**Toys on Amazon Case Study**

**Case Study and Expectations**

Isn’t analytics just way cooler when you get to interact and play with “real and dirty” data instead of “fake and clean” data? Well, that’s exactly what we will be doing in this case study! It will be a good way to challenge and prepare you for the marketplace – note that this particular case is written by your instructor, and is designed to follow the same format and structure of several case studies utilized by companies in job interviews.

In addition, this case will give you the chance to show the instructor what you can do with R and RStudio, after the completion of several lectures in basic R over the summer and five required DataCamp practice courses. After receiving the case instructions you’re putting hands-to-keyboard and have until July 31 at 11:55pm to submit the first part of the case (**CS1**) which focuses on exploration and wrangling, and until August 19 at 11:55pm to submit the second and final part of the case (**CS2**) which builds off of the first part and focuses on programming and modeling. Per the submission instructions, please include:

**For case study part 1 (CS1):**

* Working code with documentation (Rmd and HTML, pdf, or Word – HTML is preferred)
* Documentation of metadata and data quality
  + **Things to consider**: Is the data tidy? Are key variables included? Is missing data a problem? Is the collection of additional data necessary? Is there anything else I should know about, in terms of data quality?
* Visualizations of key insights (to be included in HTML)
* Notes and comments on transformations and wrangling (what were the major things that you did to the data to get it in an appropriate format for visualization and analysis?)

**For case study part 2 (CS2):**

Everything that was included in CS1 plus:

* Any changes based on the instructor feedback and / or further visualizations, transformations, and wrangling that you conducted
* Working code with documentation illustrating (if deemed necessary and useful) basic programming skills (e.g., specific functions and / or loops)
* Results including the interpretation of linear models (only in the case that a statistical model is built)

Ready to show off your analytic skills? Let’s go!

**Task and Instructions**

**Task**

You are a data analyst at Amazon, and your supervisor just provided some data on toy products for you to examine. In this task, there is no “business problem” to solve. Rather, what is wanted from you is a thorough analysis of the data, and some key insights about toys for sale on Amazon. As you begin to explore the data, think about the following questions:

* What kinds of toys are most popular on Amazon?
* Which brands are the most dominant?
* Can you break down and predict which kinds of toys and brands get the best reviews?

These are by no means the only questions you should think about. Your supervisor suggested these high level questions as a way to get you started. If other more interesting types of questions arise as you analyze the data, feel free to pursue them and please tell your supervisor why you decided to do that! In other words, explain the logic behind your analysis, preferably in a story-telling way.

Ideally, a data product will be built out of your analysis, but this is not a necessity nor an expectation. Your supervisor understands you are just getting started at Amazon and so just do your best! The overarching goal is to be able to provide insights that will help Amazon understand which kinds of toys from which types of brands receive the best kind of reviews.

You will be looking at publicly available pre-crawled data from Amazon. If you deem it necessary, feel free to collect additional data on your own. Keep in mind that you are NOT allowed to purchase any datasets. However, it is totally okay to scrape additional data or just download other datasets publicly available that you may find useful to combine with what you already have.

This dataset has the following fields/variables:

* **product\_name**
* **manufacturer** - The item manufacturer, as reported on Amazon. Some common "manufacturers", like Disney, actually outsource their assembly line.
* **price**
* **number\_available\_in\_stock**
* **number\_of\_reviews**
* **number\_of\_answered\_questions** - Amazon includes a Question and Answer service on all or most of its products. This field is a count of how many questions that were asked actually got answered.
* **average\_review\_rating**
* **amazon\_category\_and\_sub\_category** - A tree-based, &gt;&gt;-delimited categorization for the item in question.
* **customers\_who\_bought\_this\_item\_also\_bought** - References to other items that similar users bought. This is a recommendation engine component that played a big role in making Amazon popular initially.
* **description**
* **product\_information**
* **product\_description**
* **items\_customers\_buy\_after\_viewing\_this\_item**
* **customer\_questions\_and\_answers** - A string entry with all of the product's JSON question and answer pairs.
* **customer\_reviews** - A string entry with all of the product's JSON reviews.
* **sellers** - A string entry with all of the product's JSON seller information (many products on Amazon are sold by third parties).

**Instructions**

As you start the job, realize that this is real-world, imperfect data. I recommend planning about 8-12 hours to complete each part of the assignment (i.e., 8-12 hours to complete CS1 and then again 8-12 hours to complete CS2), but it’s not timed, and you are judged on the quality of the work submitted. If you are struggling to discover what the “right” answer is, realize that usually in cases like this, there isn’t a “right” answer. Your job is to use your best judgment, make an assumption where necessary (document the assumption), and keep on going.

Overall, I am asking you to show me your R skills along with your big picture analytical skills in three areas at a basic level for each part of the assignment (CS1 and CS2), and then, in the last step, tell me what you would do next if given another chance to work on this case.

**Case Study Part 1 (CS1):**

1. **Data Quality** – bad data is a problem and you should report it when you see it
   1. Explore and understand the data but do not lose sight of the big picture
   2. Report two to three data quality insights based on your analysis
   3. Create metadata for new variables (or existing ones if necessary)
2. **Data Wrangling** – tidy the data, if necessary
   1. The dataset may have different types of variables (e.g., customer review and review date) grouped in one variable (e.g., customer review and date) – if that is the case clean the data by tidying it up!
3. **Use Visualizations for Story-Telling** – *Charts and plots must be generated from your code; not from produced in external standalone software like Excel or Tableau*
   1. Visualize information about products, brands, and reviews (what do you see?)
   2. Summarize your key insights and conclusions based on your visualizations and other descriptive analysis (e.g., summary tables)
4. **What’s Next** – I recognize that 8-12 hours isn’t a lot of time to work on the first part of a case like this. You will probably come up with a number of cool ideas from an analytical or visualization perspectives that you would like to work on but will not have the time to do so. Tell me (but don’t do any work) what you would/could do next to inform a better insight to Amazon.

**Case Study Part 2 (CS2):**

1. **Feedback** – listening to feedback is important but even more important is to address feedback when you receive it
   1. Address any comments made by the instructor in the submission
   2. Be clear about any changes made to the data or visualizations since the first submission, either based on instructor feedback or your own analysis of the data
2. **Programming** – write functions and use loops, *only and if* they are helpful
   1. Knowing how to write functions and perform loops is valuable but be mindful about them. If you don’t see a need for a function do not write one. The same applies for loops. Although you are encouraged to explore and test your programming skills in this part of the case, do not feel like you are obligated to do so. In the case that you decide to write a simple function or loop be thoughtful about it. Also consider using iteration packages, such as purrr
3. **Modeling** – *A basic model should be generated from your R code; not from SAS*
   1. If you think it would be useful to create a statistical model for this case study, do so. Please note that actually building the model is not necessary, especially if you find the data is not suitable for it. Whatever you decide to do in regard to modeling, be explicit about your decision and explain the logic behind it
4. **What’s Next** – Again, I recognize that 8-12 hours isn’t a lot of time. If you have any new ideas after having the chance to work on this dataset one more time, let me know. What would you do to improve the work?

* I will look for 4 main aspects to your submissions:
  + R skills (**CS1 and CS2**) – Is your code reusable or hardcoded? Is the report / output file (HTML) well-structured and clean? Did you provide good comments in the code?
  + Exploration and Wrangling (**CS1 and CS2**) – Did you explore a variety of ggplot2 functions and create good visualizations? Were you able to transform the data in useful ways? Did you identify any data outliers and data quality issues?
  + Programming and Modeling (**CS2**) – If you decided to write a basic function or loop, was it useful? Were you able to accurately build a linear model that provided you with useful insights? If you decided to not build a model, did you think carefully about the data and provided great ideas for model building in the future?
  + Business Solution (**CS1 and CS2**) – What kind of story are you telling? Do your visualizations/models support your story? Did you provide good insight and analysis? What recommendations do you have and are they feasible and concrete?

**Data and Tools**

*Solutions that require purchase and use of a software license that is not R and RStudio or that involve purchased access to data will not be accepted.*

**Data**

Downloading the data is simple:

1. Please access Sakai > BAN 6003 A S1 2019 > Resources > data.
2. Once in the folder, download and extract the ZIP file (Toys\_Amazon.zip).