# REDHYTE
# An Interactive Platform for Rapid Exploration of Data and Hypothesis Testing

## USER TUTORIAL


**DEVELOPED BY:**

Wei Zhong Toh[1], Limsoon Wong[2], Kwok Pui Choi[1]

[1]Faculty of Science, National University of Singapore (NUS)
[2]School of Computing, NUS

**CONTENTS**

# 1. ABOUT REDHYTE

Welcome to Redhyte.

Redhyte is an interactive platform that epitomizes the novel idea of exploratory hypothesis testing. Redhyte, short for "rapid exploration of data and hypothesis testing", is a hypothesis mining where users start off with an initial domain knowledge-driven question, which Redhyte uses to mine for relevant and interesting hypotheses. These hypotheses seek to deepen the user's understanding of his or her data. In addition, Redhyte provides basic functionalities for data visualizations, checking of parametric test assumptions, and data manipulation.

Hypothesis mining

As the term suggests, hypothesis mining is concerned with the search of interesting hypotheses from a given dataset. In order to do so, Redhyte puts together the user's domain knowledge, the well-established framework of statistical hypothesis testing, and classification techniques from data mining. To evaluate the interestingness of mined hypotheses, Redhyte utilizes a set of hypothesis mining metrics, so as to divert the user's attention to the most interesting collection of hypotheses mined by Redhyte.

Redhyte is developed using and powered by the statistical programming language R, with the actualization of the user interface made possible by the `shiny` package.

Rationale

Hypothesis testing is a technique used by many non-statistician data analysts – the idea of comparing lung cancer incidence between two subpopulations, say smokers and non-smokers, is intuitive and easy to understand. It is also easy to search through a small dataset of, say, 10 variables, (e.g. in an epidemiological study) and identify any existing statistically and practical significant phenomena and trends. However, in the current Big Data era, the search for statistical and practical significance becomes a non-trivial task – formulating and testing a small hypothesis in large dataset is both wasteful and flawed.

Using data mining techniques, Redhyte aims to regard hypothesis testing in a more comprehensive manner. The objective of Redhyte is to identify, based on the initial domain knowledge-driven question that the user had in mind, practical and insightful hypotheses.

User interface

Redhyte is fundamentally a web application that renders in a web browser, such as Google Chrome or Mozilla Firefox. Redhyte's user-facing interface is organised into tabs, with each tab housing a specific functionality that Redhyte provides. These tabs contain the settings control, data preview, data visuzalization, initial test module, test diagnostics module, context mining module, mined hypothesis formulation and scoring module, and finally log documentation.

Redhyte was developed by Wei Zhong Toh, Limsoon Wong and Kwok Pui Choi at the National University of Singapore, Faculty of Science and School of Computing, and is part of Toh's undergraduate Honours work.

To feedback regarding improvements or bugs, please drop an email to tohweizhong[at]u[dot]nus[dot]edu. We are hopeful that Redhyte, and exploratory hypothesis testing and mining, can be a good addition to the arsenal of the scientist and the data analyst, by giving them an additional tool for the rapid exploration of data.

## 2. GETTING STARTED

There are currently two main ways Redhyte can be put to use: the user may choose to install R on a personal computer, and import Redhyte's GitHub repository to use Redhyte for analysis locally. Alternatively, Redhyte has been deployed as an R shiny application at shinyapps.io, and can be accessed via the web browser.

### 2.1 Using Redhyte for analysis locally

Go to the following URL: https://github.com/tohweizhong/redhyte, and look for a "Download ZIP" option near the bottom of the page. Clicking on this option will initiate a download of the zipped folder named "redhyte-master.zip". Unzip the folder into a desired directory, and fire up Rstudio. Ensure that the `shiny` package has been installed.

Finally, to run Redhyte, open the project file in the unzipped folder. Click on "Run app" in Rstudio.

### 2.2 Using Redhyte at shinyapps.io

Simply visit the following link: https://tohweizhong.shinyapps.io/redhyte/, and Redhyte is available for analysis.

# 3. EXPLORATORY HYPOTHESIS TESTING WITH REDHYTE

## 3.1 Settings and Data preview

The first two tabs in the interface are the Settings and the Data Preview tabs. In the Settings tab, users can have specific control over how Redhyte treats the input data. Options such as file types and transposing allow certain degree of flexibility in the data format. Also housed in the Settings tab are settings used in test diagnostics, context mining and hypothesis mining, to give the user more control over the hypothesis mining process. These settings include the maximum number of classes in the categorical attributes of the dataset, the p-value threshold for switching to non-parametric tests in the Test diagnostics module, minimum classification accuracy for context mining models, number of context attributes to mine for, and class-ratio threshold for class-imbalance learning in context mining (refer to section 4.4.3), and minimum cell support for mined hypotheses. The Data preview serves as a simple functionality for users to have a quick peek at the input dataset.

