FEDERAL STATE AUTONOMOUS EDUCATIONAL INSTITUTION OF HIGHER EDUCATION
ITMO UNIVERSITY

Report on learning practice № 2

Analysis of multivariate random variables

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

Saint-Petersburg
2022

Goals

1. Plotting a non-parametric estimation of PDF in form of a histogram and kernel density function for MRV (or probability law in case of discrete MRV).
2. Estimation of multivariate mathematical expectation and variance.
3. Non-parametric estimation of conditional distributions, mathematical expectations and variances.
4. Estimation of pair correlation coefficients, confidence intervals for them and significance levels.
5. Task formulation for regression, multivariate correlation.
6. Regression model,multicollinearity and regularization (if needed).
7. Quality analysis.

Brief theoretical part

In this lab we are going to work with a multivariate distribution; In general multivariate distribution show comparisons between two or more measurements and the relationships among them. For each univariate distribution with one random variable, there is a more general multivariate distribution. For example, the normal distribution is univariate and its more general counterpart is the multivariate normal distribution. While the multivariate normal model is the most commonly used model for analyzing multivariate data, there are many more: the multivariate lognormal distribution, the multivariate binomial distribution, and so on.

Below are listed measures that we are going to work with in this lab:

- Covariance of MRV is a measure of linear dependency of two random variables.

$$K_{\xi_i\xi_j} = M[(\xi_i - M[\xi_i])(\xi_j - M[\xi_j])] =$$
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(\xi_i - M[\xi_i])(\xi_j - M[\xi_j])f_{\xi_i\xi_j}(z_i, z_j)dz_i dz_j$$

- MRV mathematical expectation:

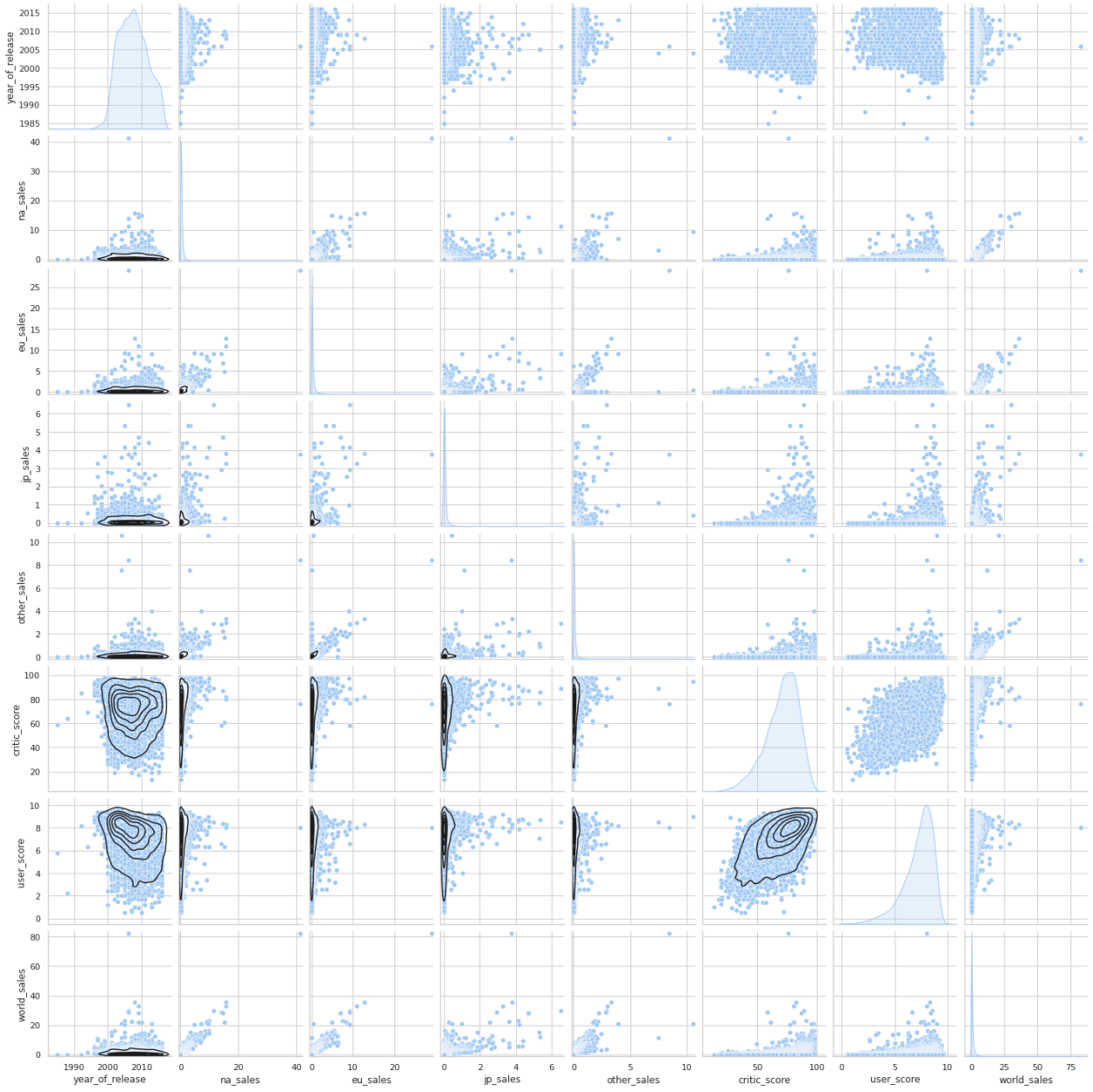
$$m_{\xi_i} = M[\xi_i] = \int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} z_i f_{\Xi}(z_1, \dots, z_n)dz_1 \dots dz_n = \int_{-\infty}^{\infty} z_i f_i(z_i)dz_i$$

