Rutgers Business School
Newark and New Brunswick

Data-Driven Attribution Modeling using Statistical Analysis

Capstone Project Final Report

Kalp Patel
Ayushi Mundhe

Supervised by:

Dr. Madhavi Chakrabarty

Fall 2019

This Capstone Project was approved by:

Advisor: _____
Dr. Madhavi Chakrabarty

Assistant Dean: _____
Goncalo Filipe

A Project Submitted in Partial Fulfillment
of the Requirements of the
Master of Information Technology and Analytics

ACKNOWLEDGEMENTS

Firstly, I wish to express my deepest gratitude to my supervisor, Dr. Madhavi Chakrabarty, for her thought out guidance and constant support.

I place on record as well, my sincerest thanks to Ayushi Mundhe/ Kalp Patel who gave a considerable amount of his time helping in this project.

My greatest thanks go to Google, for supplying us with testing data of the Google Merchandise Store.

I take this opportunity to express gratitude to all of Rutgers University faculty members and students that helped me throughout the different stages of my curriculum, for their unconditional support and attention.

TABLE OF CONTENTS

Chapter No.	Title	Page No.
1	Introduction	7
1.1	Definition	7
1.2	Types of Attribution Models	7
2	Background	11
3	Project Description	13
3.1	Understanding the requirement of sample size	13
3.2	Data Collection	14
3.3	Hypothesis Testing ANOVA	17
3.3.1	Implementation using MS Excel	17
3.4	Tukey HSD Post-Hoc Analysis	21
3.4.1	Estimating Differences of Means	22
4	Conclusion	25

LIST OF FIGURES

Figure No.	Name	Page No.
Figure 1.2.1	Last Interaction Attribution	7
Figure 1.2.2	First Interaction Attribution	8
Figure 1.2.3	Last Non-Direct Click Attribution	8
Figure 1.2.4	Linear Attribution	9
Figure 1.2.5	Time Decay Attribution	9
Figure 1.2.6	Position Based Attribution	10
Figure 2.1	An Illustration of Multi-Touch Attribution Problem	11
Figure 3.1.1	Sample Size for Data Points	13
Figure 3.2.1	Google Analytics Model Comparison Tool_1	14
Figure 3.2.2	Google Analytics Model Comparison Tool_2	15
Figure 3.2.3	Dataset 1: All Attribution Models & Channel - Referral	15
Figure 3.2.4	Dataset 1: All Attribution Models & Channel - Paid Search	16
Figure 3.2.5	Dataset 1: All Attribution Models & Channel - Social Network	16
Figure 3.3.1.1	Selection of Data Analysis from MS Excel Toolbar	17
Figure 3.3.1.2	Data Analysis Tool from MS Excel	18
Figure 3.3.1.3	Anova:Single Factor Data Selection	18
Figure 3.3.1.4	Structure of ANOVA: Single Factor	18
Figure 3.3.1.5	ANOVA: Single Factor (Dataset 1)	19
Figure 3.3.1.6	ANOVA: Single Factor (Dataset 2)	20
Figure 3.3.1.7	ANOVA: Single Factor (Dataset 3)	20
Figure 3.4.1.1	Critical Q Calculation	22
Figure 3.4.1.2	Output of Tukey HSD for Dataset 1	23

ABSTRACT

Most paths to purchase are not a straight line. There are lots of ways for customers to discover your brand, engage with it, and move further down the sales funnel. That's why attribution models have become a necessary tool for marketers looking for data to improve their campaigns.

Attribution modeling is a strategy that allows marketers to analyze and assign credit to marketing touchpoints that occur at the specific steps of the customer journey, from searching for a product online to making a purchase, and every action in between. Using attribution models helps marketers better understand which parts of their marketing efforts are driving the most leads to that part of the sales funnel.

In this project, we try to understand if there is a significant difference in Data of Google Merchandising Store using the Model Comparison Tool of Google Analytics. Here we implement Hypothesis Testing - ANOVA by comparing various models across a single channel. If there is a presence of Significant Difference after ANOVA implementation we then perform Tukey HSD(Honestly Significant Difference) Test to understand exactly which means are different i.e. because of which Attribution Models it delivers the Significant Difference. The result of our analysis would be to deliver if there is any presence of Statistical Significance in Google Merchandising Store Attribution Modelling Data.

Chapter One
INTRODUCTION

1.1 Definition

An attribution model is a rule or set of rules, that determines how credit for sales and conversions is assigned to touchpoints in conversion paths. The Multi-Channel Funnels Model Comparison Tool is used to compare how different attribution models impact the valuation of marketing channels. For example, the Last Interaction model in Analytics assigns 100% credit to the final touchpoints (i.e., clicks) that immediately precede sales or conversions. In contrast, the First Interaction model assigns 100% credit to touchpoints that initiate conversion paths.

A Multi-Channel Funnels Model Comparison Tool allows you to analyze how each model distributes the value of a conversion, but here in this project, we will try to find how significant is the data present here. There are six common attribution models: First Interaction, Last Interaction, Last Non-Direct Click, Linear, Time-Decay, and Position-Based.

1.2 Types of Attribution Models

Here the different types of attribution Models:

1. **Last Interaction Attribution:**

   Last Interaction Attribution is also referred to as "last-click" or "last-touch." As the name implies, this model gives 100% of the credit to the last interaction your business had with a lead before they convert.

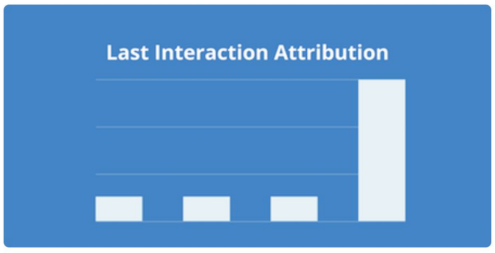


Figure 1.2.1 - Last Interaction Attribution

2. **First Interaction Attribution:**

First Interaction is similar to Last Interaction, in that it gives 100% of the credit to one-click/interaction. First Interaction (also called "First-Click") gives all of the credit for a conversion to your business' first interaction with the customer.

