DABI ASSIGNMENT 1

AY20/21 Oct Semester

DECLARATION

I declare that I am the originator of this work and that all other original sources used in this work have been appropriately acknowledged.

I understand that plagiarism is the act of taking and using the whole or any part of another person's work and presenting it as my own without proper acknowledgement.

I also understand that plagiarism is an academic offence and that disciplinary action will be taken for plagiarism."

☑ I Agree (Please Tick ✓)

My Information

Name (as in matriculation card)	Wee Kar Ghee
Admin Number	2080985A
Group (1, 2, 3 or 3)	2
Task selected (A or B)	A

For Tutor Use

Overall Grade:	
Feedback on Task Performance	
Feedback on proposed application area	

Performance of Pattern Discovery Task

Data Cleaning

For TaskA, we will be using the listings-Task A.xlsx.

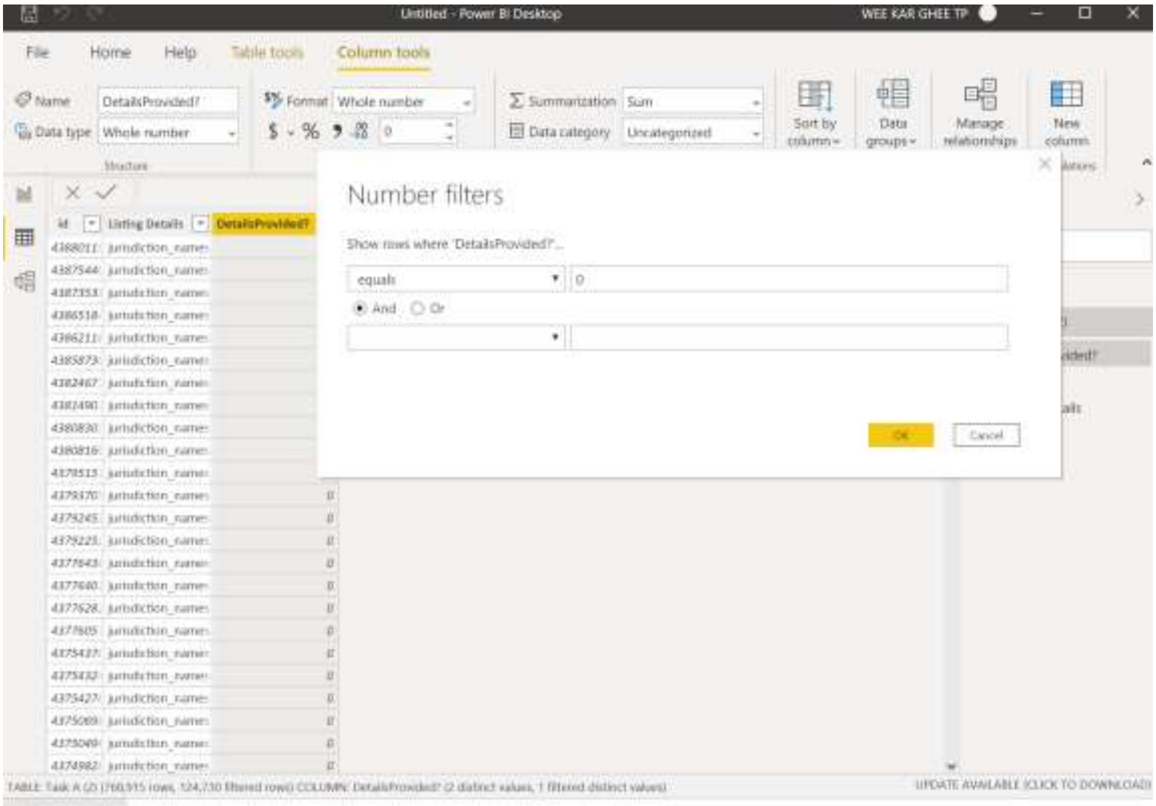
There are 768,916 rows and 3 columns.
The three variables are ‘id’, ‘Listing Details’ and ‘DetailsProvided’.
Values in the ‘DetailsProvided’ are ‘0’ and ‘1’. ‘0’ interpreted as ‘not provided’ while ‘1’ means ‘provided’.

There was initial attempt to filter ‘1’s but dataset was too big and there was lag time when attempt to run in SAS Enterprise miner as we are using via VDI. Hence, it was decided to remove the ‘1’s.

Hence, this study will based on ‘0’s with 120k+ rows.

The tool used for Data Cleaning was Microsoft Power BI.

- Open the Listings-Task A.xlsx in Power BI.
- Use Query to filter values dynamically. Set Filter Rows condition as equal to ‘0’.



Name: DetailsProvided?
Data type: Whole number
Format: Whole number
Summarization: Sum
Data category: Uncategorized

Number filters

Show rows where 'DetailsProvided?'...

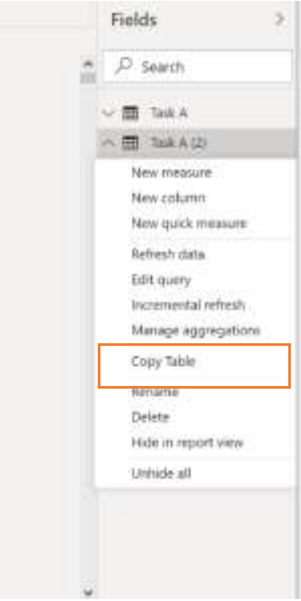
equals ▾ | 0

● And ○ Or

id	Listing Details	DetailsProvided?
4388011	jurisdiction_name	
4387544	jurisdiction_name	
4387353	jurisdiction_name	
4386518	jurisdiction_name	
4386211	jurisdiction_name	
4385873	jurisdiction_name	
4382467	jurisdiction_name	
4382490	jurisdiction_name	
4380830	jurisdiction_name	
4380816	jurisdiction_name	
4378513	jurisdiction_name	
4379370	jurisdiction_name	0
4379245	jurisdiction_name	0
4379225	jurisdiction_name	0
4377643	jurisdiction_name	0
4377640	jurisdiction_name	0
4377628	jurisdiction_name	0
4377605	jurisdiction_name	0
4375437	jurisdiction_name	0
4375432	jurisdiction_name	0
4375427	jurisdiction_name	0
4375069	jurisdiction_name	0
4375049	jurisdiction_name	0
4374982	jurisdiction_name	0

TABLE: Task A (2) (768,915 rows, 124,730 filtered rows) COLUMN: DetailsProvided? (2 distinct values, 1 filtered distinct values)

UPDATE AVAILABLE (CLICK TO DOWNLOAD)



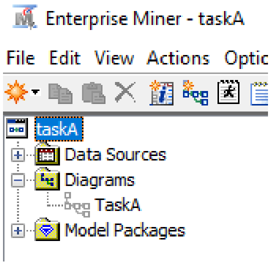
After that, click on the ‘copy table’ on the right and paste it to a new excel sheet. Name as ‘TaskA_cleaned.xlsx.’

