**PROJECT WORK**

**Theme: “**Analysis of AB test with ARPPU/ARPU calculation”

**Course: “Date Analyst” Skillbox**

**Done by:** Arina Tsoy

**Supervisor:** Karlushin. N

**Almaty, 2024**

**Contents**

[Part 1. 4](#_Toc174089719)

[**Aims of the project.** 4](#_Toc174089720)

[Part 2. 4](#_Toc174089721)

[**Resource analysis.** 4](#_Toc174089722)

[Part 3. 6](#_Toc174089723)

[**Data cleaning.** 6](#_Toc174089724)

[Part 4. 11](#_Toc174089726)

[**Usage of statistic methods.** 11](#_Toc174089727)

[Part 5. 13](#_Toc174089729)

[**Report.** 13](#_Toc174089730)

[Conclusion 17](#_Toc174089732)

[Bibliography 18](#_Toc174089733)

[References 18](#_Toc174089734)

# **Part 1.**

## **Aims of the project.**

**Theme of the Project: “Analysis of AB test with ARPPU/ARPU calculation”**

**Aim of the project:** To reveal the usefulness of further promotion

To achieve the goal of the project, the steps will be described in the practical part sequentially. In the practical part of the project, steps will be performed to analyze and clean up the data. Statistical methods such as confidence intervals will also be used in the work.

**Tasks of the practical part of project:**

1. Analysis of resources
2. Data cleaning
3. Statistic methods
4. Preparation of the report

# **Part 2.**

## **Resource analysis.**

Python and SQL are two crucial tools within the framework of data analysis. However, python was picked to carry out this work.

Python, a versatile and widely used programming language in the data science community, has gained popularity due to its extensive library support, rich ecosystem, and flexibility in handling diverse data formats. Libraries such as Pandas, NumPy, and Matplotlib empower data analysts to easily perform complex data manipulation, statistical analysis, and data visualization tasks.

#### **Key strengths of Python include:**

**1. Rich Ecosystem of Libraries:** Python's vast collection of libraries provides a wide range of specialized tools for data analysis tasks. These tools cover areas such as data manipulation, statistical analysis, machine learning, and data visualization.

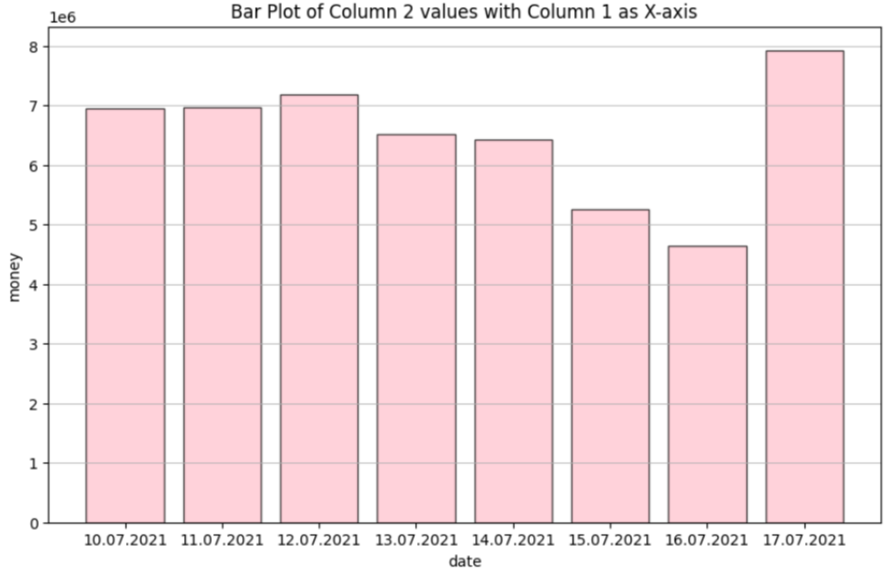
**2. Flexibility and Extensibility:** Python's flexible nature allows data analysts to integrate different data sources and perform custom data transformations. This flexibility enables the development of bespoke analytical solutions that can be tailored to meet specific business needs.

**3. Advanced Analytics Capabilities:** Python libraries provide advanced analytics tools for tasks such as predictive modeling, clustering, and natural language processing. These capabilities expand the scope of data analysis beyond traditional SQL-based methods.



Pic.1 The source code for uploading data





Pic.2 The visualization via matplotlib library in the project.

