COLERIDGE INITIATIVE

Website

Reviewing Export Files

Expectations

At the conclusion of this training, you should be able to:

1. Navigate the export application
2. Review export requests for disclosive information
3. Assess complimentary disclosure risks

Motivation and Outline

In this presentation, we will provide examples of common export requests using fake data and show you how to evaluate them in accordance with dataset-specific rules. You can apply the techniques you learn here to the files you review.

1. Describing an export request
2. Accessing export requests
3. How to assess files for disclosure review
4. Walk-through examples

What is an export request?

An export request is when an ADRF user properly requests a file to be transfered outside the ADRF for future use. These files can be code files, graph outputs, tabular outputs, or a word document. Any type of file needs to be reviewed for disclosive information. Disclosive information is anything that will not pass your disclosure rules.

What is needed in an export request?

1. Files for Export
    - The files you want outside of the ADRF.
2. Files for Export Documentation
    - These are the supporting files that contain the underlying counts, data, and code used to create the files for export.
3. Documentation Memo
    - This is generally a .txt or .doc file that contains detailed information about each file for export and its corresponding files for export documentation.

ADRF User Guide and Video Walk-Through

The user guide is a great reference for the export process or using the ADRF.
Check out the **ADRF User Guide**

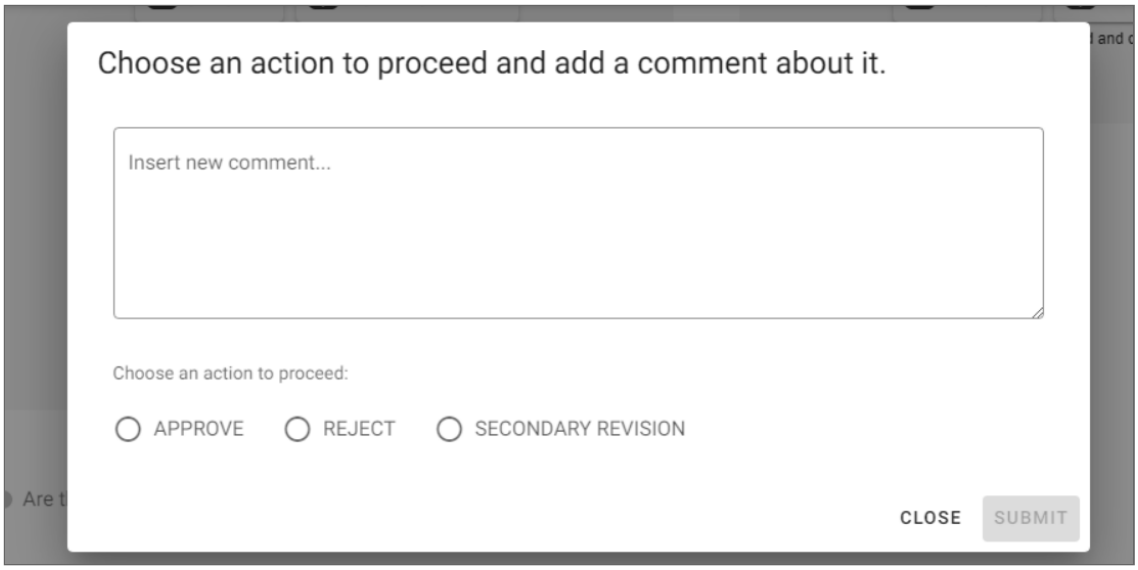
Check out the Export Module Video Walk-through
Navigate to the **ADRF**

How to review files?

