# DAVA – It’s Data Prep Time

1. Derive New Columns (10 marks)

a) Show the steps taken to prepare new column (field) called *property\_simple.*

Add a Rule Engine node downstream from the Excel Reader node. Use the Rule Engine to append a new column call property\_simple. The data in the property\_simple column is derived from property\_type column with the following expression below:

$property\_type$ LIKE "Aparthotel" => "apartment"

$property\_type$ LIKE "Apartment" => "apartment"

$property\_type$ LIKE "Bed and breakfast" => "hostel"

$property\_type$ LIKE "Boat" => "special"

$property\_type$ LIKE "Boutique hotel" => "hotel"

$property\_type$ LIKE "Bungalow" => "house"

$property\_type$ LIKE "Bus" => "special"

$property\_type$ LIKE "Cabin" => "special"

$property\_type$ LIKE "Campsite" => "special"

$property\_type$ LIKE "Condominium" => "apartment"

$property\_type$ LIKE "Guest suite" => "hostel"

$property\_type$ LIKE "Guesthouse" => "hostel"

$property\_type$ LIKE "Hostel" => "hostel"

$property\_type$ LIKE "Hotel" => "hotel"

$property\_type$ LIKE "House" => "house"

$property\_type$ LIKE "Loft" => "apartment"

$property\_type$ LIKE "Other" => "special"

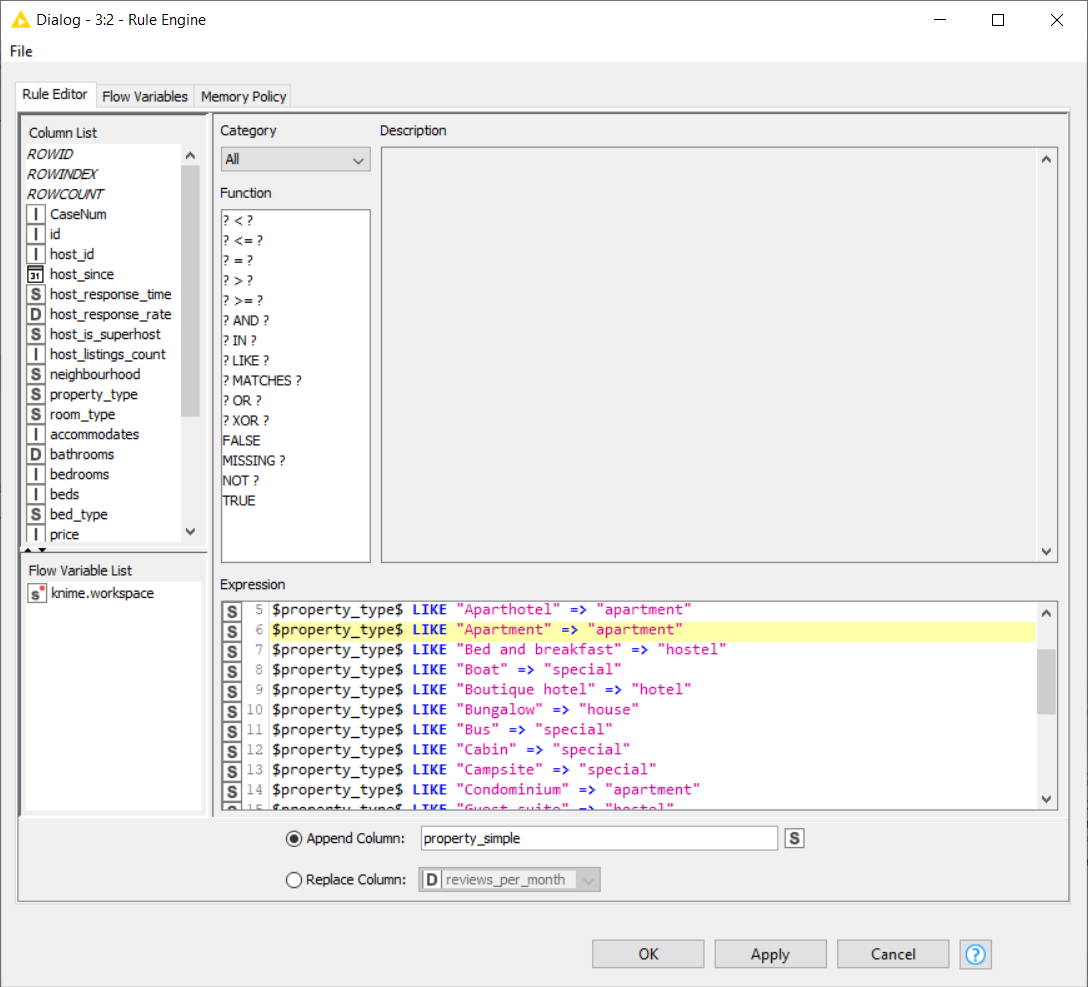
$property\_type$ LIKE "Serviced apartment" => "apartment"

$property\_type$ LIKE "Tent" => "special"

$property\_type$ LIKE "Townhouse" => "house"

$property\_type$ LIKE "Villa" => "house"

TRUE => "others"



1. Show the steps taken to prepare new column (field) called *property\_num.*

Add a Rule Engine node downstream from the 1st Rule Engine node. Use the Rule Engine to append a new column call property\_num. The data in the property\_num column is derived from property\_simple column with the following expression below:

$property\_simple$ LIKE "special" => 1

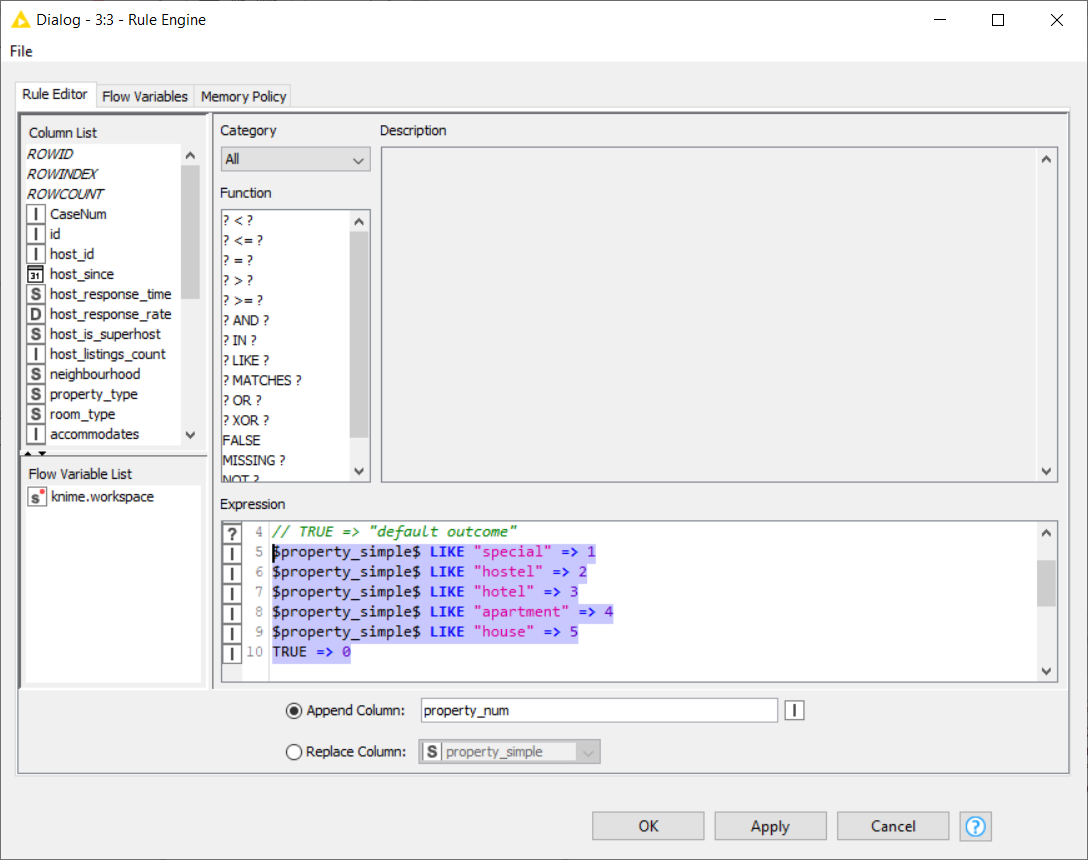
$property\_simple$ LIKE "hostel" => 2

$property\_simple$ LIKE "hotel" => 3

$property\_simple$ LIKE "apartment" => 4

$property\_simple$ LIKE "house" => 5

TRUE => 0



1. Show the steps taken to prepare new column (field) called *room\_num.*

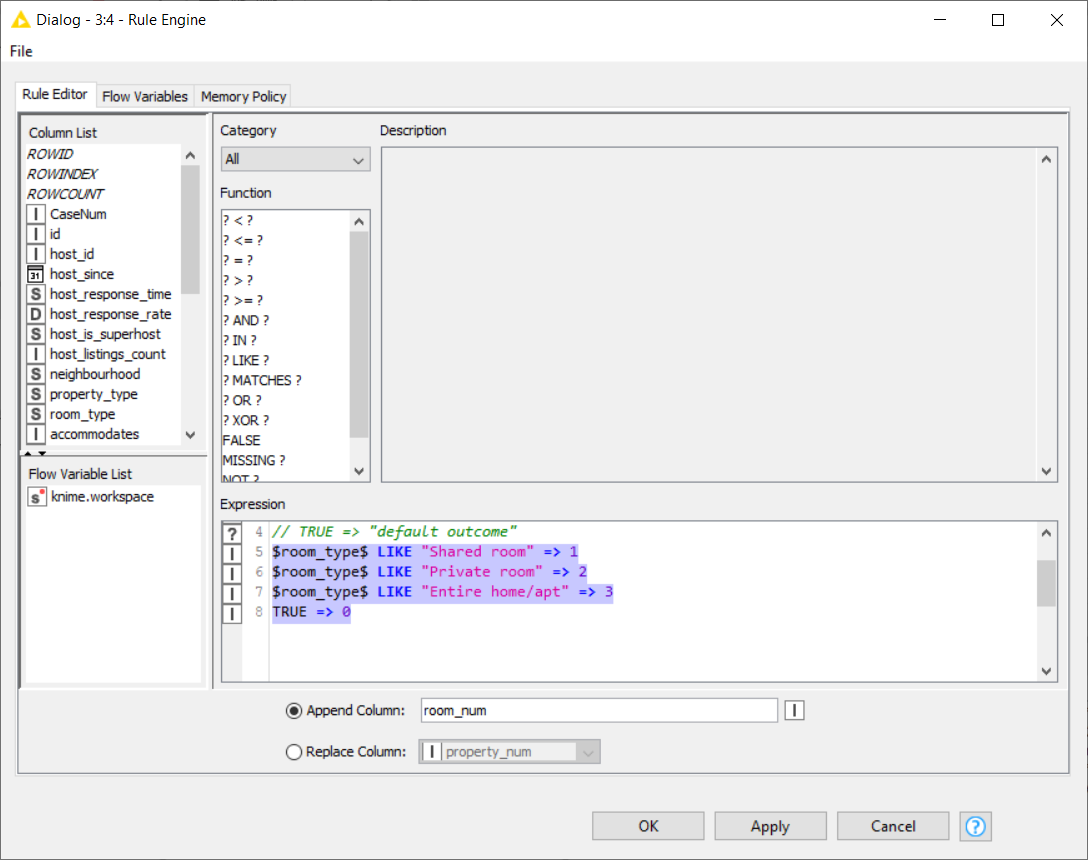
Add a Rule Engine node downstream from the 2nd Rule Engine node. Use the Rule Engine to append a new column call room\_num. Tthe data in the room\_num column is derived from room\_type column with the following expression below:

$room\_type$ LIKE "Shared room" => 1

$room\_type$ LIKE "Private room" => 2

$room\_type$ LIKE "Entire home/apt" => 3

TRUE => 0



1. Show the steps taken to prepare new column (field) called *cancellation\_num.*

Add a Rule Engine node downstream from the 3rd Rule Engine node. Use the Rule Engine to append a new column call cancellation\_num. The data in the cancellation\_num column is derived from cancellation\_policy column with the following expression below:

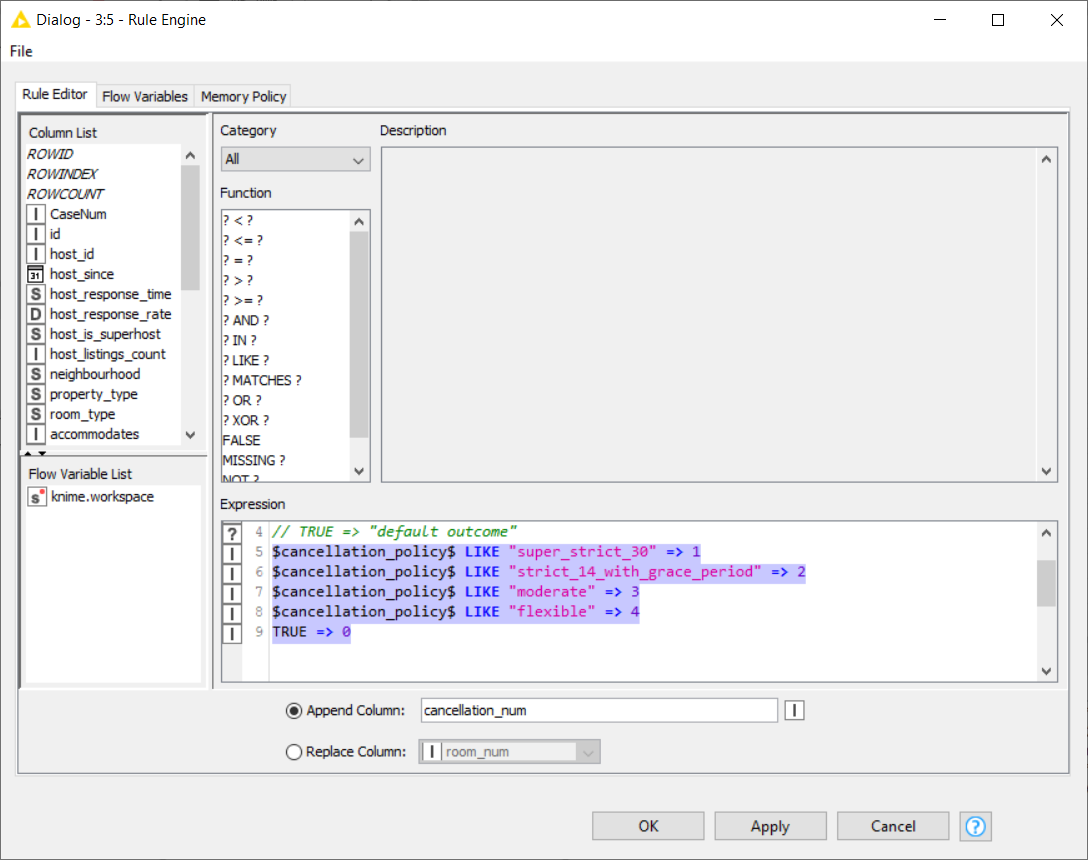
$cancellation\_policy$ LIKE "super\_strict\_30" => 1

$cancellation\_policy$ LIKE "strict\_14\_with\_grace\_period" => 2

$cancellation\_policy$ LIKE "moderate" => 3

$cancellation\_policy$ LIKE "flexible" => 4

TRUE => 0

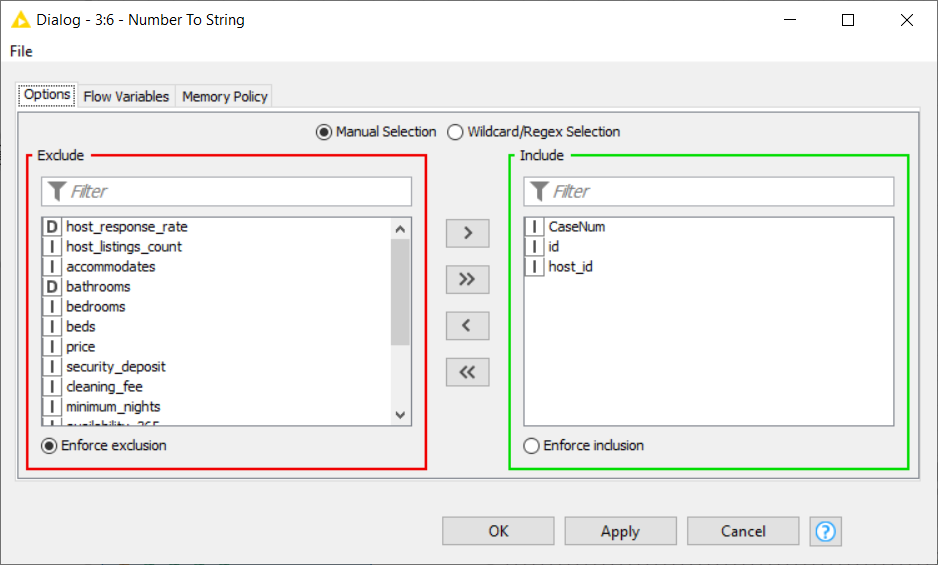


2. Data Cleaning (15 marks)

a) Show the steps taken to convert the columns *CaseNum*, *id* and *host\_id* to string type.

Add a Number to String node downstream from the 4th Rule Engine node.

Use the Number to String node to convert the *CaseNum*, *id* and *host\_id* to string type.



b) Show the steps taken to convert the column *host\_is\_superhost* from categorical to numeric data.

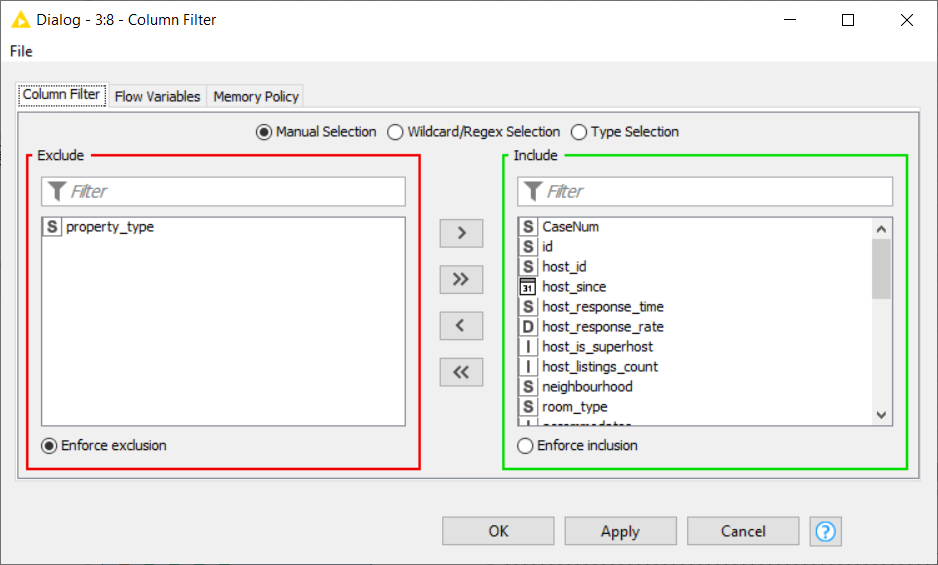
Add a Category to Number node downstream from the Number to String node.

Use the Category to Number node to convert the column *host\_is\_superhost* from categorical to numeric data.

|  |  |
| --- | --- |
|  |  |

c) Show the steps taken to remove the column (field) *property\_type* and to indicate how many columns are left.

Add a Column Filter node downstream from the Category to Number node to remove the column (field) *property\_type*. After removing *property\_type* column it left with 29 columns.