In Summary, using Python in this project is an great choice due to its versatility and the significant role that its libraries play in the analysis process. The extensive Python ecosystem with powerful libraries such as Pandas(pic.1) and Matplotlib(pic.2) provides efficient data processing, accurate calculation of confidence intervals and clear visualization of data. These tools have not only simplified complex tasks, but also expanded general analytical capabilities for working on this project.

# **Part 3.**

## **Data cleaning.**

According to the description of the study case, there are cheaters among the players — players who, by hacking the game, charge themselves large amounts of in-game currency. There is a list of known cheaters among the existing data, but there are also not yet caught cheaters whose results may affect the calculations.

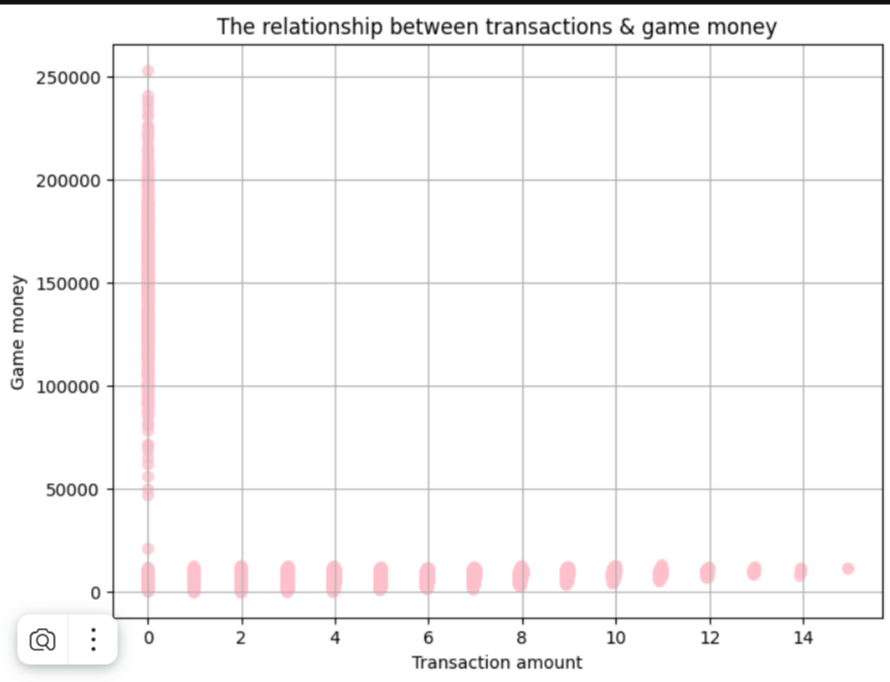
The first stage of the project task was to remove known cheaters and search for implicit cheaters. After reviewing the data frame with players information, all users marked cheater = 1 were moved into new data frame. The construction of an in-game currency boxplot among players and basic metrics still showed a high number of anomalies in the dataset. To identify implicit cheaters, first of all, it was necessary to analyze the behavior of already known hackers well.

In order to accomplish task, visualizations of in-game currency spending were created for the days from July 10 to July inclusive(Pic.2). According to the received schedule:

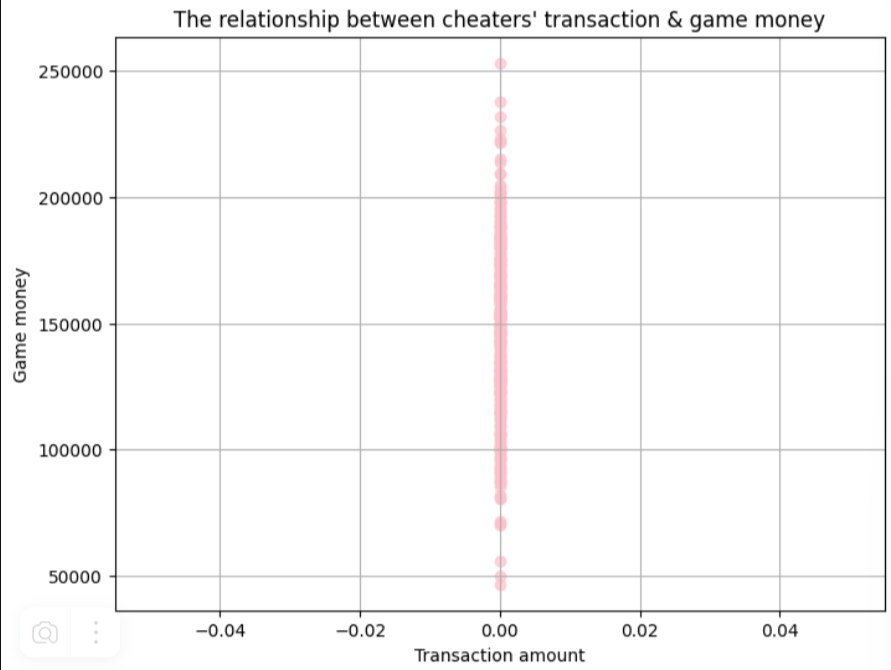
*- “The highest incidence of cheating occurred on July 17th, 2021 and July 12th, 2021, while the lowest was observed on July 16th, 2021.”* (Cheaters elimination stage)

