Final 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

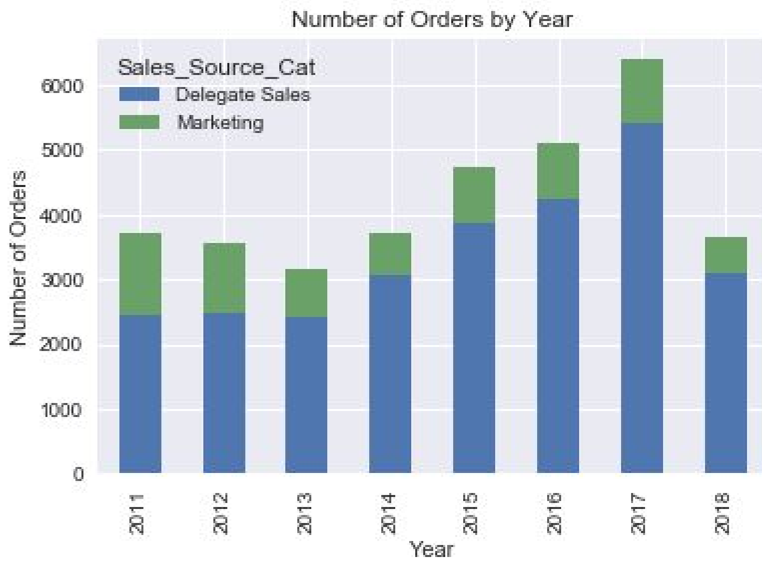
Before proceeding to EDA, a number of data type conversions took place:

- 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

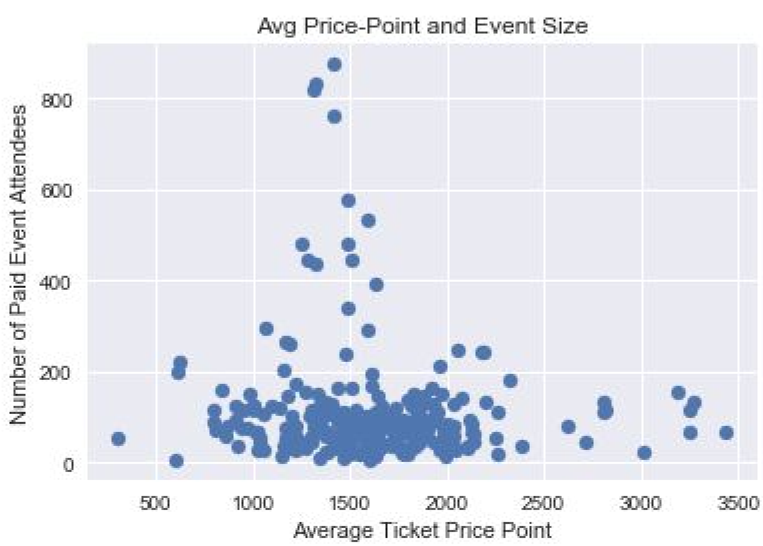
Order_Date	mean	count
2015-01-30	2130.576000	25
2018-03-30	1313.184091	22
2011-02-18	2037.397727	22
2017-01-31	1524.832500	20
2011-08-31	1437.776316	19
2011-01-14	1031.750000	19
2015-08-31	1691.861111	18
2017-03-31	1634.436111	18
2011-01-31	1873.613889	18
2011-04-29	1332.002941	17

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 messaging about the pricing, is a strong factor in the decision to purchase event tickets.