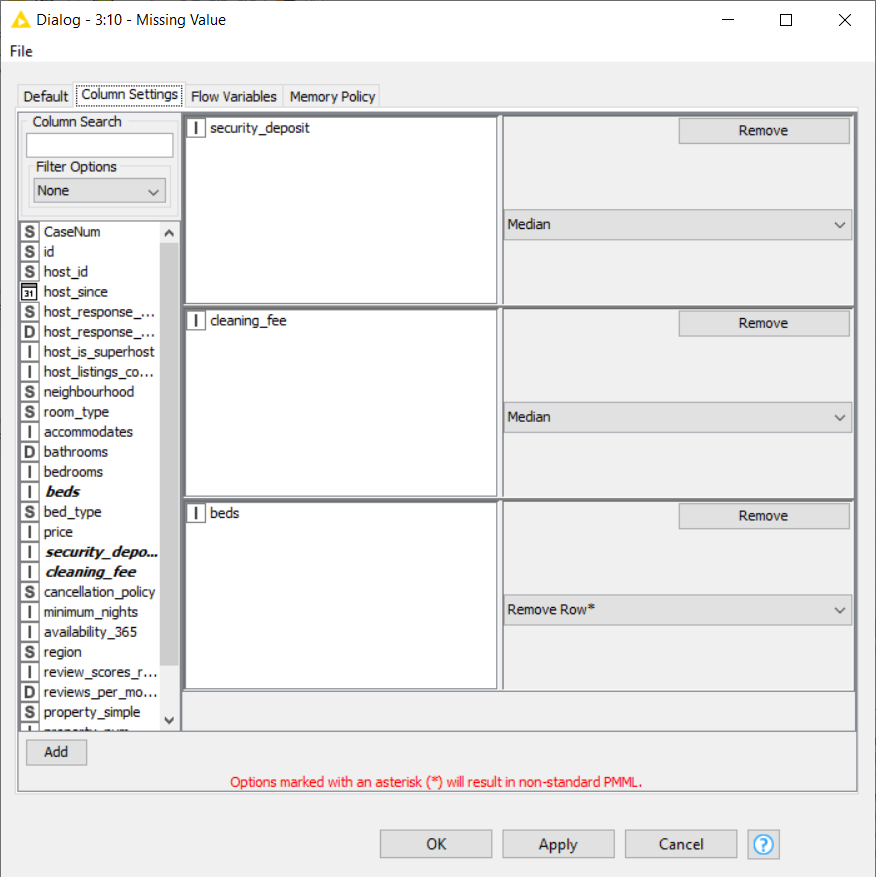


d) Show the steps taken to handle missing values in the data and to justify why you chose the method for each case.

From the statistics node it shows there are 791, 697 and 1 missing values in the security\_deposit, cleaning\_fee and beds columns respectively.

Replace the missing values in the security\_deposit and cleaning\_fee columns with their median value since their distribution curve are skewed to the right as shown in the statistics.

For the missing value in the beds column the decision was made to remove the row since there is only 1 missing value (which is less than 0.1% of the data size).



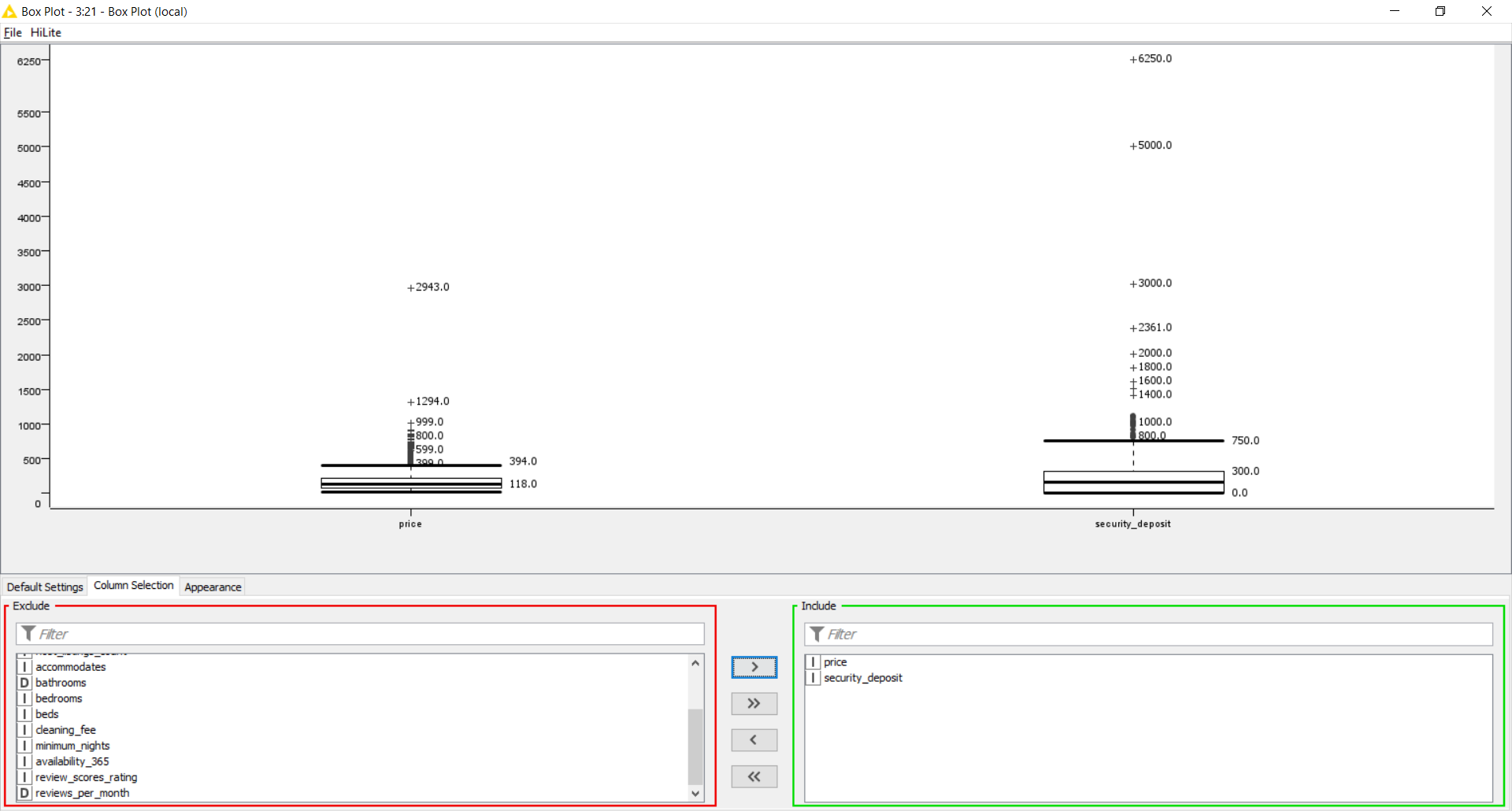
e) Show the steps taken to handle other opportunity for improving the quality of the data.

From the box-plot, it is observed that the security deposit column had 2 extreme outliers ($6250 and $5000). These two extreme outliers are:

* Only a private room type
* Located in average-class neighbourhood (not in the Southern Island upper-class property).
* Even for the highest price property ($2943) the security deposit is only $150.