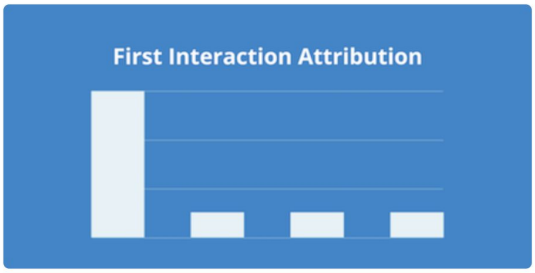


Figure 1.2.2 - First Interaction Attribution

3. **Last Non-Direct Click:**

The Last Non-Direct Click Model is a bit more helpful than a standard last-click model. 100% of the value is still assigned to a single interaction. But, with the last non-direct click, it eliminates any "direct" interactions that occur right before the conversion.

Figure 1.2.3 - Last Non-Direct Click Attribution

4. **Linear Attribution:**

With a Linear attribution model, you split credit for a conversion equally between all the interactions the customer had with your business.

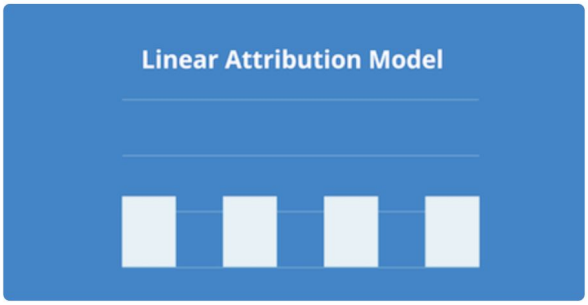


Figure 1.2.4 - Linear Attribution

5. **Time Decay Attribution:**

Time Decay attribution is similar to Linear attribution - it spreads out the value across multiple events. But unlike, Linear attribution, the Time Decay model also takes into consideration when the touchpoint occurred. Interactions that occur closer to the time of purchase have more value attributed to them. The first interaction gets less credit, while the last interaction will get the most.

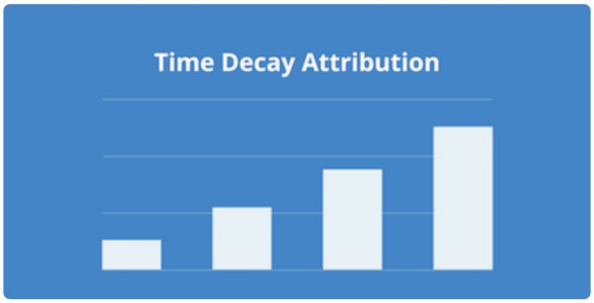


Figure 1.2.5 - Time Decay Attribution

6. **Position Based Attribution:**

The Position Based attribution model (also called U-shaped attribution) splits the credit for a sale between a prospect's first interaction with your brand and the moment they convert to a lead. 40% of the credit is given to each of these points, with the remaining 20% spread out between any other interactions that happened in the middle.

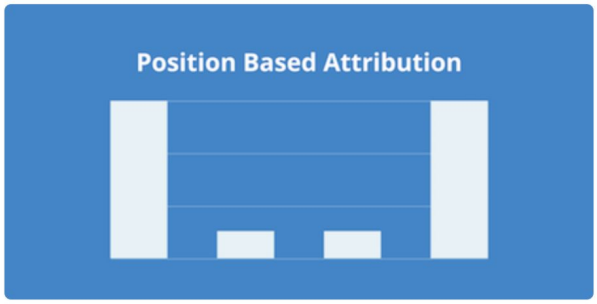


Figure 1.2.6 - Position Based Attribution

Chapter Two
BACKGROUND

Digital advertising started 16 years ago as a new media where traditional print ads can appear. When the internet continues to grow with an exploding rate, advertising industry embraced digital advertising and has made it a $40 Billion a year mega industry in the US alone. Advertising campaigns are often launched across multiple channels. Traditional advertising channels include outdoor billboards, TV, radio, newspapers and magazines, and direct mailing.

Digital advertising channels include search, online display, social, video, mobile and email. In this article, we focus on the digital advertising channels. Typically multiple advertising channels have delivered advertisement impressions to a user. When the user then makes a purchase decision or signs up to a service being advertised, the advertiser wants to determine which ads have contributed to the user's decision. This step is critical in completing the feedback loop so that one can analyze, report and optimize an advertising campaign. This problem of interpreting the influence of advertisements to the user's decision process is called the attribution problem.

The goal of attribute modeling is to pin-point the credit assignment of each positive user to one or more advertising touch point, which is illustrated in Figure 2.1.

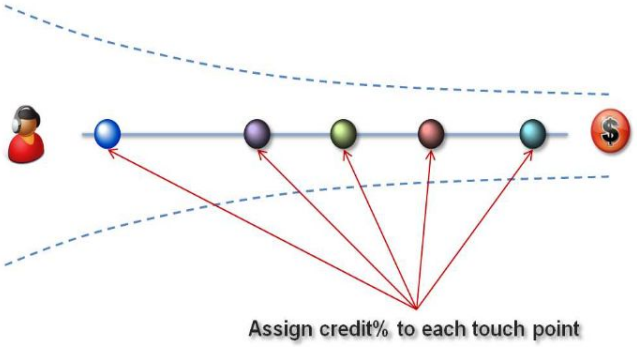


Figure 2.1 - An Illustration of Multi-Touch Attribution Problem

To determine which media channel or which ad is to be credited, initially a simple rule was developed and quickly adopted by the online advertising industry: The last ad the user clicked on before he made the purchase or sign up decision, or say, conversion, gets 100% of the credit. This "last- click win" model was extended to include "last-view win" if none of the ads were clicked within a reasonable time window before user conversion. We call both these two models "last-touch attribution" (LTA), where "touch" or touch point is defined to be any ad impression, click or advertising related interaction the user has experienced from the advertiser. The last-touch attribution model is simple. However, it completely ignores the influences of all ad impressions except the last one.

