**Project 1: Data Integration, Data Warehousing, Data Pre-processing**

**CS548 Knowledge Discovery and Data Mining – Fall 2016**

**Prof. Carolina Ruiz**

**Students:** Mu Niu & Yue Zhao

|  |  |
| --- | --- |
| **Problems** | **Scores** |
| **I. KDD Process (20 points)** | / 20 |
| **II. Data Preprocessing (65 points)** | /65 |
| **III. Feature Selection (60 points)** | /60 |
| **IV. Exploring Real Data (60 points)** | /60 |
| **V. Data Integration, Data Warehousing and OLAP (50 points)** | /50 |
| **TOTAL written report** | /255 = / 100 |

**Problem I. Knowledge Discovery in Databases (20 points) AT MOST 1 PAGE**

1. (5 points) Define knowledge discovery in databases.

Knowledge discovery in databases is an overall process of discovering useful knowledge from data.

1. (10 points) Briefly describe the steps of the knowledge discovery in databases process.

1st: Input data.

2nd: Data Preprocessing: a process which included feature selection, dimensionality reduction, normalization and data subsetting. The purpose of it is to transform the raw input data into an appropriate format for subsequent analysis

3rd: Data mining: Non-trivial extraction of implicit, previously unknown and potentially useful information from data.

4th: Postprocessing: a process which included filtering patterns, visualization, and pattern interpretation. And in the end, postprocessing will make sure only valid and useful result are incorporated into the decision support system.

5th: data output and get useful information.

1. (5 points) Define data mining.

Data mining: Non-trivial extraction of implicit, previously unknown and potentially useful information from data.

**Problem II. Data Preprocessing (65 points) AT MOST 2 PAGES**

1. (5 points) Three approaches for replacing missing values.



1. (5 points) Keeping the COUNTRY attribute?

No, since it’s not a valid variable, just names makes no attribute to result.

1. (5 points) Transforming the COUNTRY attribute.



1. (5 points) Discretizing AC-S-ED into 4 equal-width intervals. Show your work.



1. (5 points) Discretizing AC-S-ED into 4 equal-frequency intervals. Show your work.



1. (10 points) Discretizing AC-S-ED using new procedure. Show your work.



1. (30 points) Supervised discretization.
   1. (2 points) Results of discretizing LIFE-EXP with Weka.



* 1. (20 points) Describe the process followed by the supervised discretization Weka filter code in your own words. Describe the inputs received by the code, the output it produces and the process it follows to do the discretization.

At the beginning we construct a public class Discretize extends Filter implements SuoervisedFilter. Then initializes filter by “setAttributeIndiced”. After that, we need to set all kinds of options such as “getMakeBinary”, “getUseBinNumbers”, “getUseBetterEncoding”, and “getUseKononenko” and so on. When finished all setting options, we will format input instances structure. And then input instances for filtering, which instances form the total input. Also we should make sure that only numerical input will be used for future, any others may be deleted. We also need to make sure output can be call only all required instances are ready for input. The key point for this code is the algorithm for discretized. First we need to calculate cut points for splitting intervals. It is the same method with textbook by bisecting the initial values so that the intervals give minimum entropy. After that we can use Kononenko’s MDL criterion and Fayyad and Iran’s MDL criterion to test the result. Final we set the output format by “setOutputFormat” to get our final output, the split interval.

* 1. (8 points) Follow the code by hand to show how the LIFE-EXP attribute was discretized.

the LIFE-EXP attribute was discretized by the entropy formula: e = sum(P(i,j)ln(P(I,j))

67.65 is the mean of the 3rd largest value and the 4th one. We need to check all the cutting points to get the smallest entropy.

For the optimal outpoint, the split interval is (-∞, 67.65] and [67.65, +∞), the entropy for left side is: .

The entropy for right side is:.