Next is to find out which service on the property listed on Airbnb along with what other service(s).

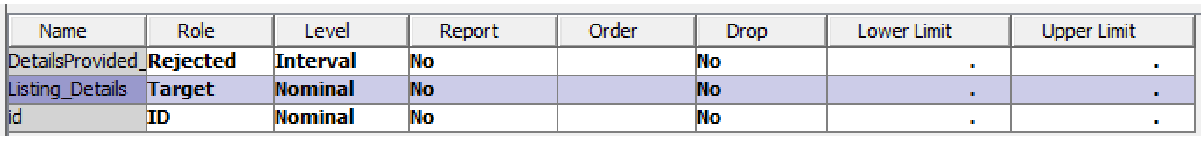
The following steps and nodes were performed:

1. Create a new SAS EM Project ‘taskA’.
2. Create a new SAS EM Diagram and read the file into SAS EM using the File Import Node.

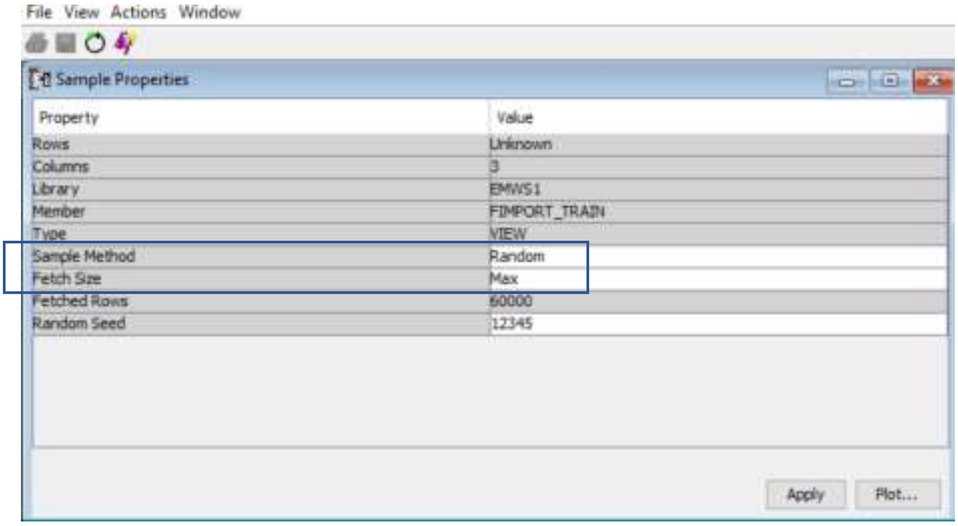
File > New Diagram (name it ***TaskA)***



The level of measurement for id and Listing_Details will be nominal as they are reference for property listing and information respectively. The DetailsProvided are values and irrelevant as it had been filtered and contained one number.



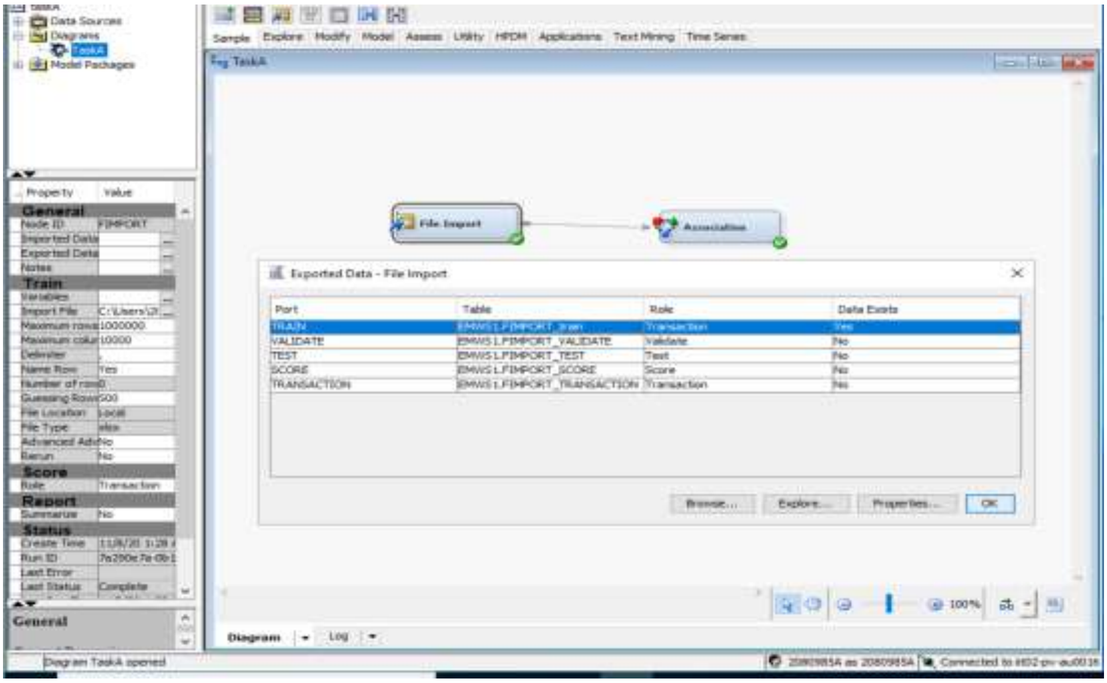
Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
DetailsProvided_	Rejected	Interval	No		No	.	.
Listing_Details	Target	Nominal	No		No	.	.
id	ID	Nominal	No		No	.	.



Property	Value
Rows	Unknown
Columns	3
Library	EMWS1
Member	FIMPORT_TRAIN
Type	VIEW
Sample Method	Random
Fetch Size	Max
Fetched Rows	60000
Random Seed	12345

In Sample Properties seen above, Sample Method and Fetch Size are changed to Random and Max respectively as dataset is big. More rows are fetched for exploration.

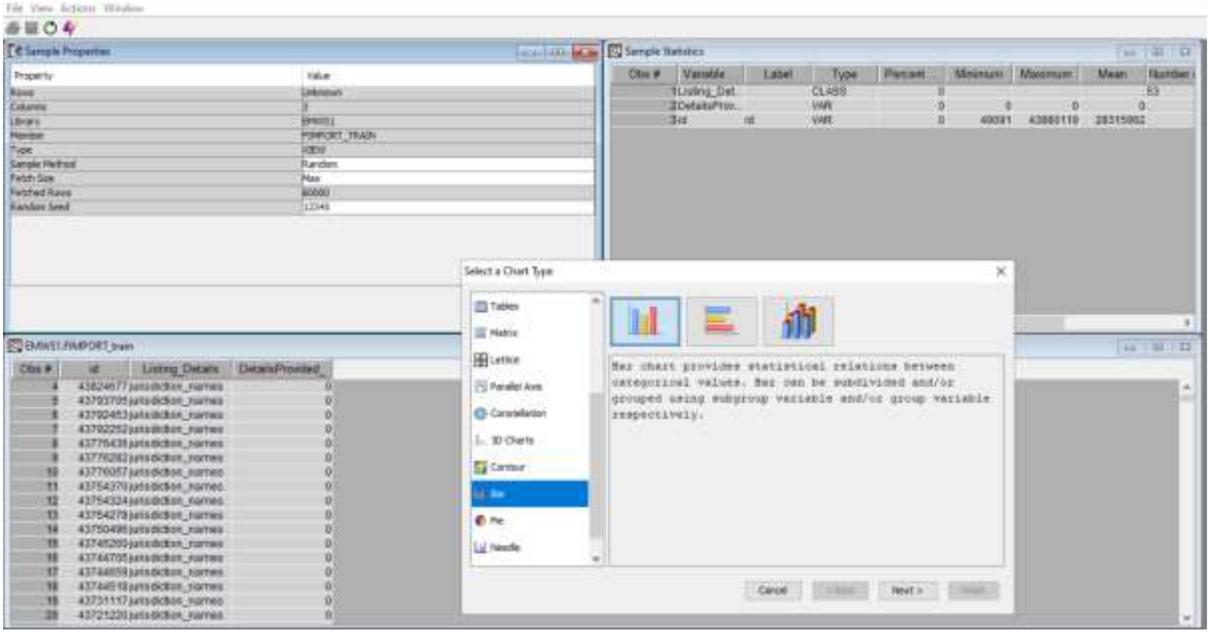
Run the File Import node. Once completed, click on the Exported Data in the property pane, select the DATA port and click on the Explore button to explore the exported data.



Property	Value
General	
Node ID	FIMPORT
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Import File	C:\Users\2(
Maximum rows	1000000
Maximum colur	10000
Delimiter	,
Name Row	Yes
Number of row	0
Guessing Rows	500
File Location	Local
File Type	xlsx
Advanced Advi	No
Rerun	No
Score	
Role	Transaction
Report	
Summarize	No
Status	
Create Time	11/8/20 1:28
Run ID	7a200e7a-0b1
Last Error	
Last Status	Complete

Port	Table	Role	Data Exists
TRAIN	EMWS1.FIMPORT_train	Transaction	Yes
VALIDATE	EMWS1.FIMPORT_VALIDATE	Validate	No
TEST	EMWS1.FIMPORT_TEST	Test	No
SCORE	EMWS1.FIMPORT_SCORE	Score	No
TRANSACTION	EMWS1.FIMPORT_TRANSACTION	Transaction	No

Diagram TaskA opened



Sample Properties

Property	Value
Rows	Unknown
Columns	3
Library	EMWS1
Member	FIMPORT_TRAIN
Type	VIEW
Sample Method	Random
Fetch Size	Max
Fetched Rows	60000
Random Seed	12345

Sample Statistics

Obs #	Variable	Label	Type	Percent ...	Minimum	Maximum	Mean	Number ...
1	Listing_Det...		CLASS	0				53
2	DetailsProv...		VAR	0	0	0	0	
3	id	id	VAR	0	49091	43880119	28315062	

EMWS1.FIMPORT_train

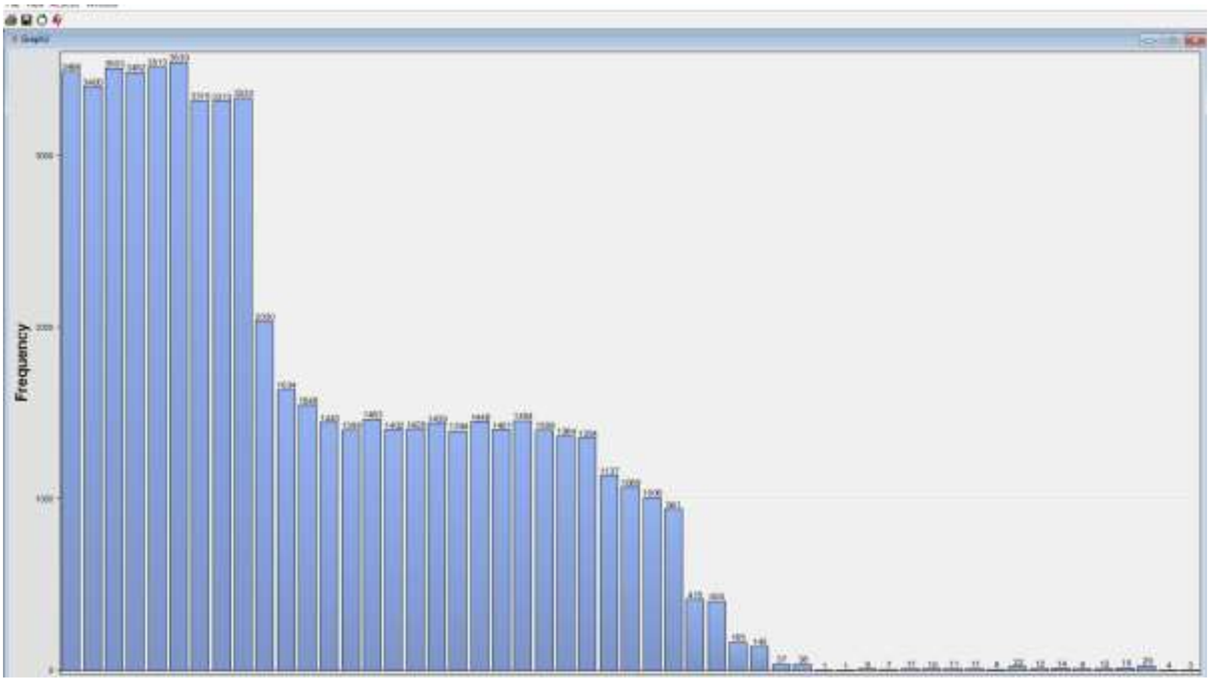
Obs #	id	Listing_Details	DetailsProvided
4	43824677	jurisdiction_names	0
5	43793705	jurisdiction_names	0
6	43792453	jurisdiction_names	0
7	43792252	jurisdiction_names	0
8	43776438	jurisdiction_names	0
9	43776282	jurisdiction_names	0
10	43776057	jurisdiction_names	0
11	43754370	jurisdiction_names	0
12	43754324	jurisdiction_names	0
13	43754278	jurisdiction_names	0
14	43750495	jurisdiction_names	0
15	43745260	jurisdiction_names	0
16	43744705	jurisdiction_names	0
17	43744659	jurisdiction_names	0
18	43744518	jurisdiction_names	0
19	43731117	jurisdiction_names	0
20	43721220	jurisdiction_names	0

