Behavioral Research and Experiment Design
Assignment 2

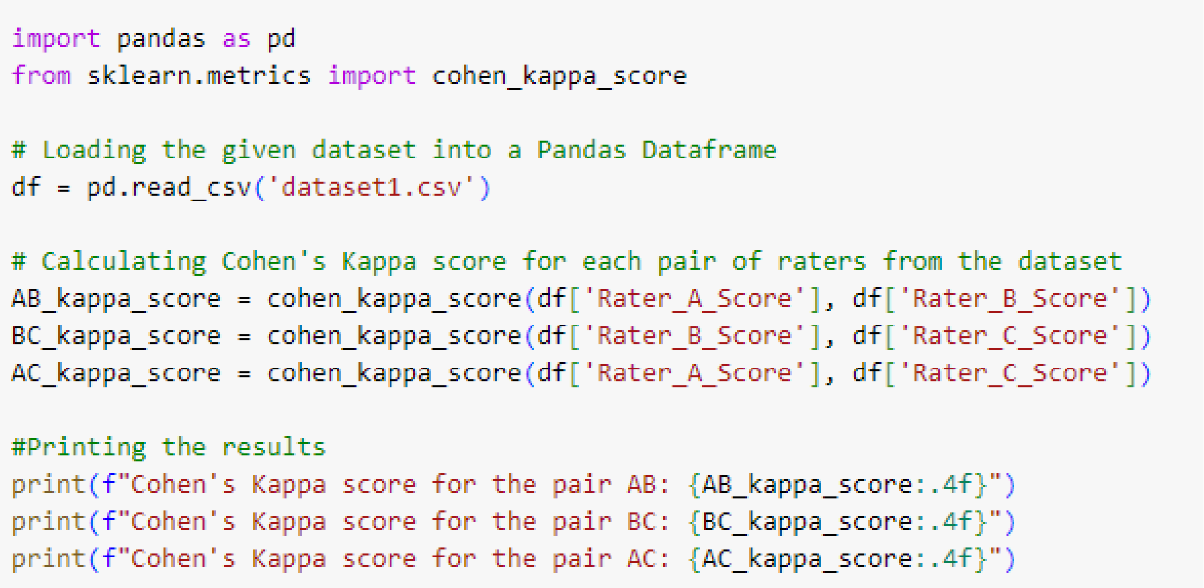
Aaryan Sharma
2020115008

Ans 1:

For the given observational study, we have data of three different raters (Rater A, Rater B and Rater C) which independently assess the aggressiveness of 50 participants on a scale of 1-5 while playing "Getting Over It with Bennett Foddy".

Now we must calculate the agreement between the three different raters. Now to calculate the agreement between different raters, we go ahead with Cohen's Kappa score which calculates the agreement between a pair of raters, and we make three different pairs i.e. A and B, B and C, and A and C. And now to calculate the Cohen's Kappa score, we use sklearn.metrics library and import the Cohen's Kappa score function, which is an inbuilt function to calculate the score.

And the code for the same is:



```
import pandas as pd
from sklearn.metrics import cohen_kappa_score

# Loading the given dataset into a Pandas Dataframe
df = pd.read_csv('dataset1.csv')

# Calculating Cohen's Kappa score for each pair of raters from the dataset
AB_kappa_score = cohen_kappa_score(df['Rater_A_Score'], df['Rater_B_Score'])
BC_kappa_score = cohen_kappa_score(df['Rater_B_Score'], df['Rater_C_Score'])
AC_kappa_score = cohen_kappa_score(df['Rater_A_Score'], df['Rater_C_Score'])

#Printing the results
print(f"Cohen's Kappa score for the pair AB: {AB_kappa_score:.4f}")
print(f"Cohen's Kappa score for the pair BC: {BC_kappa_score:.4f}")
print(f"Cohen's Kappa score for the pair AC: {AC_kappa_score:.4f}")
```

And the results we get are as follows:

- **Cohen's Kappa score for the pair AB: 0.6719**
- **Cohen's Kappa score for the pair BC: 0.0594**
- **Cohen's Kappa score for the pair AC: 0.0308**

And the value of Kappa coefficient between A and B is 0.672, which is a substantial agreement between both raters, which correlates with higher agreement between the raters. In case of the pair B and C and pair A and C, the value of Kappa coefficient is 0.0594 and 0.308 respectively. This shows that the value of Kappa coefficient between these two pair is less than 0.2, i.e. there is negligible agreement between these pairs and the there exists a high variation between the raters in both pairs. Now, we can see that A and B have a high agreement, A and C have minimal agreement and the pair B and C also have negligible agreement, which shows that the rater C is an outlier and the rater C has high variations with other raters, therefore the rater A and B are more reliable than rater C.

