Sydney Ball

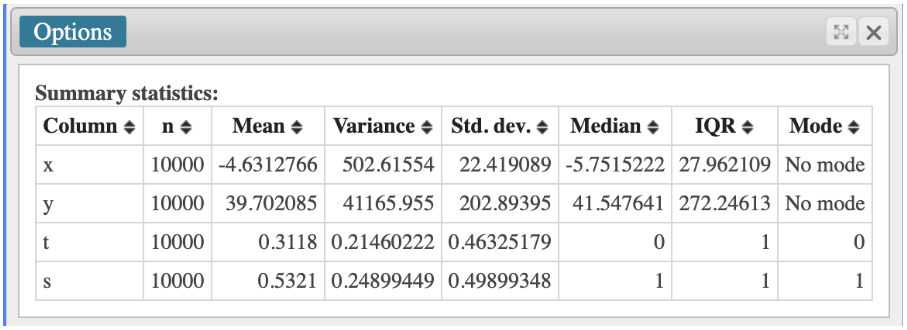
HW#1A

14 January 2025

https://www.donnadietz.com/data/newdata.php

Single Numerical Variable:

6.

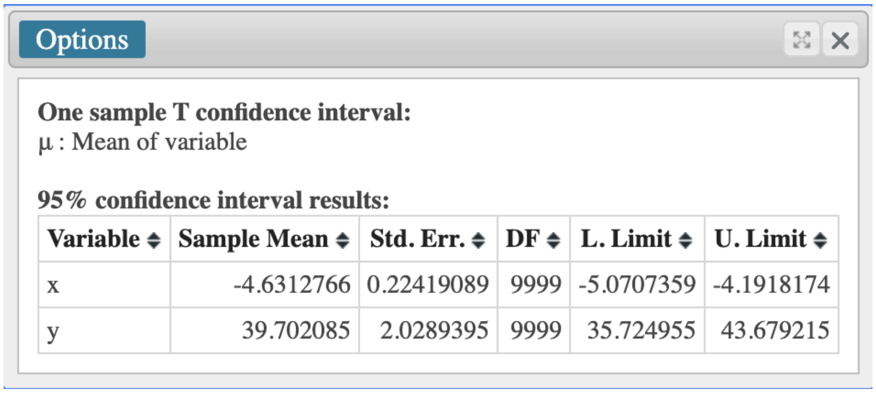


Summary statistics:

Column	n	Mean	Variance	Std. dev.	Median	IQR	Mode
x	10000	-4.6312766	502.61554	22.419089	-5.7515222	27.962109	No mode
y	10000	39.702085	41165.955	202.89395	41.547641	272.24613	No mode
t	10000	0.3118	0.21460222	0.46325179	0	1	0
s	10000	0.5321	0.24899449	0.49899348	1	1	1

The data shows that 53.21% of sample is male. Etc.

7.



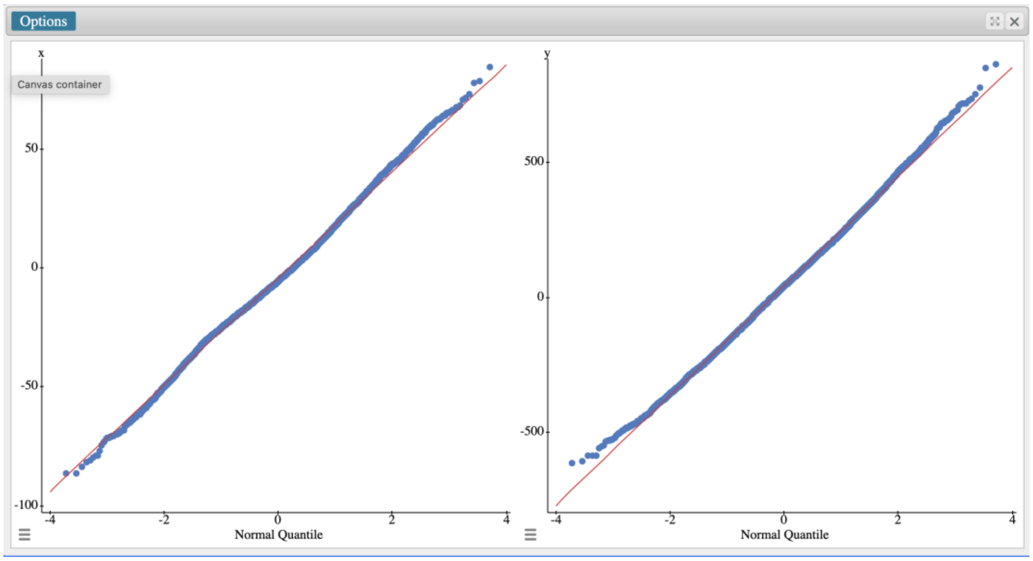
One sample T confidence interval:
μ : Mean of variable

95% confidence interval results:

Variable	Sample Mean	Std. Err.	DF	L. Limit	U. Limit
x	-4.6312766	0.22419089	9999	-5.0707359	-4.1918174
y	39.702085	2.0289395	9999	35.724955	43.679215

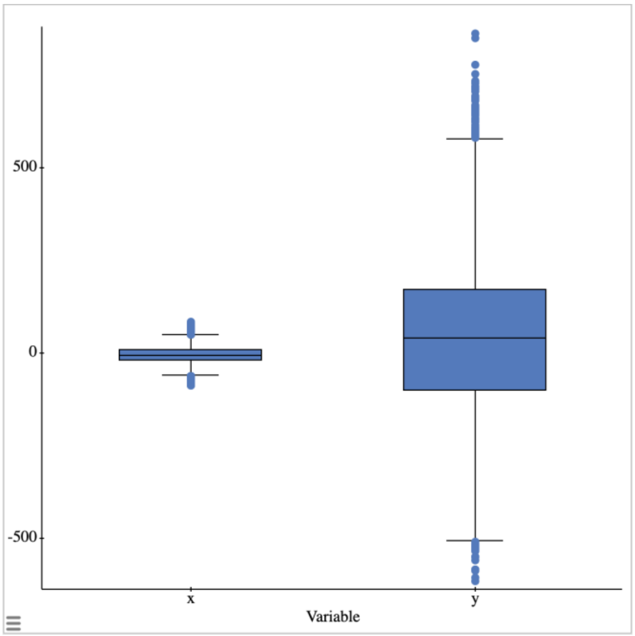
Graphics:

8.

