Cost and Advertisement Data Analysis Project

Isaac Heeseok Joo

Background

- XYZ advertises on TV
- XYZ collected data from customers if they saw XYZ’s ads and where they saw it if they have
- XYZ wants to know which network is most cost efficient and how much it costs to acquire customers

Relevant Terminology

Spend	Amount of money spent to air the ad
Lift	Number of visits attributed to the ad. Calculated by Total traffic - Expected number of natural traffic (Baseline)
Purchase	Number of purchases from the exit survey data
Count	Number of times the ad aired on a network
Cost per Visitor (CPV)	Amount of money spent per visitor (Spend/Lift)
Conversion Rate (CR)	Purchases made per visit attributed to the ad (Purchase/Lift)
Cost per Acquisition (CPA)	Amount of money spent per conversion (Spend/Purchase)

Data Provided

3 Excel Spreadsheets were provided

1. Purchase Exit Survey Data: Survey data filled by customers who visited XYZ website because of the ads and purchased the product
2. Airing Data: Data on which network and when the ads aired, and associated Spend and Lift
3. Lookup Data: Reference table on the names of network and their abbreviations to link Survey data and Airing data

Data Formatting/Data Cleaning

- Survey_data: the Excel file was designed to look better on Excel, but unusable in Python. The data was formatted to make it workable in Python

	Unnamed: 0	Unnamed: 1	Submitted Application Timestamp	Unnamed: 3	Unnamed: 4	Unnamed: 5	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	...	Unnamed: 48	Unnamed: 49	Unnamed: 50
0	NaN	NaN	2017	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
1	NaN	NaN	Q3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
2	NaN	NaN	September	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
3	Source Category	Source	2	3.0	4.0	5.0	6.0	7.0	8.0	10.0	...	21.0	22.0	23.0
4	tv_commercial	(blank)	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
5	NaN	aapka_colors	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	2.0
6	NaN	baby_first	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
7	NaN	bloomberg	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
8	NaN	cbs_sports	NaN	1.0	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
9	NaN	cnbc	NaN	1.0	1.0	NaN	NaN	NaN	2.0	NaN	...	NaN	1.0	NaN

	Source Category	Source	2017-09-02	2017-09-03	2017-09-04	2017-09-05	2017-09-06	2017-09-07	2017-09-08	2017-09-10	...	2017-10-21	2017-10-22	2017-10-23	2017-10-24	2017-10-25	2017-10-26	2017-10-27	2017-10-28	2017-10-29
0	tv_commercial	(blank)	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
1	tv_commercial	aapka_colors	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0
2	tv_commercial	baby_first	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	tv_commercial	bloomberg	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	tv_commercial	cbs_sports	0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	tv_commercial	cnbc	0	1.0	1.0	0.0	0.0	0.0	2.0	0.0	...	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
6	tv_commercial	cnn	0	0.0	3.0	0.0	0.0	3.0	3.0	0.0	...	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	tv_commercial	comedy_central	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
8	tv_commercial	dateline	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0
9	tv_commercial	dish_network	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

- Lookup_data: This reference data had duplicate column and header was not named

	Lookup table for survey response field to airings network ticker symbol	Unnamed: 1	Unnamed: 2
0	Exit Survey	Airings	Exit Survey
1	(blank)	NaN	(blank)
2	aapka_colors	NaN	aapka_colors
3	baby_first	BABY	baby_first
4	bloomberg	BLOM	bloomberg

	Source	Network
1	(blank)	NaN
2	aapka_colors	NaN
3	baby_first	BABY
4	bloomberg	BLOM
5	cbs_sports	CBSS

