Statistics for Data Analytics (B8IT152)

Solution Sheet

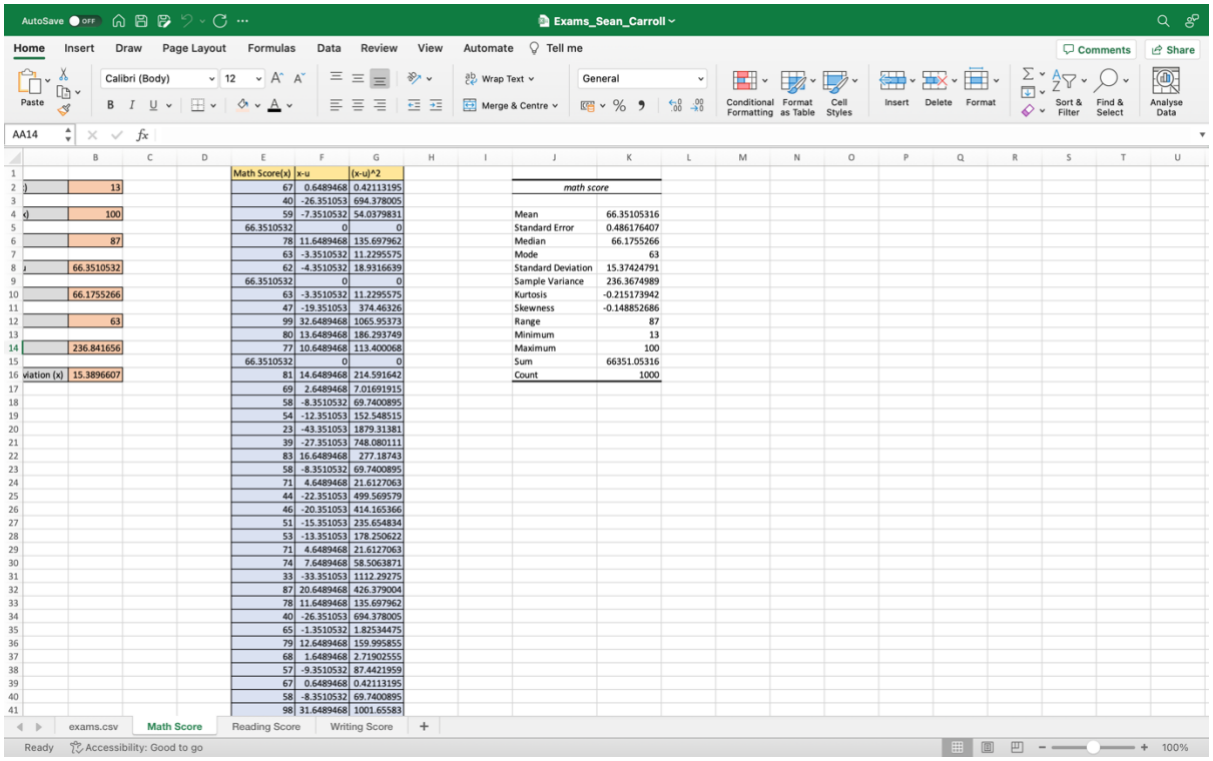
Name: Seán Carroll

Student Number: 20024157

Question 1: Using the provided dataset "exams"

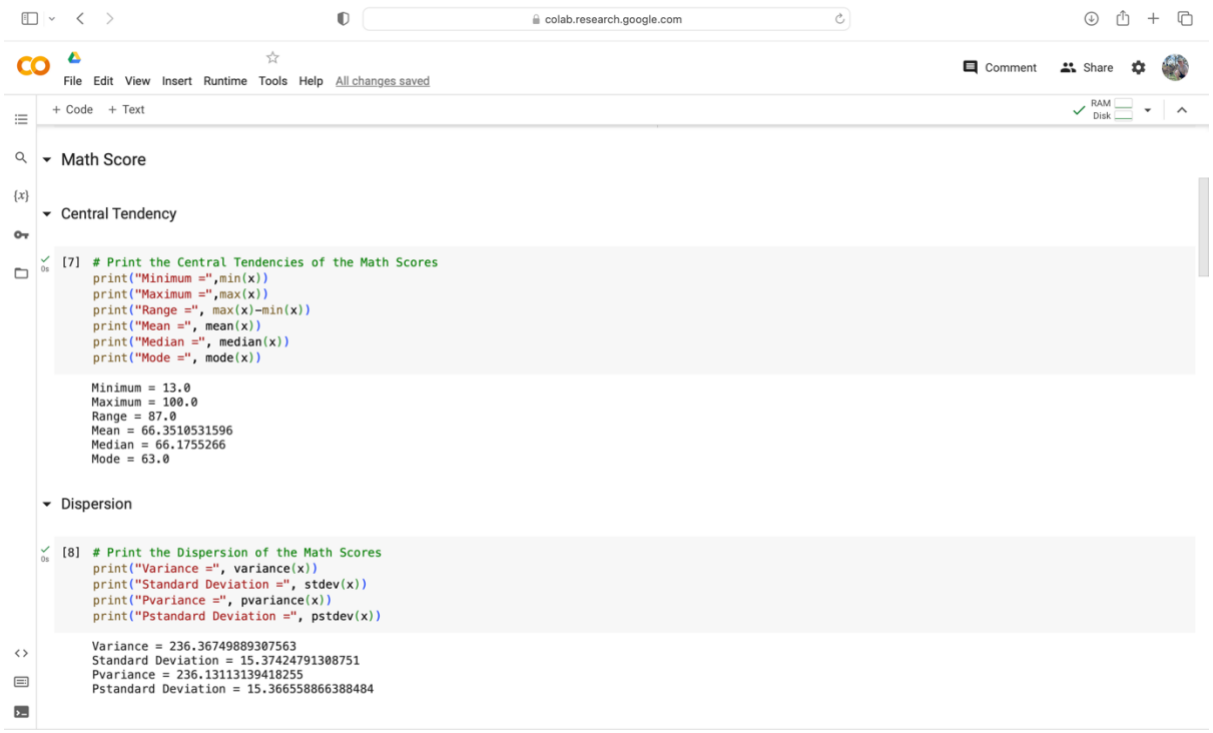
For Part(a) use python code and MS-Excel, for Part (b) use Python code

