Who’s Running for Mayor?

Understanding Voter Search Behavior on Google

Annabel Uhlman ‘22, Data Science Major Capstone

WELLESLEY

Objectives & Research:

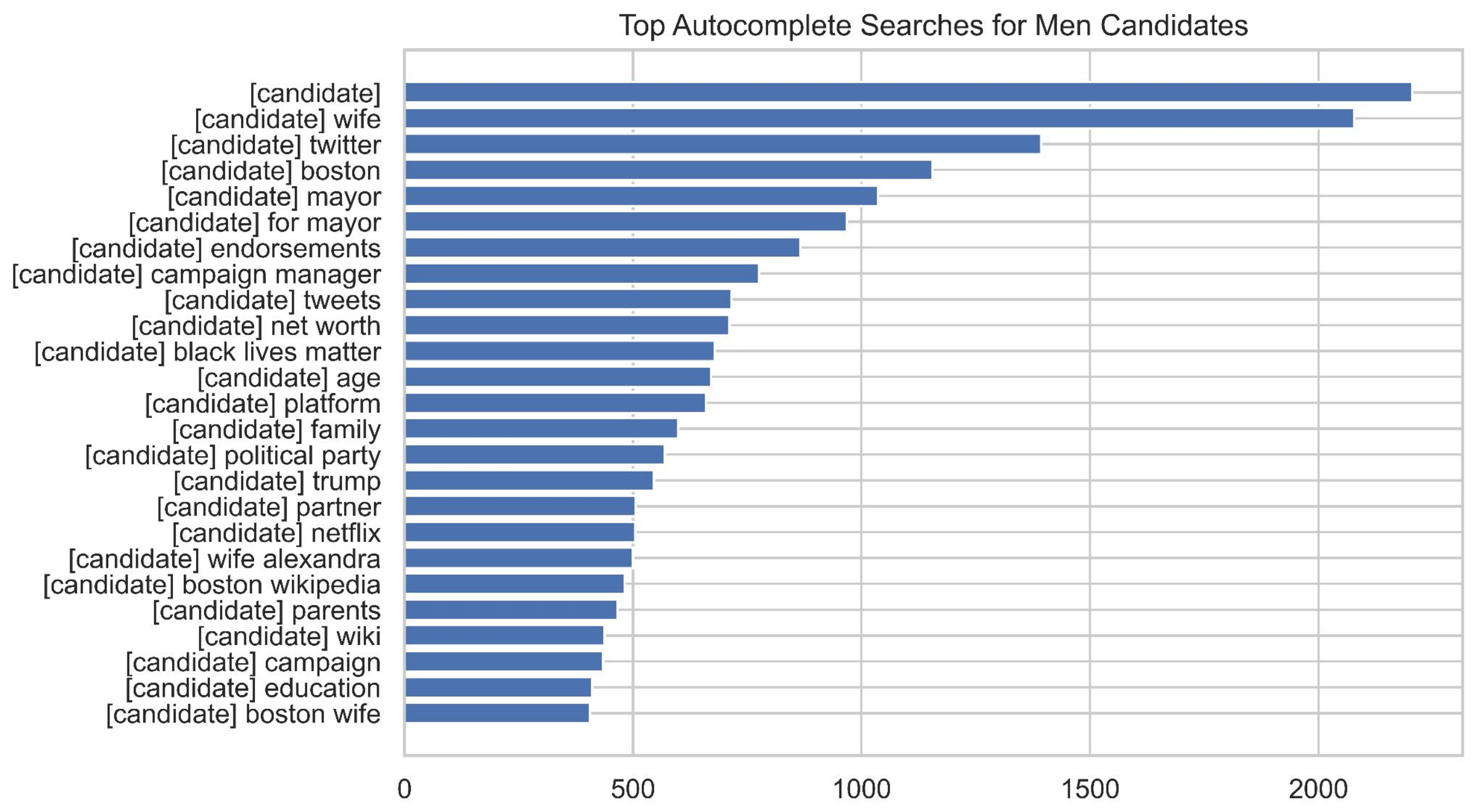
Google is one of the top sources of election news for voters and since political science literature has shown women face a disadvantage running in political elections against men, understanding what information voters look for regarding candidates is important. This is especially true since most research has focused on national elections.

