Table of Contents

1. Statistical analysis

In the following part the categories of interests presented in the introduction are analyzed: The analysis is done in Python [6], using “pandas” [7], “NumPy” [8], and the module “Stats” from the library “Scipy” [9]. Every result will be presented visually using the “Matplotlib” [10]. The code, along with the visuals will be uploaded to a “GitHub” [11] repository and the code of comments such that it can be conveniently correlated to the results is this report.

2. Student's t-test

Student's t-test is a statistical test used to test whether the difference between the response of two groups is statistically significant or not. It is any statistical hypothesis test in which the test statistic follows a Student's t-distribution under the null hypothesis [12].

This test is a way to determine whether there is any statistically significant difference between two sample groups. In the process of a t-test, a “null hypothesis” is tested - an hypothesis claiming that there is no significant difference between the samples in the test. This is done using the mean of both sample groups, “variance” [13], sample size, and a “pre-specified significance level” [14].

Two results are produced by the t-test a “t-statistic [10]” and a “p-value” [15]. The t-statistic is a value used to determine if there is a difference between means of our two groups. A small t-statistic value suggests a small difference between the group means and a large value, a large difference. A p-value is a probability value used to determine whether our results are due to chance, or due to a difference between the samples. A small p-value indicates that any difference observed between two group means is due to chance, and thus evidence to reject the null-hypothesis, while a high p-value would indicate that the differences observed between the two group means is due to a difference in the population and we can see it as an indication to accept the null-hypothesis. For the purpose of this test, the pre-specified significance level will be: 0.05. Any p-value below 0.05 indicates that the observed difference between the groups is statistically significant, and the null-hypothesis is rejected.

“Degrees-of-freedom” [16] are another value used in a t-test. This value represents the variability in our samples, and can be formally described as the “number of independent values that can vary in a statistical calculation”. This value is important to note in the result of a t-test because a small value of degrees-of-freedom will indicate that the test results are less statistically relevant, and a large value will give us more reliable results.

In the context of this project I will use a t-test to compare 3 of the 4 chosen categories. This test will uncover whether the mean value of the chosen categories differ significantly between the “Arena” and the “Music and bar” show venues. The categories of interest for the t-test are *Approximate annual income, Age range,* and *Spending habits.*

Mathematical annotations of t-test

$\overline{x}_1$ = *The mean age of Arena gusts*

$\overline{x}_2$ = *The mean age of Music and Bar gusts*

s^2 = Variance of Arena and Music and Bar age

n_1 = The number of sampels in the Arena group

n_2 = The number of sampels in the Music and Bar group

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{(s^2(\frac{1}{n_1} + \frac{1}{n_2}))}}$$

Hypothesis

H_0 = there is no difference between the group means

H_1 = there is difference between the group means

2.1 T-test assumptions

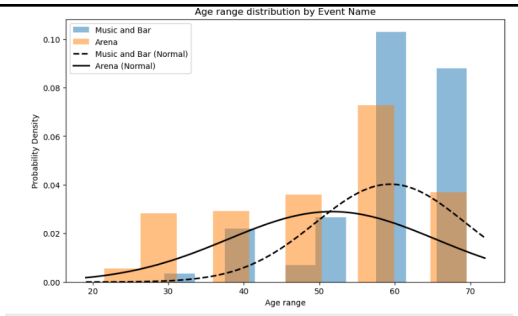
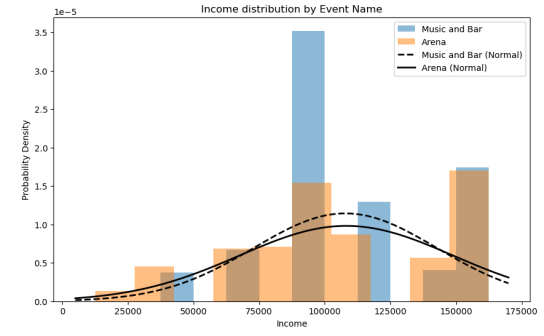
Age range
The data points for this part appear in the combined data set as intervals. The intervals are: *19-24, 25-34, 35-44, 45-54, 55-64, 65+*, *Prefer not to say, and "null"* - empty cells in the dataset. For the purposes of this test the absolute value for every interval is the average value. As an example: *19-24* will be calculated as 21.5 years of age. For the category *65+*, the average of 65 and the next interval is used: 74. For the category *Prefer not to say* the mean value of the lowest value: 19, and the largest value 74 is used.

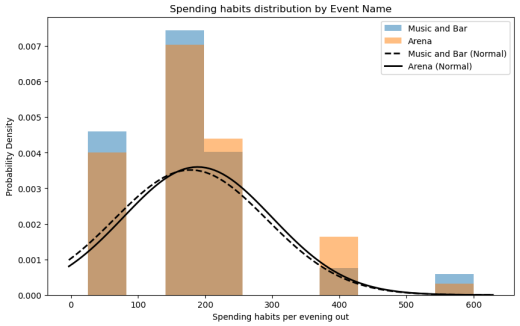
Approximate annual income
The data points for this part appear in the combined data set as intervals. The intervals are: *Less than $25,000, $25,000–$49,999, $50,000–$74,999, $75,000–$99,999,$100,000–$124,999, $150,000+*, *Prefer not to say, and "null"* - empty cells in the dataset. For the purposes of this test the absolute value for every interval is the average value. As an example: *$100,000–$124,999* will be calculated as *$112499.5*. For the category *$150,000+*, the average of $150,000 and the next interval is used: *$162499.5*. For the category *Prefer not to say* the mean value of the lowest value and the highest value is used. *Less than $25,000 was calculated* as the mean value between *$25,000* and *$0*.

Spending habits for a night out
The data points for this part appear in the combined data set as intervals. The intervals are: *$0 to $50', $51 to $100, $101 to $200, $201 to $300,$301 to $500, More than $500*,Not applicable. For the purposes of this test the absolute value for every interval is the average value. As an example: *$301 to $500* will be calculated as *$400.5*. For the category *More than $500*, $600 is used. The value *Not applicable* is discarded for this test.