Select a Chart Type

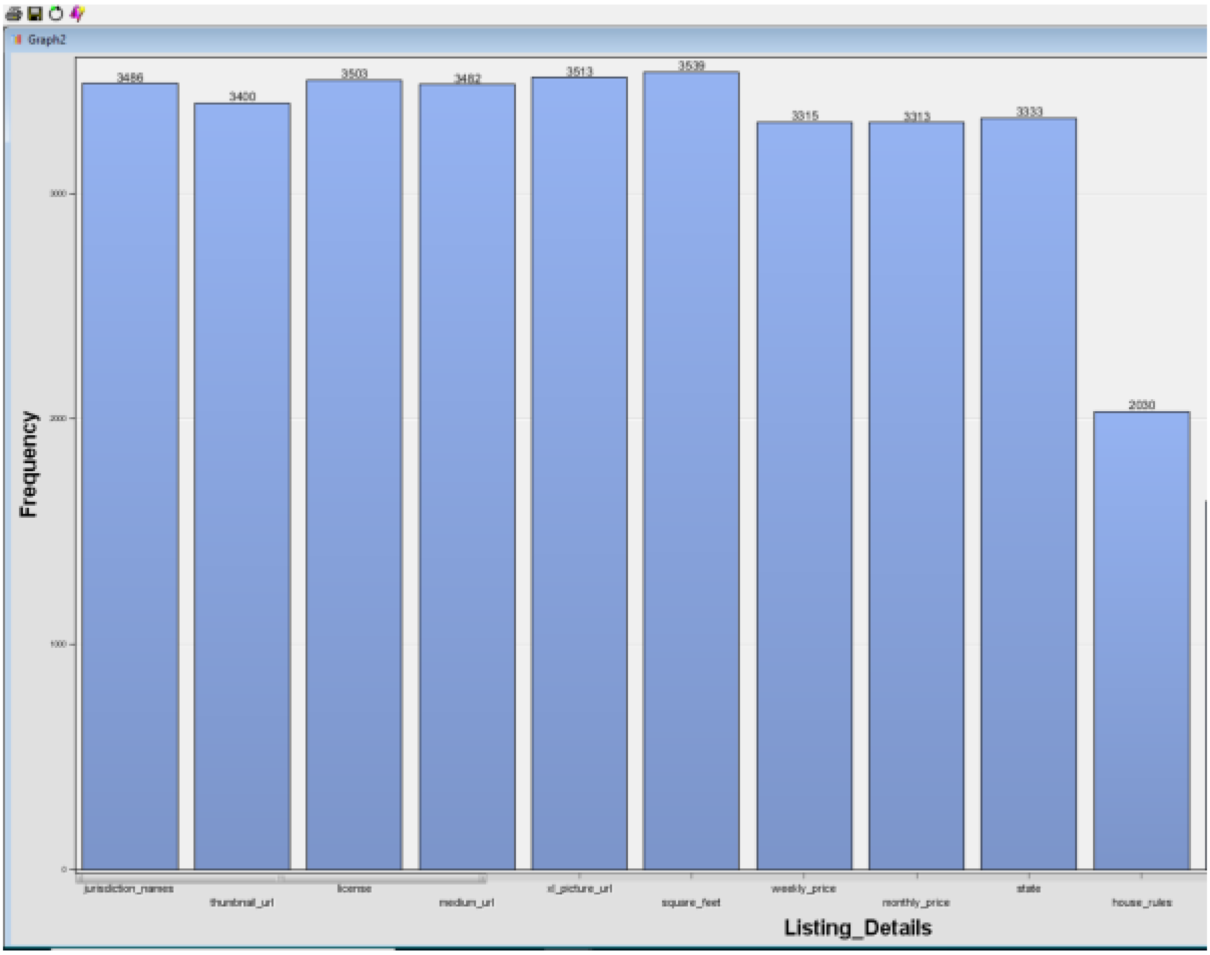
Bar chart provides statistical relations between categorical values. Bar can be subdivided and/or grouped using subgroup variable and/or group variable respectively.

Next, in resulting window, click on the Plot icon.

- Select Bar chart and click Next
- Set the Chart Role of Listing_Details to Category (X-Axis) and response statistic as Frequency and click Finish.



Overall view of all the variables in Listing_Details



Top 10 with highest frequency are :

Listing_Details	Frequency
Jurisdiction_names	3486
Thumbnail_url	3400
license	3503
Medium_url	3482
xl_picture_url	3513
square_feet	3539
weekly_price	3315
monthly_price	3313
state	3333

Connect an Association node (found in Explore group) to the File Import node.

