

Final Project Part 1

Michelle Helfman

10-30-2022

Introduction

Lotteries are everywhere. You can buy tickets at grocery stores, gas stations, airports, kiosks at the mall, everywhere... Lottery sales exceeded \$210 billion in 2021, and individuals, families, friends, and groups of co-workers want their piece of the lottery pie. I will be researching the two most popular lotteries in the United States: Powerball and Mega Millions. First, I will look into the most common numbers, jackpots, and the states where the winning tickets were sold for each lottery. Then I will compare the Powerball and Mega Millions results to each other, focusing on commonalities and differences to answer questions about which lottery to participate in, which numbers to include, and which states to reside in when purchasing tickets.

Statistical analysis of the available lottery data gives insight into the patterns and frequency of the numbers drawn and the size and frequency of jackpots. Then, the results can be displayed graphically for ease of understanding.

Research Questions

Mega Millions Analysis

1. What are the 20 most frequent individual numbers?
2. What are the 20 largest jackpots?
3. Which year had the paid out the most money since 2012?
4. How much time is between winning lottery drawings?
5. Which ten states have the most winners?

Powerball Analysis

1. What are the 20 most frequent individual numbers?
2. What are the 20 largest jackpots?
3. Which year had the paid out the most money since 2012?
4. How much time is between winning lottery drawings?
5. Which ten states have the most winners?

Mega Millions vs. Powerball

1. Compare the 10 most frequent individual numbers?
2. Compare the 10 largest winning lottery jackpots?
3. Compare the 10 states have the most winners?

Addressing The Problem

My first step will be to extract lottery information from Mega Millions and Powerball related websites. Then I will standardize the data for the 2 lotteries, including dates, numbers, and state information, then importing the results into R Markdown code to graphically display the analysis for the two lotteries. Finally, I will compare the Mega Millions to Powerball results to see the commonalities and differences between the 2 lotteries and graphically depicting those results.

Approaching The Problem

Once the data is standardized, it can be summarized, analyzed, and displayed for each lottery. Having the information consistent across both lotteries, allows for comparison between the two lotteries, and the results are displayed similarly to the individual lotteries.

Datasets (Select Description to Open Website)

The informataion was collected on October 15, 2022

Powerball drawings from Feburary 2010 to present - This data has 11 columns including the drawing date that is broken into 3 columns that will be concatenated into a usable date column. In 2012, the Power Play was introduced so prior to that date, this is missing, but will not needed for this project. Finally the resulting dataset will only include lottery drawings for January 2012 to present.

Mega Million drawings from December 2003 to present - This data has 11 columns including the drawing date that is broken into 3 columns that will be concatenated into a usable date column. Finally the resulting dataset will only include lottery drawings for January 2012 to present.

Powerball - Winning Jackpot and State information 2003 to present - This webpage has both state and jackpot information. The dataset with state information has 3 columns. The jackpot related data has 6 columns. Two columns have jackpot amounts, I will be using the Annuity Value column since this is the publicized amount.

Mega Millions - Winning Jackpot and State information from 2002 to present - This webpage has both state and jackpot information. The dataset with state information has 3 columns. The jackpot related data has 6 columns. The drawing date includes the day of the week, which will need to be removed before standardizing this particular column.

Project Packages

1. ggplot2 - For graphs
2. plotly - For graphs
3. stringr - String functions
4. plyr - For manipulating data structures
5. xlsx - Reading spreadsheets
6. tidyverse - Variety of libraries for graphs, string manipulations, and more
7. pandor - Rendering R objects in RMarkdown
8. lubridate - Working with dates
9. Additional packages will be inclued if needed.

Plots and Tables

Visualizations

1. Bar Graphs
2. Line Graphs
3. Histograms

Tables

1. Most common Individual numbers
2. Largest Jackpots

Questions For Future Steps

1. How do you depict frequency distribution?