



Automated Data Collection

PRICE Systems, LLC



Agenda

- Introduction and Motivation
- Data Collection Challenges
- Introduction to RapidMiner Capabilities
- RapidMiner - Crawling the Web
- Implementation
- Future Directions

- **Data Collection is a necessary evil for cost estimators**
 - To support the creation of Cost Estimating Relationships (CERs)
 - To support estimating by analogy
 - To support selection of input values for cost estimating models

- **Data collection is hard**
 - Data is often hard to find
 - Even when data is found it is often hard to mine as the process is tedious and time consuming
 - Data is often very noisy making it hard to understand and work with

- Original motivation for this work arose from efforts around an overhaul of the TruePlanning[®] Information Technology Services Cost Model
- Many of the models we developed for IT Services required commodity pricing information
- Commodity pricing is hard to estimate because
 - Prices are constantly changing
 - Many companies have negotiated agreements with specific vendors
 - There are many things that drive commodity pricing outside of the scope of a typical cost estimating relationship

- **Finding the right data**
 - Accurate pricing data
 - Significant technical and specification information
 - Normalization across multiple vendors

- **Keeping the data up to date**
 - Commodity prices change frequently – based on market factors, supply and demand, etc.
 - Good pricing data from last quarter is unlikely to be relevant in this quarter

- **Need a solution that is**
 - Repeatable
 - Consistent
 - Can be accomplished quickly with the push of a button
 - Can be updated regularly (monthly, quarterly, bi-annually) without extensive time investment

- Open Source application
- RapidMiner 5.3 and 7.2 currently available for download from <http://rapidminer.software.informer.com/5.3/>
<https://my.rapidminer.com/nexus/index/html#downloads>
- Licensed under GMU Affero General Public License version 3
- User friendly, graphical user interface that allows for data collection and analysis

Summary of RapidMiner Capability



■ Data Transformation

- Filtering
- Sorting
- Replacing Missing values in a data set
- Aggregation, etc.

■ Modeling

- Classification
- Clustering
- Correlation, etc.

■ Evaluation

- Validation
- Regression
- Significance, etc.

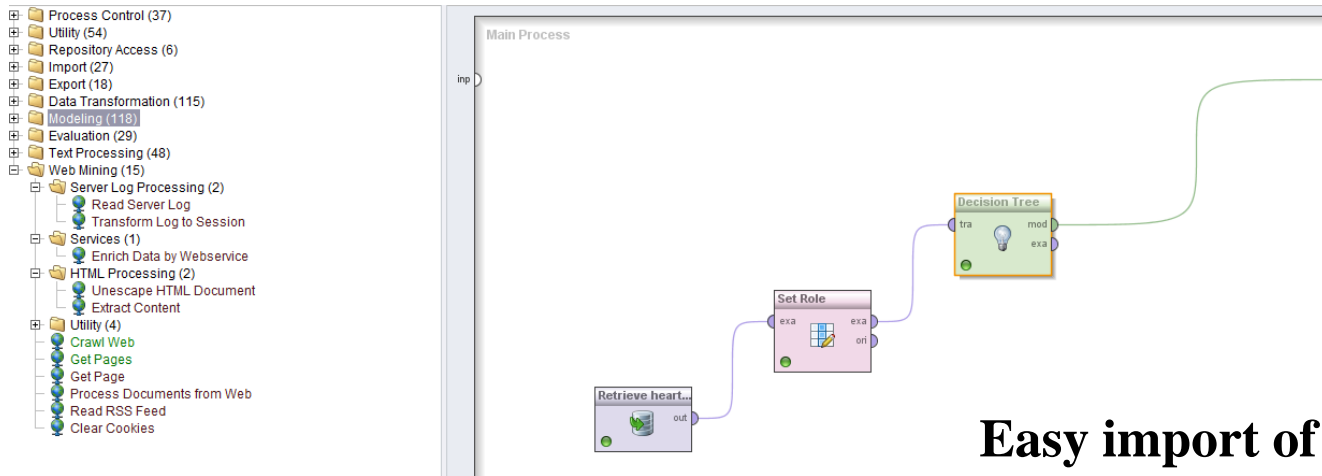
■ Text Processing

- Extraction
- Transformation
- Extraction
- Tokenization, etc.

