UNIVERSIDADE DE SÃO PAULO

FACULDADE DE ECONOMIA, ADMINISTRAÇÃO E CONTABILIDADE DE RIBEIRÃO PRETO

DEPARTAMENTO DE ADMINISTRAÇÃO

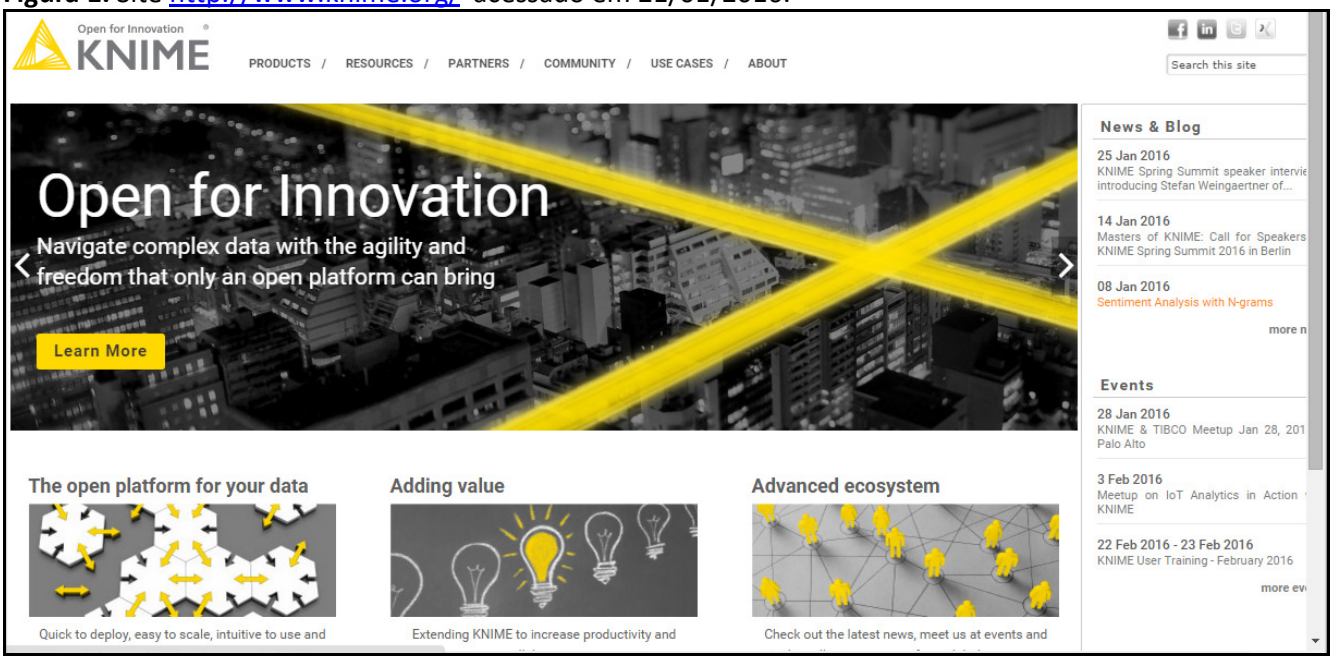
Estatística Aplicada à Administração com o software KNIME

Prof. Dr. Evandro Marcos Saidel Ribeiro

1 O software KNIME

Nesta apostila o conteúdo de Estatística Aplicada à Administração é visto com a aplicação do software KNIME (pronuncia-se "naime"). O KNIME é um software livre que proporciona acesso fácil e intuitivo para técnicas avançadas de ciência dos dados. Veja informações no site http://www.knime.org/ (Figura 1).

Figura 1. Site http://www.knime.org/ acessado em 21/01/2016.



1.1 Leitura de dados no KNIME

Os dados podem ser disponibilizados em diversos formatos. Nesta apostila são considerados arquivos gravados em planilha Excel no formato CSV (comma-separated values), que para o Excel em português consiste em arquivo com variáveis em colunas separadas por ponto e vírgula, neste caso a vírgula serve como separador de casas decimais em números reais. Para esta apostila considerei arquivo com colunas

