A	8/20
B	11/20
C	11/20
D	12/20
E	12/20
F	5/20
G	2/20
H	7/20
I	3/20
J	1/20
K	9/20
L	8/20
M	1/20

1).

{A}, {B}, {C}, {D}, {E}, {H}, {K}, {L}

AB	5/20
AC	4/20
AD	4/20
AE	4/20
AH	4/20

AK	3/20
AL	3/20
BC	7/20
BD	8/20
BE	8/20
BH	1/20
BK	5/20
BL	3/20
CD	7/20
CE	6/20
CH	3/20
CK	4/20
CL	2/20
DE	7/20
DH	2/20
DK	5/20
DL	4/20
EH	2/20
EK	4/20
EL	5/20
HK	4/20
HL	3/20

KL	2/20

{BC}, {BD}, {BE}, {CD}, {CE}, {DE}

BCD	5/20
BCE	6/20
BDE	5/20
CDE	4/20

{BCE}

2).

B->C	0.64
C->B	0.64
B->D	0.73
D->B	0.67
B->E	0.73
E->B	0.67
C->D	0.64
D->C	0.58
C->E	0.55
E->C	0.5
D->E	0.58
E->D	0.58

B->CE	0.55
C->BE	0.55
E->BD	0.5
CE->B	1
BE->C	0.75
BC->E	0.86

{CE->B}, {BE->C}, {BC->E}

3). {BCD->E}: 80%

4) {CE->B}, {BE->C}, {BC->E},

{BCD->E}, {BDE->C}, {CDE->B}, {BCD->E}, …

It is not reasonable because if an association rule has low support but high confidence, the element set in this rule is unlikely to appear.

2.

1) <{a}>, <{b}>,<{c}>, <{g}>, <{d}>, <{a}, {c}>, <{c}, {a}>, <{g}, {c}>, <{c},{g}>, <{b}, {a}>
2) In sequential pattern mining, we must consider the order of elements.

   For example, when people may buy different kinds of kitchenware at same time, but there is no specific order.

   When people rent a series of movies, they tend to rent them in order.

3.

1) **Plot 1:**

There is no obvious outlier and sudden shifts.

There is no obvious upward or downward trend.

The data show a seasonal pattern. The pattern repeats every 10 years.

Plot 2:

There is a sudden shift occurred at about 100 days

There is downward trend.

The data show a cyclic movement. The cycles don't repeat at regular intervals and don't have the same shape.

Plot 3:

There is no obvious outlier and sudden shifts.

There is upward trend.

The data show a cyclic movement. The cycles repeat at regular intervals and the variation becomes larger and larger.

Plot 4:

There are some outliers in the plot.

There is no obvious upward or downward trend.

The data show a cyclic movement. The cycles don't repeat at regular intervals and don't have the same shape. The values of variable change very frequently.

2) Yes, cyclical component is easier to estimate because cyclical movement don't need fixed period. The cyclical patterns are usually more obvious, while seasonality

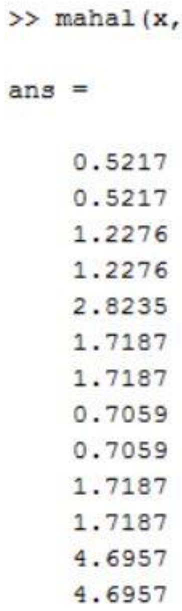
requires fixed periods.

In terms of standards, the seasonality is stricter. Also, cyclical movement and seasonality may happen together, and the cyclical movement can cover the pattern of seasonality in some extent, which make it hard to estimate seasonality.

3) Data Transformation, Multi-dimensional indexing, subsequence matching.

4.