* These assumptions will be the same through all analyses in this project.
* All results have been rounded to two decimal points

2.2 Test results - t-test

Age range Music and Bar Mean: 59.42 Arena Mean: 51.66 Music and Bar Standard Deviation: 9.91 Arena Standard Deviation: 13.775 Music and Bar Sample Size (n): 216 Arena Sample Size (n): 458 Degrees of Freedom: 672 T-statistic: 8.31 P-value: 7.31e-16	
Approximate annual income Music and Bar Mean: 107840.935 Arena Mean: 108214.575 Music and Bar Standard Deviation: 34899.71 Arena Standard Deviation: 40688.15 Music and Bar Sample Size (n): 216 Arena Sample Size (n): 458 Degrees of Freedom: 672 t-statistic: -0.12 p-value: 0.91 Missing values for income: 99	

Spending habits Music and Bar Mean: 178.01 Arena Mean: 188.77 Music and Bar Standard Deviation: 113.485 Arena Standard Deviation: 110.84 Music and Bar Sample Size (n): 208 Arena Sample Size (n): 435 Degrees of Freedom: 641 T-statistic: -1.13 P-value: 0.26 Missing values for spending habits: 152	

2.3 Interpretation: t-test

Age range:

1. The t-statistic 8.307 is larger than 0.05, this indicates that the differences between the groups of interest is significantly bigger than would have been if the difference was due to chance. In the context of the hypothesis-test, this means that we reject the null-hypothesis, the hypothesis claiming that there is no significant difference between the mean ages or the two groups of interest, the Arena group and the Music-and-Bar group.
2. The p-value is 0.00000000000000073243662975434 57, which is significantly lower than 0.05. This indicates that the differences we observe between the age groups in the test are very unlikely to be due to chance, and that we can reject the null-hypothesis.

Approximate Annual income:

1. In this test, the t-statistic is “-0.12”. This result indicates that there is a difference between the mean income of the Arena and Music-and-Bar guest, but the small value of the t-statistic indicates that the difference is not a large one. The negative value of this statistic is not a relevant factor as it is only an indication that one of the averages is lower than the second, and switching the order of analysis would result in the positive value: 0.12.
2. The p-value of 0.91 is a relatively high result. This indicates that the difference we might observe in income between the Arena guest and the Music-and-bar guest, should we sample them, would be likely due to chance and not to a statistically significant difference. In the context of hypothesis testing, we fail to reject the null-hypothesis. This means not that we accept the null-hypothesis (that the means of both groups are the same), but rather that we do not have enough evidence to suggest that there is a significant difference between the two groups' means.

Spending habits:

1. The small t-statistic value: -1.13 indicates that there is a small difference between the Arena group mean and the Music-and-Bar group mean.
2. The p-value of 0.26 is significantly higher than the threshold for this test, 0.05. This is an indication that we cannot reject the null-hypothesis and that we are confident that there is no significant difference between the two group means.

2.4 Findings: t-test

1. Based on the t-test for age, we can conclude that there is a statistically significant difference between the age range for the Arena guests, and the Music-and-Bar guest, and that should observe this difference, we can be confident that it is not due to chance.
2. Based on the t-test for approximate annual income, we see no significant difference in the means of the two groups of interest, suggesting that the approximate annual income is the same across both venues.
3. The t-test for spending habits found no significant difference in the means of the two groups of interest in regard to spending habits, suggesting that the spending habits are the same across both venues.

2.5. Student’s -test conclusion

The t-test analysis performed on the average values of the 3 categories of interest, age range, approximate annual income and spending habits reveals a lot of similarity between guests attending the Arena and guests attending the Music-and-Bar venue. One finding stands out - the age of the attending guests. We uncovered a statistically significant difference in age, the Arena group is younger than the Music-and-Bar group. The large t-value: 8.307 and the extremely small p-value indicate a large difference between the two groups' average age, and that the result is one that we are confident of.

3. Chi-square test for independence - Day of the week most interested in attending events at Main Street Summer Shows.

A chi-squared test (also chi-square or χ2 test) is a statistical hypothesis test used in the analysis of contingency tables when the sample sizes are large. In simpler terms, this test is primarily used to examine whether two categorical variables (two dimensions of the contingency table) are independent in influencing the test statistic (values within the table) [17].

This test is an hypothesis test used to determine the relationship between two *categorical variables* [18]. Categorical variables are variables that can take on only one value. These variables differ from *quantitative variables* [19] that are variables whose values are on a continuum, such as age or temperature.

A Chi-square test uses a *standard-normal-distribution* [20]. A normal-distribution that has been standardized such that the mean is 0 and the *standard deviation* [21] is 1. The values in this distribution are then squared resulting in all positive values. A critical factor for this test are the degrees of freedom, in the case of an independence test is equal to the number of categories, or rows subtract 1 times the number of columns subtract 1 (see the formula for "df" in the next section). For our test we have 7 degrees of freedom and the null-hypothesis is that there is no relationship between our selected variables for the test. The degrees of freedom determined the threshold value for the "Chi-squared statistic" [23], a value above which the null-hypothesis is rejected. This value is coupled with the p-value for this test and found in a "Chi-square distribution table" [23]. For 7 degrees of freedom and a 0.05 p-value, the chi-square-statistic value is 14.067.

In the process of a Chi-square test the expected values of both our groups of interest are compared with the observed values, or the values as counted in the data set. If the values are close together, then we can say that there is no statistically significant difference between our two groups of interest, and we accept the null-hypothesis,
In the context of this project, the null hypothesis is that the preferred day of the week for guests at the Arena and Music-and-Bar venues are independent of each other. This suggests that any differences in preference for the day of the week to attend a show between venues are not statistically significant. The significance level for this test will be 0.05.

