GSK NLP Analyst Technical Case Study

A guide to the take home assignment

Introduction

For this assignment, you will perform an analysis on a sample data set. You will return a short report of your analysis, as well as a proposal for how you would scale this project. Read the document below for more information. Send an email to Kate Farmer (Kathryn.b.farmer@gsk.com) if you have any questions

Data

You have a file called `case_study_data.tsv`. This is a tab delineated file containing 2880 rows with the following structure

Field	Type	Description
`uniqueID`	Int	A unique ID number for each review. IDs range from 999 to 232190. Ids are not sequential
`drugName`	String	The drug that the review is referencing
`review`	String	Reviews provided by patients regarding a medication. Reviews range in length from 243 to 732 characters
`date`	Date string	Date the review was collected. Format is dd-Mon-yy

Prompt

Part 1 – Analysis

Select a drug from the dataset. Imagine that a leader for that product has asked you the following question:
"What aspects do patients like about using our product, and what do people dislike?"

Use the data to write a report for this leader that will address their question.

Part 2 – Proposal

Imagine that product leaders are often asking these kinds of questions. How would you uncover insights for them in the long-term? Some questions to consider:

- What if you had access to additional datasets? Would your approach change?
- Are all products good candidates for your approach? Why or why not?
- What are the risks associated with your approach? What are the costs?
- What are the long-term benefits of your approach?

Deliverables

You will submit your work via email to Kate (Kathryn.b.farmer@gsk.com). While you may be asked follow-up questions about your response in follow-up interviews, the documents you submit should stand on their own without additional presentation or explanation.

Report

Format	Undefined - our goal is to see how you use data to tell a story, not how well you follow pre-set guidelines
Time Commitment	Approx. 3 hours (2 hour for analysis, 1 hour for report creation)
Audience	Non-technical (product leaders and scientific directors)

Proposal

Format	PDF or PPT
Time Commitment	Approx. 1 hour (we are looking for general ideas, it doesn't need to be thoroughly researched)
Audience	Technical leadership and business partners

Review Criteria

Below are some guidelines for how we will be assessing your report and proposal

Report

	Methodology	**Storytelling**	**Visualizations**
Excellent	Your analytical methods are thorough, highly accurate, and contain a variety approaches which examine the data across multiple dimensions	You deliver clear insights, with compelling recommendations which inspire informed decision making	You present visualizations which are expertly designed and provide the reader with a complete visual summary of your findings
Satisfactory	Your methods are sound and reasonable, leading to findings which are trustworthy and reliable	You share your findings with clarity, and the connection to the underlaying data is straightforward	You have relevant visualizations which accurately depict the data and emphasize your message
Poor	Your methods are unsuitable to the data, leading to inaccurate or unreliable results	Your recommendations are unsupported by the data, or your message is unclear	Your visualizations are misleading, unhelpful, or missing

Proposal

	Clarity	Feasibility	Business Value
Excellent	Your project proposal is clear, and the message from idea, to execution, to evaluation is detailed enough for a technical audience, while also being understood by business leaders	The costs and risks of your proposal are well understood and are justified for the benefit of the work. The work is attainable with minimal costs and a small team	You propose a high value solution which demonstrates insight into the stakeholder needs, with minimal overhead. You offer fast delivery of value with long-term returns
Satisfactory	Your proposal is clear to a technical audience, but may need additional explanation for a non-technical reader	Your approach is technically feasible with minimal risk and moderate cost. Risks and costs are not discussed in detail, but meet industry norms	You propose a solution which offers good value for the cost of the effort, and which would further text analytics and NLP capabilities
Poor	Your proposal is unclear, and ideas do not follow in a logical order. Your ideas are hard to follow, either because they are too abstract or the technical details cloud the intention	Your approach is unfeasible or unsustainable in a typical data science team. It may include considerable risks to budget, privacy, security, or value	You have low or poorly articulated business value associated with your project. It emphasizes technical achievement over business outcomes.