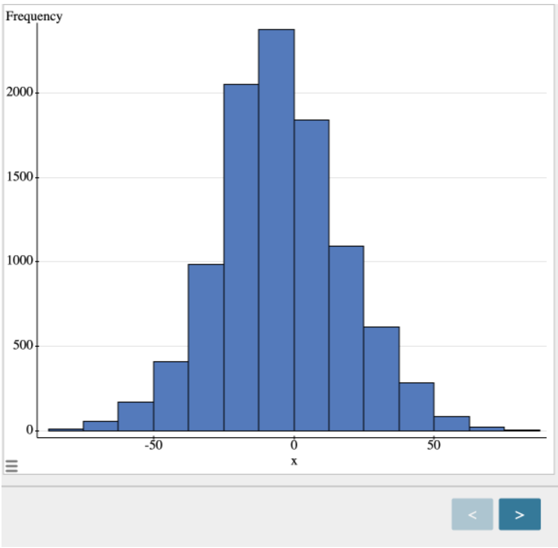


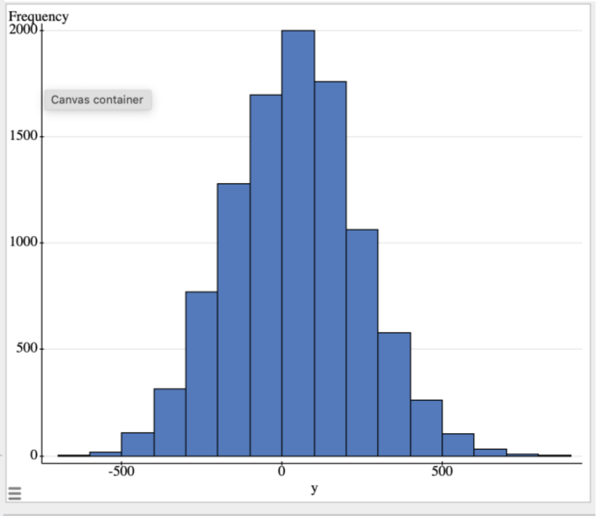
X (on the left) looks a bit better than Y (on the right), but these are generally normal distributions because they are generally fit to the red line.

9.

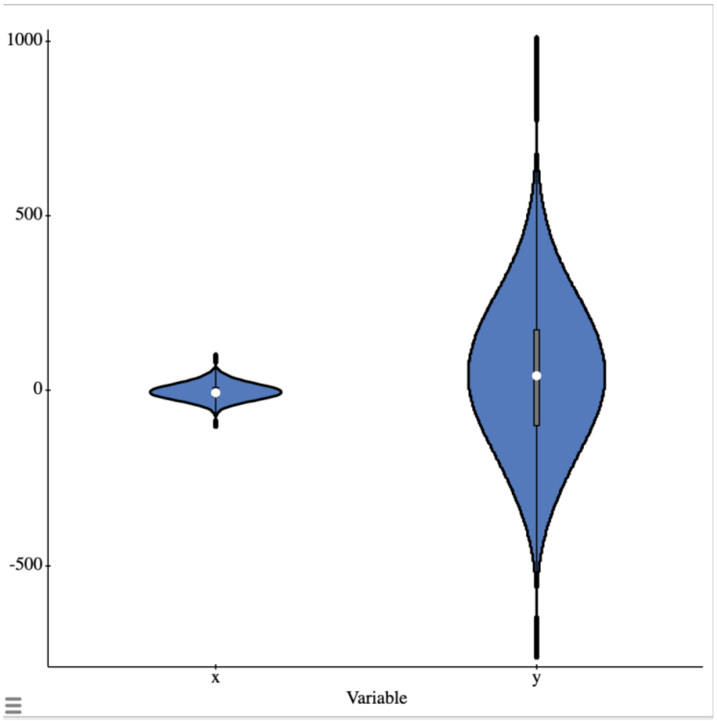


10.



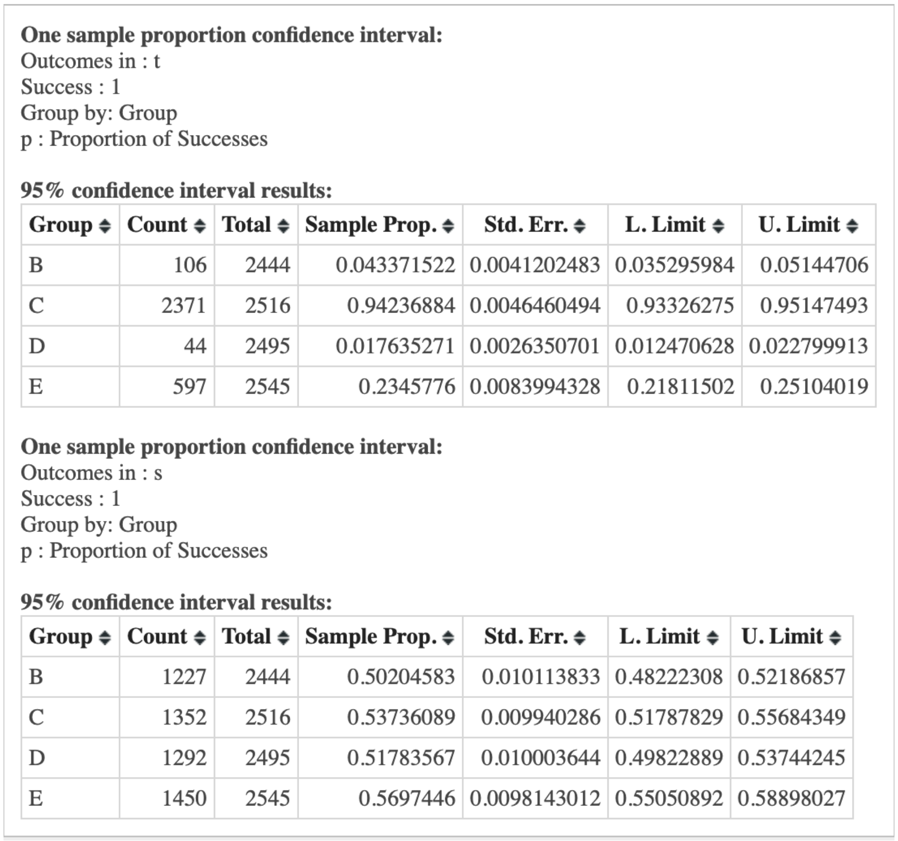


11.



Between the violin plot and the box plot, I think I would prefer the violin plot because it is a better visualization of the data distribution showing more shape than the box plots do. If the data was less normal, a violin plot would also demonstrate this by showing peaks or modes within the data.

12.



One sample proportion confidence interval:
Outcomes in : t
Success : 1
Group by: Group
p : Proportion of Successes

95% confidence interval results:

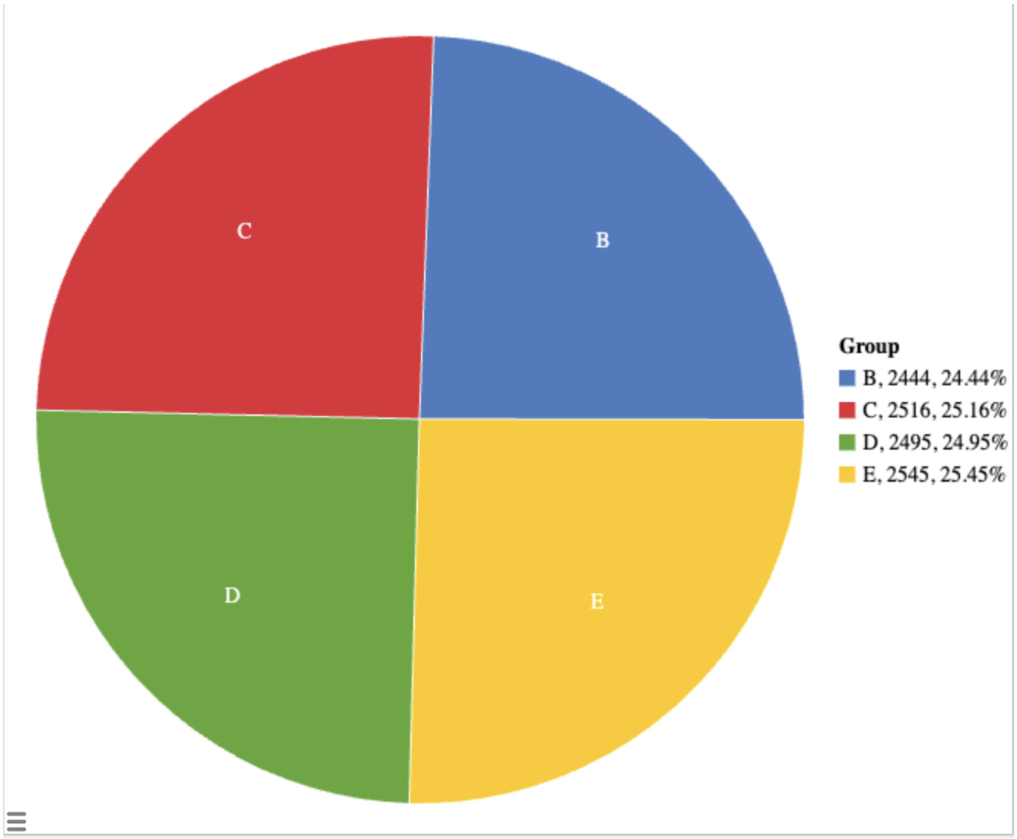
Group	Count	Total	Sample Prop.	Std. Err.	L. Limit	U. Limit
B	106	2444	0.043371522	0.0041202483	0.035295984	0.05144706
C	2371	2516	0.94236884	0.0046460494	0.93326275	0.95147493
D	44	2495	0.017635271	0.0026350701	0.012470628	0.022799913
E	597	2545	0.2345776	0.0083994328	0.21811502	0.25104019

