Rotterdam School of Economics
Rotterdam, Block III, Spring 2021

Multi-Channel Budget Allocation

Evaluating W.M. Winters Channel Strategy to
Enhance Revenue and New Customer Acquisition

593088ah

Digital Marketing Strategy

Evaluator: Xi Chen

Table of Contents

Analysis and Evaluation (Section I)

This analysis part focuses exclusively on the analysis of the given data and does not consider any attribution models. This first abstraction is important to give indications regarding important position classes, especially for Converters and Originators, because those only appear once per order. Additionally, we decided to focus on an aggregated channels, because this allows for a facilitated and more clear understanding of the situation.

Channel and Touchpoint Distribution (Q1.1)

Figure 1: Channel Touchpoints per Position Classification

Affiliate Marketing (AM), is also known as referral marketing and uses methods, such as SEO, paid search engine, email, and content marketing to drive awareness for the brand across a publishing network. In this case, AM includes “Buzz Affiliate” and “CJ”. We see a strong concentration on “Converter” and “Roster”, especially the former is strong. Decent performance in assist makes sense because AM works usually best after a user already heard of the brand / merchant. This also explains why performance in the “Originator” class is lower than for other channels. Generally, Buzz Affiliate outperforms CJ.

Display Advertisement (DA) are ads provided by CPM. Generally, it is difficult to assess the performance of ads if the campaigns and their underlying strategy, objectives, and actual creatives are not known. For the sake of simplicity, we assume that the displayed ads include

brand awareness campaigns, that target the “Originator” and “Roster” part, but also selling campaigns that drive revenue by focusing on conversions and assisting by using pain points and call-to-actions. This assumption is also supported by Figure 1 and Figure 2, showing that DA is very important across all position classes. It’s superior performance in “Roster” also displays, that retargeting was applied, meaning that same customers got targeted by the ads multiple times until they were assisted and finally converted. This as further indications for the eventual attribution strategy, that will be discussed in Section II. This analysis shows that DA seems to be the most important channel for overall success.

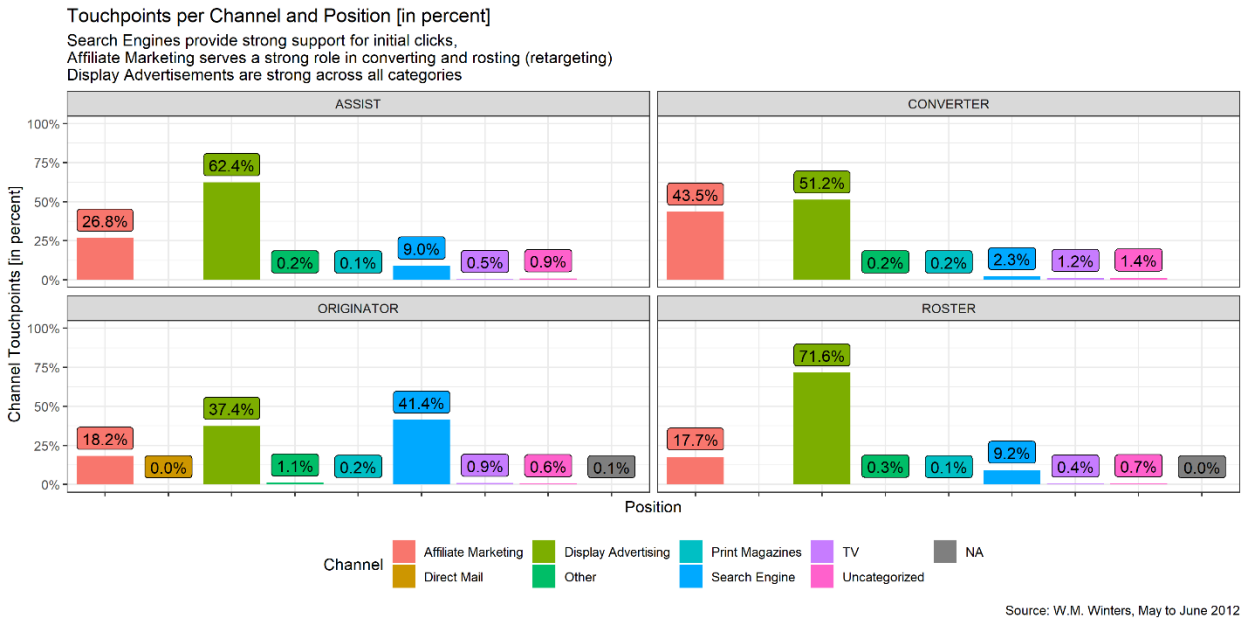

Figure 2: Share of Aggregated Channels and Position Classes

Search Engine (SE) includes different search engines using either branded or non-branded keywords. The aggregated picture can be seen in Figure 2. SE are the most important channel in terms of click origination, meaning that the most first clicks come from SE listings. Their performance across other position classes (e.g. “Converter”, “Roster”, “Assist”) remains marginal.

