STA130 Fall 2019 – T0107

Week 1: Data Visualization

(Materials used in this presentation are provided by the U of T Statistics Department)

Agenda

- Introduction
- Administrative Information
- Visualizations and Vocabulary list
  - E.g. Homework question 1
- Group discussion
- Writing example
- Short writing exercise

Introduction

- L0101, Bethany White, Monday 10am-12pm
- T0107, Vivian Ngo, Friday 10am-12pm, UC330

Administrative Information

- Tutorial is 20% of your grade
  - 10 tutorials
  - Attendance is mandatory ** (10:10 – 12)
  - 1pt homework before class, 1pt attendance, 4pts in-class exercise (writing or presentation)
  - Next tutorial is a half tutorial (mentorship program, 3%)
- Submit work to Quercus
- Tutorial is not for troubleshooting/programming help
- Tutorial IS a safe environment

Visualizations and Vocabulary list

Vocabularies when describing distributions of variables or relationships between two variables:

Bar graphs, histograms:

1. Where it is centered (towards the left, right, middle)
2. How much spread (relative to what?)
3. The tails relative to a normal distribution (fat-tailed or heavy-tailed and thin-tailed)
4. Modes: where, how many, unimodal, bimodal, multimodal, uniform
5. Symmetric, left-skewed, right-skewed
6. Outliers, extreme values
7. Frequency (which category occurred the most or least often; data concentrated near a particular value or category)

Scatterplots (bivariate or pairwise scatterplots):

1. Strong / weak relationship
2. Linear (positive or negative) / nonlinear relationship
3. Outliers (deviation from what?)
4. Any visible clusters forming
5. Each dot represents ...

Visualizations and Vocabulary list

General considerations:

- What are the most effective types of graphs to summarize information in categorical or quantitative variables?
- What does the distribution tell you about for each types of data (categorical or quantitative)?
- How do you describe a histogram or a scatterplot? (refer to this week's vocabulary list)

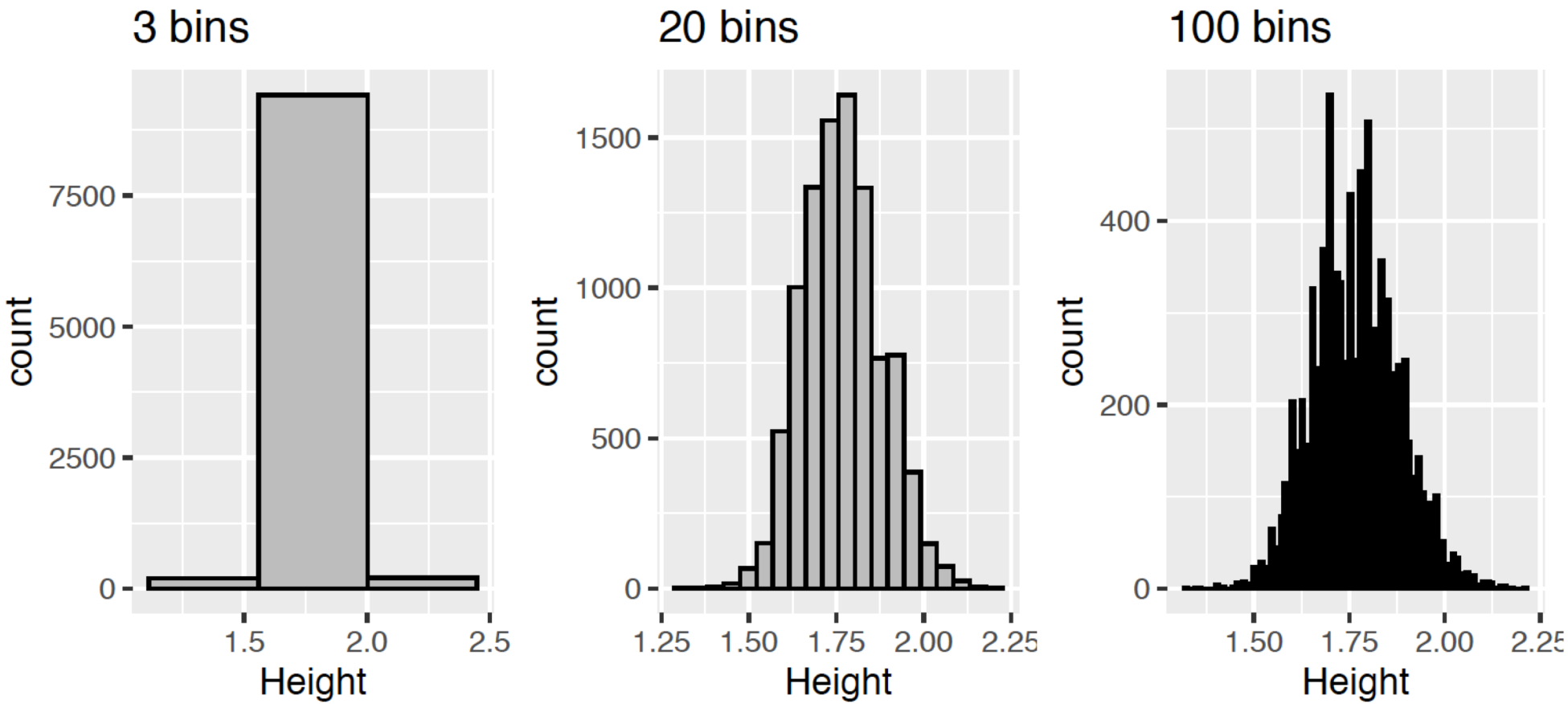
Symmetric, unimodal | Skewed right | Skewed left | Bimodal | Multimodal | Symmetric

