**Case Study Project – Video Game Categories**

**Background**: Only 23% of buyers use genre filter to find games of interest. Instead, 71% of buyers use advanced filters.

**Task**: Identify new genres (No more than 5, unless there is a concrete reason to do so) that will improve the users’ experience in finding new products.

**Business** **Focus**: Improve user experience which should improve customer retention and could possibly increase the amount of sales in the future with repeat buyers.

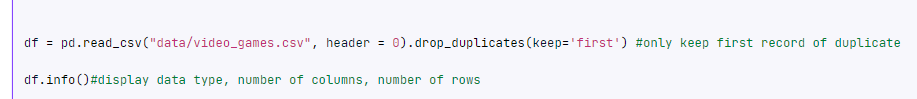
**Business** **Metric**: New genres, no more than 5. Future metrics would include returning customers (Account tracking), and A/B testing for time spent on website to make a purchase.

**Method**: Unsupervised machine learning task that needs to separate the data into a number of segments. The best choice for this task would be K-Means clustering.

**Written Report**

**Data Validation**

* All duplicate entries were removed, only the first row was kept. Resulted in 40,832 rows



* Indexes 19725, 7402, 23089, 29089, 29480, 20469 ,12867, 12844, 12141, 6848, 23468, 23408, 7762, 9955, 28167, 5410 were removed because of too many missing variables.

Graphical user interface, text, application, email

Description automatically generated

* For columns “achievements” and “num\_reviews”, all NULL values were replaced with 0’sText

  Description automatically generated
* Removed all remaining rows if any NULLs are contained. This resulted in remaining 16,671 rows.

A picture containing text

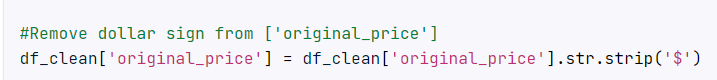
Description automatically generated

* “Multiplayer” and “mature” columns were enumerated.

Text

Description automatically generated with medium confidence

* Removed dollar sign from “original\_price” column. Changed to “float” type.



* Any column containing “Free” in “original\_price” was changed to 0. Rest of non-numeric rows were dropped. Values changed to “float”

Text

Description automatically generated

* New column “positive\_review\_count” was created, but was not used because of high correlation with num\_reviews.

Graphical user interface, text

Description automatically generated

Chart, waterfall chart

Description automatically generated

**Exploratory Analysis**

* All variables are not highly correlated, except for the created column “positive\_review\_count”, which was not used in model training because of the high correlation. Instead, “num\_reviews” was chosen because it also gives insight on how popular the game is.

Chart, waterfall chart

Description automatically generated

* Identified most popular developers – judging by amount of reviews. Since there was no sales data provided, this can be an alternative to identifying the most successful developers because of the positive correlation with positive\_review\_count– and therefore successful games. Can also be used to identify trending developers with added time filters (possibly top chart on website).Chart, bar chart, histogram

  Description automatically generated
* Identified highest rated developers with over 100,000 reviews total. This can be used in a similar manner as point above and implemented as a top chart on the website.

**Chart, bar chart, histogram

Description automatically generated**

* Identified most reviewed games, combined with time filters, this can be used as a top chart on the website -- showing currently trending games.

Chart, bar chart, histogram

Description automatically generated

* Identified highest rated games with over 100,000 total reviews. This can be used in a similar manner as point above and implemented as a top chart on the website.

**Chart, bar chart, histogram

Description automatically generated**

* Non-multiplayer games tend to be rated slightly higher. This could be due to multiple factors like toxic online community, hackers, server issues, overall game experience, bad matchmaking, etc.

Chart, radar chart

Description automatically generated

* Mature and not Mature games tend to be reviewed similarly.

Chart, radar chart

Description automatically generated

* There is a slightly negative correlation between positive reviews and number of achievements.

**Chart, scatter chart

Description automatically generated**

**Model Development**

**Model Selected**: K-Means with reduced dimensionality by applying PCA.

**Reason**: This dataset does not contain any labels and we are trying to segment the dataset to develop new “genres” to improve customer experience. Therefore, a unsupervised machine learning method is necessary. K-Means

**Data preparation**: Data will be standardized/normalized, number of components for PCA will be determined using cumulative explained variance(around 80% target), and K-means n\_clusters will be identified visually, using the “elbow” (with y-variable being Within Cluster Sum of Squares)

**Data Normalization VS Standardization**: Prior to applying PCA and then fitting the K-means model, data was normalized using MinMaxScaler and standardized using StandardScaler. In case of standardization, data is reduced to have a mean of 0 and variance as 1 – assuming that the distribution is normal within the data, while MinMaxScaler simply rescales the variables between 0 and using minimum and maximum variables. In the future, it would be better to remove the excessive outliers in the data, or use another scaler like RobustScaler, which is less sensitive to outliers.

**Model Evaluation:** There are 2 models.

* First one standardizes data using a standard scaler, applies PCA, and applies K-Means (StandardScaler -> PCA -> K-means)
* Second one normalizes data using MinMaxScaler, applies PCA, and applies K-Means

(MinMaxScaler -> PCA -> K-means)

**Model 1a(StandardScaler -> PCA -> K-means):**

1. Columns ['achievements', 'original\_price', 'percent\_positive', 'num\_reviews', 'multiplayer', 'mature'] were selected.
2. StandardScaler was applied to the data, and PCA was fit. Afterwards, Explained Variance by Component was plotted and analyzed

**Chart, line chart

Description automatically generated**

1. After reviewing the plot, 5 components were selected because it was in the 80% range, and the 6th component would overfit the model with 100% explained variance.
2. PCA was applied to data with n\_components variable as 5
3. K-means n\_clusters were optimized by graphing Within Cluster Sum of Square for a range of 1 through 14 n\_clusters.

**Chart, line chart

Description automatically generated**

1. Going off the ‘elbow’ rule, n\_clusters variable input was selected as 7. Elbow is the part of the chart that significantly reduces the slope rate (explaining diminishing returns for every additional cluster)
2. K-means data was fitted and charted with the first 2 components

Chart, scatter chart

Description automatically generated

**Model 1b(StandardScaler -> PCA -> K-means):**

1. Repeats all steps in model 1a, except K-mean n\_clusters are limited at 5 due to the limitations set by the case study. Visually, the result is not significantly different, therefore model 1b was preferred:

Chart, scatter chart

Description automatically generated

**Model 2(MinMaxScaler -> PCA -> K-means):**

1. Columns ['achievements', 'original\_price', 'percent\_positive', 'num\_reviews', 'multiplayer', 'mature'] were selected.
2. MinMaxScaler was applied to the data, and PCA was fit. Afterwards, Explained Variance by Component was plotted and analyzed:

**Chart, line chart

Description automatically generated**

1. In this case, 2 components in PCA would describe over 85% of variance. Less computationally intensive than the 5 in model 1a.
2. PCA was applied to data with n\_components variable as 2
3. K-means n\_clusters were optimized by graphing Within Cluster Sum of Square for a range of 1 through 14 n\_clusters: Chart, line chart

   Description automatically generated
4. Going off the ‘elbow’ rule, n\_clusters variable input was selected as 4. Elbow is the part of the chart that significantly reduces the slope rate (explaining diminishing returns for every additional cluster)
5. K-means data was fitted and charted with the only 2 components Chart

   Description automatically generated
   1. While reviewing results, it seems that this model identified 4 different clusters connected to 2 variables: mature and multiplayer. Graphical user interface

      Description automatically generated with medium confidence

**Summary**: Model #2 perfectly 2 filters to add: mature and multiplayer. But, more data is needed for better evaluation and model development. The type of data needed will be listed in the next part.

**Going Forward:**

* Recommended to change the rules in the database so that no blanks or nulls would be allowed in the future when entering data.
* Provide sales data with the current dataset, this might be a key variable in current/future models.
* Provide current genres of games, to possible condense/combine in a more efficient manner. Also include existing filters and maybe even their utilization % by user.
* Promote account creation for customers – to track sales. Which with enough data can be turned into a recommendation engine (users who bought X also bought Y).
* Measure amount of time spent on the website (from homepage to sale) for future A/B testing and measuring whether new features play a significant role in user experience/sales.
* Implement trending and top selling games/developers charts on the website.