Informasjon om eksamen

SOS2900 - Algoritmer, store data og samfunnsendring

- Skoleeksamen
- 23. april 2019, kl. 09:00-13:00 (4 timer)

Om eksamen

- Eksamen består av to oppgaver. Begge oppgaver må være besvart for å bestå eksamen.
- Oppgave 1 teller 50% av eksamenskarakteren og oppgave 2 teller 50% av eksamenskarakteren.
- Eksamensoppgaven er på norsk. Du kan besvare eksamenen på norsk, svensk, dansk eller engelsk.

Hjelpemidler

1. Ordbok som er levert og kontrollert i forkant av eksamen.
2. Alle R-script som er blitt brukt i seminarene

R-oppgaver

Instruksjon om datasett

Se eget vedlegg (til venstre) med beskrivelse av hvordan du laster ned datasettet og setter riktig filsti til der dataene ligger.

Datasettet du skal bruke laster du ned her: voters. Det er allerede i R-format og du trenger ikke gjøre noen endringer i datasettet før du gjør analysene.

Du åpner dataene med følgende kommando:

load("voters.RData")

- Inspera kan komme til å endre navnet på datasettet, så pass på å kopiere filnavnet (se instruksjoner i vedlagt pdf) og at filenavnet slutter på ".RData". (Hvis det mot formodning skulle oppstå problemer med å åpne datasettet ligger det også i csv-format her: voters).
- Til disse oppgaven vil du trenge følgende pakker: **randomForest** og **caret**. De er installert på maskinen, men må lastes med library-kommandoen.

Oppgave 1)

Oppgavene under (a-d) utgjør oppgave 1.

OBS! Ta vare på koden du har brukt til å besvare alle oppgavene nedenfor. Hele koden skal limes inn på neste side i oppgavesettet, som er en del av oppgave 1.

Dataene er fra en survey utført av YouGov rett etter presidentvalget i USA i 2016. Datasettet inneholder bl.a. informasjon om hvem de stemte på. Variabelen Clinton_supp er en factor-variabel med verdiene «Yes» hvis de stemte på Clinton og «No» hvis de stemte på noen andre (i tillegg til Trump var det også fire uavhengige kandidater). Full variabelliste finner du nederst i den vedlagte pdf-filen (til venstre) med instruksjoner om datasettet.

Se for deg at du jobber i en valgkamporganisasjon for neste valg. Du ønsker å ha to kampanjer: en rettet mot tidligere Clinton-stemmere og en annen for alle andre. Formålet kan f.eks. være å øke/redusere valgdeltakelse eller å overtale dem til å bytte politisk side.

Fremtidig datainnhenting vil hente inn informasjon du kan bruke til å klassifisere folk, men først trenger du å bygge en klassifikasjonsmodell for å vite hvem som skal få hvilket budskap.

a) Bruk random forest til gjøre klassifikasjonen med bruk av samtlige variable i datasettet, uten å tune på noe vis. Lag en confusion matrix. Hvor mange klassifiserer du som:

1. Riktig som Clinton-tilhengere

2. Riktig som ikke-Clinton tilhengere
3. Feil som Clinton-tilhengere
4. Feil som ikke Clinton-tilhengere

Skriv ditt svar her...

b) Det kan være kontraproduktivt hvis budskapet rettet mot Clinton-tilhengere mottas av andre grupper, men motsatt antas det å ikke ha noen effekt. Du ønsker derfor å minimere antall «false positive», mens mange «falske negative» ikke er så farlig.

Bruk tuningparametrene sampsize, mtry, og/eller ntree og se om du klarer forbedre modellen til dette formålet.
Oppgi hvilke tuning parametre du brukte for de modellene du vurderer som ga best og dårligst prediksjon.
Bruk den beste modellen til å lage en confusion matrix, og beskriv cost-ratio for denne med egne ord.

Skriv ditt svar her

c) Lag et variable importance plot. Hvilke variable har størst og minst betydning for prediksjonen?

Skriv ditt svar her

d) Variabelen **imiss_I_2016** angir hvor viktig politisk spørsmål man mener klimaendringer er. (1 er veldig viktig, 4 er lite viktig). Lag et partial dependence plot for variabelen, sett which.class= «Yes» for å få kurven til å gjenspeile sannsynligheten for å stemme Clinton.

Skriv ditt svar her

Maks poeng: 10

2 Lim inn koden

Kopier all kode fra R og lim inn her:

Skriv ditt svar her...

```
1
```

Maks poeng: 10

3 Langsvarsoppgave

Langsvarsoppgaven skrives som et essay.

