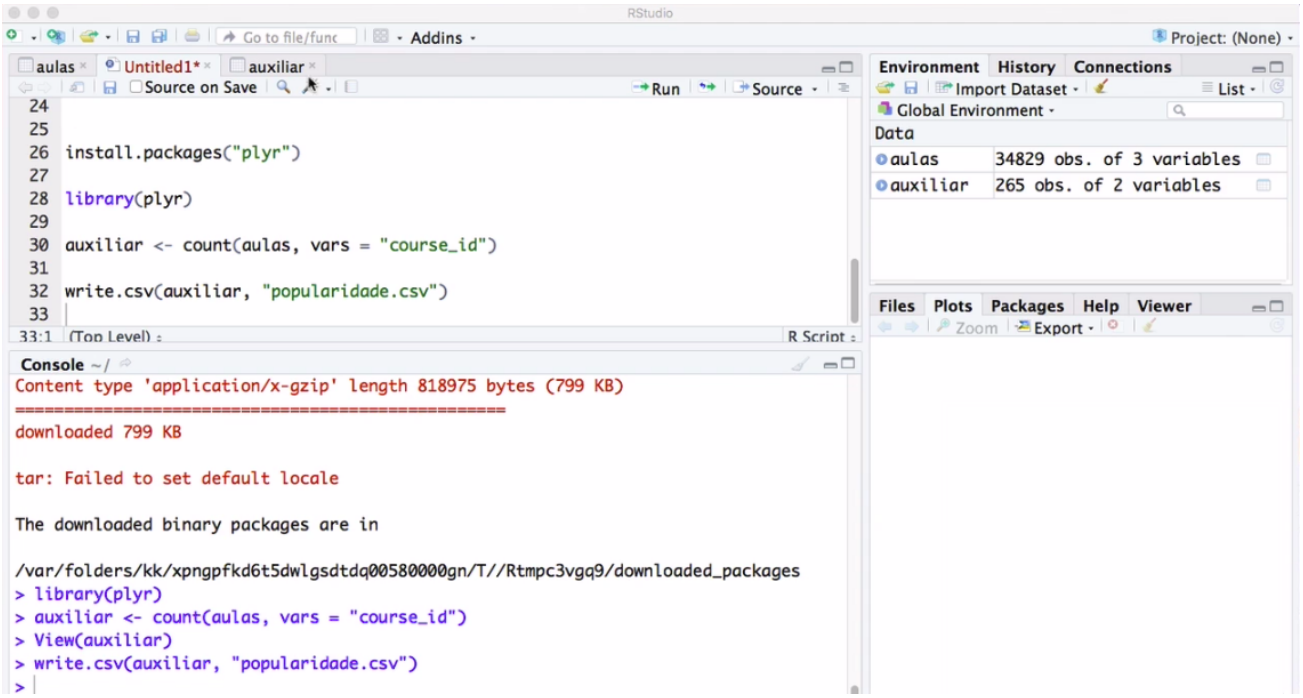
01

Criando gráficos

Transcrição

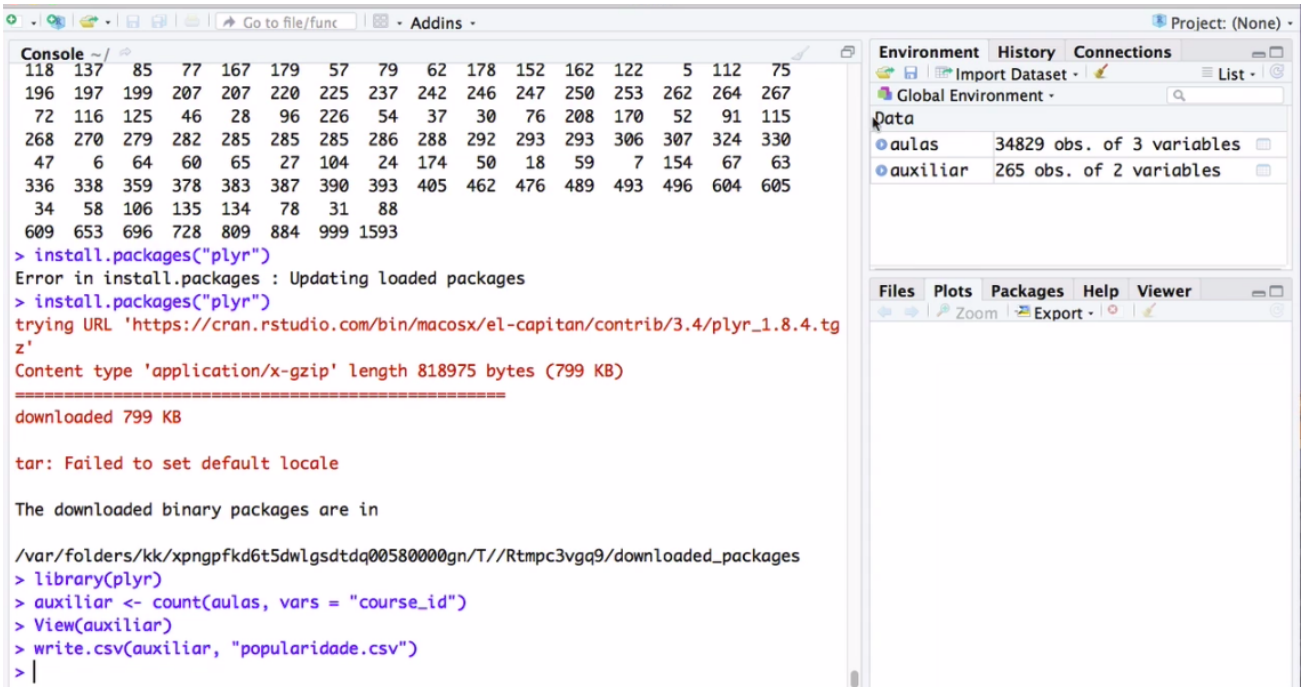
Trabalharemos com os cursos para descobrir a duração deles até a conclusão. Para isso, utilizaremos um novo banco de dados, mas antes limparemos a área de trabalho, que está com todas as informações da sessão anterior, para evitar confusão:



Para isso, fecharemos as abas de visualização:

- do objeto `auxiliar` ;
- de `aulas` ;
- o arquivo de R Script, que ao fecharmos, abrirá uma mensagem questionando se desejamos salvar.

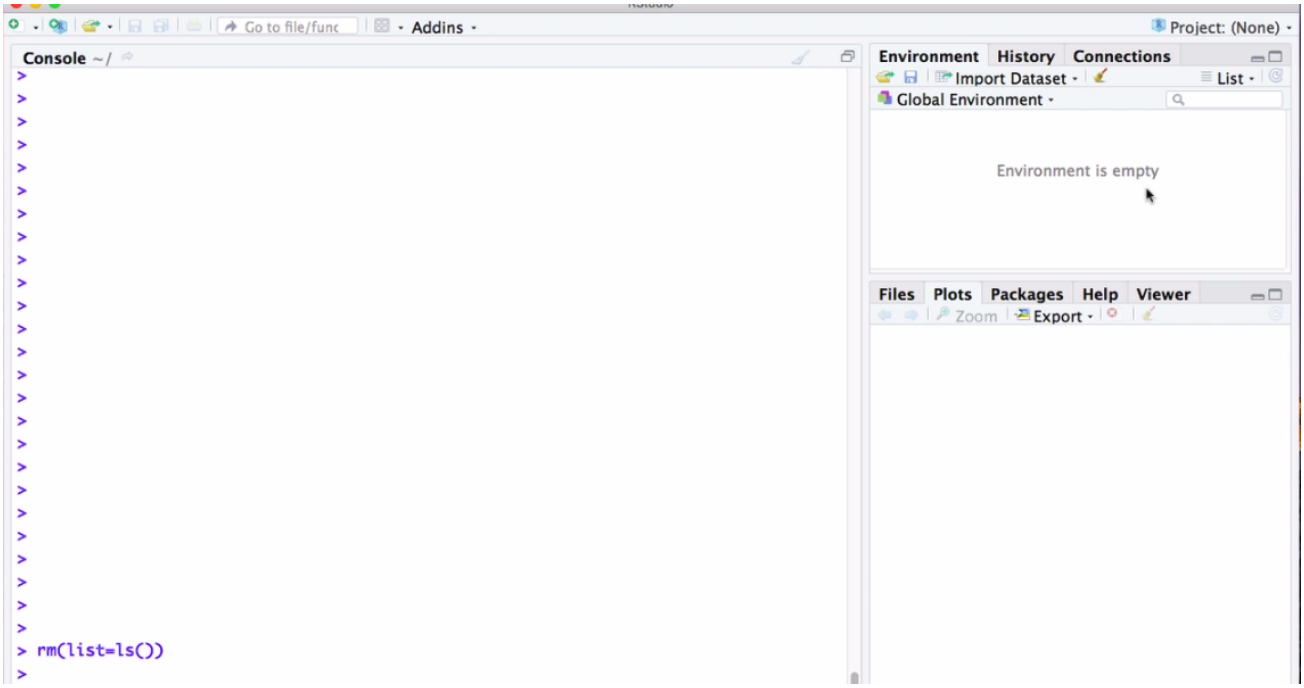
Salvaremos como `script1` com a extensão `.r` , que poderá ser aberta tanto no R quanto no RStudio, ou até mesmo em alguns editores de texto. Após fecharmos esses arquivos, a janela de Script ficará oculta.