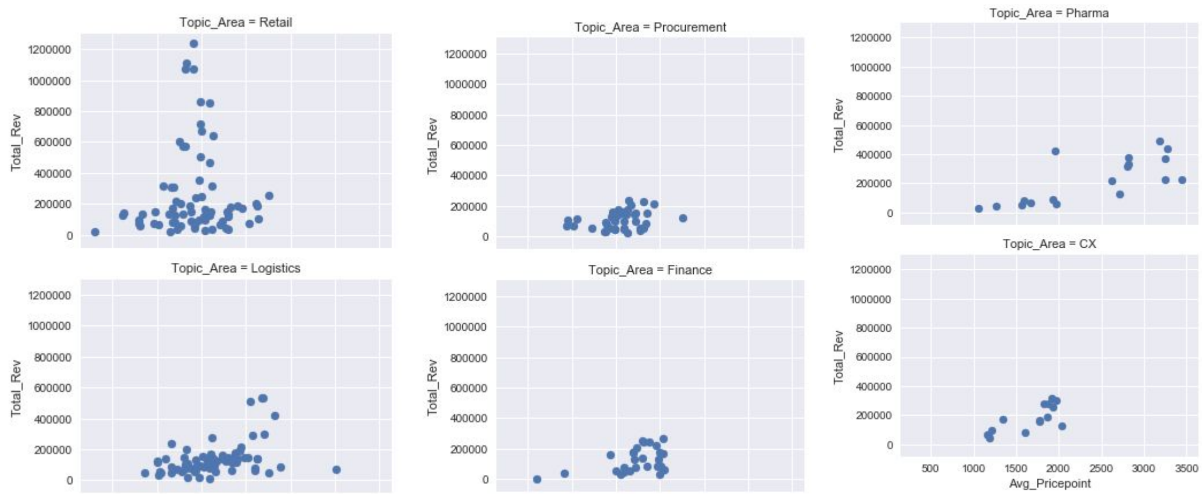
Pricing and Revenue:

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).

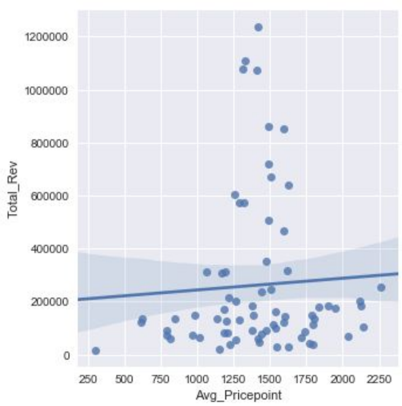


Looking at individual event subject areas showed varying demand distributions. *Additional breakdown can be found in appendix B.*



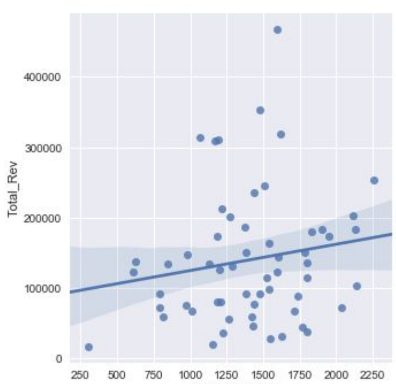
In looking at the retail events in more detail, I tried to identify a correlation between the average price increase and the overall association to increased revenue

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



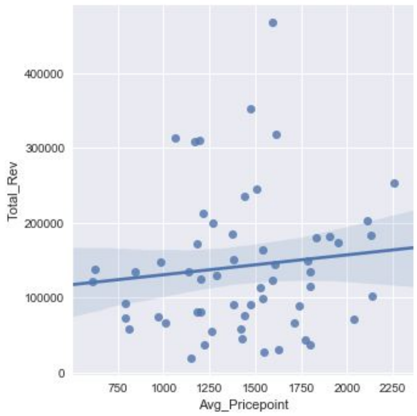
There are a number of outlier events, however, such as those that have a greater natural demand. Those were removed to see what a more typical event correlation might be:

```
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
```



Machine Learning:

After cleaning up the data for a linear regression and creating dummy variables for event types and seniority of registrant, an initial linear regression model gave a baseline model, which we've subsequently attempted to improve:

OLS Regression Results

Dep. Variable:	Total_Net_Price	R-squared:	0.599
Model:	OLS	Adj. R-squared:	0.598
Method:	Least Squares	F-statistic:	594.7
Date:	Mon, 04 Mar 2019	Prob (F-statistic):	0.00
Time:	17:51:22	Log-Likelihood:	-1.3463e+05
No. Observations:	17997	AIC:	2.694e+05
Df Residuals:	17951	BIC:	2.697e+05
Df Model:	45		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	68.8719	50.950	1.352	0.176	-30.996	168.740
Group_Size	-34.7636	1.564	-22.228	0.000	-37.829	-31.698
Days_Ahead_of_Event	-1.4576	0.070	-20.724	0.000	-1.595	-1.320
Total_Dels	0.7192	0.235	3.064	0.002	0.259	1.179
Total_Rev	-13.8399	7.255	-1.908	0.056	-28.060	0.381
Booking_Pattern_Comparison	0.3031	0.087	3.499	0.000	0.133	0.473
Total_Guests	-0.6846	0.067	-10.197	0.000	-0.816	-0.553
Avg_Cost_of_Acquisition	-0.2782	0.082	-3.383	0.001	-0.439	-0.117
Mktg_Dels	2.4522	0.693	3.539	0.000	1.094	3.810
Mktg_Rev	13.8355	7.255	1.907	0.057	-0.385	28.056
Mktg_Price_Point	0.1263	0.018	7.209	0.000	0.092	0.161
PP__act	0.0027	0.000	5.475	0.000	0.002	0.004
Sales_Dels	-0.2905	0.498	-0.583	0.560	-1.268	0.686
Sales_Rev	13.8405	7.255	1.908	0.056	-0.380	28.061
Sales_perc_of_ttl_rev	3.6173	0.858	4.215	0.000	1.935	5.300
Sales_Price_Point	0.1158	0.042	2.728	0.006	0.033	0.199
Num_Active_Inq	0.0002	0.076	0.002	0.998	-0.150	0.150
Active_Inq_Del	-0.7710	0.486	-1.585	0.113	-1.724	0.182
Num_Passive_PDF	0.0845	0.014	5.920	0.000	0.057	0.112
Passive_PDF_Del	-0.6714	0.355	-1.891	0.059	-1.367	0.025
Total_EQ_Rev	13.8402	7.255	1.908	0.056	-0.380	28.061
EQ_Price_Point	0.2573	0.041	6.226	0.000	0.176	0.338
EQ_perc_of_ttl_Rev	2.9773	0.809	3.681	0.000	1.392	4.563
Spex_Rev	-6.557e-05	2.83e-05	-2.318	0.020	-0.000	-1.01e-05
Spex_Last_Year	-6.268e-05	1.61e-05	-3.904	0.000	-9.42e-05	-3.12e-05
Num_Spex_EQs	0.0222	0.019	1.165	0.244	-0.015	0.059
Num_Spex_Props	0.0746	0.213	0.350	0.726	-0.343	0.492
Props_Last_Year	0.3274	0.182	1.795	0.073	-0.030	0.685
Num_TMs	0.0081	0.013	0.637	0.524	-0.017	0.033
Num_SPKRs	1.1335	0.299	3.787	0.000	0.547	1.720
Avg_Pricepoint	0.3799	0.085	4.449	0.000	0.213	0.547
CX	25.6570	16.031	1.600	0.110	-5.766	57.080
Finance	41.0102	16.004	2.563	0.010	9.641	72.379
HR	-131.0859	21.712	-6.038	0.000	-173.643	-88.529
Logistics	28.9191	11.716	2.468	0.014	5.955	51.884
Pharma	78.7483	23.122	3.406	0.001	33.427	124.069
Procurement	83.9195	13.178	6.368	0.000	58.090	109.749
Retail	-58.2963	11.560	-5.043	0.000	-80.956	-35.637
C-Level	26.8527	14.440	1.860	0.063	-1.451	55.156
Consultant	-24.8530	37.816	-0.657	0.511	-98.975	49.269
Director	20.6079	12.298	1.676	0.094	-3.498	44.714
Manager	4.2354	12.027	0.352	0.725	-19.339	27.810
Other	-8.3233	12.341	-0.674	0.500	-32.513	15.867
VP	50.3522	13.731	3.667	0.000	23.438	77.266
Delegate_Sales	-185.8159	27.289	-6.809	0.000	-239.306	-132.326
ACD Conversion	-241.6378	35.301	-6.845	0.000	-310.832	-172.444

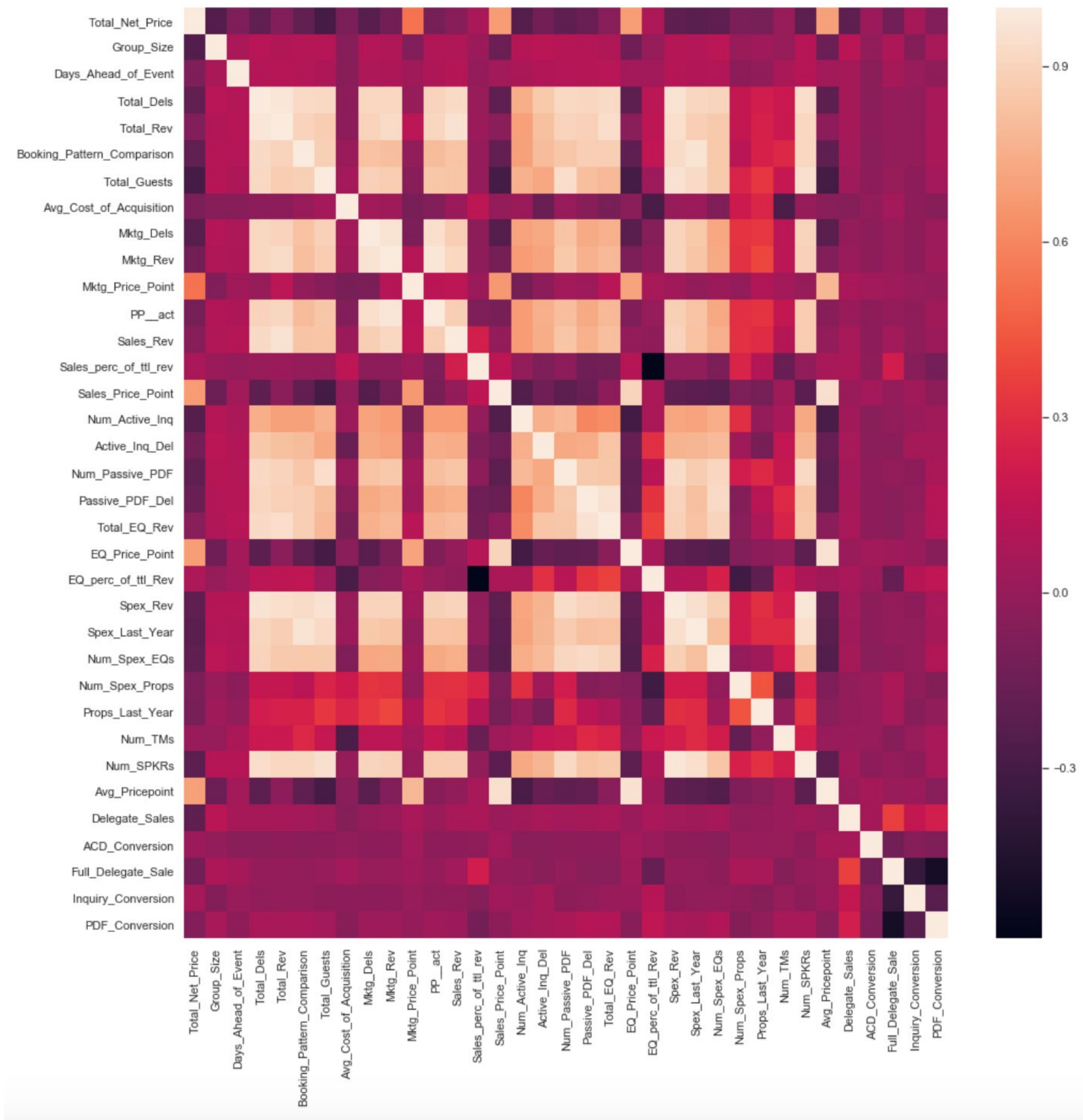
The root mean square error of this initial model was $429, which is nearly 25% of the average, or baseline, conference passes. Subsequent models were created to attempt to improve the RMSE.