And we got the smallest entropy for this cutting point:

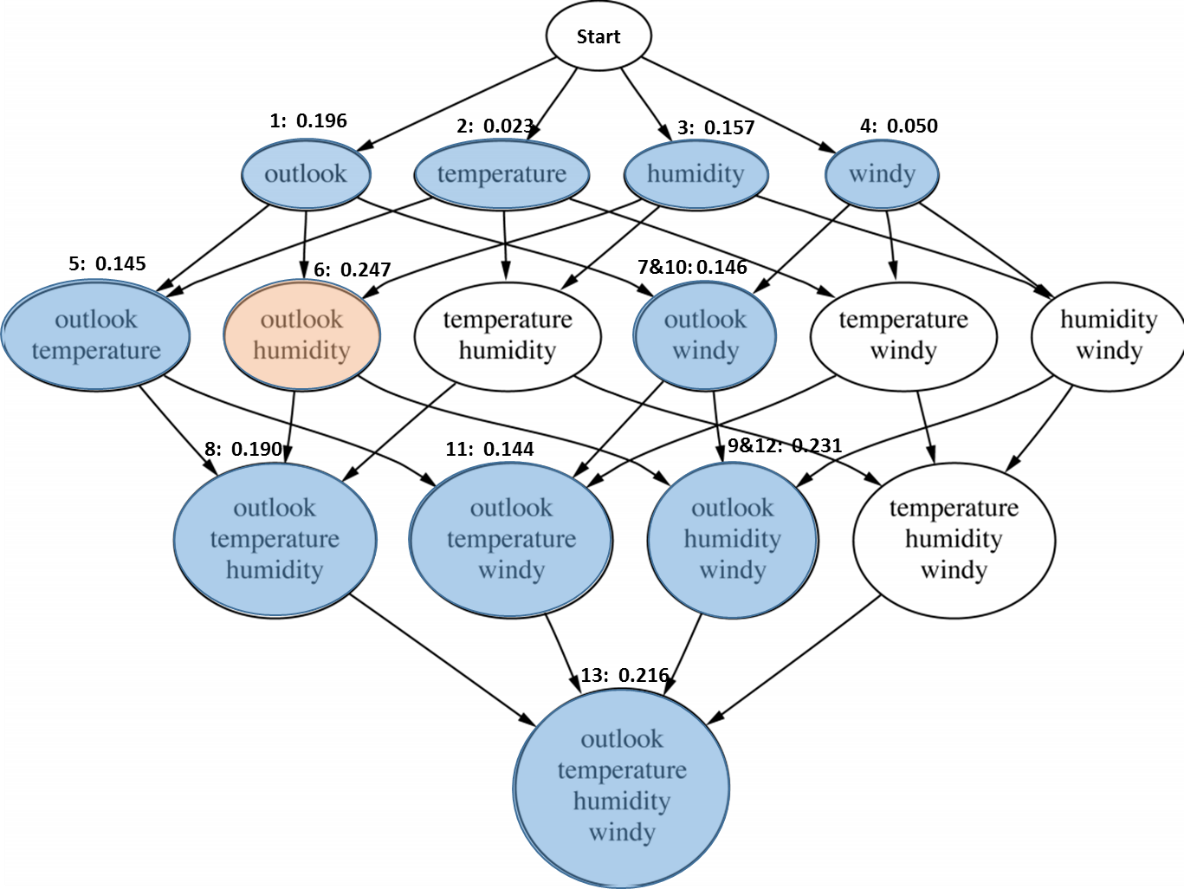
**Problem III. Feature Selection (60 points) AT MOST 1.5 PAGES**

1. (5 points) Attributes selected by CfsSubsetEval over the weather.arff dataset.

Attributes selected by Weka is outlook and humidity.

1. Weka code implementing CfsSubsetEval:
   1. (5 points) Initial subset of attributes under consideration. Forward or backward search?

Initial subset has no attributes, and it’s forward research

* 1. (25 points) Show step by step the process followed by CfsSubsetEval. Use the lattice below to show neatly the order in which it considers the subsets and the goodness value of each of the subsets considered
  2. (25 points) Use the CfsSubsetEval formulas to calculate the goodness of the "best" (sub)set of attributes considered. Show your work.

The best subset chosen by Weka and our algorithm is “outlook”, denoted as “A1” and “humidity”, denoted as A3, and the target attribute “Play” denoted as “C”.

