Shark Tank (continued)

In the first part of this lab, you started to explore the Shark Tank data set (https://dlsun.github.io/pods/data/sharktank.csv).

In this second part, you will explore more complex questions by making multivariate visualizations. The questions in this part are deliberately more vague because we want you to get practice with:

- translating real world questions into data questions
- telling a story with the data.

Your answers may differ from your friend's, and that's okay!

Read in the data into a Pandas `DataFrame` and clean the data like you did in the previous part of this lab.

In [98]:
```
import pandas as pd
from plotnine import *
```

In [99]:
```
df = pd.read_csv("https://dlsun.github.io/pods/data/sharktank.csv")
df
```

Out[99]:

	Season	No. in series	Company	Deal	Industry	Entrepreneur Gender	Amount	Equity	Corco
0	1.0	1.0	Ava the Elephant	Yes	Healthcare	Female	$50,000	55%	
1	1.0	1.0	Mr. Tod's Pie Factory	Yes	Food and Beverage	Male	$460,000	50%	
2	1.0	1.0	Wispots	No	Business Services	Male	NaN	NaN	N
3	1.0	1.0	College Foxes Packing Boxes	No	Lifestyle / Home	Male	NaN	NaN	N
4	1.0	1.0	Ionic Ear	No	Uncertain / Other	Male	NaN	NaN	N
...	...	...	...	...	...	...	...	...	
490	6.0	28.0	You Kick Ass	Yes	Children / Education	Female	$100,000	10%	N
491	6.0	29.0	Shark Wheel	Yes	Fitness / Sports	Male	$225,000	8%	N
492	6.0	29.0	Gato Cafe	No	Uncertain / Other	Female	NaN	NaN	N
493	6.0	29.0	Sway Motorsports	Yes	Green/CleanTech	Male	$300,000	20%	N
494	6.0	29.0	Spikeball	Yes	Fitness / Sports	Male	$500,000	20%	N

495 rows × 17 columns

In [100…

```
df[["Corcoran", "Cuban", "Greiner", "Herjavec", "John", "O'Leary", "Harrington"
df
```

Out[100]:

	Season	No. in series	Company	Deal	Industry	Entrepreneur Gender	Amount	Equity	Corc
0	1.0	1.0	Ava the Elephant	Yes	Healthcare	Female	$50,000	55%	
1	1.0	1.0	Mr. Tod's Pie Factory	Yes	Food and Beverage	Male	$460,000	50%	
2	1.0	1.0	Wispots	No	Business Services	Male	NaN	NaN	
3	1.0	1.0	College Foxes Packing Boxes	No	Lifestyle / Home	Male	NaN	NaN	
4	1.0	1.0	Ionic Ear	No	Uncertain / Other	Male	NaN	NaN	
...	...	...	...	...	...	...	...	...	
490	6.0	28.0	You Kick Ass	Yes	Children / Education	Female	$100,000	10%	
491	6.0	29.0	Shark Wheel	Yes	Fitness / Sports	Male	$225,000	8%	
492	6.0	29.0	Gato Cafe	No	Uncertain / Other	Female	NaN	NaN	
493	6.0	29.0	Sway Motorsports	Yes	Green/CleanTech	Male	$300,000	20%	
494	6.0	29.0	Spikeball	Yes	Fitness / Sports	Male	$500,000	20%	

495 rows × 17 columns

In [101…

```
#Cleaning the data in our Amount and Equity Columns
df["Amount"]= df["Amount"].astype(str).str.replace("$", "")
df["Amount"] = df["Amount"].astype(str).str.replace(",", "")
df["Amount"] = df["Amount"].astype(float).fillna(0)

df["Equity"] = df["Equity"].astype(str).str.replace("%", "").fillna(0)
df["Equity"] = df["Equity"].astype(float) / 100
```