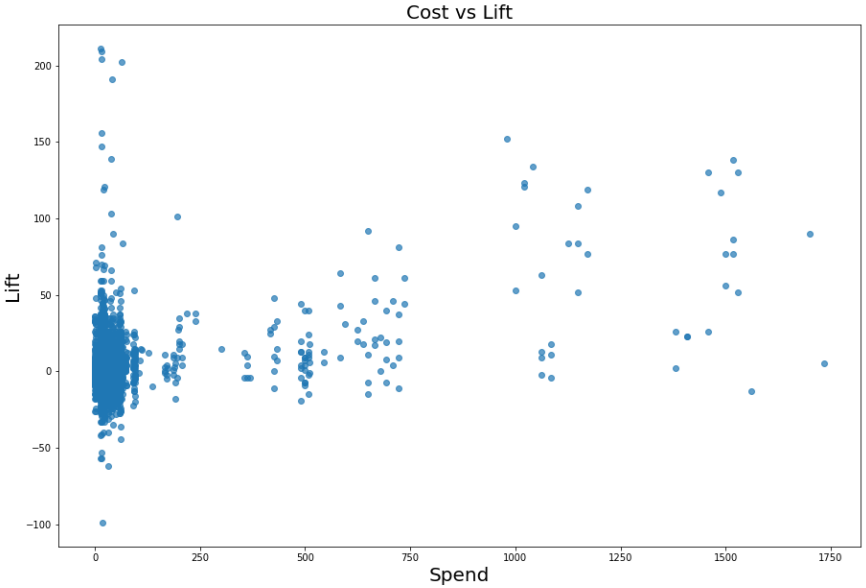
Data Formatting/Data Cleaning

- Airing_data: It was already in a workable format for Python. Airing data did not need formatting, but there are some points to note about this data
  - ZEETV and CNBCWORLD do not seem to label their program
  - MSNB is missing one program name, but it seems to be a simple error
- Program names are not relevant in this case study, but a note for future analysis involving individual programs

	Company	Date/Time ET	Rotation	Creative	Network	Spend	Lift	Program
0	Company XYZ	2017-10-30 22:50:00	MSNB Weekday Prime	ISCICODE0015H	MSNB	980.0	152	THE LAST WORD WITH LAWRENCE O'DONNEL
1	Company XYZ	2017-10-30 22:27:50	HIST Everyday Prime (mirrored)	ISCICODE0015H	HIST	1500.0	77	PAWN STARS
2	Company XYZ	2017-10-30 21:42:20	TWC Everyday Prime	ISCICODE0015H	TWC	300.0	15	WEATHER HACKS
3	Company XYZ	2017-10-30 21:17:22	MSNB Weekday Prime	ISCICODE0015H	MSNB	1020.0	123	THE RACHEL MADDOW SHOW
4	Company XYZ	2017-10-30 20:28:46	MSNB Weekday Prime	ISCICODE0015H	MSNB	1020.0	121	ALL IN WITH CHRIS HAYES
5	Company XYZ	2017-10-30 11:11:12	WILO Cricket	ISCICODE0015H	WILO	0.0	0	CRICKET
6	Company XYZ	2017-10-30 11:10:04	WILO Cricket	ISCICODE0015H	WILO	0.0	-4	CRICKET
7	Company XYZ	2017-10-30 11:08:58	WILO Cricket	ISCICODE0015H	WILO	0.0	9	CRICKET
8	Company XYZ	2017-10-30 11:08:14	WILO Cricket	ISCICODE0015H	WILO	0.0	9	CRICKET
9	Company XYZ	2017-10-30 10:19:56	WILO Cricket	ISCICODE0015H	WILO	0.0	-4	CRICKET