Social Media Sites only include three data points overall, meaning that no reliable indications can be inferred. Other channel touchpoints, especially direct (print) mailings, magazines, or TV play an inferior role, as can be seen in Figure 1 and Figure 2. That offline and uncategorized channels perform below their online counterparts may be not surprising, because of the nutrition industry focus of M.W. Winters and the overall digital multi-channel strategy.

Differences in Conversion and Sales (Q1.2)

We calculate the time to convert as the difference between *Positiontime* and *Orderdatetime*. The conversion time is expressed in hours to get a better understanding of within-day conversions. We express the results as conversion time averages on channel level. Channels categorized as “Other” or “Uncategorized” and Direct Mailings have an inferior conversion time performance. Offline channels, such as TV and Print Magazines show good results, especially if they serve as “Converters”. Search Engine (SE) is a strong “Converter”, but has an average performance on the “Originator” side. This is important to keep in mind, because SE is the channel that serves most often as the initiator. Affiliate Marketing (AM) shows low conversion times if it acts as “Converter”, but suffers from higher conversion times if it acts as a conversion. Display Advertising (DA) seems to be the best overall channel in terms of conversion times, because it has very short conversion times when it acts as “Originator” and “Converter”.

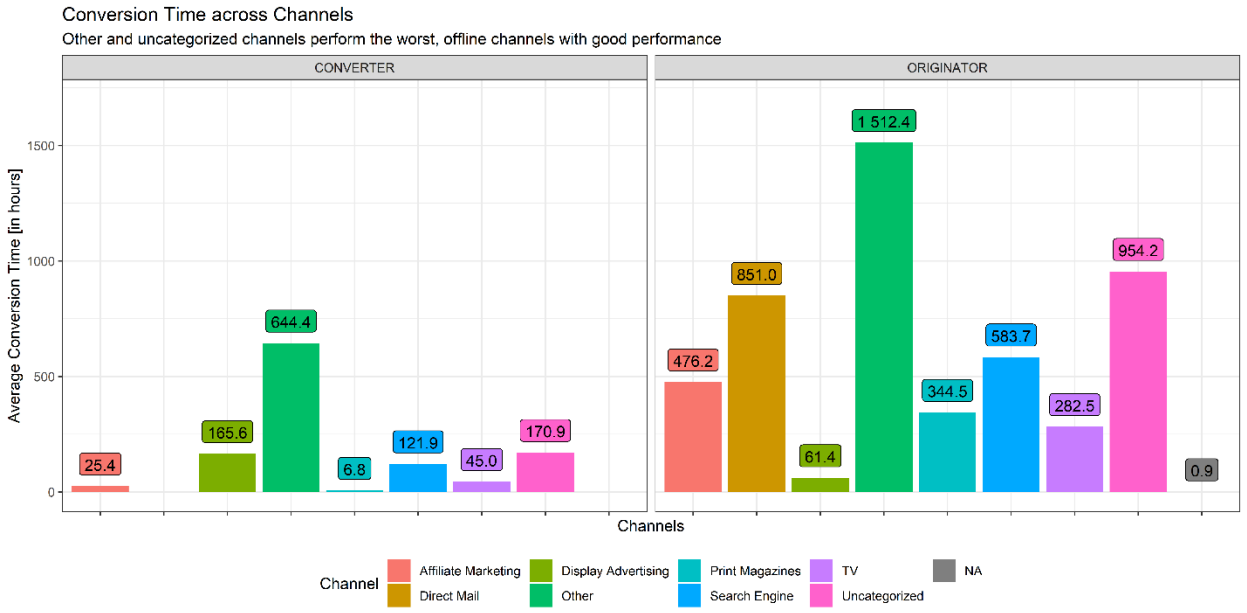

Figure 3: Distribution of Conversion Time across Interval and within first 24 hours

As shown in Figure 1 and Figure 2 before, Affiliate Marketing and Display Advertising were the biggest contributors to converting touchpoints, with all other channels having almost no real contribution. This fact is also shown in Figure 4, representing the revenue achieved in the respective position class and channel. Same holds true for the “Originators”, the class in which Search Engine perform the best in terms of revenue, followed by Display Advertising (CPM) and Affiliate Marketing (Buzz Affiliate, CJ). Other channels are not noteworthy, because of their marginal shares in generated revenue. This figure supports the previous hypothesis of DA being

the most important channel. Not only does it provide the highest average revenue across its "Originator" or "Converter" role, but also it shows the lowest conversion times across all channels.

