hw01

March 3, 2022

Name: Allan Gongora

Section: 0131

1 Homework 1: Causality and Expressions

Please complete this notebook by filling in the cells provided. Before you begin, execute the following cell to load the provided tests.

[1]:
```
pip install gofer-grader
```

```
Requirement already satisfied: gofer-grader in /opt/conda/lib/python3.7/site-
packages (1.1.0)
Requirement already satisfied: jinja2 in /opt/conda/lib/python3.7/site-packages
(from gofer-grader) (3.0.3)
Requirement already satisfied: tornado in /opt/conda/lib/python3.7/site-packages
(from gofer-grader) (6.1)
Requirement already satisfied: pygments in /opt/conda/lib/python3.7/site-
packages (from gofer-grader) (2.11.2)
Requirement already satisfied: MarkupSafe>=2.0 in /opt/conda/lib/python3.7/site-
packages (from jinja2->gofer-grader) (2.0.1)
Note: you may need to restart the kernel to use updated packages.
```

[2]:
```
# These lines load the tests.

from gofer.ok import check
```

Recommended Reading: - What is Data Science - Causality and Experiments - Programming in Python

1) For all problems that you must write explanations and sentences for, you **must** provide your answer in the designated space. This can include:
   A) Sentence reponses to questions that ask for an explanation
   B) Numeric responses to multiple choice questions
   C) Programming code
2) Moreover, throughout this homework and all future ones, please be sure to not re-assign variables throughout the notebook! For example, if you use `max_temperature` in your answer to one question, do not reassign it later on. Otherwise, you will fail tests that you thought you were passing previously!

Once you're finished, select "Save and Checkpoint" in the File menu. Your name and course section number should be in the first and last cell of the assignment. Be sure you have run all cells with code and that the output from that is showing. Then click "Print Preview" in the File menu. Print a copy from there in pdf format. (This means you right click and choose print and choose "save as pdf" from your printer options.) You will need to submit the pdf in Canvas by the deadline.

The gopher grader output and/or output from your coding are essential to helping your instructor grade your work correctly and in a timely manner.

Files submitted that are missing the required output will lose some to all points so double check your pdf before submitting.

1.1 1. Scary Arithmetic

An ad for ADT Security Systems says,

> "When you go on vacation, burglars go to work [...] According to FBI statistics, over 25% of home burglaries occur between Memorial Day and Labor Day."

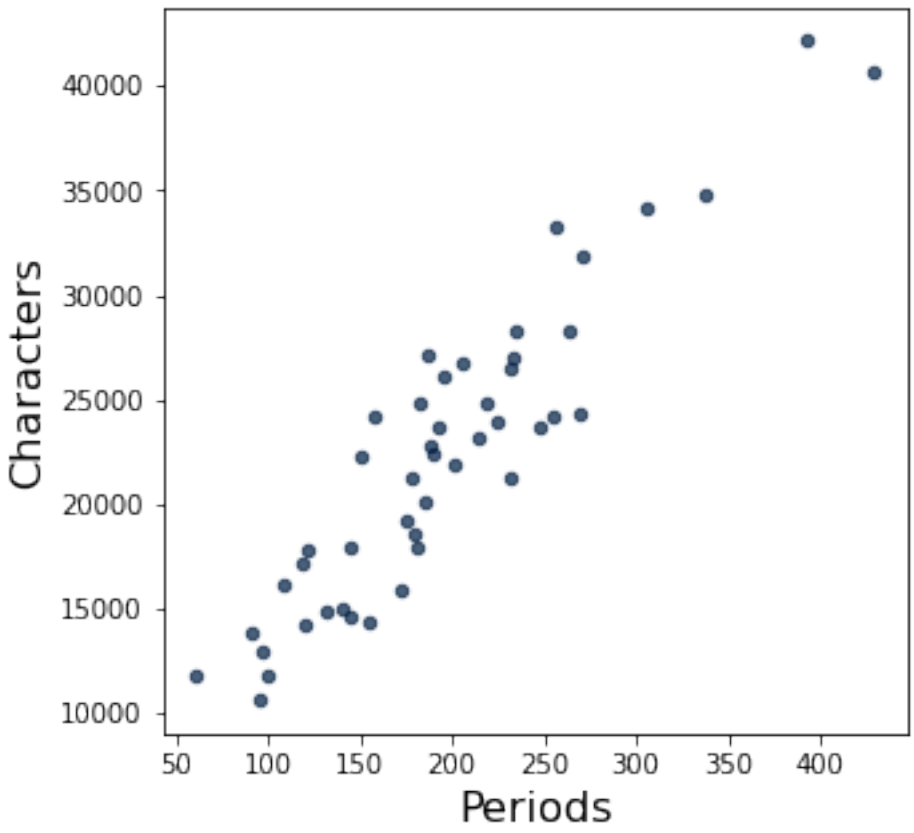
Question 1.1 Does the data in the ad support the claim that burglars are more likely to go to work during the time between Memorial Day and Labor Day? Please explain your answer.

