Survey Analysis of CPP Students

Cyrus Davis, David Kaufman, Eugene Monforte, Max Wang

March 13, 2025

Abstract:

As statistical consultants, we were tasked to assess data from a dataset containing responses of Cal Poly Pomona (CPP) students regarding their experiences and environments. The data was comprised of 1,250 students, who responded to 17 categorical questions. This report provides a series of statistical tests to identify if the questions postulated specific distributions, examine the relationship among questions and if observed values were statistically different from expected values.

Introduction:

A survey was conducted over a five-day period at the California State Polytechnic University, Pomona by 10 student surveyors. The purpose of the survey was to collect student views on important issues related to their CPP environments. The survey queried students on transportation topics and communication facilities.

Our objective is to analyze the 17 categorical questions through a series of statistical tests: Chi Squared Test, Goodness of Fit Test, the Chi Squared Test of Independence and Z-Tests. The Goodness of Fit test will determine if a question result suggest a specific distribution, such as the binomial distribution. The test of Independence will determine if two data results are related, and the Z tests will test proportions among various groups. The discussions of each test will include brief interpretations along with a statement of the hypothesis we are testing.

Methods:

Throughout this report, we will be using R Software to conduct the specific tests: Chi Squared Test of Independence (TOI), Chi Squared Goodness of Fit (GOF) Test, and Z Tests for proportions. Both the GOF test and TOI tests, each will be done using the R function: chisq.test() but for a TOI test, a table will be constructed between two questions based on valid responses that are not NA values. For the Z tests, the prop.test() will be used. In addition, we will default to the "correct = FALSE" parameter, to disengage R's automatic adjusting mechanism due to the size of the dataset (n=1,250). The tests have been divided into 3 main stages.

Stage 1:

In stage 1, relates to Q#01 (Are you male or female), Q#04 (Did you come to CPP because of...), Q#13 (What makes CPP stand out from other universities?). For Q#01, we conducted a GOF test to determine if a binomial distribution was appropriate at $p_i = 0.50$ where i = a (male) and b (female). HOI test whether the proportion of females were significantly greater than 0.50 and compared the p-values between the 2 tests. For Q#04, a GOF test was conducted to determine whether a binomial distribution of $p_i = 0.20$ is a suitable model for describing the number of students for each of the category i: a (closeness to home), b (someone's recommendation), c (Beautiful Campus), d (good academic reputation) and e (others). Q#13, had the choices: a (the faculty), b (the education), c (the campus), d (it does not stand out), where each choice was set to their respective probability, $p_a = 0.10$ and $p_b = p_c = p_d = 0.30$. A GOF test is conducted to

determine if the probabilities were reasonable in describing the number of students in each category. A Z test conducted to determine whether the "Wow" category, comprises of a, b, and c, was greater than 2/3 of the selected students.

Each test has the code format:

__.test(counts, probability, correct = FALSE),

where ___ is either "prop" or "chisq" used and the probablity is based on which category is being set for testing.

Stage 2:

Stage two consists of a series of tests comparing the three questions: Q#01, Q#03 (Was CPP your first choice for college), and Q#04. There was a comparison between Q#01 and Q#03 on whether they are related, a comparison between Q#01 and Q#04 to explore the question if the reasons why student came to CPP varied among gender. In addition, a Z test was used for Q#03 and Q#04 to compare the proportion of students who came to CPP based on its academic reputation to the proportion of respondents who answered "yes" on Q#03.

Each of the following will be constructed using as table. The r code format conducting this was using: table(a, b) where a = Q#01, Q#03, and Q#04, and b = Q#03, Q#04, and Q#03.

Stage 3:

Stage 3 has consisted of Q#07 and Q#08. For this stage, we will divide both questions into 3a and 3b. For 3a, we compared Q#02 (What is your class standing) and Q#12 (How many times a week do you attend CPP). 3b was coupled with Q#07 (Do you use the e-mail stations at school?), Q#7a (How often?) and Q#7b (At what location do you most often use them).

Stage 3a:

A TOI was conducted to test whether the two questions were independent. Afterwards two new datasets were created, (each with their respective tests). A TOI test was done to determine the relationship between two variables. Two auxiliary datasets were derived using Q02 and Q12. Correlation coefficients were commuted, and a Z test was conducted to determine whether the coefficients were significantly different from zero.

