Assignment 7:
Clustering

Deadline: Saturday, 9 June 2012 (late submission is not allowed)

This assignment will contribute 12 points to your final grade.

Description

In this assignment, you will analyze a sales record dataset and determine if fraudulent reports have been filed by sales persons. You will perform this by detecting outliers in the reported transaction history. The data set is available in the assignment page (sales.RData) as a R object—loading it into R using the load function will give you a data frame named sales. This data frame has five columns: ID- identifier of sales person, Prod- identifier of product sold, Quant- quantity of the product sold, Val- total value obtained by selling the product, and Insp- a label indicating if the record is fraudulent (fraud), okay (ok) or have not been checked manually (unkn). Each transaction (record) is a report filed by a particular sales person about a particular product.

Given the limited amount of organizational resources available to scrutinize these reports, our objective is to identify a limited number of potential fraud attempts for manual reviewing. We will do this by detecting outliers that do not fit typically observed norms.

Task 1: Data exploration

Load the data set into R using load("sales.RData"). The data set will become available in a data frame named sales. Find answers to each of the following questions, list them in the report, and discuss the R code you wrote to do it.

1. How many unique sales persons and products are represented in the data set?
2. How many records have no quantity information, or no value information, or both?
3. Determine the distribution of Insp labels.
4. Are all sales persons performing equally well in terms of the number of reports filed? Are all products equally “hot” in terms of quantity sold? Create informative plots to show your conclusions.
5. Compute the unit price in each record and append it to the sales data as a sixth column. Unit price is defined as \( \frac{Val}{Quant} \). Ideally the unit price of a product should stay the same over a short period of time. We may assume a normal distribution for the unit price of a product, meaning most of the time the product is sold at a particular price, while sometimes sales persons may sell it at a lower or higher price to achieve commercial goals. However, too much deviation could be indicative of fraud or error!
6. What is the median unit price of each product? The aggregate function will come in handy here. Remember, there are NA values in the data that should be ignored (typically done using na.rm = T as an argument).
7. Using the median unit price of products, determine the 5 most expensive and 5 cheapest products. Observe the variability in the prices of the cheapest and the most expensive products. Do this by creating a boxplot of the prices of the products, one for each. A boxplot will show you the first quartile, median and third quartile prices (do a google if you do not know what these terms mean). What differences do you observe in the two plots?
8. Is everyone doing equally good in sales? Find out the top 100 and bottom 2000 sales persons based on the total value they sold. What percentage of the total sales value is contributed by the top 100 and what percentage is contributed by the bottom 2000? Anything interesting!?

Answer to all questions can be generated fairly easily in R. If you are writing more than 5-6 lines of code for a question (most are just 2 lines), then you are not using the power of R.

**Task 2: Data cleanup**

You should have noticed in Question 2 of the previous task that there are missing values in the data set. Perform the following steps to clean the data.

1. Remove all records that has both `Quant` and `Val` missing. How many did you remove?
2. Determine if there is any product with a missing value of `Quant` in all the records where it appears. If so, remove those records. State in your report which products were removed, if at all.
3. Determine if there is any product with a missing value of `Val` in all the records where it appears. If so, remove those records. State in your report which products were removed, if at all.
4. Recompute the median unit price of each product using only records that are not labeled as `fraud` (you cannot trust the numbers in such records). Use these prices to replace the missing value (it would be either `Quant` or `Val`) in the records. Remember the relationship: $\text{UnitPrice} = \frac{\text{Val}}{\text{Quant}}$. Use the ceiling function to avoid fractional `Quant`.
5. Compute and replace the `NA` in unit price of those records that earlier had a missing quantity or value.
6. Remove all records that are labeled as `fraud`.

**Task 3: Outlier ranking**

Our assumption for outlier detection is that the unit price of a product should not deviate significantly from its median price. Any record in which it happens should be flagged. The problems with this approach lies in determining the meaning of “significant”, as well as the lack of a severity measure for the flagged records. Outlier ranking seeks to address this by ranking each record in terms of a score; the higher the score, the more likely it is an outlier. You will test two different methods to do this.