No it doesn't because the majority (75%) of buglaries occur during the rest of the year.

1.2 2. Characters in Little Women

In lecture, we counted the number of times that the literary characters were named in each chapter of the classic book, Little Women. In computer science, the word "character" also refers to a letter, digit, space, or punctuation mark; any single element of a text. The following code generates a scatter plot in which each dot corresponds to a chapter of Little Women. The horizontal position of a dot measures the number of periods in the chapter. The vertical position measures the total number of characters.

[3]:
```
# This cell contains code that hasn't yet been covered in the course,
# but you should be able to interpret the scatter plot it generates.

from datascience import *
from urllib.request import urlopen
import numpy as np
%matplotlib inline

little_women_url = 'https://www.inferentialthinking.com/data/little_women.txt'
chapters = urlopen(little_women_url).read().decode().split('CHAPTER ')[1:]
text = Table().with_column('Chapters', chapters)
Table().with_columns(
    'Periods',    np.char.count(chapters, '.'),
    'Characters', text.apply(len, 0)
    ).scatter(0)
```



Question 2.1 Around how many periods are there in the chapter with the most characters? Assign either 1, 2, 3, 4, or 5 to the name `characters_q1` below.

1. 250
2. 390
3. 440
4. 32,000
5. 40,000

[4]:
```
characters_q1 = 2
```

Question 2.2 Which of the following chapters has the most characters 1. The chapter with about 60 periods 2. The chapter with about 350 periods 3. The chapter with about 440 periods per period?

Part a) Use python mathematics coding to calculate the characters per period. Parb b) Assign either 1, 2, or 3 to the name `characters_q2` below.

[5]:
```
### Part a
```

[6]:
```
char_per_period = Table().with_columns(
    'Periods',    np.char.count(chapters, '.'),
```

```
    'Characters', text.apply(len, 0)
    )
char_per_period = char_per_period.with_columns(
    "char_per_period", char_per_period["Characters"] /␣
 ↪char_per_period["Periods"]
)
```

[7]:
```
char_per_period.where(char_per_period["Periods"]).sort("Periods")
```

[7]:
```
Periods | Characters | char_per_period
61      | 11824      | 193.836
91      | 13873      | 152.451
95      | 10600      | 111.579
96      | 12902      | 134.396
100     | 11757      | 117.57
109     | 16135      | 148.028
118     | 17182      | 145.61
120     | 14221      | 118.508
121     | 17806      | 147.157
131     | 14851      | 113.366
… (37 rows omitted)
```

[ ]:
```

```

[ ]:
```

```

[8]:
```
chapters_per_period_60 = char_per_period.where(char_per_period["Periods"],␣
 ↪61)["char_per_period"][0]
print('The chapter with about 60 periods had', chapters_per_period_60,␣
 ↪'chapters per period')
```

```
The chapter with about 60 periods had 193.8360655737705 chapters per period
```

[9]:
```
chapters_per_period_350 = char_per_period.where(char_per_period["Periods"],␣
 ↪337)["char_per_period"][0]
print('The chapter with about 350 periods had', chapters_per_period_350,␣
 ↪'chapters per period')
```

```
The chapter with about 350 periods had 103.13353115727003 chapters per period
```

[10]:
```
chapters_per_period_440 = char_per_period.where(char_per_period["Periods"],␣
 ↪429)["char_per_period"][0]
print('The chapter with about 60 periods had', chapters_per_period_440,␣
 ↪'chapters per period')
```

```
The chapter with about 60 periods had 94.68764568764568 chapters per period
```

[11]:
```
### Part b
```

Assign either 1, 2, or 3 to the name `characters_q2` below, to reflect which chapter had the most characters per period. 1. The chapter with about 60 periods 2. The chapter with about 350 periods 3. The chapter with about 440 periods

[12]:
```
char_per_period.where(char_per_period["Periods"], 61)["char_per_period"][0],␣
↪char_per_period.where(char_per_period["Periods"],␣
↪337)["char_per_period"][0], char_per_period.
↪where(char_per_period["Periods"], 429)["char_per_period"][0]
```

[12]:
```
(193.8360655737705, 103.13353115727003, 94.68764568764568)
```

[13]:
```
characters_q2 = 1
```

Check that your answers are in the correct format. This test *does not* check that you answered correctly; only that you assigned a number successfully in each multiple-choice answer cell.

