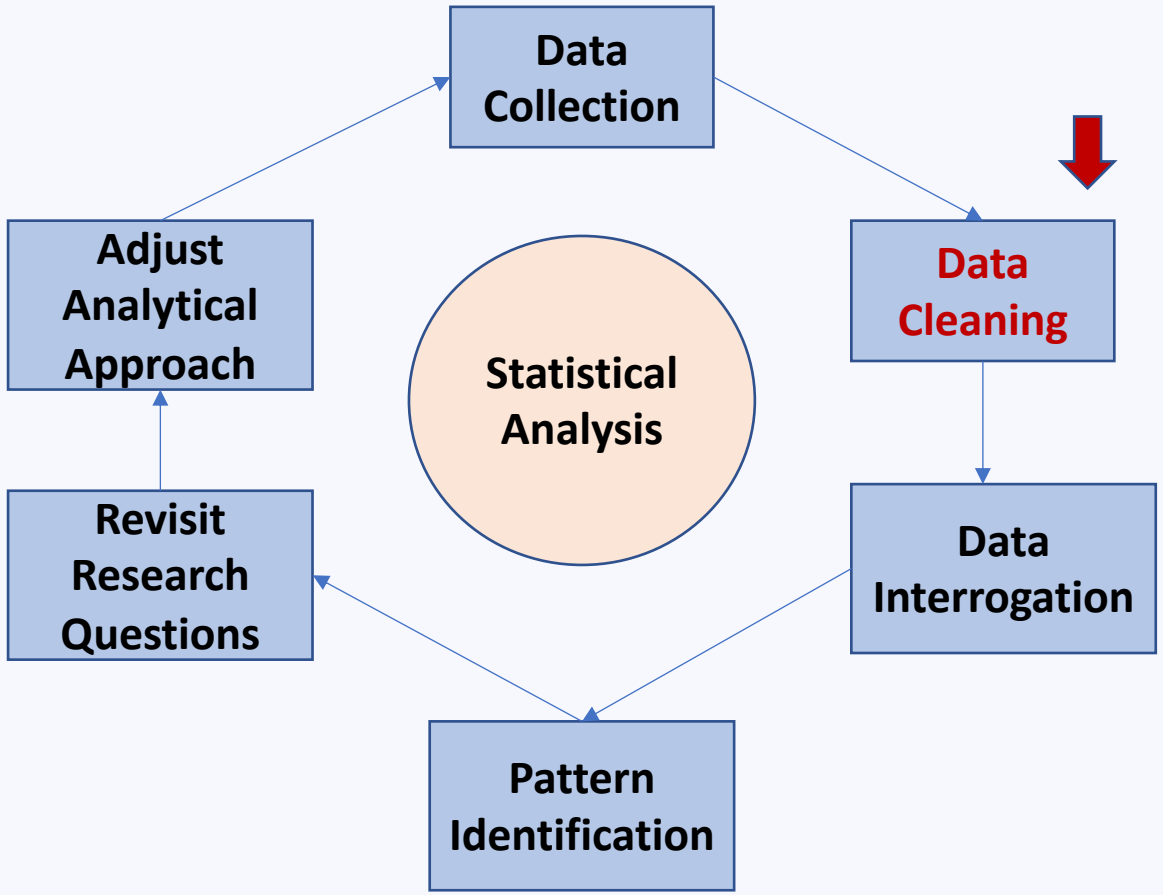
Data Cleaning



Data Cleaning

- Deduplication
  - A respondent has submitted more than one survey
- Disqualification
  - Respondent answered questions randomly
  - Respondent chose the same answer for each question
  - Respondent answered inconsistently

Data Cleaning

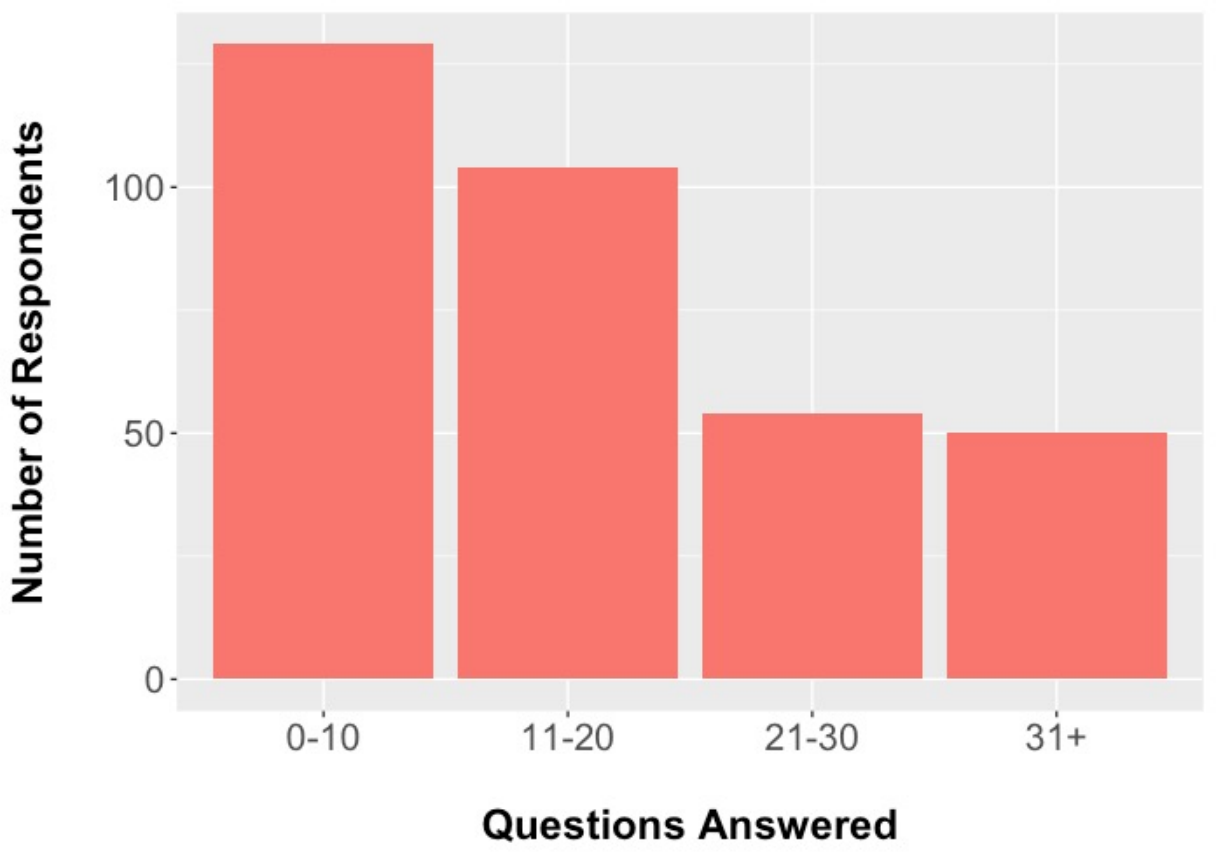
- Skip Logic Enforcement
  - Manual recoding of responses
- Finalization
  - Extraneous answers removed
  - Check for completeness of administrative data
  - One survey complete from each respondent
  - **Determine what is a complete survey**