• Check the properties for the Association node: ○ Maximum Item = 4 – ***This limits our rules to contain at most 4 items and will eliminate some of the rules produced / explored.***

○ Minimum Confidence = 50% - ***This will eliminate some of the rules which are not useful.***

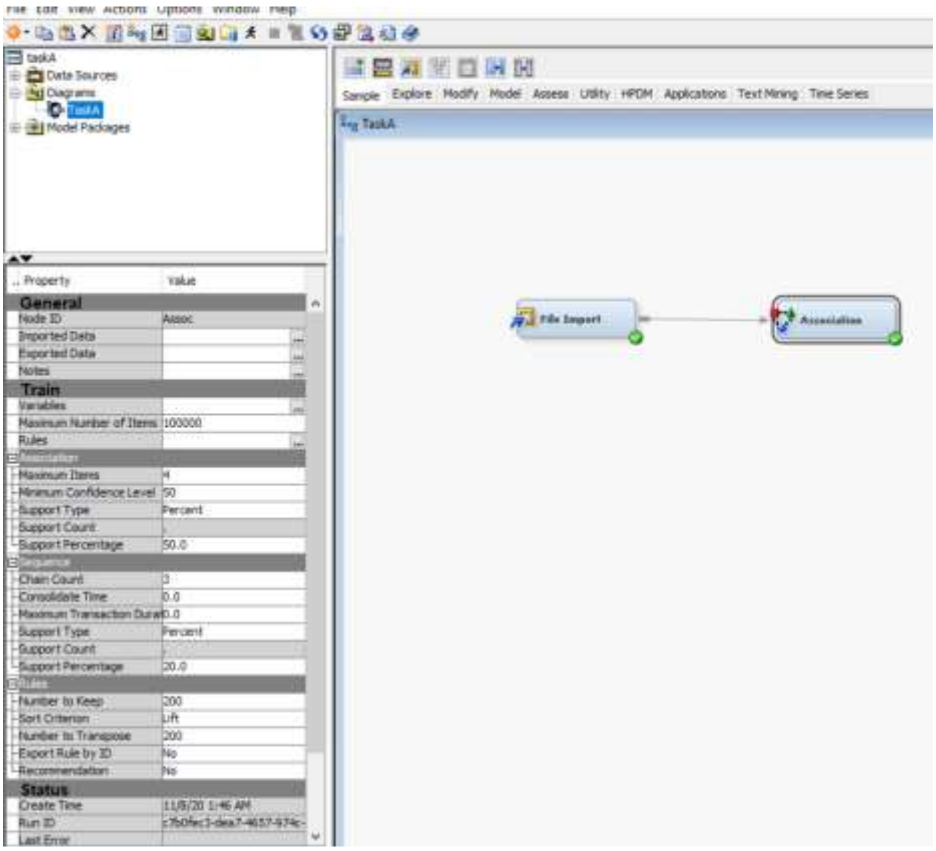
○ Support Type = percent

○ Minimum Support = 50% - ***This will eliminate some of the rules which are not interesting and allow us to focus on interesting rules that are applicable to larger group of customers.***

○ Maximum Rules to Keep = 200

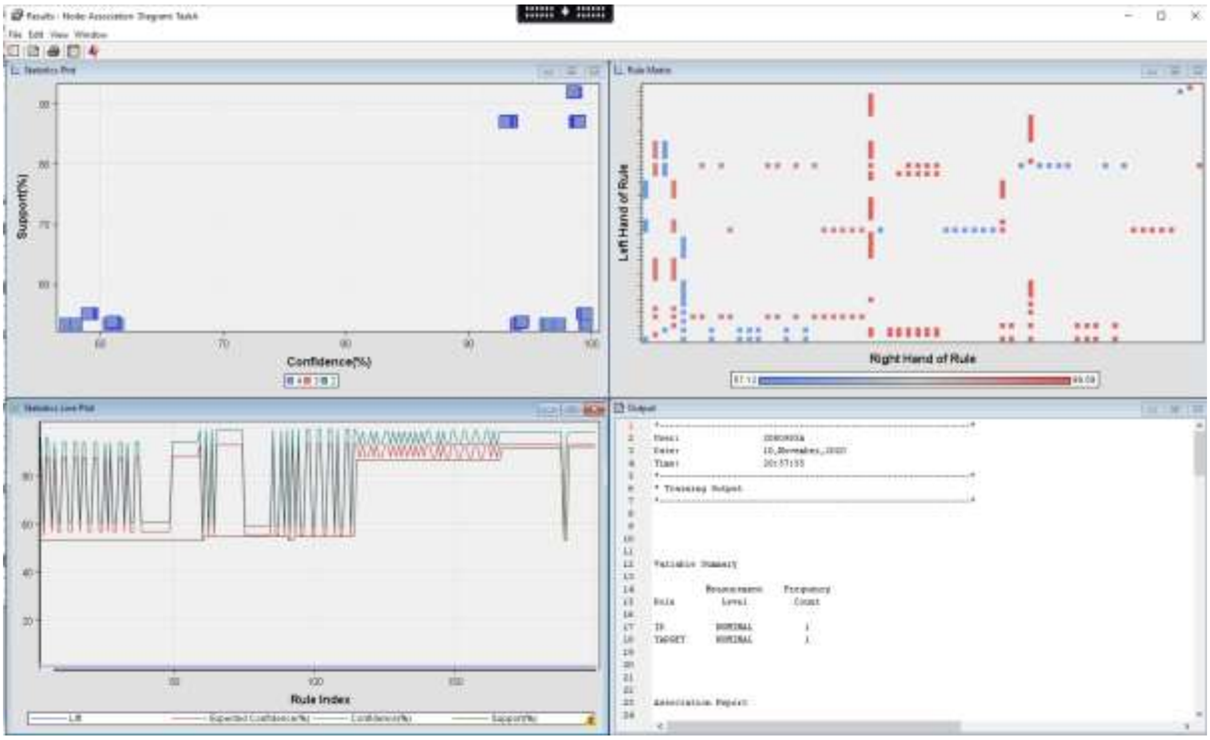
○ Sort Criteria = Lift – ***This will sort the rules by their informativeness.***

○ Leave other properties as their default values


File Edit View Actions Options Window Help
taskA
Data Sources
Diagrams
TaskA
Model Packages
Sample Explore Modify Model Assess Utility HPDM Applications Text Mining Time Series
TaskA
File Import
Association

