Milestone Report

Problem Statement:

An organization that has tens of thousands of clients a year, across many product (event) categories, is sitting on a lot of data. I initially was curious to identify what opportunities can be found in analyzing the data that could contribute to the business growth of the business. While the data indeed showed a number of interesting patterns, which will be discussed briefly below, it became apparent that the available data presents an interesting pricing optimization problem. Stated briefly: *Are the event prices that are being charged from attendees revenue maximizing?*

Dataset Description:

The dataset is of orders from a company that organizes dozens of annual B2B conferences of varying sizes and across several industries. While there was a single source - the company's CRM - there were two different data sets that were eventually merged to form the final analysis and model. The two data sets are: 1. Individual orders, this includes ticket price, time, type, etc. and 2. Event performance metrics, this includes metrics like event revenue, number of attendees, number of paid attendees and more.

A number of cleaning and wrangling operations were necessary:

- A few orders included erroneous symbols or currency symbols - these were cleaned up
- A small number of negative orders were removed
- A unique *'id'* variable was created from first names, last names and emails to eliminate the latter and proceed with the analysis without sensitive user data
- New variable was created to attribute sales source by registration type in the form of a new categorical series
- A new variable was created to denote the number of attendees in a group registration (the value of 1 for an individual registration with no group)
- A new seniority variable was created and populated by keywords from the individual job titles

Before proceeding to EDA, a number of data type conversions took place:

- Variables such as order type and order source were converted to categorical
- Order and event dates were converted to datetime
- Price data was converted to floats

More information on variable definitions and transformations can be found in the [Capstone proposal](#).

Initial Findings:

There are two ways in which attendees can register for the events: 1. passively - online, chat or via customer service or 2. actively - through a consultative phone conversation with a sales rep. Seems the growth in attendee registrations has come mainly through the latter.

I became curious about the motivation behind registrations, while pricing plays a major part in a purchasing decision, there could be other elements at play, such as scarcity, exclusivity and social pressure. In order to try and distill some of the key reasons, I looked at the top registration days for ticket purchases.

All Orders

Order_Date	mean	count
2018-01-31	1341.435957	141
2018-04-30	1254.289841	126
2017-01-31	1390.507265	117
2018-02-28	1103.890000	103
2017-02-28	1404.298600	100
2017-03-31	1354.684105	95
2017-09-29	1212.224839	93
2016-08-31	1114.764767	86
2015-01-30	1805.966395	86
2017-03-30	1099.386588	85

Passive Orders Only

Order_Date	mean	count
2015-01-30	2130.576000	25
2018-03-30	1313.184091	22
2011-02-18	2037.397727	22
2017-01-31	1524.832500	20
2011-08-31	1437.776316	19
2011-01-14	1031.750000	19
2015-08-31	1691.861111	18
2017-03-31	1634.436111	18
2011-01-31	1873.613889	18
2011-04-29	1332.002941	17

It became apparent that the last day of the month is a day with many ticket purchases. There could be two reasons for this: 1. Discounting - discounts often end on the last day of the month, which would offer a strong incentive for purchase and 2. Internal commission deadlines - As we previously saw, the majority of sales come through a sales team and their incentives also tended to align with monthly deadlines.
In order to distill which of the two was more likely at play, I ran the numbers based on passive orders only - this allowed us to ignore the commission deadlines as they do not apply to such sales. It became apparent that pricing, perhaps with added messaning about the pricing, is a strong factor in the decision to purchase event tickets.

The next step in developing a pricing model was trying to understand if there are types of attendees that would be willing to pay more for tickets. Looking at NAICS coding did present a distinguishing breakdown in which pharmaceutical and bit-tech companies did seem to be able to spend more on tickets. It is important to note that there are many event types and target attendees so it is important to review these figures in relation to the event subject area - this analysis was subsequently done for the four key event subject areas: Retail, Logistics and Procurement.

Account_SIC	mean	count
Bio-Technoloy	2615.150000	10
Research and Development in the Social Sciences	2579.173077	13
IQ - Nanotech research and development	2578.635185	27
Drugs and Druggists' Sundries Merchant Whole424210	2283.964286	35
Pharmaceutical Preparation Manufacturing	2236.649179	207
Research and Development in the Physical, En541710	2234.038462	26
Pharmaceutical and Medicine Manufacturing	2210.934615	65
Advertising/Marketing	2162.774773	176
Biotechnology & Drugs	2105.762647	34
Robotics	2077.650000	11