- MRV variance:

$$D_{\xi_i} = M[(\xi_i - M[\xi_i])^2] = \int_{-\infty}^{\infty}(z_i - M[\xi_i])^2 f_i(z_i)dz_i$$

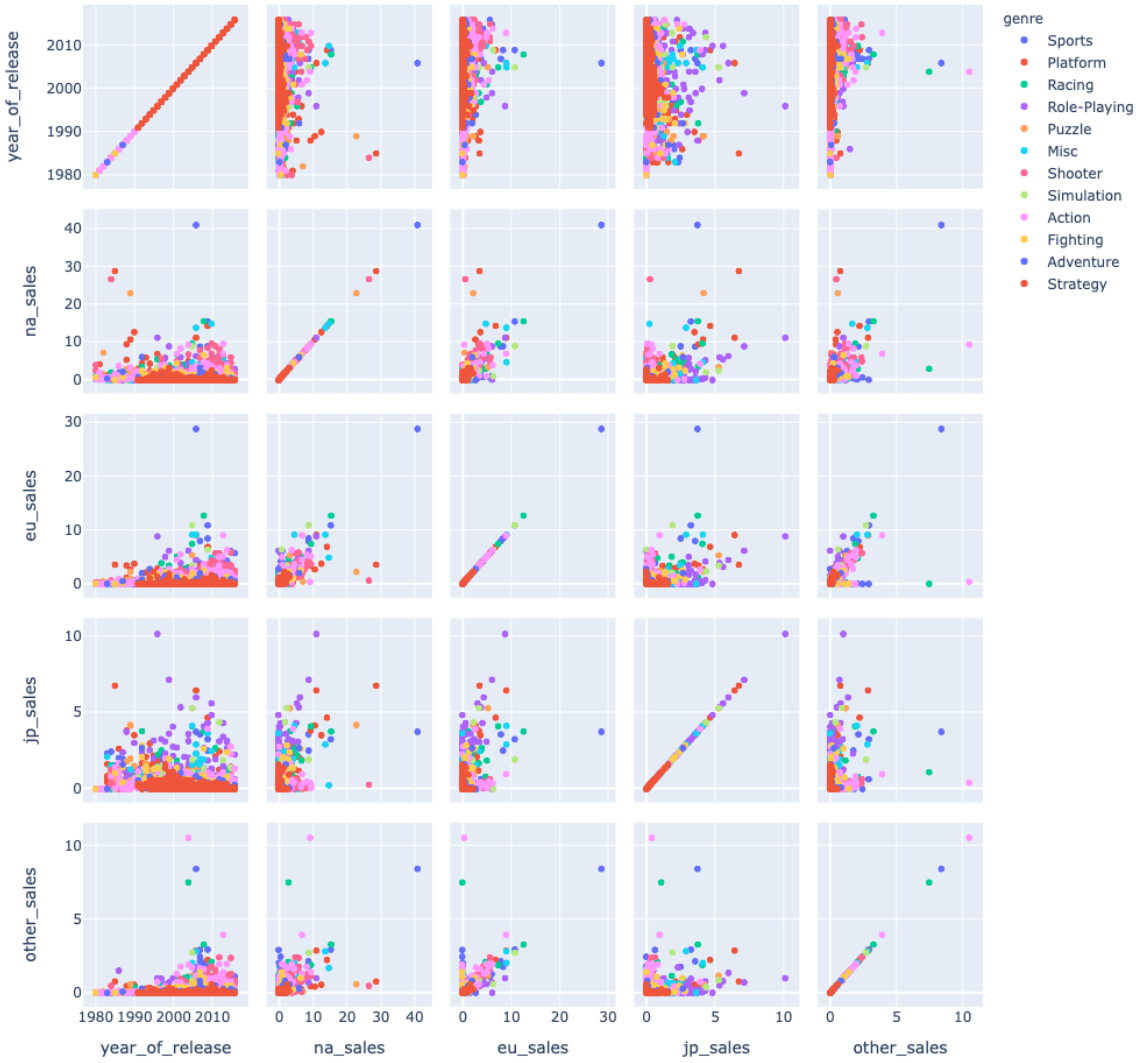
Results

Plotting a non-parametric estimation.

