TRỰC QUAN HÓA DỮ LIỆU – CQ2018/32

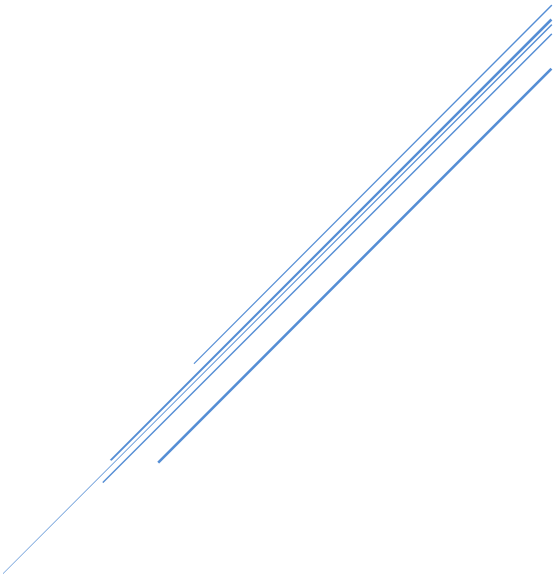
GVHD: TS Bùi Tiến Lên

Bài tập cá nhân 1

Thông tin sinh viên:

MSSV: 18120254

Họ tên: Nguyễn Huy Tú



Khoa Công nghệ thông tin

Trường Đại học Khoa học Tự nhiên

Đại học Quốc gia TP.HCM

MỤC LỤC

THÔNG TIN SINH VIÊN

1. Thông tin sinh viên.

Họ tên: Nguyễn Huy Tú.

Mã số sinh viên: 18120254.

Email: 18120254@student.hcmus.edu.vn

2. Các công việc đã thực hiện.

STT	SV thực hiện	Tên công việc	% Hoàn thành
1	18120254	Bài tập 2.1	100%
2	18120254	Bài tập 2.2	100%

BÁO CÁO

1. Bài tập 2.1: improve this table.

***STEP 1**: Review the data in Figure 2.1a. What observations can you make? Do you have to make any assumptions when interpreting this data? What questions do you have about this data?*

Hình 2.1a cho thấy cơ cấu khách hàng và doanh thu theo các loại hạng. Cơ bản, ta có thể thấy:

- Khách hàng **hạng C** chiếm thị phần **nhiều nhất.**
- Tuy nhiên, khách hạng **hạng B** lại đem lại **nhiều lợi nhuận nhất**.
- Ngoài ra, khách **hạng A+** chỉ chiếm số lượng bằng 1/30 khách hạng B nhưng đem lại **doanh thu giá trị** hơn ¾ giá trị khách hạng B mang lại.

Có vẻ khách hàng hạng C thuộc phân khúc bình dân, tuy giá trị từng cá nhân mang lại không cao nhưng chiếm số đông. Trong khi đó khách hạng A, A+ thuộc phân khúc cao cấp hơn, số lượng ít nhưng mang lại nhiều giá trị. Nếu có thể tập trung phát triển ở 2 loại này thì lợi nhuận mang lại sẽ rất lớn. Khách hàng ở phân khúc cao cấp cũng có xu hướng trung thành hơn so với bình dân.

Nhược điểm của bảng dữ liệu này là chưa được sắp xếp theo % *Doanh thu* nên nhìn vào, ta dễ nhầm lẫn rằng khách hàng *Hạng A* chiếm phần nhiều thị phần nhất.

***STEP 2:** Consider the layout of the table in Figure 2.1a. Let's assume you've been told this information must be communicated in a table. Are there any changes you would make to the way the data is presented or the overall manner in which the table is designed? Download the data and create your improved table.*

