# Research Questions

Omaha Girls Rock is a local non-profit organization that seeks to use music to grow the young girls in the community. By teaching them music and giving them outlets to perform, they seek to create identities of self. As part of this program, the organization conducts surveys before and after the girls participate in the program to see how their skills have improved. These questions ask the girls to rate their likelihood of doing various activities, such as wearing the kind of clothes they like even if they are different from others to making friends their age, on a scale of 1-7 (1 being the least likely to 7 being the most). Additionally, there were three statements where they were to either agree or disagree to a varying degree. We want to analyze these results to make recommendations on how Omaha Girls Rock can improve their programs.

First, we would like to see if there was any significant difference in how the girls answered the questions before and after their program based on if they were in the older group (10-12) or the younger group (13-16). We believe that the girls should show growth for each response, regardless of what age group they are a part of. This information would tell them which age groups they need to focus on in building the skills that are associated with that given question. For example, if girls in the younger age group showed that significantly they did not improve or was worse at being able to express their feelings to another kid, the organization could work with the girls 10-12 on how to express themselves more.

Second, we would like to examine the results to see if there were any significant differences between how the girls answered the survey based on their race. Again, we expect there to be no significant differences in how the girls answer each question regarding their race. If we do find a significant difference though, Omaha Girls Rock would then know how to better cater their activities to the community. No organization wants to show that a specific group is being treated unequally compared to others. This will let them know of any blind spots they might have and work to correct them.

# Data

**Data Background**

The data in the dataset pertains to a survey of girls, ages 8-14, before and after participation in the OGR Academy offered by the Omaha Girls Rock Organization. The dataset consists of age, Years at Camp, Race/Ethnicity, and Zip Code of the participants in the program.

**Data Processing**

After familiarizing ourselves with the raw survey data, we noticed some data quality errors, such as missing data, different representations of the same field (i.e., 1 vs 1st and 2 vs 2nd in the same field), class imbalance (race/ethnicity), and non-numeric answers. We underwent a data processing step before deciding to do any analysis on the data, so that we are only working with easy-to-use and clean data.

***Many Tables***

* Instead of having 8 datasets to do analysis on, all the data was combined to a single table, where fields such as age\_group, year, and test\_type were added as new fields. These new fields allowed us to identify what each data observation represents (i.e., Older Group, 2019, Pre-Test). We can filter on these fields to subset the data in any way we want, while only needing to read from and work with a single table.

***Missing Data***

* There were some instances where a question did not have a score for a given observation. Since the data we are working with is ordinal, we replaced and missing data with the mode of that question. Since the amount of missing data only makes up < 1% of the total dataset, we do not risk altering the true results of the survey.

***Data Representation***

* In the *Years at Camp* field, some survey results include the superscript of the year, (1st, 2nd, 3rd), while others did not (1, 2, 3). Since it is appropriate to cast this field as an integer, all superscripts were removed and set to an integer type.

***Class Imbalance***

* At first look into the race/ethnicity field, there were 17 unique groups of race/ethnicity. There was no standardized way to collect this data in the survey, therefore there were differences in the way a race is referenced and the order of races if an individual is multi-racial. During data processing, a cleaning procedure is done to reclassify race/ethnicity by standardizing all these different factors. This left us with 12 standardized unique groups.
* However, classes were still highly unbalanced: 250 Caucasian, 53 Black, and the remaining 10 groups all have less than 10 data points. To overcome this, race/ethnicity was further grouped into 4 primary groups: Caucasian, Black, Multi-Racial, and Other.

***Non-numeric answers***

* The last three questions in the survey do not contain numeric answers, instead they are categorized as strings with 6 levels from strongly disagree to strongly agree. We used a numeric dictionary ranging from 1 to 6 to enumerate these values into numbers to aid numerical analysis.

# Descriptive Statistic

**Descriptive Statistics Analysis 1: Relevant Attributes for Analysis (Race/Ethnicity and Age)**

Chart, bar chart

Description automatically generated

Looking at Race/Ethnicity groups in the data, it is clear that an imbalance exists between races, Caucasians being the dominate race in the survey data. Since our group will be looking into race/ethnicity we will have to consider this imbalance in our analysis and be aware of the affects that class imbalance has on the results

Chart, bar chart

Description automatically generated

The second focus of our study deals with age and age groups, by analyzing the graph above we can ensure that class imbalance will not be an issue between age groups. Interestingly, there is a much more significant reduction in older group participants between 2018 and 2019. This could indicate the there were more students who “graduated” from the program in 2018 than girls who joined the program in years after that graduating group.

**Descriptive Statistics Analysis 2: Distribution of Mean Scores and Variation**

Chart, histogram

Description automatically generated

The above chart depicts the average (mean) scores across all questions in the study.

Chart, histogram

Description automatically generated

Chart, box and whisker chart

Description automatically generated

* Question 15 is the only question where we see the range of the girls answers got worse on the post test

Table

Description automatically generated

* Something cool here

Chart, line chart

Description automatically generated

# Other Info

One outside data source that we potentially could use when analyzing this data set is using the Omaha census data from the last census in 2021. This will allow us to see the current demographics of the city and see how the survey participants line up. If the percentage of participants of a certain race or ethnicity is smaller than the census data, we might want to see if the difference is significant. From our initial look of the census data, we see that the Caucasian has the highest percentage at 75.5%, followed by Black at 12.1%. This does align well with our survey participant breakdown (2021).

