FEDERAL STATE AUTONOMOUS EDUCATIONAL INSTITUTION OF HIGHER EDUCATION
ITMO UNIVERSITY

Report on learning practice № 1

Analysis of univariate random variables

Performed by:
Denis Zakharov,
Maxim Shitilov,
Sapelnikova Ksenia,
Vdovkina Sofia,
Academic group J4132c, J4133c

Saint-Petersburg
2022

Goals

Step 1. Choose a subsample with main variables for your further analysis. Then for each of them do the following tasks.

Step 2. You need to make a non-parametric estimation of PDF in the form of histogram and using kernel density function (or probability law in case of discrete RV).

Step 3. You need to make an estimation of order statistics and represent them as “box with whiskers” plot.

Step 4. Find one or several theoretical distributions that could describe your sample on a basis of non-parametric analysis results.

Step 5. Estimate parameters of chosen distributions using methods of maximum likelihood and least squares method.

Step 6. Validate your estimated parameters using QQ biplots.

Step 7. Estimate correctness of fitted distributions using at least 2 statistical tests.

Brief theoretical part

In this lab we are going to work with a univariate distribution; It’s a probability distribution with only one random variable. This is in contrast to a multivariate distribution, the probability distribution of a random vector (consisting of multiple random variables).

To visualize our result we will use several graphical plots:

- *Histogram* as way to display frequency characteristics of random values: $f^{*}(x) = \frac{n}{\Delta x n},\ x - \frac{\Delta x}{2} < x_i < x + \frac{\Delta x}{2}$
- *Box with whiskers* the box plot where there can be lines (which are called whiskers) extending from the box indicating variability outside the upper and lower quartiles.

Kolmogorov–Smirnov test (K-S test or KS test) is a nonparametric test of the equality of continuous (or discontinuous), one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution (one-sample K–S test), or to compare two samples (two-sample K–S test). In essence, the test answers the question "What is the probability that this collection of samples could have been drawn from that probability distribution?" or, in the second case, "What is the probability that these two sets of samples were drawn from the same (but unknown) probability distribution?". It is named after Andrey Kolmogorov and Nikolai Smirnov.

Results

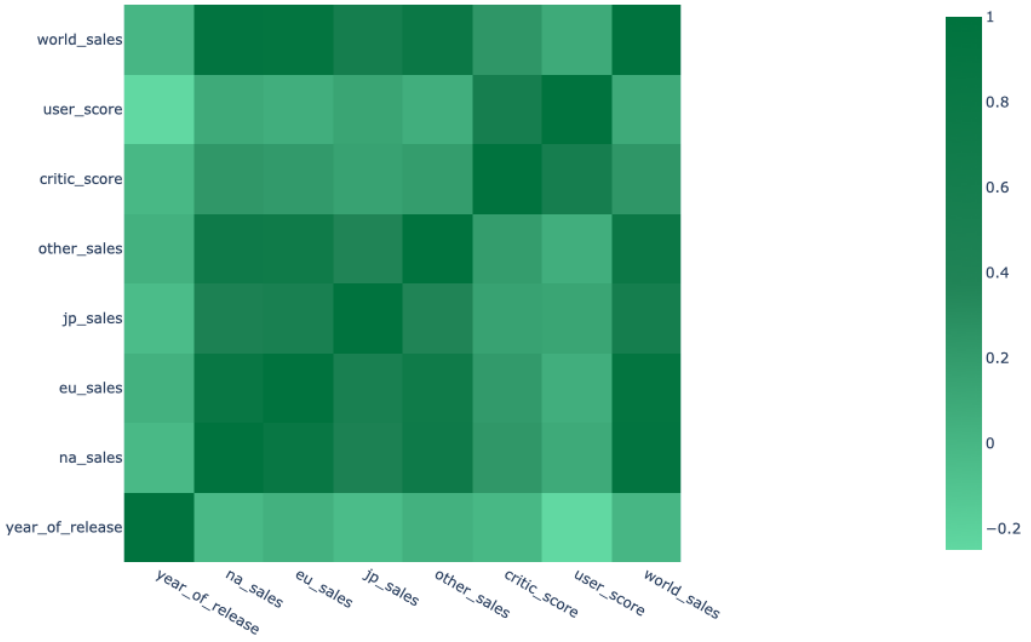
1. Substantiation of chosen subsample;

The game industry remains one of the fastest growing and most advanced industries in digital technology. Today game development is a unique interdisciplinary profession that combines technical and creative activities and does not limit the specialist in the format of work. Game company, indie development, freelancing - all these are available for a game creator. The level of development of the labor market has reached the stage when most companies are not just looking for a competent programmer or designer, but for a versatile specialist.

This dataset simply allows us to make a research on patterns of a successful game using data from Metacritic. There are 16_400+ rows in the dataset, so why assume that it's suitable for the job.

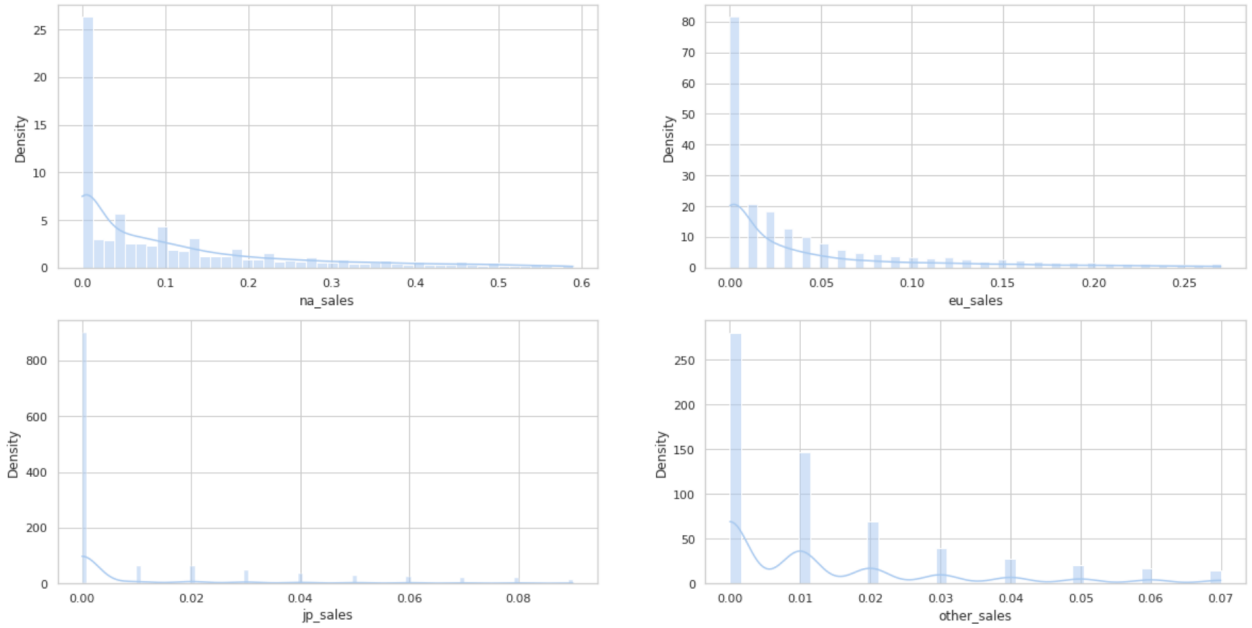
Data description:

- `Name` — game title;
- `Platform` — computing platform an environment in which a piece of software is executed.;
- `Year_of_Release` — year of the;
- `Genre` — video game genre is an informal classification of a video game based on how it is played rather than visual or narrative elements;
- `NA_sales` — sales in North America (millions of copies sold);
- `EU_sales` — sales in Europe (millions of copies sold);
- `JP_sales` — sales in Japan (millions of copies sold);
- `Other_sales` — sales in other countries (millions of copies sold);
- `Critic_Score` — critics' score (maximum 100);
- `User_Score` — user rating (maximum 10);
- `Rating` — rating from Entertainment Software Rating Board that determines the rating of computer games and assigns them the appropriate age category.

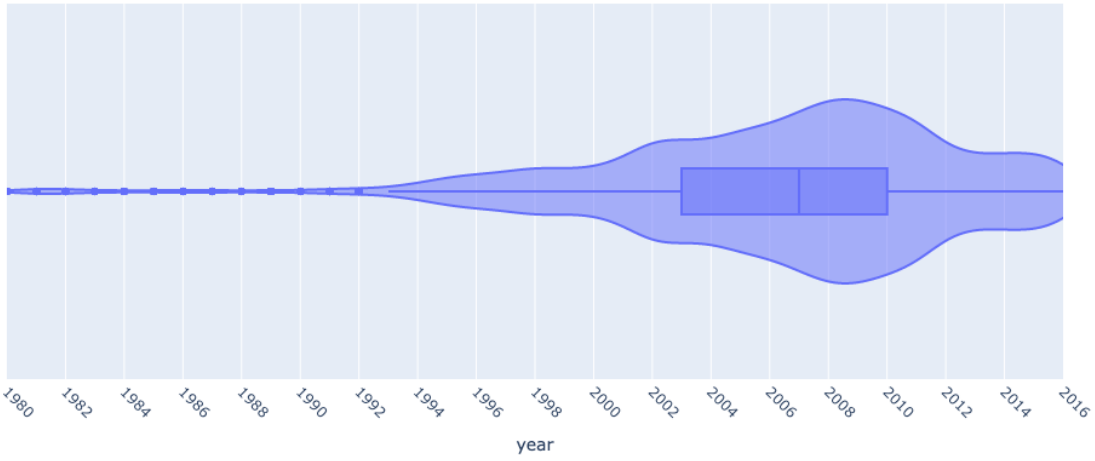


Pic. 1 — Heatmap of the correlation features

2. Plotting a non-parametric estimation of PDF in form of a histogram and using kernel density function (or probability law in case of discrete RV)



Pic. 2 — Non-parametric estimation of PDF

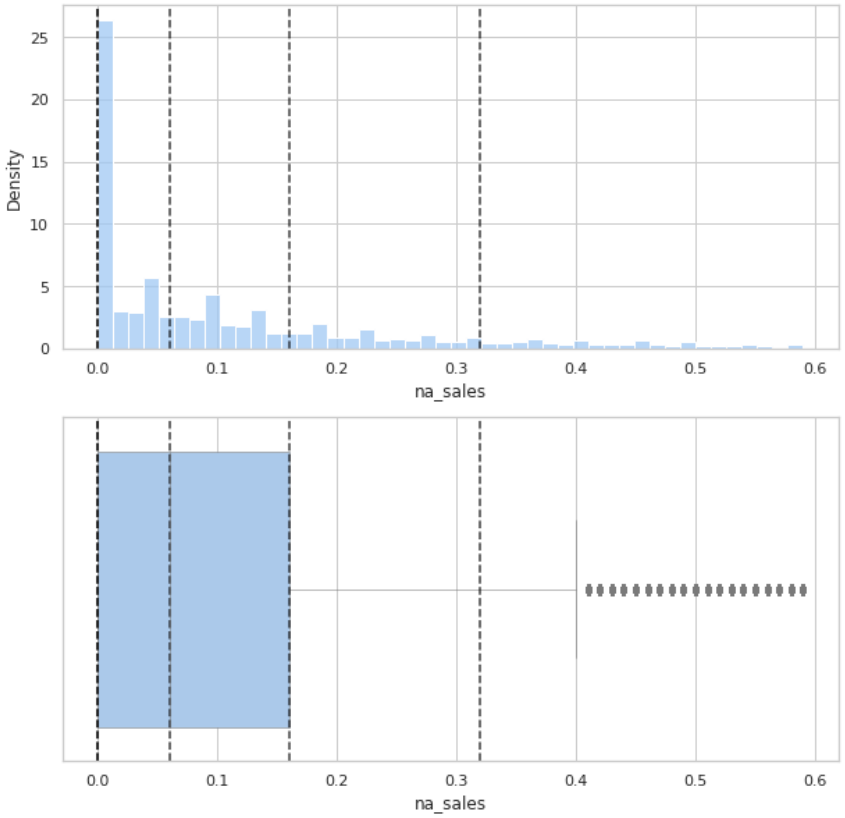


Pic. 3 — Non-parametric estimation of PDF (violinplot)

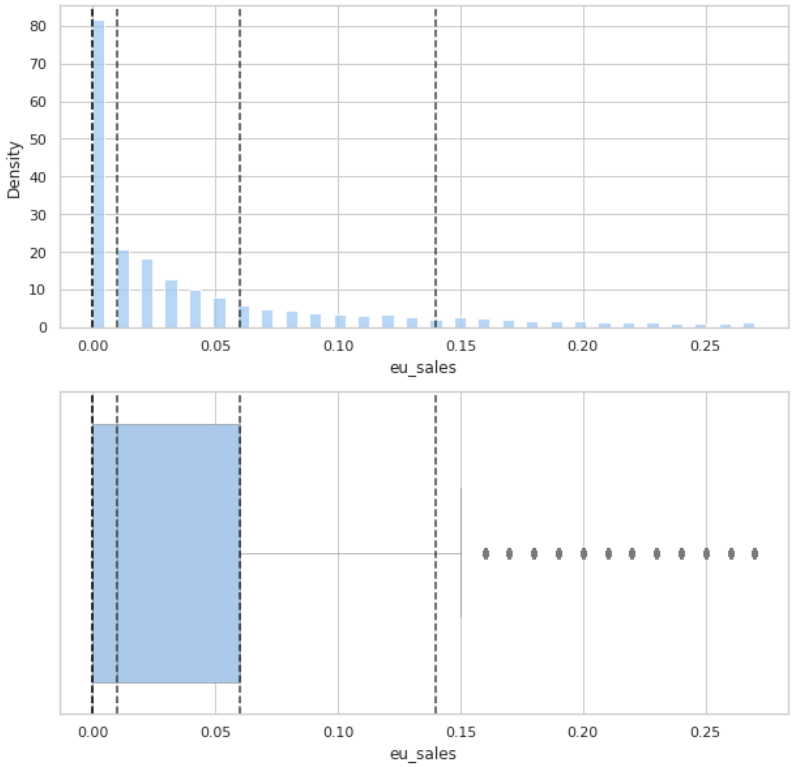
To be the more precisely we can zoom in violinplot and its nested boxplot that the exact process started at 2007.

