Adaptive Biotechnologies Data Case Study

Background:

You are on the Data Analytics team of a life-sciences company. This week, you have received several requests from various internal stakeholders. We ask that you complete these requests yourself, but you may use the internet or other resources for help – provided you list your sources you used when turning in your work.

Requests

1. A common stereotype of Thanksgiving is the inevitable post-meal nap. There are a few hypotheses why one might be tired after a Thanksgiving meal: https://www.scientificamerican.com/article/fact-or-fiction-does-turkey-make-you-sleepy/

   Your company is hoping to create a product that helps adults stay awake after their Thanksgiving meal. Using the article and the attached dataset (https://github.com/fivethirtyeight/data/tree/master/thanksgiving-2015), pick one hypothesis for what causes post-Thanksgiving fatigue (and therefore the target of your product). Pose 5-10 questions that might assess the viability and scope of this project. Please answer these questions in a memorandum for the executive team in charge of this initiative.

   Example questions (you may use these or create your own):
   - What fields in the dataset allow us to examine the validity of our hypothesis?
   - Which geographies contain the highest frequency of people susceptible to this hypothesis?
   - What are the common characteristics of our average target consumer?
   - What are the limitations of the data provided?
   - What is your conclusion on the viability of this product?

   Usage of both visuals and text commentary is encouraged (any visualization tool of your choice). Management is busy and is looking for a clear, concise summary, not an overcomplicated analysis.

2.

a. One of your colleagues has left on a vacation. While they're gone, a stakeholder approaches you about a project your colleague was helping them with. Unfortunately, all they have is a SQL query that they cryptically sent to the stakeholder before their flight boarded (bad practice!!). As you figure out what's going on, comment on the function of each line of the following query and provide a hypothesis of what the goal of the query might be.

```
select
prd.provider_fullname
,count(ord.order_name)
,count(DISTINCT [Patient Unique Patient ID])
,count(osf.SpecimenDimKey)
,SUM(CASE WHEN ord.order_status_type IN ('Completed','Active') then 1 else 0 END)
,DATEDIFF(DAY,prd.[Provider Implied or Actual Create Date PT],GETDATE())
from OrderDim ord
LEFT JOIN OrderFact orf
on ord.OrderDimKey=orf.OrderDimKey
LEFT JOIN AccountDim ad
on orf.AccountDimKey=ad.AccountDimKey
LEFT JOIN ProviderDim prd
on orf.ProviderDimKey=prd.ProviderDimKey
LEFT JOIN PatientDim pad
on orf.PatientDimKey=pad.PatientDimKey
INNER JOIN OrderSpecimenFact osf
on orf.OrderDimKey=osf.OrderDimKey

WHERE (ord.order_icd_group='ALL' or
order_icd10_codes LIKE '%C90.01%')
and ord.order_category_type='Diagnostic'
GROUP BY prd.provider_fullname, DATEDIFF(DAY,prd.[Provider Implied or Actual Create
Date PT],GETDATE())
ORDER BY DATEDIFF(DAY,prd.[Provider Implied or Actual Create Date PT],GETDATE())
ASC
```

b. If the last join was changed to a left join, what might be a possible change in the output? If I wanted to use a left join but replicate the results of the inner join, what statement could I add to the WHERE statement?

3. The Lab Operations group is reviewing last month's performance of the lab and wants to analyze the efficiency of the Biological Specimen Management (BSM) team. BSM oversees the sample accessioning process - the arrival/unpacking of samples and the preparation of samples for storage or processing (https://www.genologics.com/blog/sample-accessioning/).

   Sample accessioning is a multi-stage process and a sample must travel through various stages in order to be accessioned. Each stage may have a unique machine, operator, time of day, location in the lab or other characteristics. The sequence of steps a sample may travel through depends on the sample type (e.g. FFPE slides, Plasma, etc.), temperature, genomic material concentration, specimen source (e.g. tissue, peripheral blood, etc.), retrieval date and other sample properties. Accessioned samples are either immediately processed or stored for future use.

   Questions:
   - What KPIs might describe the performance of the lab? Why?
   - What type of data do you think you would need to acquire in order to complete this task?
   - What would be your general course of action from start to finish to complete this project?
   - How do you think your performance completing this project may be measured?