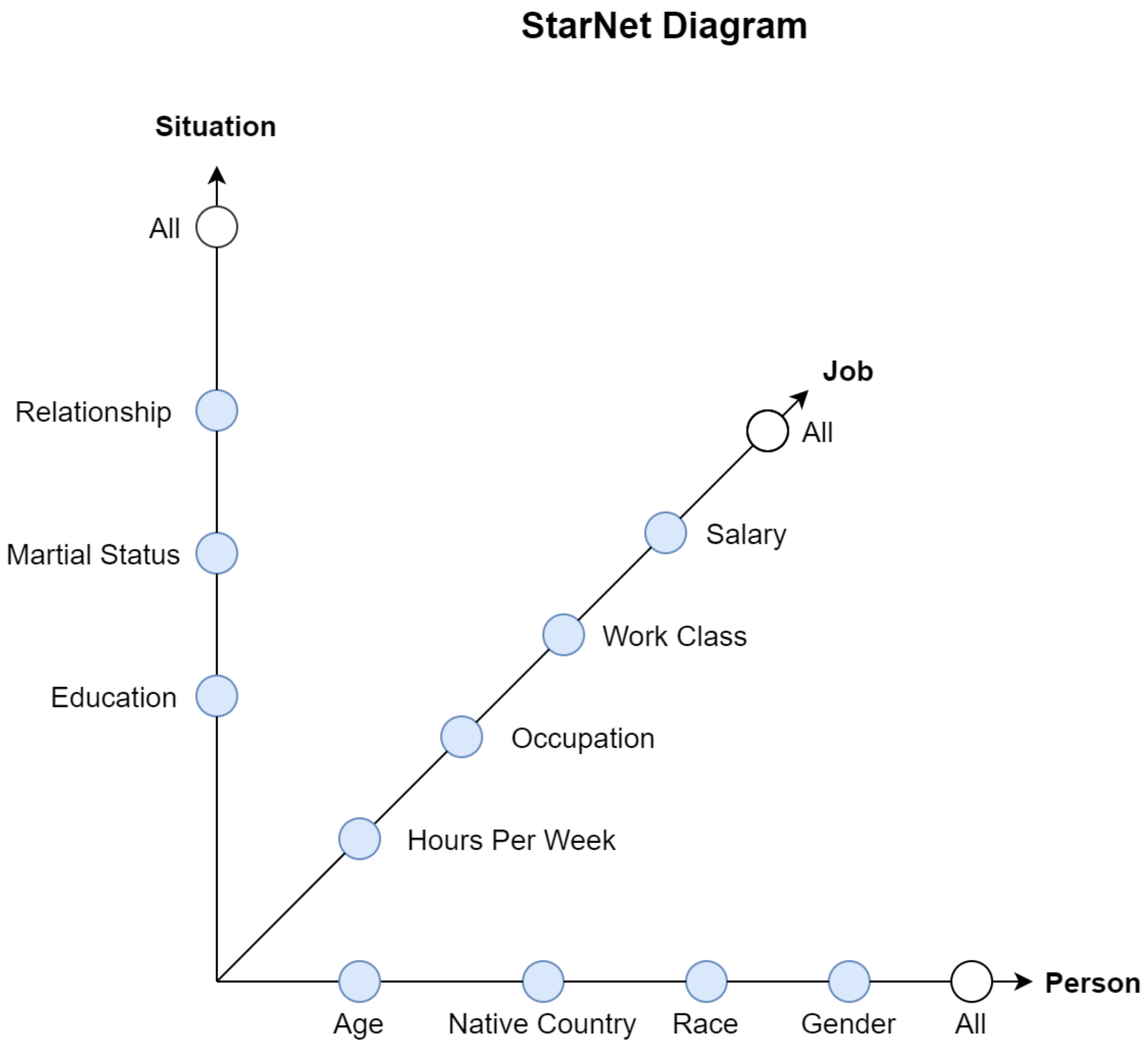
CITS3401 Mid Project Submission – US Adult Income Dataset
Gregory Edmonds 21487148

