Data Analysis Toolkit

Arts and Wellbeing Indicators Project

UF | UNIVERSITY of FLORIDA

ART WORKS.

National Endowment for the Arts
arts.gov

CULTURE BUILDS FLORIDA
FLORIDA DEPARTMENT OF STATE
DIVISION OF CULTURAL AFFAIRS

Center for ARTS IN MEDICINE
UNIVERSITY OF FLORIDA / COLLEGE OF THE ARTS

Table of Contents

UF | UNIVERSITY of FLORIDA

ART WORKS.

National Endowment for the Arts
arts.gov

CULTURE BUILDS FLORIDA
FLORIDA DEPARTMENT OF STATE
DIVISION OF CULTURAL AFFAIRS

Center for ARTS IN MEDICINE
UNIVERSITY OF FLORIDA / COLLEGE OF THE ARTS

Introduction

The State of Florida Division of Cultural Affairs has partnered with the University of Florida Center for Arts in Medicine on the three-phase project to develop a set of indicators for associating the arts with wellbeing at the community level. The Arts and Wellbeing Indicators project supports the Division's strategic goal of promoting healthy, vibrant, and thriving communities.

The mission of the State of Florida Division of Cultural Affairs (DCA), as stated in its strategic plan for 2015-2020, is to "Advance, support, and promote arts and culture to strengthen the economy and quality of life for all Floridians." Further, the plan asserts a goal to "promote healthy, vibrant, thriving communities". The Arts and Wellbeing Indicators project is a step toward strengthening that mission, and a step toward documenting that Florida's investments in the arts have positive health impacts on Florida's communities. This work aligns with the DCA's commitment to advancing arts and culture in the State of Florida and makes it possible to provide important data to arts advocates and arts organizations in keeping with its strategic goal to "collect, distill, and disseminate current information that advances arts and culture in Florida."

The Arts and Wellbeing Indicators model is a tool for assessing the associations between arts participation and wellbeing in communities. It is important to note that association is distinct from correlation or causation, and that the Indicators model does not identify a direct cause and effect relationship between arts participation and wellbeing. The model includes the primary domains of wellness, arts, and community. Wellness encompasses health and quality of life; the arts domain encompasses participation, access, value, infrastructure, and investment; and the domain of community encompasses civic involvement, satisfaction with leadership, openness, safety, social capital, and satisfaction with community.

A single 24-question survey, which takes an average of 10 minutes to complete, was developed to assess each of the model's variables. Over the project's second and third phases, the survey was tested in nine Florida counties. An array of surveying methods, including paper and pencil, telephone and electronic methods, were tested and assessed for cost-effectiveness. The project also assessed and tested the reliability of survey outcomes, with overall findings confirming positive associations between arts participation and wellbeing, and the feasibility of the instrument for assessing these outcomes.

This toolkit is designed to guide management and analysis of Arts and Wellbeing Indicators survey data. This toolkit follows from the Arts and Wellbeing Indicators Survey Data Collection Toolkit. This toolkit will provide guidance for how to manage and analyze the survey data in SAS (Statistical Analysis Software). This toolkit also provides resources on how to read and interpret SAS output. This toolkit is intended for those who have experience in statistical programming and analyses (e.g., data analyst, statistician, etc.)

This toolkit provides step-by-step guidance on how to do a variety of data analyses used after collecting data concerning the Arts and Wellbeing Indicators survey. These analyses answer the following questions:

- What are the demographics of those who completed the Arts and Wellbeing Indicators survey?
- What are the differences between those that participate in formal arts versus informal arts versus no participation?
- How do you assess aesthetics, civic involvement, leadership, openness, social capital, and safety in communities and how do they associate with arts participation?

Materials Needed

- After data collection, datasets should be downloaded from whichever data management system has been used (SurveyMonkey, Qualtrics, etc.).
    - The recommended format is in excel: csv file.
    - If you are using multiple modes of data collection: It is recommended to stratify datasets depending on data collection mode. For example, if the data is collected on paper, this should be separated by file or category, from data collected electronically.
- Data Dictionary/Data Codebook and SAS codes
    - Please see **Appendix A** for how the indicators group into constructs.
    - Please see **Appendix B** for an example of a data dictionary.
    - The files needed also include the Codebook, Data Management SAS code and Data Analyses SAS code
        - Please note that the SAS codes are based on datasets from Phase III of the Arts and Wellbeing survey data collection. **It is recommended to alter these codes to fit your unique datasets and their locations on your computer**.
    - The SAS codes are also documented step-by-step, from importing datasets from Excel into creating an analytical dataset and analytical data tables.
    - All SAS codes and codebooks are available by request by contacting the University of Florida Center for Arts in Medicine: cam@arts.ufl.edu.
- Statistical Programming Software
    - This toolkit is for analyses conducted in:
        - SAS

https://www.sas.com/en_us/home.html

Data Management in SAS

If you have several datasets due to different modes of collection, it is recommended to use SAS to merge your data appropriately. Please refer to the Data Management SAS code.