Ans 2:

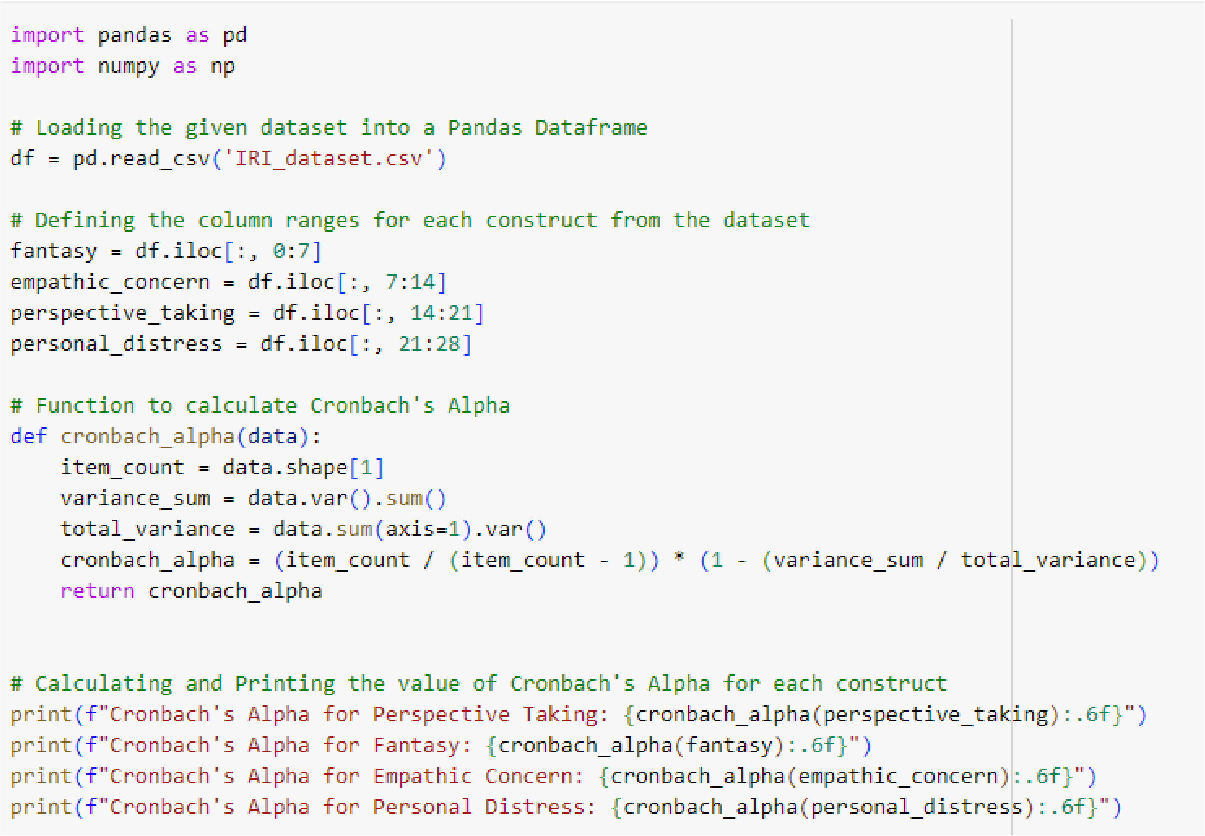
For the given dataset, which contains the responses based on a Likert scale to measure various aspects of empathy. We have 28 questions in the questionnaire which aims to assess the four different constructs i.e. perspective taking, fantasy, empathic concern and personal distress with 7 questions to assess each construct.

Now, we need to analyse the dataset and examine the internal consistency of the questionnaire. Therefore, we will go ahead with calculating the Cronbach's Alpha, which is a statistic used to measure the internal consistency and reliability of a set of survey questions. It is a coefficient that quantifies how closely a set of items are closely related. The value of Cronbach's Alpha ranges from 0 to 1, with higher values indicate a higher consistency.

$$\alpha = \frac{N\bar{c}}{\bar{v} + (N-1)\bar{c}}$$

Where N is equal to number of items, c̅ is the average inter-item covariance among the items and Ṽ equals average variance.

Now, to calculate the Cronbach's Alpha for the internal consistency of the dataset, we go ahead with the following code:



```python
import pandas as pd
import numpy as np

# Loading the given dataset into a Pandas Dataframe
df = pd.read_csv('IRI_dataset.csv')

# Defining the column ranges for each construct from the dataset
fantasy = df.iloc[:, 0:7]
empathic_concern = df.iloc[:, 7:14]
perspective_taking = df.iloc[:, 14:21]
personal_distress = df.iloc[:, 21:28]

# Function to calculate Cronbach's Alpha
def cronbach_alpha(data):
    item_count = data.shape[1]
    variance_sum = data.var().sum()
    total_variance = data.sum(axis=1).var()
    cronbach_alpha = (item_count / (item_count - 1)) * (1 - (variance_sum / total_variance))
    return cronbach_alpha


# Calculating and Printing the value of Cronbach's Alpha for each construct
print(f"Cronbach's Alpha for Perspective Taking: {cronbach_alpha(perspective_taking):.6f}")
print(f"Cronbach's Alpha for Fantasy: {cronbach_alpha(fantasy):.6f}")
print(f"Cronbach's Alpha for Empathic Concern: {cronbach_alpha(empathic_concern):.6f}")
print(f"Cronbach's Alpha for Personal Distress: {cronbach_alpha(personal_distress):.6f}")
```

And the output for the given code is following:

```
•  Cronbach's Alpha for Perspective Taking: 0.738509
•  Cronbach's Alpha for Fantasy: 0.765548
•  Cronbach's Alpha for Empathic Concern: 0.810251
•  Cronbach's Alpha for Personal Distress: 0.770565
```

Here the Cronbach’s Alpha score for Perspective Taking, Fantasy and Personal Distress is **`0.738509, 0.765548 and 0.770565`** respectively, where the score of between 0.7 and 0.8 is acceptable, whereas the Cronbach’s Alpha score for Empathic Concern is **`0.810251,`** which is a good score. Therefore we can say that the dataset contains acceptable internal consistency and is a reliable dataset.