Frequencies

Q01		Q02		Q03			
A	B	A	B	A	B	C	D
90	70	80	80	40	34	20	22

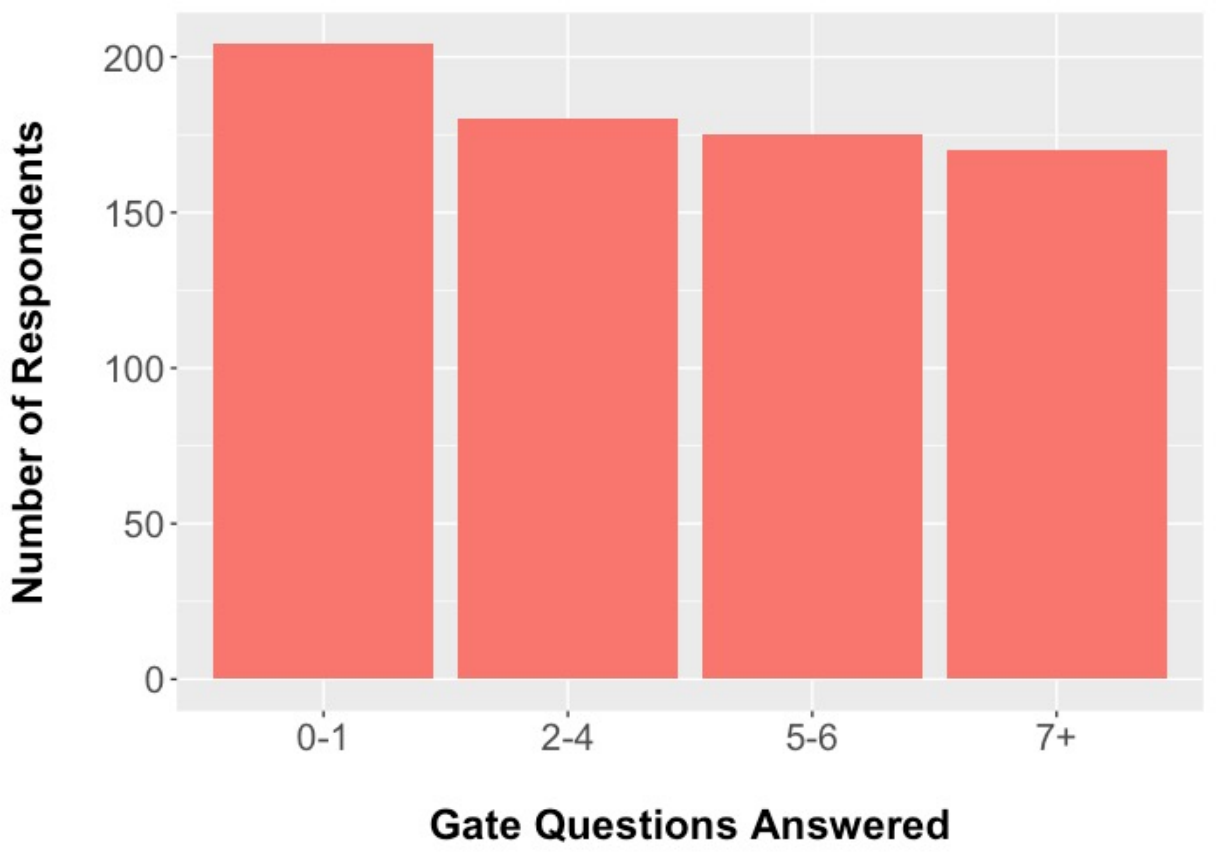
Data Cleaning and Survey Completion Metrics

- What is considered a completed survey?
  - Respondent progressed to the final question
  - All questions have been answered
  - A certain threshold of questions have been answered
  - A calculated score
    - Weighted importance of questions
  - Any question has been answered