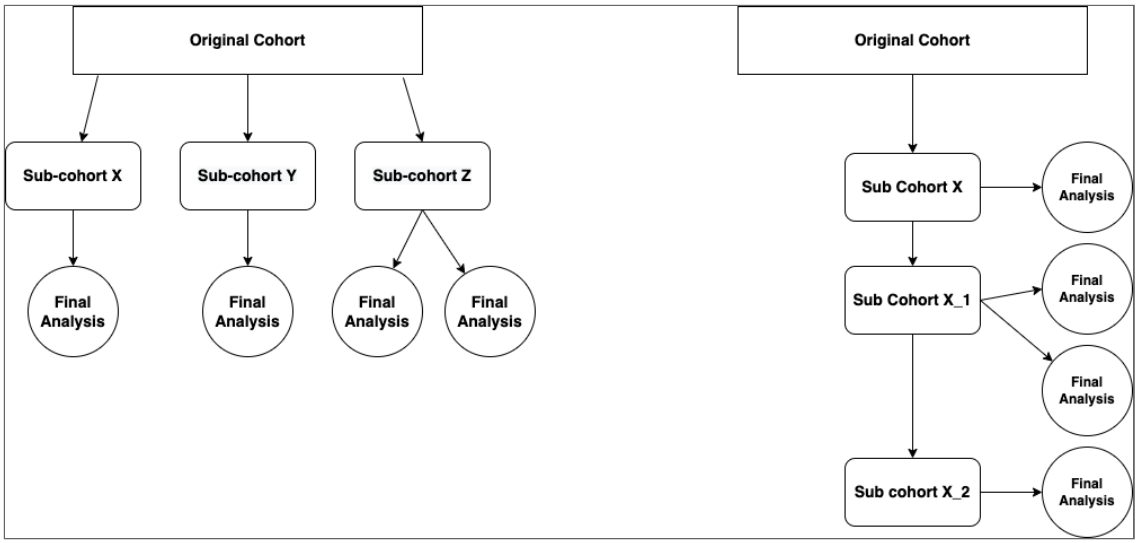
1. Look at what is included in the export, is there anything missing?
    - Evaluate the documentation memo
        - Do the export files have the appropriate supporting files?
        - Is the researcher describing their cohort and output tables correctly?
        - If missing supporting files, documentation files, or export files then send the export request back to the researcher.
2. Look at files and underlying counts.
    - Check the underlying counts, do they pass disclosure review?
    - How are the statistics created? Are the underlying counts 10 or greater?
        - Fuzzy percentiles, means, sums, etc...

- Is there a code file present? Is it free from references of data?

3. Complimentary disclosure
    - Are any of the export files created from the same cohort? Or, created from multiple subsets?
        - If so, need to assess for complimentary disclosure.


Choose an action to proceed and add a comment about it.
Insert new comment...
Choose an action to proceed:
APPROVE
REJECT
SECONDARY REVISION
CLOSE
SUBMIT

Understanding complimentary disclosure


Original Cohort
Sub-cohort X
Sub-cohort Y
Sub-cohort Z
Final Analysis
Final Analysis
Final Analysis
Final Analysis
Original Cohort
Sub Cohort X
Sub Cohort X_1
Sub cohort X_2
Final Analysis
Final Analysis
Final Analysis
Final Analysis

Define the rules

- General rules: No cell should have fewer than 10 observations unless otherwise designated. Always report the total number of observations.
    - Aggregation: If a table contains sensitive cells (i.e. fewer than 10 observations), users can aggregate (collapse) those categories.
- Suppression: When sensitive cells still occur and no further grouping is appropriate, the procedure is to suppress the cell (remove its value), and then suppress other cells to stop the first cell from being determined, also known as secondary suppression.
    - Secondary suppression: secondary suppression is the suppression of other cells or marginal totals in the table so that the suppressed cell(s) cannot be recalculated.

- Percentages: Report the number for both the numerator and denominator. Output needs to be suppressed where either, or both, of the counts used to calculate the percentage, proportion, or ratio have been suppressed. Round percentages calculated from unweighted counts to 1 decimal place. Do not report 0 or 100%.
- Percentiles: Do not report exact percentiles. Users can calculate a fuzzy median by averaging the true 45 and 55 percentiles. This logic can be used when calculating any percentile.
- Maxima and Minima: Suppress maximum and minimum values in general. Top-coded values may be considered for release.
- Cell values: round all reported values to the nearest sensible units.
- Weighted Data: Report both unweighted and weighted counts.