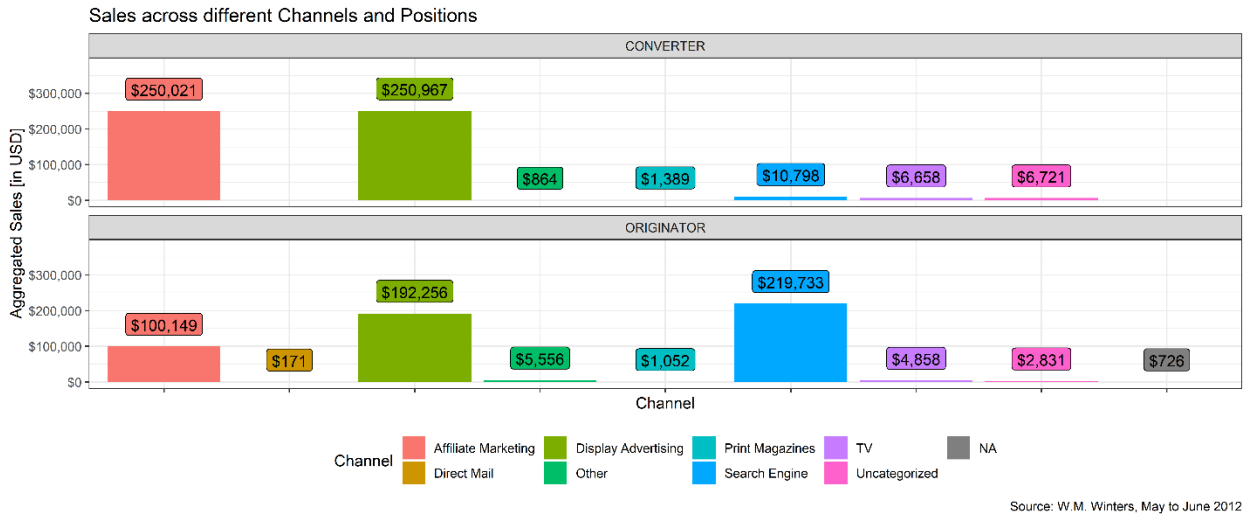

Figure 4: Sales across Converter and Originators and across Channels

Differences in Customer Behavior (Q1.3)

New customers convert faster across all positions (Figure A1, Appendix) and across all channels (Figure 5). The reduction is important for AM, DA, and SE because of their sales weight.

