Raunak Sarbajna and Christoph F. Eick

COSC 3337 *“Data Science I”* Fall 2023

Problem Set1

Individual Tasks[[1]](#footnote-1)

Second Draft

Last Updated: September 5, 4pm

Task1: Exploratory Data Analysis for a Video Sales Dataset



Remark: This is a first, somewhat preliminary specification of Task1; there might be still ninor changes, corrections and additions in the next 5 days. More details submission instructions should also be available by Sept. 18 or earlier.

Task1 Due: Saturday, Sept. 23, 11:59p (electronic Submission)

Tentative weight: about 23-30% of the points allocated with the course’s ProblemSet tasks.

Responsible TA: Raunak

**Learning Objectives**:

1. Learn how to manage and preprocess datasets and how to compute basic statistics and to create basic data visualizations (using R/Python or other tools)
2. Learn how to interpret popular displays, such as histograms, scatter plots, box plots, density plots,…
3. Get some practical experience in exploratory data analysis
4. Learn how to create background knowledge for a dataset
5. Learn to distinguish expected from unexpected results in data analysis and data mining—in general, this task is quite challenging, as it requires background knowledge with respect to the employed data mining technique, and also practical experience.

**The Video Game Sales Dataset is a non-exhaustive collection of the sales performance of videogames from various gaming platforms, accurate up to the year 2016. The data was collated from a web scrape of VGChartz Video Games Sales and review scores from Metacritic.** The goal of this project is to perform exploratory data analysis for the [*Processed Video Game Sales Dataset*](https://gist.githubusercontent.com/RaunakDune/4381ef89d4dc3c66459b1fd80bae254d/raw/b96afe240cbb3efa08935e3328ac786ec962cd9e/Videogame_Sales_2016_Processed.csv)which is a modification of the [original Video Game Sales Dataset](https://www.kaggle.com/datasets/sidtwr/videogames-sales-dataset?select=Video_Games_Sales_as_at_22_Dec_2016.csv). The original is a (15+1)D dataset and the *Processed Video Games Dataset* is a (15+2)D dataset with several attributes factorized, rows with empyt columns dropped and a nominal Global\_sales\_target attribute added; the attributes of this dataset and their domains are listed below:

* + - 1. Name,
      2. Platform,
      3. Year\_of\_Release {1985,…,2016},
      4. Genre,
      5. Genre\_factor {0,…,11}
      6. Publisher,
      7. NA\_Sales [0, 41.36],
      8. EU\_Sales [0, 28.96],
      9. JP\_Sales [0, 6.5],
      10. Global\_Sales [0.01, 82.53,],
      11. Critic\_Score [13, 98],
      12. Critic\_Count [3, 113],
      13. User\_Score [5, 96],
      14. User\_Count [4, 10665],
      15. Developer,
      16. Rating,
      17. GS\_Category [high, medium, low]

The first 3 examples of the dataset are listed below:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Name | Platform | Year\_of\_Release | Genre | Genre\_factor | Publisher | NA\_Sales | EU\_Sales |
| New Super Mario Bros. | DS | 2006 | Platform | 2 | Nintendo | 11.28 | 9.14 |
| Animal Crossing: Wild World | DS | 2005 | Simulation | 8 | Nintendo | 2.5 | 3.45 |
| Brain Age 2: More Training in Minutes a Day | DS | 2005 | Puzzle | 5 | Nintendo | 3.43 | 5.35 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| JP\_Sales | Global\_Sales | Critic\_Score | Critic\_Count | User\_Score | User\_Count | Developer | Rating | GS\_Category |
| 6.5 | 29.8 | 89 | 65 | 85 | 431 | Nintendo | E | high |
| 5.33 | 12.13 | 86 | 57 | 87 | 242 | Nintendo | E | high |
| 5.32 | 15.29 | 77 | 37 | 71 | 19 | Nintendo | E | high |

The values of the class attribute GS\_Category have been computed from the Global\_sales attribute as follows: [0.01-0.1)🡪low, [0.1-1.0}🡪medium, [1.0-82.53}🡪high; In general, we are interested in predicting Attributes 10 and 17 using the other attributes; that is, we like to predict which of the factors amongst genre/platform/publishers/review scores contribute most to the overall sales of a game. Another subject we are interested in is finding relationships between the attributes in the dataset, and to understand what factors influences successful games the most.

Task1 Subtasks:

Apply the following exploratory data analysis techniques **using R [Preferred]/Python** to your dataset:

1. Use the *Processed Video Game* dataset created by the TA or clean up the raw dataset yourself!
2. Compute the covariance matrix for each pair of the following attributes:

Critic\_Score

User\_Score

NA\_Sales

JP\_Sales

**Global\_Sales**

Next, compute the correlations for each of the pairs of attributes. Interpret the statistical findings! **3 points**

1. Create scatter plot for the attributes Critic\_score/Critic\_Count and User\_Score/Global\_Sales Interpret the scatter plot**! 3 points**
2. Create histograms for Platform and User\_Count attributes for High, Medium and Low GS\_category classes; interpret the obtained histograms. **6 points**
3. Create box plots for the User\_Score/Critic\_Score attributes for the instances of the 3 GS\_category class— low/medium/high — and a fourth box plot for all instances in the dataset. Interpret and compare the box plots for each attribute! **4 points**
4. Create supervised scatter plots for the following 3 pairs of attributes using the GS\_category as a class variable: Critic Score/NA\_Sales, NA\_Sales/User Score and Critic Score/User Score. Use different colors for the class variable. Interpret the obtained plots; in particular, address what can be said about the difficulty in predicting the Global Sales Target and the distribution of the instances of the three classes. **6 points**
5. Create 2 density plots for the instances of the 3 GS\_category classes in the Critic\_Score/User\_score space. Compare the density plots! **6 points**
6. Create a table which reports the frequency of associations of the 12 genres with the three classes of the GS\_Category attribute. Create histograms for the Global\_Sales attribute for the instances of each of the 12 genres. Interpret the table and the histograms you created. **8 points**
7. Create a new dataset *Z*-*Processed Video Games* from the *Processed Video Games* dataset by transforming the Year, Critic\_Score, Critic\_Count, User\_Score, User\_Count attributes into z-scores. Fit a linear model that predicts the values of the Global\_Sales attribute using the 5 z-scored, continuous attributes as the independent variables. Report the R2 of the linear model and the coefficients of each attribute in the obtained regression function. What do the obtained coefficients tell you about the importance of each attribute for predicting a successful game? **6points**
8. Create 3 decision tree models with 20 or less nodes for the dataset (both intermediate and leaf nodes count; do not submit models with more than 20 nodes!); use the GS\_Category attribute as the class variable, and use the remaining attributes of the dataset excluding attribute Global\_Sales of the dataset when building the decision tree model. Explain how the 3 decision tree models were obtained! Report the training accuracy and the testing accuracy of the submitted decision trees. Interpret the learnt decision tree. What does it tell you about the importance of the chosen attributes for the classification problem? **11 points**
9. Write a conclusion (at most 13 sentences!) summarizing the most important findings of this task; in particular, address the findings obtained related to predicting a successful game (high global sales) using attributes 1-14. If possible, write about which attributes seem useful for predicting high video game sales and what you as an individual can learn from this dataset! **6 points (and up to 4 extra points)**

Remark: About 30-40% of the Task1 points will be allocated to interpreting statistical findings and visualizations!

**Submission Guidelines Task1[[2]](#footnote-2)**: When you submit your task 1 for problem set 1, you should submit a compressed (zipped) folder that contains a word file/pdf that displays your graphs and your interpretations. Each interpretation should use complete sentences to describe your findings. Also in the folder, you should include all files used to complete your tasks, such as your R or python files. If you have doubts what to submit send Raunak an e-mail.

1. Collaboration with other students in the course is not allowed! [↑](#footnote-ref-1)
2. More detailed submission instructions for Task1 will be added to this specification by Sept. 18, 2022 the latest. [↑](#footnote-ref-2)