Alternatively, the concept of multi-touch attribution (MTA) model has been recently proposed, where more than one touch point can each have a fraction of the credit based on the true influence each touch point has on the outcome, i.e., user's conversion decision.

Chapter Three
PROJECT DESCRIPTION

In order to understand the presence or absence of Significant Difference in the Data of Google Merchandising Store taken from Model Comparison Tool of Google Analytics, we performed Statistical Analysis Methods such as ANOVA & Tukey HSD.

3.1 Understanding the requirement of sample size

Sample size is always determined to detect some hypothetical difference. It takes huge samples to detect tiny differences but tiny samples to detect huge differences, so you have to specify the size of the effect you are trying to detect.

With reference to the Figure 3.1, table abridged from 9-26 of by Bausell and Li

ES	N	Total N
0.3	87	348
0.5	39	156
0.7	17	68
0.8	13	52
1.0	9	36
1.5	5	20
2.0	4	16
3.0	3	12

Figure 3.1.1 - Sample Size for Data Points

We assume an effect size (ES) = 0.5, resulting in atleast 39 data points per condition to get an appropriate result of ANOVA.

Model → Channel ↓	Last Interaction	Time Decay	Linear	First Interaction	Position Based	Last Non-Direct Click
Referral	39	39	39	39	39	39

Social Network	39	39	39	39	39	39
Paid	39	39	39	39	39	39

So based on the above table, we considered taking monthly data for 4 years which delivers a sample size of 48 data points per condition.

3.2 Data Collection

Exported data from Google Analytics for the last 4 years based on a monthly basis i.e. from October 2015 to September 2019 month wise for all the attribution models across 3 channel individuals. Various attribution models are Last Interaction, Time Decay, Linear, First Interaction, Position-Based & Last Non-Direct Click.

For e.g. Figure 3.2.1 & Figure 3.2.2 represents data from Google Analytics Model Comparison Tool page for the month of September 2019 consisting all the models & channels.

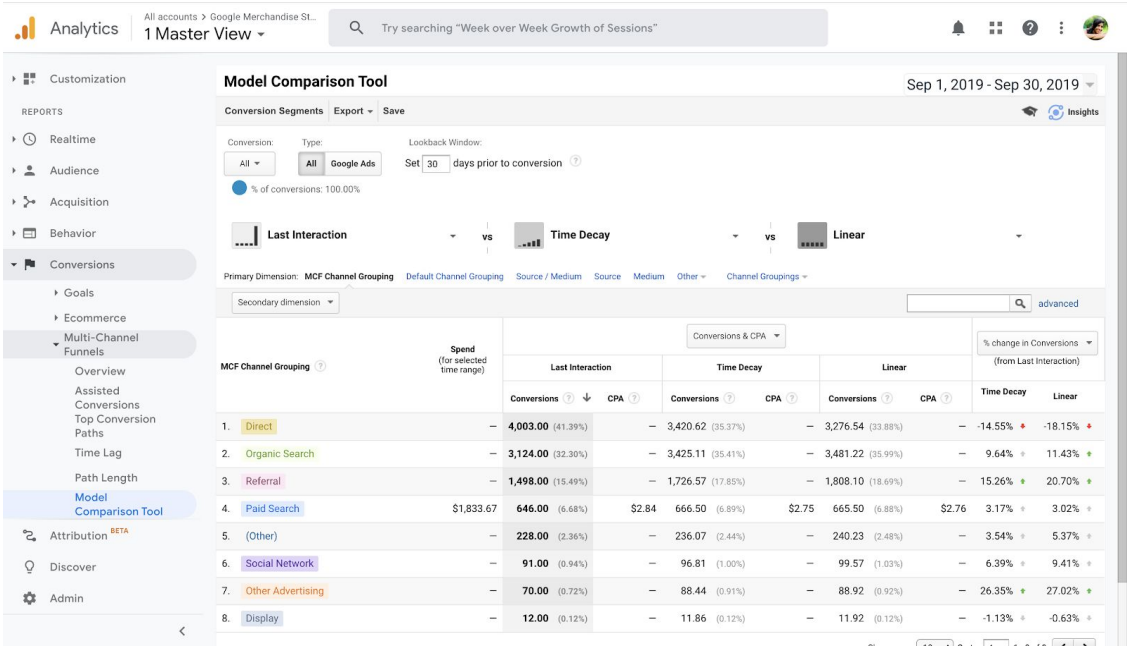


Figure 3.2.1 - Google Analytics Model Comparison Tool_1

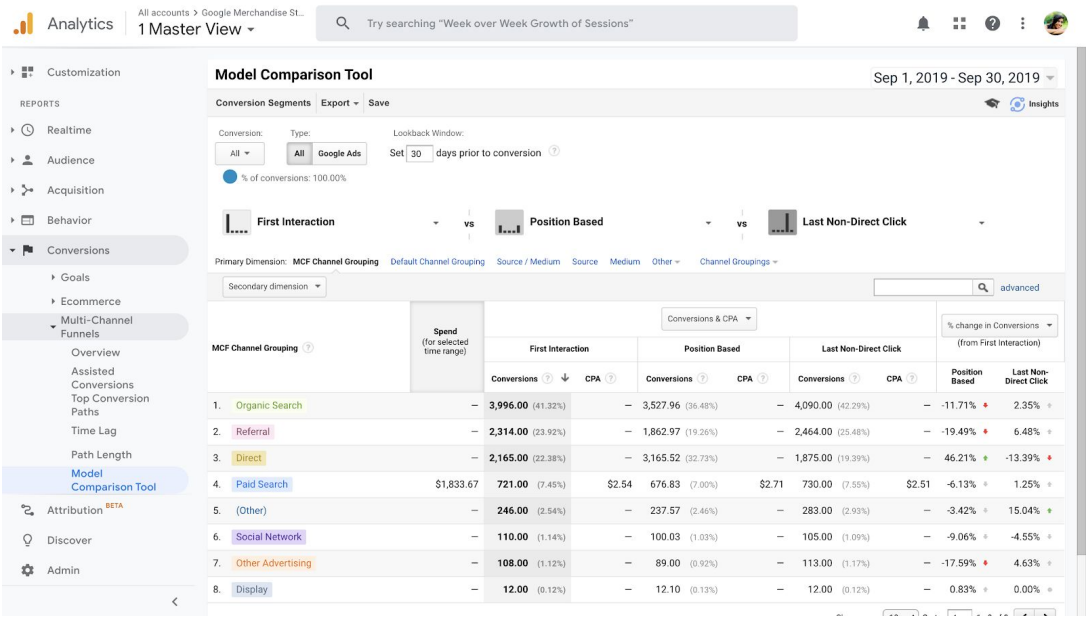


Figure 3.2.2 - Google Analytics Model Comparison Tool_2

The data above was exported and stored in Spreadsheet for further implementation.