Export Example 1

We want to export this table:

- Documentation memo:
    - export_1.csv
        - Average yearly wages for individuals their first year after exiting TANF.
        - export_1_data.csv
        - exports.ipynb

A data.frame: 5 × 2

avg_wage	year
<dbl>	<dbl>
39029	2015
40257	2016
56987	2017
75908	2018
89032	2019

We need to see proof of the underlying counts:

A data.frame: 5 × 3

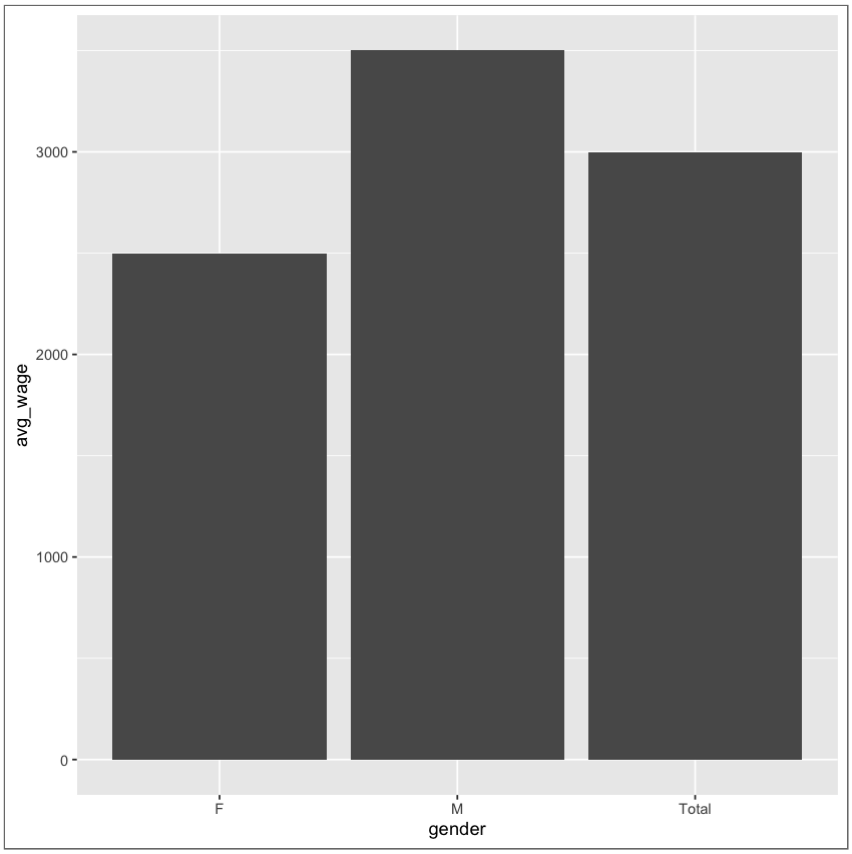
avg_wage	year	count_ssn
<dbl>	<dbl>	<dbl>
39029	2015	767
40257	2016	890
56987	2017	543
75908	2018	987
89032	2019	231

We can release this table because all counts are greater than or equal to 10.

Export Example 2

We want to export this figure:

- Documentation memo:
    - export_2.png
        - Counts of genders for those that exited TANF between the years 2015 and 2019.
        - export_2_data.csv
        - exports.ipynb


3000
2000
1000
0
avg_wage
F
M
Total
gender

We need to show the underlying counts:

A data.frame: 3 × 3

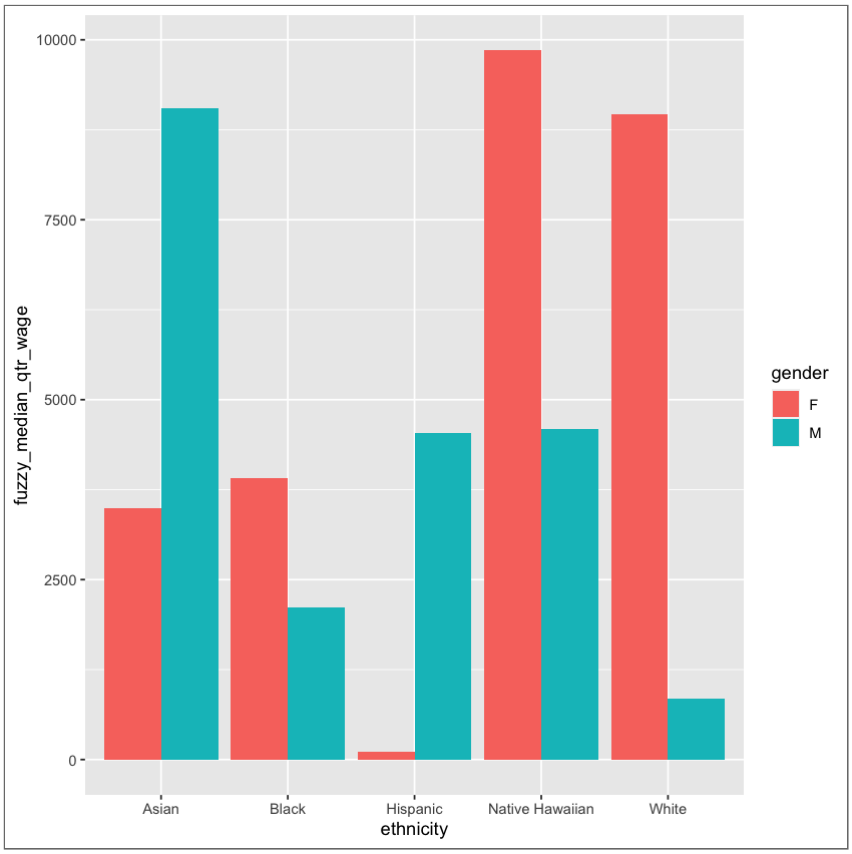
gender	avg_wage	count_ssn
<fct>	<dbl>	<dbl>
M	3500	10
F	2500	5
Total	3000	15