[14]:
```
check('tests/q2.py')
```

[14]:
```
<gofer.ok.OKTestsResult at 0x7fc00a3d1f50>
```

To discover more interesting facts from this plot, read Section 1.3.2 of the textbook.

1.3 3. Names and Assignment Statements

Question 3.1 When you run the following cell, Python produces a cryptic error message.

[15]:
```
4 = 2 + 2
```

```
  File "/tmp/ipykernel_208/2912417615.py", line 1
    4 = 2 + 2
          ^
SyntaxError: can't assign to literal
```

Choose the best explanation of what's wrong with the code, and then assign 1, 2, 3, or 4 to `names_q1` below to indicate your answer.

1. Python is smart and already knows `4 = 2 + 2`.
2. `4` is already a defined number, and it doesn't make sense to make a number be a name for something else. In Python, "`x = 2 + 2`" means "assign `x` as the name for the value of `2 + 2`."
3. It should be `2 + 2 = 4`.
4. I don't get an error message. This is a trick question.

[16]:
```
names_q1 = 2
```

Question 3.2 When you run the following cell, Python will produce another cryptic error message.

[17]:
```
two = 3
six = two plus two
```

```
  File "/tmp/ipykernel_208/3853341378.py", line 2
    six = two plus two
                  ^
SyntaxError: invalid syntax
```

Choose the best explanation of what's wrong with the code and assign 1, 2, 3, or 4 to `names_q2` below to indicate your answer.

1. The `plus` operation only applies to numbers, not the word "two".
2. The name "two" cannot be assigned to the number 3.
3. Two plus two is four, not six.
4. Python cannot interpret the name `two` followed directly by a name that has not been defined.

[18]:
```
names_q2 = 4
```

Question 3.3 When you run the following cell, Python will, yet again, produce another cryptic error message.

[19]:
```
x = print(5)
y = x + 2
```

```
5
```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
/tmp/ipykernel_208/2727161758.py in <module>
      1 x = print(5)
----> 2 y = x + 2