■ Web Mining

- Server Log Processing
- HTML Processing
- Web Crawling
- Read RSS Feed, etc.

Drag and Drop Interface to Build Processes



Easy import of data from Excel

Data View Meta Data View Plot View Advanced Charts Annotations

ExampleSet (138 examples, 0 special attributes, 8 regular attributes)

Row No.	Age	Marital_Stat...	Gender	Weight_Cat...	Cholesterol	Stress_Man...	Trait_Anxiety	2nd_Heart...
1	60	2	0	1	150	1	50	Yes
2	69	2	1	1	170	0	60	Yes
3	52	1	0	0	174	1	35	No
4	66	2	1	1	169	0	60	Yes
5	70	3	0	1	237	0	65	Yes
6	52	1	0	0	174	1	35	No
7	58	2	1	0	140	0	45	No
8	59	2	1	0	143	0	45	Yes
9	60	2	0	0	139	0	45	No
10	51	1	1	0	174	1	40	No
11	52	1	0	0	189	1	65	No
12	70	2	1	1	147	1	50	Yes
13	52	2	1	2	160	0	40	Yes
14	74	3	1	2	178	0	75	Yes
15	64	2	1	2	236	1	80	Yes
16	69	2	0	1	146	1	50	Yes
17	58	2	0	0	141	0	45	No
18	68	1	0	0	172	0	60	No

Countless ways to visualize your data

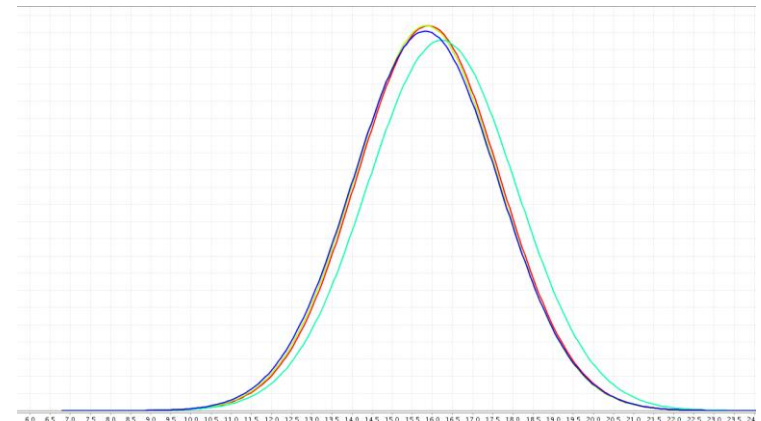
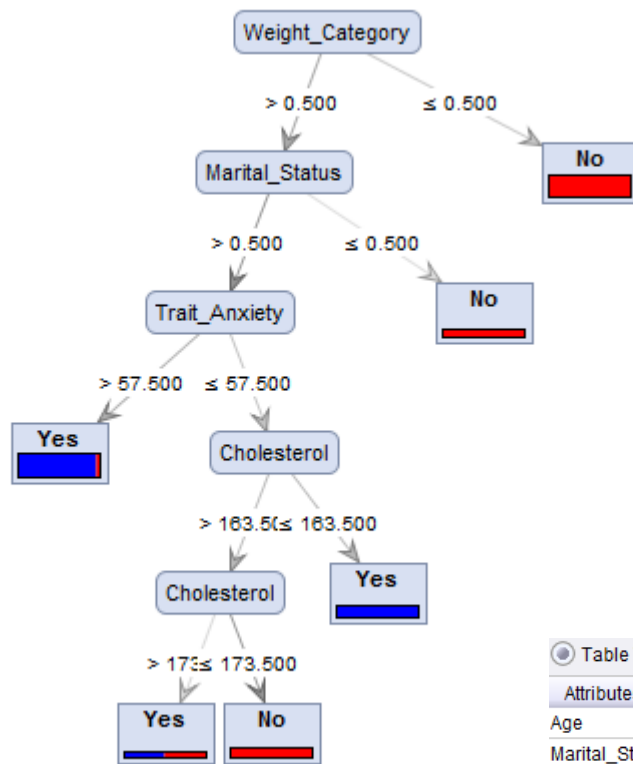