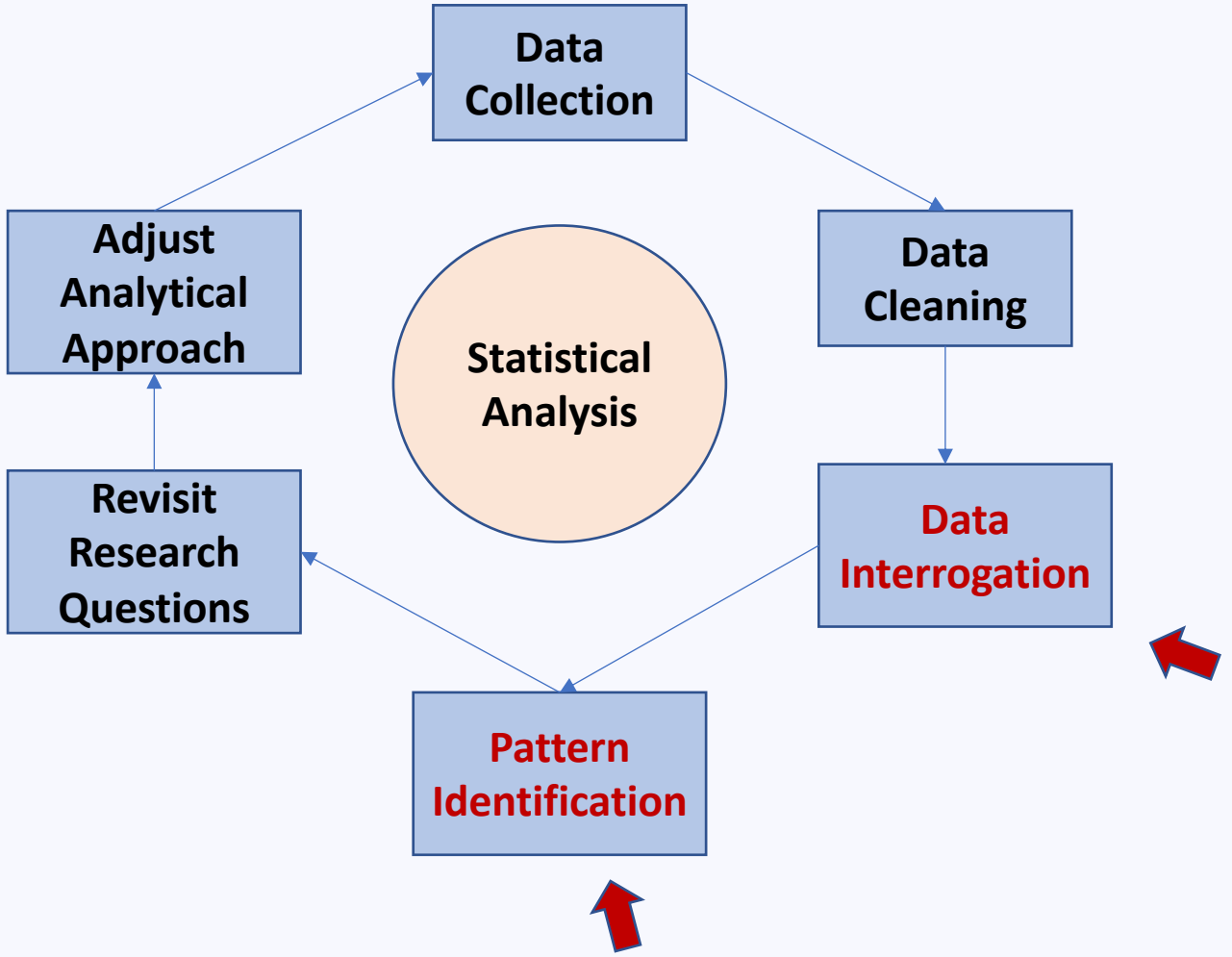


Data Cleaning and Survey Completion Metrics

- What is considered a completed survey?
  - Respondent progressed to the final question
  - All questions have been answered
  - A certain threshold of questions have been answered
  - A calculated score
    - Weighted importance of questions
  - Any question has been answered

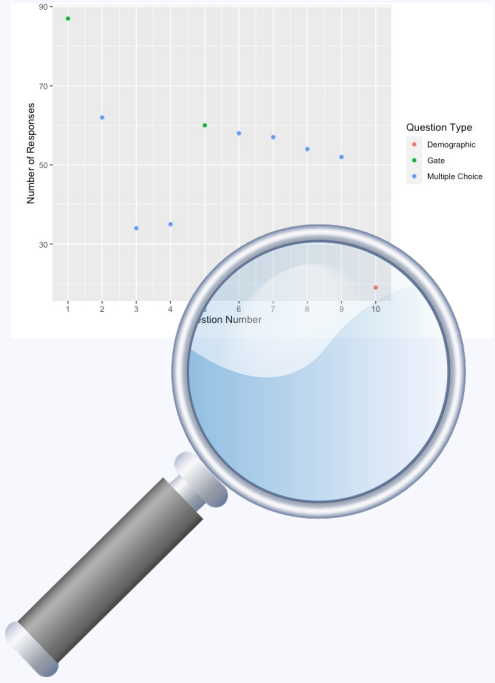


Exploratory Data Analysis



Exploratory Data Analysis

How do we get a final dataset?



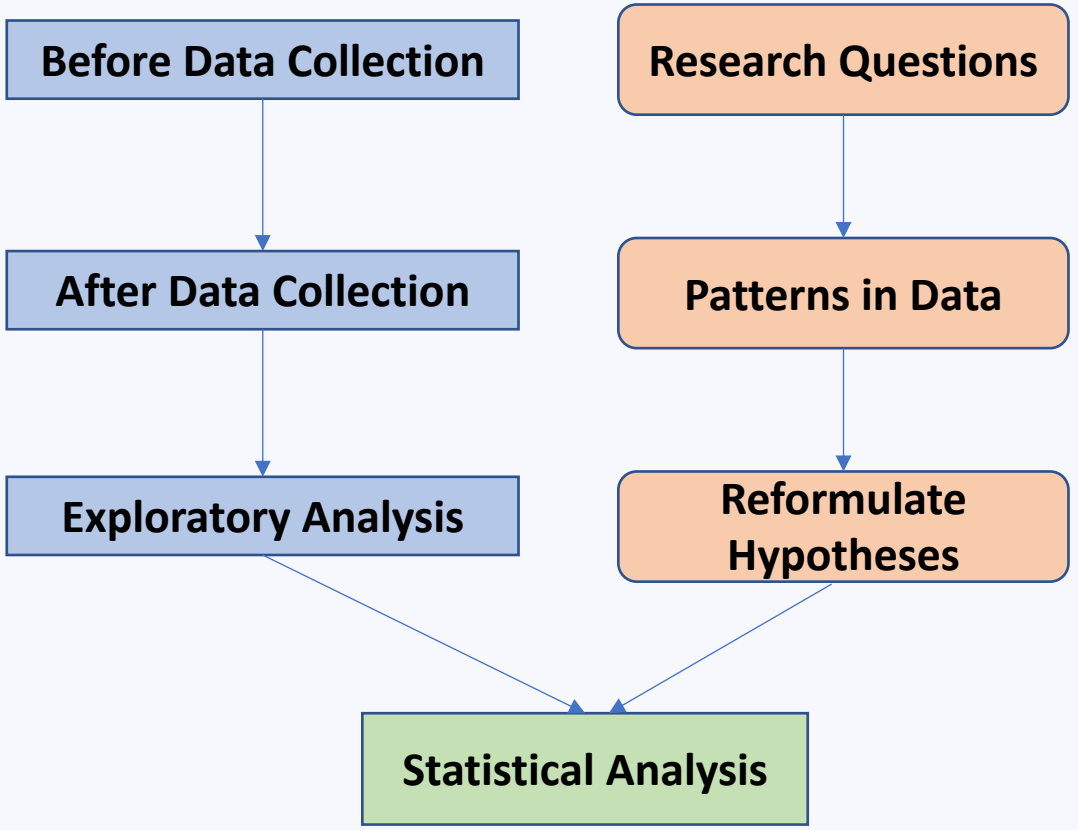
Are there consistent themes in our questions?

What trends should we investigate?

How do we interpret our results?

