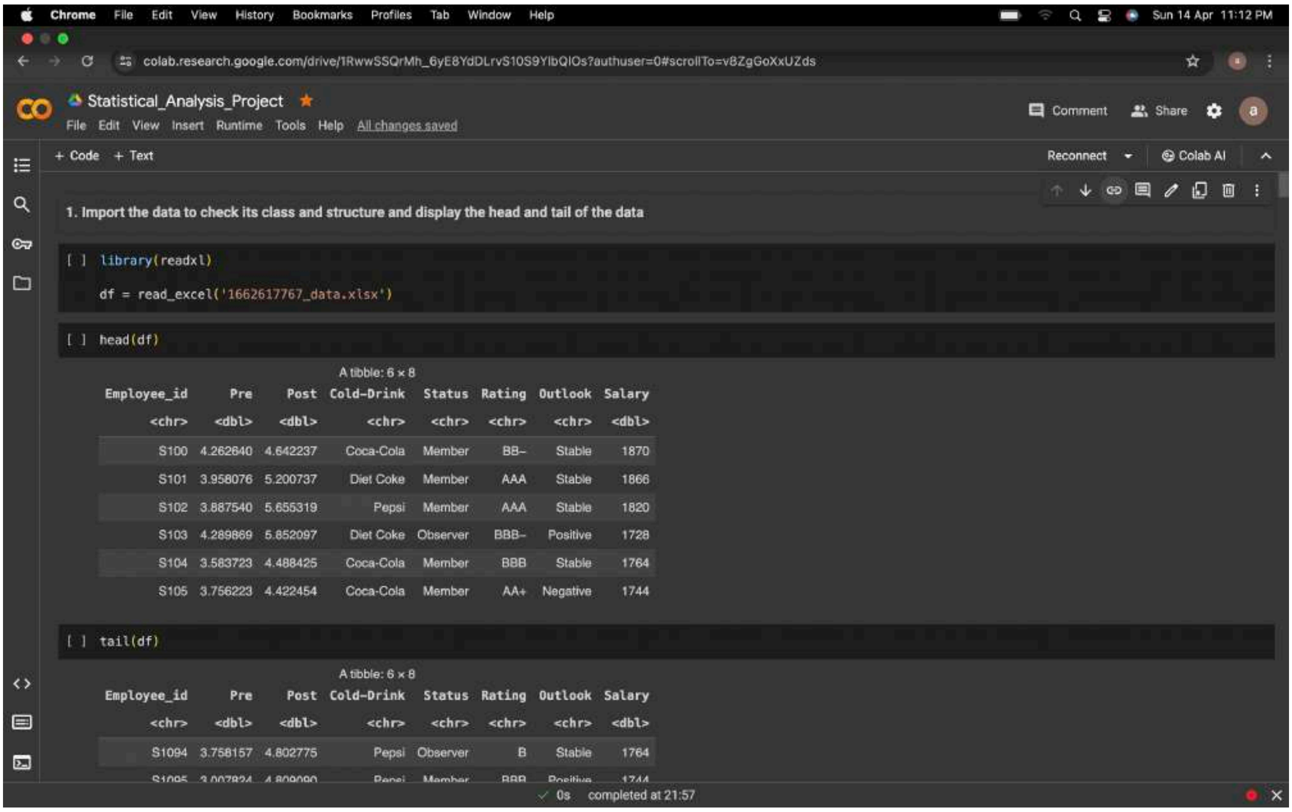
1. Import the data to check its class and structure and display the head and tail of the data



1. Import the data to check its class and structure and display the head and tail of the data

```
library(readxl)

df = read_excel('1662617767_data.xlsx')
```

```
head(df)
```

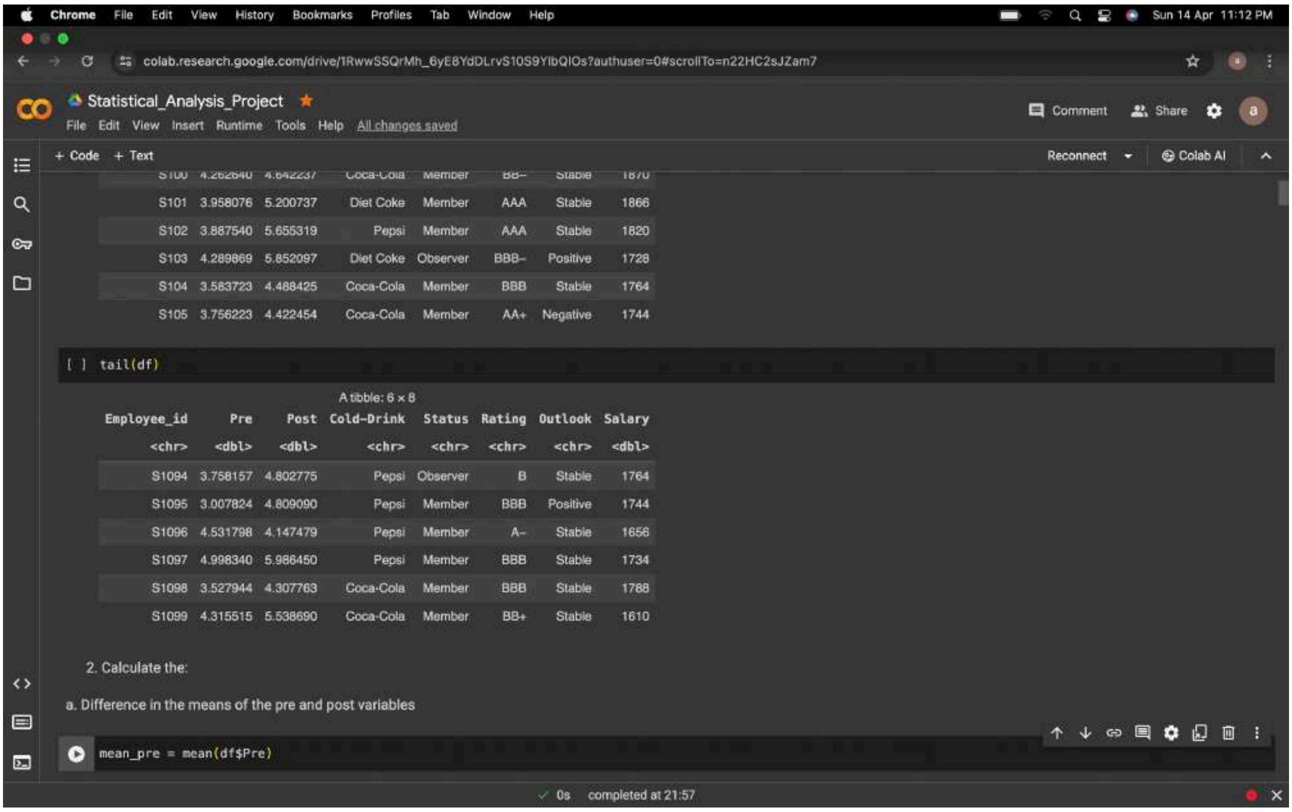
A tibble: 6 × 8

Employee_id	Pre	Post	Cold-Drink	Status	Rating	Outlook	Salary
<chr>	<dbl>	<dbl>	<chr>	<chr>	<chr>	<chr>	<dbl>
S100	4.262640	4.642237	Coca-Cola	Member	BB-	Stable	1870
S101	3.958076	5.200737	Diet Coke	Member	AAA	Stable	1866
S102	3.887540	5.655319	Pepsi	Member	AAA	Stable	1820
S103	4.289869	5.852097	Diet Coke	Observer	BBB-	Positive	1728
S104	3.583723	4.488425	Coca-Cola	Member	BBB	Stable	1764
S105	3.756223	4.422454	Coca-Cola	Member	AA+	Negative	1744

```
tail(df)
```

A tibble: 6 × 8

Employee_id	Pre	Post	Cold-Drink	Status	Rating	Outlook	Salary
<chr>	<dbl>	<dbl>	<chr>	<chr>	<chr>	<chr>	<dbl>
S1094	3.758157	4.802775	Pepsi	Observer	B	Stable	1764



S101	3.958076	5.200737	Diet Coke	Member	AAA	Stable	1866
S102	3.887540	5.655319	Pepsi	Member	AAA	Stable	1820
S103	4.289869	5.852097	Diet Coke	Observer	BBB-	Positive	1728
S104	3.583723	4.488425	Coca-Cola	Member	BBB	Stable	1764
S105	3.756223	4.422454	Coca-Cola	Member	AA+	Negative	1744

```
tail(df)
```

A tibble: 6 × 8

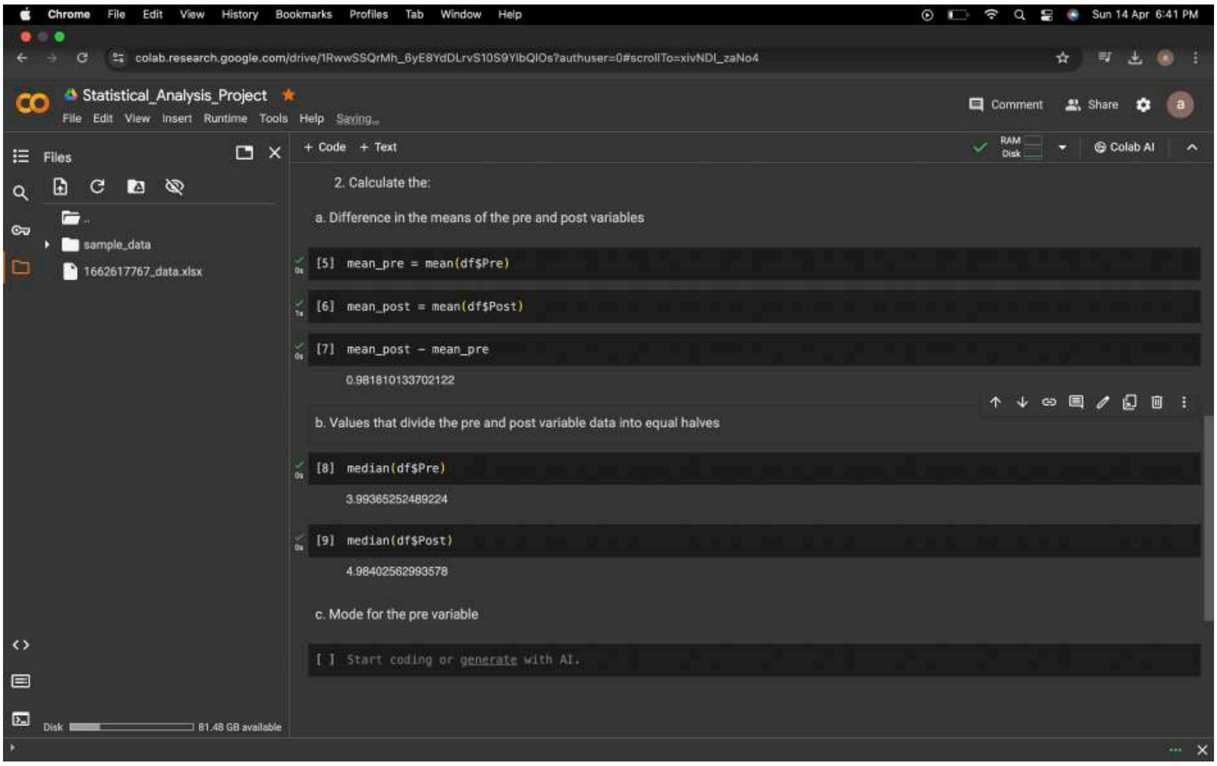
Employee_id	Pre	Post	Cold-Drink	Status	Rating	Outlook	Salary
<chr>	<dbl>	<dbl>	<chr>	<chr>	<chr>	<chr>	<dbl>
S1094	3.758157	4.802775	Pepsi	Observer	B	Stable	1764
S1095	3.007824	4.809090	Pepsi	Member	BBB	Positive	1744
S1096	4.531798	4.147479	Pepsi	Member	A-	Stable	1656
S1097	4.998340	5.986450	Pepsi	Member	BBB	Stable	1734
S1098	3.527944	4.307763	Coca-Cola	Member	BBB	Stable	1788
S1099	4.315515	5.538690	Coca-Cola	Member	BB+	Stable	1610

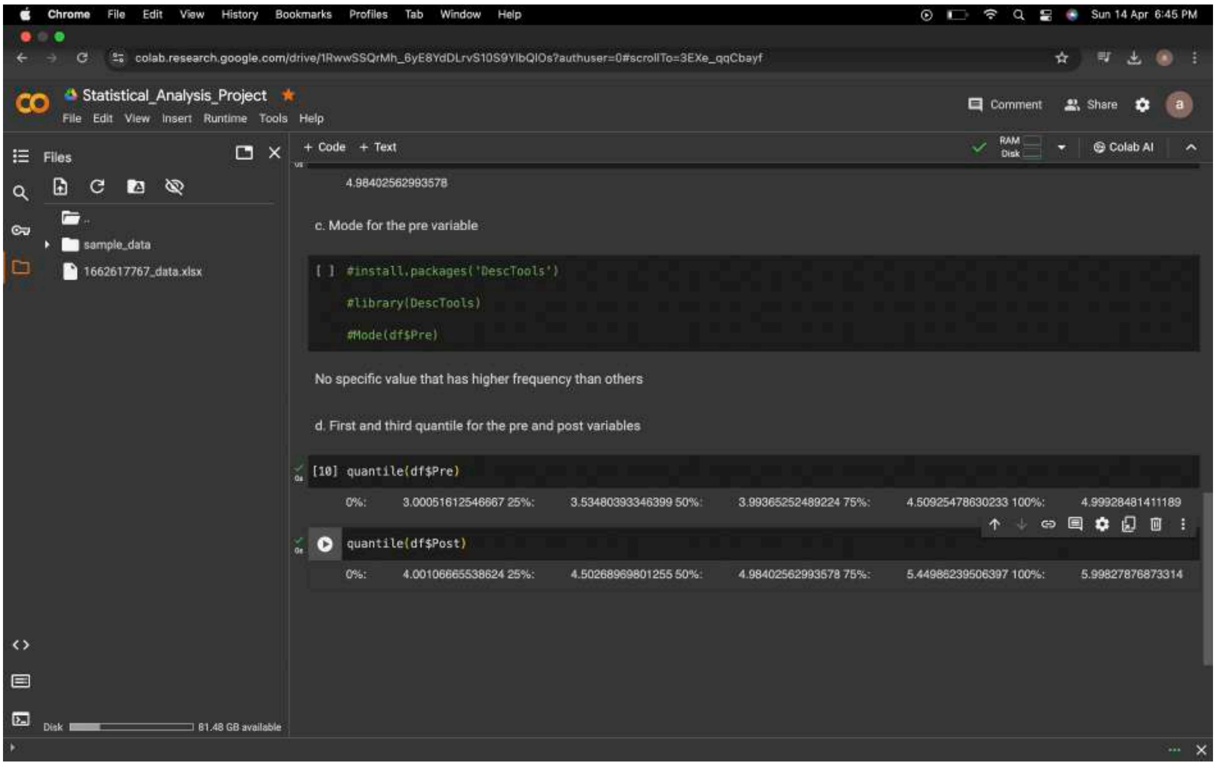
2. Calculate the:

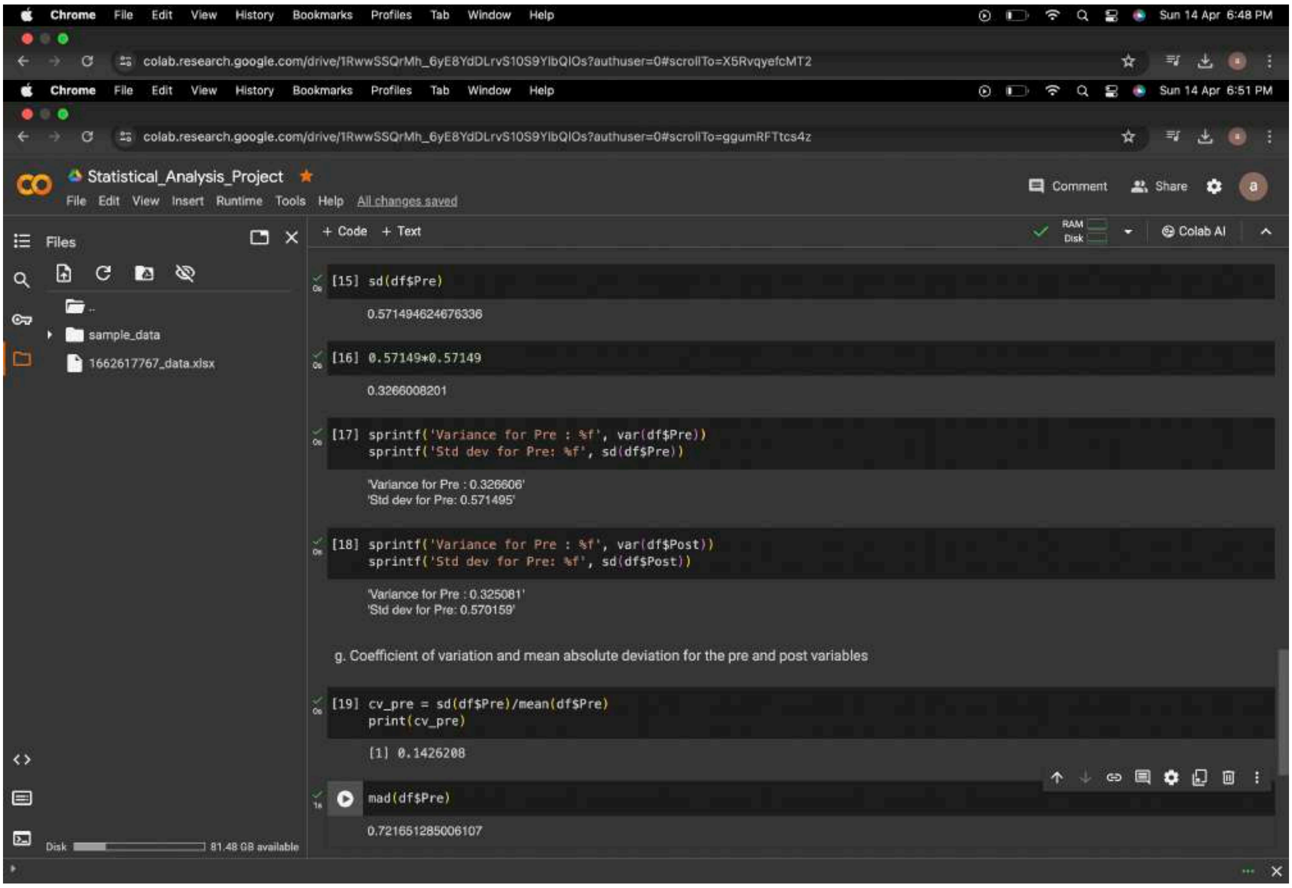
a. Difference in the means of the pre and post variables

```
mean_pre = mean(df$Pre)
```