3. Order statistics estimation and its representation as.

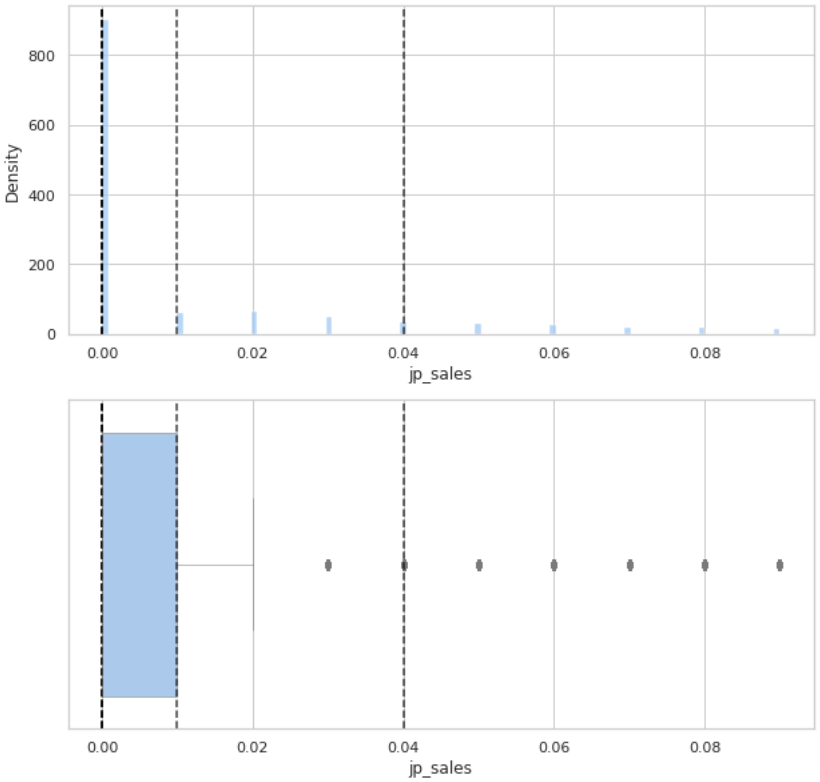
Estimation of order statistics



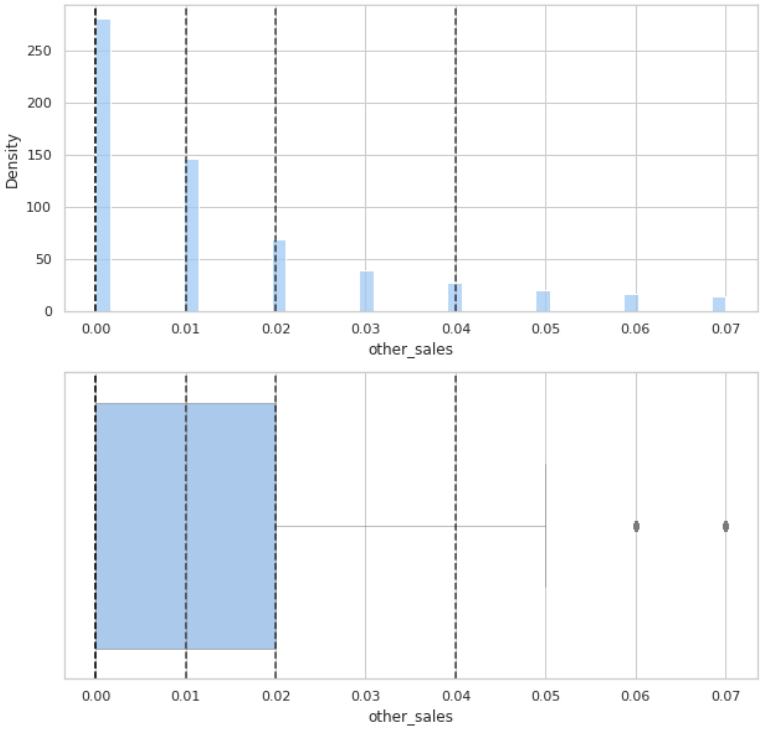
Pic. 4 — Estimation of order statistics by na_sales



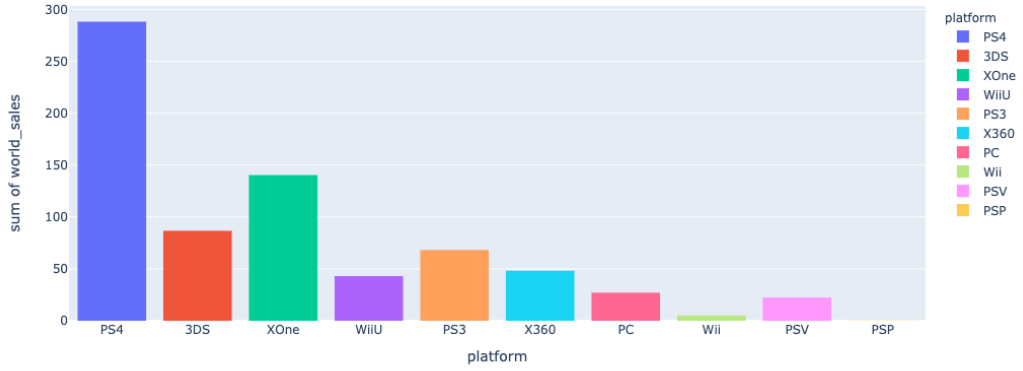
Pic. 5 — Estimation of order statistics by eu_sales



