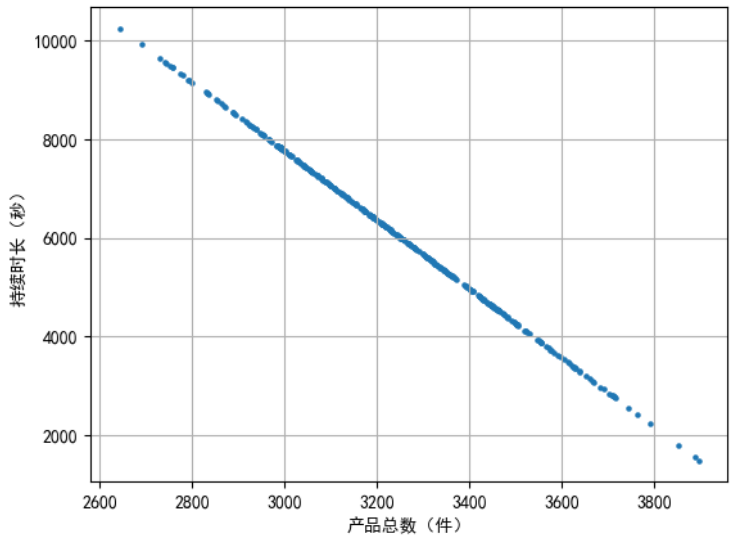

图 11. 故障发生的持续时间与生产线产量的散点图

由图可知，故障发生的持续时间与生产线产量几乎呈严格的负相关性，为更严谨的验证结论，我们编程得到结果，其中，故障发生的持续时间与生产线产量相关性系数 r=-0.9999993，进行显著性检验，p 值为 0，因此我们可以认为 r=-1。

假设故障发生的持续时间为 y，生产线产量为 x，进一步计算得 y 与 x 的关系如下：

$$y = -6.99x + 28728.95$$

至此，我们得到了影响生产线产量的重要因素。

综上分析，A1 与 A4 对产品合格率的影响最大，其次是 A3，A2 对产品合格率的影响最少；而产量与故障时间呈现显著负相关。