Exploratory Data Analysis

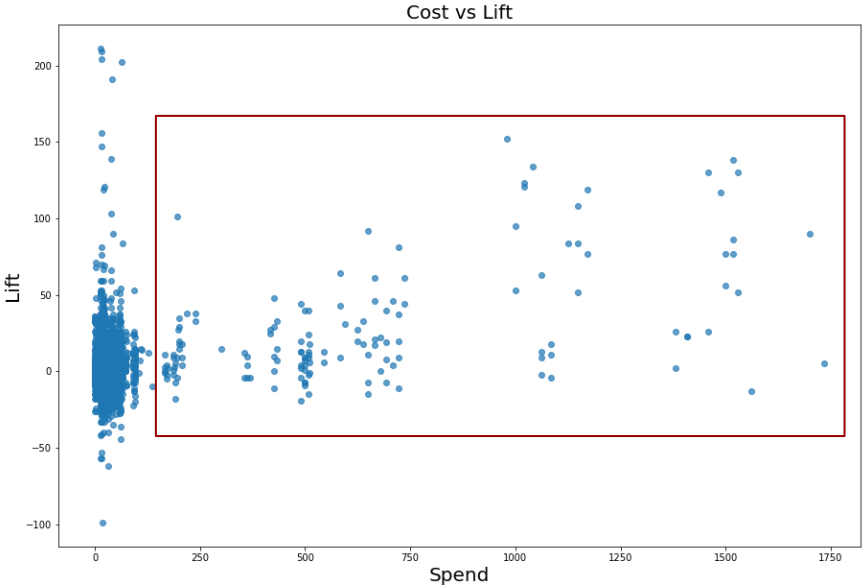
Cost vs Lift



- Spend vs Lift plot to see if there is any correlation
- Large number of points concentrated near the origin
- Huge deviation in lift near (Spend = 0) making data noisy

Exploratory Data Analysis

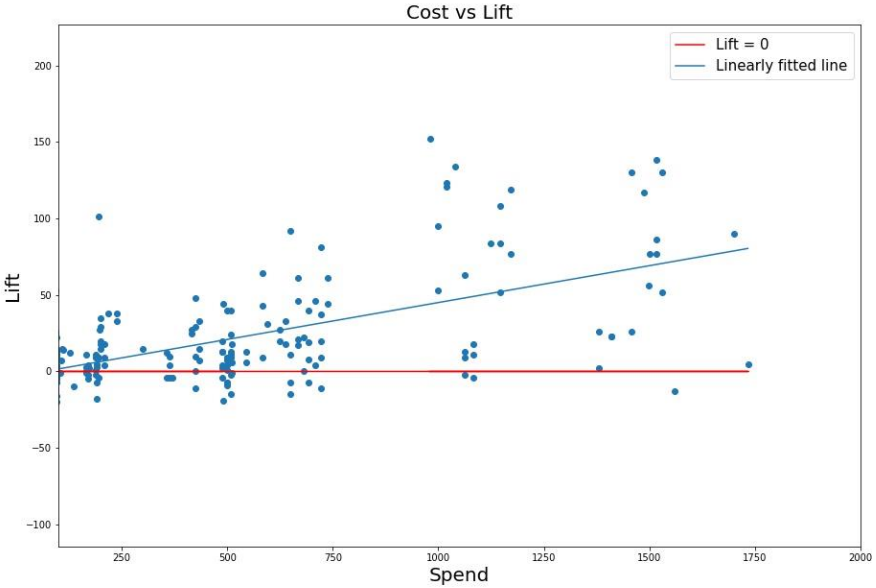
Cost vs Lift



- Spend vs Lift plot to see if there is any correlation
- Large number of points concentrated near the origin
- Huge deviation in lift near (Spend = 0) making data noisy

Exploratory Data Analysis

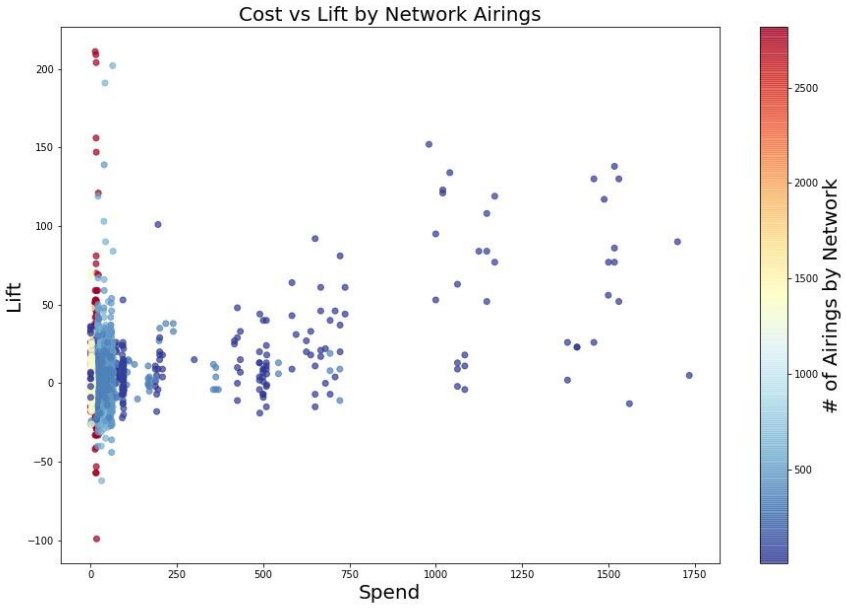
Cost vs Lift Zoomed In



- Zoomed in to look at points (Spend > 100)
- Noise is significantly reduced as Spend increases
- Increasing Spend provides reasonable causality for Lift to increase
- Weak but apparent positive correlation between Spend and Lift
- This is intuitive since more money you spend on ads will most likely bring more people to the website

