For this assignment, we decided to use famous 'Adult' dataset from UCI Machine Learning Repository, which was extracted, cleaned and formatted from 1994 census bureau database. This dataset contains 48842 tuples with multiple dimensions: age, workclass, fnlwgt( calculated weights), education, education\_num, marital\_status, occupation, relationship, race, sex, capital\_gain, capital\_loss, hours\_per\_week, native\_country (<http://archive.ics.uci.edu/ml/datasets/Adult>).

Our goal of this visualization is to help people learn more about census dataset and give them flexibility to dig into it in an interactive way.

The dataset

Here I list overall diagrams about each category and several relationships.

The most important feature of this dataset is that, as you can see, most of the variables are nominal or ordinal (or can easily be transferred to). Interval and Ratio values cannot be eyeballed directly, which needs to calculated by percentage or something else transformation. For simplicity, we decided to show direct values and relationships among those variables rather than generate more numeric variables (which is a great extension but need more time).

achine generated alternative text:
age 
workclass 
fnlwgt 
Valid 
Mismatched 
Empty 
5M 
117 , 822 
237 , 
32.561 
Valid 
Mismatched 
Empty 
Btate-gov 
Self-emp-inc 
*ederal -gov 
thout -pay 
'Never-worked 
32.561 
1,836 
1,298 
Valid 
Mismatched 
Empty 
Minimum 
Lower quartile 
Median 
Upper quartile 
Maximum 
1.0M 
32.561 
1.5M 
12,285 
178,356 
education 
Valid 
Mismatched 
Empty 
ers 
oc-voc 
•soc-acdm 
•th-8th 
*of-school 
•zth 
Doctorate 
5th -6th 
llst-4th 
'preschool 
1,382 
1,067 
Minimum 
Lower quartile 
Median 
Upper quartile 
Maximum 

achine generated alternative text:
education_num 
marital_status 
occupation 
Valid 
Mismatched 
Empty 
Minimum 
Lower quartile 
Median 
Upper quartile 
Maximum 
32.561 
Valid 
Mismatched 
Empty 
*parated 
dowed 
arried-spouse-absent 
'Married-AF-spouse 
32.561 
1,025 
418 
23 
Valid 
Mismatched 
Empty 
sales 
Other-serv• 
-c eaners 
"r*ng-fishing 
*tective-serv 
priv-house-serv 
'Armed-Forces 
32.561 
3770 
3,650 
3,2gs 
2,002 
1,843 
1,370 
relationship 
Valid 
Mismatched 
Empty 
*her-relative 
3446 
1, s68 

achine generated alternative text:
capital_gain 
Valid 
Mismatched 
Empty 
IAmer-Indian-Eskimo 
10ther 
32.561 
o 
Valid 
Mismatched 
Empty 
capital_loss 
32.561 
Valid 
Mismatched 
Empty 
50k 
Minimum 
Lower quartile 
Median 
Upper quartile 
Maximum 
32.561 
100k 
, 999 
Valid 
Mismatched 
Empty 

achine generated alternative text:
hrs_per_week 
native_country 
Valid 
Mismatched 
Empty 
(phi lippines 
IGermany 
ICanada 
'Puerto-Rico 
'El -Salvador 
'India 
lcuba 
'England 
IJamaica 
Isouth 
'China 
lltaly 
'Dominican-Republic 
etnam 
IGuatema1a 
'Japan 
'Poland 
income 
Valid 
Mismatched 
Empty 
Minimum 
Lower quartile 
Median 
Upper quartile 
Maximum 
32.561 
32.561 
Valid 
Mismatched 
Empty 

Here are some relationships among some variables:

achine generated alternative text:
Educatbn 
11K 
10K 
Income 
e=50K 
>50K 
1st4th 5th.6th 7th.8th 
Bachelors Doctorate HS-grad Masters Preschool 
10th 
11th 
12th 
Assoc- 
acdm Assoc-voc 
um of Number of Records for each Education. Color shows details about Income. The data is filtered on Race, which keeps Amer-Indian-Eskimo, Asian-Pac-islander, 
prOf_ 
school college 
Black, Other and 

achine generated alternative text:
Number of Records 

achine generated alternative text:
Number of Records 

achine generated alternative text:
Number of Records 

achine generated alternative text:
1500 
1000 
Adm- 
clerical 
Armed- 
Forces 
Craft- 
repair 
Exec- 
managerial 
Farming- 
fishing 
Handlers- 
cleaners 
Occupation 
Machine- 
op-inspct 
Other- 
service 
Sales 
hous„ 
specialty 
Protective.. 
Tech- 
support 
Transport- 
moving 

It is clearly to see that there are some patterns and relationships among different variables, such as occupation against sex, marital status against income etc. Since those variables are highly integrated, it is useful to use an interactive graph to show all those variables.

Visualization Decision

From above description, we think Sankey Diagram is a proper choice: 1. this diagram can show different dimensions in one picture. 2. User can interact with this diagram to see more details.

User Story

People who want to use our diagram curious about the relationships and patterns among different dimensions in a census dataset. He/she might ask "What is the relationship between education level and income level?" And then add more dimensions into former relationship, like "What is the relationship or pattern among education level, sex, race, maritus status and income level?". Our visualization can offer them a lot flexibility to choose whatever relationships they want to see. The basic process of user story is: user can click on some variable (male in sex for example), and then can see related values of other dimensions (PICTURE!). User can then choose select or hover to the next dimension to see anything changed. Another option for users is that they can drag and drop one dimension into another place to 'sort' the diagram and show more details. (PICTURE!)

There are more other advanced interactions, like zooming in when clicking one subset of a variable, and value filtering etc. but have no time to finish them.