Stage 3b:

A Z test was done for Q3b to determine whether the proportion of students who use e-mail stations at CPP were significantly greater than the proportion of students who do not use the e-mail stations. All hypothesis testing was done at the alpha level, $\alpha = 0.05$, where the null hypothesis H_0 and alternative hypothesis H_a. See the following format on the next page:

Goodness of Fit Test (GOF):

H_0: $p_i = x$ where x = probability vs H_a: at least one p_i is different

X^2 = ____ at df = k – 1 where k = # of categories

p-value = ____

Test of Independence (TOI):

H_0: Q#A and Q#B are independent vs H_a: Q#A and Q#B are not independent

X^2 = ____ at df = (r – 1)(c – 1) where r = # of rows of table, c = # of columns of table

p-value = ____

Z Test:

H_0: p_1 __ p_2 vs H_a: p_1 __ p_2

Where p_1 is the testing condition and p_2 is the set value or question.

" __" is determined if it is greater (>), less than (<) or equal to (=) for H_0 and

H_1 has the opposite sign

Z = ____ p-value = ____

Results:

Stage 1: Q#01, Q#04, Q#13

Q#01 (Are you male or female?) has the choices: a (male) and b (female). To test whether a binomial distribution of p_i (i = a, b) = 0.50 is a suitable model for the number of students, we were required to gather the number of students. The table below displays the number of students (male and female). As there were no "NA" responses, we continued with the GOF test for the binomial distribution to see if the proportion of female students is greater than 0.50.

Q#01 Responses:

Choice	Male (a)	Female (b)	Total
Frequency	615	635	1250

GOF test for binomial distribution:

$$H_0\text{: } p_a = p_b = 0.50 \quad \text{vs} \qquad H_a\text{:At least one } p_i\ (i = a, b) \text{ is different}$$

$X^2 = 0.32$, with df = 2 – 1 = 1, p-value = 0.5716

GOF test for female students:

$$H_0\text{: } p_b \leq 0.50 \qquad \text{vs} \qquad H_a\text{: } p_b > 0.50$$

$X^2 = 0.32$**,** df = 1, p-value = 0.2858

Since both the p-values for the GOF test for binomial distribution and test for female students are greater than 0.50, we do not reject the null hypothesis. Since the null hypothesis was not rejected at the alpha level, we can say that Q#01 may follow a binomial distribution, and the proportion of female students is about 0.50.

Observing the p-values for the GOF tests, the p-values are different. This is because the GOF test for binomial distribution is a two tailed test and the second GOF test is a one-sided test. This is confirmed by doubling the p-value of the GOF test for female students. 0.2858(2) = 0.5716.

For Q#04 (Did you come to CPP because of), we have the following choices: a (closeness to home), b (Someone's recommendation), c (Beautiful campus), d (A school of good academic reputation) and e (Others). When cleaning the data, we observed there are 3 "NA" responses. These were subtracted from the GOF test we conducted (from a sample size of 1,250 to 1,247 student responses). The 1,250 sample responses are recorded in the table on the next page.

Q#04 Valid Reponses:

Choice	Frequency
Closeness to home (a)	369
Someone's recommendation (b)	169
Beautiful campus (c)	67
A school of good academic reputation (d)	334
Others (e)	308
Total	1247

For the GOF test, we would let $p_i \; where \; i = a, \; b, \; \ldots \; , \; e$ be the probability that the reason why a student came to CPP belongs to the $i \; th$ category. Each $p_i = 0.20$ a the following GOF test is performed on the 1,247 responses, where k = 5 for 5 choices.

GOF Test for Q#04:

H_0: $p_a = p_b = p_c = p_d = p_e = 0.20$ vs H_a: At least one p_i is different from $0.20 X^2 = 259.14$, df = k – 1 = 5 –1 = 4; p-value = 2.2e-15

Since the p-value is less than the alpha value of 0.05, we rejected the null hypothesis. We can conclude that at least one of the probabilities for the question choices for Q04 is different.

