序号:　　成绩:

数据分析程序设计

课程报告

学号: 20123101

姓名: 李昀哲

专业: 智能科学与技术

题目：用 Pandas 揭秘美国选民的总统喜好

赛题介绍:

本项目主要带领学习者利用 Python 进行数据分析以及数据可视化，包含数据集的处理、数据探索与清洗、数据分析、数据可视化四部分，利用 pandas、matplotlib、wordcloud 等第三方库带大家玩转数据分析。

针对数据新人开设的实战练习专场，以有趣主题作为实践场景，提供详尽入门教程以及 DSW 算力资源，手把手教你学习 Python 语法。天池希望此赛事成为高校备受热捧的 Python 实战课程，帮助更多学生掌握 Python 技能、增加实战项目经验。

赛题以数据处理、数据分析、数据可视化为任务，数据集可以报名参与比赛后，查看操作指南进行下载使用，该数据来自 FEC 平台的公开数据集，源数据包含 3 张数据表，经过 baseline 数据处理步骤后合并为一张表，总数据量 75w+条，包含 8 列变量信息。

鼓励结合各类方法对数据进行个性化的分析和可视化。

利用 Pandas 分析美国选民总统喜好度

摘要：本文基于 Pandas 数据处理工具，以 2022 年美国联邦选举委员会 FEC (Federal Election Commission)官网最新数据为切入点，将“捐款信息”、“候选人委员会信息”、“候选人信息”相关联，提取、归纳出有用知识，以此分析美国不同地区选民对不同总统候选人的喜好度。采用多种可视化方法直观地展示分析结果，使用信息熵公式计算选民对候选人的喜好或胜选概率。分析结果结合各类算法，可以用于未来总统竞选成功率的预测。

关键字：数据分析；数据可视化；美国选民喜好；信息熵

1. 引言

美国选民体系正在逐渐进入一个选民日益活跃的党派主义时代，特别是在总统选举中，选民的角色日益重要[1]，2020 年美国总统选举投票率达到 1910 年以来的历史最高点，约为 66.9%，同时还在不断呈现上升趋势，如图 1 所示。选民的积极参与表明了其喜好对于总统选举起到了重要作用，有助于预测新的总统当选人。因此，通过相关数据，分析选民对总统候选人的喜好有极高的研究价值和研究前景。

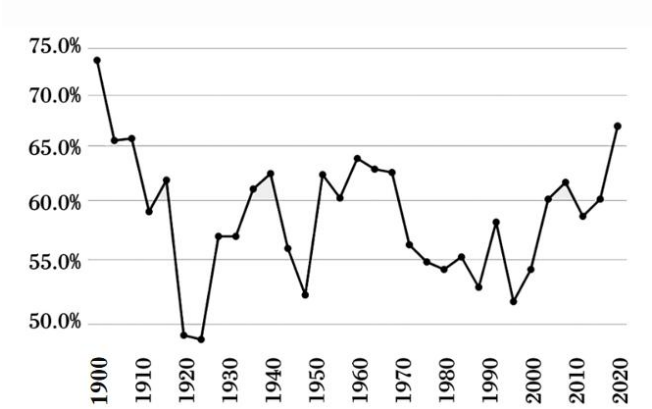

图 1 1900-2020 年美国总统大选选民投票率示意图

选民喜好主要通过反向分析他们对于总统候选人委员会的捐款信息，归纳各种分类结果得到捐款排行，每一位候选人对应一个或多个选举人委员会，选民对于委员会的捐款可以认为是对某个候选人的“喜好”或“倾向”。

根据上述“喜好”的定义，很自然地将问题聚焦于选民对候选人委员会的捐款，对于捐款的分析又可以拆分为对各种分类结果下捐款排名的分析，各种选民对美国总统的偏向程度等子问题，例如：分析捐款人的党派，由党派捐款总金额排行、对应候选总统所属党派判断党派捐款上的优劣势等。

选民捐款信息、候选人委员会信息、总统候选人信息均为公开数据，可在美国联邦选举委员会 FEC (Federal Election Commission)官网查询和下载。但官网数据较为复杂、冗长，人类很难通过肉眼观察直接、直观地得出结论，因此引入 Pandas 数据分析工具，对多份较大规模的数据进行整合、关联，提取有价值的信息进行统计和分析。通过对捐款各个子问题进