Table View
 Pairwise Table
 Plot View
 Annotations

Attributes	Age	Marital_Stat...	Gender	Weight_Cat...	Cholesterol	Stress_Man...	Trait_Anxiety	2nd_Heart_...
Age	1	0.427	0.076	0.402	0.395	-0.359	0.667	-0.499
Marital_Stat	0.427	1	-0.033	0.065	0.274	-0.292	0.238	-0.365
Gender	0.076	-0.033	1	0.449	0.191	-0.241	0.082	-0.318
Weight_Cate	0.402	0.065	0.449	1	0.398	-0.347	0.501	-0.731
Cholesterol	0.395	0.274	0.191	0.398	1	-0.406	0.579	-0.507
Stress_Man	-0.359	-0.292	-0.241	-0.347	-0.406	1	-0.321	0.439
Trait_Anxiety	0.667	0.238	0.082	0.501	0.579	-0.321	1	-0.483
2nd_Heart_	-0.499	-0.365	-0.318	-0.731	-0.507	0.439	-0.483	1

Multiple Regression Models, Weighting and Analysis

Attribute	Coefficient	Std. Error	Std. Coeffi...	Tolerance	t-Stat	p-Value	Code
ittribute_1	2.053	1.677	1.618	0.904	1.224	0.307	
ittribute_2	2.007	1.531	1.721	0.930	1.310	0.257	
ittribute_3	-5.631	1.425	-4.936	0.886	-3.951	0.000	****
ittribute_4	3.253	0.874	2.808	0.953	3.720	0.000	****
ittribute_8	-1.191	0.532	-0.753	0.910	-2.241	0.030	**
ittribute_9	0.943	0.637	0.627	0.850	1.479	0.181	
ittribute_10	-0.381	0.586	-0.246	0.802	-0.649	0.523	
ittribute_11	0.247	0.665	0.139	0.701	0.371	0.714	
ittribute_12	1.027	0.538	0.575	0.777	1.908	0.069	*
ittribute_13	0.422	0.536	0.218	0.851	0.788	0.439	
ittribute_14	-0.361	0.485	-0.200	0.955	-0.745	0.464	
ittribute_15	0.257	0.494	0.165	0.996	0.521	0.608	
ittribute_16	-0.256	0.480	-0.157	1.000	-0.532	0.600	
ittribute_17	-0.696	0.502	-0.441	1.000	-1.387	0.219	
ittribute_18	1.007	0.469	0.582	1.000	2.149	0.038	**
ittribute_19	-0.732	0.404	-0.374	0.961	-1.811	0.087	*
ittribute_20	0.615	0.473	0.287	0.925	1.300	0.263	
ittribute_21	-0.623	0.464	-0.264	0.896	-1.341	0.242	
ittribute_22	0.647	0.468	0.265	0.941	1.384	0.221	
ittribute_23	-0.506	0.479	-0.196	0.970	-1.057	0.428	
ittribute_24	1.156	0.490	0.411	0.995	2.358	0.022	**

attribute	
duration	0
pension	0.062
education-allowance	0.064
shift-differential	0.185
bereavement-assistance	0.185
working-hours	0.227
col-adj	0.258
vacation	0.271
wage-inc-3rd	0.298
standby-pay	0.412
contrib-to-dental-plan	0.424
longterm-disability-assistance	0.535
wage-inc-2nd	0.547
contrib-to-health-plan	0.676
statutory-holidays	0.685
wage-inc-1st	1