Figure 3.2.3 → **Dataset 1**: All Attribution Models against Single Channel - Referral.

Figure 3.2.4 → **Dataset 2**: All Attribution Models against Single Channel - Paid Search.

Figure 3.2.5 → **Dataset 3**:All Attribution Models against Single Channel - Social Network.

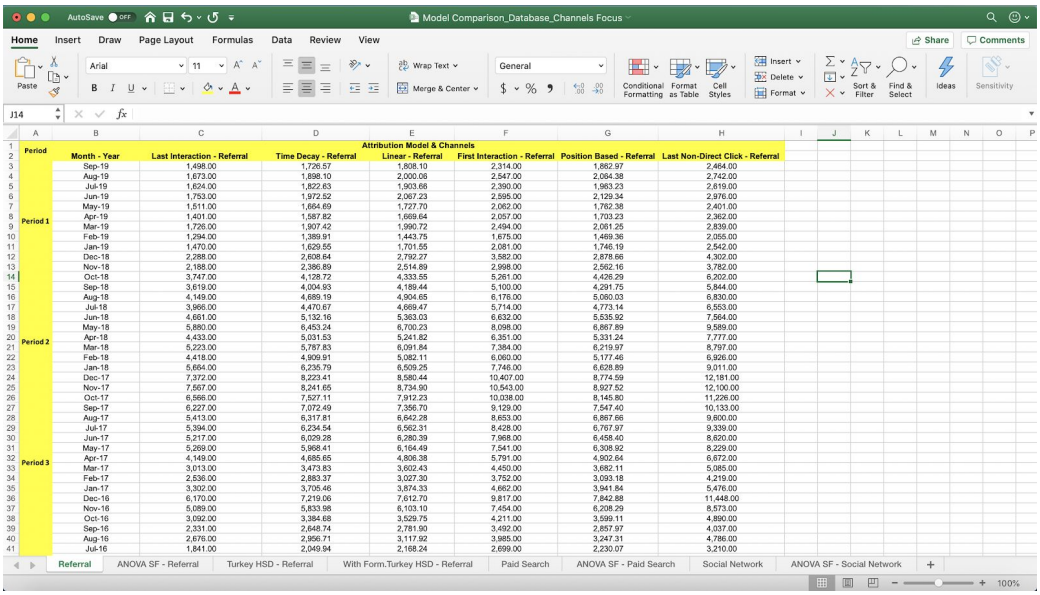
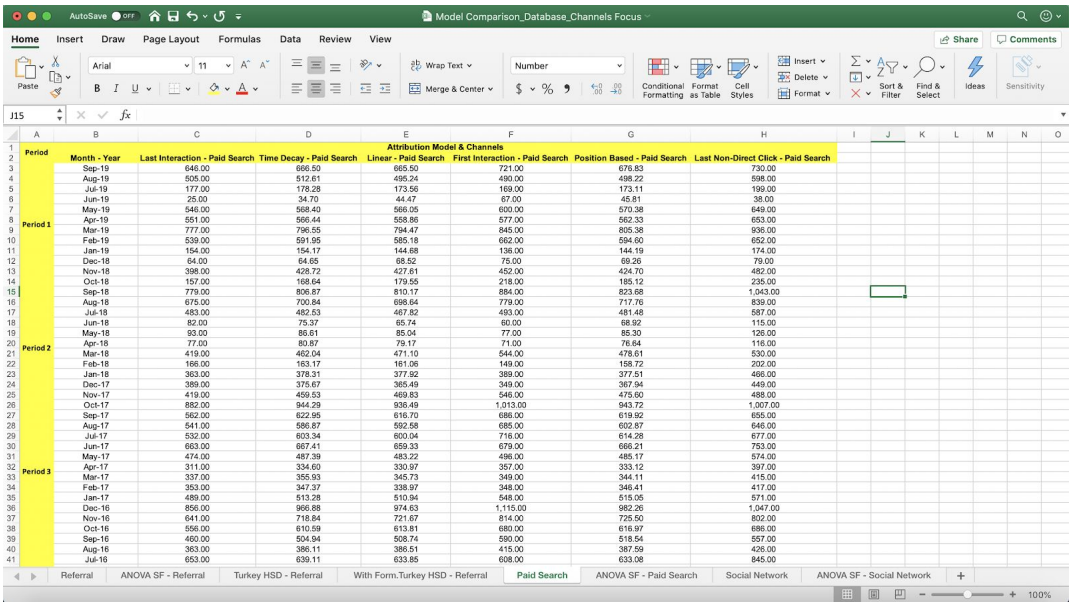
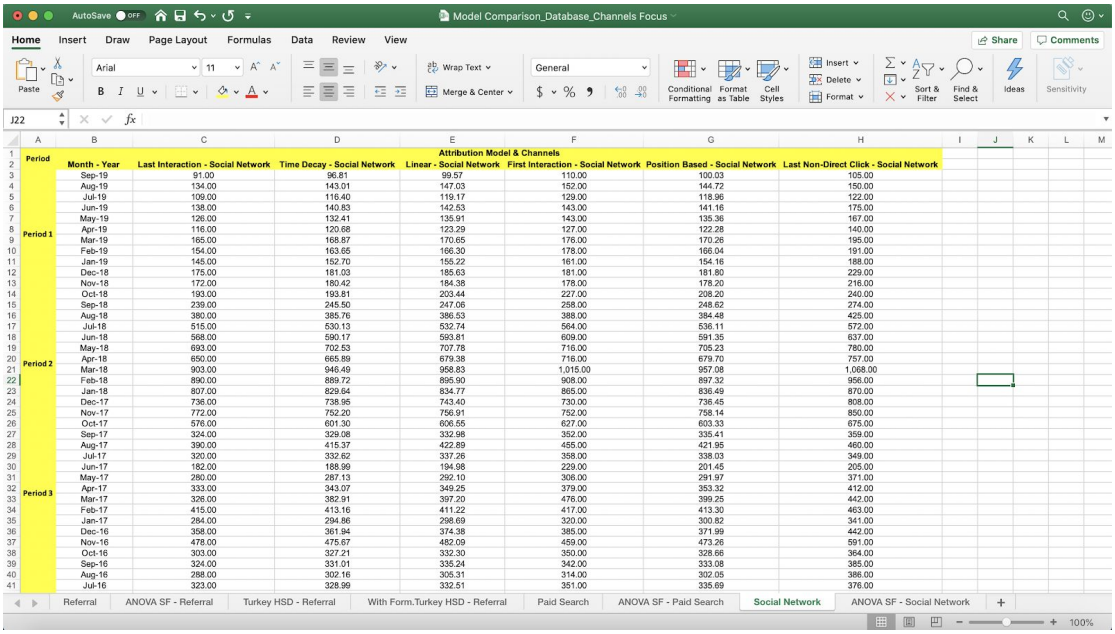