TypeError: unsupported operand type(s) for +: 'NoneType' and 'int'
```

Choose the best explanation of what's wrong with the code and assign 1, 2, 3, or 4 to `names_q3` below to indicate your answer.

1. You cannot add the letter x with 2.
2. The `print` operation is meant for displaying values to the programmer, not for assigning values!
3. Python doesn't want `y` to be assigned.
4. What error message?

[20]:
```
names_q3 = 2
```

Check that your answers are in the correct format. This test *does not* check that you answered correctly; only that you assigned a number successfully in each multiple-choice answer cell.

[21]:
```
check('tests/q3.py')
```

[21]:
```
<gofer.ok.OKTestsResult at 0x7fc0501d85d0>
```

1.4 4. Job Opportunities & Education in Rural India

A study at UCLA investigated factors that might result in greater attention to the health and education of girls in rural India. One such factor is information about job opportunities for women. The idea is that if people know that educated women can get good jobs, they might take more care of the health and education of girls in their families, as an investment in the girls' future potential as earners. Without the knowledge of job opportunities, the author hypothesizes that families invest less in their women, and instead, invest in their men.

The study focused on 160 villages outside the capital of India, all with little access to information about call centers and similar organizations that offer job opportunities to women. In 80 of the villages chosen at random, recruiters visited the village, described the opportunities, recruited women who had some English language proficiency and experience with computers, and provided ongoing support free of charge for three years. In the other 80 villages, no recruiters visited and no other intervention was made.

At the end of the study period, the researchers recorded data about the school attendance and health of the children in the villages.

Question 4.1 Which statement best describes the *treatment* and *control* groups for this study? Assign either 1, 2, or 3 to the name `jobs_q1` below.

1. The treatment group was the 80 villages visited by recruiters, and the control group was the other 80 villages with no intervention.
2. The treatment group was the 160 villages selected, and the control group was the rest of the villages outside the capital of India.
3. There is no clear notion of *treatment* and *control* group in this study.

[22]:
```
jobs_q1 = 1
```

Question 4.2 Was this an observational study or a randomized controlled experiment? Assign either 1, 2, or 3 to the name `jobs_q2` below.

1. This was an observational study.
2. This was a randomized controlled experiment.
3. This was a randomized observational study.

[23]:
```
jobs_q2 = 2
```

Question 4.3 The study reported, "Girls aged 5-15 in villages that received the recruiting services were 3 to 5 percentage points more likely to be in school and experienced an increase in Body Mass Index, reflecting greater nutrition and/or medical care. However, there was no net gain in height. For boys, there was no change in any of these measures." Why do you think the author points out the lack of change in the boys?

Hint: Remember the original hypothesis. The author believes that educating women in job opportunities will cause families to invest in their women more.

Because his hypothesis included "families invest less in their women, and instead, invest in their men" if this was true we should expect that there was a negative change in those measures for men

Check that your answers are in the correct format. This test *does not* check that you answered correctly; only that you assigned a number successfully in each multiple-choice answer cell. Be sure that you answered the free response question as that will be graded manually later.