PolynomialRegression

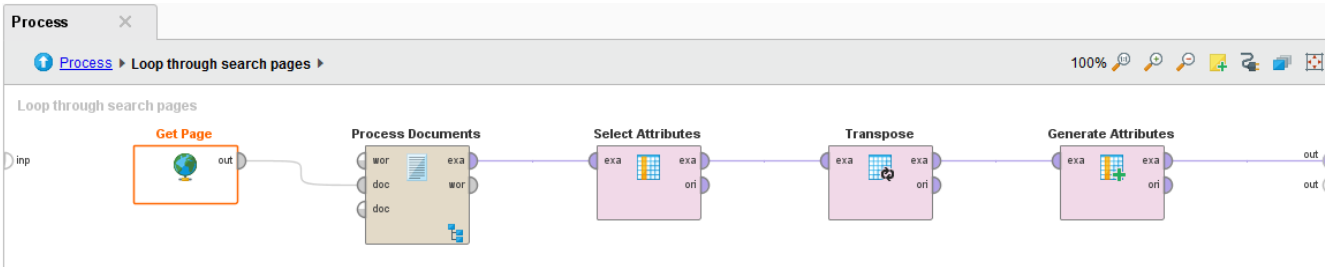
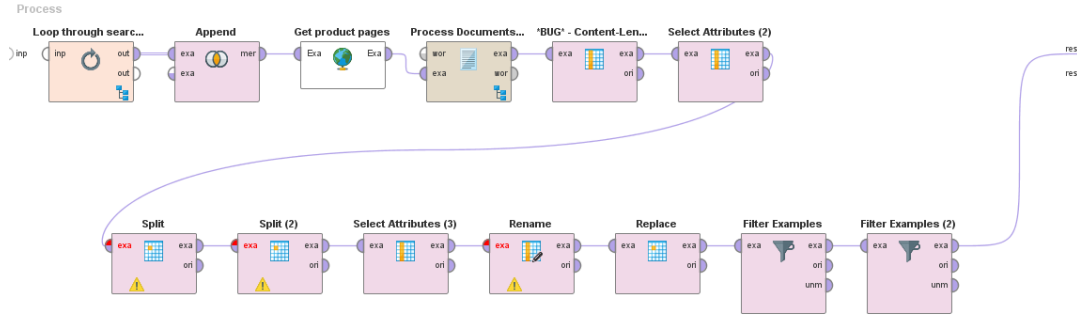
```

25.320 * a1 ^ 1.000
+ 3.697 * a2 ^ 2.000
+ 1.814 * a3 ^ 2.000
- 2.192 * a4 ^ 1.000
- 0.673 * a5 ^ 2.000
- 83.828
    
```

Starting with a webpage that is setup where you can view list of products:

- **Get Page**
 - Sends RapidMiner out to product page URL
 - Example: http://www.tigerdirect.com/applications/category/category_slc.asp?page=1&Nav=|c:4935|&Sort=3&Recs=10
- **Process Documents**
 - Uses Xpaths to grab product links
- **Get Pages**
 - Sends RapidMiner out to each product link that it grabbed in Process Documents
- **Process Documents from Data**
 - Uses Xpaths to grab product specifications on each product page

Implementation to Support Commodity Pricing Data Collection



Parameters window for the 'Get Page' process:

- url:
- random user agent
- user agent:
- connection timeout:

Process Documents window with an 'Extract Information' dialog box open. The dialog lists attribute names and their corresponding XPath queries:

attribute name	query expression
product1	//h.h3[class="itemName"]h.a[href][1]
product2	//h.h3[class="itemName"]h.a[href][2]
product3	//h.h3[class="itemName"]h.a[href][3]
product4	//h.h3[class="itemName"]h.a[href][4]
product5	//h.h3[class="itemName"]h.a[href][5]
product6	//h.h3[class="itemName"]h.a[href][6]
product7	//h.h3[class="itemName"]h.a[href][7]
product8	//h.h3[class="itemName"]h.a[href][8]
product9	//h.h3[class="itemName"]h.a[href][9]
product10	//h.h3[class="itemName"]h.a[href][10]
product11	//h.h3[class="itemName"]h.a[href][11]

Combining into one process

The screenshot displays a workflow automation interface with a process diagram and a parameter list dialog.

Process Diagram:

- Category URLs:** An orange box with an 'in' port and an 'out' port.
- Make selection:** A purple box with 'exa' and 'ori' ports.
- Subprocess:** An orange box with 'in' and 'out' ports.
- Handle Exception (2):** An orange box with 'in' and 'out' ports, containing a warning icon.
- Process Documents...:** A grey box with 'wor' and 'exa' ports.
- Select Subprocess (2):** An orange box with 'in' and 'out' ports.
- Replace (11):** A purple box with 'exa' and 'ori' ports, containing a warning icon.
- Remove unwanted a...:** A purple box with 'exa' and 'ori' ports, containing a warning icon.
- Write Excel:** A purple box with 'inp' and 'thr' ports, containing a warning icon.