For Q#13 (What makes CPP stand out from other universities?), it had the answers choices: a (the faculty), b (the education), c (the campus), d (it does not stand out). From the 1250 responses, we had 4 NA's these won't be included in our tests so we would be testing only on 1246 responses, which is 99.7% of the 1250 response. The following table gives the 1246 samples:

Choice	Frequency
The faculty (a)	115
The education (b)	409
The campus (c)	351
It doesn't stand out (d)	371
Total	1246

From here, we will conduct two tests: a GOF test for the student categories and a Z-test to determine whether a combined "Wow" category of choices (a), (b), and (c) were greater than 2/3 (0.67). For the first test, we set the probabilities, $p_i \; (i = a, b, c, d)$, each were set as: $p_a = 0.10$, $p_b = p_c = p_d = 0.30$ based on the rounded relative frequencies of the table set up. We conducted a GOF test to determine whether test set probabilities were suitable for describing the 4 student categories where k = 4.

GOF test for Q#13 Student Categories:

$$H_0: p_a = 0.10 \; and \; p_b = p_c = p_d = 0.30 \;\; \text{vs} \qquad H_a: \text{At least } \underline{\text{one}}\ p_i \text{ is different}$$

$$X^2 = 4,\ df = k - 1 = 3 - 1 = 2,\ \text{p-value} = 0.2615$$

Since p-value = .2615 >.05, we fail to reject H_0. The assumed probabilities reasonably describe what students think stands out about Cal Poly Pomona. 10% of students think the faculty is the reason CPP stands out, while 30% of students think the reason is education and 30% of students think it's the campus. 30% of students think that Cal Poly Pomona does not stand out from other universities.

Another test was conducted where categories (a), (b) and (c) were combined into an 875 numbered category. A table is listed giving the combined "Wow" category:

Choice	Frequency
WOW (a, b, c)	875
It doesn't stand out (d)	371
Total	1246

From here a Z-test was conducted on 875 "Wow" category to the 1246 table sample with the probability, p = 2/3 (0.67).

Z Test:

$$H_0: p \leq \frac{2}{3} \;\; \text{vs} \qquad H_1: p > \frac{2}{3}$$

$$X^2 = 5.8602,\ df = k - 1 = 2 - 1 = 1,\ \text{p-value} = 0.007743$$

$\frac{2}{3}$ of Cal Poly Pomona students selected chose Cal Poly Pomona because of WOW (faculty, education, campus). This implies less than $\frac{1}{3}$ of students selected, think that Cal Poly Pomona does not stand out from other universities.

Stage 2: Test between two questions: Q#01, Q#03, and Q#04

We first tested whether Q#01 and Q#03 (Was CPP your first choice for college?), were related. Between the two samples Q#03 had 8 NA responses so these won't be part of the 1250 samples. We created a table based on the remaining 1242 samples from the responses to each Q#01 and Q#03. Q#01 has the responses: a (male) and b (female) whereas Q#03 has the responses a (yes) and b (no). Below lists a 2x2 table from the 1242 valid responses, where each section lists the number of students who answered male/female and yes/no to choosing CPP as their first college. Next page lists the table of both questions.

Q#01 and Q#03 Valid Responses:

Q#01/Q#03	a (yes)	b (no)	Total (Q#03)
a (male)	323	288	611
b (female)	326	305	631
Total (Q#01)	649	593	1242

A TOI test was conducted in the following hypothesizes and p-value:

TOI test for Q#01 and Q#03:

H_0: Q#01 and Q#03 are independent (gender and if CPP was students first choice, are not related)

H_A: Q#01 and Q#03 are not independent (gender and if CPP was students first choice, are related)

$$X^2 = 0.17921, \text{ df} = (r-1)(c-1) = (2-1)(2-1) = 1*1 = 1, \text{ P-value} = 0.6721$$

Since the p-value is greater than α = 0.05, this fails the to reject H_0 . The selected students' gender and whether or not Cal Poly Pomona was their first choice are not related.

A TOI test between Q#01 and Q#04 was conducted. To test for independence between the two, a table was constructed between the two questions. Q#01 had 2 responses while Q#04 had 5: a (closeness to home), b (someone's recommendation), c (beautiful campus), d (a school of good academic reputation, and e (others). Since Q#04 had 1247 valid responses, a 2 x 5 table format was created with these responses.