I conclude these 2 extreme outliers as typo error. The decision is to change them to the median value ($150) instead of mean value since security\_deposit data is skewed to the right (Use Rule Engine to make the change).

|  |  |
| --- | --- |
|  |  |

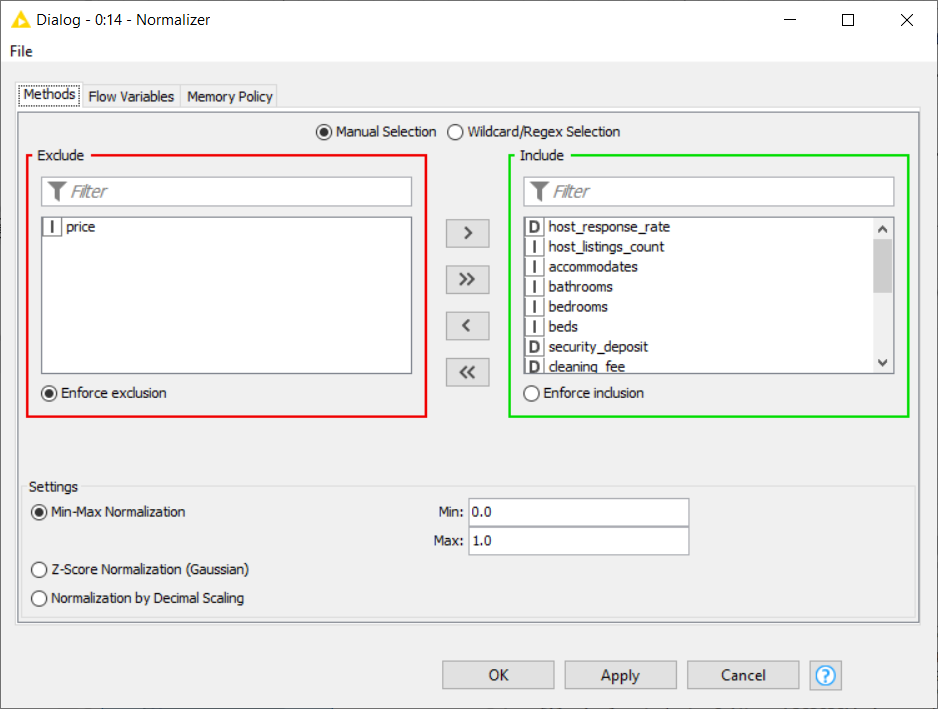


# DAVA – Airbnb Analytics – Clustering or Regression?

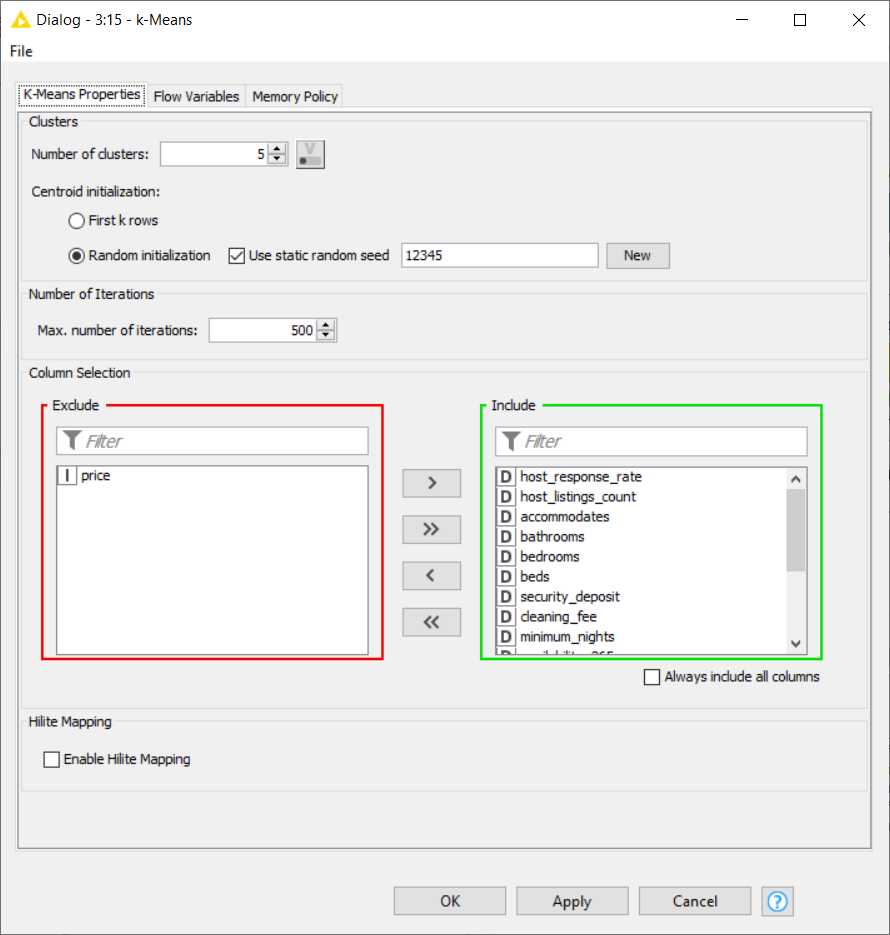
3. The option I have selected is: Option 1 (Clustering)

a) Steps taken: (10 marks)