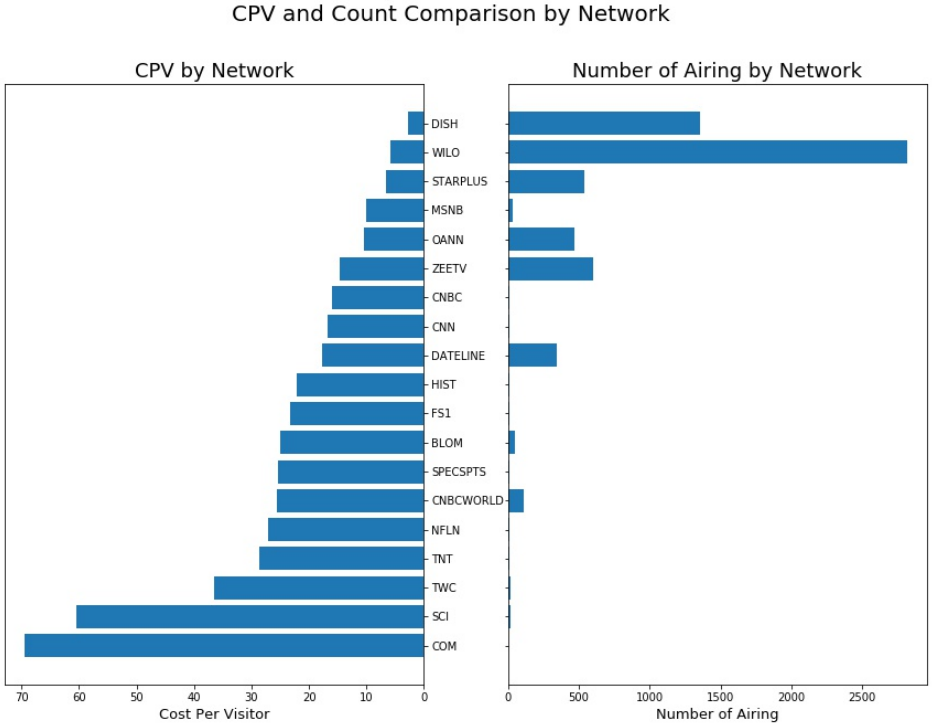
Exploratory Data Analysis

Relationship between Noise and Number of Airing



- Color indicates number of times the ad aired on a network
- Red - More Airing
- Blue - Less Airing
- Increase in number of airing results in more noise in Lift
- Also can see that less money spent correlates to more airing
- This makes sense because low spending leads to high number of airing, which could result in large deviation in lift

Cost Per Visitor (CPV)



- Cost per Visitor (CPV) is amount of money spent per visitor (Spend/Lift)
- More airing naturally leads to more lift
- WILO has large amount of airing and this shear number of airing could skew the metric
- Networks like WILO and DISH seem to return highest efficiency by costing less per visitor, but when you look at the number of these networks aired the ad, it confirms previous statement where less cost -> more airing, thus causing more efficient CPV

CPV of entire dataset
10.80

Data formatting for CR and CPA

- New dataframe was created for convenience since we have to work across two datasets
- Pandas merging function works like JOIN in SQL
- Merged Survey_data and Airing_data on ‘Source’ and ‘Network’ by using reference from Lookup_data
- The data where networks have Airing data were labeled as ‘Real’, and the data that includes networks that have Purchase data but not the Airing data were labeled as ‘Estimate’
- Missing Data: Used average Conversion Rate and Cost Per Acquisition from Real data to obtain missing ‘Spend’ and ‘Lift’ values for networks in Estimate data instead of using average of total spend or lift from Airing data
- Average CR and CPA calculated from Real data were used to get reasonable spend and lift correlating to the purchase number of the networks
- In this report, we only deal with existing Real data, but estimated data is available in the Jupyter Notebook