This graph cannot be released. For this to pass disclosure review, we need to redact the average wage for the values `F` and `Total` or redact the values `M` and `F`.

Export Example 3

We want to export this figure:

- Documentation memo:
    - export_3.png
        - The fuzzy median wage for ethnicity broken down by gender. The column `n` is the total distinct count of people in our cohort.
        - export_3_data.csv
        - exports.ipynb


10000
7500
5000
2500
0
fuzzy_median_qtr_wage
Asian
Black
Hispanic
Native Hawaiian
White
ethnicity
gender
F
M

We need to show the underlying counts:

A data.frame: 10 × 5

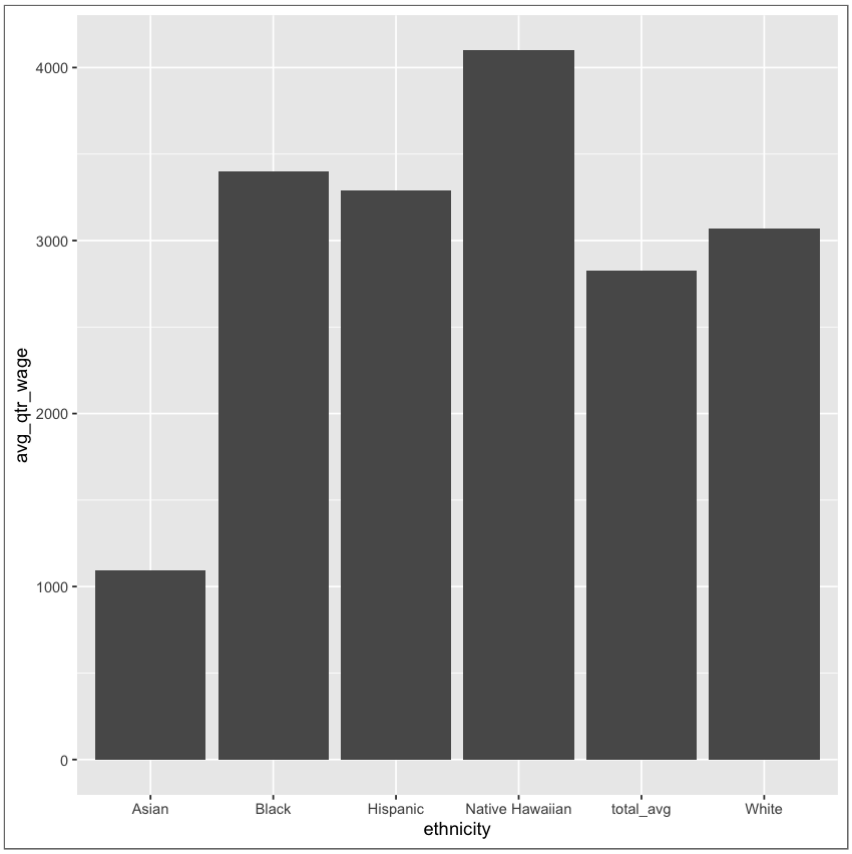
gender	ethnicity	median_qtr_wage	fuzzy_median_qtr_wage	n_counts
<fct>	<fct>	<dbl>	<dbl>	<dbl>
M	Asian	9043	9056	94
F	Asian	3486	3499	984
M	Black	2134	2109	89
F	Black	3904	3911	485
M	Hispanic	4567	4530	98
F	Hispanic	124	109	9
F	Native Hawaiian	9860	9850	45
M	Native Hawaiian	4598	4587	204
F	White	8954	8967	42
M	White	854	849	2

As it currently is, this graph will not pass disclosure review. We need to redact the values for Hispanic female and White male.