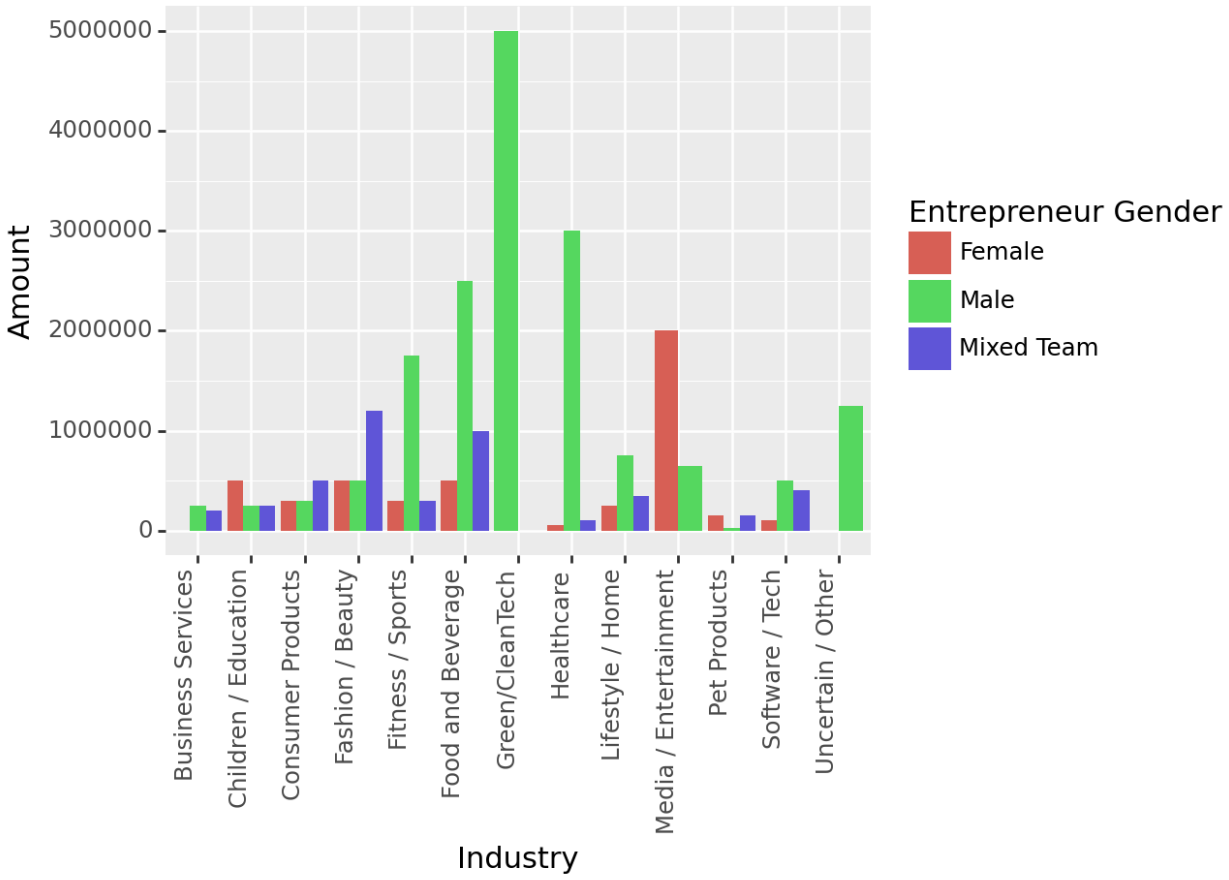
Question 1. How Does Funding Rate Depend on Industry and Gender?

Use the grammar of graphics to make a visualization that explores how funding rate depends on the industry of the startup and the gender of the founders.

In [102…

```
#Relationship between amount, gender, and industry
(ggplot(df, aes(x = "Industry", y = "Amount", fill = "Entrepreneur Gender"))
```

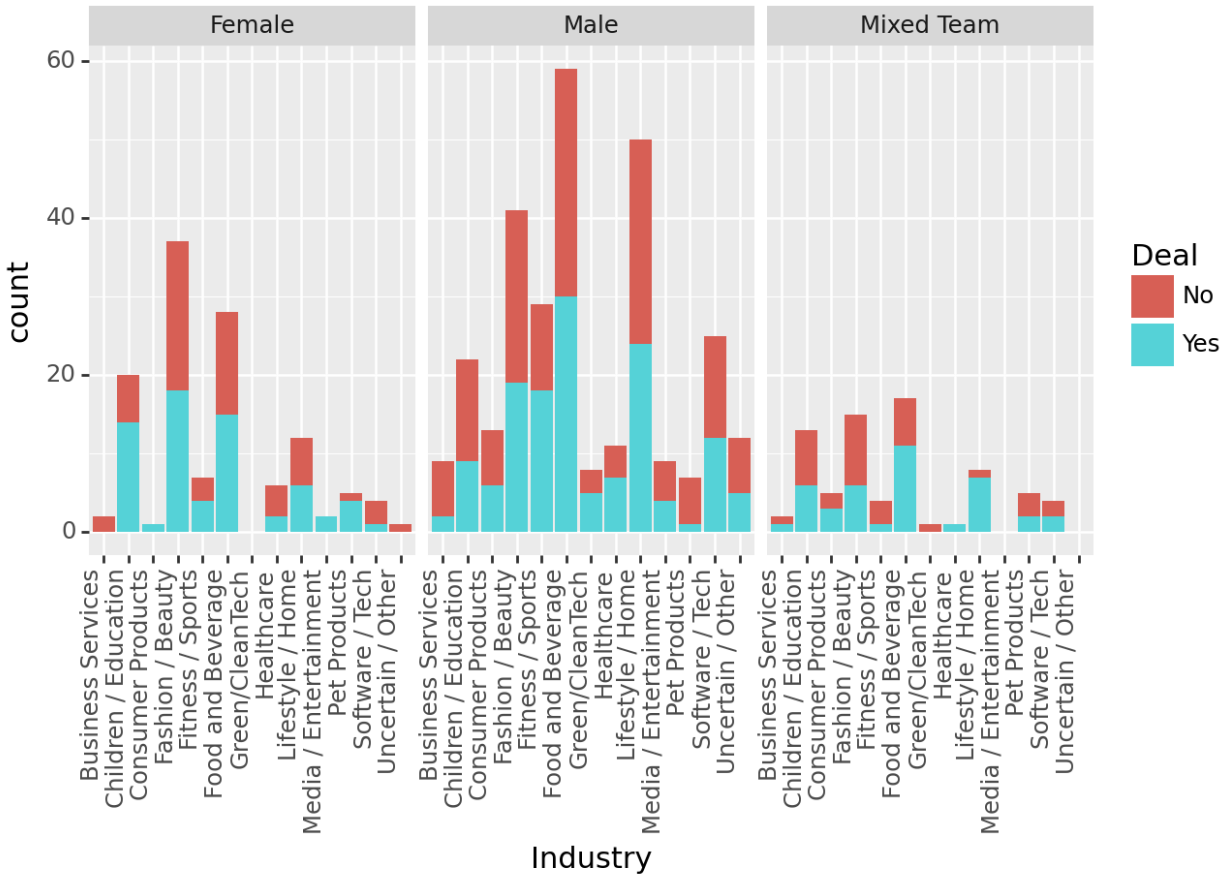
```
+ geom_col(position = "dodge")
+ theme(axis_text_x=element_text(rotation=90, hjust=1))
)
```