2. Calculate the:
    a. Difference in the means of the pre and post variables
    b. Values that divide the pre and post variable data into equal halves
    c. Mode for the pre variable
    d. First and third quantile for the pre and post variables
    e. Range of the pre and post variables
    f. Variance and standard deviation for the pre and post variables
    g. Coefficient of variation and mean absolute deviation for the pre and post variables
    h. Interquartile range of the pre and post variables







```
[15] sd(df$Pre)
```

0.571494624676336

```
[16] 0.57149*0.57149
```

0.3266008201

```
[17] sprintf('Variance for Pre : %f', var(df$Pre))
     sprintf('Std dev for Pre: %f', sd(df$Pre))
```

'Variance for Pre : 0.326606'
'Std dev for Pre: 0.571495'

```
[18] sprintf('Variance for Pre : %f', var(df$Post))
     sprintf('Std dev for Pre: %f', sd(df$Post))
```

'Variance for Pre : 0.325081'
'Std dev for Pre: 0.570159'

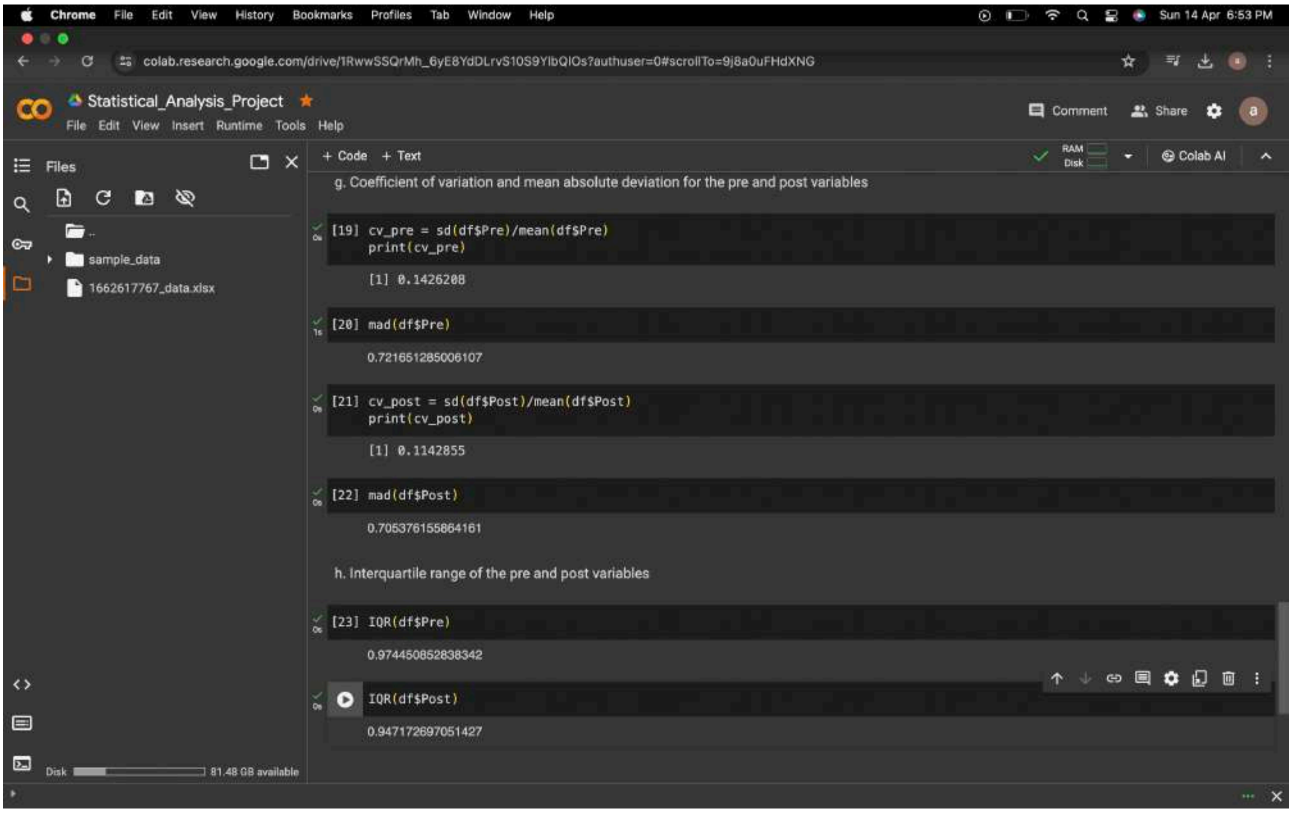
g. Coefficient of variation and mean absolute deviation for the pre and post variables

```
[19] cv_pre = sd(df$Pre)/mean(df$Pre)
     print(cv_pre)
```

```
[1] 0.1426208
```

```
mad(df$Pre)
```

0.721651285006107



g. Coefficient of variation and mean absolute deviation for the pre and post variables

```
[19] cv_pre = sd(df$Pre)/mean(df$Pre)
     print(cv_pre)
```

```
[1] 0.1426208
```

```
[20] mad(df$Pre)
```

0.721651285006107

```
[21] cv_post = sd(df$Post)/mean(df$Post)
     print(cv_post)
```

```
[1] 0.1142855
```

```
[22] mad(df$Post)
```

0.705376155864161

h. Interquartile range of the pre and post variables

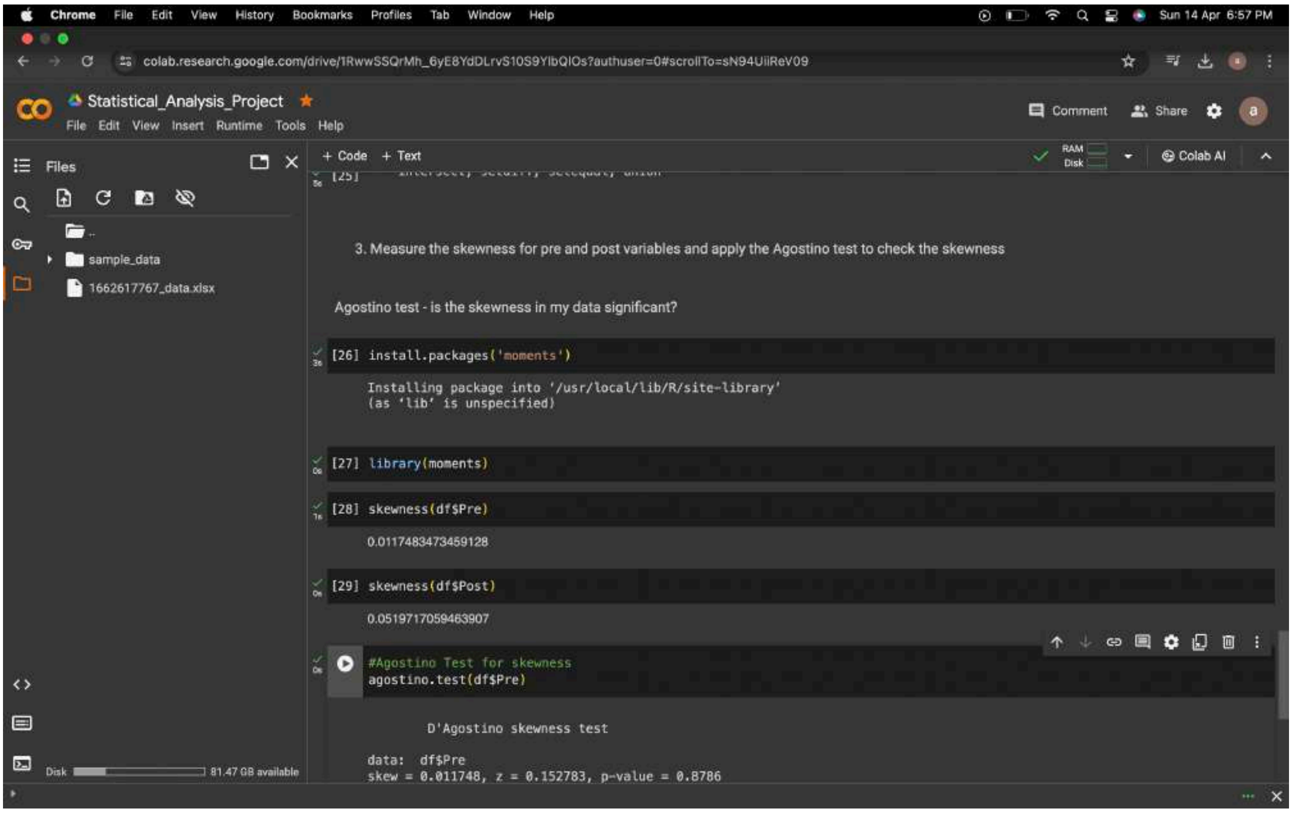
```
[23] IQR(df$Pre)
```

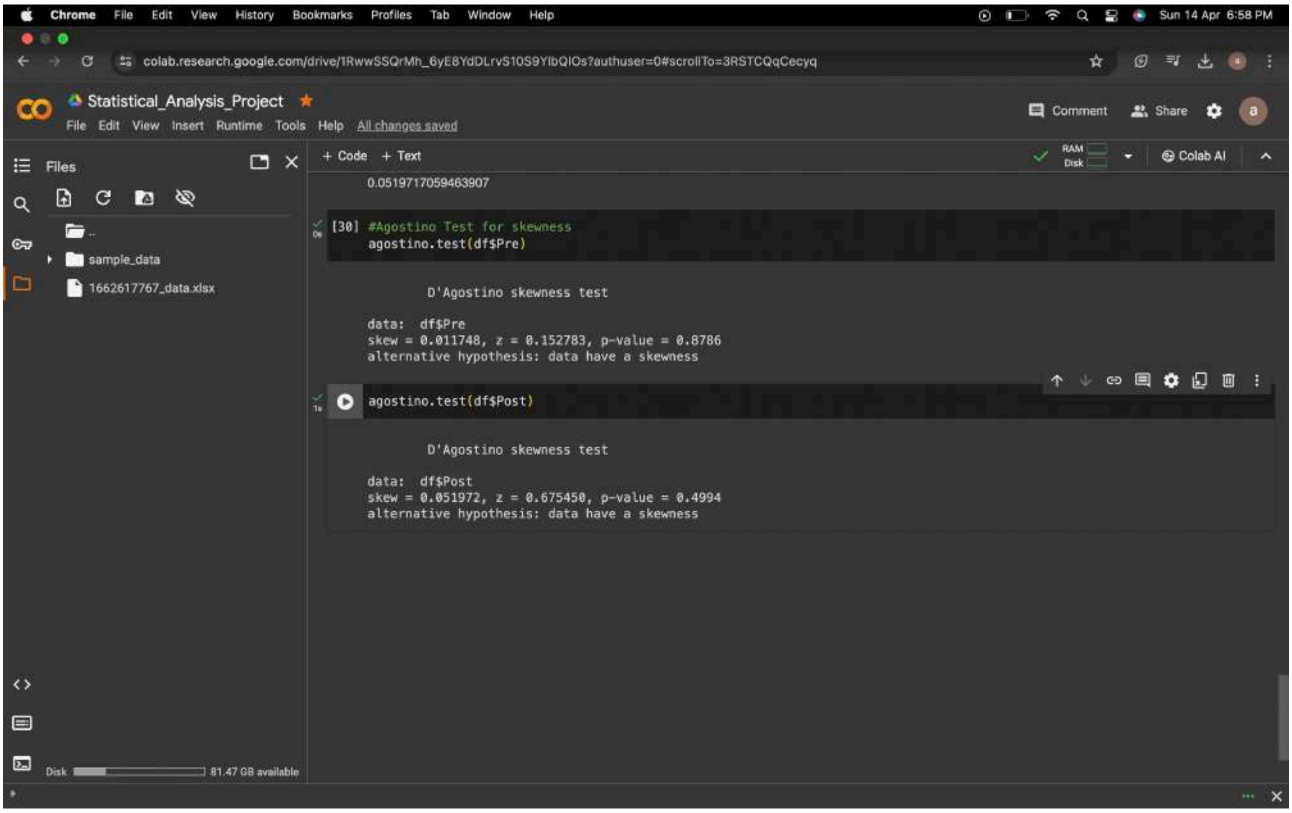
0.974450852838342

```
IQR(df$Post)
```

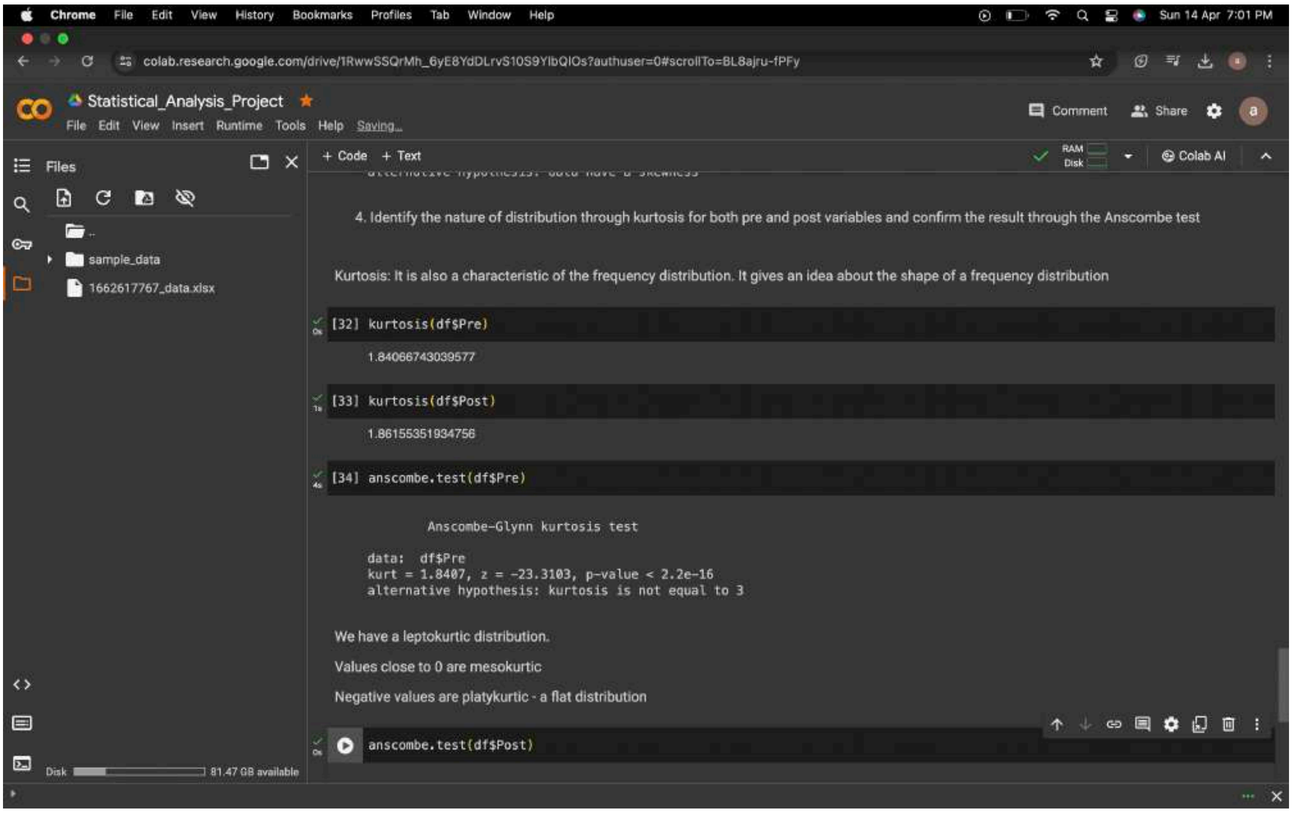
0.947172697051427

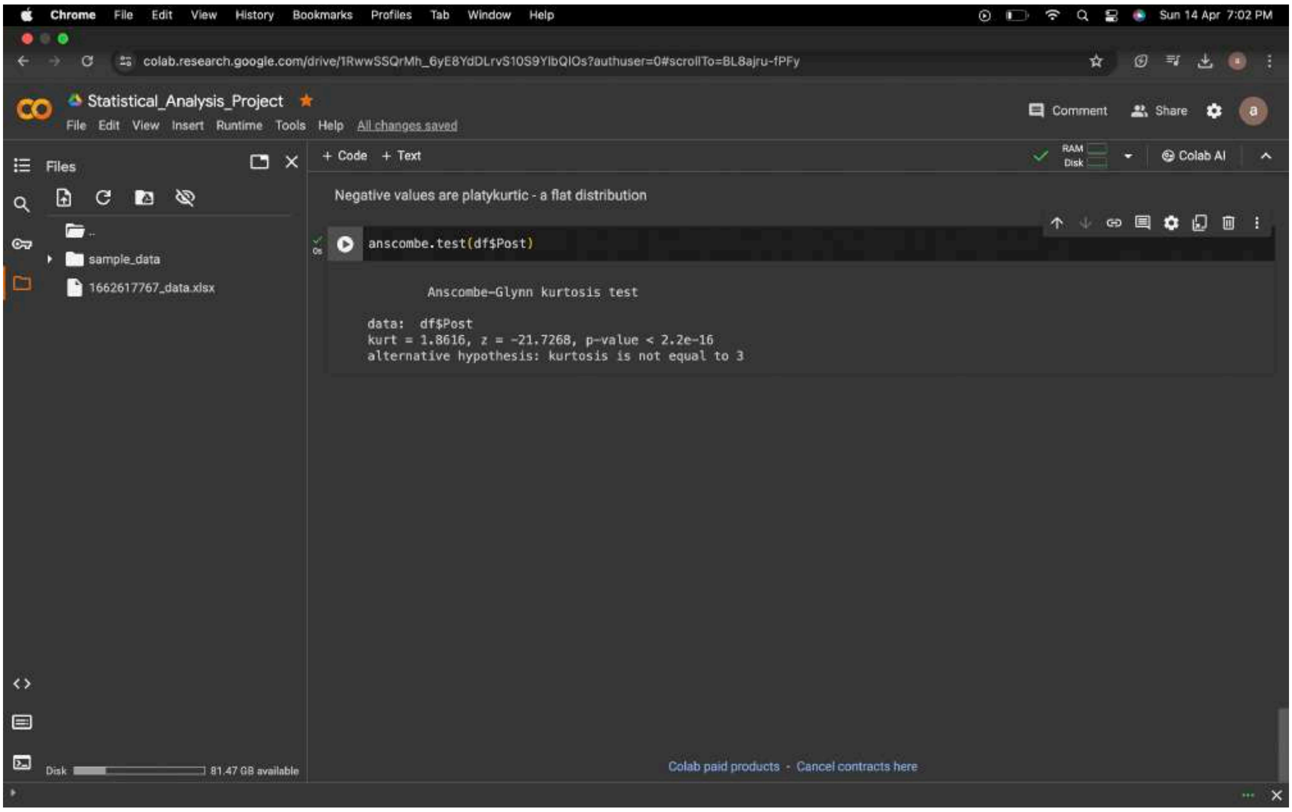
3. Measure the skewness for pre and post variables and apply the Agostino test to check the skewness



