University of Toronto School of Continuing Studies
SCS 3252 Big Data Management Systems & Tools
Term Project

2019 Canadian Federal Election

CALCULATING THE RESULTS USING APACHE SPARK

ZLATA IZVALAVA

Project

- The 2019 Canadian federal election was held on October 21, 2019, to elect members of the House of Commons to the 43rd Canadian Parliament.
- The objective of this project is to process the validated results of the 2019 Federal Election using Apache Spark and explore the data.
- Data processing was done in a Databricks notebook (Scala)

Dataset

October 21, 2019 Canadian Federal Election: The latest results for all electoral districts

https://enr.elections.ca/DownloadResults.aspx

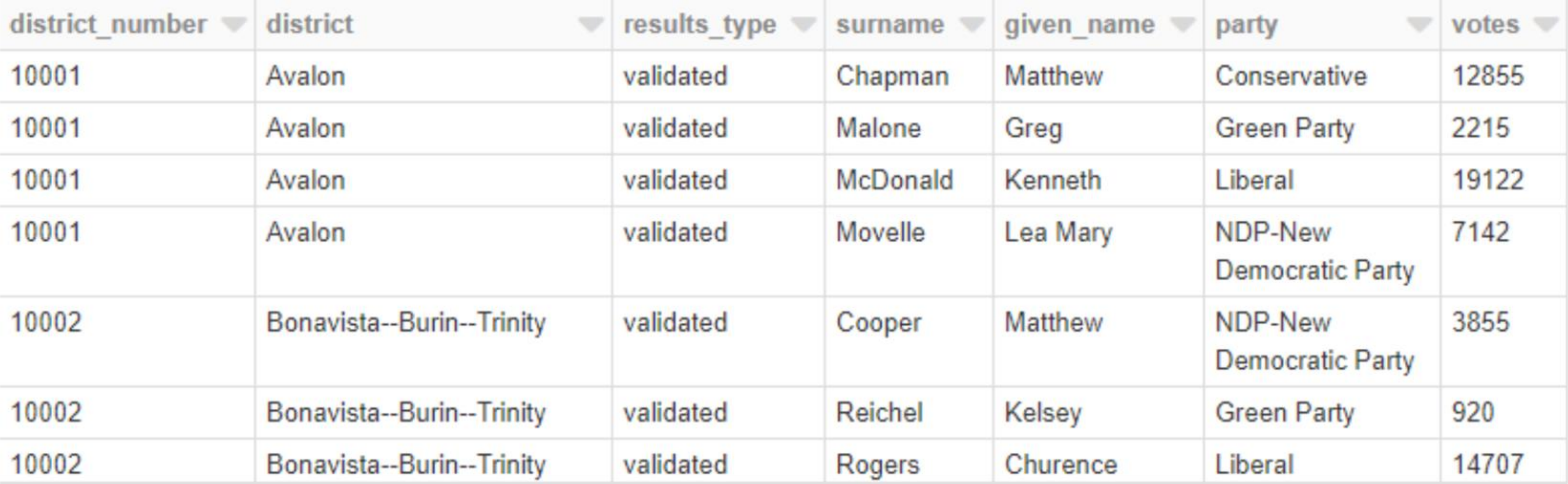
- The file contains 4292 rows of data with results for all candidates in each federal electoral district
- There are 338 federal electoral districts in Canada
- The dataset includes both preliminary and validated results

Approach

- Read the data from txt file (tab-delimited format) into a DataFrame
  - using sqlContext.read.format("com.databricks.spark.csv")
- Data preparation
  - Select rows with validated results only
  - ‘Middle name’ column contains null values
  - Remove the column with middle names and some other columns
  - Rename the rest of the columns (“withColumnRenamed” function)
- Explore the popular vote
- Determine the candidates and parties that won the seats in the House of Commons
- Find seat breakdown by province/territory

Data Sample

The DataFrame after pre-processing:



district_number	district	results_type	surname	given_name	party	votes
10001	Avalon	validated	Chapman	Matthew	Conservative	12855
10001	Avalon	validated	Malone	Greg	Green Party	2215
10001	Avalon	validated	McDonald	Kenneth	Liberal	19122
10001	Avalon	validated	Movelle	Lea Mary	NDP-New Democratic Party	7142
10002	Bonavista--Burin--Trinity	validated	Cooper	Matthew	NDP-New Democratic Party	3855
10002	Bonavista--Burin--Trinity	validated	Reichel	Kelsey	Green Party	920
10002	Bonavista--Burin--Trinity	validated	Rogers	Churence	Liberal	14707

Popular Vote

Total amount of validated votes = 18,171,636

Conservative	6239510
Liberal	6019097
NDP-New Democratic Party	2903789
Bloc Québécois	1387030
Green Party	1189631
People's Party	294104
Independent	72547
Christian Heritage Party	18901

Popular Vote

Conservative: 34.3%

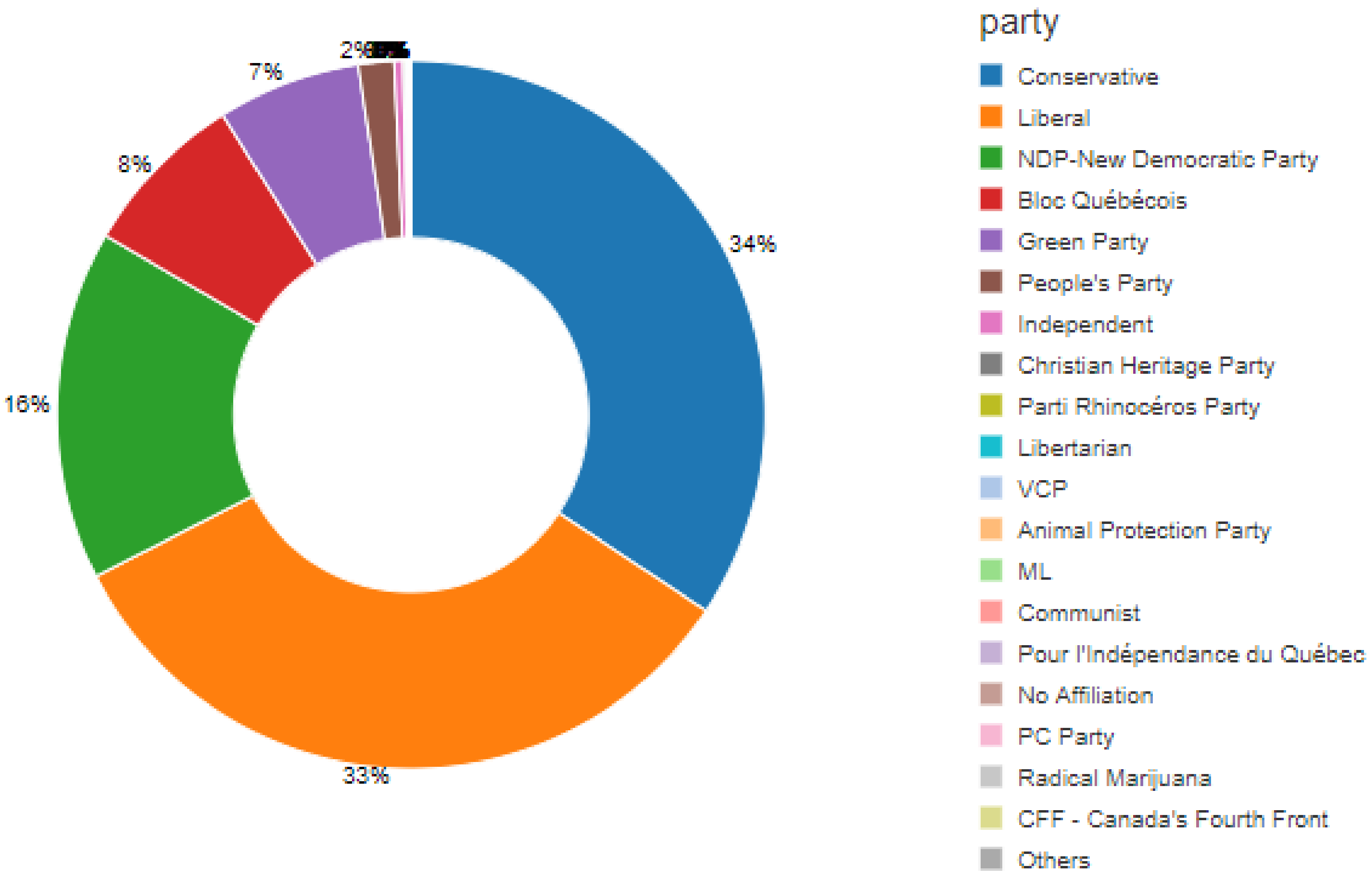
Liberal: 33.1%

NDP: 16%

Bloc Québécois: 7.6%

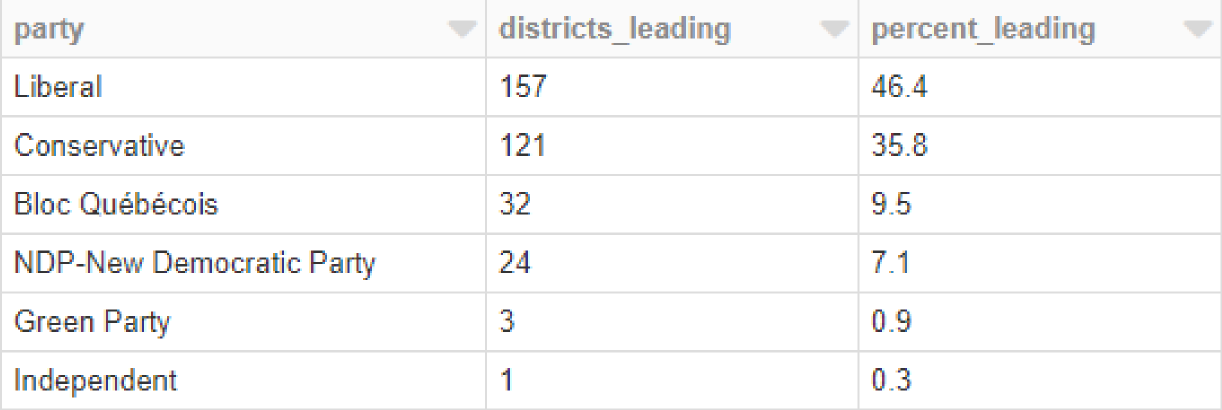
Green Party: 6.5%

People’s Party: 1.6%



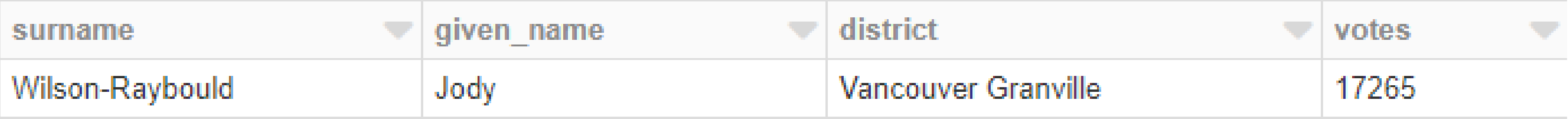
Seats Breakdown

- Find the candidate who received the most votes in each electoral district:
  - Use a window function to partition the data by district number, order the candidates in each district by number of votes, then take the first row in each district
- For each party find the number of districts where it is leading (number of seats in Parliament)



party	districts_leading	percent_leading
Liberal	157	46.4
Conservative	121	35.8
Bloc Québécois	32	9.5
NDP-New Democratic Party	24	7.1
Green Party	3	0.9
Independent	1	0.3

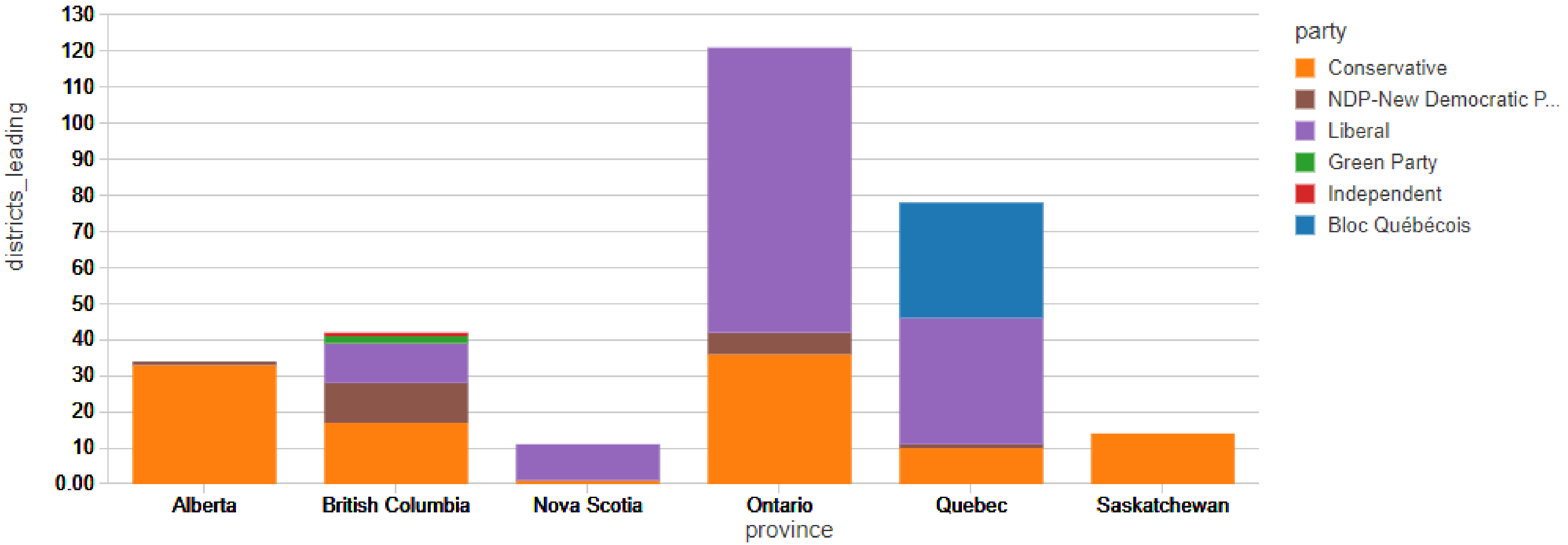
- Independent candidate:



surname	given_name	district	votes
Wilson-Raybould	Jody	Vancouver Granville	17265

Results for provinces

- Read another text file into a DataFrame (it contains province/territory names and district names)
- Join 2 tables
- Find how many seats different parties won by province/territory



THANK YOU!