One sample proportion confidence interval:
Outcomes in : s
Success : 1
Group by: Group
p : Proportion of Successes

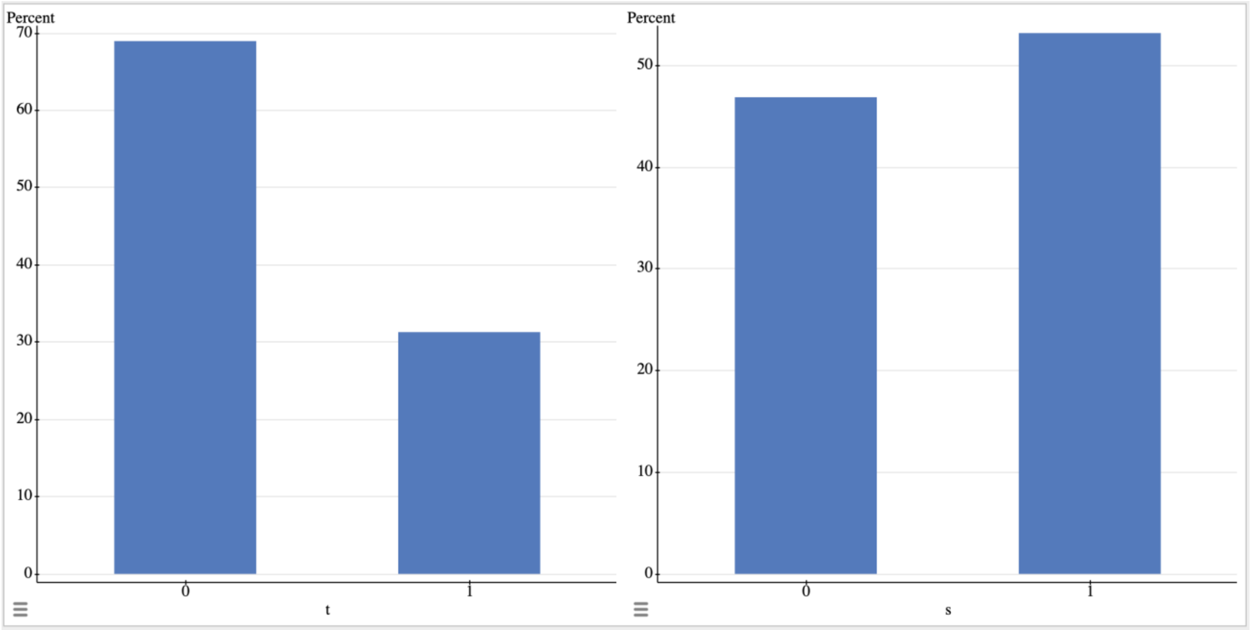
95% confidence interval results:

Group	Count	Total	Sample Prop.	Std. Err.	L. Limit	U. Limit
B	1227	2444	0.50204583	0.010113833	0.48222308	0.52186857
C	1352	2516	0.53736089	0.009940286	0.51787829	0.55684349
D	1292	2495	0.51783567	0.010003644	0.49822889	0.53744245
E	1450	2545	0.5697446	0.0098143012	0.55050892	0.58898027

13.

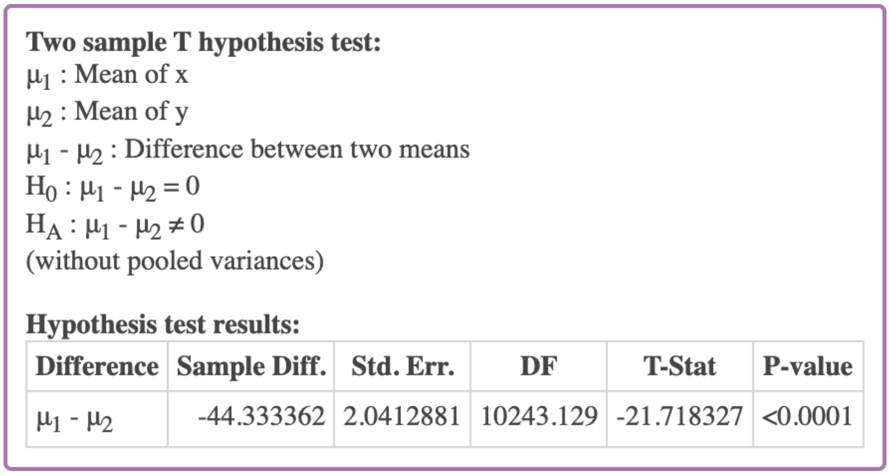

C
B
D
E
Group
🟦 B, 2444, 24.44%
🟥 C, 2516, 25.16%
🟩 D, 2495, 24.95%
🟨 E, 2545, 25.45%

14.


Percent
70
60
50
40
30
20
10
0
0
1
t
Percent
50
40
30
20
10
0
0
1
s

Two Unpaired numerical variables:

15.



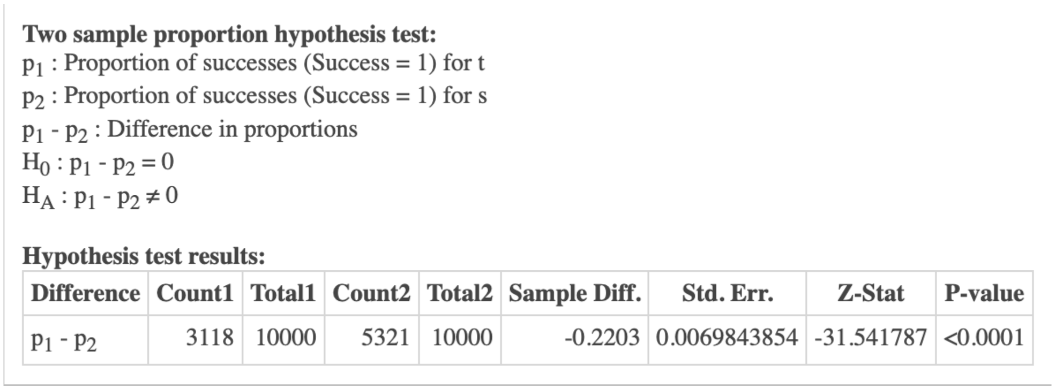
Two sample T hypothesis test:
μ_1 : Mean of x
μ_2 : Mean of y
$\mu_1 - \mu_2$: Difference between two means
$H_0 : \mu_1 - \mu_2 = 0$
$H_A : \mu_1 - \mu_2 \neq 0$
(without pooled variances)

Hypothesis test results:

Difference	Sample Diff.	Std. Err.	DF	T-Stat	P-value
$\mu_1 - \mu_2$	-44.333362	2.0412881	10243.129	-21.718327	<0.0001

Two Unpaired categorical variables:

16.