行实验，得到了较为直观的分析结果；利用信息熵公式，结合数据，计算候选人胜选的概率，根据结果得出结论：Sullivan Dan 是受到大部分选民青睐的总统候选人，Demings Val 也有较大的潜力。

2. 方法

2.1 数据预处理

2.1.1 多表合并

本项目采用的数据集为更具时效性的 2022 年最新候选人、选举人委员会、选民捐款信息，摒弃了“天池”平台提供的对于 2020 年总统竞选的较早数据。数据集中各个表的信息如表 1 所示，各数据表中的数据信息项目较多，且十分分散，很难直接从单张表中得到有价值的信息，因此需要根据“共有数据信息项”进行合并。

表 1 数据集信息

数据表名称	描述
ccl22	委员会信息：候选人 ID、选举年份，委员会 ID 等；
weball22	候选人信息：候选人 ID、姓名、候选人委员会、政党等；
Itcont_2022_20220629_20220817	个人捐款信息：姓名、所属职业、所在州、委员会 ID 等；

Pandas 库中的 ***pandas.merge()***方法可以将两个 DataFrame 类型数据表按照指定的规则进行连接，通过 ***left/right***, ***on***, ***left_on***, ***right_on***, ***how***, ***sort*** 等参数的使用，最后拼接成一个新的、理想的 DataFrame 类型数据表，以便进行后续的数据分析操作。

这里将 **ccl22** 数据表作为合并过渡表，其中存在和 **weball22** 关联的“候选人 ID”，和 **Itcont_2022_20220629_20220817**（2022 年 6 月 29 日至 2022 年 8 月 17 日捐款信息）关联的“委员会 ID”，通过 ***pandas.merge()***方法进行合并，得到易于分析、完整性较强的数据表。

2.1.2 缺失值处理

个人或企业每天会产出或者收集大量的数据，机器学习和人工智能等学科的发展使得数据的获取越来越便捷，同时数据的重要性愈发凸显，数据质量也逐渐引起人们的重视。其中数据缺失问题常常发生，甚至难以回避。因此，处理缺失值是分析、实验前必不可少的预处理步骤，处理主要分为两个方向：删除和插补，在删除和插补中又分为多种方法以处理不同的数据情形，如对于本数据表中的“分类数据”，常使用“将缺失值作为新的分类”、“多重插补”、“逻辑回归”等方法，图 2 详细列举了各种缺失值处理的方法。

目前，针对图 2 中的插补方法，研究人员也在尝试引入神经网络、构建学习模型对缺失值进行填补，目前的研究主要有基于自联想神经网络(Auto-Associative Neural Network, AANN)对于属性关联建模填补缺失值 [2]、基于置信度对不完整数据建模填补缺失值 [3] 等，这使得在数据质量难以保障的情况下，减少缺失值对研究的影响。

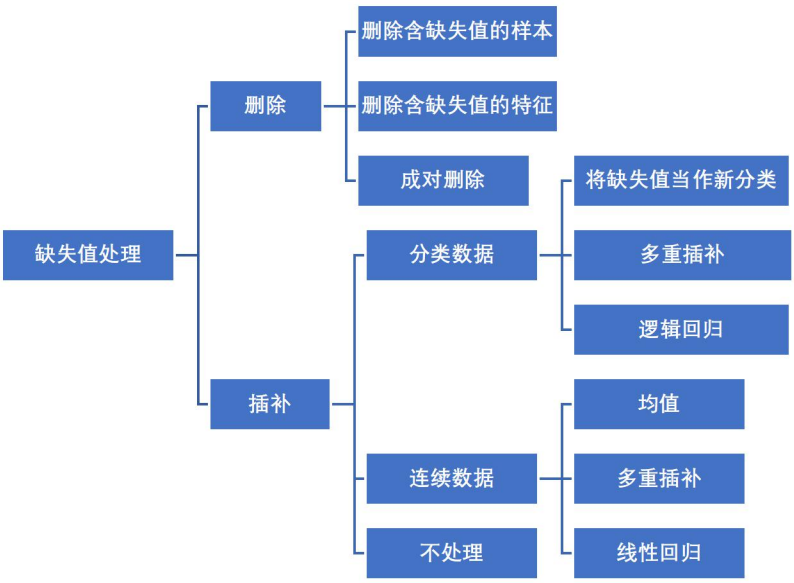

图 2 缺失值方法处理

根据本数据集的特征，缺失内容大多为职业（OCCUPATION）、雇主（EMPLOYER），故结合实际可能发生的“未调查到选民雇主信息”、“选民没有实际工作”的情况，因此这里对于缺失值的处理，选择仅将缺失值作为一个新的分类 ***NOT PROVIDED*** 比较合适。