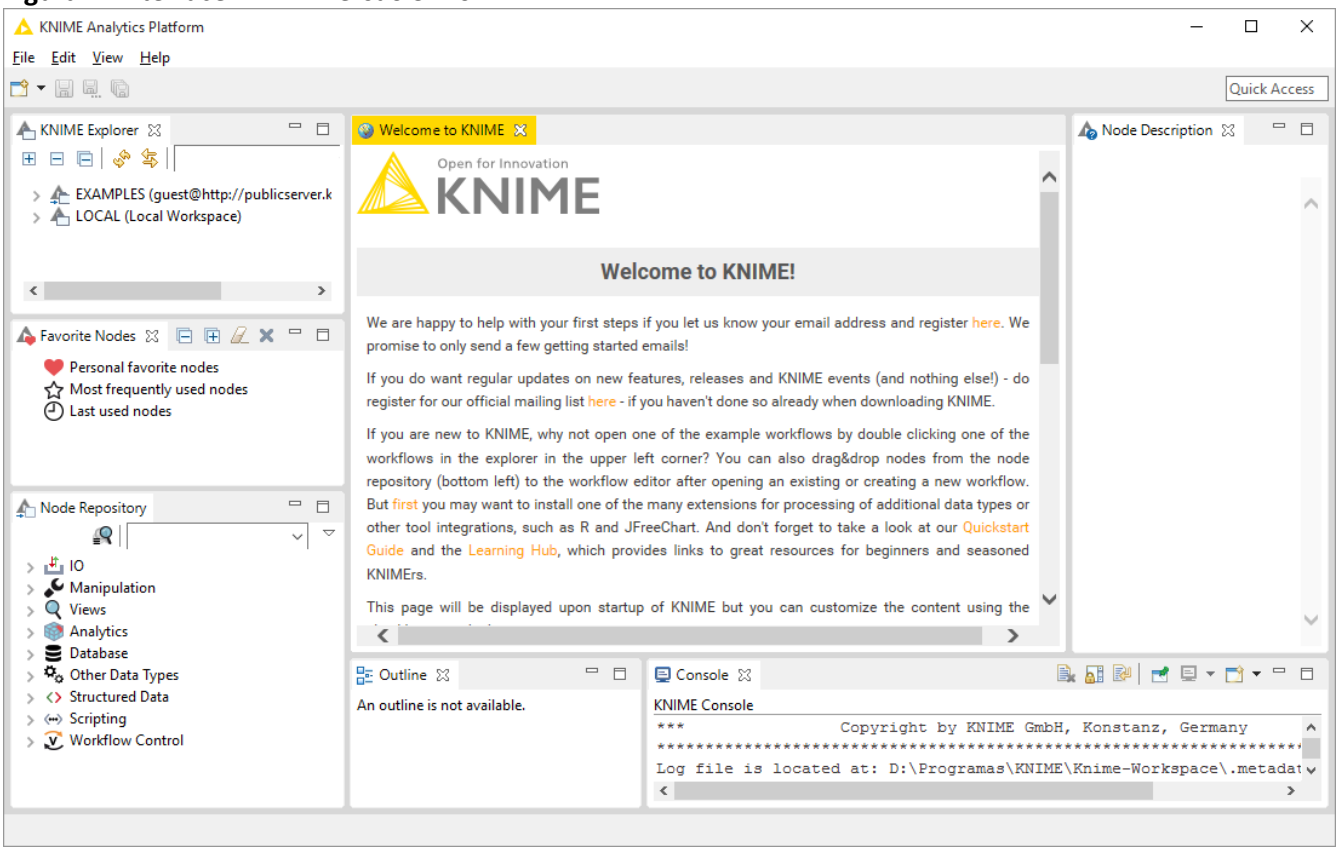
separadas por virgulas e as casas decimais são pontos. Este arquivo CSV está disponibilizado no site STOA da disciplina:

Cap01_Corrar_etal_2007.csv

• **Exercício 1**: Abrir o software KNIME e ler os dados contidos no arquivo Cap01_Corrar_etal_2007.csv.

Clicando no ícone o KNIME é inicializado, apresentando a interface (versão 3.1.0) vista na Figura 2.

Figura 2. Interface KNIME versão 3.1.0.



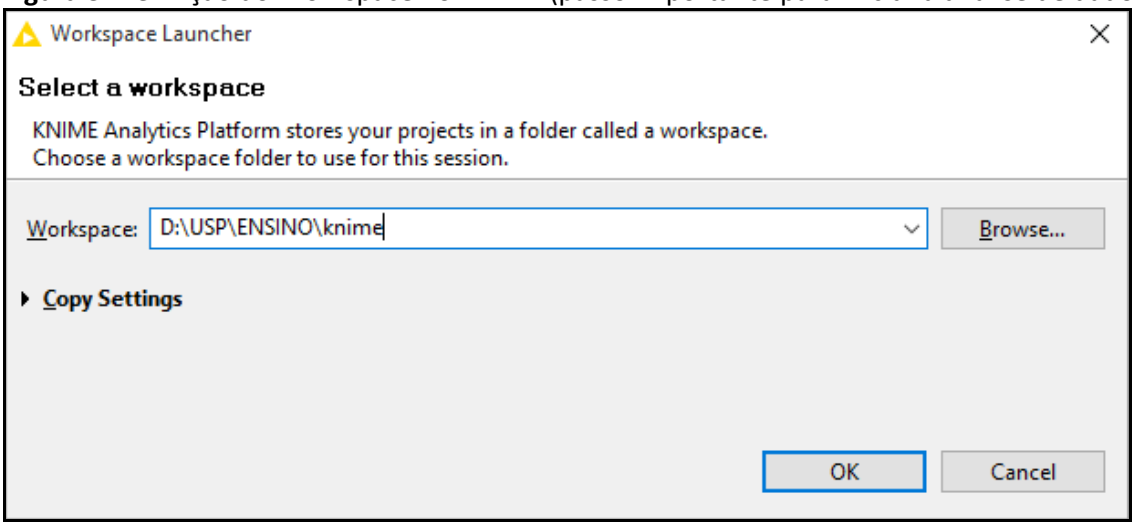
Nesta apostila serão apresentados procedimentos necessários para realização de análises estatísticas. Para conhecer melhor o software é recomendável seguir os passos disponibilizados no “KNIME Quickstart Guide”.