By the method to calculate entropies, we get

Then the correlations between attributes are:

Finally, the goodness of the "best" (sub)set of attributes is:

**Problem IV. Exploring Real Data (65 points) AT MOST 3 PAGES**

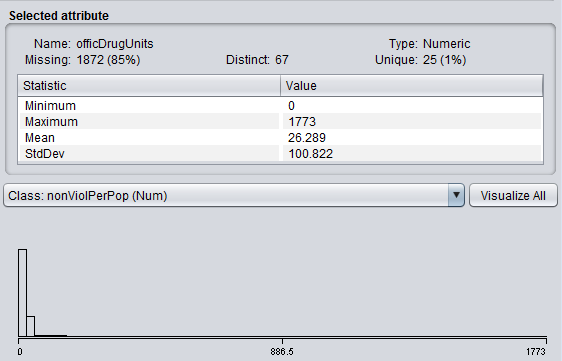
1. Dataset Exploration
   1. (5 points) Two observations of good things about the dataset and two observations of bad things about the dataset (include visualizations).

Good things: 1. The dataset have enough samples.

2. The attributes are explained well

Bad things: 1. Massive missing values

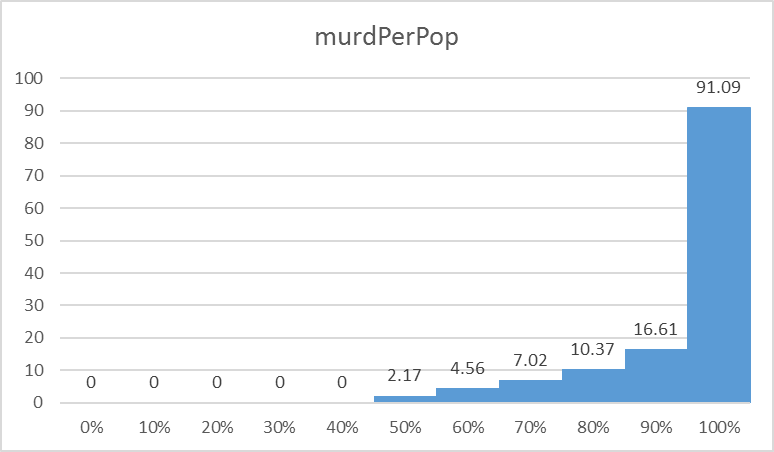
2. There are some redundant attributes and some confused value.