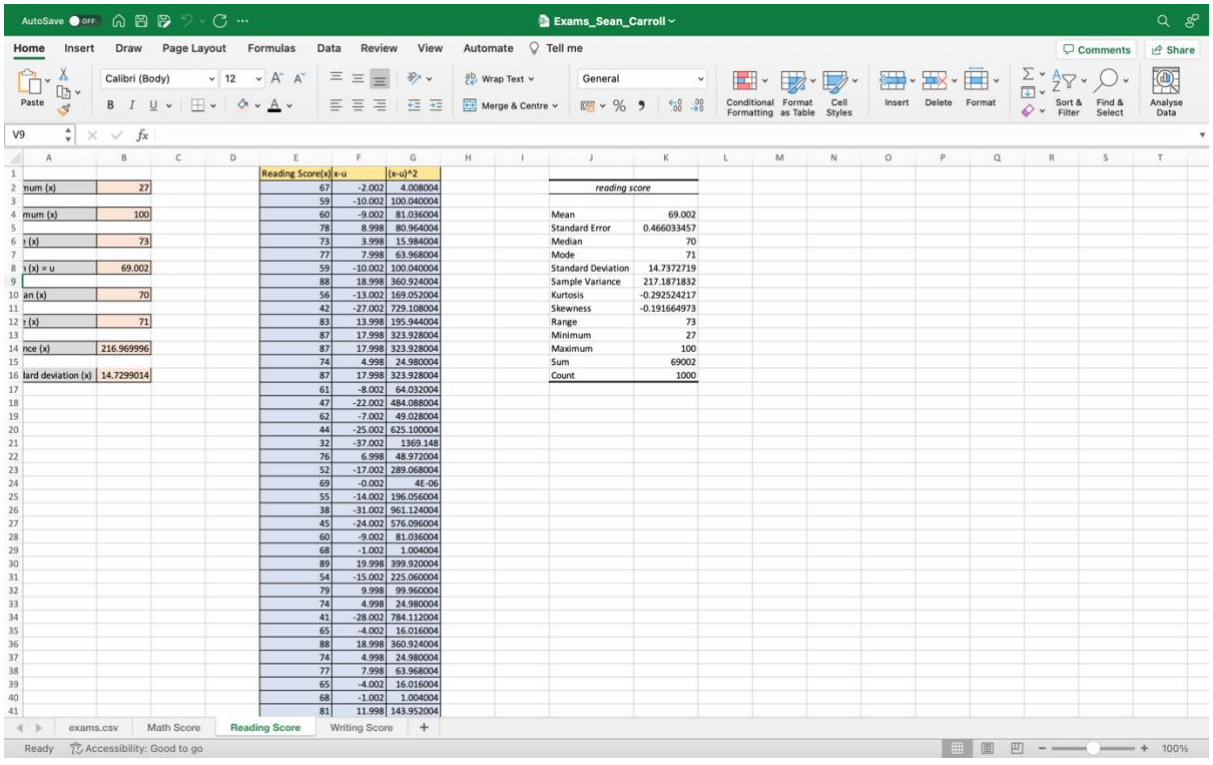
1. a) Compute central and dispersion measures for the attributes (math score, reading score).