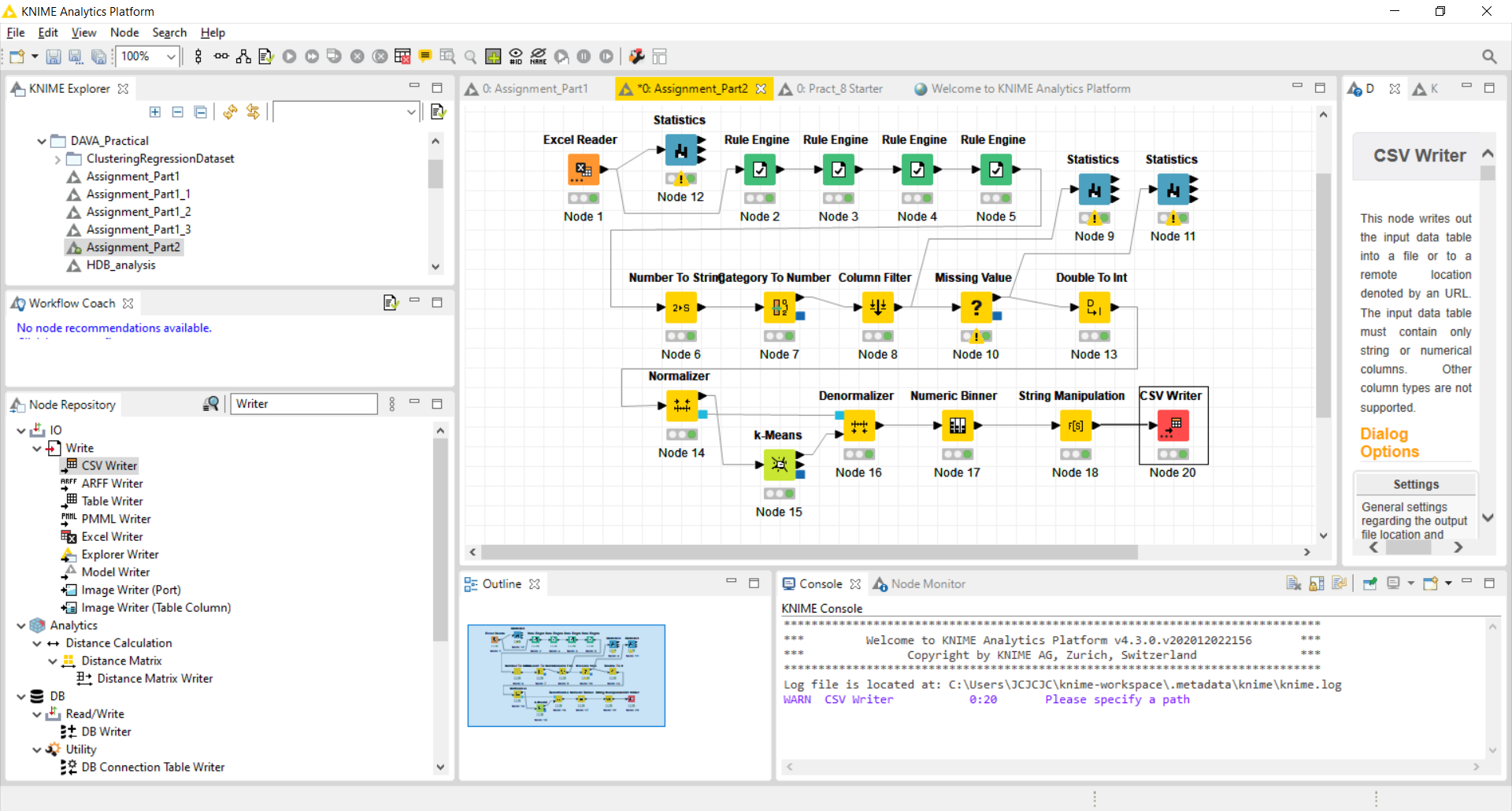
1. Normalize the fields to reduce the effects of unequal ranges across the fields before performing k-Means clustering. Use the Min-Max Normalization, where Min = 0.0 and Max = 1.0, except for the price.



1. Perform k-Means clustering.
   * Number set to 5
   * Use Random Initialization (12345)
   * Set number of iterations to 500
   * Exclude the price to see if the properties can be group into cluster base on their feature.



1. Perform a denormalization to reverse the earlier normalization process together with the k-Means output (that is the cluster column).

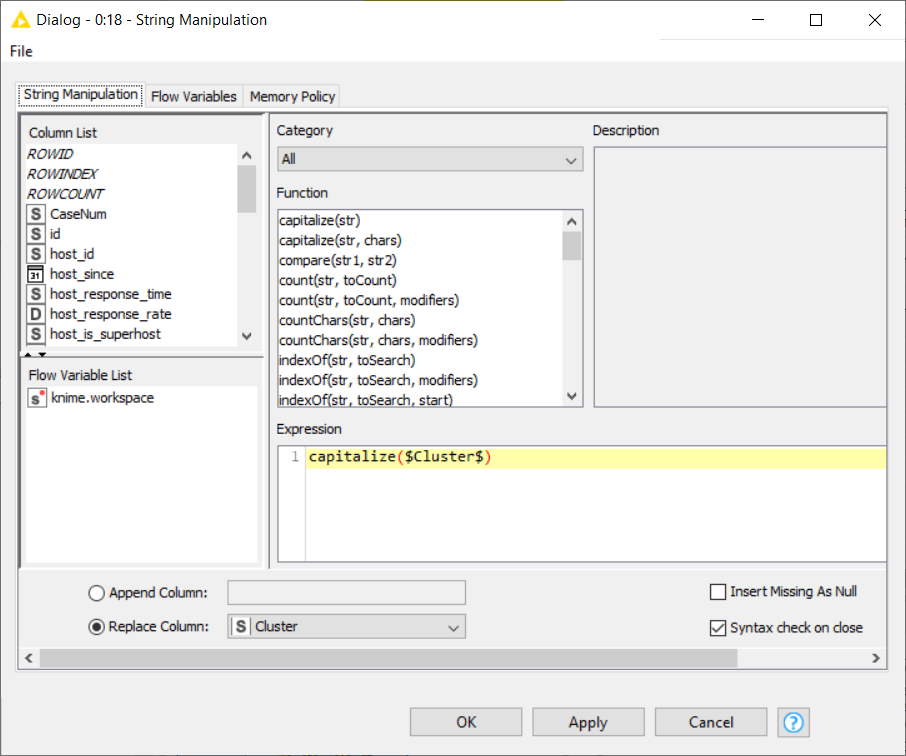


1. Add a Numeric Binned after the denormalization to categorize the price into 5 bins (Low, Mid-Low, Mid, Mid-High and High) base on the prices as shown in the diagram below.

