Exploratory Data Analysis for Machine Learning

IBM Machine Learning -Project 1

Iheb Chaabane

December 2021

About the Data

The Data originally came from Liza Wood via Steam Spy. She recently published a blog post on her data analysis of this video game data.

The Data were split before this Analysis: 82% train and 18% test.

Data Dictionary:

variable	class	description
number	double	Game number
game	character	Game Title
release_date	character	Release date
price	double	US Dollars + Cents
owners	character	Estimated number of people owning this game.
developer	character	Group that developed the game
publisher	character	Group that published the game
average_playtime	double	Average playtime in minutes
median_playtime	double	Median playtime in minutes
metascore	double	Metascore rating

Plan:

01 Data Overview

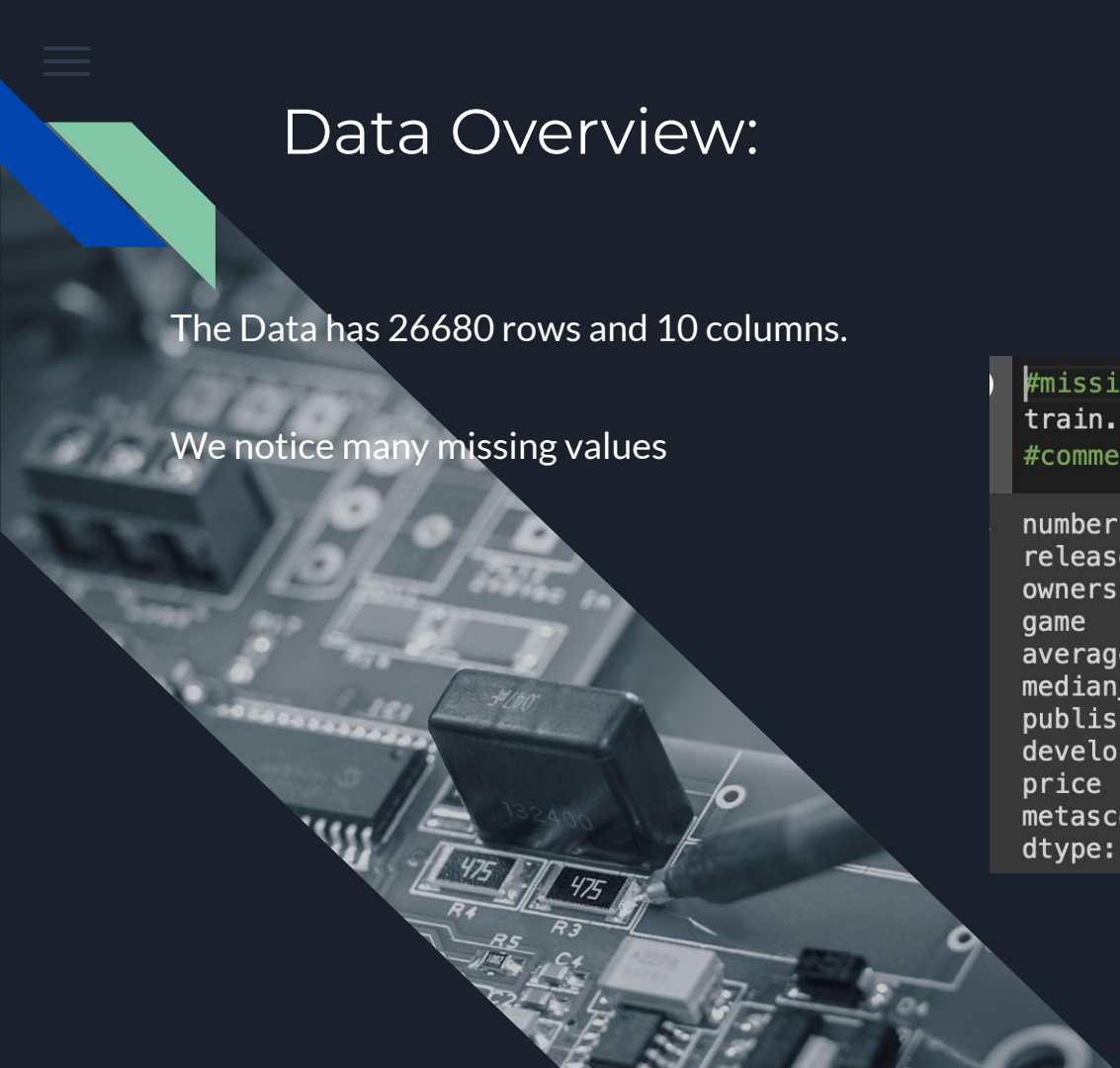
02 Data Cleaning and Feature Engineering

