1. What are the names of the people in your group?

Tim Abramov

1. What research topic or question are you interested in?

Analyzing and visualizing data about video games.

1. What type of project will this be? Will you be analyzing a data set, creating a visual narrative of data, using R to create simulations, etc.?

I plan to do most of data cleanup, analysis, and visualization using R.

1. Do you plan to collect your own data, or do you hope to use a data set that already exists (or is simulated)? If you plan to use a pre-existing data set, have you already found it?

Dataset was provided to me by a Professor.

1. Describe your likely data set.
   1. List the unit of study in your data. For example, if your data is in a spreadsheet, does each row represent one person (or one movie, or one country, etc.)?

Each row in the dataset represents an individual title in Steam store. Some are not actually games, those will need to be removed for my research.

* 1. How big is the data? How many observations are there?

Dataset consists of 13,357 records. It represents a full snapshot of titles available on Steam at that time (December 12, 2016).

* 1. List the variables you will measure or obtain data about, including units when appropriate.

Dataset consists of merged data from two main resources:

1. Steam – biggest game-distributing platform for Windows, Linux, and Mac OS (for PC)

2. Steam Spy – service that collects and stores number of games purchased on Steam, also provides estimates on how many of people who purchased a specific game are players.

Variables:

Metacritic rating – Metacritic provides a rating service for video games, movies, television, shows, and music. It uses a proprietary weighted average of various professional critics and publications, scaled from 1 to 100, and is one of the most frequently used metrics for video game quality.

Category and Genre variables – represent different classifications (for instance “Action” for genre, and “Single Player” for category) of the game, set by game developers. A game may have multiple categories and genres that it’s in.

Owners count and player estimate – data from Steam Spy (discussed earlier).

Price – price of a game at the time the data was gathered. No averaging was done on this variable.

Detailed description – full description of the game on the store page.

Screenshot count – number of “pictures” in games description.

Supported languages – a list of supported languages. There is also an indication on what languages are partially supported, might be of use.

Recommendation count – count of user reviews on the game. What is important is that this number doesn’t differentiates positive and negative reviews, it is all-in-one.

Some other variables: Name, Release date, Controller support, Is free.

1. Statistical questions: List several specific questions that can be answered using the variables in your data set.

* Not all games received a Metacritic score, I would like to find out what differentiates games that got this score from those that didn’t. My hypothesis would be that games that were more “public” (reached bigger audience) most likely have gotten one, but there might be some other reasons that I can’t really see without investigating this question.
* Another question I would like to investigate is: does length of game description or number of game screenshots have a measurable effect on sales of the game. I would expect that there is no distinct pattern between those, but I have a hunch that there might be a range of numbers that does differ from majority of games in their game description length and/or number of screenshots that also have significantly different sale figures.
* Is there an association between how many copies were sold and price of a game? This is interesting question to me since, on one side a AAA game is sold usually for highest “acceptable” price (usually if game costs 40-60 a pretty high chance it was created by AAA studio(s)) and also needs to sell many copies to make some profit; and on the other side even if game cost is pretty low ($5 for instance) it still can sell a lot of copies. So, the question can show, in some respect, what kind of studios (big vs small) are more successful in game development.
* One more question I would like to investigate is: does relative success of a game mean that most likely it’s localized in most popular languages and are there any exceptions for this assumption?

1. What is your background reading plan? Give a short list of topics you need to research for your project.

* Need to find out how to clean/pre-process the data in R and how it’s usually done
* Get myself acknowledged with the research that was already done in this area
* Most likely will need to research some analysis techniques to help me answer my questions, since some of the analysis might not be something that we learned in class (maybe to find a multivariate relationship for instance)