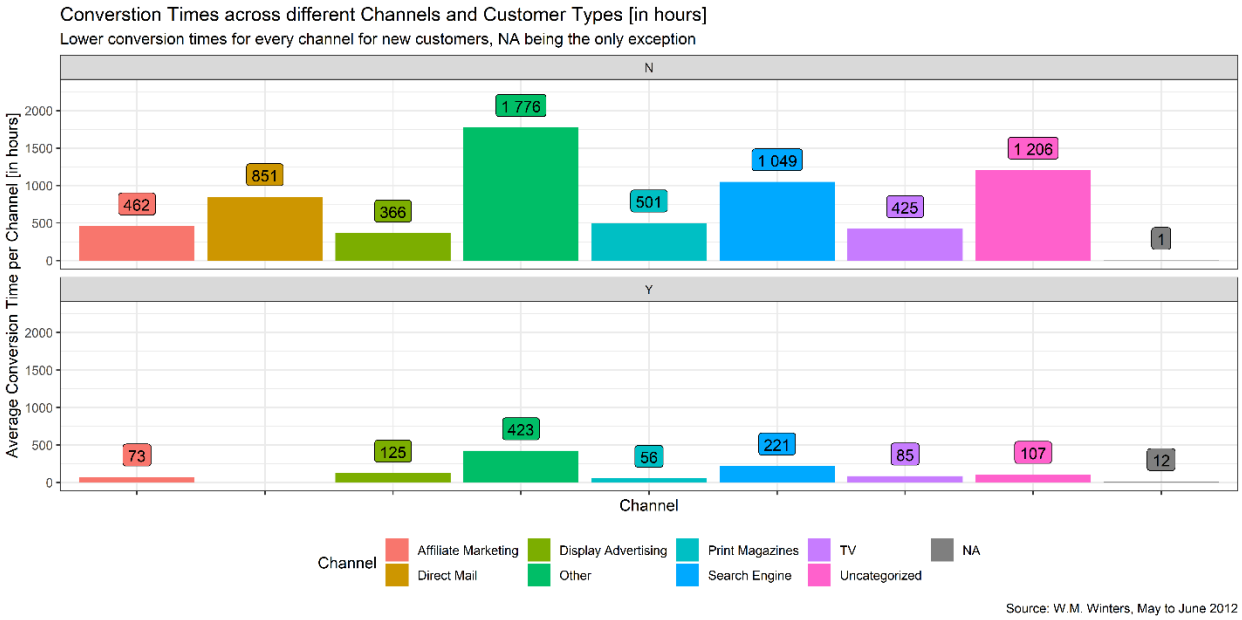

Figure 5: Average Conversion Time for Customer Types, across Channels

In terms of sales, new customers also spend on average significantly more (Figure A2, Appendix) and across all channels (Figure 6), with the "NA" channel being the only exception, because there

are too few data points to make any reliable suggestions. Strong differences in sales can be seen for all channels. But comparing the total number of touchpoints and revenue, these indications support the thesis that AM, DA, and SE channels are the three most important channels.

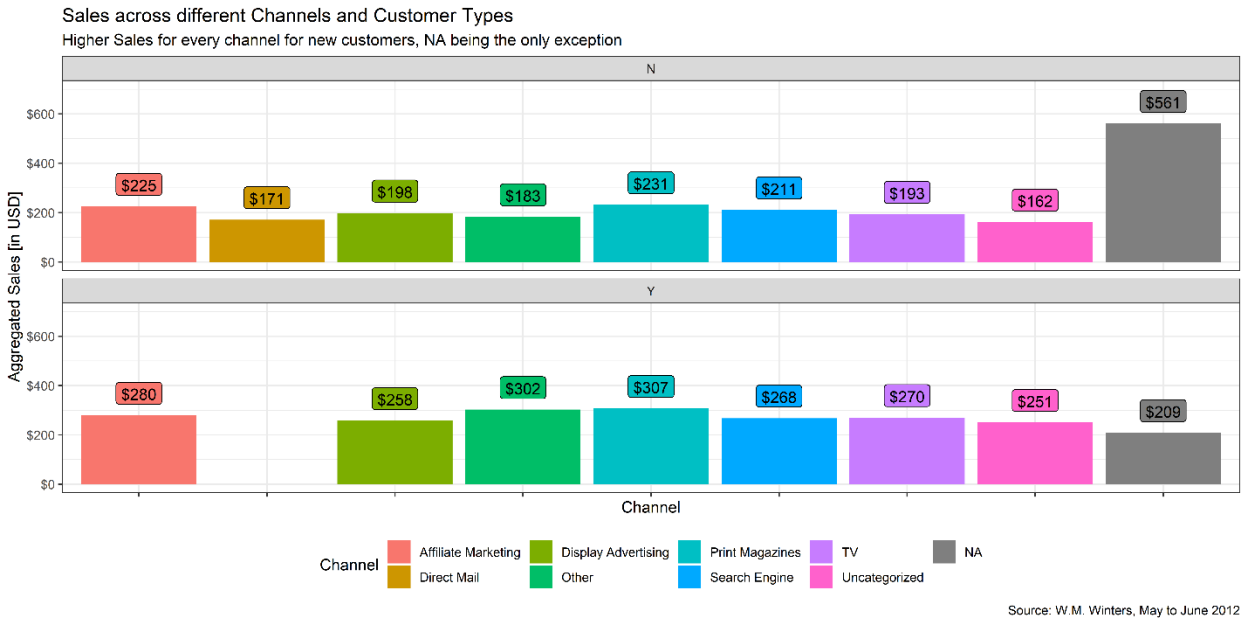

Figure 6: Average Sales for Customer Types, across Channels

When comparing the touchpoints specifically for “Originators” and “Converters”, it can be observed that in general SE, DA, and MA have a higher overall importance. This is logical, because we also have a biased data set with more touchpoints for new customers (5,570 vs. 4,549), which also refers to more new customers that actually converted (1,296 vs. 874). The only mentionable thing for a latter attribution strategy is, that the Affiliate Marketing has a considerably higher importance for converting new customers than old ones, as can be seen in Figure 7, which will have important budget allocation indications the eventual marketing strategy.