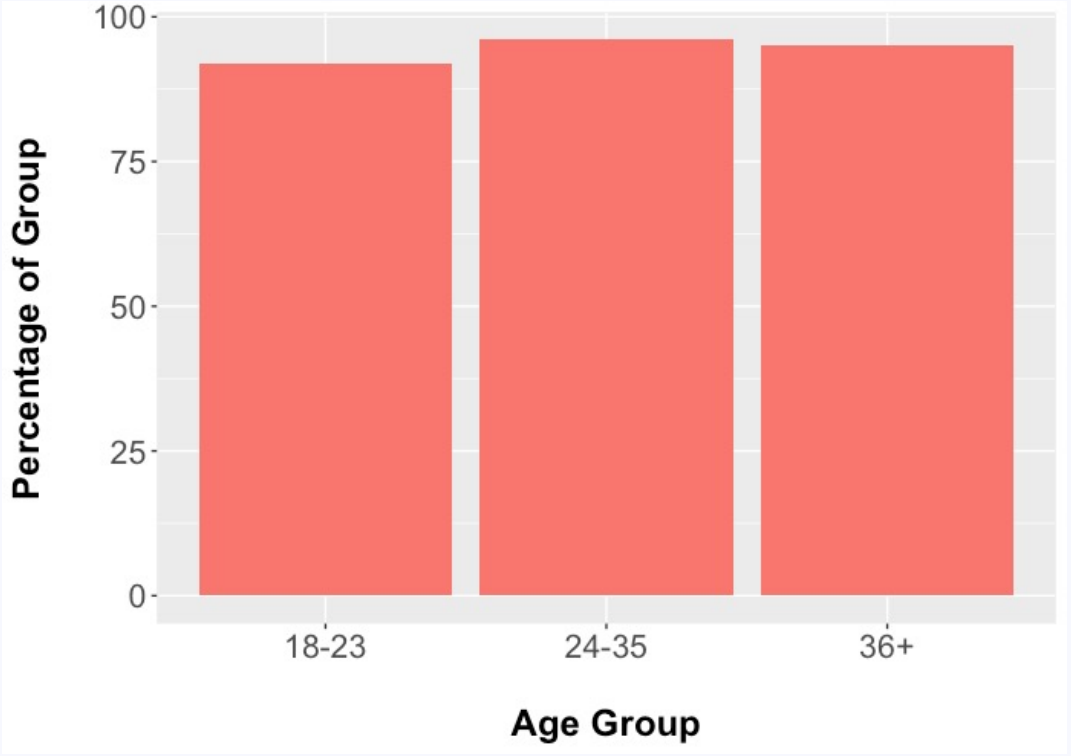
Exploratory Data Analysis

- Revisit Initial Research Questions
  - Are we interested in patterns among demographic subgroups?
    - Did enough respondents complete the survey?
  - Which experiences are respondents evaluating in the survey?
  - Is there a time-sensitive program being evaluated?
  - Are respondents representative of the population?



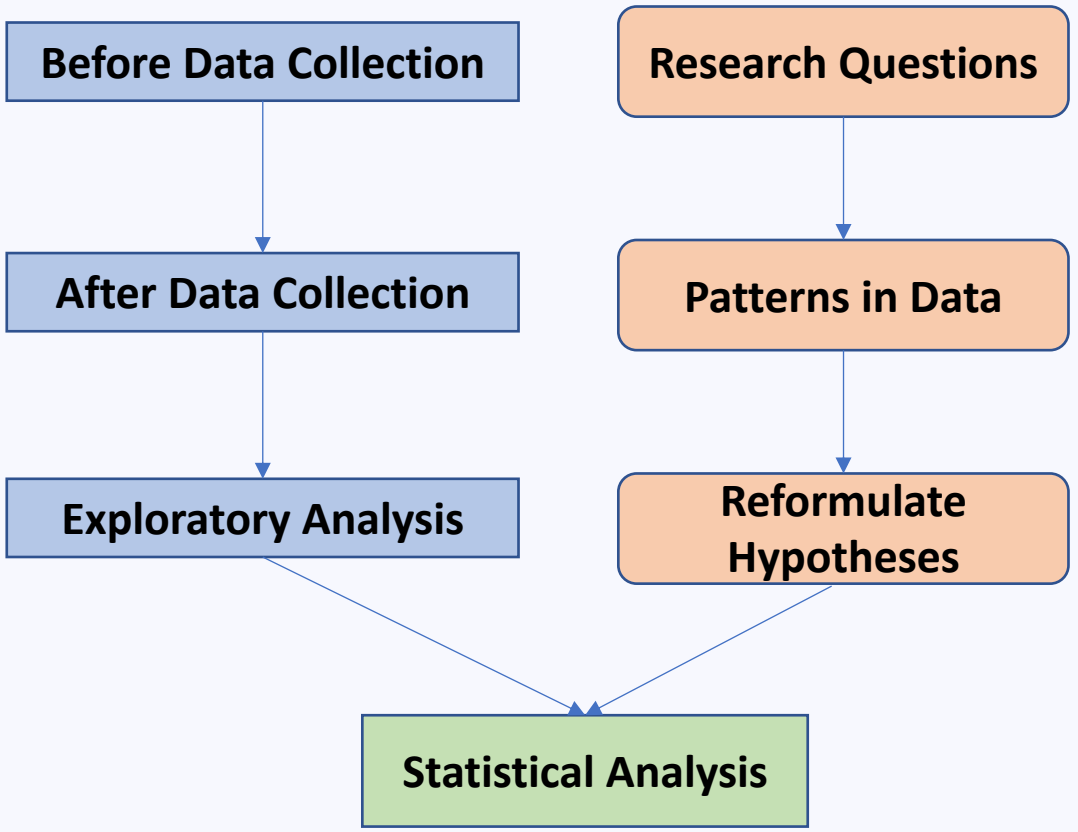
Exploratory Data Analysis

- Revisit Initial Research Questions
  - Are we interested in patterns among demographic subgroups?
    - Did enough respondents complete the survey?
  - Which experiences are respondents evaluating in the survey?
  - Is there a time-sensitive program being evaluated?
  - Are respondents representative of the population?



