MA 385 Midterm Project

Salwa Jeries
(Worked with Candice Belcher + Ethan Simpson)

1) a. Irrelevant variables can be determined by assessing the p-values of each one. With a p-value greater than 0.05, we can assume that the variable is irrelevant because this means that the probability under the null hypothesis of getting a more extreme or equal value for that variable is high. Looking at the p-values of the traits for the Relationship Satisfaction Score, we can safely say that Openness and Extroversion are statistically irrelevant, as they have p-values of 0.709 and 0.313 respectively. We could tentatively say that Agreeableness is statistically irrelevant since it has a p-value of 0.073, although this is really close to the 0.05 soft limit we set. The traits of Neuroticism and Conscientiousness would be statistically relevant since they have very low p-values, below 0.05.

b. The coefficients of each of the variables in the Relationship Satisfaction Score equation demonstrates the impact of each one on the resulting score. A positive coefficient results in a positive impact on the score, while a negative coefficient results in a negative impact. Based on this, we can assume that traits like Openness, Conscientiousness, and Agreeableness should be maximized because their positive coefficients would result in an increased in the Relationship Satisfaction Score. On the other hand, traits like Neuroticism and Extraversion should be minimized since their negative coefficients have a negative impact on the overall score.

c. R^2 tells us how much the model explains the variability in the Relationship Satisfaction Score, or in other words “goodness of fit”. The R^2 score for this dataset is 94.35%, which means that 94.35% of the variability can be explained by the model

d. The residual plots provided alongside the data for this question have a relatively “normal” distribution shape, for both the Histogram and Normal Probability Plot. This suggests that the regression assumptions are correct. Looking at the Versus Plots also helps determine if the regression assumption are met. A randomly distributed group of points that is centered around 0 verifies this, which is what we see in the Versus plots for this data. There are no distinct outliers, and there appears to be enough randomness in the values to demonstrate heteroscedasticity.

e. The high R^2 value of 94.35% for this model suggests the overall adequacy of the linear model that we built. Because R^2 demonstrates the “goodness of fit” of the model for out data relationship, the high R^2 value solidifies this adequacy.

f. Your projected Relationship Satisfaction Score is calculated below, with your “Big Five” traits data used to perform this calculation. Your score, based on the Relationship Satisfaction Equation, is 53.62. Based on the metrics provided earlier in the question, since this score is below 70, the model suggests that you are not emotionally in a good place for a “good relationship”. Earlier, we noted that traits like Openness, Conscientiousness, and Agreeableness are traits that should be maximized in order to maximize the Satisfaction Score. Your Conscientiousness score of 100 is already maximized, but your Openness and Agreeableness are 70 and 60, respectively. Focusing on maximizing these traits as much as possible would help to increase your score. Further, we noted that Extroversion and Neuroticism have a negative impact on Satisfaction Score. Neuroticism especially has a very large negative impact on the score, and your score of 90 is extremely high and concerning. For your sake and for the sake of improving the Satisfaction Score, focusing on minimizing these traits would significantly improve your score. Working on improving these traits in their respective magnitudes will provide the best method of improving your Relationship Satisfaction Score. However, you should always ask yourself if you feel mentally and emotionally ready to be in a relationship first and foremost, regardless of what a “Satisfaction Score” tells you!

2) a. To visually depict the arsenic concentration data, I generated a histogram of all the concentrations and their frequencies. The histogram and the python code to generate this are provided below.

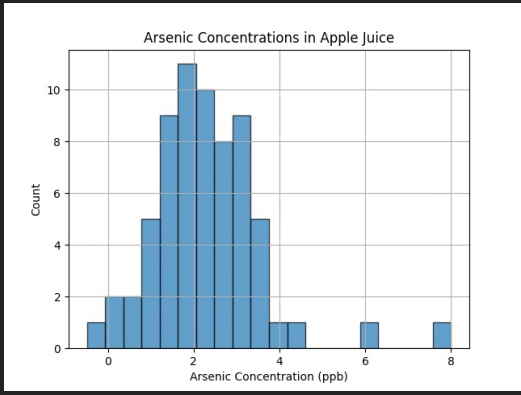

Fig 1: Arsenic Concentrations in Apple Juice Histogram

```python
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from scipy import stats

np.set_printoptions(suppress=True)      # Turn off scientific notation format for data output

# Concentration Data
concentrations = np.array([ 1.49, …
years = np.array([ 2019, …

# Generate histogram of concentrations
plt.hist(concentrations, bins=20, edgecolor='k', alpha=0.7)
plt.xlabel('Arsenic Concentration (ppb)')
plt.ylabel('Count')
plt.title('Arsenic Concentrations in Apple Juice')
plt.grid(True)
plt.show()      # Display histogram
```