[24]:
```
check('tests/q4.py')
```

[24]:
```
<gofer.ok.OKTestsResult at 0x7fc007ad8e50>
```

1.5 5. Differences Between Universities

Question 5.1 Suppose you'd like to *quantify* how *dissimilar* two universities are, using three quantitative characteristics. The US Department of Education data on UW and Cal describes the following three traits (among many others):

Trait	UW	Cal
Average annual cost to attend ($)	13,566	13,707
Graduation rate (percentage)	83	91
Socioeconomic Diversity (percentage)	25	31

You decide to define the dissimilarity between two universities as the maximum of the absolute values of the 3 differences in their respective trait values.

Using this method, compute the dissimilarity between UW and CAL. Name the result `dissimilarity`. Use a single expression (a single line of code) to compute the answer. Let Python perform all the arithmetic (like subtracting 91 from 83) rather than simplifying the expression yourself. The built-in `abs` function takes absolute values.

*** It is ok here to just input the values into the absolute value function and the other function you will use. You don't have to call the values from the table. But you MUST complete this using Python coding***

[25]:
```
dissimilarity = max([abs(13566 - 13707), abs(83 - 91), abs(25 - 31)])
dissimilarity
```

[25]:
```
141
```

Use the cell below to test for formatting (in this case, that dissimilarity is a number)

[26]:
```
check('tests/q5_1.py')
```

[26]:
```
<gofer.ok.OKTestsResult at 0x7fc050615410>
```

1.6 6. Nearsightedness Study

Myopia, or nearsightedness, results from a number of genetic and environmental factors. In 1999, Quinn et al studied the relation between myopia and ambient lighting at night (for example, from nightlights or room lights) during childhood.

Question 6.1 The data were gathered by the following procedure, reported in the study. "Between January and June 1998, parents of children aged 2-16 years [...] that were seen as outpatients in a university pediatric ophthalmology clinic completed a questionnaire on the child's light exposure both at present and before the age of 2 years." Was this study observational, or was it a controlled experiment? Explain.

Observational, they simply reported the results of the survey. They didn't change any variables.

Question 6.2 The study found that of the children who slept with a room light on before the age of 2, 55% were myopic. Of the children who slept with a night light on before the age of 2, 34% were myopic. Of the children who slept in the dark before the age of 2, 10% were myopic. The study concluded that, "The prevalence of myopia [...] during childhood was strongly associated with ambient light exposure during sleep at night in the first two years after birth."

Do the data support this statement? You may interpret "strongly" in any reasonable qualitative way.

Yes. The data presented seems to suggest that a decrease in ambient light when sleeping in the first 2 years after birth also decreases myopia.

Question 6.3 On May 13, 1999, CNN reported the results of this study under the headline, "Night light may lead to nearsightedness." Does the conclusion of the study claim that night light causes nearsightedness?

No. It states that it may, that is to say the current data suggests this but we require more data to verify the claim.

Question 6.4 The final paragraph of the CNN report said that "several eye specialists" had pointed out that the study should have accounted for heredity.

Myopia is passed down from parents to children. It's reasonable to suppose that myopic parents are more likely to leave lights on in their children's rooms than other parents. In what way do you think this might have affected the data?