Then the relationship between transactions & game money scatter plot was made with correlation between in game money and real currency(pic.3). According to the following data:

*- “It has been observed that cheaters show a transaction amount of 0.00, indicating they have not made any real-money purchases for in-game cash¶”.* (Cheaters elimination stage)



Pic.3 The relationship between transactions & game money scatter plot 1



Pic.3 The relationship between transactions & game money scatter plot 2

After calculating the main metrics (mean, median, minimum, maximum values, standard deviation, etc.) and gathering analysis from visual metrics it was already possible to see the most possible description of the cheater:

1. The number of real money transfers is zero with high rates of in-game currency (However, in this case this is not one true option, since it must be taken into account that players can receive in-game currency If they win tournaments)
2. High indicators of in-game currency on the last day of the promotion on the 17th and low or zero indicators of in-game currency on the 16th.

Cheater’s elimination was accomplished via obtaining basic metrics and then testing IQR & 3-Sigma methods.

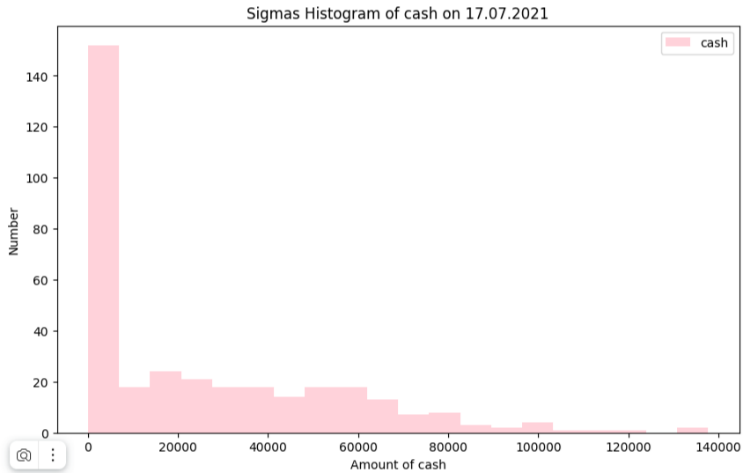
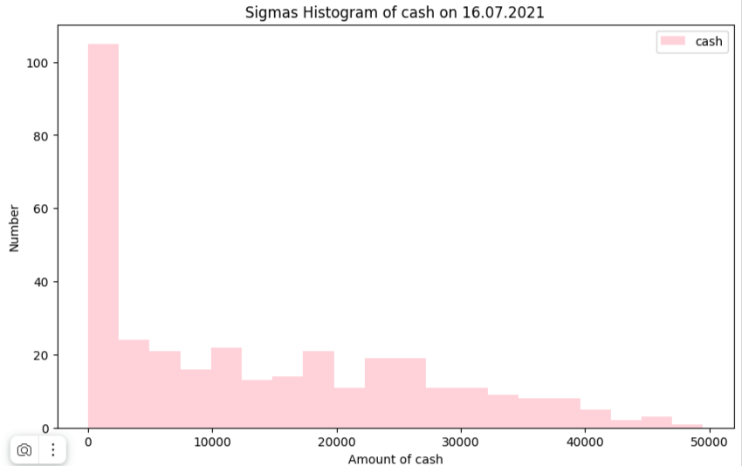
The basic metric analysis showed:

1. The presence of outliers, particularly the maximum value of 240,950, suggests that some users have accumulated substantially more in-game currency than most others, potentially due to special behaviors.
2. The distribution of “game\_cash” is likely skewed to the right, with a long tail of users holding large amounts of in-game currency.

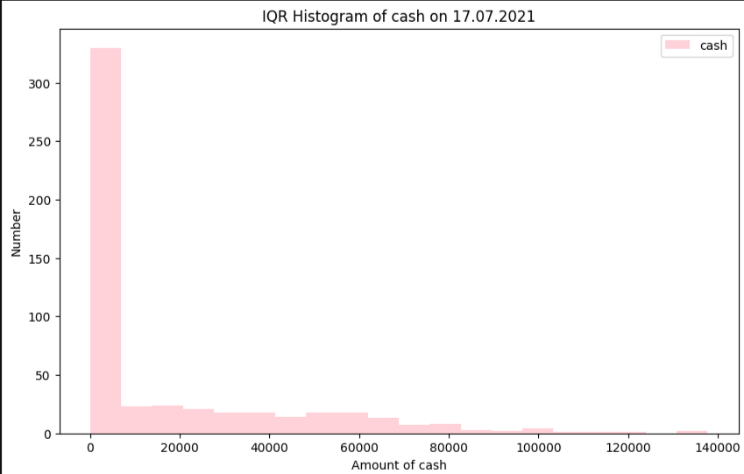
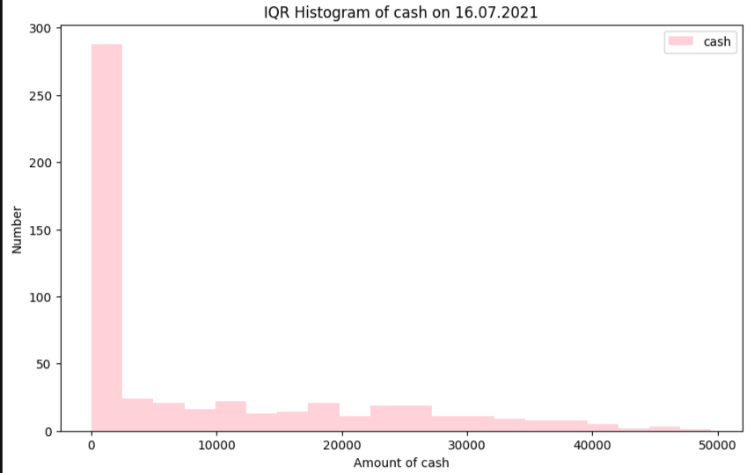
Then the IQR & 3-Sigma methods were run (Pic.4). The results then were compared via histograms where results were presented by the most distinguishing days 16th and 17th (Pic.5, Pic.6)

**

Pic.4 3-Sigma & IQR methods



Pic.5 The 3-sigma method

****

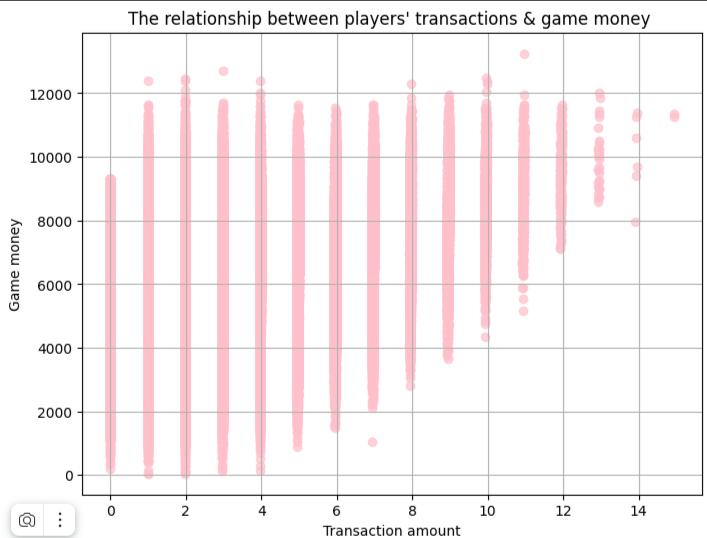
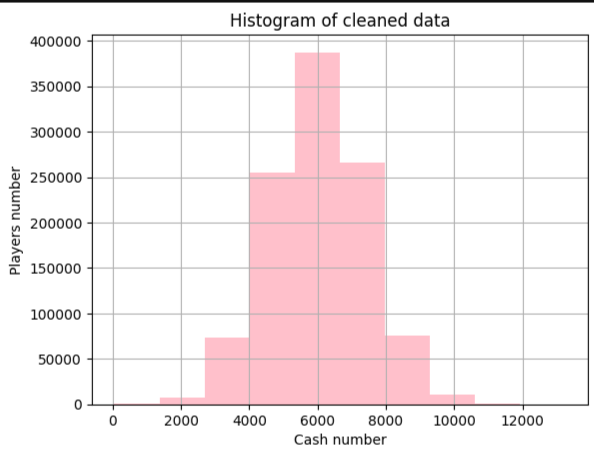
Pic.6 The IQR method