Figure 3.2.3 - Dataset 1: All Attribution Models & Channel - Referral



Period	Month - Year	Last Interaction - Paid Search	Time Decay - Paid Search	Linear - Paid Search	First Interaction - Paid Search	Position Based - Paid Search	Last Non-Direct Click - Paid Search
		Attribution Model & Channels					
	Sep-19	646.00	666.50	665.50	721.00	676.83	730.00
	Aug-19	505.00	512.61	495.24	490.00	498.22	598.00
	Jul-19	177.00	178.28	173.56	169.00	173.11	199.00
	Jun-19	25.00	34.70	44.47	67.00	45.81	38.00
	May-19	546.00	568.40	566.05	600.00	570.38	649.00
Period 1	Apr-19	551.00	566.44	558.86	577.00	562.33	653.00
	Mar-19	777.00	796.55	794.47	845.00	805.38	936.00
	Feb-19	539.00	591.95	585.18	662.00	594.60	652.00
	Jan-19	154.00	154.17	144.68	136.00	144.19	174.00
	Dec-18	64.00	64.65	68.52	75.00	69.26	79.00
	Nov-18	398.00	428.72	427.61	452.00	424.70	482.00
	Oct-18	157.00	168.64	179.55	218.00	185.12	235.00
	Sep-18	779.00	806.87	810.17	884.00	823.68	1,043.00
	Aug-18	675.00	700.84	698.64	779.00	717.76	839.00
	Jul-18	483.00	482.53	467.82	493.00	481.48	587.00
	Jun-18	82.00	75.37	65.74	60.00	68.92	115.00
	May-18	93.00	86.61	85.04	77.00	85.30	126.00
Period 2	Apr-18	77.00	80.87	79.17	71.00	76.64	116.00
	Mar-18	419.00	462.04	471.10	544.00	478.61	530.00
	Feb-18	166.00	163.17	161.06	149.00	158.72	202.00
	Jan-18	363.00	378.31	377.92	389.00	377.51	466.00
	Dec-17	389.00	375.67	365.49	349.00	367.94	449.00
	Nov-17	419.00	459.53	469.83	546.00	475.60	488.00
	Oct-17	882.00	944.29	936.49	1,013.00	943.72	1,007.00
	Sep-17	562.00	622.95	616.70	686.00	619.92	655.00
	Aug-17	541.00	586.87	592.58	685.00	602.87	646.00
	Jul-17	532.00	603.34	600.04	716.00	614.28	677.00
	Jun-17	663.00	667.41	659.33	679.00	666.21	753.00
	May-17	474.00	487.39	483.22	496.00	485.17	574.00
Period 3	Apr-17	311.00	334.60	330.97	357.00	333.12	397.00
	Mar-17	337.00	355.93	345.73	349.00	344.11	415.00
	Feb-17	353.00	347.37	338.97	348.00	346.41	417.00
	Jan-17	489.00	513.28	510.94	548.00	515.05	571.00
	Dec-16	856.00	966.88	974.63	1,115.00	982.26	1,047.00
	Nov-16	641.00	718.84	721.67	814.00	725.50	802.00
	Oct-16	556.00	610.59	613.81	680.00	616.97	686.00
	Sep-16	460.00	504.94	508.74	590.00	518.54	557.00
	Aug-16	363.00	386.11	386.51	415.00	387.59	426.00
	Jul-16	653.00	639.11	633.85	608.00	633.08	845.00