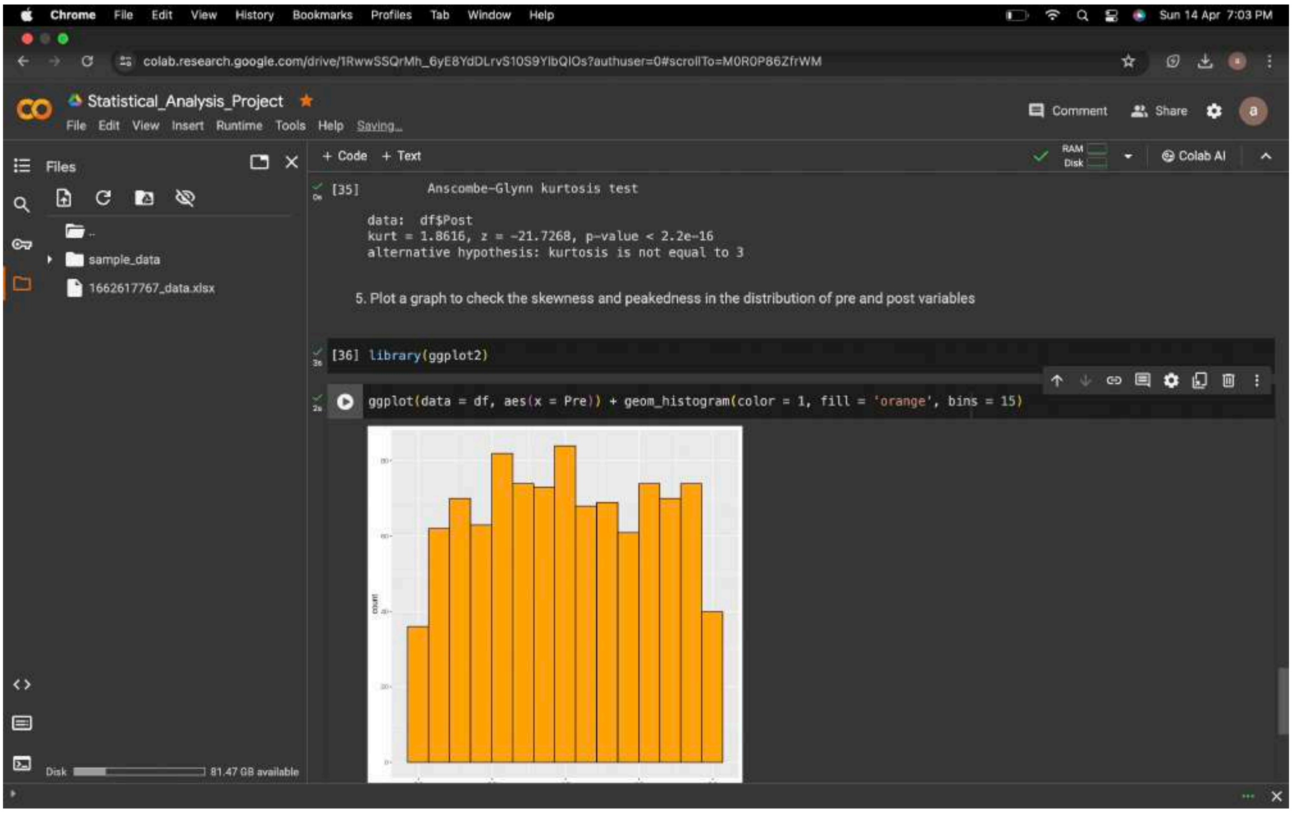


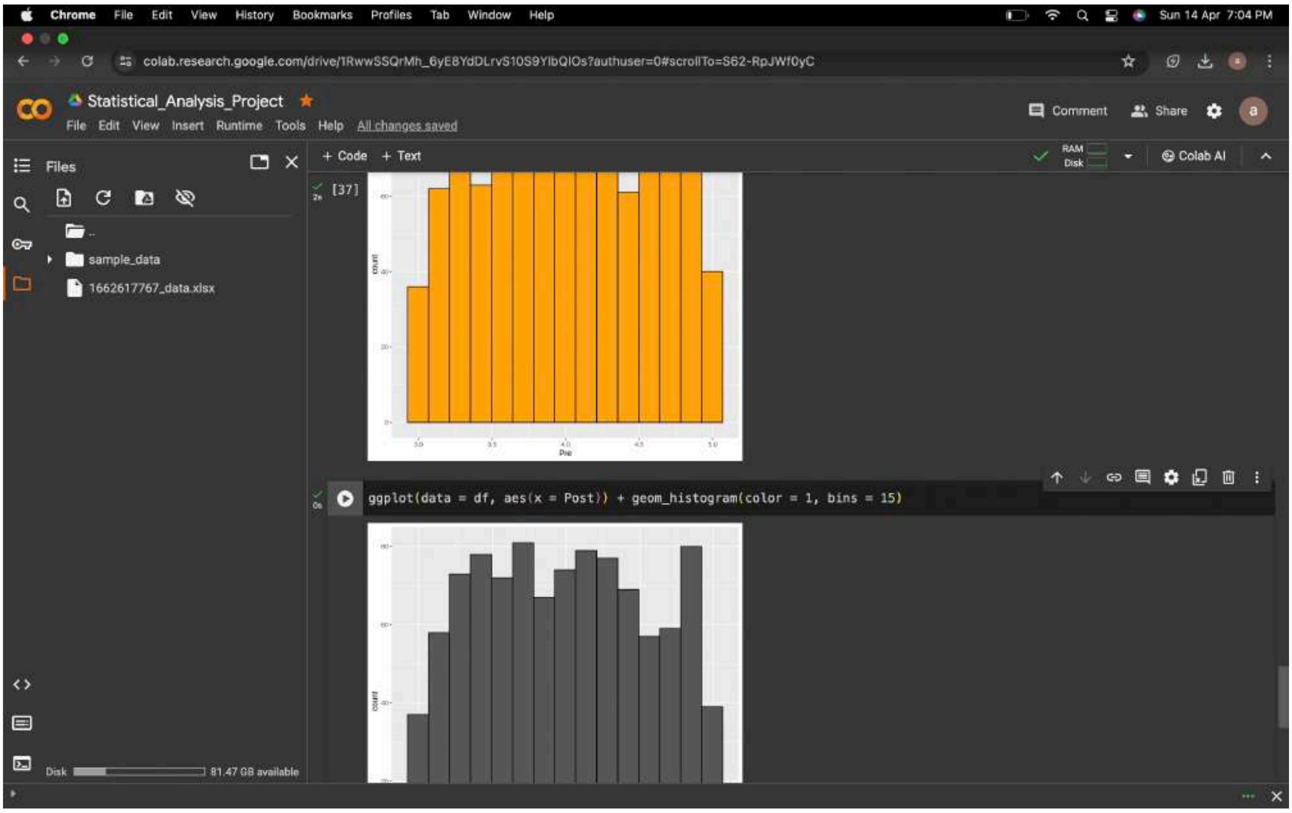
4. Identify the nature of distribution through kurtosis for both pre and post variables and confirm the result through the Anscombe test