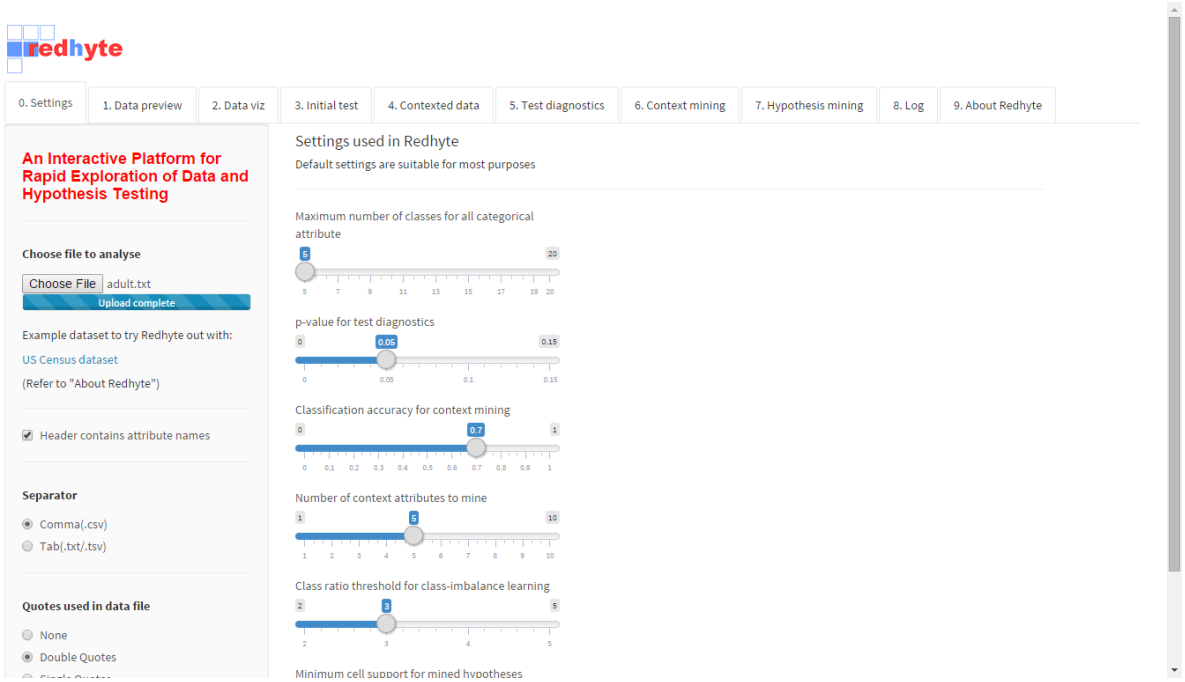
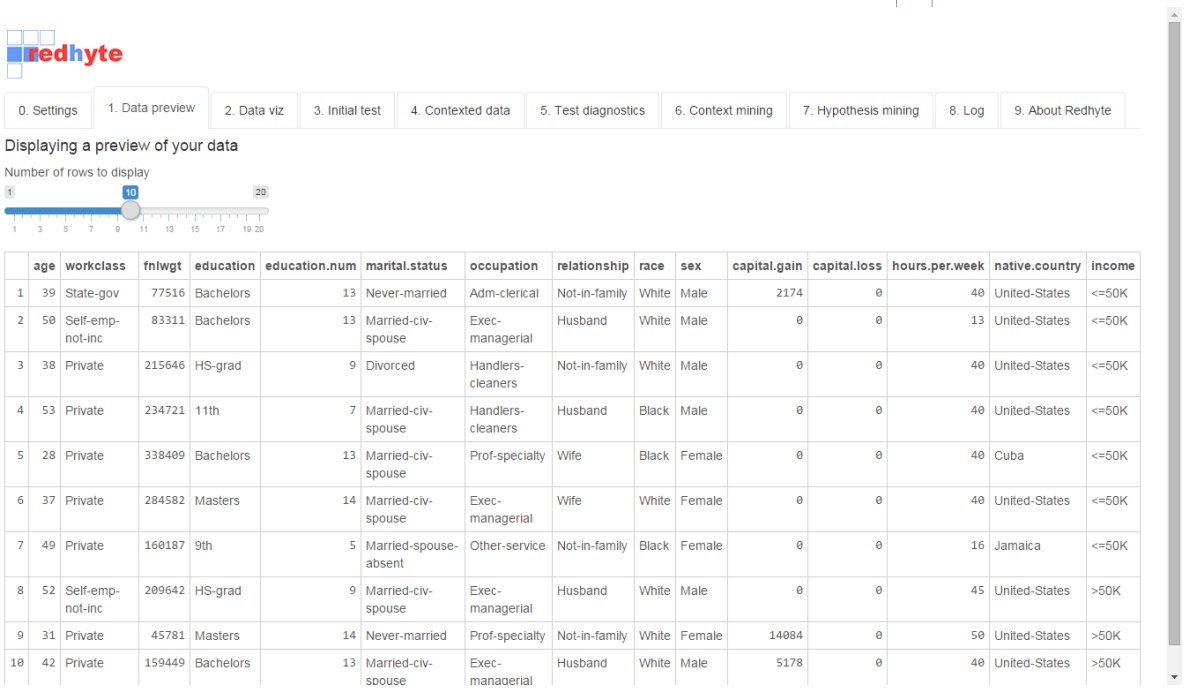