Multi-Channel Attribution (Section II)

As we have seen in the previous parts, channels have different importance for customer acquisition of revenue growth, depending on which position classification is applied. If first clicks are valued more than other position classes, SE grows in importance relative to other channels. The opposite is true if we look exclusively at converters, a class SE plays an insignificant role

compared to AM and DA. But, all these assumptions did not consider the touchpoints between the first and last-click, and their importance to generate revenue or acquiring new customers.

Figure 7: Difference in Touchpoints per Customer Type and Channel

Multi-Channel Attribution Strategy (Q2.1)

In the beginning, we only compare three basic attribution strategies, namely even, first-click, and last-click attribution. The first attributes to every touchpoint within a conversion journey the same share in revenue. The second attributes the whole revenue share to the first click so mainly "Originator", and the third to the last click, meaning that "Converters" play the most important role.

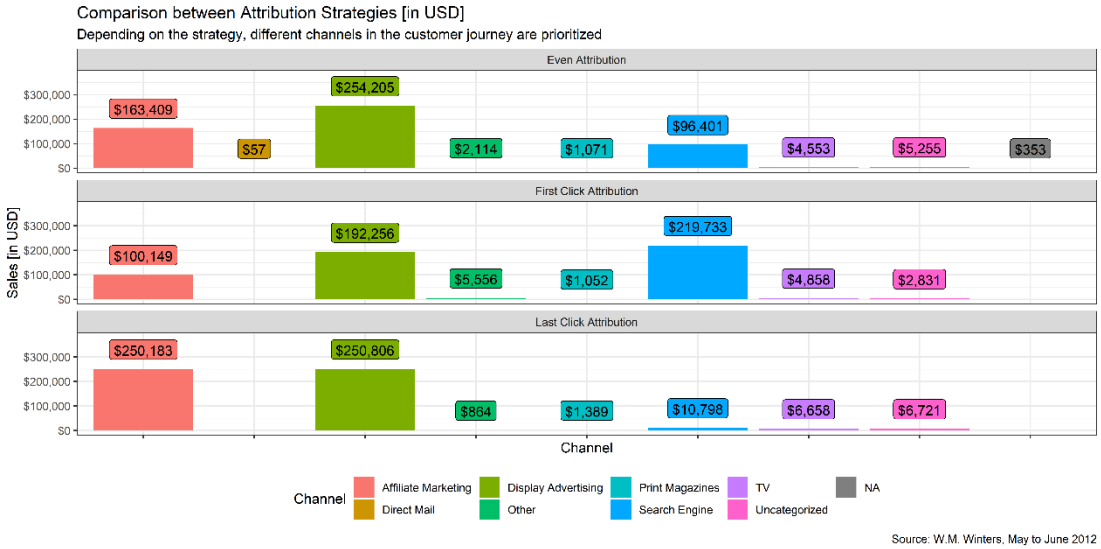

Figure 8: Comparison between Attribution Strategies

Recommendation for Best Attribution Strategy (Q2.2)

As we saw before in the channel comparisons, last-click attribution is not a suitable model, because it only weighs in favor of the converting channels. But, as was illustrated in Figure 2, 41.4% of all first clicks come from Search Engines. Opting for a last-click strategy would therefore get rid of 41.4% of initial clicks and therefore harm the effectiveness of the campaign greatly.

The opposite argument applies to the first-click model. This model attributes too much revenue to the initial channels. Even though SE is responsible for 41.4% of first clicks, and DA only for 37.4%, DA and AM are more important across all different position classes, from roster, to assist, to ultimately converter. Looking at converters and the results from Figure 2, we see that AM is responsible for 43.5% and DA for 51.2% of the converting customers. Therefore, an attribution that skews the allocated budget towards SE leads to bias in multi-channel result interpretation.

The best model among the three is the even attribution. It does not suffer from the same biases and considers the importance of Assist and Roster as important position classes. Rightfully, it also assumes a suitable hierarchy between SE, AM, and DA with the former being the last important of the three and DA being the most important channel (see Figure 8). The only downsides is that it biases the results for channels, which will have multiple “Roster” positions.

