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 themselves 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 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 what 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. The Mann-Whitney U Test was used to compare the outliers with the non-outliers. Finally, each features’ distribution was independently compared to the sample visually and the result file created for the client.

# B. Project Execution

Rapid Application Development process was employed to deliver this project. Consisted of a design session, code, review, and iterations until client was satisfied and a sign-off was obtained.

The project was originally slated to take one hundred ours and start on June 24th. Instead, the client wanted to finish before the July 4th weekend, so the start time was moved up to June 10th. After starting the project, it was realized that the tasks were not as complicated as first thought and they took much less time than originally expected. The project was therefore completed in thirty and a half hours. Almost a third of the original estimate. The largest areas of variance were the design meeting and client review sessions. Also, there were no other iterations needed and the client signed off on the first draft. See the deliverables and timeline below.

*Project Goals, Objectives, and Deliverables*

* Goal 1: The goal is to select a set of users with the most influence among the others.
  + Objective 1.0: Client discussions centering around plan, design, and expectations.
    - Deliverable 1.0.1: Meeting notes and transcript with client.
    - Deliverable 1.0.2: Project plan and timelines.
  + Objective 1.1: Acquire the data.
    - Deliverable 1.1.1: The data will be downloaded in its original CSV format.
    - Deliverable 1.1.2: The data will be saved in parquet format for faster processing.
  + Objective 1.2: Analyze available features and clean the data.
    - Deliverable 1.2.1: Features will be selected with descriptive statistics, correlation, and basic understanding of the data structure used to get a sense of the data.
  + Objective 1.3: Sample enough data to run machine learning algorithms on.
    - Deliverable 1.3.1: A sample of data with the highest impactful users will be selected and stored in memory.
    - Deliverable 1.3.2: The sample will be much smaller than the original for faster processing during machine learning.
  + Objective 1.4: Reduce the data to two dimensions and visualize for an understanding of the outliers.
    - Deliverable 1.4.1: PCA will be used to reduce dimensions to two dimensions and a scatter plot displayed to show the pattern of our sample.
    - Deliverable 1.4.2: t-SNE will be used to reduce to two dimensions and a scatter be displayed of the results to see how it agrees with the PCA version.
  + Objective 1.5: Create an Isolation Forest model.
    - Deliverable 1.5.1: An isolation forest model shall be created using the relevant features and appropriate contamination parameter.
    - Deliverable 1.5.2: A pickle file shall be saved of the model to disk.
  + Objective 1.6: Select the outliers and ensure count is below one thousand.
    - Deliverable 1.6.1: A dataset shall be created of the outliers predicted by the model.
    - Deliverable 1.6.2: The dataset shall be saved as an Excel file per the client.
  + Objective 1.7: Meet with clients and deliver the result.
    - Deliverable 1.7.1: Obtain sign-off on work or feedback about revisions.
    - Deliverable 1.7.2: Wrap up project or deliver a new plan based on feedback.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Milestone or deliverable | Duration  (hours or days) | Projected  start date | Anticipated  end date | Deviation |
| Deliverable 1.0.1 | 16 hours | *June 24th 2024* | *June 25th 2024* | *This project was moved up to start on th June 10th. Only took 4 hours to meet with client.* |
| Deliverable 1.0.2 | 24 hours | *June 26th 2024* | *June 28th 2024* | *Project planning took less time. Only 4 hours.* |
| Deliverable 1.1.1 | 2 hours | *July 1st 2024* | *July 2nd 2024* | *File acquisition only took 0.5 hours.* |
| Deliverable 1.1.2 | 2 hours | *July 1st 2024* | *July 2nd 2024* | *Saving data to parquet only took 0.5 hours.* |
| Deliverable 1.2.1 | 8 hours | *July 1st 2024* | *July 2nd 2024* | *There was no deviation here. It was as expected.* |
| Deliverable 1.3.1 | 7.5 hours | *July 3rd 2024* | *July 3rd 2024* | *Sampling the data only took 2 hours.* |
| Deliverable 1.3.2 | 0.5 hours | *July 3rd 2024* | *July 3rd 2024* | *Verifying the sample was smaller did not deviate.* |
| Holiday | 0 hours | *July 4th 2024* | *July 7th 2024* | *Started earlier, should not factor in.* |
| Deliverable 1.4.1 | 4 hours | *July 8th 2024* | *July 8th 2024* | *PCA visual took 2 hours.* |
| Deliverable 1.4.2 | 4 hours | *July 8th 2024* | *July 8th 2024* | *t-SNE visual to 2 hours.* |
| Deliverable 1.5.1 | 15 hours | *July 9th* | *July 10th 2024* | *Isolation forest only took 4 hours to train and validate.* |
| Deliverable 1.5.2 | 1 hours | *July 10th 2024* | *July 10th 2024* | *Only took 0.5 hours to save to pickle file.* |
| Deliverable 1.6.1 | 6 hours | *July 11th 2024* | *July 11th 2024* | *Only took 2 hours to review the results.* |
| Deliverable 1.6.2 | 2 hours | *July 11th 2024* | *July 11th 2024* | *Only took 0.5 hours to save the final file.* |
| Deliverable 1.7.1 | 6 hours | *July 12th 2024* | *July 12th 2024* | *Meeting with clients was not necessary after sending them the results.* |
| Deliverable 1.7.2 | 2 hours | *July 12th 2024* | *July 12th 2024* | *No further planning was needed as the client signed off on the work.* |

