Análise dos concursos da

Mega-Sena

­­

Universidade Federal Rural de Pernambuco

Computação para Análise de Dados – 2018.1

Bruno Henrique Pereira Marques

Danielly de Moura Borba Queiroz

Recife, 12 de Agosto de 2018

# Introdução

No Brasil, milhões de pessoas jogam nas loterias federais, seja com baixa frequência ou tem o costume de apostar. Assediadas pelos altos valores das estimativas do prêmio, ainda sim, sabem que a probabilidade de acertar as seis dezenas da Mega-Sena é de 1 em 50.063.860 (cinquenta milhões, sessenta e três mil e oitocentos e sessenta), apostando em apenas um jogo com 6 dezenas que custa R$3,50. É possível reduzir esta proporção para 1 em 10.003 (dez mil e três) porém, esta aposta custará R$17.517,50 [Ref. 1].

O valor arrecadado em cada concurso da Mega-Sena é dividido nas seguintes partes: 54,7% para Programas Sociais do Governo e 45,3% para a Premiação. Ainda há um desconto de 13,8% da parte da premiação para o imposto de renda. Por fim, o resultado desta operação é dividido em: 35% para 6 acertos, 19% para 5 acertos, 19% para 4 acertos, 22% acumulado para o próximo concurso e 5% acumulado para o concurso no final do ano. Apesar da porcentagem final do prêmio ser baixa, o número de apostadores tem aumentado [Ref. 2].

Através dos dados históricos de todos os concursos da maior loteria do Brasil [Ref. 3], realizados desde o primeiro (em 11/03/1996) até o último concurso (número 2000) do ano de 2017, esperamos revelar dados como: crescimento do número de ganhadores, crescimento do valor dos prêmios, estado e região com maior número de ganhadores, estado e região com maior valor premiado, e relações entre a ocorrência das dezenas.

Para explorar os dados da Mega-Sena, será utilizado o arquivo no formato HTM disponibilizado no site oficial da loteria [Ref. 1]. Esta é a única forma que os dados são encontrados. A lista a seguir apresenta as principais variáveis contidas no arquivo.

* Número do concurso
* Data do concurso
* Primeira dezena sorteada
* Segunda dezena sorteada
* Terceira dezena sorteada
* Quarta dezena sorteada
* Quinta dezena sorteada
* Sexta dezena sorteada
* Ganhadores da sena
* Cidade dos ganhadores
* Estado dos ganhadores
* Valor de rateio da sena
* Valor acumulado

A proposta deste trabalho é apresentar gráficos de interesse, representados por ano, por estado, por região, pelos valores dos prêmios e com base nas dezenas sorteadas. Esses gráficos ajudarão a apresentar o crescimento da loteria no decorrer dos anos, como o cenário está montado de acordo com os ganhadores no Brasil e como a ocorrência das dezenas está distribuída. Além disso, será apresentada previsões de estimativas de prêmio, valores acumulados e estimativa para a Mega-Sena da virada.

# Pacotes requeridos

A tabela seguinte apresenta os pacotes que são necessários para executar o código sem erros.

|  |  |
| --- | --- |
| Nome do Pacote | Descrição |
| Tidyr | Pacote para organização de dados. |
| Dplyr | Pacote para organização de dados. |
| Maps | Desenha linhas e polígonos conforme um banco de dados de mapas. |
| Maptools | Leitura de ESRI *shapefiles* de mapas. |
| Mapproj | Gerencia mapas, projeções e *grids*. |
| Ggmap | Gerencia mapas baseados em ggplot2, com suporte a Google Maps. |
| Rgdal | Conjunto de operações para projeções. |

# Preparação dos dados

Os dados inicialmente foram baixados pelo link [Ref.1] onde se encontram no formato HTML. Cada variável está localizada e uma coluna do código HTML conforme a figura 1.