Figure 1: Screenshot of the Settings tab in Redhyte




| | age | workclass | fnlwgt | education | education.num | marital.status | occupation | relationship | race | sex | capital.gain | capital.loss | hours.per.week | native.country | income |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 39 | State-gov | 77516 | Bachelors | 13 | Never-married | Adm-clerical | Not-in-family | White | Male | 2174 | 0 | 40 | United-States | <=50K |
| 2 | 50 | Self-emp-not-inc | 83311 | Bachelors | 13 | Married-civ-spouse | Exec-managerial | Husband | White | Male | 0 | 0 | 13 | United-States | <=50K |
| 3 | 38 | Private | 215646 | HS-grad | 9 | Divorced | Handlers-cleaners | Not-in-family | White | Male | 0 | 0 | 40 | United-States | <=50K |
| 4 | 53 | Private | 234721 | 11th | 7 | Married-civ-spouse | Handlers-cleaners | Husband | Black | Male | 0 | 0 | 40 | United-States | <=50K |
| 5 | 28 | Private | 338409 | Bachelors | 13 | Married-civ-spouse | Prof-specialty | Wife | Black | Female | 0 | 0 | 40 | Cuba | <=50K |
| 6 | 37 | Private | 284582 | Masters | 14 | Married-civ-spouse | Exec-managerial | Wife | White | Female | 0 | 0 | 40 | United-States | <=50K |
| 7 | 49 | Private | 160187 | 9th | 5 | Married-spouse-absent | Other-service | Not-in-family | Black | Female | 0 | 0 | 16 | Jamaica | <=50K |
| 8 | 52 | Self-emp-not-inc | 209642 | HS-grad | 9 | Married-civ-spouse | Exec-managerial | Husband | White | Male | 0 | 0 | 45 | United-States | >50K |
| 9 | 31 | Private | 45781 | Masters | 14 | Never-married | Prof-specialty | Not-in-family | White | Female | 14084 | 0 | 50 | United-States | >50K |
| 10 | 42 | Private | 159449 | Bachelors | 13 | Married-civ-spouse | Exec-managerial | Husband | White | Male | 5178 | 0 | 40 | United-States | >50K |

Figure 2: Screenshot of the Data preview tab in Redhyte

### 3.2 Data visualization

The next tab in the interface houses the Data visualization tab. Here, users can select two attributes from the input data, using which Redhyte renders the appropriate statistical graphics for visualisations, such as histograms, barplots, scatterplots, boxplots, and spineplots. The type of statistical graphic rendered depends solely on the type of selected attributes. For example, if the selected attributes are both numerical, a scatterplot is rendered. If the selected attributes are each numerical and categorical, boxplots are rendered, as in Figure 3.

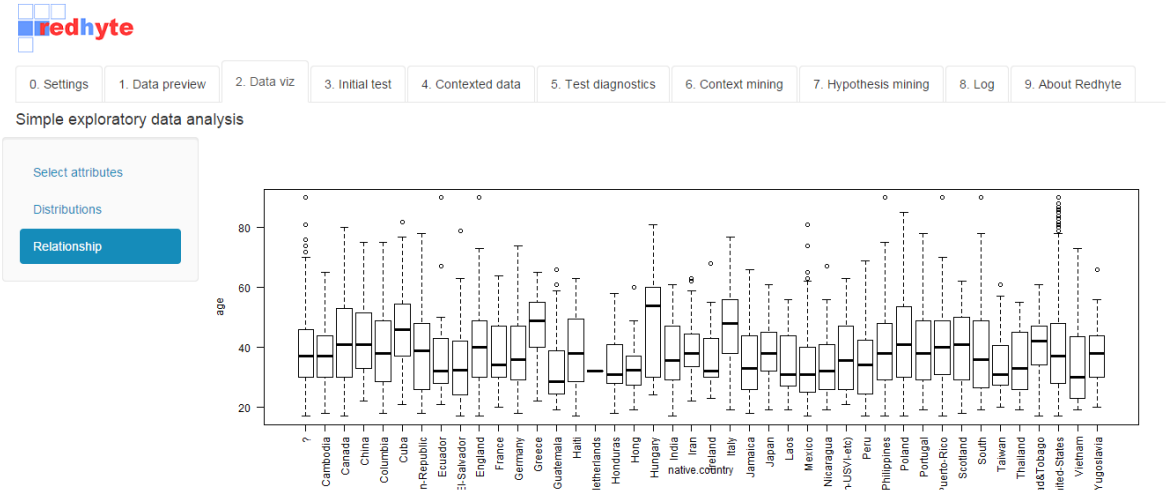



Figure 3: Screenshot of the Data visualization tab in Redhyte

### 3.3 Initial test and Test diagnostics, Contexted data

Following the Data visualization tab is the Initial test module, where users set up their initial hypothesis. After the initial hypothesis is set up, the relevant table(s) and test(s) are rendered and conducted. The following module is the Test diagnostics module, within which diagnostic tests such as the F-test and/or hypothesis analysis is done. Using Redhyte up till this point in the framework may already be sufficient for some users, as the hypothesis and test that they were interested in would be sufficiently addressed by the Initial test and the Test diagnostics module.