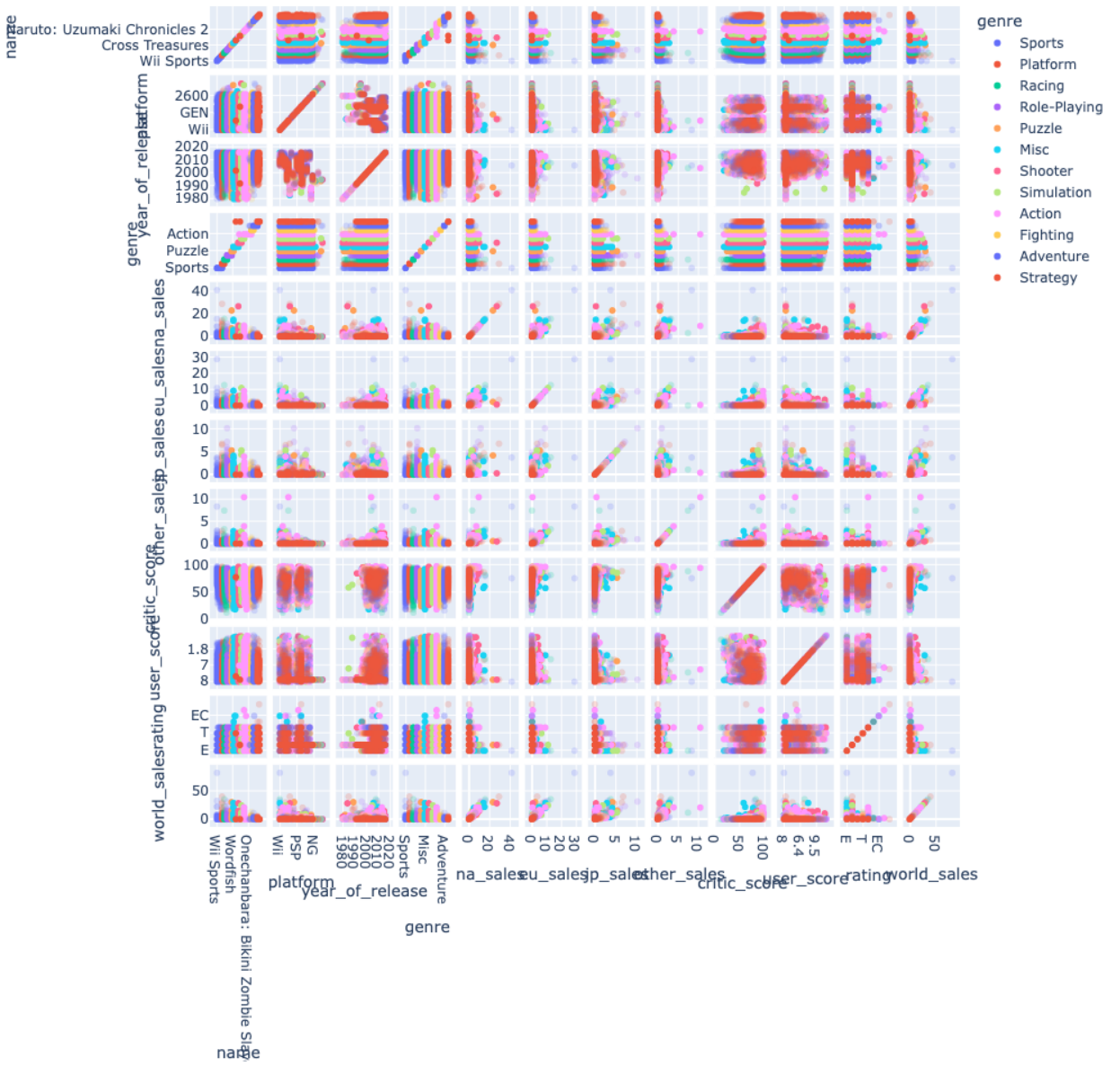


Pic. 1 — Non-parametric estimation of PDF

On the heatmap and sub-histograms we can see a huge increase of game production between 2008 and 2010 years and that most popular genre is Action.



Pic. 2 — Estimation of PDF



Pic. 3 — Estimation of PDF

This parametric plot shows that with the years Platform genre becomes the most popular, and the roots of this comes from Japan. Strategy is the least popular genre among the listed.

Estimation of multivariate mathematical expectation and variance.

	count	mean	std	min	25%	50%	75%	max
year_of_release	16444.0	2006.486256	5.875525	1980.0	2003.00	2007.00	2010.00	2016.00
na_sales	16444.0	0.264012	0.818378	0.0	0.00	0.08	0.24	41.36
eu_sales	16444.0	0.145930	0.506716	0.0	0.00	0.02	0.11	28.96
jp_sales	16444.0	0.078487	0.311100	0.0	0.00	0.00	0.04	10.22
other_sales	16444.0	0.047594	0.188005	0.0	0.00	0.01	0.03	10.57
world_sales	16444.0	0.536023	1.558786	0.0	0.06	0.17	0.47	82.54

Pic. 4 — Estimation of multivariate mathematical expectation (second col)

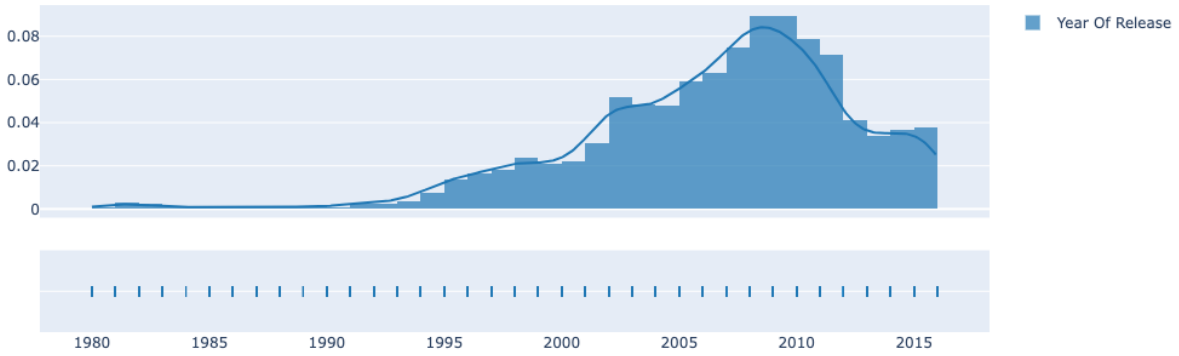
Despite Worlds Sales the highest sales can be seen in North America.

	.3
year_of_release	34.52179614079422
na_sales	0.6697423013725736
eu_sales	0.2567614829661589
jp_sales	0.096783382364229…
other_sales	0.035346024495358…
world_sales	2.4298131165200507

Pic. 5 — Estimation of multivariate variance

World Sales has the highest spread of sales, probably we can assume that it depends on the population of the country.

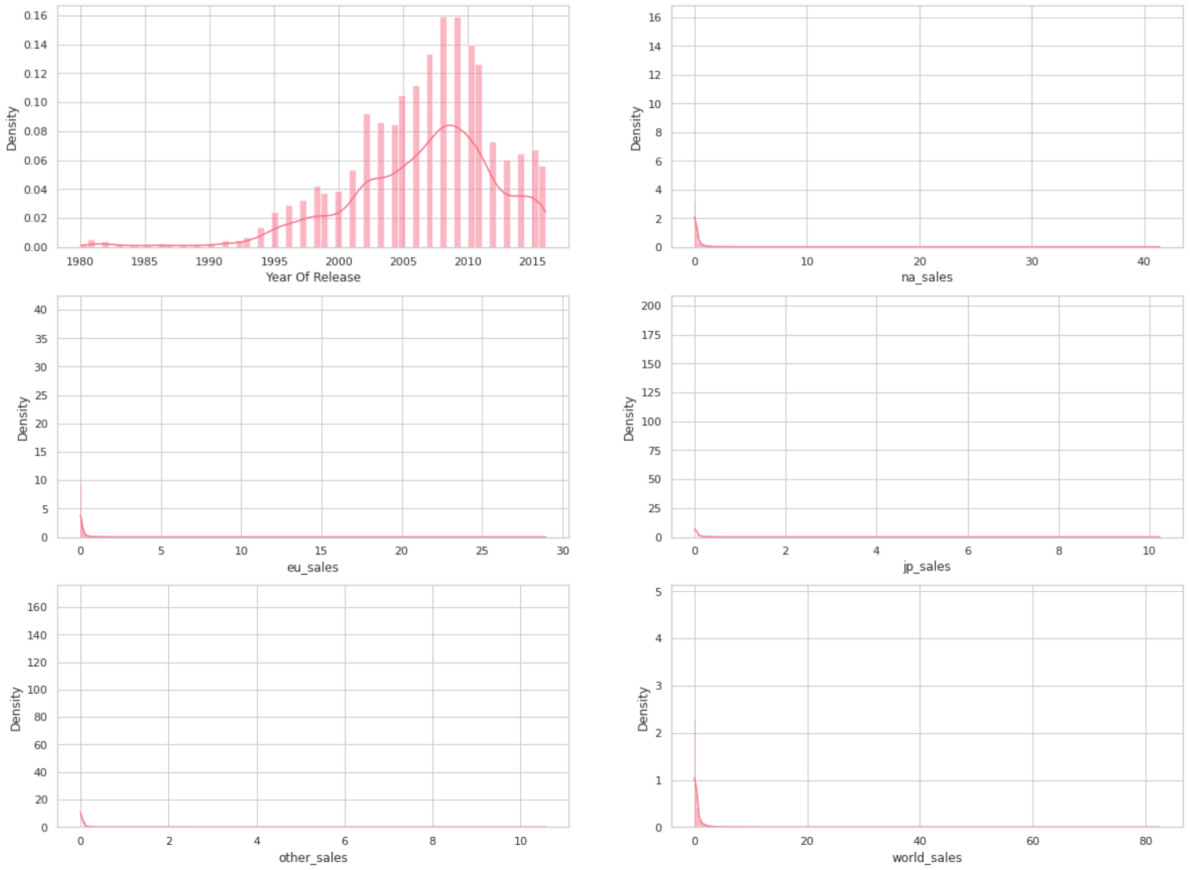
Non-parametric estimation of conditional distributions, mathematical expectations and variances.



Pic. 6 — Distribution of Year Of Release (plotly)

Pic. 7 — Distributions of different sales regions zoomed in (plotly)

Because World Sales usually starts lately than in NA it's more spread through the time, and it's more possible that the trend of the success game is that it's was developed and started selling in NA.



Pic. 8 — All of the provided distributions (seaborn)

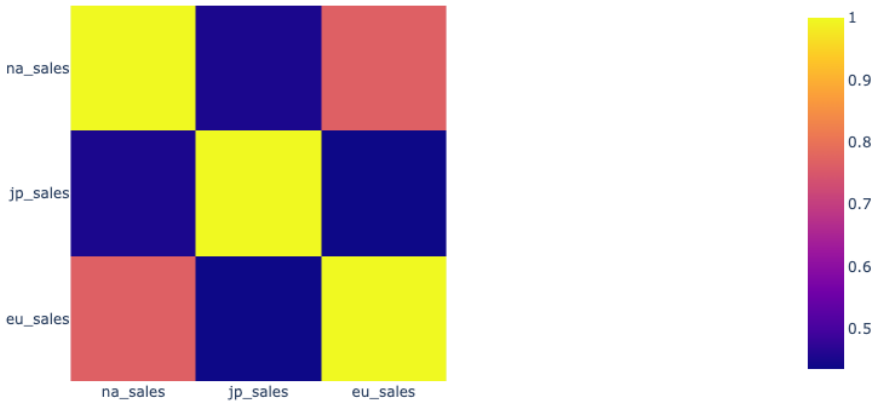
	year_of_release = 2000	year_of_release = 2005	year_of_release = 2010	year_of_release = 2015
genre	nan	nan	nan	nan
na_sales	0.385686274509804	0.3169928825622776	0.49475638051044085	0.30886877828054293
eu_sales	0.2470588235294118	0.15427046263345195	0.3020417633410673	0.27647058823529413
jp_sales	0.11049019607843137	0.06802491103202846	0.05844547563805104	0.053981900452488706
other_sales	0.053823529411764715	0.055249110320284706	0.10266821345707658	0.08601809954751131
world_sales	0.7970588235294119	0.5945373665480427	0.9579118329466356	0.7253393665158372
critic_score	72.17647058823529	70.20284697508897	69.11600928074246	73.07692307692308
user_score	7.53627450980392	7.514590747330961	6.883294663573086	6.863800904977376
rating	nan	nan	nan	nan

