Question 01

Suppose the numbers of calories in 10 different brands of chocolate milk of $244mL$ are: $164, 182, 176, 149, 184, 190, 160, 139, 175, 148.$ Assume these numbers are the observed values from a random sample of ten independent normal random variables with mean μ and variance σ^2, both unknown. Find a 95 confidence interval for the mean calories μ.

Confidence intervals are calculated using the formula:

$$\bar{X} \pm k\frac{\sigma}{\sqrt{n}}$$

When the population standard deviation is unknown, the sample standard deviation, s, is used instead.

The sample standard deviation of a sample is found using:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

For this sample:

$$s = \sqrt{\frac{(164-166.7)^2 + (182-166.7)^2 + (176-166.7)^2 + (149-166.7)^2 + (184-166.7}{}}$$

When the population mean is unknown, the t-distribution is used to calculate the critical value. Thus, the formula for the confidence interval becomes:

$$\bar{X} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

To calculate the critical value $t_{\alpha/2,n-1}$, we can use the inverse cumulative distribution function of the t-distribution with $n-1$ degrees of freedom and a significance level of $\alpha/2$:

$$t_{\alpha/2,n-1} = t_{0.025,9} = 2.262$$

The sample mean is calculated as:

$$\bar{X} = \frac{164 + 182 + 176 + 149 + 184 + 190 + 160 + 139 + 175 + 148}{10} = 166.7$$

Substituting the values into the formula for the confidence interval:

$$166.7 \pm 2.262 \times \frac{16.9}{\sqrt{10}} \approx \boxed{(154.762, 178.638)}$$

Question 02

> Let $X_1, \ldots, X_n$ be a random sample from the normal distribution with unknown mean μ and known variance σ^2. How large a random sample must be taken so that 90% confidence interval has length less than 0.02σ?

The length of the confidence interval is given by:

$$2 \times k \times \frac{\sigma}{\sqrt{n}}$$

Where k is the critical value of the distribution.

Since the population variance is known, extra variability is already accounted for and the z-distribution can be used for the test statistic. The critical value for a 90% confidence interval is:

$$z_{0.05} = 1.645$$

Substituting the values into the formula for the confidence interval length:

$$2 \times 1.645 \times \frac{\sigma}{\sqrt{n}} < 0.02\sigma$$

Solving for n:

$$\sqrt{n} > \frac{2 \times 1.645}{0.02} = 164.5 \implies n > 164.5^2 \approx 27060.25$$

Therefore, a random sample of at least $\boxed{27061}$ observations must be taken to ensure that the 90% confidence interval has a length less than 0.02σ.

Question 03

> Consider the setting in question 1, except that we now assume a known variance of $\sigma^2 = 16$. Suppose we wish to test the hypotheses: $H_0 : \mu = 170, H_A : \mu \neq 170$. Determine whether the test rejects H_0 at significance 0.05.

Since the variance is known, the z-distribution can be used to calculate the test statistic:

$$z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{166.7 - 170}{\frac{4}{\sqrt{10}}} = -1.75$$

We can find the probability of observing a value less than or equal to -1.75 in the standard normal distribution:

$$P(Z \leq -1.75) = 0.0401$$

Thus, the minimum probability such that the null hypothesis can be rejected is 0.0401. Since $0.0401 < 0.05$, we can reject the null hypothesis H_0 at a significance level of 0.05.

Question 04

Suppose that nine observations are selected at random from the normal distribution with unknown mean μ and unknown variance σ^2, and for these nine observations it is found that $X_n = 20$ and $\sum ni = 1(X_i - X_n)^2 = 70$. Find p-value of the test with hypotheses: $H_0 : \mu \leq 18, H_A : \mu > 18$.

The variance is given as:

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = 70$$

To find the sample standard deviation, we can use the formula:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

Substituting the values:

$$s = \sqrt{\frac{70}{9-1}} = \sqrt{\frac{70}{8}} = \sqrt{8.75} = 2.96$$

To find the test statistic, we should use the t-distribution since the population variance is unknown. The value is given by:

$$t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{20 - 18}{\frac{2.96}{\sqrt{9}}} = \frac{2}{0.987} = 2.03$$

The probability of obtaining this test statistic or a more extreme value (since H_0 is $\mu \leq 18$, “more extreme” means greater than 2.03) can be found using the t-distribution with $n - 1 = 8$ degrees of freedom:

$$P(T > 2.03) \approx 0.038$$

The minimum probability such that the null hypothesis can be rejected is 0.038. Therefore the p-value of the test is $\boxed{0.038}$.