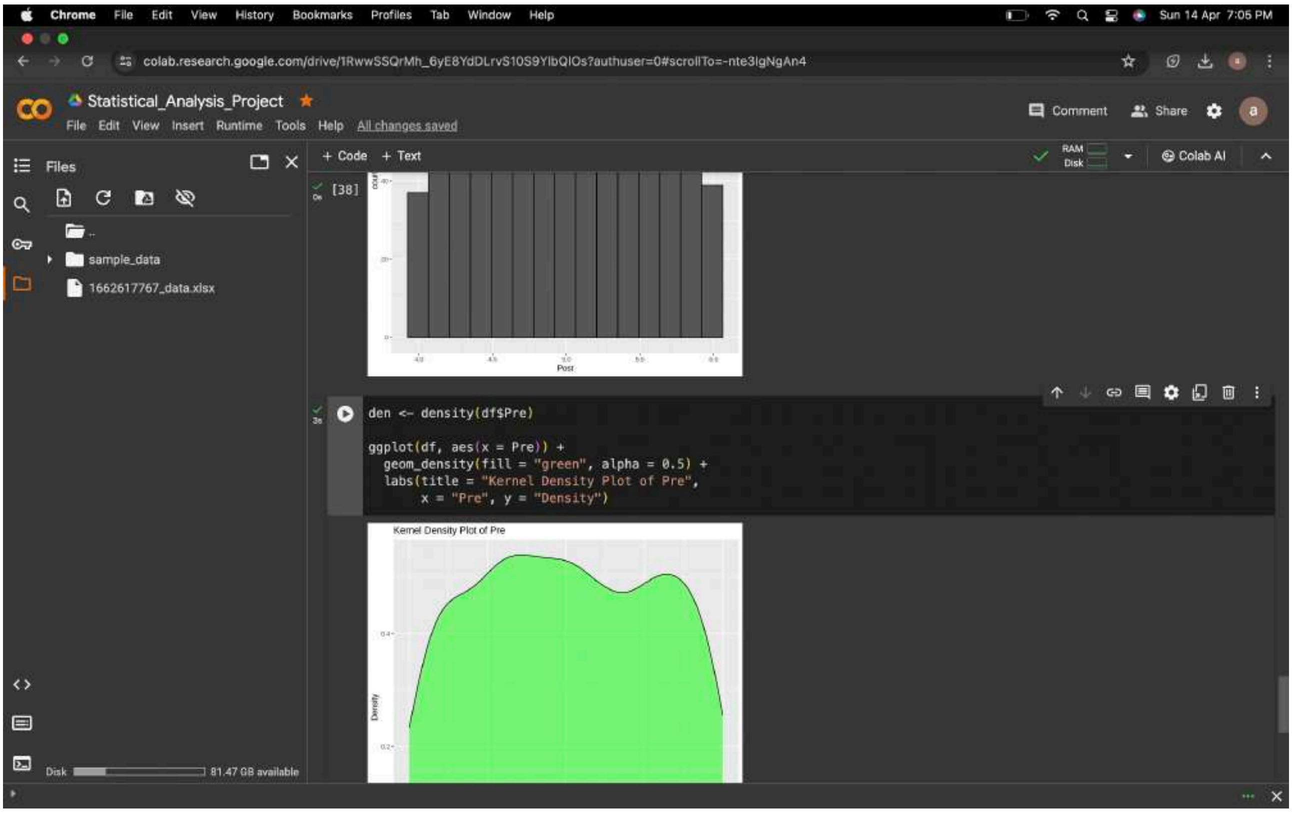


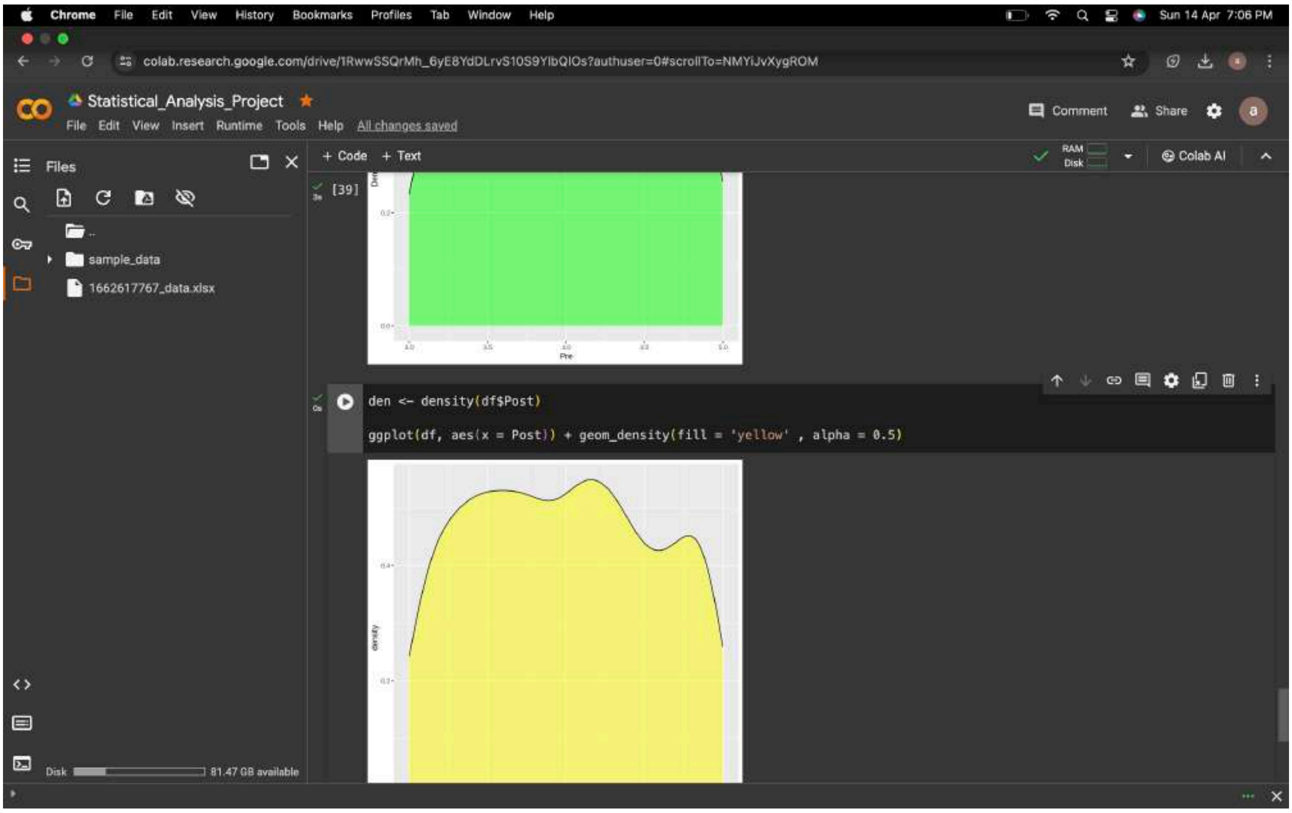


5. Plot a graph to check the skewness and peakedness in the distribution of pre and post variables

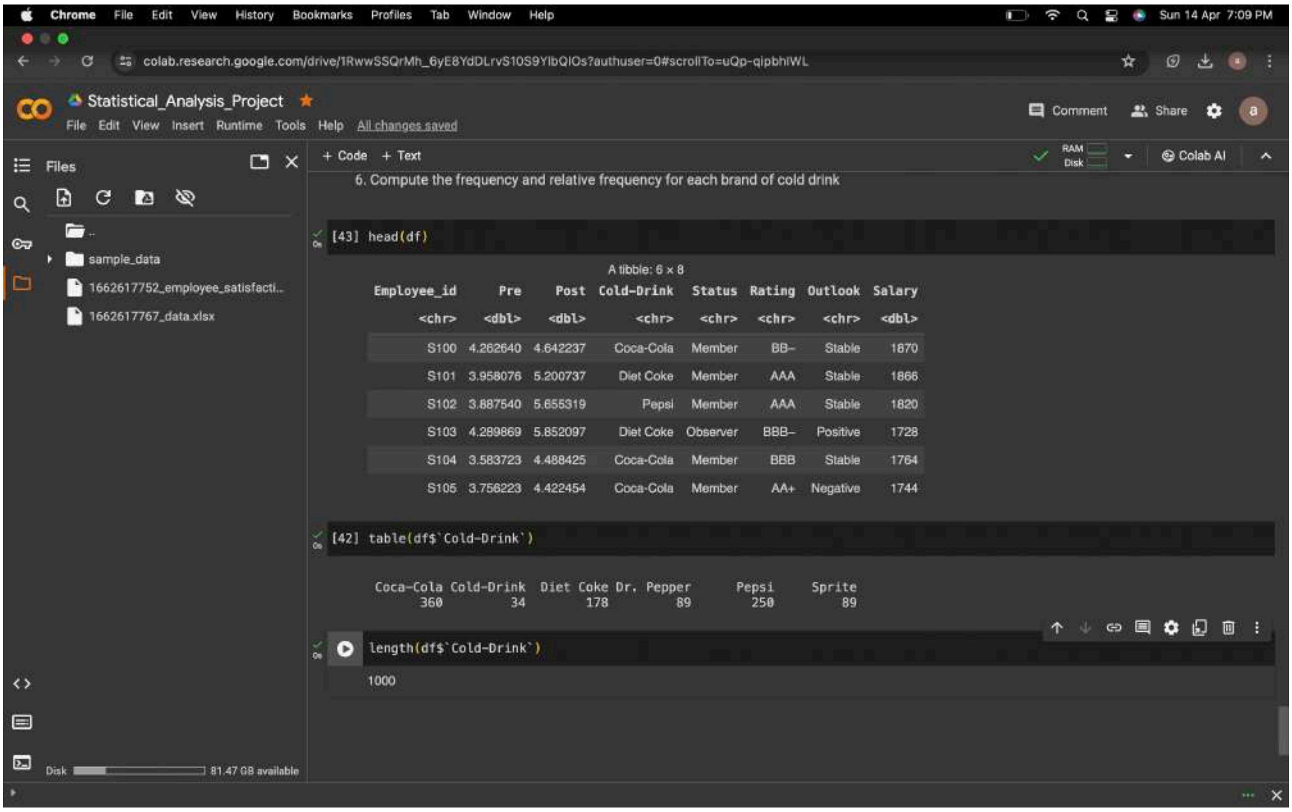


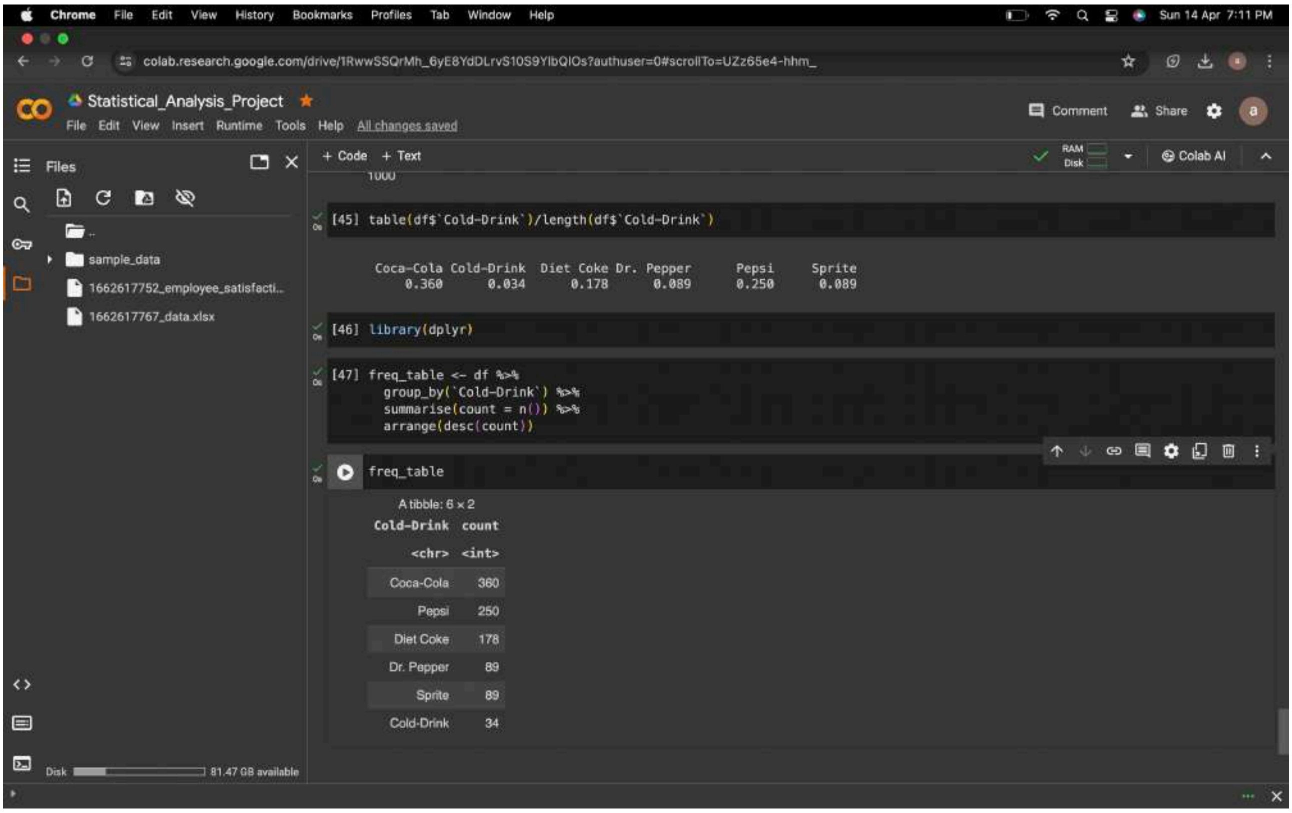



Chrome File Edit View History Bookmarks Profiles Tab Window Help
Sun 14 Apr 7:05 PM
colab.research.google.com/drive/1RwwSSQrMh_6yE8YdDLrvS10S9YlbQIOs?authuser=0#scrollTo=-nte3lgNgAn4
Statistical_Analysis_Project
File Edit View Insert Runtime Tools Help All changes saved
Comment Share
+ Code + Text
Colab AI
Files
sample_data
1662617767_data.xlsx
Disk 81.47 GB available
[38]
Post
den <- density(df$Pre)
ggplot(df, aes(x = Pre)) +
geom_density(fill = "green", alpha = 0.5) +
labs(title = "Kernel Density Plot of Pre",
x = "Pre", y = "Density")
Kernel Density Plot of Pre
Density


Chrome File Edit View History Bookmarks Profiles Tab Window Help
Sun 14 Apr 7:06 PM
colab.research.google.com/drive/1RwwSSQrMh_6yE8YdDLrvS10S9YlbQIOs?authuser=0#scrollTo=NMYiJvXygROM
Statistical_Analysis_Project
File Edit View Insert Runtime Tools Help All changes saved
Comment Share
+ Code + Text
Colab AI
Files
sample_data
1662617767_data.xlsx
Disk 81.47 GB available
[39]
Pre
den <- density(df$Post)
ggplot(df, aes(x = Post)) + geom_density(fill = 'yellow' , alpha = 0.5)
density

6. Compute the frequency and relative frequency for each brand of cold drink





Diet Coke	178
Dr. Pepper	89
Sprite	89
Cold-Drink	34

```
[49] rel_freq_table <- df %>%
       group_by(`Cold-Drink`) %>%
       summarise(count = n()) %>%
       mutate(rfreq = count/sum(count))
```

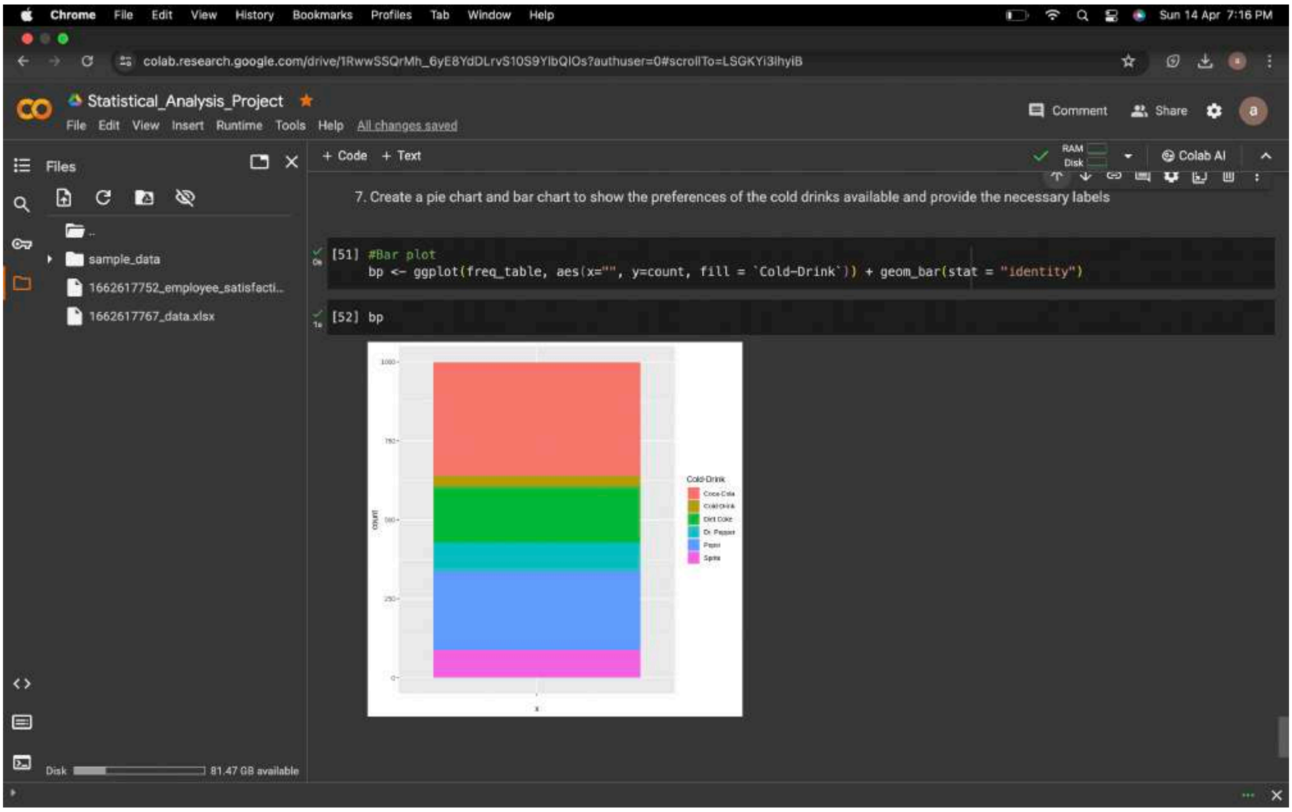
```
[50] rel_freq_table
```

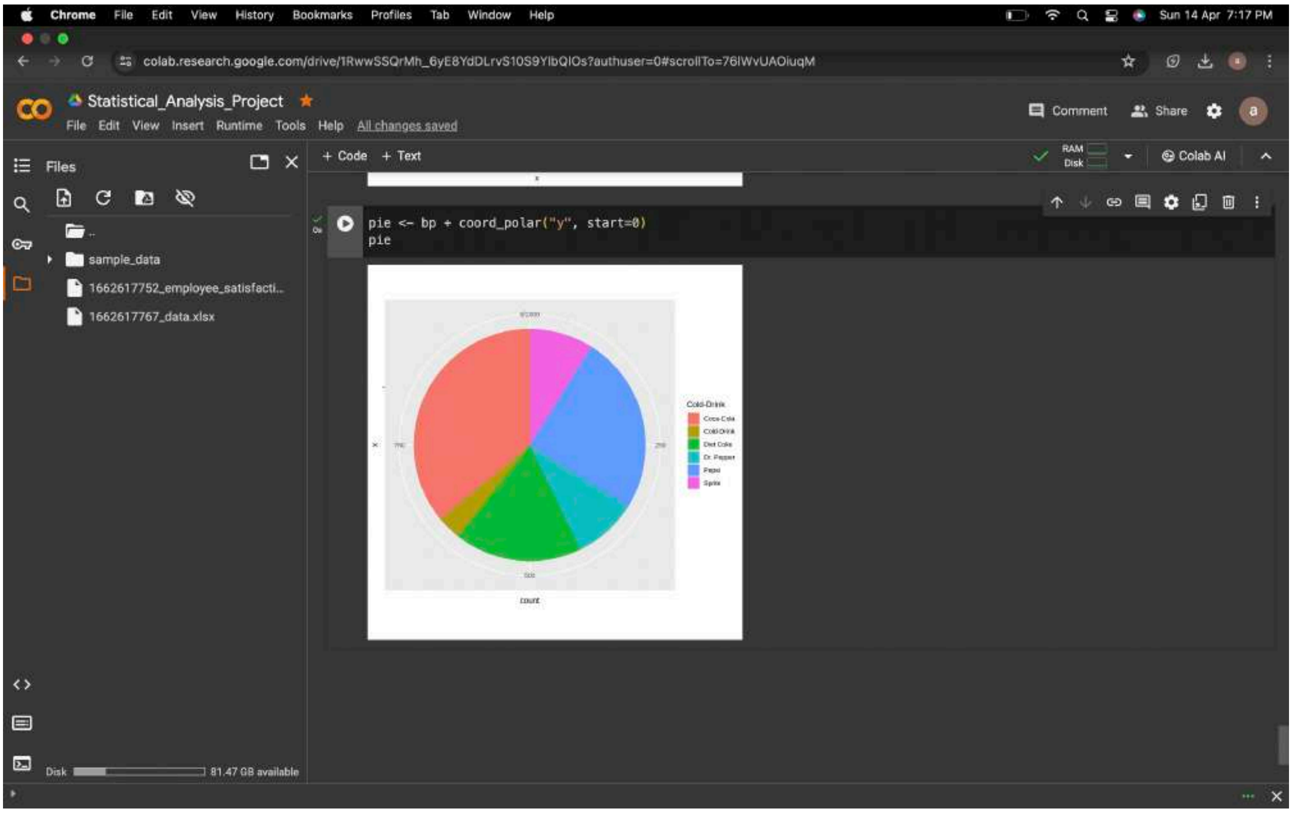
A tibble: 6 × 3

Cold-Drink	count	rfreq
<chr>	<int>	<dbl>
Coca-Cola	360	0.360
Cold-Drink	34	0.034
Diet Coke	178	0.178
Dr. Pepper	89	0.089
Pepsi	250	0.250
Sprite	89	0.089

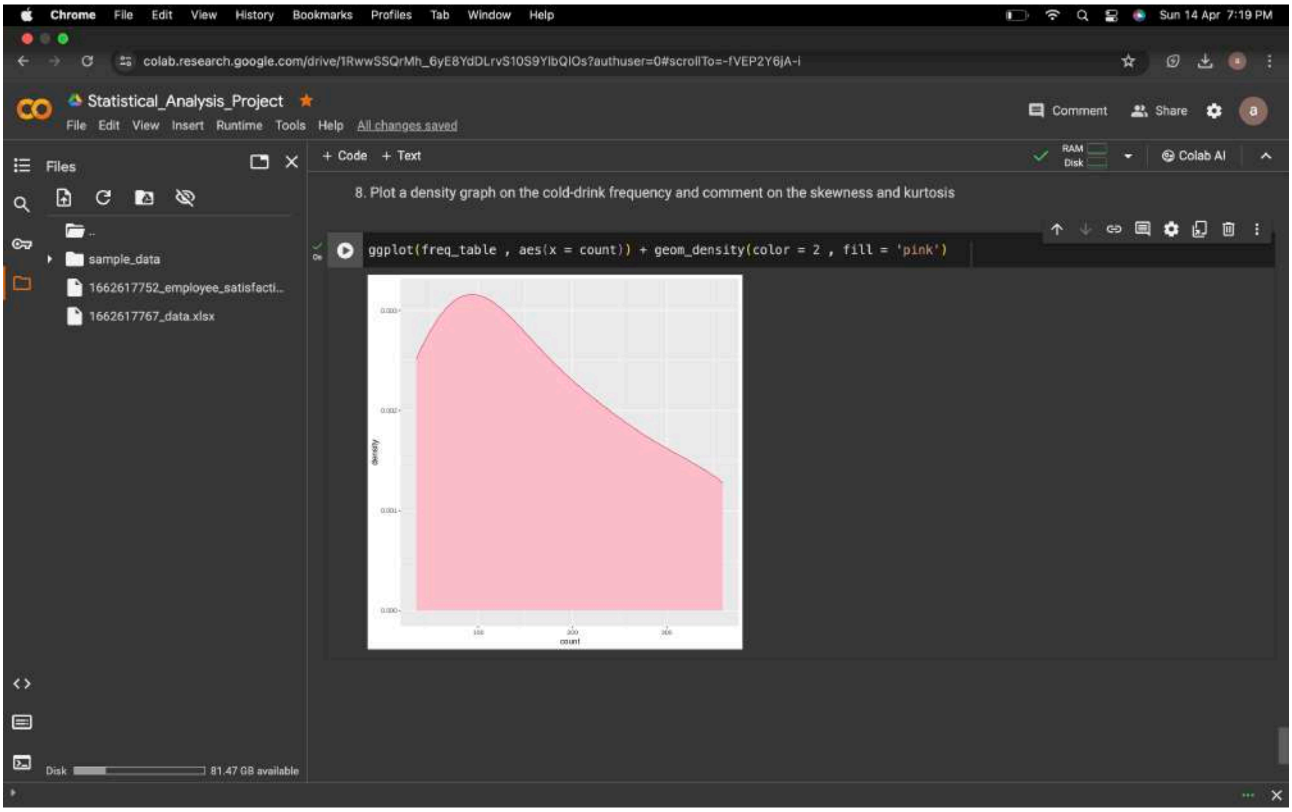
7. Create a pie chart and bar chart to show the preferences of the cold drinks available and provide the necessary labels

7. Create a pie chart and bar chart to show the preferences of the cold drinks available and provide the necessary labels