Pic. 9 — Conditional mathematical expectations

	year_of_release = 2000	year_of_release = 2005	year_of_release = 2010	year_of_release = 2015
genre	nan	nan	nan	nan
na_sales	0.3306564550572703	0.38386812472643533	1.3156784870231482	0.2826446236116824
eu_sales	0.14767047175305767	0.38838993028463403	0.3601930308099067	0.3420647593582888
jp_sales	0.09311757911085222	0.17942657335337886	0.03817176388064533	0.026082254216371868
other_sales	0.004061473500291205	0.031101631238066238	0.043321468731452005	0.0287386178527355
world_sales	1.1431278974956316	2.815433029795548	3.6447058619759347	1.4501268161250511
critic_score	258.97845078625505	173.3170685291263	214.47023147898344	153.3895104895105
user_score	1.9876810328091639	1.8907314721423996	1.9554877245993636	2.0409563965446322
rating	nan	nan	nan	nan

Pic. 10 — Conditional variances

Estimation of pair correlation coefficients, confidence intervals for them and significance levels.



Pic. 11 — Correlation coefficients

Thus, sales in North America have a high correlation with sales in Japan.

```
Correlation Coefficient: 0.4511619582349975
Significance Level: 0.0
Confidence Interval: (0.07373169923535364, 0.08324227303416715)
```

Pic. 12 — Correlation between `jp_sales` and `na_sales`

As a proof for information that heatmap shows we calculate a confidence interval for a correlation between `jp_sales` and `na_sales` and its same number.

Task formulation for regression, multivariate correlation.

We are going to predict "sales". Let's generate dummies for some task.

```
pd.get_dummies(df['genre'])
```

	Action	Adventure	Fighting	Misc	Platform	Puzzle
0	0	0	0	0	0	0
1	0	0	0	0	1	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	1
6	0	0	0	0	1	0
7	0	0	0	1	0	0
8	0	0	0	0	1	0

16444 rows x 13 columns Jump to top Jump to bottom

Pic. 13 — Dummy data

Regression model,multicollinearity and regularization (if needed).

```
from sklearn.model_selection import train_test_split

# Split data

df_X = data.drop(columns='critic_score')

df_y = data['na_sales']

X_train, X_test, y_train, y_test = train_test_split(df_X, df_y, test_size=0.33,
random_state=42)

from sklearn.linear_model import LinearRegression

# Teach regressor

cls_lr = LinearRegression()

cls_lr.fit(X_train, y_train)

# Predict

y_pred= cls_lr.predict(X_test)
```

	.3
0	0.36
1	0.139999999999999…
2	0.219999999999999…
3	0.1999999999999998
4	0.179999999999999…
5	0.119999999999999…
6	0.54
7	-2.63677968348474…
8	1.4800000000000004

Pic. 14 — `y_pred` output of the first 9 rows

In order to do a multicollinearity analysis, we can do several things

- *Remove some attributes based on correlations*
- *Do regularization with the help of:*
  - *Lasso.*
  - *Ridge.*

For further work we decided to use L1-regularization (Lasso) that will protect us against unnecessary features

```
[ 0.90617408 -0.         -0.         -0.          0.          0.
 -0.          0.         -0.          0.         -0.          0.
 -0.          0.          0.         -0.        ]
```

Pic. 15 — Lasso regularization

	.3
0	1.0000000000000013
1	0.0
2	0.0
3	0.0
4	0.0
5	0.0
6	0.0
7	0.0
8	0.0

16 rows x 2 columns

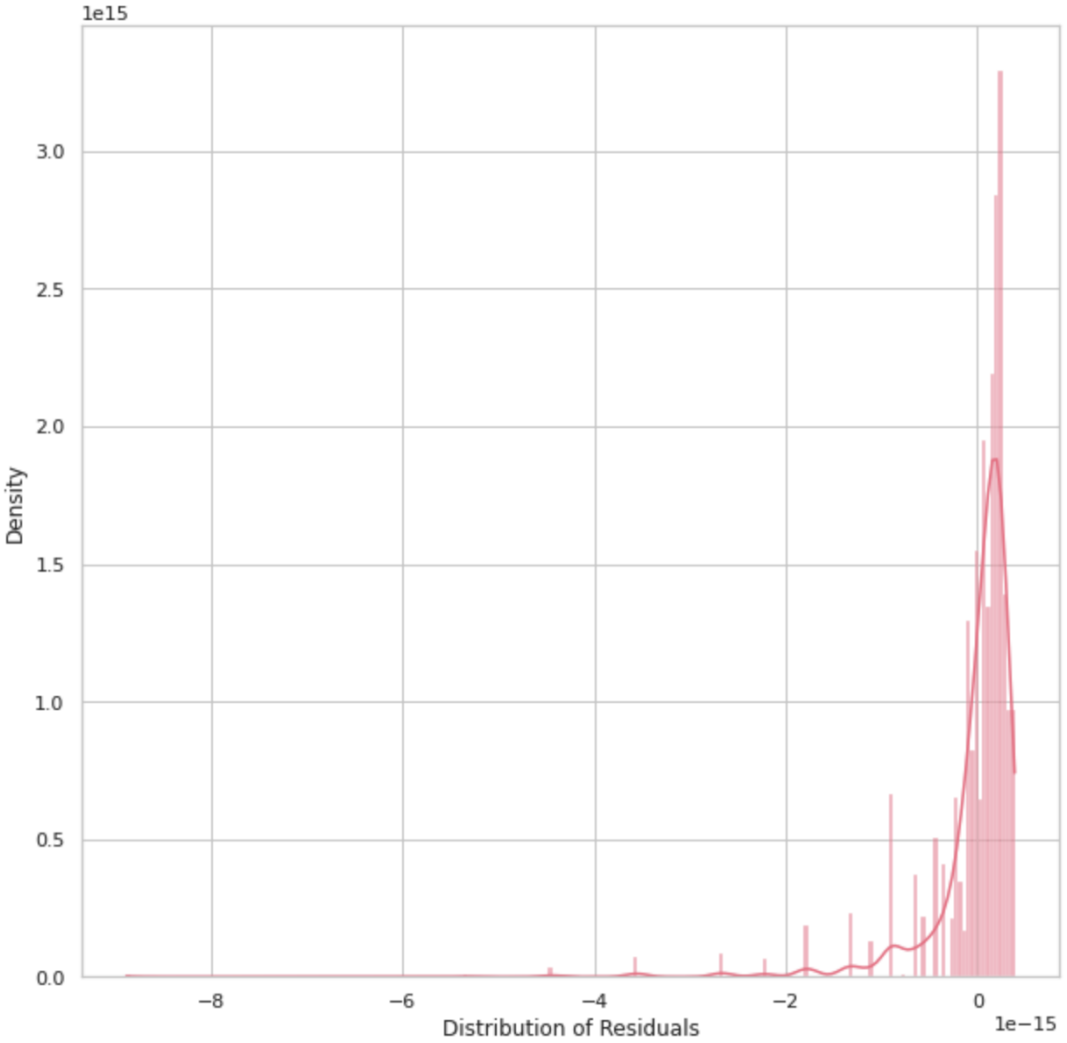
```
Mean absolute error with lasso =  0.036928802088366 35
Mean squared error with lasso =  0.0054036472053904 87
Mean absolute error with aic lasso =  5.292289744278036e-16
```