Data formatting for CR and CPA

Metric Function

- Created a function called ‘metric(df)’ that calculates and appends Cost per Visitor, Conversion Rate, and Cost per Acquisition after it went through cleaning and formatting process
- Average of Spend, Lift, CPV, CR, and CPA are calculated by dividing number of times the ad aired on a network (ex. Average Spend = Average amount of money spent per airing of ad)

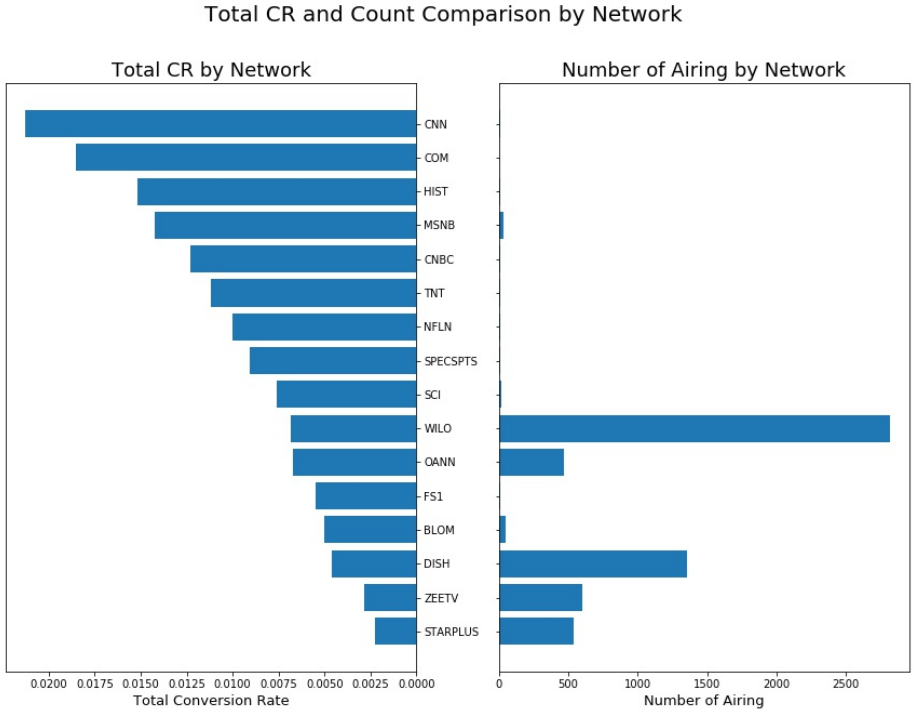
Data formatting for CR and CPA

- Formatting produces following multiindex dataframe with ‘Source’, ‘Network’, ‘Date/Time ET’ as indices

Source	Network	Date/Time ET	Purchase	Spend	Lift	Count	Avg Spend	Avg Lift	Total CPV	Avg CPV	Total CR	Avg CR	Total CPA	Avg CPA
bloomberg	BLOM	2017-09-30	1.0	4966.72	199	53	93.711698	3.754717	24.958392	0.470913	0.005025	0.000095	4966.720000	93.711698
cnbc	CNBC	2017-09-30	5.0	6481.42	406	14	462.958571	29.000000	15.964089	1.140292	0.012315	0.000880	1296.284000	92.591714
cnn	CNN	2017-09-30	13.0	9159.60	507	7	1308.514286	72.428571	18.066272	2.580896	0.025641	0.003663	704.584615	100.654945
		2017-10-31	10.0	8954.75	574	6	1492.458333	95.666667	15.600610	2.600102	0.017422	0.002904	895.475000	149.245833
comedy_central	COM	2017-10-31	2.0	7501.25	108	7	1071.607143	15.428571	69.456019	9.922288	0.018519	0.002646	3750.625000	535.803571
dish_network	DISH	2017-09-30	4.0	2513.09	976	1150	2.185296	0.848696	2.574887	0.002239	0.004098	0.000004	628.272500	0.546324