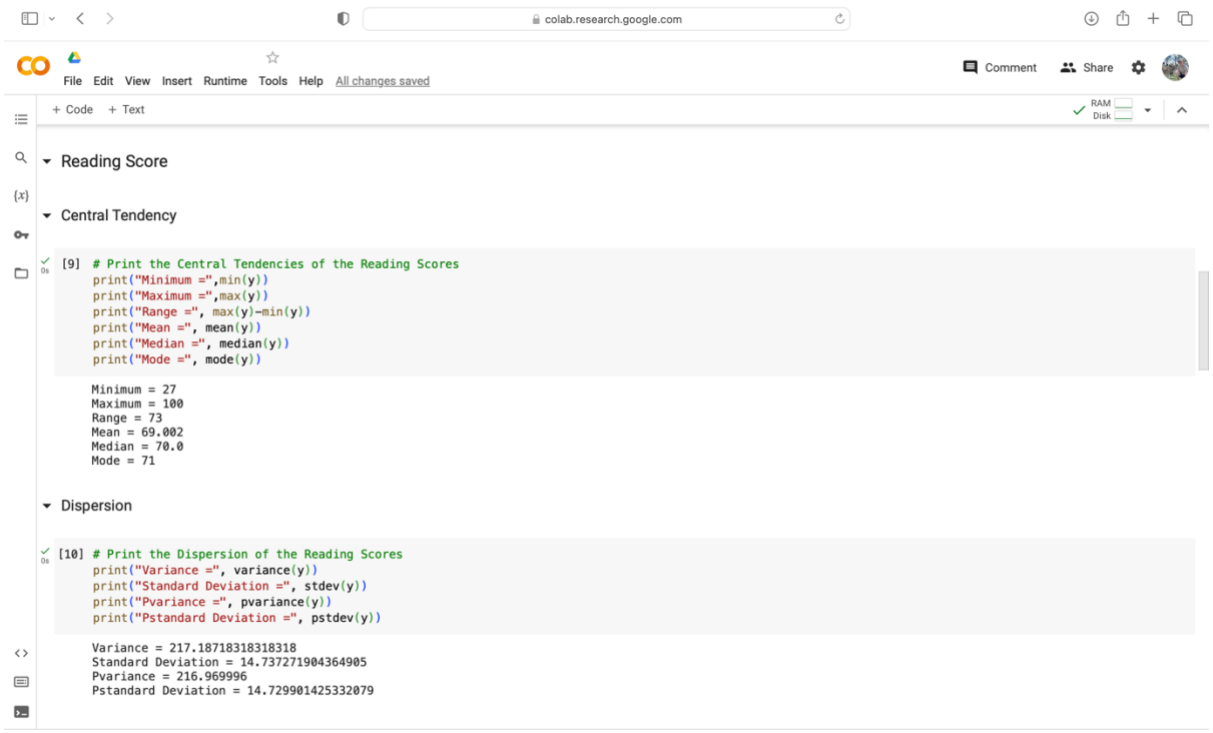


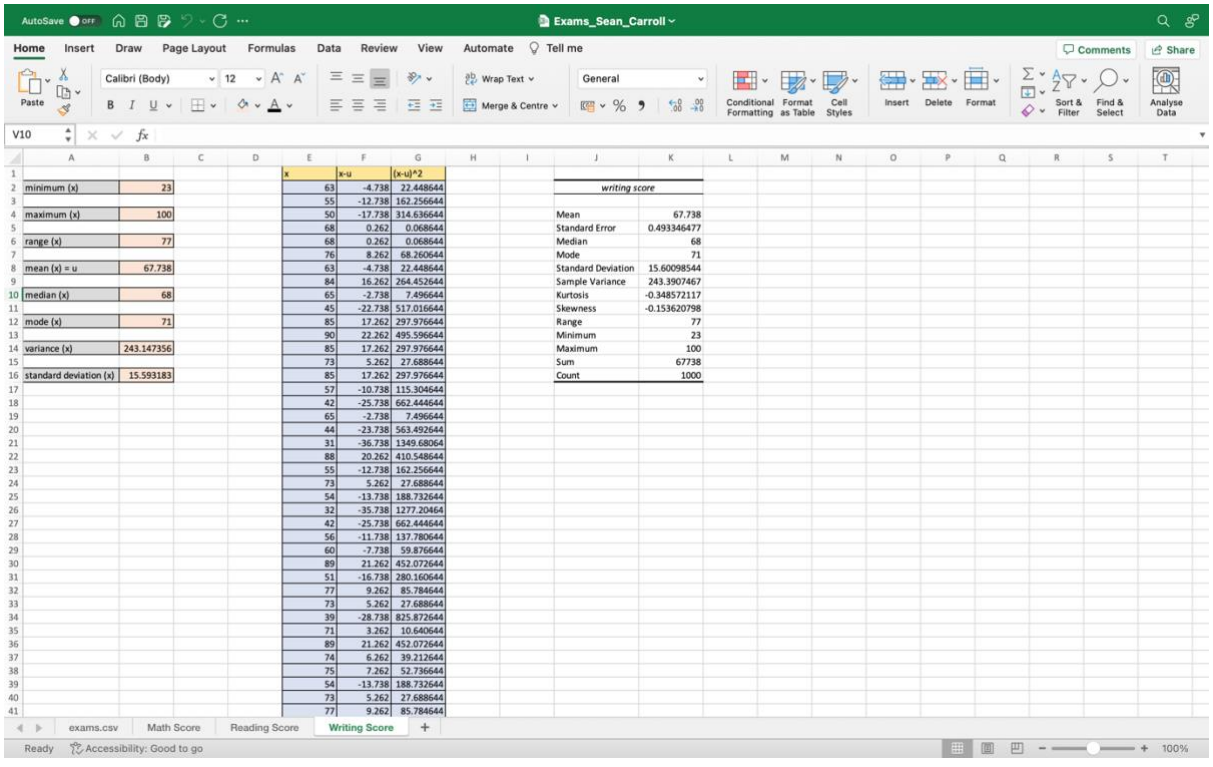
Math Score(x)	x-u	(x-u)^2
67	0.6489468	0.42113195
40	-26.351053	694.378005
59	-7.3510532	54.0379831
66.3510532	0	0
78	11.6489468	135.697962
63	-3.3510532	11.2295575
62	-4.3510532	18.9316639
66.3510532	0	0
63	-3.3510532	11.2295575
47	-19.351053	374.46326
99	32.6489468	1065.95373
80	13.6489468	186.293749
77	10.6489468	113.400068
66.3510532	0	0
81	14.6489468	214.591642
69	2.6489468	7.01691915
58	-8.3510532	69.7400895
54	-12.351053	152.548515
23	-43.351053	1879.31381
39	-27.351053	748.080111
83	16.6489468	277.18743
58	-8.3510532	69.7400895
71	4.6489468	21.6127063
44	-22.351053	499.569579
46	-20.351053	414.165366
51	-15.351053	235.654834
53	-13.351053	178.250622
71	4.6489468	21.6127063
74	7.6489468	58.5063871
33	-33.351053	1112.29275
87	20.6489468	426.379004
78	11.6489468	135.697962
40	-26.351053	694.378005
65	-1.3510532	1.82534475
79	12.6489468	159.995855
68	1.6489468	2.71902555
57	-9.3510532	87.4421959
67	0.6489468	0.42113195
58	-8.3510532	69.7400895
98	31.6489468	1001.65583

math score	
Mean	66.35105316
Standard Error	0.486176407
Median	66.1755266
Mode	63
Standard Deviation	15.37424791
Sample Variance	236.3674989
Kurtosis	-0.215173942
Skewness	-0.148852686
Range	87
Minimum	13
Maximum	100
Sum	66351.05316
Count	1000

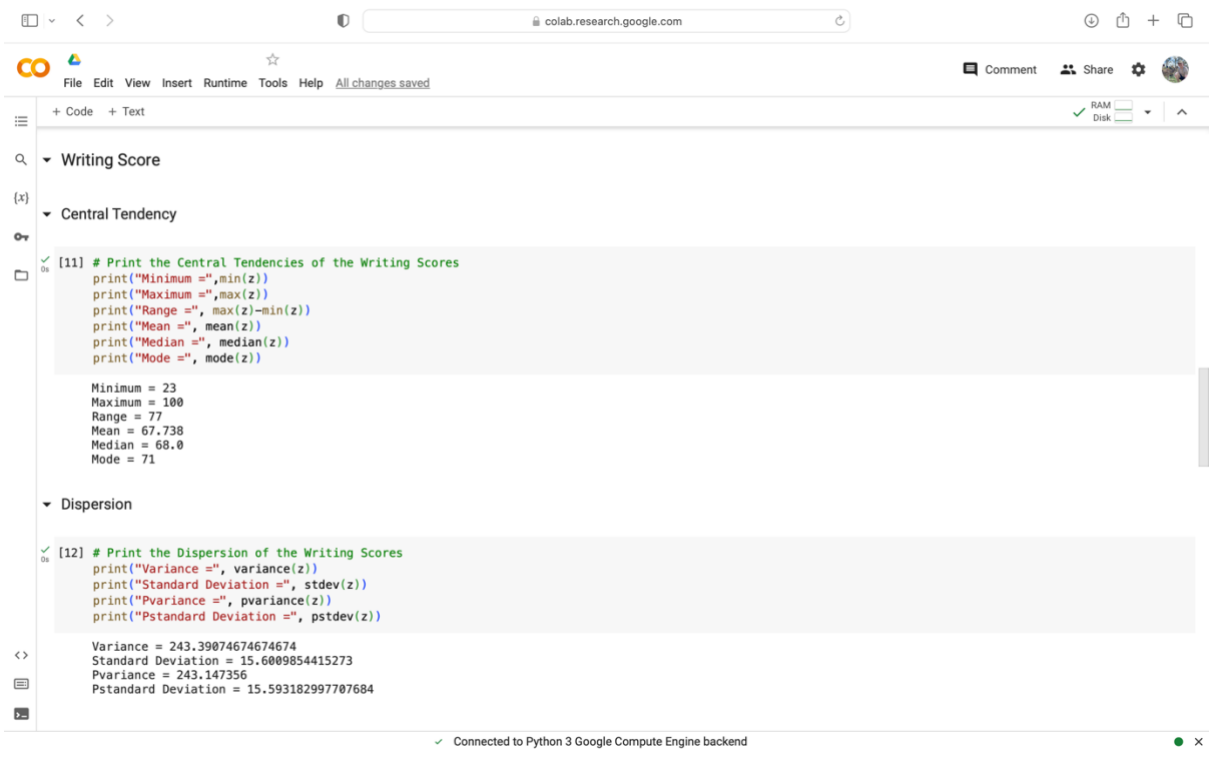

colab.research.google.com
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
RAM
Disk
Math Score
Central Tendency
[7] # Print the Central Tendencies of the Math Scores
print("Minimum =",min(x))
print("Maximum =",max(x))
print("Range =", max(x)-min(x))
print("Mean =", mean(x))
print("Median =", median(x))
print("Mode =", mode(x))
Minimum = 13.0
Maximum = 100.0
Range = 87.0
Mean = 66.3510531596
Median = 66.1755266
Mode = 63.0
Dispersion
[8] # Print the Dispersion of the Math Scores
print("Variance =", variance(x))
print("Standard Deviation =", stdev(x))
print("Pvariance =", pvariance(x))
print("Pstandard Deviation =", pstdev(x))
Variance = 236.36749889307563
Standard Deviation = 15.37424791308751
Pvariance = 236.13113139418255
Pstandard Deviation = 15.366558866388484
Connected to Python 3 Google Compute Engine backend