Mathematical annotation of Chi-square test

Chi-square test

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

χ^2 *is the Chi − squared distribution*

O = the observed data

E = the expected value

Degrees of freedom formula

$$df = (r-1)(c-1)$$

c − the degrees of freedom

r − number of rows

c − number of columns

Hypothesis

H_0 = the preferred day of the week for guests at the Arena and Music − and − Bar venues are independent of each other

H_1 = the preferred day of the week for guests at the Arena and Music − and − Bar venues

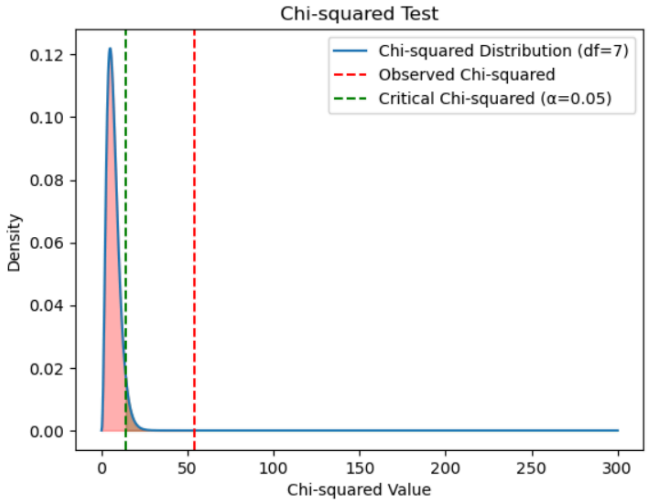
are not independent of each other

* This category is unique in this analysis due to the possible answers to the survey design. In this case, multiple answers could have been given, resulting in a very large amount of unique answers due to the large array of various combinations for this question (65 unique combinations). For this reason, the count of every phrase in the question was considered and not the combination of phrases. An example is: for the answer "Monday through Wednesday Daytime,Monday through Wednesday Evening,Thursday through Friday Daytime,Thursday through Friday Evening", each value separated by a comma, was counted as a preference and not the full answer as one preference.

3. 1 Result - Chi-squared test of independence - Day of the week most interested in attending events at Main Street Summer Shows.

Chi-squared contingency table

Day of the Week	Arena - frequency	Music and Bar - frequency	Total	Expected Joint frequency - Arena	Expected Joint frequency - Music and Bar
Monday through Wednesday Daytime	28	40	68	42.98	23.095
Monday through Wednesday Evening	91	82	173	109.34	58.756
Thursday through Friday Daytime	32	45	77	48.66	26.15
Thursday through Friday Evening	212	124	336	212.335	114.117
Weekend Mornings	59	26	85	53.72	28.87
Weekend Afternoons	169	86	255	142.2	76.42
Weekend Evenings	376	156	532	336.23	180.685
I'm not sure	46	34	80	50.56	27.17
Total	1013	539	1567	995.045	535.26



Chi-squared statistic: 54.31
P-value: 2.04e-09
Degrees of freedom: 7

3.2 Interpretation: chi-test of independence

1. A large Chi-squared statistic value of 54.31 is an indication of a statistically significant similarity between both categories in the test. This value is compared with the value found in the chi-squared-distribution-Table of 14.067, a value corresponding with our 7 degrees of freedom and the threshold for the p-value test.
2. An extremely small p-value of 2.042676712083393e-09 (0.000000002042676712083393 in decimal form), is significantly smaller than our threshold for the test - 0.05 and is strong evidence against the null-hypothesis.

3.3 Findings: chi-test of independence - this chi-square test suggests strong evidence against the null-hypothesis. This means that the guests attending both event venues, the Arena and the Music-and-Bar venue, do not differ significantly in their preference of day of the week most interested in attending events at Main-Street-Summer-Shows.

4. Linear regression

In statistics, linear regression is a statistical model which estimates the linear relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables) [24].

In general terms, linear-regression analysis will uncover how closely correlated one or more variables are to another variable in our data. Two variables are used for this analysis, an “independent variable” [25], and a “dependent variable” [25]. A close correlation between these variables suggests that the independent variable is a strong predictor of the dependent variable, and thus can be thought of as a “predictor”. Three statistical values will be used in this analysis, *“*slope*”* [26]*, “*p-value*”* and *“*r-squared*”* [27]. The slope in the context of linear regression is a representation of the rate of change in the dependent variable in response to the rate of change in the independent variable. A positive slope value would indicate an increase in the dependent variable as the independent variable increases, while a negative slope value would indicate a decrease in the value of the dependent variable as the independent variable increases.

The p-value in this context is interpreted in the same way as in the t-test, as a way to test the null-hypothesis stating that there is no linear dependency between the variables in the model.

R-squared, also called “the correlation coefficient”. This is a masseuse that suggests how much of the variability in the dependent variable can be explained by the independent variable. The value is between -1 and +1. A value of -1 will suggest a perfect negative correlation between the variables at hand, while a value of +1 would suggest a perfectly positive relationship. Following from this, is that an r-squared value of 0 would suggest no linear relationship or dependency between the values in the model.

Mathematical notation of the linear regression model

$$y = \beta_0 + \beta_1 X + \varepsilon$$

y *is the dependent variable*

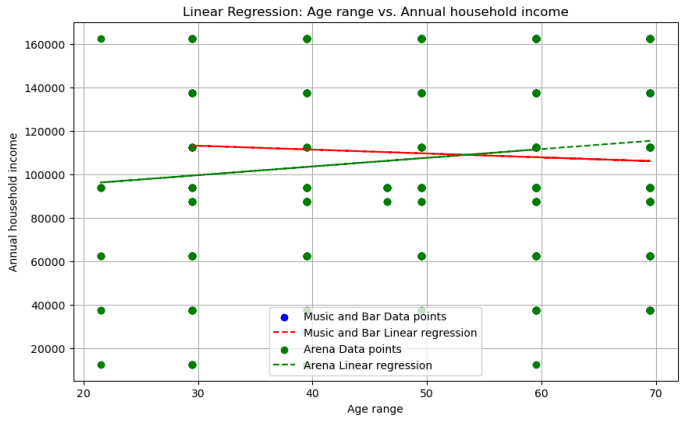
β_0 *is the intercept*

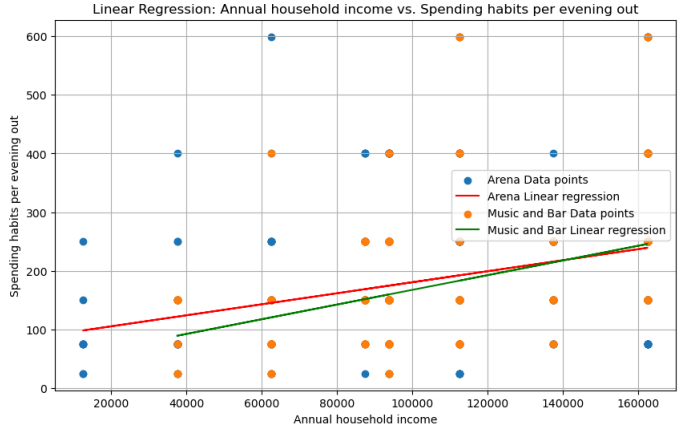
β_1 *is the slope*

X *is the independent variable*

ε *is the error*

4.1 Linear regression - results

Age vs income	
Arena Slope: 399.38 Intercept: 87583.05 R-squared: 0.02 P-value: 0.01 Degrees of Freedom: 456 Number of data points used: 458 **Music and Bar** Slope: -178.67 Intercept: 118457.098 R-squared: 0.01 P-value: 0.46 Degrees of Freedom: 214 Number of data points used: 216	