Out[102]: <Figure Size: (640 x 480)>

In [103…

```
(ggplot(df, aes(x = "Industry", fill = "Deal"))
+ geom_bar()
+ facet_wrap("Entrepreneur Gender") #Facetting on Entrepreneur Gender
+ theme(axis_text_x=element_text(rotation=90, hjust=1))
)
```



Out[103]: <Figure Size: (640 x 480)>

Write one paragraph summarizing your conclusions from your visualization.

The first visualization above displays how funding rate (or amount) can depend on the industry of the product and the entrepreneur's gender. Based on this data, it appears that the most profitable industry to pitch depends on the genders of the entrepreneurs on the team. To illustrate this, the industry with the highest amount of investment for all-female teams is the media and entertainment industry while that for all-male teams is the Green/CleanTech industry. For mixed-gender teams, the industry with the most potential payoff and investment was the Fashion/Beauty industry. These differences based on the entrepreneur's gender are most likely due to a combination of gender-based stereotypes and norms that have been perpetuated over extended periods of time and efforts to reverse those stereptypes. To elaborate on this, mixed-gender teams most likely saw the most investment in the fasion and beauty industry due to the promising idea of men and women working together in the fashion world and making it more equitable. Conversely, all-male teams are likely the ones who receive the most investment for the Green/CleanTech industry because the world of tech has historically been largely male-dominated. At the same time, all-female teams were the groups that saw the most investment in the children/education industry most likley because the idea of women being motherly and becoming teachers has been perpetuated for centuries.

The second visualization above is a display faceted on the deal (whether the shark agreed to invest or not) and showing the frequency of how many times the company either recived funding or not based on the industry and entrepreneur gender of that company. Based on the data shown above, it appears that the industry with the highest number of successful deals for all-male teams was the Green/CleanTech industry, and all-male companies in this industry were funded more often than not. Additionally, the all-female teams tended to see the most deals with funding decisions in the fasion and beauty industry. This is most likely due to the historical social stereotype that fashion and beauty are more feminine fields and the tech industry has been dominated by men since its creation. All in all, these generalizations reveal that the industry with the most successful funding rate and amount of investment depends on the gender distribution of the company.

Write one paragraph explaining your design choices in making your visualization. (For example, why did you map a variable to that aesthetic?)

I decided to create the visualizations I did and label the axes in this way due to the dependence of each variable on the other and maximizing the clarity of the data. For example, I used the problem statement to see that our task was to see which how the funding rate was dependent on the industry and entrepreneur (this could also be thought of as examining how these two factors influence the funding rate of each product). This signified that the proportions of the funding decisions (the funding rates) should be on the y-axis because this is the variable that we are aiming to see the result of. This therefore implied that we would be needing to exmaine the funding rates for each industry and separate this evaluation based on gender.

For the first visualization, I decided to plot the amount on the y-axis to see if there were differences in the correlations between the funding rate and the amount funded. This first visualization grants a side-by-side bar plot of the entrepreneur genders and the industries, to allow for an examination of which industries had the highest payoffs for each gender distribution.