That analysis led to such conclusion:

*- “The results are quite similar, both IQR and 3-sigma methods identified anomalies in game cash.*

*According to the charts of the distribution of game cash by day: 07/17/2021 peak of cheater activity, 07/16/2021 - decline in cheater activity, the results of the IQR method give the most coincidences.”* (Cheaters elimination stage)

After elimination data in the scatter plot looks more natural, while the data in the histogram appears symmetric.(Pic.7)



Pic.7 Data after removing anomalies

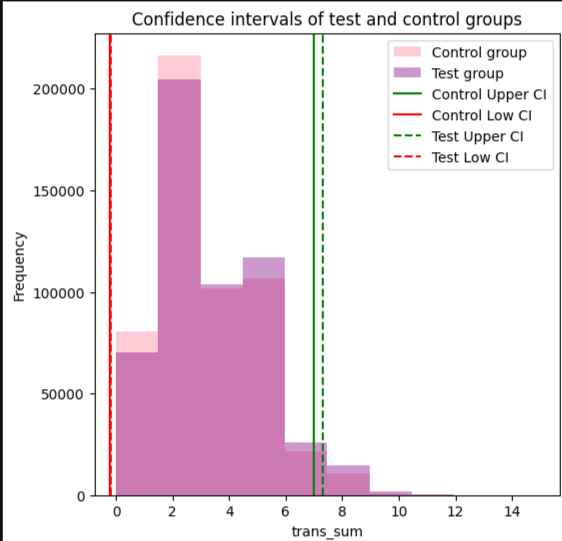
# **Part 4.**

## **Usage of statistic methods.**

The next major process of the project was to calculate confidence intervals. To perform that stage, data frames with information about the platforms, AB groups and the cleaned data from the previous stage were directly involved.

Test and control data frames were created for the first confidence intervals. The Scipy library with “the stats.norm.interval” function, the average and standard deviation for the column with real money were defined and used for the calculation.(Pic.8)





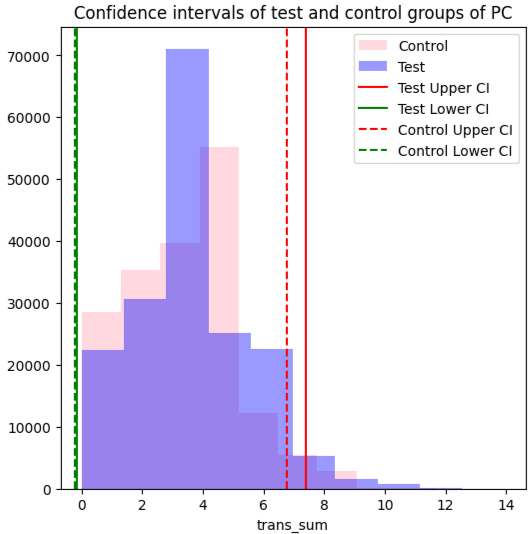
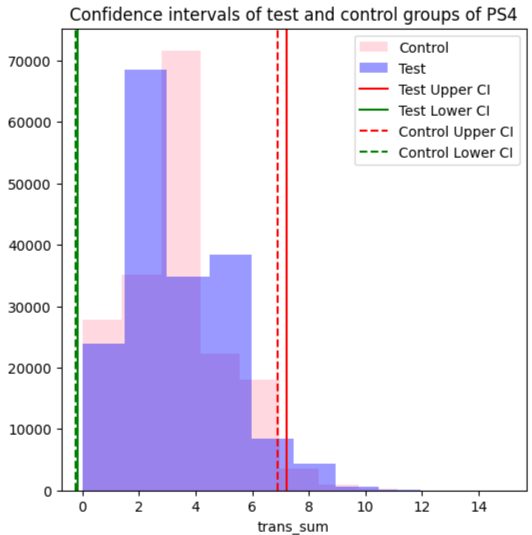
Pic.8 Confidence intervals function and diagram for test and control groups.