AutoSave
Exams_Sean_Carroll
Home Insert Draw Page Layout Formulas Data Review View Automate Tell me
Comments
Share
Paste
Calibri (Body)
12
Wrap Text
General
Merge & Centre
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
Sort & Filter
Find & Select
Analyse Data
V9
num (x) 27
mum (x) 100
(x) 73
(x) = u 69.002
an (x) 70
(x) 71
nce (x) 216.969996
lard deviation (x) 14.7299014
Reading Score(x) x-u (x-u)^2
67 -2.002 4.008004
59 -10.002 100.040004
60 -9.002 81.036004
78 8.998 80.964004
73 3.998 15.984004
77 7.998 63.968004
59 -10.002 100.040004
88 18.998 360.924004
56 -13.002 169.052004
42 -27.002 729.108004
83 13.998 195.944004
87 17.998 323.928004
87 17.998 323.928004
74 4.998 24.980004
87 17.998 323.928004
61 -8.002 64.032004
47 -22.002 484.088004
62 -7.002 49.028004
44 -25.002 625.100004
32 -37.002 1369.148
76 6.998 48.972004
52 -17.002 289.068004
69 -0.002 4E-06
55 -14.002 196.056004
38 -31.002 961.124004
45 -24.002 576.096004
60 -9.002 81.036004
68 -1.002 1.004004
89 19.998 399.920004
54 -15.002 225.060004
79 9.998 99.960004
74 4.998 24.980004
41 -28.002 784.112004
65 -4.002 16.016004
88 18.998 360.924004
74 4.998 24.980004
77 7.998 63.968004
65 -4.002 16.016004
68 -1.002 1.004004
81 11.998 143.952004
reading score
Mean 69.002
Standard Error 0.466033457
Median 70
Mode 71
Standard Deviation 14.7372719
Sample Variance 217.1871832
Kurtosis -0.292524217
Skewness -0.191664973
Range 73
Minimum 27
Maximum 100
Sum 69002
Count 1000
exams.csv
Math Score
Reading Score
Writing Score
Ready
Accessibility: Good to go
100%

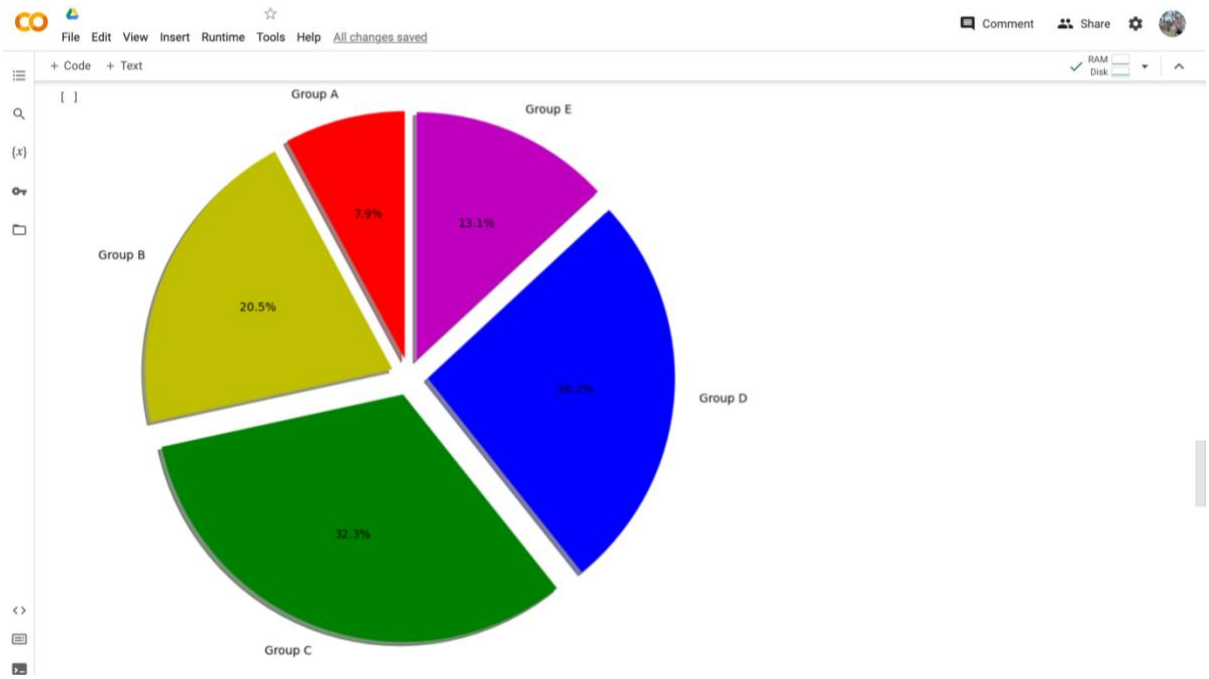




	x	x-u	(x-u)^2
1	63	-4.738	22.448644
2	55	-12.738	162.256644
3	50	-17.738	314.636644
4	68	0.262	0.068644
5	68	0.262	0.068644
6	76	8.262	68.260644
7	63	-4.738	22.448644
8	84	16.262	264.452644
9	65	-2.738	7.496644
10	45	-22.738	517.016644
11	85	17.262	297.976644
12	90	22.262	495.596644
13	85	17.262	297.976644
14	73	5.262	27.688644
15	85	17.262	297.976644
16	57	-10.738	115.304644
17	42	-25.738	662.444644
18	65	-2.738	7.496644
19	44	-23.738	563.492644
20	31	-36.738	1349.68064
21	88	20.262	410.548644
22	55	-12.738	162.256644
23	73	5.262	27.688644
24	54	-13.738	188.732644
25	32	-35.738	1277.20464
26	42	-25.738	662.444644
27	56	-11.738	137.780644
28	60	-7.738	59.876644
29	89	21.262	452.072644
30	51	-16.738	280.160644
31	77	9.262	85.784644
32	73	5.262	27.688644
33	39	-28.738	825.872644
34	71	3.262	10.640644
35	89	21.262	452.072644
36	74	6.262	39.212644
37	75	7.262	52.736644
38	54	-13.738	188.732644
39	73	5.262	27.688644
40	77	9.262	85.784644
41			


colab.research.google.com
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
RAM
Disk
Writing Score
Central Tendency
[11] # Print the Central Tendencies of the Writing Scores
print("Minimum =",min(z))
print("Maximum =",max(z))
print("Range =", max(z)-min(z))
print("Mean =", mean(z))
print("Median =", median(z))
print("Mode =", mode(z))
Minimum = 23
Maximum = 100
Range = 77
Mean = 67.738
Median = 68.0
Mode = 71
Dispersion
[12] # Print the Dispersion of the Writing Scores
print("Variance =", variance(z))
print("Standard Deviation =", stdev(z))
print("Pvariance =", pvariance(z))
print("Pstandard Deviation =", pstdev(z))
Variance = 243.39074674674674
Standard Deviation = 15.6009854415273
Pvariance = 243.147356
Pstandard Deviation = 15.593182997707684
Connected to Python 3 Google Compute Engine backend