Question 05

An experiment is carried out to see if there is any relation between a person's age and whether the person actively uses social media. Suppose that 100 people, 18 years of age or older, are selected at random, and each person is classified according to whether or not they are between 18 and 30 years of age and also according to whether or not they actively use social media. The observed numbers are given in the table below. Test the hypothesis that there is no relationship between a person's age and whether they actively use social media.

	Active social media user	Not active social media user	Total
Between 18 and 30	18	26	44
Over 30	8	48	56
Total	26	74	100

The test is:

H_0 : There is no relationship between a person's age and whether they actively use social med

H_A : There is a relationship between a person's age and whether they actively use social medi

We will use a significance level of 0.05.

Since we are testing categorical data, we can use the chi-squared test.
The test statistic is given by:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Where:

- O_{ij} is the observed frequency in cell (i, j)
- E_{ij} is the expected frequency in cell (i, j)
- r is the number of rows
- c is the number of columns

Since the events are independent, the expected frequency for each cell in the table can be calculated as:

$$E_{ij} = \frac{R_i \times C_j}{N}$$

Where:

- R_i is the total of row i
- C_j is the total of column j
- N is the total number of observations.

Table of expected frequencies:

	Active social media user	Not active social media user	Total
Between 18 and 30	11.44	32.56	44
Over 30	14.56	41.44	56
Total	26	74	100

Substituting the values into the formula for the chi-squared test statistic:

$$\chi^2 = \frac{(18-11.44)^2}{11.44} + \frac{(26-32.56)^2}{32.56} + \frac{(8-14.56)^2}{14.56} + \frac{(48-41.44)^2}{41.44} \approx 9.078$$

The degrees of freedom for the chi-squared test is given by:

$$df = (r-1) \times (c-1) = (2-1) \times (2-1) = 1$$

The p-value for the chi-squared test can be found using the chi-squared distribution with 1 degree of freedom:

$$P(\chi^2 > 9.078) \approx 0.0026$$

Since $0.0026 < 0.05$, we reject the null hypothesis H_0 at a significance level of 0.05. Therefore, there is evidence to suggest that there is a relationship between a person's age and whether they actively use social media.

Question 06

> What is the most common speaker occupation in the dataset?

```python
import pandas as pd
import json

CSV = "https://raw.githubusercontent.com/cpethe/TED_Talks/master/ted_main.csv"
df = pd.read_csv(CSV)

df['speaker_occupation'].value_counts().idxmax()
# >>> 'Writer'
```

The most common speaker occupation in the dataset is `Writer`.

Question 07

> Drop a column from the dataframe that is uninformative (information already contained in other columns), so that the dataframe no longer contains that column.

```python
if 'name' in df.columns:
  df.drop('name', axis=1, inplace=True)
```

The column `name` is uninformative because it is just a concatenation of the `main_speaker` and the `title` columns and therefore redundant. Therefore, I have dropped the `name` column from the dataframe.

Question 08

Get the rows corresponding to talks about climate change.

```
talks_abt_climate_change = df[df["tags"].apply(lambda x: "climate change" in x)]
talks_abt_climate_change
```

