Information Visualization

CHECKPOINT II: Data cleaning and processing

G14-A

1. Initial Dataset

Our initial dataset consisted of data from 6 different data sources, as described in Checkpoint I. The first dataset (database.csv), a table with 9965 lines, contains Oscar winners and nominees from 1927 to 2015. The second dataset (us_presidents.csv), a table with 46 lines, contains information about each presidential term, including the start and end dates and the political party of the president. The third dataset (Oscars-demographics-DFE.csv), a table with 442 lines, contains demographic informations about Oscars winners, such as their ethnicity and birthplace. The following datasets were webscrapped from the pages mentioned in checkpoint I. The fourth dataset (ethnicityActorsAndDirectors.csv), a table with 2104 rows, contains the ethnicity for all winners and nominees for performance awards and Best Director Award. The fifth dataset (black_milestones.csv), a table with 40 lines, contains black milestones and conflicts per year. The sixth dataset (new_lgbt.csv), a table with 253 lines, contains LGBT nominees per year.

Here is a data sample of the first dataset (database.csv):

Year, Ceremony, Award, Winner, Name, Film
1927/1928,1,Actor,null,Richard Barthelmess,The Noose

2. Selected/Derived Data

We have one csv file to answer each of our 6 questions. For our **1st question**, we select "orientation" from dataset 6 and "party" from dataset 2. We have a derived measure, "count_orientation", which is the number of people with that sexual orientation nomineed while that political party was in charge. For our **2nd question**, we select "birthplace" from dataset 3 and "party" from dataset 2. We have a derived measure, "count_birthplace", which is the number of people who were born in that place nomineed while that political party was in charge. For our **3rd question**, we select "Year" and "Winner" from dataset 1 and "race_ethnicity" from dataset 4. We have a derived measure, "ratio", which is the percentage of people with that ethnicity who were nominated or won in that year. For our **4th question**, we select "race_ethnicity" from dataset 4, "party" from dataset 2, and "Winner" from dataset 1. We have a derived measure, "count_ethnicity", which is the number of people with that ethnicity who won (Winner=1) or were nominated (Winner=0) while that political party was in charge. For our **5th question**, we select "Year" from dataset 1, "Winner" from dataset 1, "race_ethnicity" from dataset 4 and "milestone" from dataset 5. We also have "sentiment", which was filled by hand. We have a derived measure, "ratio", which counts the percentage of people with that ethnicity who won (Winner=1) or were nominated (Winner=0) in that year. For our **6th question**, we select "Year" from dataset 1. We have another attribute, "Gender", which was filled by hand because we didn't find a dataset with it. We have a derived attribute, "ratio", which has the percentage of persons with that gender nominated for the award of Best Director in that year.

3. Data abstraction

When it comes to the types of the generated datasets all the 6 of them are multidimensional tables. The following attributes will be used, some in more than one table: "**orientation**": Nominal variable (Sexual orientation of the nominees); "**party**": Nominal variable (Political party in power

in US); "**count_orientation**": Continuous variable (Number of nominees with that sexual orientation); "**birthplace**": Nominal variable (Country where winners were born); "**count_birthplace**": Continuous variable (Number of winners born in that country when that political party was in power); "**count_ethnicity**": Continuous variable (Number of candidates with that ethnicity when that political party was in power); "**year**": Sequential variable (Year of the Oscars ceremony); "**winner**": Nominal variable (Tells whether the ratio is about Oscars winners or just nominees); "**race_ethnicity**": Nominal variable. (Ethnicity of the Oscars candidates); "**ratio":** Ratio variable (Ratio for an ethnicity of Nominees/Winners); "**gender**": Nominal variable (gender of the Oscar nominees); "**milestone**": Nominal variable (Events related with racism); "**sentiment**": Nominal variable (Tells whether milestone is positive or negative). The ratio attribute is a sequential scale attribute.

4. Dataset processing

To clean the datasets, we used PDI transforms, that are: removing columns, sort the dataset by a column, lowercasing column names, join datasets through common columns, created new datasets, added sentinel values to empty cells.

The datasets for questions which are not stated below were obtained by joining the datasets mentioned in 3, without problems like missing values. For **Q1**, we joined datasets 2 and 5 by year. First, dataset 2 was processed so that the attribute "start" was just the year instead of the whole date. A left outer join was used, with dataset 5 on the left and dataset 2 on the right, which filled the years with the start date. The remaining years were hand filled using the party in the start date. For **Q3**, we joined datasets 1 and 4 by year, and grouped by year, winner and race_ethnicity. There was a problem with the year: until 1934, the "Year" column of database.csv had two years, for example, 1928/1929, which was fixed by hand. Regarding the "Winner" attribute, it initially had value 1 if the person won the award, otherwise it was null. This was solved using the "If field value is null" step from PDI, replacing the null values by 0, then proceeded to calculate percentages for the ratio. For Q5, we joined dataset 1, 4 and 5 by year. Then added a new column, create a derived measure "ratio" that shows the percentage of winners and nominees by ethnicity, besides that created column "Sentiment" associated to the column "milestone to describe if the black history milestone was positive and negative, also added a sentinel value (-1) to the year where no milestone (and sentiment) occurred.

5. Mapping (Data sample / Questions)

Q3: How does the ethnicity of Oscar winners and nominees change over time?

Year, Winner, race_ethnicity, ratio

1972, 0, Black, 15.0

This sample from Q3.csv allows to see how the percentage of non-white nominees/winners changes with time

Q5: How does the ethnicity of winners and nominees change with racial conflicts?

Race, ratio, winner, year, milestone, sentiment

white,90.0, 0, 1954,Brown v. Board Of Education, May 17, 1954, Acceptance

black,5.0, 0, 1954,Brown v. Board Of Education, May 17, 1954, Acceptance

hispa,5.0, 0, 1954,Brown v. Board Of Education, May 17, 1954, Acceptance

white,100.0, 1, 1954,Brown v. Board Of Education, May 17, 1954, Acceptance

This sample from Q5.csv allows to see that in a year where there was a black acceptance milestone, there were 5% nominees who were black but didn't win.