TEMASEK POLYTECHNIC
SCHOOL OF INFORMATICS & IT
SPECIALIST DIPLOMA IN BUSINESS ANALYTICS
AY2021/2022 APR SEMESTER TERM A

DATA ANALYTICS FOR BUSINESS INSIGHTS (CBA1C09)

ASSIGNMENT 1

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)	Goh Aik Hong Jonathan
Admin Number	2081973F
Group	Group 4
Task selected (A or B)	B

For Tutor Use

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

Performance of Pattern Discovery Task

Data Exploration:

Importing the csv to Jupyter Notebook, a quick check reveals that there are 7323 rows and 106 columns.

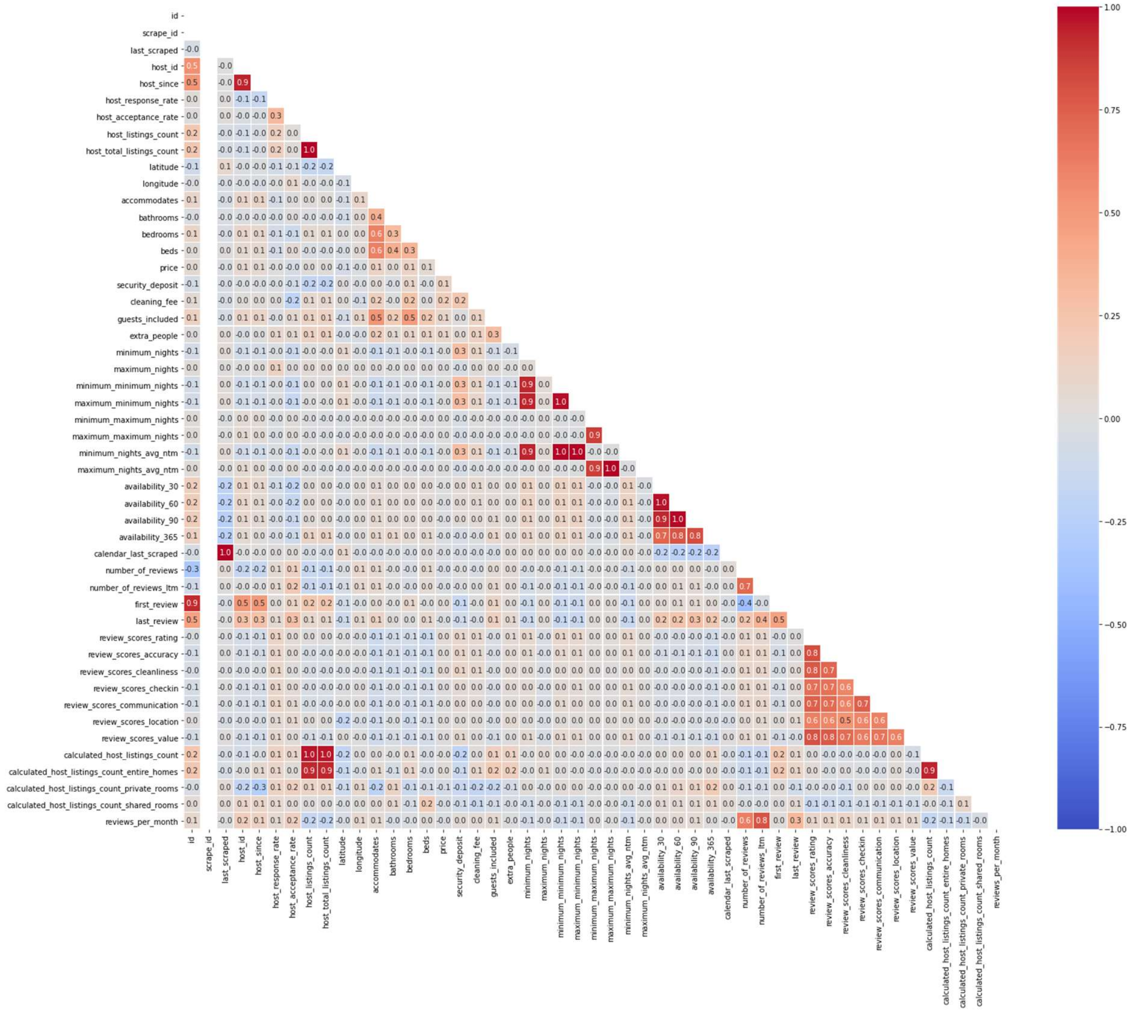
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7323 entries, 0 to 7322
Data columns (total 106 columns):
 #   Column                                        Non-Null Count  Dtype
---  ------                                        --------------  -----
 0   id                                            7323 non-null   int64
 1   listing_url                                   7323 non-null   object
 2   scrape_id                                     7323 non-null   float64
 3   last_scraped                                  7323 non-null   int64
 4   name                                          7319 non-null   object
 5   summary                                       6964 non-null   object
 6   space                                         5340 non-null   object
 7   description                                   7041 non-null   object
 8   experiences_offered                           7323 non-null   object
 9   neighborhood_overview                         4354 non-null   object
 10  notes                                         3995 non-null   object
 11  transit                                       4396 non-null   object
 12  access                                        4435 non-null   object
 13  interaction                                   4036 non-null   object
 14  house_rules                                   3149 non-null   object
 15  thumbnail_url                                 0 non-null      float64
 16  medium_url                                    0 non-null      float64
 17  picture_url                                   7323 non-null   object
 18  xl_picture_url                                0 non-null      float64
```

As mentioned in the assignment given, not all fields may be useful for analysis. For a start, features that contain a large number of rows with null values (>4000) are dropped.

