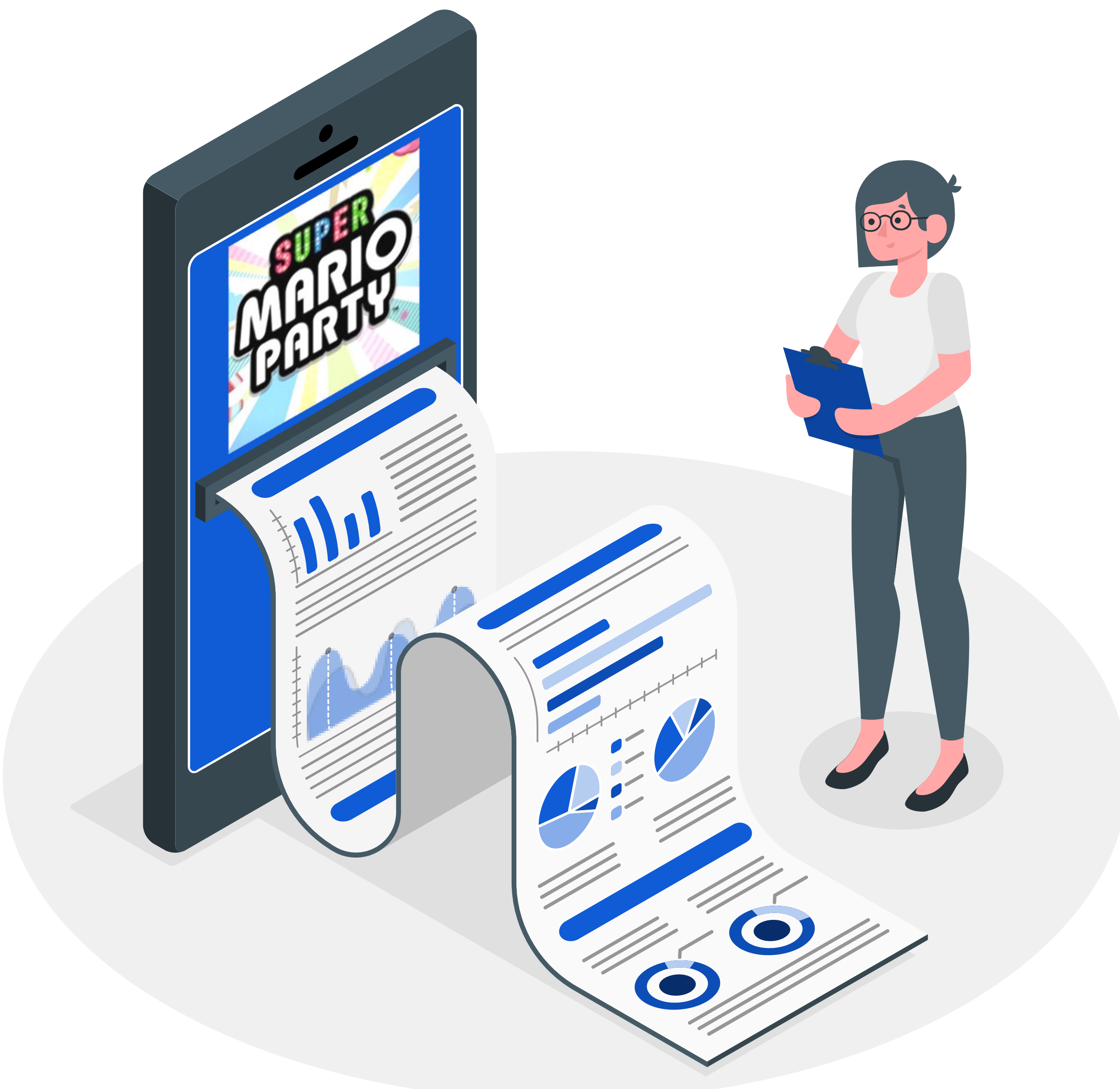
# Game On: Investigating the Impact of Team Changes and AI Integration on Dash and Dine's Players' Performance

Rowshanak HosseinzadehAttar

**11/20/2023**



# Table of contents

# 1. Introduction and motivation

The landscape of collaborative environments, both in the digital and real-world spheres, continually evolves with the integration of artificial intelligence (AI) and advancements in team dynamics. In this context, our study delves into a captivating exploration of the effects of team composition changes and AI integration on player performance within the Dash and Dine mini-game of Super Mario Party. The Dash and Dine mini-game serves as a microcosm of collaborative endeavors, allowing us to scrutinize the impact of key variables—namely, the introduction of AI teammates and alterations in team composition—on the overall performance of players.

Our primary focus is to ascertain whether team changes and the introduction of AI components result in discernible shifts in player performance. The initial phase witnessed human-only teams engaging in the Dash and Dine mini-game, in the second phase, teams underwent transformations—some members were substituted with AI entities, while others experienced changes in human composition. The control group remained untouched, allowing us to measure performance changes across the spectrum.

To assess the significance of performance differences between teams in each phase, a statistical framework was applied, including the utilization of Analysis of Variance (ANOVA) to explore variance between group means. Subsequently, Tukey's Honestly Significant Difference (HSD) post hoc test, a robust method for multiple comparisons, was employed to identify specific group differences. Through the application of these statistical tools, the investigation aimed to discern not only the existence of differences but also to identify groups exhibiting distinct performance patterns. Additionally, a thorough descriptive analysis was conducted.

The intriguing outcome of our analysis reveals that, while AI-integrated teams performed better in both phases, the control group, with a consistent human composition, exhibited the most substantial overall improvement. This observation suggests that familiarity and team synergy cultivated over the course of the first phase played a pivotal role in enhancing performance. Unraveling these patterns sheds light on the nuanced dynamics of collaborative teams, offering valuable insights into the interplay between human and artificial intelligence.

It is essential to underscore that the modest scale of this study imposes limitations on the extent of its generalizability. Nevertheless, the consistent enhancement in performance observed when AI became involved prompts valuable reflections on the potential of AI in collaborative settings.

## 2.1. Detailed Description of the Problem

The primary task in this study is to analyze the impact of team composition changes and the integration of artificial intelligence (AI) on player performance in the Dash and Dine mini-game of Super Mario Party. The study includes two distinct phases: the initial phase where human-only teams engage in the Dash and Dine mini-game, and the second phase where teams undergo transformations, including AI substitutions and changes in human composition. At the second phase teams are divided into 3 main groups, groups with no changes are labeled “control”, groups with AI substitutional are labeled "ai” and groups with human substitutional are labeled "newhire”.

## 2.2. Research Questions

Through this report the fallowing questions will be answered:

- How does the introduction of AI teammates influence player performance in the Dash and Dine mini-game?
- What is the effect of altering team composition on overall player performance?
- Are there discernible differences in performance patterns between teams with AI integration, human-only teams, and teams with unchanged compositions?

## 2.3. Technical Overview of Data

Data Source: The data originates from the gameplay of the Dash and Dine mini-game in Super Mario Party.

Collection Method: The number of total collected ingredients in each round of the mini-game is reported for each team.

Variables:

“totalingred”: The key variable is "totalingred," representing the total number of ingredients collected in each round. The scale level of the "totalingred" variable is quantitative, measured as counts of collected ingredients.

“team_id” and “group”: each team, consisting of four players has an unique “id”, using this variable it can be discovered how each team changed in the second phase. This variable is an integer.  “group” variable indicates how each team changed, there are three possibilities for this variable: ai, newhire, control. This variable is string.

Other variables: “phase” is an integer, it can be either 1 or 2, for each phase. “round” it is an integer from 1-12 for each round.

Missing Values: A preliminary examination indicates no missing values in the key variables.

# 3. Methods

## 3.1. overview

Working with a set of data, in order to find a pattern or extract a result, has a certain cycle known as CRISP-DM. The first three phases are "business understanding", "data understanding" and "data pre-processing". Which includes answering the questions: What is our problem? What data do we already have to solve that problem? What data do we need? How to get those data? How to clean and sort them so that they can be used for our specific needs?

These three steps have already been completed in this case. The problem has been explained in the previous section, and the data is almost ready for use.

## 3.3. Pre-processing

In this case, the given data set appears clean and ready to use, but there are still a few steps to be taken. This study used the Python programming language on the Jupyterlab platform. As a first step, the CSV file must be transformed into a data frame. A DataFrame is a data structure that organizes data into a 2-dimensional table of rows and columns. Dataframes are one of the most common data structures used in modern data analytics because they are a flexible and intuitive way of storing and working with data. Many necessary and useful functions can be implemented on data frames with Python programming, such as ".shape", ".head", ".tail", and others.

The next step is to examine the data in a general manner. By using the "data.shape" function, it can be seen that there are '600 rows and 5 columns, and by using "data.head(20)", the first 20 rows are shown. It is also important to identify the type of value each variable has, which can be determined by using the function "data.dtypes". It can be seen here that, except for "group", all the other variables are "int64" which is a good thing since many of the functions we will use in the future work with integers. While "group" is an object, it can be transformed into [-1,0,1] since it has 3 unique values, but only if it is necessary. It is then revealed by using the "groupby" function that 20 teams are labelled ai, 20 as control, but only 15 as newhire. Next, it is time to check if there are null blocks using ".isna()". Fortunately, there are none.

### 3.4.1. Visualizing

Data frames have the advantage of having the" describe" function. The "describe" function calculates all the important statistical indicators like sum, mean, min, max, Q1,Q2,Q3, and it is clear that the average is the same as Q2, so we have a generic view of the data. As useful as this information are, they would be better understood and more effective if they were visualized. Therefore, some of the most popular plots were used in this step, including a plot box

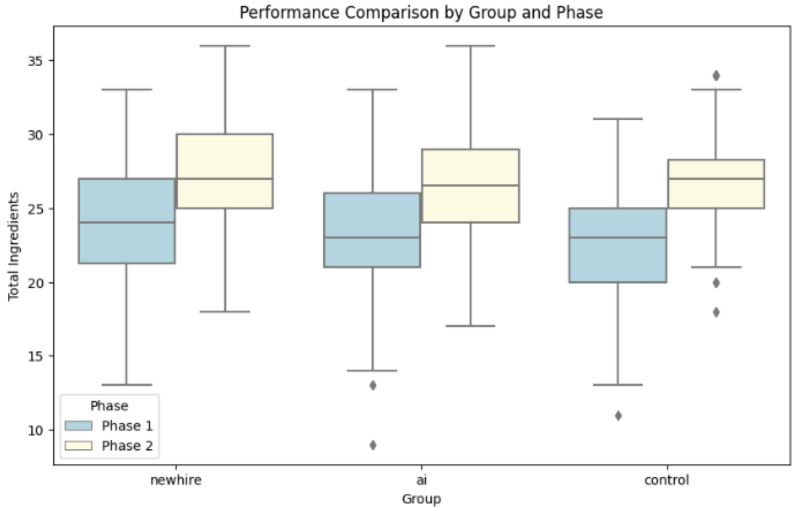



*Figure 1: Plot Box*

and a line plot, to gain a better perspective and check for outliers.

### 3.4.2. Plot Box

In general, Box plots are used to visualize numerical data by showing their quartiles graphically. This function provides a summary of a dataset's distribution and displays its minimum, first quartile, median, third quartile, and maximum, which are calculated with “describe”. Initially, it seems clear that all groups performed better at the second phase since there are 6 boxplots, for each phase and each group. Inside the box is a line representing the median, or the middle value of the dataset, so comparing the middle line, we can see that all groups made an improvement. The box represents the interquartile range (IQR), which is the range between the first quartile (Q1) and the third quartile (Q3). The length of the box indicates the spread of the middle 50% of the data. The whiskers extend from the box to the minimum and maximum values within a specified range. They give an indication of the data's spread beyond the interquartile range. In general, data points that fall outside whiskers are considered outliers. As we can see, there are some outliers. Outliers are not noise, so we should keep them. [Fig 1] The interesting thing here is, you can see that the AI group only has outliners in the first phase, which was played by humans, so if they accidentally mess up in some rounds, that's fair enough! Human mistakes are one of the common reasons for outliners in a scenario where humans are involved, but in the control group with humans playing both phases, outliners are in both phases which is completely normal, as explained. The data does not need any further cleaning, integration, reduction, or transformation at this point.

### 3.4.3. Line Plot

As part of the second phase of the problem, we must determine how the performance of these three groups varies, before proceeding with a numerical test. A line plot can be used for a semi-final decision. Using a line plot, you can visualize data by connecting data points called markers with straight lines. Like many other graphs, a line plot has two axes. Typically, the horizontal axis (X-axis) represents the independent variable, here phases, while the vertical axis (Y-axis) represents the dependent variable, here totalingred. The equation for the line is Y=mx+b, where Y is the dependent variable, X is the independent variable, m is the slope of the line, and b is the y-intercept [Fig 2]. Technically, this plot is not the most suitable choice for this problem, but the slope of each line, which represents each group, indicates which group has improved the most, a steeper slope indicates more improvement.

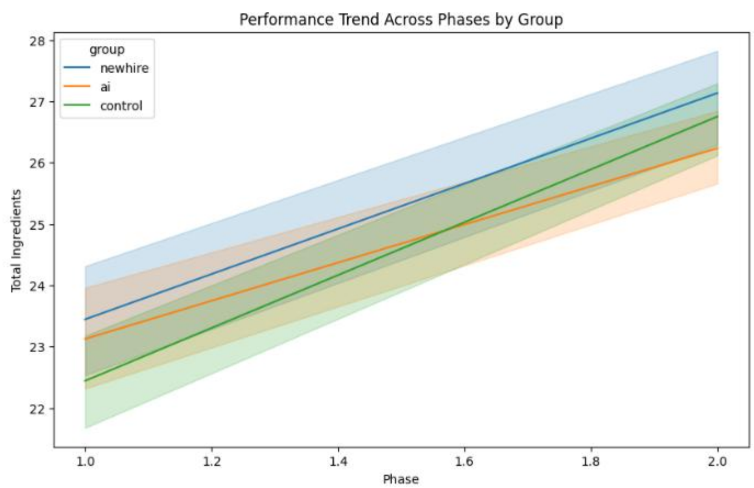



*Figure 2: Line Plot*

## 3.5. Before ANOVA

Although the "AI" and "control" groups have 20 teams, the "newhire" group has only 15. To achieve a fair result, 15 teams from both "ai" and "control" are chosen using the "random.choice" function in nampy.

### 3.6.1. ANOVA

The next step is to select a method to determine which group has improved the most and how their performance has changed over these two phases. In this case, the "ANOVA" method was used. "ANOVA" is a statistical method used to analyze the differences among group means in a sample. ANOVA is used when there are three or more groups to compare means. It helps determine whether there are any statistically significant differences between the means of these groups, which is similar to our problem. ANOVA's advantages can be explained as Efficiency, Flexibility, and Overall Comparison, as ANOVA not only indicates if there is a significant difference, but also identifies which groups are different from each other through post hoc tests, in which the details will be explained later. The disadvantage of this method is that it can be susceptible to outliers, which may unduly affect the results, which here is not going to be an issue since only a few outliers have been found.

### 3.6.2. ANOVA’s Formula

The formula for a one-way ANOVA involves calculating the F-statistic, which is used to test whether there are statistically significant differences between the means of three or more independent groups.

- Total Sum of Squares (SST):
  - $SST = \sum\sum(xij - \bar{x})2$

    Where:

    $x_{ij}$ is an individual data point.

    $\bar{x}$ is the overall mean of all data points.

- Between-Group Sum of Squares (SSB):
  - $SSB = \sum ni(\bar{x}i - \bar{x})2$
    Where:
    $n_i$ is the number of observations in the $i-th$ group.
    $x_i$ is the mean of the $i-th$ group.
- Within-Group Sum of Squares (SSW):
  - $SSW = \sum\sum(xij - \bar{x}i)2$
- Degrees of Freedom (df):
  - $df_{total}$ = N−1 (where N is the total number of observations).
  - $df_{Between}$= k−1 (where k is the number of groups).
  - $df_{within}$ = N−K
- Mean Square (MS):
  - $MS_{Within} = \frac{SSW}{df_{Within}}$
  - $MS_{Between} = \frac{SSB}{df_{Between}}$
- F-statistic:
  - $F = \frac{MS_{Between}}{MS_{Within}}$

- Decision Rule:
  - Reject the null hypothesis if $F > F_{\alpha,df_{Between},df_{Within},}$ where $F_{\alpha,df_{Between},df_{Within},}$ is the critical value from the F-distribution.

## 3.7.1. HSD

Tukey's Honestly Significant Difference "HSD" test is a post hoc test often used in conjunction with Analysis of Variance "ANOVA". Tukey's HSD is used to identify which specific groups differ significantly from each other when the ANOVA indicates that there are significant differences among groups. In situations with more than two groups, it is particularly useful in finding pairs of groups that show significant mean differences.

## 3.7.2. HSD Formula

Tukey's HSD test calculates a critical value, often denoted as $q$ , which is then compared to the studentized range statistic $(q_{\alpha,k,df_{Within}})$.

- Studentized Range Statistic:
  - $q_{\alpha,k,df_{Within}} = q.\left(\frac{MS_{Between}}{n}\right)^{1/2}$
- Critical Value ($q$):
  - The critical value depends on the desired significance level ($\alpha$), the number of groups ($k$), and the degrees of freedom within groups ($df_{Within}$).
- Decision Rule: Reject the null hypothesis for pairs of group means where the absolute difference is greater than $q_{\alpha,k,df_{Within}}$.

| | sum_sq | df | F | PR(>F) |
|---|---|---|---|---|
| **C(phase)** | 1859.266667 | 1.0 | 115.142524 | 1.855914e-24 |
| **C(group)** | 83.344444 | 2.0 | 2.580719 | 7.666372e-02 |
| **C(phase):C(group)** | 23.544444 | 2.0 | 0.729042 | 4.828504e-01 |

*Table 1 : Result of ANOVA test*

# 4. Evaluation

The results of the preliminary tests will be used in the evaluation process to decide if there is a significant difference in the number of ingredients collected for the three groups in phase 2, and if so, which groups are different.

## 4.1. Determine Overall Significance

Using ANOVA, we can determine whether there is a statistically significant difference in means between groups. If the ANOVA test yields a significant result (i.e., p-value < 0.05), it indicates that at least one group's mean is different from the others. If ANOVA yields a non-significant result (p-value > 0.05), it suggests that there is not enough evidence to conclude a significant difference between the groups.

### 4.2.1. Results of the ANOVA

Effect of Phase (C(phase)):

Significant Effect: Yes (p < 0.05)

Explanation: The factor "Phase" (C(phase)) has a significant effect on the number of collected ingredients.

Effect of Group (C(group)):

Significant Effect: No (p > 0.05)

Explanation: The factor "Group" (C(group)) does not appear to have a significant effect on the number of collected ingredients.

Interaction Effect (C(phase):C(group)):

Significant Effect: No (p > 0.05)

Explanation: The interaction effect between "Phase" and "Group" (C(phase):C(group)) is not statistically significant (p = 0.4828504). This implies that the effect of "Phase" on the number of collected ingredients does not depend on the specific "Group" and vice versa. [Table 1]

Clearly, there is a significant difference between each group in these two phases, and while all three groups improved significantly, the reasons for each improvement were different. [Fig3]

### 4.2.2. Justifying the outcome

#### 4.2.2.1. "ai" group

The average number of points collected by this group at the first phase was 23.44, which was best among the other groups, while at the second phase it was 27.13.[Fig3]AI players are likely to make fewer mistakes, act more accurately and play faster than human players, so the "ai" group performed better at the