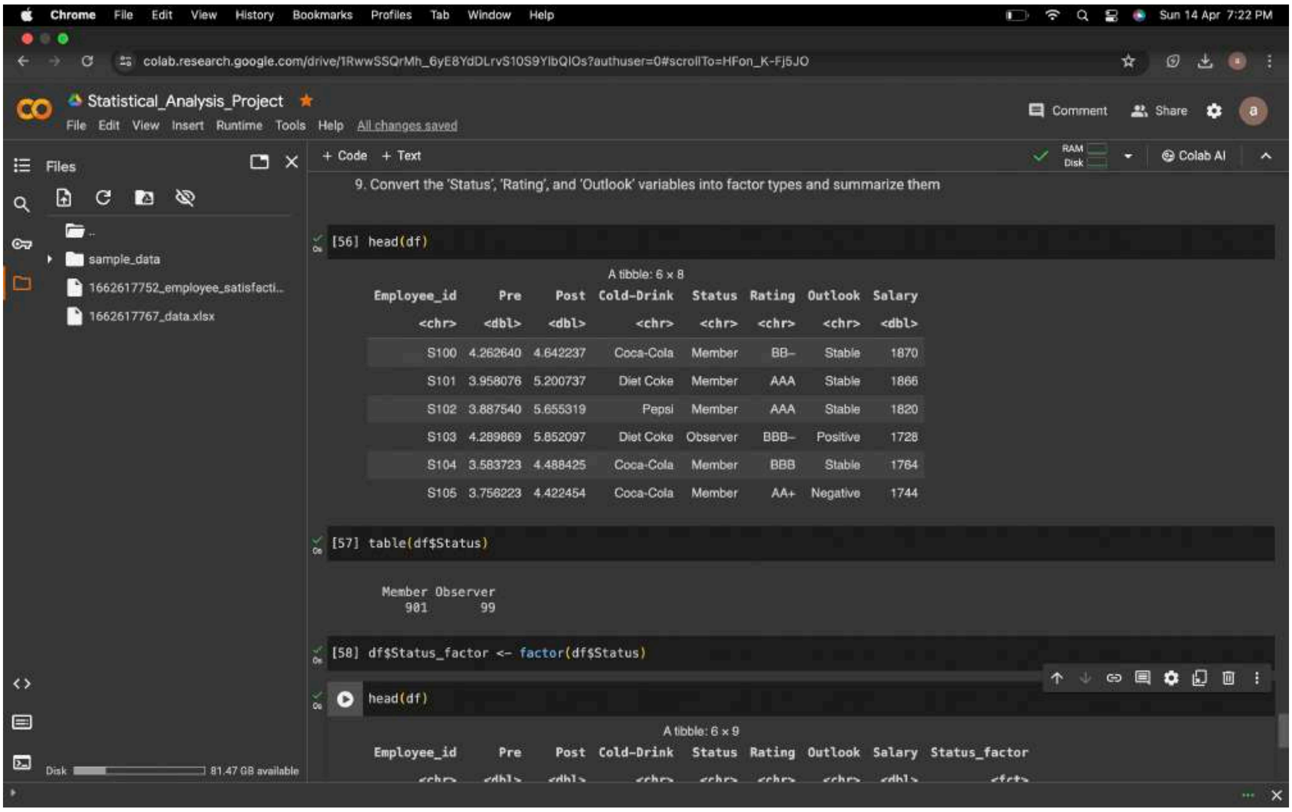


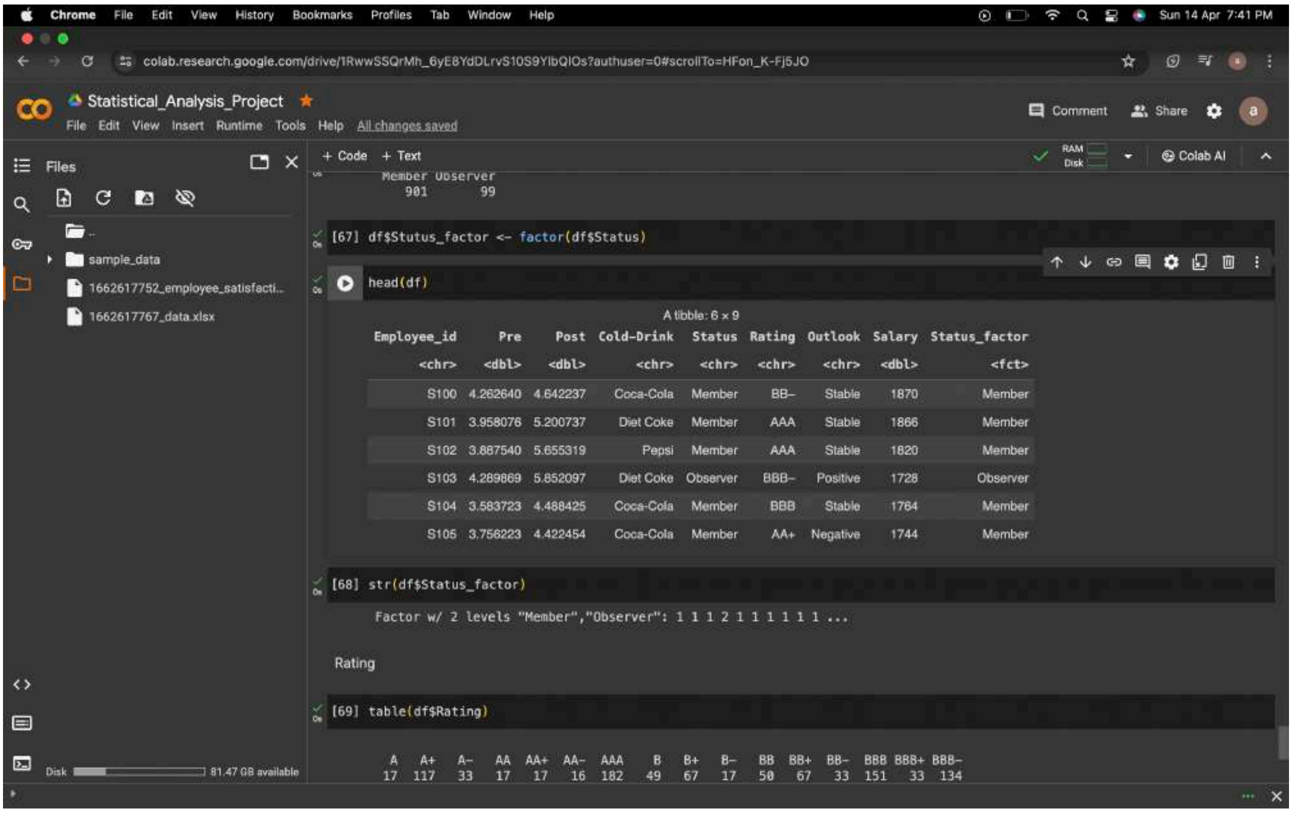


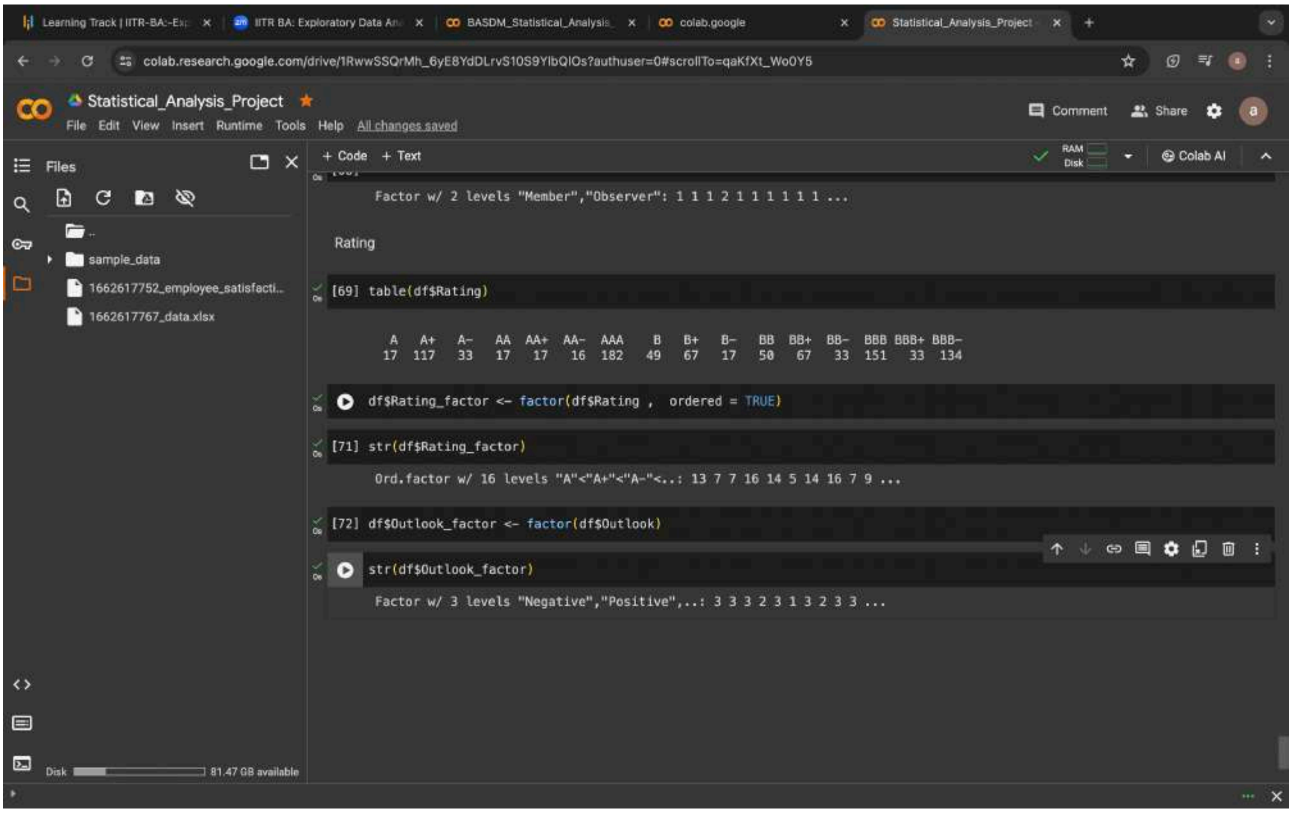
8. Plot a density graph on the cold-drink frequency and comment on the skewness and kurtosis



9. Convert the ‘Status’, ‘Rating’, and ‘Outlook’ variables into factor types and summarize them

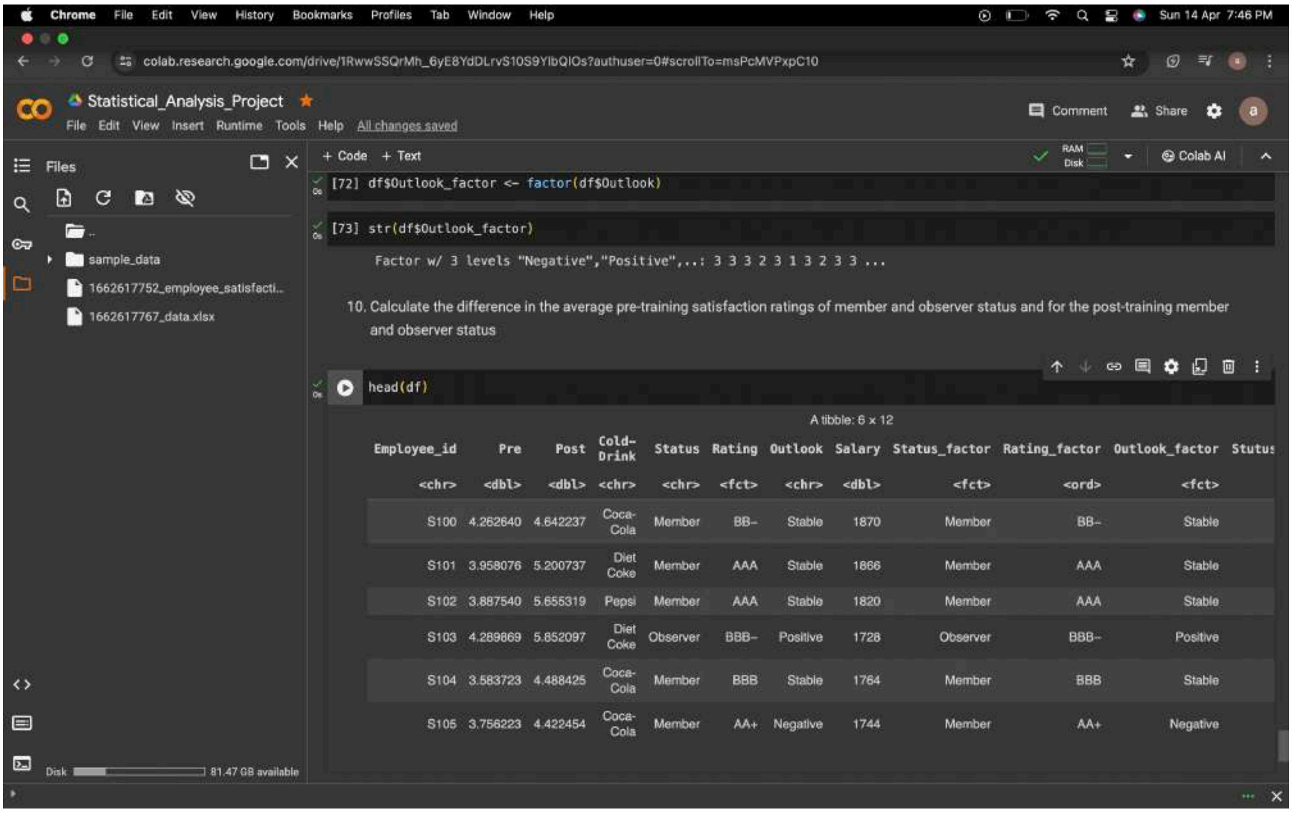


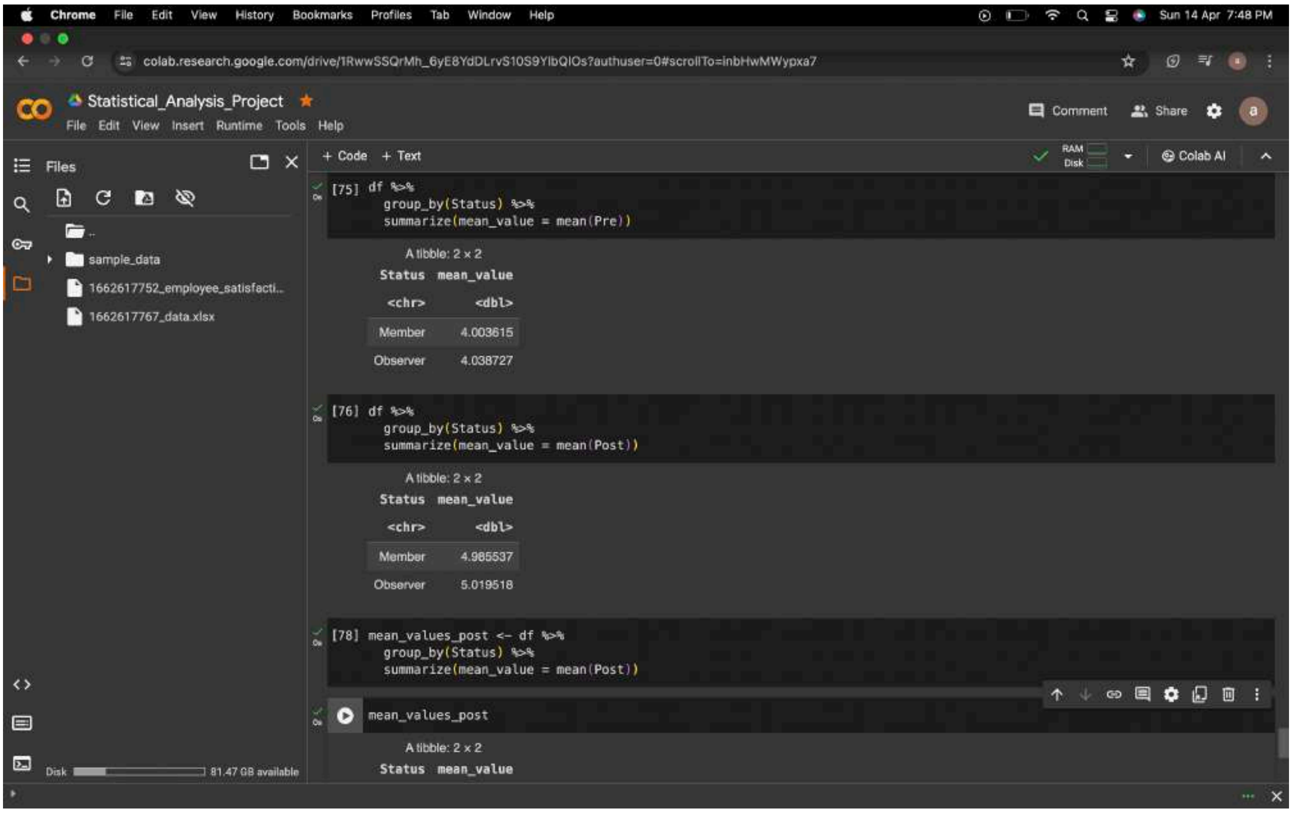




Statistical_Analysis_Project
File Edit View Insert Runtime Tools Help All changes saved
Files
sample_data
1662617752_employee_satisfacti...
1662617767_data.xlsx
Factor w/ 2 levels "Member","Observer": 1 1 1 2 1 1 1 1 1 1 ...
Rating
[69] table(df$Rating)
A A+ A- AA AA+ AA- AAA B B+ B- BB BB+ BB- BBB BBB+ BBB-
17 117 33 17 17 16 182 49 67 17 50 67 33 151 33 134
df$Rating_factor <- factor(df$Rating , ordered = TRUE)
[71] str(df$Rating_factor)
Ord.factor w/ 16 levels "A"<"A+"<"A-"<..: 13 7 7 16 14 5 14 16 7 9 ...
[72] df$Outlook_factor <- factor(df$Outlook)
str(df$Outlook_factor)
Factor w/ 3 levels "Negative","Positive",..: 3 3 3 2 3 1 3 2 3 3 ...
Disk 81.47 GB available

10. Calculate the difference in the average pre-training satisfaction ratings of member and observer status and for the post-training member and observer status





```
mean_values_post
```