According to the diagram, the confidence intervals (CIs) for the control and test groups overlap significantly.

*- “Based on the metrics the percentiles and range (min to max) are quite similar between the two datasets, suggesting that the overall distributions of trans\_sum are alike. Also there is a slight difference in means, with test dataset having a slightly higher average trans\_sum and slightly higher variability (standard deviation) compared to control dataset.”* (Confidence intervals)

The next step was analyzing metrics and confidence intervals for each platform(PS4, PC, XBox) in test and control groups that resulted in such outcomes and graphics (Pic.9):

*-“ The test groups on all platforms have higher Mean values compared to the control groups, which may indicate the impact of the changes being tested. The standard deviation and wide confidence intervals indicate significant variability in the data. A large number of observations (Length) makes the samples representative, and low values of the standard error of the mean indicate high accuracy in estimating the mean.”* (Confidence intervals)

** **

# 

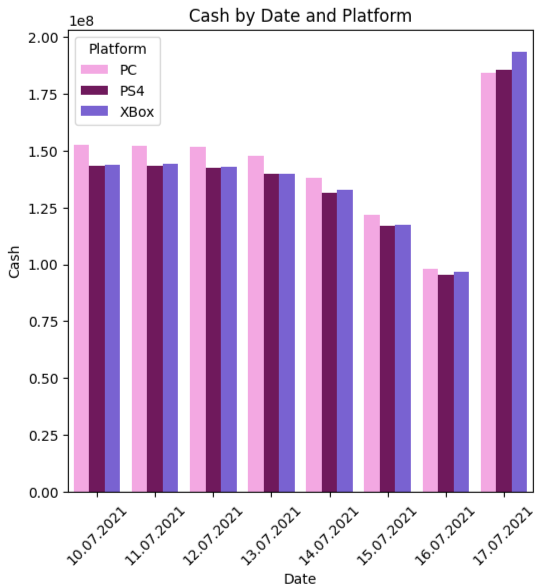
Pic.9 CI for each platform

Considering the analysis, the overlapping confidence intervals for each platform indicate that there is no statistically significant difference between them. This suggests that the promotion did not lead to distinguishable changes in transaction sums across the platforms

# **Part 5.**

## **Report.**

In order to write a complex repot of accomplished work different programs such as Power BI, Excel, Python were used (ARPU&ARPPU).The in game cash analysis showed which platform was in favor and in what days were the ups and downs in value.(Pic.10)



Pic.10 In game cash by date and platforms

The barplot of above conveys information regarding to the in game cash by date and platform. The first three days in game money parameter was holding steady to the same exact position with following decline as it come closer to the end day of promotion. That dip on 16th of July may be a result of networking issues or the end of the working week. However, the last day remarked as the peak of raise in game money. That result may be considered due to the last day of promotion or a weekend when people have more desire and spare time for playing on platforms.

Considering cash amount by platforms the PC leading was noticeable during all days but the last one when cash quantity from XBox suddenly surged. The PS4 values were gradualy declining from the start to the 16th of July with following rise on the last day.