Fig 2: Python code to generate histogram

The histogram is right skewed, meaning that most of the arsenic concentration data is clustered on the left. This displays that the vast majority of the arsenic concentration samples taken have low ppb levels. There are some extreme or outlier samples with higher-than-normal ppb levels. These are what generate the right skewed shape of the histogram, as these outliers are located on the right and create the “tail” of the graph. Because of this skewed shape, we can also assert that the data does not follow a normal distribution.

b. For 2019, the confidence interval’s upper limit is about 3.09. The following year (2020), it decreases to 2.07. However, every year following 2020 increases, reaching about 3.91 in 2023. This suggests that the arsenic concentrations seem to be increasing in more recent years. Therefore, apple juice appears to be getting less safe over time. Still, the interval appears to widen slightly over time, meaning that there is a wider range of arsenic levels over time and not necessarily just higher concentrations.

```
Mean:  2.2980303030303033
Standard Dev:  1.2855184164638012

Year 2019 - Confidence Interval:  (2.0534089484720432, 3.093649875057369)  Mean: 2.573529  Std: 1.011607  n: 17
Year 2020 - Confidence Interval: (1.3091626404529815, 2.0708373595470184)  Mean: 1.690000  Std: 0.411785  n:  7
Year 2021 - Confidence Interval:   (1.207032194936529, 2.432967805063471)  Mean: 1.820000  Std: 1.061632  n: 14
Year 2022 - Confidence Interval:    (1.5385478758826001, 2.7414521241174)  Mean: 2.140000  Std: 0.995297  n: 13
Year 2023 - Confidence Interval:   (1.790848262903348, 3.914485070429985)  Mean: 2.852667  Std: 1.917396  n: 15
```

Figure 3: Confidence intervals data for each year (2019 – 2023)

```python
# Sort data by year
sorted_data = {
    2019: [ 1.49, …
    2020: [ 1.96, …
    2021: [ 0.46, …
    2022: [ 4.27, …
    2023: [ 3.25, …
}

confidence_level = 0.95     # Set confidence level to 95%

# Print overall dataset stats
print()
print('Mean: ', np.mean(concentrations))
print('Standard Dev: ', np.std(concentrations, ddof=1))
print()

# Calculate confidence interval for each year
for year, data in sorted_data.items():
    n = len(data)
    mean = np.mean(data)
    standard_dev = np.std(data, ddof=1)
    t_critical = stats.t.ppf(1 - (1 - confidence_level) / 2, df=(n - 1))
    m_o_e = t_critical * (standard_dev / np.sqrt(n))    # Margin of Error
    confidence_interval = (mean - m_o_e, mean + m_o_e)

    # Print stats
    print("Year %4d - Confidence Interval: %40s  Mean: %5f  Std: %5f  n: %2d" % (year, confidence_interval, mean, standard_dev, n))

print()
```

Figure 4: Python code to calculate confidence intervals for each year

c. Since in general, p > 0.05, we fail to reject the null hypothesis (H0). The only year where we reject H0 is in 2020, in which the level of arsenic concentration was vastly different (seen in the confidence intervals, too). Therefore, in general, this data matches the generally accepted consensus of 2.25 ppb for each year.

Hypothesis Test:

$H_0: \bar{x} = 2.25$ ppb
$H_1: \bar{x} \neq 2.25$ ppb

$\alpha = 0.05 \longrightarrow$ 95% confidence level

T-Test:

$2019 \longrightarrow t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{2.57 - 2.25}{1.01/\sqrt{17}} = 1.306 \longrightarrow p = 0.1 > 0.05$

$2020 \longrightarrow t = \frac{1.69 - 2.25}{0.41/\sqrt{7}} = -3.73 \longrightarrow p = 0.005 < 0.05$

$2021 \longrightarrow t = \frac{1.82 - 2.25}{1.06/\sqrt{14}} = -1.54 \longrightarrow p = 0.075 > 0.05$

$2022 \longrightarrow t = \frac{2.14 - 2.25}{0.99/\sqrt{13}} = -0.41 \longrightarrow p = 0.1 > 0.05$