# C. Data Collection Process

## Recommendations.csv is a tabular file that has the most records which identify whether the user recommends the game, how many people found the recommendation helpful or funny, the review date, and number of hours user played the game. This is the file that we will be using for the project. The data was downloaded from Kaggle collected directly from the Steam website using screen scraping by the author. The CSV file was imported using Pandas and then saved as a parquet file in case needed to re-run the process.

## This dataset has the number of recommendations made along with the number of products and whether the user recommended it or not. Along with that it has important measures for how many other people thought the recommendation was helpful and/or funny. This is appropriate because we are looking for those users that have the most influence.

## Our data is exceptionally clean and up to date. There are no missing values. It was curated well. There are a few issues that will need to be taken care of. One, there are values in the columns that are zero, and those rows should be removed. Two, some total hours played do not seem like they could be accurate. No other cleaning was different than planned and the items above were executed before running any processes. No other privacy, security, or governance issues came up that were unexpected.

## C.1 Advantages and Limitations of Data Set

One advantage of this data set is its shear coverage of data. It has been curated well and there are over fourteen million users that have been collected from the Steam website itself since the beginning of the platform. Another is that Steam captures pieces of information that are very useful for analysis for this particular use case and that is the “helpful” and “funny” columns. Every time someone makes a comment, others can say whether they found it helpful and/or funny.

A disadvantage of this data is there is not much in the way identifiable information in which to glean even more insights. For example, we cannot stratify by gender or region. These would be great to characterize the users and potentially weed them out based on certain attributes.

# D. Data Extraction and Preparation

The data was downloaded from Kaggle [here](https://www.kaggle.com/datasets/antonkozyriev/game-recommendations-on-steam?resource=download&select=recommendations.csv) using a browser. It was then saved in a Jupyter Notebook project. The CSV file was uploaded using Pandas and then saved as a parquet file in case the process needed to be re-run. Pandas is a common library for loading data from many different file types. It can consume CSV and parquet. Parquet if processed very quickly compared to CSV and consumes much less space. Thus, the time to load is reduced and space saved.

After obtaining the data, it was examined using Pandas methods, info() and describe(). Info() produces the schema of the data along with missing values and types. Describe() produces a table of descriptive statistics for each numerical type. After viewing the describe output, it was determined that any value with 0 was to be removed. Also, hours played were too high in some cases, so those were removed as well. The data was then visualized for each column using the hist() function of Panadas. A histogram was rendered for each column and every distribution was skewed heavily right which aligns well with the mean and medians provided by the describe() function.

Finally, the data was summarized on all the columns to the user\_id level resulting in over fourteen million user records. After the summarization, the data was examined for the appropriateness of features and a PCA component was created to summarize those features. Then, a sample of the right most five percent selected based on that new column. At this point the data is clean, selected, and ready to be further analyzed.

# E. Data Analysis Process

## E.1 Data Analysis Methods

Raw data was explored, summarized, and cleaned. A single PCA component was created from the relevant features. Everything right of the ninety fifth percentile was selected as our primary sample using the single PCA component.

An Isolation Forest algorithm was created with the appropriate contamination rate set and features to obtain enough users for our final sample and labeled them as outliers or not. The contamination rate we chose was appropriate to find enough outliers to satisfy our requirements.

Finally, the Mann-Whitney U Test was used to compare the outliers with the non-outliers. This was an appropriate tool to compare two skewed distributions. A p-value of zero was found and which was less than the alpha of .05 which gave compelling evidence in support of the alternative hypothesis, thus concluding analysis and delivering the users. For completeness, the test was run and visualized for each feature independently.

## Summary of Statistical and Mchine Learning Methods

**Standardization**

* Unsupervised statistical method for scaling data. Fitted before use of PCA.

**PCA**

* Dimension reduction unsupervised.
* Principal Component Analysis algorithm
* Features included in the fitting of the PCA component will rely on correlation metrics of less than 0.60 and removal of features that do not make sense. Feature importance can be viewed to see how each feature is weighed. Social features were heavily weighted, and the others removed. The process was repeated once more to determine what the features were.
* No benchmark for this. In a different scenario, the number of components could be determined by the amount of variance explained by the components. This could be viewed using a scree plot, but this is not necessary for this application as only one PCA component was needed to decide the tail of the data.

**Percentile Method**

* n = (P/100) x N where P=percentile and N = number of values in the dataset sorted from greatest to least using our single PCA component.
* The percentile method was used to sample everything right-of and include the 95th percentile.
* The count of the sample size was measured to ensure it was large enough.
* Greater than 10,000 was sufficient.

**Isolation Forest**

* Unsupervised
* Isolation Forest algorithm
* The count of outliers is important to measure as the number of users to be returned is less than a thousand.
* Between 500 and 1,000 users should be sufficient. We approximately returned ~700 users.

**Mann-Whitney U Test**

* The Mann-Whitney U test was used to validate that the sample of outliers was greater than the non-outliers.
* The metrics computed from the Mann-Whitney U test are the U-Statistic and the p-value.
* An alpha of 0.05 was used to measure this.
  + H0: There is no difference in the means between the outlier users from the non-outlier users.
  + H1: The outlier user means are greater than the non-outlier user means.

Standardization is recommended for PCA. PCA operates under the assumption that the data is normally distributed. The data was heavily skewed right. PCA is helpful here to gain a single value that maintained the relative positioning based on multiple features.

Once the user sample was selected, Isolation Forest was used to separate out the strong cases from the others. The contamination parameter was set appropriately to ensure we had enough users in the final output. Isolation Forests is a well-known method for discovering outliers in a dataset. PCA and t-SNE with two components was used to visualize the outlier vs non-outliers and it was observed there was a clear separation of the data. More importantly the outliers were more on the right side of the distribution. After verifying the distributions were different with eth Mann-Whitney U test, the data was bootstrapped to get a normal distribution, so the differences could be visualized.

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

*Tools*

* **Windows 11**
  + Advantage: Almost everyone is familiar with and has it
  + Disadvantages: Requires a license and not open source
* **Windows Subsystem for Linux (WSL) with Ubuntu 20.04.6 LTS**
  + Advantage: Great for isolating your main OS from your process thus not crashing the main OS.
  + Disadvantage: Steep learning curve.
* **Miniconda**:
  + Advantage: Great for creating separate environments for running different versions of python and packages.
  + Disadvantage: Miniconda is a bare bones approach and requires knowledge of what you need upfront. Conda could be a better choice as most all tools will be installed by default and has a user interface to simplify packages and environments.
* **Python 3.10.14**
  + Advantage: Very recognized as a data science platform.
  + Disadvantage: Requires programming skills and practice. This version may be missing some bug fixes, optimizations, or other helpful newer features.
* **GitHub**
  + Advantage: Great for storing, versioning, and sharing your work.
  + Disadvantage: Large file storage can cost, and the learning curve can be steep.
* **Jupyter Notebook**
  + Advantage: Very well recognized notebook environment and well documented.
  + Disadvantage: Knowing how to launch and manage this as a server can be challenging.
* **Matplotlib**
  + Advantage: Great for quick visualizations.
  + Disadvantage: Does not produce pretty output by default. Can take more work.
* **NumPy**
  + Advantage: Great for managing arrays and the math functions to support them.
  + Disadvantage: If only relying on this, then a lot of code would be required to manage.
* **Pandas**
  + Advantage: Easily load data from a variety of data source types.
  + Disadvantage: Can be slow with large data.
* **Scikit-learn**
  + Advantage: Easy to implement classical machine learning.
  + Disadvantage: Not great for deep learning scenarios and can be slow with large datasets.
* **Seaborn**
  + Advantage: Produces nice looking visuals by default.
  + Disadvantage: Less control over finer details of the visual. Must use Matplotlib for that.
* **SciPy**
  + Advantage: Has many statistical functions.
  + Disadvantage: Some algorithms are slower when compared to other libraries.
* **Polars**
  + Advantage: Fast loading of data and manipulation.
  + Disadvantage: So different from Pandas that you must relearn data frames. Luckily there is a to\_pandas() function, but of course than the Pandas data frame will perform worse.
* **OpenPyXL**
  + Advantage: Easily export a data frame to an Excel file.
  + Disadvantage: Can be complex for new users and slow with larger datasets.

*Techniques*

* **Descriptive Statistics**
  + Advantage:Well defined and easy to implement.
  + Disadvantage:Can oversimplify results and sensitive to outliers.
* **Dimension Reduction**
  + Advantage: Summarize multiple features to less or a single dimension.
  + Disadvantage: Can potentially cause data loss which could cause results to be invalid.
* **Machine Learning**
  + Advantage: Can handle a lot of data and features and produce insights impossible for humans to do manually.
  + Disadvantage: Can be harder to interpret and training can take a while.
* **Bootstrapping**
  + Advantage: Can take an abnormal distribution and make it more normal.
  + Disadvantage: Can take a while to perform and ensuring you have enough samples is important and not the same for every problem.

## 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.