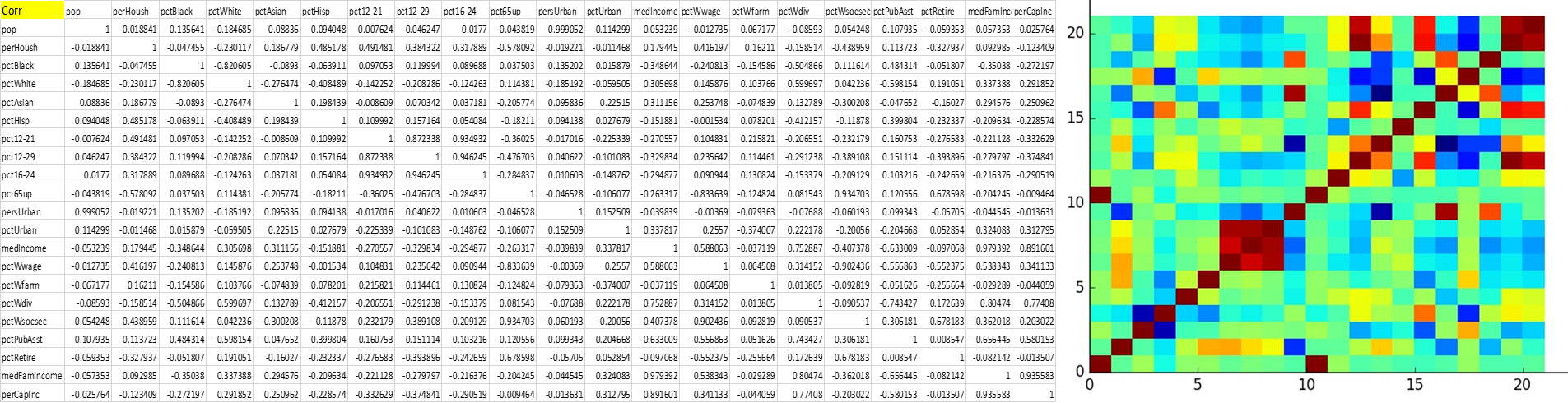


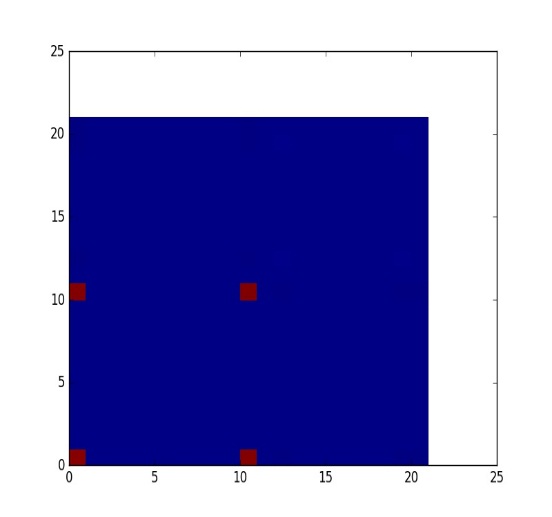
* 1. Percentiles for the *murdPerPop* attribute



Histogram of the *murdPerPop* attribute



* 1. (10 points) Covariance matrix (5 points) and heatmap visualization (5 points)
  2. (10 points) Correlation matrix (5 points) and heatmap visualization (5 points)



(5 points) Two attributes you’d remove, if any, based on these two matrices and why. Explain your answer.

#13, medIncom and #16, pctWdiv.

Since these two attributes have high correlations with other attributes, they can be represented by other attributes.

1. (10 points) Dimensionality Reduction.
   1. PCA results in Weka (5 points). How many dimensions are obtained? How much of the variance is explained? First new component/attribute. Interesting observations.

The original dataset has 121 dimensions (string attributes and attributes with massive missing values are removed).

Two dimensions are obtained after PCA, 98.36% of the variance are explained.

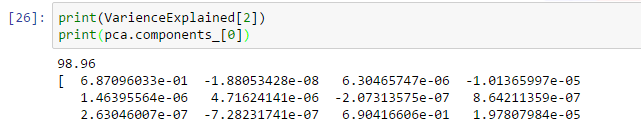
First linear combinations:







* 1. PCA results in Python (5 points) How many dimensions are obtained? How much of the variance is explained? First new component/attribute. Interesting observations.

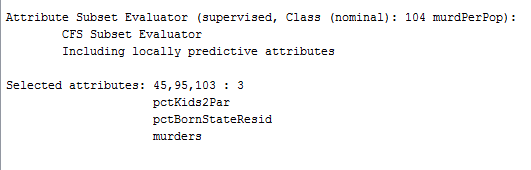


From the result we can see that PCA result in python is exactly the same with PCA result in Weka.

1. (10 points) Feature Selection.
   1. CFS results in Weka (5 points). Which attributes were selected? Interesting observations.

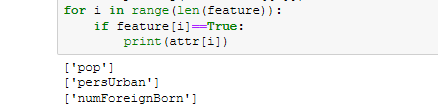
The original dataset has 120 dimensions (string attributes and attributes with massive missing values are removed).

Weka result:



The meaning of the selected attributes is highly different from each other, we can barely find selected attributes

* 1. CFS results in Python (5 points). Which attributes were selected? Interesting observations.