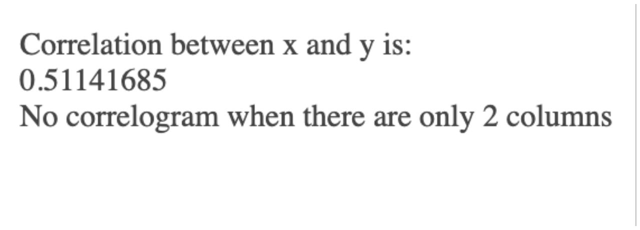
Two sample proportion hypothesis test:
p_1 : Proportion of successes (Success = 1) for t
p_2 : Proportion of successes (Success = 1) for s
$p_1 - p_2$: Difference in proportions
$H_0 : p_1 - p_2 = 0$
$H_A : p_1 - p_2 \neq 0$

Hypothesis test results:

Difference	Count1	Total1	Count2	Total2	Sample Diff.	Std. Err.	Z-Stat	P-value
$p_1 - p_2$	3118	10000	5321	10000	-0.2203	0.0069843854	-31.541787	<0.0001

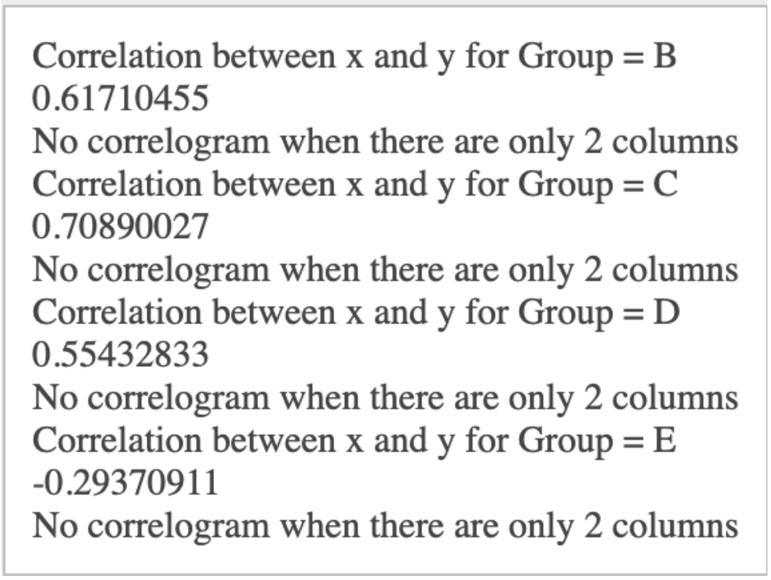
Two Paired Numerical Variables:

17.



Correlation between x and y is:
0.51141685
No correlogram when there are only 2 columns

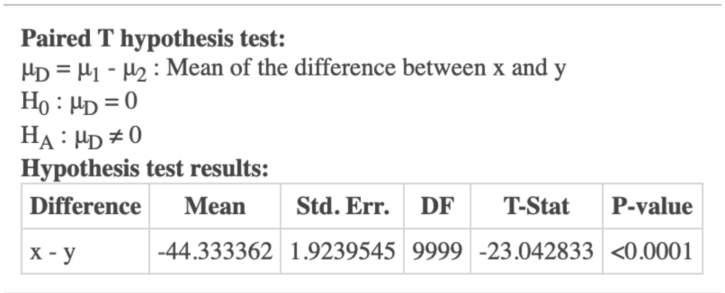
18.



```
Correlation between x and y for Group = B
0.61710455
No correlogram when there are only 2 columns
Correlation between x and y for Group = C
0.70890027
No correlogram when there are only 2 columns
Correlation between x and y for Group = D
0.55432833
No correlogram when there are only 2 columns
Correlation between x and y for Group = E
-0.29370911
No correlogram when there are only 2 columns
```

It seems that the strongest correlations within the data are between x and y in group C with a .71 or strong positive relationship. The weakest correlation between x and y was in group E with a weak negative correlation of -.29.

19.



Paired T hypothesis test:

$\mu_D = \mu_1 - \mu_2$: Mean of the difference between x and y

$H_0 : \mu_D = 0$

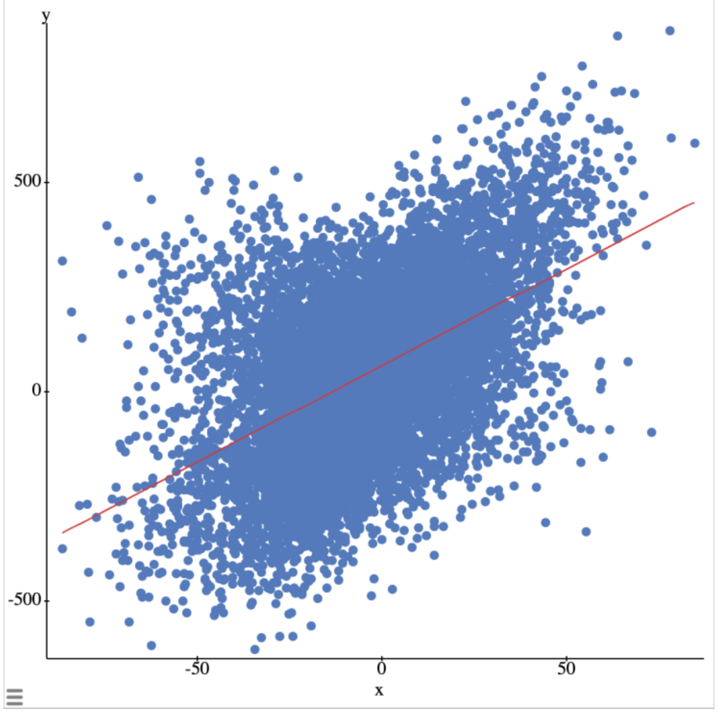
$H_A : \mu_D \neq 0$

Hypothesis test results:

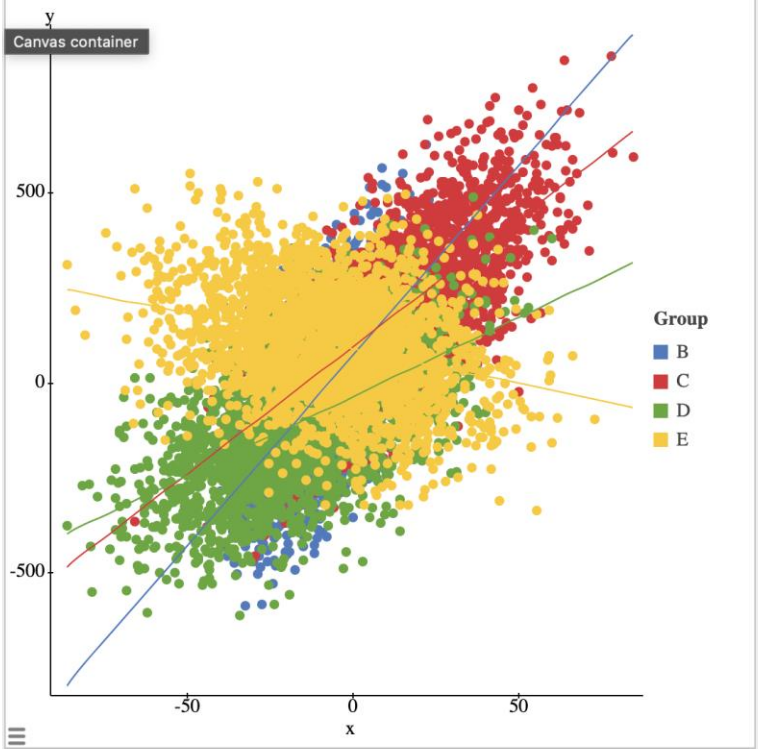
Difference	Mean	Std. Err.	DF	T-Stat	P-value
x - y	-44.333362	1.9239545	9999	-23.042833	<0.0001

Graphics:

20.