Ans 3:

For the following list of reasons, we can say that the given questionnaire is inappropriate for assessing the customer satisfaction at a local supermarket:

1. The usage of **double negatives** in q3: “Don't you think our prices are not unreasonable?”, where the question makes it challenging for the respondent to interpret the question and answer appropriately. Such question with double negative can be replace with a straightforward question such as “Do you think our prices are reasonable”, where the respondent has more clarity, and the answer would be more objective and binary in nature.
2. The questionnaire contains **leading questions** which can influence the responses of the customers by asking them polarised and leading questions instead of neutral questions. For example, in q1 (How satisfied are you with our store's excellent service and friendly staff?) the questionnaire establishes the store’s service as excellent and the staff’s behaviour as friendly. The similar leading questions can be seen in q5 (How much do you like our store's fresh produce section, which is the best in town?) and q6 (Our store's prices are competitive, right?), where both the questions are not at all neutral in nature.
3. In q4 (How many times did you visit our store last year? If you can't remember, please guess.), the questionnaire aims to **tax the respondent’s memory** and **in the options the division is not discrete** in nature. For example, if you have visited the store 10 times, then you can go ahead with option a as well as option b.
4. Moreover, the options in q2 present a **Likert scale**. However, scale doesn’t have equal extremities in the scale. On one end, we have extremely likely, whereas the other extremity is very unlikely instead of extremely unlikely.

In the questionnaire aimed towards assessing the customer satisfaction, we need to have a certain number of subjective and open-ended questions which allows the respondents to share any specific incident, which is also absent from the questionnaire. Moreover, the questionnaire is not detailed enough and does not provide the respondent with the information of why the data is being collected and how the data would be used further. Moreover, the subjects in the question is limited and repetitive, therefore we should not go ahead with the given questionnaire.

Ans 4:

For each of the given experiment designs, we will follow the following sampling methods:

a. We need to investigate the effectiveness of a new drug on patients with a specific medical condition, where we aim to collect data from patients from various age groups from different geographic regions, therefore we would go ahead with multi-stage **stratified sampling.**
Stratified sampling involves dividing the whole population into strata based on certain characteristics, and then randomly selecting samples from each stratum. In the given experiment design, we would create strata based on **age groups** (11-20, 21-30, 31-40 … so on) to ensure representation from different age groups and to represent different **geographic regions** in every stratum, we would create sub-strata based on different states or (regions like north, east, west etc) to ensure geographical diversity. This **multi-stage stratified sampling** method would allow us to account for diversity in age groups as well as geographic diversity.

b. To assess products' quality on the assembly line in a manufacturing company, while ensuring that the sampled products are representative of the production process, we should go ahead with **systematic sampling,** as this sampling method ensures that samples are selected at regular intervals, which can be seen as a representative of the throughout production process.
To implement systematic sampling in an assembly line, we would create a list of or sequence of products in assembly line and then we would select every nth item for quality assessment. For example, if the assembly line produces 10000 items/hour and we must select 100 items for sample, then we would select every 100th item for quality assessment. Also, we will have a random starting point in the line, so we can avoid bias towards any production phase/cycle.

c. As we want to ensure comprehensive feedback from students of different academic majors and years of study, the most appropriate sampling method would be **stratified sampling**, where we would divide the student community into different strata based on their **academic major and year of study**, and then randomly select the samples from these strata to statistically represent the student community.
Firstly, we create the strata based on the **academic majors** (Computer Science, Electronics, Natural Sciences etc.), which ensures that we have stratification across academic disciplines. Furthermore, we stratify each of the major stratum based on the **year of study** (freshman, sophomore, junior, senior etc) to get insights into the potential variation across years. Then we will proportionally select students from each stratum (major and year), which would result in an actual representation of the student community.

d. To efficiently collect feedback from the customers regarding their online shopping experience for an e-commerce company which has a country-wide customer base, we would like to go ahead with **multi-level sampling** method with a short and targeted surveys, which ensures a good representation of the customer base while minimizing the respondent's burden and maximizing the response rate among the customers.
We will first go ahead with **cluster sampling** where we divide the customers in pre-existing demographic boundaries (such as districts, states etc). Then we can have another level of **stratified sampling** where we divide the customer base into different strata based on age and sex. And then we randomly select a representative sample of regions, ensuring proportionality to the customer distribution, and within each such region we can then randomly select the customers from the company's database and send them direct links for surveys while incentivising the customers for response in form of minor cashbacks or coupons.