Figure 3.2.4 - Dataset 2: All Attribution Models for Channel - Paid Search



Period	Month - Year	Last Interaction - Social Network	Time Decay - Social Network	Linear - Social Network	First Interaction - Social Network	Position Based - Social Network	Last Non-Direct Click - Social Network
		Attribution Model & Channels					
	Sep-19	91.00	96.81	99.57	110.00	100.03	105.00
	Aug-19	134.00	143.01	147.03	152.00	144.72	150.00
	Jul-19	109.00	116.40	119.17	129.00	118.96	122.00
	Jun-19	138.00	140.83	142.53	143.00	141.16	175.00
	May-19	126.00	132.41	135.91	143.00	135.36	167.00
Period 1	Apr-19	116.00	120.68	123.29	127.00	122.28	140.00
	Mar-19	165.00	168.87	170.65	176.00	170.26	195.00
	Feb-19	154.00	163.65	166.30	178.00	166.04	191.00
	Jan-19	145.00	152.70	155.22	161.00	154.16	188.00
	Dec-18	175.00	181.03	185.63	181.00	181.80	229.00
	Nov-18	172.00	180.42	184.38	178.00	178.20	216.00
	Oct-18	193.00	193.81	203.44	227.00	208.20	240.00
	Sep-18	239.00	245.50	247.06	258.00	248.62	274.00
	Aug-18	380.00	385.76	386.53	388.00	384.48	425.00
	Jul-18	515.00	530.13	532.74	564.00	536.11	572.00
	Jun-18	568.00	590.17	593.81	609.00	591.35	637.00
	May-18	693.00	702.53	707.78	716.00	705.23	780.00
Period 2	Apr-18	650.00	665.89	679.38	716.00	679.70	757.00
	Mar-18	903.00	946.49	958.83	1,015.00	957.08	1,068.00
	Feb-18	890.00	889.72	895.90	908.00	897.32	956.00
	Jan-18	807.00	829.64	834.77	865.00	836.49	870.00
	Dec-17	736.00	738.95	743.40	730.00	736.45	808.00
	Nov-17	772.00	752.20	756.91	752.00	758.14	850.00
	Oct-17	576.00	601.30	606.55	627.00	603.33	675.00
	Sep-17	324.00	329.08	332.98	352.00	335.41	359.00
	Aug-17	390.00	415.37	422.89	455.00	421.95	460.00
	Jul-17	320.00	332.62	337.26	358.00	338.03	349.00
	Jun-17	182.00	188.99	194.98	229.00	201.45	205.00
	May-17	280.00	287.13	292.10	308.00	291.97	371.00
Period 3	Apr-17	333.00	343.07	349.25	379.00	353.32	412.00
	Mar-17	326.00	382.91	397.20	476.00	399.25	442.00
	Feb-17	415.00	413.16	411.22	417.00	413.30	463.00
	Jan-17	284.00	294.86	298.69	320.00	300.82	341.00
	Dec-16	358.00	361.94	374.38	385.00	371.99	442.00
	Nov-16	478.00	475.67	482.09	459.00	473.26	591.00
	Oct-16	303.00	327.21	332.30	350.00	328.66	364.00
	Sep-16	324.00	331.01	335.24	342.00	333.08	385.00
	Aug-16	288.00	302.16	305.31	314.00	302.05	386.00
	Jul-16	323.00	328.99	332.51	351.00	335.69	376.00

Figure 3.2.5 - Dataset 3: All Attribution Models for Channel - Social Network

3.3 Hypothesis Testing - ANOVA

"The one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups."

The one-way ANOVA compares the means between the groups you are interested in and determines whether any of those means are statistically significantly different from each other. Specifically, it tests the null hypothesis:

$$Ho: \mu 1 = \mu 2 = \mu 3 = \ldots\ldots\ldots = \mu k$$

where μ = group mean and k = number of groups.

If, however, the one-way ANOVA returns a statistically significant result, we accept the alternative hypothesis (HA), which is that there are at least two group means that are statistically significantly different from each other.

In order to understand if returned out shows the significant difference or not, following conditions are followed:

1) If p-value < alpha (Accept) and consider f-value result:
    a) F-value < F-critical (No significant difference i.e. Ho Accept)
    b) F-value > F-critical (Significant difference i.e. Ho Reject & H1 Accept)
2) If p-value > alpha (Reject), No Significant difference and discard f-value result.

3.3.1 Implementation using MS Excel:

Dataset 1: All Attribution Models for Channel - Referral

Step A:

Data → Data Analysis → ANOVA Single Factor

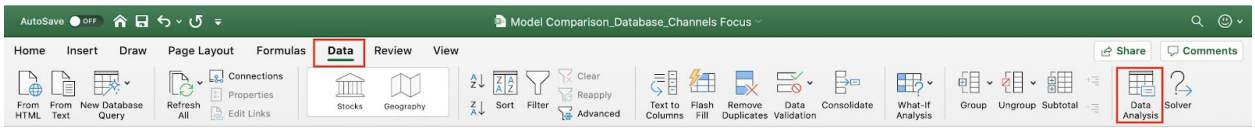


Figure 3.3.1.1 - Selection of Data Analysis from MS Excel Toolbar

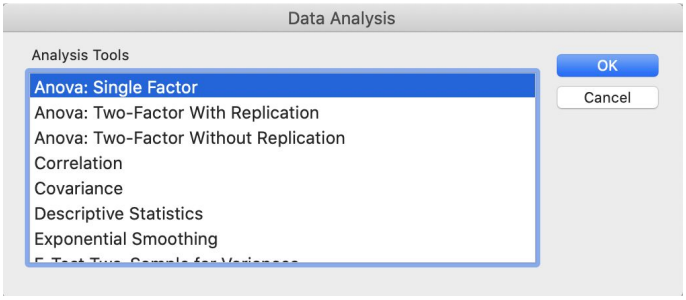


Figure 3.3.1.2 - Data Analysis Tool from MS Excel

Step B:

a) Select the Input Range

b) Grouped By: Column

c) Labels in First Row

d) Alpha = 0.05

e) Output: New Worksheet

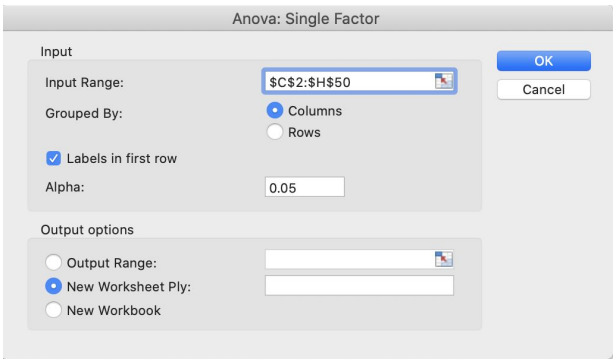


Figure 3.3.1.3 - Anova:Single Factor Data Selection

Step C:

Here, Figure 3.3.1.4 delivers the structure of an ANOVA output.