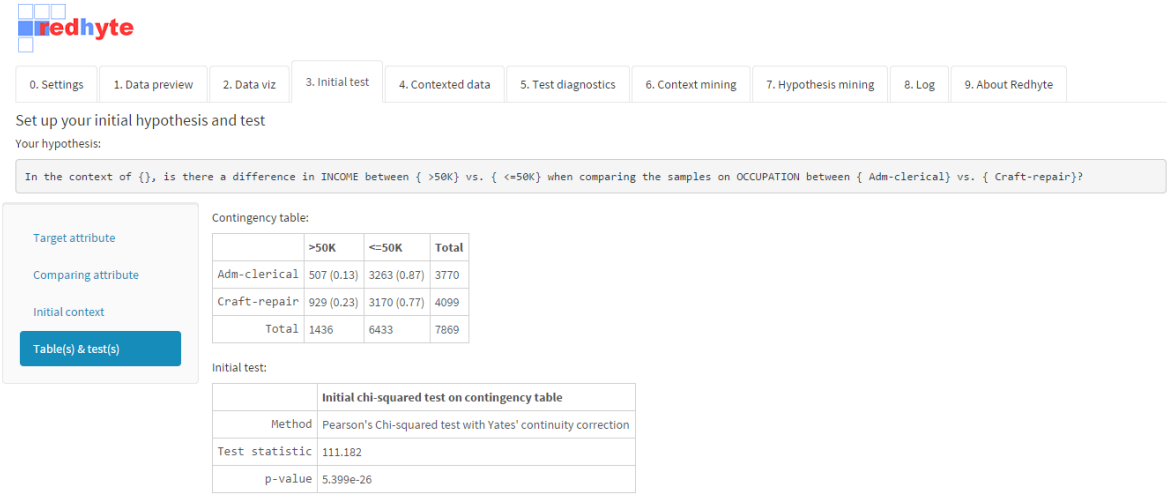
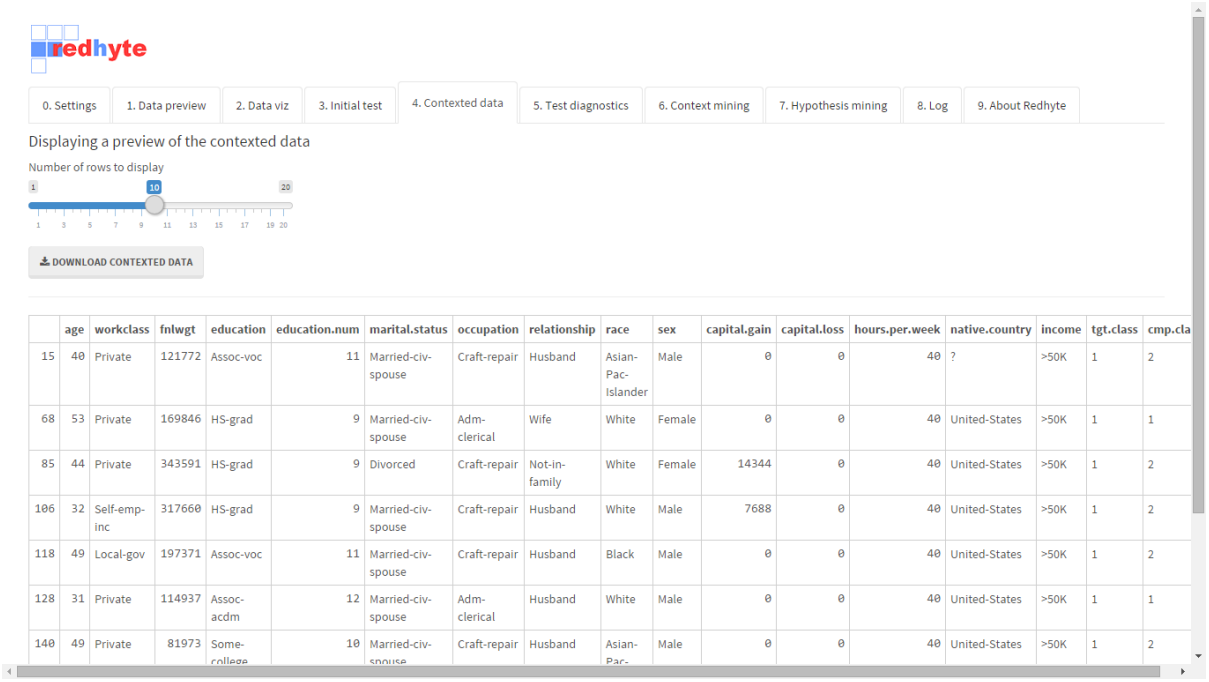



Figure 4: Screenshot of the Initial test module in Redhyte

The Contexted data tab allows users to have a quick look at the subset of the original input data that is relevant to the initial hypothesis ("contexted" simply means the addition of context items into a hypothesis, which makes the hypothesis more specific and the underlying dataset relevant to the hypothesis smaller). Furthermore, the less programming-savvy data analyst may make use of the Initial test module to do some simple subsetting of the original data, and download the data subset from the Contexted data tab for analysis in another platform or software.




