Final Project Part 2

Michelle Helfman

11-06-2022

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 winnings per year?
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.

Data Importing and Clensing

Data Import

1. The lottery numbers for both the Mega Millions and Powerball drawings were downloaded as CSV files from the Texas Lottery website.
2. Winning jackpot information for the Mega Millions lottery, including jackpot amounts, dates of the winning drawings, and the states where the winners reside, is copied from the official Mega Millions website. The information is then pasted into an excel spreadsheet, and a CSV file is created and additional commas in the winner name were removed.
3. Winning jackpot information for the Powerball lottery, including jackpot amounts, dates of the winning drawings, and the states where the winners reside, is copied from the official Powerball website. The information is then pasted into an excel spreadsheet, and a CSV file is created and additional commas in the winner name were removed.
4. The CSV files were loaded into a SQL Server database.
5. Cleaned and standardized data will be loaded into R dataframes for selecting %>% *specific columns* %>% filtering %>% grouping.

Data Cleaning and Standardization

1. Stored Procedures were created to standardize the lottery drawing information for both the Mega Millions and Powerball lotteries. The drawing dates were initially separated into the day, month, and year. The 1st thing is to create a combined date (mm/dd/yyyy), then use a cursor to create rows of just game_name, drawing_date, and one number_drawn. By adding a game name column, both lotteries can be combined later into 1 dataset but still seen as two separate lotteries. Finally, create a dataset for each lottery with information starting at 1/1/2012
2. the Mega Millions and Powerball state-related records required no additional changes to the data, but a game name column was added so the records can be combined later.
3. The Mega Millions jackpot records required removing the day of the week and the abbreviations at the end of the day and converting this date to mm/dd/yyyy format. A game name column was added.
4. The Powerbuilder jackpot records required no additional changes to the data, but a game name column was added.
5. Lottery jackpot information was then added to the Mega Millions and Powerball state records to create records with jackpot winner totals by state.

Questions About Importing.

1. Can 2 files be loaded into 1 dataframe at the same time or does each one have to be loaded into separate dataframes and combined using rbind()?
2. Can 2 sheets in an Excel spreadsheet be loaded into 1 dataframe at the same time or does each one have to be loaded into separate dataframes and combined using rbind()?

Comments About Data Cleaning and Standardization

Start with the pieces that will become keys to tie the various tables or datasets together. Ids (like customer numbers or codes) must match so the information is meaningfully combined. This may require the use of a cross-reference table.
Then start standardizing dates, numbers (especially with decimal places), and any information used as filters, in groupings, or presented in reports. Always keep in mind where the data is coming from and going to; indicators are needed to identify the system/dataset where the information belongs.

Final Data Samples

```
## Set the working directory
setwd("C:/Masters Degree/DSC 520 Statistics for Data Science/Final  Project")

## Mega Million Numbers Drawn
mm_numbers_df <- read.csv("Data Samples/Megamillions_Numbers_sample.csv")
mm_numbers_df
```

```
##        game_name drawing_date number_drawn
## 1  Mega Millions     1/3/2012           15
## 2  Mega Millions     1/3/2012           36
## 3  Mega Millions     1/3/2012            2
## 4  Mega Millions     1/3/2012            3
## 5  Mega Millions     1/3/2012           22
## 6  Mega Millions   10/14/2022           41
## 7  Mega Millions   10/14/2022           22
## 8  Mega Millions   10/14/2022           26
## 9  Mega Millions   10/14/2022            9
## 10 Mega Millions   10/14/2022           44
```

```
## Powerball Numbers Drawn
pb_numbers_df <- read.csv("Data Samples/Powerball_Numbers_sample.csv")
pb_numbers_df
```

```
##    game_name drawing_date number_drawn
## 1  Powerball     1/4/2012           50
## 2  Powerball     1/4/2012           35
## 3  Powerball     1/4/2012           21
## 4  Powerball     1/4/2012           47
## 5  Powerball     1/4/2012           46
## 6  Powerball   10/12/2022           59
## 7  Powerball   10/12/2022           30
## 8  Powerball   10/12/2022           42
## 9  Powerball   10/12/2022           14
## 10 Powerball   10/12/2022           41
## Mega Million Winning Jackpots
mm_jackpot_df <- read.csv("Data Samples/Megamillions_Jackpot_sample.csv")
mm_jackpot_df
```

```
##        game_name   win_date    jackpot
## 1  Mega Millions  1/24/2012   71000000
## 2  Mega Millions  3/30/2012  640000000
## 3  Mega Millions   5/4/2012  118000000
## 4  Mega Millions  5/15/2012   25000000
## 5  Mega Millions  5/29/2012   32000000
## 6  Mega Millions   7/3/2012   86000000
## 7  Mega Millions  1/28/2022  421000000
## 8  Mega Millions   3/8/2022  126000000
## 9  Mega Millions  4/12/2022  106000000
## 10 Mega Millions  4/15/2022   20000000
## 11 Mega Millions  7/29/2022 1337000000
## 12 Mega Millions 10/14/2022  494000000
```

```
## Powerball Winning Jackpots
pb_jackpot_df <- read.csv("Data Samples/Powerball_Jackpot_sample.csv")
pb_jackpot_df
```

```
##    game_name   win_date   jackpot
## 1  Powerball  12/5/2012  50000000
## 2  Powerball 12/12/2012  50000000
## 3  Powerball 12/19/2012  50000000
## 4  Powerball 12/26/2012  50000000
## 5  Powerball   2/6/2013 217200000
## 6  Powerball  3/23/2013 338300000
## 7  Powerball  10/4/2021 699800000
## 8  Powerball   1/5/2022 632600000
## 9  Powerball  2/14/2022 185300000
## 10 Powerball  4/27/2022 473100000
## 11 Powerball  6/29/2022 366700000
## 12 Powerball   8/3/2022 206900000
```

```
## Mega Million Jackpots by State
mm_state_df <- read.csv("Data Samples/Megamillions_States_with_Jackpots_sample.csv")
mm_state_df
```

```
##       game_name      state number_winners state_total_jackpots
## 1 Mega Millions California             36           7204000000
## 2 Mega Millions   Illinois             14           5328000000
## 3 Mega Millions  Minnesota              1            212000000
## 4 Mega Millions   New York             41           7708000000
## 5 Mega Millions      Texas             13           2218000000
## 6 Mega Millions  Wisconsin              1            238000000
```

```r
## Powerball Jackpots by State
pb_state_df <- read.csv("Data Samples/Powerball_States_with_Jackpots_sample.csv")
pb_state_df
```

```
##   game_name      state number_winners state_total_jackpots
## 1 Powerball California             11           4613800000
## 2 Powerball   Illinois              2             87500000
## 3 Powerball  Minnesota             22           1174000000
## 4 Powerball   New York             12           2263600000
## 5 Powerball      Texas              2            604100000
## 6 Powerball  Wisconsin             19           1919900000
```