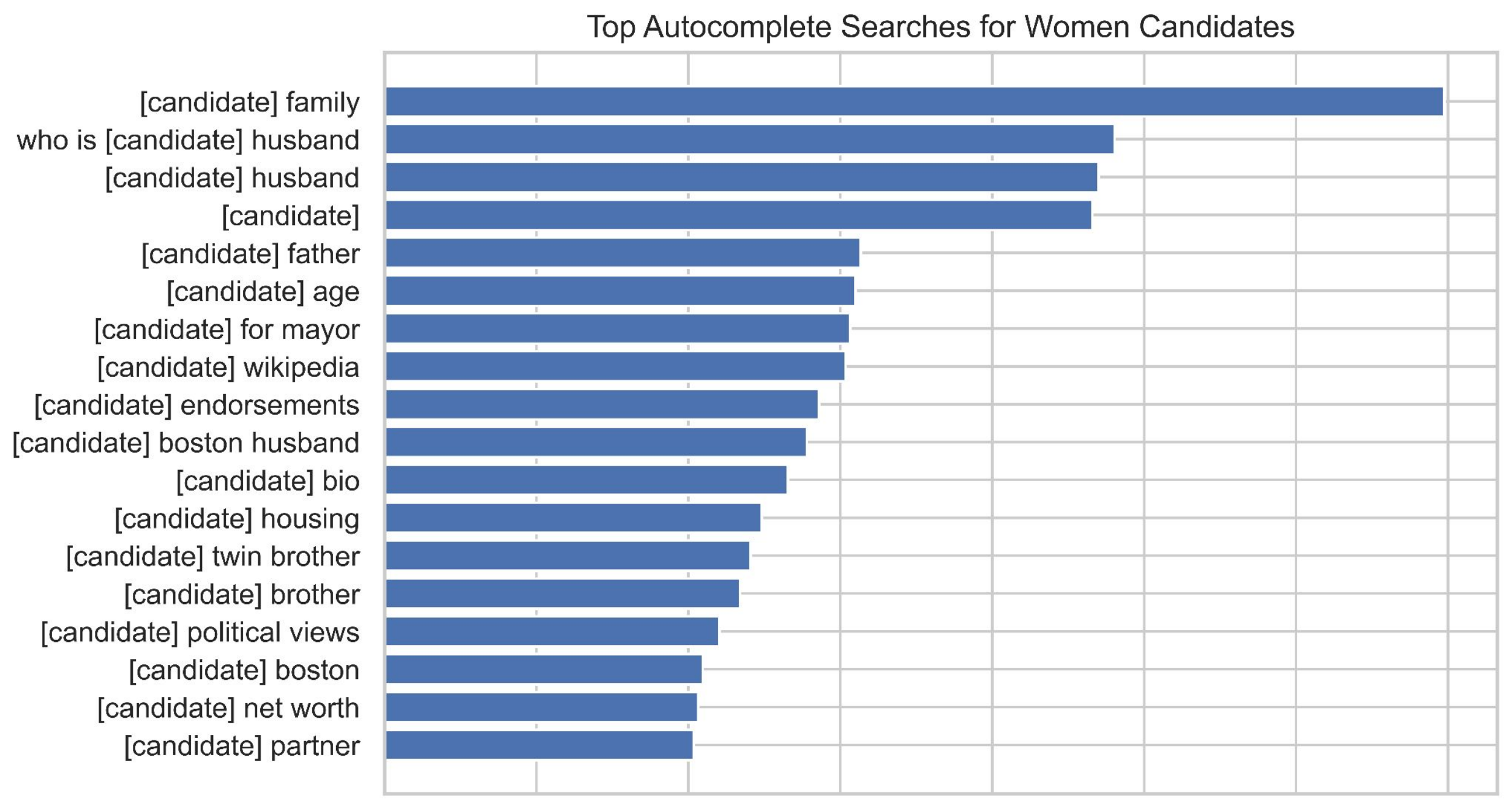
Question:

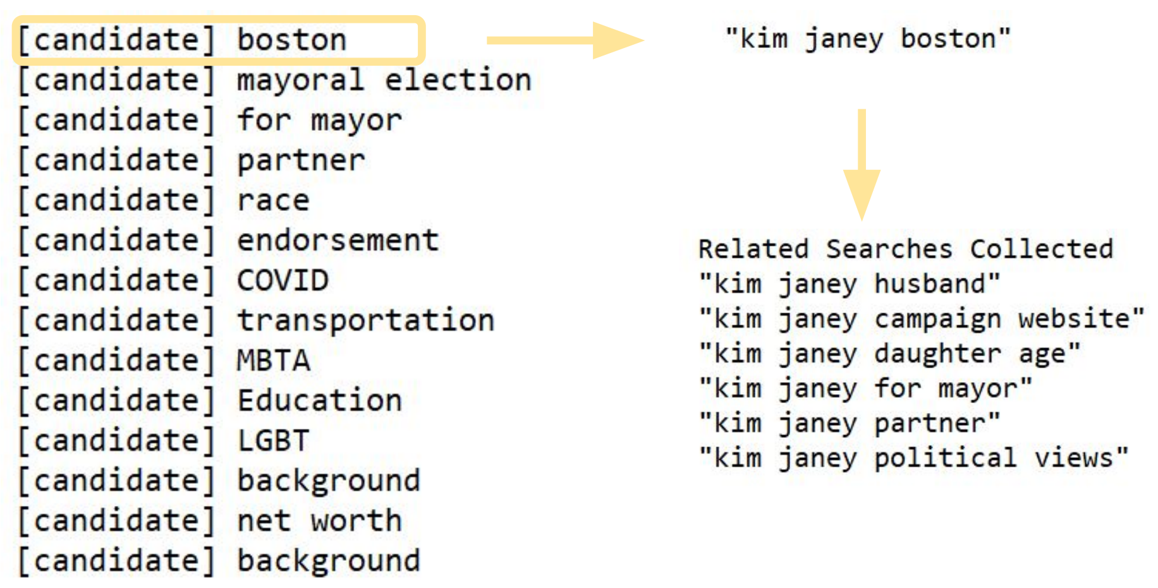
To what extent does Google suggest more gendered autocomplete and related results about women candidates?