| | age | workclass | fnlwgt | education | education.num | marital.status | occupation | relationship | race | sex | capital.gain | capital.loss | hours.per.week | native.country | income | tgt.class | cmp.cla |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 15 | 40 | Private | 121772 | Assoc-voc | 11 | Married-civ-spouse | Craft-repair | Husband | Asian-Pac-Islander | Male | 0 | 0 | 40 | ? | >50K | 1 | 2 |
| 68 | 53 | Private | 169846 | HS-grad | 9 | Married-civ-spouse | Adm-clerical | Wife | White | Female | 0 | 0 | 40 | United-States | >50K | 1 | 1 |
| 85 | 44 | Private | 343591 | HS-grad | 9 | Divorced | Craft-repair | Not-in-family | White | Female | 14344 | 0 | 40 | United-States | >50K | 1 | 2 |
| 106 | 32 | Self-emp-inc | 317660 | HS-grad | 9 | Married-civ-spouse | Craft-repair | Husband | White | Male | 7688 | 0 | 40 | United-States | >50K | 1 | 2 |
| 118 | 49 | Local-gov | 197371 | Assoc-voc | 11 | Married-civ-spouse | Craft-repair | Husband | Black | Male | 0 | 0 | 40 | United-States | >50K | 1 | 2 |
| 128 | 31 | Private | 114937 | Assoc-acdm | 12 | Married-civ-spouse | Adm-clerical | Husband | White | Male | 0 | 0 | 40 | United-States | >50K | 1 | 1 |
| 140 | 49 | Private | 81973 | Some-college | 10 | Married-civ-spouse | Craft-repair | Husband | Asian-Pac- | Male | 0 | 0 | 40 | United-States | >50K | 1 | 2 |

Figure 5: Screenshot of the Contexted data tab in Redhyte

### 3.4 Context mining

The Context mining module first allows users to remove attributes that should not be included in the context mining procedure, e.g. duplicated, redundant, or irrelevant attributes. After context mining is completed, the confusion matrices of the classification models, a list of mined context attributes, and variable importance plots (to be elaborated later) of the models are rendered. Redhyte also allows users to get a quick glance of the class distributions of the mined context attributes, with respect to the initial hypothesis.

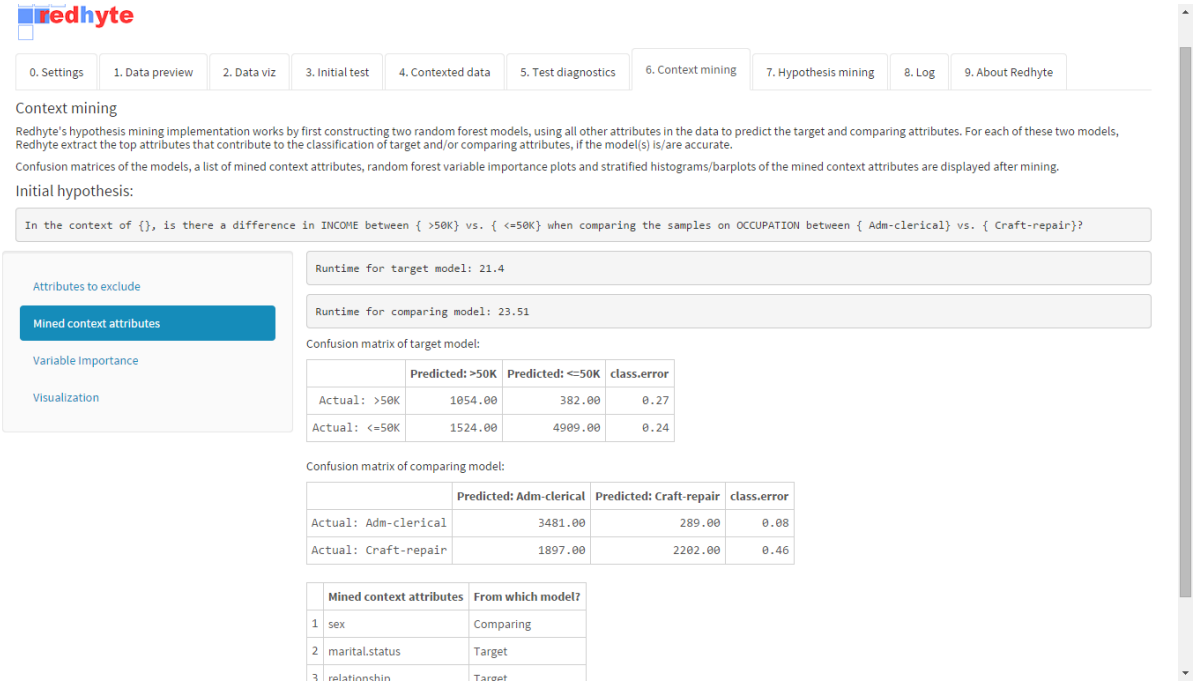



Figure 6: Screenshot of the Context mining module in Redhyte

## 3.5 Hypothesis mining

Using the mined context attributes, Redhyte generates a list of mined hypotheses, suitably scored by the hypothesis mining metrics. In the Hypothesis mining module, users can rank the mined hypotheses according to the hypothesis mining metrics, such as the difference lift and the independence lift. Users can also quickly identify the mined hypotheses in which Simpson's Reversals occurred; this cannot be easily done without the use of statistical programming in data analysis.