1）centroid (3,5)



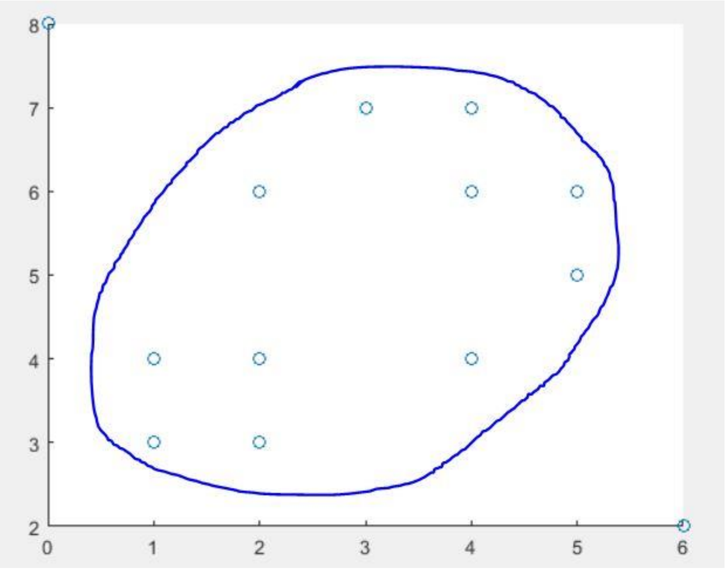
```
>> mahal(x,x)

ans =

    0.5217
    0.5217
    1.2276
    1.2276
    2.8235
    1.7187
    1.7187
    0.7059
    0.7059
    1.7187
    1.7187
    4.6957
    4.6957
```

(0,8) and (6,2) are outliers

2)



3) The LOF may behave differently because the sample space is small so that LOF is very sensitive to the K

5

1） Transparency

DOD principles mentioned transparency in “Traceable”: “including with transparent and auditable methodologies, data sources and design procedures and documentation.” This satisfies the requirement of transparency, but it didn’t mention if the origin of data source should be notified.

2） Individual Participation

DOD principles didn’t mention “individual Participation”. The principles focus on how the use and development side of AI should follow ethics, but it didn’t specify how to ensure the rights of individuals whose data is collected by AI.

3） Purpose Specification

In “Reliable”, DOD principles set requirement on “Purpose Specification”: “the department's AI capabilities will have explicit, well-defined uses”. This meets the standard in FIPP.

4） Data Minimization

DOD principles didn’t show any requirement on “Data Minimization”. The principles only showed that data source must be transparent, but it didn’t declare the limitation of data collection and storage.

5） Use Limitation

Though the principles said, “The department's AI capabilities will have explicit, well-defined uses”, there is no clear limitation on the use of AI. Defining use is not equal to Limiting use.

6） Data Quality and Integrity

DOD principles showed that data source must be transparent and auditable, but it didn’t ensure the accuracy and correctness of data. Thus, the principles didn’t meet the requirement of “data quality and integrity”

7） Security

The “Governable” mentioned the “the ability to detect and avoid unintended consequences”, but it put more emphasis on the use of AI than the security of data collected by AI.

8） Accountability and Auditing

The first principle “Responsible” declared “DOD personnel will exercise appropriate levels of judgment and care while remaining responsible for the development, deployment and use of AI capabilities.” This principle clearly define the responsible party, which satisfied “Accountability and Auditing”.

2 Data Understanding

2.1 Collect Initial Data

Initial Data Collection Report

The dataset is acquired from IPUMS-USA: Version 8.0 Extract of 1940 Census for U.S. Census Bureau Disclosure Avoidance Research. This dataset included a full extract of 1940 Census. The dataset is retrieved from IPUMS-USE: https://usa.ipums.org/usa/1940CensusDASTestData.shtml

2.2 Describe Data

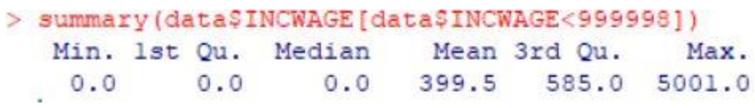
Data Description Report

There are 1351732 records in the dataset with 177 fields including YEAR, SAMPLE, SERIAL, NUMPREC, etc. The detailed meaning of fields and values can be found on https://usa.ipums.org/usa-action/variables/group?id=income

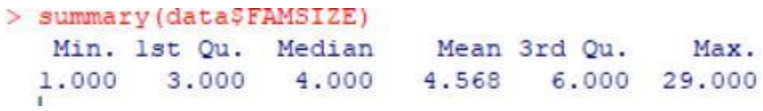
The dataset included the income, work, birthplace and many other personal information that is helpful for the project.