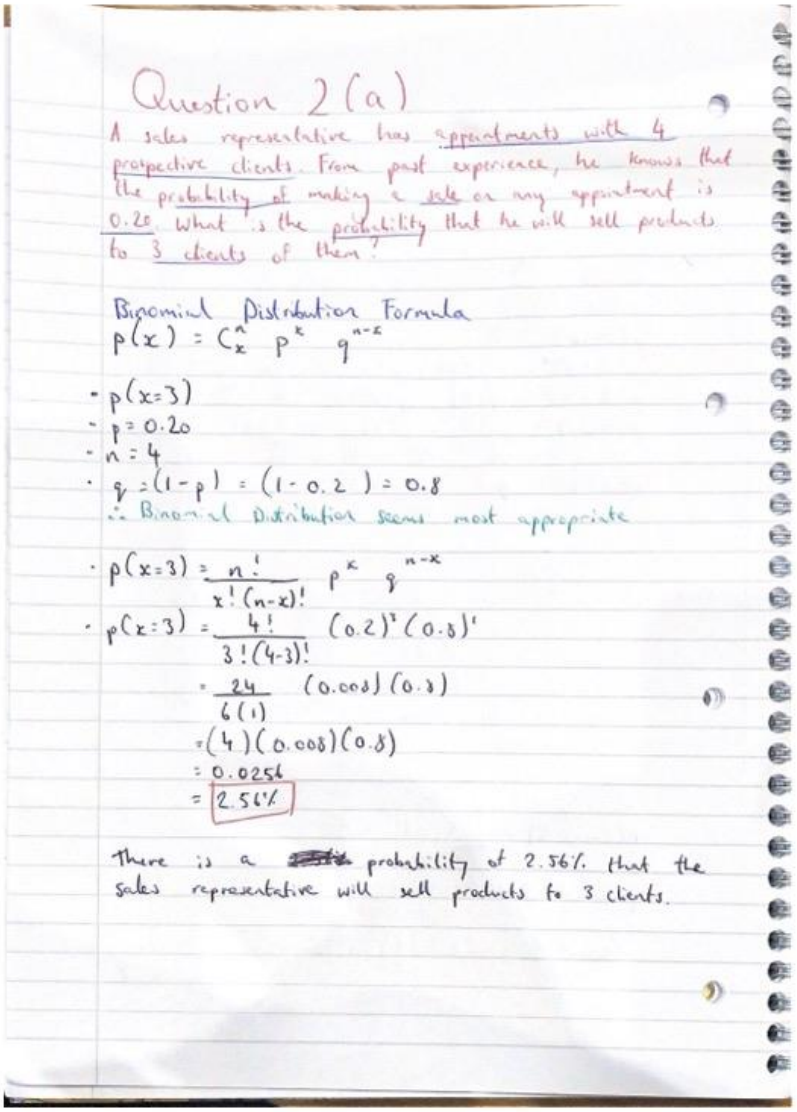
2. b) Describe the data attribute “race/ethnicity” using the appropriate graph type.



Question 2

Solve the following questions and explain which distribution seems most appropriate for each.

a) A sales representative has appointments with 4 prospective clients. From past experience, he knows that the probability of making a sale on any appointment is 0.20. What is the probability that he will sell products to 3 clients of them?



Question 2 (a)

A sales representative has appointments with 4 prospective clients. From past experience, he knows that the probability of making a sale on any appointment is 0.20. What is the probability that he will sell products to 3 clients of them?

Binomial Distribution Formula

$p(x) = C_x^n \, p^x \, q^{n-x}$

- $p(x=3)$
- $p = 0.20$
- $n = 4$
- $q = (1-p) = (1-0.2) = 0.8$

∴ Binomial Distribution seems most appropriate

- $p(x=3) = \frac{n!}{x!(n-x)!} \, p^x \, q^{n-x}$
- $p(x=3) = \frac{4!}{3!(4-3)!} (0.2)^3 (0.8)^1$