21.

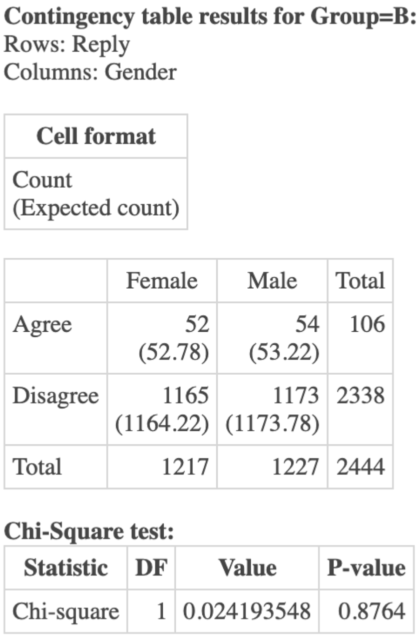


As seen in question 18, the correlation of x and y between the group is now graphically observed. We can see that once again group C in red and group B in blue have very strong and positive correlations between x and y. Inversely, group E in yellow, has a weak negative correlation shown by its downward sloping line. Generally, however, we see B, C, and D have positive and similar correlations, but we would say that red probably has the best fit line.

Two paired categorical variables:

Statistics:

22.



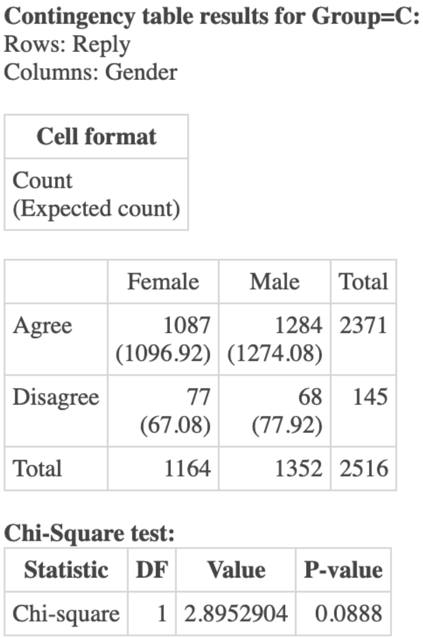
Contingency table results for Group=B:
Rows: Reply
Columns: Gender

Cell format
Count (Expected count)

	Female	Male	Total
Agree	52 (52.78)	54 (53.22)	106
Disagree	1165 (1164.22)	1173 (1173.78)	2338
Total	1217	1227	2444

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	1	0.024193548	0.8764



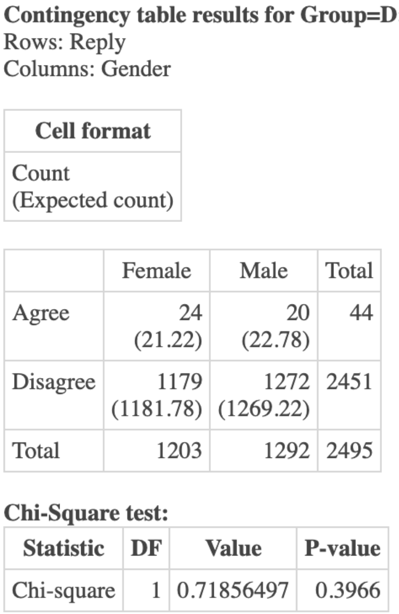
Contingency table results for Group=C:
Rows: Reply
Columns: Gender

Cell format
Count (Expected count)

	Female	Male	Total
Agree	1087 (1096.92)	1284 (1274.08)	2371
Disagree	77 (67.08)	68 (77.92)	145
Total	1164	1352	2516

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	1	2.8952904	0.0888



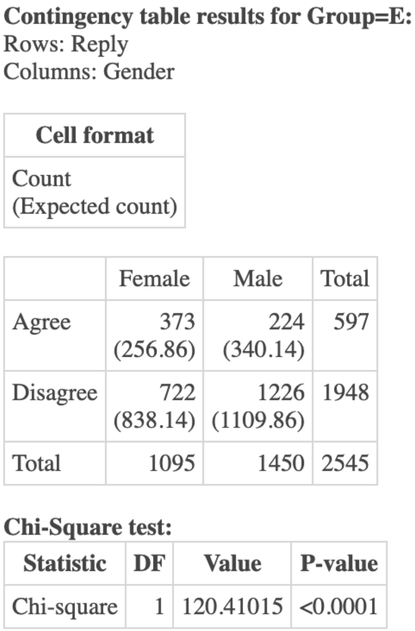
Contingency table results for Group=D:
Rows: Reply
Columns: Gender

Cell format
Count (Expected count)

	Female	Male	Total
Agree	24 (21.22)	20 (22.78)	44
Disagree	1179 (1181.78)	1272 (1269.22)	2451
Total	1203	1292	2495

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	1	0.71856497	0.3966



Contingency table results for Group=E:
Rows: Reply
Columns: Gender

Cell format
Count (Expected count)

	Female	Male	Total
Agree	373 (256.86)	224 (340.14)	597
Disagree	722 (838.14)	1226 (1109.86)	1948
Total	1095	1450	2545

Chi-Square test:

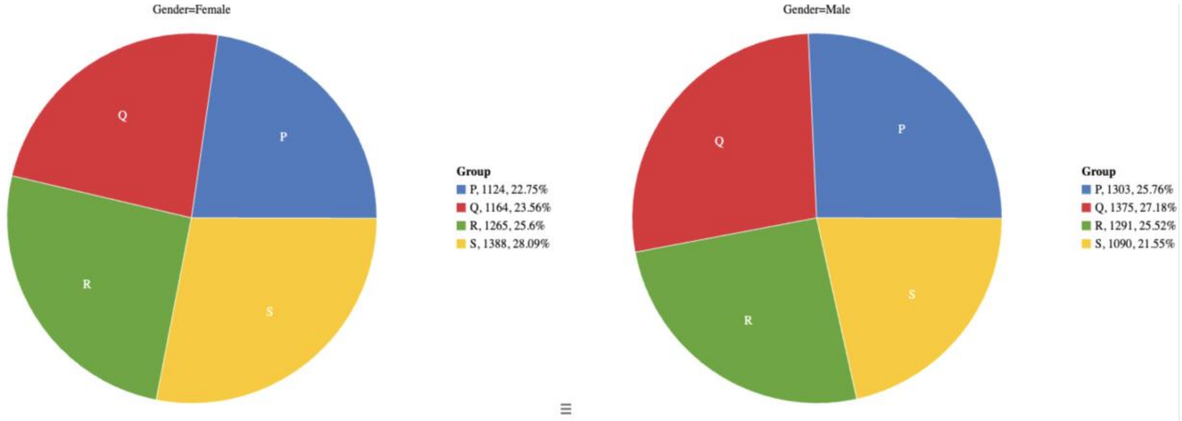
Statistic	DF	Value	P-value
Chi-square	1	120.41015	<0.0001

The results show that females in group E were more likely to agree than males in group E are. Because of the low p-value of <0.0001, it's unlikely that the difference in the proportions is due to sampling error.