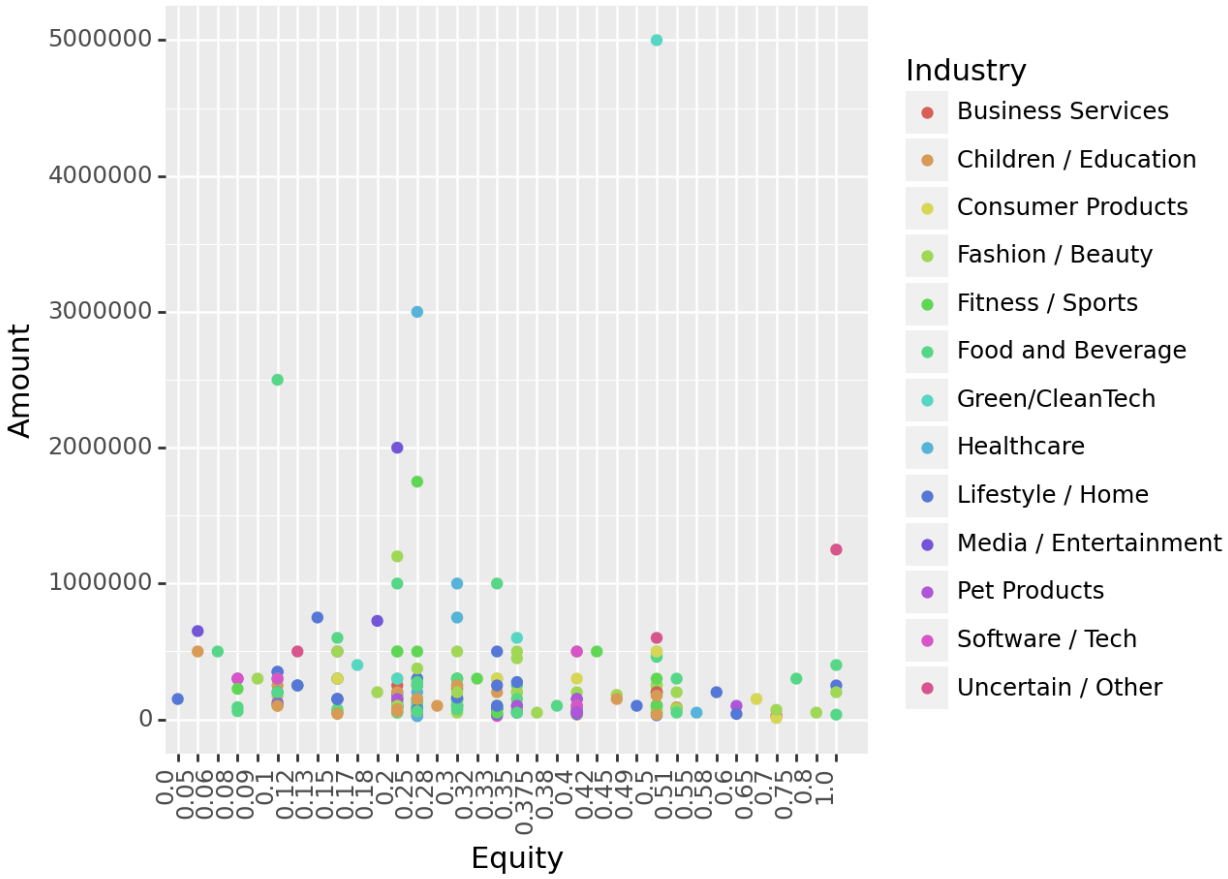
Question 2. What Determines Amount Funded?

What factors are associated with the amount funded? Make a multivariate visualization that addresses this question.

In [104…
```
df["Equity"] = df["Equity"].astype("category")
```

In [105…
```
(ggplot(df, aes(x = "Equity", y = "Amount", color = "Industry"))
+ geom_point()
+ theme(axis_text_x=element_text(rotation=90, hjust=1))
)
```

```
/usr/local/lib/python3.10/dist-packages/plotnine/layer.py:364: PlotnineWarnin
g: geom_point : Removed 250 rows containing missing values.
```



Out[105]: <Figure Size: (640 x 480)>

Write one paragraph summarizing your conclusions from your visualization.

This multivariate scatterplot shown above demonstrates the relationship between the Equity of a product and the amount of investment it receives. I have also included coloring for the specific industries on this visualization. The data displayed here reveals that the amount invested in a product tends to go down as its equity increases-there is a weak positive correlation between the two. The healthcare and Green/CleanTech industries are the ones that sees the strongest corrletion in this, as there are notable increases in the amount invested as the equity increases. Conversely, the software/tech industry and any products whose industries are unknown see the opposite pattern, with the amount invested increasing as the equity increases (directly proportional relationship). These findings indicate that the equity of a company's product has a weak correlation to the amount of funding it receives, and this correlation varies depending on the industry the product falls under. This is likely because the most profitable indsutries are ever-changing based on the time of investment, and the sharks will take into account the share of the company they receive when deciding on an amount to invest.

Write one paragraph explaining your design choices in making your visualization. (For example, why did you map a variable to that aesthetic?)

I decided to use a multivariate scatter plot to demonstrate how the amount invested in a product relates to its equity level because this most clearly demonstrates correlation and proportionality in relationships between two variables. I additionally decided to color on each individual industry to more clearly see if there were any distinctions in the patterns demonstrated for each individual industry. The two variables I chose were the amount invested and the equity level because I recalled how in Lab 2A, these two variables determined the valuation of a particular company and therefore had an impact on how much money was being invested in each product.