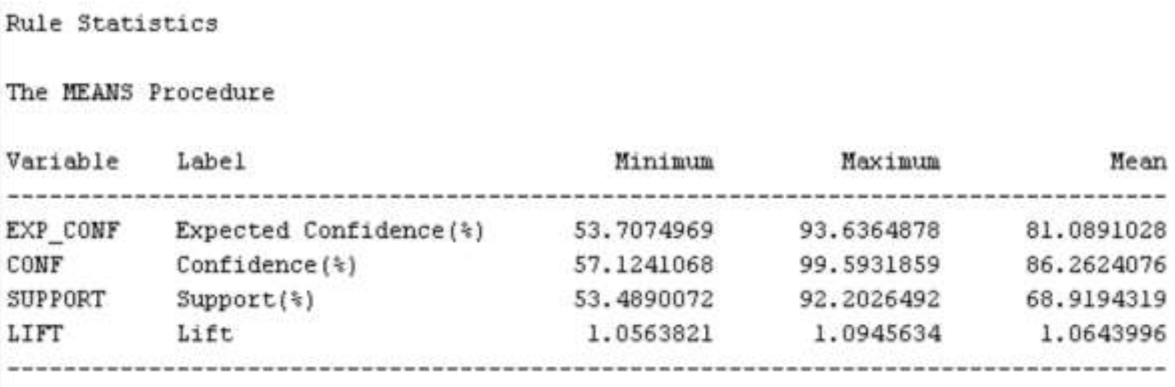
Property	Value
General	
Node ID	Assoc
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Maximum Number of Items	100000
Rules	...
Association	
Maximum Items	4
Minimum Confidence Level	50
Support Type	Percent
Support Count	1
Support Percentage	50.0
Sequence	
Chain Count	3
Consolidate Time	0.0
Maximum Transaction Durat	0.0
Support Type	Percent
Support Count	1
Support Percentage	20.0
Rules	
Number to Keep	200
Sort Criterion	Lift
Number to Transpose	200
Export Rule by ID	No
Recommendation	No
Status	
Create Time	11/8/20 1:46 AM
Run ID	c7b0fec3-dea7-4657-974c-
Last Error	

Run the Association node, once completed, view the results.
1. Statistics Plot
2. Statistics Line Plot
3. Rule Matrix
4. Output windows visible.



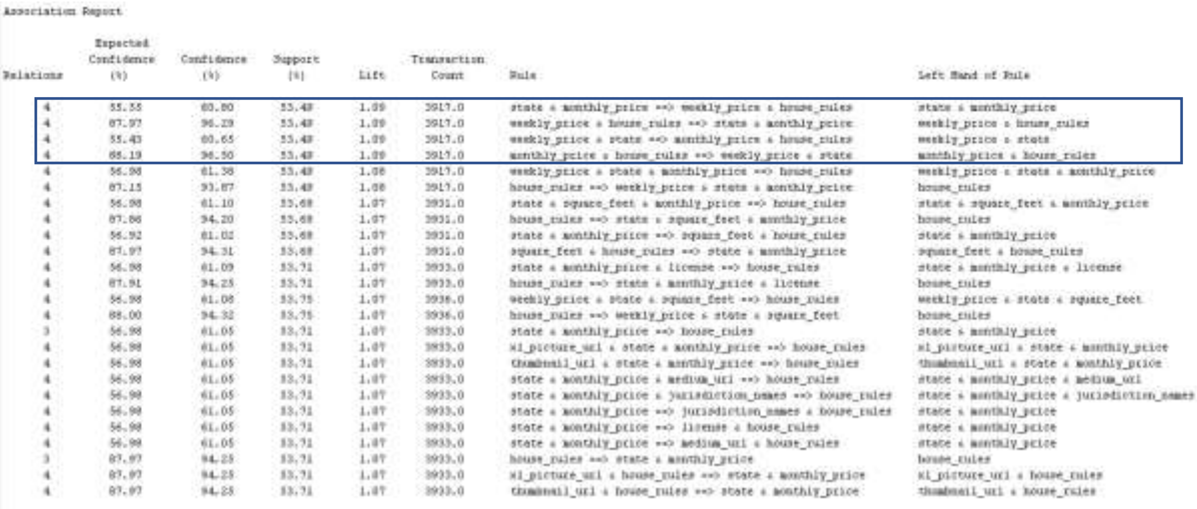
```
*------------------------------------------------------------*
User:               [illegible]
Date:               10,November,2020
Time:               20:57:55
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

            Measurement    Frequency
Role           Level          Count

ID            NOMINAL           1
TARGET        NOMINAL           1




Association Report
```

Interpretation of the Results



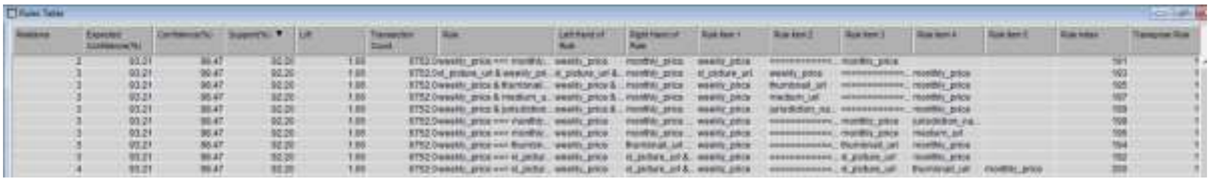
```
Rule Statistics

The MEANS Procedure

Variable    Label                        Minimum        Maximum           Mean
------------------------------------------------------------------------------
EXP_CONF    Expected Confidence(%)    53.7074969     93.6364878     81.0891028
CONF        Confidence(%)             57.1241068     99.5931859     86.2624076
SUPPORT     Support(%)                53.4890072     92.2026492     68.9194319
LIFT        Lift                       1.0563821      1.0945634      1.0643996
------------------------------------------------------------------------------
```

As seen from Association Report below, for the 1st 4 rules in the box, Lift is the highest. However, Support and Confidence are very low almost to minimum. Thus, Rules Table was explored.