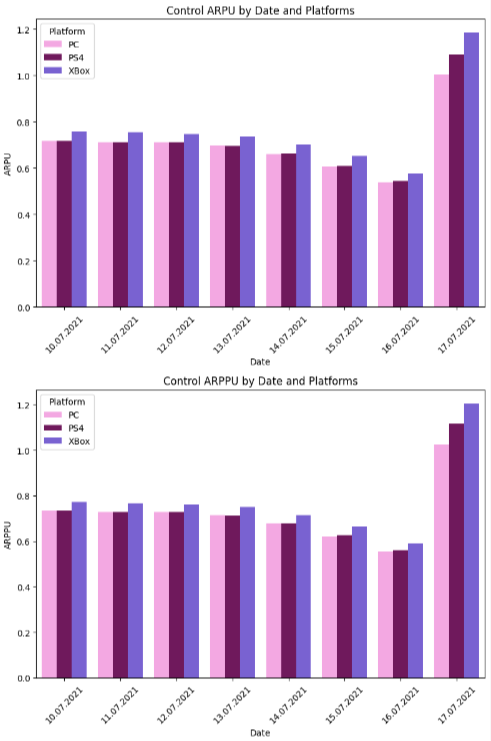
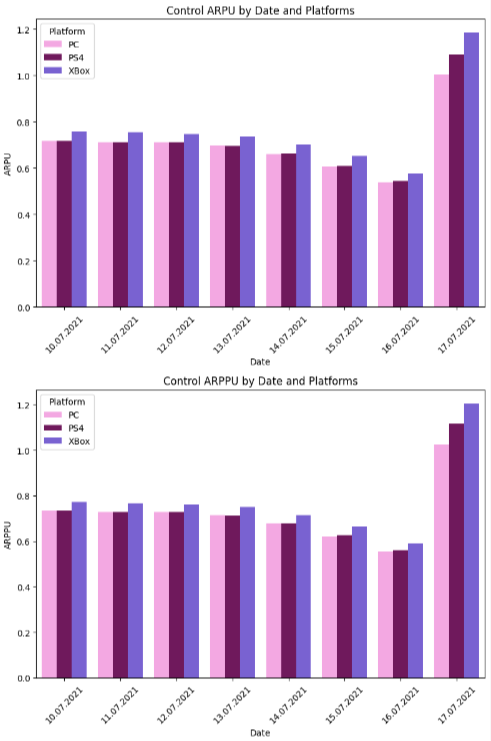
In the report comparison and calculation of ARPU/ARPPU were significant as well. Metrics were founded by the basic formulas:

**ARPPU= Total Revenue​ / Total Number of Paying Users**

**ARPU= Total Revenue / Total Number of Users​**

Data was visualized for brief analysis in pandas using seaborn and matplotlib library (Pic.11):



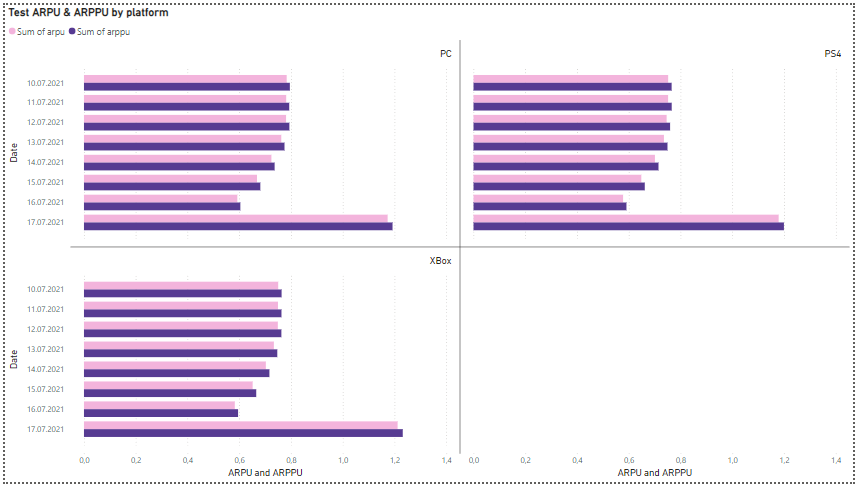


Pic.11 Code and diagrams for control group in pandas

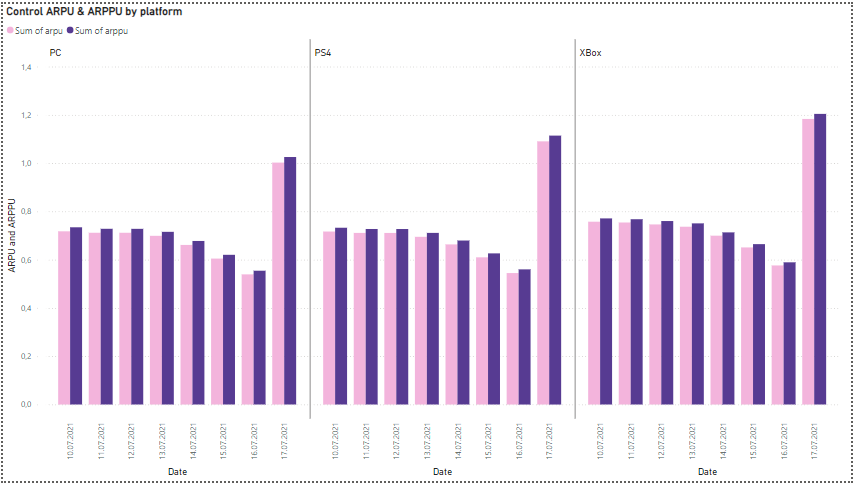
More convenient visuals were made in Power BI using clustered bar chart and clustered column chart. Were made 7 pieces within 3 in total. The first one reflects the results of ARPU/ARPPU by groups (Pic.12), the second one presents test ARPU/ARPPU by platform and date (Pic.13) while the last one shows control ARPU/ARPPU by platform and date (Pic.14). ( (Power BI visuals))