TOI test between Q#01 and Q#04:

Q#01/Q#04	a	b	c	d	e	Total (Q#04)
a (male)	184	80	26	206	119	615
b (female)	185	89	41	128	189	632
Total (Q#01)	369	169	67	334	308	1247

H_0: There is no difference between a student's reason among gender

H_a: There is a difference for at least one choice.

$$X^2\text{: } 37.74, \text{ df} = (r-1)(c-1) = (2-1)(5-1) = 1(4) = 4, \text{ p-value} = 1.268e\text{-}07$$

Given the p-value is less than the alpha value of 0.05, the null hypothesis is rejected. We can conclude that for at least one reason there is a significant difference between males and females.

A TOI test was then conducted for Q#03 and Q#04. We compared if the proportions of students who selected CPP based on reputation to students who selected Cal Poly as their first choice was larger than 30%. A R prop.test() function was used for this analysis.

Z test between Q#03 and Q#04:

Q#03/Q#04	a	b	c	d	e	Total (Q#04)
a	204	7	23	209	125	568
b	163	80	44	125	180	592
Total (Q#01)	367	87	67	334	305	1160

H_0: The difference between these two stated groups is 30%.
H_a: The difference between the groups is greater than 30%

$$X^2 = 469.35, \text{ df} = 1, \text{ p-value} = 2.2e\text{-}16$$

Given the displayed p-value is significantly less than our alpha value of .05 we summarily reject the null hypothesis and conclude that the gap between these two proportions is larger than 30%. Given the sample estimate, the difference is approximately 50%.

Stage 3:

Stage 3a: (Q#02 and Q#12)

For this stage, we are looking at if the two questions, Q#02 (What is your class standing?) and Q#12 (How many times a week do you attend CPP?) are related. To find this relationship, a table was created based on the 5 responses of Q#02 (a ~ e) and 6 responses (a ~ f). Q#02 responses are a (Fr. = Freshmen), b (So. = Sophomore), c (Jr. = Junior), d (Sr. = Senior), and e (Others). Q#12 has the responses: a (1), b (2), c (3), d (4), e (5), f (Others). Q#02 didn't have any NA responses, but Q#12 has 7 NA responses. A 5 x 6 table format was created based on the 1243 valid responses and TOI tests was conducted.

Q#02 and Q#12 Valid Responses:

Q#02/Q#12	a	b	c	d	e	f	Total (Q#12)
a	12	13	104	89	214	32	464
b	6	7	112	63	90	10	288
c	7	14	49	66	111	3	250
d	15	9	33	53	86	12	208
e	2	4	3	6	16	2	33
Total (Q#02)	42	47	301	277	517	59	1243

TOI Test for Q#02 and Q#12:

H_0: Q#02 and Q#12 are independent vs H_a: Q#02 and Q#12 are not independent

$$X^2 = 89.194, \text{ df} = (r-1)(c-1) = (5-1)(6-1) = 4*5 = 20, p= 1.025e-10$$

Since the p-value = 1.025e-10 is less than the alpha value, we reject the null hypothesis. We can conclude that questions: Q#02 and Q#12 are not independent.

Another test was conducted where we removed the "Others" category from both Q#02 and Q#12. This reduces the table format to a 4x6 format where only 1153 samples were recorded.

Another TOI test was conducted and calculates a p-value of 3.076e-09. Since this p-value (3.076e-09) is less than the 0.05 level, we reject the null hypothesis and conclude that Q#02 and Q#12 are not independent without the "others" category. The 4x6 table and the TOI test of this new table is listed on the next page where r = 4 and c = 6.

TOI Test for Q#02 and Q#12 without "Others":

Q#02/Q#12	a	b	c	d	e	Total (Q#12)
a	12	13	104	89	214	432
b	6	7	112	63	90	278
c	7	14	49	66	111	247
d	15	9	33	53	86	196
Total (Q#02)	40	43	298	271	501	1153

H_0: Q#02 and Q#12 are independent vs H_a: Q#02 and Q#12 are not independent

$$X^2 = 64.716, \text{ df} = (r-1)(c-1) = (4-1)(5-1) = 3(4) = 12, \text{ p-value} = 3.076e-09$$

For the Q02 and Q12 since they are an ordinal scale and not nominal scale, we replaced the following a, b, c, d, and e with the corresponding values of 1, 2, 3, 4, and 5. We wanted to look at the correlation between the two questions. The table for the new nominal scale between the two questions is listed below but the Pearson's correlation coefficient (r) is -0.03953358. This indicates a very weak relationship between the two questions since this coefficient is near zero. To test the validity of this correlation, a correlation test was conducted with the results on the next page.