Visualizations and Vocabulary list

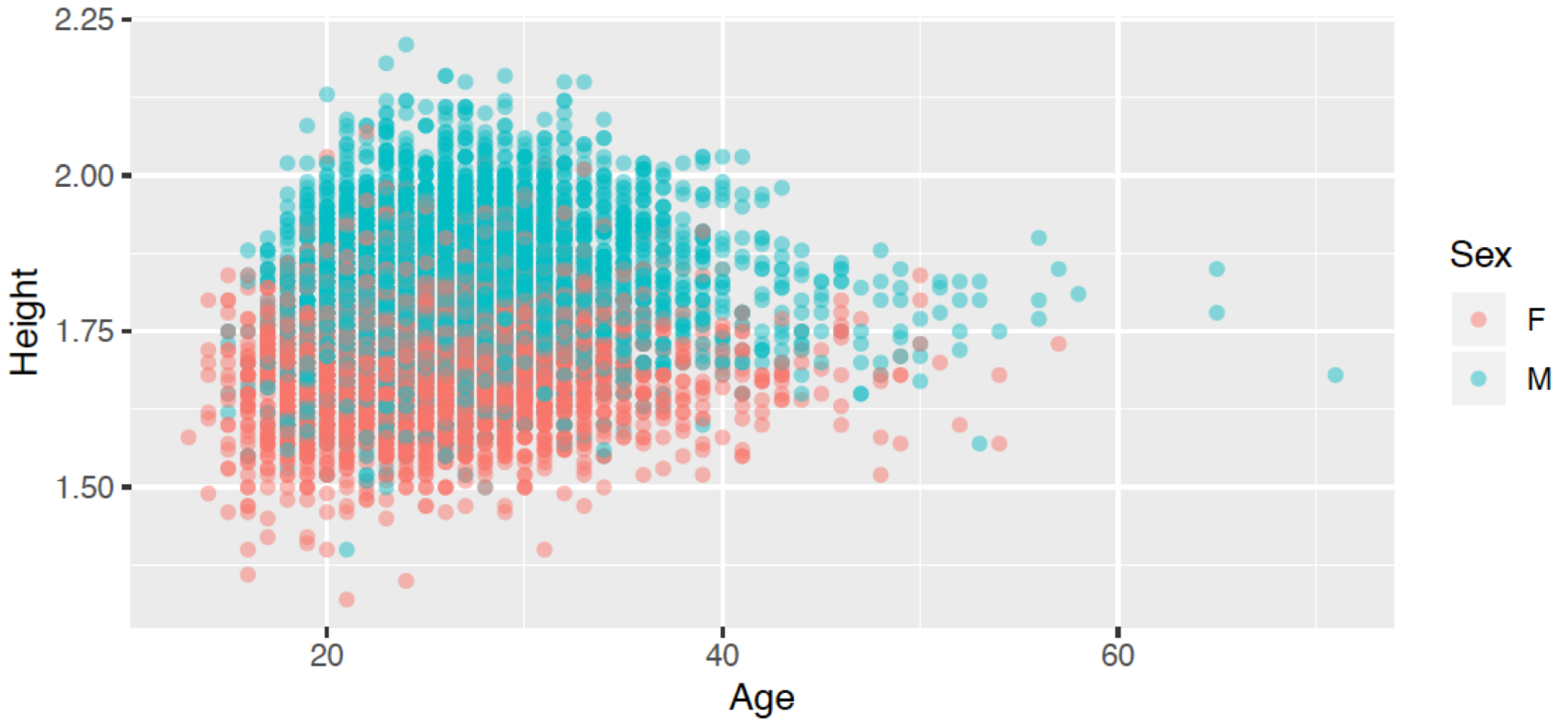
Questions specific to homework question 1 (for example):

