DATASCI 101 Fall 2024 HW 1

Due: Wednesday, September 11 at 9pm
Your name: Brad McNew
Your uniqname: mcnew

Remember, you are allowed (and encouraged!) to work with other students on this assignment. All students must write their own answers, but you are free to discuss your approaches and assist each other with crafting the solutions.

Question 1 (5 points)

Read the article "Can we really teach machines to smell?".

For each of the main skills of data science identify a place in the article where the skill is used or mentioned. For each answer, write two or three sentences explaining what the skill is in general, and how it appears in the article.

Data Collection

Data collection involves gathering information from various sources to use in analysis.
In Tewari's class students collect data by smelling various molecules and recording their perceptions. This data is essential for training machine learning models to predict scents based on chemical structures.

Exploration

Exploration involves analyzing collected data to understand its structure, patterns, and relationships.
Students in Tewari's class explore the data they collected by experimenting with smell blotters and documenting what they think they are smelling. This helps them understand the complexities of odor perception and how these can be modeled using machine learning techniques.

Inference

Inference is the process of drawing conclusions from the data, often using statistical methods or models. It involves predicting outcomes or determining the relationship between variables.
The machine learning model created by Tewari and his students makes inferences by predicting human odor perception from a molecule's chemical structure.

Communication

Communication involves effectively conveying data science findings to different audiences, using visualizations, reports, or presentations.
Tewari's students learn to communicate their findings through their honors thesis projects, class presentations, and documentation of their research work. They also attend guest lectures and discussions.

Ethics

Ethics in data science involves consideration of the moral implications of data collection, analysis, and application.
Tewari's lectures cover the ethical issues in public discourse such as bias in AI models and the ethical implications of using machine learning in sensitive applications.

Question 2 (5 points)

Visit the UCSD Library Data Science Projects Guide.

Question 2.a (2 points)

Review the embedded slides on how to start a data science project. Think about a data science project you might like to perform after completing this class. If you need help thinking of ideas, use the tools to generate areas of study and data types listed below.

Write a paragraph explaining your idea and why you think it would be fun or informative project. If you need some additional inspiration, you can look at the Guided Projects section.

A data science project I might like to perform after completing this class is a project where I analyze the relationship between gut microbiome diversity and different diets or health conditions using publicly available microbiome datasets. The study of the gut microbiome is a rapidly growing field with significant implications in many aspects of human health. Understanding the microbiome's composition and its relationship with diet and health conditions can provide insights into disease prevention, personalized nutrition, and treatment strategies. Using data science techniques to analyze microbiome data allows for uncovering hidden patterns, correlations, and potential causations that may not be evident through other analysis methods. Findings from this project can serve as a foundation for further research.

Question 2.b (2 points)

Review the example projects listed at the bottom of the projects guide. Select one example to summarize. Explain how the 5 skills of data science are used, or if you did not observe the use of a particular skill explain what where that skill could have been applied.

The projects “Football (soccer) match outcome prediction focuses on predicting the outcome of football matches (home win, away win, or draw) by analyzing various match-related data points. This involves handling missing and imbalanced data, building machine learning models to predict the probability of each outcome, and evaluating the models’ performance.

Data Collection - This project involves collecting data from multiple sources such as sports databases, APIs, and historical records.

Exploration - Exploring the data will help to understand the data structure, detect patterns, and identify correlations between different variables such as home-field advantage, player form, or team strategies

Inference - Models such as logistic regression, decision trees, random forests, or neural networks are used to predict match outcomes.

Communication - Communicating the model’s results and findings are especially important if the results are intended for stakeholders like sports analysts, bettors, or fans. Visualizations can effectively communicate model performance.

Ethics - Predictive models used for betting can raise ethical issues such as promoting gambling or exploiting vulnerable individuals.

Question 2.c (1 points)

Visit the Michigan Data Science Team’s list of projects from a previous semester. By clicking on the file icon, you can see some results of this project. Select a particular project and write a brief summary of the project and why it attracted your attention.

The project I selected was the Real vs Fake Faces project. The project does not make a true deep fake detector because it is too complicated for a single semester. Instead it solely focuses on determining whether a picture of a face is fake or not. The project uses neural networks and computer vision to accomplish this goal. This project attracted my attention because as AI image generation keeps advancing it can be hard to tell what photos are real and fake on the internet. This could raise serious ethical issues as in the near future real videos of people and fake videos or people could look almost exactly the same raising the question if we will even be able to trust videos for evidence of crimes.

Question 3 (5 points)

University of Michigan and Eastern Michigan University faculty recently published a paper in the Journal of Racial and Ethnic Health Disparities titled, “Understanding the Intersectionality of COVID-19 Racism, Mental Distress, Alcohol Use, and Firearm Purchase Behavior Among Asian Americans.”