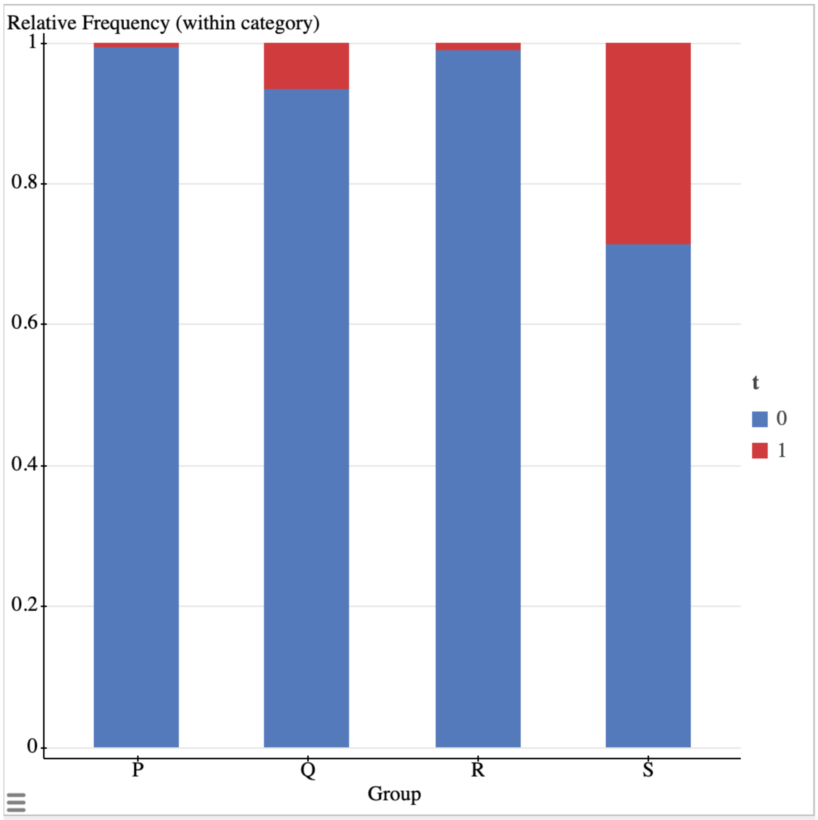
Alternatively, the other chi-squared results have higher p-values of .87, .08, and .39 (> .05) and so a significant conclusion cannot be made based on these outputs. We cannot conclude that males or females from groups B, C, or D are more or less likely to agree or disagree to the statements.

Graphics:

23.



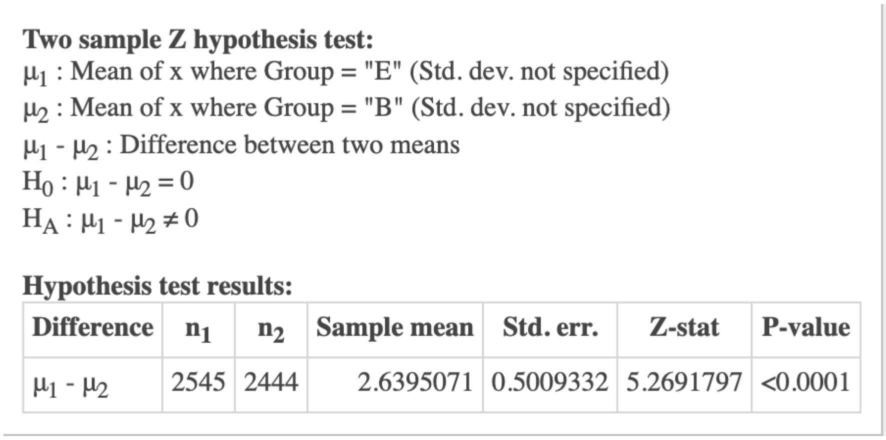
24.



Paired: one numerical and one categorical variable:

Statistics:

25.



Two sample Z hypothesis test:

μ_1 : Mean of x where Group = "E" (Std. dev. not specified)

μ_2 : Mean of x where Group = "B" (Std. dev. not specified)

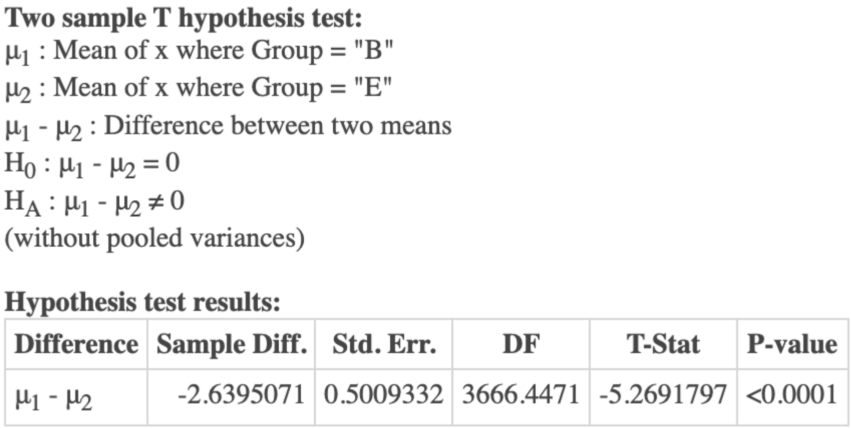
$\mu_1 - \mu_2$: Difference between two means

$H_0 : \mu_1 - \mu_2 = 0$

$H_A : \mu_1 - \mu_2 \neq 0$

Hypothesis test results:

Difference	n_1	n_2	Sample mean	Std. err.	Z-stat	P-value
$\mu_1 - \mu_2$	2545	2444	2.6395071	0.5009332	5.2691797	<0.0001



Two sample T hypothesis test:

μ_1 : Mean of x where Group = "B"

μ_2 : Mean of x where Group = "E"

