Andrew Balaschak CPTS 475 - Homework 1 8/31/2023

In [ ]:

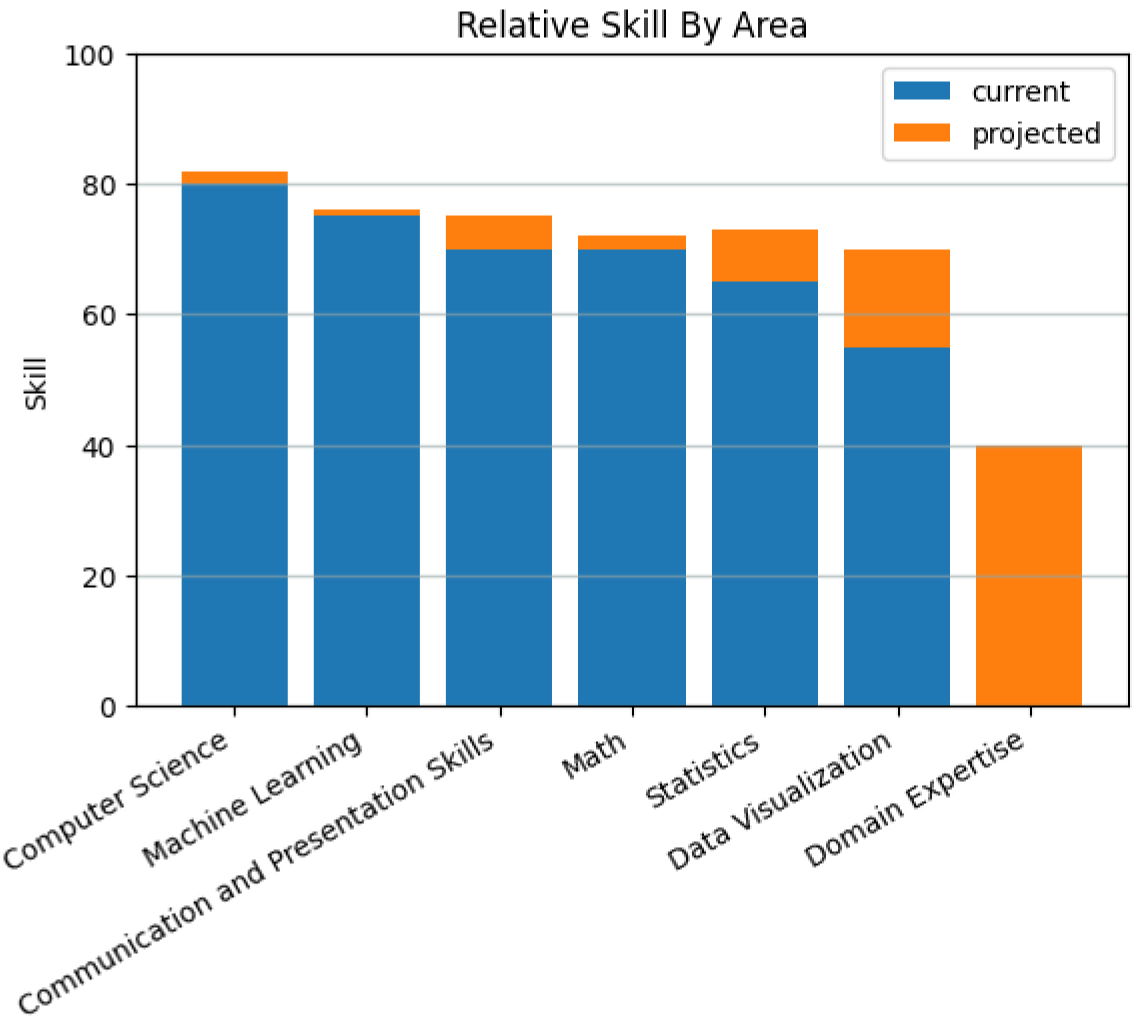
```
import matplotlib.pyplot as plt

fig, ax = plt.subplots()

# Data
data_pre = [80, 75, 70, 70, 65, 55, 0]
data_post = [2, 1, 5, 2, 8, 15, 40]
labels = ["Computer Science", "Machine Learning", "Communication and Presentation S

# Formatting
ax.set_ylabel("Skill")
ax.set_title("Relative Skill By Area")
fig.autofmt_xdate()
plt.grid(color='#95a5a6', linestyle='-', linewidth=1, axis='y', alpha=0.6)
plt.ylim([0, 100])

# Plotting
plt.bar(labels, data_pre, label = "current")
plt.bar(labels, data_post, bottom = data_pre, label = "projected")
ax.legend()
plt.show()
```



1.a. I ordered the colums in my chart by the amount of skill I estimated I currently have, with the amount of skill I am projected to gain as a tie-breaker if two colums have the same current value. I close to use a stacked bar chart because it allows easy visualization of how much skill I believe I should gain from this course, as opposed to constructing two separate bar charts that would be harder to compare visually

1.b. I think it would be interesting to have a couple specific programming languages as skills in this bar chart, specifically Python and R since that is what we will be using in this class. I don't believe any of the skills on the chart should be removed.

2.a. Dhar differentiates data science from statistics in that a data scientist must have a broad skillset involving not only statistics, but also computer science, machine learning, and problem formulation. The data used in data science is often unstructured, as opposed to structured data used by statisticians, and must be interpreted and made sense of using code.

2.b. For the "hard" sciences, causal models may be extracted from large amounts of data. For the "soft" sciences, predictive models can narrow down the set of potential causes and then prompt further research. On the whole, the amount of data we have allows machine learning algorithms to notice correlations between variables and outcomes that can be expanded upon. These correlations do not imply causation and are a way of passively looking at the data rather than conducting a scientific experiment, but they can serve as the spark for said experimental research. The large amount of data also significantly helps to resolve model misspecification and sample size issues.

2.c. Headline: "Pulling Something From Nothing" Summary: "Using machine learning, data scientists are able to recognize subtle patterns that would otherwise go unnoticed" "Humans can easily find patterns from one or two input variables, but computers can analyze hundreds of inputs"