Here is the abstract from the paper:

> Firearm-related injuries are a major public health concern in the USA. Given the increased racism endured by Asian Americans during the COVID-19 pandemic, the current study aims to investigate the direct and indirect effects of racism, mental distress, and substance use on firearm purchase among Asian Americans. To fulfill this purpose, we collected data from a national sample of 916 Asian Americans in 2021. The study results showed that Asian Americans’ racism experience is directly related to increased mental distress, substance abuse, and firearm purchase. Both mental distress and alcohol use were also linked to firearm purchase. It was found that racism links to more mental distress and increased alcohol use, which in turn link to increased firearm purchases. The findings add new information on how racism can have compounded effects on mental distress and alcohol use in addition to firearm-related risk behavior among Asian Americans and posing serious public health concerns.

You do not need to read the paper to answer this question, but if you would like to study it more, you can access it through the University’s library proxy.

Question 3.a (1 point)

Based on the abstract, how large was the sample in this study and what were the units? What population do you think we could imagine these units were drawn from?

The sample size in this study was 916 units.
The units were Asian Americans who participated in the study
The population from which these units were likely drawn could be the general population of Asian Americans living in the U.S. in 2021.

Question 3.b (2 points)

The authors report the following variables were measured on the units:

> Characteristics of Study Participants

> Respondents provided their age, gender, race/ethnicity, education level, income, and marital status. Age was measured as a continuous variable and reported as four categories in the demographic table (Table 1). Marital status was measured by married, living as married (i.e., cohabiting couples), divorced, widowed, separated, and single, whereas the first two responses were coded as “married” and the rest responses as “not married.”

What did the authors mean that “age was measured as a continuous variable?” Explain why if a reader only looked at Table 1 and saw the “four age categories,” the reader might not know that age was measured on a continuous scale?

The authors mean that age was initially collected with a precise number that could take any value within a range such as 25.5 years or 45 years.
If a reader only looked at Table 1 and saw the four age categories, the reader might not realize that the original data was more detailed. Instead, the reader could mistakenly believe that age was collected as a categorical variable instead of a continuous variable.

What kind of variable is marital status?
Marital status is a categorical variable, specifically a nominal variable since the categories (married, divorced, widowed, separated, and single) have no meaningful order or ranking.

Question 3.c (2 points)

To measure racism experiences by the participants, the authors used the following technique:

Racism was assessed with three measures: (1) direct experiences of racial discrimination, … (2) perceptions of cultural racism…, and (3) response of racism was measured by the four-item Anticipatory Racism-Related Stress Scale (ARRSS) that assessed participants’ psychological and behavioral reactions to an impending race-related event. The ARRSS was adapted from Utsey’s Prolonged Activation and Anticipatory Race-Related Stress Scale (PARS) [34]. Participants rated each item using the following scale: 1 = less than once a year, 2 = a few times a year, 3 = at least once a month, 4 = a few times a month, 5 = at least once a week, and 6 = almost every day.

What kind of variable is the ARRSS? Suppose you observe a person who scores 5 on the scale and another person who scores 3. Could you conclude that that the first person has 2 more units of racism experienced than the first?
ARRSS is an ordinal variable since it contains 6 categories with meaningful order but the intervals between the numbers are not interpretable in a numerical sense.
We cannot conclude that the first person has experiences exactly 2 more units of racism than the second person. This is because the scale is ordinal, so the difference between the scale points are not guaranteed to be equal. For example the difference between score 1 (“less than once a year”) and score 2 (“a few times a year”) is not guaranteed to be the same as the

difference between score 4 and score 5. So all we can say is that a score of 5 experienced more race-related stress than a score of 3.

Question 4 (5 points)

Bring up a browser with UMGPT. UMGPT provides access to a variety of AI “large language model” tools, such as ChatGPT. Ultimately, we will use a more focused Python AI assist (Copilot) to help us with Python programming, but the general purpose tools can also be quite adept at understanding and explaining Python code.

Question 4.a

Copy and paste the following Python code into a chat. Ask ChatGPT to explain the code.

```
x = 10
y = 5
sum_result = x + y
product_result = x * y
```

Was the result from ChatGPT useful? Given our discussion of Python in lecture, was the explanation accurate?
Yes it is useful because it breaks the code down into explainable portions so that it is easier to understand. The explanation it gave me was also accurate given our discussion.

Question 4.b

Ask ChatGPT how you can get access to the third course this student took:

```
student = {
    "name": "John Doe",
    "age": 21,
    "courses": ["Math", "Science", "English"]
}
```

After reading ChatGPT’s response, create some python code to create the variable `jd_third` that contains the third course taken by John Doe.

```
jd_third  = student[“courses”][2]
```

Question 4.c

Ask ChatGPT to help you understand this error message:

```
>>> student["major"]
Traceback (most recent call last):
  File "<string>", line 1, in <module>
KeyError: 'major'
```

Why is this error occurring?

This error is occurring because the student dictionary doesn't have a key named "major" so you can't access student[major]

Question 4.d

Here is the hypotenuse function we covered in class:

```
def hypotenuse(a, b):
    return (a**2 + b**2)**0.5
```

Ask ChatGPT to write documentation for this function. What kind of input does ChatGPT think a and b should be? Ask Python to tell you more about this data type. Explain in your own words what kinds of values we can put in a and b.

The params a and b should be numbers that can be squared together such as int or float. Therefore we can put whole numbers (int) and numbers with decimal points (float).