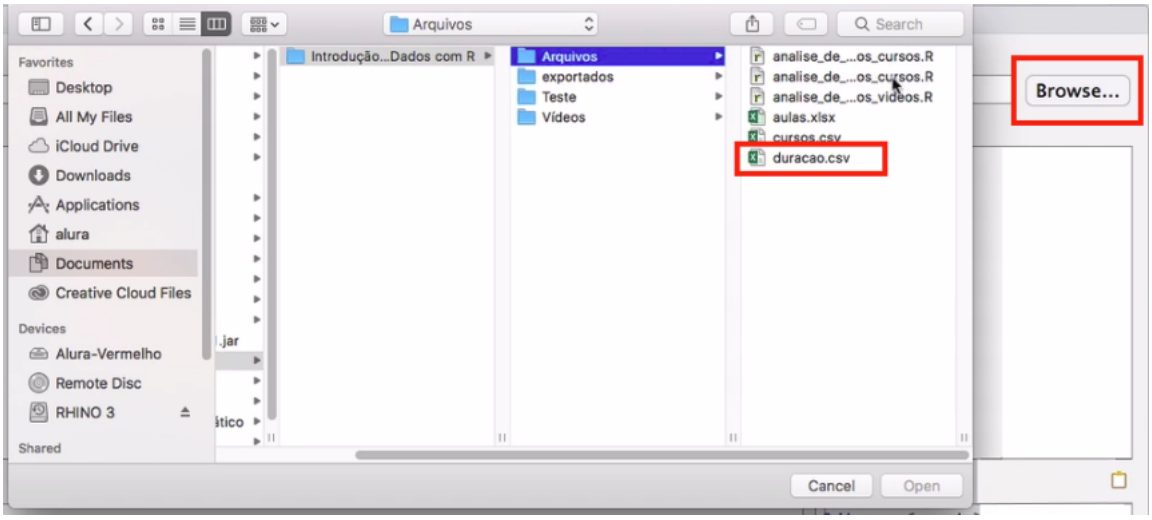
Ainda poderemos limpar o Console, por meio do comando `rm`, abreviação de *remove*. Entre parênteses, especificaremos `list=ls()`:

```
rm(list=ls())
```

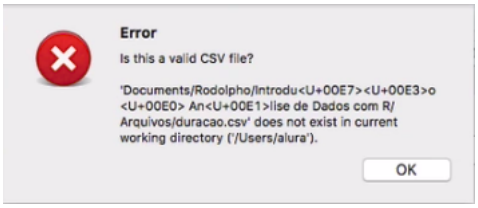
Assim, limparemos a **lista** do Console.



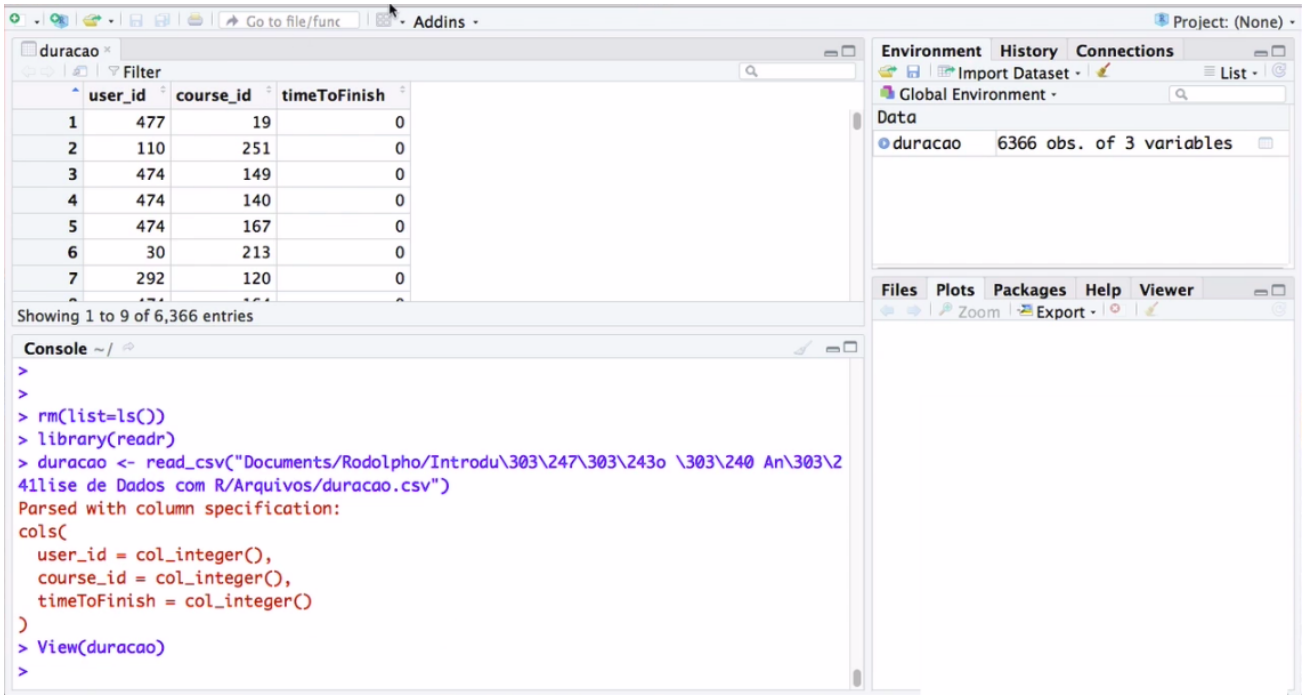
A área de trabalho está pronta para uma nova análise! Importaremos um novo banco de dados `duracao.csv`, com informações sobre o tempo que os alunos demoram para concluir os cursos. Clicaremos em "*Import Dataset* > *From Text (readr)...* > *Browse*" e selecionaremos "*From Text (readr)...*", pois agora trabalharemos o arquivo `duracao.csv`, considerado um arquivo de texto, e não `.xlsl`, do Excel.



Após a seleção, abre-se uma mensagem de erro para confirmar se o arquivo é `.csv` :

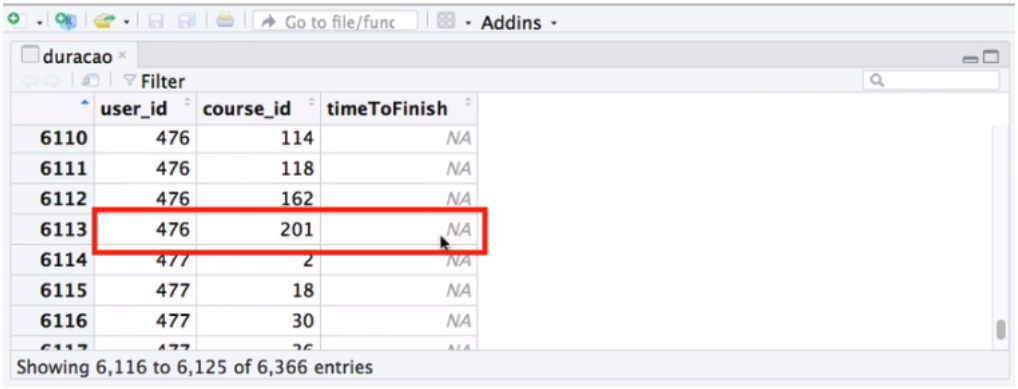


Podemos clicar em "OK" e, em seguida, em "*Import*", para importar o banco de dados. O banco de dados será exibido na parte superior à esquerda do programa. Nele, veremos as variáveis `user_id` com códigos dos alunos, `course_id` com códigos de cursos e `timeToFinish` , com que ainda não trabalhamos.



Essa nova variável, nomeada pelo desenvolvedor do banco de dados, informa quantos dias o aluno demorou para concluir um curso. Se lermos a primeira linha da tabela, por exemplo, obteremos a informação de que o aluno `477` levou `0` dias para concluir o curso `19` . Ou seja, ele começou e terminou o curso no mesmo dia, algo que acontece em muitos cursos.

Nas linhas abaixo, veremos que os dias aumentam; o aluno `45` fez o curso `181` em `5` dias. Seguindo para as próximas linhas da tabela, encontraremos que o aluno `476` , que fez vários cursos na empresa, não concluiu (`NA`) o curso `201` .



`NA` é sigla de "*Not Available*", "Não Disponível" em inglês. Isso indica que no banco de dados, não há número de dias naquela variável, e que o aluno começou e não terminou, ou ainda que não tinha terminado o curso quando a empresa nos passou o banco de dados.

De qualquer forma, somente a empresa tem acesso a essa informação. Discutiremos isso futuramente. Agora, abriremos uma nova janela R Script para colocar os novos comandos, clicando no ícone verde com sinal de mais (`+`) branco, localizado no canto superior esquerdo do programa.

Digitar `timeToFinish` toda vez que quisermos trabalhar com a variável poderia dificultar a análise. O nome dela possui uma escrita que a linguagem entende, mas é difícil para digitar, pois não há espaçamento e requer a utilização da tecla "Shift" a todo momento, começando com letra minúscula e no meio passa a ter letras maiúsculas no começo de cada palavra - isso é um padrão no desenvolvimento de banco de dados, sobretudo em SQL, por conta da legibilidade.

Usaremos um nome mais fácil para trabalharmos no RStudio; no novo documento de R Script que abrirmos, digitaremos `duracao` e atribuiremos (`<-`) a ela a função `rename()` , de renomear, que está no pacote `plyr` . Assim, estaremos solicitando que a variável seja renomeada. Especificaremos o banco de dados em que a variável se localiza, no caso `duracao` , e após a vírgula (`,`) digitaremos `replace` — comando de substituição, sinal de igual (`=`) e, em seguida, passaremos as variáveis que queremos renomear.

Ao passarmos mais de uma variável, utilizamos um comando chamado `c` , ou "Criação de um Vetor", em que especificaremos, entre parênteses, as variáveis que desejamos renomear, com seus respectivos novos nomes. Reparem que colocaremos os nomes antigos e novos **entre aspas** (`"`) e entre elas inserimos sinal de igual (`=`).

> **Atenção**: dependendo da máquina, a navegação com aspas muda, e poderão ser confundidas com o acento trema (`"`). Por isso, devemos ter cuidado na hora de digitar os nomes e sempre fechar as aspas após abri-las.

Para renomearmos outras variáveis, inseriremos vírgula e abrimos aspas. Alteraremos `"course_id"` para `curso` e a variável que mais nos interessava, `timeToFinish` , para `dias` , que será uma informação interessante para nós, com nome curto e fácil para trabalharmos. O comando ficará da seguinte forma:

```
duracao <- rename(duracao, replace = c("user_id"="aluno","course_id"="curso","timeToFinish"="dia
```

Executaremos para certificar que não cometemos erros de sintaxe.

```
> duracao <- rename(duracao, replace = c("user_id"="aluno","course_id"="curso","timeToFinish"="
```

Funcionou! No Console, tivemos o retorno correto. Caso apareça alguma mensagem de erro, basta voltar e corrigir. Agora, o banco de dados está com as variáveis renomeadas. Se clicarmos em "duracao" na janela superior direita, na janela ao lado esquerdo, o banco de dados será exibido com os novos nomes, mais intuitivos, das colunas:

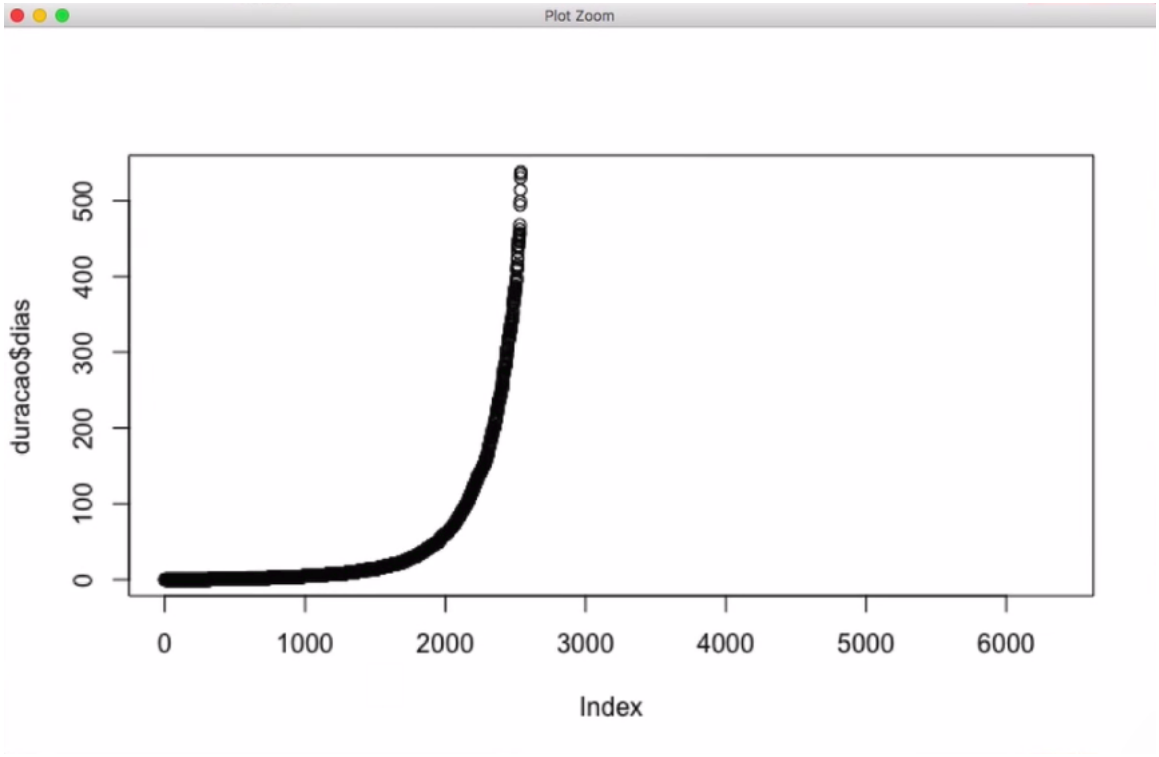
- `user_id` aparece como `aluno` ;
- `course_id` aparece como `curso` ;
- `timeToFinish` aparece como `dias` .

Analisaremos os cursos e quantos dias os alunos demoram para concluí-los. Vamos tentar criar um gráfico com esses dados, pois anteriormente vimos que trabalhar com eles de forma bruta é complicado. No RStudio, a função `plot()` é utilizada para criar gráficos. Digitaremos no novo arquivo de R Script, com o banco de dados `duracao` , seguido de cifrão (`$`) e `dias` , entre parênteses.

```
plot(duracao$dias)
```

Desta forma, pedimos para que o programa plote quantos dias os cursos levam para ser concluídos. A função `plot()` escolhe um gráfico — a princípio, não sabemos qual — a partir das informações que fornecemos, e do tipo de dados. Ao clicarmos em "Run" na janela inferior direita, será aberta uma janela de gráfico, com o padrão de exibição do RStudio para gráficos.

Podemos aplicar Zoom e expandir a janela para analisarmos melhor o gráfico:

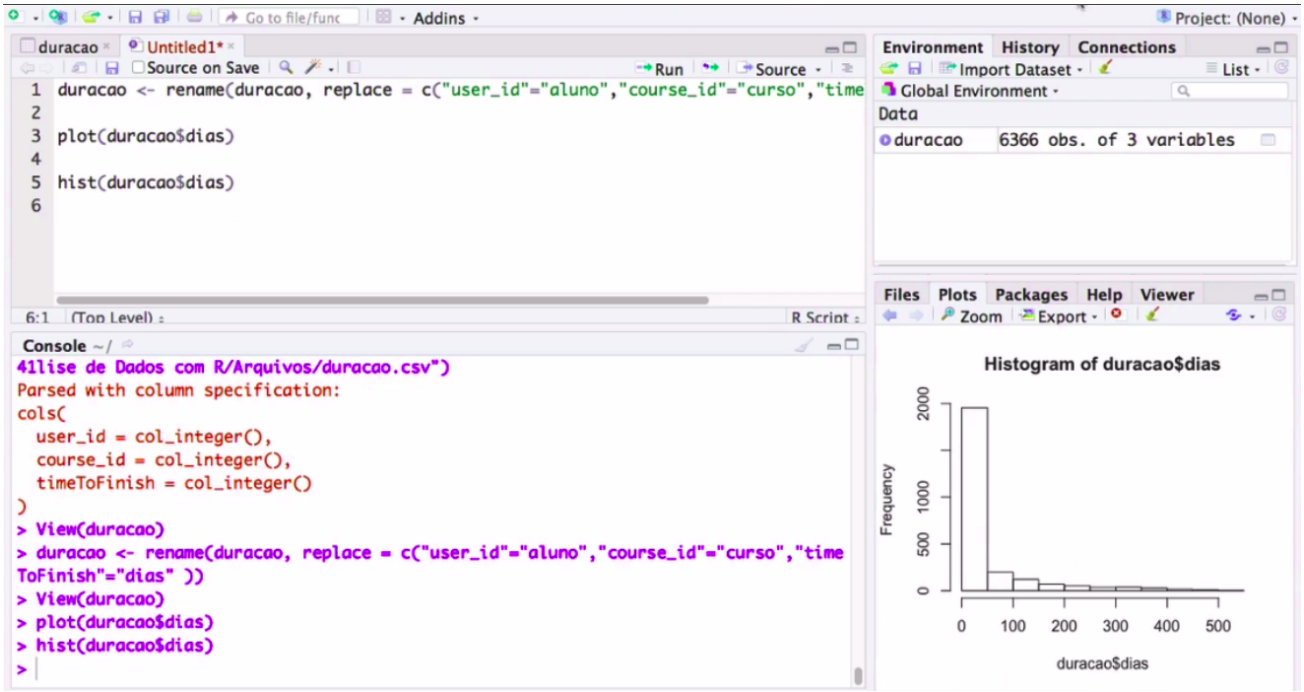


Nele, temos um índice ("*Index*", na horizontal), um posicionamento das observações, e na vertical, a duração em dias ("duracao$dias"). Não conseguimos extrair nenhuma informação a partir desse gráfico, e constatamos que não foi uma boa ideia deixar para o programa escolher o gráfico para a análise.

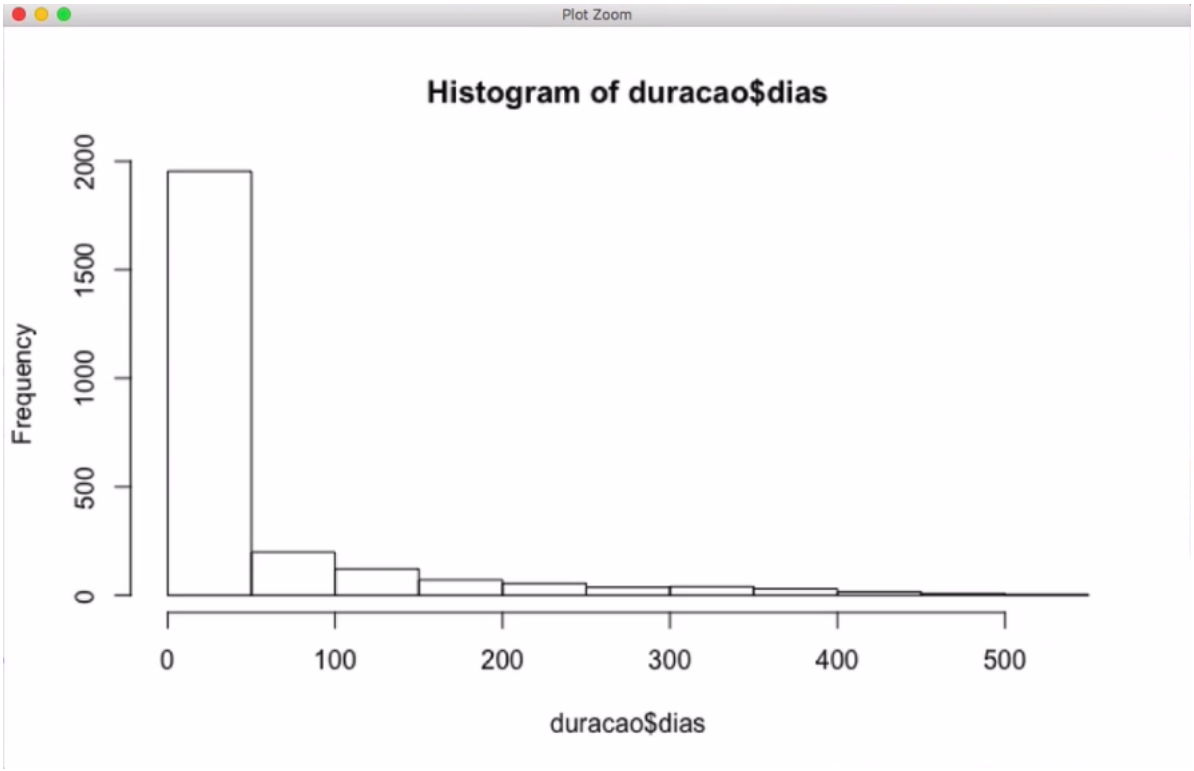
Existe um gráfico melhor, do qual poderemos extrair mais informação sobre a distribuição dos dias, que é o nosso objetivo. Poderemos quantificar os cursos e os dias necessários para conclusão. O gráfico adequado para essa análise é o histograma, cujo comando no RStudio é `hist` :

```
hist(duracao$dias)
```

Ao ser executado, teremos um gráfico informativo:



E com o zoom, teremos uma melhor visualização:



Conseguimos visualizar a distribuição da duração dos cursos. No eixo horizontal observamos a variável `dias`, que vai de `0` a mais de `500`. No eixo vertical, temos a `frequência`, e por meio dela podemos enumerar quantas vezes os cursos que duraram `0` dias apareceram na amostra, por exemplo. São quase `2000` deles realizados em um único dia. Depois, há aproximadamente `250` cursos que duraram entre `50` e `100` dias. E assim o número vai diminuindo. No entanto, ainda no final do eixo horizontal encontraremos alguns cursos que levaram mais de `500` dias para a conclusão.