Age vs spending habits	
Arena Slope: 1.02 Intercept: 135.53 R-squared: 0.02 P-value: 0.01 Degrees of Freedom: 433 Number of data points used: 435 **Music and Bar** Slope: -0.35 Intercept: 198.99 R-squared: 0.0009 P-value: 0.6725 Degrees of Freedom: 206 Number of data points used: 208	
Income vs spending habits	
Arena Slope: 0.0009 Intercept: 86.59 R-squared: 0.120 P-value: 8.33e-14 Degrees of Freedom: 433 Number of data points used: 435 **Music and Bar** Slope: 0.001 Intercept: 42.29 R-squared: 0.152 P-value: 5.61e-09 Degrees of Freedom: 206 Number of data points used: 208	

4.2 Interpretation: linear regression

Age vs income

1. A slope of 399.38 suggests a positive relationship between age and income for the Arena group. This means that as age increases, so does the income.
2. An r-squared value of 0.0183 is an indication that age can explain only 1.83% of the income for the Arena group. This leaves 98.17% of the variability in the approximate-annual-income-unexplained. Thus we can conclude that age is a weak predictor of income for the Arena group.
3. A p-value of 0.0037 is an indication that the linear relationship observed, the slope, and the r-squared value, are statistically significant, as our p-value is under the threshold of 0.05, and evidence that we can reject the null-hypothesis, the hypothesis claiming that there is no linear dependency between our selected variables.
4. A slope of “-178.67” indicates that there is a negative relationship between age and approximate annual income for the Music-and-bar group. This means that as age increases, the income decreases.
5. An r-squared value of 0.0026 is an indication that 0.26% of the variability in income can be described by the age.
6. A p-value of 0.458 is much larger than the threshold used for this test: 0.05. Such a large value is strong evidence that we can accept the null-hypothesis and claim that age is not a good predictor for income at the Music-and-Bar group.

Age vs spending habits

1. A slope of 1.025 indicates a positive relationship between age and spending habits of the guest attending shows in the Arena.

2. An r-squared value of 0.0161 indicates that 1.61% of the variation of age range can predict the expense habits for a night out of the Arena guests. This leaves 98.39% of the variation in spending habits unexplained.
3. A p-value of 0.00814 is an indication that the linear dependency between age and spending habits is statistically significant, and thus we can reject the null-hypothesis.
4. A slope of "-0.352" suggests a negative relationship between age and spending habits for the Music-and-bar group. This means that as the age increases, the spending habit per night-out decreases.
5. An r-squared value of 0.000869 is an indication that 0.087% of the variation in age can explain the variation in spending habits for the Music-and-Bar guests.
6. A p-value of 0.673 is above the 0.05 threshold and an indication that the relationship between age and spending habits is not statistically significant and we can not reject the null-hypothesis.

Income vs spending habits

1. A slope of 0.000939 indicates a very small positive relationship between income and spending habits for the Arena group. This is a very small value and very close to 0 suggesting that the rate of change is not a large one.
2. An r-squared value of 0.121 indicates that 12.1% of the variation in spending can be explained by the variation in income. While this proportion is not a large one, it is finding some significance.
3. A p-value of 8.33e-14 (0.00000000000008332680812942 41 in decimal form) is an extremely small one indicating a very statistically significant relationship between income and spending-habits and we can reject the null-hypothesis.
4. A slope of 0.01 indicates a small positive relationship between income and spending habits from the Music-and-Bar group. This is a small value and an indication that the rate of change is not very small.
5. A r-squared value of 0.152 indicates that 15.2% of the variation in income can explain the variation in spending habits. This is a finding of moderate significance.
6. A p-value of 5.61e-09 (0.000000005614032180028272 in decimal form) is an extremely small value and suggest that the linear relationship between income and spending habits for the Music-and-Bar group is statistically significant, and we can reject the null-hypothesis.

4.3 Findings: linear regression

1. For the Arena group we uncovered a statistically significant positive linear relationship between age and income, but due to the small r-squared value, we see a weak linear dependence between age and income. This means that age is a weak predictor of income.
2. A small decrease in income can be observed as the age increases, however the small r-squared value and the large p-value is an indication that age is a weak predictor of income at the Music-and-Bar group.
3. The small p-value and the positive slope value suggest a statistically significant positive relationship between age and spending habits, but the small r-squared value suggests that age alone is a weak predictor of spending habits for the Arena guests.
4. A negative slope suggests that as age increases for the Music-and-Bar group, the spending habits decrease. However the extremely small r-squared value and the large p-value suggests a very small linear dependency between age and spending habits and indicate that we cannot reject the null hypothesis for this group.
5. The r-squared value for this model coupled with the very small p-value, are an indication that income can be used as a predictor of spending habits for the Arena guests. 12.1% of the variability in spending can be attributed to the annual income. The positive value of the slope would indicate a small rise in spending as the age increases.
6. The finding for the Music-and-Bar group is similar to the Arena group. We see a relatively large r-squared value and an extremely small p-value. These results along with a small slope, would indicate 15.2% of the variability in spending habits for this group can be explained by the annual income, and that as age increases, the spending slightly increases.