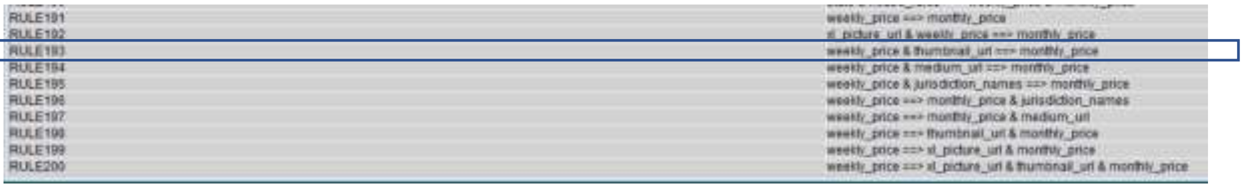
```
Association Report

             Expected
            Confidence  Confidence   Support            Transaction
Relations       (%)         (%)        (%)      Lift       Count      Rule                                                   Left Hand of Rule

    4          55.55       60.80      53.49     1.09      3917.0      state & monthly_price ==> weekly_price & house_rules   state & monthly_price
    4          87.97       96.29      53.49     1.09      3917.0      weekly_price & house_rules ==> state & monthly_price   weekly_price & house_rules
    4          55.43       60.65      53.49     1.09      3917.0      weekly_price & state ==> monthly_price & house_rules   weekly_price & state
    4          88.19       96.50      53.49     1.09      3917.0      monthly_price & house_rules ==> weekly_price & state   monthly_price & house_rules
    4          56.98       61.38      53.49     1.08      3917.0      weekly_price & state & monthly_price ==> house_rules   weekly_price & state & monthly_price
    4          87.15       93.87      53.49     1.08      3917.0      house_rules ==> weekly_price & state & monthly_price   house_rules
    4          56.98       61.10      53.68     1.07      3931.0      state & square_feet & monthly_price ==> house_rules    state & square_feet & monthly_price
    4          87.86       94.20      53.68     1.07      3931.0      house_rules ==> state & square_feet & monthly_price    house_rules
    4          56.92       61.02      53.68     1.07      3931.0      state & monthly_price ==> square_feet & house_rules    state & monthly_price
    4          87.97       94.31      53.68     1.07      3931.0      square_feet & house_rules ==> state & monthly_price    square_feet & house_rules
    4          56.98       61.09      53.71     1.07      3933.0      state & monthly_price & license ==> house_rules        state & monthly_price & license
    4          87.91       94.25      53.71     1.07      3933.0      house_rules ==> state & monthly_price & license        house_rules
    4          56.98       61.08      53.75     1.07      3936.0      weekly_price & state & square_feet ==> house_rules     weekly_price & state & square_feet
    4          88.00       94.32      53.75     1.07      3936.0      house_rules ==> weekly_price & state & square_feet     house_rules
    3          56.98       61.05      53.71     1.07      3933.0      state & monthly_price ==> house_rules                  state & monthly_price
    4          56.98       61.05      53.71     1.07      3933.0      xl_picture_url & state & monthly_price ==> house_rules xl_picture_url & state & monthly_price
    4          56.98       61.05      53.71     1.07      3933.0      thumbnail_url & state & monthly_price ==> house_rules  thumbnail_url & state & monthly_price
    4          56.98       61.05      53.71     1.07      3933.0      state & monthly_price & medium_url ==> house_rules     state & monthly_price & medium_url
    4          56.98       61.05      53.71     1.07      3933.0      state & monthly_price & jurisdiction_names ==> house_rules  state & monthly_price & jurisdiction_names
    4          56.98       61.05      53.71     1.07      3933.0      state & monthly_price ==> jurisdiction_names & house_rules  state & monthly_price
    4          56.98       61.05      53.71     1.07      3933.0      state & monthly_price ==> license & house_rules        state & monthly_price
    4          56.98       61.05      53.71     1.07      3933.0      state & monthly_price ==> medium_url & house_rules     state & monthly_price
    3          87.97       94.25      53.71     1.07      3933.0      house_rules ==> state & monthly_price                  house_rules
    4          87.97       94.25      53.71     1.07      3933.0      xl_picture_url & house_rules ==> state & monthly_price xl_picture_url & house_rules
    4          87.97       94.25      53.71     1.07      3933.0      thumbnail_url & house_rules ==> state & monthly_price  thumbnail_url & house_rules
```

Rules Table- with Lift 1.06, Support 92.20 (high), Confidence 98.47 (high) and relations of 3, it was decided to use Rule 193 for discussion.



Rules Table

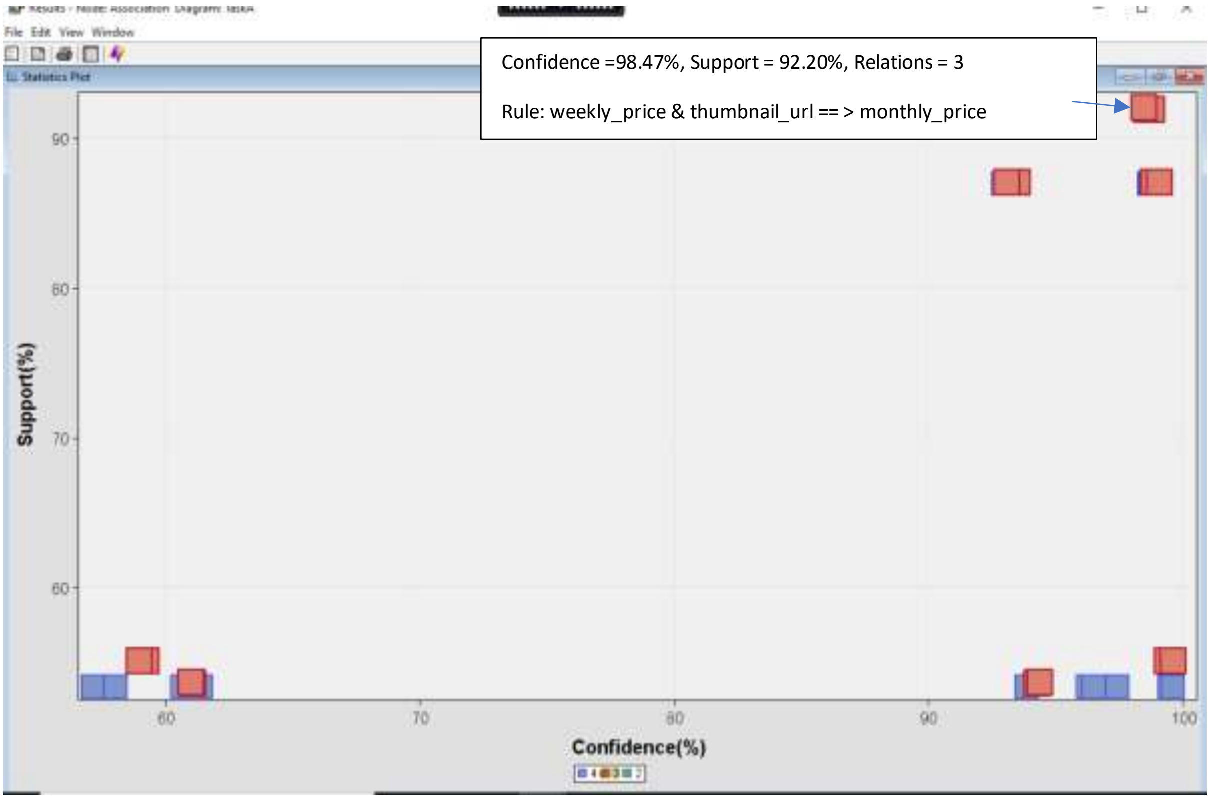
Relations	Expected Confidence(%)	Confidence(%)	Support(%) ▼	Lift	Transaction Count	Rule	Left Hand of Rule	Right Hand of Rule	Rule Item 1	Rule Item 2	Rule Item 3	Rule Item 4	Rule Item 5	Rule Index	Transpose Rule
3	93.21	98.47	92.20	1.06	8752.0	weekly_price ==> monthly...	weekly_price	monthly_price	weekly_price	================...	monthly_price			191	1
3	93.21	98.47	92.20	1.06	8752.0	xl_picture_url & weekly_pri...	xl_picture_url &...	monthly_price	xl_picture_url	weekly_price	================...	monthly_price		192	1
3	93.21	98.47	92.20	1.06	8752.0	weekly_price & thumbnail...	weekly_price &...	monthly_price	weekly_price	thumbnail_url	================...	monthly_price		193	1
3	93.21	98.47	92.20	1.06	8752.0	weekly_price & medium_u...	weekly_price &...	monthly_price	weekly_price	medium_url	================...	monthly_price		194	1
3	93.21	98.47	92.20	1.06	8752.0	weekly_price & jurisdiction...	weekly_price &...	monthly_price	weekly_price	jurisdiction_na...	================...	monthly_price		195	1
3	93.21	98.47	92.20	1.06	8752.0	weekly_price ==> monthly...	weekly_price	monthly_price ...	weekly_price	================...	monthly_price	jurisdiction_na...		196	1
3	93.21	98.47	92.20	1.06	8752.0	weekly_price ==> monthly...	weekly_price	monthly_price ...	weekly_price	================...	monthly_price	medium_url		197	1
3	93.21	98.47	92.20	1.06	8752.0	weekly_price ==> thumbn...	weekly_price	thumbnail_url ...	weekly_price	================...	thumbnail_url	monthly_price		198	1
3	93.21	98.47	92.20	1.06	8752.0	weekly_price ==> xl_pictur...	weekly_price	xl_picture_url &...	weekly_price	================...	xl_picture_url	monthly_price		199	1
4	93.21	98.47	92.20	1.06	8752.0	weekly_price ==> xl_pictur...	weekly_price	xl_picture_url &...	weekly_price	================...	xl_picture_url	thumbnail_url	monthly_price	200	1