There are 87 rows corresponding to talks about climate change:

```
      comments                                      description  duration  \
1          265  With the same humor and humanity he exuded in ...       977
25         184  Legendary scientist David Deutsch puts theoret...      1140
38          57  Arctic explorer Ben Saunders recounts his harr...      1083
51         499  Given $50 billion to spend, which would you so...      1001
54         203  Speaking as both an astronomer and "a concerne...      1046
...        ...                                              ...       ...
2478        31  Anab Jain brings the future to life, creating ...       881
2486        12  Rivers are one of nature's most powerful force...       668
2488        26  Climate change is real, case closed. But there...       787
2497        17  Corals in the Pacific Ocean have been dying at...       434
2534         2  What the astronauts felt when they saw Earth f...       725

               event   film_date  languages      main_speaker  num_speaker  \
1            TED2006  1140825600         43           Al Gore            1
25    TEDGlobal 2005  1121299200         29     David Deutsch            1
38           TED2005  1109203200         26      Ben Saunders            1
51           TED2005  1107302400         32     Bjorn Lomborg            1
54    TEDGlobal 2005  1121299200         29       Martin Rees            1
...              ...         ...        ...               ...          ...
2478         TED2017  1492992000         10         Anab Jain            1
2486         TEDxPSU  1393718400         12         Liz Hajek            1
2488         TED2017  1492992000         10       Kate Marvel            1
2497         TED2017  1492992000         12  Kristen Marhaver            1
2534       TEDxSkoll  1491523200          1    Benjamin Grant            1

      published_date                                            ratings  \
1         1151367060  [{'id': 7, 'name': 'Funny', 'count': 544}, {'i...
25        1158019860  [{'id': 9, 'name': 'Ingenious', 'count': 269},...
38        1161735060  [{'id': 7, 'name': 'Funny', 'count': 80}, {'id...
51        1167696660  [{'id': 3, 'name': 'Courageous', 'count': 283}...
54        1168992660  [{'id': 1, 'name': 'Beautiful', 'count': 214},...
...              ...                                                ...
2478      1497884701  [{'id': 1, 'name': 'Beautiful', 'count': 47}, ...
2486      1499957123  [{'id': 10, 'name': 'Inspiring', 'count': 11},...
2488      1500303942  [{'id': 24, 'name': 'Persuasive', 'count': 20}...
2497      1501253483  [{'id': 23, 'name': 'Jaw-dropping', 'count': 1...
```

```
2534      1504814438  [{'id': 10, 'name': 'Inspiring', 'count': 46},...

                                          related_talks    speaker_occupation  \
1     [{'id': 243, 'hero': 'https://pe.tedcdn.com/im...      Climate advocate
25    [{'id': 2237, 'hero': 'https://pe.tedcdn.com/i...     Quantum physicist
38    [{'id': 2292, 'hero': 'https://pe.tedcdn.com/i...       Arctic explorer
51    [{'id': 248, 'hero': 'https://pe.tedcdn.com/im...    Global prioritizer
54    [{'id': 167, 'hero': 'https://pe.tedcdn.com/im...        Astrophysicist
...                                                 ...                   ...
2478  [{'id': 2858, 'hero': 'https://pe.tedcdn.com/i...    Futurist, designer
2486  [{'id': 2424, 'hero': 'https://pe.tedcdn.com/i...          Geoscientist
2488  [{'id': 1763, 'hero': 'https://pe.tedcdn.com/i...     Climate scientist
2497  [{'id': 2385, 'hero': 'https://pe.tedcdn.com/i...  Coral reef biologist
2534  [{'id': 2511, 'hero': 'https://pe.tedcdn.com/i...                Author

                                                   tags  \
1     ['alternative energy', 'cars', 'climate change...
25    ['climate change', 'cosmos', 'culture', 'envir...
38    ['climate change', 'culture', 'exploration', '...
51    ['AIDS', 'Africa', 'business', 'choice', 'clim...
54    ['astronomy', 'climate change', 'complexity', ...
...                                                 ...
2478  ['AI', 'algorithm', 'cities', 'climate change'...
2486  ['TEDx', 'ancient world', 'climate change', 'e...
2488  ['Anthropocene', 'biosphere', 'climate change'...
2497  ['TED Fellows', 'animals', 'biology', 'climate...
2534  ['TEDx', 'art', 'climate change', 'environment...

                                                  title  \
1                           Averting the climate crisis
25          Chemical scum that dream of distant quasars
38                     Why did I ski to the North Pole?
51         Global priorities bigger than climate change
54                           Is this our final century?
...                                                 ...
2478           Why we need to imagine different futures
2486  What rivers can tell us about the earth's history
2488  Can clouds buy us more time to solve climate c...
2497              Why I still have hope for coral reefs
2534         What it feels like to see Earth from space

                                                    url    views
1     https://www.ted.com/talks/al_gore_on_averting_...  3200520
25    https://www.ted.com/talks/david_deutsch_on_our...  1096862
38    https://www.ted.com/talks/ben_saunders_skis_to...   745231
51    https://www.ted.com/talks/bjorn_lomborg_sets_g...  1391142
```

```
54    https://www.ted.com/talks/martin_rees_asks_is_...  2121177
...                                                 ...      ...
2478  https://www.ted.com/talks/anab_jain_why_we_nee...  1259603
2486  https://www.ted.com/talks/liz_hajek_what_river...  1031716
2488  https://www.ted.com/talks/kate_marvel_can_clou...   907844
2497  https://www.ted.com/talks/kristen_marhaver_why...   956539
2534  https://www.ted.com/talks/benjamin_grant_what_...   646174

[87 rows x 16 columns]
```

Question 09

Get the rows corresponding to 10 most lengthy talks with at least 10 million views or at least 3000 comments