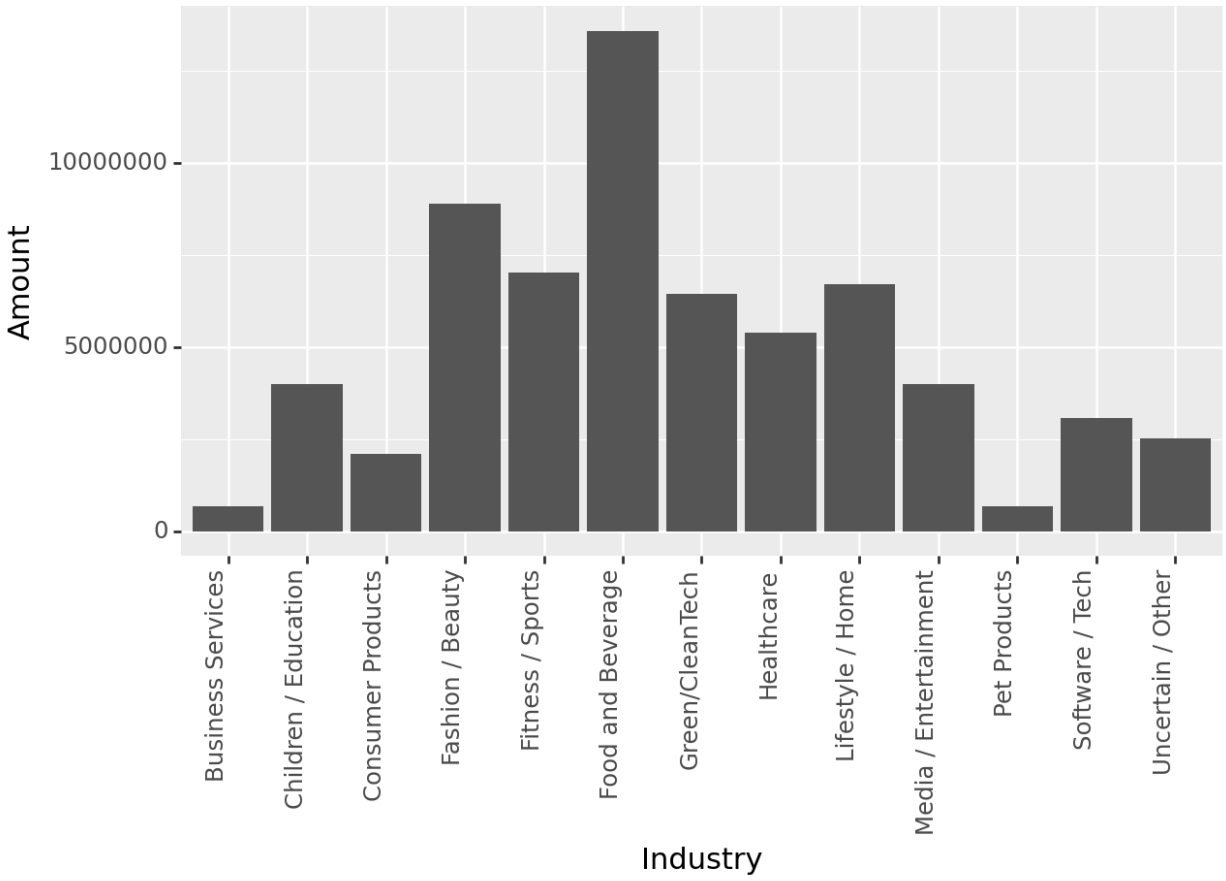
Question 3. Help Your Friends!

Your friends are going on the next season of *Shark Tank*. They have reached out to you for advice about what kind of company to pitch and which sharks to target.

Use the data to give them some concrete advice about how to maximize their chances of being funded.

In [106…

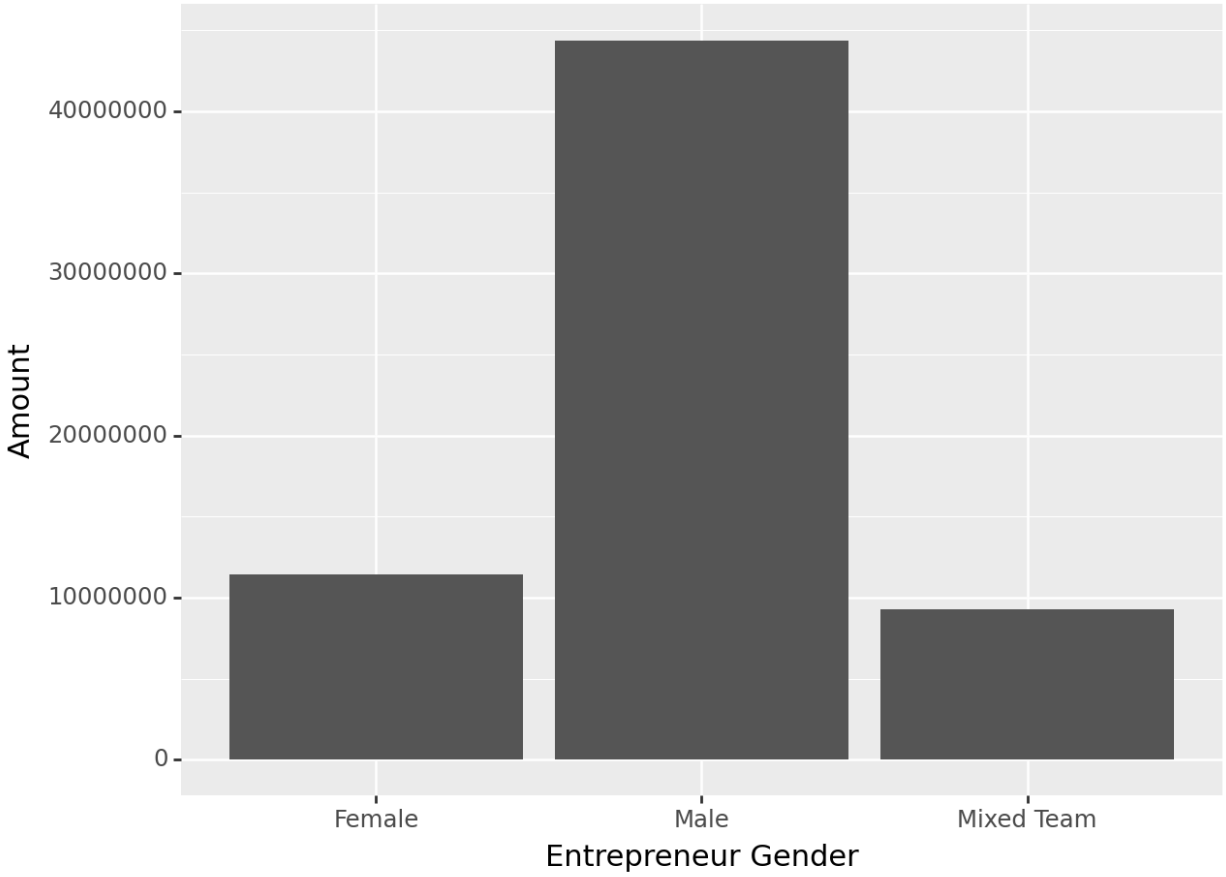
```
# Relationship between Industry and Amount
(ggplot(df, aes(x = "Industry", y = "Amount"))
+ geom_col()
+ theme(axis_text_x=element_text(rotation=90, hjust=1))
)
```



