

For each question, describe how you would make it testable and translate it from a general question into something statistically rigorous.

1. You work at an e-commerce company that sells three goods: widgets, doodads, and fizzbangs. The head of advertising asks you which they should feature in their new advertising campaign. You have data on individual visitors' sessions (activity on a website, pageviews, and purchases), as well as whether or not those users converted from an advertisement for that session. You also have the cost and price information for the goods.

Question: Which one to feature in the new advertising campaign; widgets, doodads, or fizzbangs?

Data: individual visitors' sessions

- activity on the website
- pageviews
- purchases

whether the individuals converted from an advertisement
cost
price

Actions	Widgets	Doodads	Fizzbangs
Activity on the website			
Pageviews			
Purchases			
Conversion from advertisement?			
Cost			
Price			
Cost/Price Ratio			

I would arrange a dataframe in a fashion as illustrated above. Then, I would assign the appropriate value in the various empty spaces. I

would then tell the head of advertising that the answer depended on what the company wanted to advertise. Does the company want to advertise the product that is the most popular or does the company want to advertise the product that is the least popular?

2. You work at a web design company that offers to build websites for clients. Signups have slowed, and you are tasked with finding out why. The onboarding funnel has three steps: email and password signup, plan choice, and payment. On a user level you have information on what steps they have completed as well as timestamps for all of those events for the past 3 years. You also have information on marketing spend on a weekly level.

Question: Why have signups slowed?

Data: email and password signup, plan choice, payment, timestamps, marketing spend on a weekly level (52 weeks x 3 years)

	Email Signup	Password Signup	Plan Choice	Payment	Complete
Timestamp					
Marketing Spend					
Season of Year					

I would arrange a dataframe as illustrated above. I would then analyze the data to see if there is a correlation to marketing spend and complete signups. I would also see where people are taking the most time. I would then ask the employer to see if there are any issues with the GUI where the people are taking the most time. I would also try and see if there was a correlation to time of the year. This would be my initial analysis.

3. You work at a hotel website and currently the website ranks search results by price. For simplicity's sake, let's say it's a website for one city with 100 hotels. You are tasked with proposing a better ranking system. You have session information, price information for the hotels, and whether each hotel is currently available.

Question/Task: Propose a better ranking system.

Data: session information, price information, availability

Hotel	Activity on the website	Pageviews	Bookings	Price	Availability

I would arrange a dataframe in the above framed fashion. I would then calculate the ratios of a column value in the "Activity", "Page views" and "Bookings" for each hotel (i.e. if hotel "A" had 1000 pageviews, then that would be divided by the total number of pageviews of hotels for the site, and a value would be obtained; that value would then be stored in an assigned order and then that value would be ranked from 1 to 100) and the same thing would be done for each value in every column aforementioned. Then, I would get the total of the 3 columns as a value and use that value for rankings.

4. You work at a social network, and the management is worried about churn (users stopping using the product). You are tasked with finding out if their churn is atypical. You have three years of data for users with an entry for every time they've logged in, including the timestamp and length of session.

Question: What is going on with churn?

Data: 3 years worth of data - user, entry, timestamp, length of session.

First I would define a churn episode by setting a length of period from the last use of the product by the user outward to about 6 weeks. Then I would calculate the number of churns per day achieved, per

week achieved, per month achieved. I would then graphically see if there are distinct periods of maximum churns compared with other time periods.