Q#02 and Q#12 In Nominal Scale:

Q#02/Q#12	1	2	3	4	5	Total (Q#12)
1	12	13	104	89	214	432
2	6	7	112	63	90	278
3	7	14	49	66	111	247
4	15	9	33	53	86	196
Total (Q#02)	40	43	298	271	501	1153

Correlation test for Q#02 and Q#12:

$$H_0: \textit{The correlation coefficient is } 0 \ (\textit{nocorrelation, independence})$$

$$H_a: \textit{The correlation coefficient is not } 0 \ (\textit{there is a correlation})$$

T = -1.3423, df = n – 2 = 1153 – 2 = 1151, p-value = 0.1798

Since the p-value = 0.1798, this is greater than the alpha level, so we don't reject the null hypothesis. There is not enough evidence to say that there is a correlation between the two questions, and say they are independent.

Stage 3b: (Q#07, Q#07a, Q#07b)

For this stage, we tested Q#07 of the survey. We did a Z-test on whether the proportion of students who use email stations at schools were significantly greater than the proportion of students who do not use email stations at schools. A R prop.test () function was utilized. For this analysis, there were 701 who used email vs. 535 who did not. 14 NA responses would not be included in this report and only 1236 responses were analyzed

Z-test for Q#07 on proportion of students using emails:

Q#07	yes (a)	no (b)	Total
Frequency	701	535	1236

H_0: The proportions of students who used email stations are the same as the proportion of students who do not use email at school.

H_a: The proportions of these two groups are not the same.

X^2: 22.294, df = 1; p-value = 1.169e-06

Given the displayed p-value is less than our alpha value of 0.05, we reject the null hypothesis of equality and conclude there is a statistically significant difference among the two student groups who use the campus emails and those who don't.

We would have done an analysis on Q#07a (How often? Answer this if you chose "yes" above) and Q#07b (At what location do you most often use them? Answer this if you chose "yes" above), but these are inaccurate. 14 respondents didn't answer the question and that raises some problems with further analysis when doing these tests. Q#07a has 9 NA respondents and Q#07b had 8 NA respondents. This would raise problems with testing because we don't know who of the 14 answered "yes" to Q#07 and the same when continuing testing for Q#07a and Q#07b.

Conclusion:

The analysis conducted across multiple surveys among the CPP student population provides valuable insights into the distributions and relationships among students' responses. The results of the Goodness-of-Fit (GOF) tests, Test of Independence (TOI) tests, and Z-tests confirm or reject various hypotheses regarding the demographic and preference-related factors at Cal Poly Pomona.

For Q#01, the GOF test for the binomial distribution suggest that the student gender distribution follows an approximately equal distribution for male and female students. The p-values from the GOF tests were not significant, indicating that we do not reject the null hypothesis, and the assumption of a binomial distribution remains reasonable. Looking at Q#04, the GOF test shows us with a significant p-value, which makes us reject the null hypothesis, that there is not an equal proportion for the students' reason of choosing the college. This means that at least one of the categories is not 20 percent of proportion.

For Q#13, we want to see if the percentage of why the campus stands out is because of the facility, the education, the campus, and it does not stand out with the respective percentages (10, 30, 30, 30). GOF test was used and had fail to reject null, which tells us that the percentage that we had given was reasonable. For the second part we want to combine the facility, the education, and the campus to be "Wow". We want to see if this "Wow" is more than 2/3 of the students' reason for choosing the campus. The null was also rejected, which means there is evidence to support that 2/3 of the students surveyed was because of "Wow".

Tests between the 3 questions: Q#01 (Are you male or female?), Q#03 (What makes CPP stand out from other universities?), and Q#04 (Was CPP your first choice for college?) were conducted. Between Q#01 and Q#04, a TOI test reveals the students' gender and the chose of college was not related. For the relationship between Q#01 and Q#03, there is evidence to support that males and females have different reasons choosing CPP. The last thing, we want to look at is for if CPP was the first choose and why it stands out. A TOI test reveals there is evidence that 30% of people think that the school has good academic reputation as their first school

The relationship between Q#02 (What is your class standing?) and Q#12 (How many times a week do you attend CPP?) reveals that these questions are independent. A TOI test returns with us to conclude that the two questions are not independent, but another TOI was used reveals that there is evidence where they are not independent. Something to note is that since Q02 and Q12 are in ordinal scale we can give the responses a scale where a, b, c, d, and e is with the corresponding values of 1, 2, 3, 4, and 5. A calculated Pearson's correlation coefficient (r) indicates a very weak relationship of -0.03953358. The relationship between the two questions is really low if non-existent since this coefficient is near zero which was confirmed by the correlation test that there is no correlation, and they are independent.