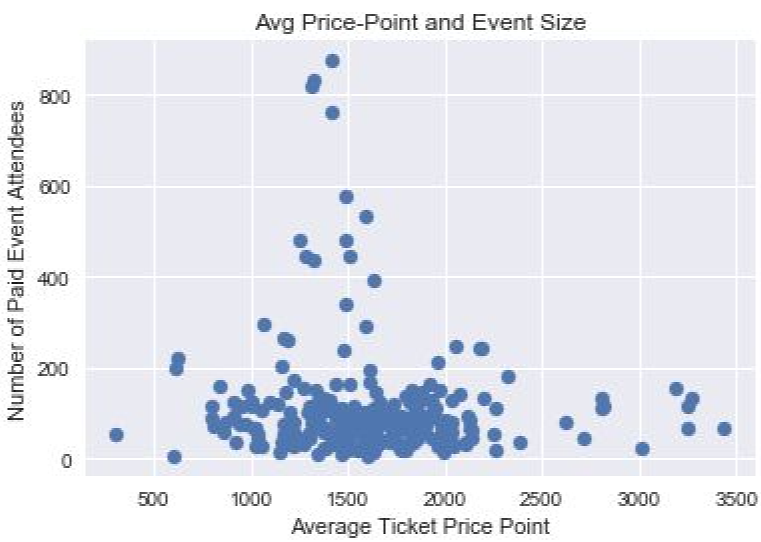
Along with a company type, another element of importance in distinguishing a type of registrant is their seniory. Two things were explored here: the source of the registration and the type of attendee. In terms of types of attendees, the data showed that while the events do attract many senior attendees (VPs and above), these attendees tend to come in the form of event speakers. Alternatively, while director level titles pay for events, the majority of paid attendees tend to be managers. In looking for the source of purchases, we see that more senior level attendees are more likely to choose a passive registration channel. *Chart and table displayed in appendix A.*

It could be interesting to incorporate seniority into the demand model as it is possible that seniority can be inversely correlated with the willingness to purchase a ticket at a certain rate.

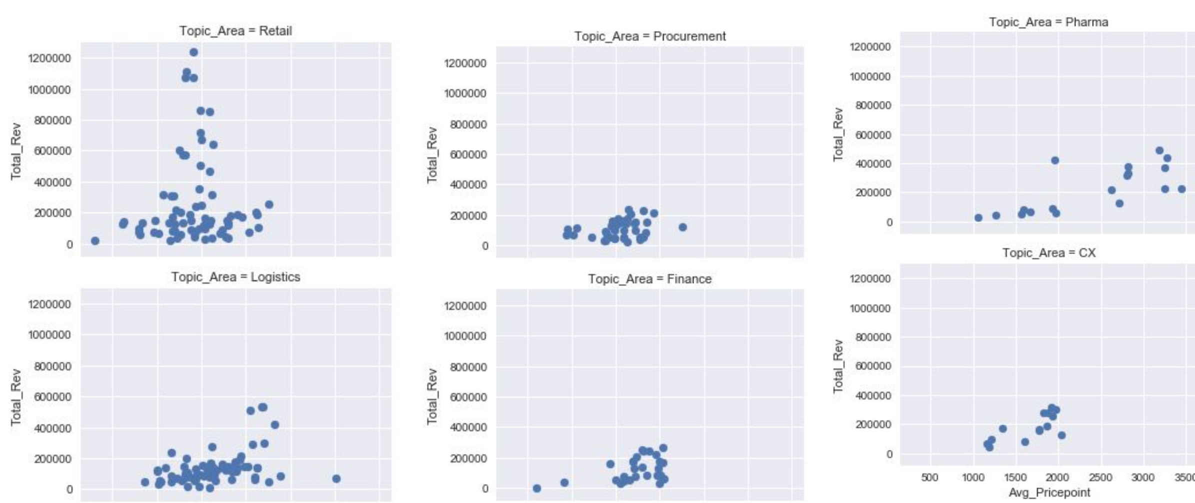
Pricing and Revenue:

Initial EDA was focused on understanding the correlation between average event prices and event revenue. The assumption was that there could be a point of diminishing return, at which demand would drop and the number of total ticket purchases, even at the higher rate, would not result in as high of an overall revenue.

Looking at all events showed a central tendency of average event prices in the $1,000-$2,000 range, while the number of paid attendees varied greatly within this range. There were also a number of outliers in terms of ticket price (<$700 or >$2,500) and in terms of paid attendees (events with more than 400 paid attendees).