Source	*SS*	*df*	*MS*	*F*	*Sig.*
Between	SS_b	k-1	MS_b	MS_b/MS_w	*p* value
Within	SS_w	N-k	MS_w		
Total	$SS_b + SS_w$	N-1			

Figure 3.3.1.4: Structure of ANOVA: Single Factor

And here Figure 3.3.1.5 delivers the actual ANOVA output for the above selected data range.

	A	B	C	D	E	F	G
1	**Anova: Single Factor (All Attribution Models and Channel - Referral)**						
2							
3	SUMMARY						
4	*Groups*	*Count*	*Sum*	*Average*	*Variance*		
5	Last Interaction - Referral	48	181308	3777.25	3678186.745		
6	Time Decay - Referral	48	221320.2	4610.8375	11158524.04		
7	Linear - Referral	48	209088.99	4356.020625	4848744.974		
8	First Interaction - Referral	48	253887	5289.3125	7186939.453		
9	Position Based - Referral	48	214205.3	4462.610417	5074406.46		
10	Last Non-Direct Click - Referral	48	291724	6077.583333	9446640.589		
11							
12							
13	ANOVA						
14	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
15	Between Groups	156280972.1	5	31256194.41	4.530600894	0.000546055	2.246015189
16	Within Groups	1945491786	282	6898907.043			
17							
18	Total	2101772758	287				
19							

Figure 3.3.1.5 - ANOVA: Single Factor (Dataset 1)

Step D:

As per the conditions of ANOVA, we check if the output delivered shows statistically significant difference or not.

As in the Figure 3.3.1.5,

P-Value < Alpha i.e. 0.000546055 < 0.05 and so we consider F-Value result for further analysis.

F-Value > F-Critical i.e. 4.530600894 > 2.246015189 so it shows the presence of Statistically Significant Difference and so we Reject Ho and Accept H1.

Step E: Conclusion

There is Statistical Significant Difference in comparison of All Attribution Models with Single Channel - Referral.

Step F:

Similarly we perform the above steps for other two datasets i.e. and came up with the following conclusions:

Dataset 2: All Attribution Models for Channel - Paid Search

	A	B	C	D	E	F	G
1	**Anova: Single Factor (All Attribution Models and Channel - Paid Search)**						
2							
3	SUMMARY						
4	*Groups*	*Count*	*Sum*	*Average*	*Variance*		
5	Last Interaction - Paid Search	48	35864	747.1666667	1303967.333		
6	Time Decay - Paid Search	48	37346.93	778.0610417	1358123.941		
7	Linear - Paid Search	48	37146.22	773.8795833	1349083.227		
8	First Interaction - Paid Search	48	39230	817.2916667	1448054.381		
9	Position Based - Paid Search	48	37404.06	779.25125	1365691.593		
10	Last Non-Direct Click - Paid Search	48	42435	884.0625	1691388.102		
11							
12							
13	ANOVA						
14	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
15	Between Groups	560754.0355	5	112150.8071	0.079013676	0.995414895	2.246015189
16	Within Groups	400266503.2	282	1419384.763			
17							
18	Total	400827257.2	287				

Figure 3.3.1.6 - ANOVA: Single Factor (Dataset 2)

As in the Figure 3.3.1.6,

P-Value > Alpha i.e. 0.995414895 > 0.05 and so we discard F-Value result for further analysis and conclude that there is No Statistical Significant Difference and we Accept Ho.

Dataset 3: All Attribution Models for Channel - Social Network

	A	B	C	D	E	F	G
1	**Anova: Single Factor (All Attribution Models and Channel - Social Network)**						
2							
3	SUMMARY						
4	*Groups*	*Count*	*Sum*	*Average*	*Variance*		
5	Last Interaction - Social Network	48	34250.2	713.5458333	2180735.135		
6	Time Decay - Social Network	48	35085.97	730.9577083	2232080.206		
7	Linear - Social Network	48	35369.34	736.86125	2245490.173		
8	First Interaction - Social Network	48	36635	763.2291667	2339540.819		
9	Position Based - Social Network	48	35403.47	737.5722917	2254638.717		
10	Last Non-Direct Click - Social Network	48	37821	787.9375	2322864.23		
11							
12							
13	ANOVA						
14	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
15	Between Groups	167226.7338	5	33445.34677	0.014782093	0.999922069	2.246015189
16	Within Groups	638041416.2	282	2262558.213			
17							
18	Total	638208642.9	287				

Figure 3.3.1.7 - ANOVA: Single Factor (Dataset 3)

As in the Figure 3.3.1.7,

P-Value > Alpha i.e. 0.999922069 > 0.05 and so we discard F-Value result for further analysis and conclude that there is No Statistical Significant Difference and we Accept Ho.

3.4 Tukey HSD Post-Hoc Analysis

If ANOVA test shows that the means aren't all equal i.e. there is a statistical significant difference, the next step is to determine which means are different, to our level of significance.

In order to understand because of which models it has this difference we implement Tukey HSD (Honestly Significant Difference) Test.

The easiest thing is to compute the confidence interval first, and then interpret it for a significant difference in means (or no significant difference).

- If the endpoints of the Confidence Interval have the same sign (both positive or both are negative), then 0 is not in the interval and you can conclude that the means are different.
- If the endpoints of the Confidence Interval have opposite signs, then 0 is in the interval and you can't determine whether the means are equal or different.

Here we compute that confidence interval similarly to the confidence interval for the difference of two means, but using the q distribution:

$$\bar{x}_i - \bar{x}_j \pm q(\alpha, r, df_w)\sqrt{\frac{MS_w}{2}\cdot\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

where $\bar{x}i$ and $\bar{x}j$ are the two sample means, ni and nj are the two sample sizes, MSw is the within-groups mean square from the ANOVA table, and q is the critical value of the studentized range for α, the number of treatments or samples r, and the within-groups degrees of freedom dfW. The square-root term is called the standardized error.