|  |  |
| --- | --- |
|  |  |

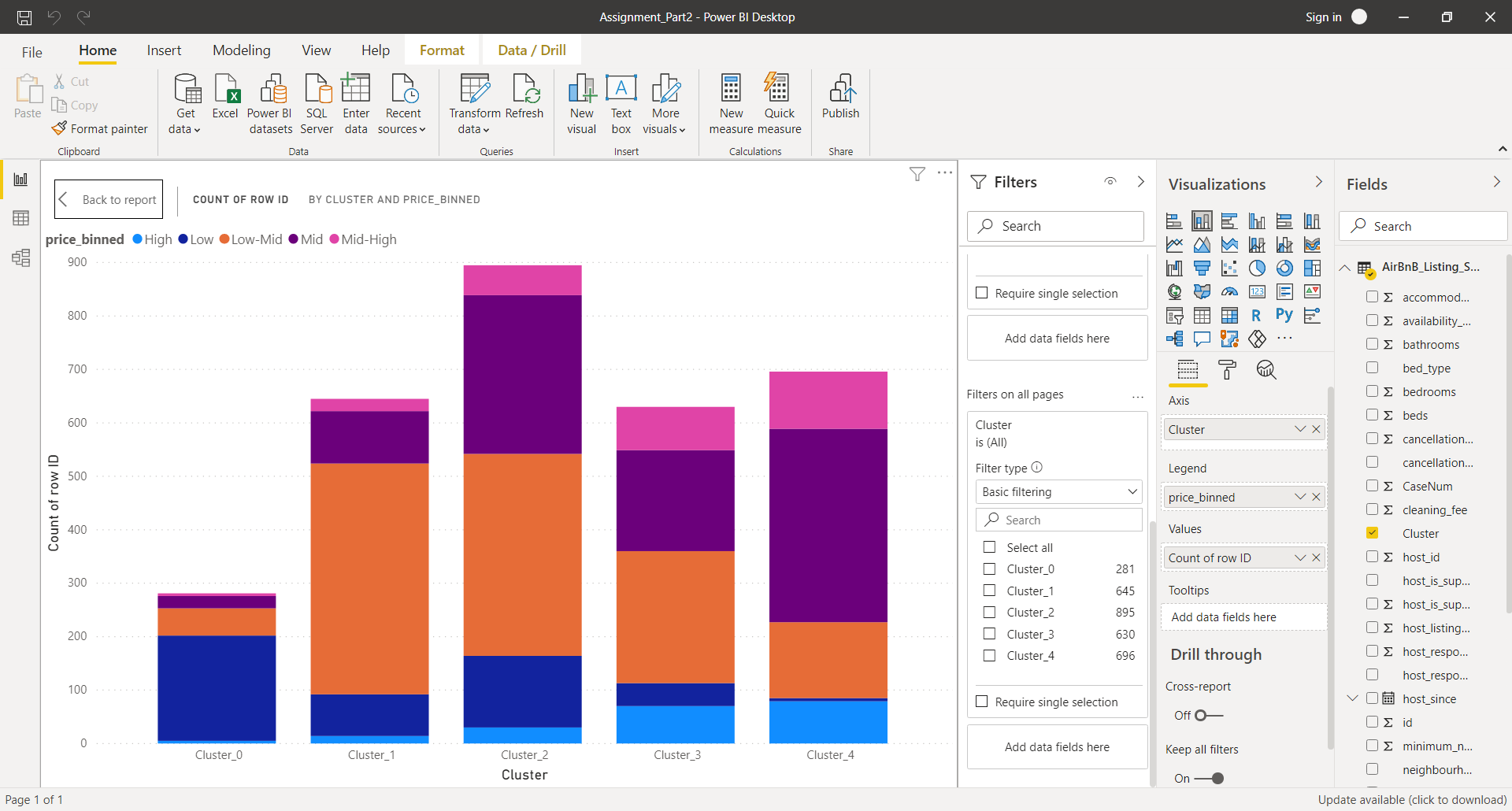
Output : The last 2 columns is the Cluster and price\_binned.

1. Perform a String Manipulation after the Numeric Binned node to capitalize the cluster ids. (from cluster\_0, cluster\_1, cluster\_2 to Cluster\_0, Cluster\_1, Cluster\_2).



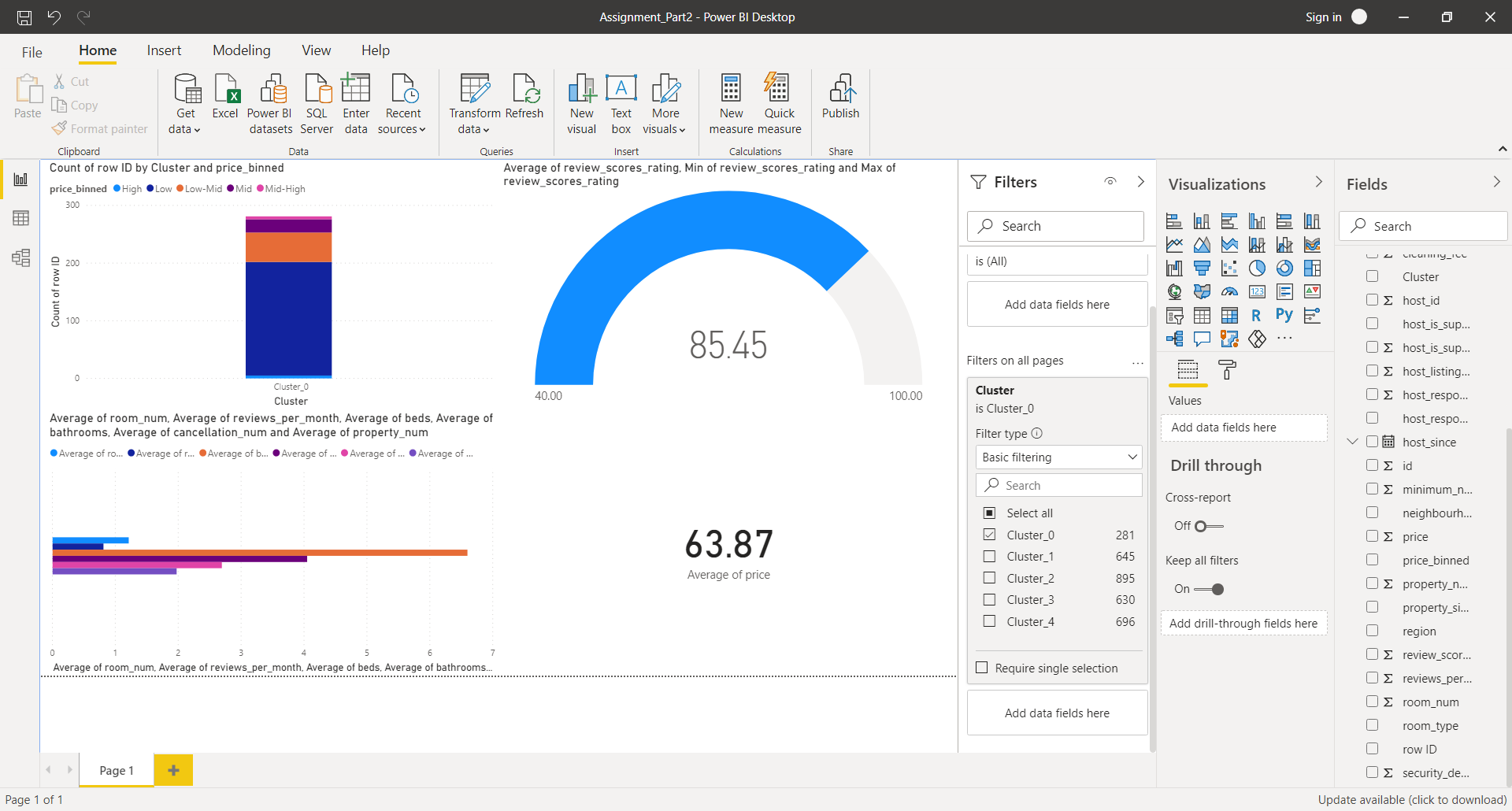
1. Save the output files into a CSV format to be used by PowerBI tool for analysis.

b) My Findings / Explanation: (15 marks)