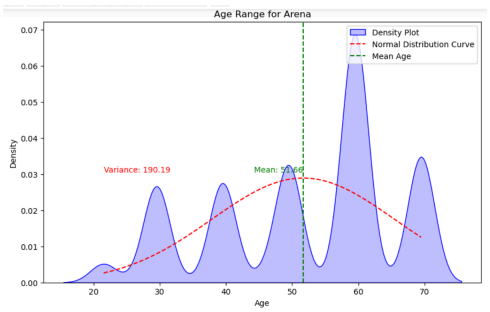
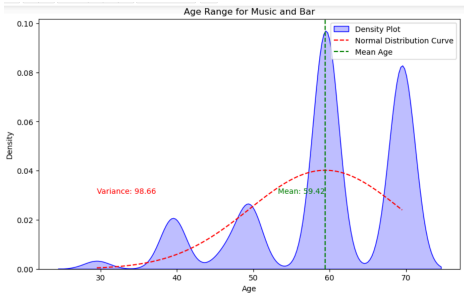
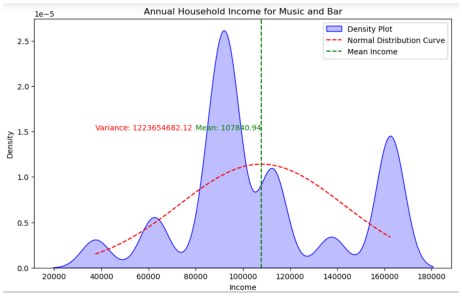
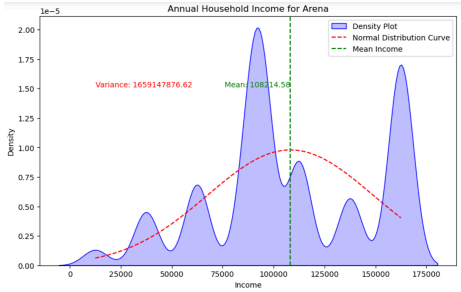
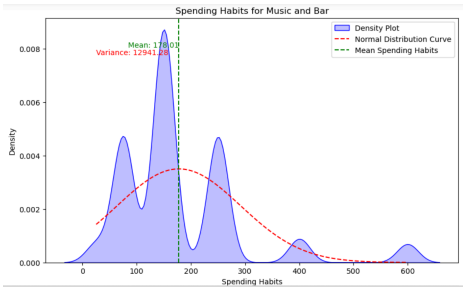
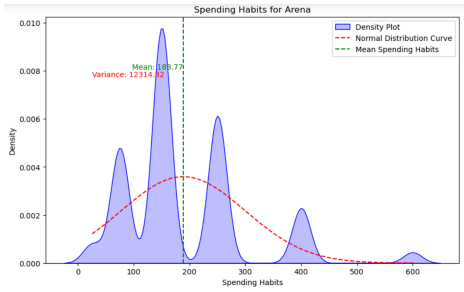
4.4 Conclusion - linear regression

The linear regression model uncovered a lot of similarity between the Arena guest and the Music-and-Bar guests. In this analysis we looked for linear dependency between age and income, age and spending habits and income and spending habits. The first two models uncovered results that are statistically insignificant, and are an indication that age is a weak predictor of both income and spending. The third model uncovered that approximate annual income is a good predictor of spending habits for both groups of interest. However, the spending changes very slightly as income is increased, so this result would not translate into larger differences in spending habits for guests attending both event venues as the age of guests increases.

5. Frequency analysis and averages

In this section of the project an analysis of averages and variability of the guests attending the Arena show venue and the Music-and-bar show venue will be performed using *Probability density function (PDF)* [5] method. This will uncover the means, frequency and variance of the observed data.

5.1 Results - Frequency analysis and averages

Age range - Music and Bar Number of samples: 216 Mean age for Music and bar: 59.4 years Mean age for both groups: 51.3 years **Age range - Arena** Number of samples: 458 Mean age for Arena: 51.6 years Mean age for both groups: 51.3 years	 
Approximate annual income - Arena Number of samples: 458 Mean annual income Arena: 108214.6 Mean annual income for both groups: $108226.5 **Approximate annual income - Music and Bar** Number of samples: 216 Mean annual income Music-and-Bar: $107840.9 Mean annual income for both groups: $108226.5	 
Spending habits for a night out - Arena Number of samples: 458 Mean spending habits value for Arena: $188.8 Mean spending habits value for both groups: $182 **Spending habits - Music and Bar** Number of samples: 458 Mean spending habits value for Music and Bar: $178	 

Mean spending habits value for both groups: $182	

5.2 Interpretation: frequency analysis and averages

Age range

1. We see an approximate 8 year difference in age between our two groups. The t-test analysis performed in section one of this part of the project, suggests that this is a statistically significant difference.
2. The variance values would suggest more diversity in the arena group than at the Music-and-Bar group.

Approximate annual income

1. We see a small difference of approximate $374 in annual income between both groups. The t-test performed on this data in section 1, confirms that this small difference is statistically insignificant.
2. The variance in the Arena group is larger than in the Music-and-Bar group, suggesting more diversity in this respect in the former, than in the latter.

Spending habits

1. The average value in spending habits between both groups is small, approximately $10. The t-test performed in section 1 of this part of the project confirmed that this is not a statistically significant difference.
2. We can observe a larger variance value for the Music-and-Bar group than for the Arena group.

6.3 Findings: frequency analysis and averages

1. This analysis uncovered the average age of the Arena group and the Music-and-Bar group. This result is statistically significant as confirmed by the t-test performed in section 1 (page 4) of this project.
2. We see no statistically significant difference between our two groups of interest in terms of the approximate-annual-income average. We do see a difference in variance, that could suggest more diversity in the Arena crowd.
3. Both groups' spending habits do not vary significantly, and the Music-and-bar group has more variability in terms of spending habits.
4. We can see a preference for weekend-evening shows, and a general preference towards the evenings-events and towards weekend-events. This result is statistically significant as confirmed by the chi-square-test-for-independence.

5.4 Conclusion - frequency analysis and averages

This section of the project looked at the averages and variability of the guests attending the Arena show venue and the Music-and-bar show venue. Probability density function plots were used to display the findings. The t-tests performed in section 1 of the statistical analysis chapters served to confirm the statistical significance of the finding.

The findings are that the average age differs between the groups: Arena at 51.6 years of age, and Music-and-Bar at 59.4 years. Income and spending habits can be seen statistically the same for both groups, approximate-annual-income of $108226.5 and spending habits for a night out: $182.

6. Statistical analysis conclusion

This statistical analysis was aimed at deriving statistically significant information about the guests attending the two event venues at Main-Street-Summer-Shows. This was done as part of the efforts undertaken by the marketing department at Main-Street-Summer-Shows, in the attempt to uncover the profiles of the guests attending both venues. It has been assumed by the marketing department based on observation that there might be unique characteristics between the guests attending the two main venues, the Arena venue and the Music-and-Bar venue. should the analysis have uncovered two distinct groups, the marketing strategy would be divided into two distinct markets, and if no significant difference would be found, the market for both venues would be the same.

The categories of interest were, age, approximate-annual-income, spending habits for a night out, and the day of the week preferred to attend a show. The statistical methods used for this analysis were student's t-test, linear regression, chi-square test for independence and a probability density function.

The result show the following:

Profile Category	Arena	Music and Bar	Sample size	Population size	Confidence level
Age	51.6 years	59.4	674	7582	95%
Income	$108226.5	$108226.5	674	7582	95%
Spending habits	$182	$182	643	7582	95%
Preferred day of the week	Weekend evenings	Weekend evenings	1567	7582	95%

We can see one profile of guests when it comes to approximate annual income, spending habits and preferred day of the week for attending a show. The groups, however, differ in the age range. The average age has been derived in section 3 (page 11) of this project and was verified using the student's t-test (page 4) as statistically significant difference.

8. Report conclusion

This demographic analysis looked at 4 categories out of 43 appearing as answers for a survey sent out to 7582 guests of Main-Street-Summer-Shows, and 1424 answers received from the guests. These categories have been chosen by the marketing department at the organization as among the most relevant for marketing purposes. Two datasets were combined, one of the survey, and the second from a ticketing platform, containing the event attended, using the email address as the unique identifier. This method enables us to correlate the survey answers to the event venues of interest.

Aside from the age range, no significant difference has been found. Further analysis can be performed on the remaining categories to uncover more precisely the profile of the guests, as a step toward more intimate knowledge of the target audience at Main-Street-Summer-Shows.

Sources:

1. Adam Smith, "The Wealth Of Nations," Book IV Chapter VIII, volume ii, page 660, paragraph 49.
2. Wikipedia contributors. "Student's t-test." Wikipedia, The Free Encyclopedia. [https://en.wikipedia.org/wiki/Student%27s_t-test]
3. Wikipedia contributors. "Chi-squared test." Wikipedia, The Free Encyclopedia. [https://en.wikipedia.org/wiki/Chi-squared_test]
4. Wikipedia contributors. "Linear regression." Wikipedia, The Free Encyclopedia. [https://en.wikipedia.org/wiki/Linear_regression]
5. Wikipedia contributors. "Probability density function." Wikipedia, The Free Encyclopedia. [https://en.wikipedia.org/wiki/Probability_density_function]
6. Python Software Foundation. "Python: About." Python, [https://www.python.org/doc/essays/blurb/]
7. Contributors to NumPy. "What is NumPy?" NumPy. [https://numpy.org/doc/stable/user/whatisnumpy.html]
8. McKinney, Wes, et al. "pandas documentation." pandas, [https://pandas.pydata.org/docs/]
9. Virtanen, Pauli, et al. "SciPy: Open source scientific tools for Python." SciPy, [https://docs.scipy.org/doc/scipy/]
10. Hunter, John D. "Matplotlib: Visualization with Python." Matplotlib, [https://matplotlib.org/stable/users/index.html]
11. SciPy community. "scipy.stats.ttest_ind." SciPy, [https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html]
12. Wikipedia contributors. "Null hypothesis." Wikipedia, The Free Encyclopedia. [https://en.wikipedia.org/wiki/Null_hypothesis]
13. Wikipedia contributors. "Variance." Wikipedia, The Free Encyclopedia. [https://en.wikipedia.org/wiki/Variance]
14. Wikipedia contributors. "Statistical significance." Wikipedia, The Free Encyclopedia. [https://en.wikipedia.org/wiki/Statistical_significance]
15. Wikipedia contributors. "P-value." Wikipedia, The Free Encyclopedia. [https://en.wikipedia.org/wiki/P-value]
16. Wikipedia contributors. "Degrees of freedom (statistics)." Wikipedia, The Free Encyclopedia. [https://en.wikipedia.org/wiki/Degrees_of_freedom_(statistics)]
17. Wikipedia contributors. "Chi-squared test." Wikipedia, The Free Encyclopedia. [https://en.wikipedia.org/wiki/Chi-squared_test]
18. Wikipedia contributors. "Categorical variable." Wikipedia, The Free Encyclopedia. [https://en.wikipedia.org/wiki/Categorical_variable]
19. Wikipedia contributors. "Quantitative research." Wikipedia, The Free Encyclopedia. [https://en.wikipedia.org/wiki/Quantitative_research]
20. Wikipedia contributors. "Normal distribution." Wikipedia, The Free Encyclopedia. [https://en.wikipedia.org/wiki/Normal_distribution]
21. Wikipedia contributors. "Standard deviation." Wikipedia, The Free Encyclopedia. [https://en.wikipedia.org/wiki/Standard_deviation]
22. Wikipedia contributors. "Chi-squared test." Wikipedia, The Free Encyclopedia. [https://en.wikipedia.org/wiki/Chi-squared_test]
23. MedCalc Software Ltd. "Chi-square table." MedCalc, [https://www.medcalc.org/manual/chi-square-table.php]
24. Wikipedia contributors. "Dependent and independent variables." Wikipedia, The Free Encyclopedia. [https://en.wikipedia.org/wiki/Dependent_and_independent_variables]
25. Wikipedia contributors. "Coefficient of determination." Wikipedia, The Free Encyclopedia. [https://en.wikipedia.org/wiki/Coefficient_of_determination]
26. Wikipedia contributors. "Slope." Wikipedia, The Free Encyclopedia. [https://en.wikipedia.org/wiki/Slope]
27. Eric Measuring Devices. "Sensitivity and Specificity." Easy Surf, [https://www.easysurf.cc/scintd.htm]

Appendix

Capstone Project Proposal

AIGS1012 - Professional Portfolio
Submitted to: Babatunde Giwa, PhD
Submitted by: Ilan Goldfarb

Table of contents

Disclaimer:
The following project has been depersonalized in order to keep personal data secure. This is part of the non disclosure agreement involved in getting the data and working on it.

Introduction

Adam Smith (1723–1790), In his monumental book, *An Inquiry into the Nature and Causes of the Wealth of Nations* (1776)***,*** states the following: "Consumption is the sole end and purpose of all production; and the interest of the producer ought to be attended to only so far as it may be necessary for promoting that of the consumer" [1].

In this, Smith emphasizes two things. The first is that production of goods and services are aimed at consumption, or in more modern terms, to be sold, and the second, that the attention of the producer should be focused on the customer as far as it is necessary for the product to be sold. In modern terms this can be seen as an attempt of a producer to get familiar with their client and thus tailoring the products to their desires, which results in selling more produce.

An intimate knowledge of the target market is an advantage for the development of a business and for increasing revenue, as it allows us to focus efforts..

To make informed decisions at an organization, data, historical and more general, can be used. The following is a “Market Demographic Analysis" based on historical-financial data and demographic data gathered from surveys sent out to clients of Main Street Summer Shows. The analysis is mostly descriptive in nature and is aimed at describing in an intuitive and in an easily understood way, the demographic characteristics of Main Street Summer Shows clientele.

Project Objectives

- To derive demographic data from the survey supplied by the organization and uncover the profile of the guests attending shows in two main venues.
- Compare both event venues using statistical methods.
- Create a Power Bi dashboard showcasing the results.

1. To derive demographic data from the survey supplied by the organization and uncover the profile of the guests attending shows in two main venues. The profile will consist of the age range, annual income, spending habits for an "evening out," and preferred day of the week to attend a show. These demographic characteristics were chosen as they are most useful, in my opinion, as key points for marketing purposes in the goal of advertising shows in the places where the profiled people reside.

All of the data for these come from an Excel file composed of 1424 rows, each row a response to the survey, and 52 columns, the questions asked in the survey.

1.1 - The location of the relevant data - in the survey data set:

a. Column DB - *Please specify your age range:*
Possible answers: 19-24, 25-34, 35-44, 45-54, 55-64, 65+.

b. Column EN - *What is your approximate annual household income?*
Possible answers: Less than $25,000, $25,000 - $49,999, $50,000 - $74,999, $75,000 - $99,999, $100,000 - $124,999, $125,000 - $149,999, $150,000+, Prefer not to say.

c. Column CX - *Not including your experience at Main Street Summer Shows, how much do you typically spend on an average evening out, including items like entertainment tickets, dinner, childcare, parking, taxi, hotel?*
Possible answers: $0 to $50, $51 to $100, $101 to $200, $201 to $300, More than $300.

d. Column AN - *What day of the week - and time of day - would you be most interested in attending events at Main Street Summer Shows? Check all that apply.*
Possible answers: Monday through Wednesday Daytime, Monday through Wednesday Evening, Thursday through Friday Daytime, Thursday through Friday Evening, Weekend Mornings, Weekend Afternoons, Weekend Evenings, I'm not sure.

2. Compare both event venues using statistical methods:

2.1 Statistical methods:

a. Quantitative differences between categories. How age, income and so on differ between the two event venues.
b. Central limit theorem to describe the results: mean, median and mode for each category, and a normal distribution.
c. Time permitting, probability calculator with age, income, spending habits and preferred day of attending an event.

Description of the data

1. Survey data: an XLSX file composed of 1424 rows, each row a response to the survey, and 52 columns, the questions asked in the survey.
2. Ticketing data: an XLSX file composed of 7582 transactions and 17 columns of details about the transactions.
3. Structures data sets for the purpose of the Capstone Project:
    a. A combined data set of the first two data sets.
    b. Six data sets of combined data with the count of the number of shows attended per guest(based on the email address).

Solution Approach

1. Cross both data sets using the email as the unique ID. This part will be done using Python and Excel.
    a. The email address in the survey data is in column EQ
    b. The email address in the ticketing data is in column B
2. Build a dashboard in Power BI as a descriptive tool to visualize the data on hand. The visualization will be derived from the data sets above and will consist of the categories:
    a. Age Range

   b. Annual Income
   c. Expense habits for a "night out"
   d. Preferred day of the week for a show
3. Locate the demographic characteristics most strongly correlated with both show venues. This part will be done using the Power BI dashboard visualization.
4. Compare the results for both venues using statistical analysis with “Central Limit Theorem” - mean, median, mode, standard deviation. This part will be done using Power BI, and Excel.
5. Implement a renaming and depersonalization of the data to ensure the privacy and confidentiality of the guests and the organization. This part will be done in Power BI and Excel and verified excel, and manually.

Deliverables

1. A Power Bi dashboard describing the demographic data as stated in the previous section, part 2 showcasing the results for both venues.
   a. The results are descriptive results of the categories from part 2 of the “solution approach” section of this document - this part will mostly be of the “count” of categories and “percentage of grand total” per category described.
   b. The results of the statistical analysis by both venues.
2. Submitting all the Python code involved in deriving the results of the project. This will be done as a GitHub directory as to specifications of the project submission instructions.
3. A Power-Point slide show describing the process and results in a way that can be easily understood both to people with a “technical background” and to people without a technical background.
   a. For the purpose of this project, “people with a technical background” are proposed to be people who have had a post secondary education in mathematics, or in fields that are strongly correlated with mathematics, generally in the “Natural Science”. “People without a technical background” will be supposed to be people whose mathematical education is of high-school-level.

Project description

In an organization called Main Street Summer Shows. This organization offers, along with other services, music shows in two main venues. These shows are an important source of revenue for this organization. Both venues are different in size and capacity, as well as in types of services offered, and attract, so it is assumed by the owners, a different demographic group of people.

Provided by the organization is one dataset of 7582 tickets sold over the summer for various shows in both event venues, supplied by the ticketing software, along with approximately 1424 survey answers from a random sample of guests, asking general questions about the experience during the shows, and questions of demographic nature, 48 questions in total.

Both data sets, the ticketing set, and the survey one have the email address of the users, and by crossing both data sets, it should be possible to combine the survey data with the ticketing data by using the email address as the unique identifier.

For statistical purposes our population is the 7582 emails from the ticketing platform, and our sample size will consist of the 1424 survey answers. Every category subjected for analysis will have its own sample size that could differ from the total sample size. This phonominum might occur because some questions in the survey were not answered by everoun (“prefer not to say” as an example), resulting in “null” values .

Problem Definition

1. What are common demographic characteristics, or profiles of people attending shows in both venues?
2. Do the profiles differ by event venue?

Risks and Constraints

1. General time management.

   Solution approach:

   a. create “work breakdown structure” with clear timlins, and follow it closely.
   b. Speak to members from other teams to make sure that nothing is missed.
   c. Work efficiently and without delay.
   d. Contact the course instructor with any problems and clarifications needed.