As we concluded in Step 3 that Dataset 1: All Attribution Models for Channel - Referral has Statistical Significant Difference so here with implementing Tukey HSD we will proceed with

understanding exactly which means are different and because of which this Statistical Significant Difference showed up.

3.4.1 Estimating Differences of Means

Step A: Each row compares one pair of treatments. If you have r treatments, there will be r(r−1)/2 pairs of means.

Here, r = 6

Therefore r(r-1)/2 = 6(6-1)/2 = 15 i.e. 15 pairs as shown in Figure 3.4.1.2

Step B: Calculating point estimate of difference which is nothing more than the difference or the two sample means i.e. $\bar{x}i - \bar{x}j$ as shown in Figure 3.4.1.2

Step C: Calculating critical q, from the confidence interval formula q(α,r,dfW).

It depends on the number of treatments and total number of data points, not on the individual treatments, so it's the same for all rows in any given experiment.

Here, α = 0.05, r = 6 & dfw = 282

Calculating critical q online, gives the output as shown in Figure 3.4.1.1.

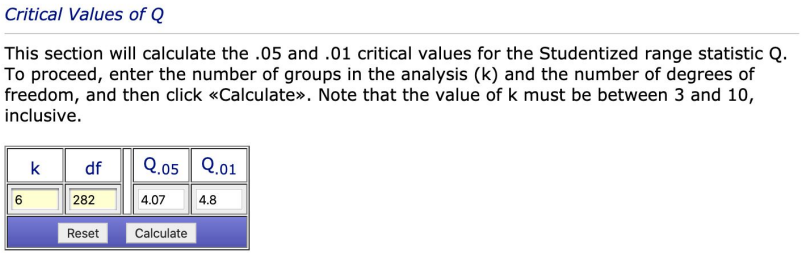


Figure 3.4.1.1 - Critical Q Calculation

As calculating critical q at α = 0.05, therefore we consider Q at 0.05 = 4.07

Therefore Critical q = 4.07

Step D: The standardized error is the square root term from Tukey's Formula of Confidence Interval.

$$\sqrt{\frac{MS_w}{2}\cdot\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

Here, as we each treatment has 48 data points, and so the standardized error is the same for every pair of means: sqrt([(MSW/2)·(1/6+1/6)])= sqrt([(6898907.043/2)·(1/48 + 1/48)] = 379.1137429

Therefore Standardized Error = 379.1137429 as shown in Figure 3.4.1.2

Step E: The endpoints of the confidence interval, as usual, are the point estimate plus or minus the critical q times the standardized error.

Critical q * standardized error = 4.07 × 379.1137429 = 1542.992934

The difference of means is $\bar{x}i-\bar{x}j$, so the endpoints of the confidence interval **"($\bar{x}i-\bar{x}j$) - (Critical q * standardized error)"** and **"($\bar{x}i-\bar{x}j$) + (Critical q * standardized error)"** as shown in Figure 3.4.1.2

Models	Xi - Xj	Critical q q(a,r,dfw)	Standerdized Error	95% C.I.for Mi - Mj		Significance at 0.05
Last Interaction - Time Decay	833.59	4.07	379.1137429	2,376.58	709.41	
Last Interaction - Linear	578.77	4.07	379.1137429	2,121.76	964.22	
Last Interaction - First Interaction	1,512.06	4.07	379.1137429	3,055.06	30.93	
Last Interaction - Position Based	685.36	4.07	379.1137429	2,228.35	857.63	
Last Interaction - Last Non-Direct Click	**2,300.33**	**4.07**	**379.1137429**	**3,843.33**	**757.34**	**Yes**
Time Decay - Linear	254.82	4.07	379.1137429	1,288.18	1,797.81	
Time Decay - First Interaction	678.48	4.07	379.1137429	2,221.47	864.52	
Time Decay - Position Based	148.23	4.07	379.1137429	1,394.77	1,691.22	
Time Decay - Last Non-Direct Click	1,466.75	4.07	379.1137429	3,009.74	76.25	
Linear - First Interaction	933.29	4.07	379.1137429	2,476.28	609.70	
Linear - Position Based	106.59	4.07	379.1137429	1,649.58	1,436.40	
Linear - Last Non-Direct Click	**1,721.56**	**4.07**	**379.1137429**	**3,264.56**	**178.57**	**Yes**
First Interaction - Position Based	826.70	4.07	379.1137429	716.29	2,369.70	
First Interaction - Last Non-Direct Click	788.27	4.07	379.1137429	2,331.26	754.72	
Position Based - Last Non-Direct Click	**1,614.97**	**4.07**	**379.1137429**	**3,157.97**	**71.98**	**Yes**

Figure 3.4.1.2 - Output of Tukey HSD for Dataset 1

Step F: Here in the Figure 3.4.1.2, the output shows that the means are different for the following models and results in Statistical Significant Difference:

Last Interaction - Last Non-Direct Click

Linear - Non-Direct Click

Position-Based - Last Non-Direct Click

Chapter Four
CONCLUSION

In this project, we conclude that the Google Merchandising Store Data from Model Comparison Tool on Google Analytics has a **"Significant Difference"** for **"Channel - Referral"** over the last "4 years" because of the difference in means in following Models:

"Last Interaction - Last Non-Direct Click

Linear - Non-Direct Click

Position-Based - Last Non-Direct Click"

We came to this conclusion by using Statistical Analysis Methods such as ANOVA & Tukey HSD (Honest Significant Difference).

REFERENCES

1. **https://analytics.google.com/analytics/web/#/report/bf-roi-calculator/a54516992w87479473p92320289/**
2. **https://agencyanalytics.com/blog/marketing-attribution-models**
3. **https://www.graphpad.com/support/faq/how-can-i-determine-needed-sample-size-for-an-experiment-to-be-analyzed-by-two-way-anova/**
4. **https://brownmath.com/stat/anova1.htm#HSD**
5. **http://www.vassarstats.net/tabs.html#q**
6. **http://www0.cs.ucl.ac.uk/staff/w.zhang/rtb-papers/data-conv-att.pdf**