03 Hypothesis Testing

Data Overview:

The Data has 26680 rows and 10 columns.

We notice many missing values



```
#missing values
train.isnull().sum().sort_values()
#comment :there are a lot of missing values :
```

```
number                0
release_date          0
owners                0
game                  3
average_playtime      9
median_playtime      11
publisher            84
developer           128
price              2541
metascore         19527
dtype: int64
```

Data Cleaning

Removing features that are not useful to determine the target value: release date ,game… and rows with missing values.

Feature Engineering

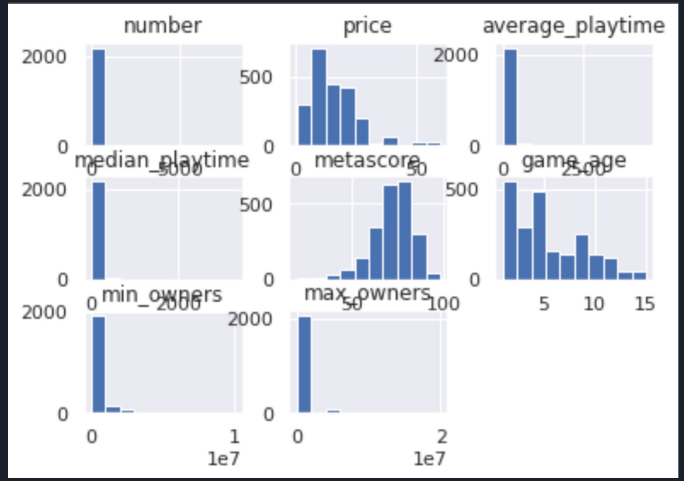
Instead we create new features such as “game_age” derived from release date, “max_owners” and “min_owners”.

Converting categorical variables to dummies: one hot encoding.

Skewness:

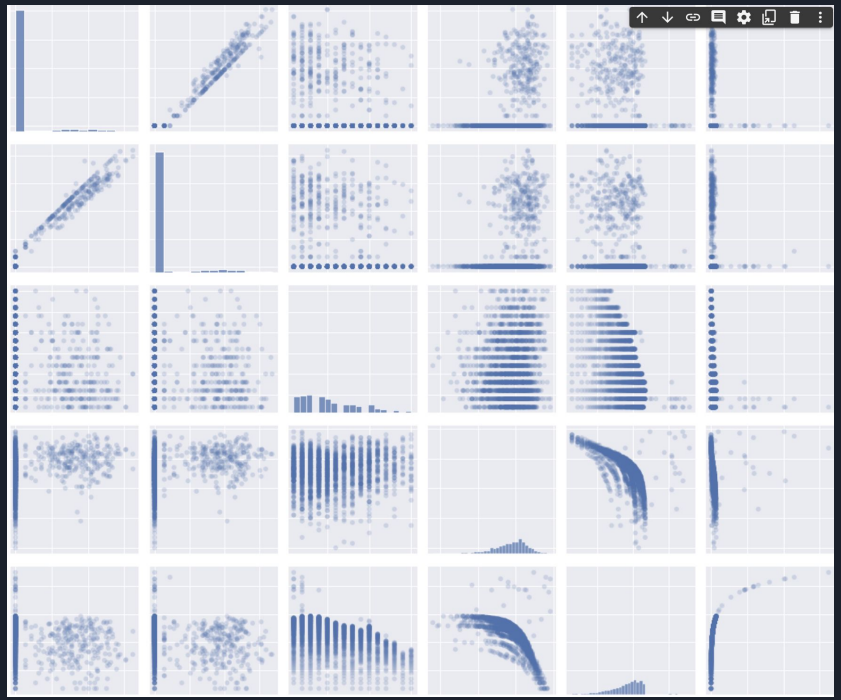
Most of the features are skewed and for that we perform a log transformation to these features.