Out[106]: <Figure Size: (640 x 480)>

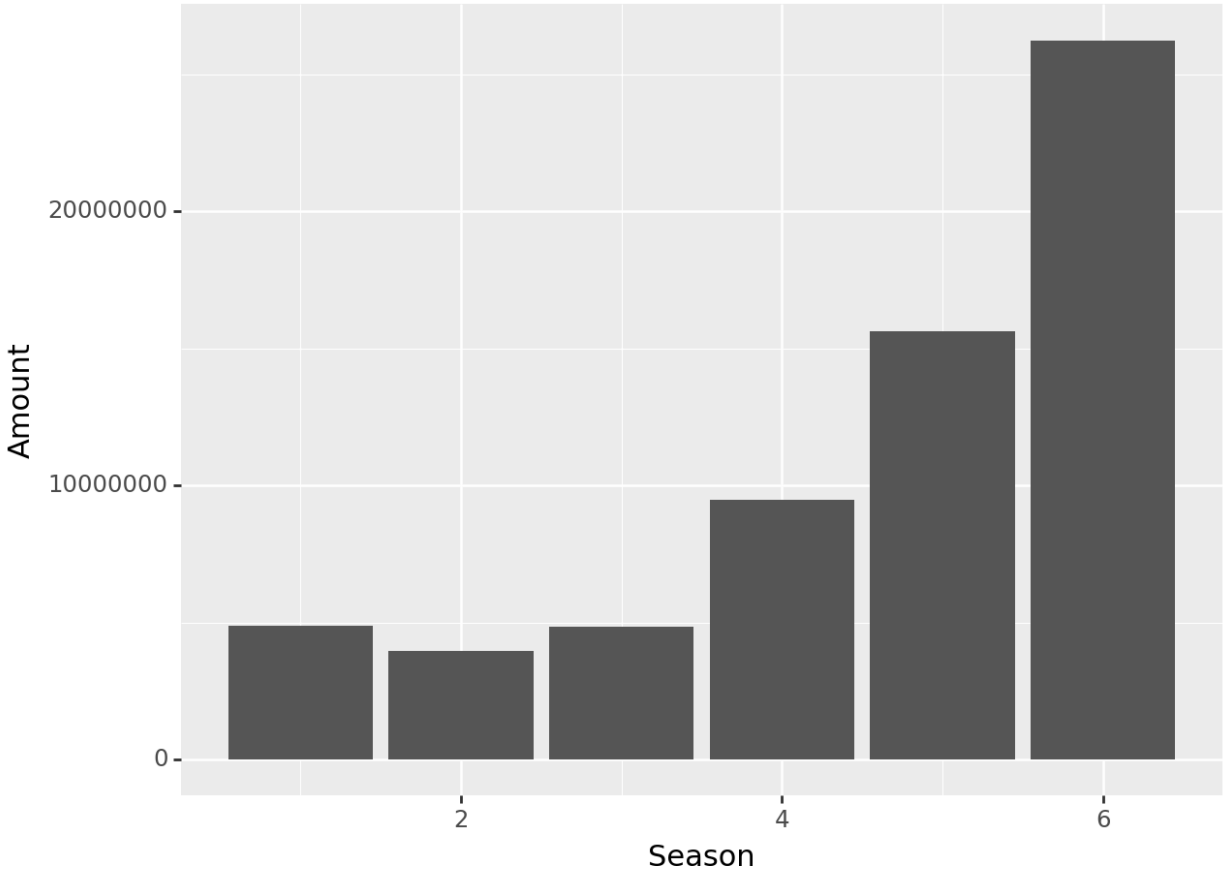
In [107…
```
df["Equity"] = df["Equity"].astype('category')
```

In [108…
```
# Relationship between Entrepreneur Gender and Amount
(ggplot(df, aes(x = "Entrepreneur Gender", y = "Amount"))
+ geom_col()
)
```



Out[108]: <Figure Size: (640 x 480)>

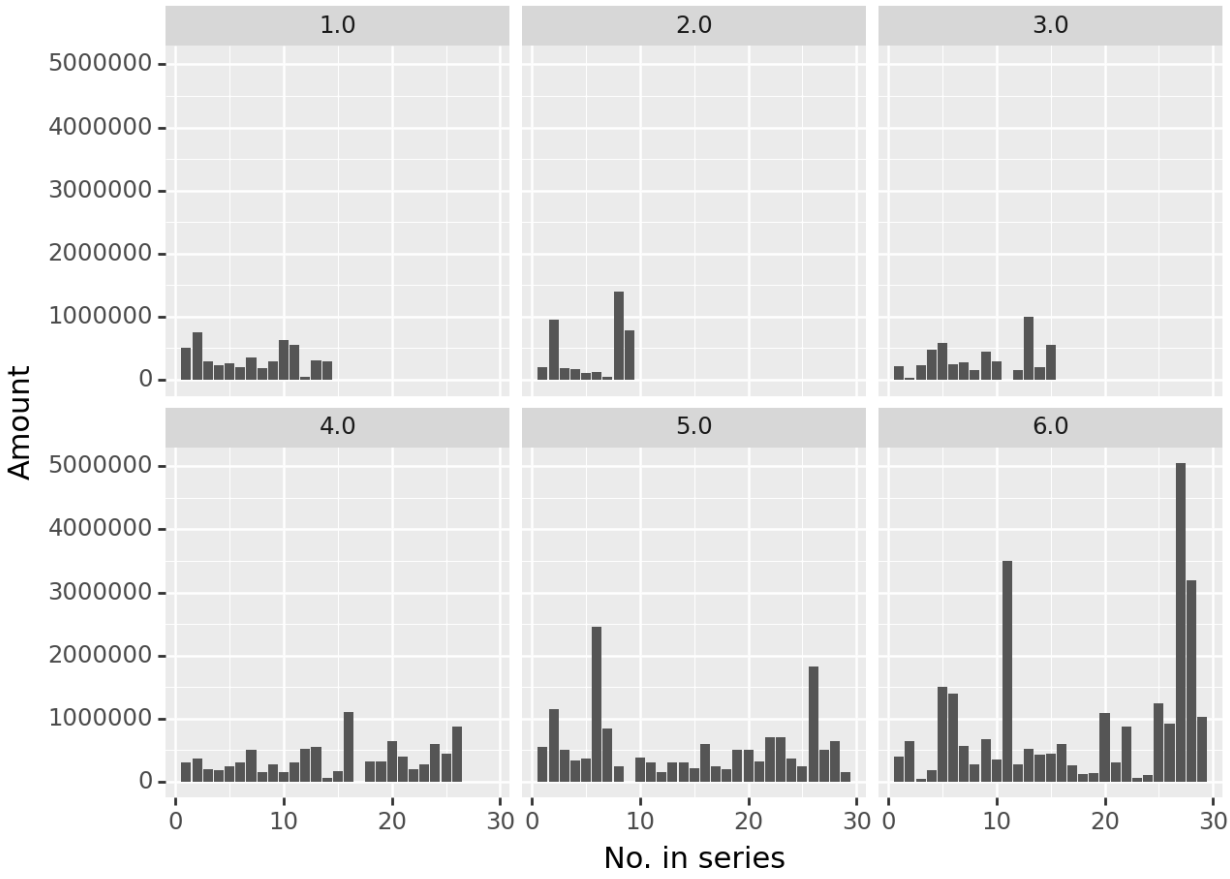
In [109…
```
#Amounts Invested Across the Seasons
(ggplot(df, aes(x = "Season", y = "Amount"))
+ geom_col()
)
```



Out[109]: <Figure Size: (640 x 480)>

In [110…

```
#Number in Series versus Amount
(ggplot(df, aes(x = "No. in series", y = "Amount"))
+ geom_col()
+ facet_wrap("Season")
)
```



Out[110]: <Figure Size: (640 x 480)>

In [111…
```
marginal_gender = df['Entrepreneur Gender'].value_counts(normalize = True) #Co
marginal_gender
```

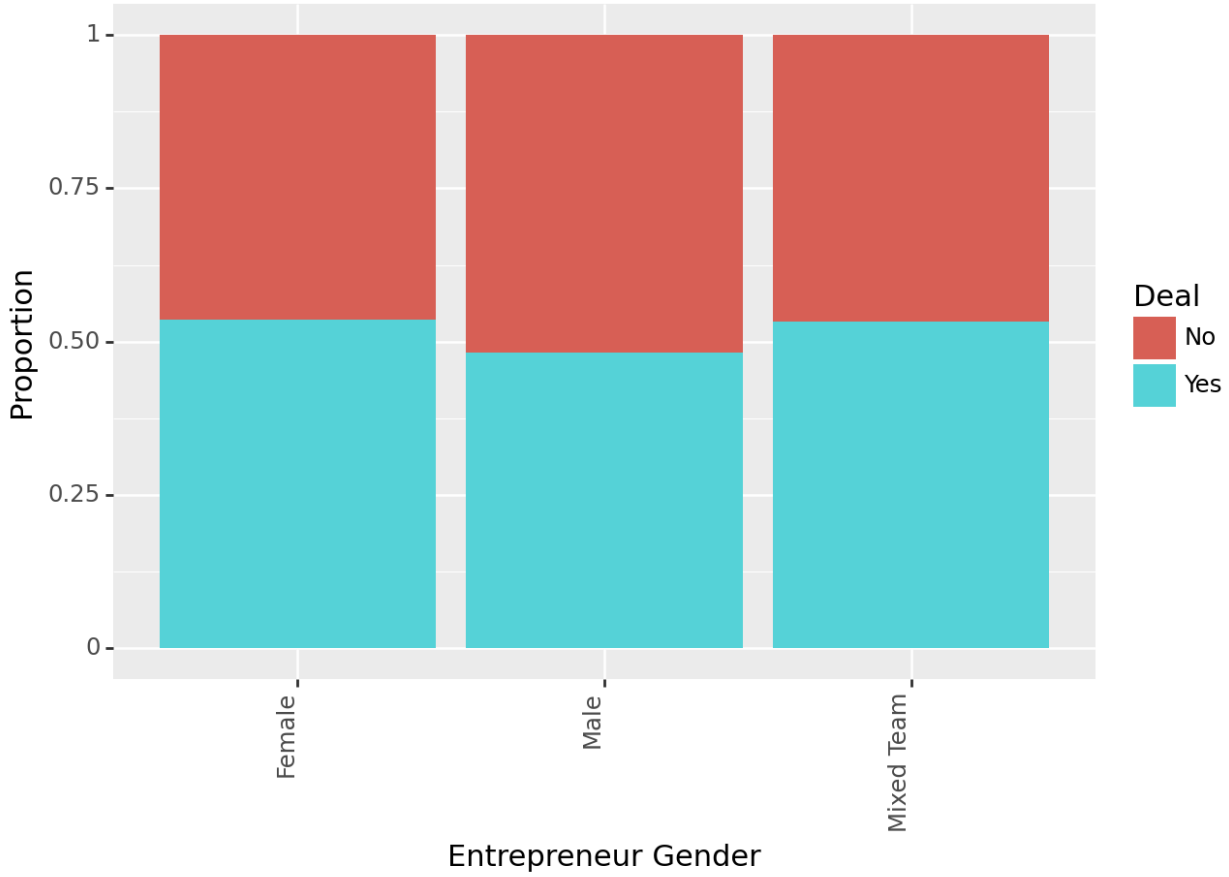
Out[111]:
```
Entrepreneur Gender
Male          0.595960
Female        0.252525
Mixed Team    0.151515
Name: proportion, dtype: float64
```

In [112…
```
#Obtaining a table of the conditional probabilities of funding for each gender
joint_funding_gender = df[["Deal", "Entrepreneur Gender"]].value_counts(normal
joint_funding_gender.divide(marginal_gender)
```

Out[112]:

Entrepreneur Gender	Female	Male	Mixed Team
Deal			
No	0.464	0.518644	0.466667
Yes	0.536	0.481356	0.533333

In [ ]:
```
from plotnine import *
#Making a visualization of the funding rates conditioned on entrepreneur gende
(ggplot(df, aes(x = "Entrepreneur Gender", fill = "Deal"))
+ geom_bar(position = "fill")
+ ylab("Proportion")
+ theme(axis_text_x=element_text(rotation=90, hjust=1)) #code obtained from Sta
)
```



Out[ ]: <Figure Size: (640 x 480)>

Write an e-mail to your friends summarizing your advice.

Dear friends,

I hear you are interested in making a company pitch on the next season of Shark Tank! How exciting! I wanted to let you know I've run some data on the past 6 seasons of Shark Tank and have found some patterns in which companies are recieving the highest amounts of investmnents. Based on the data visualizations I have generated, it appears that the sharks have gotten more generous over the seasons, as there has been a steady increase in the amount of money invested since Season 3 of the show. It also seems that the show has gained more episodes per season as time has gone on, and the most amount of money is invested in the last few episodes in the season or just after the first 30% of the season has passed.

Additionally, the amount of money received and the funding rate (whether sharks agree to fund or not) is dependent on the industry and gender distribution of each company. All-male teams have received the most amount of investment in dollars, though conditioning the data on the Entrepreneur Gender showed me that sharks are more likley to fund an all-female or mixed-gender team. Additionally, the most profitable industries for mixed-gender teams were lifestyle/home and food/beverage while fashion/beauty was the industry most likely to receive funding for all-female teams. For all-male teams, the food/beverage industry saw the greatest number of successful funds.

Best of luck, and hope that helps!

Submission Instructions

- Restart this notebook and run the cells from beginning to end.
  - Go to Runtime > Restart and Run All.

In [ ]:
```
# @markdown Run this cell to download this notebook as a webpage, `_NOTEBOOK.h

import google, json, nbformat

# Get the current notebook and write it to _NOTEBOOK.ipynb
raw_notebook = google.colab._message.blocking_request("get_ipynb",
                                                      timeout_sec=30)["ipynb"]
with open("_NOTEBOOK.ipynb", "w", encoding="utf-8") as ipynb_file:
  ipynb_file.write(json.dumps(raw_notebook))

# Use nbconvert to convert .ipynb to .html.
!jupyter nbconvert --to html --log-level WARN _NOTEBOOK.ipynb

# Download the .html file.
google.colab.files.download("_NOTEBOOK.html")
```

- Open `_NOTEBOOK.html` in your browser, and save it as a PDF.
  - Go to File > Print > Save as PDF.
- Double check that all of your code and output is visible in the saved PDF.
- Upload the PDF to Gradescope.
  - Please be sure to select the correct pages corresponding to each question.