Another data source we may reference is an age breakdown of Omaha’s citizens. Unlike the U.S. Census, this breaks the age groups down farther into 5-year periods. As with our race data source, this is to make sure that our population of survey participants is like that of the population of Omaha as a whole. It is about equal in the number of kids between 10-14 and 15-19 in the Omaha region, similarly to how the number of girls in the older and younger groups are similar. We might need to see if the difference is significant, but that will come after comparing our results for the research question (2020).

Lastly, a data source we may want to use is the average income of a household for each zip code in Omaha. This may come into effect when looking at our results to see if financial status of each survey participants could also be affecting their survey results. Families with more resources have more privileges to help their kids develop and grow more substantially than those that do not.

# Statistical Methods

To find the result for the given questions we use statistical methods. Statistical methods involved in carrying out a study include the planning, designing, analysis, inference of meaningful interpretation, and reporting of the research findings. There are few statistical methods like T-test, Anova, Z-test etc...

* For young age and old age (research question1) we use Two-sample t-test statistic test. The two-sample t-test is used to determine whether two population means are equal or not. Here we are assuming the older group and younger group are equal as null hypothesis and not equal as alternative hypothesis.

H0: Older Group = Younger Group

Ha: Older Group! = Younger Group

To see if there are any significant differences in how the girls from the two different age groups answered the survey before and after the program, we plan to use a completely randomized design with our two groups. For each question, we will use the difference of the girls’ answers on their post-test from their pre-test answers. Our hypothesizes are:

We plan to run this method with a confidence interval of 95%. If after running our experiment, we find that we can reject our null hypothesis, we can take the differing pairs and run independent random samples for each survey question to see in which questions the populations have significantly differently changes in answers.

To check if there are any significant differences in how girls of different races answered the survey before and after the program, we again will use a completely randomized design since we have 4 different sub-populations of girls: Caucasian, Black, Multi-racial, and Other. For each question, we will use the difference of the girls’ answers on their post-test from their pre-test answers. Our hypothesizes are:

We will continue working with a confidence interval of 95%. If after running our experiment, we find that we can reject our null hypothesis, we will then run Fisher’s LSD Procedure, comparing each population with each other to determine which ones are different. From here, we will then again compare the differing population pairs for each survey question with independent random samples to find which questions were significantly different for those populations.

Two-Factor Factorial Experiment (ANOVA Table)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Factor??** | **Treatments (Delta Between Pre and Post)** | | | | | | | | |
| Q15 | Q16 | Q17 | Q18 | . | . | . | . | . |
| Younger | 1.5 |  |  |  |  |  |  |  |  |
| Younger | 1 |  |  |  |  |  |  |  |  |
| Younger | -0.5 |  |  |  |  |  |  |  |  |
| Older | 1 |  |  |  |  |  |  |  |  |
| Older | .07 |  |  |  |  |  |  |  |  |
| Older | 1.2 |  |  |  |  |  |  |  |  |

|  |  |
| --- | --- |
| Question 15 | |
| Younger Group | Older Group |
| 6 | 6 |
| 7 | 7 |
| 2 | 5 |
| 3 | 2 |
| 5 | 4 |
| 4 | 7 |
| 3 | 5 |

|  |  |
| --- | --- |
| Question 16 | |
| Younger Group | Older Group |
| 6 | 6 |
| 7 | 7 |
| 2 | 5 |
| 3 | 2 |
| 5 | 4 |
| 4 | 7 |
| 3 | 5 |

Question 2

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Factor??** | **Treatments (Delta Between Pre and Post)** | | | | | | | | |
| Q15 | Q16 | Q17 | Q18 | . | . | . | . | . |
| Cauc | 1.5 |  |  |  |  |  |  |  |  |
| Cauc | 1 |  |  |  |  |  |  |  |  |
| black | -0.5 |  |  |  |  |  |  |  |  |
| black | 1 |  |  |  |  |  |  |  |  |
| multi | .07 |  |  |  |  |  |  |  |  |
| multi | 1.2 |  |  |  |  |  |  |  |  |
| other |  |  |  |  |  |  |  |  |  |
| other |  |  |  |  |  |  |  |  |  |

# References

Infoplease. (2000). *Omaha, NE demographic statistics*. Infoplease. Retrieved October 18, 2022, from https://www.infoplease.com/us/census/nebraska/omaha/demographic-statistics

United States Census Bureau. (2021). *U.S. Census Bureau quickfacts: Omaha City, Nebraska*. United States Census Bureau. Retrieved October 19, 2022, from https://www.census.gov/quickfacts/omahacitynebraska

ZipAtlas.Com. (2022). *Median Household Income in Omaha, NE by Zip Code*. Median household income in Omaha, NE by ZIP code. Retrieved October 18, 2022, from http://zipatlas.com/us/ne/omaha/zip-code-comparison/median-household-income.htm