Myndighetene ønsker å bekjempe kriminalitet og vurderer det som at forebygging er langt mer effektivt enn strafferettslige tiltak. I fagmiljøer er man enige om at tidlig intervensjon er det mest effektive, det vil si å sette i verk tiltak så tidlig i barndommen som mulig.

Det foreslås derfor å bygge en prediksjonsmodell som forutser i voksen alder basert på egenskaper ved foreldrene ved fødselstidspunktet. Med andre ord: blinke ut barn som er i høyrisikogruppe allerede ved spedbarnsalder. Hva slags algoritme som skal brukes spesifiseres ikke, men det overlates til eksperter i maskinlæring og kunstig intelligens å utvikle dette verktøyet. Data om foreldrene tenkes å samles inn fra administrative registre (utdanning, skatt osv), kundeforhold, helseregistre, og datatrafikk osv.

Oppgave:
Flere av pensumbidragene drøfter prinsipielle problemstillinger knyttet til hvordan bruk av slike algoritmer kan slå skjevt ut for ulike grupper som det bør tas eksplisitt stilling til. Ta utgangspunkt i eksempler eller prinsipielle diskusjoner fra pensum og vurder i hvilken grad tilsvarende vil gjelde i dette tilfellet. Hvilken betydning har utformingen av tiltak for hvordan prediksjonsmodellen lages? Foreslå to ganske forskjellige typer tiltak som i praksis ville blitt lagt under ulike sektorer (f.eks. justisfeltet, helsetjeneste eller barnevern). Vurder og begrunn hvordan cost-ratios bør settes.

Ordgrense: maks. 3000 ord (ca. 4 sider)

SOS2900 1 Algoritmer, store data og samfunnsendring

Skriv ditt svar her...