Before evaluating the VIFs and addressing collinearity, there was an attempt to look at a model solely with some of the key features which would logically have the largest effects on the price paid:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        Total_Net_Price   R-squared:                       0.439
Model:                            OLS   Adj. R-squared:                  0.438
Method:                 Least Squares   F-statistic:                     739.8
Date:                Mon, 04 Mar 2019   Prob (F-statistic):               0.00
Time:                        17:51:29   Log-Likelihood:            -1.3765e+05
No. Observations:               17997   AIC:                         2.753e+05
Df Residuals:                   17977   BIC:                         2.755e+05
Df Model:                          19
Covariance Type:            nonrobust
==============================================================================================
                                coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                      1660.6272     16.554    100.316      0.000    1628.180    1693.074
Group_Size                  -42.6307      1.815    -23.491      0.000     -46.188     -39.073
Total_Dels                    0.5481      0.080      6.821      0.000       0.391       0.706
Days_Ahead_of_Event          -1.0273      0.082    -12.500      0.000      -1.188      -0.866
CX                          256.0893     18.671     13.716      0.000     219.491     292.687
Finance                     315.2602     19.187     16.431      0.000     277.651     352.869
Pharma                     1351.5533     19.678     68.682      0.000    1312.982    1390.125
Logistics                   260.5486     12.860     20.261      0.000     235.342     285.755
HR                         -387.3481     25.355    -15.277      0.000    -437.046    -337.650
Procurement                 110.5695     15.004      7.370      0.000      81.161     139.978
Delegate_Sales             -153.7055      5.514    -27.877      0.000    -164.513    -142.898
Full_Delegate_Sale          -93.3789      8.268    -11.294      0.000    -109.585     -77.173
Booking_Pattern_Comparison   -0.6910      0.052    -13.362      0.000      -0.792      -0.590
Total_Guests                 -2.0744      0.048    -43.442      0.000      -2.168      -1.981
Mktg_Dels                    -2.2637      0.300     -7.536      0.000      -2.852      -1.675
Num_Active_Inq                0.3499      0.057      6.164      0.000       0.239       0.461
Spex_Rev                      0.0005    1.7e-05     27.728      0.000       0.000       0.001
Director                     52.2137     10.322      5.058      0.000      31.981      72.446
Manager                      13.2002      9.587      1.377      0.169      -5.592      31.992
VP                          125.7505     13.479      9.330      0.000      99.331     152.170
Delegate_Sales             -153.7055      5.514    -27.877      0.000    -164.513    -142.898
==============================================================================
Omnibus:                     3724.419   Durbin-Watson:                   1.033
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            18785.460
Skew:                           0.913   Prob(JB):                         0.00
Kurtosis:                       7.660   Cond. No.                     2.48e+21
==============================================================================
```

While it is interesting to see that nearly all of the features had a statistically significant effect on the final price and that the baseline (constant) was close to the average price-point, the explanatory strength of this model decreased and do did the RMSE. The RMSE of this model went up to $507, meaning our error took us even further from the actual price paid.

As mentioned, the next step was to address collinearity. Firstly, the dummy variables were removed and subsequently by looking at a correlative heatmap (below) and the Variance Inflation Factors, some of the highly collinear variables were removed from the model.



The outcome of this exercise was a model in which all variables had low VIFs, although the explanatory strength of the model was lower than the baseline linear regression model:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        Total_Net_Price   R-squared:                       0.170
Model:                            OLS   Adj. R-squared:                  0.169
Method:                 Least Squares   F-statistic:                     282.5
Date:                Mon, 04 Mar 2019   Prob (F-statistic):               0.00
Time:                        17:52:44   Log-Likelihood:            -1.4117e+05
No. Observations:               17997   AIC:                         2.824e+05
Df Residuals:                   17983   BIC:                         2.825e+05
Df Model:                          13
Covariance Type:            nonrobust
============================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                    1983.2474     17.022    116.510      0.000    1949.882    2016.612
Group_Size                -60.9438      2.197    -27.736      0.000     -65.251     -56.637
Days_Ahead_of_Event        -0.6802      0.099     -6.853      0.000      -0.875      -0.486
Avg_Cost_of_Acquisition    -0.7263      0.087     -8.362      0.000      -0.897      -0.556
Num_Active_Inq             -1.6651      0.077    -21.703      0.000      -1.816      -1.515
Active_Inq_Del              3.2628      0.235     13.910      0.000       2.803       3.723
Num_Spex_EQs               -0.0659      0.009     -7.392      0.000      -0.083      -0.048
Num_Spex_Props              0.9358      0.228      4.100      0.000       0.488       1.383
Props_Last_Year            -1.2605      0.145     -8.703      0.000      -1.544      -0.977
Num_TMs                     0.0167      0.015      1.087      0.277      -0.013       0.047
ACD_Conversion           -251.5956     33.310     -7.553      0.000    -316.886    -186.305
Full_Delegate_Sale       -358.4752     13.707    -26.152      0.000    -385.343    -331.608
Inquiry_Conversion       -217.8789     16.965    -12.843      0.000    -251.133    -184.625
PDF_Conversion           -334.9936     15.076    -22.220      0.000    -364.545    -305.442
==============================================================================
Omnibus:                     4184.018   Durbin-Watson:                   0.699
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            15244.757
Skew:                           1.137   Prob(JB):                         0.00
Kurtosis:                       6.894   Cond. No.                     1.02e+04
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.02e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

RMSE went further up as well, to over $600.

There would be one more attempt to improve the linear regression models by looking only at the data of registrants through the marketing (online) channels, this in order to remove the human discounting bias of sales representatives over the phone. The resulting RMSE of $552 showed that this discounting bias might not have been a major issue and unfortunately an improvement on the base linear model did not happen through these adjustments.

Random Forest:

Initial regressions using an out-of-the-box Random Forest gave similar results to the a linear model, with a score of .53 - very similar to the R-squared of the linear model. Hyperparameter tuning ensued using GridSearch. Finally, it seemed that a deeper tree (depth of 15) and more trees as part of the forest (100 estimators), gave a model that had a close fit with the test data:

```
from sklearn import tree
from sklearn.model_selection import GridSearchCV

parameters = {'max_depth':range(3,20), 'n_estimators':[50,100]}
clf = GridSearchCV(RandomForestRegressor(random_state=1), parameters, n_jobs=4)
clf.fit(X=X_train, y=y_train)
tree_model = clf.best_estimator_
print (clf.best_score_, clf.best_params_)
print (tree_model.score(X_test, y_test))
```

```
0.6942490063407305 {'max_depth': 15, 'n_estimators': 100}
0.7045193845165232
```

```
y = orders_merged_delegate_primary['Total_Net_Price']
X = orders_merged_delegate_primary.drop('Total_Net_Price', 1)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=1)
regr = RandomForestRegressor(max_depth=15, n_estimators=100, random_state=1)
regr.fit(X_train, y_train)

y_pred = regr.predict(X_test)
y_true = y_test.values

rms = np.sqrt(mean_squared_error(y_true, y_pred))
rms
```

```
377.47535356025656
```

Indeed, this model also had a lower RMSE than any of the previous linear models or the initial attempts to model using a RandomForest.

Business Case:

Getting attendees to pay a price that is closer to the price paid by others with similar characteristics to them could offer a significant financial opportunity. While charging each person a different price for attendance might not be feasible, making some pricing decision can indeed have a large impact.

Just how large of an impact?

Calculating the RMSE from the average price paid per event instance as compared to the price paid for the ticket results in a price difference of $542 on average. Comparing this to our models RMSE of $377 shows a potential improvement of up to 30% from a pricing optimization that taken into account the models features.

Some immediate business recommendations include:

- Tickets sold offline show strong discounting - incentivizing attendees to purchase passes online can result in an increase of over $150 per ticket holding all else constant.