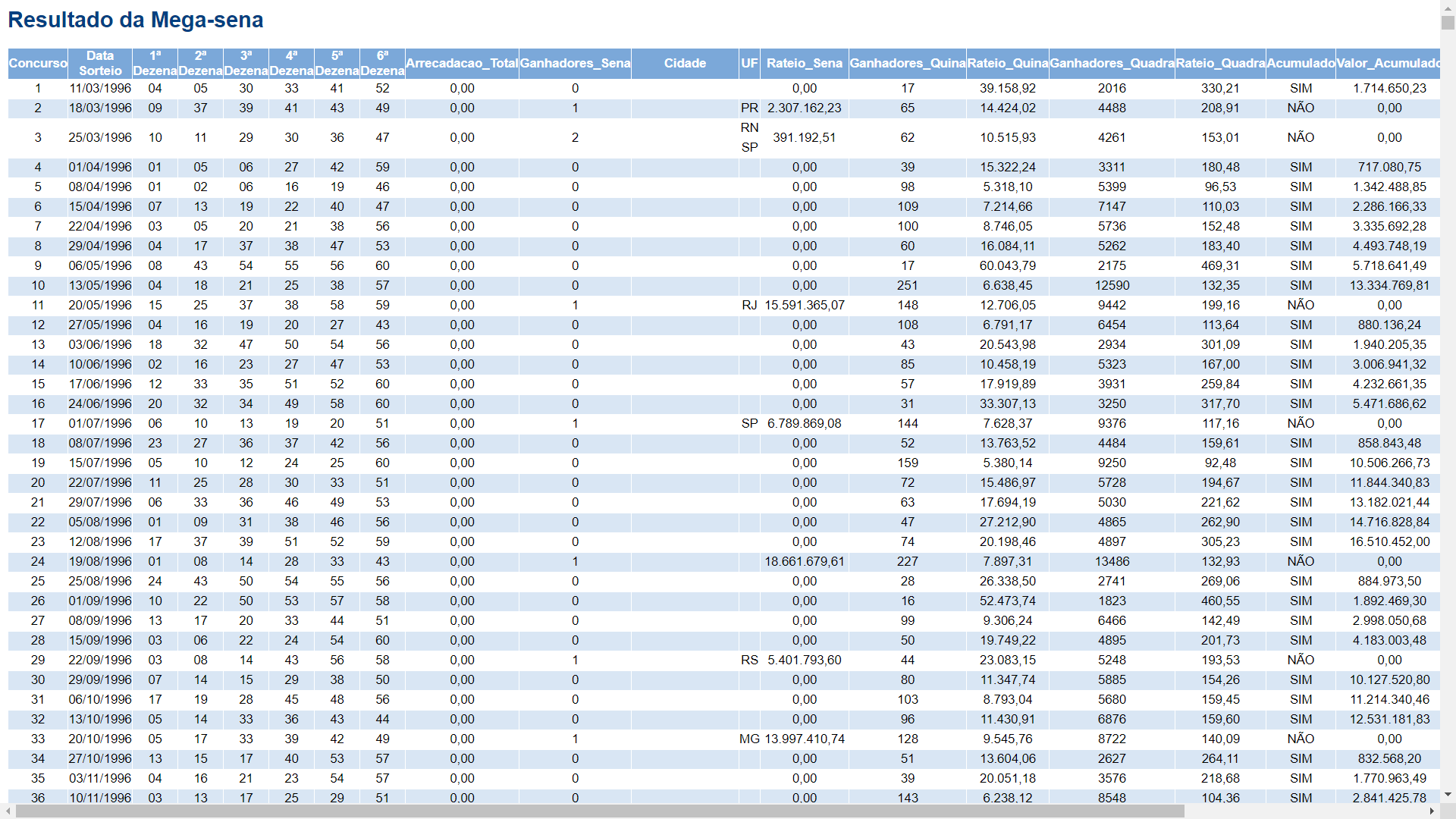


Figura 1: Tabela HTML com os dados da Mega-Sena.

Em seguida, foi realizado uma conversão manualmente do código HTML para CSV utilizando a ferramenta Excel 2013. Dessa forma, passa a ser possível carregar os dados para processamento com a linguagem R. A figura 2 apresenta parte dos dados carregados na ferramenta RStudio. Alguns problemas de conversão são visíveis como a linha 4, por exemplo, a qual está toda preenchia com ***NA***, valores de UF separados por vírgula e a formatação da moeda Rateio\_Sena.

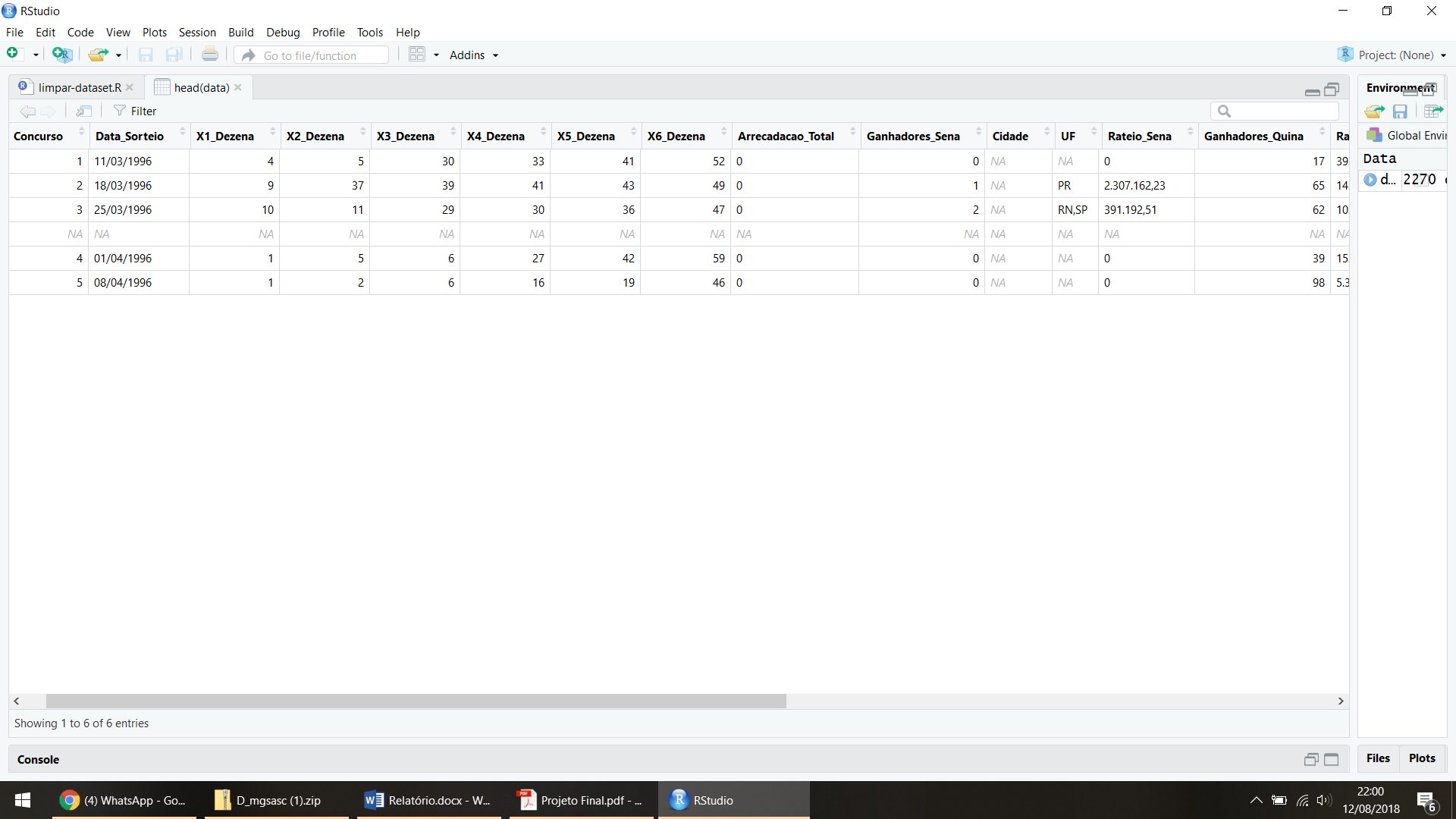
­­

Figura 2: Tabela CSV convertida.

Por fim, a remoção, substituição e formatação dos valores foram realizados para que se o manuseio com os dados fique mais coerente. A lista abaixo apresenta a ordem das operações finais da preparação dos dados.

* 1. Remoção das linhas com ***NA***;
  2. Conversão das datas formatadas como *String* para variáveis numéricas separadas (Dia, Mês e Ano);
  3. Conversão da moeda formatada como *String* para valor numérico decimal;
  4. Conversão dos valores SIM e NAO (*String*) para TRUE e FALSE (*Logical*);
  5. Agrupamento das variáveis X***n***\_Dezenas (***n*** é a ordem a qual a dezena foi sorteada) para um *dataframe* organizado com número do concurso, ordem do sorteio e dezena – figura 3;
  6. Separação dos valores da variável UF para um *dataframe* organizado com número do concurso e UF – figura 4;
  7. Criação de um *dataframe* organizado para as regiões do Brasil – figura 5;
  8. Persistência dos *dataframes* em quatro arquivos CSV.

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Figura 3: clean-data-dezenas.csv | Figura 4: clean-data-estados.csv | Figura 5: clean-data-regioes.csv |

# Análise exploratória dos dados

# Conclusões

# Referências

Ref. 1 – Caixa Econômica Federal – Mega-Sena.

<http://loterias.caixa.gov.br/wps/portal/loterias/landing/megasena/>

Ref. 2 – G1 – Número de ganhadores da Mega da Virada foi recorde; veja estatísticas.

<https://g1.globo.com/economia/noticia/numero-de-ganhadores-da-mega-da-virada-foi-recorde-veja-estatisticas.ghtml>

Ref. 3 – IQ Sorteios – Conheça as três loterias mais populares do Brasil.

<https://www.iq360.com.br/sorteios/artigos/as-loterias-mais-populares-do-brasil>