Export Example 4

We want to export this figure:

- Documentation memo:
    - export_4.png
        - Average quarterly wage broken down by ethnicity.
        - export_4_data.csv
        - exports.ipynb


4000
3000
2000
1000
0
avg_qtr_wage
Asian
Black
Hispanic
Native Hawaiian
total_avg
White
ethnicity

We need to show the underlying counts:

A data.frame: 7 × 3

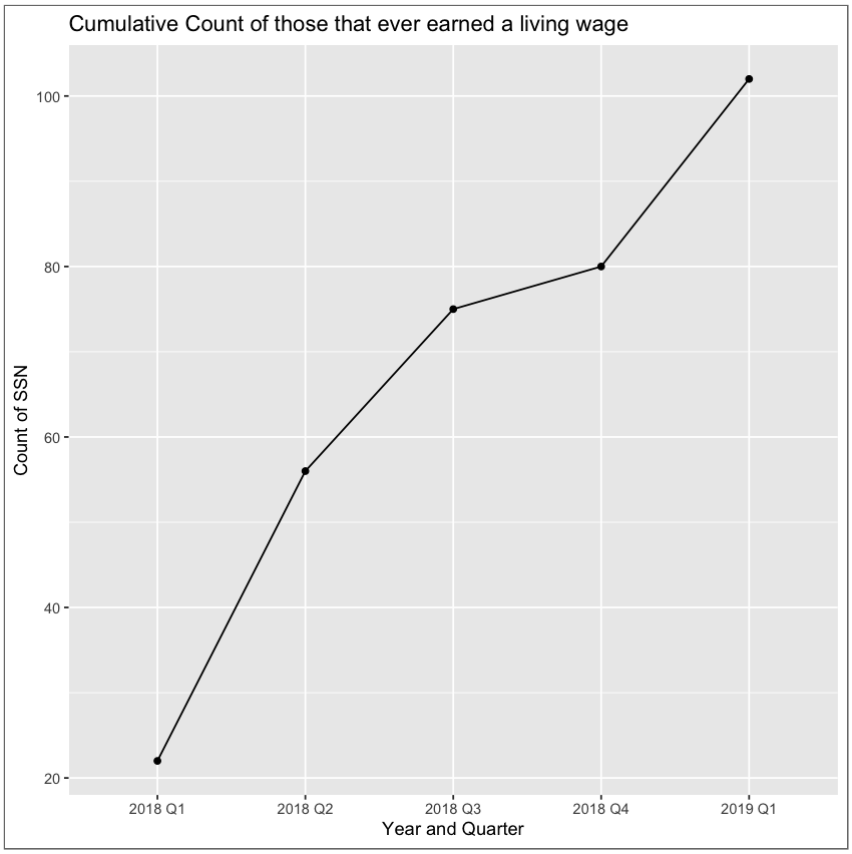
ethnicity	fuzzy_median_wage	count_ssn
<fct>	<dbl>	<dbl>
Hispanic	3287	10
Native Hawaiian	4098	59
Asian	1094	109
White	3068	32
Black	3398	11
other	2001	9
total_avg	2824	230

This graph will not pass disclosure review. The graph does not contain the fuzzy median wage value for the ethnicity category `other`. However, it does contain the total average for every ethnicity category, so we need to redact the `total_avg` median wage value.

