**ETL and Powerball**

By

Hunter Cash, Amy Claman, Heather Trinh

**Data Sources: Extraction and Transformation**

We used a flat file from Kaggle.com, pb\_winning\_numbers\_03-18-2017.csv, which represents Powerball winning numbers dating back to 1997. This data was originally acquired from powerball.com. We downloaded the csv file from Kaggle and used pandas to clean up our data primarily by removing data prior to 01/01/2010. Kaggle data included data up until March of 2017. Because most of our other data sources didn’t have data prior to 2010, we decided that 2010 would be our earliest data and March 2017 would be our latest data. This data represents the crux of our chosen topic. It gives us a baseline of winning numbers, the powerball number and a power play option.

We scraped the Texas Lottery website, txlottery.org, to supplement the Kaggle data. The additional data that we collected from this website was the estimated jackpot. We felt it was important for potential future analysis to track the estimated jackpot in addition to the actual jackpot when won. The data scraped from this website required some cleanup on the Draw Date column. The column imported several text strings that were not straight dates. After cleaning that, the Draw Date column needed to be converted to a datetime object. This data was then appended to the original Kaggle data, ultimately adding a new estimated jackpot column.

We also scraped the Portal Seven website for Powerball winner data. We added actual jackpot prize amount, winner name and what state had the winning ticket to our dataset. This dataset, again, needed to be trimmed down to our working dates (01/01/2010 - 03/17/2017). Otherwise, the transformation of this data was straightforward.

Finally, we utilized lottoreport.com to get the ticket sales date for each draw date. We created a flat, excel spreadsheet with this information. The number of tickets sold data needed to be changed to a numeric value.

The data transformations were necessary to create clean data with date columns and numeric values as needed. We plan to join our tables on the powerball draw date, so we wanted to ensure consistency especially with our primary key.

**Data Sources: Loading and Database choice**

Ultimately, we decided to create 2 SQL database tables. One for the basic powerball data including the draw date, the numbers selected, the estimated jackpot for that game and the number of tickets sold. Another table to hold the winners, as winners don’t occur for each and every powerball drawing. The winner table holds our win date, our winner name, the state in which the winner won (bought the ticket) and the jackpot amount.

We felt that because we had varying types of data that could be queried upon in different scenarios, that we should utilize a SQL database. We planned to hold data in several tables which also led to this decision.

The schema of our powerball database is as follows:

Database: lottery

Tables: powerball, winners

Powerball schema: DrawDate, WB1, WB2, WB3, WB4, WB5, PB, PP, estimated\_jackpot, tickets\_sold

Winners schema: WinningDate, JackpotPrize, Winner, State

Powerball.DrawDate and winners.WinningDate have a join relationship.

**Hypothetical Use Cases**

Is there a certain powerball number that comes up more often than others? Are there any trends in frequency of powerball numbers that result in wins?

What states hold the most Powerball winners? What is a typical Powerball winning?

Do chances increase/decrease as the Powerball estimated jackpot amount increase? We can compare tickets sales for varying estimated winnings.