Conversion Rate

CR vs Number of Airing

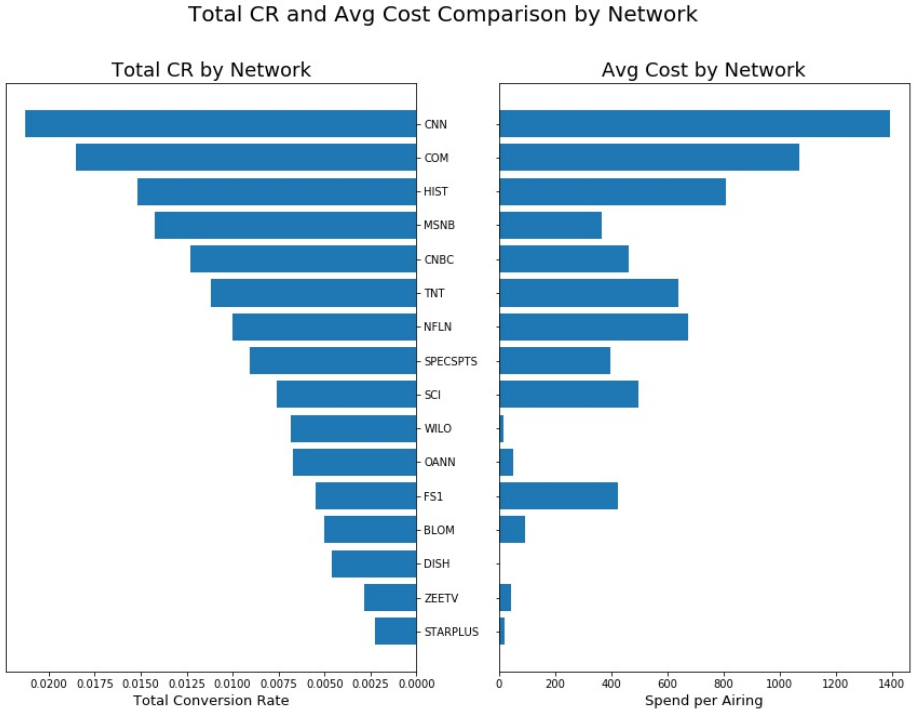


- Conversion Rate (CR) is the Purchase made per visitor by the ads (Purchase/Lift)
- CR does not seem to correlate with the amount of airing
- CNN, COM, and HIST have shown greatest CR with few airings
- WILO, DISH seem to be inefficient compared to the number of airings