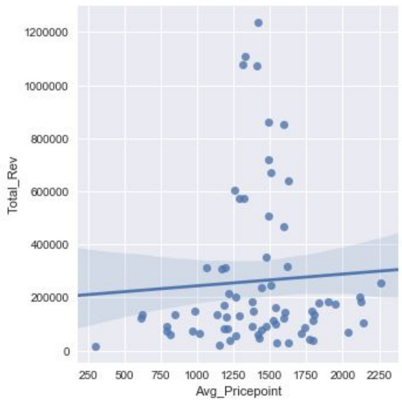


Looking at individual event subject areas showed varying demand distributions. *Additional breakdown can be found in appendix B.*



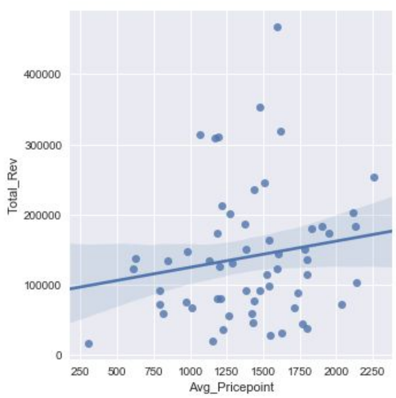
In looking at the retail events in more detail, I tried to identify a correlation between the average price increase and the overall association to increased revenue

```
44.20399685876672 199656.7614655343
For Retail Events - On average, a pricepoint increase of $100 is associated with added revenue of 4420.4 dollars
```



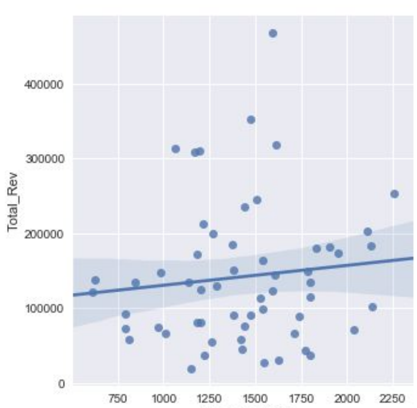
There are a number of outlier events, however, such as those that have a greater natural demand. Those were removed to see what a more typical event correlation might be:

```
37.405379663222256 87172.44871670128
After removing the outlier events - for Retail Events - on average, a pricepoint increase of $100 is associated with
added revenue of 3740.54 dollars
```



Lastly, a number of events had a very low average price-point, which means that they were driven mostly by comps. Removing these shows an even lower association:

```
26.491319779672047 104210.4962551559
After removing the outlier events - for Retail Events - on average, a pricepoint increase of $100 is associated with
added revenue of 2649.13 dollars
```



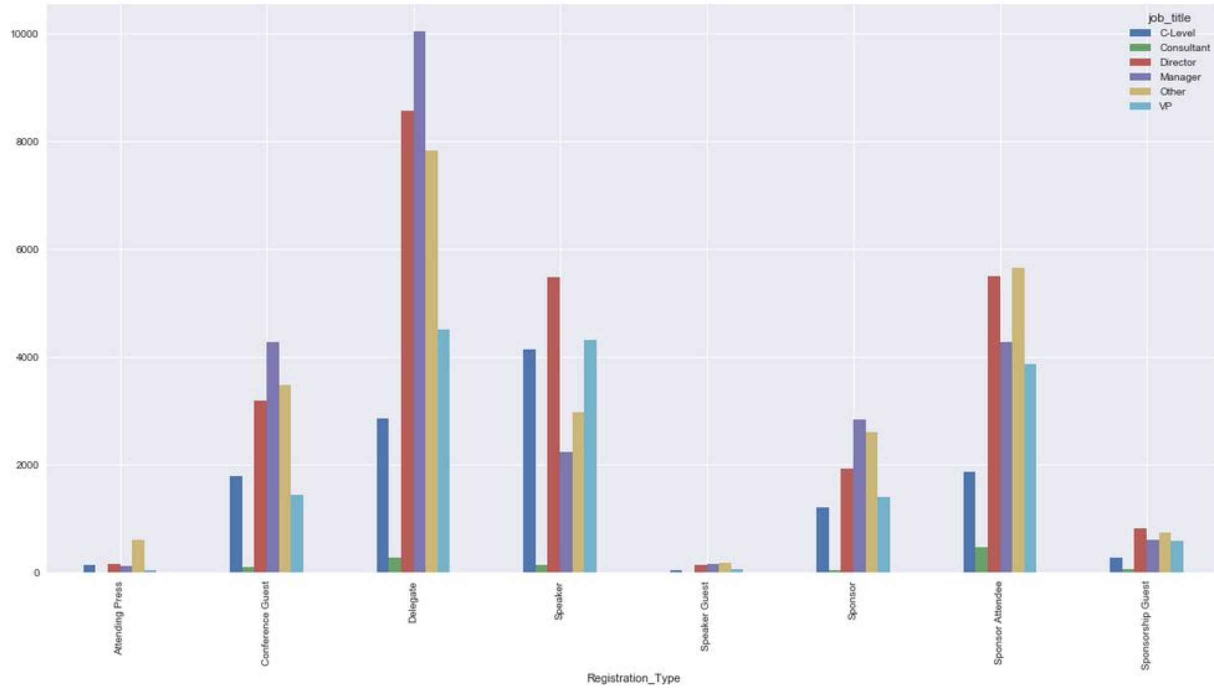
Next Steps:

In order to develop a revenue maximizing pricing strategy it will be important to develop a demand model that would consider event subject area, historical event size, type of attendee, time of purchase and other similar metrics.