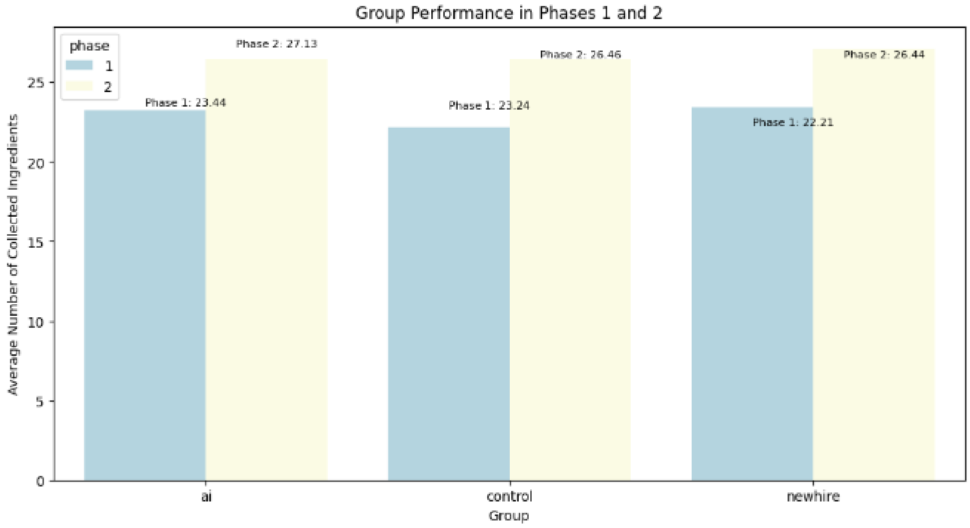



*Figure 3: Group Performance Comparison Across Phases*

second phase compared to the first phase, just as it was expected from the beginning.

#### 4.2.2.2. "control" group

The average number of points collected by this group at the first phase was 23.24, which was the second best among the other groups, while at the second phase it was 26.46.[Fig 3] Surprisingly, the "control" group made a significant progress even though the players remained the same in both phases. In this case, the improvement can be justified as follows, after getting familiar with the game during the first phase, the players became more experienced, so they performed better during the second phase by using their previous experience.

#### 4.2.2.3. "newhire" group

The average number of points collected by this group in the first phase was 22.21, which was worst among the other groups, while in the second phase it was 26.44.[Fig3] Based on the first phase's result, it is easy to conclude that the majority of these teams aren't very good at this game compared to other teams. As some players were replaced in phase two, it is possible that some better players joined the group, and the players who stayed used their previous experience from phase one to achieve a better result.

### 4.3. The overall result

Due to different causes, all groups showed noticeable improvement in the second phase. [Fig3]

### 4.4. Comparing across groups based on ANOVA result

Based on the ANOVA results, it can be concluded that although each group improved in the second phase compared to the first phase, there was no significant difference between groups.[Fig 3] In other words, the effect of replacing part of the team with a new member or artificial intelligence does not appear to lead to significantly different outcomes in terms of the number of collected ingredients in the second phase of the experiment. Initially, it was expected that when AI players joined the "ai" group their performance would be noticeably different from other groups, but the ANOVA test suggests this is not the case. The “ai“group had the highest average score in the second phase,

| group1 | group2 | meandiff | p-adj | lower | upper | reject |
|---|---|---|---|---|---|---|
| ai | control | -0.5222 | 0.5021 | -1.618 | -0.1346 | False |
| ai | newhire | 0.4389 | 0.6143 | -0.6569 | 1.5346 | False |
| control | newhire | 0.9611 | 0.099 | -0.1346 | 2.0569 | False |

*Table 2: HSD result*

27.13, compared to other groups. However, it isn't even one point higher than the "control" group with an average of 26.46.[Fig3]

## 4.5. Comparing across groups based on HSD result

HSD test yields the same results. The "Reject" column is False for all comparisons (AI vs. control, AI vs. newhire, control vs. newhire) after implementing the HSD test. This indicates that, after adjusting for multiple comparisons, there is no significant difference in the means between these groups.

Although there was no significant difference in means between these groups, it is still worth measuring which group improved the most and if it is as expected at first that the "ai" group improved the most.