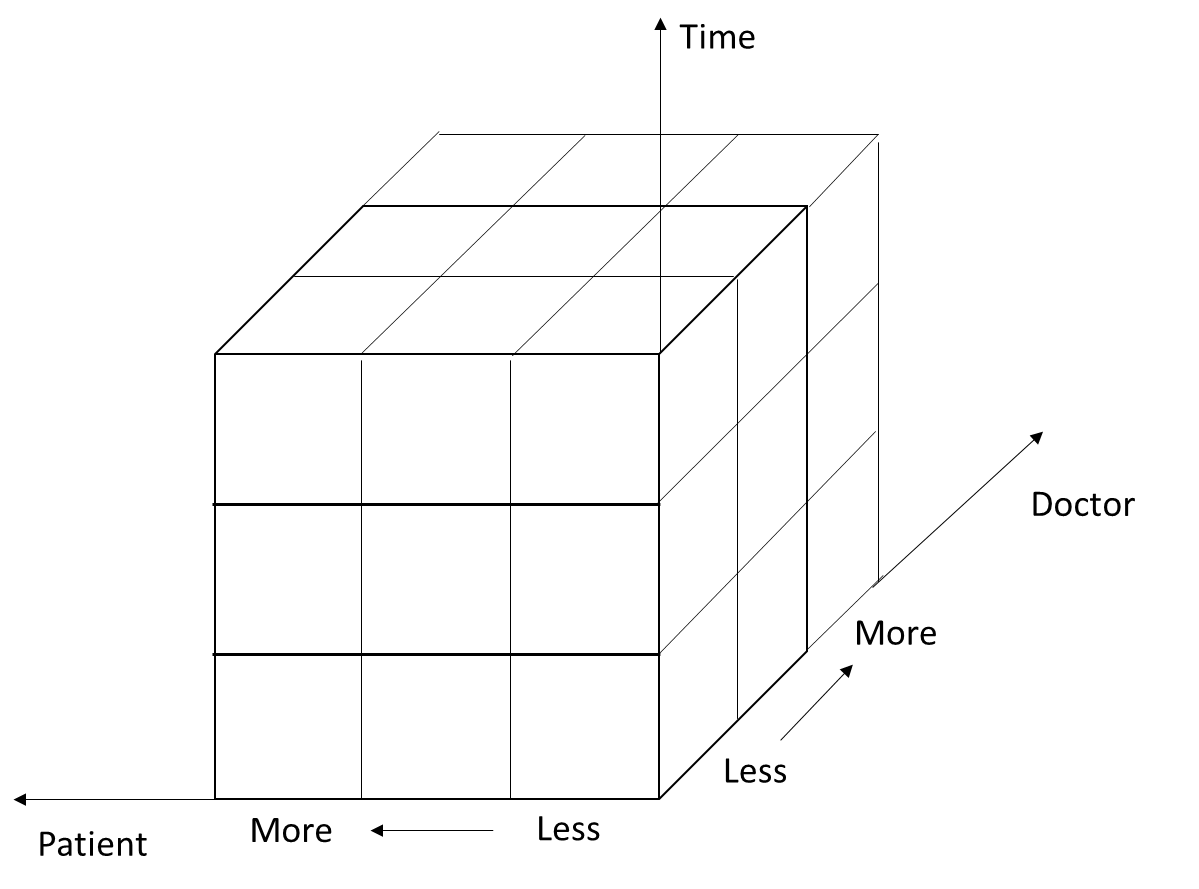
We used K best subset selection in python, and turned out to get different result than Weka.

**Problem V. Data Integration, Data Warehousing and OLAP (50 points) AT MOST 2.5 PAGES**

1. (10 points) Describe the main differences between the mediation approach and the data warehousing approach for data integration.

Mediation approach only keep the latest data while data warehousing approach keeps all data that comes to it

1. Data warehouse consisting of the three dimensions *time*, *doctor*, and *patient*, and the two measures *count* and *charge*, where *charge* is the fee that a doctor charges a patient for a visit.
   1. (5 points) Depiction of this data warehouse as a cuboid / multidimensional array.



* 1. (5 points) Sequence of specific OLAP operations to go from the base cuboid [day, doctor, and patient] to the total fee collected per doctor in 2014. Explain your answer.