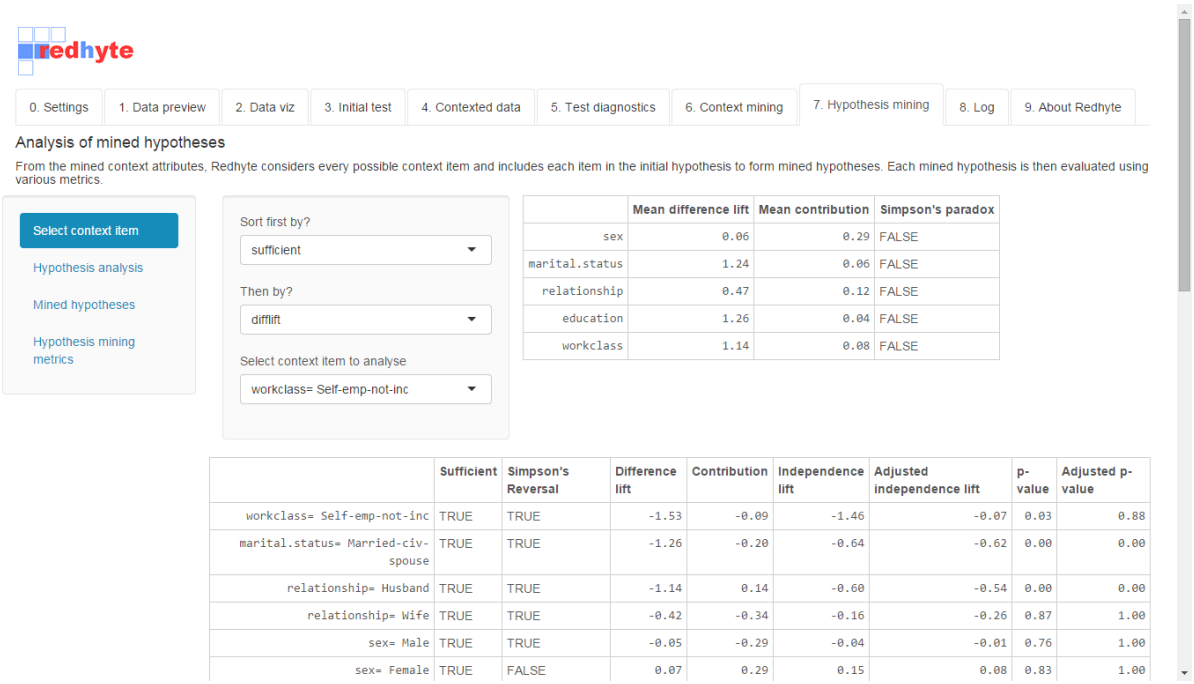



Figure 7: Screenshot of the Hypothesis mining module in Redhyte

Based on the hypothesis mining metrics, users can select mined hypotheses that are deemed interesting for analysis. A comparison between the initial and the selected mined hypotheses (Figure 8) allows the user to quickly identify the rationale behind the (lack of) interestingness of the selected mined hypothesis, be it directed shrinkage (refer to section 4.5.1) or insufficient support. Finally, advanced users may wish to investigate the behaviour of the hypothesis mining metrics, using scatterplots of the various metrics.

