Data Scientist for Business Intelligence Take-Home Test

You are a Data Scientist at Collective Health and were asked to (1) run summary statistics on healthcare plan membership and claims data and (2) develop an analysis of the drivers of total healthcare spend. You will be analyzing the cost and utilization for two employer groups -- Mickey Mouse Inc. and Donald Duck Co. -- that are currently self-funding their medical benefits. You are given the first four months of data for 2020 and looking to answer the following questions to analyze the population.

Using your preferred method of data analysis (SQL, Python, R, etc.), please answer the following questions.

IMPORTANT: Make sure to show all the work leading up to your answer (ex. any code you wrote) and consider your presentation and how easy it is to follow your logic and results.

Questions

1. How many subscribers were effective on the PPO medical plan for each month for Donald Duck Co.?
2. What is the distribution of claim dollars (employer-paid basis) across various coverage tiers?
3. Compare the allowed PMPM (per member per month) due to respiratory conditions between the two employer groups.
4. Are the claim paid amounts statistically significantly different between the two companies? For simplicity, you may assume that paid claims fit a normal distribution.
5. We want to identify the drivers of medical costs by analyzing the total allowed amount accrued per member for both employer groups.
    a. Who are the top 5 large claimants (based on the total allowed amount) for Donald Duck Co. across the entire time captured in the data? Your response should include member ID, gender, relationship, and the total allowed amount.
    b. We want to identify the drivers of the total allowed amount per claimant over the entire period. Please run a multivariate OLS regression with the total allowed

dollars across all claims per member (same methodology as 5a above before filtering for top 5 claimants) as the dependent variable and major clinical condition and employer as the independent variables. What is the goodness of fit for this model? Assuming there is an omitted variable bias, how would you improve the model?

6. List and describe any data issues you found along with assumptions/fixes you had to make to answer the above questions.

Data Description

Attached are two datasets (csv files) showing membership and claims for two employers for 4 months (2020-01 ~ 2020-04). Please see below for the description of each field in the two datasets:

Membership

member_id	Unique member identifier number
employer_name	Name of employer
gender	Gender of member
effective_month	Month that the member is effective for (i.e. if a member was on an effective plan for all 4 months, that member will have 4 rows showing 4 different months in this field)
coverage_tier	Employee only, employee + spouse, employee+child(ren), employee+family
relationship_medical	Shows whether the covered member is the subscribe (employee), spouse, or child
plan_type	Shows what plan the member is covered for (medical, pharmacy, dental, vision)
plan_name	Name of covered plan

Claims

member_id	Unique member identifier number
employer_name	Name of employer
claim_id	Unique claim identifier number

line_number	Each claim can have multiple claim lines. This field shows the line number for each claim
benefit_category	Claim service category
major_clinical_condition	Broader indication of clinical condition
minor_clinical_condition	More granular indication of clinical condition
claim_status	Shows whether the claim was finalized (and paid) or denied
claim_grouping_major	Broader definition of service category (inpatient, outpatient, professional)
claim_grouping_minor	More granular definition of service category
is_in_network	Network status
service_month	Month that the claim incurred on
allowed_amount	Dollar amount that provider was paid for (includes employer paid cost and member cost sharing)
paid_amount	Dollar amount that employer paid for the claim