- What type of distribution does height have?
- Is there an association between height and age? Height and sex?
- What types of figures (that we've learned so far) would be appropriate for this question? Why or why not?

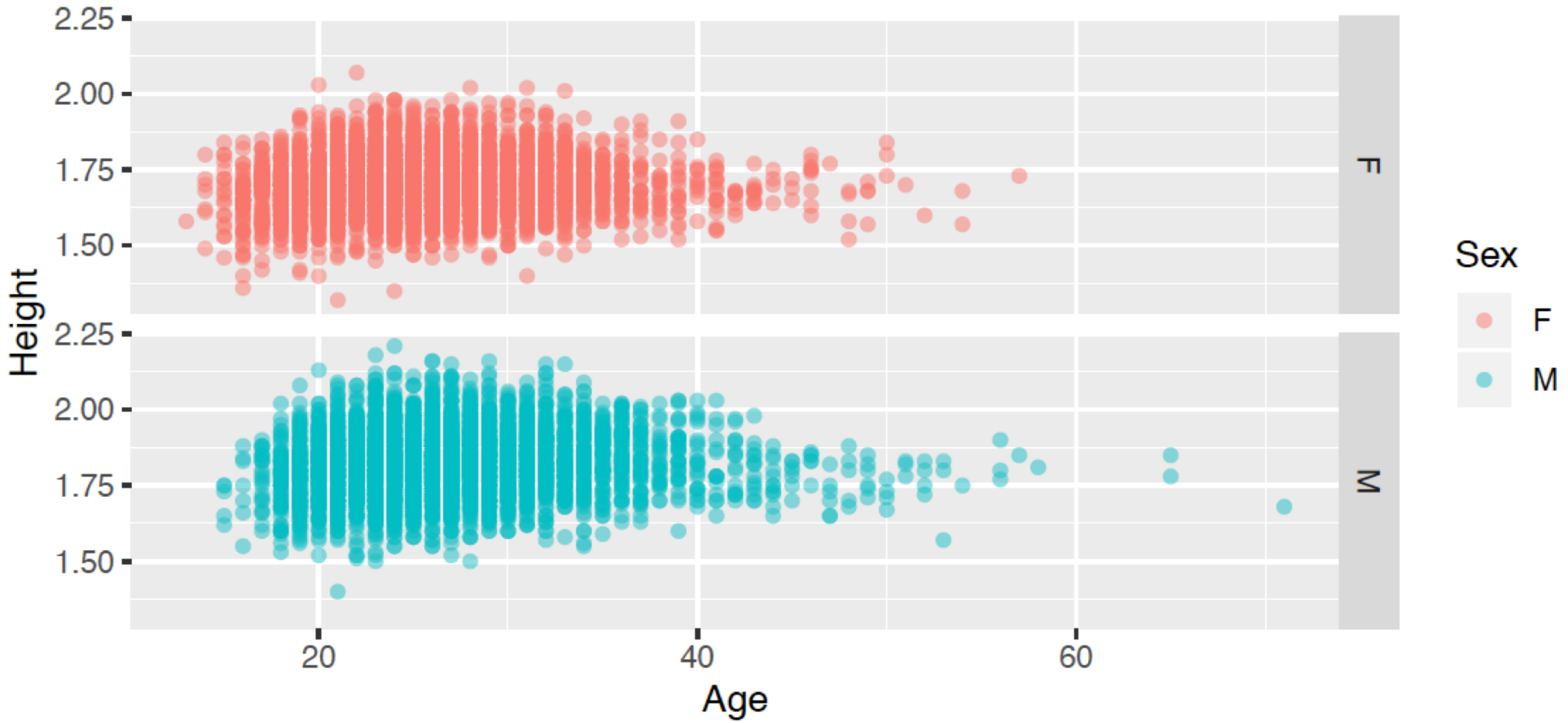
Homework question #1: Olympics 2012 dataset