Parameters Panel:

- Category URLs (Generate Data by User Specification):** Includes 'attribute values' and 'set additional roles' sections, each with an 'Edit List' button.

Edit Parameter List: attribute values Dialog:

This dialog defines the attributes and their values in the single example returned.

attribute n...	attribute value
Tablets	"http://www.tigerdirect.com/applications/category/category_slc.asp?page=1&Na...
Laptops	"http://www.tigerdirect.com/applications/category/category_slc.asp?page=1&Na...
Workstations	"http://www.tigerdirect.com/applications/category/guidedSearch.asp?CatId=6&e...
Servers	"http://www.tigerdirect.com/applications/category/category_slc.asp?page=1&Na...
Printers	"http://www.tigerdirect.com/applications/category/guidedSearch.asp?CatId=21&e...
Storage	"http://www.tigerdirect.com/applications/category/category_slc.asp?page=1&Na...
Routers	"http://www.tigerdirect.com/applications/category/category_slc.asp?page=1&Na...
Switches	"http://www.tigerdirect.com/applications/category/category_slc.asp?page=1&Na...

Buttons: Add Entry, Remove Entry, Apply, Cancel.

Process to find relevant pages when webpage has various product links to follow:

- **Crawl Web**

- Start at the highest level URL for each product

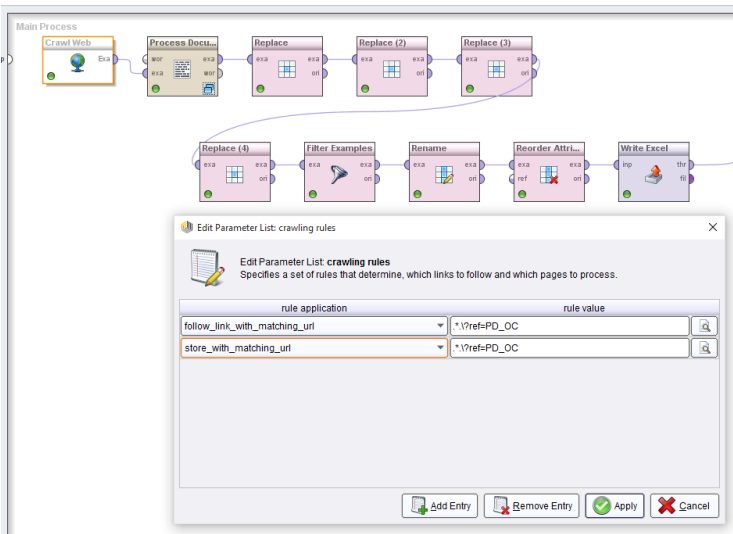
- Example: <http://www.dell.com/us/business/p/laptops>

- Crawling rules will send RapidMiner out to any URL matching your search

- **Process Documents from Data**

- Uses Xpaths to grab product specifications on each product page

Implementation to Support Commodity Pricing Data Collection



Crawl Web

url:

crawling rules:

write pages into files

add pages as attribute

max pages:

max depth:

domain:

delay:

max threads:

max page size:

user agent:

Edit Parameter List: crawling rules

Edit Parameter List: crawling rules
Specifies a set of rules that determine which links to follow and which pages to process.

rule application	rule value
follow_link_with_matching_url	*\?ref=PD_OC
store_with_matching_url	*\?ref=PD_OC

Extract Inform...

Edit Parameter List: xpath queries

Edit Parameter List: xpath queries
Specifies a list of attribute names and their corresponding XPath queries. See the operator documentation for details on XPath.

attribute name	query expression
Name	normalize-space(//*[@id='mastheadPageTitle']/text())
Unit Purchase Price	//h:span[contains(., 'Starting at') and @class='startText']/h:sp
Processor	string(//h:h5[contains(., 'Processor')]/following-sibling:h:div/h
Operating_System	string(//h:h5[contains(., 'Operating System')]/following-sibling
Memory	string(//h:h5[contains(., 'Memory')]/following-sibling:h:div/sp
Hard_Drive	string(//h:h5[contains(., 'Hard Drive')]/following-sibling:h:div/h
Type of Device	string('laptop')
Memory_GB	string(//h:h5[contains(., 'Memory')]/following-sibling:h:div/sp
Hard_Drive_GB	string(//h:h5[contains(., 'Hard Drive')]/following-sibling:h:div/h
Level	string('2')
Definition Name	string('End User Services New Projects')

Extract Information

query type:

attribute type:

xpath queries:

namespaces:

ignore CDATA

assume html

Extract Information

Row No.	URL	Name	Definition Name	Hard_Drive	Hard_Drive...	Level	Memory	Memory_GB	Operating_System	Processor	Type of Device	Unit Purchase Price
1	http://www.d...	Latitude 13 (3	End User Services	M.2 128GB S	128	2	4GB (1x4G) :	4	Windows 10 Pro 64t	Pentium DC	laptop	699
2	http://www.d...	Latitude 12 (7	End User Services	M.2 128GB S	128	2	4GB (1x4GB) 4		Windows 10 Pro, 64-	6th Generat	laptop	1049
3	http://www.d...	Latitude 13 (7	End User Services	M.2 128GB S	128	2	4GB LPDDR 4		Windows 7 Professi	Intel® Core [™]	laptop	1299
4	http://www.d...	New Inspiron	End User Services	500GB 5400	500	2	4GB Single C	4	Windows 10 Home,	Intel® Pentii	laptop	499
5	http://www.d...	New Inspiron	End User Services	500GB 5400	500	2	4GB Single C	4	Windows 10 Home,	Intel® Pentii	laptop	449
6	http://www.d...	New Inspiron	End User Services	256GB Solid	256	2	8GB Dual C	8	Windows 10 Home €	6th Generat	laptop	749
7	http://www.d...	New Inspiron	End User Services	256GB Solid	256	2	8GB Dual C	8	Windows 10 Home €	6th Generat	laptop	749
8	http://www.d...	Precision 15 :	End User Services	500GB 2.5 ir	500	2	8GB (2x4GB) 8		Windows 7 Professi	Intel® Core [™]	laptop	999
9	http://www.d...	Precision 15 :	End User Services	500GB 2.5 ir	500	2	8GB (2x4GB) 8		Windows 7 Professi	Intel® Core [™]	laptop	1399
10	http://www.d...	XPS 15 Lapto	End User Services	500GB 7200	200	2	8GB (1x8G) :	8	Windows 10 Home,	6th Generat	laptop	999

Implementation to Support Commodity Pricing Data Collection



- Processes have been created to crawl Dell, HP and TigerDirect for pricing and performance data for:
 - Laptops
 - Workstations
 - Tablets
 - Printers
 - Storage Devices
 - Servers
 - Other Supporting Hardware
- These processes create Excel files that are directly importable into the IT Hardware TrueFindings® database
- This database can be updated in several hours to support monthly or quarterly updates of the database

	Value
1 Start Date	
2 Device Information	
3 Type of Device	Laptop
4 Service Level	3.00
5 Number of Deployments	Custom - Yearly
6 Quantity Per Next Higher Level	1.00
7 Purchase or Lease	Purchase
8 Service Options	In-House
9 Project Details	
10 Organizational Productivity	1.000
11 Purchase Inputs	
12 Unit Purchase Price	710.03
13 Inventory	
14 Unit Lifetime	
15 Supporting Details	
16 Software Price per Unit	
17 Annual Training per End User	
18 Annual Support per End User	
19 Training per Unit Delivered by Help Desk Analyst	

Name	Value	Method	Type
Laptops - Unit Purchase Price	710.031578947...	Distribution	Mean
Printers - Unit Purchase Price	1552.25946327...	Distribution	Mean
Routers - Unit Purchase Price	6406.29338842...	Distribution	Mean
Servers - Unit Purchase Price	1573.89285714...	Distribution	Mean
Storage Devices - Unit Purchase Price	1369.10394265...	Distribution	Mean
Switches - Unit Purchase Price	389.513089005...	Distribution	Mean
Tablets - Unit Purchase Price	932.086206896...	Distribution	Mean
Workstations - Unit Purchase Price	783.866310160...	Distribution	Mean

Selected Column

Search TrueFindings Database

Your Search

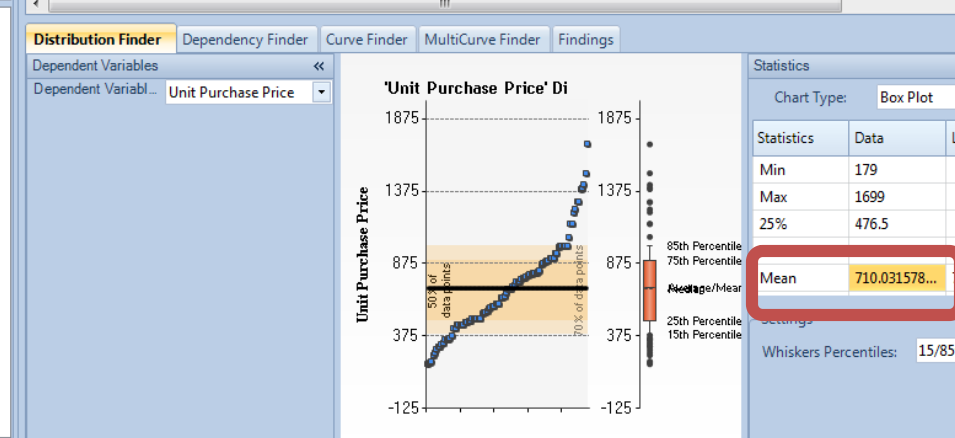
- Type of Device
- laptop

Keyword

- Characteristics
 - Definition Name
 - Hard_Drive
 - Memory
 - Operating_System
 - Processor
 - URL
- Performance
 - Number of Devices (95)
 - Unit Purchase Price (95)
 - External Integration Complexity (95)
 - Total Weight (95)
 - Hard_Drive_GB (95)
 - Memory_GB (95)

Data (95 rows)

	Name	Definition Name	Type of Device	Number of De...	Unit Purcha...	External Integr...	Total Weight	Ha
1	HP Chromebo...	End User Servi...	laptop	1	179	1	0	16
2	HP Chromebo...	End User Servi...	laptop	1	189	1	0	16
3	HP Chromebo...	End User Servi...	laptop	1	199	1	0	16
4	ASUS 11.6" Eee...	End User Servi...	laptop	1	241	1	0	32
5	Lenovo ThinkP...	End User Servi...	laptop	1	259	1	0	12
6	HP Chromebo...	End User Servi...	laptop	1	279	1	0	16
7	Acer Aspire E1 ...	End User Servi...	laptop	1	302	1	0	50
8	Acer Aspire E5...	End User Servi...	laptop	1	302	1	0	50



Implementation to Support Commodity Pricing Data Collection

- These data points can also be accessed via the File New Template Search in TruePlanning 16.0 for immediate drag and drop into a project file.

The screenshot illustrates the process of adding commodity pricing data to a project file in TruePlanning 16.0. It shows the 'New' menu on the left, a search for 'IT Hardware' in the top right, a file explorer view of the 'IT Hardware' folder containing various device models, and a project details table on the right.

	Value
1 Start Date	
2 Device Information	
3 Type of Device	Tablet
4 Number of Devices	1.00
5 Service Options	In-House
6 Project Details	
7 Organizational Productivity	1.000
8 Purchase Inputs	
9 Unit Purchase Price	399.00
10 Supporting Details	
11 Training per Device	0.00
12 Training per Unit Delivered by Help Desk Analyst	0.00
13 Setup and Installation Time per Device	1.00
14 Work by Systems Administrator	50.00%
15 IT Manager Time per Unit	10.00%
16 Integration Information	
17 Number of Operational Hours	0.00
18 Recovery Time Objective	4.00
19 External Integration Complexity	1.00
20 Total Weight	0.000

- PRICE intends to update and extend the IT Hardware Commodity database
 - Regular updates with the existing processes that have been developed
 - Developing new processes based on user requirements & suggestions
- We are currently investigating the feasibility of creating similar processes to support the IT Software pricing requirements
 - This may be problematic because many software applications require calls to the vendor for quotes – we're hopeful we may be able to find sources
- We are considering extending this project to include commodity prices for electronic components to support Microcircuit cost estimation
- We would like to apply this expertise to develop custom solutions for clients based on their specific purchasing processes