É necessário verificar o local de trabalho (Workspace) e se for o caso redefinir o Workspace para um local mais adequado. Quando o software é iniciado, o Workspace será aquele definido na instalação, mas o Workspace pode ser alterado por:

File > Switch Workspace > Other

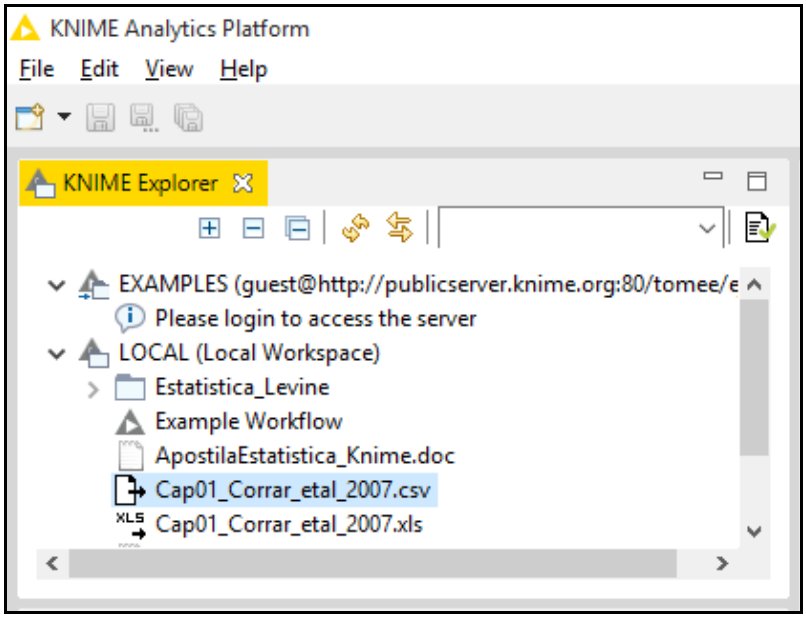
No caso desta apostila o Workspace foi definido para: D:\USP\ENSINO\knime (veja Figura 3).

Figura 3. Definição do Workspace no KNIME (passo importante para iniciar a análise de dados).



Observe, no NIME Explorer (que fica no canto superior esquerdo da interface do KNIME), que o arquivo a ser analisado (Cap01_Corrar_etal_2007.csv) está neste Workspace (veja Figura 4 o destaque em azul-claro).

Figura 4. Detalhe do KNIME Explorer com alguns arquivos presentes no "Local Workspace". Note o arquivo Cap01_Corrar_etal_2007.csv marcado em azul-claro.



As análises são feitas em termos de "nós" conectados formando um fluxo de análise (workflow). Assim, uma análise estatística consiste num workflow (esquema com nós conectados). Existem vários tipos de nós, que podem ser vistos no repositório de nós, como apresentado na Figura 5. Pode ser visto na Figura 5 que existem nós do tipo Entrada e Saída (I/0), Manipulação, Visualização, Análises, Base de Dados, Outro Tipo de Dados, ...

Para iniciar a análise deve-se criar um novo workflow: File > New > KNIME >. Veja a Figura 6.

Figura 5. Detalhe do Repositório de nós da interface do KNIME.

