**Outliers Detection**

Download Datasets

*This data project has been used as a take-home assignment in the recruitment process for the data science positions at Physician Partners.*

**Assignment**

In this exercise, we are providing you with an example of financial data (all numbers are fictional).

You need to create an algorithm that can find outliers in this data by one column / several columns. E.g. some members have extremely high costs in the current month and your solution should be able to detect such records. Think about features and how you would explain them to business people.

**Data Description**

The dataset **sfr\_test.csv** contains fictional financial information about customers of the company.

Below is the description of the selected columns from this dataset:

* **member\_unique\_id** - member's ID
* **gender** - member's gender
* **dob** - member's date of birth
* **eligible\_year** - year
* **eligible\_month** - month
* **affiliation\_type** - doctor's type
* **pbp\_group** - health plan group
* **plan\_name** - health plan name
* **npi** - doctor's ID
* **line\_of\_business** - health plan type
* **esrd** - True if patient is on dialysis
* **hospice** - True if patient is in hospice

The remaining columns contain various financial indicators. **All values in these financial columns contain a $ sign.**

**Practicalities**

Please provide a Jupyter notebook describing your approach as a result of this task.

Since it is an open-ended task, your solution will need to include your thought process, including data exploration, outlier detection approaches, results, and interpretation. Your code must be well-structured, and all outputs (numerical and visual) are properly documented and easily readable. Always keep in mind that the final document would be read by someone with a business background, so be sure to clearly explain any technical terms and metrics. You should spend no more than 2-3 hours on this task.