Export Example 5

We want to export this figure:

- Documentation memo:
    - export_5.png
        - A cumulative count of individuals that ever earned a living wage from 2018 quarter 1 to 2019 quarter 1.
        - export_5_data.csv
        - exports.ipynb


Cumulative Count of those that ever earned a living wage
Count of SSN
100
80
60
40
20
2018 Q1
2018 Q2
2018 Q3
2018 Q4
2019 Q1
Year and Quarter

We need to show the underlying counts:

A data.frame: 5 × 2

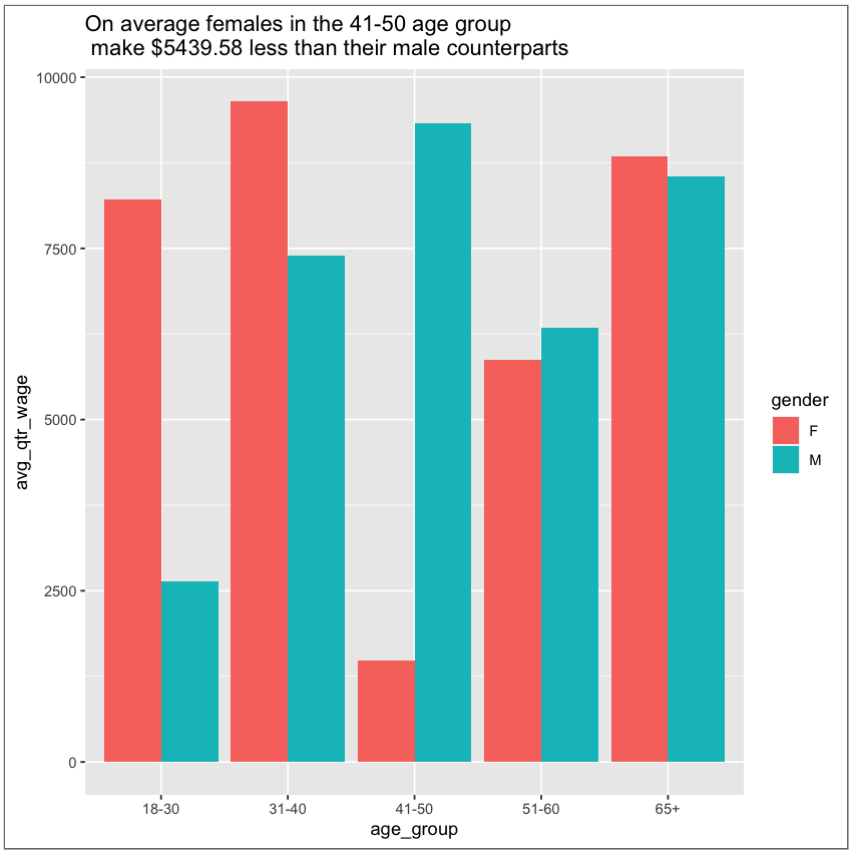
year_qtr	count_ssn
<fct>	<dbl>
2018 Q1	22
2018 Q2	56
2018 Q3	75
2018 Q4	80
2019 Q1	102

This graph will not pass disclosure review. Since this graph is representing a cumulative count, we need to focus on the differences between the quarters. Between quarters 3 and 4 in 2018 there is a difference of 5.

Export Example 6

We want to export this figure:

- Documentation memo:
    - export_6.png
        - Plotting the average quarterly wage by age group and gender. This export uses the same subset as export_7.png and export_8.png.
        - export_6_data.csv
        - exports.ipynb


On average females in the 41-50 age group
make $5439.58 less than their male counterparts
10000
7500
5000
2500
0
avg_qtr_wage
18-30
31-40
41-50
51-60
65+
age_group
gender
F
M

We need to show the underlying counts:

A data.frame: 10 × 4

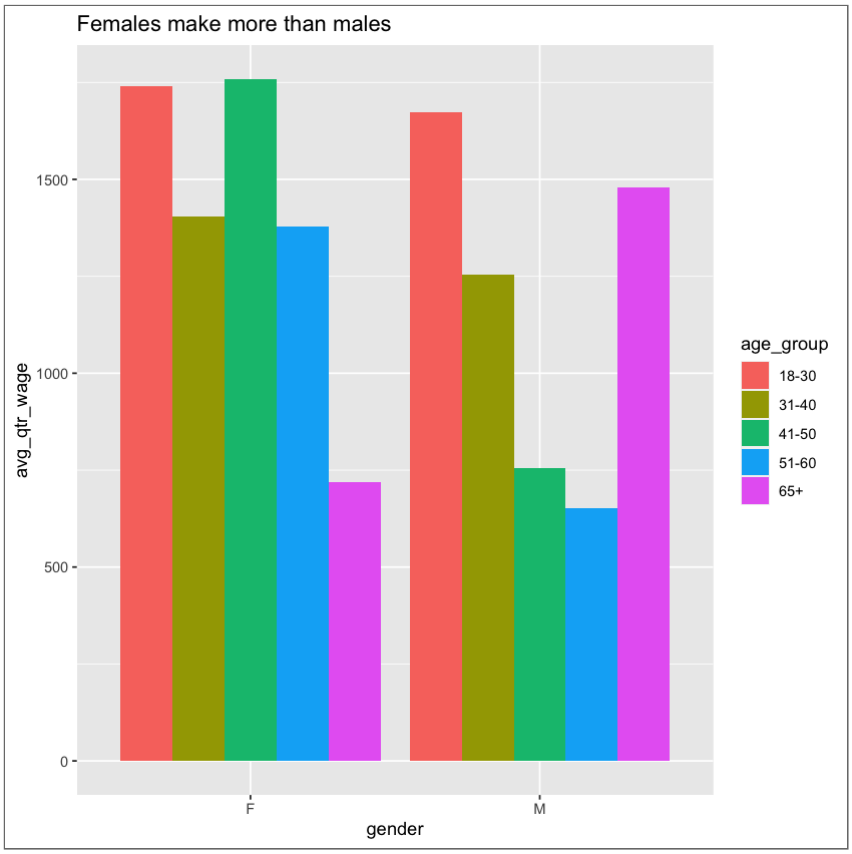
age_group	gender	avg_qtr_wage	total_n
<fct>	<fct>	<int>	<dbl>
18-30	M	2633	466
31-40	M	7388	340
41-50	M	9333	442
51-60	M	6341	553
65+	M	8545	721
18-30	F	8212	530
31-40	F	9642	808
41-50	F	1475	193
51-60	F	5868	312
65+	F	8845	775

This graph will not pass review. The title of the graph contains disclosive information. Is this a sensible unit to round to?

Export Example 7

We want to export this figure:

- Documentation memo:
    - export_7.png
        - Average wages broken down by gender. This export uses the same subset as export_6.png and export_8.png.
        - export_7_data.csv
        - exports.ipynb


Females make more than males
avg_qtr_wage
1500
1000
500
0
F
M
gender
age_group
18-30
31-40
41-50
51-60
65+

We need to show the underlying counts:

A data.frame: 10 × 5

age_group	gender	avg_qtr_wage	degree	n
<fct>	<fct>	<int>	<fct>	<int>
18-30	M	1673	Associates	10
31-40	M	1253	Associates	41
41-50	M	755	Associates	31
51-60	M	652	Associates	5
65+	M	1480	Associates	32
18-30	F	1741	Associates	14
31-40	F	1405	Associates	39
41-50	F	1759	Associates	16
51-60	F	1378	Associates	35
65+	F	719	Associates	37

This graph will not pass disclosure review. The value for males in the age group 65+ needs to be redacted.

Export Example 8

We want to export this figure:

- Documentation memo:
    - export_8.csv
        - the count of gender by age group. The total_n_by_age_gender is the unique count of individuals in this sample. This export uses the same subset as export_6.png and export_7.png.
        - export_8_data.csv
        - exports.ipynb

```
`summarise()` ungrouping output (override
with `.groups` argument)
```

A tibble: 5 × 2

age_group	total_n_by_age_gender
<fct>	<int>
18-30	1000
31-40	1152
41-50	639
51-60	869
65+	1500

This export will not pass disclosure review because the total count for gender by age group contains more observations than the wages by age and gender bar plot. There is a fundamental issue with the cohort that the researcher needs to address.

Next Steps

- At this point we have files that will pass disclosure and files that will not pass disclosure. We will reject this export request and list all the disclosure issues and send to the researcher. They will correct the issues and resubmit the export.

Final Notes and General Rules

- If you reject an export, let Coleridge know because the export will come back to us. We will also need to reject it.
- Keep in mind the disclosure unit of interest. In most cases it will be the SSN value.
- If a researcher submits multiple requests within a the same time frame, reject the exports and have the researcher submit all files as 1 export. This makes it easier to assess for complimentary disclosure.
- When reviewing code, make sure the code does not contain references to data or statistical results.
- If you have any questions, please reach out.