Tailor-Made Attribution Strategy for Revenue Growth (Q2.3)

Additionally, the even attribution model has a further drawback: It assumes that every touchpoint is of equal importance to convert a user and therefore, to generate revenue. Even though this might be closer to reality than the first- or last-click attribution, it is difficult to check without knowing the actual costs of the different channels. Also, it is not in line with the goal of “increasing revenue”, because it will weigh mid-positions, such as “Roster” unproportionally, because of their repeated appearance. But assuming a target audience, which is defined as an audience for that the probability of purchase is the highest, is defined, the ROAS for such accurate techniques are lower than to display the ads in the first place (WordStream, 2020). Also, convincing a user with each repeated step is easier, because the brand and product offerings are most likely known, if compared with “Originators”. This means that “Roster” and “Assist” should play a smaller role.

Additionally, without a first- or a last-click an order would not happen at all, meaning that no revenue would be generated. Therefore, we build a custom position-based model that prioritizes the first and last-click more than the touchpoints in between, especially to target the downside of

biasing results for retargeting. This is necessary, because of the different importance of channels when comparing their role as “Converters” and “Originators”. Additionally, this model gives us a weight, which we can optimize to find the most effective weight for first, last-click and consequently for the channels weights in between. The position-based weights are:

- If only two positions (Originator + Converter) = 50% each
- First position (always Originator) = k %
- Last position (always Converter) = k %
- Every position in between (Roster, Assist) = (1 – 2k) / number of mid-positions

This means that the higher k, the higher the attribution towards “Originators” and “Converters”, and the lower the attribution towards “Roster” and “Assist”. These weights are optimized in order to fix the problems of even attribution, especially regarding the attribution to “Assist” and “Roster”. We opt for optimization, because the suggested weights do not consider the specifics of this data set (Neilpatel, 2020). The specifics of the weight optimization can be found in Figure A4, Appendix. An important finding is, that regardless of the weight, the achieved revenue stays the same. Higher weight towards “Originators” and “Converters” leads to an increasing allocation to SE and AM channels, but this increase is countered by a decrease in attributed revenue for DA. Therefore, we set the weight, k, to 30% to balance the effect between the position classes. This balance is important in order to prioritize AM and DA, due to their influence on conversion time reduction and sales increase, regardless of customer type (see Figure 5 to 7). Another reason is to balance the effect of SE, a channel which has overall limited but important benefits for customer origination. Optimizing these factors also optimizes the revenue gathered via these channels.

In Figure 9, we compare the even attribution with the optimized position-based model. Compared to Figure 2, we achieve a higher total attribution to converter and originator, while reducing the effect of the other position classes respectively. Despite the low effectiveness and importance of the remaining channels, these entries cannot be deleted, as it would severely impact the data, by reducing the overall revenue and making attribution models incomparable. A way to deal with this will be suggested in the last section of this resport. In Figure 9 we see an important feat of the proposed attribution model: we change the weighs between the position classes. Overall, we decreased the attribution for “Assist” and “Roster” positions based on our cost assumption. Additionally, we increased the revenue attribution towards the critcial customer journey points: the first- and last-clicks. Upon closer inspection, we also see a closer share between SE and DA in the “Originator” part, which is closer to the actual touchpoint splits, we have seen in Figure 2. Figure 10 indicates, that in total a smaller share is attributed to DA (46.0% vs. 48.2%), because

we reduce its priority for “Assist” and “Roster” while increasing it for positions that actually play a bigger part in generating revenue At the same time we increase the weight for AM and SE to increase revenue potential and conversions. The aggregated channel weights are in Figure 10.