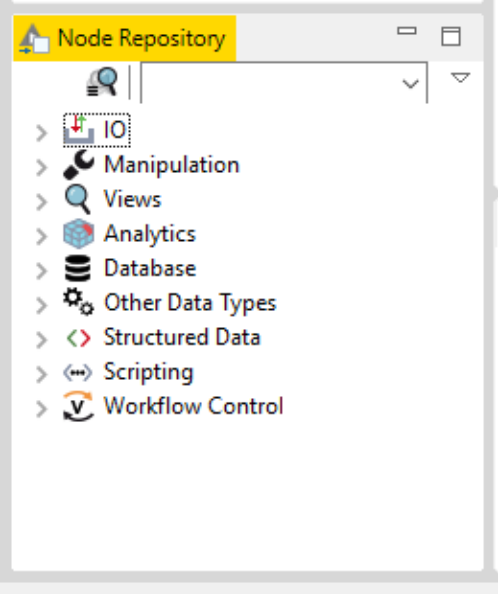
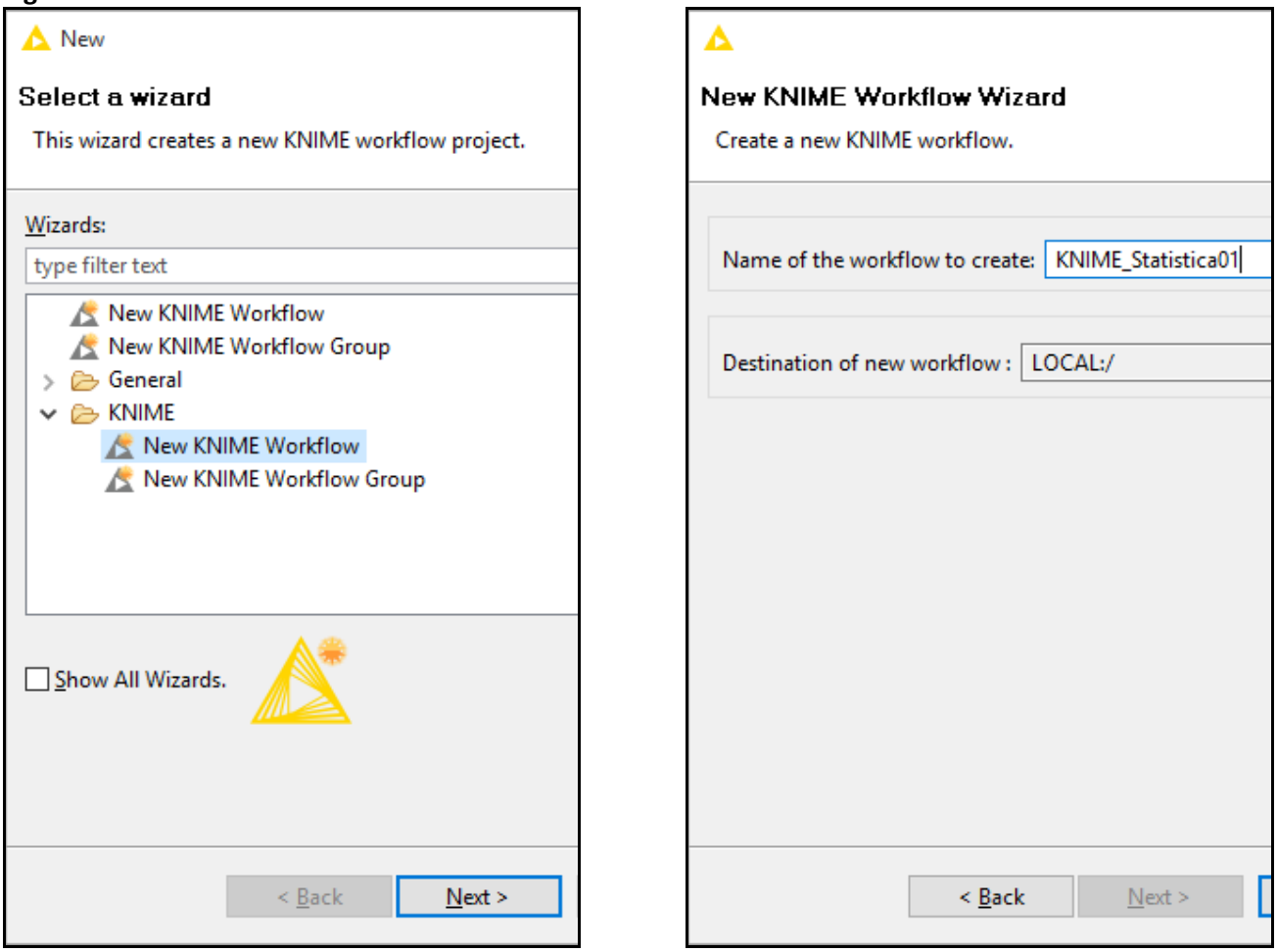
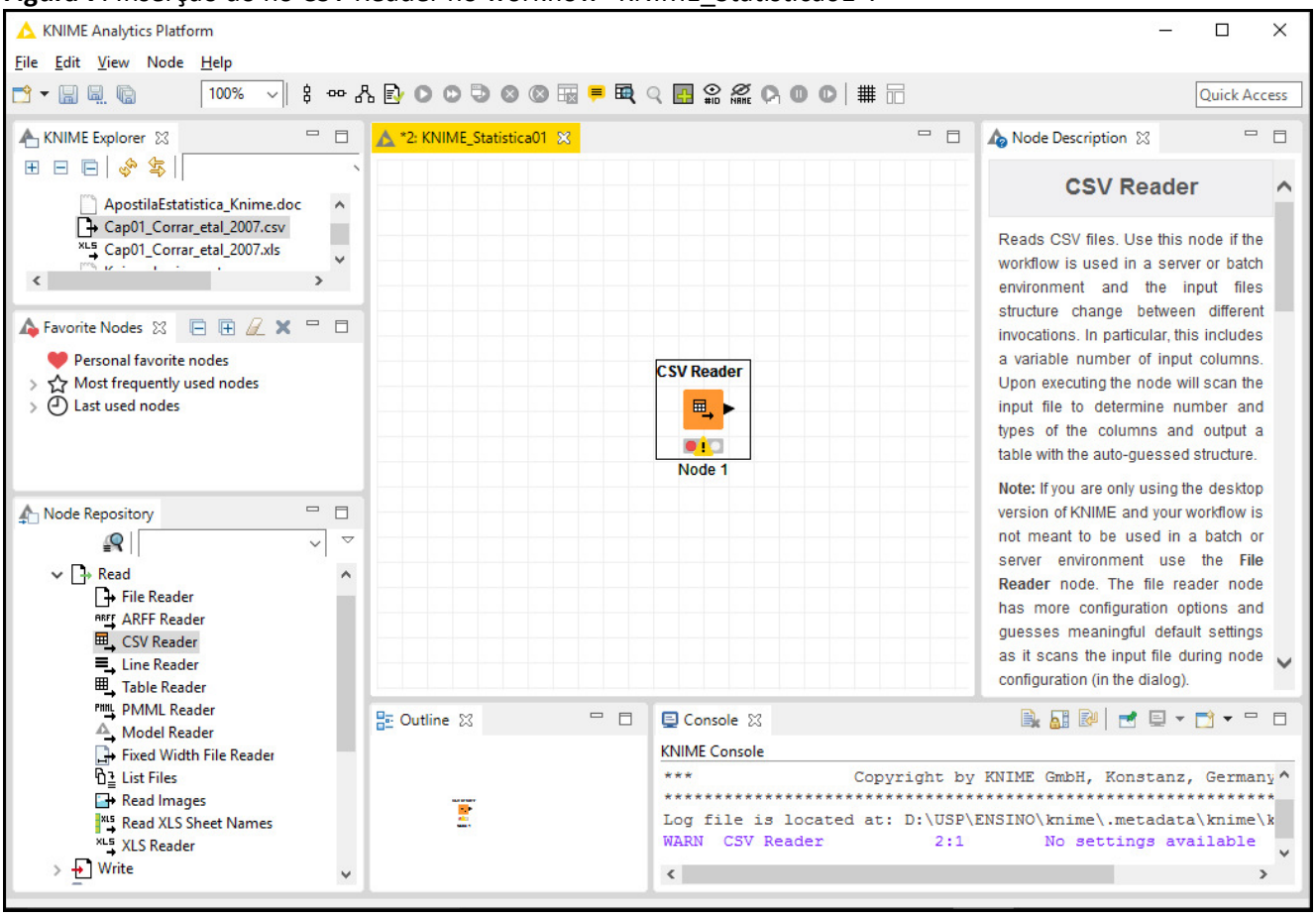


Figura 6. Iniciando um novo workflow no KNIME.



O primeiro nó a ser considerado é o de leitura de dados. Clicando duas vezes no nó CSV Reader o nó aparece na janela na qual será desenhado o workflow da análise. Veja Figura 7.

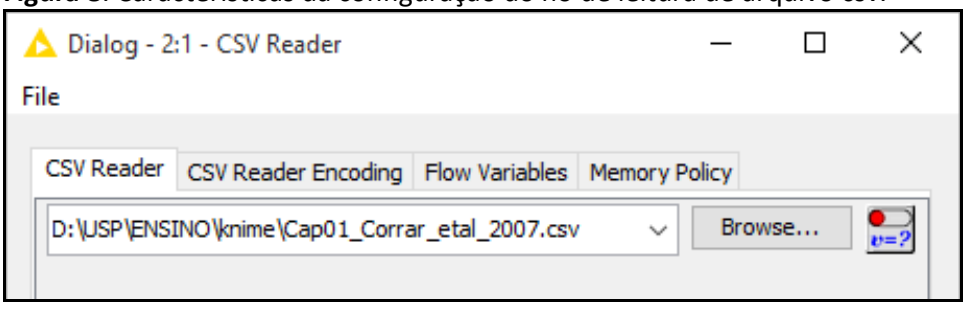
Figura 7. Inserção do nó CSV Reader no workflow "KNIME_Statistica01".



Note que o nó "CSV Reader" que aparece no centro da Figura tem a luz vermelha acesa e um ponto de exclamação (amarelo) indicando que o nó deve ser configurado. Clique no nó com o botão direito do mouse (ou clique no nó e aperte a tecla F6) para configurar o nó.

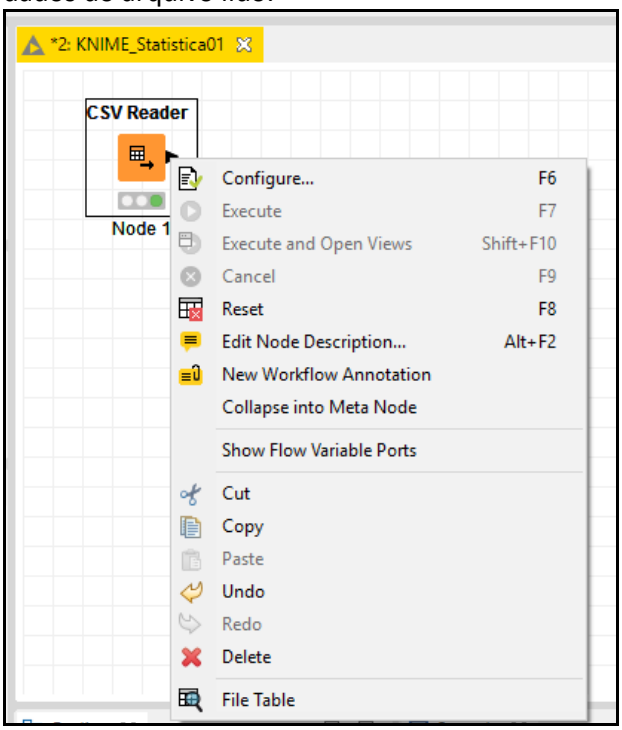
Verifique as características da leitura na janela apresentada (veja Figura 8). Modifique o delimitador de acordo com o tipo de arquivo que você gravou em csv. Atenção para vírgulas como casas decimais. Pode ser necessário editar o arquivo csv para que as casas decimais sejam ponto ao invés de vírgula.

Figura 8. Características da configuração do nó de leitura de arquivo csv.



Após configurar o nó clique em “OK” e depois de retornar ao workflow selecione o nó e clique no triangulo verde (Execute Selected – F7). Note se a luz do nó fica alterada de vermelha para verde. Se a luz ficou verde então o nó foi executado com sucesso e a leitura foi feita. O resultado pode ser observado na saída do nó, clicando com o botão direito observe a opção “File Table” (Figura 9).

Figura 9. Opções para o nó “CSV Reader”. Note a última opção “File Table” que apresenta a tabela com os dados do arquivo lido.



Após selecionar a opção “File Table” observe o resultado na Figura 10 para as 12 primeiras empresas da base de dados.

Figura 10. Tabela do arquivo “Cap01_Corrar_etal_2007.csv” obtida pelo nó “CSV Reader”. Detalhe das 12 primeiras linhas do arquivo.

File Table - 2:1 - CSV Reader

File

Table "Cap01_Corrar_etal_2007.csv" - Rows: 100 | Spec - Columns: 9 | Properties | Flow Variables

Row ID	CAP	TAM	PL	AC	PC	AP	ARLP	PELP	VLL
1	Capital Aberto	Pequena	63685	30475	41400	79300	5004	40098	0.046
2	Capital Fechado	Pequena	89430	53000	43125	128100	25020	17604	0.076
3	Capital Fechado	Média	81300	35775	74175	125050	43368	33252	0.095
4	Capital Fechado	Pequena	79945	30475	31050	118950	8340	26406	0.019
5	Capital Aberto	Grande	105690	60950	58650	68625	7506	58680	0.025
6	Capital Fechado	Média	65040	25175	44850	147925	27522	18582	0.027
7	Capital Aberto	Pequena	89430	59625	60375	115900	20016	44988	0.03
8	Capital Fechado	Média	69105	29150	48300	105225	35028	12714	0.001
9	Capital Aberto	Grande	63685	39750	60375	115900	13344	53790	-0.011
10	Capital Fechado	Média	81300	42400	63825	132675	29190	39120	-0.013
11	Capital Aberto	Pequena	65040	37100	34500	88450	13344	23472	0.007
12	Capital Aberto	Média	75880	33125	51750	126575	18348	38142	0.032

Este é o final do Exercício 1. ■

1.2 Estatísticas descritivas no KNIME

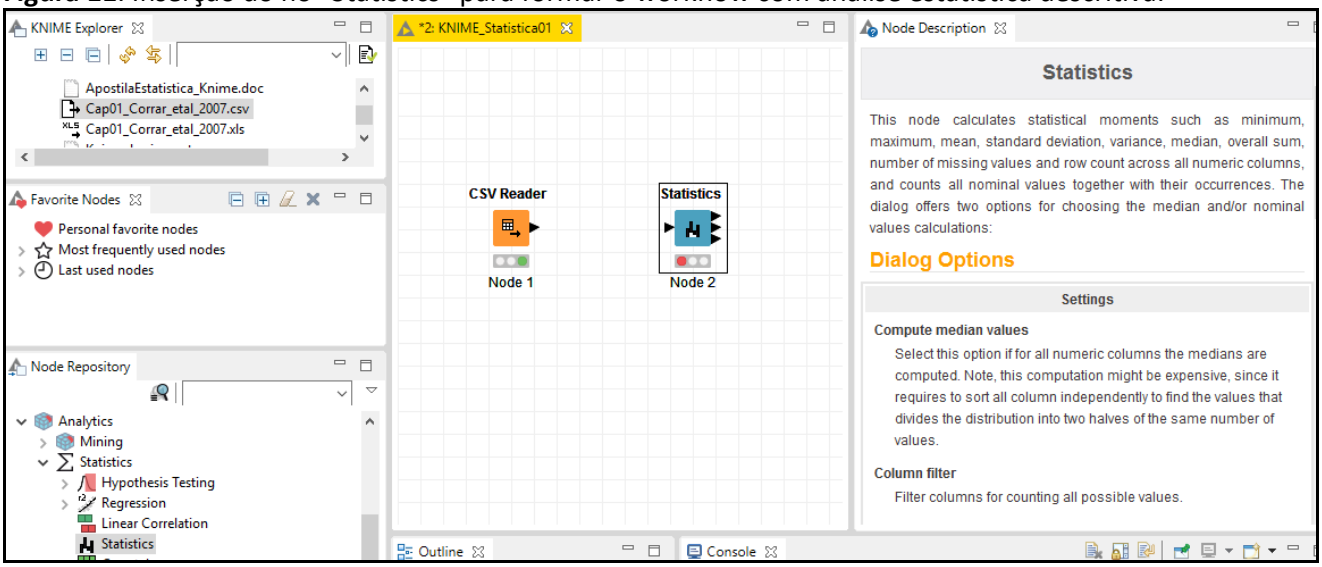
Após a leitura dos dados será explorada a obtenção de estatísticas descritivas tais como, média, desvio padrão, mínimo, máximo, frequência de ocorrência, ... das variáveis presentes na base de dados.

• **Exercício 2**: Estatísticas descritivas para as variáveis do arquivo Cap01_Corrar_etal_2007.csv,

Vamos considerar o mesmo workflow iniciado no Exercício 1. Para fazer a análise estatística descritiva das variáveis temos que considerar o repositório de nós.

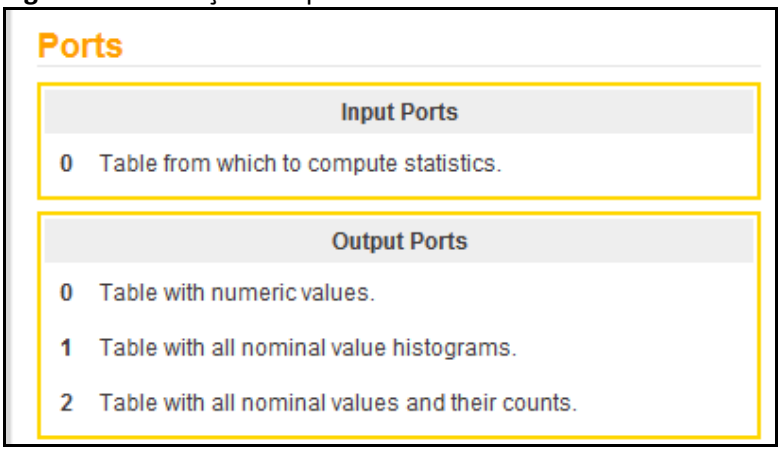
No repositório, no conjunto "Analytics" são disponibilizados nós em dois subconjuntos: Mining e Statistics. No subconjunto Statistics temos o nó Statistics. Vamos então inserir este nó no workflow "arrastando" o mesmo para um local próximo ao nó "CSV Reader". O resultado é apresentado na Figura 11.

Figura 11. Inserção do nó "Statistics" para formar o workflow com análise estatística descritiva.



Note na Figura 11, que o nó "Statistics" tem uma entrada (que será utilizada para entrar com a base de dados, e três saídas. A descrição completa do nó pode ser vista no quadro à esquerda (Node Description). A descrição das portas pode ser vista neste quadro e é apresentada na Figura 12.

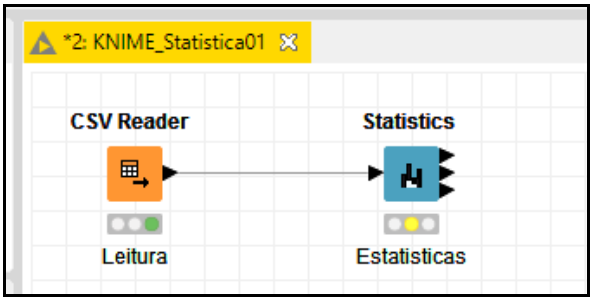
Figura 12. Descrição das portas do nó "Statistics".

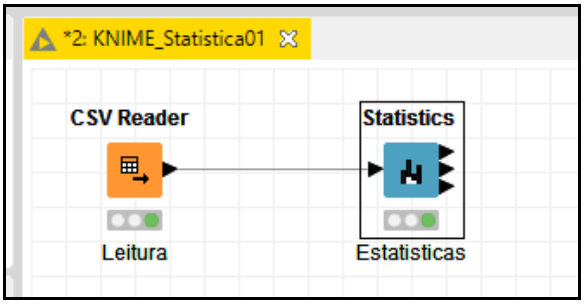


Assim, as estatísticas descritivas para as variáveis contidas no arquivo em análise podem ser obtidas conectando o nó de leitura de base de dados ao nó “Statistics” e depois e executando o workflow para então observar os resultados nas portas de saída.

Note que na Figura 11 o nó “Statistics” apresenta luz vermelha, indicando que o nó não tem entrada definida. Após definir a entrada a luz muda para amarela. Indicando que existe dados de entrada mas ainda não foram produzidas saídas (veja Figura 13). Quando executado (botão verde com triângulo) o nó passa para luz verde (veja Figura 13)

Figura 13. Workflow para análise estatística descritiva.





Os resultados das estatísticas descritivas podem ser acessados clicando com o botão direito no nó “Statistics”. As últimas três opções referem-se às três saídas. As Figuras 14 a 16 apresentam os resultados de forma resumida.

Figura 14. Porta de saída 1 do nó “Statistics”. Estatísticas descritivas para as variáveis numéricas: Mínimo, máximo, média, desvio padrão variância, ...

Statistics Table - 2:2 - Statistics (Estatisticas)

File

Table "default" - Rows: 7 | Spec - Columns: 16 | Properties | Flow Variables

Row ID	Column	Min	Max	Mean	Std. deviation	Variance	Skewness	Kurtosis	Overall sum
PL	PL	33,875	111,110	71,245.9	15,312.136	234,461,505.242	0.194	0.092	7,124,590
AC	AC	14,575	60,950	35,311.25	10,213.827	104,322,252.21	0.493	0.107	3,531,125
PC	PC	12,075	79,350	50,249.25	12,942.796	167,515,965.341	-0.363	0.15	5,024,925
AP	AP	56,425	152,500	106,094.25	24,257.341	588,418,596.654	-0.204	-0.885	10,609,425
ARLP	ARLP	1,668	45,036	19,715.76	9,971.795	99,436,685.76	0.469	-0.509	1,971,576
PELP	PELP	5,868	59,658	34,735.2	12,441.72	154,796,403.636	0.034	-0.729	3,473,520
VLL	VLL	-0.117	0.097	0.017	0.031	0.001	-0.772	3.944	1.695

Figura 15. Porta de saída 2 do nó "Statistics". Histogramas para variáveis nominais.

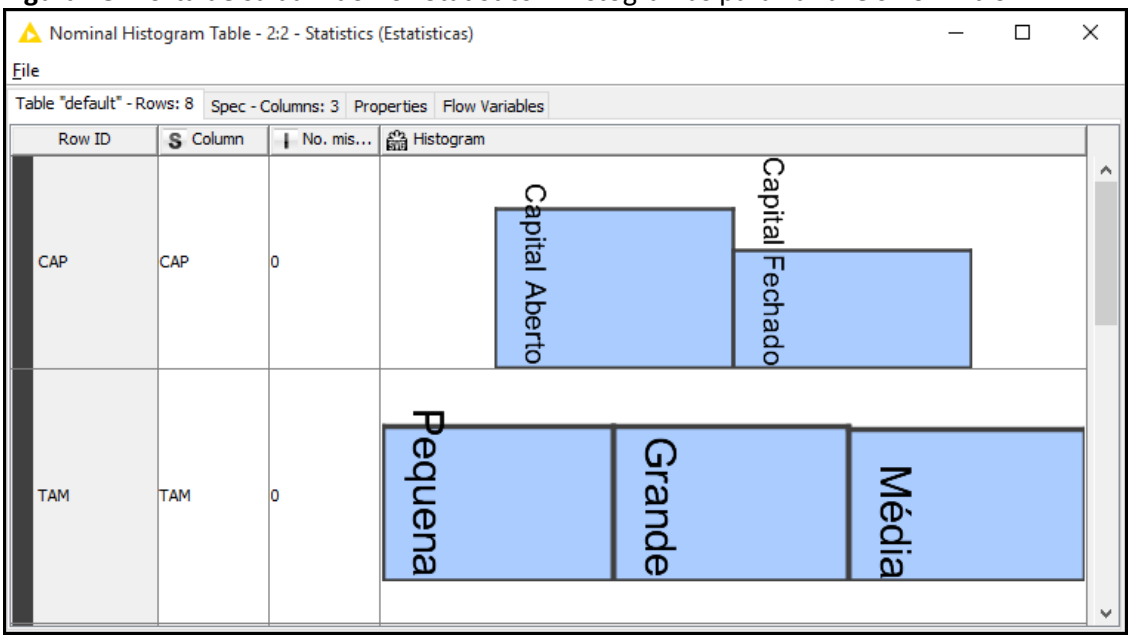
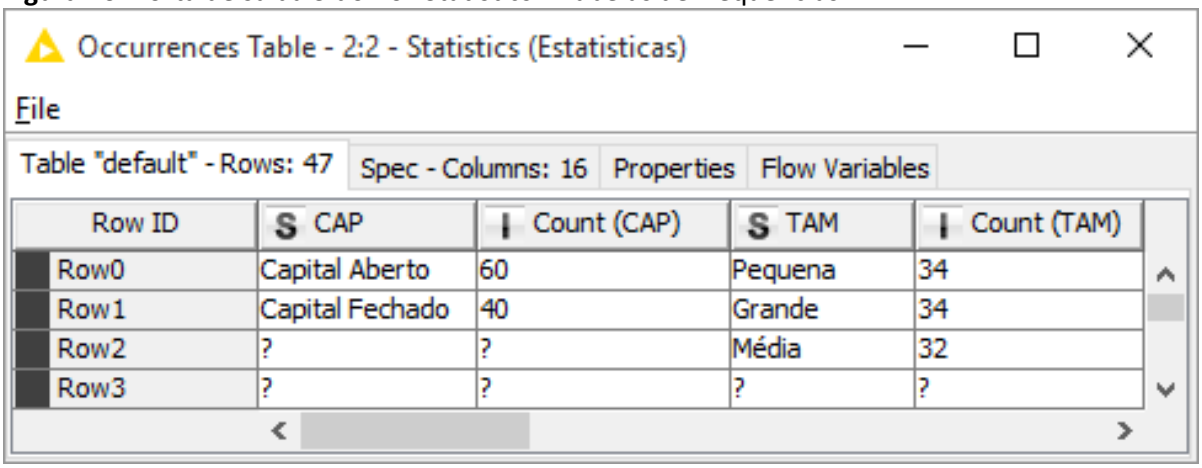


Figura 16. Porta de saída 3 do nó "Statistics". Tabelas de frequências.



Row ID	CAP	Count (CAP)	TAM	Count (TAM)
Row0	Capital Aberto	60	Pequena	34
Row1	Capital Fechado	40	Grande	34
Row2	?	?	Média	32
Row3	?	?	?	?

Assim, as principais estatísticas descritivas são obtidas pelo KNIME com o workflow descrito na Figura 13.

Este é o final do Exercício 2. ■

A obtenção de estatísticas descritivas através do workflow mais simples, que consiste em apenas dois nós conectados (Figura 13), conclui os exemplos de introdução ao software KNIME. As próximas seções são dedicadas ao conteúdo de disciplinas de Estatística Aplicada à Administração desenvolvido com o software KNIME.