Pandas 库中提供了 ***dataframe.fillna()***方法，其中方法的调用者 dataframe 为某个具体的 DataFrame 类型实例，填充内容可以是常数或字典，还可以根据需要 ***method*** 参数用前一个或下一个非缺失值进行填充。

2.2 数据分析及可视化

数据分析主要将会从各个数据项的捐款排名入手，分析选民党派、各个候选人获得的捐款总额及人数、最热门的两位候选人一个月内每一天的捐款信息、捐赠的职业分布以及捐款人所属州的分布情况，使用 ***SeaBorn*** 进行图表可视化分析、***folium*** 进行热力图分析，通过饼图、折线图、直方图、热力图等图表直观地展现数据结果，便于直观地观察和进行后续的拓展研究。

2.3 信息熵公式预测胜选率

本想法受决策树中应用信息熵、信息增益率筛选节点属性的启发 [4]，认为信息熵公式不失为一个可以用于判断选民喜好程度的依据。

在信息论中，随机离散事件出现的概率存在着不确定性或混乱程度。通常，一个信源发送出什么符号是不确定的，衡量它可以根据其出现的概率来度量。概率大，出现机会多，不确定性小；反之不确定性就大。不确定性函数 ***Entropy*** 是概率 ***P*** 的减函数；两个独立符号所产生的不确定性应等于各自不确定性之和。因此每一位选民的捐款可以看作是“消息”，每一个捐款事件 ***x*** 的信息量可以反映某选民对某候选人的喜好不确定度，信息量公式如下，

$$h(x) = -\log_2 p(x) \tag{1}$$

再将所有选民捐款抽象为一个随机变量 $\boldsymbol{X}$，它的所有可能取值的信息量的期望就称为信息熵，根据信息熵公式，计算“消息”的不确定性，不确定性越大，受选民喜好的概率越低，或胜选概率越低。

为了衡量上述信息的不确定性，信息学之父香农(Shannon)引入了信息熵的概念，并给出了计算信息熵的数学公式，本文的计算也将基于这个公式展开：

$$Entropy(X) = -\sum_{i=1}^{n} p(x_i)\log p(x_i) \tag{2}$$

3.实验结果及分析

3.1 数据预处理

根据 2.1 所述方法，使用 ***pandas.read_csv()***方法读入数据，各数据表的部分详细信息如表 2(a), 2(b), 2(c)所示，每张数据表都具有冗长、繁杂的信息，但并非所有数据列对分析选民喜好有价值，因此根据 2.1.1 多表合并方法，采用 **ccl** 数据表中的 CAND_ID 关联 **weball22** 数据表、CMTE_ID 关联捐款信息数据表，进行合并操作，得到较精简的捐款信息，且数据列的选择做到了尽可能的全面。

表 2(a)　weball22 数据集部分信息

CAND_ID	CAND_NAME	…	CAND_PTY_AFFILIATION	TTL_RECEIPTS	…
H2AK00200	CONSTANT, CHRISTOPHER	…	DEM	164637.90	…
H2AK01158	PELTOLA, MARY	…	DEM	1912264.52	…
H2AK01240	WOOL, ADAM L	…	DEM	16217.07	…
H2AK00218	REVAK, JOSHUA CARL	…	REP	121841.00	…
…	…	…	…	…	…

表 2(b)　ccl 数据集部分信息

CAND_ID	CAND_ELECTION_YR	FEC_ELECTION_YR	CMTE_ID	CMTE_TP	…
C00713602	2019	2022	C00712851	O	…
H0AK00105	2020	2022	C00607515	H	…
H0AL01055	2022	2022	C00697789	H	…
H0AL01063	2020	2022	C00701557	H	…
…	…	…	…	…	…

表 2(c)　个人捐款数据集部分信息

CMTE_ID	AMNDT_IND	RPT_TP	TRANSACTION_PGI	IMAGE_NUM	…
C00556506	A	Q2	G2022	202208169525409046	…
C00556506	A	Q2	G2022	202208169525408999	…
C00556506	A	Q2	G2022	202208169525409014	…

C00590489	A	Q2	P2022	202207159521961995	…
…	…	…	…	…	…

合并后的数据表包含候选人姓名、捐款人姓名、捐款人所属州、捐款人雇主、捐款人职业、捐款人捐款总数、捐款人捐款日期、候选人党派八类信息。通过 ***dataframe.info()***方法查看整体数据信息。

其中 Non-Null 表示某个数据列中非空的值，在 326055 组数据中，雇主（EMPLOYER）、职业（OCCUPTION）两项数据列的缺失值情况较多，因此根据 2.1.2 缺失值处理方法进行处理，将缺失值看作是新的类 ***NOT PROVIDED*** 插补。

经过预处理后的数据表如表 3 所示，保留了尽可能完整的、有价值的数据项，且对于缺失值进行了插补，并对于捐款日期进行更便于理解的形式，即:“月/日/年”改为了“年/月/日”。

表 3 多表合并后数据信息

CAND_NAME	NAME	STATE	EMPLOYER	OCCUPATION	TRANS_AMT	TRANS_DT	CAND_PTY
CARL, JERRY LEE, JR	DUNN, TIM	TX	CROWN QUEST OPERATING	CFO	2900	2022630	REP
CARL, JERRY LEE, JR	HOWEL DIANNE M.	TX	MILTOPE	PRESIDENT	500	2022630	REP
CARL, JERRY LEE, JR	POARCH BAND OF CREEK INDIANS	AL	NOT PROVIDED	NOT PROVIDED	2900	2022630	REP
HARVEY-HALL, PHYLLIS	BISSOO, MIRANDA	AL	NOT PROVIDED	NOT PROVIDED	500	2022630	DEM
SEWELL, TERRI A.	MCCOY, MOYER	DC	PMI GLOBAL SERVICES	MANAGER EXTERNAL AFFAIRS	350	2022629	DEM
…	…	…	…	…	…	…	…

3.2 数据分析及可视化

3.2.1 捐款人党派情况

对于一个具体的 DataFrame 对象而言，可以调用 ***dataframe.groupby()***方法获取其中某项数据列的全部数据，再调用 ***sum()***函数，对“捐款金额”（TRANSACTION_AMT）数据进行求和，结果如表 4 所示，美国的主要党派为民主党和共和党，因此该结果并不令人意外且符合实际。党派在选举上有着决定作用,可以说什么党的票数多大概率就对应了什么党派候选人会胜选，从结果中不难看出民主党（DEM）选民的捐款总额远超共和党（REP）选民的捐款总金额，但距离 2024 年大选还有很长一段时间，且本数据集仅是一个月的信息，因此并不能直接决定胜选情况。通过网络资料发现，现任美国总统为民主党人拜登，因此捐款总数民主党占据较大优势也一定程度上和现任总统党派有关。

表 4 党派捐款信息表

CAND_PTY_AFFILIATION	TRANSACTION_AMT
DEM	97,503,585
REP	29,409,549
DFL	500,940

3.2.2 捐款人对总统候选人的喜好情况

我们将捐款人对候选人的捐款看作是对某个候选人的“喜好”，因此采用类似的方法 ***dataframe.groupby("CAND_PTY_AFFILIATION").sum()***，对每个总统候选人所获得的捐款总额进行分析，如表 5 所示，在一个月中，7 名候选人获得的捐款总额超过 1,000,000 美元，其中前 4 名更是超过了 2,000,000 美元。但对于获得捐款最多的 MERCER, LEE 获得的捐款数额过于庞大，甚至占到了民主党派一个月内总捐款数额的 66%，且数额是整数，因此怀疑该候选人的捐款信息可能出现异常。

表 5 候选人获捐款总额

CAND_NAME	TRANSACTION_AMT
MERCER, LEE	64,000,000
SULLIVAN, DAN	5,092,362
DEMINGS, VAL	2,705,822
HASSAN, MARGARET WOOD	2,481,407
KELLY, MARK	1,790,196
RUBIO, MARCO	1,507,452
PELOSI, NANCY	1,135,776

表 6 候选人获捐款人数

CAND_NAME	DEVOTION_NUM
SULLIVAN, DAN	71,427
DEMINGS, VAL	43,867
RUBIO, MARCO	21,961
KELLY, MARK	21,916
HASSAN, MARGARET WOOD	14,831
OCASIO-CORTEZ, ALEXANDRIA	8,669
SANDERS, BERNARD	7,084

单独获取 MERCER LEE 所获捐款信息，查看发现，他获得的捐款记录仅有一条，同时捐款人是他自己，可能是将自己的资产全部捐给自己作为参选的筹码。这种行为对于上述的怀疑是成立的，也很自然地驱使我们思考：不能仅将候选人所获捐款的数额作为选民喜好的判断标准。因此引入候选人所获捐款的人数作为辅助标准，通过选出 DataFrame 类型对象中的*“CAND_NAME”*进行统计，作为人数的分析依据，如表 6 所示。

对于表格信息，在图 3(a), 3(b)中进行了直观的可视化分析，现任美国阿拉斯加州联邦参议员 SULLIVAN DAN 和佛罗里达州众议员、现任总统拜登的副手 DEMINGS VAL 无论在获得捐款金额上，还是获得捐款人数上，都占据较有利的地位，二者可被认为是受选民喜爱的两位候选人。

类似地，对捐款人的职业分布进行分析。分析前可以根据“民主党所获捐款总额稍占优势”这个前提进行猜测，民主党代表“穷人”和“中产阶级”利益，而共和党代表“富人”利益，因此一个月内的捐款信息应当以平民或社会身份地位较低的职业为主。由于表格信息并不直观，

因此此处利用 ***SeaBorn*** 工具进行可视化分析，如图 3(c), 3(d)所示，退休选民及无雇主选民的捐款金额和人数占据多数，符合之前的猜测和其余数据分析结果。

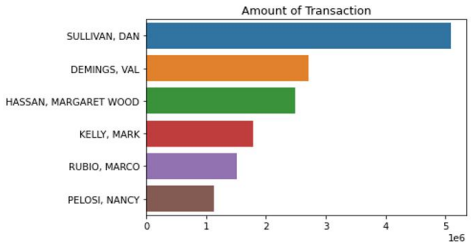

图3(a) 候选人获捐款总额

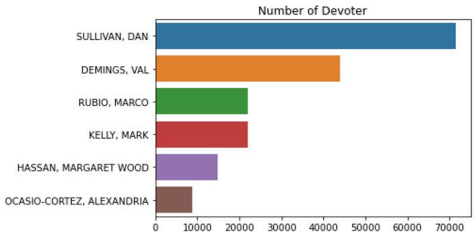

图3(b) 候选人获捐款总人数

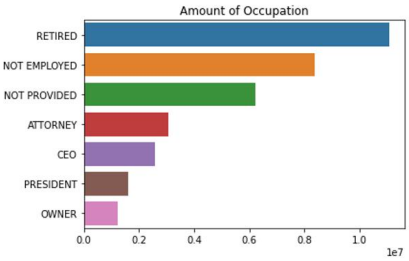

图3(c) 候选人获捐款各职业总额

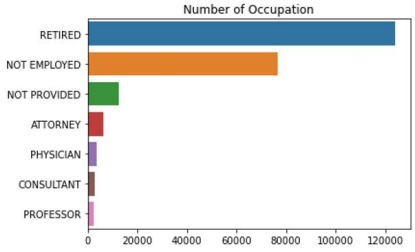

图3(d) 候选人获捐款各职业总人数

图 3 候选人获捐款数额、职业信息

根据美国选举机制，各个州的总捐款数和对应的人数不仅可以反映该州的选票将归于哪一位总统候选人，也能侧面反映该州的经济发展程度，具有社会意义，通过两种热力图的分析，直观地体现了各个州在一个月内的捐款数额，图 4 (a)从上至下依次根据色块颜色，区分了各个州的数额多少；图 4 (b)中红色的纽约州（NY）和橙色的加州（CA）、佛罗里达州（FL）色块最具区分度，也是捐款最多的三个州，其余绿色的部分，由深至浅，数额递减。

根据往年数据和美国经济发展来看，这三大州的独占鳌头并不令人意外，佛罗里达的经济在近年呈现显著的上升，有赶超加州的趋势。捐款数额方面，猜测可能也与候选人中有在佛罗里达州较有声望的众议员、拜登副手 DEMINGS VAL 有关，或许也会为她在后续的大选中助力。

图 4 (a)

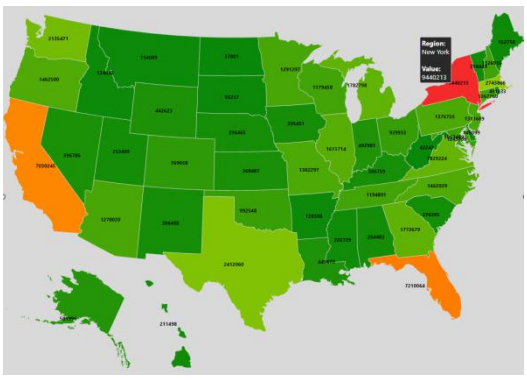

图 4 (b)

图 4 热力图分析各州捐款数额

SULLIVAN DAN 和 DEMINGS VAL 在此份数据中，毫无疑问是两大热门人选，因此对二者在一个月中的获捐款情况进行了详细的分析，如图 5 所示直观展示了从 2022 年 6 月 29

日到 2022 年 7 月 31 日期间的获捐款信息，图中不难看出，从折线图分析，SULLIVAN DAN 更胜一筹。图 6 详细展示了 SULLIVAN DAN 获得各州捐款的分布，饼图呈现的分布较为平均，各州都有相应的贡献，也反映了其良好的群众基础。

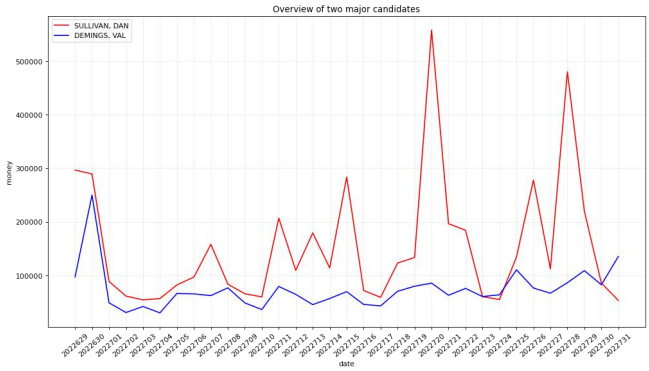

图 5 热门候选人获捐款信息对比

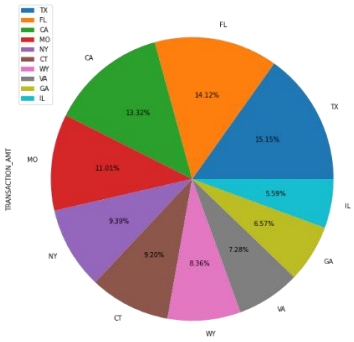

图 6 SULLIVAN DAN 获各州捐款分布

3.3 信息熵公式预测胜选率

表 7 候选人获捐款人数

	获支持州总数	支持选民总数
SULLIVAN DAN	**48**	**2833**
DEMINGS VAL	46	2384

美国共有 50 个州 1 个华盛顿特区，捐款人总数为 13553 人。根据 2.3 中的公式（2）和表 7 数据，可计算得出各总统“获支持州总数”的信息熵和“支持选民总数”的信息熵，即 $\boldsymbol{p}(X)_{\boldsymbol{STATE}}$，$\boldsymbol{p}(X)_{\boldsymbol{OCCUPATION}}$：

$$\boldsymbol{p}(\text{SULLIVAN DAN}, 51, 48)_{\boldsymbol{STATE}} = -\left(\frac{48}{51}\right)\log_2\left(\frac{48}{51}\right) - \left(\frac{3}{51}\right)\log_2\left(\frac{3}{51}\right) = 0.323 \quad (3)$$

$$\boldsymbol{p}(\text{SULLIVAN DAN}, 13553, 2833)_{\boldsymbol{OCCUPATION}} = -\left(\frac{2833}{13553}\right)\log_2\left(\frac{2833}{13553}\right) - \left(\frac{10720}{13553}\right)\log_2\left(\frac{10720}{13553}\right) = 0.740 \quad (4)$$

$$\boldsymbol{p}(\text{DEMINGS VAL}, 51, 46)_{\boldsymbol{STATE}} = -\left(\frac{46}{51}\right)\log_2\left(\frac{46}{51}\right) - \left(\frac{5}{51}\right)\log_2\left(\frac{5}{51}\right) = 0.463 \quad (5)$$

$$\boldsymbol{p}(\text{DEMINGS VAL}, 13553, 2384)_{\boldsymbol{OCCUPATION}} = -\left(\frac{2384}{13553}\right)\log_2\left(\frac{2384}{13553}\right) - \left(\frac{11169}{13553}\right)\log_2\left(\frac{11169}{13553}\right) = 0.671 \quad (6)$$