****

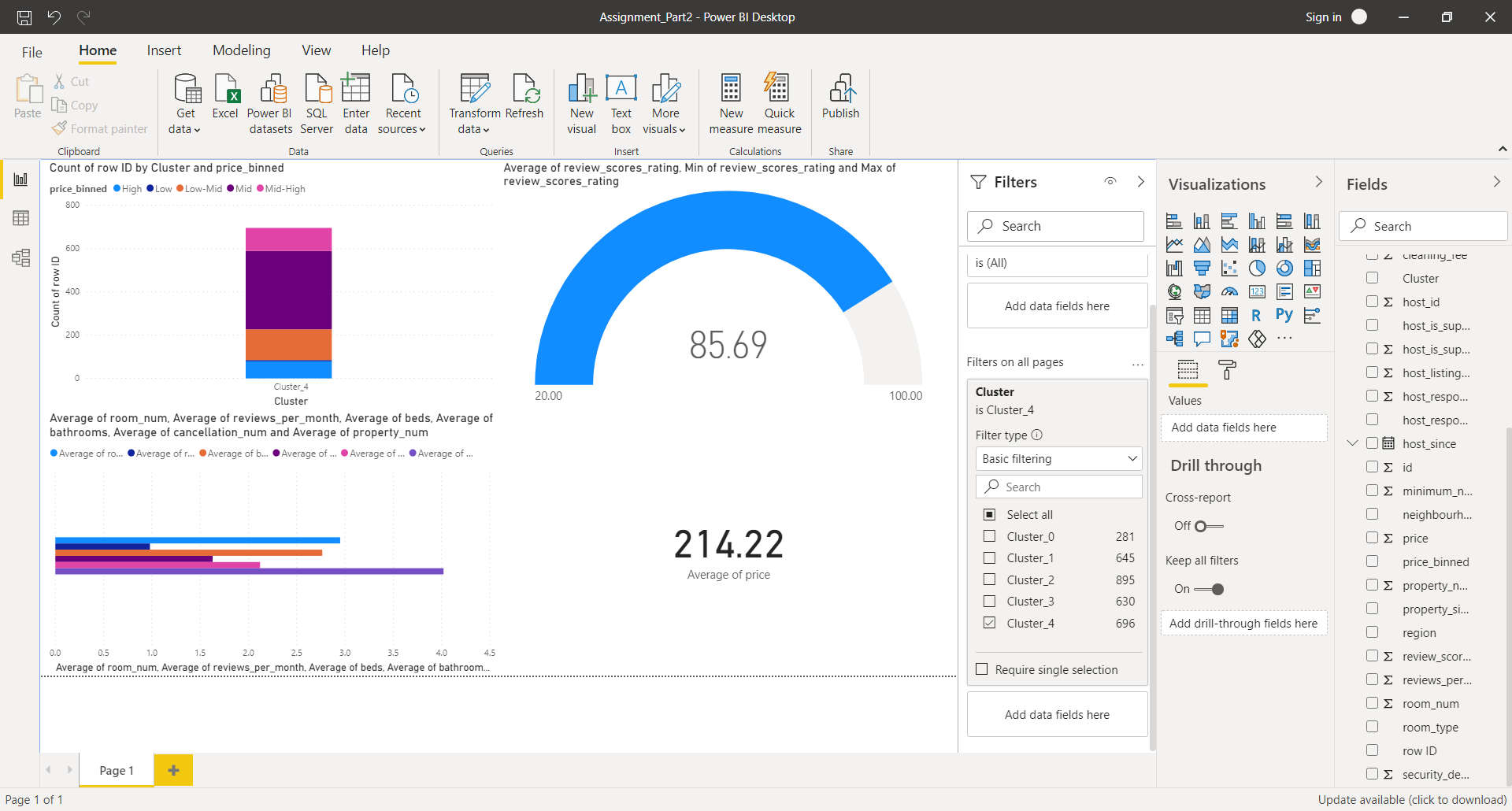
* Base on the chart above, it is possible to group the properties by their features into 5 non-overlapping groups.
* Base on the mixture of property price for each cluster, the cluster does not solely base on the price. **I conclude there is no high degree of correspondence between the cluster and price.**

**Filter : Cluster\_0**



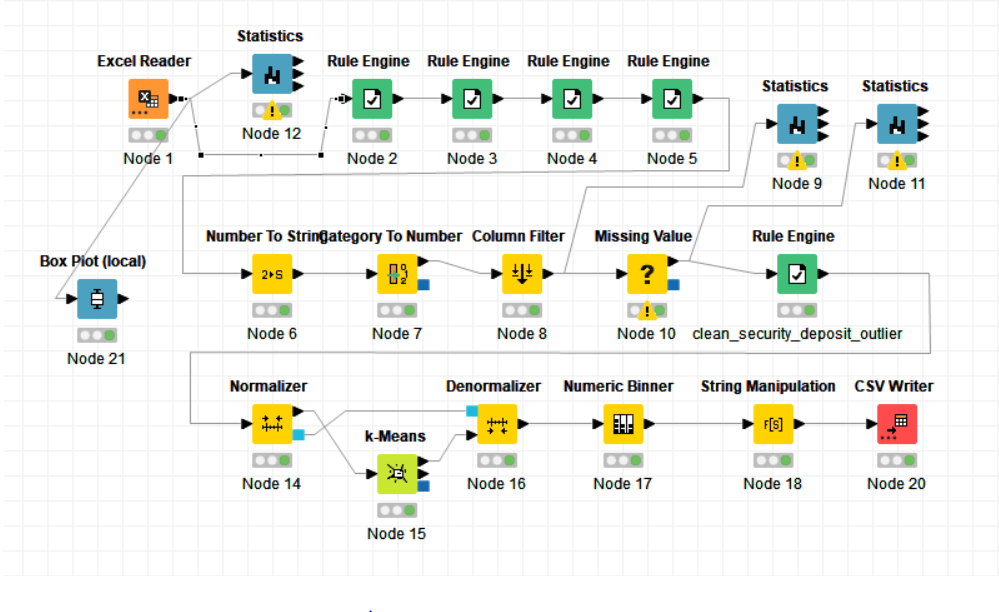
* Lowest reviews rating score among the 5 clusters.
* Lowest average price.
* Properties has low number of rooms
* Properties has high number of beds
* Low average property\_num (that is, it has higher concentration of “special” and “hostel“ type of properties under this cluster)

**Filter : Cluster\_4**



* Average price is the highest.
* Review rating score is low compare to other clusters (except Cluster\_0) even though is most expensive.
* Property has high number of rooms.
* High average property\_num (that is, it has higher concentration of “house” and “apartment” type of properties under this cluster)

# Screenshot of my overall KNIME workflow is shown below.



\*\*\*\*\* END OF PART 2 \*\*\*\*\*