INSURIFY - QUANTITATIVE ANALYST TAKE HOME INTERVIEW

Please use any programming language you would like to answer the following questions.

Submission Instructions:

Part 1: Submit code + answers to the questions (in pdf form)
Part 2: Submit code + answers to the questions (in pdf form)
Part 3: Submit code + figure and short explanation (in pdf form)
Part 4: Submit code and a few different examples of your code in action. No need to use the same inputs as the given examples

Some of these questions will be vague (that is intentional). There are not always clear cut right or wrong ways to do things at Insurify. Do your best to answer the questions using the data.

If there is anything ambiguous about the questions themselves, you can email blake@insurify.com for clarification.

Section I: Looking at Demographics + Medical Charges

data-set: insurance.csv

The original data for this exercise are here. I have taken a sample and slightly modified it:
https://www.kaggle.com/mirichoi0218/insurance

Please do not look at any of the kernels for this data-set (as I am very familiar with them) and answer the following questions yourself.
However, feel free to use Google as a general resource

1. Read in Data and report summary statistics (mean + std / frequency) for age, sex, bmi, children, smoker, and charges) by region.

2. How would you characterize this population? Use figures/tables to support you answer

3. In this sample, is female age different from male age?

4. Is there a difference in smoking rates between those who have kids and those who do not?

5. Is there a difference in smoking rates between regions?

6. Are there any instances of high collinearity in this data-set?

7. A coworker wants to know whether:
   - being male affects medical cost
   - being a smoker affects medical cost
   - what is the effect of each additional year on medical cost

Build a model(s) to answer this. Please detail any assumptions you make / how you checked them.

8. What are some of the limitations of the model(s) that you built?

9. If a hospital was looking to minimize its cost, what population should it target based on your analysis?

10. A Co-worker asks you whether you should use AIC, BIC, or R-squared to evaluate one model over another. Explain to them (in layman's terms)
each of these metrics and why you should use one over the other.

11. If this was a time-series panel as opposed to cross-sectional data, how would you have changed your model?

12. Your boss comes to you and says we wants to limit patients that may cost more than 50K. You don't need to write code to do this,
but outline how you could create a model that would take a new patient's characteristics and output the probability that their medical charges would be over 50K.

How would you evaluate the effectiveness of your model?
Once your boss gets your model, he/she sees that your model outputs probabilities.
He/She then asks you what probability cut-off should we use to exclude patients (ie if prob is above X, we exclude them. Tell us what X should be)

Section II: Checking Conversion Rates

On 9/5/2018, we implemented a new feature on our website aimed at increasing conversion. Unfortunately, for this particular product change, we were
unable to A/B test it. However, we do have data on people who came to our flow and whether they converted. It is your job to determine whether the
product change improved conversion rates

Based on the above findings, how would you recommend that we proceed as a company with product improvement?

Use the following data: conversion_rates.csv

Here is the data dictionary:

NOTE: THIS DATA IS FAKE

Date - Date they came to our site
male - whether the person is a male
age - age of the person
has_insurance - person currently has auto insurance
came_from - The place they came to our site from
reached_end - person reached end of flow and submitted an application

Section III - Visualizing Data

NOTE: THE DATA IS FAKE

You are given two data-sets.

Data-set 1 comes from our internal database (names_id_age.csv)
Columns:

Column Name	Column Description
Id	Person ID
Name	Encrypted Name of Person
Lead_ID	Numeric ID that we generate when a person clicks out of our site to an insurance carrier
Lead Type	Either A, B, or C. A indicates highest intent lead (most likely to buy) and C indicates lowest intent

Data-set 2 (lead_sale_stats.csv) comes from partners and tells us which leads became sales and how much we made off of that sale

Lead_sale_stats.csv

Column Name	Column Description
lead_id	Some Partners are in form {lead_type}_{lead_id} and others are {lead_id}_{lead_type} where lead_id matches data-set 1 and lead_type matches the lead type from data-set 1
Bought_policy (0 / 1)	Whether a person bought a policy. Equals 1 for people who bought policy
policy_amount	Amount of money that we made from the sale

Come up with a single figure that uses the data to help us determine how we can grow as a business. Produce a single figure (with a line or two description if you would like) to help our executive team grow the business

Section IV - Close Rate Statistics

NOTE: This is Fake Data

File: agency_close_rates.csv

Column Name	Column Description
period	Date
Agency Name	Name of Agency
Leads	Number of people who clicked out to that Agency's website from Insurify
Sales	Number of People who made Sales

Write code that produces close_rate statistics where close_rate = sales / leads

Here are your inputs:

Input Parameter	Explanation
list_of_agencies (list)	The list of agencies that I want close rate for. If list is empty, I want all of the agencies
start_date (string)	Do not look at data before this date. If blank use the first available date for each agency.
end_date (string)	Do not consider data after this date. If blank use any data after the start_date
aggregated (True or False)	If True, print out average close_rate for each agency in agency_list looking at data between start_date and end_date. If False, produce a time-series plot with date as x-axis and close_rate as the y-axis for each agency

Examples

So if you were to write a method called get_close_rates here is what it would look like:

Input:
get_close_rates(list_of_agencies=[], start_date=None, end_date=None, aggregated=True)
Output:
Prints out average close rate for each agency using all available data

Input:
get_close_rates(list_of_agencies=[Agency_A], start_date=2019-05-01, end_date=None, aggregated=True)
Output:
Print average close_rate for Agency A looking at data on and after 2019-05-01

Input:
get_close_rates(list_of_agencies=[Agency_A], start_date=2019-05-01, end_date=None, aggregated=False)
Output:
Produce time-series plot for Agency A where the x-axis is date and y-axis is close_rate for dates after 2019-05-01

Input:
get_close_rates(list_of_agencies=[Agency_A, Agency_C], start_date=None, end_date=2019-05-10, aggregated=False)
Output:
Produce time-series plot for Agency A where the x-axis is date and y-axis is close_rate for dates before 2019-05-10 for Agency A and Agency C