MSIN0093: Business Strategy and Analytics
Individual Assignment
Due: 29 November 2019 10:00

Instructions

Please complete each of the questions in the following assignment. If you refer to external work, cite appropriately. Indicate the word count at the end of each answer. **All work is to be completed individually.** PDF submissions are strongly preferred.

1 HHI Index (max 100 words)

All else equal, would you rather compete in an industry where the Herfindahl-Hirschman Index, HHI, is 0.25 or 0.75? Briefly explain your answer.

2 Descriptive Analytics (max 200 words)

You have a dataset on select movies (moviedata.csv) that were released in the United States from 2001 to 2006. The dataset contains the following variables:

Variable	Description
movie	Name of movie
metascore	Aggregate Metascore
rating	Movie rating
runtime	Length of movie (in minutes)
genre1	Primary genre listed
release_date	Movie release date
release_year	Year of movie release
rat_avg	Average user rating on Metacritic
rat_count	Total user ratings on Metacritic
rat_poscount	Count of positive ratings from users on Metacritic
rat_mixcount	Count of mixed ratings from users on Metacritic
rat_negcount	Count of negative ratings from users on Metacritic
prodbudget	Production budget
adbudget	Marketing and advertising budget
franchise	Franchise detail
source	Source material
domgross	Domestic gross (in USD)
intlgross	International gross (in USD)
distributor	Name of Distributor

Produce one well designed data visualization using the data that revealed something interesting to you. In a few sentences explain the key takeaway from the visualization.

3 Setting up a Survey (max 200 words)

As head of HR, you ask one of your data analysts to run a survey to guage employee satisfaction of a planned change to the paid time off policies. Your analyst makes the following suggestion: "I think we should run a small pilot before distributing our survey to the full sample. That way, we can (a) calculate the mean satisfaction among the pilot respondants. We can also (b) check whether the respondants are interpreting the questions the way we intended." Do you think each of these justifications for running the pilot is valid? Briefly explain your answer.

4 Prediction Model Takeaways (max 250 words)

You work at an insurance company. The data scientist at the company has built a prediction model using the characteristics of drivers and their cars to predict the likelihood of getting into an accident. The most pronounced result in the model is that yellow cars are highly predictive of being in an accident. Your colleague looks at the results and says the following: "I can imagine us acting on this in two ways. First, we can send out notices to all our customers suggesting they avoid yellow cars when making future automotive purchases. Second, we can raise insurance prices on customers that have yellow cars." How would you respond to each of the two suggestions? In other words, what is your view on each of the two applications of the model?

5 Causal Interpretations (max 250 words)

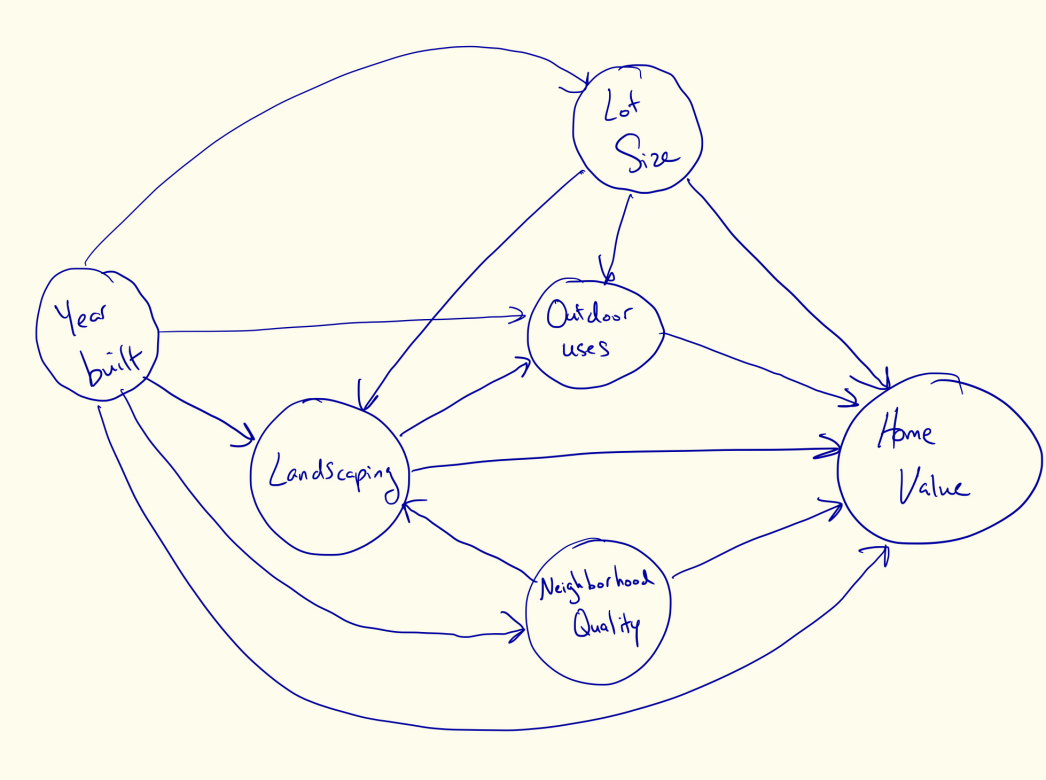
You work as a data analyst at a real estate company. Management has asked the analytics team to come up with a list of actions that homeowners can take to increase the value of their homes. Using a dataset of a random sample of homes in the region the real estate company operates in, your colleague runs the following regression:

$$\textbf{value} = \beta_0 + \beta_1(\textbf{landscaping}) + \beta_2(\textbf{year\_built}) + \beta_3(\textbf{outdoor\_uses}) + \beta_4(\textbf{lot\_size}),$$

where *landscaping* is a score on the quality of the landscaping work done in the outdoor spaces (on a scale of 1-100), *value* is the value of the house (in log pound sterling), *year_built* is the year the home was originally built, *outdoor_uses* is a dummy variable that equals one if the outdoor space can accommodate activities, such as sports or barbecues, and *lot_size* is the size of the property (in square feet). She claims that the relationship between *landscaping* and *value* is causal, and thus argues that making investments in landscaping cause home values to increase.

You sketch out your model of the world using a DAG that is described in Figure 1, which includes a variable, *neighborhood_quality*, that measures the value of nearby homes.

Figure 1: Your Causal Model of Landscaping Quality on Home Value



Based on your DAG what is your assessment of your colleague's claim regarding the relationship between *landscaping* and *value*?

6 Impact of IT on Employee Productivity (max 1,000 words)

Your company, Beto Enterprises, operates call centers. Among Beto's clients are several companies in the Fortune 500. Call agents at Beto take customer service calls on behalf of client firms and attempt to troubleshoot the problems that customers are facing. The company has set up call centers around the world, with locations in Austin (USA), Bangalore (India), Bucharest (Romania), and Buenos Aires (Argentina). Each call center employs approximately 1,000 agents.

Recently, the company purchased and deployed new software for agents to use during calls in the Austin location. The intent of the software was to speed up the average time an

agent needed to spend on a call, one of the key performance metrics the company uses to measure agent productivity.

You are the chief data analyst at Beto. Management has asked you to measure the impact of the software on time spent on calls. After privately grumbling that management did not let you run an experiment, you look at how the software was deployed to all the agents. You discover that due to a technical error in the purchase of the software licenses, just over half the agents did not have access to the software for the first four weeks of the deployment period.

You collect a weekly dataset of the average time that agents spent on calls for four weeks prior to and after the deployment (agent_software.csv). The dataset contains the following variables:

Variable	Description
agent_id	an anonymized agent identifier
week_num	the week of the data period (ranges from 1 to 8, and the software was rolled out on a Sunday night between weeks 4 and 5)
got_software	=1 if the agent received one of the licenses and was using the new software during that week, =0 otherwise
years_experience	the number of years of experience of the agent
female	=1 if agent identifies as female, =0 otherwise
degree	=1 if the agent has a university degree (or higher), =0 otherwise
avg_timeoncall	the average number of minutes spent on all calls during the week

1. Describe the design you would employ in order to assess the causal impact of the software on average time spent on calls by agents. Explain why this approach is appropriate, given the context and data.

2. Use the dataset (agent_software.csv) to provide an estimate of the causal impact of software adoption on call times. Show the results from all your analysis. Make sure your presentation of results is done in a manner that is suitable for a report to management.

3. How would you report these results to management? How would you contextualize the size of the effect? Are there any limitations worth noting in the report?

4. How did you incorporate the data on gender, education, and work experience?

5. If you had just looked at the difference before and after getting the software (for the agents who eventually received the software), by how much would you have over- or under-estimated the impact of the software?

7 Experiment Proposal (max 1,000 words)

You have joined the data science team of a large online retailer. (Choose either Amazon or Alibaba. If you choose a different company, indicate the company at the start of your answer.)

The head of the data science team gives you your first assignment: "At our company, we depend on experimentation to find ways to improve performance. I would like you to propose and design an experiment to test one particular aspect of our business. Your proposal can be related to the company's user interface, products, customer experience, or another area of the company. I would like to know why you are interested in the aspect you choose, your proposed alternative (i.e. the experimental treatment), and how you would go about conducting the test."

Produce a writeup outlining your proposed experiment.[1] Your response should include the following elements:

1. Identification of the aspect of the business or operations that you would like to test and an explanation for your interest.

2. A definition of the overall evaluation criterion (OEC) and justification for your choice.

3. A description of the current implementation (i.e. the control) and your proposed alternative (i.e. the treatment). Justify your proposal. Make sure your proposed intervention is clear and be as creative as you would like in describing it (verbal descriptions, sketches, screenshots, etc. are all acceptable).

4. What is the population of interest and describe how you would randomly assign subjects of that population into the treatment and control.

5. The null and alternative hypotheses.

6. State the level of significance and the desired power of your test. Briefly explain what level of significance and power are.

7. Any problems you anticipate might arise during the running of your experiment and initial ideas on how to mitigate those potential problems.

8. What are the policy implications for your company if the null hypothesis is rejected? What are the implications if the null hypothesis is not rejected?

[1] You do not need any data nor do you need to run the experiment.