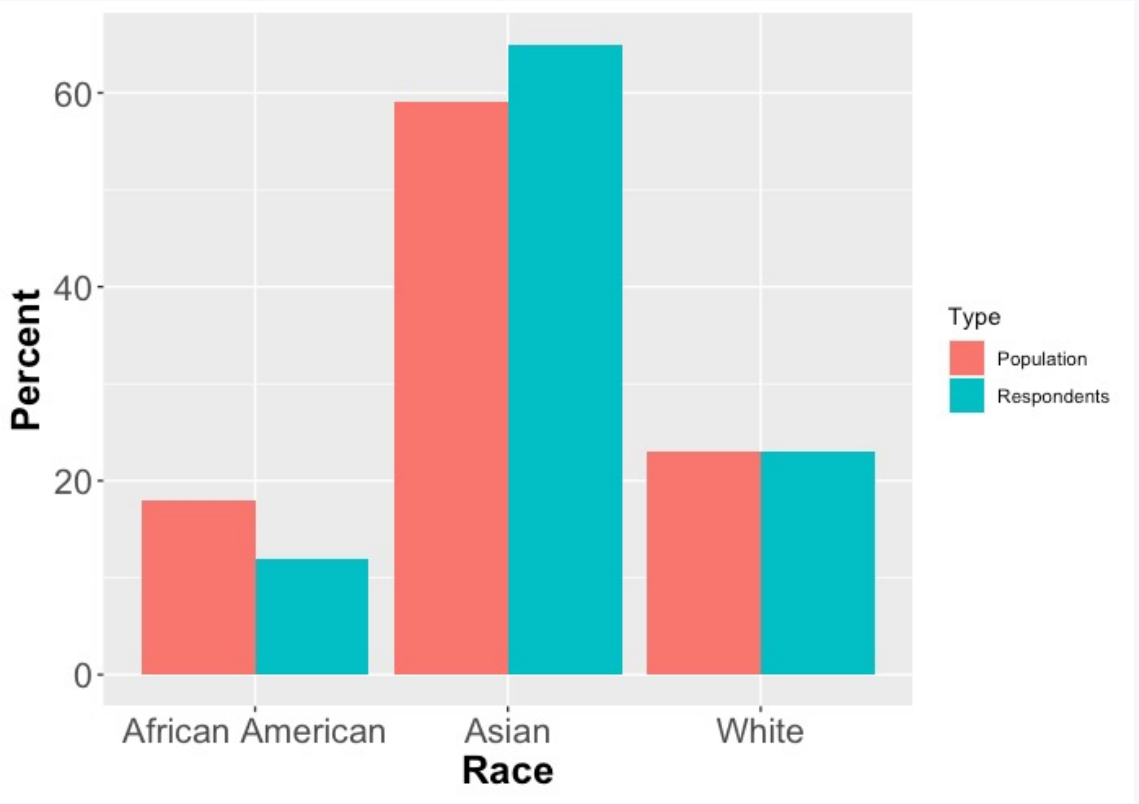
Exploratory Data Analysis

- Revisit Initial Research Questions
  - Which components of our services are most important?
  - What information can we gain from the survey that is actionable?
  - How do we gain information from our survey to create or modify a new program?



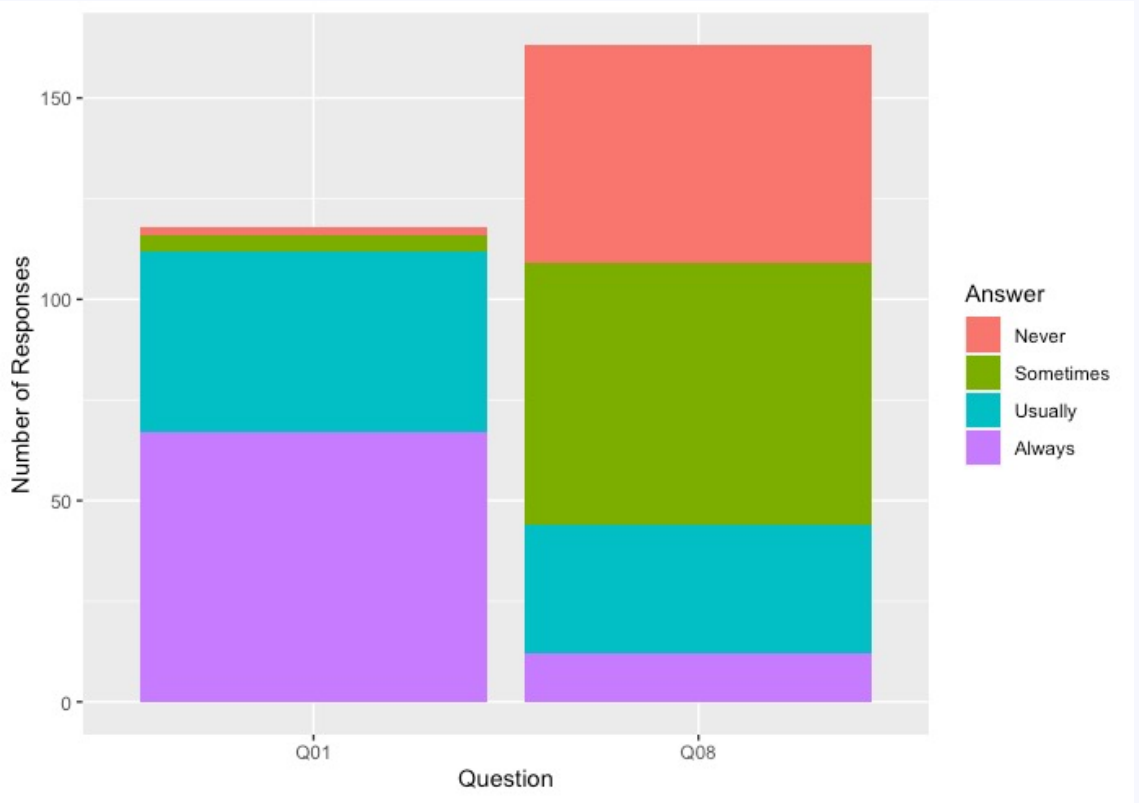
Survey Response Rates

- Vary depending on population
  - Can vary by subgroup
  - Reveal how representative the respondents are
- There is no response rate less than 100% that guarantees representativeness
- Effects may go in two directions:
  - Subgroups may inherently answer questions differently
  - Unrepresentative respondents may provide less valid results



Answer Choice Counts

- Simplest method to view results
- Can be represented as counts or proportions
- Shows the relative popularity of answer choices for questions with the same scale



Crosstabs

- Reveal how those who responded to one questions responded to another
  - Can reveal hidden subgroups and preferences within the data
- Do not require advanced statistics
- Do require a large number of respondents to extract meaningful results

- Identify errors in survey design
- Can reveal trends in the data that would likely be found in a formal hypothesis test
- May also reveal multicollinearity/confounds in demographic variables
  - Suggests that a more advanced statistical technique will be needed to separate the effects of collinear independent variables