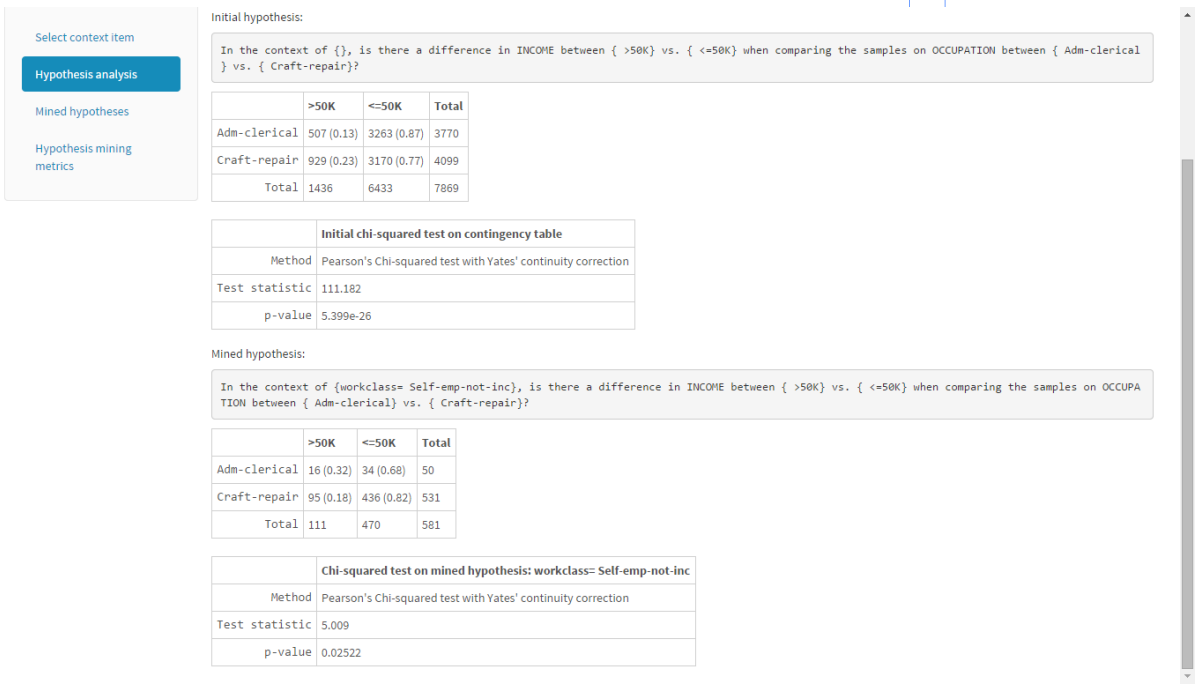
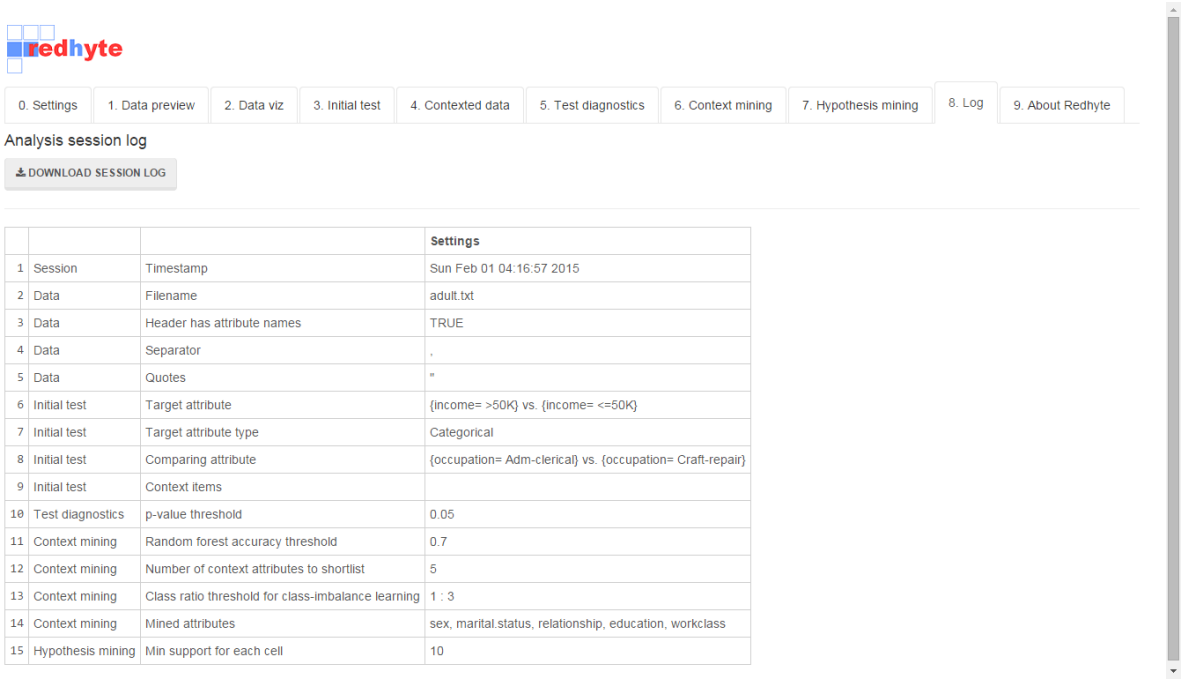



Figure 8: Screenshot of the Hypothesis analysis functionality in the Hypothesis mining module

### 3.6 Log

Finally, the Log documents all settings used in a particular analysis session, and allows users to quickly profile the analysis session. The log can be downloaded as a .csv file, and shared amongst collaborators for reproducibility of hypothesis mining and analysis results.




| | | | Settings |
|---|---|---|---|
| 1 | Session | Timestamp | Sun Feb 01 04:16:57 2015 |
| 2 | Data | Filename | adult.txt |
| 3 | Data | Header has attribute names | TRUE |
| 4 | Data | Separator | , |
| 5 | Data | Quotes | " |
| 6 | Initial test | Target attribute | {income= >50K} vs. {income= <=50K} |
| 7 | Initial test | Target attribute type | Categorical |
| 8 | Initial test | Comparing attribute | {occupation= Adm-clerical} vs. {occupation= Craft-repair} |
| 9 | Initial test | Context items | |
| 10 | Test diagnostics | p-value threshold | 0.05 |
| 11 | Context mining | Random forest accuracy threshold | 0.7 |
| 12 | Context mining | Number of context attributes to shortlist | 5 |
| 13 | Context mining | Class ratio threshold for class-imbalance learning | 1 : 3 |
| 14 | Context mining | Mined attributes | sex, marital.status, relationship, education, workclass |
| 15 | Hypothesis mining | Min support for each cell | 10 |

Figure 9: Screenshot of the Log tab in Redhyte

## 4. EXAMPLES / CASE STUDIES

In this section, we give an illustration on how Redhyte can be used to lead the user to interesting mined hypotheses, using the UC Berkeley admissions and the *adult* dataset. We use the following hypotheses as illustrations:

Table 10: Hypotheses from the UC Berkeley admission and the *adult* dataset

| | |
|---|---|
| Hypothesis A | In the context of {}, is there a difference in ADMIT between {Admitted} vs. {Rejected} when comparing the samples on GENDER between {Male} vs. {Female}? |
| Hypothesis B | In the context of {race= White}, is there a difference in INCOME between { >50K} vs. { <=50K} when comparing the samples on OCCUPATION between { Adm-clerical} vs. { Craft-repair}? |

### 4.1 UC Berkeley admissions

Based on hypothesis A, the initial test suggests the relationship between admission numbers and gender is significant ($p < 0.05$), with the males being more likely to be admitted into the university than females:

Table 11: Contingency table of Hypothesis I

| | **Admitted** | **Rejected** | **Total** |
|---|---|---|---|
| **Males** | 1198 (44.5%) | 1493 (55.5%) | 2691 |
| **Females** | 557 (30.4%) | 1278 (69.6%) | 1835 |
| **Total** | 1755 | 2771 | 4526 |

However, stratifying by various departments (departments A to F) gives different conclusions. In particular, inserting the context item {Dept = A} gives the following contingency table, with a p-value less than 0.05:

Table 12: Contingency table of mined hypothesis with {Dept = A}

| **{Dept = A}** | **Admitted** | **Rejected** | **Total** |
|---|---|---|---|
| **Males** | 512 (62.1%) | 313 (37.9%) | 825 |
| **Females** | 89 (82.4%) | 19 (17.6%) | 108 |
| **Total** | 601 | 332 | 933 |

Clearly, a Simpson's Reversal has taken place. When considering the admission rate of a particular department A, females are favoured for admission than males. This is contrary to the conclusion given by the initial test. Setting up the above hypothesis in Redhyte will allow the user to easily arrive at Table 12 in a matter of seconds.

### 4.2 Adult

Based on hypothesis B in Table 10, the initial test suggests that the relationship between income and occupation is significant ($p < 0.05$), with white administrative clerks earning more than white craft repairers, as shown in Table 13.

Table 13: Contingency table of Hypothesis III

| | Income > 50K | Income <= 50K | Total |
|---|---|---|---|
| **Administrative clerks** | 439 (14.2%) | 2645 (85.8%) | 3084 |
| **Craft repairers** | 844 (22.8%) | 2850 (77.2%) | 3694 |
| **Total** | 1283 | 5495 | 6778 |

Using the default settings, Redhyte identifies five mined context attributes after context mining, namely sex, relationship, workclass, education, and education.num. In particular, considering the context items {Sex = Male}, {Sex = Female} and {Workclass = Self-emp-not-inc} leaves us with the following contingency tables:

Table 14: Contingency table of mined hypothesis with {Sex = Male}

| {Sex = Male} | Income > 50K | Income <= 50K | Total |
|---|---|---|---|
| **Administrative clerks** | 251 (24.2%) | 787 (75.8%) | 1038 |
| **Craft repairers** | 829 (23.5%) | 2695 (76.5%) | 3524 |
| **Total** | 1080 | 3482 | 4562 |

Table 15: Contingency table of mined hypothesis with {Sex = Female}

| {Sex = Female} | Income > 50K | Income <= 50K | Total |
|---|---|---|---|
| **Administrative clerks** | 188 (9.2%) | 1858 (90.8%) | 2046 |
| **Craft repairers** | 15 (8.8%) | 155 (91.2%) | 170 |
| **Total** | 203 | 2013 | 2216 |

The above illustrates an exact instance of a Simpson's Paradox, with both genders resulting in reversals of trends. This is also an example hypothesis mined by Liu *et al*'s hypothesis mining system, used as a case study in Liu *et al*.

Table 16: Contingency table of mined hypothesis with {Workclass = Self-emp-not-inc}

| {Workclass = Self-emp-not-inc} | Income > 50K | Income <= 50K | Total |
|---|---|---|---|
| **Administrative clerks** | 16 (34.8%) | 30 (65.2%) | 46 |
| **Craft repairers** | 90 (18%) | 409 (82%) | 499 |
| **Total** | 106 | 439 | 545 |

The hypothesis mining metrics (refer to https://sites.google.com/site/tohweizhongcv/redhyte, section 4.5.1) evaluated on these three items are as follows:

Table 17: Hypothesis mining metrics evaluated for the selected context items

| Context items | Difference lift | Contribution | Independence lift | Adjusted independence lift | p-value |
|---|---|---|---|---|---|
| **{Sex = Male}** | -0.08 | -0.31 | -0.06 | -0.02 | 0.69 |
| **{Sex = Female}** | -0.04 | 0.31 | -0.09 | -0.05 | 0.98 |
| **{Workclass = Self-emp-not-inc}** | -1.94 | -0.11 | -1.89 | -0.05 | 0.01 |

Based on Hypothesis B, the default settings in Redhyte is used to illustrate the above, and to generate 27 other mined hypotheses, suitably scored and ranked using the hypothesis mining metrics, for the user to inspect. The user is also able to alter the settings in Redhyte – for instance, increasing the number of context attributes to mine for, to suit analysis purposes. All of these analysis results are entirely reproducible, with the help of the session log documentation.