Next, it is observed that both categorical variables and numerical variables are present.

ID, name, URL and most of the date columns were dropped as they are deemed to be not useful for further analysis.

Visualizing the correlation of the numerical features on a heatmap, further dropping can be done by removing features that have high collinearity.



For the categorical features, those that have only one unique value are dropped since there is no variance. Features that contain long text strings are also dropped since NLP will not be part of this assignment. For the remaining categorical features, dummy variables were created.

	count	unique	top	freq
listing_url	7323	7323	https://www.airbnb.com/rooms/38388737	1
name	7319	6763	City-centered 1BR apartment *BRAND NEW*	12
summary	6964	4341	A beautiful and spacious apartment equipped with the following room amenities: -Designer bed frames with Queen Size Mattress and quality linens -Designer dining table with beautiful chairs -Kitchen with Fully cooking utensils, Rice cooker, Stove -Refrigerator and Freezer -Comprehensive Cooking Utensils & Cutlery -Hair Dryer -Iron and Ironing Board -Washing Machine cum Dryer -Air-Conditioning with Individual Controller	253
space	5340	3118	A beautiful and spacious apartment equipped with the following room amenities: -Designer bed frames with Queen Size Mattress and quality linens -Designer dining table with beautiful chairs -Kitchen with Fully cooking utensils, Rice cooker, Stove -Refrigerator and Freezer -Comprehensive Cooking Utensils & Cutlery -Hair Dryer -Iron and Ironing Board -Washing Machine cum Dryer -Air-Conditioning with Individual Controller	214
description	7041	5093	A beautiful and spacious apartment equipped with the following room amenities: -Designer bed frames with Queen Size Mattress and quality linens -Designer dining table with beautiful chairs -Kitchen with Fully cooking utensils, Rice cooker, Stove -Refrigerator and Freezer -Comprehensive Cooking Utensils & Cutlery -Hair Dryer -Iron and Ironing Board -Washing Machine cum Dryer -Air-Conditioning with Individual Controller A beautiful and spacious apartment equipped with the following room amenities: -Designer bed frames with Queen Size Mattress and quality linens -Designer dining table with beautiful chairs -Kitchen with Fully cooking utensils, Rice cooker, Stove -Refrigerator and Freezer -Comprehensive Cooking Utensils & Cutlery -Hair Dryer -Iron and Ironing Board -Washing Machine cum Dryer -Air-Conditioning with Individual Controller Arrived SG text us one hour in advance CHECK IN and CHECK OUT TIME Our check in time is 1500 hrs and check out time is 1200 hrs EARLY CHECK IN & LATE CHECK	141
experiences_offered	7323	1	none	7323
calendar_updated	7323	79	3 months ago	974
has_availability	7323	1	t	7323
requires_license	7323	1	f	7323
instant_bookable	7323	2	f	4227
is_business_travel_ready	7323	1	f	7323
cancellation_policy	7323	5	strict_14_with_grace_period	4664
require_guest_profile_picture	7323	2	f	7289
require_guest_phone_verification	7323	2	f	7276

The final dataset has 59 columns and is exported as a csv again.