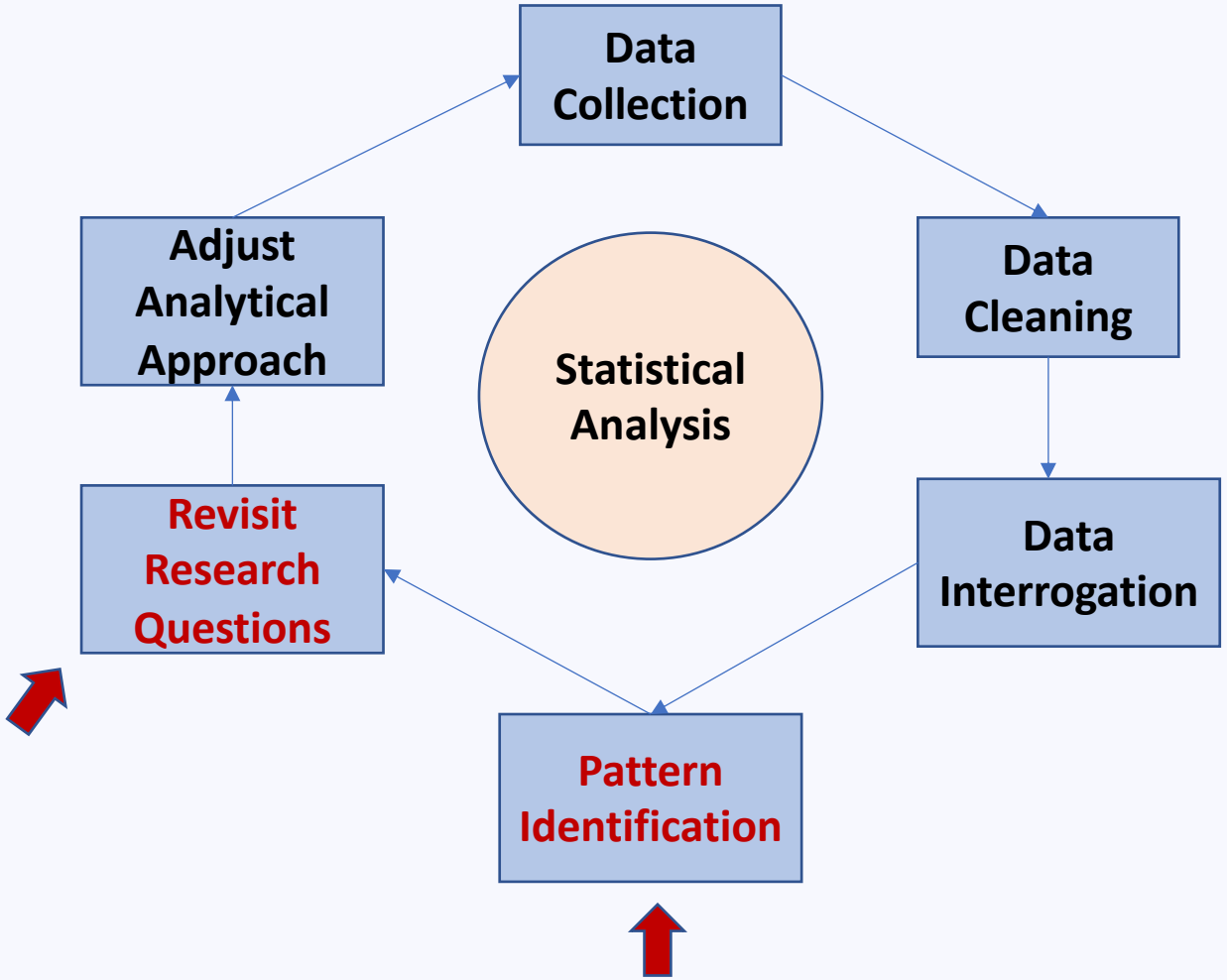
Crosstabs

	Q01: How many times did you visit the library this year?			
Q08: How satisfied are you with library services?	0	1-2	3-5	6+
Very Satisfied		15	23	30
Satisfied		25	14	8
Neither Satisfied nor Dissatisfied		4	5	2
Dissatisfied		8	3	1
Very Dissatisfied		6	1	0

Crosstabs

	Q34: What is your age?			
Q08: How satisfied are you with library services?	18-23	24-35	36-64	65+
Very Satisfied	76	54	20	30
Satisfied	34	9	18	8
Neither Satisfied nor Dissatisfied	12	2	5	5
Dissatisfied	5	1	6	12
Very Dissatisfied	1	0	7	8

Item Analysis



Survey Validity & Reliability

- Validity
  - Does the survey measure what we intend for it to measure?
- Reliability
  - Will the survey produce the same results if given a second time?
  - Will each respondent understand the question in the same way?

Reliability Statistic: Cronbach's Alpha

- Not typically used for overall survey reliability, but for composite development

$$\alpha = \frac{N\bar{c}}{\bar{v} + (N-1)\bar{c}}$$

Composites

- Improve interpretability
  - Generate themes that can easily be explained in reports
  - Individual questions become components
- Easier to track changes over time
- Less reliant on specific question wording
- Improves construct validity
  - Possibility to compare against other thematic composites
- Generated after data collection

Customer Service Composite

1. How long did it take to get a response to your inquiry?
2. When you spoke with an agent, was he/she able to help you solve the problem?
3. How often did you contact customer service?
4. Was the agent polite when speaking with you?

Item Analysis with Cronbach's Alpha

- For composite development
  - A set of question that have one theme or are associated with one outcome measure
    - (customer service, overall satisfaction, etc.)
- Measure not a statistical test
  - Internal variance and inter-item covariance terms
- Heuristic to select questions for thematic composites

$$\alpha = \frac{N\bar{c}}{\bar{v} + (N-1)\bar{c}}$$

Item Analysis with Cronbach's Alpha

- Alpha statistic represents the internal consistency of a set of questions (composite)
- Item-total correlation provides a coherence metric for one question compared against all others
- Statistical software will also provide the alpha statistic in the case when a question is removed
  - Same information as item-total correlation

Reliability statistics

Cronbach's alpha
0.866

	Corrected item-total correlation	Cronbach's alpha if item deleted
Q1	0.830	0.820
Q2	0.682	0.839
Q3	0.746	0.831
Q4	0.494	0.893
Q5	0.700	0.838
Q6	0.682	0.840

Cronbach's alpha	Internal consistency
$\alpha \geq 0.9$	Excellent
$0.9 > \alpha \geq 0.8$	Good
$0.8 > \alpha \geq 0.7$	Acceptable
$0.7 > \alpha \geq 0.6$	Questionable
$0.6 > \alpha \geq 0.5$	Poor
$0.5 > \alpha$	Unacceptable

Cronbach’s Alpha vs. Pearson Correlation

- Cronbach’s Alpha
  - Measure of reliability
  - Used to make a decision about thematic composites
  - No straightforward method for hypothesis testing
  - Comparison among multiple (2+) questions
- Pearson Correlation
  - Measure of correlation
  - Used to determine covariance among questions
  - Used in hypothesis testing
  - Comparison only between two questions

Individual Question: Scoring Methods

- Represent answer choices scales in numeric form
- Proportional Scoring
  - Score of each answer choice is set relative to the score for the whole scale
- Top Box Scoring
  - Only the most positive answer choices are given a nonzero value

Proportional Scoring
Never (0), Sometimes (0.33), Usually (0.67), Always (1)

Top Box Scoring
Very Satisfied (1), Satisfied (1), Neither Satisfied nor Dissatisfied (0), Dissatisfied (0), Very Dissatisfied (0)

Proportional vs. Top Box

- Proportional scoring allows for comparisons between questions with different scales

- Top Box scoring collapses the scale into a more digestible form for reporting
  - Particularly useful when there are few negative responses

Proportional Scoring
Never (0), Sometimes (0.33), Usually (0.67), Always (1)