Format

Words: 0/3000

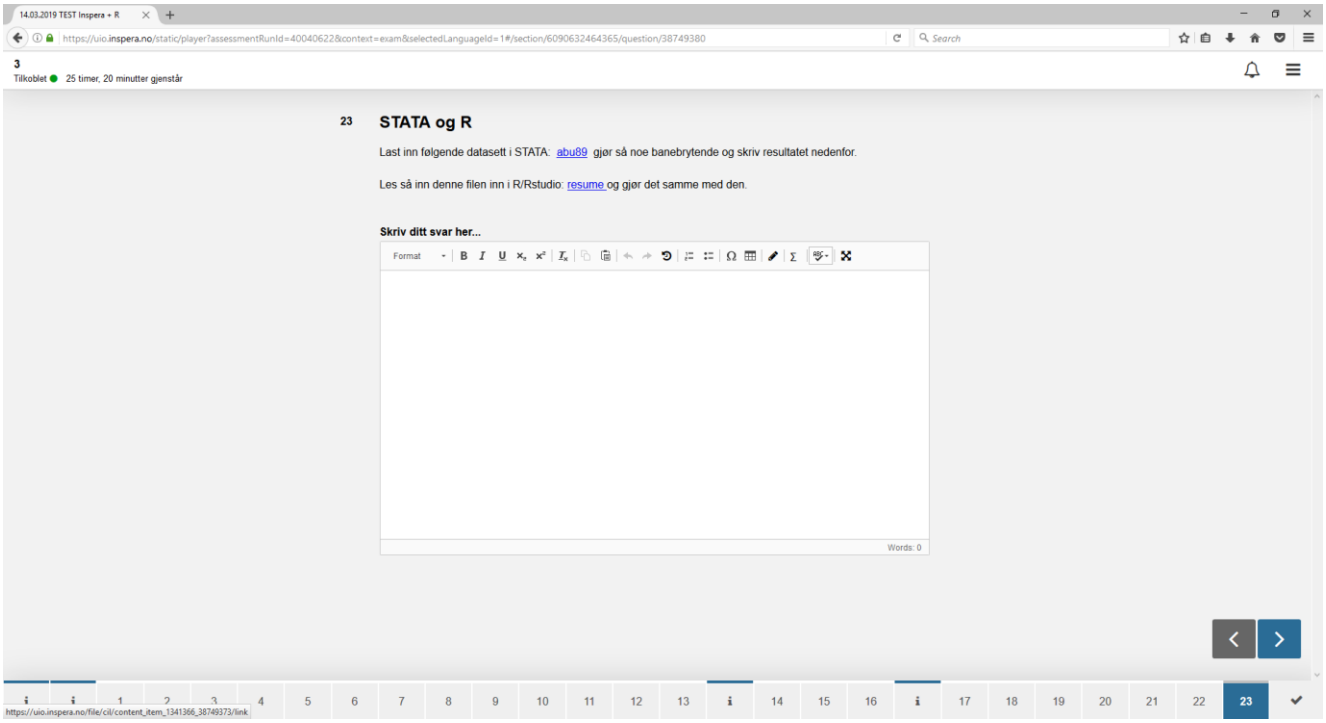
Maks poeng: 10

Question 1
Attached

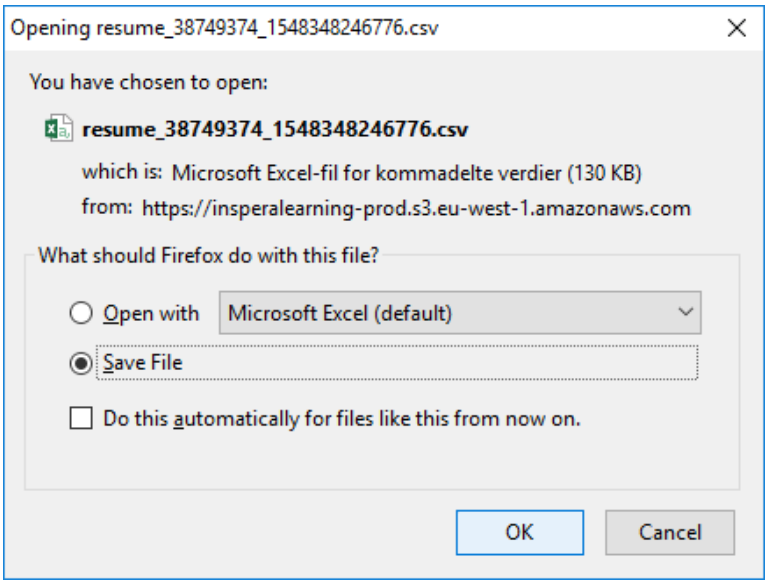
Hvordan åpne datasett i R og RStudio for eksamen i SOS2900

R og RStudio åpnes ved å dobbelklikke på ikonet på desktopen på maskinen. For å se desktopen må du minimere nettleseren Inspera er åpen.

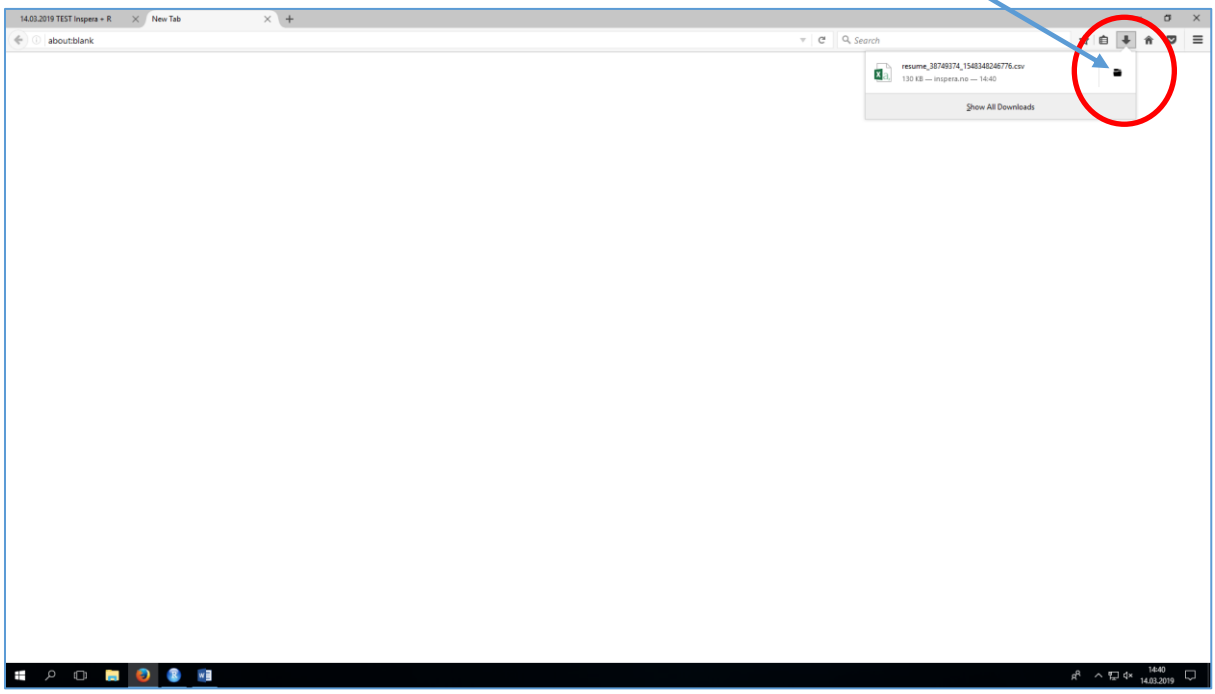
1) Du laster ned angitt datasettet fra Inspera. Trykk på lenken. Den er blå i oppgaveteksten.