Inclusion Criteria

Please refer to **Appendix C** for an example of the inclusion criteria that was implemented for Phase III of the Arts and Wellbeing Indicators Survey.

In order to build the statistical model, it is important to first conduct some preliminary data cleaning procedures. It is recommended to verify all respondent data before conducting analyses:

- Assess all data for missingness. It is recommended to only include observations that have at least half non-missing responses.

- In addition to significant missing responses, only include observations that have responses for zip-code of residence (question three on the survey), especially if the research question is focused on county or site-specific population.

    - o **Ensure that the zip-codes are within Florida and your area of interest** (i.e. county). Please refer to the Data Management SAS code, which merges the Florida zip code spreadsheet (Florida Zip Codes by City and County.xlsx) with the survey dataset.

- Include observations that have responses for arts participation (question 21_9 and 21_10 on the survey)

    - o These questions are used to create the predictor variable of arts participation.

- Include observations that have responses for **all** of the following demographics: age, gender, marital status, race/ethnicity, education and income

    - o These demographics are important to characterize the sample that is being surveyed.

Data Cleaning Procedures

1. In preparing to undergo statistical modeling, it is required to understand the variations in each of the variables. For this purpose, frequencies for all variables would have to be calculated.

- These variables include participation in the arts, self-reported health indicators measured through the standardized measures (PROMIS Global Short Form and the Short Flourishing Scale), the community vitality indicators (Aesthetics, leadership, openness to diversity, social offerings, civic involvement, social capital, safety) and socio-demographic variables (age, gender, race, ethnicity, education, income, and zip code).

- This univariate descriptive data allowed you to find errors in data entry (such as erroneous numbers for the variables).

- This step also allowed you to understand if there were any variables that displayed unusual or interesting variations. For instance, in the Phase III analyses, some of the items were assessed as yes vs. no while other questions were assessed with a Likert scale where respondents had to rate from very good/ good/ neither good or bad/ bad/ very bad such as those in the aesthetics and openness variables.

- Researchers can get a general understanding of the data and where to consolidate responses by running a univariate analysis. If it is found that these categories have few responses (<5), it is suggested to collapse categorical responses (i.e. very good and good| neither good or bad| bad and very bad)

2. Based on the prior steps, creation of new variables based on those already assessed is possible. For example, in Phase III, there are two questions to assess for arts participation—one for formal and one for informal arts participation. A new variable was created that reflect:

- Participation in both the formal and informal arts
- Participation in formal arts only
- Participation in informal arts only
- No participation in any arts activity

This will allow you to assess whether respondents had participated in various types of art forms with a single variable.

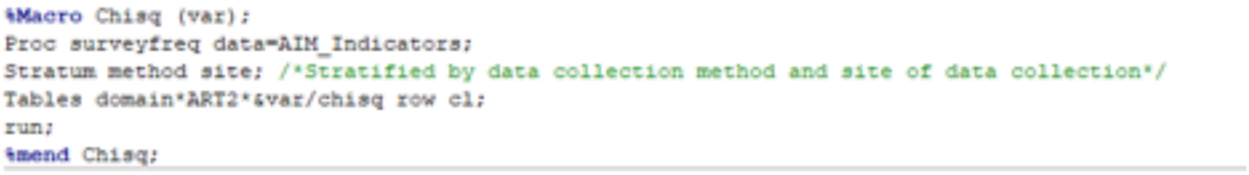
3. If standardized assessments are used, it is required that the calculation provided in the scoring guide of the assessment be strictly followed. For Phase III, based on the coding guide provided for the standardized tools, calculation of the total Physical health and total mental health (PROMIS scale) scores was carried out. Similarly, it is recommended to calculate total well-being score for the Short Flourishing Scale as given.