The last thing we look at is Q#07 (Do you use the e-mail stations at school?), Q#7a (How often? Answer this if you chose "yes" above), and #Q7b (At what location do you most often use them? Answer this if you chose "yes" above) The first thing we want to look at is if the amount of people that us e-mail station is greater than those who do not. We use a z-test for this and reject the null hypothesis and see that there is evidence that more people use the e-mail station then those who don't. As for the responses to Q#07a and Q#07b we should not use the used in further analysis since we the answers are invalid. This is because we had people answer 7a or 7b after answering they did not use the e-mail stations which should not happen.

Appendices:

Q#01 and Q#02:

```
# Q1, let p(i) be the probablity of the sudent in genter belonging to i = a,b

# 1.1) test on binomoial distribtion with p(i) = 0.50 for
#  number of student in 2 categories

# count male and female
male_count <- sum(q01 == "a", na.rm = TRUE) # na.rm = remove NA
female_count <- sum(q01 == "b", na.rm = TRUE) # na.rm = remove NA

male_count                          # 615
female_count                        # 635
sum = male_count + female_count     # 1250

# probablities for p(a) and p(b)
prob <- c(0.50, 0.50)

# store counts
counts <- c(male_count, female_count)

# chisq test for counts based on probablities
chisq.test(counts, p = prob, correct = FALSE)

# 1.2) test wheter prop of female is greater than 0.50
# prop.test( want, total, probablity, alternative = "(less,greater)", correct = False)

prop.test(female_count, sum, p = 0.5, alternative = "greater", correct = FALSE)
```

```
> chisq.test(counts, p = prob, correct = FALSE)

        Chi-squared test for given probabilities

data:  counts
X-squared = 0.32, df = 1, p-value = 0.5716
```

```
> prop.test(female_count, sum, p = 0.5, alternative = "greater", correct = FALSE)

        1-sample proportions test without continuity correction

data:  female_count out of sum, null probability 0.5
X-squared = 0.32, df = 1, p-value = 0.2858
alternative hypothesis: true p is greater than 0.5
95 percent confidence interval:
 0.4847491 1.0000000
sample estimates:
    p
0.508
```

```
# 2) Based on Q04, let p(i) [i = a,b, ... ,e] be the probability why a student
# came to CPP belong to i. Test binomial distribution with p(i) = 0.20

close2home <- sum(q04 == "a", na.rm = TRUE) # na.rm = remove NA
Some_Rec <- sum(q04 == "b", na.rm = TRUE) # na.rm = remove NA
Beaut_Camp <- sum(q04 == "c", na.rm = TRUE) # na.rm = remove NA
Good_Rep <- sum(q04 == "d", na.rm = TRUE) # na.rm = remove NA
other <- sum(q04 == "e", na.rm = TRUE) # na.rm = remove NA

close2home      # 369
Some_Rec        # 169
Beaut_Camp      # 67
Good_Rep        # 334
other           # 308

# sum of questions
sum2 <- close2home + Some_Rec + Beaut_Camp + Good_Rep+ other
sum2           # 1247

# put counts in table
count2 <- c(close2home, Some_Rec, Beaut_Camp, Good_Rep, other)

# probabilities
prob2 <- c(0.20, 0.20, 0.20, 0.20, 0.20)

# test
chisq.test(count2, p = prob2, correct = FALSE)
```

```
> chisq.test(count2, p = prob2, correct = FALSE)

        Chi-squared test for given probabilities

data:  count2
X-squared = 259.14, df = 4, p-value < 2.2e-16
```