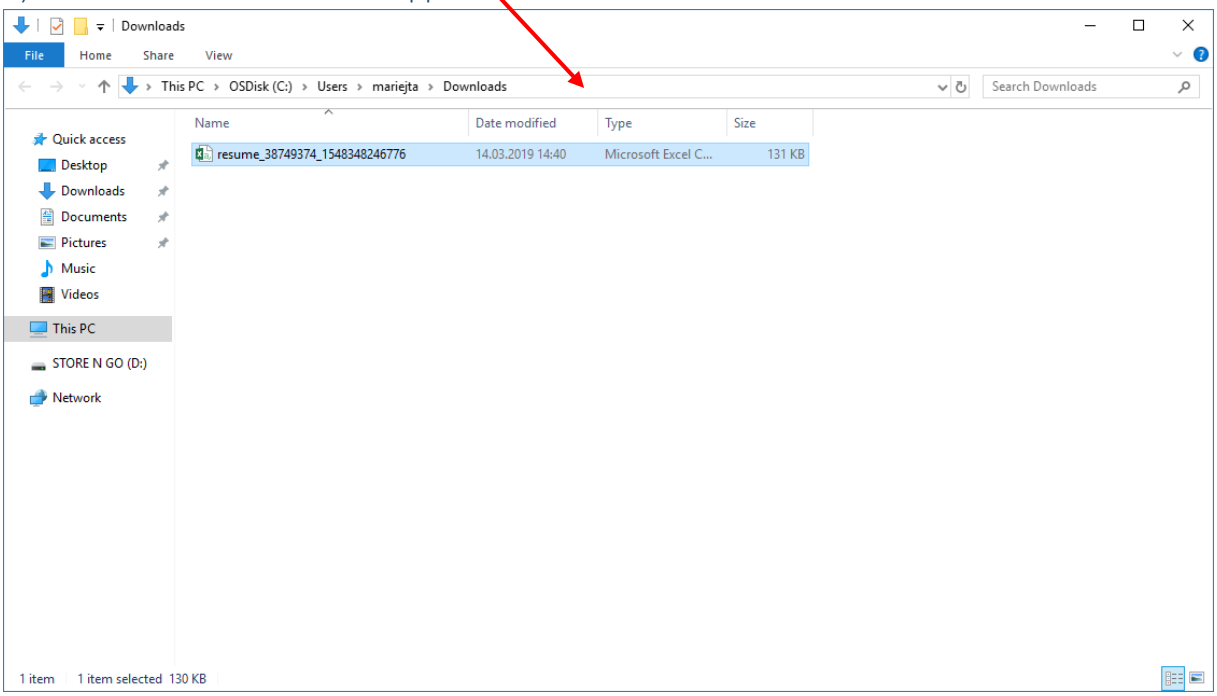
2) Velg ‘save file’. Filen lagres automatisk i mappen Downloads



