Developing Data Science Skills Using Call of Duty® Data – Overview of Examples

Matt Slifko (mds6457@psu.edu), Department of Statistics
The Pennsylvania State University

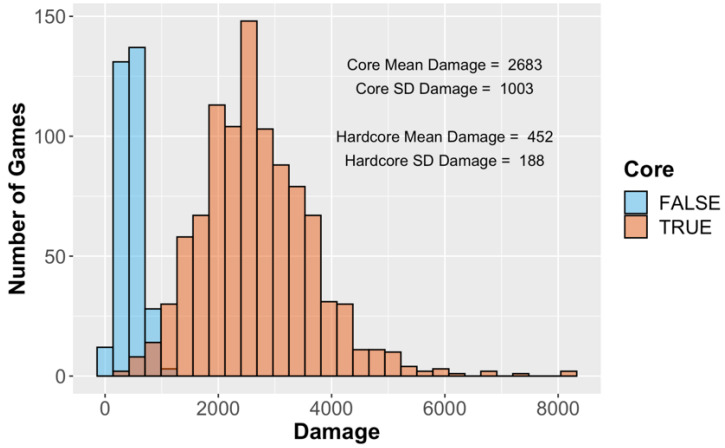
You can find a data dictionary, the 5 data sets, R code, and more thorough discussions by visiting https://github.com/matthewdslifko/CallOfDutyProject. This handout is meant to give a quick preview of ideas spanned by the examples.

Example 1: Exploratory Data Analysis (EDA) Process and Processing Character Data

This example provides an opportunity to discuss the following ideas:

1. data visualization
2. the process of data exploration
3. complementary roles of visualization and numerical summaries
4. annotating plots with text
5. character string processing and new variable creation
6. multivariable thinking

Students are given background information about the variable Damage and asked to explore its distribution. Upon learning that the distribution is bimodal, students are asked to consider other variables that might affect the distribution. Providing students with additional information about the GameType variable leads to the plot shown below.



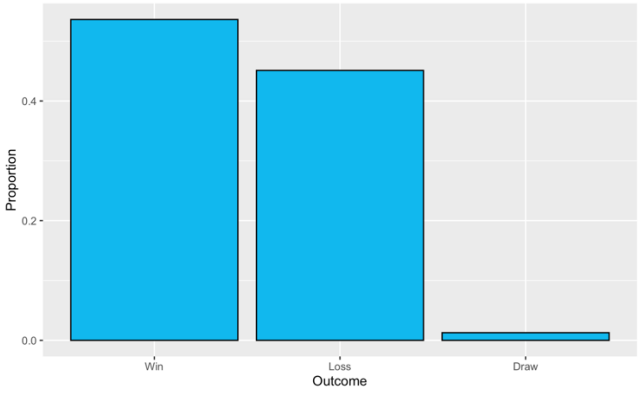
Example 2: Processing Character Data, Importance of Checking Calculations, and Warnings

This example provides an opportunity to discuss the following ideas:

1. character string processing and new variable creation
2. if/else with more than 2 conditions
3. importance of manually checking calculations
4. data visualization and reordering factors for visualization
5. exploring WARNING messages
6. BONUS: discuss whether there is a difference between proportions of wins and losses OR is the difference just random variation?

Students are given background information about the character variable Result (e.g., “100-97”) and asked to convert it to usable information to answer the research question: Is the player's team more likely to win, lose, or draw? A common mistake is used to illustrate the importance of checking calculations. (The table shows incorrect PlayerResult values due to comparing character.) After correcting the mistake, the proportion of wins, losses, and draws is visualized. Applying the code to the more complex dataset (CODGames2) leads to warnings that we did not previously encounter due to differences in format of some Result values.

	Result <chr>	PlayerScore <chr>	OpponentScore <chr>	PlayerResult <chr>
1	100-97	100	97	Loss
2	76-89	76	89	Loss
3	100-92	100	92	Loss
4	80-100	80	100	Win



Example 3: Combining Data Tables and Data Quality Issues

This example provides an opportunity to discuss the following ideas:

1. Joining tables (with no common variable names)
2. Filtering observations according to a condition
3. Data visualization
4. Adding means to boxplot
5. Categories with small *n*
6. Handling typographical errors in data entry