Top Box Scoring
Very Satisfied (1), Satisfied (1), Neither Satisfied nor Dissatisfied (0), Dissatisfied (0), Very Dissatisfied (0)

Composite Scoring

- More complex than individual question scoring
  - Missing responses
    - When to disqualify an individual's composite score
  - Weighting possibilities
  - Imputation
- Abstracted from the original survey results
  - More easily interpreted and useful for year-over-year comparisons

Q01(Sometimes, 0.33); Q02(True, 1); Q03(Satisfied, 0.75); Q04(False, 0) =

(0.33 + 1 + 0.75 + 0) / 4 = 0.52

Q01(Sometimes, 0.33); Q02(NA); Q03(Satisfied, 0.75); Q04(False, 0) =

(0.33 + 0.75 + 0) / 3 = 0.36

Composite Scoring

- More complex than individual question scoring
  - Missing responses
    - When to disqualify an individual's composite score
  - Weighting possibilities
- Abstracted from the original survey results
  - More easily interpreted and useful for year-over-year comparisons
    - Can change question text but keep same composite

Half-Scoring

- At least half of the individual items must be answered for the composite score to be recorded

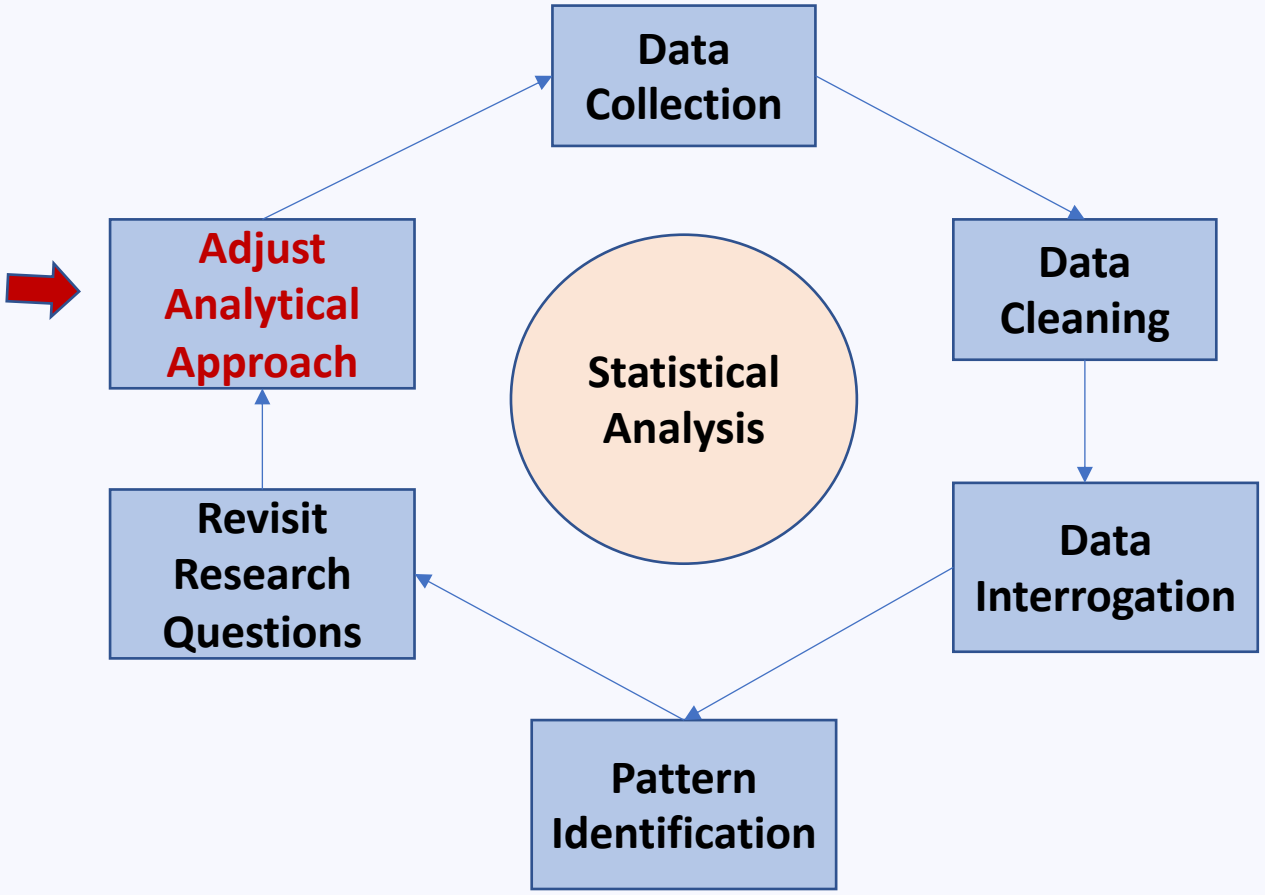
Proportional Scoring

- All items are included in the composite and scored based on the size of the scale

Top Box

- All items are included in the composite, but only the choices are collapsed into 0-1 scores

Statistical Analysis



Mean Comparisons: t-tests

- Compare mean scores
- Hypothesis Testing
  - Year-over-year changes
  - Comparisons between subgroups
  - Comparisons between composites

$$t = \frac{m - \mu}{s/\sqrt{n}}$$

t = Student's t-test

m = mean

μ = theoretical value

s = standard deviation

n = variable set size

Pearson Coefficient

- Among the most useful statistics in survey research
  - Used in hypothesis testing
- Correlation between survey questions
- Uncover which question(s) drive performance for a specific service
  - Often used to measure correlation between a specific question and overall satisfaction

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

Pearson Coefficient

	Visit Frequency	Customer Service	Opening Hours	Software Availablity	Study Rooms	Overall Satisfaction
Visit Frequency	1	0.76	0.9	0.23	-0.1	0.6
Customer Service	0.76	1	0.6	0.1	0.14	0.9
Opening Hours	0.9	0.6	1	0.4	0.95	0.65
Software Availability	0.23	0.1	0.4	1	0.01	0.32
Study Rooms	-0.1	0.14	0.95	0.01	1	0.57
Overall Satisfaction	0.6	0.9	0.65	0.32	0.57	1

Overall Satisfaction Question: On a scale from 1 to 10, with 10 being the most satisfied, how satisfied are you with the library's current services?