Pic. 6 — Estimation of order statistics by jp_sales

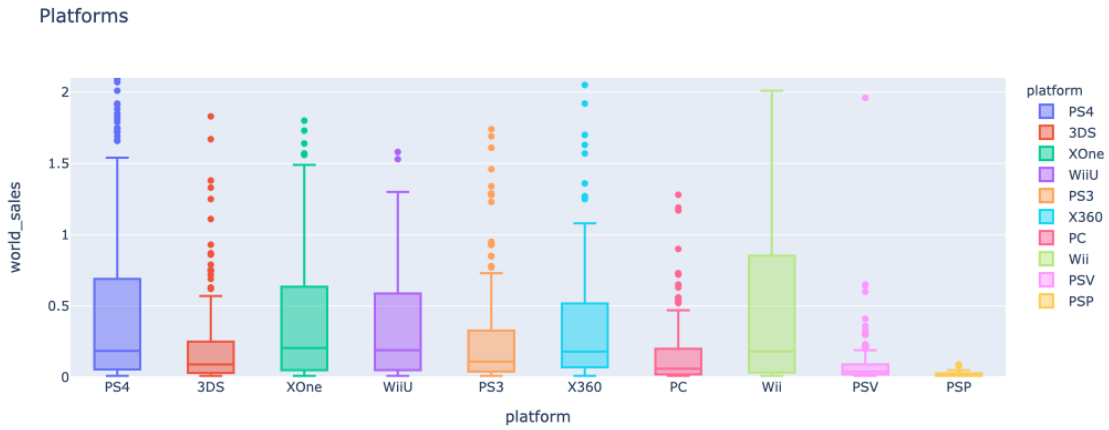


Pic. 7 — Estimation of order statistics by na_sales



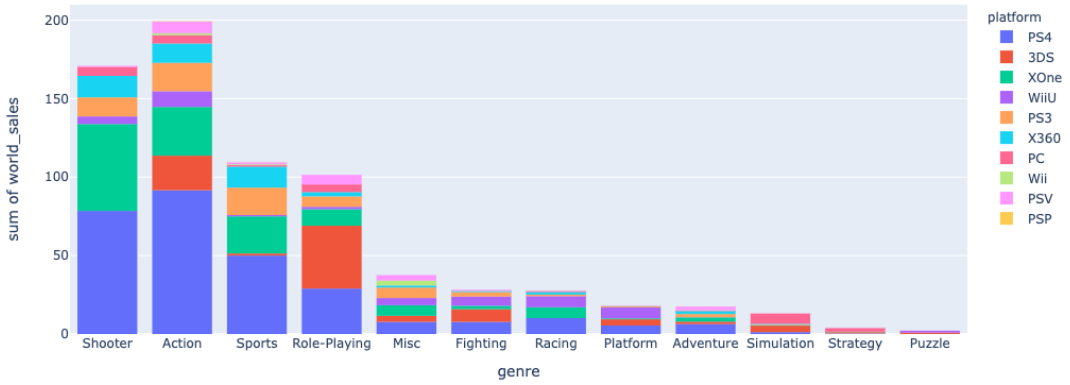
Pic. 8 — World sales distribution by Game Platforms (histogram)

The most popular platform for game release was PS4 during the analysis that we are working on and the least productive in game world-sales is PSP (we also made a sub-research that it's peak of sales was in 2006 from where it's can bee seen a soft decrease till the 2014 which is the year when hardware shipments of the PSP ended worldwide) which are both platforms from Sony Interactive.

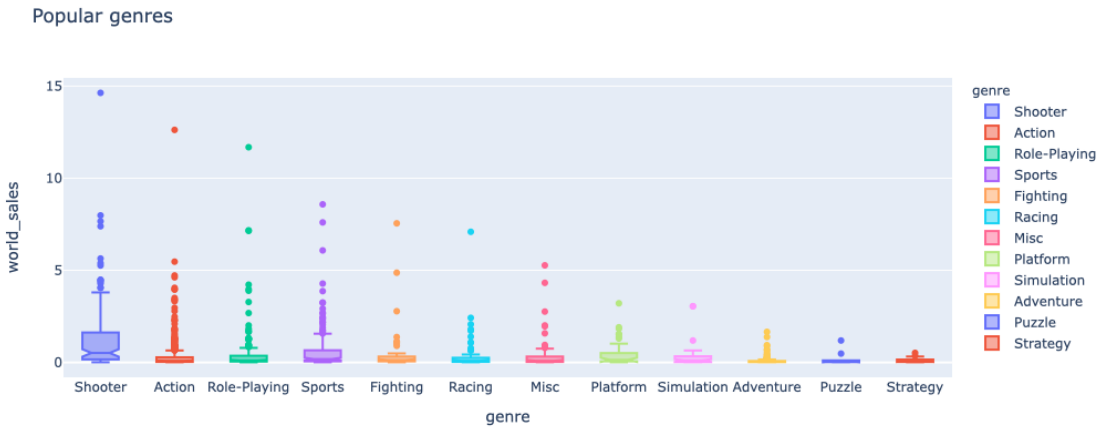


Pic. 9 — World sales distribution by Game Platforms (box with whiskers)

PS4 has a lot of outliers, that can be seen as sold games, so they have a high sailing number of copies.



Pic. 10 — World sales distribution by Game Genres (histogram)



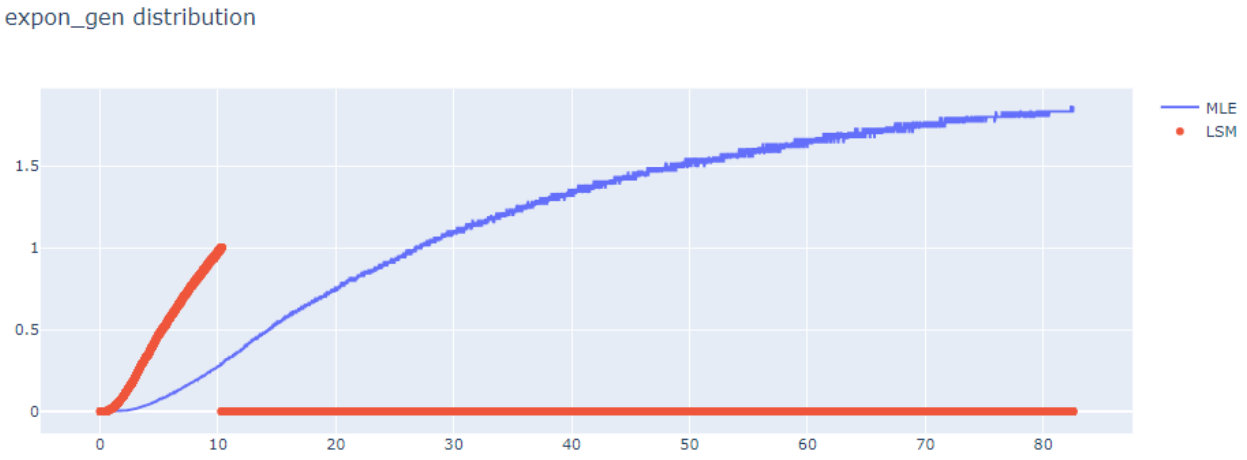
Pic. 11 — World sales distribution by Game Genres (box with whiskers)

Shooters and Actions are the most popular genres that can be seen from the histogram, however we can also see some anomalies in Action games.

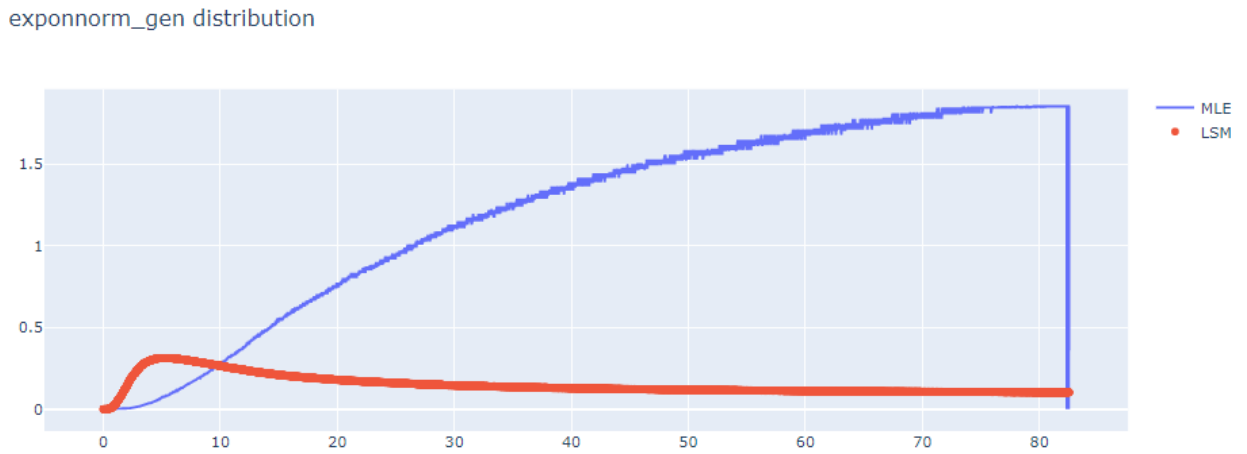
4. Selection of theoretical distributions that best reflect empirical data;

```
from scipy.stats.distributions import expon, exponnorm, exponweib, exponpow

distributions = [expon, exponnorm, exponweib, exponpow]

x = np.linspace(np.min(df['world_sales']), np.max(df['world_sales']))
```

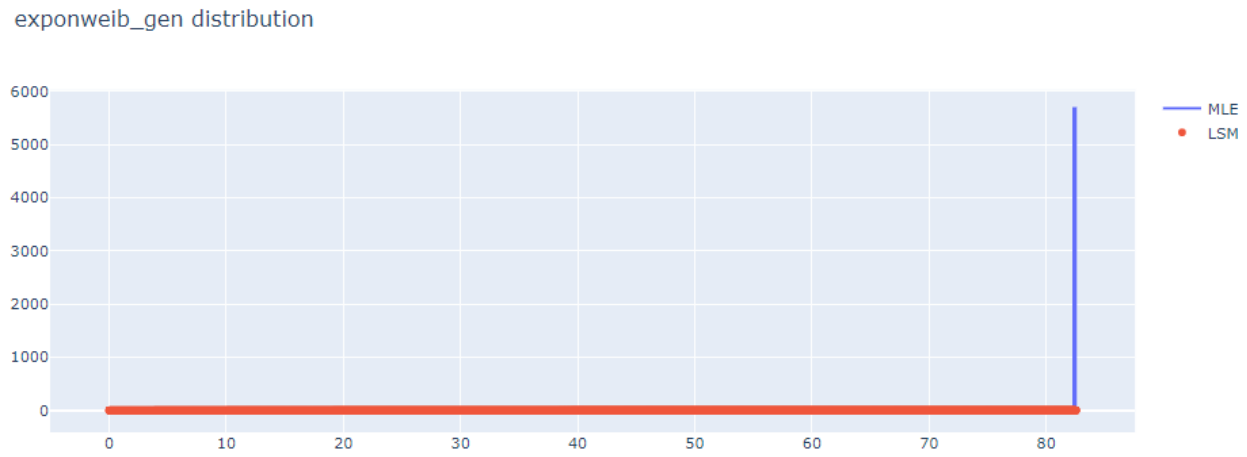
5. Estimation of random variable distribution parameters using maximum likelihood technique and LS methods.



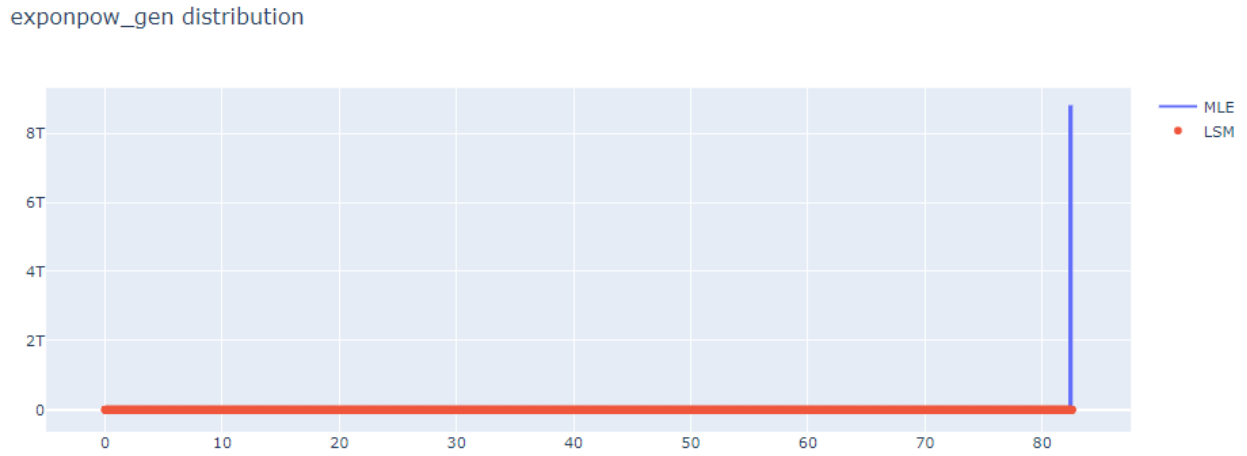
Pic. 12 — An exponential continuous random variable distribution



Pic. 13 — An exponentially modified Normal continuous random variable distribution