A tibble: 2 x 2

Status	mean_value
<chr>	<dbl>
Member	4.985537
Observer	5.019518

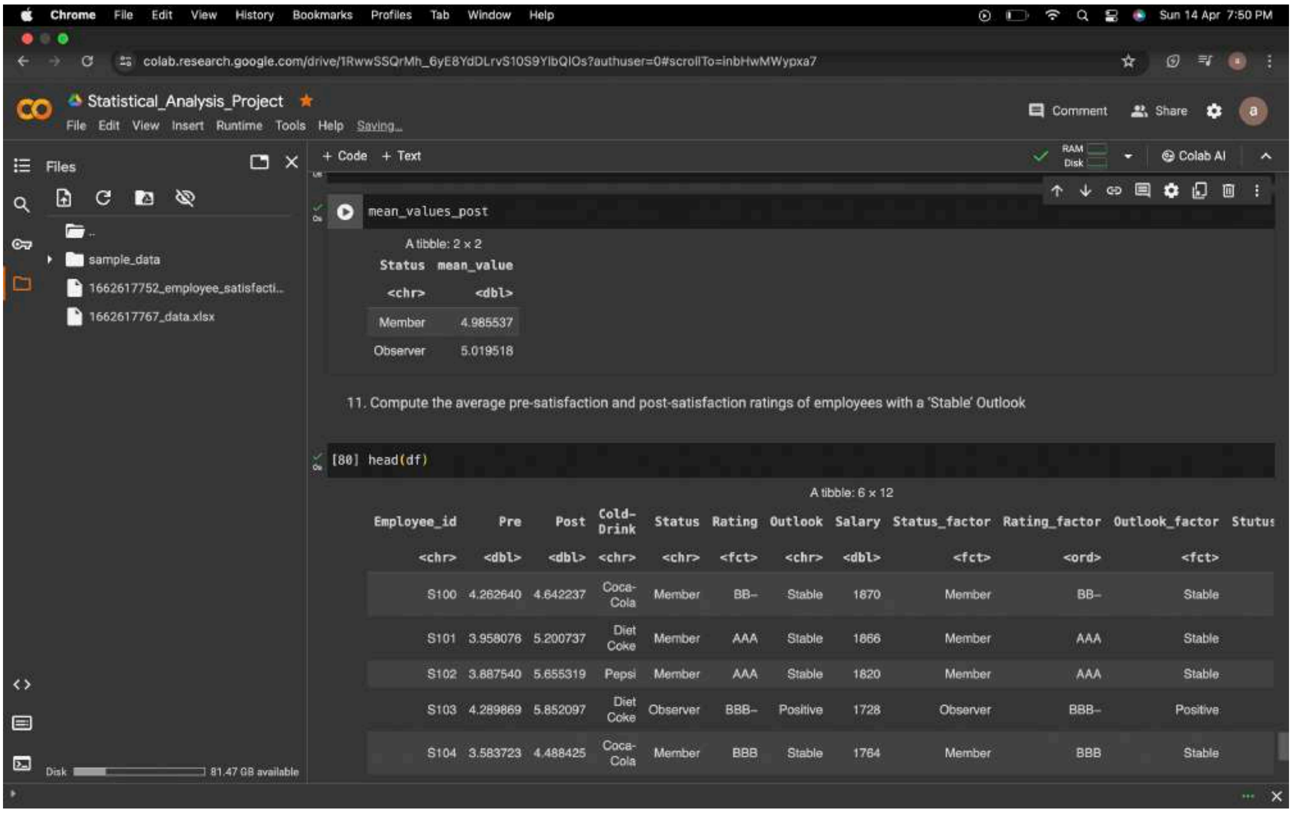
11. Compute the average pre-satisfaction and post-satisfaction ratings of employees with a 'Stable' Outlook

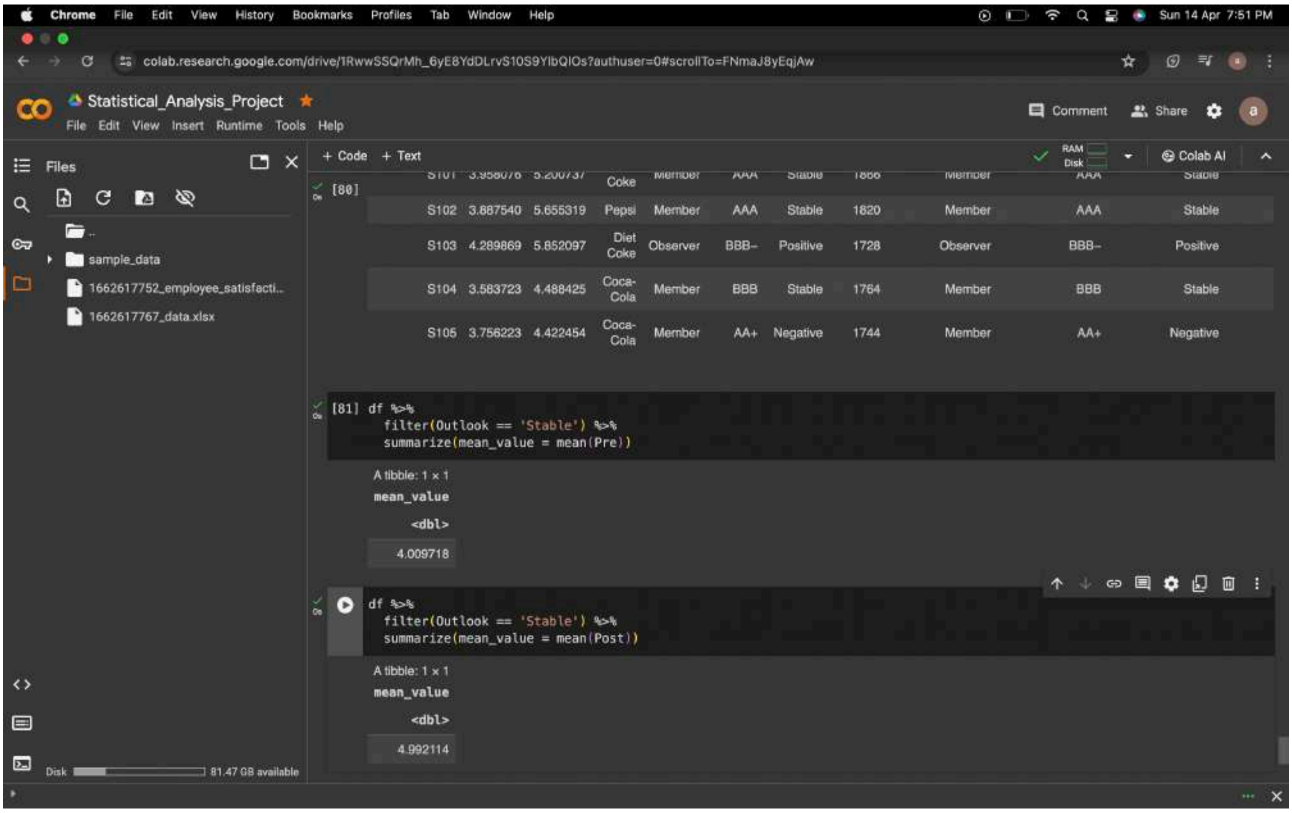
```
[80] head(df)
```

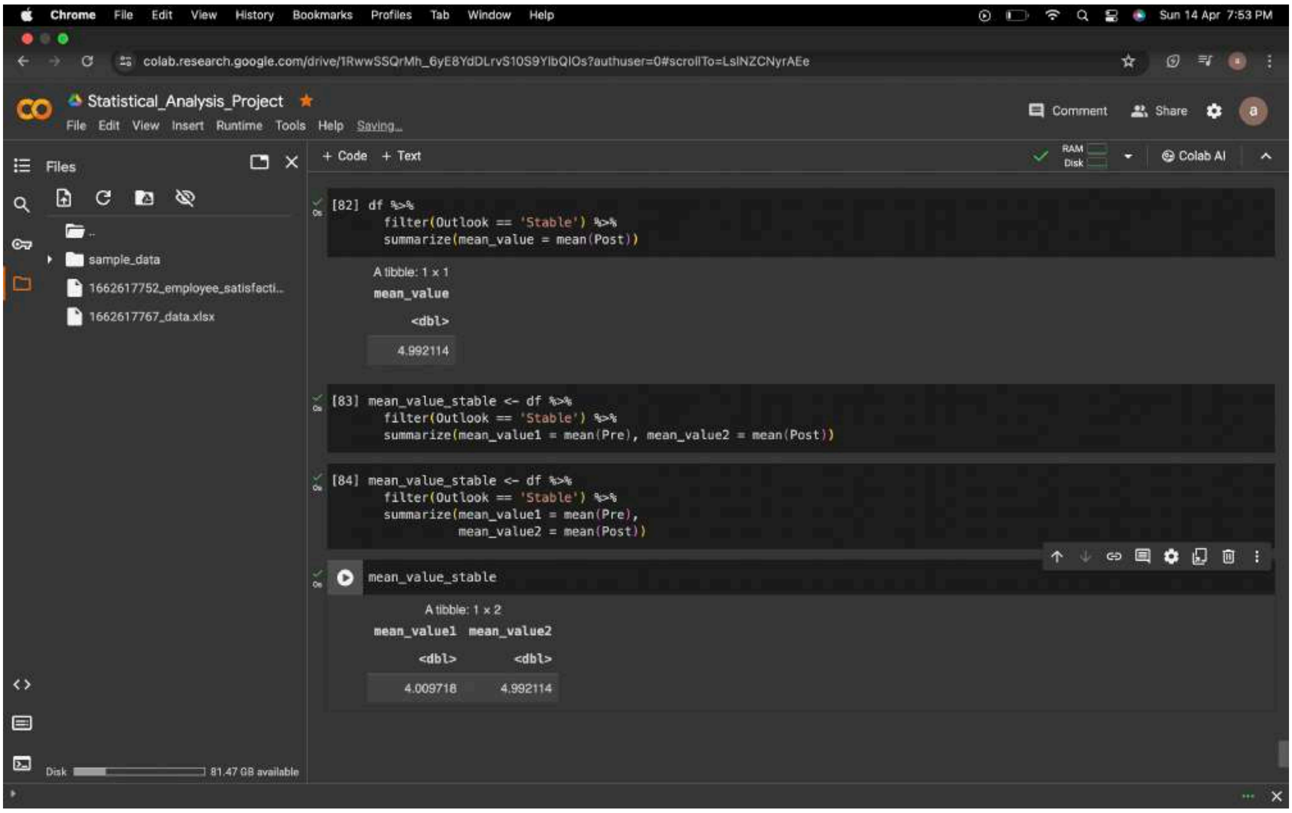
A tibble: 6 x 12

Employee_id	Pre	Post	Cold-Drink	Status	Rating	Outlook	Salary	Status_factor	Rating_factor	Outlook_factor	Stutus
<chr>	<dbl>	<dbl>	<chr>	<chr>	<fct>	<chr>	<dbl>	<fct>	<ord>	<fct>	
S100	4.262640	4.642237	Coca-Cola	Member	BB-	Stable	1870	Member	BB-	Stable	
S101	3.958076	5.200737	Diet Coke	Member	AAA	Stable	1866	Member	AAA	Stable	
S102	3.887540	5.655319	Pepsi	Member	AAA	Stable	1820	Member	AAA	Stable	
S103	4.289869	5.852097	Diet Coke	Observer	BBB-	Positive	1728	Observer	BBB-	Positive	
S104	3.583723	4.488425	Coca-Cola	Member	BBB	Stable	1764	Member	BBB	Stable	

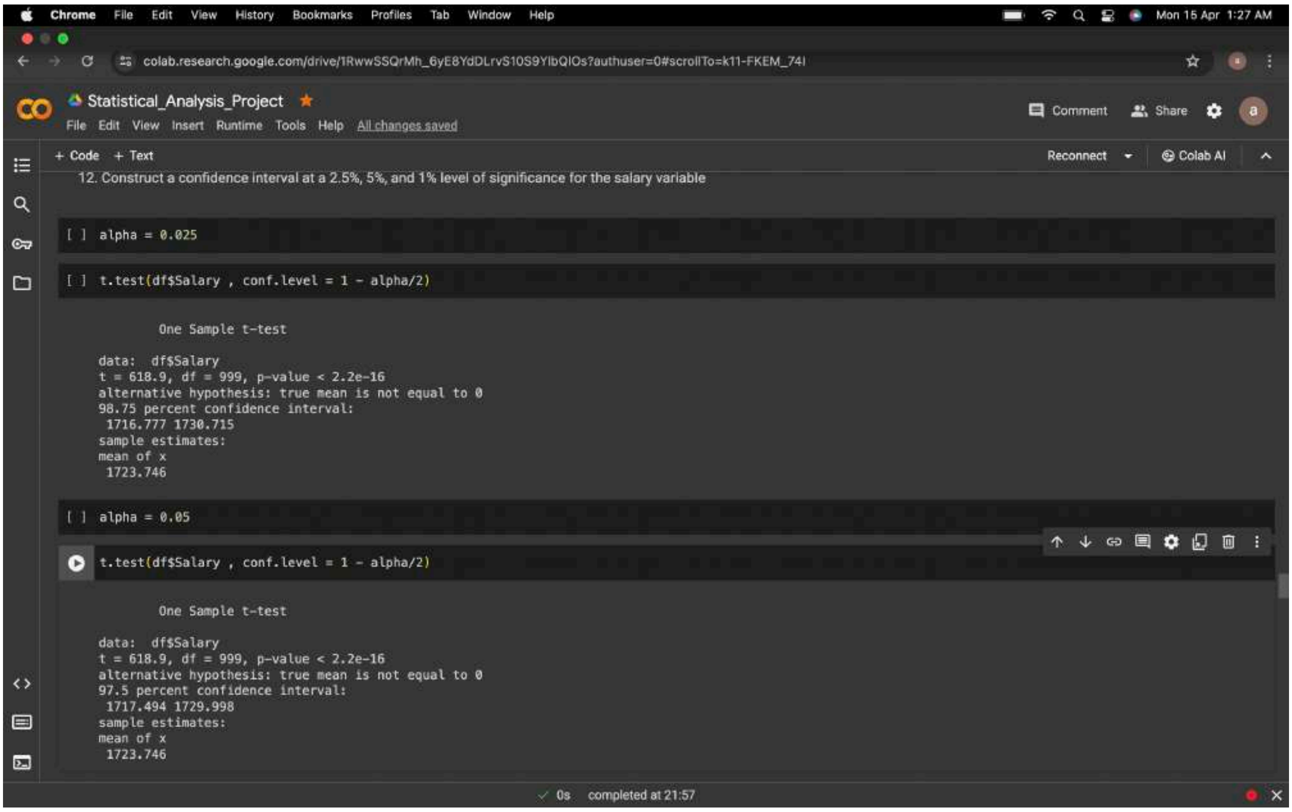
11. Compute the average pre-satisfaction and post-satisfaction ratings of employees with a 'Stable' Outlook