Pearson Coefficient

- Among the most useful statistics in survey research
  - Used in hypothesis testing
- Correlation between survey questions
- Uncover which question(s) drive performance for a specific service
  - Often used to measure correlation between a specific question and overall satisfaction

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

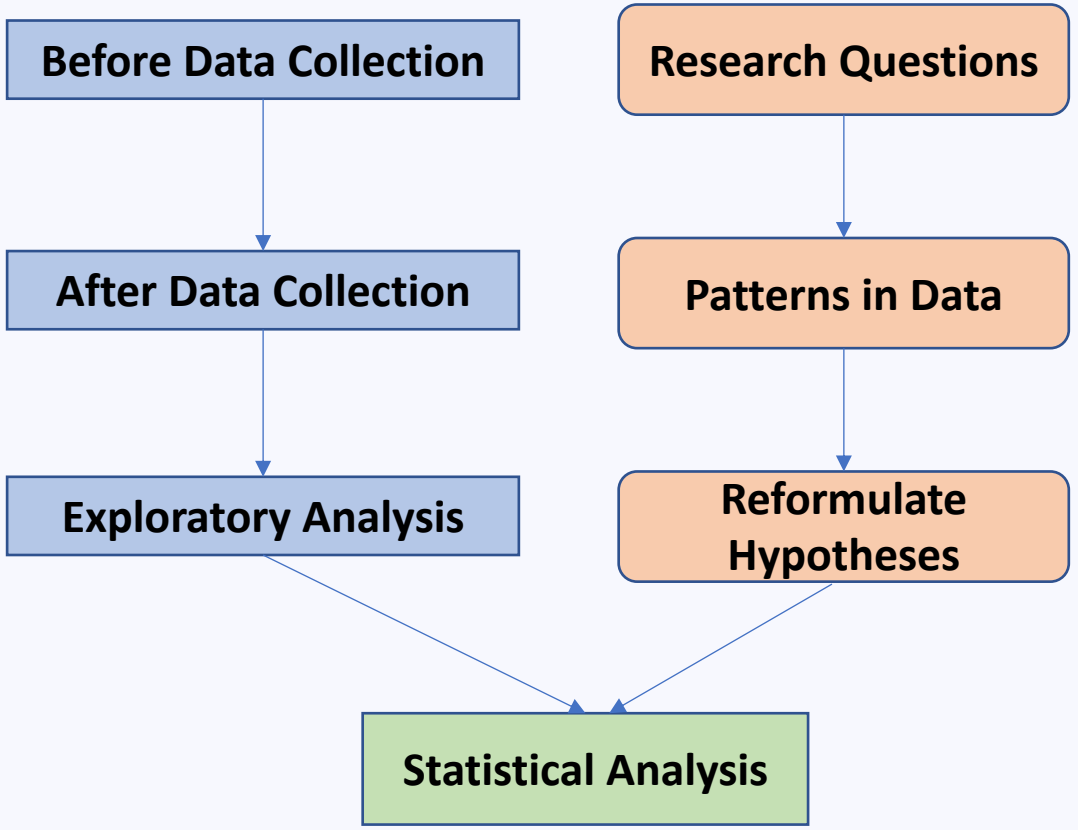
$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

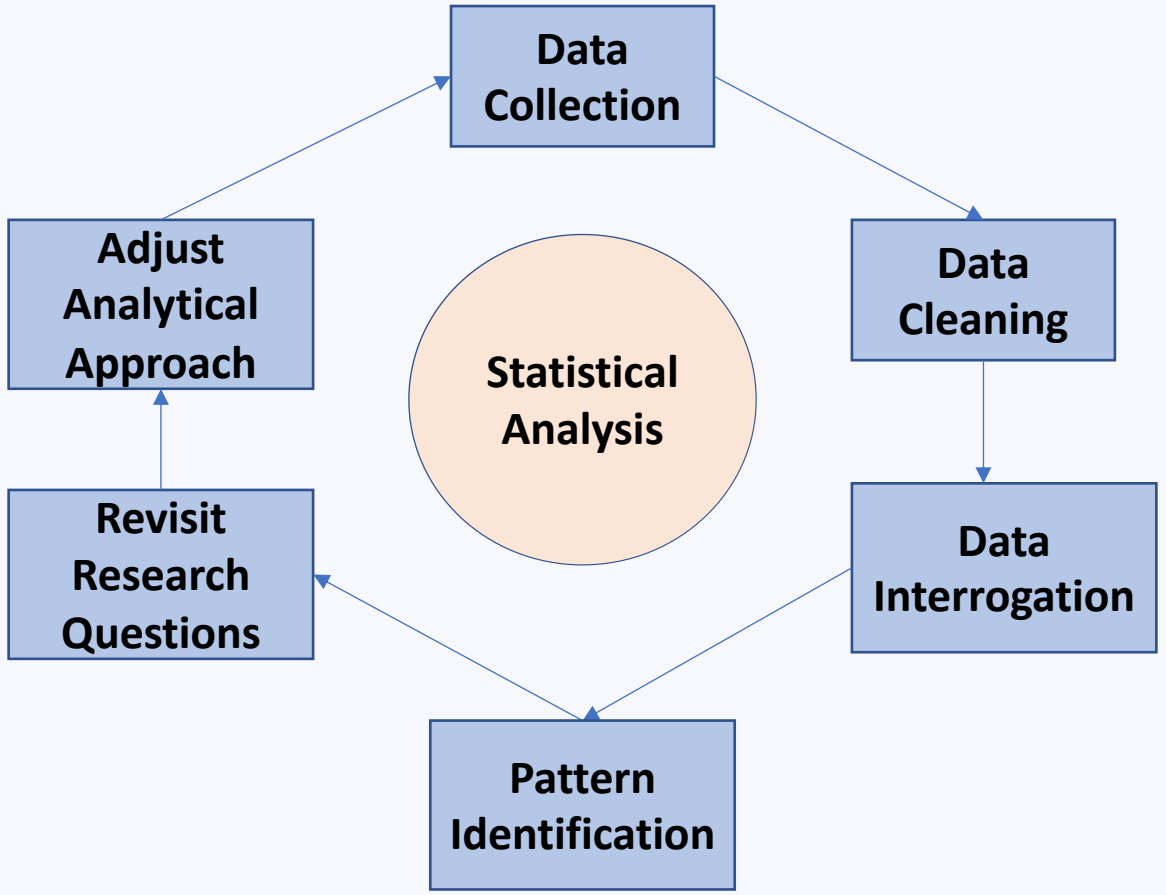
Answering Research Questions

- Theory
  - Services can be divided into components that drive performance on specific measures
- Representativeness
  - Demographics
- Change over Time
  - New programs or interventions
- Interpretability
  - Communication with stakeholders



Conclusions

- Data cleaning and deduplication preserve survey integrity
- Exploratory data analysis provides a chance to detect trends and gain intuition
- Composite formation improves both reliability and validity
- Hypothesis testing based on expert knowledge



Fall Workshop: Survey Reporting

- Regression analysis
- Preparation for future survey cycles
  - Reliability testing
  - Survey structure review and question editing
- Report structure and development
  - Visualization principles
  - Interpretation of findings
- Dashboarding and written reports