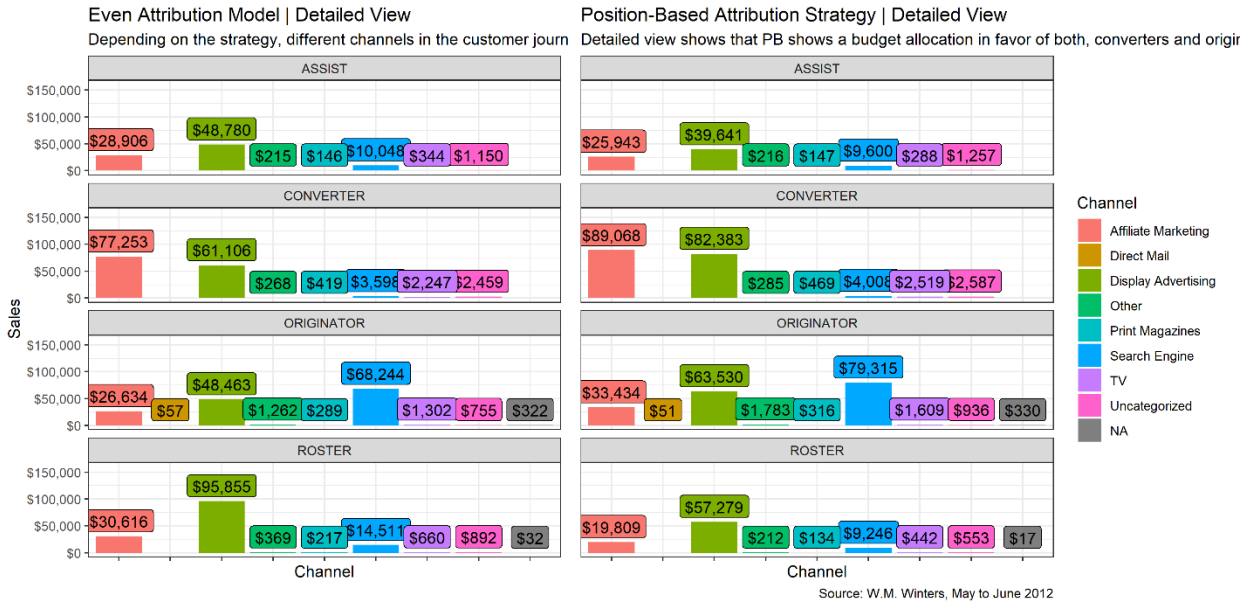

Figure 9: Comparison between Even-Click and Position-Based Model across Classes

Tailor-Made Attribution Strategy for Customer Growth (Q2.4)

As was observed in Figure 5, new customers are converted more effectively via the Affiliate Marketing (AM) channel, but conversion by the Display Advertisements (DA) channel lead to a higher average increase in revenue, if compared to existing customers. This is why balancing both channels is important, a flaw of the even attribution model, which is corrected by the position-based model, especially if we check the attribution difference for DA and AM at “Converters” between both attribution models (Figure 9). We do not opt for a higher weight (k), because this would change the allocation distribution between SE, DA, and AM even further, a distribution that is already resembling the one in Figure 7.

As we have seen, the shorter conversion time also translates into a slightly different distribution: New customers require on average also fewer positions to be fully converted, as shown in Figure A5, Appendix. Also, most new customers can be found around position zero and three. Thus, we assume that a stronger focus on first- and last-clicks, is the strategy to increase revenue but also to acquire more new customers. Conclusively, this means that the position-based model from 2.3. is also the best choice in order to acquire new customers, because it prioritizes the first- and last-

clicks. An even higher weight (k) would also diminish the importance of “Roster” and “Assist” channels even further, and therefore put an overly strong focus on the first- and last-clicks.

Final Recommendations (Q2.5)

Without knowing the cost structure of the multi-channel set-up, we rely only on the effectiveness of the current strategy in terms of touchpoints, customer types, and achieved revenue to define the precise budget splits. On the one hand, we understood that offline and other channels were not as effective as DA, MA, and SE. Problems ranging from data availability, to tracking problems of general disadvantages of offline solutions, and the suboptimal classification of these channels lead to the conclusion that those channels should be dropped for future marketing strategies. On the other hand, we saw ranging effectiveness of SE, AM, and DA, depending on the position class or position in the customer journey and the targeted customer type. Despite different possible contributions of the respective channels, we introduced an weight-optimized position-based attribution model that accounts for all the biases of the standard attribution methods. These biases can be summarized as bias towards first- or last-click, bias towards mid-positions, such as “Roster” if the number of touchpoints becomes large (> four), and generally revenue allocation to channels that are not sought after by new customers.

Considering this, there are two scenarios for the budget allocation: a) from 2.3 and b) without any ineffective channels, with the latter (b) being the final recommendation for the budget allocation. This recommendation also includes the growth trend that builds on targeting new customers.