2.3 Explore Data

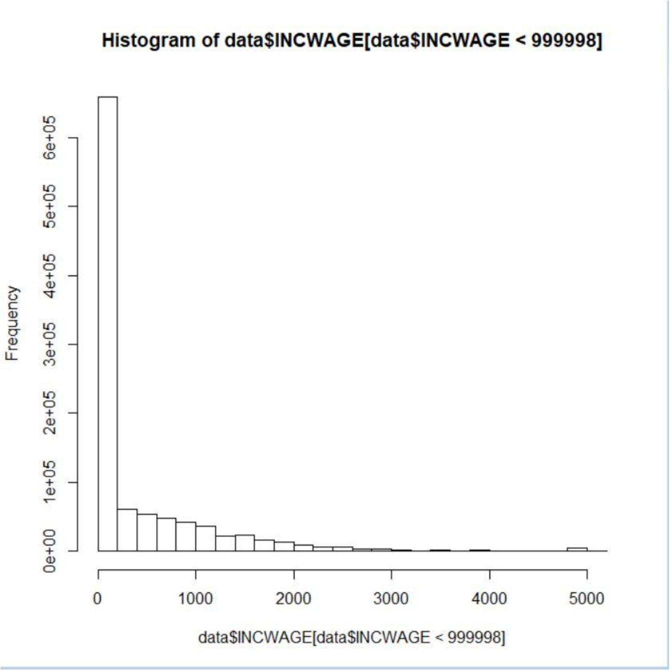
To achieve the goal of the project, we are interested in the income wage and family size in the dataset. Here are simple summaries of INCWAGE and FAMSIZE in 1940 Census:

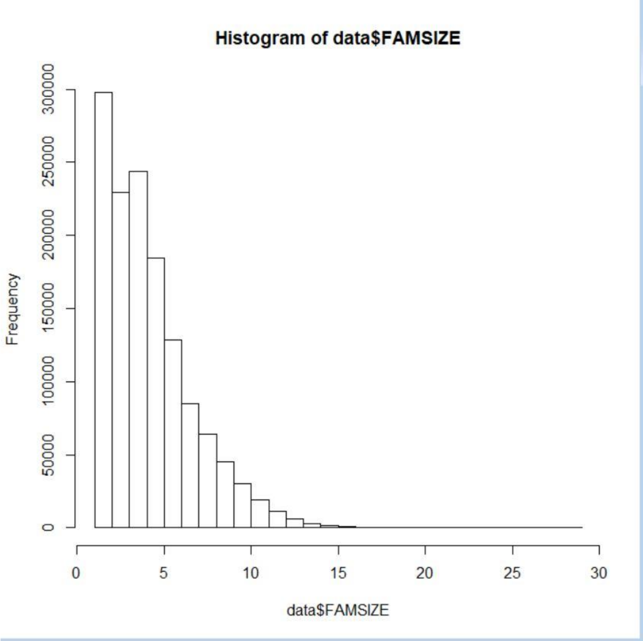


```
> summary(data$INCWAGE[data$INCWAGE<999998])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0     0.0     0.0   399.5   585.0  5001.0
```



```
> summary(data$FAMSIZE)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   3.000   4.000   4.568   6.000  29.000
```





Based on this data, we can analyze the poverty rate, and other fields like age and race are also easy to acquire in the dataset.

2.4 Verify Data Quality

The samples of data are 100% selected from IPUMS-USA 1940 database. The dataset is also used by The U.S. Census Bureau. Therefore, we are confident that

this data is authoritative and representative.

3.1 Select Data

Based on our project goals, we need to analyze characteristics such as income, age, race, gender, family size, etc. Here are part of the characteristics and their corresponding variable name in the data set.

Characteristics	Variable Name in dataset
Income	INCWAGE
Age	AGE
Race	RACE
Gender	SEX
Family Size	FAMSIZE
...	...

3.2 Clean Data

In the dataset, some variables have special code to represent missing, N/A or other problems. For example, in INCWAGE (income wage), 999999 = N/A, 999998 = missing. To clean the data, we must filter these special codes for our further analysis. Codes Description of all variables can be acquired at https://usa.ipums.org/usa-action/variables/