I know nothing of nearsightedness so it's hard to say; however, just including a question about the parent's nearsightedness would account for this variable so i think it should have been included. It may show that hereditary myopia is correlated to parents leaving lights on at night which would lessen the claim that you're at an increased risk for myopia if you have lights on from 0-2 years old

1.7 7. Studying the Survivors

The Reverend Henry Whitehead was skeptical of John Snow's conclusion about the Broad Street pump. After the Broad Street cholera epidemic ended, Whitehead set about trying to prove Snow

wrong. (The history of the event is detailed here.)

He realized that Snow had focused his analysis almost entirely on those who had died. Whitehead, therefore, investigated the drinking habits of people in the Broad Street area who had not died in the outbreak.

Question 7.1 What is the main reason it was important to study this group?

1) If Whitehead had found that many people had drunk water from the Broad Street pump and not caught cholera, that would have been evidence against Snow's hypothesis.
2) Survivors could provide additional information about what else could have caused the cholera, potentially unearthing another cause.
3) Through considering the survivors, Whitehead could have identified a cure for cholera.

[27]:
```
# Assign survivor_answer to 1, 2, or 3
survivor_answer = 1
```

[28]:
```
check('tests/q7_1.py')
```

[28]:
```
<gofer.ok.OKTestsResult at 0x7fc007ab5ed0>
```

Note: Whitehead ended up finding further proof that the Broad Street pump played the central role in spreading the disease to the people who lived near it. Eventually, he became one of Snow's greatest defenders.

1.8 8. Submission

Once you're finished, select "Save and Checkpoint" in the File menu. Your name and course section number should be in the first and last cell of the assignment. Be sure you have run all cells with code and that the output from that is showing.

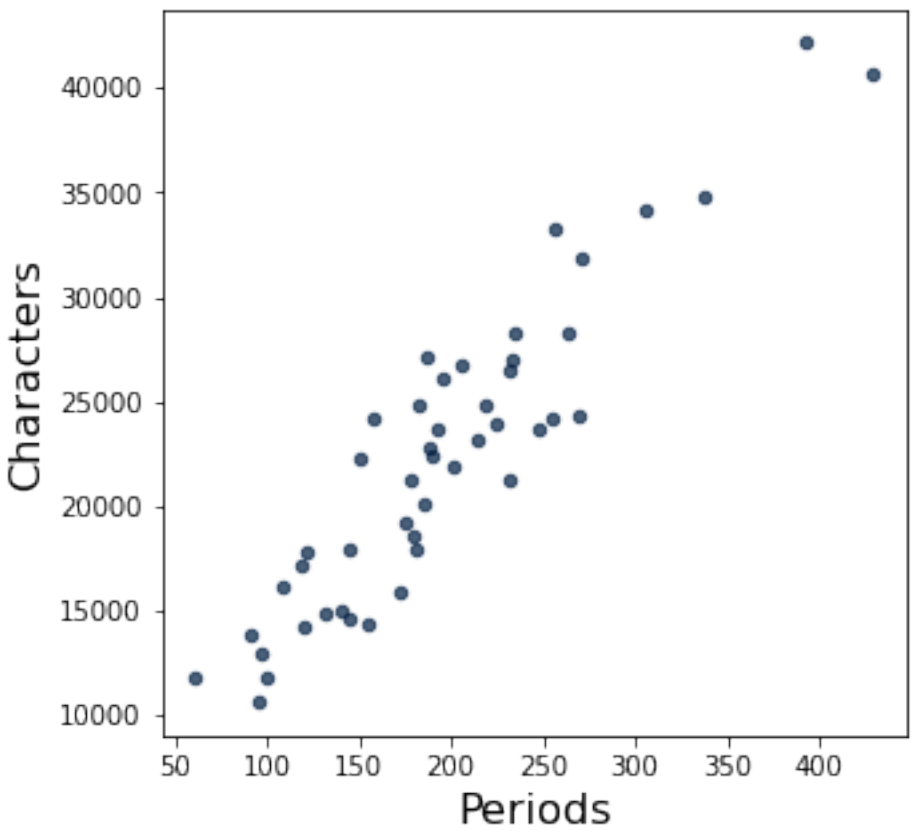
Double check that you have completed all of the free response questions as the auto-grader does NOT check that and YOU are responsible for knowing those questions are there and completing them as part of the grade for this homework. When ready, click "Print Preview" in the File menu. Print a copy from there in pdf format. (This means you right click and choose print and choose "save as pdf" from your printer options.) You will need to submit the pdf in Canvas by the deadline.

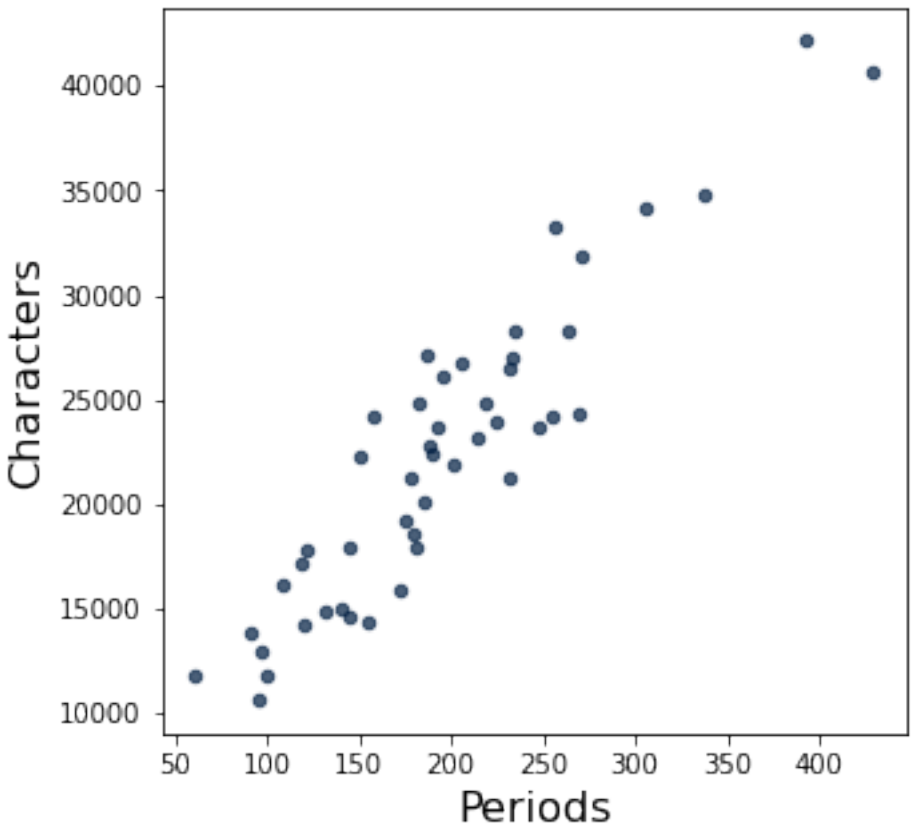
The gopher grader output and/or output from your coding are essential to helping your instructor grade your work correctly and in a timely manner.

Files submitted that are missing the required output will lose some to all points so double check your pdf before submitting.

[29]:
```
# For your convenience, you can run this cell to run all the tests at once!
import glob
from gofer.ok import grade_notebook
if not globals().get('__GOFER_GRADER__', False):
    display(grade_notebook('hw01.ipynb', sorted(glob.glob('tests/q*.py'))))
```

```
The chapter with about 60 periods had 193.8360655737705 chapters per period
The chapter with about 350 periods had 103.13353115727003 chapters per period
The chapter with about 60 periods had 94.68764568764568 chapters per period
5
['tests/q2.py', 'tests/q3.py', 'tests/q4.py', 'tests/q5_1.py', 'tests/q7_1.py']
Question 1:

<gofer.ok.OKTestsResult at 0x7fc007891290>

Question 2:

<gofer.ok.OKTestsResult at 0x7fc0077e0bd0>

Question 3:

<gofer.ok.OKTestsResult at 0x7fc0077e0a90>

Question 4:

<gofer.ok.OKTestsResult at 0x7fc0077e0f50>

Question 5:

<gofer.ok.OKTestsResult at 0x7fc0077e0c10>

Question 6:

<gofer.ok.OKTestsResult at 0x7fc00aa01890>

Question 7:

<gofer.ok.OKTestsResult at 0x7fc007a0d9d0>

1.0
```


Characters
Periods
40000
35000
30000
25000
20000
15000
10000
50
100
150
200
250
300
350
400



Name: Allan Gongora

Section: 0131