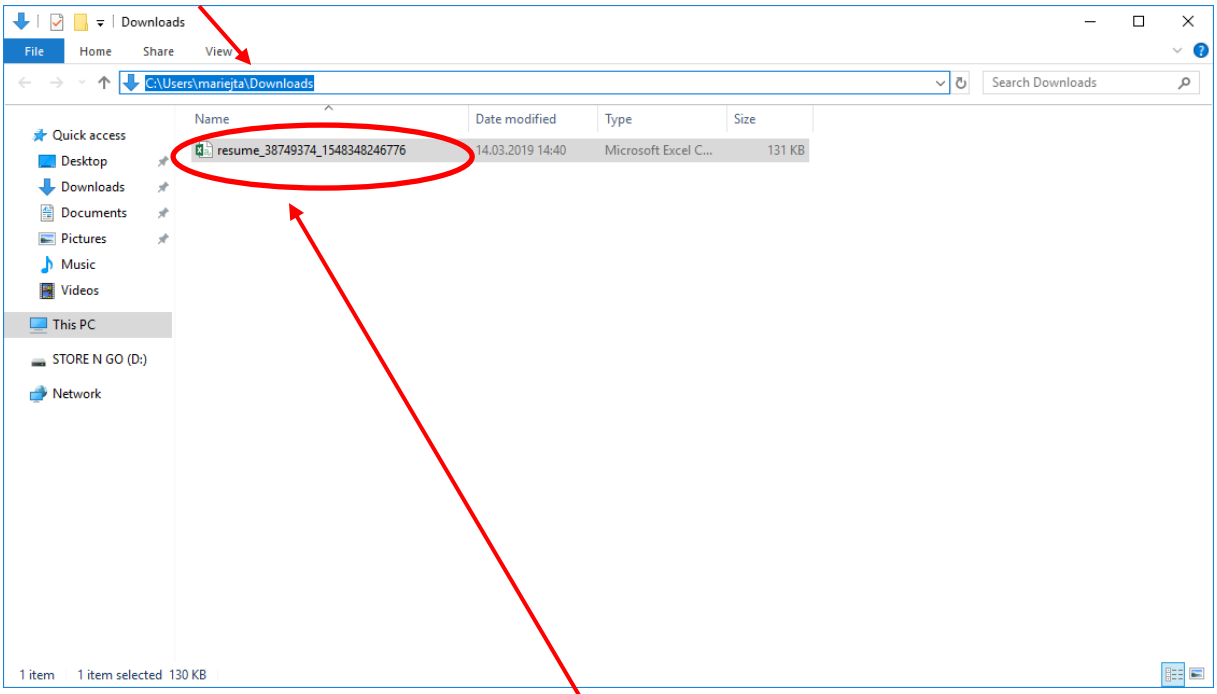
3) For å åpne filen, trykk på ↓ i nettleseren. Trykk så på filsymbolet for å åpne Downloads. Her ser du at filen er lagret.