$2023 \longrightarrow t = \frac{2.85 - 2.25}{1.92/\sqrt{15}} = 1.22 \longrightarrow p = 0.1 > 0.05$

d. There are several factors that may contribute to the tighter confidence interval of (1.207032194936529, 2.432967805063471) in 2021. The sample size of 14 data points, larger than many of the other years, can result in more precise results than a smaller sample size. This could lead to the narrower interval, versus a smaller sample size that may have a very wide range of data points already. This year also has one of the higher standard deviations (1.062), which is an indication that values in this sample set have less variability. In other words, the data values are more concentrated together, resulting in a tighter confidence interval. Further, a sample mean (1.82) that is closer to the population mean (2.29) results in a tighter confidence interval.

e. Values that are not physically possible should not be considered when performing analysis on the dataset, so these values should be removed. This means that the negative concentration values should be taken out of the data set. Outliers should also be removed from the dataset's considerations, as they are not representative of the majority. We can use a z-score to determine if certain value are outliers or not. We can use the following equation to evaluate the possible outlier values, 6 and 8:

$$z = \frac{x - \mu}{\sigma}$$

We can use the sample mean of 2.298 and standard deviation of 1.286 (calculated in the program) to calculate the z-scores for each of these values:

$$z = \frac{6 - 2.298}{1.286} = 2.879 \qquad z = \frac{8 - 2.298}{1.286} = 4.434$$

For values with a z-score outside the range of (-3, 3), we can assert that these values are outliers. For the value of 6, we got a z-score of 2.879, which is within this range of (-3, 3) so we should keep the value as it is statistically relevant. However, the value of 8 has a z-score of 4.434, which is well beyond the range of (-3, 3) and therefore we can treat it as an outlier and consider it statistically irrelevant to the dataset. We can remove this value from the rest of the calculations.

3) PART 1:

The following categories are provided in the dataset that are included in calculating the ranking of each of the schools:

- Teaching - this indicates how effective the teaching of the faculty is for the student population
- Research - relates to everything involved with research, including funding, amount of research published, and opportunities available to students to participate in and contribute to this research
- International Reputation - the reputation of the school to the general public, especially as seen by other schools, employers, research founders, etc.
- Service - public community service opportunities available, volunteer work, etc.
- Student Income - average amount of income students make after graduating with a degree from this school
- Num Students - total number of students that attend this school
- Student Staff Ratio - ratio of how many students attend the school to the number of faculty/ professors
- Percent International - percentage of the student population that are international students
- Percent Female - percentage of the student population that are female

PART 2:

In order to determine how to focus the university's efforts to increase the national ranking, it would be beneficial to compare UA's scores for the above categories to the national averages. Being above the national average will improve UA's rank as well as provide enticing statistics for potential incoming students.

Category	National Mean	UA
Teaching	41.55	27.2
Research	36.20	19.1
International Reputation	49.89	31
Service	69.07	54.2
Student Income	46.47	37.5
NumStudents	22901	35457
StudentStaffRatio	14.3	20.9
% International	13	4
% Female	54	59

This table compares the UA data from the provided dataset to the national means of these categories among universities in the US. UA has scores below the national mean for seven of the nine categories: Teaching, Research, International Reputation, Service, Student Income, and Percent International. This will significantly impact UA's ranking in a negative way, and the fact that it is below average on such a large number of categories makes it less appealing to potential incoming students. In regards to the three pillars of higher education (Teaching, Research, and Service), UA is below average for all of these categories. To improve these categories, it would be beneficial to improve the faculty at the university and listen to the students' feedback on the teaching performance of the professors. Research is also subjective to the professors at the university that are conducting said research, so there should be some balance of teaching performance and research contributions evaluated to determine the overall performance and value of the professors. Research can also be promoted by ensuring funding is sufficient, and focusing on research topics that are more beneficial to the public and that provide students with plenty of opportunities. Service can be increased by encouraging student organizations that participate in volunteer work, as well as promoting service opportunities as a university in general.

PART 3:

International Reputation appears to be more in line with the core values of the university. International reputation refers to the view of the school as seen by other universities, the general public, employers, and research funders. Having a better international reputation indicates that the rest of the categories are also higher. For example, it signifies better research which means that research funders are more likely to continue funding or provide new funding if they know that the money will go towards well-reviewed, published research projects/papers. This also means that the faculty and staff have a good reputation as good teachers, raising the Teaching score. Student Income is still relevant towards the national ranking, but this is not symbolic of the core values of the university. In addition, a higher international reputation itself will improve student income since more employers will value the education of the university, and therefore be willing to pay higher starting salaries to students that graduate from the school.