Structured machine learning regressions could help predict price of ticket paid based on a number of criteria, such as time of purchase before the event, event type and more. Such prediction might take the form of a time series to understand whether a ticket price is likely to go up or down as we get closer to the event date.

Appendix A: Attendee types and purchase source

Bar chart of attendee seniority by type of attendee:

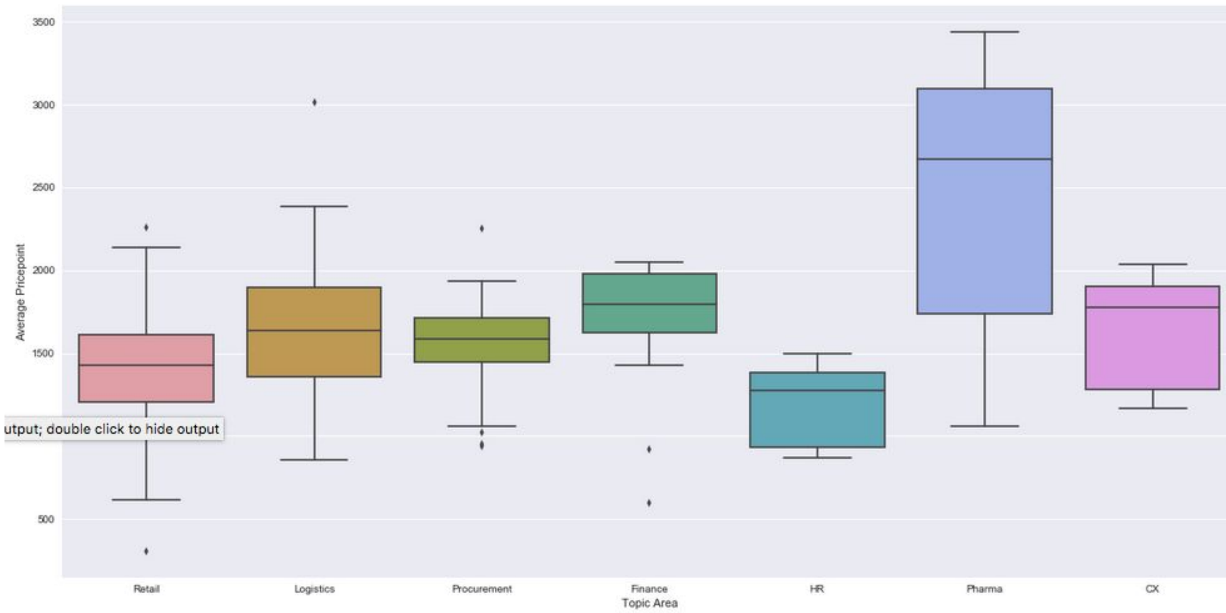


Attendee seniority by source of purchase:

	index	Sales_Source_Cat	job_title	Registration_Type	as_percentage
0	0	Delegate Sales	C-Level	3871	9.68
1	1	Delegate Sales	Consultant	270	0.68
2	2	Delegate Sales	Director	9582	23.96
3	3	Delegate Sales	Manager	12307	30.78
4	4	Delegate Sales	Other	9213	23.04
5	5	Delegate Sales	VP	4741	11.86
6	6	Marketing	C-Level	1214	9.60
7	7	Marketing	Consultant	166	1.31
8	8	Marketing	Director	3184	25.18
9	9	Marketing	Manager	2769	21.90
10	10	Marketing	Other	3474	27.48
11	11	Marketing	VP	1837	14.53
12	12	Production	C-Level	4192	21.11
13	13	Production	Consultant	139	0.70
14	14	Production	Director	5614	28.27
15	15	Production	Manager	2383	12.00
16	16	Production	Other	3151	15.86
17	17	Production	VP	4383	22.07
18	18	Sponsorship	C-Level	3072	9.72
19	19	Sponsorship	Consultant	508	1.61
20	20	Sponsorship	Director	7417	23.48
21	21	Sponsorship	Manager	7101	22.48
22	22	Sponsorship	Other	8236	26.07
23	23	Sponsorship	VP	5257	16.64

Appendix B: Attendee types and purchase source

Box plot of average price-point by topic area:



Box plot of delegate revenue by topic area: