Anomaly Detection Techniques to Find Influential Users

Shawn L. Watts

Western Governors University

Table of Contents

[A. Project Highlights 4](#_Toc1868160802)

[B. Project Execution 5](#_Toc2011082291)

[C. Data Collection Process 5](#_Toc1191971028)

[C.1 Advantages and Limitations of Data Set 6](#_Toc10980602)

[D. Data Extraction and Preparation 6](#_Toc512824111)

[E. Data Analysis Process 6](#_Toc1571068809)

[E.1 Data Analysis Methods 6](#_Toc1409685802)

[E.2 Advantages and Limitations of Tools and Techniques 7](#_Toc737998485)

[E.3 Application of Analytical Methods 7](#_Toc825242289)

[F Data Analysis Results 7](#_Toc1939531306)

[F.1 Statistical Significance 7](#_Toc109772385)

[F.2 Practical Significance 8](#_Toc1419440648)

[F.3 Overall Success 8](#_Toc1210557849)

[G. Conclusion 9](#_Toc1495356842)

[G.1 Summary of Conclusions 9](#_Toc613065263)

[G.2 Effective Storytelling 9](#_Toc16322062)

[G.3 Recommended Courses of Action 9](#_Toc660897025)

[H Panopto Presentation 9](#_Toc2048387211)

[References 12](#_Toc1410715514)

[Appendix A 12](#_Toc1695981861)

[Title of Appendix 13](#_Toc1655579738)

[Appendix B 13](#_Toc1686431257)

[Title of Appendix 14](#_Toc334098601)

[Appendix C 14](#_Toc1371871733)

[Title of Appendix 15](#_Toc38061623)

[Appendix D 15](#_Toc724305577)

[Title of Appendix 16](#_Toc1812292947)

# A. Project Highlights

Which users are influencing others the most with their reviews and recommendations of video games on the Steam platform?

In this fictitious scenario based on real a real company, Valve Corporation and Steam gaming platform need help discovering user patterns in their recommendation and review systems data to better understand who their most influential users are. They have asked for a compact list of the most influential users for study by their selves and other game development companies that use the Steam platform to sell their games. The hope is that this will help game companies develop better games and marketing strategies to increase sales. Since this data is not known currently, they have asked to do some discovery and provide them a list of users with proper rational and methods. The value of this study could improve success for all parties.

The scope of the project will be limited to acquiring de-identified user data with minimal features. The only hard deliverable here is the final list of users in Excel format, no more than one thousand. If the client is not satisfied, a review process will ensue and a new iteration will begin. Bringing in identifiable data and doing further analysis will be the responsibility of the client. These attributes are not provided and thus would be impossible to analyze.

Rapid Application Development was used as the project management strategy. Data was collected that was needed to complete the project. A CSV file was downloaded with helpof a browser from the Kaggle website. A WSL Ubuntu environment was used to isolate our workspace in Windows 11. Miniconda was utilized to create a python and Jupyter Notebook environment for performing and displaying the analysis. Standard data science python packages were used including, Pandas, Polars, Seaborn, Matplotlib, Scikit-Learn, and full list below:

* **Windows 11**: This is the main operating system of my laptop.
* **Windows Subsystem for Linux (WSL) with Ubuntu 20.04.6 LTS**: I want to isolate my process from my main OS. I will run the project in a Linux virtual machine.
* **Miniconda**: This is a free installer for a minimal version of conda. This is a package and environment management system. We will want to create our conda environment separate from any others. We will also need it to install Jupyter Notebooks and other python packages.
* **Python 3.10.14**: This is the python version we will use for this project.
* **GitHub**: GitHub is where we will keep all our work. We need some place to store our code and changes and be able to roll back to a previous version if needed.
* **Jupyter Notebook**: This tool helps present your data process, visualizations, and run most any python script needed for this project. Our EDA and final file creation will be all here.
* **Matplotlib**: For visualization of data.
* **NumPy**: It has many mathematical functions that will be useful for data analytics.
* **Pandas**: It helps read and store data in a structure that is ready to be consumed by python.
* **Scikit-learn**: This is a popular machine learning library for python, and we will use it for PCA, t-SNE, and Isolation Forests.
* **Seaborn**: This extends Matplotlib with a richer feature set and a much better presentation.
* **SciPy**: This has many uses. Among them is the test we plan to use to compare to skewed distributions. Another is the bootstrapping we will use to simulate normal distributions for our skewed data.
* **Polars**: A faster implementation and alternative for large data sets. It will be used to import the raw data.
* **OpenPyXL**: Needed to write our file to Excel. It extends Pandas and adds an Excel method to export the data frame.

Raw data was explored and assessed to understand it and what was needed to be done. The data was summarized at the user level. Cleaning was completed after assessment. A Single PCA component was created over the relevant features and the right five percent of users were sampled based on the PCA component.

An Isolation Forest model was created with the appropriate contamination rate set and features to obtain enough users for our final sample and label them as outliers or not.

Finally, the Mann-Whitney U Test to compare the outliers with the non-outliers was applied. A p-value of 0.00 was found to be less than 0.05 in support of the alternative hypothesis. After assessing visually each feature independently using the bootsrap methodoly, its was concluded that the final list was of high quality.

# B. Project Execution

**Rubric B:** The summary accurately addresses how the execution of *each* element differed from the plan developed in part B of Task 2.

This part summarizes the project’s execution and any variances from the execution plan provided in Task 2. “Each” element for which to summarize variances are provided in the task directions as follows:

* The project plan (see Task 2 section B1).
* The project planning methodology (see Task 2 section B3).
* Project timeline and milestones (see Task 2 section B4).

If the element did not vary from Task 2, explain that there were no variances and why. You can reuse content from Task 2 as needed. You are allowed and encouraged to reuse any of your work. The similarity check will not check for plagiarism against your content.

# C. Data Collection Process

**Rubric C:** The discussion of the data selection and collection process includes specific details for *each*of the given elements.

This part summarizes the data selection, collection, processing, and governance issues. The “given” elements to summarize are provided in the task directions as follows:

* How the data selection and collection differed from your plan (see Task 2 section D3).
* How you handled any obstacles encountered while collecting your data (see Task 2 section D4).
* How you handled any unplanned data governance issues (see Task 2 section D5).

If the element did not vary from Task 2 or you did not encounter any unplanned issues, explain why. You can reuse content from Task 2 as needed. You are allowed and encouraged to reuse any of your work. The similarity check will not check for plagiarism against your content.

## C.1 Advantages and Limitations of Data Set

**Rubric C1:** The discussion addresses *both*the advantages and limitations of the data set, including specific examples of *both*. *All* discussed advantages and limitations apply to the data set that was used.

* Discuss at least one advantage of your data set, and provide at least one example of an advantage.
* Discuss at least one disadvantage of your data set, and provide at least one example of an advantage.

# D. Data Extraction and Preparation

**Rubric D:** The submission explains *both* the data extraction and data preparation processes, including details on why the processes were appropriate for the data. The explanation includes the tools and techniques that were used for *both*processes.

Discuss the data extraction and preparation process and their appropriateness (see Task 2 part D), and discuss the tools used for both extractions. If little or no extraction or processing was necessary, explain why.

# E. Data Analysis Process

## E.1 Data Analysis Methods

**Rubric E1:** The submission includes a thorough and accurate description of data analysis methods that are appropriate for the intended goals of the project.

Name and describe each method used to analyze the data. Include the method(s) used to support your hypothesis in Task 3 part F (below; also see Task 2 part C). Explain why each method was appropriate.

## E.2 Advantages and Limitations of Tools and Techniques

**Rubric E2:** The submission accurately discusses the advantages and the limitations of the tools and techniques used to analyze the data.

Discuss at least one advantage and one limitation of each tool used during the analysis (see Task 2 section C3).

## E.3 Application of Analytical Methods

**Rubric E3:** The submission includes a thorough step-by-step explanation of how the analytical methods were applied to the data and how *all* assumptions or requirements were verified.

Describe the steps used to complete each method used for the data analysis in part F of Task 3 (see Task 2 section C2). Describe the requirements for each method and how they were verified.

# F Data Analysis Results

## F.1 Statistical Significance

**Rubric F1:** A thorough evaluation of the statistical significance of the analysis is provided, and the evaluation uses accurate calculations.