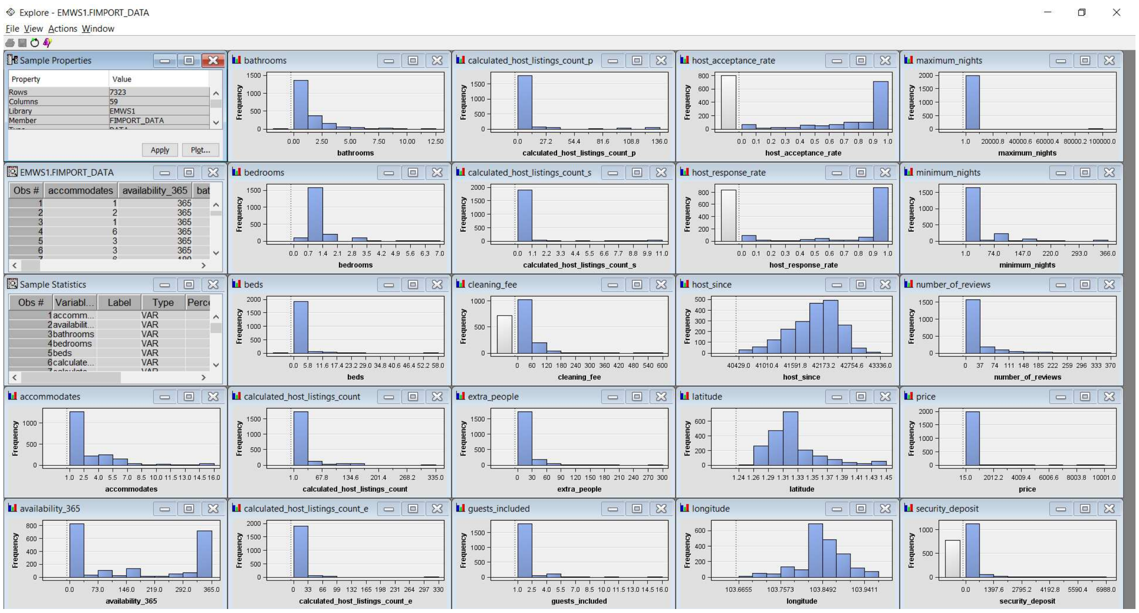
The file was then imported into SAS Enterprise. However, upon checking the variables of the import node, it was observed that a few features were renamed as VAR59, VAR60, etc. Further investigation reveals that SAS Enterprise may have a limit on the number of characters each feature can have and some of the dummy variables created had rather long names that were shortened and caused naming conflicts. This was resolved by going back to the Jupyter Notebook and shortening the names of the affected features.

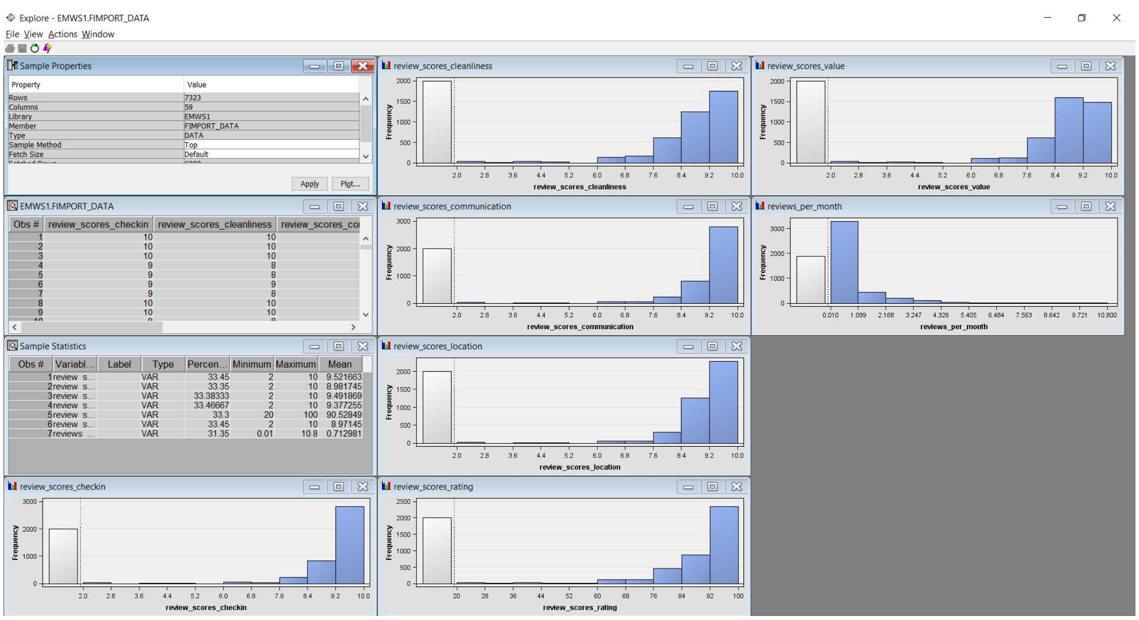
The updated dataset was then exported into a csv again and imported into SAS with no issues this time.