# 

Pic.12 ARPU/ARPPU by groups



Pic.13 test ARPU/ARPPU by platform and date



Pic.14 control ARPU/ARPPU by platform and date

As in in game cash analysis the peak of values is on the last day of promotion and remarkeble decline is on the 16th of July. ARPU and ARPPU for PC and PS4 stayed almost the same for 6 days but on the last day PS4 took a lead. The highest parameters were from XBox platform throughout whole ab testing.

Comparing results of test and control groups the difference is highly visible. The test PS4 and XBox values were quite the same for 4 days and then started gradually declining to the 16th of July. The PC ARPU/ARPPU were steadily declining however still remained as highest parameter amoung other platforms until 17.07.2021 On the last day in both ARPU and ARPPU barplots XBox leaped as in the control group.

To store and further use new data was gathered and structured in Excel’s Pivot tables. ( (Pivot tabels))

# **Conclusion**

According to the analysis, the campaign did not succeed, despite a slight increase in ARPU/ARPPU metrics, the overlapping of confidence intervals indicates the absence of a statistically significant difference between the groups, which means that the observed improvement could be accidental.

However, in addition for success analyze of the AB test, it would be useful to compare the flow of new customers or the behavior of old players, calculate the conversion rate to figure out how many people clicked on the proposed promotion. Based on the results, analyze the importance of promotion’s design. Could be possible to implement some change in layout. For instance, to add a catchier phrase, add bright colors or change the size of the buttons so that it is more convenient for users to make a purchase for the first time. Also, would be nice to conduct a survey of users about which armor or any other attribute the promotion is most expected for, and then conduct a new AB test.

In addition, another idea would be testing the campaign on holidays or weekends, since at this time most people are not busy with work, and students do not attend educational institutions and have more free time for games and accomplished analysis in this project showed a positive result in in game cash and ARPPU/ARPU on weekend.

# **Bibliography**

# **References**

Tsoy, A. (n.d.). *ARPU&ARPPU.* [NOOTE\Материалы для финальной работы (2)\Материалы для итоговой работы\final\arpu.ipynb](NOOTE/Материалы%20для%20финальной%20работы%20(2)/Материалы%20для%20итоговой%20работы/final/arpu.ipynb)

Tsoy, A. (n.d.). *Cheaters elimination stage.* [NOOTE\Материалы для финальной работы (2)\Материалы для итоговой работы\final\cheaters elimination.ipynb](NOOTE/Материалы%20для%20финальной%20работы%20(2)/Материалы%20для%20итоговой%20работы/final/cheaters%20elimination.ipynb)

Tsoy, A. (n.d.). *Confidence intervals.* [NOOTE\Материалы для финальной работы (2)\Материалы для итоговой работы\final\CI.ipynb](NOOTE/Материалы%20для%20финальной%20работы%20(2)/Материалы%20для%20итоговой%20работы/final/CI.ipynb)

Tsoy, A. (n.d.). *Pivot tabels.* [NOOTE\Материалы для финальной работы (2)\Материалы для итоговой работы\final\a\_platforms\_pivot.xlsx](NOOTE/Материалы%20для%20финальной%20работы%20(2)/Материалы%20для%20итоговой%20работы/final/a_platforms_pivot.xlsx)

[NOOTE\Материалы для финальной работы (2)\Материалы для итоговой работы\final\b\_platforms\_pivot.xlsx](NOOTE/Материалы%20для%20финальной%20работы%20(2)/Материалы%20для%20итоговой%20работы/final/b_platforms_pivot.xlsx)

Tsoy, A. (n.d.). *Power BI visuals.* [NOOTE\Материалы для финальной работы (2)\Материалы для итоговой работы\final\arpu\_arppu.pbix](NOOTE/Материалы%20для%20финальной%20работы%20(2)/Материалы%20для%20итоговой%20работы/final/arpu_arppu.pbix)

*.*