**A quartile-based approach**

Compute the first quartile, second quartile (also known as median), and third quartile of the unit prices corresponding to each product. Let’s denote them as $l_p$, $m_p$ and $h_p$ for product $p$. Look into the `boxplot.stats` function. The outlier score of a record $x$ with $\text{Prod} = P$ and unit price $U$ is given as

$$OS_{qba}(x) = \begin{cases} \frac{|U - m_P|}{h_p - l_p}, & h_p \neq l_p \\ \frac{|U - m_P|}{m_P}, & h_p = l_p \end{cases}$$

Create a plot showing the `Val` vs. `Quant` of all records (use log scale). Add to this plot (use different color), the `Val` and `Quant` of 1000 most prominent outlier records. Discuss your observations on any trend you see on these outlier records. What fraction of these 1000 outlier records have already been marked `ok`? Is that an unexpected result?
A clustering-based approach

Outlier ranking can also be performed using the dendogram created by an agglomerative clustering algorithm. Recall that the hclust function returns a member called merge that holds a history of the merging process (how instances were merged at each step of the algorithm). Note that outliers are relatively different than regular instances. Hence, they are unlikely to be merged with other instances in the early steps of the clustering process. If it happens, it would most likely be with another outlier. When they are finally merged with a different cluster (in a later step), the other cluster is expected to have already become significantly well populated (comparatively larger in size). Therefore, this approach computes the outlier score using cluster sizes. At each step $i$ of the merging, two clusters $clst_{left}$ and $clst_{right}$ are merged. Assign a score $s_i(x)$ to every record in $clst_{right}$ and $clst_{left}$ as

$$s_i(x) = \begin{cases} 
\max \left( 0, \frac{\text{size}[clst_{left}] - \text{size}[clst_{right}]}{\text{size}[clst_{left}] + \text{size}[clst_{right}]} \right), & x \in clst_{right} \\
\max \left( 0, \frac{\text{size}[clst_{right}] - \text{size}[clst_{left}]}{\text{size}[clst_{left}] + \text{size}[clst_{right}]} \right), & x \in clst_{left} 
\end{cases}$$

For all other records not in either $clst_{right}$ or $clst_{left}$, the $s_i$ is zero. Once the $s_i$ values are computed for each merging step, the outlier score of record $x$ is computed as

$$OS_{cba}(x) = \max_i s_i(x).$$

To use the above process, first determine the records corresponding to a product. Using the unit price of those records, apply hclust. Use the “ward” method when calling hclust. Compute $OS_{cba}$ for all those records (you will have to write some code R). Repeat the steps for every other product. Create similar plots as in the case of the quartile-based approach. Do the methods agree in the most of the top 1000 outliers? What records are typically chosen as top outliers?

Report

Your report is NOT a collection of statements corresponding to the discussion asked from you. It must be complete in itself, primarily including the problem you are solving, description of the data, the method employed, relevant R code, the experiments performed, the results you observed, informative (with labels, captions, and the remaining glitter!) and readable plots, and most importantly, a discussion along the lines of what is asked in the tasks. You are strongly encouraged to include your own observations and discussions outside of what is asked. Tell me a story!

Grading

This assignment will be graded out of 120 points. The points split is as follows:

<table>
<thead>
<tr>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
<th>Report</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>20</td>
<td>40</td>
<td>30</td>
</tr>
</tbody>
</table>

The points you receive in each task will depend on how accurately you complete the task. In addition, your R code should be clean and well-commented. Sticking to directions is important! Additional explorations (outside of what is asked in the tasks) is most welcome. The report will be graded based on content, structure, formatting (I do not want to see half of a page empty because the plot does not fit there — reorganize!!), spelling, grammar, and other usual characteristics of a good scholarly document.

Submission

Submit the pdf report (not the Latex or Lyx source) and the R code into Blackboard by Saturday, 9 June 2012, 11:59 PM. Late submissions are NOT allowed for this assignment.

You must work alone on this assignment.