---

```
df[
    (df["views"] >= 10_000_000) | (df["comments"] >= 3000)
].sort_values(by="duration", ascending=False).head(10)
```

Output view:

```
      comments                                        description  duration  \
96        6404  Richard Dawkins urges all atheists to openly s...      1750   
644       3356  Questions of good and evil, right and wrong ar...      1386   
1940      1355  "Public shaming as a blood sport has to stop,"...      1346   
5          672  Tony Robbins discusses the "invisible forces" ...      1305   
29         970  Dan Gilbert, author of "Stumbling on Happiness...      1276   
1346      2290  Body language affects how others see us, but i...      1262   
837       1927  Brené Brown studies human connection -- our ab...      1219   
596        296  In this highly personal talk from TEDMED, magi...      1219   
4          593  You've never seen data presented like this. Wi...      1190   
262        669  First, Keith Barry shows us how our brains can...      1189   

               event   film_date  languages     main_speaker  num_speaker  \
96           TED2002  1012608000         42  Richard Dawkins            1   
644          TED2010  1265846400         39       Sam Harris            1   
1940         TED2015  1426723200         41  Monica Lewinsky            1   
5            TED2006  1138838400         36     Tony Robbins            1   
29           TED2004  1075680000         43      Dan Gilbert            1   
1346  TEDGlobal 2012  1340668800         51        Amy Cuddy            1   
837      TEDxHouston  1275782400         52      Brené Brown            1   
596      TEDMED 2009  1256601600         34     David Blaine            1   
4            TED2006  1140566400         48     Hans Rosling            1   
262          TED2004  1075680000         28      Keith Barry            1   

      published_date                                            ratings  \
96        1176689220  [{'id': 3, 'name': 'Courageous', 'count': 3236...   
644       1269249180  [{'id': 8, 'name': 'Informative', 'count': 923...   
1940      1426894031  [{'id': 3, 'name': 'Courageous', 'count': 8668...   
5         1151440680  [{'id': 7, 'name': 'Funny', 'count': 1102}, {'...   
29        1159229460  [{'id': 7, 'name': 'Funny', 'count': 1728}, {'...   
1346      1349103608  [{'id': 23, 'name': 'Jaw-dropping', 'count': 3...   
837       1293115500  [{'id': 10, 'name': 'Inspiring', 'count': 2144...   
596       1263889320  [{'id': 22, 'name': 'Fascinating', 'count': 91...   
4         1151440680  [{'id': 9, 'name': 'Ingenious', 'count': 3202}...   
262       1216366800  [{'id': 2, 'name': 'Confusing', 'count': 273},...   
```

```
                                          related_talks  \
96    [{'id': 86, 'hero': 'https://pe.tedcdn.com/ima...
644   [{'id': 666, 'hero': 'https://pe.tedcdn.com/im...
1940  [{'id': 2073, 'hero': 'https://pe.tedcdn.com/i...
5     [{'id': 229, 'hero': 'https://pe.tedcdn.com/im...
29    [{'id': 944, 'hero': 'https://pe.tedcdn.com/im...
1346  [{'id': 605, 'hero': 'https://pe.tedcdn.com/im...
837   [{'id': 1391, 'hero': 'https://pe.tedcdn.com/i...
596   [{'id': 310, 'hero': 'https://pe.tedcdn.com/im...
4     [{'id': 2056, 'hero': 'https://pe.tedcdn.com/i...
262   [{'id': 1821, 'hero': 'https://pe.tedcdn.com/i...

                               speaker_occupation  \
96                         Evolutionary biologist
644                   Neuroscientist, philosopher
1940                              Social activist
5     Life coach; expert in leadership psychology
29                 Psychologist; happiness expert
1346                          Social psychologist
837                      Vulnerability researcher
596                 Illusionist, endurance artist
4            Global health expert; data visionary
262                                      Magician

                                                   tags  \
96    ['God', 'atheism', 'culture', 'religion', 'sci...
644   ['culture', 'evolutionary psychology', 'global...
1940  ['communication', 'media', 'social media', 'su...
5     ['business', 'culture', 'entertainment', 'goal...
29    ['TED Brain Trust', 'brain', 'choice', 'cultur...
1346  ['body language', 'brain', 'business', 'psycho...
837   ['TEDx', 'communication', 'culture', 'depressi...
596     ['biology', 'magic', 'medicine', 'performance']
4     ['Africa', 'Asia', 'Google', 'demo', 'economic...
262     ['brain', 'entertainment', 'illusion', 'magic']

                                         title  \
96                            Militant atheism
644         Science can answer moral questions
1940                        The price of shame
5                         Why we do what we do
29         The surprising science of happiness
1346  Your body language may shape who you are
837                 The power of vulnerability
596        How I held my breath for 17 minutes
```

```
4              The best stats you've ever seen
262                                Brain magic

                                                    url     views
96    https://www.ted.com/talks/richard_dawkins_on_m...   4374792
644   https://www.ted.com/talks/sam_harris_science_c...   3433437
1940  https://www.ted.com/talks/monica_lewinsky_the_...  11443190
5     https://www.ted.com/talks/tony_robbins_asks_wh...  20685401
29    https://www.ted.com/talks/dan_gilbert_asks_why...  14689301
1346  https://www.ted.com/talks/amy_cuddy_your_body_...  43155405
837   https://www.ted.com/talks/brene_brown_on_vulne...  31168150
596   https://www.ted.com/talks/david_blaine_how_i_h...  15601385
4     https://www.ted.com/talks/hans_rosling_shows_t...  12005869
262   https://www.ted.com/talks/keith_barry_does_bra...  13327101
```