```

The CONTENTS Procedure

Data Set Name        EMWS1.FIMPORT_DATA          Observations          7323
Member Type          DATA                        Variables             59
Engine               V9                          Indexes               0
Created              12/05/2021 17:28:34         Observation Length    472
Last Modified        12/05/2021 17:28:34         Deleted Observations  0
Protection                                       Compressed            NO
Data Set Type                                    Sorted                NO
Label
Data Representation  WINDOWS_64
Encoding             wlatin1  Western (Windows)

```

Distribution of numerical features:

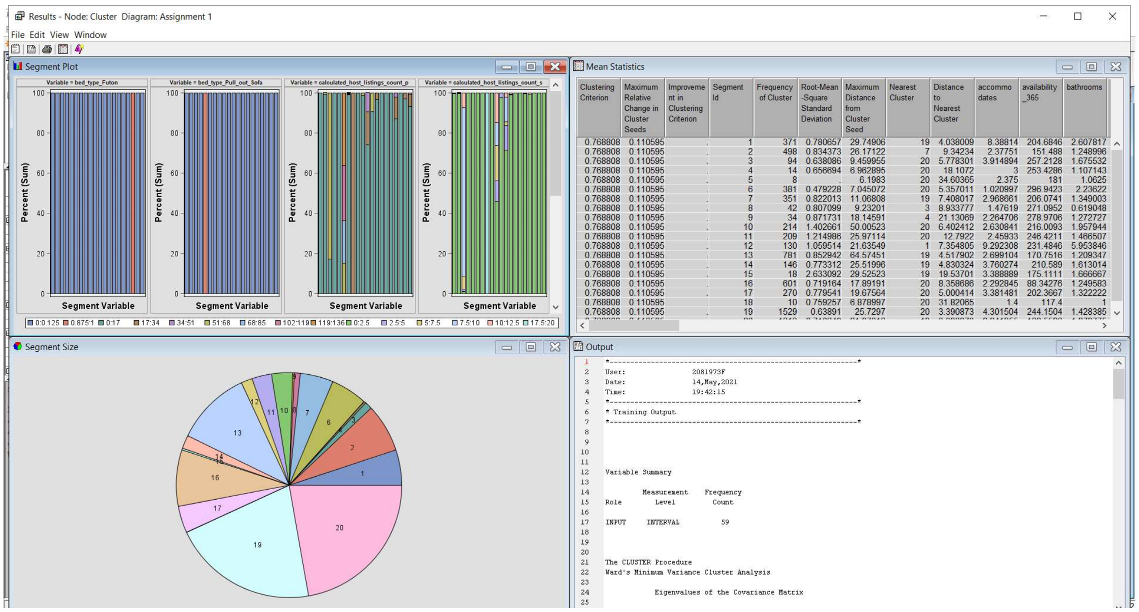
Using the explore function of the numerical features, the following histograms were obtained. It is observed that there seems to be an outlier point for the maximum nights in the [90000, 100000] bin. Upon further checking, there are a few more outlier values (>9000). These can be removed by filtering.



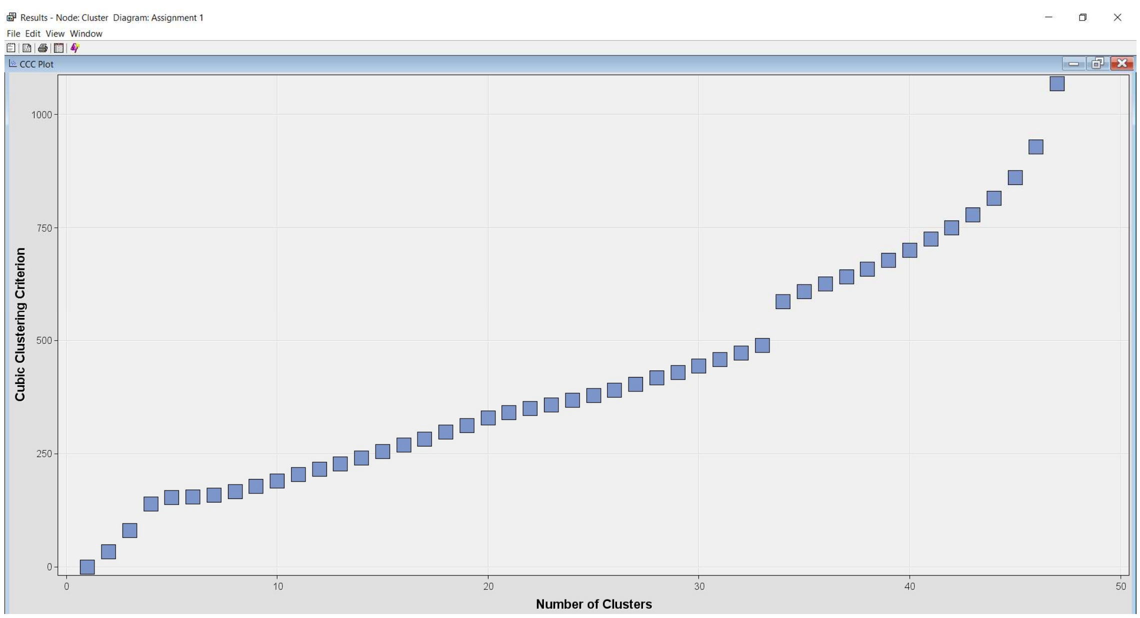


Clustering and Segment Profile:

Clustering node was ran with the following results:

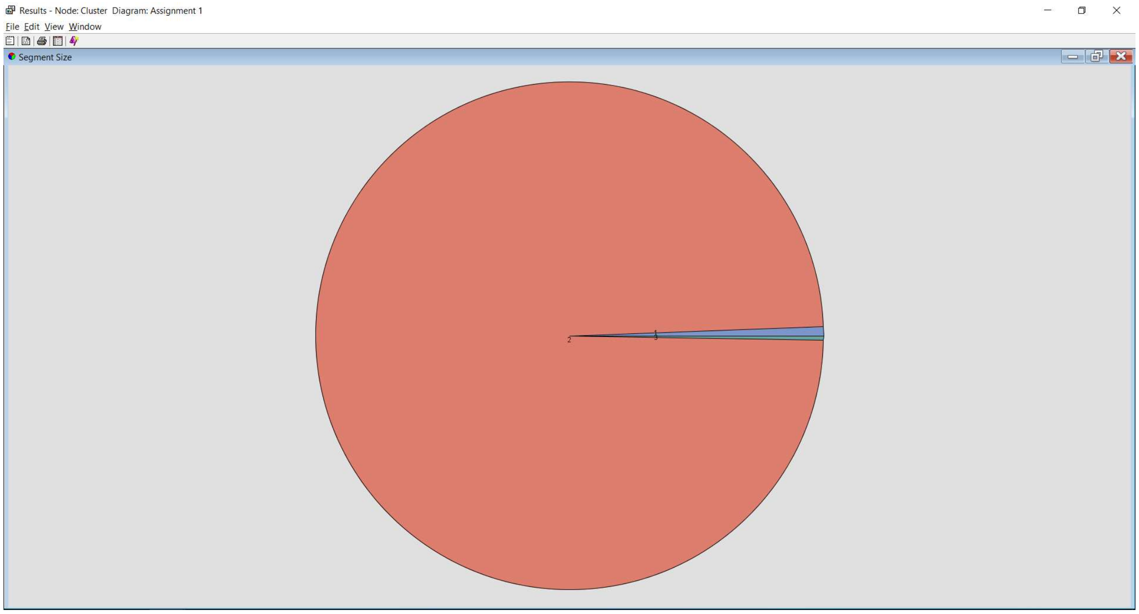


20 segments can be observed based on the results.

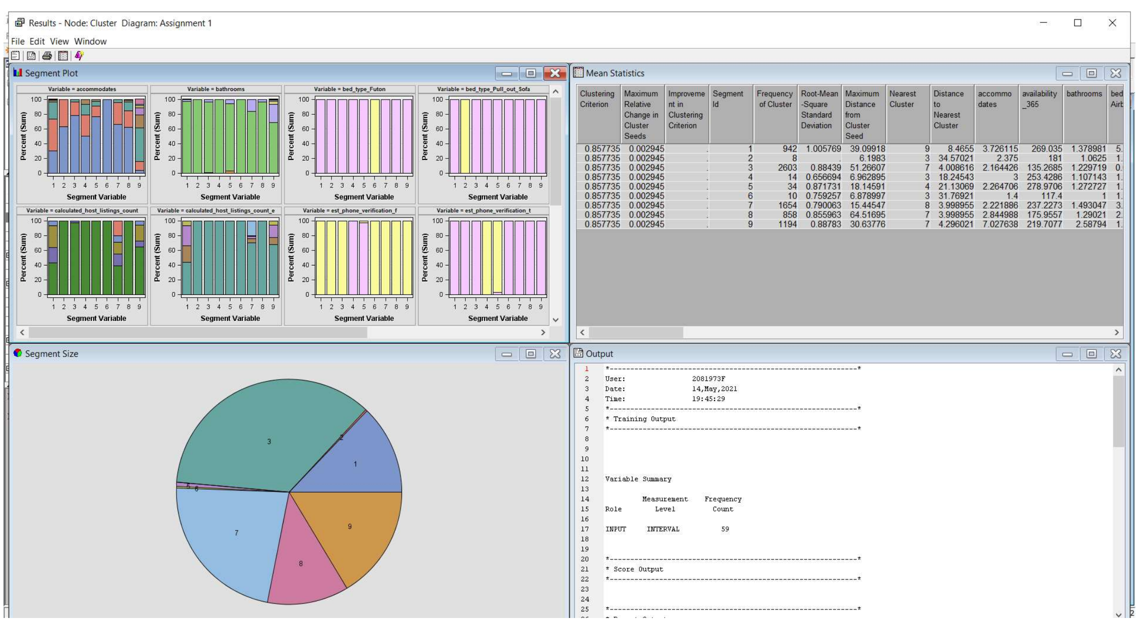


Based on the CCC plot, an ideal number of clusters cannot be determined.