Uncovering New Information

By counting and summarizing the information will give insight into the patterns and frequency of the numbers drawn and the size and frequency of jackpots for each of the two lotteries. Comparing and contrasting the 2 lotteries will show what number and winning states they have in common and which lottery has the largest jackpots.

Looking At The Data

This information will be grouped together by lottery (Mega Millions and Powerball).
Individual lottery numbers will be counted. Jackpots will be looked at individually, summed up by year and state. The states where the winners reside are counted. The frequency between winning lotteries will averaged, the minimum and maximum time will be taken and normalized for graphing.

Slicing, Dicing, and Joining The Data

The information will be grouped by lottery (Mega Millions and Powerball). The individual numbers (count), state jackpots totals (sum), lottery winners by state (top 20), and the largest jackpots (top 20) The time between lotteries will be averaged, the minimun and maximum taken. The 2 lotteries will be combined into dataframes by the most common drawn numbers, largest jackpots, and states with the most winners to look at what is in common. New variables will be created for counts and totals and new datafames will be created for records used in the graphs and tables used for analysis.

Summarizing The Information

The information will be counted by the drawn numbers and states where the winners reside. Jackpots will be summarized by year and state and time between winners will be normalized and plotted. Taking the individual lottery results and then comparing the results to each other.

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. pander - Rendering R objects in RMarkdown
8. lubridate - Working with dates
9. Additional packages will be included, if needed.

Plots and Tables

Visualizations

1. Bar Graphs (ggplot/geom_bar) - i. Compare top 10 states with the most winners: x axis - State, y axis - Number of Winners, Legend - Lottery (Mega Millions and Powerball) ii. Compare top 10 most frequent numbers: x axis - Number Drawn, y axis - Number of Times Drawn, Legend - Lottery (Mega Millions and Powerball)
2. Line Graphs (ggplot/geom_line) - i. Total winnings by year for each lottery: x axis - Year, y axis - Total Winnings ii Compare both lotteries total winnings per year: x axis - Year, y axis - Total Winnings, Legend - Lottery (Mega Millions and Powerball)
3. Histograms (ggplot/geom_histogram) - Time Between Winners per Lottery:
   x axis - Time Between Winning Lotteries, y axis - Frequency

Tables - by Lottery (Mega Millions and Powerball)

1. 20 Most common Individual numbers - number drawn and count
2. 20 Largest Jackpots - largest jackpots and drawing dates
3. 10 States with the most winners state - states and number of winners

Questions For Future Steps

1. How do you depict frequency distribution?
2. How can I incorporate machine learning? Is there an appropriate location where machine learning would be meaningful in predicting which lottery numbers to choose?

Incorporate Machine Learning

After reading the upcoming assignments on Machine Learning, if there is an appropriate spot in predicting future lottery numbers I would like to incorporate this information.