$= \frac{24}{6(1)} (0.008)(0.8)$

$= (4)(0.008)(0.8)$

$= 0.0256$

$= 2.56\%$

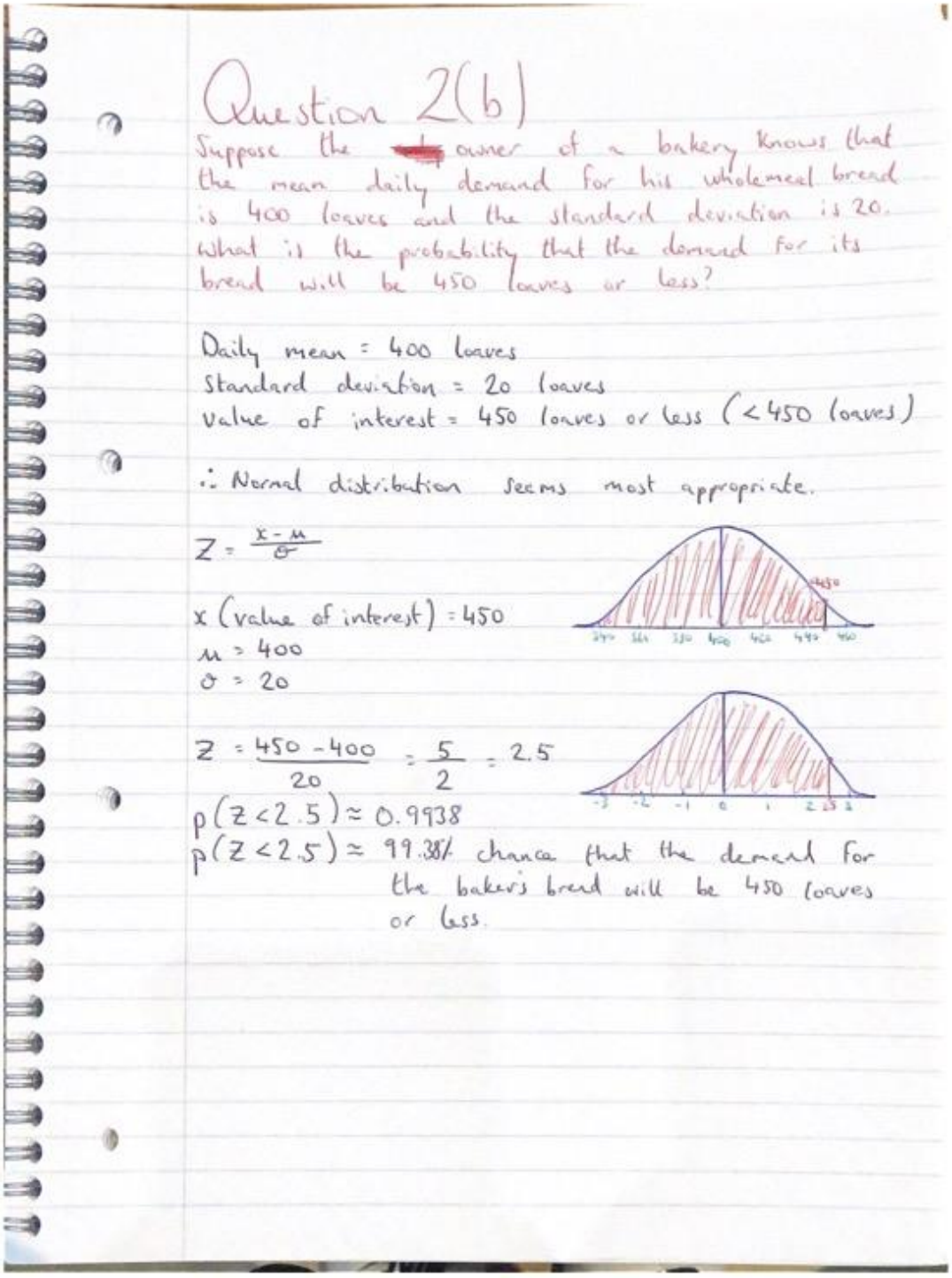
There is a probability of 2.56% that the sales representative will sell products to 3 clients.

Question 2

a) A sales representative has appointments with 4 prospective clients. From past experience, he knows that the probability of making a sale on any appointment is 0.20. What is the probability that he will sell products to 3 clients of them?

```
[19] # Import binom and norm from scipy module
     from scipy.stats import binom, norm

     # Calculate binomial probability
     probability = binom.pmf(k=3, n=4, p=0.2)

     # Display the probability
     print('The probability that he will sell products to 3 clients is', format(probability, '.4f'), '=', format(probability * 100, '.2f'), '%')

     The probability that he will sell products to 3 clients is 0.0256 = 2.56 %
```

1. b) Suppose the owner of a bakery knows that the mean daily demand for his wholemeal bread is 400 loaves and the standard deviation is 20. What is the probability that the demand for its bread will be 450 loaves or less?



Question 2(b)

Suppose the owner of a bakery knows that the mean daily demand for his wholemeal bread is 400 loaves and the standard deviation is 20. What is the probability that the demand for its bread will be 450 loaves or less?

Daily mean = 400 loaves
Standard deviation = 20 loaves
Value of interest = 450 loaves or less (< 450 loaves)

∴ Normal distribution seems most appropriate.

$Z = \frac{x - \mu}{\sigma}$

x (value of interest) = 450
$\mu = 400$
$\sigma = 20$

$Z = \frac{450 - 400}{20} = \frac{5}{2} = 2.5$

$p(Z < 2.5) \approx 0.9938$
$p(Z < 2.5) \approx$ 99.38% chance that the demand for the baker's bread will be 450 loaves or less.

b) Suppose the owner of a bakery knows that the mean daily demand for his wholemeal bread is 400 loaves and the standard deviation is 20. What is the probability that the demand for its bread will be 450 loaves or less?