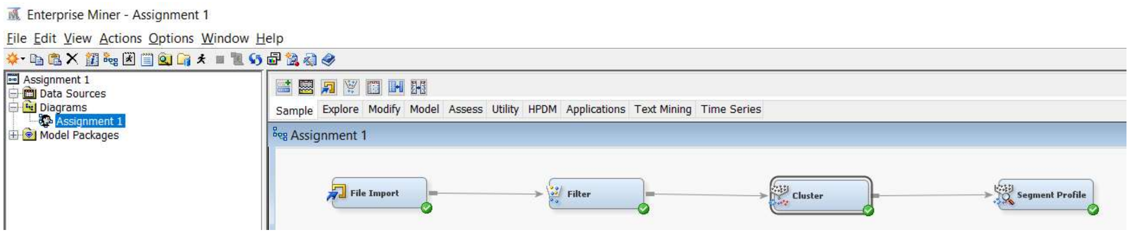
Defining the number of clusters to be 4, the clustering node is re-run. However, the segment size returned as follows. As such, k-means of 4 is does not segment the dataset well.



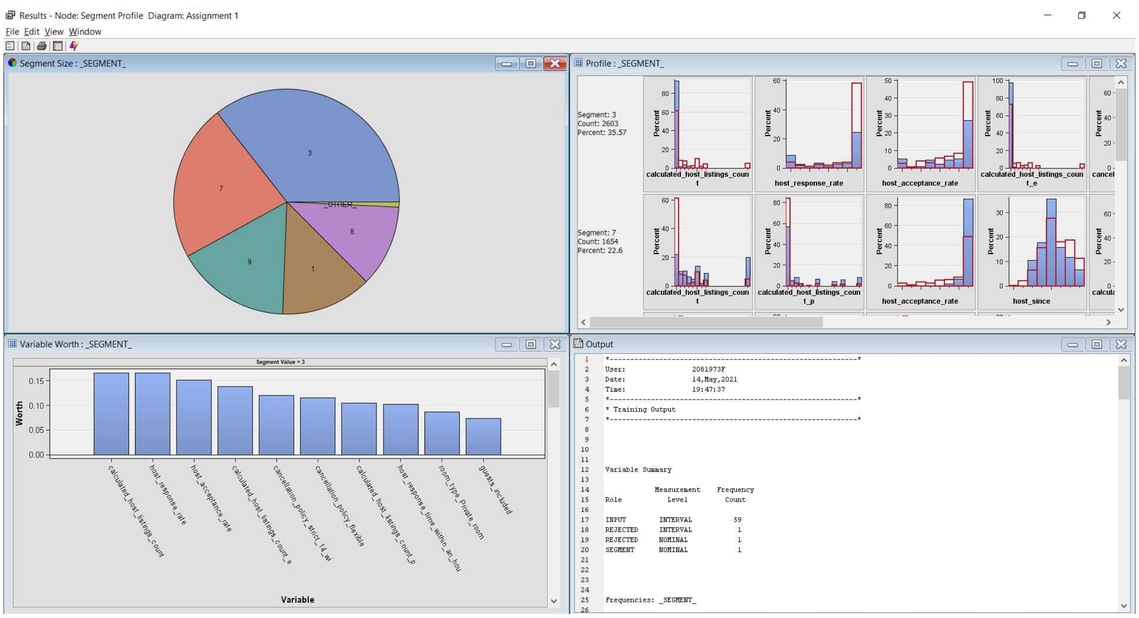
Increasing the number of clusters by one a time, the clustering node is re-run until the largest segment size does not take up an overwhelming majority. The final number of clusters determined by this method is 9.



A segment profile node is then added for further analysis of the segments. The finalized diagram in SAS is as below



With the results as follows when ran:

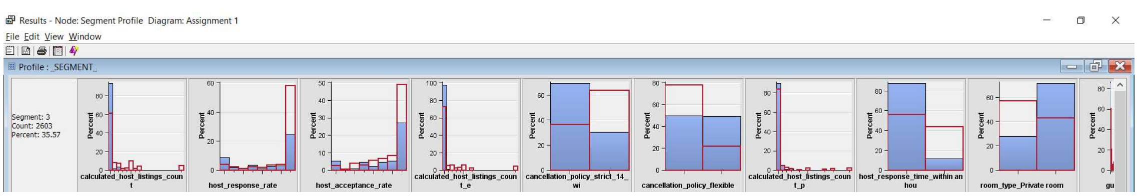


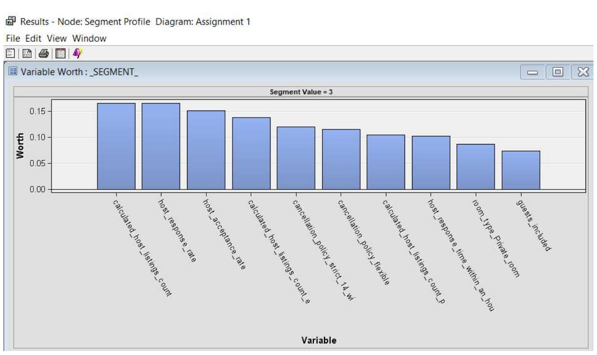
Segments 1, 3, 7, 8 and 9 can be further analysed while the smaller segments have been put under “Others”.

