Minicurso: Ciência de dados: uma introdução a linguagem "R"

Palestrantes: ROBERT ARMANDO ESPEJO e
ANTONIO CARLOS CANTEIRO DORSA

O presente curso objetivou a importância do uso da linguagem "R", desde rotinas simples até as mais complexas, envolvendo a instalação de programas e a instalação de pacotes. Posteriormente foram exemplificados comandos básicos e rotineiros, correspondente a entrada, processamento e saída de dados através de gráficos.

A linguagem "R" é utiliza um software totalmente gratuito, de fácil aprendizagem, amplamente utilizado no mercado e pela academia, com uma enorme quantidade de pacotes e uma ótima ferramenta para a criação de gráficos e de relatórios.

O "R" é dividido entre o "*R base*" e o "*RStudio*", a primeira função serve como uma linha para inserção de comandos, e com isso, o novo "programador" encontrava um caminho inicial mais complexo e limitava os que possuíam maior experiência.

Com a chegada do "*RStudio*", apresentou-se uma interface gráfica mais moderna, com diversas funcionalidades e, assim, facilitando o dia a dia.

O "*RStudio*" é dividido em quatro partes principais:

a) **Editor de Código:** Nesse editor é onde escrevemos e editamos os comandos, é provavelmente a parte que mais utilizaremos;
b) ***Console:*** No console é mostrado a maioria dos resultados do comando, sendo possível executar os comandos com mais rapidez. Nesse local também é possível o uso da ajuda pelo comando **"?"**.
c) ***Environment* e *History*:** No *Environment* e *History*, ficam guardados todos os objetos que foram criados na sessão. Na aba *History* é criado um histórico de comandos que foram utilizados.
d) ***Files, Plots, Packages, Help* e *Viewer*:** Na aba *Files* teremos uma navegação dos arquivos do computador. Aqui poderemos definir o diretório de trabalho. Na aba *Plots*, mostra os gráficos gerados, e também, possibilita a exportação dos mesmos com alguns formatos, como .pdf, .png. Na aba *Package*, estão listados os pacotes que foram instalados pelo usuário. Podemos verificar quais estão em uso, e carregar algum outro necessário para a sua análise do momento. Na aba *Help*, é a ajuda do *RStudio*, podendo aqui sanar as dúvidas de cada pacote. E, por fim, a aba *Viewer* onde visualizamos o conteúdo *web*.

O curso de "R" ministrado na SBPC, mostrou os conceitos básicos da ferramenta, desde o uso do software para calculadora, o entendimento de algumas funções do *Console*, a criação de *Scrips* simples, como salvar o *Script*, como colocar comentário nos *scripts*, e esses comentários não aparecerem no texto quando finalizado, usando o comando (#). Num segundo momento a criação de objetos (variáveis), passando pelo uso de algumas funções básicas (raiz quadrada, média, mediana, arredondamento, quantis, variância, soma dos valores, etc). Em outro momento, o conceito dos pacotes (*package*) faz-se necessário, que são usados para auxiliar as diversas linhas de pesquisa, como medicina, biologia, gráficos, estatística, econometria e outras. Um comando é apresentado para a instalação, como também uma facilitação que o *RStudio* apresenta, que são os atalhos para a busca do pacote desejado.

O momento seguinte é talvez um dos mais esperados pelo usuário, ou seja, a leitura dos dados, que podem ser desde dados estruturados (bem separados e organizados), dados não-estruturados (onde não se tem uma estrutura previsível) e os dados semiestruturados (que possuem organização fixa, mas não um padrão de estrutura linha/coluna).

O passo seguinte é definir o local para os dados (*Working directory*), que é uma pasta de trabalho para o uso do software "R" e de preferência, exclusiva do R. Em seguida, escolhemos os pacotes para a leitura dos dados que serão usados. Alguns pacotes o "*R base"* já possui, como o *readr, tidyverse* e outros.

A manipulação dos dados é o passo seguinte, onde após a leitura e o carregamento desses dados, precisamos dominar algumas técnicas para a manipulação, ou seja, como executar algum modelo econométrico, como visualizar os dados num gráfico, qual variável escolher, como preencher a tabela com dados faltantes (NA), como usar as funções de leitura de dados e outras.

O *RStudio* também oferece etapas como a limpeza de dados (deixando no formato ideal a ser usado), separando variáveis, unindo-as, e também usando para leitura de dados de texto, união de dados, cruzamento de dados. Para assim, chegarmos na fase final da análise, e assim a necessidade da geração de arquivos, gráficos, planilhas, pdf e outros.

Para a parte de visualização de dados, o pacote mais desenvolvido para tal fim, é o (*ggplot*). E muitos *scripts* livres na web são encontrados para que o usuário encontre seu gráfico ideal. O pacote "*ggplot"* oferece uma série de tipos de gráficos, e são usados com a opção (*geom)*: Dispersão (*geom_point*), gráfico de bolhas (*geom_point*), gráfico de barras (*geom_bar*), Histograma (*geom_histograma*), Boxplot (*geom_boxplot*), Densidade (*geom_density*), Gráfico de linhas (*geom_line*) e também oferece as edições para cada linha de comando, como a utilização de camadas, escalas, eixos, como também mapear algum elemento estético para cada variável, entre outros. No pacote temos a opção dos *subplots*, para uso em gráficos de séries temporais por exemplo, e entre outros.

Nesse trabalho especificamos uma parte do vasto uso da ferramenta *RStudio*, e esperamos termos contribuído para a comunidade da **71ª Reunião Anual da SBPC.**