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

6. 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.

7. Your boss comes to you and says we want 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: These data are 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: These data are fake**

You are given two data-sets.

File 1: names\_id\_age.csv, comes from our internal database

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 |

File 2: lead\_sale\_stats.csv, comes from partners and tells us which leads became sales and how much we made off of that sale

|  |  |
| --- | --- |
| 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