Howework #1

我们提供了真实的数据挖掘场景（个贷违约预测）的数据，见 data_10000.csv，要求根据该数据完成第一、二、三题。

一、数据属性练习（2 分）

对于下面列出的每一个数据属性，从 {Nominal, Ordinal, Interval, Ratio} 四个选项中选出最合适的类型，并说明理由

1. total_loan（贷款数额）
2. year_of_loan（贷款年份）
3. grade（贷款级别）
4. loan_id（贷款记录唯一标识）

二、计算统计信息（2 分）

1. total_loan (贷款数额) 是否存在缺失值？请对可能的缺失值进行填充 (使用均值填充)。（注：下面二、三道题目的 total_loan 都使用填充后的结果。）
2. 计算 total_loan（贷款数额）的五数概括，并画出盒图
3. interest（网络贷款利率）和 scoring_low（借款人在信用评分系统所属的下限范围）两个特征之间是否存在相关性？请给出理由

三、数据预处理与可视化（3 分）

1. 绘制 work_year（就业年限）和 employer_type（所在公司类型）的分布情况，给出柱状图
2. 考虑 monthly_payment（分期付款金额），使用等深分箱分成 10 个箱，并画出每个箱包含的人数的直方图
3. 考虑 scoring_low（借款人在信用评分系统所属的下限范围），使用等宽分箱分成 10 个箱，并画出每个箱包含的人数的直方图

四、文本数据的表示（3 分）

我们提供了纽约时报的部分新闻语料，在文件夹 nyt_corp0/中，每一个文件表示一篇文档。使用任意一种编程语言，完成如下练习:

1. 文档的表示: 根据语料内容构造词典，然后将语料中的每篇文档都表示成词典上的 tf-idf 向量。

2. 词语的表示: 使用先前构造的词典，计算词语的共现矩阵，从而得到词语的共现向量。(共现矩阵的一个元素 C(i, j)表示词语 i 和词语 j 共同在文档中出现的次数)

3. 文档距离计算与分析: 任选一篇文档，使用 tf-idf 向量找出与它欧式距离最近/余弦相似度最高的各 5 篇文档，并简单分析这 10 篇文档是否与它内容相似。

4. 词语距离计算与分析: 任选一个词典中的词，使用共现向量找出与它欧式距离最近/余弦相似度最高的各 5 个词，并简单说明这 10 个词是否与它意思相近。

提交说明: 需要提交源代码与报告。报告中简单说明 1)、2)的实现思路，如词典的构造细节，然后写清楚 3)、4)的结果与分析。

注意: 请不要调用第三方的库来直接生成文档和词语的表示。