Students are given background information about variables Score, FullPartial, and PrimaryWeapon. They are asked to answer the research question: For complete matches in CODGames1, does the player's score depend on the weapon class used? Since the weapon class labels are contained in a different dataset, students should combine tables to get the additional information. Side-by-side boxplots, along with the means, are used to investigate. Applying the code to the more complex dataset (CODGames2) leads to issues with typographical errors (e.g., AK 47 vs AK-47) and prompts discussion as to how to handle these challenges

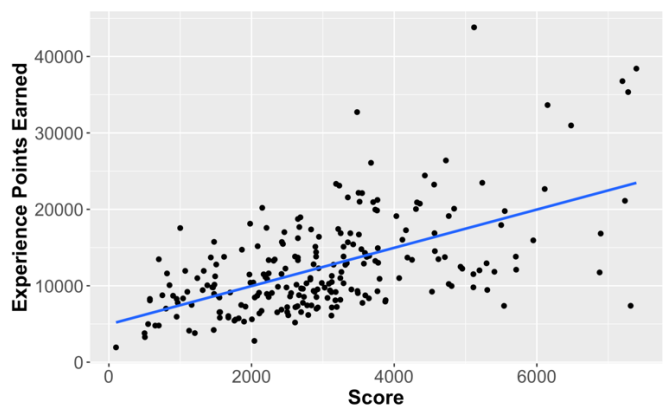

```
##    PrimaryWeapon
## 1          AK 47
## 2          AK-47
## 3         AK-47u
## 4            AUG
## 5   Combat Knife
## 6         DMR 14
## 7        FARA-83
## 8         FFAR 1
## 9          Gallo
## 10         Groza
```

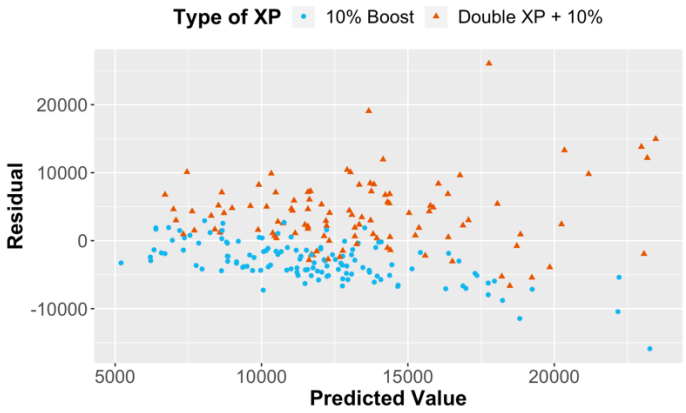
Example 4: Regression

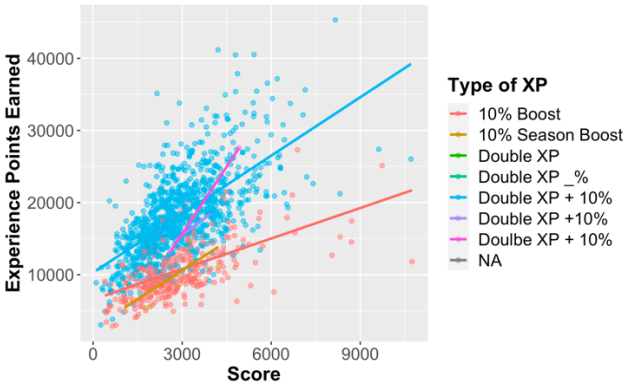
This example provides an opportunity to discuss the following ideas:

1. linear regression and checking assumptions
2. residuals
3. multivariable thinking and lurking variables
4. indicator variables
5. interaction
6. comparing models
7. outliers
8. statistics as an investigative process
9. implementation issues on new data

Students are given background information about eXperience Points (XP) in the game and are asked to answer the research question: What is the relationship between the player's score and the XP earned? Initial exploration suggests a linear association may be reasonable. Students are asked to discuss other variables that could affect the relationship. If no one mentions it, students are given information about another variable called XPType that indicates when a player elected to use a Double XP token. Incorporating this extra information into a plot of residuals vs predicted values suggests some patterns. In addition to systematically over/under predicting, there may be issues with increasing variance and some potential outliers to discuss. Updating the initial scatterplot motivates a conversation about adding an indicator variable and interactions terms, but also comparing models. Finally, applying the code to the more complex dataset (CODGames2) leads to a variety of challenges with typos and missing data.







Details and further explanations regarding all examples may be found in the Examples folder on the GitHub page or by emailing Matt (mds6457@psu.edu).