Q#03:

```
# Q03

#3-1

#from table
observed = c(115, 409,351,371)

expected = c(.1,.3,.3,.3)

chisq.test(observed, expected, correct=FALSE)

#3-2

library(dplyr)

#creating WOW category

merged <- survey %>%
  mutate(q13 = case_when(
    q13 == "a" ~ "WOW",
    q13 == "b" ~ "WOW",
    q13 == "c" ~ "WOW",
    TRUE ~ q13  # Keeps other values unchanged
  ))

WOW_table <- table(merged$q13) %>% sort(decreasing = TRUE)
WOW_table

prop.test(875, 1246, p = 0.67, alternative = "greater", correct = FALSE)

>     WOW_table

WOW   d
875 371
>     prop.test(875,1246, p=.67,alternative= "greater",correct=FALSE)

        1-sample proportions test without continuity correction

data:  875 out of 1246, null probability 0.67
X-squared = 5.8602, df = 1, p-value = 0.007743
alternative hypothesis: true p is greater than 0.67
95 percent confidence interval:
 0.6805196 1.0000000
sample estimates:
        p
0.7022472
```

```
#4

tableQ4 <- table(survey$q01,survey$q03)

tableQ4



chisq.test(tableQ4,correct=FALSE)
```

```
> chisq.test(tableQ4,correct=FALSE)

        Pearson's Chi-squared test

data:  tableQ4
X-squared = 0.17921, df = 1, p-value = 0.6721
```

Q#04 (above) and Q#05 (below)

```
#homework 5

hm1extract <- select(clean,q01,q04)

hm2extract <- hm1extract %>% filter(q01 != 'NA')
hm3extract <- hm2extract %>% filter(q04 != 'NA')

contingency_table_hm <- table(hm3extract$q01, hm3extract$q04)

contingency_table_hm

chi_sq_hm1 <- chisq.test(contingency_table_hm)

chi_sq_hm1
```

```
> contingency_table_hm

     a   b   c   d   e
  a 184  80  26 206 119
  b 185  89  41 128 189
> chi_sq_hm1 <- chisq.test(contingency_table_hm)
> chi_sq_hm1

        Pearson's Chi-squared test

data:  contingency_table_hm
X-squared = 37.74, df = 4, p-value = 1.268e-07
```

Q#06:

```
#homework 6

extract_both <- select(clean,q03,q04)

# set-up first choice

#note table command does not count #na's and sums the other variables

fc_total <- table(extract_both$q03)

# a - cal poly first choice 649

#b - cal poly second choice 593

fc_col <- cbind(fc_total)

#set up reputation

q04total <- select(extract_both,q04)

rep_total <- q04total %>% filter(q04=='d')

#334
rep_sum = nrow(rep_total)

rep_other <- q04total %>% filter (q04!='NA')

rep_other_sum = nrow(rep_other)

#913

rep_diff = rep_other_sum - rep_sum

rep_col <- cbind(rep_other_sum, rep_diff)

t(rep_col)

result <- prop.test(x=c(649+593), n=c(649+334+593+913), p=.30,alternative= "greater",correct=FALSE )

print(result)
```

```
>   t(rep_col)
             [,1]
rep_other_sum 1247
rep_diff       913
>   result <- prop.test(x=c(649+593), n=c(649+334+593+913), p=.30,alternative= "greater",correct=FALSE )
>   print(result)

        1-sample proportions test without continuity correction

data:  c(649 + 593) out of c(649 + 334 + 593 + 913), null probability 0.3
X-squared = 469.35, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is greater than 0.3
95 percent confidence interval:
 0.4825208 1.0000000
sample estimates:
        p
0.4989956
```