CR of entire dataset
0.007540

Conversion Rate

CR vs Average Cost

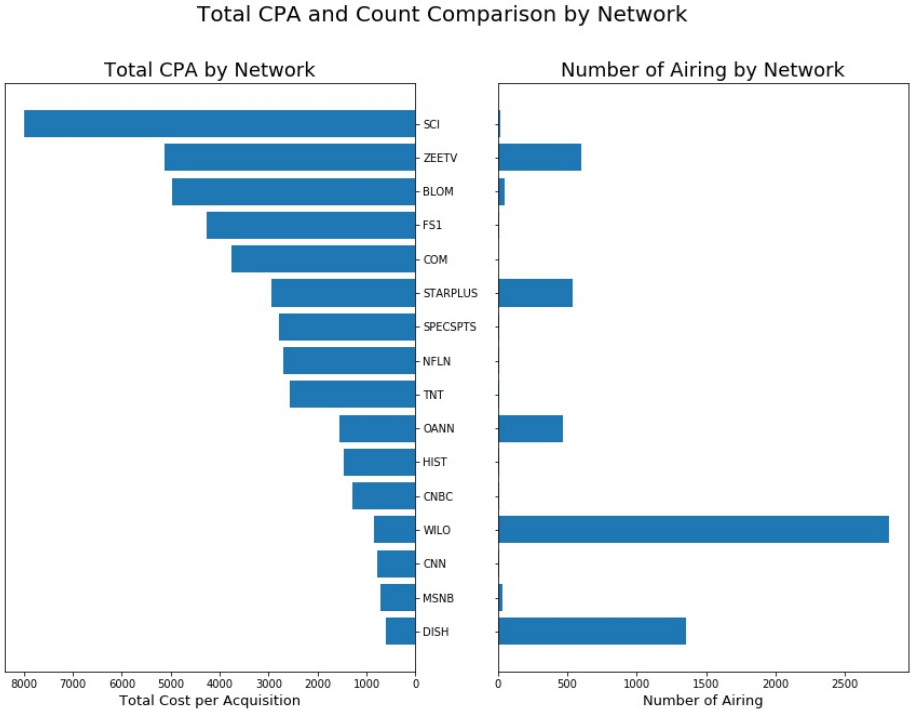


- Here we compare Total Conversion Rate with the amount of money spent per airing
- There is a definite correlation between CR and Average Spend
- More money spent on ads probably means the network is watched by many, thus higher price
- Network with more audience also means higher likelihood of purchases from the ads
- In this case, MSNB and WILO yields good efficiency based on the money spent

CR of entire dataset
0.007540

Cost per Acquisition

CPA vs Number of Airing

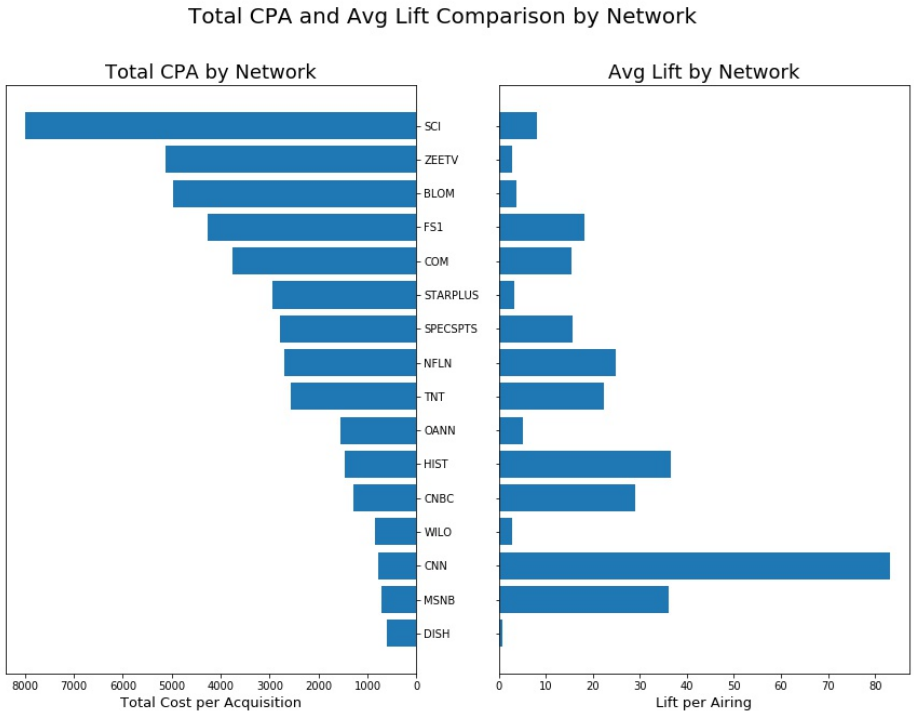


- Cost per Acquisition (CPA) is amount of money spent on each conversion (Spend/Purchase)
- CPA also does not seem to correlate with amount of airing
- CNN and MSNB seem to give great CPA with low number of airing
- SCI and FS1 cost a lot per acquisition, and does not give great CR