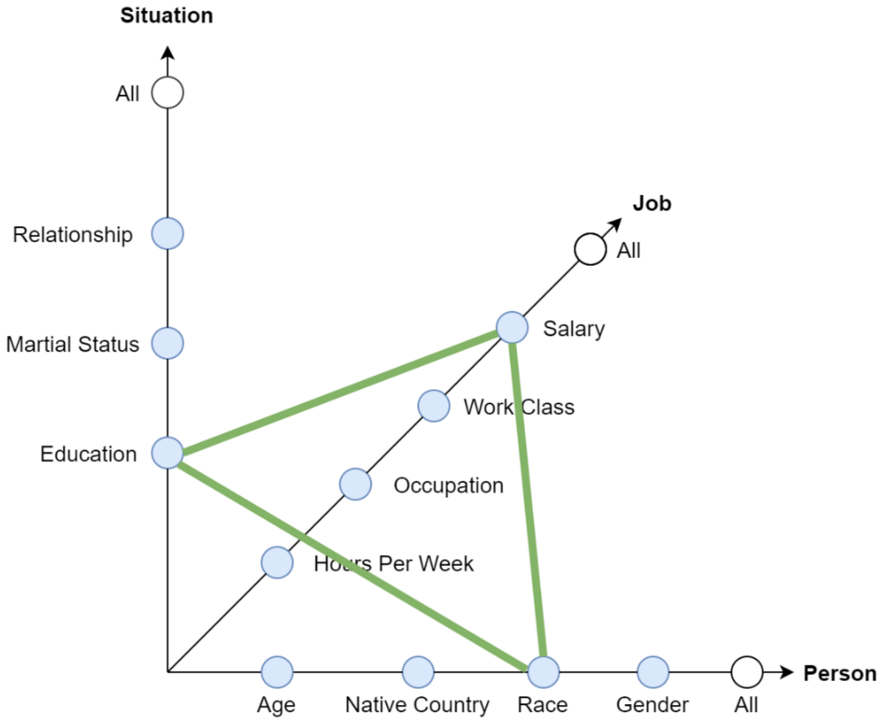
Business Queries

1. What number of people whose race is not white and whose highest education is HS-grad have a salary <=50K?
2. How do the salaries of those who are male and single compare with those who are male and married?
3. How does the salary of people under 30 with children compare with people under 30 without children?
4. How does the salary of people who have a bachelors education differ between males and females?
5. What age has the highest number of never-married people with a salary >50K?

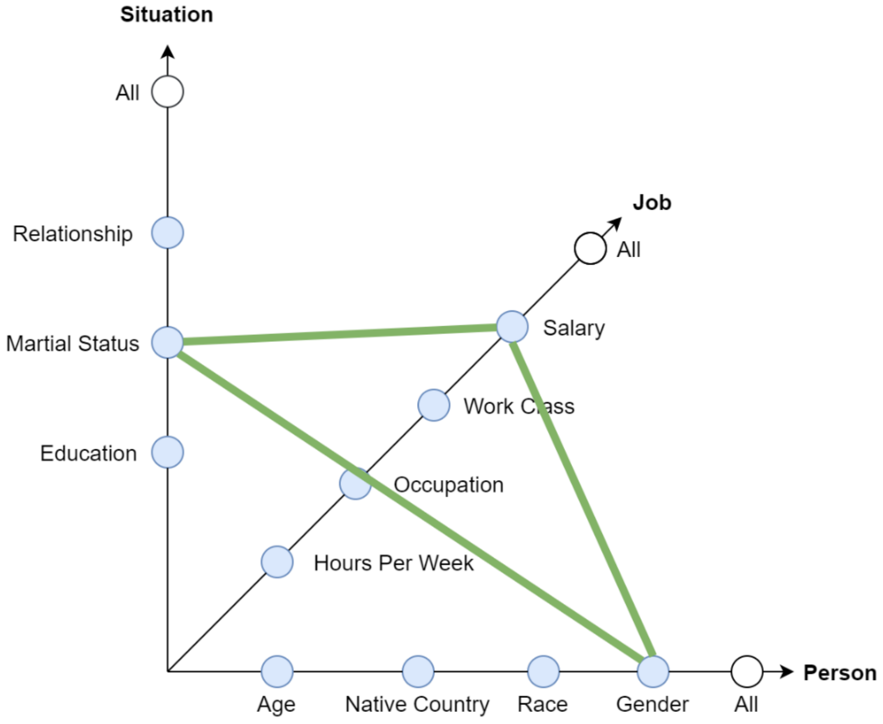


All five queries can be answered using the StarNet design. Each query contains one footprint in each dimension, and all include Salary. This is shown in the examples below.

StarNet of Query 1



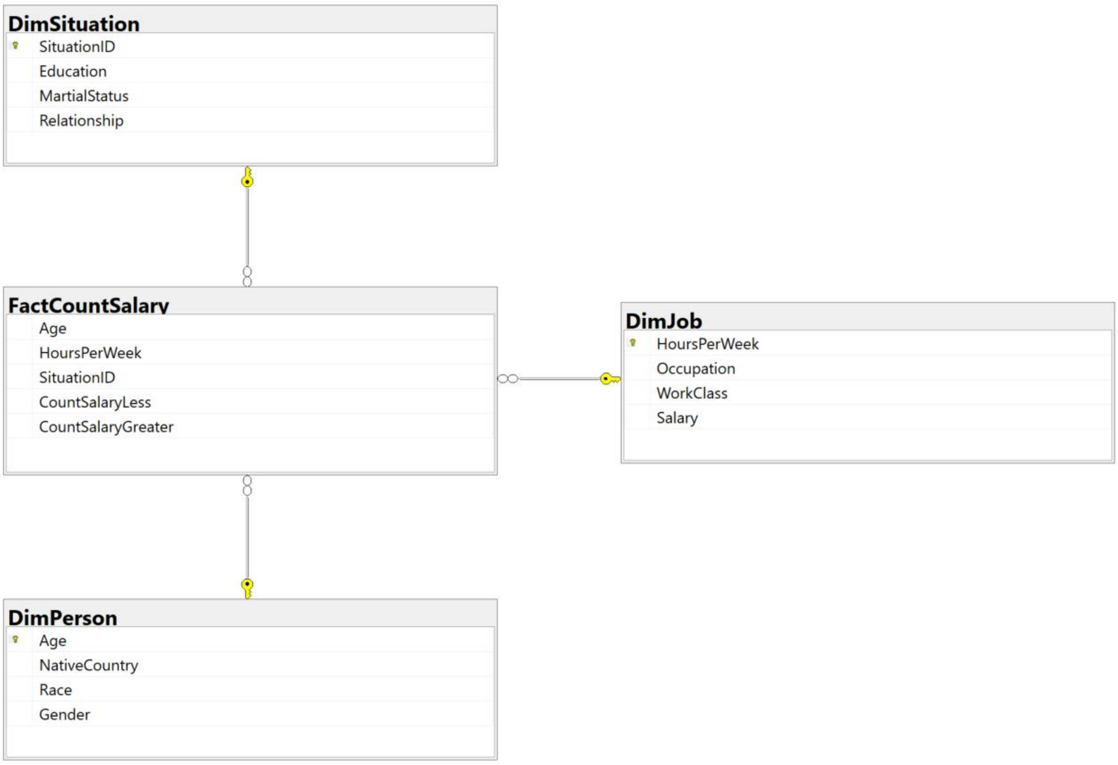
StarNet of Query 2



Due to the nature of the data, I designed to make the hierarchies categorical instead of hierarchal. The number of options each footprint could have determine where they are in the hierarchy. For example, HoursPerWeek and Age both sit at the lowest level of their dimensions as they are continuous numbers. Likewise, Salary and Gender sit at the top as they only have two options each.