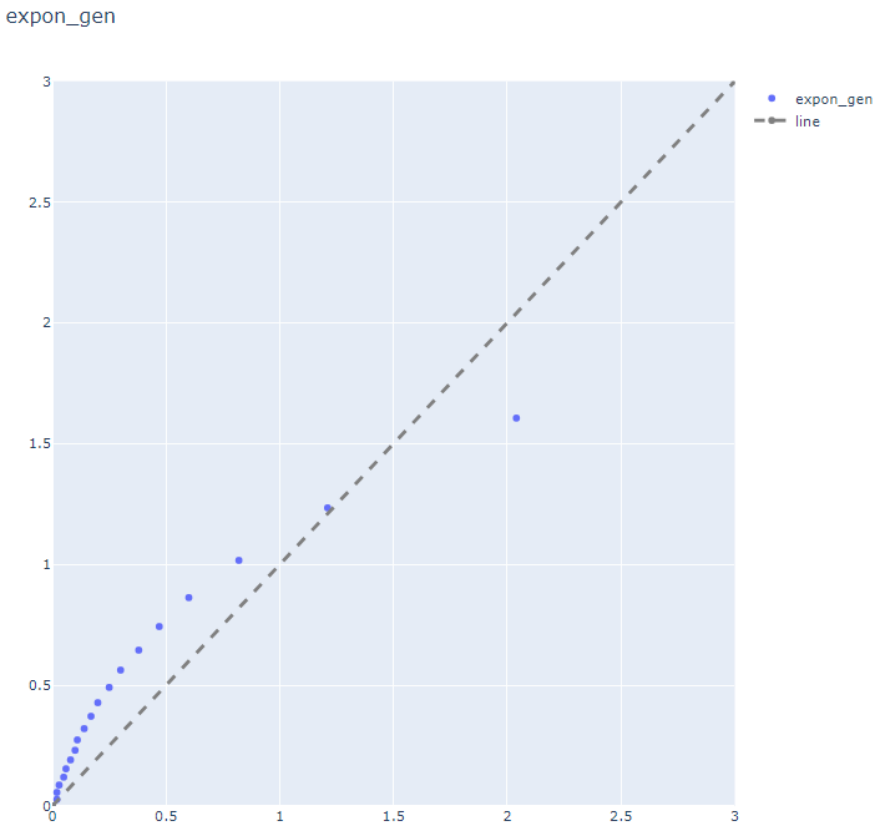
Pic. 14 — An exponentially Weibull continuous random variable distribution



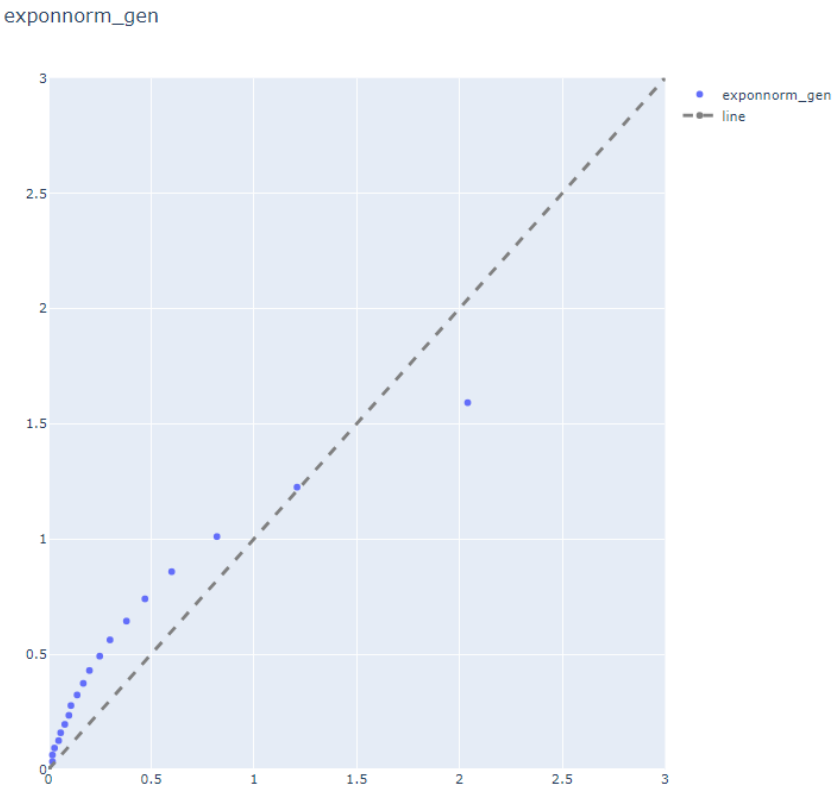
Pic. 15 — An exponentially power continuous random variable distribution

An exponentially modified Normal distribution seems to fit the chosen variable. As a result of research we can consider LSM as the worst method for exponential distribution. MLE, on the other hand, does quite the opposite by perfectly fitting in it, which is shown on the pictures above.

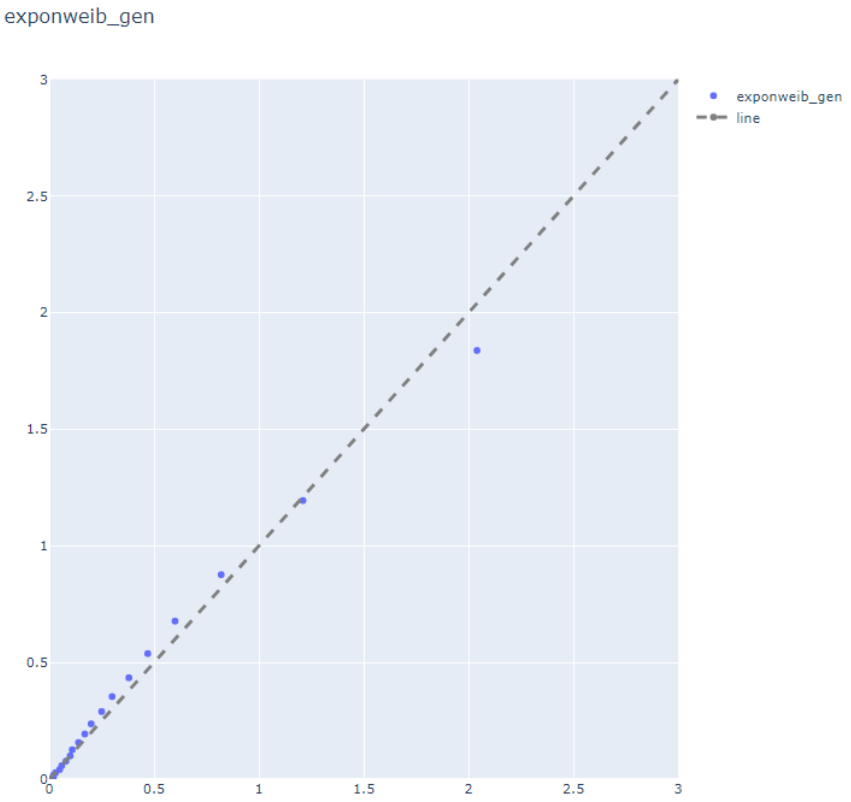
6. Validation of empirical and theoretical distributions using quantile biplots.



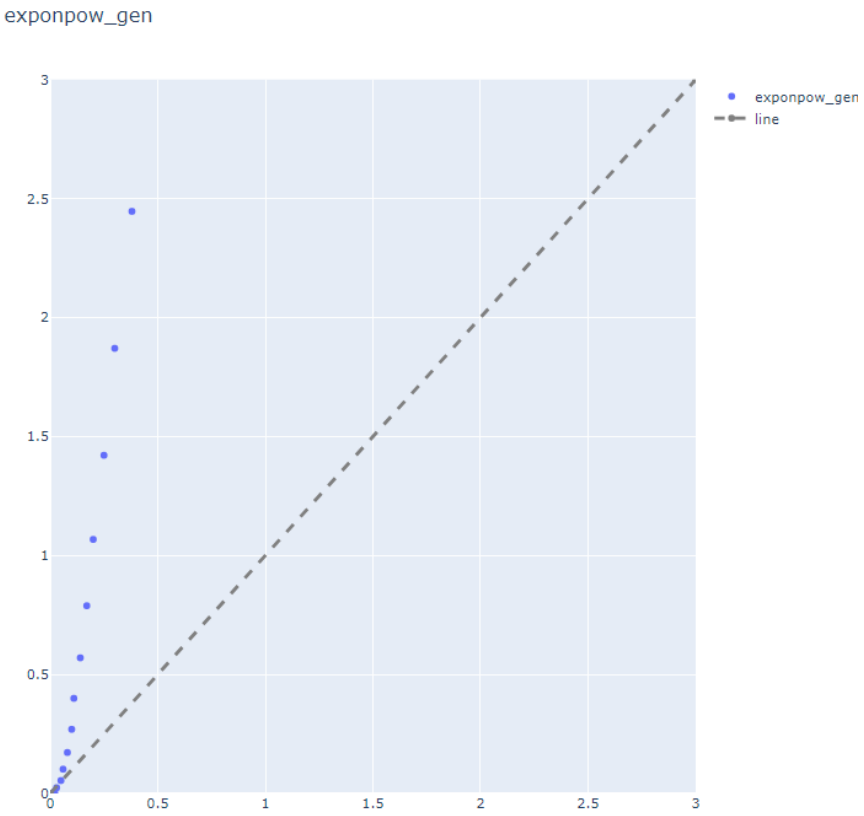
Pic. 16 — QQ biplots parameters estimation (MLE)



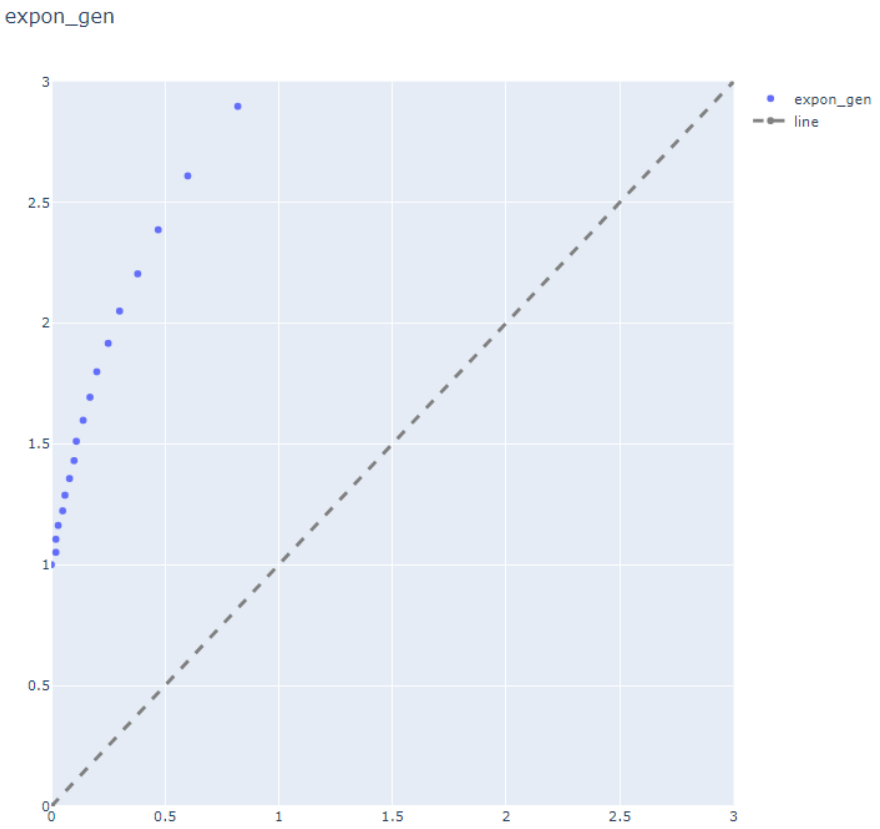
Pic. 17 — QQ biplots parameters estimation (MLE)



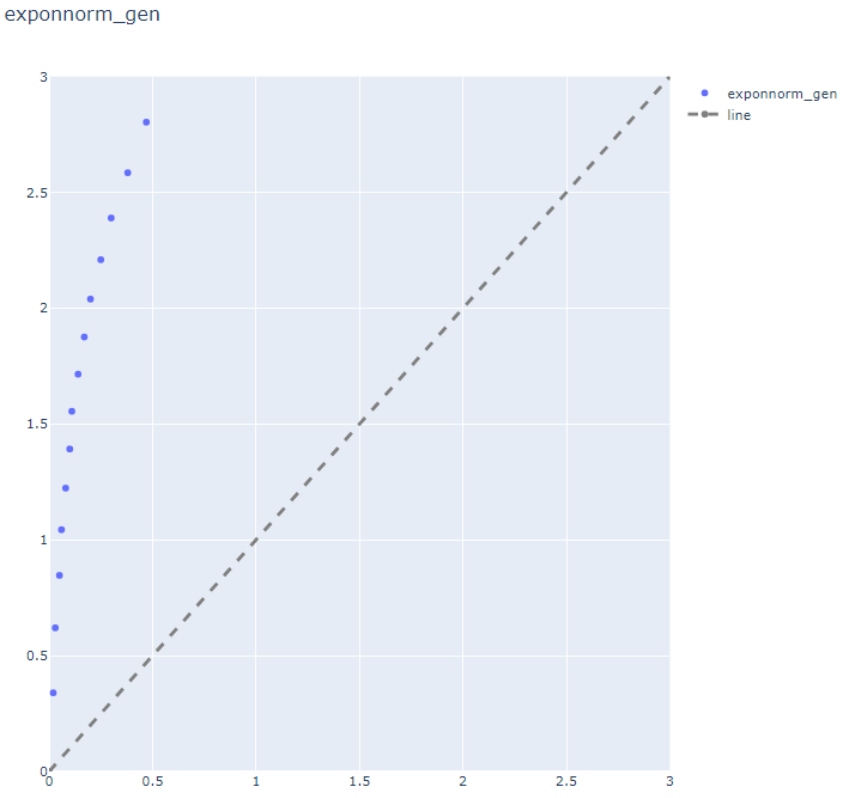
Pic. 18 — QQ biplots parameters estimation (MLE)



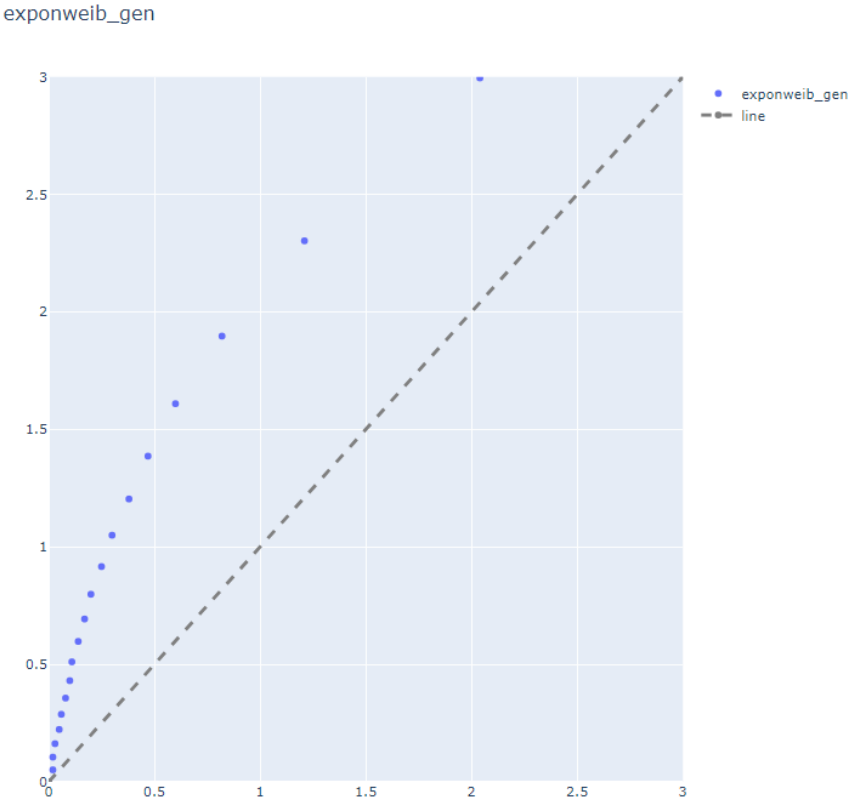
Pic. 19 — QQ biplots parameters estimation (MLE)



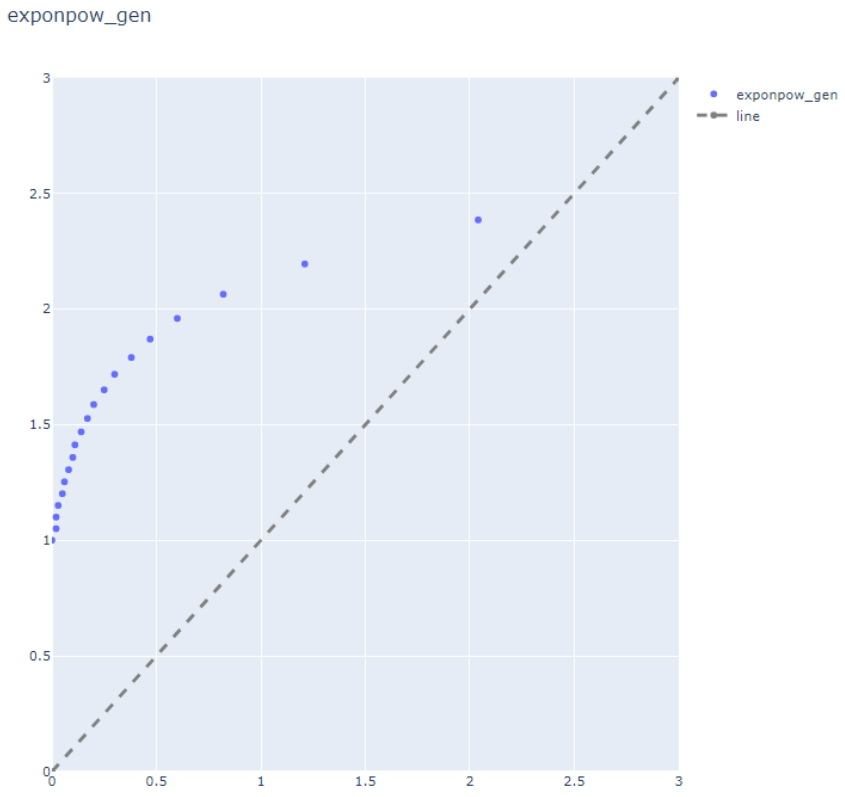
Pic. 20 — QQ biplots parameters estimation (LSM)



Pic. 21 — QQ biplots parameters estimation (LSM)



Pic. 22 — QQ biplots parameters estimation (LSM)



Pic. 23 — QQ biplots parameters estimation (LSM)

Judging by the graphs, the maximum likelihood method did better than the least squares method. Exponentially Weibull distribution fits best. (See Pic. 13)

7. Statistical tests (2 at least).

```
expon_gen :
KstestResult(statistic=nan, pvalue=nan)
Power_divergenceResult(statistic=74536.6929661791, pvalue=0.0)

exponnorm_gen :
KstestResult(statistic=0.49640910885425105, pvalue=0.0)
Power_divergenceResult(statistic=74536.6929661791, pvalue=0.0)

exponweib_gen :
KstestResult(statistic=0.7714849359517226, pvalue=0.0)
Power_divergenceResult(statistic=74536.6929661791, pvalue=0.0)

exponpow_gen :
KstestResult(statistic=0.7378329097040996, pvalue=0.0)
Power_divergenceResult(statistic=74536.6929661791, pvalue=0.0)
```

Pic. 24 — MLT

```
expon_gen :
KstestResult(statistic=0.8810880240484631, pvalue=0.0)
Power_divergenceResult(statistic=74536.6929661791, pvalue=0.0)

exponnorm_gen :
KstestResult(statistic=0.8810880240484631, pvalue=0.0)
Power_divergenceResult(statistic=74536.6929661791, pvalue=0.0)

exponweib_gen :
KstestResult(statistic=0.8810880240484631, pvalue=0.0)
Power_divergenceResult(statistic=74536.6929661791, pvalue=0.0)

exponpow_gen :
KstestResult(statistic=0.8810880240484631, pvalue=0.0)
Power_divergenceResult(statistic=74536.6929661791, pvalue=0.0)
```

Pic. 25 — LSM

A small p-value indicates that the observations are inconsistent with the Null hypothesis. That means that the data doesn't seem to be close to any chosen distribution.

Conclusions

As a result of the work, we've worked with several tools of visualization: Histogram to see the distribution of game releases by years, boxplots to output games distribution by platforms and by genres.

Appendix

DataLore: site. – URL: https://datalore.jetbrains.com/notebook/RemqSkuJwmr1PM4Gc3cBqB/2db3tlCHNUwdln1bxvtbVT/ (circulation date: 07.11.2022)