- There is a significant variation in pricing between events in different industries. For example, some models have shown that HR events charge $360 less than a similar

purchased in a different vertical. It would be important to evaluate the opportunities in this market and test how they would impact the demand.

Future Scope of Work:

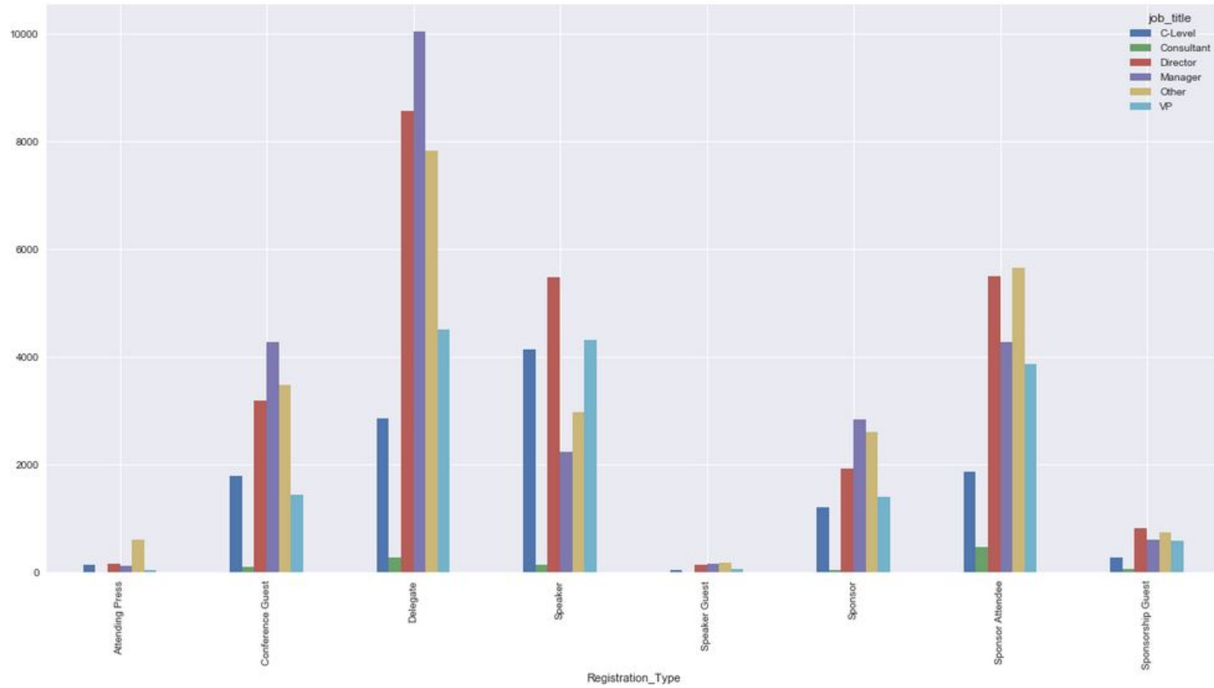
One of the downsides of these models is that they looked at passes purchased without considering the full demand equation and taking into consideration those people that inquires about events and did not purchase. Doing demand analysis using such data would be good next step in identifying what markets might have pricing power and could do with a price increase.

If time and resources would allow, considerations for continued work include:

- Monitoring the prices that are coming through by the utilization of the model. Are they improving the average prices and overall revenue?
- Increased data set that would include inquiries, to get a better idea of the demand model
- More quantitative data to be included in the model from other analytics sources - for example Google Analytics (visits, page conversions, etc…) and company analytics (size, emploress, resources, etc…) as these would contribute to a greater demand and ability to pay.
- The biggest return for such analysis might be on the sponsorship side. A pricing optimization/demand analysis would be fruitful to understand optimal sponsorship behavior on events given company and event characteristics.
    - Some questions to be answered:
        - Do more speakers at an event contribute to greater sponsorships?
        - Does seniority of attendees have a major impact?
        - What type of inquiry behavior contribute to sponsoring an event (downloading brochure, attendee list, etc..)

Appendix A: Attendee types and purchase source

Bar chart of attendee seniority by type of attendee:

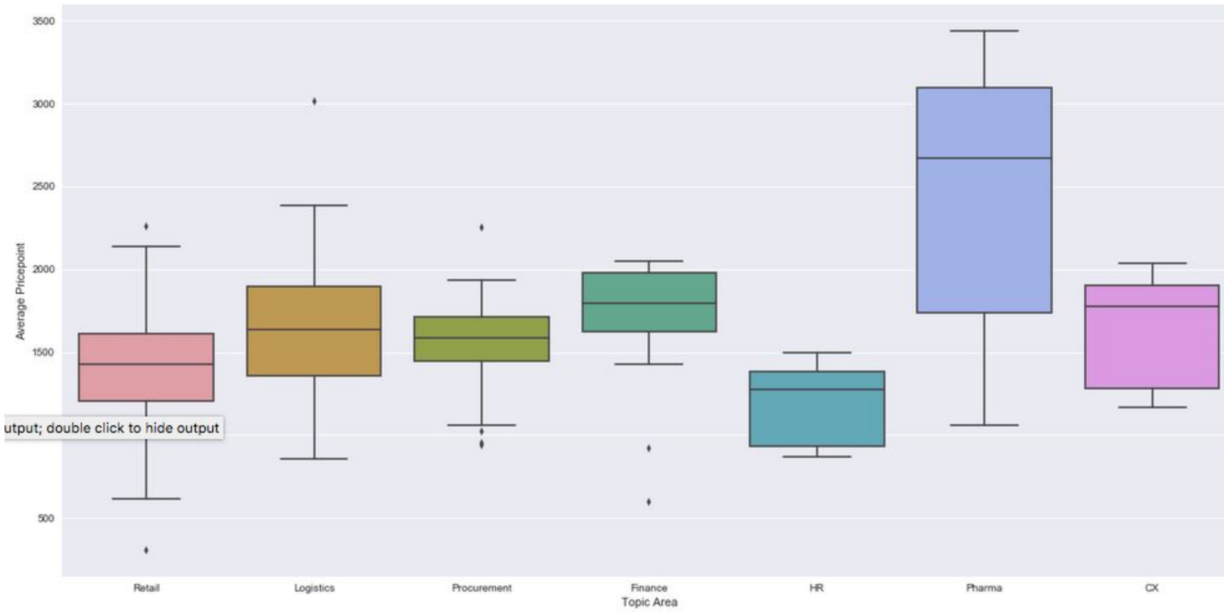


Attendee seniority by source of purchase:

	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: