Data Mining: US Adults Work Hours

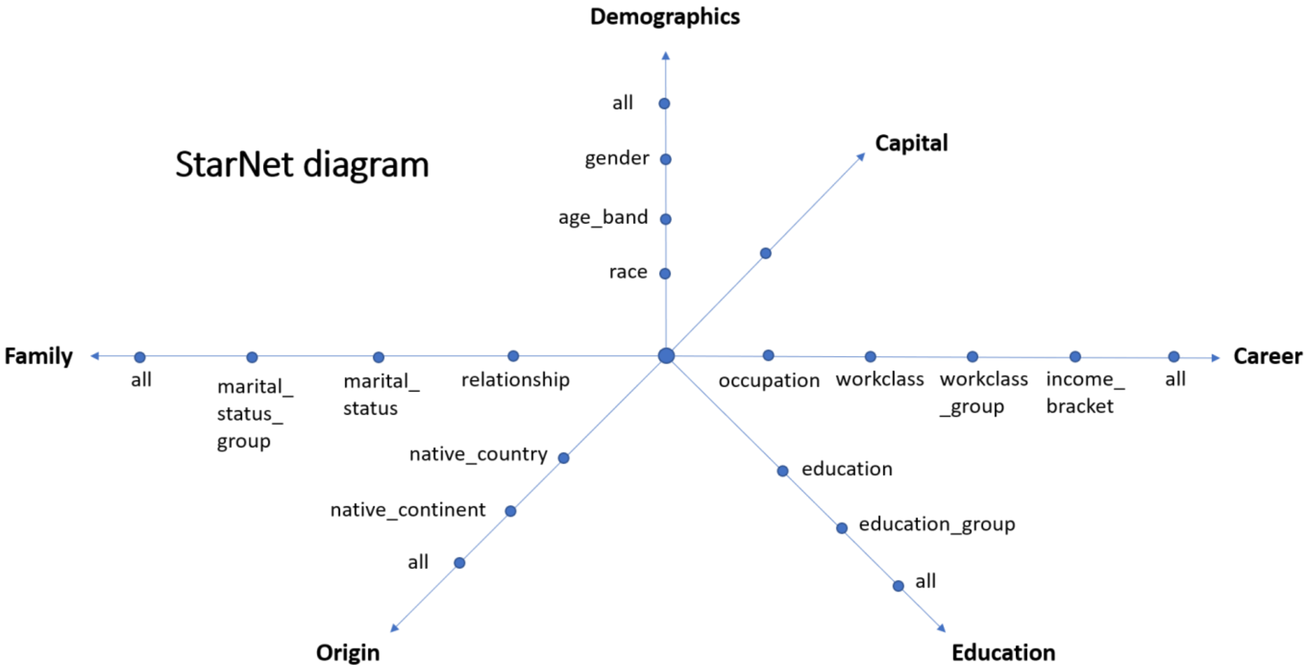
1. Introduction

Research suggests that a healthy average work hour should not be more than 39 hours per week (Dinh et al., 2017). However, this is not always attainable even for the developed country like the United States. In this project, data warehouse design and implementation were carried out using the US Adult Income dataset to model a business process **US adults' work hours**, where we would like to see how status, background, career, wage or any other characteristics and conditions differ their average work hours per week.

There are essentially many business questions arising from the data which can be investigated further in the following sections

i. Is average work hours of US adults conform with their wages for all work classes?
ii. How does education level affect weekly work hours for certain income bracket?
iii. Are males' and females' work hours different for all age group?
iv. Do people from different marital status have different number of work hours?
v. Do people originated outside North America have higher or lower work hours per week when comparing to those native to North America?

2. Design

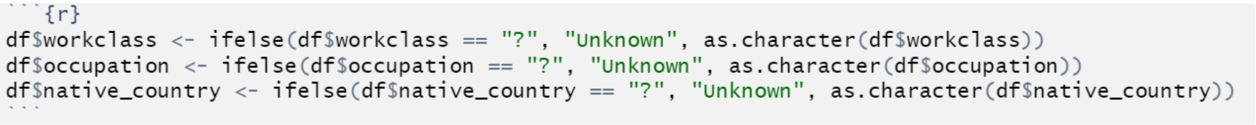

StarNet diagram

To answer the business queries in the above section, a StarNet query model for the US Adults data warehouse was constructed with 6 radial lines, representing concept hierarchies for the dimensions *Capital, Career, Demographic, Education, Family* and *Origin*. Each line consists of footprints representing abstraction levels of the dimension.

3. Data Preprocessing (ETL process)

i. Data cleaning

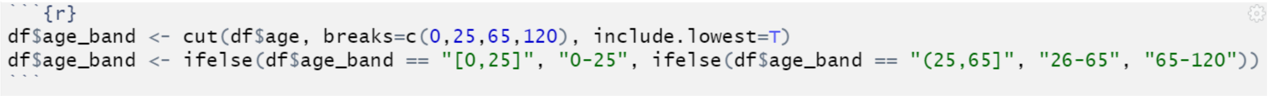
- dealing with uninformative values *in workclass, occupation and native_country* attributes: revalue “?” to “Unknown”.



~~~
```{r}
df$workclass <- ifelse(df$workclass == "?", "Unknown", as.character(df$workclass))
df$occupation <- ifelse(df$occupation == "?", "Unknown", as.character(df$occupation))
df$native_country <- ifelse(df$native_country == "?", "Unknown", as.character(df$native_country))
```
~~~

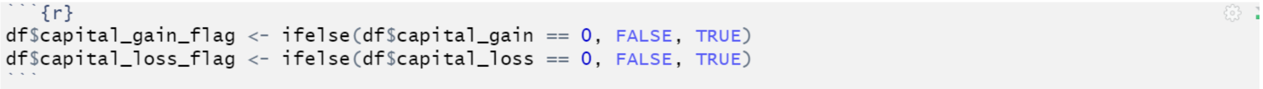
ii. Discretizing numeric attributes

- *age* to *age_band*, where age is grouped in range 0-25, 26-65 and 66-120 (assume that maximum age = 120).



~~~
```{r}
df$age_band <- cut(df$age, breaks=c(0,25,65,120), include.lowest=T)
df$age_band <- ifelse(df$age_band == "[0,25]", "0-25", ifelse(df$age_band == "(25,65]", "26-65", "65-120"))
```
~~~

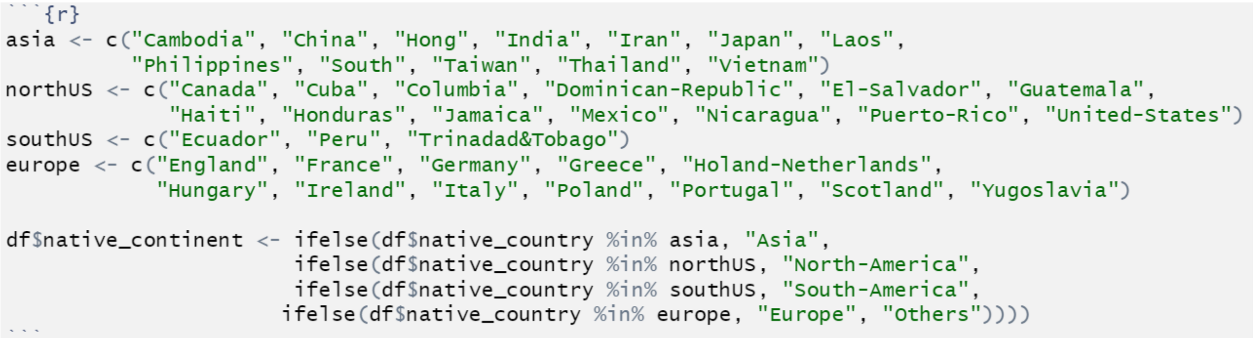
- *capital_gain* and *capital_loss* to *capital_gain_flag* and *capital_loss_flag*, where 0 is mapped to FALSE and other values greater than 0 to TRUE.



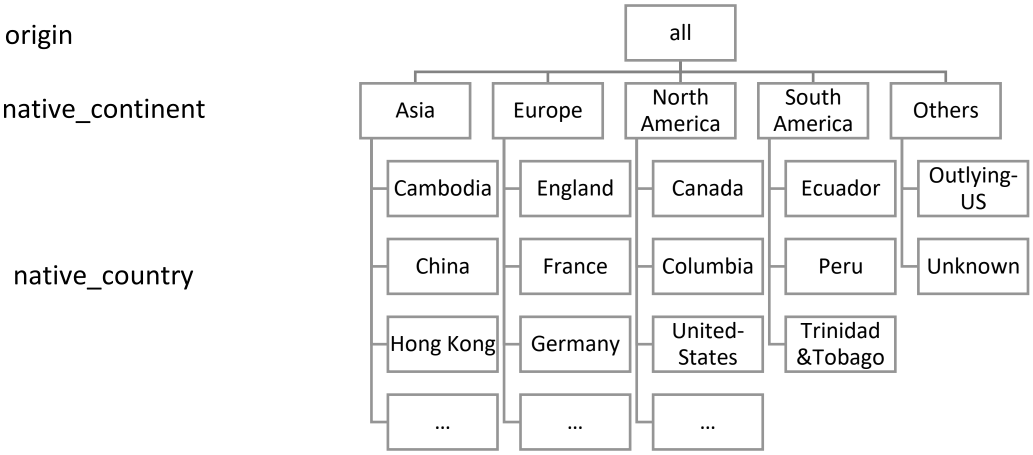
~~~
```{r}
df$capital_gain_flag <- ifelse(df$capital_gain == 0, FALSE, TRUE)
df$capital_loss_flag <- ifelse(df$capital_loss == 0, FALSE, TRUE)
```
~~~

iii. Concept hierarchies' generation

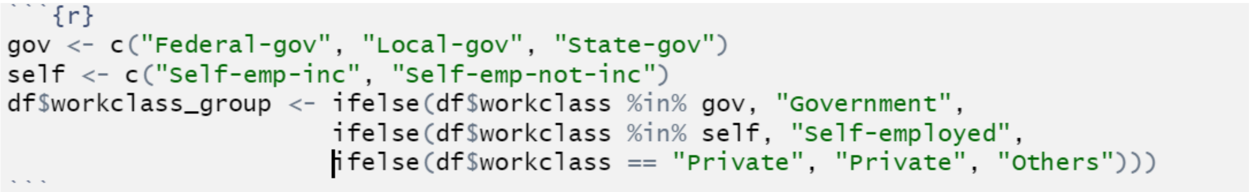
- set-grouping hierarchies:
  - native_country < native_continent



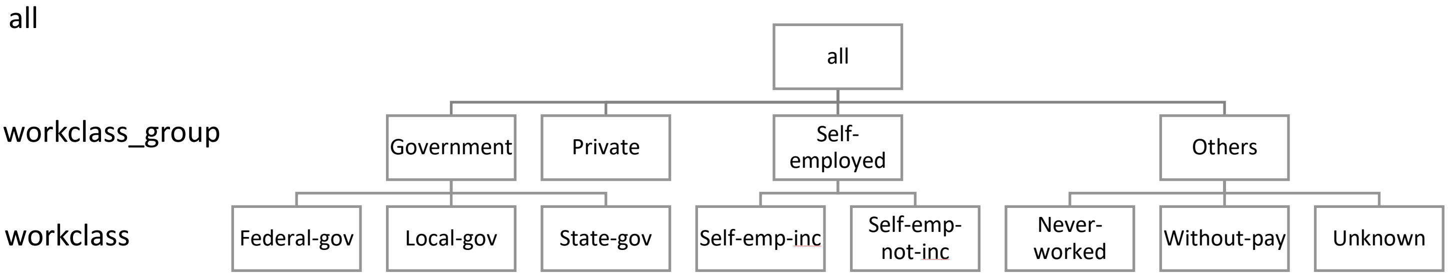
~~~
```{r}
asia <- c("Cambodia", "China", "Hong", "India", "Iran", "Japan", "Laos",
          "Philippines", "South", "Taiwan", "Thailand", "Vietnam")
northUS <- c("Canada", "Cuba", "Columbia", "Dominican-Republic", "El-Salvador", "Guatemala",
             "Haiti", "Honduras", "Jamaica", "Mexico", "Nicaragua", "Puerto-Rico", "United-States")
southUS <- c("Ecuador", "Peru", "Trinadad&Tobago")
europe <- c("England", "France", "Germany", "Greece", "Holand-Netherlands",
            "Hungary", "Ireland", "Italy", "Poland", "Portugal", "Scotland", "Yugoslavia")

df$native_continent <- ifelse(df$native_country %in% asia, "Asia",
                        ifelse(df$native_country %in% northUS, "North-America",
                        ifelse(df$native_country %in% southUS, "South-America",
                       ifelse(df$native_country %in% europe, "Europe", "Others"))))
```
~~~



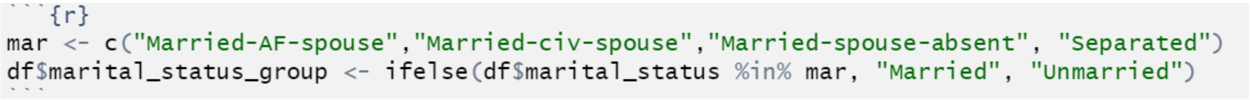
- ○ workclass < workclass_group



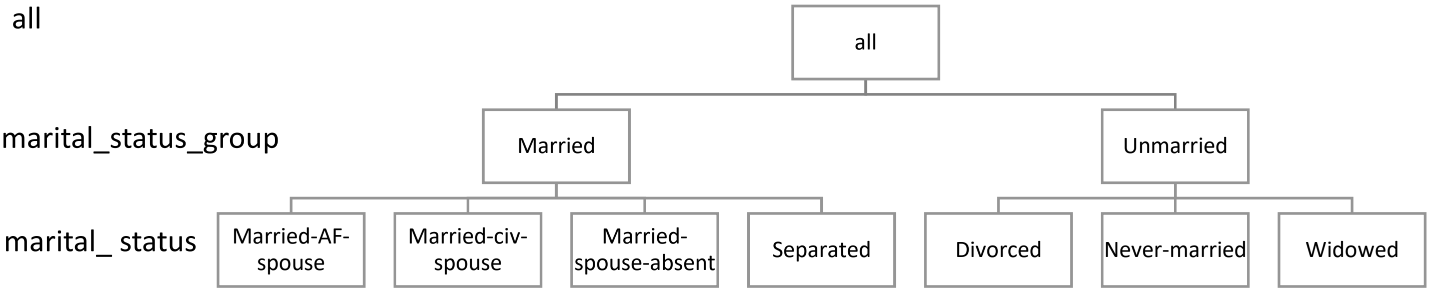
~~~
```{r}
gov <- c("Federal-gov", "Local-gov", "State-gov")
self <- c("Self-emp-inc", "Self-emp-not-inc")
df$workclass_group <- ifelse(df$workclass %in% gov, "Government",
                      ifelse(df$workclass %in% self, "Self-employed",
                      ifelse(df$workclass == "Private", "Private", "Others")))
```
~~~



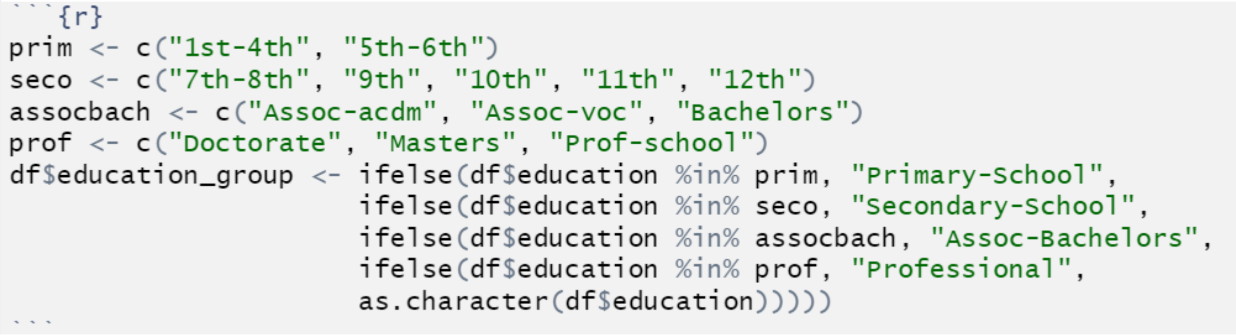
- ○ marital_status < marital_status_group



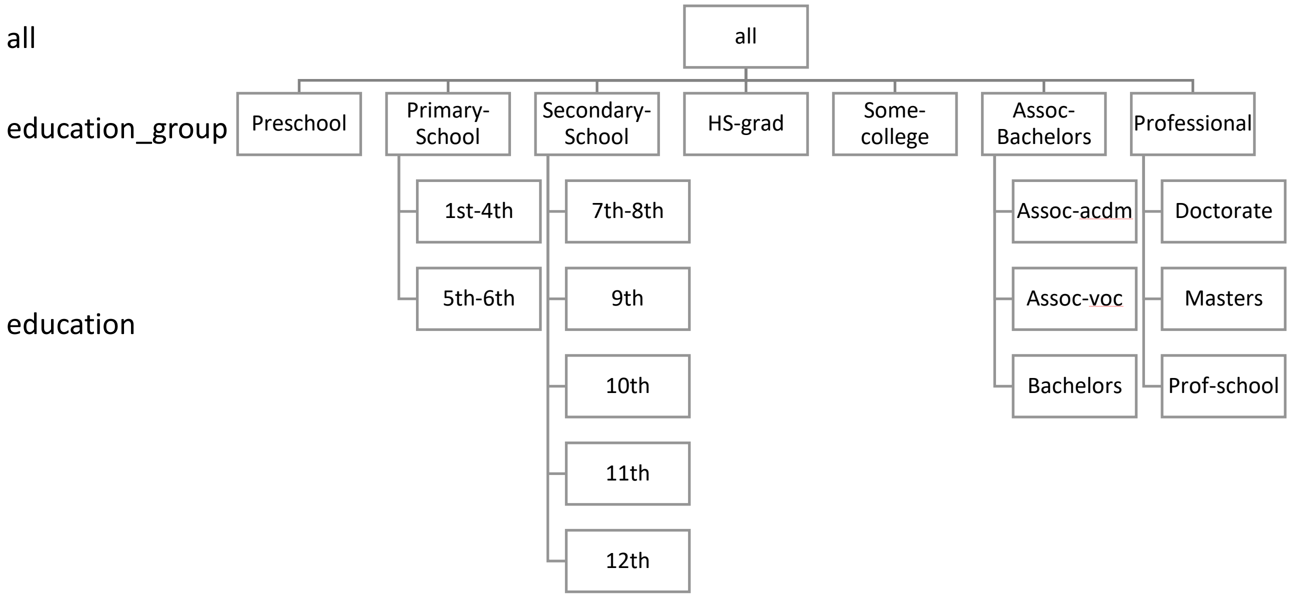
~~~
```{r}
mar <- c("Married-AF-spouse","Married-civ-spouse","Married-spouse-absent", "Separated")
df$marital_status_group <- ifelse(df$marital_status %in% mar, "Married", "Unmarried")
```
~~~



- ○ education < education_group



~~~
```{r}
prim <- c("1st-4th", "5th-6th")
seco <- c("7th-8th", "9th", "10th", "11th", "12th")
assocbach <- c("Assoc-acdm", "Assoc-voc", "Bachelors")
prof <- c("Doctorate", "Masters", "Prof-school")
df$education_group <- ifelse(df$education %in% prim, "Primary-School",
                      ifelse(df$education %in% seco, "Secondary-School",
                      ifelse(df$education %in% assocbach, "Assoc-Bachelors",
                      ifelse(df$education %in% prof, "Professional",
                      as.character(df$education)))))
```
~~~

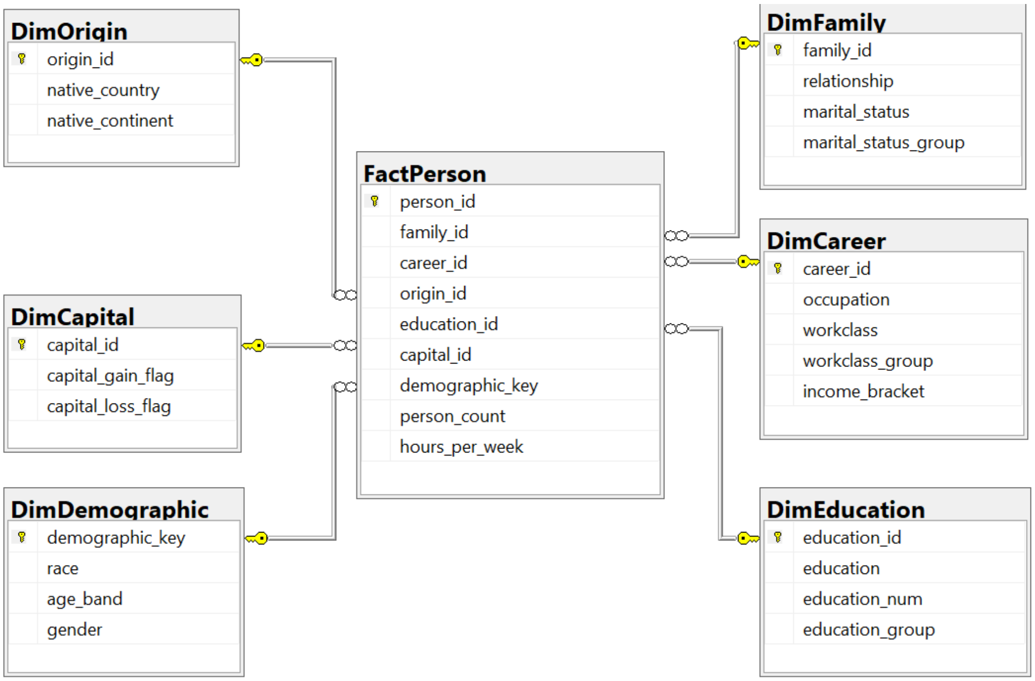


iv. **Create separate .csv files for building and populating data**

6 dimension + 1 fact tables were created using *dplyr* package and then exported as csv files, *DimCapital.csv, DimCareer.csv, DimDemographic.csv, DimEducaiton.csv, DimFamily.csv, DimOrigin.csv* and *FactPerson.csv.*

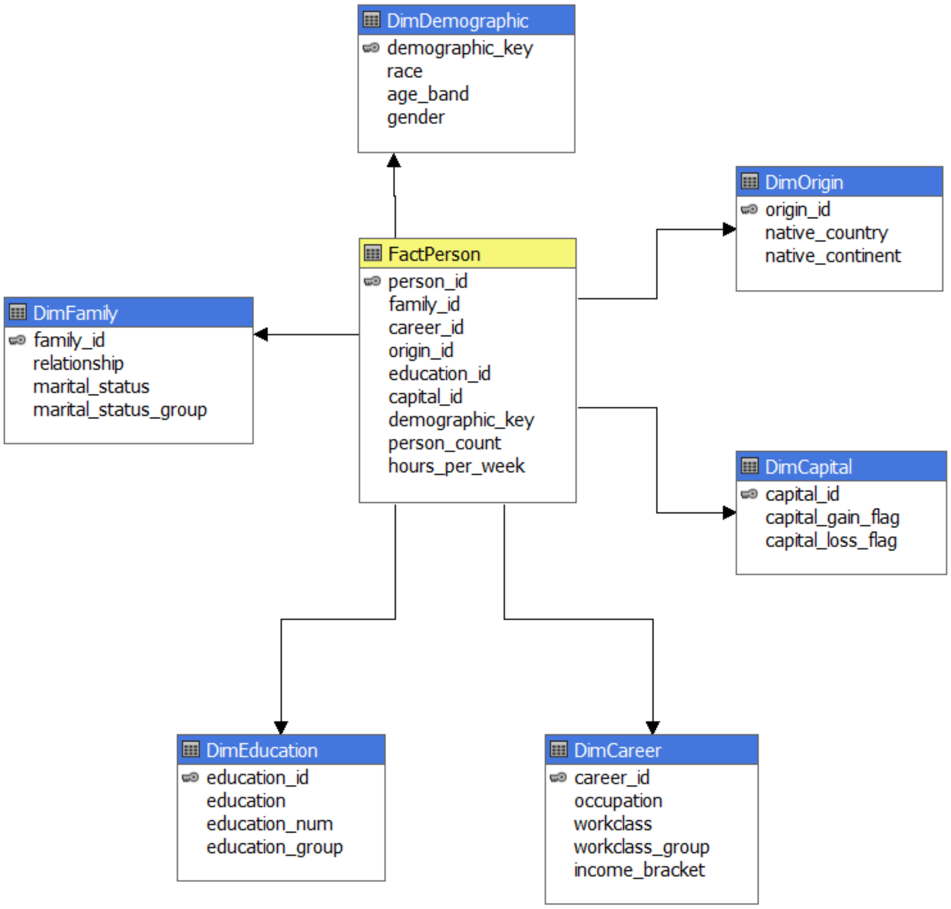
4. Implementation

Once the csv files were created, a star schema was implemented using script from *Create_tables.sql*, and populated using *Bulk_insert_data.sql*. The ER diagram generated in SSMS is as follows:



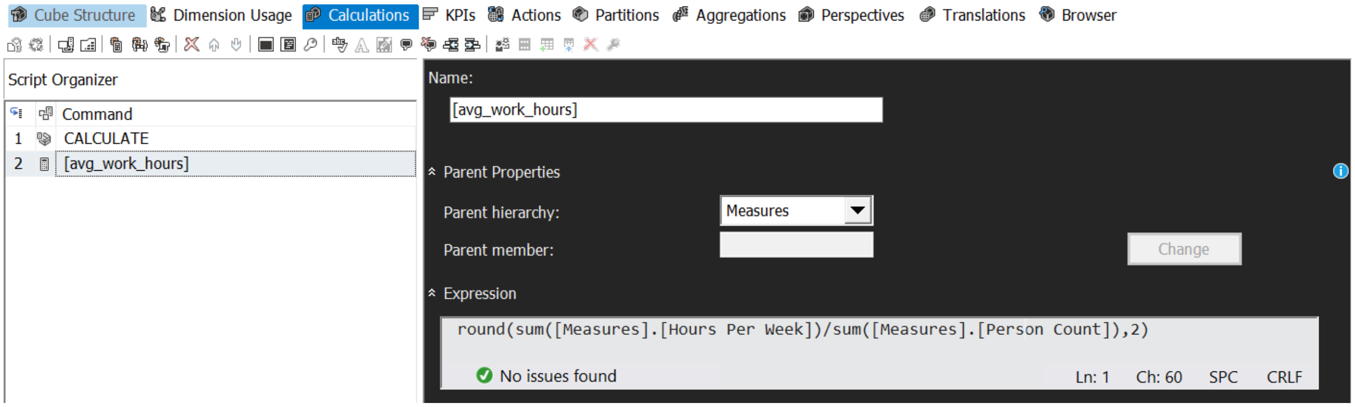
Next, SQL Server Data Tools was used to build a multi-dimensional analysis service solution together with a cube designed to answer the business queries.

Concept hierarchies of each dimension were built to match the StarNet designed in the first section. The solution project file, *US_income.sln,* and its folder is submitted together in the .zip file. The cube diagram generated is as follows:



Since the fact attribute I chose is not an additive fact, I have also created new calculated measure called *avg_work_hours* using MDX query to achieve the objective.

The script I used is adapted from [Lesson 6: Defining Calculated Members](#) available on Microsoft SQL Server 2008 Documentation.



Script Organizer

Command
1 CALCULATE
2 [avg_work_hours]

Name:
[avg_work_hours]

Parent Properties

Parent hierarchy: Measures

Parent member:

Expression

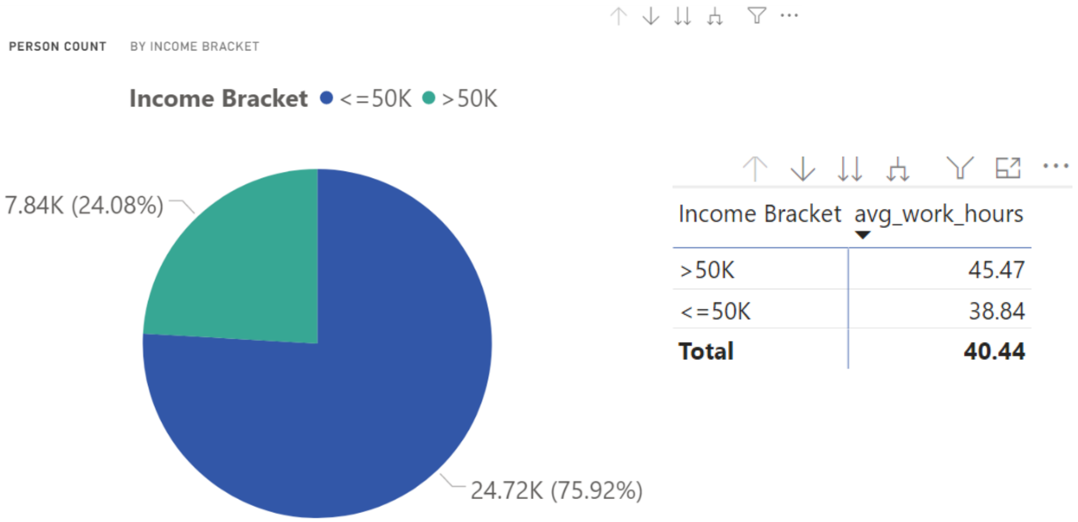
```
round(sum([Measures].[Hours Per Week])/sum([Measures].[Person Count]),2)
```

No issues found Ln: 1 Ch: 60 SPC CRLF

5. Usage of Data Warehouse

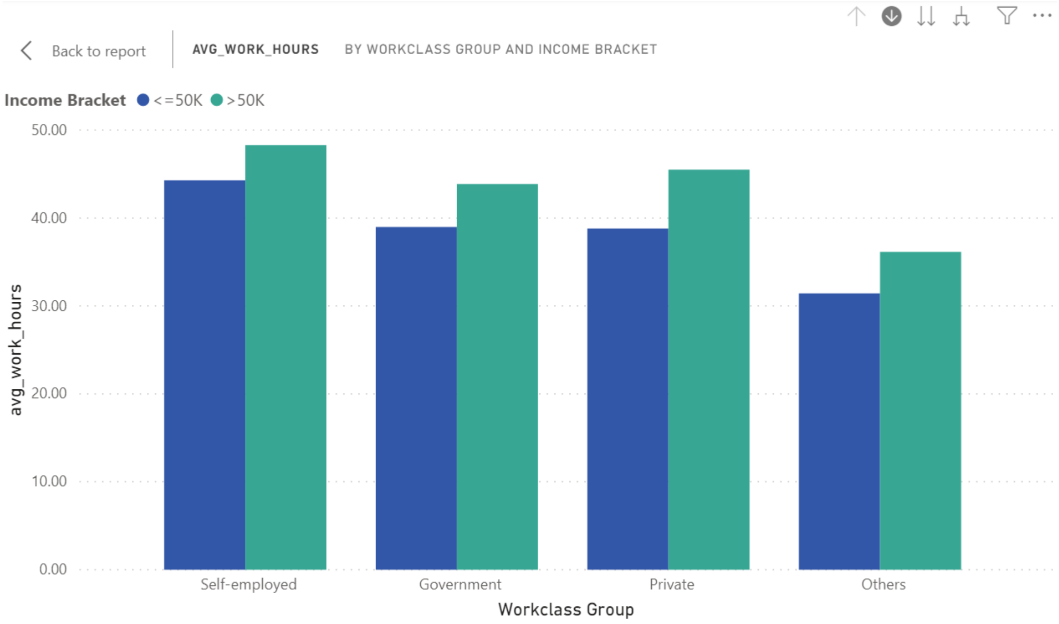
With the use of Power BI, we are now be able to visualize the data returned from the business queries.

a. Is average work hours of US adults conform with their wages for all work classes?

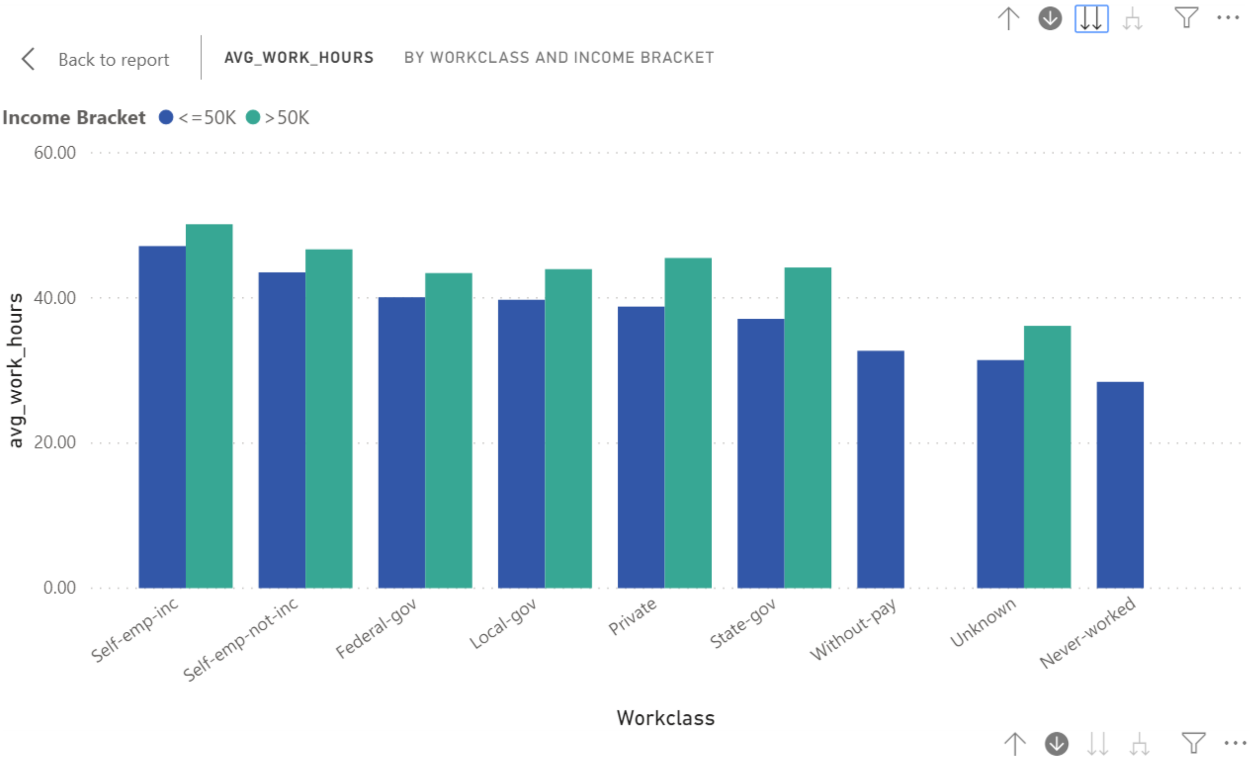


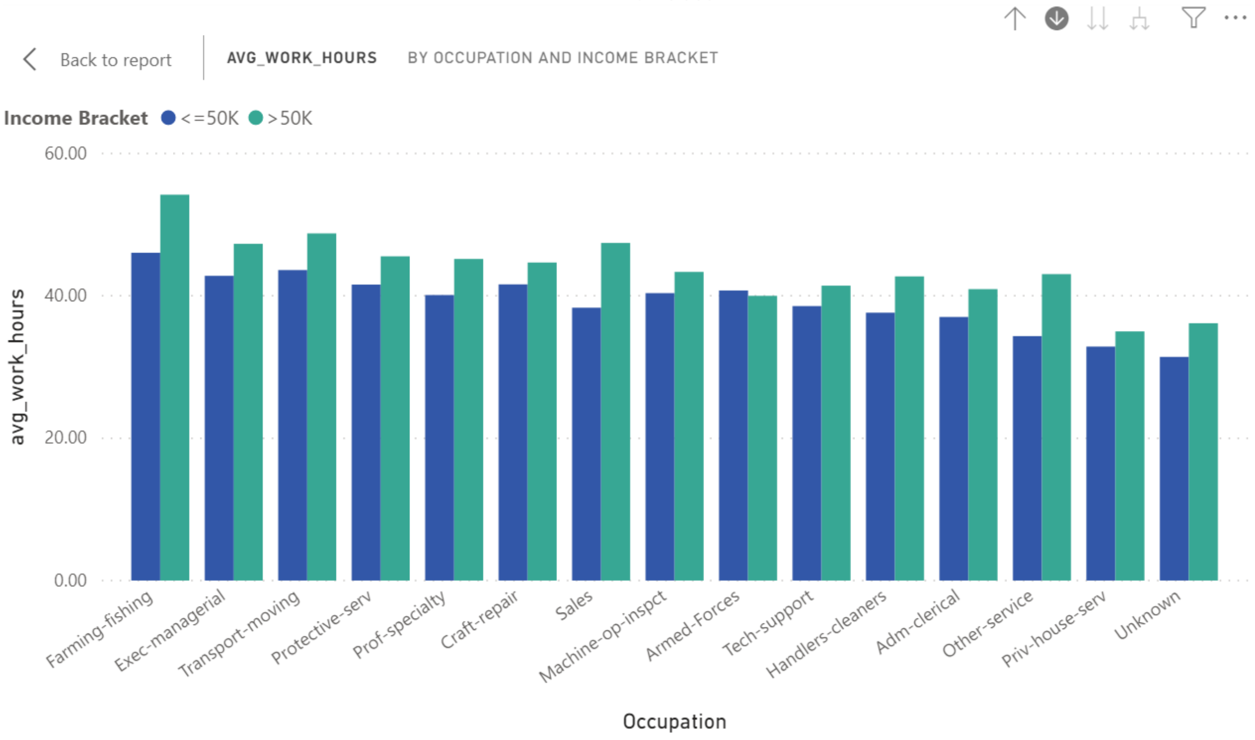
Income Bracket	avg_work_hours
>50K	45.47
<=50K	38.84
Total	**40.44**

From the pie chart, there is about 24% of US population with income of greater than 50K, while another 76% of the population are with income of lower than 50K. The real question is: do people who works longer hours tend to have higher income or vice versa? The table on the right-hand side suggests that the average work hours of adults with higher income (>50K) is considerably higher than that of adults with less income (<=50K). However, does this fact apply to people from all work classes? The below bar chart suggests that people with higher income usually work longer hours than those with lower income for all work classes, while self-employed adults tend to work the longest.

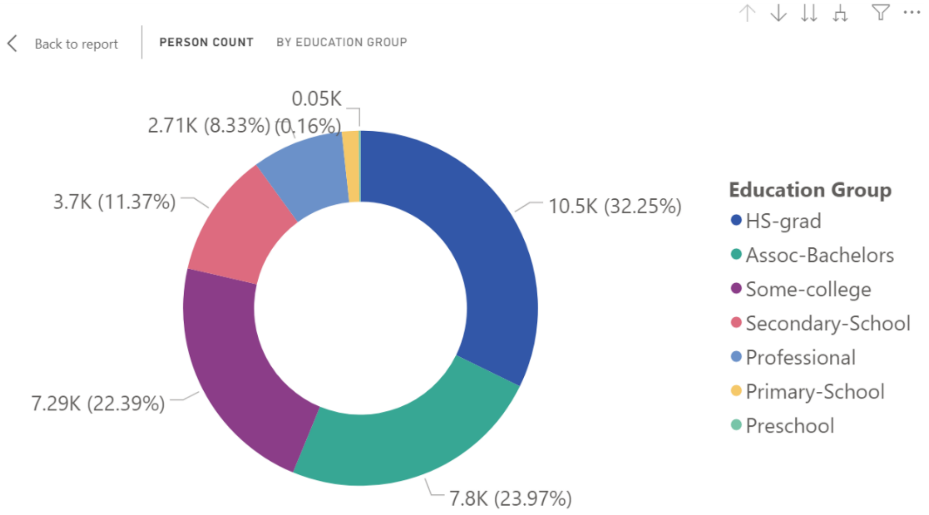


Drilling down a bit further to *workclass* and *occupation level*, we can see that people who work longer hours tend to have more income for all work classes. However, people in Armed-Forces occupation tend to experience the opposite.

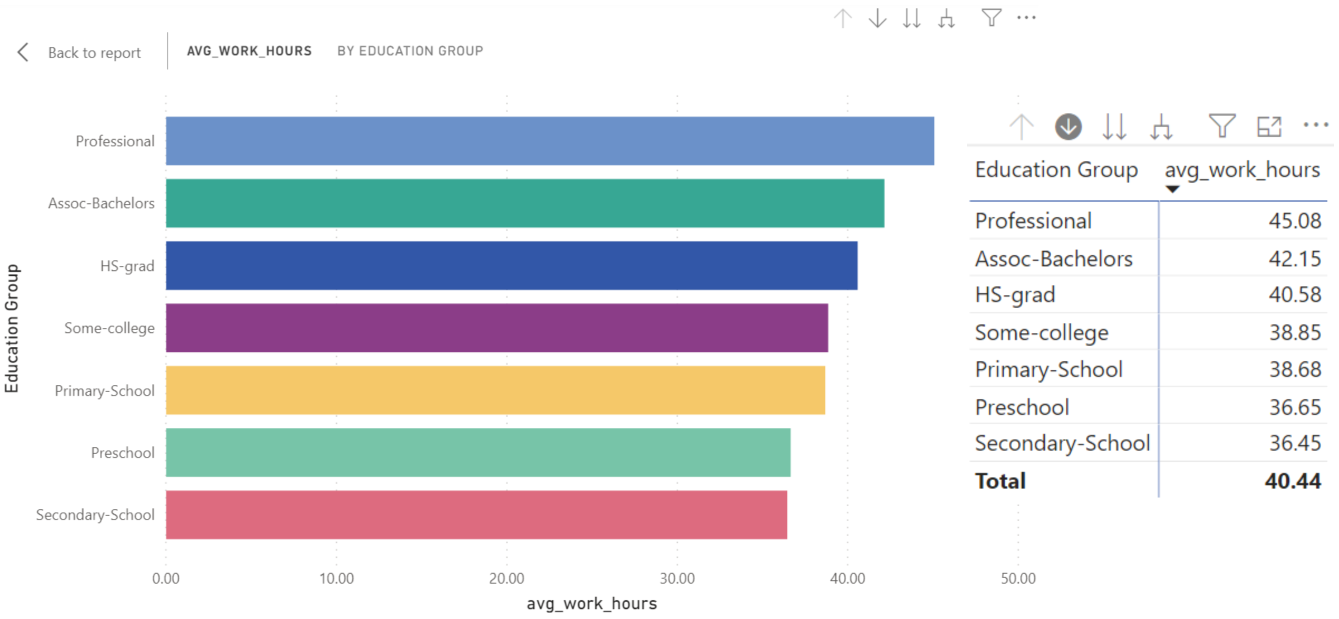




b. How does education level affect weekly work hours for certain income bracket?

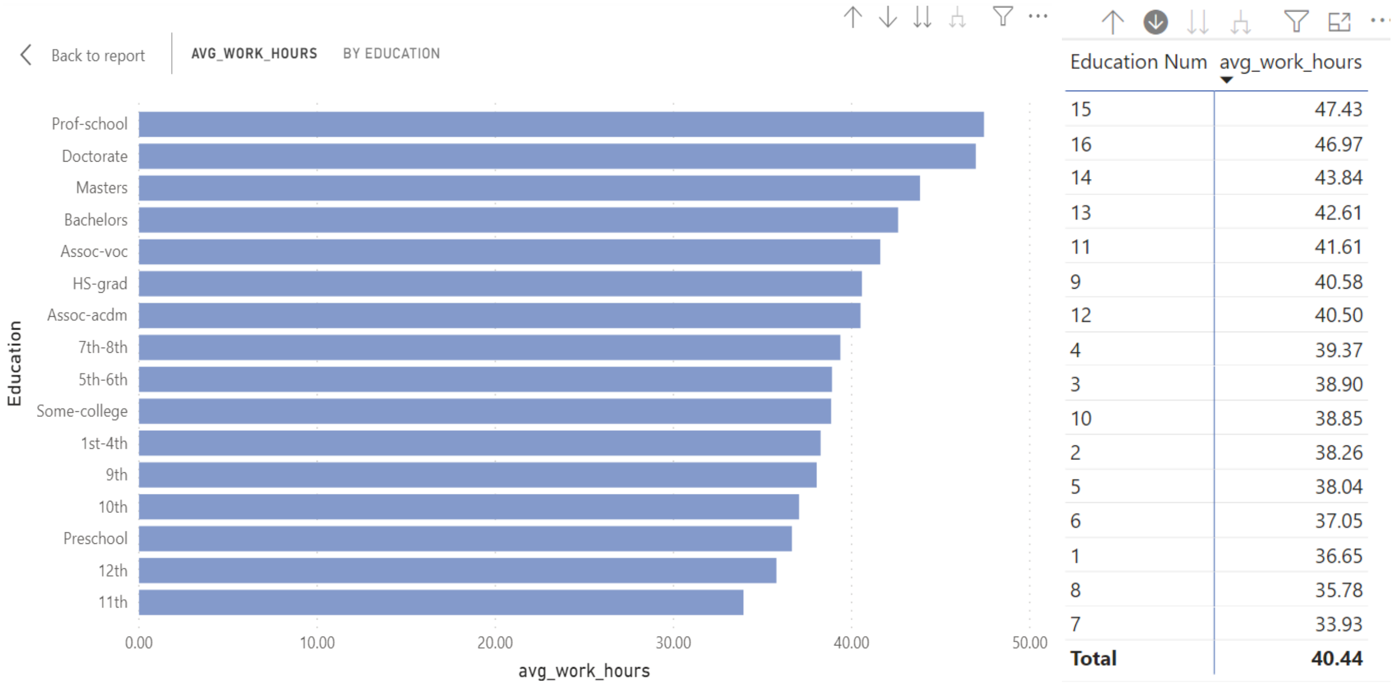


From the pie chart, most of the population have either a high school diploma (32.35%), associate-bachelor's degree (23.97%), or graduated from some college (22.39%). Only 8.33% of population graduated with professional degree (Doctorate, Master or Prof-school). However, they seem to work the longest (approximately 45 hours/week) comparing to other education groups as seen on the bar chart and the table below.



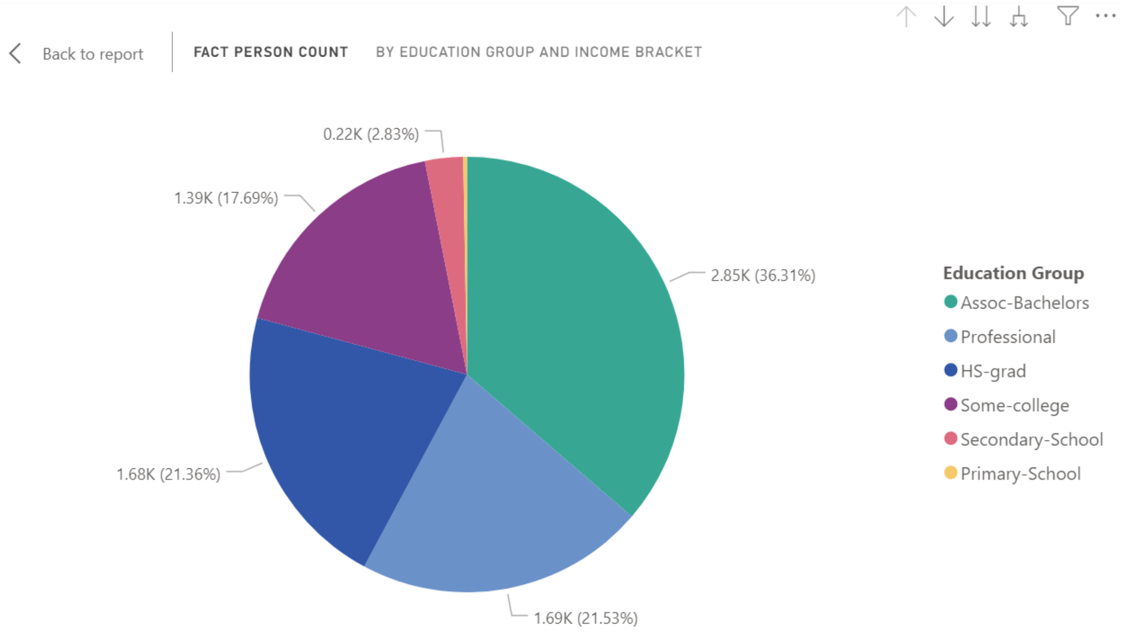
Education Group	avg_work_hours
Professional	45.08
Assoc-Bachelors	42.15
HS-grad	40.58
Some-college	38.85
Primary-School	38.68
Preschool	36.65
Secondary-School	36.45
Total	**40.44**

It seems like the higher the education level, the longer the number of hours they work per week. Notice that the number of work hours of people with high school diploma, associate-bachelor's degree and professional degree are all greater than the work-hour threshold recommended by the scientist. Let drilling down further to see whether this is the case for *education* and *education_num*.



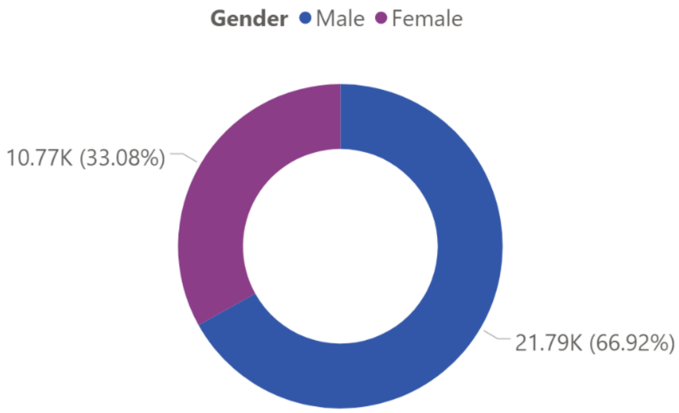
Education Num	avg_work_hours
15	47.43
16	46.97
14	43.84
13	42.61
11	41.61
9	40.58
12	40.50
4	39.37
3	38.90
10	38.85
2	38.26
5	38.04
6	37.05
1	36.65
8	35.78
7	33.93
Total	**40.44**

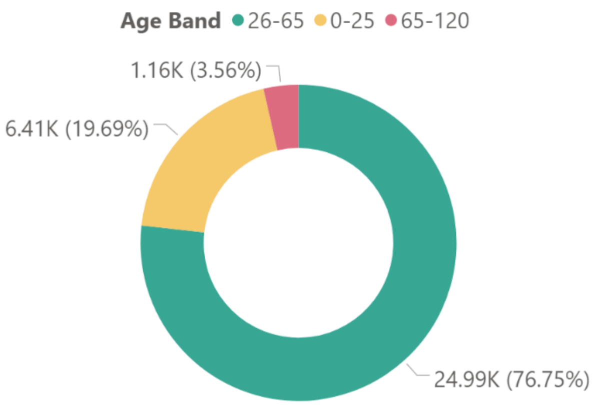
The bar chart and table above suggest that the more people study, the longer they work per week. At this state, we must be wondering why people study hard for a professional degree to work tirelessly 45 hours a week for approximately 40-50 years afterwards? Well, they probably get higher paid!



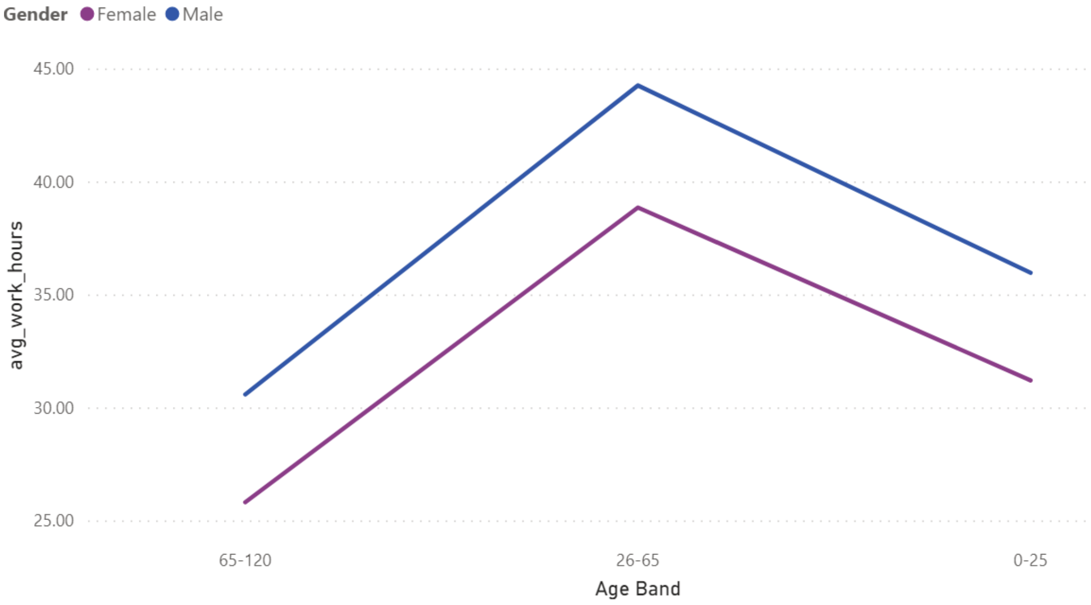
The pie chart above represents the percentage of people from each education group with income of greater than 50K. We can see that more than half of the population with higher paid are with either high school diploma, associate-bachelor's degree, or professional degree.

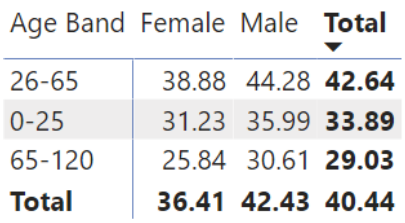
c. Are males' and females' work hours different for all age group?





The left pie chart suggests that there are more males than females in the US workplace back in 1994 as we can see that about two-thirds of the US adults income data population are males. The right pie chart represents the percentage of population from each age group. Most of the population (76.75%) were aged between 26-65 (working age) which is sensible.

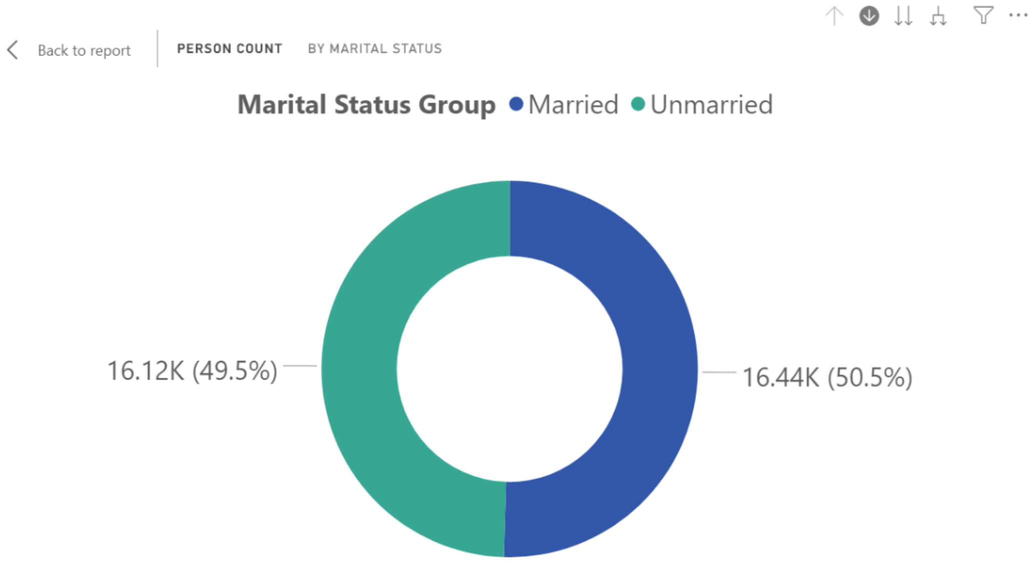




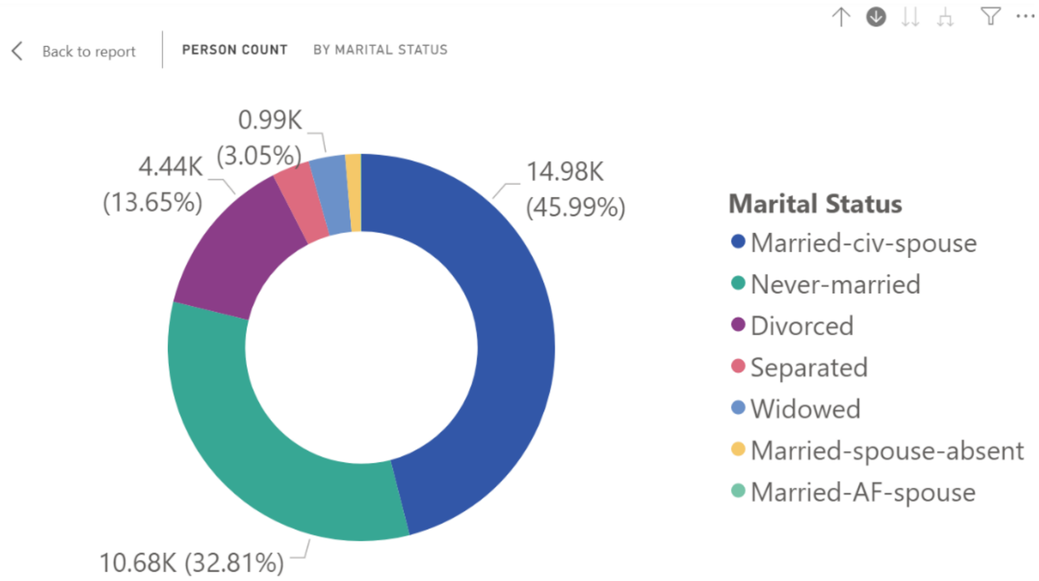
Age Band	Female	Male	**Total** ▼
26-65	38.88	44.28	**42.64**
0-25	31.23	35.99	**33.89**
65-120	25.84	30.61	**29.03**
Total	**36.41**	**42.43**	**40.44**

The above line chart illustrates the average work hours of males and females, respectively. We can see that males actually work longer than female for all age groups, while the average work hours of elders aged between 65-120 tend to be the lowest.

d. Do males and females from different marital status have different work hours?

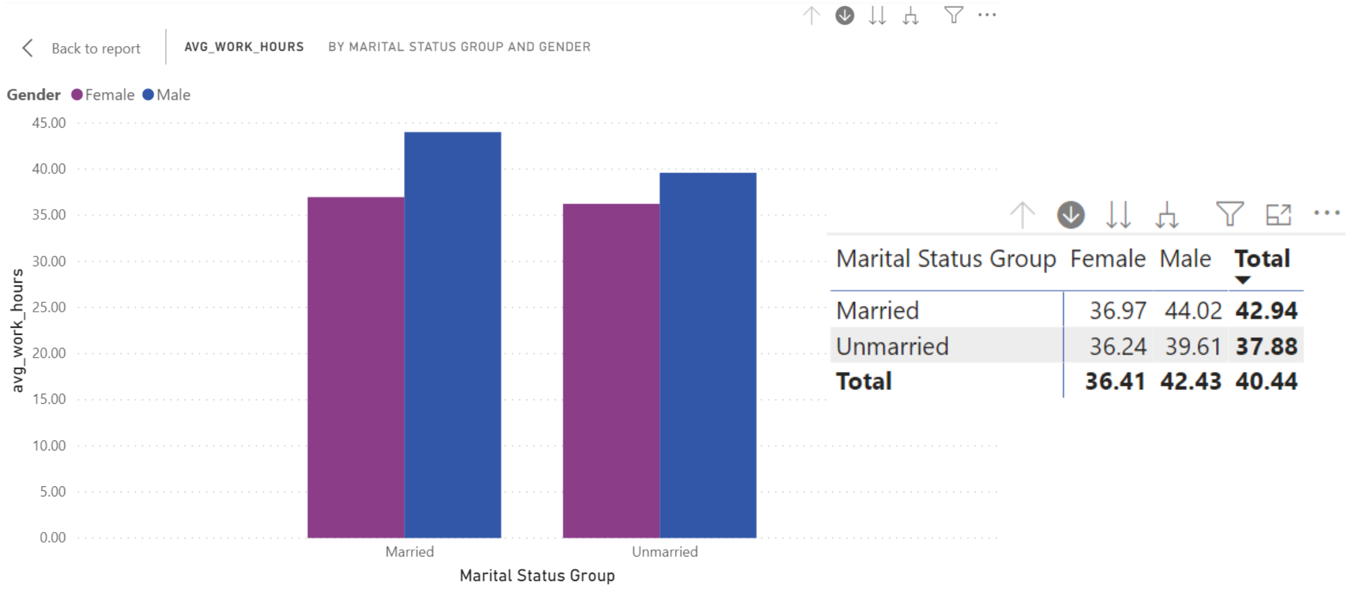


From the above pie chart, there are about the same number of married and unmarried people in the dataset. Let us drill down further to see which marital status actually has the highest proportion.



With the above pie chart, we can see that most of the dataset population had civil marriage with their spouse (46%), never married (32.81%), or divorced (13.65%), while less than 10% of the population are separated, widowed, or married with other conditions.

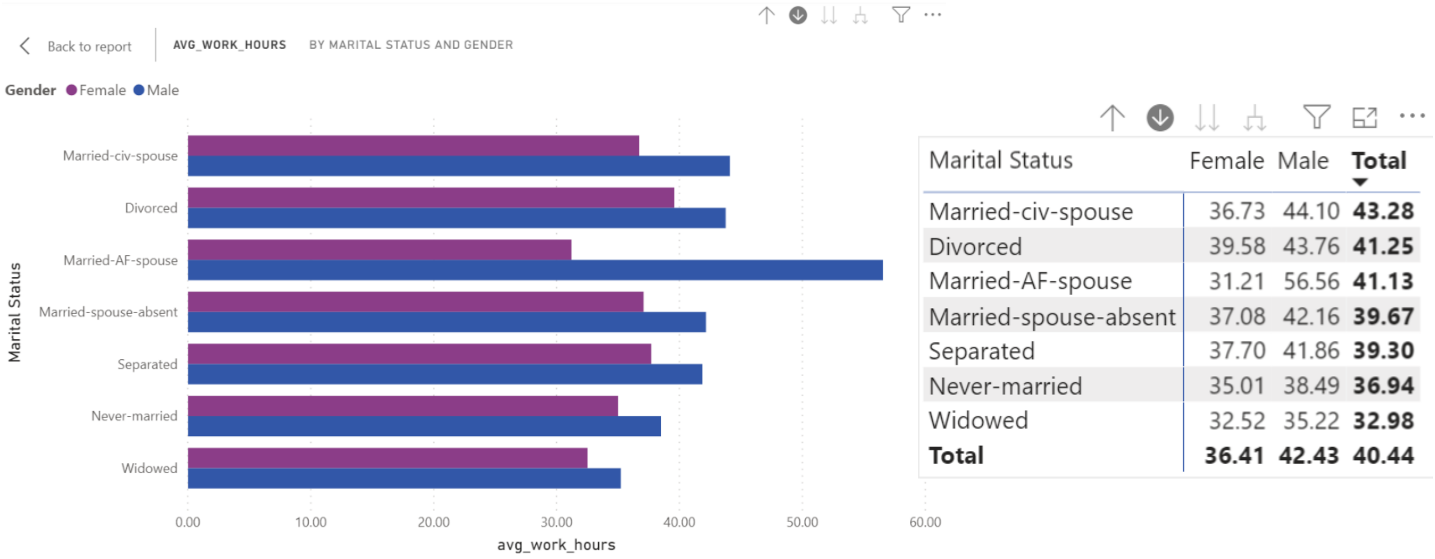
Next, we will have a look into the average work hours of people from different marital status and gender.



Marital Status Group	Female	Male	Total
Married	36.97	44.02	**42.94**
Unmarried	36.24	39.61	**37.88**
Total	**36.41**	**42.43**	**40.44**

The above table and bar chart suggest that married people work longer than unmarried people with average work hours of 42.94 hours/week, while married men tend to work the longest having average work hours of 44.02 hours/week.

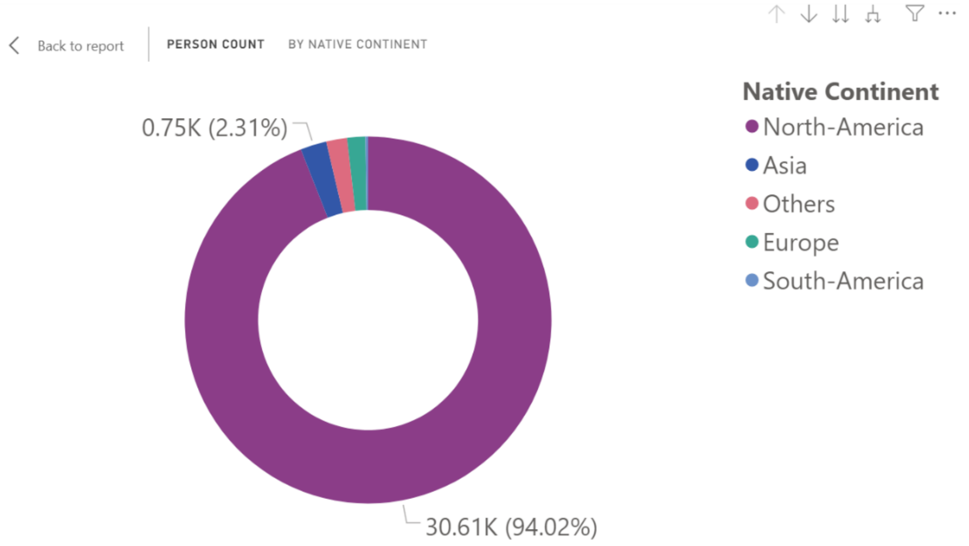
Drilling down further, the below table and bar chart suggest about the same information, showing that married males, on average, have extended hours of work, especially males who are married to armed forces spouse (Married-AF-spouse) with average work hours of 56.56 hours/week, which is significantly different than females with the same marital status



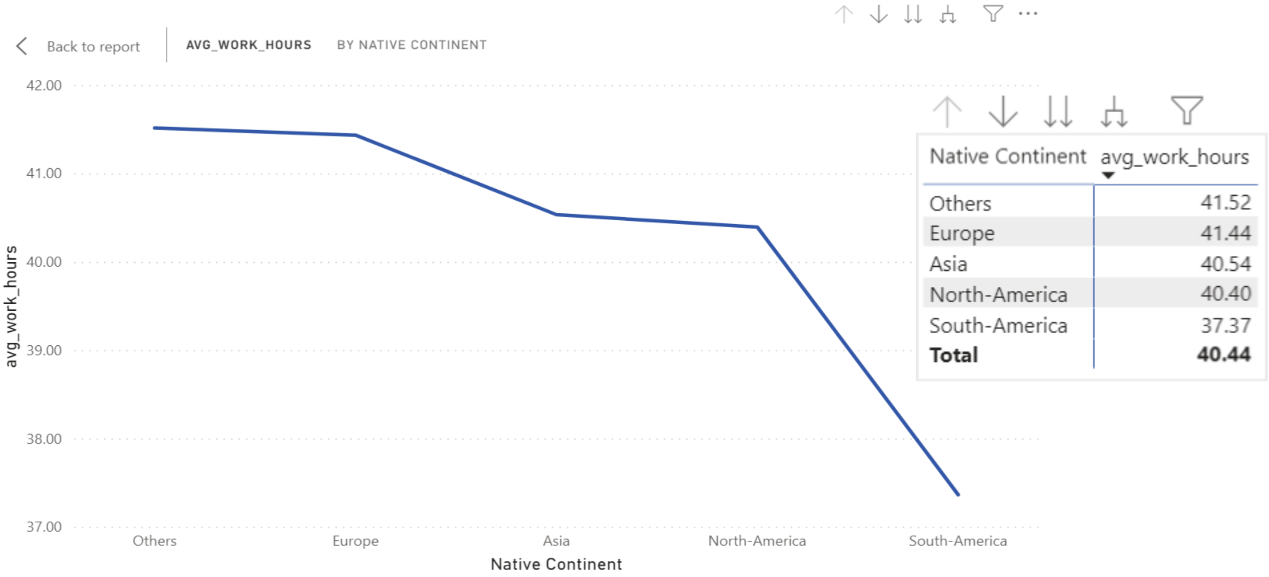
Marital Status	Female	Male	Total
Married-civ-spouse	36.73	44.10	**43.28**
Divorced	39.58	43.76	**41.25**
Married-AF-spouse	31.21	56.56	**41.13**
Married-spouse-absent	37.08	42.16	**39.67**
Separated	37.70	41.86	**39.30**
Never-married	35.01	38.49	**36.94**
Widowed	32.52	35.22	**32.98**
Total	**36.41**	**42.43**	**40.44**

e. Do people originated outside North America have higher or lower work hours per week when comparing to those native to North America?

It’s a controversial fact that immigrants work harder to survive in the country that is not their homeland. Now, we will see whether this is the case.

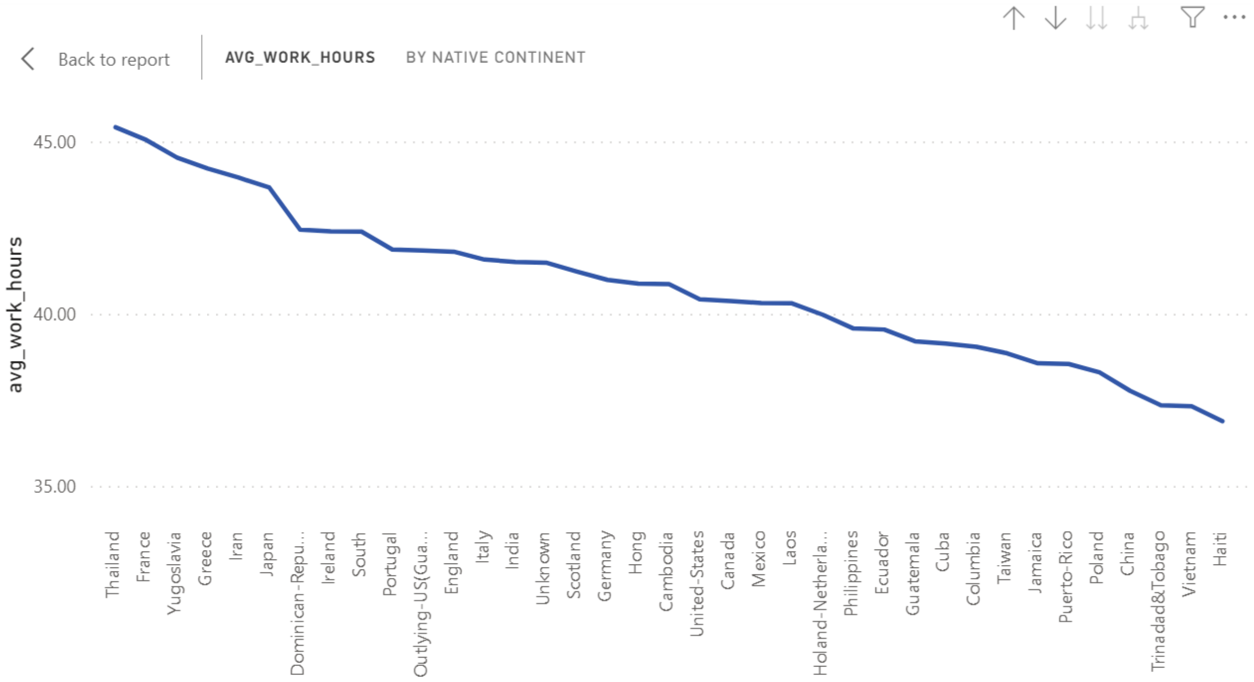


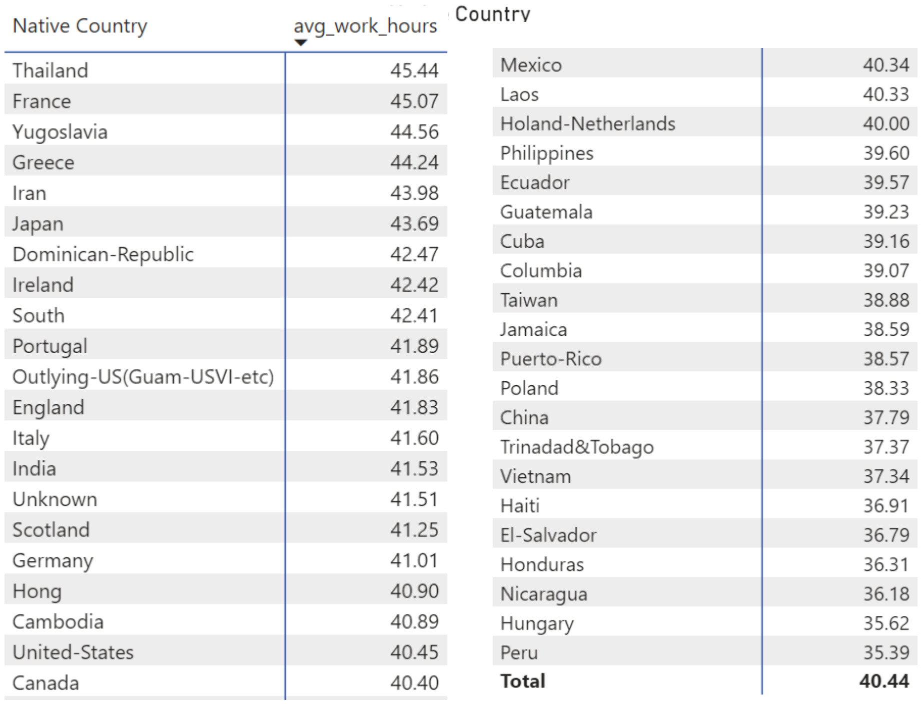
Since this is a US income data, most of the population (94.02%) are from North America, while another 6% are from Asia, Europe, South America and other continents.



Native Continent	avg_work_hours
Others	41.52
Europe	41.44
Asia	40.54
North-America	40.40
South-America	37.37
Total	**40.44**

The line chart and table above illustrate the average work hours of people originated from each continent. People from other continents and Europe tend to have highest average work hours of 41.52 and 41.44 per week, respectively. People originated from Asia and North America have about the same average work hours of 40.54 and 40.40 per week, respectively. Lastly, people from South America tend to have lowest average work hours of 37.37 per week.





Native Country	avg_work_hours
Thailand	45.44
France	45.07
Yugoslavia	44.56
Greece	44.24
Iran	43.98
Japan	43.69
Dominican-Republic	42.47
Ireland	42.42
South	42.41
Portugal	41.89
Outlying-US(Guam-USVI-etc)	41.86
England	41.83
Italy	41.60
India	41.53
Unknown	41.51
Scotland	41.25
Germany	41.01
Hong	40.90
Cambodia	40.89
United-States	40.45
Canada	40.40
Mexico	40.34
Laos	40.33
Holand-Netherlands	40.00
Philippines	39.60
Ecuador	39.57
Guatemala	39.23
Cuba	39.16
Columbia	39.07
Taiwan	38.88
Jamaica	38.59
Puerto-Rico	38.57
Poland	38.33
China	37.79
Trinadad&Tobago	37.37
Vietnam	37.34
Haiti	36.91
El-Salvador	36.79
Honduras	36.31
Nicaragua	36.18
Hungary	35.62
Peru	35.39
Total	**40.44**

Drilling down to native country level, we can see that people originated from Thailand had the highest weekly work hours (45.44), while France (45.07) and Yugoslavia (44.56) came as the second and the third place, respectively. However, people originated from Nicaragua, Hungary and Peru tend to have the lowest weekly work hours with average work hours of 36.18 per week, 35.62 per week and 35.39 per week, respectively.

6. References

Verbeeck, J. (2017, June). Analysis Services (SSAS) Multidimensional Design Tips

Relations and Hierarchies. Retrieved from https://www.sqlshack.com/analysis-services-ssas-multidimensional-design-tips-relations-hierarchies/.

Microsoft SQL Server 2008 Documentation. (2012). Lesson 6: Defining Calculated Members. Retrieved from https://docs.microsoft.com/en-us/previous-versions/sql/sql-server-2008-r2/ms166568%28v%3dsql.105%29.

Hila, DG. (2011). Avoid visible attribute hierarchies for attributes used as levels in user-defined hierarchies. Retrieved from https://hdg-thefish.blogspot.com/2011/12/avoid-visible-attribute-hierarchies-for.html

Dinh, H., Strazdins, L., and Welsh, J. (2017, March). Hour-glass ceilings: Work-hour thresholds, gendered health inequities. https://doi.org/10.1016/j.socscimed.2017.01.024