Homework question #1: Olympics 2012 dataset



Homework question #1: Olympics 2012 dataset

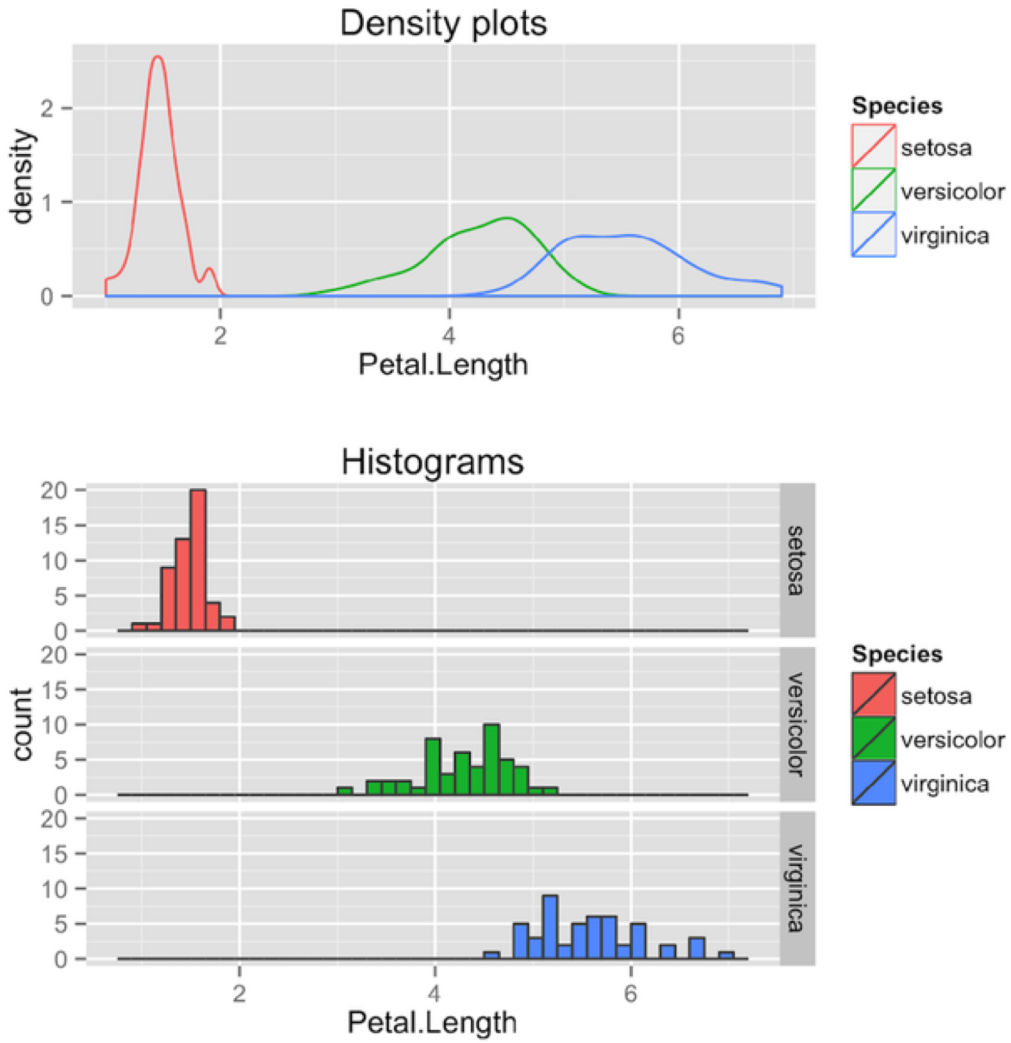


Group Discussion

- **Ice breaker!**
- What do you notice about the number of bins a histogram has, its shape and precision?
- In Question 1e, you could have presented both sexes in the same plot or presented them on separate plots? What are some considerations for which presentation you may want to choose (e.g. what are the pros and cons of each one)?
- If presenting two plots side by side, what are some things to consider to ensure they are comparable and reader-friendly?

Writing example

Iris dataset from R



Writing example

- Finding a way to lead a reader through a visual
- Describe what the graphs are telling us (x-axis, y-axis labels should be clear, etc)
- Come up with a “story” of main results
- Provide figures to support the “story”

Writing example

A possible writing template:

- Give some context to the variables you are graphing based on what you know about the dataset (units and types of variables involved should be clear).

Either:

- Give the most striking features of the graphs (contrast or similarity).
- Synthesize these features and make a conclusion based on these features.

Or:

- Make a statement or conclusion based on your impression.
- Explain each of the features of the graphs (contrast or similarity) that support your statement or conclusion.

Writing example #1

- The *petal length of* Iris setosa distributes differently from Iris versicolor and Iris virginica. The density plot/histogram of petal length of Iris setosa has a sharp peak while the other two have a flatter distribution.

Writing example #2

- We looked at the petal length of *Iris*. Specifically, Iris versicolor and Iris virginica, despite having different centres, have similar spread in terms of their petal length. Interestingly, the shape of distribution also differs between the species. We conclude that the petal length of Iris setosa, Iris versicolor and Iris virginica are different in terms of their centre, spread and shape.

Writing Example #3

- The *petal length in c.m.* of *50 samples from each of three species of Iris* (*Iris setosa, Iris virginica and Iris versicolor*) was investigated/examine/summarized. The graph suggested that *distribution of petal length is species dependent*. In particular, petal length of Iris setosa is shown to be less variable than Iris versicolor and Iris virginica. However, despite Iris versicolor having on average longer petals than Iris virginica, the range of petal length is similar for these two species. Further, the shape of distribution also differs according to species, with Iris setosa more or less symmetric about its centre, and versicolor and virginica skewed to the left and right, respectively.

Short writing exercise

- Write a short paragraph to describe coherently the graphs you produced and structure these graphs to tell a story. Use at least 3 graphs from question 2 to support the story.
- Submit on Quercus

STA130 Writing Activity Rubric

Student's Name: _____ Grade (/4): _____

	4 (Excellent)	3 (Good)	2 (Adequate)	1 (Poor)