Background:

Two mayoral elections in Boston and NYC in 2021 with diverse candidate pools (22 candidates: 7 women, 13 men, 1 nonbinary candidate; 8 white candidates, 14 POC)







```
[candidate] boston
[candidate] mayoral election
[candidate] for mayor
[candidate] partner
[candidate] race
[candidate] endorsement
[candidate] COVID
[candidate] transportation
[candidate] MBTA
[candidate] Education
[candidate] LGBT
[candidate] background
[candidate] net worth
[candidate] background
```

```
"kim janey boston"

Related Searches Collected
"kim janey husband"
"kim janey campaign website"
"kim janey daughter age"
"kim janey for mayor"
"kim janey partner"
"kim janey political views"
```

Example of Seed Query to Search Process

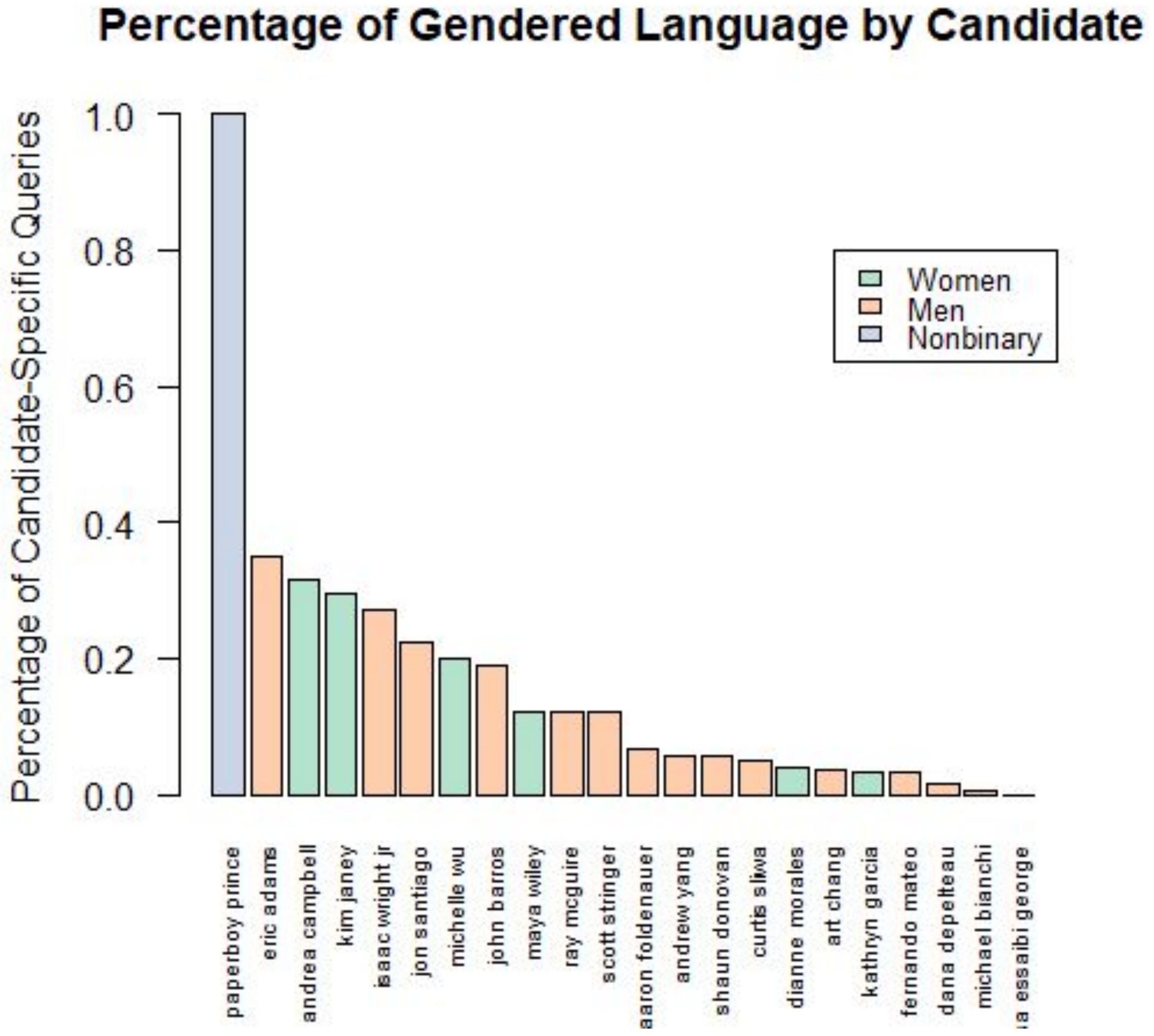
Methodology:

Data Collection

From May to November we searched a set of seed queries daily and collected autocomplete & related searches, totaling about 150 queries a day. We downloaded Search Engine Result Pages for all queries. We had a total of 188,772 search results (122,057 from Boston, 66,715 from NYC)

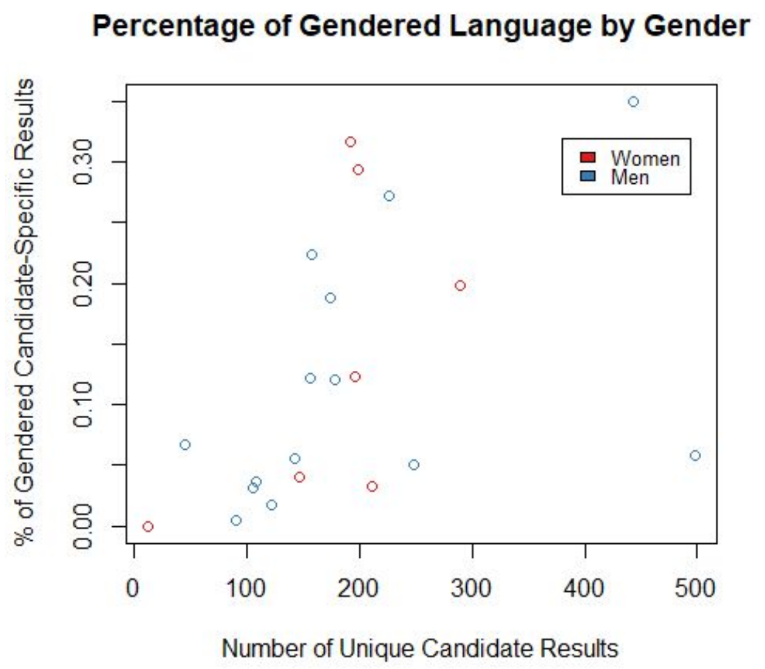
Analysis

- Term-Frequency Inverse Document Frequency (TF-IDF) Analysis to identify which words were relatively more frequent in search results.
- Jaccard Similarity Index to assess changes in search result corpus over election span
- Binomial Logistic Regression model to see if the percentage of gendered language in search results about a candidate is predictable.



TF-IDF Analysis:

- Created a gendered lexicon to identify gendered search results. There is a higher percentage of gendered language in results about men than about women (16.75% of queries relating to men had gendered language, where 26.25% of those about women did)
- The top 20 unigrams with the highest TF-IDF scores showed gendered language in the top 2-3 positions for women candidates in both cities
- Most popular searches have some gendered words (family, husband, father,
 etc.) in the top searches for women, but not as high for men



Binomial Logistic Regression Analysis:

- To fit assumptions, had to exclude outlier of nonbinary candidate (# candidates=21)
- Created a dataset with one row per candidate
- Compared two models aiming to predict the percentage of gendered queries for each candidate. Both models’ variables included i) city the candidate ran in; ii) the number of total search results about them; iii) the number of unique search results about them
  - ➢ Model A included gender as a variable while model B did not in aims of assessing if gender is an important predictor
  - ➢ Model A AIC: 7115.22; Model B AIC: 7157.382
  - ➢ Model A BIC: 7121.487; Model B BIC: 7162.604
  - ➢ P-value of gender variable was significant $p<0.001$
  - ➢ City was not significant

Discussion:

- Multiple levels of analysis showed difference in gendered queries between women and men
- Difficult to generalize beyond this dataset; limited sample size
- Future research to include race, and analyze Search Engine Result Pages (SERPs)