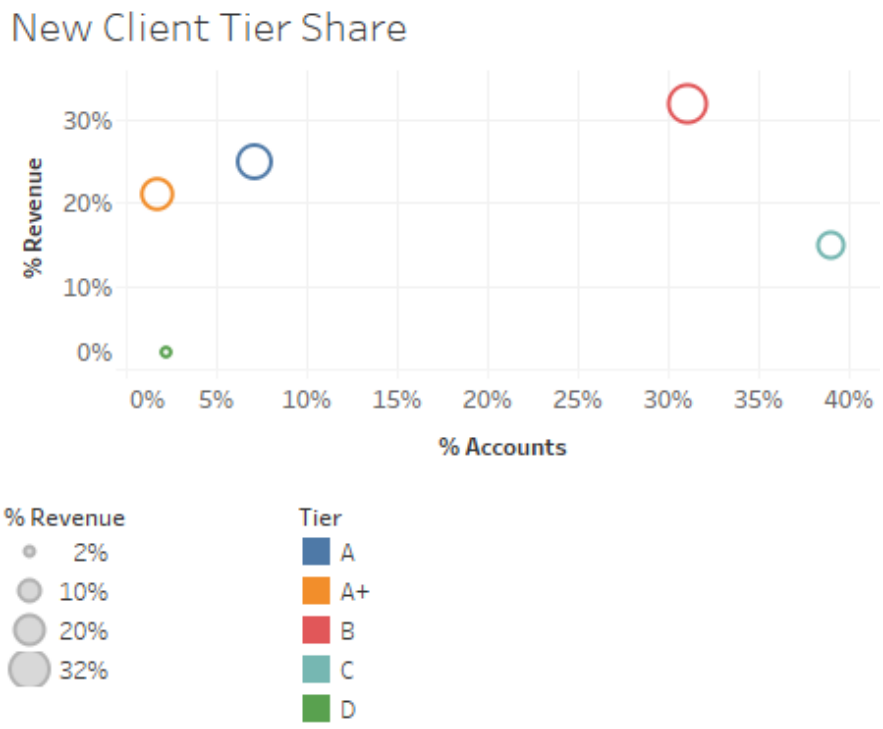
Bảng dữ liệu trên được sắp xếp theo *Loại hạng* người dùng, đây là một biến độc lập, thông tin không mấy quan trọng bằng các biến phụ thuộc như *Doanh thu* hay *Số lượng tài khoản*. Vì vậy, ta nên **sắp xếp** lại bảng trên theo chủ đề muốn diễn tả.

Ngoài ra, ta có thể **lược bỏ** cột *#of Accounts* và *Revenue* vì cả 2 thông tin cột này đều được thể hiện qua các cột %. Ta chỉ giữ lại nếu muốn biểu diễn số liệu chi tiết.

Tier	*% Accounts*	*% Revenue*
B	31.07%	32%
A	7.08%	25%
A+	1.75%	21%
C	39.06%	15%
D	2.21%	2%

STEP 3: *Let's assume the main comparison you want to make is between how accounts are distributed across the tiers compared to how revenue is distributed and that you have the freedom to make bigger changes. How would you visualize this data? Create a graph in the tool of your choice.*

Để thể hiện mối quan hệ giữa doanh thu và hạng tài khoản - mối tương quan đó là mạnh hay yếu, tích cực hay tiêu cực, tuyến tính hay phi tuyến tính, em sẽ sử dụng **scatter plot**. Biểu đồ này cũng rất tốt để xác định các điểm ngoại lệ và các khoảng trống có thể có trong dữ liệu.



2. Bài tập 2.2: visualize!

STEP 1: *Apply* ***heatmapping*** *to the second column of values.*

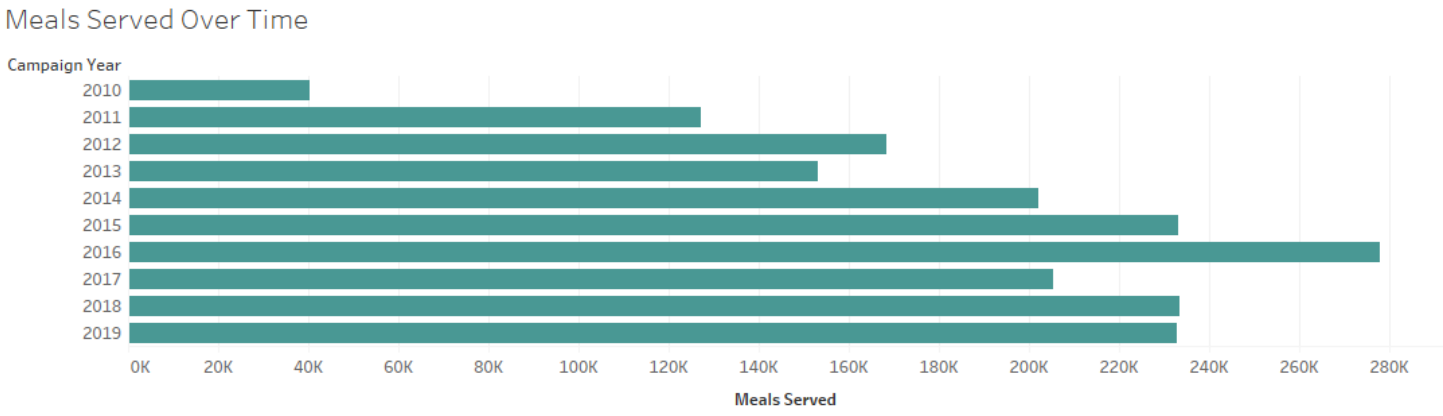
Meals Served Over Time

Campaign Year	
2010	40,139
2011	127,020
2012	168,193
2013	153,115
2014	202,102
2015	232,897
2016	277,912
2017	205,350
2018	233,389
2019	232,797

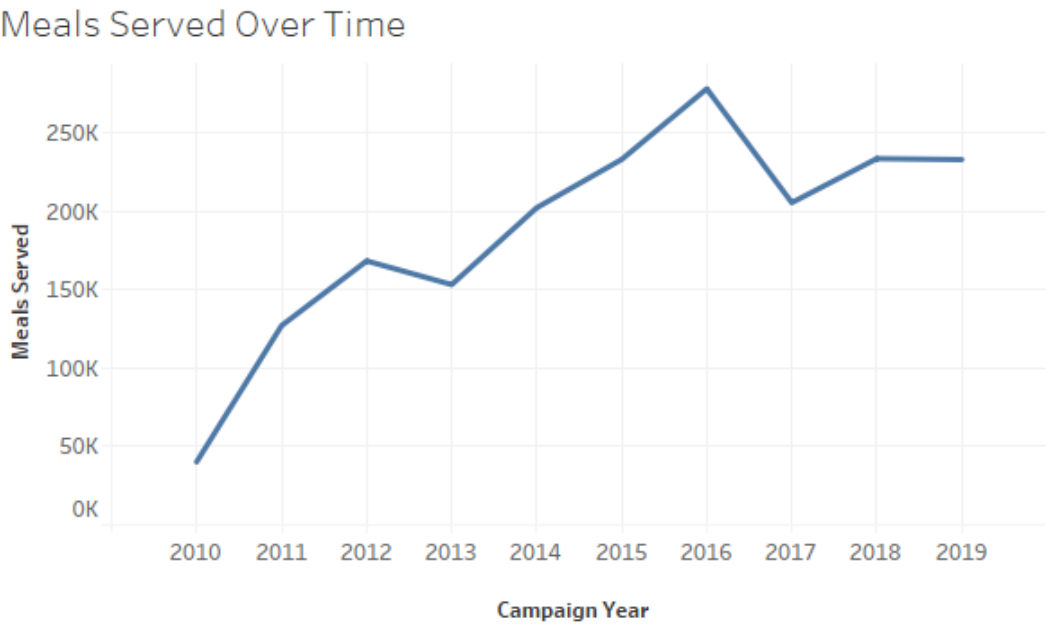
Meals Served

40,139 277,912

STEP 2: *Create a* ***bar graph.***



STEP 3:** Create a **line graph.



***STEP 4:** Choose: **which of the visuals you've created do you like best?** Are there any other ways you would graph this data?*

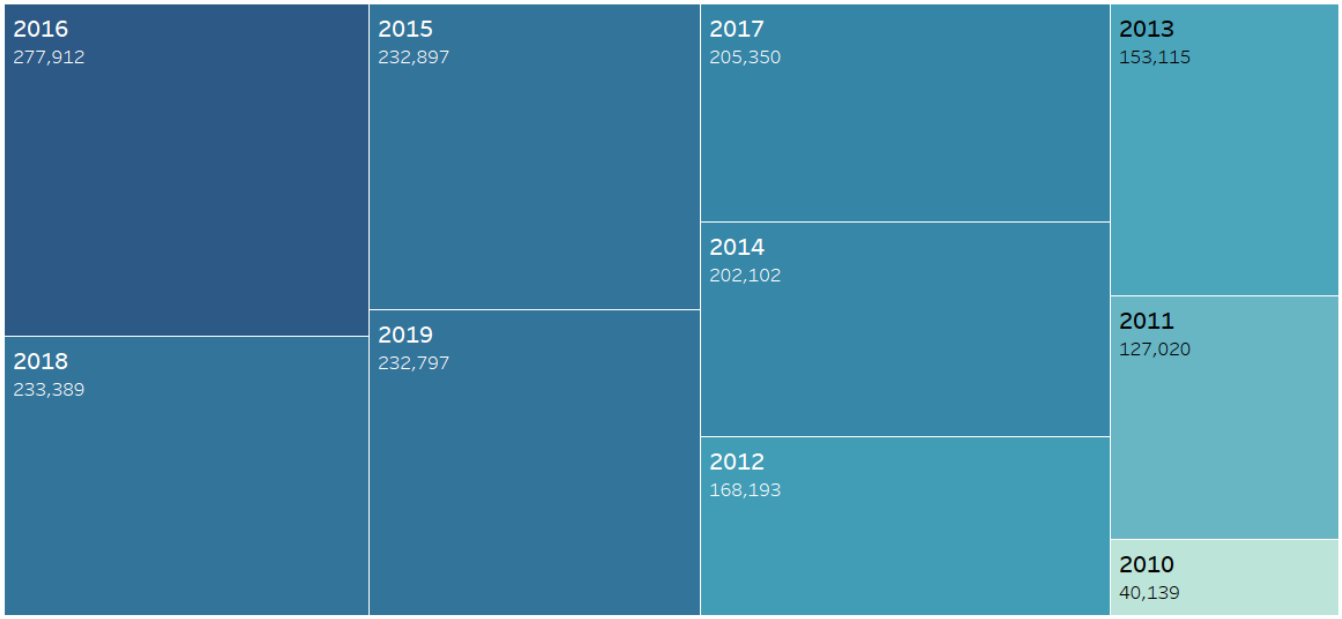
Tùy vào mục đích cụ thể mà ta có thể chọn biểu đồ phù hợp để thực hiện trực quan hóa dữ liệu. Với ví dụ trên ta có thể thấy:

- **Bar graph**: thể hiện được đầy đủ thông tin. Hiệu quả hơn nếu dùng để so sánh.
- **Line graph**: thể hiện được xu hướng, sự thay đổi số lượng bữa ăn theo thời gian. Biểu đồ này khiến ta cảm thấy giá trị này là liên tục chứ không rời rạc.
- ***Heatmap***: thích hợp với thể hiện độ đo, dễ đọc qua 1 lần nhìn. ➔ Em chọn biểu đồ này vì qua đó ta có nhiều câu chuyện để kể: *thấy được sự thay đổi qua các năm; năm nào cao nhất, thấp nhất. Màu sắc đẹp cũng là điểm nổi bật của loại biểu đồ này.*

Một cách khác để biểu diễn dữ liệu này là sử dụng ***treemap***.

Về tổng quan, *treemap* khá giống *heatmap*. Tuy nhiên, khác với *heatmap*, dữ liệu phải được *sắp xếp* và *lược bỏ giá trị rỗng* trước khi vẽ. Em sẽ dùng biểu đồ này trong trường hợp muốn nhấn mạnh về cực đại và cực tiểu của dữ liệu:

Meals Served Over Time



Campaign Year and sum of Meals Served. Color shows sum of Meals Served. Size shows sum of Meals Served. The marks are labeled by Campaign Year and sum of Meals Served.

Meals Served
40,139 277,912