RULE191	weekly_price ==> monthly_price
RULE192	xl_picture_url & weekly_price ==> monthly_price
RULE193	weekly_price & thumbnail_url ==> monthly_price
RULE194	weekly_price & medium_url ==> monthly_price
RULE195	weekly_price & jurisdiction_names ==> monthly_price
RULE196	weekly_price ==> monthly_price & jurisdiction_names
RULE197	weekly_price ==> monthly_price & medium_url
RULE198	weekly_price ==> thumbnail_url & monthly_price
RULE199	weekly_price ==> xl_picture_url & monthly_price
RULE200	weekly_price ==> xl_picture_url & thumbnail_url & monthly_price

Open file in Excel for quick review of the attributes and values to understand the dataset.

- Statistics Plot



Confidence =98.47%, Support = 92.20%, Relations = 3

Rule: weekly_price & thumbnail_url == > monthly_price

Rule 193: weekly_price & thumbnail_url == > monthly_price

Antecedent: weekly_price & thumbnail_url
Consequent: monthly_price

Description of variables:

- weekly_price = price per week
- thumbnail_url = image of property
- monthly_price = price per month

Support 92.20%

- This rule is applicable to 92.20% of all transactions. (i.e. Out of all transactions, 92.20% have both weekly_price & thumbnail_url and monthly_price together)

Confidence 98.47%

- Means that the higher the confidence, the stronger the rule is.
- We are confident that when hosts did not provide details on weekly price and image of property, 98.47% will also not provide the monthly price.

Lift = 1.06

- Range of lift is from 1.05 to 1.09. The mean is at 1.06.
- This rule has mean value.
- This indicates that hosts who do not share or display the weekly price and the image of Property are 1.06 times more likely **not** to share or display the monthly price.

Recommendations for Business

Airbnb's tagline - You don't need to go far to find what matters.

With reference to the tagline, this is what I assumed of the Airbnb's vision:

- information must be easily accessible.
- One-stop service.
- All required information has to be in website.
- Information have to be detailed and clear.

Recommendations for consideration are as follows:

Display:

- Host should shared beautiful images of property in Airbnb website. Colours and pictures can attract customers who are searching and browsing in Airbnb website. Hence, browsing time can be prolonged.
- Weekly price & monthly price should also be provided. Lack of information is likely to 'turn off' customers' interest.

Packaging:

- Airbnb can encourage hosts to display good deals for weekly and monthly bookings.
- Returned customers to be given freebies. For example, booked for 6 nights get 1 night free.

Promotion:

- Send customised emails, promotional messages and notification for customers who had previously made bookings via Airbnb.

Application of Technique in Non-retail Setting

Association rule mining can be applied to healthcare and research industry in the area of understanding protein sequences.

Protein sequences

Proteins are important constituents of cellular machinery of any organism. Recombinant DNA technologies have provided tools for the rapid determination of DNA sequences and, by inference, the amino acid sequences of proteins from structural genes [1].

Proteins are sequences made up of 20 types of amino acids. Each protein has a unique 3-dimensional structure, which depends on amino-acid sequence; slight change in sequence may change the functioning of protein. The heavy dependence of protein functioning on its amino acid sequence has been a subject of great anxiety.

Lot of research has gone into understanding the composition and nature of proteins; still many things remain to be understood satisfactorily. It is now generally believed that amino acid sequences of proteins are not random.

Nitin Gupta, Nitin Mangal, Kamal Tiwari, and Pabitra Mitra [9] have deciphered the nature of associations between different amino acids that are present in a protein. Such association rules are desirable for enhancing our understanding of protein composition and hold the potential to give clues regarding the global interactions amongst some particular sets of amino acids occurring in proteins. Knowledge of these association rules or constraints is highly desirable for synthesis of artificial proteins.

References (if applicable)

https://www.researchgate.net/publication/238525379_Association_rule_mining-_Applications_in_various_areas

[1] C. Branden and J. Tooze, “Introduction to Protein Structure”, Garland Publishing inc, New York and London, 1991.

[9] N. Gupta, N. Mangal, K. Tiwari and P. Mitra, “Mining Quantitative Association Rules in Protein Sequences”, In Proceedings of Australasian Conference on Knowledge Discovery and Data Mining – AUSDM, 2006

***** END OF ASSIGNMENT 1 *****