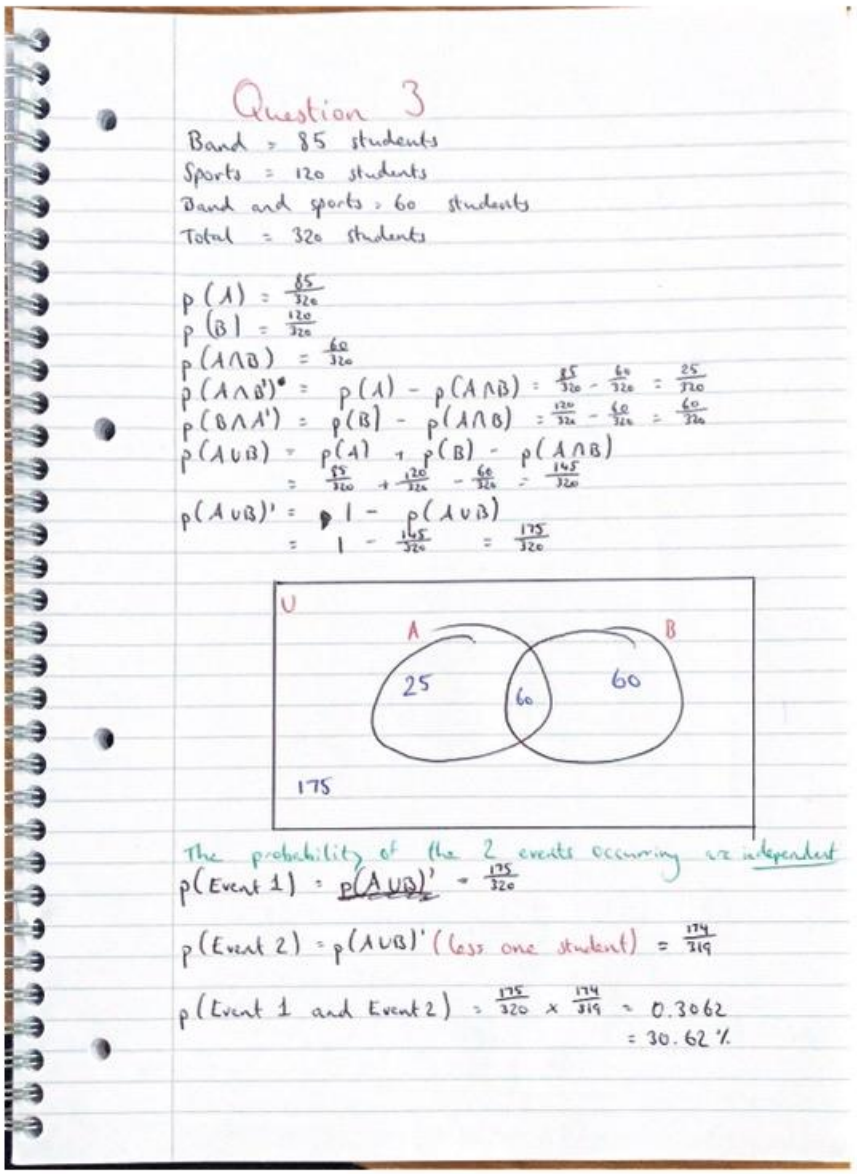
```
[20] # Variables
     mean = 400
     std_dev = 20
     demand_for_bread = 450

     # Calculate the z-score (standard normal distribution)
     z_score = (demand_for_bread - mean) / std_dev

     # Use the cumulative distribution function (.cdf) to find the probability
     probability = norm.cdf(z_score)

     # Display the probability
     print("Probability that the demand for bread will be 450 loaves:", format(probability, '.4f'), '=', format(probability * 100, '.2f'), '%')
```

```
Probability that the demand for bread will be 450 loaves: 0.9938 = 99.38 %
```

Question 3

Solve the following questions.

a) In a school of 320 students, 85 students are in the band, 120 students are on sports teams, and 60 students participate in both activities. If two students are selected randomly, what is the probability that both students will not participate in any activities?



Question 3

Band = 85 students
Sports = 120 students
Band and sports = 60 students
Total = 320 students

$P(A) = \frac{85}{320}$
$P(B) = \frac{120}{320}$
$P(A \cap B) = \frac{60}{320}$
$P(A \cap B')^* = P(A) - P(A \cap B) = \frac{85}{320} - \frac{60}{320} = \frac{25}{320}$
$P(B \cap A') = P(B) - P(A \cap B) = \frac{120}{320} - \frac{60}{320} = \frac{60}{320}$
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
$= \frac{85}{320} + \frac{120}{320} - \frac{60}{320} = \frac{145}{320}$
$P(A \cup B)' = 1 - P(A \cup B)$
$= 1 - \frac{145}{320} = \frac{175}{320}$

The probability of the 2 events occurring are independent

$P(\text{Event 1}) = P(A \cup B)' = \frac{175}{320}$

$P(\text{Event 2}) = P(A \cup B)'$ (less one student) $= \frac{174}{319}$

$P(\text{Event 1 and Event 2}) = \frac{175}{320} \times \frac{174}{319} = 0.3062$
$= 30.62\%$

Question 3

a) In a school of 320 students, 85 students are in the band, 120 students are on sports teams, and 60 students participate in both activities. If two students are selected randomly, what is the probability that both students will not participate in any activities?

```
# Variables
total_students = 320
students_in_band = 85
students_on_sports_teams = 120
students_in_both = 60

# Probability of the studets not paticipating
not_participating = total_students - (students_in_band + students_on_sports_teams - students_in_both)

# Probability that both students will not participate in any activities
probability_not_participating = (not_participating / total_students) * ((not_participating - 1) / (total_students - 1))

# Display the probability
print("Probability that both students will not participate in any activities:", format(probability_not_participating, '.4f'), '=', format(probability_not_p
```

```
Probability that both students will not participate in any activities: 0.2983 = 29.83 %
```