Q#07:

```
#Q7)

# Based on Q#02 (What is your class standing?) and Q#12 (How many times a week do you attend CPP?), do
# the following

#7-1. Test whether the two variables are independent.

# data need are Q2 class standing and Q12 how many time do you attend CPP

temp_table = table(survey_data$q02, survey_data$q12)
temp_table
heatmap(temp_table,  scale = "row")
chisq.test(temp_table, correct = FALSE)


#7-2. Now create the dataset that has entries

# (a) ~ (d) only for Q#02 and

# (a) ~ (e) only for Q#12. Test whether the two variables are related or not.

cleaned_temp = survey_data

print("-----------------------------------------------")
unique(cleaned_temp$q02)
cleaned_temp$q02 = ifelse(cleaned_temp$q02 %in% c("e"), NA, cleaned_temp$q02)
print("Cleaned")
unique(cleaned_temp$q02)


print("-----------------------------------------------")
unique(cleaned_temp$q02)
cleaned_temp$q12 = ifelse(cleaned_temp$q12 %in% c("f"), NA, cleaned_temp$q12)
print("Cleaned")
unique(cleaned_temp$q12)

temp_table2 = table(cleaned_temp$q02, cleaned_temp$q12)
temp_table2

chisq.test(temp_table2, correct = FALSE)
```

```
> temp_table

     a   b   c   d   e   f
  a 12  13 104  89 214  32
  b  6   7 112  63  90  10
  c  7  14  49  66 111   3
  d 15   9  33  53  86  12
  e  2   4   3   6  16   2
> heatmap(temp_table,  scale = "row")
> chisq.test(temp_table, correct = FALSE)

        Pearson's Chi-squared test

data:  temp_table
X-squared = 89.194, df = 20, p-value = 1.025e-10
```

```
> temp_table2

     a   b   c   d   e
  a 12  13 104  89 214
  b  6   7 112  63  90
  c  7  14  49  66 111
  d 15   9  33  53  86
> chisq.test(temp_table2, correct = FALSE)

        Pearson's Chi-squared test

data:  temp_table2
X-squared = 64.716, df = 12, p-value = 3.076e-09
```

```r
#7-3. For this part, you will re-create the dataset. If you see the entries for Q#02 (class standing)
#   and those for Q#12 (number of times a student attends) from the dataset of C12-2 above,
#   both are in the 'ordinal' measurement scale rather than 'nominal' scale. For the class standing,
#   assign 1 for "(a)", 2 for "(b)", 3 for "(c)", and 4 for "(d)".
# For the number of times a student attends, assign 1 for "(a)",
#   2 for "(b)", 3 for "(c)", 4 for "(d)", and 5 for "(e)". Using the newly created dataset,
# compute the (Pearson's) correlation coefficient between the two variables.
# Interpret the value of the correlation coefficient: strength and direction
#   in the context of class standing and number of times a student attends.

hash_map <- c("a" = 1, "b" = 2, "c" = 3, "d"= 4, "e"= 5)
new_values_02 <- hash_map[cleaned_temp$q02]

new_values_12 <- hash_map[cleaned_temp$q12]

#cleaned_temp$q02
unique(new_values_02)
unique(new_values_12)

temp_table2 = table(new_values_02, new_values_12)
temp_table2
cor(new_values_02, new_values_12, use = "complete.obs", method = "pearson")


#7-4. Test whether the correlation coefficient is significantly different from zero.
# That is to test H0 : the correlation coefficient is equal to zero (i.e., independence).

# Null Hypothesis (H₀): The correlation coefficient (ρ) is 0 (no correlation, independence).

# Alternative Hypothesis (H₁): The correlation coefficient (ρ) is not 0 (there is a correlation).

cor_test <- cor.test(new_values_02, new_values_12, method = "pearson")
cor_test
```

```
> temp_table2
             new_values_12
new_values_02   1   2   3   4   5
            1  12  13 104  89 214
            2   6   7 112  63  90
            3   7  14  49  66 111
            4  15   9  33  53  86
> cor(new_values_02, new_values_12, use = "complete.obs", method = "pearson")
[1] -0.03953358
> cor_test <- cor.test(new_values_02, new_values_12, method = "pearson")
> cor_test

        Pearson's product-moment correlation

data:  new_values_02 and new_values_12
t = -1.3423, df = 1151, p-value = 0.1798
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.09704401  0.01823997
sample estimates:
        cor
-0.03953358
```

Q#08:

```
# homework 8

extract_a <- select(cpp_data,q07)

extract_b <- extract_a %>% filter(q07== 'a')

extract_c <- extract_a %>% filter(q07== 'b')

701

yes_email <- nrow(extract_b)

no_email <- nrow(extract_c) # 535

#columns minus na's

col_q07<- select(cpp_data, q07) col_q7_noNA<- col_q07 %>% filter(q07!='NA') total_colq7 <- nrow(col_q7_noNA)

result1 <- prop.test(yes_email, total_colq7, alternative='greater', correct=FALSE)

print(result1)
```

```
> print(result1)

        1-sample proportions test without continuity correction

data:  yes_email out of total_colq7, null probability 0.5
X-squared = 22.294, df = 1, p-value = 1.169e-06
alternative hypothesis: true p is greater than 0.5
95 percent confidence interval:
 0.5438491 1.0000000
sample estimates:
        p
0.5671521
```