Interpretation of the Results

Segment 3:

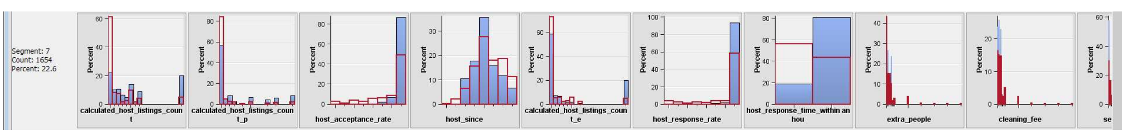
Profile	2603 which make up 35.6% of population.
Interpretation of profile by features from the highest to lowest variable worth:	
Host Listings Count Variable Worth: 0.166	The majority of the hosts in this segment had a lower listing count as compared with others within the dataset.
Host Response Rate Variable Worth: 0.166	The majority of the hosts in this segment had a lower response rate as compared with others within the dataset.
Host Acceptance Rate Variable Worth: 0.151	The majority of the hosts in this segment had a lower acceptance rate as compared with others within the dataset.
Host Listings Count that are Entire Homes Variable Worth: 0.139	The majority of the hosts in this segment had a lower listing count of entire homes as compared with others within the dataset.
Cancellation Policy 14 days with Grace Period Variable Worth: 0.121	The majority of the hosts in this segment did not have a strict 14 day cancellation policy as compared with others within the dataset.
Flexible Cancellation Policy Variable Worth: 0.115	The majority of the hosts in this segment have a flexible cancellation policy as compared with others within the dataset.
Host Listings Count that are Private Rooms Variable Worth: 0.105	The majority of the hosts in this segment had a lower listing count of private rooms as compared with others within the dataset.
Hosts that Respond within an hour Variable Worth: 0.103	The majority of the hosts in this segment did not respond within an hour as compared with others within the dataset.
Listing Type that are Private Rooms Variable Worth: 0.087	The majority of the room type in this segment is private rooms as compared to others within the dataset.
Listings that have Guests Included Variable Worth: 0.074	The majority of the hosts in this segment have a lower number of guests included as compared with others within the dataset.

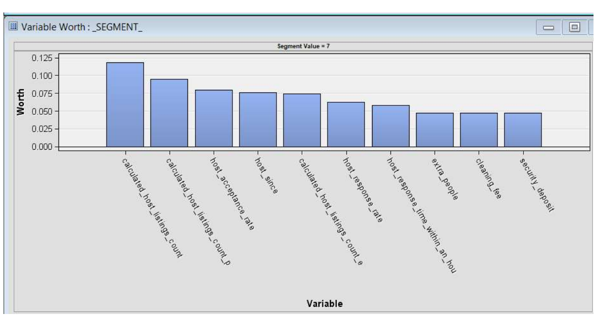




Segment 7:

Profile	1654 which make up 22.6% of population.
Interpretation of profile by features from the highest to lowest variable worth:	
Host Listings Count Variable Worth: 0.118	The majority of the hosts in this segment had a higher listing count as compared with others within the dataset.
Host Listings Count that are Private Rooms Variable Worth: 0.095	The majority of the hosts in this segment had a higher listing count of private rooms as compared with others within the dataset.
Host Acceptance Rate Variable Worth: 0.080	The majority of the hosts in this segment had a higher acceptance rate as compared with others within the dataset.
Host Since Variable Worth: 0.076	The majority of the hosts in this segment have been on Airbnb for a longer time as compared with others within the dataset.
Host Listings Count that are Entire Homes Variable Worth: 0.074	The majority of the hosts in this segment had a larger listing count of entire homes as compared with others within the dataset.
Host Response Rate Variable Worth: 0.063	The majority of the hosts in this segment had a higher response rate as compared with others within the dataset.
Hosts that Respond within an hour Variable Worth: 0.058	The majority of the hosts in this segment responded within an hour as compared with others within the dataset.
Prices for Extra People in the listing Variable Worth: 0.048	The majority of the listings in this segment charged higher than the lowest range for extra persons as compared with others within the dataset.
Cleaning Fee Variable Worth: 0.047	The majority of the listings in this segment charged lower cleaning fees as compared with others within the dataset.
Security Deposit Variable Worth: 0.047	The majority of the listings in this segment had lower security deposits as compared with others within the dataset.