Content Clarity	The context and connection to the problem are clear and all variables/concepts were mentioned. Details are appropriate for the intended audience.	Some context was provided and all variables/concepts were mentioned. Some aspects were not clear or not appropriate for the audience.	Very little context was provided and only some variables/ concepts were mentioned. The content was generally not appropriate for the intended audience.	No context. No mention of any variables/ concepts covering relevant to course materials. Entirely inappropriate for the intended audience.
Statistical Accuracy	The choice of methods is logical and appropriate to the research question(s); analyses are complete; correct description of statistical approach.	The choice of methods is appropriate and logical to the research question(s); some non-essential analyses are missing or minor errors in analytical approach or description of approach.	The choice of methods is somewhat appropriate, may not be entirely logical or best choice for answering the research question(s); some analyses are missing/incorrect or incorrectly described.	The choice of methods is inappropriate for the research question(s); essential analyses are missing. Description of method is entirely inaccurate.
Structure	Well organized, follows a logical structure.	The organization follows some logical structure.	Some structure but difficult to follow.	There is no structure, very difficult to follow.
Conclusion	There is a clear central idea and the conclusion is correct and appropriate for the audience and research question.	A conclusion is present. The conclusion might be incorrect or somewhat inappropriate for the audience or research question.	The conclusion is weak not well supported, inaccurate, or inappropriate for the audience or research question.	The conclusion is missing or is completely incorrect, not relevant to the research question, or entirely inappropriate for the audience.
Transitions	Effective and appropriate use of words and phrases to enhance the flow and signal transitions.	Good use of words and phrases to control the flow and signal transitions; 1 or 2 inappropriate uses of transitions.	Some use of transitional words and phrases, sometimes used inappropriately; e.g. overuse of words, wrong choice of word, etc.)	Lack of transitions and a poor flow between sentences/ideas.
Vocabulary	Accurate use of statistical terms and phrases, appropriate use of terms given the intended audience.	Good use of statistical terms and phrases, mostly appropriate language for the intended audience.	Demonstrated effort to incorporate statistical terms and phrases, but some were used inaccurately or not appropriate for the intended audience.	Completely inaccurate use of statistical terms and phrases. Vocabulary inappropriate for the intended audience.
Writing Mechanics	Answer written as a coherent paragraph. Sentences are clear and complete. No (or very minor) spelling, punctuation or grammatical errors.	Answer is mostly coherent and written in paragraph form with complete sentences. Some minor spelling, punctuation or grammar mistakes.	Frequent spelling, punctuation and/or grammatical errors; however, still mostly understandable.	Answer written in bullet point form and/or sentence fragments. Answer is not easy to understand.