1. Roll up from day to month to year

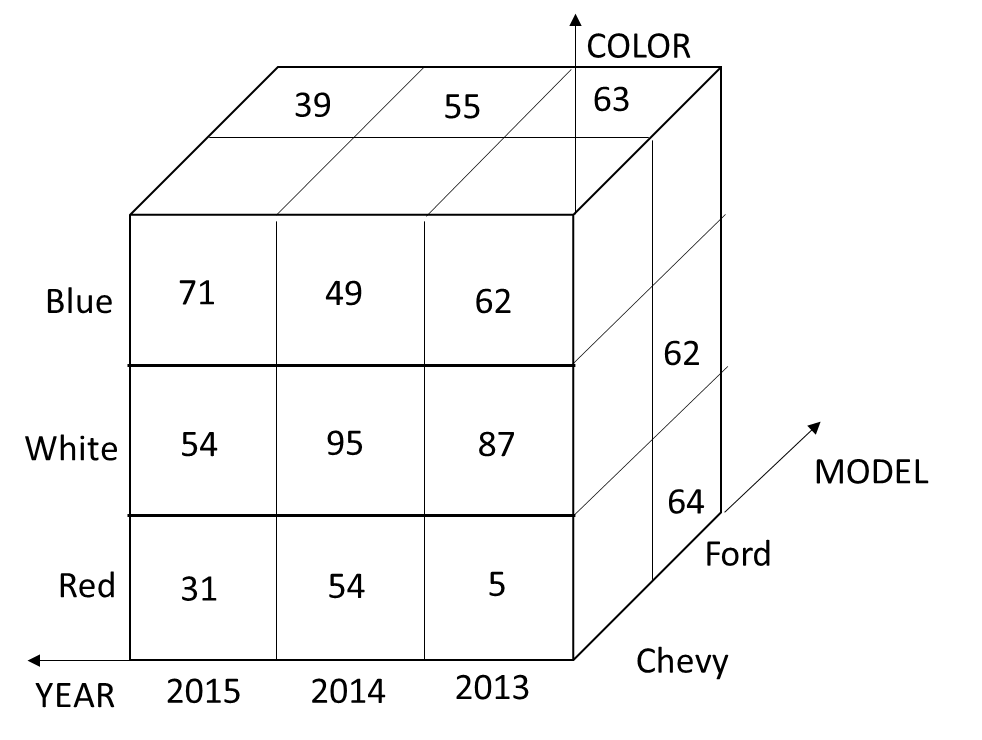
2. Slice for year = “2014”

3. Roll up on patient from individual patient to all

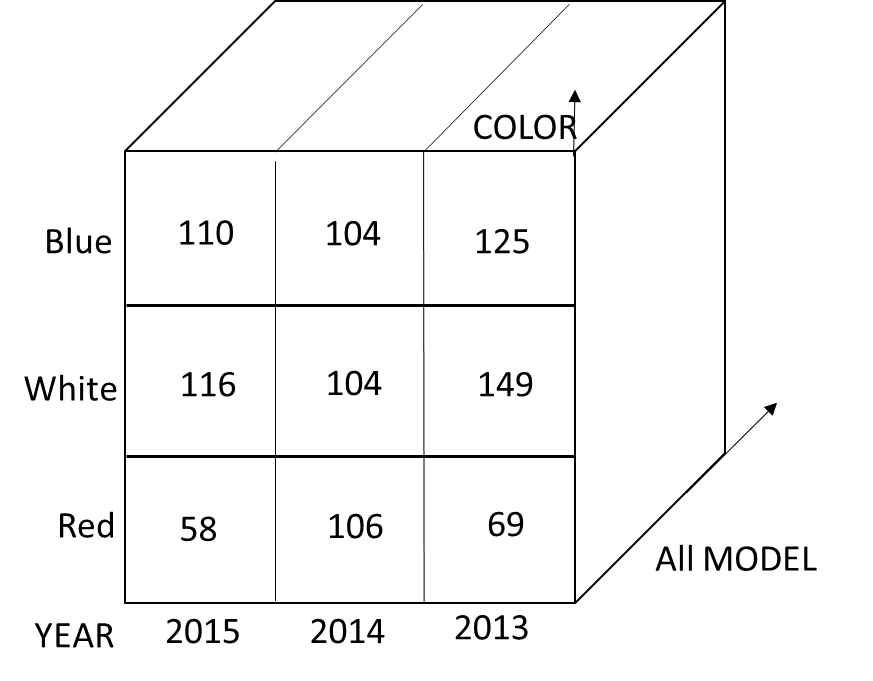
4. Slice for patient = “all”