Logistic Regression Model Building Process in SAS

1. If new variables are created, bivariate descriptive analyses (i.e. frequencies, chi-square and t-tests) need to be run to check if the newly created variables show any unusual variations. Additionally, cross tabulations between the predictor variables and the outcome variable and other variables should be conducted to arrive at an insight into the bivariate relationships between any two variables.

    - For example, it is recommended to conduct bivariate analyses for the newly created arts participation variable and the socio-demographic variables of interest such as race, gender, age etc. This would guide in understanding if and why certain variables would lose significance in the logistic regression model. Individual variables and the arts participation should be cross-tabulated following which chi-squares (for categorical variables) and t-tests (for continuous variable of age) were used to identify factors significantly associated with informal or formal arts participation in the last 12 months.

        i. In Phase III, sampling occurred across different methods (paper, electronic) and in different counties. It is important to account for complex survey procedures (e.g. accounting for stratification using PROC SURVEYFREQ and PROC SURVEYMEANS in SAS). For example (as programmed in the **Data Analyses SAS code**):



```
%Macro Chisq (var);
Proc surveyfreq data=AIM_Indicators;
Stratum method site; /*Stratified by data collection method and site of data collection*/
Tables domain*ART2*&var/chisq row cl;
run;
%mend Chisq;
```

2. Then, predictors of interest can be classified into theoretically distinct groups. By building the models within those groups first, it will enable researchers to view how related variables work together. For instance, in Phase III, the objective was to understand whether participation in the arts was associated with health and well-being among those across Florida. Potential sets of variables included:

- Demographics (age, education, race, gender, socio-economic status)

- Arts participation (informal, formal, both formal and informal)

- Community vitality indicators (aesthetics, leadership, openness to diversity, social offering's, civic involvement, social capital, safety)

Note: Often, as seen in Phase III, the variables within a group are correlated, but not so much across groups. If everything is included in the model at once, it is hard to find any relationships. Therefore, each group could be built separately first, followed by building theoretically meaningful models with a solid understanding of how the pieces fit together.

3. In the model building process, bivariate logistic regression models have to be conducted to calculate unadjusted odds ratios between each (or each of the main) main outcome variable(s) and other variables of interest.

- In Phase III, bivariate logistic regression analyses were conducted for the community vitality indicators (aesthetics, leadership, openness to diversity, social offering's, civic involvement, social capital, and safety) and demographic covariates (age, education, race, gender, socio-economic status) and the outcome variable of arts participation as described earlier.

- For reference, a standard approach to model building can be applied (please see the Data Analyses Resources section of the toolkit); only those variables in univariate models which were statistically associated (at p-value <0.05) with the outcome variable should be retained in the model building process.

4. Then, multiple logistic regression models would have to be conducted to explore factors associated with the outcome variable and the predictor variables. In Phase III, arts participation was a four level variable, therefore, no participation in any art activity in the past 12 months was the reference group for participation in informal arts only in the past 12 months, for participation in formal arts only in the past 12 months and participation in both formal and informal arts activities in the past 12 months.

- In the multiple (multivariable?) model, variables are required to be entered following an order based on the researchers need as well as variables found to be of significance in prior studies. In Phase III, variables were included in the following order: socio-demographics, health variable followed by community vitality indicators. Assessment of multi-collinearity was carried out to identify highly associated independent variables before their inclusion into the models.

- Variables known to be highly correlated should be inserted into the models independently, and retained when found to be significant when controlling for other covariates.

- Following the inclusion of each variable, the decision to either drop or retain a variable should be based on whether its coefficient differed significantly from 0 (adjusting for the effects of the other variables), whether removal of the variable altered the remaining coefficients of other terms in the model by more than 20% and considering change in the overall fit of the model was improved by its addition. In some cases, a variable is retained in a model for statistical significance but later can be insignificant when other variables were dropped. In these few cases, the variable that became insignificant is retained in the model.

5. When reporting results of the crude and adjusted multiple logistic regression analyses, adjusted odds ratios (aORs) with 95% confidence intervals (95% CI) are to be reported.

Linear Regression Modeling in SAS

Given the research question, it may be beneficial to conduct multiple linear regression. If this is of interest, it is recommended to create a dummy variable of the four-level arts participation variable in order to assess the linear relationship between arts participation and PROMIS global physical health score, PROMIS global mental health score, and Short Flourishing Score.

- Similar to the logistic regression model building process, these linear regression analyses could only include indicators that are significantly different across levels of arts participation. The referent group is no participation in the arts in the last 12 months.

