# Cota Data Case Interview 2018

## Instructions

The goal of this exercise is for our team at Cota to review your data sense & problem-solving approach and for you to get a feel for some of the foundational day-to-day data analysis required for this role. Feel free to use whichever open-source analytical tools and software libraries you are most comfortable with (R, Python, RMarkdown, Jupyter Notebook, SQL, etc.)

As working with unclean data is a pre-requisite for most analytical roles, we've intentionally introduced data errors that you will need to identify and correct. Then you will perform some basic calculations, aggregations and visualizations on the cleaned dataset to answer some case questions.

The data case and write-up should take between 2 to 4 hours total depending on your analytic experience. If you find you are stuck on one question, skip it and move on to the next one. Please don't spend more than 4 hours total on this exercise; we value your time and one of our team values is working efficiently.

---

## Submission

Please respond back to **tommywu@cotahealthcare.com** by **6pm, Sunday, 4/1/2018** with the following attachments:

1. A short 1-2 page write-up with your answers to the case questions as a word document or PDF. Where appropriate, please briefly explain your problem-solving approach in addition to the answer itself. Include the graphics and tables used in your analysis (they may exceed the page limit).

2. The code you used to answer each case question consolidated into one script, markdown file or notebook. Comment your code to explain what you did. Please also specify what operating environment, software version, and packages you used so that we may replicate your analysis.

By design, there are no numerically "perfect" answers to this case; your submission will be judged much more on your problem-solving approach and the elegance of your code.

---

## Scenario

Dr. Dai is a world-renowned expert on malignant micro-teratomas of the mediastinum, a rare, incurable cancer only detectable in advanced stages. She received NIH funding to track survival outcomes for 5 years of patients diagnosed in her clinic between January 1st, 2007 and January 1st, 2012. Dr. Dai is not a data scientist herself and has previously had to rely on summer interns that meant well but did not have the best data maintenance practices. The submission deadline for renewing her funding is coming up next week and she has hired you, a data science consultant, to make sense of the data her clinic has collected so far and provide some insights for the research proposal.

**In the attachment "simulated_patients.csv**" you will find a dataset of terminally-ill cancer patients. Each row should represent a unique patient in the dataset.

**Table 1: simulated_patients.csv**

| Column_Name | Data_Type | Description |
|---|---|---|
| patient_id | string | Patient identifier |
| patient_name | string | Full name |
| sex | string | Gender [male \| female] |
| ethnicity | string | Ethnicity [White \| Black \| Asian \| Hispanic] |
| age_at_diagnosis | integer | Age at cancer diagnosis |
| smoking_status | string | Smoking status [smoker \| non-smoker] |
| treatment | string | Type of cancer treatment received: patients may be treated with standard chemotherapy, a new targeted therapy, or be placed on active surveillance (no treatment given). |
| date_of_diagnosis | date | Date patient was diagnosed with disease |
| date_of_final_obs | date | Date of final observation for the patient |
| final_obs_status | string | Patient status at final observation date [expired \| alive] |

**Additionally, you were provided the "simulated_revenue.csv" dataset.** This was pulled from a separate billing system untouched by the interns and is error free.

**Table 2: simulated_revenue.csv**

| Column_Name | Data_Type | Description |
|---|---|---|
| patient_id | string | Patient identifier |
| date_of_charge | date | Date of charge |
| charge_amount | numeric | Amount charged ($) |

## Case Questions

1. Healthcare data sets are known to have errors and certain anomalies. While it may not be possible to detect and rectify all the errors, careful inspection of the data can help rectify some obvious errors. Examine the "simulated_patients.csv" dataset. If you detect any errors or anomalies, state why you think it is an error and the methods and assumptions you used to fix the error. Please use this rectified data set for further analysis**. [Spend no more than 45 minutes on this question]**

2. Using appropriate graphs and/or tables, summarize gender distribution for patients in the different ethnic groups. Is gender distribution associated with ethnicity?

3. Using appropriate graphs and/or tables, summarize age distribution for patients in the active surveillance group and patients on treatment (not on active surveillance). Is age distribution associated with being on active surveillance or not?

4. Using appropriate graphs and/or tables, investigate if there are any other strong associations between patient characteristics. What, if anything, did you discover? **[Spend no more than 20 minutes on this question].**

5. What is the average amount charged for patients who are 65 years or older at diagnosis? Among those charged $100,000 or more, what proportion of the patients were 65 years or older at diagnosis?

6. Dr. Dai is interested to see what factors if any are associated with the physician's choice of targeted therapy vs. chemotherapy. What, if anything, did you discover?

7. What is the median time to death in the sample? (For questions 6,7, and 8: time is measured from the date of diagnosis)

8. What is the probability that a patient will be alive for at least 1 year?

9. Is survival different for patients treated with chemotherapy vs. patients treated with targeted therapy?