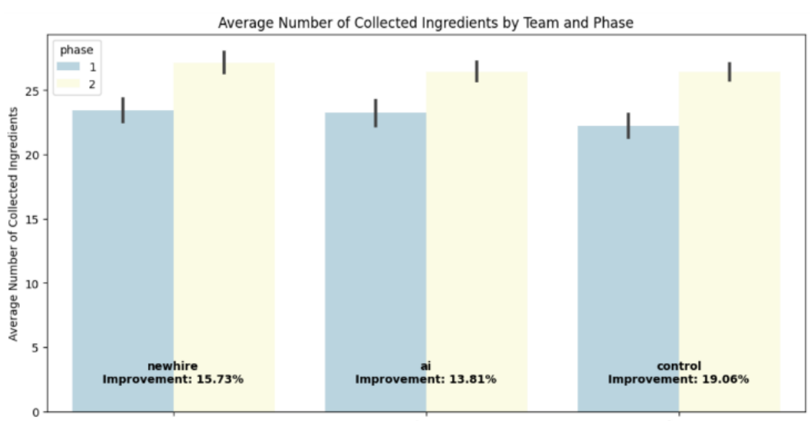



*Figure 4: Percentage Improvement Analysis Across Groups*

This chart [Fig4] shows that the "control" group, with 19.06% progress, improved the most. In second place is "newhire" with 15.73%, and the surprising third place goes to "ai" with only 13.81% improvement. As a result, I would justify the result on the basis of the numbers and my own opinion in this case. Due to two reasons, the "control" group improved the most. Due to their opportunity to play the game twice, the players in these teams performed much better in the second phase because they got to know one another and formed a sense of belonging. Knowing each other's strengths and weaknesses in the first phase may have helped them support each other in the second phase. The second aspect is that each player gained some experience during the first phase, and by using that experience in the second phase, they were able to act more effectively.

# 5. Summery

The analysis aimed to investigate the effects of replacing part of a team with artificial intelligence (AI) in the mini-game Dash and Dine of the Super Mario Party video game. The study involved teams of four players playing the mini-game six times in phase 1, followed by changes in team composition in phase 2, where some teams had a new team member (newhire), some had an AI substitution (ai), and others remained unchanged (control). The number of total collected ingredients was recorded for each game.

The descriptive analysis revealed that each group showed improvement in the second phase, with the "control" group demonstrating the highest average improvement of 19.07%. However, the mean differences between the groups were not substantial.

The statistical analysis using ANOVA indicated no significant difference in the number of collected ingredients between the three groups in phase 2. The Tukey HSD post hoc test supported this, revealing no significant differences in mean differences between any two groups.

Further exploration of individual group improvements in the second phase highlighted that the control group exhibited the most considerable improvement (19.07%), followed by the "newhire" group (15.73%) and the AI group (13.44%).

In conclusion, while both the "newhire" and "ai" groups showed improvements in the second phase, the "control" group demonstrated the highest improvement. This suggests that, in this specific context, the addition of AI did not significantly outperform the other groups. However, it's essential to consider the limitations, such as the relatively small sample size, which prevents generalization of the findings. The results emphasize the importance of teamwork and collaboration in achieving better outcomes, even in the presence of AI.

# 6. Bibliography

## 6.1. Software and Packages:


Python Software Foundation. (2021). Python Programming Language - Version 3.8. https://www.python.org/

Statsmodels (Python library for ANOVA):

Author(s): Seabold, Skipper, and Josef Perktold

Title: Statsmodels: Econometric and statistical modeling with Python

URL:https://www.statsmodels.org/stable/index.html

Scipy.stats (Python library for Tukey's HSD):

Title: Scipy.stats documentation

URL:https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.tukeyhsd.html

NumPy:

Author(s): Travis Olliphant et al.

Year: 2006

Title: NumPy: array computation library for Python

URL: https://numpy.org/

Pandas:

Author(s): Wes McKinney and Contributors

Year: 2008

Title: Pandas - Python Data Analysis Library

URL: https://pandas.pydata.org/

Statsmodels:

Author(s): Seabold, Skipper, and Josef Perktold

Year: 2010

Title: Statsmodels - Econometric and statistical modeling with Python

URL:https://www.statsmodels.org/stable/index.html

Seaborn:

Author(s): Michael Waskom

Year: 2012

Title: Seaborn: statistical data visualization

URL:https://seaborn.pydata.org/

Matplotlib:

Author(s): John D. Hunter

Year: 2003

Title: Matplotlib: A 2D plotting library

URL: https://matplotlib.org/