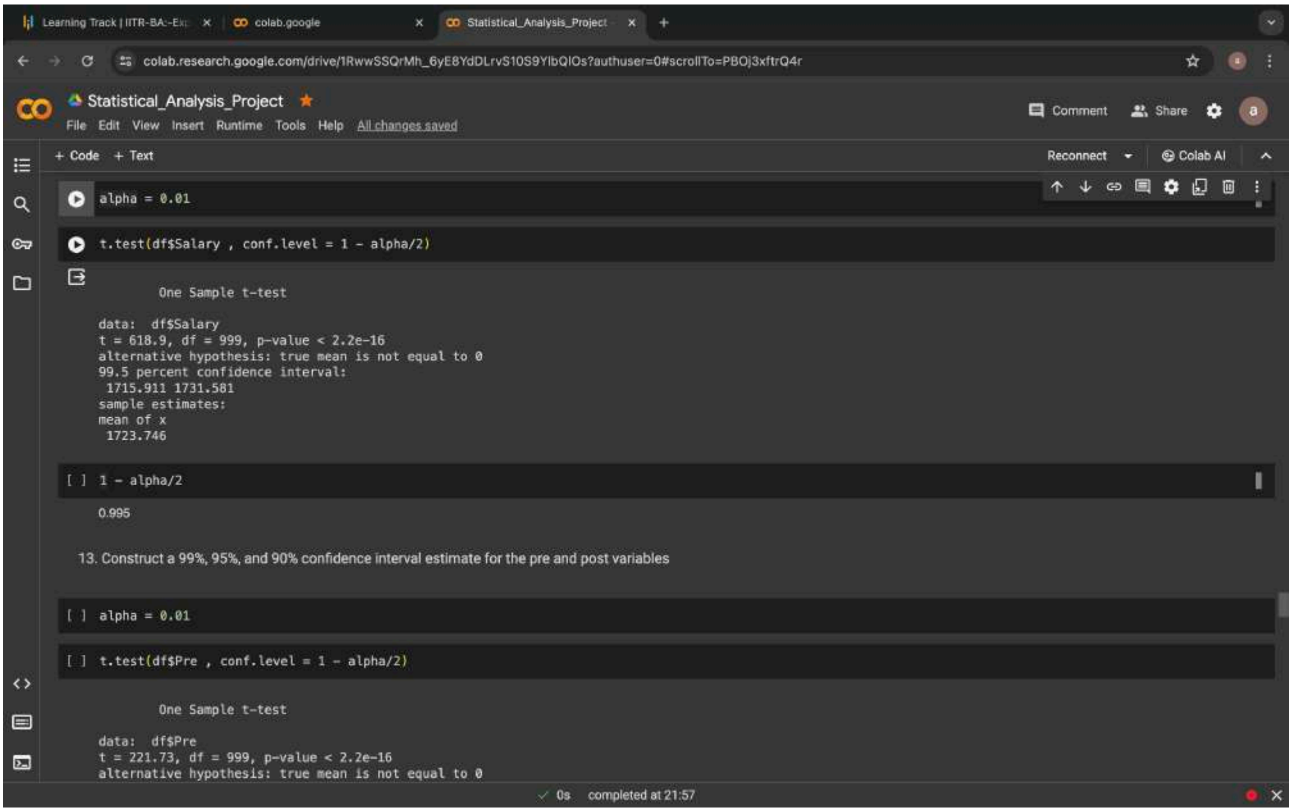




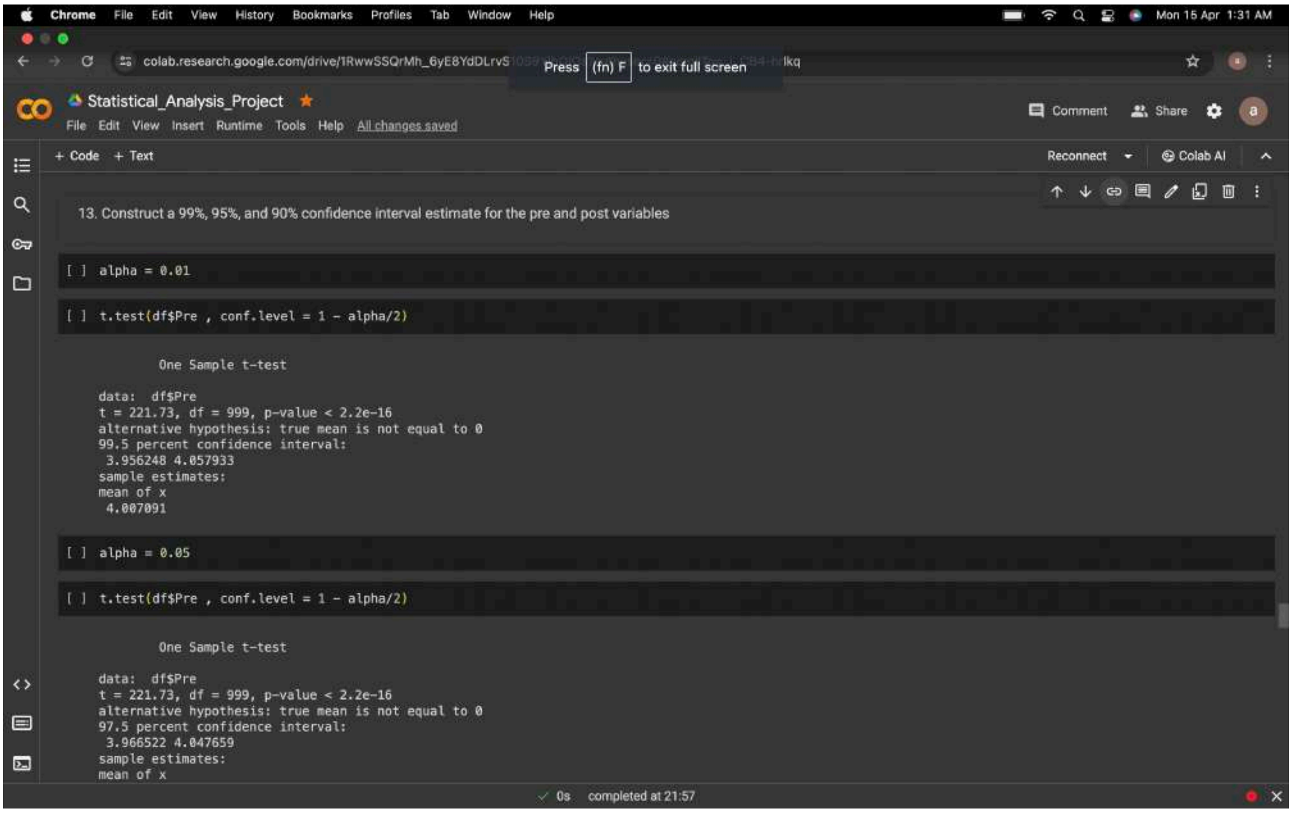

Chrome File Edit View History Bookmarks Profiles Tab Window Help
Sun 14 Apr 7:53 PM
colab.research.google.com/drive/1RwwSSQrMh_6yE8YdDLrvS10S9YlbQIOs?authuser=0#scrollTo=LsINZCNyrAEe
Statistical_Analysis_Project
File Edit View Insert Runtime Tools Help Saving...
Comment
Share
Files
sample_data
1662617752_employee_satisfacti...
1662617767_data.xlsx
+ Code + Text
RAM
Disk
Colab AI
[82] df %>%
filter(Outlook == 'Stable') %>%
summarize(mean_value = mean(Post))
A tibble: 1 x 1
mean_value
<dbl>
4.992114
[83] mean_value_stable <- df %>%
filter(Outlook == 'Stable') %>%
summarize(mean_value1 = mean(Pre), mean_value2 = mean(Post))
[84] mean_value_stable <- df %>%
filter(Outlook == 'Stable') %>%
summarize(mean_value1 = mean(Pre),
mean_value2 = mean(Post))
mean_value_stable
A tibble: 1 x 2
mean_value1 mean_value2
<dbl> <dbl>
4.009718 4.992114
Disk 81.47 GB available

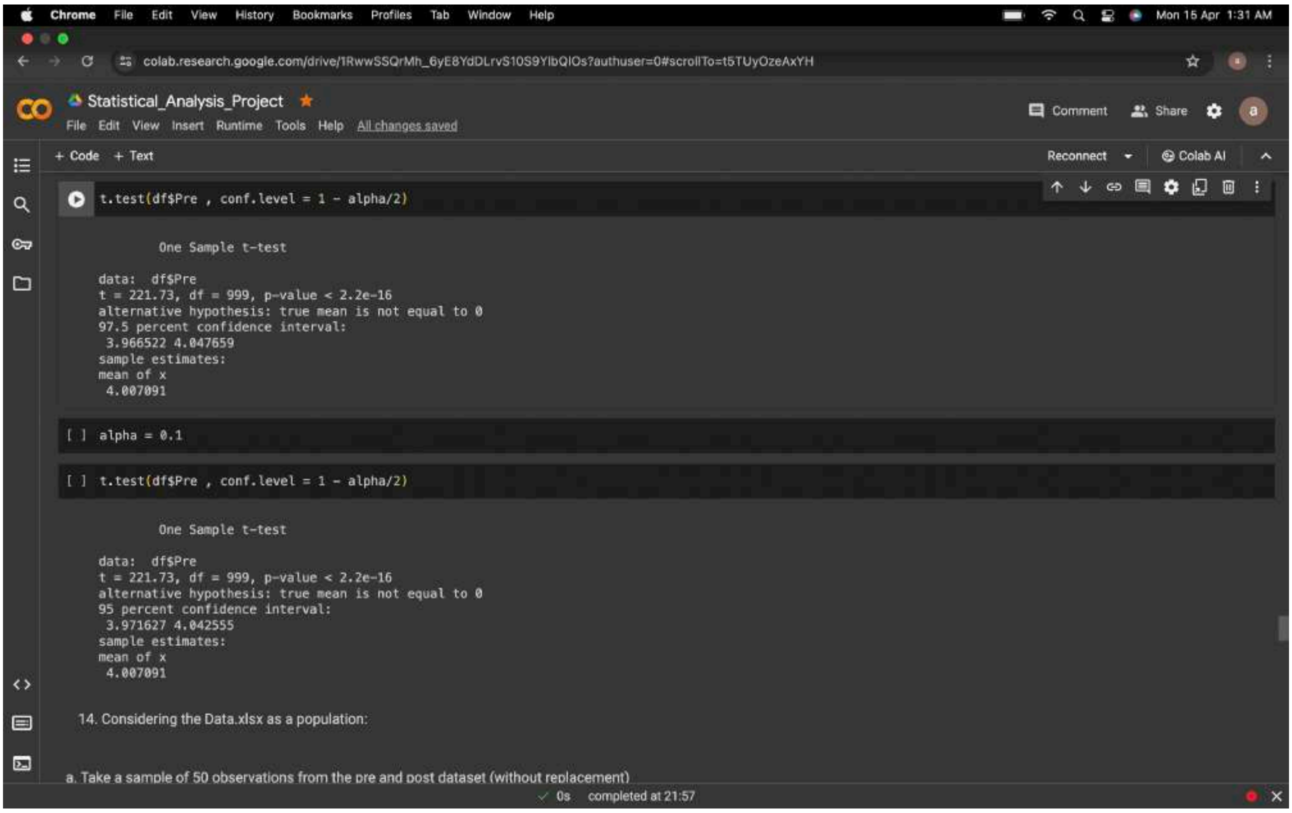
12. Construct a confidence interval at a 2.5%, 5%, and 1% level of significance for the salary variable





13. Construct a 99%, 95%, and 90% confidence interval estimate for the pre and post variables





Chrome File Edit View History Bookmarks Profiles Tab Window Help Mon 15 Apr 1:32 AM

colab.research.google.com/drive/1RwwSSQrMh_6yE8YdDLrvS10S9YlbQIOs?authuser=0#scrollTo=n-P0ntHru5Vu

Statistical_Analysis_Project

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text Reconnect Colab AI

3.599272	4.576912
4.476553	4.025431
3.295642	4.530308
4.166979	5.943382
4.799632	5.084983
3.314744	4.375508
4.046347	4.115508
4.751366	4.573402
3.848501	4.545653
3.092518	4.804599
3.533836	5.144599
3.675000	5.281465
3.051942	5.587990

b. Construct a null hypothesis to examine whether the sample (50 observations) mean score of pre and post variables is significantly different from the population (1000 observations)

H0: Mean of pre for population = mean of pre for sample

H0: Mean of post for population = mean of post for sample

c. Compute corresponding Z values for pre and post variables in the sample

✓ 0s completed at 21:57

S100	4.262640	4.642237	Coca-Cola	Member	BB–	Stable	1870	Member	BB–	Stable	Member
S101	3.958076	5.200737	Diet Coke	Member	AAA	Stable	1866	Member	AAA	Stable	Member
S102	3.887540	5.655319	Pepsi	Member	AAA	Stable	1820	Member	AAA	Stable	Member
S103	4.289869	5.852097	Diet Coke	Observer	BBB–	Positive	1728	Observer	BBB–	Positive	Observer
S104	3.583723	4.488425	Coca-Cola	Member	BBB	Stable	1764	Member	BBB	Stable	Member
S105	3.756223	4.422454	Coca-Cola	Member	AA+	Negative	1744	Member	AA+	Negative	Member

```
#random_seed = 10
sample(df$Pre, 50, set.seed(10) , replace = FALSE)
```

3.976724433247 · 4.51862944476306 · 3.28275757003576 · 4.95312619907781 · 3.51982954517007 · 3.91128359735012 · 3.58252657949924 · 3.84184411447495 · 4.83535359520465 · 3.86871136259288 · 4.0130110620521 · 3.76542822411284 · 4.05111684091389 · 3.48400401137769 · 4.07847984740511 · 3.92829271033406 · 4.99834007397294 · 4.19021744793281 · 4.55032476363704 · 3.8037439561449 · 3.33822550578043 · 3.65714739030227 · 4.84340020827949 · 3.25410187104717 · 4.40114601748064 · 3.48685706825927 · 4.33826791774482 · 3.23117155442014 · 4.60246529104188 · 4.79963213996962 · 4.07577075762674 · 3.03743795119226 · 3.91871575312689 · 3.02240114891902 · 4.65731397131458 · 3.8169745602645 · 4.51704967068508 · 3.75117048528045 · 4.99154957616702 · 4.25125201698393 · 3.63965039048344 · 3.03269549971446 · 4.32452464243397 · 3.44686652068049 · 3.55948808044195 · 4.19852462178096 · 3.40928723243997 · 3.36347079463303 · 3.45726076187566 · 4.52105776034296

✓ 0s completed at 21:57

Chrome File Edit View History Bookmarks Profiles Tab Window Help Mon 15 Apr 1:32 AM

colab.research.google.com/drive/1RwwSSQrMh_6yE8YdDLrvS10S9YlbQIOs?authuser=0#scrollTo=n-P0ntHru5Vu

Statistical_Analysis_Project

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text Reconnect Colab AI

c. Compute corresponding Z values for pre and post variables in the sample

Pre

```
[ ] # Compute Z-values for pre variable
    pre_sample_mean <- mean(sampled_values$Pre)
```

```
[ ] pre_sample_sd <- sd(sampled_values$Pre)
```

```
[ ] pre_sample_sd
```

0.566559022092012

```
[ ] pre_pop_mean <- mean(df$Pre)
```

```
[ ] pre_pop_sd <- sd(df$Pre)
```

```
[ ] pre_pop_mean
```

4.00709069415415

```
[ ] #Compute Z-value
    n = 50 #Sample size

    z_pre = (pre_sample_mean - pre_pop_mean) / (pre_pop_sd / sqrt(n))
```

```
[ ] z_pre
```

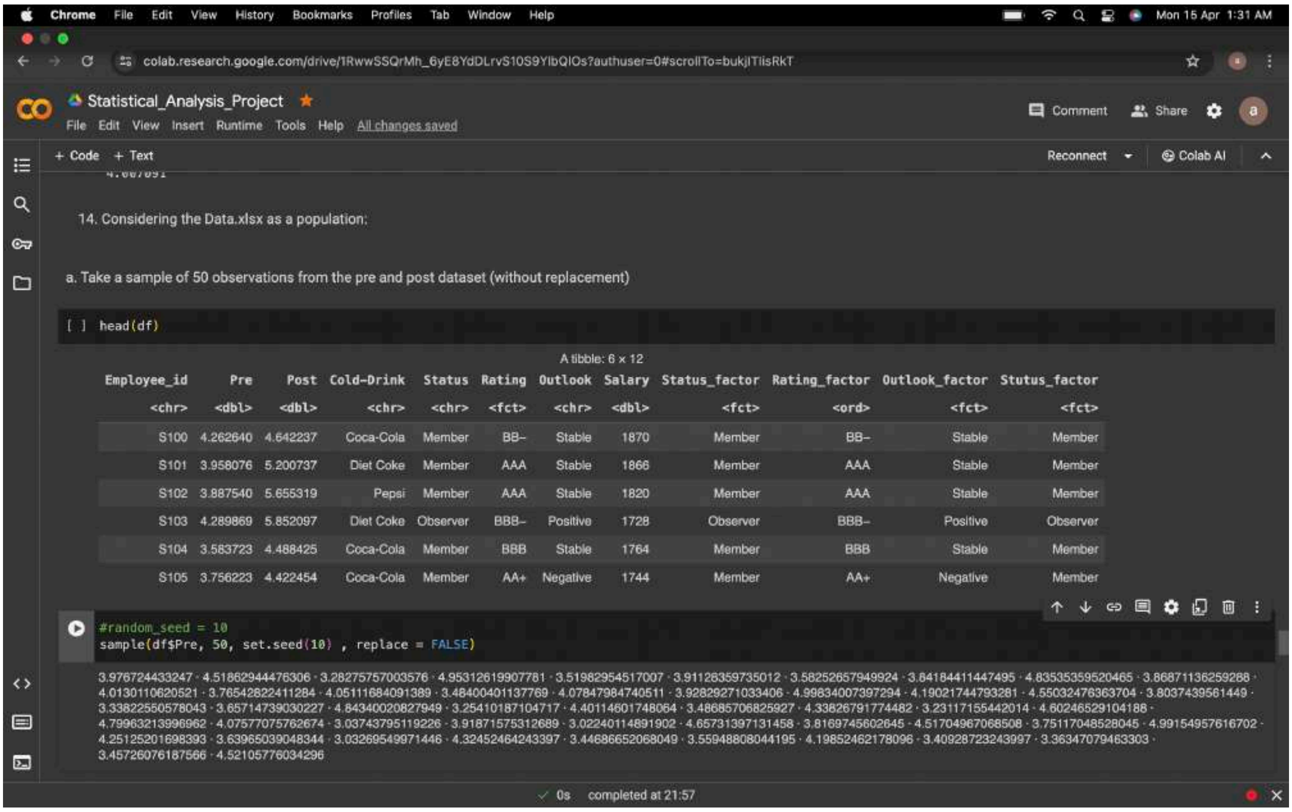
✓ 0s completed at 21:57
✓ 0s completed at 21:57

14. Considering the Data.xlsx as a population:

   a. Take a sample of 50 observations from the pre and post dataset (without replacement)

   b. Construct a null hypothesis to examine whether the sample (50 observations) mean score of pre and post variables is significantly different from the population (1000 observations)

   c. Compute corresponding Z values for pre and post variables in the sample



14. Considering the Data.xlsx as a population:

a. Take a sample of 50 observations from the pre and post dataset (without replacement)