This section should report the results of the planned statistical test(s) or model(s) from Task 2 section C4. For at least one analytic method provide all the items for a *statistical test* or a *model*:

For *statistical tests*, provide the following information:

* The null hypothesis (the opposite of your hypothesis).
* The name of the statistical test.
* The metric(s) generated from that test (e.g., a t-stat and the derived probability).
* The *alpha* value (denoted α; usually 1% or 5%) given in Task 2 section C4.
* The conclusion drawn, e.g., “There is sufficient evidence to reject the null hypothesis and support the claim that (your hypothesis).”

For *models*, provide the following information:

* The type of model, e.g., supervised regression, supervised classification, etc.
* The algorithm(s) and process used to develop the model.
* The metric(s) used to assess performance.
* The benchmark for the success of the above metric is given in Task 2 section C4
* The conclusion is drawn from that metric and how it supports or does not support your hypothesis.

## F.2 Practical Significance

**Rubric F2:** A thorough and accurate evaluation of the practical significance of the data analytics solution is provided, and the evaluation is supported by specific examples.

Discuss the practical significance of the results from F1. This can repeat what you wrote in Task 2 section C5 adjusting as necessary according to the results. Practical significance refers to how meaningful your findings are in practical application. Results are practically significant when the difference is large enough to be meaningful in real life. This is subjective; so try your best to frame the results as a success.

Include an example of how the client might apply your work discussed in section F1.

## F.3 Overall Success

**Rubric F3:** A thorough and accurate evaluation of the overall success and effectiveness of the project is provided.

Based on the results presented in F1 and F2, discuss how the project was successful. This section may repeat content from sections F1, and F2 above, and Task 2 section B6.

# G. Conclusion

## G.1 Summary of Conclusions

**Rubric G1:** The conclusions drawn from the analysis are summarized and are appropriate and logically consistent with the data set, chosen analytic methods, and stated goals of the project.

Summarize your conclusions resulting from the entire project. This section can combine, repeat, and expand on content from throughout Task 3.

## G.2 Effective Storytelling

**Rubric G2:** The explanation includes logical reasons why the chosen tools and graphical representations for visually communicating the findings support effective storytelling.

Summarize your visualizations and how they support effective storytelling. Discuss all graphical representations and the tools used for the development.

## G.3 Recommended Courses of Action

**Rubric G3:** The submission recommends *2*courses of action, and *both* are logical, based on the findings, follow logically from the analysis, and directly address the research question or organizational need of the project.

Provide TWO recommendations based on the results (section F2) and conclusion (section G1) of your data analysis. Explain how each recommendation relates to the research question or organizational need given in Task 2 section A1 or redefined in Task 3 part A.

# H Panopto Presentation

**Rubric H:** A link to a Panopto recording is provided, and the summary includes *each* of the given elements. The summary is appropriate for an audience of data analytics peers.

Provide a link to your Panopto video.

* You are required to use Panopto, and you need to first request access, [Panopto Access](https://wgu.hosted.panopto.com/). Gaining permission may take up to 48 hours.
* Your recording should include your voice, but showing your face is not required. If you need special accommodations, please contact [Assessment Services](mailto:assessmentservices@wgu.edu?subject=D195%20task%203%20Panopto%20Video) and alert your assigned course faculty.
* There is no minimum length, but it is not meant to be long. Depending on your project, approximately 5-15 minutes is a good length.

The video should provide the viewer with a bird’s-eye view of your project and how you conducted your analysis. Here are the points you should cover:

1. A summary of your research question or organizational need.
2. A summary demonstration of the functionality of any code you used for your data analytics solution.
3. An outline of the findings and implications of your analysis.

The video should summarize your question and findings. You can think of this as a “water cooler” version of your report. The evaluator should be able to watch the video and understand your project’s purpose and main argument. Then, for step 2, bring your code (or software analysis) on screen and step through how you conducted your analysis. Unfortunately, we don’t have an example video. However, I never see the video rejected unless it’s missing step 2 above.   
 

# References

No sources were cited.

# Appendix A

# Title of Appendix

Put any supporting material in these appendices. Add additional or delete superfluous appendices as needed.

# Appendix B

# Title of Appendix

Put any supporting material in these appendices. Add additional or delete superfluous appendices as needed.

# Appendix C

# Title of Appendix

Put any supporting material in these appendices. Add additional or delete superfluous appendices as needed.

# Appendix D

# Title of Appendix

Put any supporting material in these appendices. Add additional or delete superfluous appendices as needed.