The result are not that satisfying .We can try sqrt transformation or even box cox transformation.



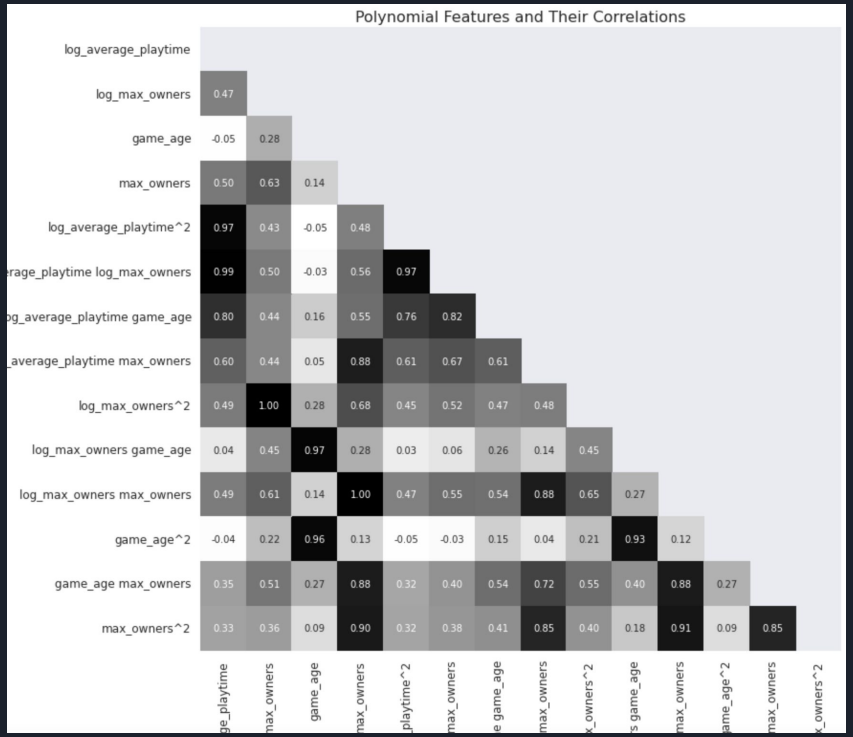
Pair Plot

- No strong linear relation between features except for between average playtime and median playtime.
- We can try adding polynomial features
- Linear Regression might not be a good fit for this data set.



Polynomial Features:

After adding some polynomial features we draw the correlation matrix below



Hypothesis Testing

We can perform many hypothesis testing like :

- `H0 : We have a normally distributed target`
- `H1: We haven't`

We can simply Apply D'agostino Test : if p-value above 0.05 we reject the null hypothesis which is the case.

```
from scipy.stats.mstats import normaltest # D'Agostino K^2 Test
```

```
[ ] normaltest(small_df.metascore.values)
```

```
NormaltestResult(statistic=206.25949858059397, pvalue=1.626743061994561e-45)
```

p-value<0.05 ===> we reject the null hypothesis

Conclusion:

As shown in the analysis linear regression may not perform well.

Notebook can be found here :https://colab.research.google.com/drive/11277MB62WgLKYXvELSITRaAcLL9pgLOC#scrollTo=Z53fZPd3ybzt