Video Games Project

Client is a small games developer who is planning next release and want to use Data to decide what kind of game to make.

They have 2 datasets with historical information on top selling games ***OF ALL TIME*** and want to understand what TYPES of games sell a lot of copies. In particular they are looking for the direction to take their company in.

Day 1 (Thursday)

Initial thoughts:

They want me to use the Data which is historical which raises some questions:

How old is it? Are there any repetitions? Have trends moved on? Can you determine that from the data? Can you compare games that are on both files and determine if there is a trend upwards or downwards towards that type of game? Is there a price point which determines popularity? Can it be linked to reviews? Are reviews on games fairly consistent? Do people base their purchase in a review? Do any products defy reviews? Can I build a model that at least identifies the relationship between the different variables that might determine or indicate volume of sales? Is there a relationship between how good a game is and sales or is it something else? Does it help if there have been series predecessors regardless of the review, essentially are you buying a franchise? Is there room at a later date to include a dashboard for the future that shows the impact of different variables on a video games sales model? Can it later be predictive or what else does it require? How much cleaning does it require? Do I need a few cleaned scripts in order to analyse different things particularly comparing historical information? Seeing as they are small gaming company, are they looking to develop the most sophisticated game or the one with greater potential as realistically they don’t have the resource for scalability yet to grow too fast? Can I determine or predict potential for a series based on first release? Product maturity lifecycle is the ideal model but it comes with inherent risks.

Tasks:

1. Create a private repo in GitHub and load initial stuff that will eventually include files such as: raw data, cleaning scripts, clean data, data analysis, picture files with graphs and other visuals, a readme file, further considerations, a markdown presentation and a PowerPoint or pdf presentation.
2. Go over my CodeClan texts on dplyr and cleaning to make sure I am asking the right questions of the data. Think about how I want it to look in order to manipulate it best. I am particularly thinking of pivot long for example but there are others: mutate being another.

Day 2 (Friday)

A shorter day due to family commitments

Tasks:

1. Determine existing data variables and what new ones I need (mutate).
2. Go over the business question again and think about what kind of graphs I would like to create
3. Mock up some wire framed potential graphs and see if they answer business questions
4. Start noting the risks to the client regarding growth and building a top selling product.

Day 3: (Sunday, I looked after my son on Saturday)

Tasks:

1. None achieved as family commitments meant I was unable to do anything.

Day 4: Monday

Tasks:

1. Plot some graphs
2. Various meetings, look after my son with my wife at work

Day 4: Tuesday

Tasks:

1. Determine existing data variables and what new ones I need (mutate).
2. Go over the business question again and think about what kind of graphs I would like to create
3. Mock up some wire framed potential graphs and see if they answer business questions
4. Start noting the risks to the client regarding growth and building a top selling product.
5. Categorise MVP and extensions

So I have determined that to complete my MVP, I must demonstrate the following in my presentation.

1. A quick look at the underlying data: outliers, missing data, high level overview

What measurements are the variables in? What is the code for the ratings? Critic\_count and critic\_score

1. Demonstrate my approach about dealing with the data e.g.

Joining tables, further cleaning (explained plainly), geographic market comparisons, historical trends using average sales per year over a period, missing values (and if there is a relationship between them and anything) and how I have dealt with them), don’t bother explaining pivot longer

1. Look at the top 50 sellers or top something sellers and examine a relationship model
2. Look at the which markets contribute most to the overall global sales
3. Determine if there is a relationship between sales and having a rating

These are extensions:

1. Perhaps run a k mean cluster exploratory analysis on this
2. Try grouping games together with sequels (and prequels) to see if there is a relationship between sales and that (this part could go on for a while)
3. Is there relationship between genre and sales and do certain markets prefer certain genres? Have any genres increased in popularity over time
4. Is the platform important, is the publisher? Have some been better over time or done less well?

Day 5 (panic stations) Wednesday

I feel like I haven’t really achieved anything yet so I am making a roadmap for the rest of the day with my presentation due tomorrow morning.

|  |  |
| --- | --- |
| Look at headings in both datasets and see what they share | Done |
| Look at missing values in both datasets | Done |
| Clean the names to lower case, clean everything else | Done |
| Join the existing datasets | failed |
| Do I need to pivot longer on sales volumes? Try it and see. | Do I need to pivot longer on sales volumes? Try it and see. |
| Text mine on the common names |  |
| Consider grouping by franchise to form a separate clean dataset |  |
| Filter on all unscored games. Is there are an age cut off? |  |
| For 2019 missing scores, is there a way you can calculate based on other variables? And for 2016? |  |
| Try calculating missing scores | Tried |
| Create a dataset missing out lines without scores |  |
| Create csv’s of all new datasets | done |
| Check whether I can plot ggplot with different geospatial sales volumes vs genres console vs is publisher important use some stats vs esrb rating. save as ggplot |  |
| Create average sales volumes per year since creation |  |
| In analysis section start plotting graphs |  |
| Create any relevant tables |  |
| Begin powerpoint presentation | begin |
| Try a gg model thing |  |
| Think of an opening question/ theme |  |
| Start filling stuff in |  |
| Write a script of notes for powerpoint |  |
| Practise a few times |  |
| Source pictures |  |
| Any surrounding contextual analysis? |  |
| Draw some conclusions |  |

**Times for tomorrow @channel:**  
Masood: 10:30 - 10:45 (+5 min questions)  
Sandra: 10:50 - 11.05 (+5 min questions)  
5 min break  
Elly: 11:15 - 11:30 (+5 min questions)  
Ric: 11:35 - 11:50 (+5 min questions)  
5 min break  
Nehali: 12:00 - 12:15 (+ 5 min questions)  
Chris: 12:20 - 12:35 (+ 5 min questions)  
Amit: 12:40 - 12:55 (+ 5 min questions)