综合两个概率，可以通过加权平均的方法计算，权值$\boldsymbol{\alpha}$和$\boldsymbol{\beta}$的设置可以根据需求由其他算法得到，此处二者都取 0.5，

$$\boldsymbol{p}(X) = \boldsymbol{\alpha}\cdot(1-\boldsymbol{p}(x)_{\boldsymbol{STATE}}) + \boldsymbol{\beta}\cdot(1-\boldsymbol{p}(x)_{\boldsymbol{OCCUPATION}}) \quad (7)$$

SULLIVAN DAN 胜选（受选民喜好）的概率为:

$$[(1-0.323)+(1-0.740)] / 2 = 0.4685 \quad (8)$$

DEMINGS VAL 胜选（受选民喜好）的概率为:

$$[(1-0.463)+(1-0.671)]\ /\ 2=0.433 \tag{9}$$

4. 总结

本文主要通过一系列数据处理方法从约 750,000 条数据中，获得简洁性较强，数据项价值较高的数据表用于进一步的分析。分析过程中，通过饼图、直方图、折线图、热力图等多种图表、信息熵理论，对各类数据项进行了分析，得到如下结论：

(1) 由于民主党选民总体基数较大，且现任总统为该党派，因此民主党选民捐款总额暂时高于共和党选民；

(2) 现任美国阿拉斯加州联邦参议员 Sullivan Dan 和佛罗里达州众议员、现任总统拜登的副手 Demings Val 无论在所获捐款总额和所获捐款人数上，都有较大优势，因此认为二者是较受选民喜爱的候选人；

(3) 捐款职业分布上以退休选民和无雇主选民为主，大多数代表民主党，社会地位相对较高的“律师”位列表中第四位，该分布也符合(1)中民主党选民捐款总额较多的情况；

(4) 捐款人所属州的分布上主要集中于纽约州（NY）、加州（CA）、佛罗里达州（FL），三者也是美国经济较为发达的州；候选人中较为热门的 Demings Val 正是出自佛罗里达州，且曾任 27 年警察局局长，在当地声誉较好，猜测会在后续得到更大支持；

(5) Sullivan Dan 和 Demings Val 在一个月中所获捐款每日情况上，Sullivan Dan 较为占优，但数据仅局限于一个月，并不能通过局部结论得到未来的全局结论，因此在未来可以通过过往数据基础上不断添加新数据，进行更为普遍的分析；

(6) Sullivan Dan 在所获州的支持上，分布较为平均，并非集中于某个或某几个州，且他并不出自，因此认为他的群众基础较好，受选民喜好度较高；

(7) 通过信息熵计算两位候选人受选民喜好的概率，Sullivan Dan 以 3%的优势略高于 Demings Val；

综合上述分析，Sullivan Dan 在本数据中是较受选民青睐的总统候选人，同时 Demings Val 也具备一定的潜力。

由于选民喜好程度可以由捐款数额和人数决定，因此未来工作可以聚焦于热门候选人在较长时间段内获得捐款的连续数据进行分析，利用回归算法进行拟合[5,6]，预测胜选概率。

参考文献

[1] 穆若曦.极端对立的政党与日趋分裂的社会--论政治极化下美国选民的政党认同[J].当代世界与社会主义,2022(03):141-149.

[2] 朱金冲.基于神经网络的属性关联建模与缺失值填补[D].大连理工大学,2021.

[3] 毛丽雯.基于置信度的不完整数据建模与缺失值填补[D].大连理工大学,2021.

[4] 杜秀丽,姜晓虎,孙晨瞳,于正.基于方向性多重假设检验和信息熵的函数型数据聚类新方法[J/OL].南京师大学报(自然科学版):1-9[2022-11-08].

[5] C.Cortes,V.Vapnik.Support vector networks[J].Machine Learning,1995,20(3):273-297.

[6] 陈伟杰.最小二乘法原理及其在实验曲线拟合中的应用分析[J].辽宁科技学院学报,2014,16(04):33-34,37.