Selecting these two largest groups identified, the notable differences are:

Segment 3	Segment 7
Lower listing count (including entire homes and private rooms)	Larger listing count (including entire homes and private rooms)
Lower response and acceptance rate	Higher response and acceptance rate
Did not respond within the hour	Responds within the hour

The profile of the remaining groups are as follows:

Segment 9 (16.3%):

Listings in this segment generally had more bathrooms, bedrooms and beds, are able to accommodate more people, had a higher number of guests included, were priced higher, were less of the private room and more of the entire home/apartment type, had hosts with a larger number of entire homes listed but lower overall listing count as compared to the others within the dataset.

Segment 1 (12.87%):

Listings in this segment generally had newer hosts with a larger number of entire homes listed and an overall lower listing count, lower host acceptance rates and lower review scores across the different categories as compared to the others within the dataset.

Segment 8 (%):

Listings in this segment generally had Superhosts, higher host acceptance rates, higher number of reviews (both overall and per month), higher review scores across the different categories and hosts that have a lower listing count as compared to the others within the dataset.

Recommendations for Business

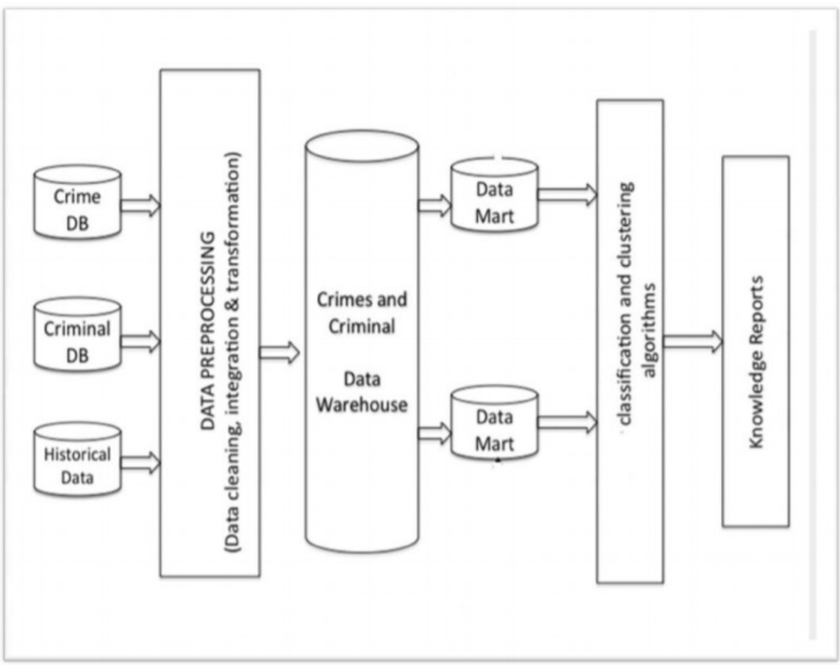
Segment	Recommendation
3	To improve the business, Airbnb could reach out to the hosts of these listings encourage higher response and acceptance rates, as well as more prompt replies.
7	The listings in this segment have been around for longer and have hosts that have a larger listing count. Judging by the higher response and acceptance rates as well as quicker response times and lower cleaning fees and security deposits, these listings could be professionally managed properties. Airbnb should maintain good customer relationships with these hosts and encourage them to keep up the good work.
9	The listings in this segment had a larger number of entire homes/apartments that are able to accommodate a larger number of persons. Airbnb can encourage hosts to consider the option of splitting up the property to take in more than one group of visitors at a time as it may be harder to find larger groups of tourists that can rent out the entire property.
1	The listings in this segment generally had newer hosts that have lower review scores across the various categories such as communication, check-in, cleanliness and location. As these hosts are newer, they may appreciate assistance from Airbnb to tackle these issues. Communications, check-in and cleanliness issues could conceivably be resolved by engaging professional property management services.
8	Similar to segment 7, this group of listings generally had good features and Airbnb should maintain good customer relationships with these hosts and encourage them to keep up the good work.

Application of Technique in Non-retail Setting

One of the applications of clustering is criminal profiling.

Referencing a paper from the Bells University of Technology in Nigeria, http://ijarcsse.com/Before_August_2017/docs/papers/Volume_6/4_April2016/V6I4-01407.pdf:

One challenge that all law-enforcement and intelligence-gathering agencies face is the ability to analyse large volumes of crime data accurately and efficiently. In clustering algorithms, Euclidean distance is used to measure the similarity. Similar objects are nearer while objects from other groups are further away. With clustering, agencies can use clustering techniques to discover patterns of crime that may otherwise go unnoticed to predict the occurrence of those crimes so as to aid in their reduction/prevention, link related crimes or narrow down suspects in an investigation, etc. The flow is illustrated in the following diagram:



A safer tomorrow can be achieved globally that is brought about by the application and enhancement of clustering techniques.

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