- When reporting results of the crude and adjusted multiple linear regression analyses, beta estimates, p-values and model R-squares are appropriate to report. Please see below for an example of a data table.

	crude			adjusted*			Adjusted (including all art participation groups)		
	β	P	Model R^2	β	P	Model R^2	β	P	Model R^2
Global Physical Health									
No Participation	-2.30	0.0012	0.007120	-2.30	0.0013	0.04794	ref	ref	0.05239
Informal Arts Participation Only	2.47	0.0210	0.004096	-1.77	0.0805	0.04307	0.38	0.7506	
Formal Arts Participation Only	0.10	0.8545	0.000022	-0.39	0.4764	0.04131	1.78	0.7506	
Both Informal and Formal Arts Participation	1.38	0.0022	0.006255	1.59	0.0004	0.04902	2.64	0.0003	

*Adjusted for: Gender, Age, Race/Ethnicity, education, income, and health checkup in last 12 months

Limitations to Analyses

Sample Size and Power

Please note that if the sample size is smaller, this would cause underpowered analyses. Issues with power will produce models not able to detect association between arts participation and other indicators. A smaller sample also means limited adjusted models. With an increase in sample, more parameters (variables) are able to be added to regression modeling.

Appendix A. Questions and Constructs

Construct	Survey Questions
Wellness	• In general, would you say your health is: • In general, how would you rate your physical health? • In general, how would you rate your mental health, including your mood and your ability to think? • In past 7 days: How often have you been bothered by emotional problems such as feeling anxious, depressed, or irritable? • In past 7 days: How would you rate your fatigue (tiredness) on average? • In past 7 days: How would you rate your pain on average?
Access to Care	• Do you have health insurance right now? • The availability and accessibility of quality healthcare
Healthcare Utilization	• Have you had a routine physical examinations or health check-up in the past twelve months?
Quality of Life	• In general, would you say your quality of life is: • In general, how would you rate your satisfaction with your social activities and relationships? • In general, please rate how well you carry out your usual social activities and roles. • To what extent are you able to carry out your everyday physical activities such as walking, climbing stairs, carrying groceries, or moving a chair? ***Short Flourishing Scale*** • I lead a purposeful and meaningful life • My social relationships are supportive and rewarding • I am engaged and interested in my daily activities • I actively contribute to the happiness and well-being of others • I am competent and capable in the activities that are important to me • I am a good person and live a good life • I am optimistic about my future • People respect me ***Arts and Wellbeing Indicators Original Questions*** • Do you think that the arts or creative activity currently contributes to your personal quality of life? • Do you think that the arts or creative activity currently contributes to your community's quality of life?

Arts Participation

*Arts and Wellbeing Indicators Original Question*s

- Attended any art activity in community in last 12 months
- Participated in any hands-on creative activity in last 12 months
- Participated in any recreational activities in last 12 months
- Approximately how many times in the past (30 days, twelve months) have you attended any arts activity in or near your community?
- Approximately how many times in the past (30 days, twelve months) have you participated in these hands-on creative activities?
- Approximately how many times in the past (30 days, twelve months) have you attended these recreational activities?

Access

- The availability and accessibility of arts and cultural opportunities, such as theater, museums, and music

Value

Infrastructure

Investment

- AEP V Data

Community

Civic Involvement
What residents give to the community

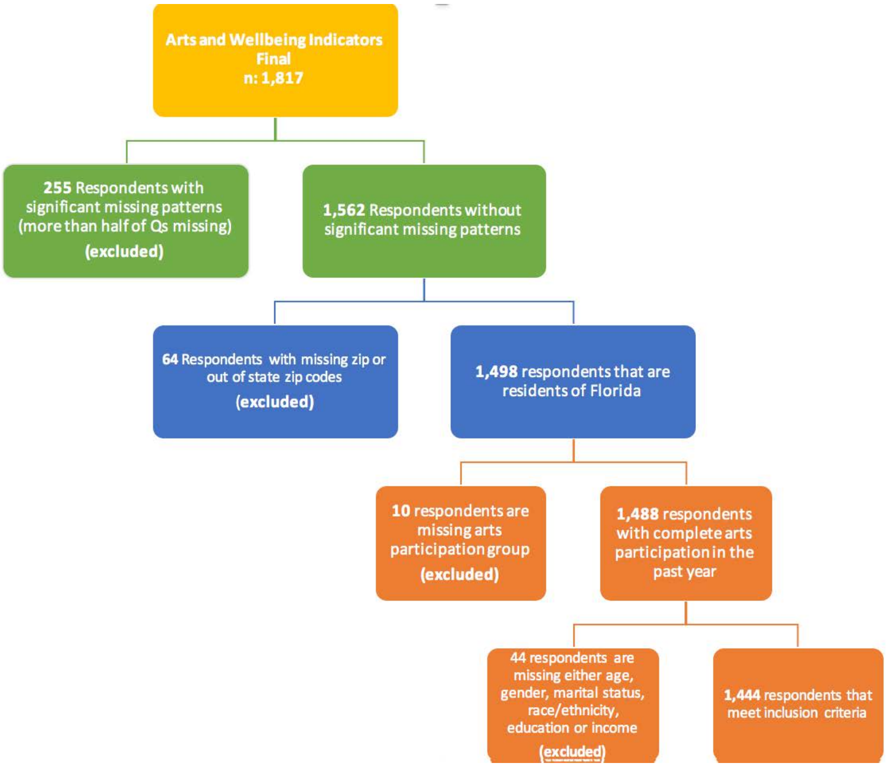
- Performed local volunteer work
- Attend local community meetings
- Voted in last local election
- Work with residents to make change
- Donated money to help a local organization
- Gave money or food to an individual in need
- Gave shelter to an individual in need
- Participated in an activity at social club/support group/church

Satisfaction with Leadership	• Rate leadership of elected officials • How much of the time do you think you can trust your local government (city/county) and elected officials to represent your interests?
Openness	• Good place for talented college graduates • Good place for immigrants • Good place for racial and ethnic minorities • Good place for gays and lesbians • Good place for family with children • Good place for senior citizens • Good place for young adults without children • Good place for artists • Good place for people with disabilities
Safety	• Safe place to live
Social Capital	• Close friends in this area are also friends with each other • Family in community • Close friends in community • Spend time with neighbors • How much people care about each other • Being a good place to meet people and make friends
Satisfaction with Community (Community Vitality Indicators)	How would you rate the following in relation to the county in which you currently live? *Aesthetics* • The availability of outdoor parks, playgrounds, and trails • The beauty or physical setting • Having a vibrant nightlife with restaurants, clubs, bars, etc. *Social Offerings* • The availability of social community events, such as festivals, picnics, parades, and street fairs

Appendix B: Example of Data Dictionary

Telephone Survey Variables	Paper and Electronic Survey Variables	SAS Code Variables	Question	Codes
respnum$	IDNew	_n_		
q1	Q1	Age	What is your age?	1, 0-17 years | 2, 18-35 years | 3, 36-59 years | 4, 60 years or older
q2	Q2	Gender	What is your gender?	1, Male | 2, Female | 3, Other
q3	Q3	Zipcode	What is your Zip Code (5-digit) in the county in which you live now?	
q4	Q4	Relshp_status	What is your relationship status?	1, Married | 2, Not Married | 3, In a relationship, living together | 4, In a relationship, not living together| 5, Divorced or separated | 6, I don't know | 7, I don't want to answer this
q5	Q5	Hispanic	Are you of Hispanic, Latino or Spanish origin?	1, Yes (Hispanic, Latino, or Spanish) | 2, No (Hispanic, Latino, or Spanish) | 3, I don't know

Appendix C. Example of Data Management Flowchart


Arts and Wellbeing Indicators Final
n: 1,817
255 Respondents with significant missing patterns (more than half of Qs missing)
(excluded)
1,562 Respondents without significant missing patterns
64 Respondents with missing zip or out of state zip codes
(excluded)
1,498 respondents that are residents of Florida
10 respondents are missing arts participation group
(excluded)
1,488 respondents with complete arts participation in the past year
44 respondents are missing either age, gender, marital status, race/ethnicity, education or income
(excluded)
1,444 respondents that meet inclusion criteria

Data Analyses Resources

Free Training in SAS:

https://support.sas.com/training/us/sp1.html

Example of SAS programming of logistic regression models:

https://stats.idre.ucla.edu/sas/dae/logit-regression/

How to interpret odds ratios:

https://stats.idre.ucla.edu/sas/faq/how-do-i-interpret-odds-ratios-in-logistic-regression/

Information on Linear Regression Analyses:

https://stats.idre.ucla.edu/stata/dae/multivariate-regression-analysis/

Information on Logistic Regression Analyses: Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.