$\mu_1 - \mu_2$: Difference between two means

$H_0 : \mu_1 - \mu_2 = 0$

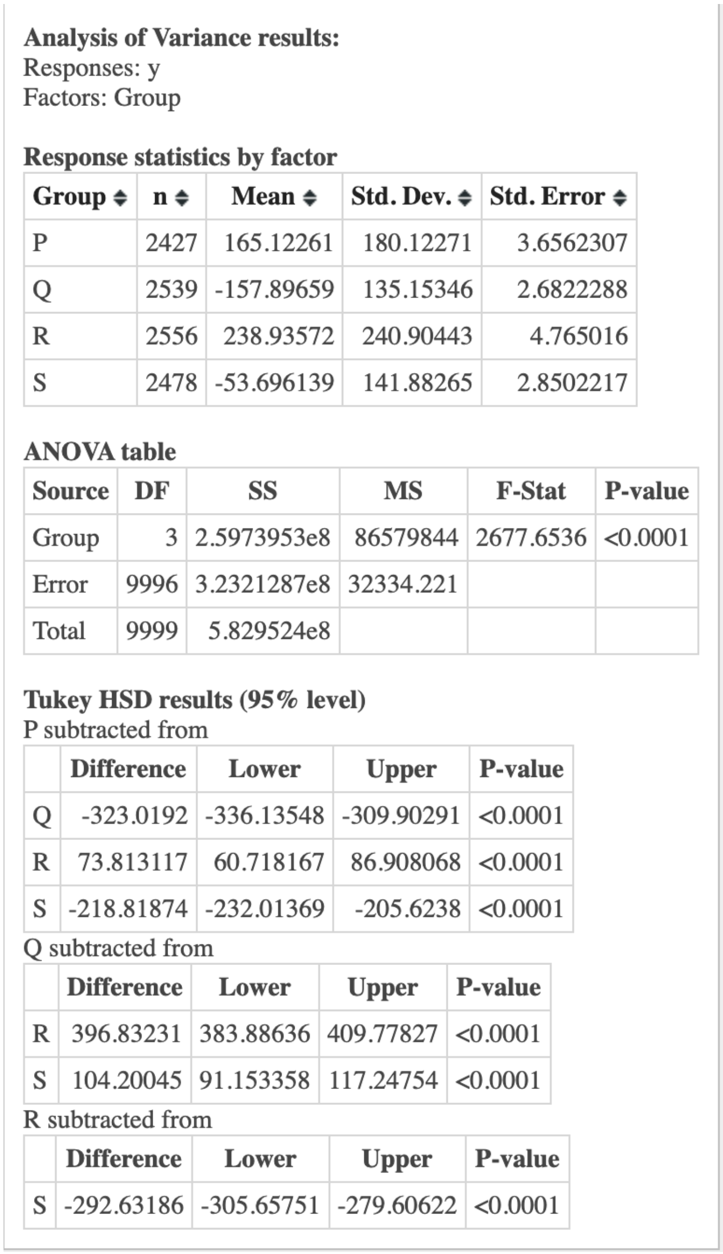
$H_A : \mu_1 - \mu_2 \neq 0$

(without pooled variances)

Hypothesis test results:

Difference	Sample Diff.	Std. Err.	DF	T-Stat	P-value
$\mu_1 - \mu_2$	-2.6395071	0.5009332	3666.4471	-5.2691797	<0.0001

26.



Analysis of Variance results:
Responses: y
Factors: Group

Response statistics by factor

Group	n	Mean	Std. Dev.	Std. Error
P	2427	165.12261	180.12271	3.6562307
Q	2539	-157.89659	135.15346	2.6822288
R	2556	238.93572	240.90443	4.765016
S	2478	-53.696139	141.88265	2.8502217

ANOVA table

Source	DF	SS	MS	F-Stat	P-value
Group	3	2.5973953e8	86579844	2677.6536	<0.0001
Error	9996	3.2321287e8	32334.221		
Total	9999	5.829524e8			

Tukey HSD results (95% level)
P subtracted from

	Difference	Lower	Upper	P-value
Q	-323.0192	-336.13548	-309.90291	<0.0001
R	73.813117	60.718167	86.908068	<0.0001
S	-218.81874	-232.01369	-205.6238	<0.0001

Q subtracted from

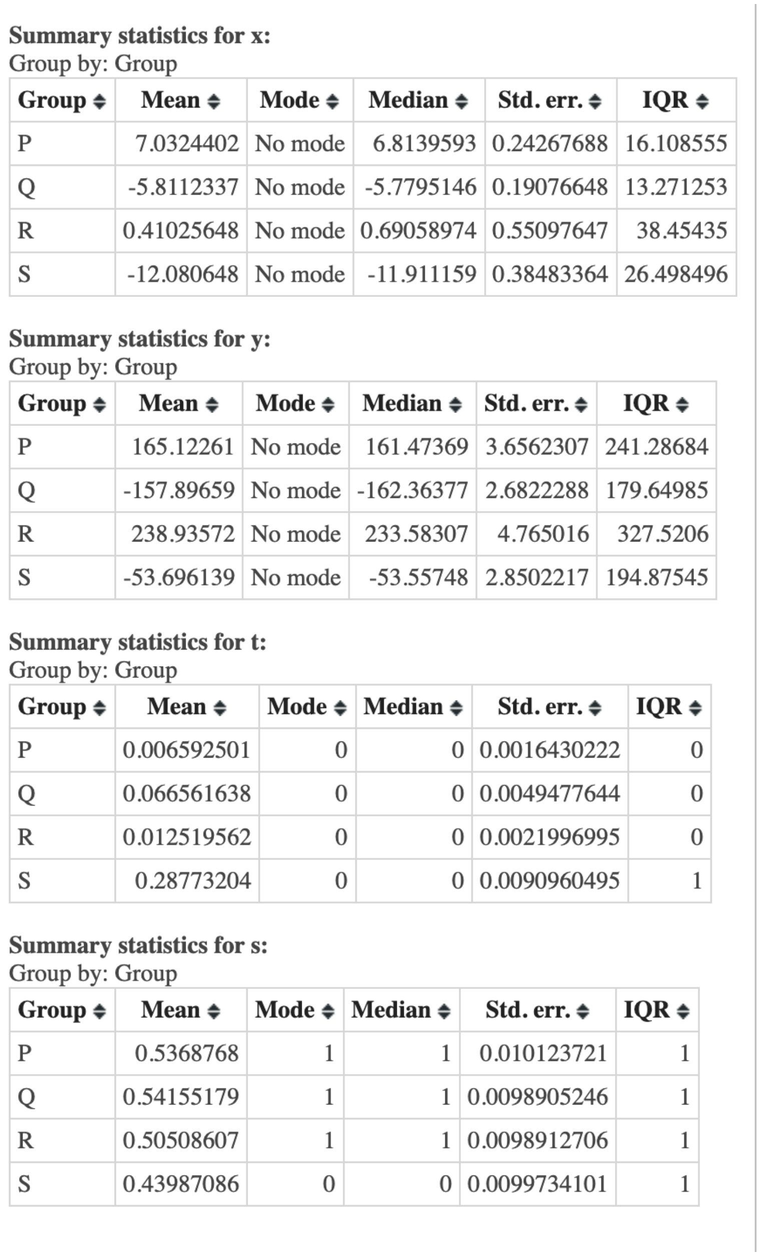
	Difference	Lower	Upper	P-value
R	396.83231	383.88636	409.77827	<0.0001
S	104.20045	91.153358	117.24754	<0.0001

R subtracted from

	Difference	Lower	Upper	P-value
S	-292.63186	-305.65751	-279.60622	<0.0001

Within my data, I found that all my pairwise comparisons are highly significant with p-values of <0.0001. Given the very low p-values I can conclude that all the pairwise comparisons are statistically significant. We are quite positive that these are different groups because all the p-values between groups are significant.

27.