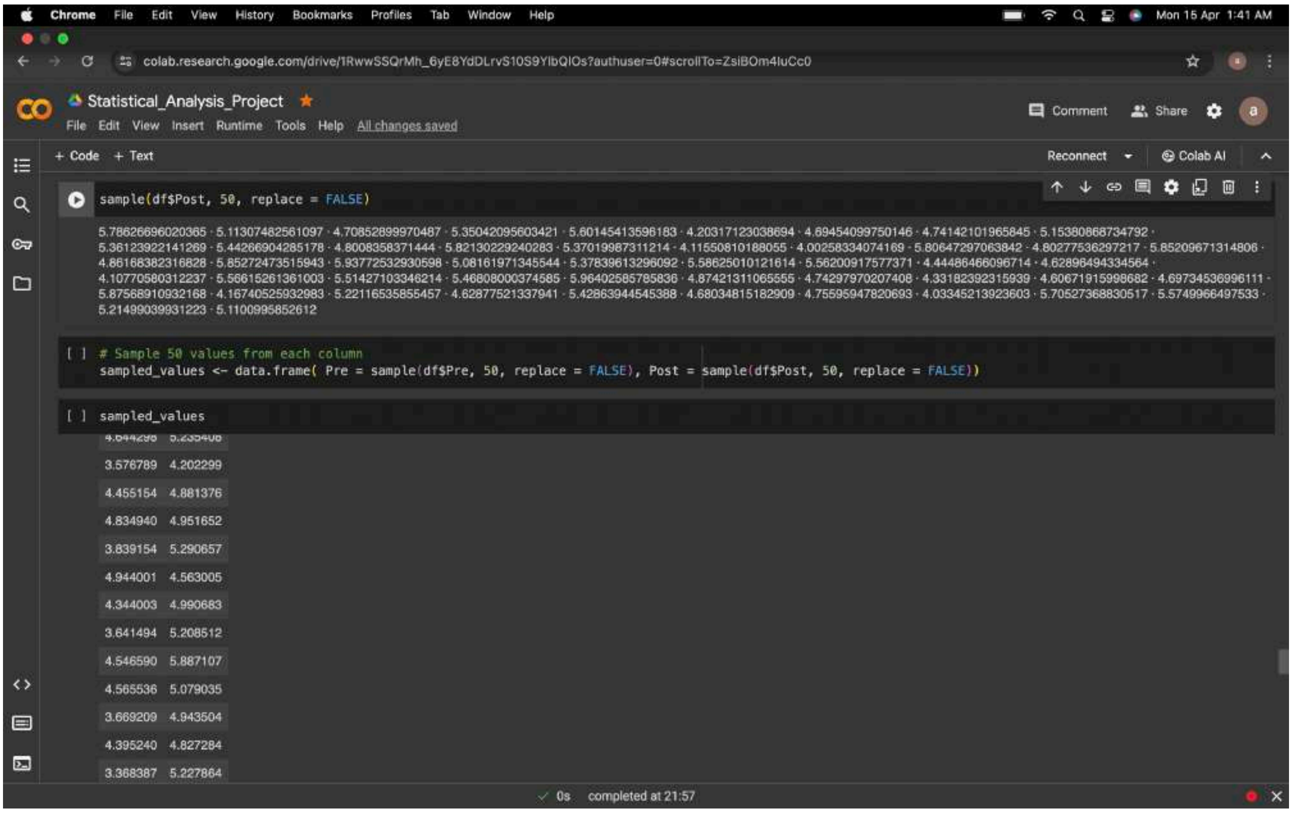
```
head(df)
```

A tibble: 6 × 12

Employee_id	Pre	Post	Cold-Drink	Status	Rating	Outlook	Salary	Status_factor	Rating_factor	Outlook_factor	Stutus_factor
<chr>	<dbl>	<dbl>	<chr>	<chr>	<fct>	<chr>	<dbl>	<fct>	<ord>	<fct>	<fct>
S100	4.262640	4.642237	Coca-Cola	Member	BB-	Stable	1870	Member	BB-	Stable	Member
S101	3.958076	5.200737	Diet Coke	Member	AAA	Stable	1866	Member	AAA	Stable	Member
S102	3.887540	5.655319	Pepsi	Member	AAA	Stable	1820	Member	AAA	Stable	Member
S103	4.289869	5.852097	Diet Coke	Observer	BBB-	Positive	1728	Observer	BBB-	Positive	Observer
S104	3.583723	4.488425	Coca-Cola	Member	BBB	Stable	1764	Member	BBB	Stable	Member
S105	3.756223	4.422454	Coca-Cola	Member	AA+	Negative	1744	Member	AA+	Negative	Member

```
#random_seed = 10
sample(df$Pre, 50, set.seed(10) , replace = FALSE)
```

3.976724433247 · 4.51862944476306 · 3.28275757003576 · 4.95312619907781 · 3.51982954517007 · 3.91126359735012 · 3.58252657949924 · 3.84184411447495 · 4.83535359520465 · 3.86871136259288 · 4.0130110620521 · 3.76542822411284 · 4.05111684091389 · 3.48400401137769 · 4.07847984740511 · 3.92829271033406 · 4.99834007397294 · 4.19021744793281 · 4.55032476363704 · 3.8037439561449 · 3.33822550578043 · 3.65714739030227 · 4.84340020827949 · 3.25410187104717 · 4.40114601748064 · 3.48685706825927 · 4.33826791774482 · 3.23117155442014 · 4.60246529104188 · 4.79963213996962 · 4.07577075762674 · 3.03743795119226 · 3.91871575312689 · 3.02240114891902 · 4.65731397131458 · 3.8169745602645 · 4.51704967068508 · 3.75117048528045 · 4.99154957616702 · 4.25125201698393 · 3.63965039048344 · 3.03269549971446 · 4.32452464243397 · 3.44686652068049 · 3.55948808044195 · 4.19852462178096 · 3.40928723243997 · 3.36347079463303 · 3.45726076187566 · 4.52105776034296



```
sample(df$Post, 50, replace = FALSE)
```

5.78626696020365 · 5.11307482561097 · 4.70852899970487 · 5.35042095603421 · 5.60145413596183 · 4.20317123038694 · 4.69454099750146 · 4.74142101965845 · 5.15380868734792 · 5.36123922141269 · 5.44266904285178 · 4.80083583714444 · 5.82130229240283 · 5.37019987311214 · 4.11550810188055 · 4.00258334074169 · 5.80647297063842 · 4.80277536297217 · 5.85209671314806 · 4.86168382316828 · 5.85272473515943 · 5.93772532930598 · 5.08161971345544 · 5.37839613296092 · 5.58625010121614 · 5.56200917577371 · 4.44486466096714 · 4.62896494334564 · 4.10770580312237 · 5.56615261361003 · 5.51427103346214 · 5.46808000374585 · 5.96402585785836 · 4.87421311065555 · 4.74297970207408 · 4.33182392315939 · 4.60671915998682 · 4.69734536996111 · 5.87568910932168 · 4.16740525932983 · 5.22116535855457 · 4.62877521337941 · 5.42863944545388 · 4.68034815182909 · 4.75595947820693 · 4.03345213923603 · 5.70527368830517 · 5.5749966497533 · 5.21499039931223 · 5.1100995852612

```
# Sample 50 values from each column
sampled_values <- data.frame( Pre = sample(df$Pre, 50, replace = FALSE), Post = sample(df$Post, 50, replace = FALSE))
```

```
sampled_values
```

3.576789	4.202299
4.455154	4.881376
4.834940	4.951652
3.839154	5.290657
4.944001	4.563005
4.344003	4.990683
3.641494	5.208512
4.546590	5.887107
4.565536	5.079035
3.669209	4.943504
4.395240	4.827284
3.368387	5.227864

4.476553	4.025431
3.295642	4.530308
4.166979	5.943382
4.799632	5.084983
3.314744	4.375508
4.046347	4.115508
4.751366	4.573402
3.848501	4.545653
3.092518	4.804599
3.533836	5.144599
3.675000	5.281465
3.051942	5.587990

b. Construct a null hypothesis to examine whether the sample (50 observations) mean score of pre and post variables is significantly different from the population (1000 observations)

H0: Mean of pre for population = mean of pre for sample

H0: Mean of post for population = mean of post for sample

c. Compute corresponding Z values for pre and post variables in the sample

Pre

c. Compute corresponding Z values for pre and post variables in the sample

Pre

```
# Compute Z-values for pre variable
pre_sample_mean <- mean(sampled_values$Pre)
```

```
pre_sample_sd <- sd(sampled_values$Pre)
```

```
pre_sample_sd
```

0.566559022092012

```
pre_pop_mean <- mean(df$Pre)
```

```
pre_pop_sd <- sd(df$Pre)
```

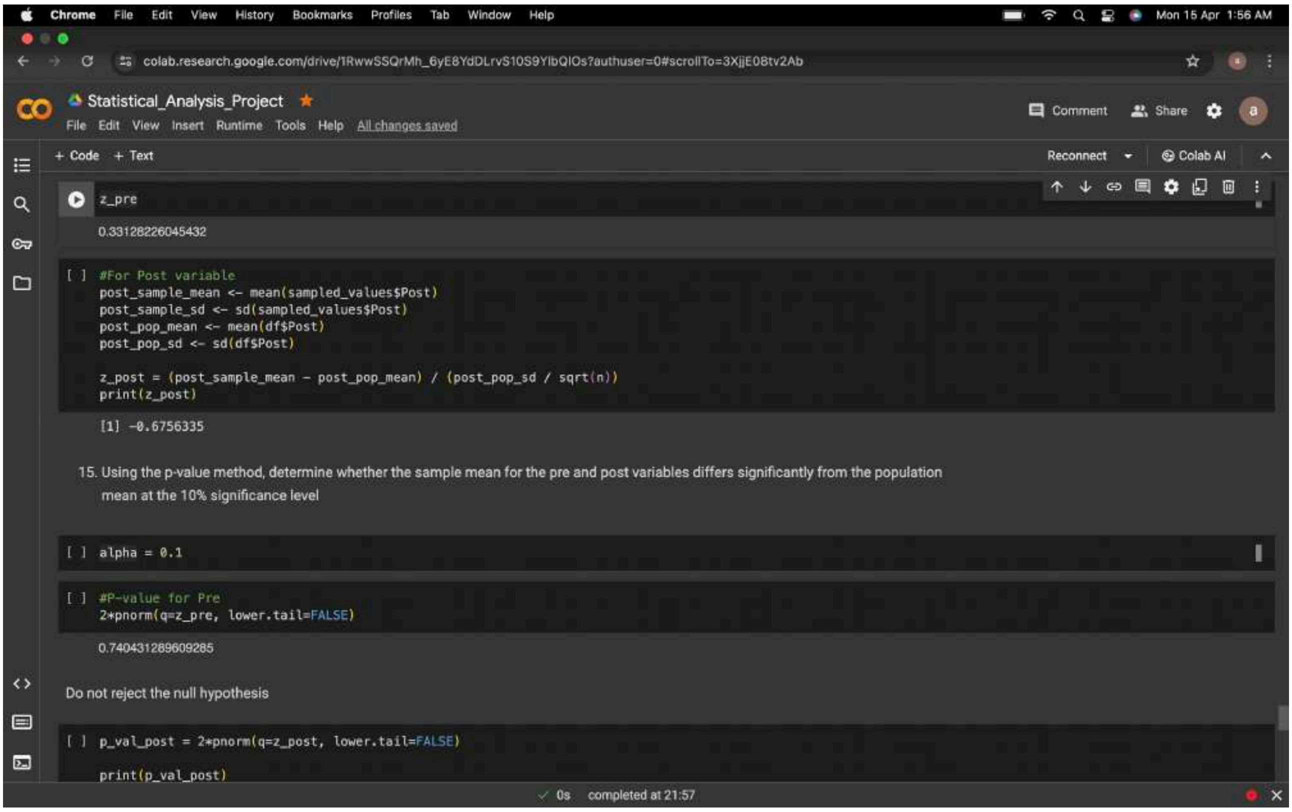
```
pre_pop_mean
```

4.00709069415415

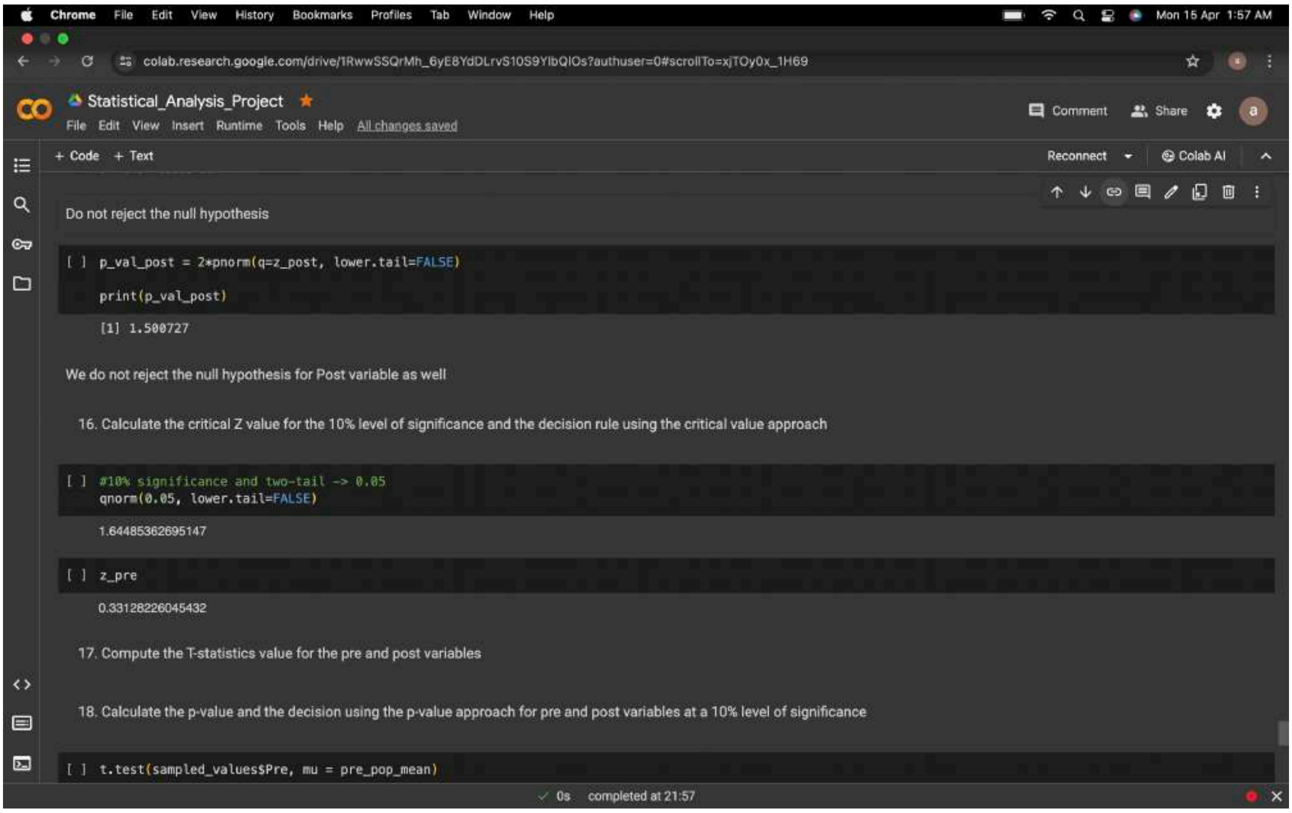
```
#Compute Z-value
n = 50 #Sample size

z_pre = (pre_sample_mean - pre_pop_mean) / (pre_pop_sd / sqrt(n))
```

```
z_pre
```

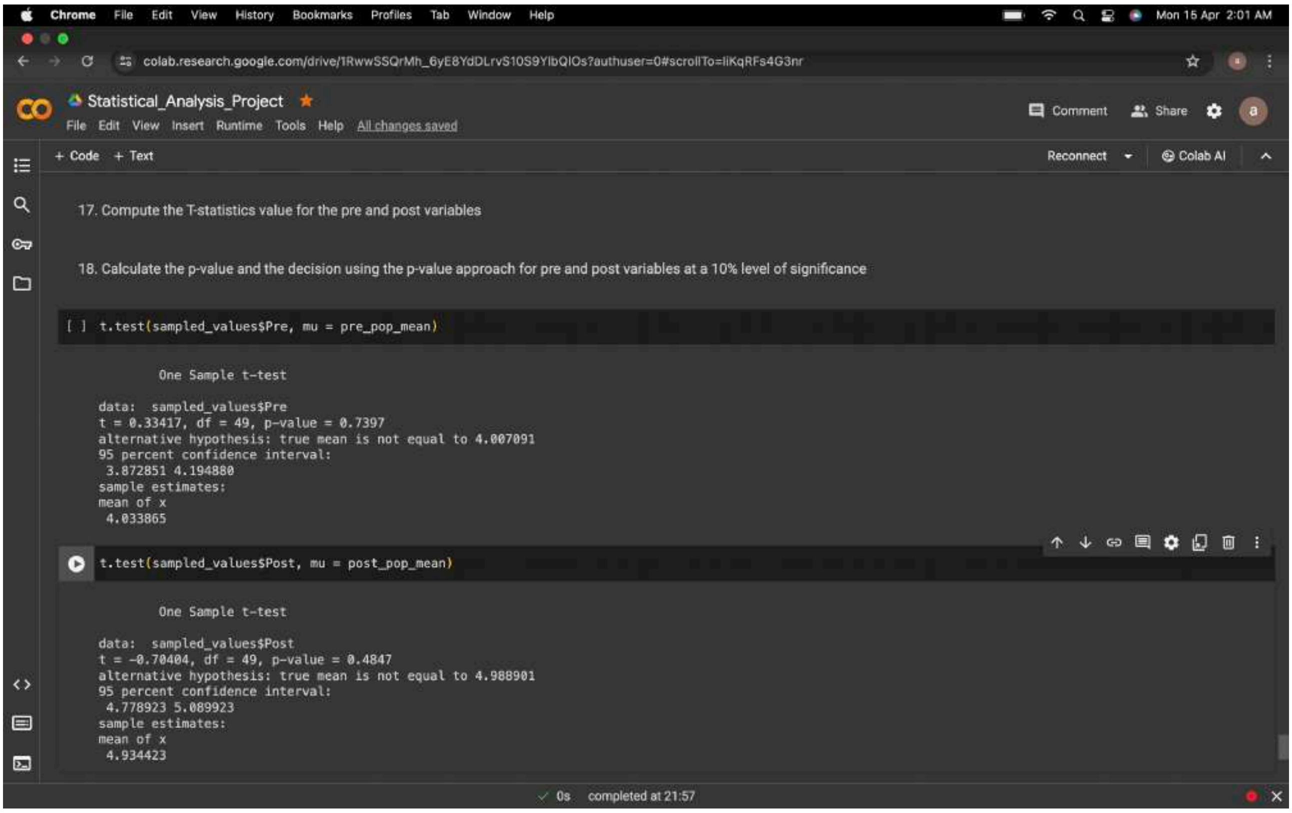


15. Using the p-value method, determine whether the sample mean for the pre and post variables differs significantly from the population mean at the 10% significance level



16. Calculate the critical Z value for the 10% level of significance and the decision rule using the critical value approach

17. Compute the T-statistics value for the pre and post variables
18. Calculate the p-value and the decision using the p-value approach for pre and post variables at a 10% level of significance



19. Calculate the critical T value for the level of significance of 10% and the decision rule using the critical value approach