CPA of entire dataset
1350.21

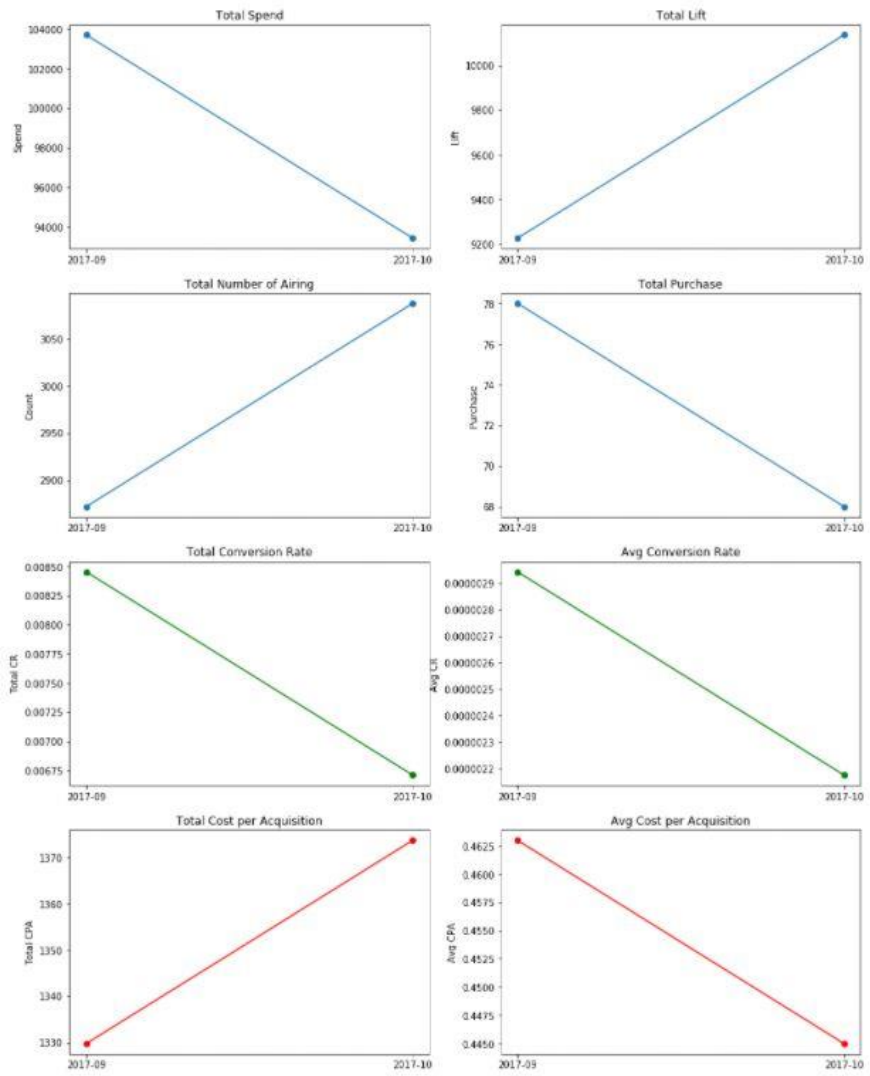
Cost per Acquisition

CPA vs Average Lift



- Here we compare CPA with Lift per airing
- There seems to be better correlation
- More Lift means more visitors to the website who saw the ads
- More visitors naturally results in more purchases
- Network with large audience, like CNN and MSNB causes more Lift with fewer spending and airing

CPA of entire dataset
1350.21



Data by Monthly

- Metrics are plotted along time in monthly intervals
- Blue - Fundamental Metrics
- Green - Conversion Rate
- Red - Cost per Acquisition
- Since the data was collected only over 2 months, there is not much insight we could draw
- This data would be useful if data is continuously collected over several years to give sense of trend

Recommendation

Analysis Results

- If it does not cost to air, then always air since more number of airing leads to more lift
  - Although they have weak CR, WILO and DISH usually do not cost much but yields high lift by the shear number of airings, thus providing low CPA
- Target mainstream network with large audience
  - MSNB provides great CR and CPA with relatively low cost per airing and low number of airing
  - Although average cost per airing of CNN and HIST are relatively high, they provide best CR and some of the lowest CPA. CNN provides highest average lift per airing and highest CR
- Avoid high cost, low conversion network
  - Although COM has high CR, it has the highest CPV and relatively high CPA
  - SCI and FS1 have two of the highest CPV and CPA, but CR is mediocre
  - ZEETV has high CPA but low CR

Invest More	Keep Airing	Invest Less
CNN, MSNB, CNBC, HIST	WILO, DISH	COM, SCI, ZEETV, FS1

Recommendation

Survey Questions

- First question does not seem necessary if we are only focusing on TV ads
  - It is okay if planning on advertising on other platform
  - Make it true/false question to know if the customer watched the ads on TV or not
- Second question could be conditional based on the answer to the first question
  - If first question was true, second question is a mandatory question to prevent blank data
- XYZ should have access to the network where their ads were aired
  - Limit the choices to network they have aired for uniform data across Survey data and Airing data
- Let customer indicate when they saw the ads
  - This provides more accurate link between Survey data and Airing data