**13.5 2012 Federal Election Commission Database**

**1. Giới thiệu dữ liệu FEC**

Phần này giới thiệu tập dữ liệu từ Ủy ban Bầu cử Liên bang Hoa Kỳ (Federal Election Commission – FEC), ghi lại toàn bộ các khoản quyên góp cá nhân cho chiến dịch tranh cử Tổng thống Mỹ năm 2012. Dữ liệu được lưu trong tệp CSV có kích thước lớn, chứa hơn một triệu bản ghi, phản ánh chi tiết tên người đóng góp, nghề nghiệp, nhà tuyển dụng, bang sinh sống, ứng cử viên nhận tiền và số tiền quyên góp. Mục tiêu của tác giả là hướng dẫn người đọc cách xử lý và phân tích một tập dữ liệu thực tế, quy mô lớn bằng pandas – từ việc đọc, lọc, đến trực quan hóa và rút ra thông tin ý nghĩa.

**2. Đọc dữ liệu và khám phá ban đầu**

fec = pd.read\_csv("datasets/fec/P00000001-ALL.csv", low\_memory=False)

print("=== 2. Đọc dữ liệu và khám phá ban đầu ===")

print(fec.info())

Sau khi tải dữ liệu, tác giả tiến hành đọc tệp CSV bằng lệnh pd.read\_csv() với tham số low\_memory=False để tránh cảnh báo do kích thước lớn. Tiếp theo, ông kiểm tra cấu trúc dữ liệu thông qua phương thức info(), giúp xác định số lượng cột, kiểu dữ liệu và giá trị thiếu. Một số cột được quan tâm nhất là cand\_nm (tên ứng cử viên), contbr\_nm (tên người quyên góp), contbr\_occupation (nghề nghiệp), contbr\_employer (nhà tuyển dụng), contb\_receipt\_amt (số tiền quyên góp) và contbr\_st (bang cư trú). Việc khám phá này là bước đầu tiên quan trọng trong quy trình phân tích dữ liệu, giúp người học hiểu rõ cấu trúc và nội dung của tập dữ liệu thực tế.

**3. Lọc dữ liệu theo hai ứng cử viên chính**

fec = fec[fec["contb\_receipt\_amt"] > 0]

fec = fec[fec["cand\_nm"].isin(["Obama, Barack", "Romney, Mitt"])]

print(fec["cand\_nm"].value\_counts())

Tác giả nhận thấy tập dữ liệu chứa thông tin về nhiều ứng cử viên khác nhau, nhưng để tập trung phân tích sâu, ông chỉ chọn hai ứng cử viên nổi bật nhất của cuộc bầu cử năm 2012 là Barack Obama và Mitt Romney. Dữ liệu được lọc bằng cách sử dụng phương thức isin() trên cột cand\_nm và chỉ giữ lại các khoản quyên góp có giá trị dương (contb\_receipt\_amt > 0). Bước lọc này giúp dữ liệu trở nên gọn gàng, dễ so sánh và tránh nhiễu từ các khoản âm (hoàn tiền hoặc điều chỉnh sai).

**4. Chuẩn hóa dữ liệu đảng phái**

parties = {"Obama, Barack": "Democrat", "Romney, Mitt": "Republican"}

fec["party"] = fec["cand\_nm"].map(parties)

print(fec[["cand\_nm", "party"]].drop\_duplicates())

Để phục vụ cho các bước phân tích sau, tác giả tạo thêm một cột mới có tên là party, biểu thị đảng phái của ứng cử viên. Thông qua ánh xạ (map), Barack Obama được gán vào nhóm “Democrat” (Dân chủ), còn Mitt Romney thuộc “Republican” (Cộng hòa). Việc bổ sung thông tin này giúp dễ dàng thực hiện so sánh theo đảng và trực quan hóa dữ liệu một cách trực quan, đồng thời thể hiện cách dùng pandas để thêm thuộc tính mô tả có ý nghĩa vào tập dữ liệu.

**5. Phân tích quyên góp theo nghề nghiệp**

def get\_top\_amounts(group, key, n=5):

totals = group.groupby(key)["contb\_receipt\_amt"].sum()

return totals.nlargest(n)

grouped = fec.groupby("cand\_nm")

print(grouped.apply(get\_top\_amounts, "contbr\_occupation", n=7))

A graph with numbers and lines

AI-generated content may be incorrect.

Trong phần này, tác giả tập trung vào cột contbr\_occupation để tìm hiểu những nhóm nghề nghiệp nào quyên góp nhiều nhất cho mỗi ứng cử viên. Sử dụng groupby() kết hợp với một hàm tổng hợp (apply()), ông tính tổng số tiền quyên góp theo từng nghề nghiệp. Kết quả cho thấy những người làm trong lĩnh vực luật (attorneys) đóng góp nhiều cho Barack Obama, trong khi nhóm người nghỉ hưu (retired) lại là nguồn quyên góp lớn cho Mitt Romney. Phân tích này cho thấy mối liên hệ thú vị giữa nghề nghiệp và xu hướng chính trị.

**6. Phân tích quyên góp theo nhà tuyển dụng**

print(grouped.apply(get\_top\_amounts, "contbr\_employer", n=10))

Tương tự với nghề nghiệp, tác giả tiếp tục nhóm dữ liệu theo cột contbr\_employer để xác định các tổ chức, doanh nghiệp hoặc cơ quan có nhân viên quyên góp nhiều cho từng ứng cử viên. Kết quả cho thấy những công ty công nghệ lớn như Google hay Microsoft có nhiều nhân viên quyên góp cho Obama, trong khi các tập đoàn tài chính như Goldman Sachs lại ủng hộ Romney nhiều hơn. Qua đó, ta có thể thấy xu hướng khác biệt giữa các ngành nghề và lĩnh vực kinh tế trong việc ủng hộ chính trị.

**7. Phân tích theo mức đóng góp**

fec\_mrbo = fec.copy()

bins = np.array([0, 1, 10, 100, 1000, 10000, 100000, 1000000, 10000000])

labels = pd.cut(fec\_mrbo["contb\_receipt\_amt"], bins)

grouped = fec\_mrbo.groupby(["cand\_nm", labels])

bucket\_sums = grouped["contb\_receipt\_amt"].sum().unstack(level=0)

normed\_sums = bucket\_sums.div(bucket\_sums.sum(axis="columns"), axis="index")

print(normed\_sums.head())

normed\_sums[:-2].plot(kind="barh", figsize=(8, 5))

plt.title("Phân bố tỉ lệ quyên góp theo mức đóng góp")

plt.xlabel("Tỉ lệ (%)")

plt.ylabel("Khoảng giá trị đóng góp (USD)")

plt.show()

A graph with blue and orange bars

AI-generated content may be incorrect.

Tác giả chia giá trị quyên góp thành các khoảng (bins) khác nhau, từ 0 đến trên 1 triệu đô la, sau đó sử dụng pd.cut() để phân loại từng khoản quyên góp vào nhóm tương ứng. Bằng cách nhóm theo ứng cử viên và khoảng giá trị, ông tính tổng số tiền trong mỗi nhóm và biểu diễn kết quả bằng biểu đồ thanh ngang. Kết quả cho thấy Barack Obama nhận được phần lớn các khoản quyên góp nhỏ (dưới 100 USD), trong khi Mitt Romney nhận được nhiều quyên góp lớn (trên 1.000 USD). Sự khác biệt này phản ánh chiến lược vận động và cơ cấu người ủng hộ của hai chiến dịch.

**8. Phân tích theo bang**

grouped\_state = fec.groupby(["cand\_nm", "contbr\_st"])["contb\_receipt\_amt"].sum().unstack(level=0).fillna(0)

print(grouped\_state.head())

Phần cuối cùng của phân tích tập trung vào sự phân bố quyên góp theo từng bang (contbr\_st). Bằng cách nhóm dữ liệu theo ứng cử viên và bang, tác giả tính tổng số tiền đóng góp ở mỗi khu vực. Kết quả cho thấy Obama nhận được nhiều sự ủng hộ tài chính từ các bang ven biển và thành thị như California, New York, trong khi Romney có thế mạnh ở các bang miền Trung và miền Nam, nơi cử tri Cộng hòa chiếm ưu thế. Phân tích này cung cấp góc nhìn địa lý rõ rệt về sự khác biệt trong mạng lưới ủng hộ chính trị.