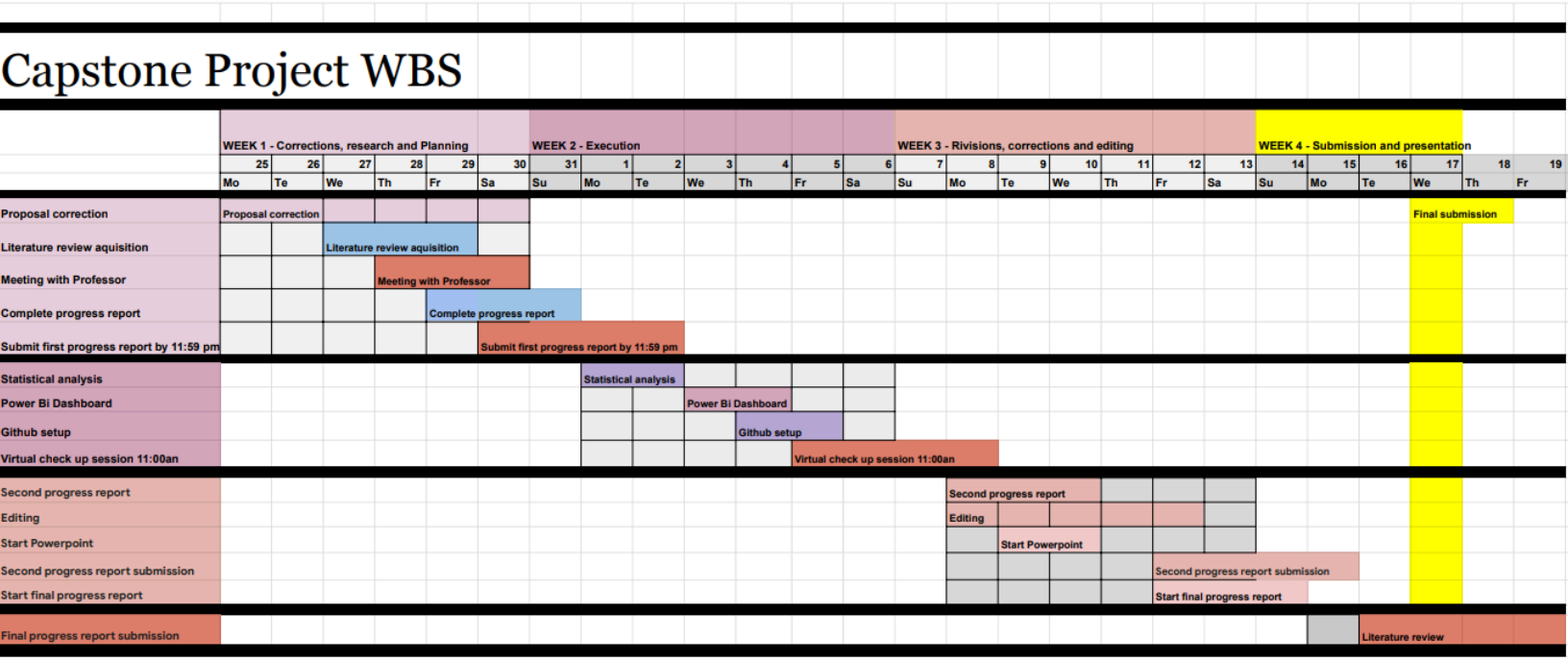
2. Low statistical significance of results.

   Solution approach:

   a. Data augmentation

.

Work Breakdown Structure



Proposed Methodology

1. Acquiring the data will be done in the process of the first meeting with the marketing team at Main Street Summer Shows and via email.
2. Combining the data from the survey with the tiketing data will be done using modules in Python: “Pandas”, “Numpy”, and verification of the accuracy of the results will be done in Excel, and Power Bi.
3. The descriptive dashboard will be created in Power Bi.
4. Statistical analysis will be done using Python and by hand.
5. Documents will be created using Google Docs.

Tools and Resources

1. Python - Python is a computer programming language often used to build websites and software, automate tasks, and analyze data [2].
2. Numpy - NumPy is the fundamental package for scientific computing in Python [3].
3. Pandas - Pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language [4].
4. Power Bi - Power BI is a collection of software services, apps, and connectors that work together to turn your unrelated sources of data into coherent, visually immersive, and interactive insights [5].
5. Microsoft Excel - Microsoft Excel is a spreadsheet editor developed by Microsoft for Windows, macOS, Android, iOS and iPadOS. It features calculation or computation capabilities, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications (VBA). Excel forms part of the Microsoft 365 suite of software [6].
6. Sample size calculator: This free sample size calculator determines the sample size required to meet a given set of constraints [7].

Summary

This report described the process by which I intend to extract information from two data sets in an attempt to derive useful demographic information for Main Street Summer Shows about the people attending shows at two event venues. This project will be done in a Capstone project framework at the school of business at Loyalist College, in the program “Artificial Intelligence and Data Science”.

The demographic analysis will be focused on 4 categories derived from the combined data sets: age range, annual income, spending habits for a “night out”, preferred day of the week for attending a show. These categories have been selected from the 43 questions asked in the survey, because I am of the opinion that these are the best data points derived from the combined data sets, for assisting in better-targeted-marketing of the population attending programs at Main Street Summer shows.

After the first stage, the descriptive stage of the project (see part 2, 3 and 5 of the Solution Approach section), further statistical analysis will be done (see 2.1 of the Project objectives section) using Central Limit Theorem techniques to obtain the significance of the results, or how accurately does the information in this project describe the guest of the two event venues.

Sources:

1. Smith, A. (1776). The Wealth of Nations. Book IV, Chapter VIII, v. ii, p. 660, para. 49.
2. Python Software Foundation. (n.d.). Python. Retrieved from https://www.python.org/doc/essays/blurb/
3. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). What is NumPy?. Retrieved from https://numpy.org/doc/stable/user/whatisnumpy.html
4. McKinney, W., & Others. (n.d.). pandas - Python Data Analysis Library. Retrieved from https://pandas.pydata.org/docs/
5. Microsoft. (n.d.). Power BI overview. Retrieved from https://learn.microsoft.com/en-us/power-bi/fundamentals/power-bi-overview
6. Wikipedia. (n.d.). Microsoft Excel. Retrieved from https://en.wikipedia.org/wiki/Microsoft_Excel
7. Calculator.net. (n.d.). Sample Size Calculator. Retrieved from https://www.calculator.net/sample-size-calculator.html