STA130 Writing Activity Rubric

Student's Name: _____ Grade (/4): _____

	4 (Excellent)	3 (Good)	2 (Adequate)	1 (Poor)
Content Clarity	The context and connection to the problem are clear and all variables/concepts were mentioned. Details are appropriate for the intended audience.	Some context was provided and all variables/concepts were mentioned. Some aspects were not clear or not appropriate for the audience.	Very little context was provided and only some variables/ concepts were mentioned. The content was generally not appropriate for the intended audience.	No context. No mention of any variables/ concepts covering relevant to course materials. Entirely inappropriate for the intended audience.
Statistical Accuracy	The choice of methods is logical and appropriate to the research question(s); analyses are complete; correct description of statistical approach.	The choice of methods is appropriate and logical to the research question(s); some non-essential analyses are missing or minor errors in analytical approach or description of approach.	The choice of methods is somewhat appropriate, may not be entirely logical or best choice for answering the research question(s); some analyses are missing/incorrect or incorrectly described.	The choice of methods is inappropriate for the research question(s); essential analyses are missing. Description of method is entirely inaccurate.
Structure	Well organized, follows a logical structure.	The organization follows some logical structure.	Some structure but difficult to follow.	There is no structure, very difficult to follow.
Conclusion	There is a clear central idea and the conclusion is correct and appropriate for the audience and research question.	A conclusion is present. The conclusion might be incorrect or somewhat inappropriate for the audience or research question.	The conclusion is weak not well supported, inaccurate, or inappropriate for the audience or research question.	The conclusion is missing or is completely incorrect, not relevant to the research question, or entirely inappropriate for the audience.
Transitions	Effective and appropriate use of words and phrases to enhance the flow and signal transitions.	Good use of words and phrases to control the flow and signal transitions; 1 or 2 inappropriate uses of transitions.	Some use of transitional words and phrases, sometimes used inappropriately; e.g. overuse of words, wrong choice of word, etc.)	Lack of transitions and a poor flow between sentences/ideas.
Vocabulary	Accurate use of statistical terms and phrases, appropriate use of terms given the intended audience.	Good use of statistical terms and phrases, mostly appropriate language for the intended audience.	Demonstrated effort to incorporate statistical terms and phrases, but some were used inaccurately or not appropriate for the intended audience.	Completely inaccurate use of statistical terms and phrases. Vocabulary inappropriate for the intended audience.
Writing Mechanics	Answer written as a coherent paragraph. Sentences are clear and complete. No (or very minor) spelling, punctuation or grammatical errors.	Answer is mostly coherent and written in paragraph form with complete sentences. Some minor spelling, punctuation or grammar mistakes.	Frequent spelling, punctuation and/or grammatical errors; however, still mostly understandable.	Answer written in bullet point form and/or sentence fragments. Answer is not easy to understand.