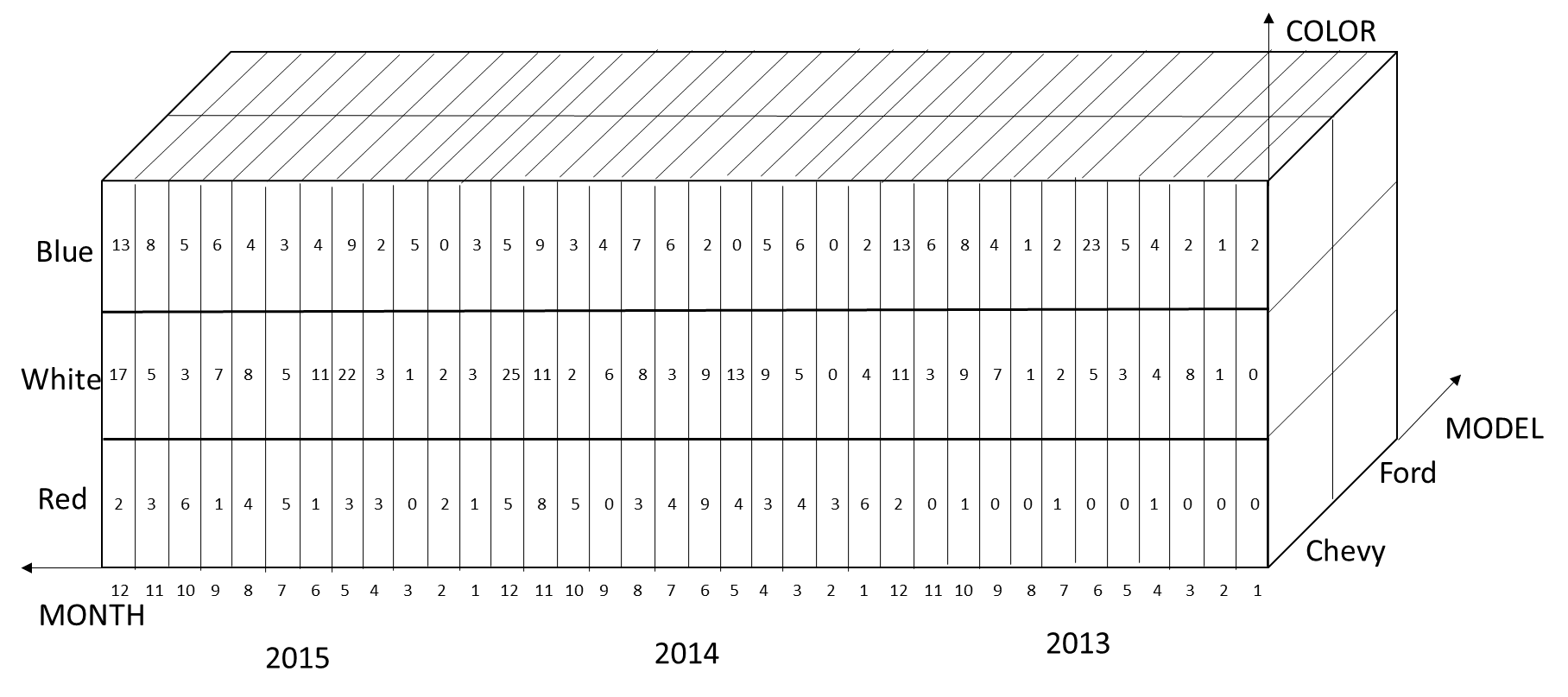
5. Get the list of total fee collected by each doctor in 2014