Pic. 16 — Metrics with lasso

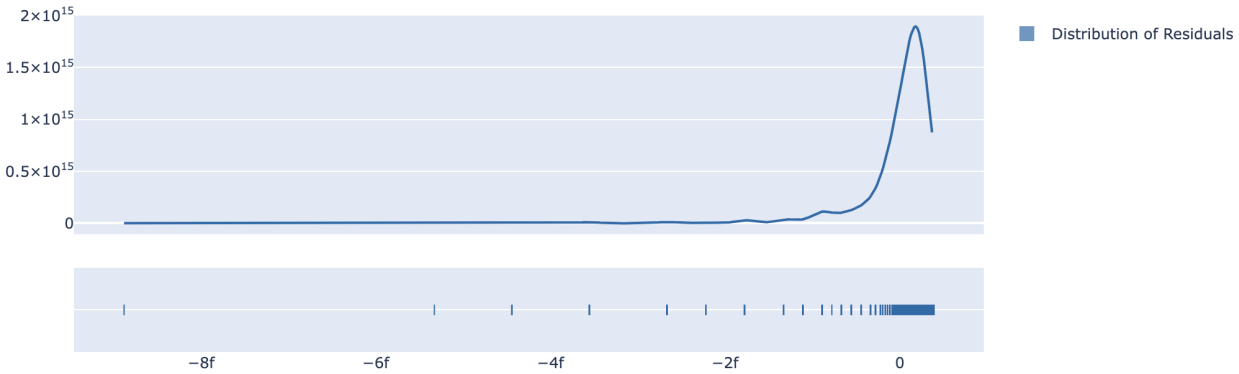
```
MAE 2.761617957328316e-16

MAPE 0.07099109131403394

MSE 2.729927951224047e-31

RMSE 5.224871243604045e-16

r2 score  1.0
```

Pic. 17 — Metrics

Regularization shows better performance in errors which are less than hundreds (Pic. 14), however without it errors are quite big which (Pic. 15) is bad for the ML-tasks.



Pic. 18 — Residuals distribution



Pic. 19 — Same residuals distribution but in plotly with subplot

	0	1
na_sales	1.000000e+00	1.000000e+00
genre_Action	-1.492573e-16	1.730990e-17
genre_Adventure	-1.705250e-16	1.841318e-16
genre_Fighting	-1.535784e-16	1.448574e-16
genre_Misc	-2.780377e-17	2.553100e-16
genre_Platform	-6.661445e-18	2.952490e-16
genre_Puzzle	-2.580843e-16	1.601266e-16
genre_Racing	-6.496825e-17	1.734156e-16
genre_Role-Playing	-1.764620e-16	6.283758e-17
genre_Shooter	-4.233433e-16	-2.045182e-16
genre_Simulation	-2.975602e-16	2.756678e-17
genre_Sports	-1.836944e-16	1.781151e-17
genre_Strategy	-2.554121e-16	1.271917e-16
rating_E	-1.199421e-16	1.442208e-17
rating_M	-2.128650e-16	-2.891203e-17
rating_T	-1.479016e-16	-2.459492e-17

Pic. 20 — Confidence interval of regression coefficient

Average user ratings for Xbox One and PC platforms are the same Let's introduce the null and alternative hypotheses:

The average user ratings of the platforms are the same Average user ratings of the platforms are different

Level of significance:

Let's calculate the p-value. If the p-value is less than the chosen significance level (), there will be grounds to reject the null hypothesis of equality of averages in favor of the alternative hypothesis. Otherwise we conclude that the data did not allow us to reject the null hypothesis.

```
clear_score = dfa.query('user_score != "unknown"')
clear_score['user_score'] =
clear_score['user_score'].astype(float)

x_one = clear_score.query('platform == "XOne"')
pc = clear_score.query('platform == "PC"')
alpha = .05

results = stats.ttest_ind(x_one['user_score'], pc['user_score'],
equal_var=False)
```

```
print('p-value:', results.pvalue)

if (results.pvalue < alpha):
   print('There are reasons to reject the null hypothesis')
else:
   print('Not enough evidence to reject the null hypothesis')
```

We got the following results p-value equals to 0.116 and not enough evidence to reject the null hypothesis.

Conclusions

As a result of the work, we have built a pairplot heat map to show the correlation between features, and also we have worked with measures of multivariate random variables to show the scatter of certain events and proposed our theory of their causes. And the test showed that

1. With a probability of 12%, the result can be obtained by chance. There are no statistically significant differences.
2. On the available data at the 1% level of significance there is insufficient evidence to reject the null hypothesis in favor of the alternative hypothesis.
3. The average user ratings for Xbox One and PC platforms are the same.

Appendix

DataLore: site. – URL: https://datalore.jetbrains.com/notebook/RemqSkuJwmr1PM4Gc3cBqB/2db3tlCHNUwdln1bxvtbVT/ (circulation date: 14.11.2022)