Summary statistics for x:
Group by: Group

Group ⬍	Mean ⬍	Mode ⬍	Median ⬍	Std. err. ⬍	IQR ⬍
P	7.0324402	No mode	6.8139593	0.24267688	16.108555
Q	-5.8112337	No mode	-5.7795146	0.19076648	13.271253
R	0.41025648	No mode	0.69058974	0.55097647	38.45435
S	-12.080648	No mode	-11.911159	0.38483364	26.498496

Summary statistics for y:
Group by: Group

Group ⬍	Mean ⬍	Mode ⬍	Median ⬍	Std. err. ⬍	IQR ⬍
P	165.12261	No mode	161.47369	3.6562307	241.28684
Q	-157.89659	No mode	-162.36377	2.6822288	179.64985
R	238.93572	No mode	233.58307	4.765016	327.5206
S	-53.696139	No mode	-53.55748	2.8502217	194.87545

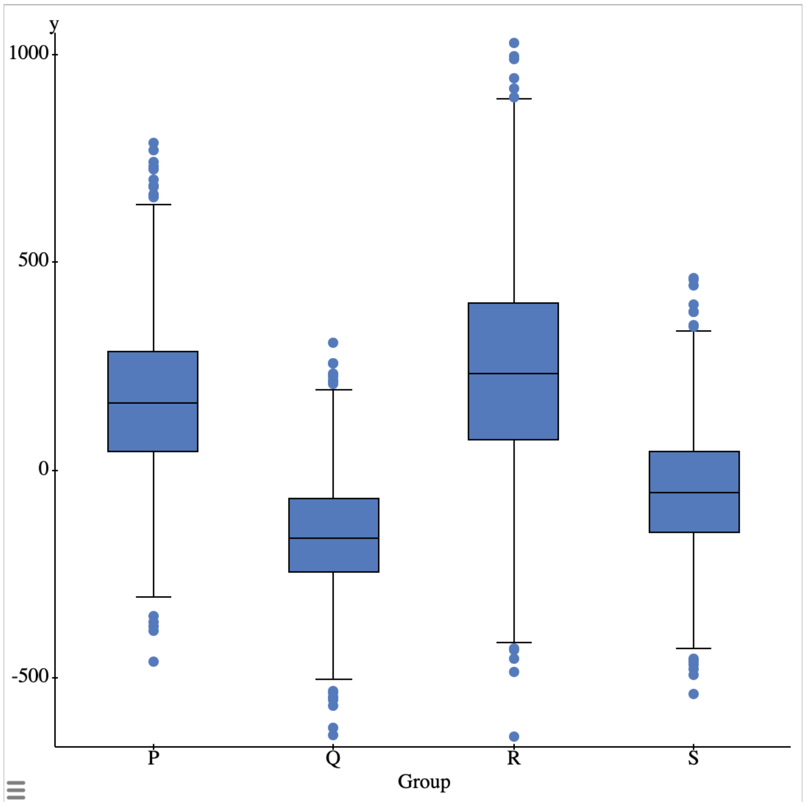
Summary statistics for t:
Group by: Group

Group ⬍	Mean ⬍	Mode ⬍	Median ⬍	Std. err. ⬍	IQR ⬍
P	0.006592501	0	0	0.0016430222	0
Q	0.066561638	0	0	0.0049477644	0
R	0.012519562	0	0	0.0021996995	0
S	0.28773204	0	0	0.0090960495	1

Summary statistics for s:
Group by: Group

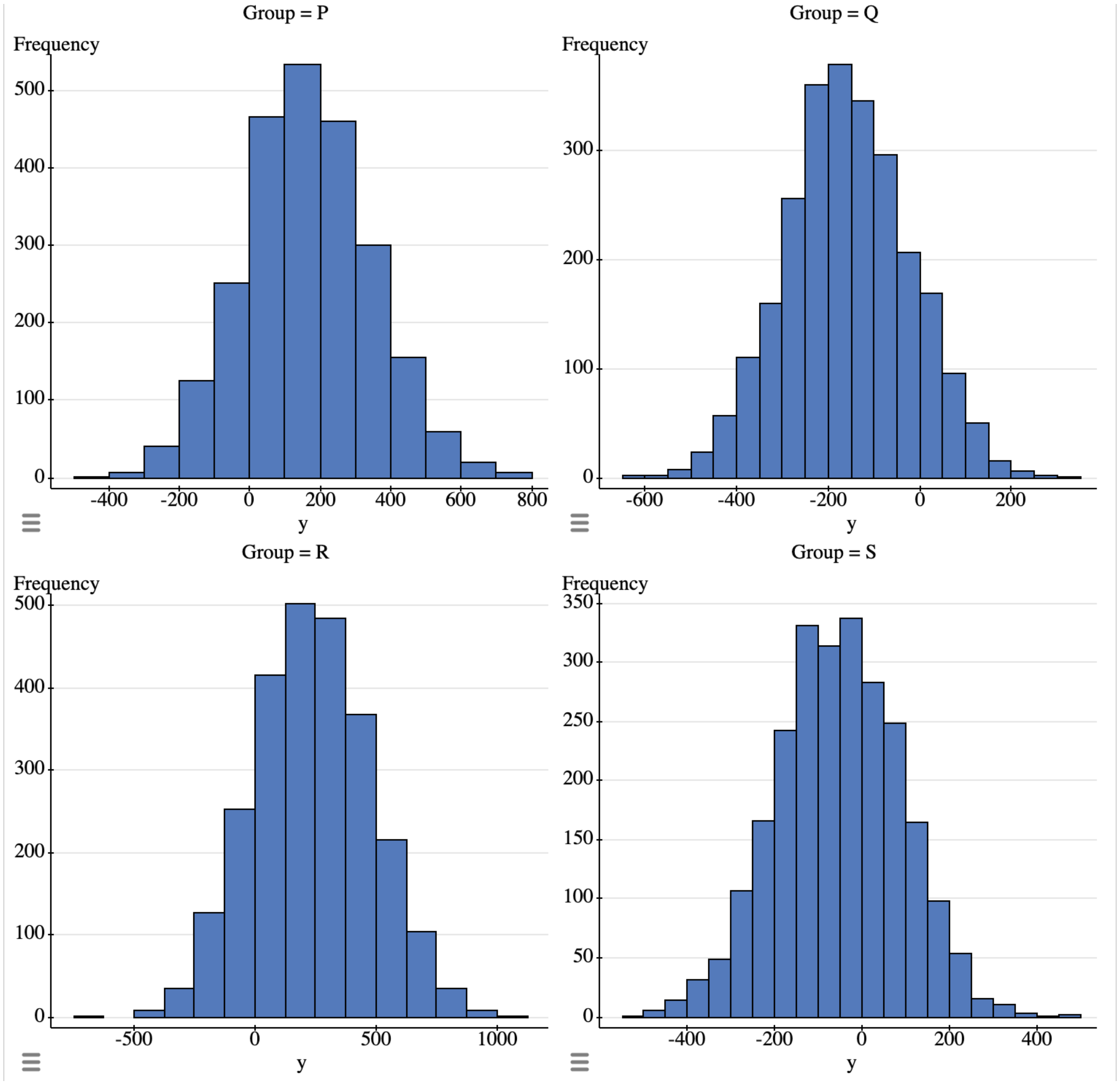
Group ⬍	Mean ⬍	Mode ⬍	Median ⬍	Std. err. ⬍	IQR ⬍
P	0.5368768	1	1	0.010123721	1
Q	0.54155179	1	1	0.0098905246	1
R	0.50508607	1	1	0.0098912706	1
S	0.43987086	0	0	0.0099734101	1

Graphics:



28. ≡

29.



This assignment was mainly just a refresher for me. I have not worked with ANOVA or Chi-squared for a few semesters. I also found the refresher of working with different types of variables like categorical and numerical to be a good and helpful practice. Working with different types of graphics and models to draw conclusions on the data is always helpful. Moreover, I found that just reading outputs from software to software can be different, so working in stat crunch was also good practice.