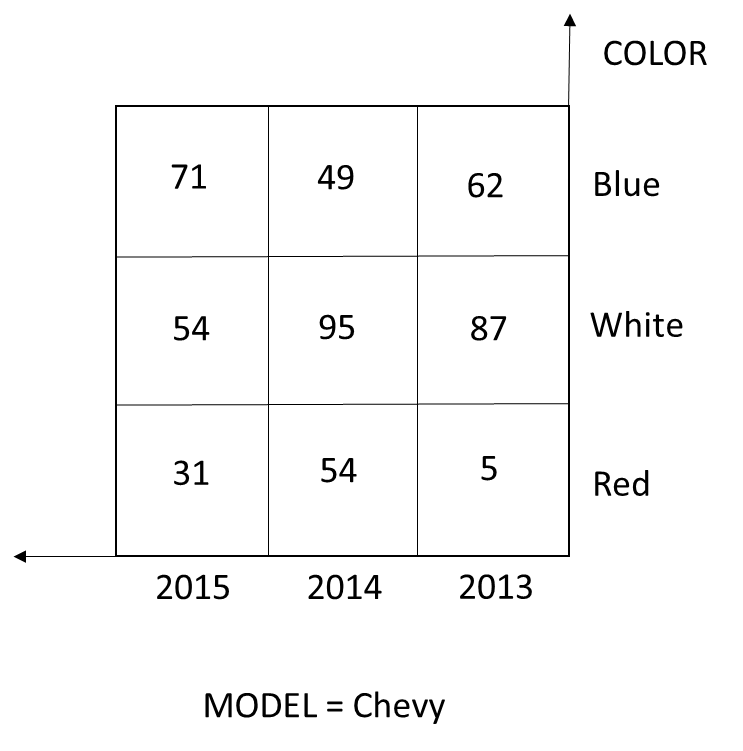
1. Data warehouse consisting of the three dimensions MODEL, YEAR, and COLOR and one measure SALE.
   1. (5 points) Depict the data as a multidimensional cuboid.



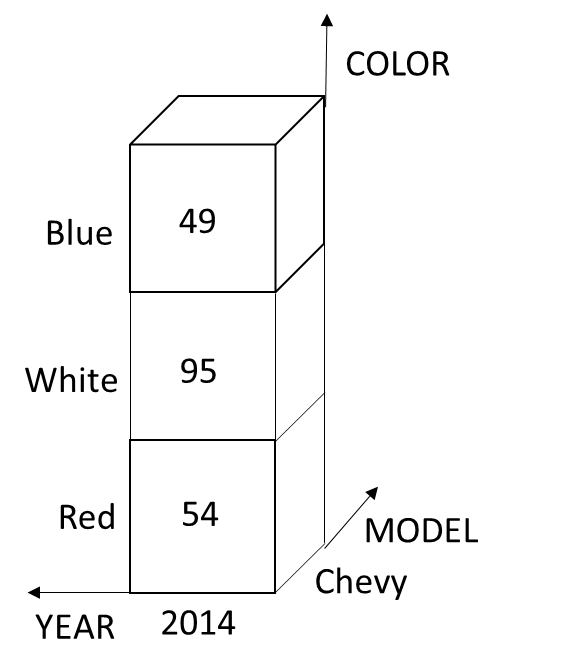
* 1. (5 points) Depict the result of rolling-up MODEL from individual models to all.



* 1. (5 points)Depict the result of drilling-down time from YEAR to month. (Although month data is not provided above, make up a couple of values to illustrate the drill-down operation.) 
  2. (5 points)Depict the result of slicing for MODEL=Chevy.



* 1. (5 points) Depict the result of dicing for MODEL=Chevy and YEAR=2014.



* 1. (5 points) Starting with the basic cuboid model, year, color, sales, what specific OLAP operations should one perform in order to obtain the total number of red cars sold? Make your sequence of operations as efficient as possible.

1. Roll up from “Year” to “All Years”

2. Slice for Color = “Red”

3. Roll up from “Model” to “All Model”

4. Get the total number of red cars sold