4) Klikk i adressefeltet for å få opp filbanen

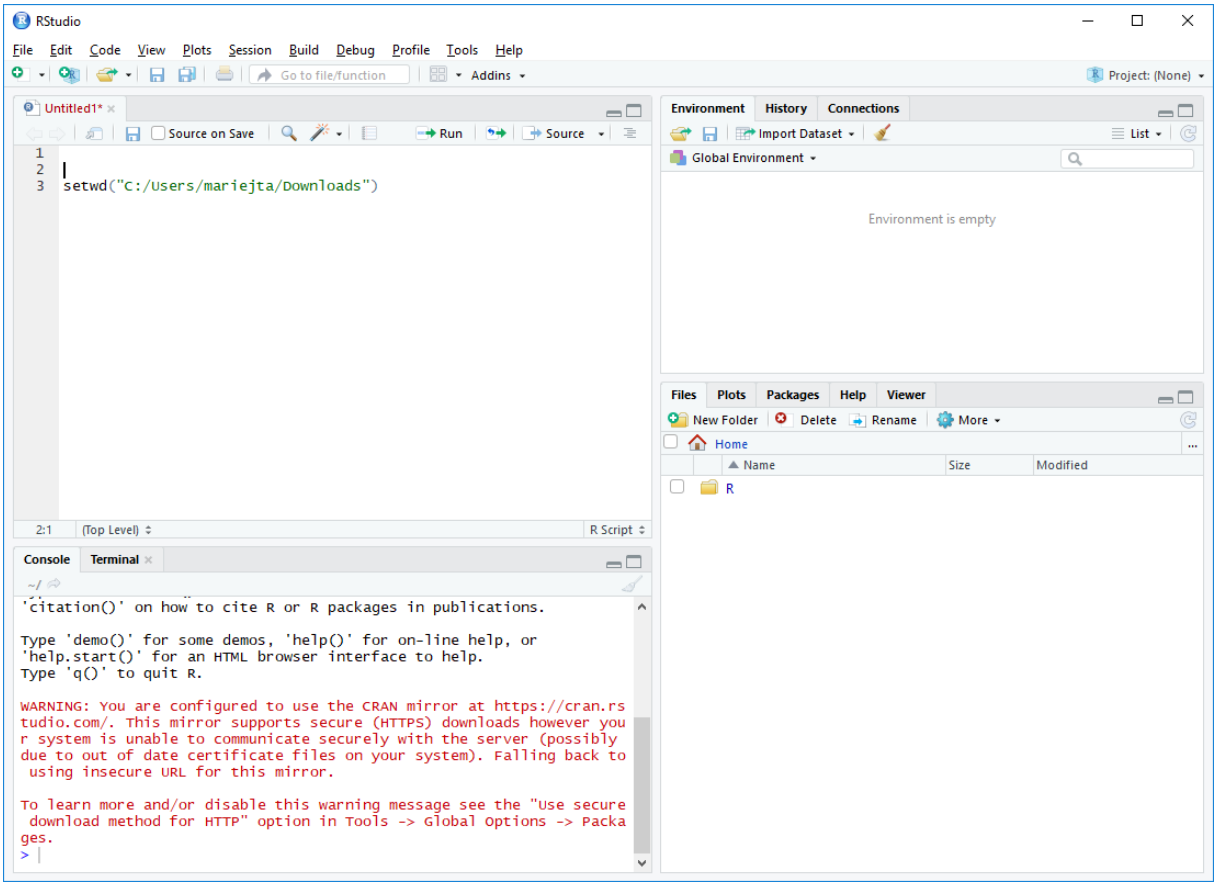


5) Kopier filstien (C:\Users\brukernavn\Downloads) over til R



6) Av rent tekniske grunner endrer Inspera filnavnet ved å legge til mange tall i navnet. Kopier filnavnet over til R.

7) Sett arbeidsområde med ved å lime inn filstien setwd-kommandoen som vist under. Husk å bytte ut '\' med '/'



Variabelliste for datasettet voters

RIGGED_SYSTEM_1_2016 Elections today don't matter; things stay the same
RIGGED_SYSTEM_2_2016 America is a fair society where everyone has the opportunity to get ahead
RIGGED_SYSTEM_3_2016 Our economic system is biased in favor of the wealthiest Americans
RIGGED_SYSTEM_4_2016 You can't believe much of what you hear from the mainstream media
RIGGED_SYSTEM_5_2016 People like me don't have any say in what the government does
RIGGED_SYSTEM_6_2016 Elites in this country don't understand the problems I am facing
track_2016 Would you say things in this country today are… (better – worse)
persfinretro_2016 Change in personal finances over past year (better – worse)
econtrend_2016 The economy is… (better – worse)
Americatrend_2016 Life in America today for people like me compared to

fifty years ago… (better - worse)

futuretrend_2016 When children today are the age you are now, do you think their standard of living will be better, about the same, or worse than yours is now?

wealth_2016 Do you feel that the distribution of money and wealth in this country is fair, or do you feel that the money and wealth in this country should be more evenly distributed among more people?

values_culture_2016 In America today, do you feel the values and culture of people like you are…

US_respect_2016 Compared with the past, would you say the US is more espected by other countries these days, less respected by other countries, or as respected as it has been in the past?

trustgovt_2016 How much of the time do you think you can trust the government in Washington to do what is right?

trust_people_2016 Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?

helpful_people_2016 Would you say that most of the time people try to be helpful, or that they are mostly just looking out for themselves?

fair_people_2016 Most people try to take advantage or try to be fair

imiss_a_2016 Issue importance - Crime

imiss_b_2016 Issue importance - The economy

imiss_c_2016 Issue importance - Immigration

imiss_d_2016 Issue importance - The environment

imiss_e_2016 Issue importance - Religious liberty

imiss_f_2016 Issue importance - Terrorism

imiss_g_2016 Issue importance - Gay rights

imiss_h_2016 Issue importance - Education

imiss_i_2016 Issue importance - Family and medical leave

imiss_j_2016 Issue importance - Health care

imiss_k_2016 Issue importance - Money in politics

imiss_l_2016 Issue importance - Climate change

imiss_m_2016 Issue importance - Social Security

imiss_n_2016 Issue importance - Infrastructure investment

imiss_o_2016 Issue importance - Jobs

imiss_p_2016 Issue importance - The budget deficit

imiss_q_2016 Issue importance - Poverty

imiss_r_2016 Issue importance - Taxes

imiss_s_2016 Issue importance - Medicare

imiss_t_2016 Issue importance - Abortion

imiss_u_2016 Issue importance - The size of government

imiss_x_2016 Issue importance - Racial equality

imiss_y_2016 Issue importance - Gender equality

Clinton_supp Dummy for å ha stemt på Clinton i sist valg