Question 10

> Which talk is the most viewed as compared to its *related talks* (the one with the maximum difference between its views and the *view count* of any of its *related talks*)?

```
import ast

def compute_related_diff(row):
    related_views = []
    for related in ast.literal_eval(row["related_talks"]):
        related_views.append(df[df["title"] == related["title"]]["views"].values[0])
    return max([row["views"] - x for x in related_views]) if related_views else 0

df["related_diff"] = df.apply(compute_related_diff, axis=1)

# Sort DataFrame by the computed difference in descending order
df = df.sort_values(by="related_diff", ascending=False)

df.iloc[0][:-1]
```

The talk with the most viewed related talk is *Do schools kill creativity?* by Sir Ken Robinson.

```
comments                                                            4553
description           Sir Ken Robinson makes an entertaining and pro...
duration                                                            1164
event                                                            TED2006
film_date                                                     1140825600
languages                                                             60
main_speaker                                                Ken Robinson
num_speaker                                                            1
published_date                                                1151367060
ratings               [{'id': 7, 'name': 'Funny', 'count': 19645}, {...
related_talks         [{'id': 865, 'hero': 'https://pe.tedcdn.com/im...
speaker_occupation                                      Author/educator
tags                  ['children', 'creativity', 'culture', 'dance',...
title                                       Do schools kill creativity?
url                   https://www.ted.com/talks/ken_robinson_says_sc...
views                                                           47227110
Name: 0, dtype: object
```

Question 11

> Which pair of features (columns) are the most correlated? Comment on whether the correlation implies causation in this case with a few sentences

---

```
max_correlated, max_correlation = max(
    [
        ((col1, col2), df[col1].corr(df[col2]))
        for col1 in df.select_dtypes(exclude="object").columns
        for col2 in df.select_dtypes(exclude="object").columns
        if col1 != col2
    ],
    key=lambda x: x[1],
)
max_correlated, max_correlation
# >> (('related_diff', 'views'), 0.9784920041585558)
```

The two features that are most correlated are `related_diff` and `views` with a correlation coefficient of `0.9784920041585558`

The correlation in this case does not imply causation. Both dates are bound by a shared timeline rather than a causal relationship. The film date doesnt cause the publication; rather, both variables are dependent on decisions made within a larger production process. The dates being sequential makes them correlated but this correlation merely relfects that they are part of a shared timeline – not that one event is causing the other.

Question 12

> Research a way to find out the significance (in terms of p-value) of the correlation of a pair of features. Try it on the pair of columns *duration* and *comments*. What is the correlation coefficient and its p-value? Comment on what this finding implies with a few sentences

---

```python
from scipy.stats import pearsonr

correlation, p_value = pearsonr(df["duration"], df["comments"])

print(f"Correlation {correlation:.4f}\np-value: {p_value:.13f}")
# >> Correlation 0.1407
# >> p-value: 0.0000000000010
```

The correlation coefficient between the `duration` and `comments` columns is `0.1407`.

The p-value is about `0.0000000000010`.

The p-value here represents the probability under the null hypothesis (the two features are uncorrelated) of obtaining a correlation as or more extreme than the one computed from the datasets.

Since the probability of obtaining a correlation of 0.1407 is ~0.0000000000010, it’s safe to reject the null hypothesis and conclude that the correlation between duration and comments is statistically significant. I.e., the correlation is not due to random chance and exists earnestly in the population of TED talks.