When choosing dimensions, I decided to split all things to do with a person between their genetic attributes and situational/circumstantial attributes. I wanted to do this to see whether a person's salary depended more on things that were in their control against things that weren't. The salary is indeed the measure of the warehouse, as shown in the ER diagram.

ER Diagram



The ER diagram is in the form of a Star schema. Each dimension has a primary key of type *int*. For the dimension DimSituation there were no continuous variables to do this. Instead, I created a column called SituationID in the DimSituation spreadsheet numbered all the way through to the end of the data. This acted as a surrogate key, and allowed a primary key of type *int*. It is the first column of the spreadsheet.

ETL Process

The ETL process involved data reduction and manipulation. Compared to the original data, I removed fnlwgt because it was a count of people that the original creators predicted belong to each row, but I believe the spreadsheet has enough data to analyse for the purpose of the warehouse – to determine how different factors affected a person's salary. I removed the education-num column for the same reason. I also removed the capital-gain and capital-loss columns as there were to many people who had no value in either of these columns. This would make the data too sparse and provide an extra load for the warehouse to process.

On inspecting the data, I found that some values had '?' instead of an actual value. Because of this, I changed all these values to *null*, and later created the sql scripts for their columns (Work Class, Occupation and Native Country) to include the *not null* attribute, so their instances would not be inserted into the database.

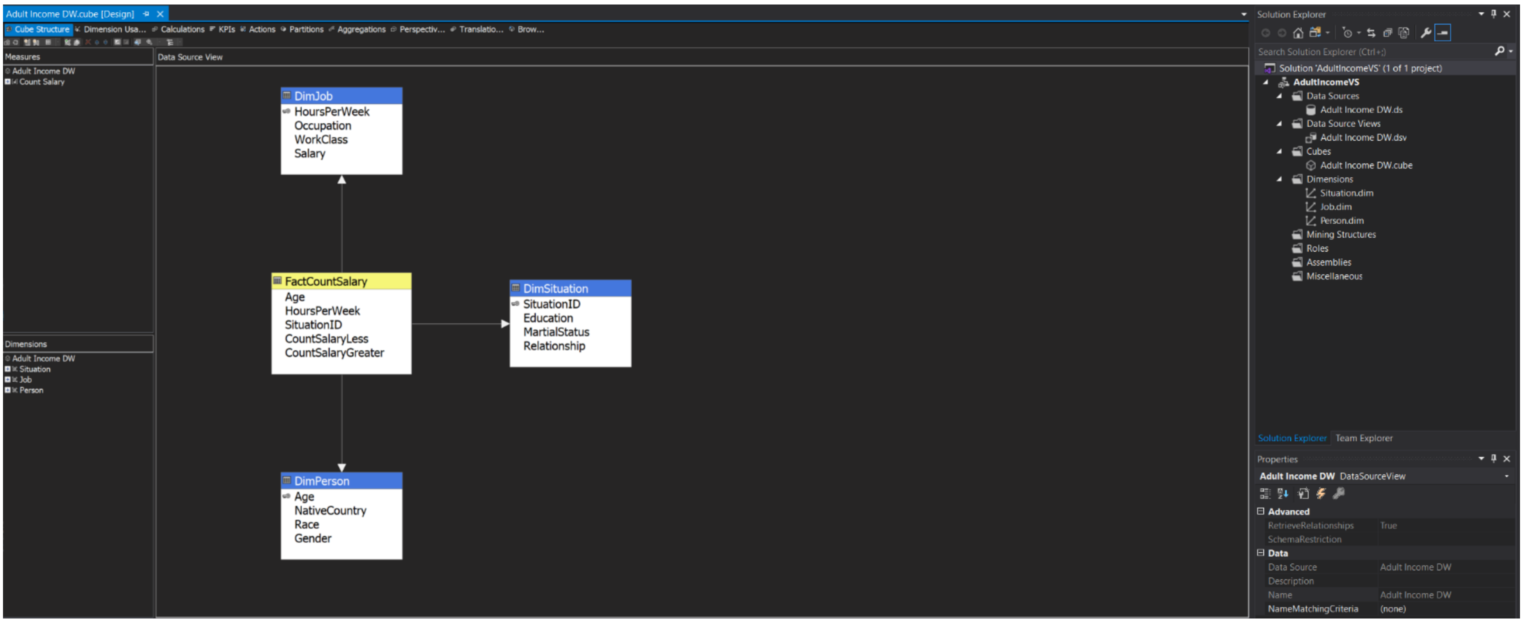
```
Create table DimPerson
(
Age int primary key identity,
NativeCountry varchar(50) not null,
Race varchar(20),
Gender varchar(10)
)
Go
```

```
Create table DimJob
(
HoursPerWeek int primary key identity,
Occupation varchar(20) not null,
WorkClass varchar(20) not null,
Salary varchar(10)
)
Go
```

Then, I reduced the spreadsheet into separate spreadsheets for each dimension to easily insert the data into MSMM.

DimJob
DimPerson
DimSituation

Cube Diagram



Measures: Adult Income DW; Count Salary

Dimensions: Adult Income DW; Situation; Job; Person

DimJob: HoursPerWeek, Occupation, WorkClass, Salary

FactCountSalary: Age, HoursPerWeek, SituationID, CountSalaryLess, CountSalaryGreater

DimSituation: SituationID, Education, MartialStatus, Relationship

DimPerson: Age, NativeCountry, Race, Gender

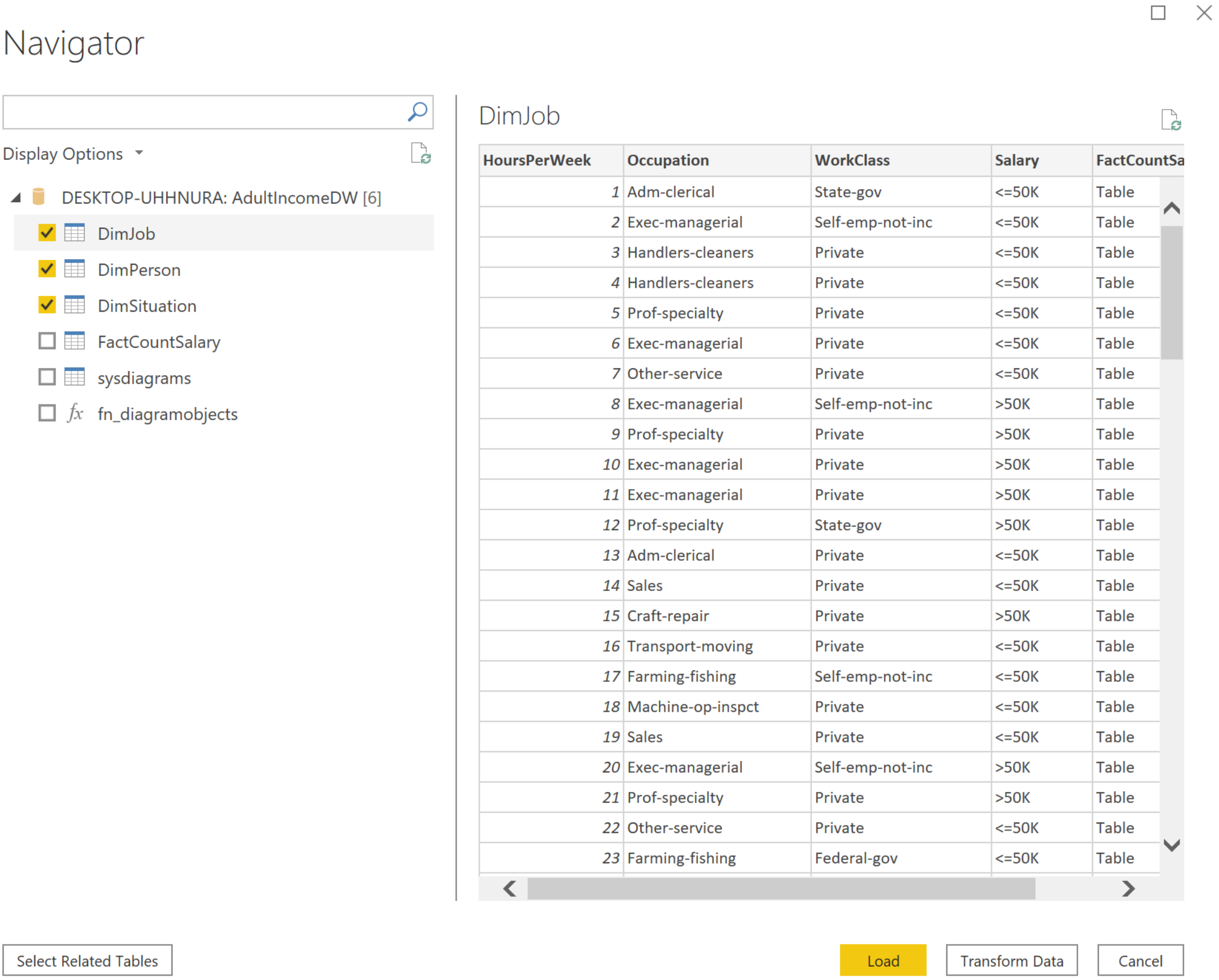
Solution 'AdultIncomeVS' (1 of 1 project)
AdultIncomeVS
Data Sources
Adult Income DW.ds
Data Source Views
Adult Income DW.dsv
Cubes
Adult Income DW.cube
Dimensions
Situation.dim
Job.dim
Person.dim
Mining Structures
Roles
Assemblies
Miscellaneous

Adult Income DW DataSourceView

Advanced
RetrieveRelationships True
SchemaRestriction
Data
Data Source Adult Income DW
Description
Name Adult Income DW
NameMatchingCriteria (none)

Power Bi

Unfortunately, I was unable to get Power Bi to work properly. When connecting to the SQL Server Analysis Services database, my graphs would be empty no matter which fields I chose. I even tried connecting to just the SQL Server database with the same results, although I was able to see data in the navigator preview using this method. I have not yet found a solution.



Navigator

DESKTOP-UHHNURA: AdultIncomeDW [6]

- ☑ DimJob
- ☑ DimPerson
- ☑ DimSituation
- □ FactCountSalary
- □ sysdiagrams
- □ fn_diagramobjects

DimJob

HoursPerWeek	Occupation	WorkClass	Salary	FactCountSa
1	Adm-clerical	State-gov	<=50K	Table
2	Exec-managerial	Self-emp-not-inc	<=50K	Table
3	Handlers-cleaners	Private	<=50K	Table
4	Handlers-cleaners	Private	<=50K	Table
5	Prof-specialty	Private	<=50K	Table
6	Exec-managerial	Private	<=50K	Table
7	Other-service	Private	<=50K	Table
8	Exec-managerial	Self-emp-not-inc	>50K	Table
9	Prof-specialty	Private	>50K	Table
10	Exec-managerial	Private	>50K	Table
11	Exec-managerial	Private	>50K	Table
12	Prof-specialty	State-gov	>50K	Table
13	Adm-clerical	Private	<=50K	Table
14	Sales	Private	<=50K	Table
15	Craft-repair	Private	>50K	Table
16	Transport-moving	Private	<=50K	Table
17	Farming-fishing	Self-emp-not-inc	<=50K	Table
18	Machine-op-inspct	Private	<=50K	Table
19	Sales	Private	<=50K	Table
20	Exec-managerial	Self-emp-not-inc	>50K	Table
21	Prof-specialty	Private	>50K	Table
22	Other-service	Private	<=50K	Table
23	Farming-fishing	Federal-gov	<=50K	Table