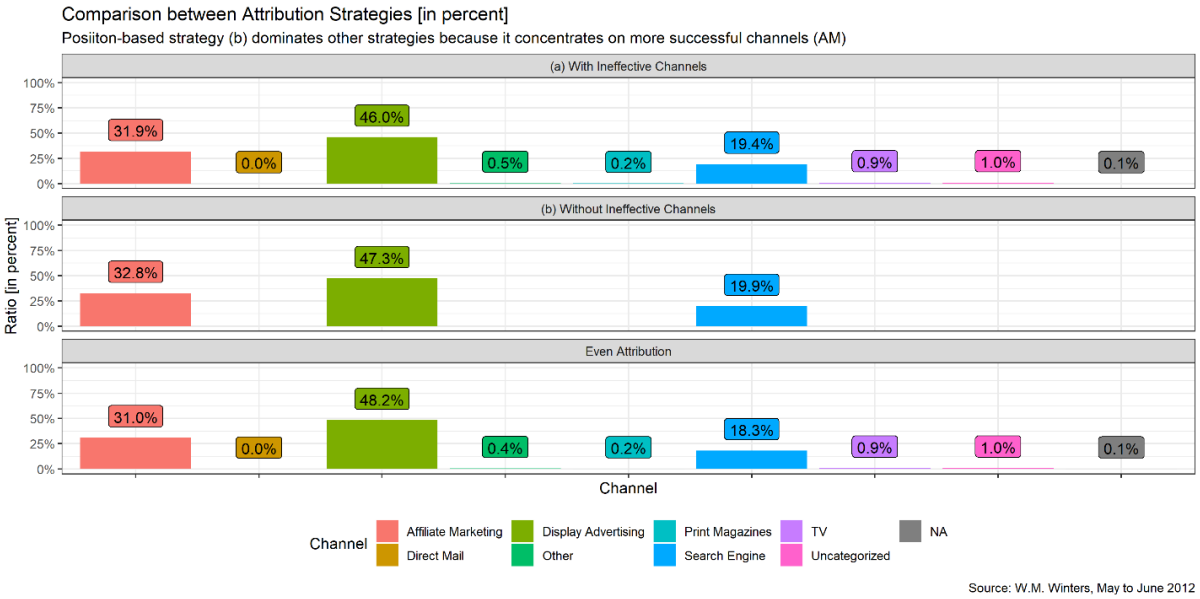

Figure 10: Final Budget Allocation (in percent, based on revenue attribution)

Sources

Neilpatel. (2020). *How to Pick the Right Analytics Attribution Model When There's No Right Answer*. https://neilpatel.com/blog/best-analytics-attribution-model/

WordStream. (2020, September 26). *7 Super-Creative, Crazy-Effective Retargeting Ad Ideas | WordStream*. https://www.wordstream.com/blog/ws/2016/04/13/retargeting-ad-ideas

Appendix

Appendix A: Supporting Graphs and Tables

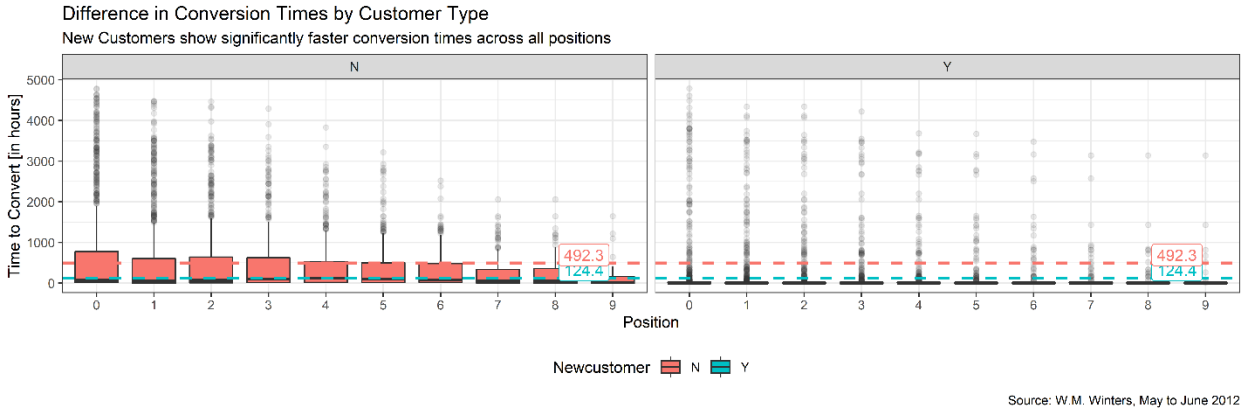

Figure A1: Difference in Conversion Times between Customer Type and along Position

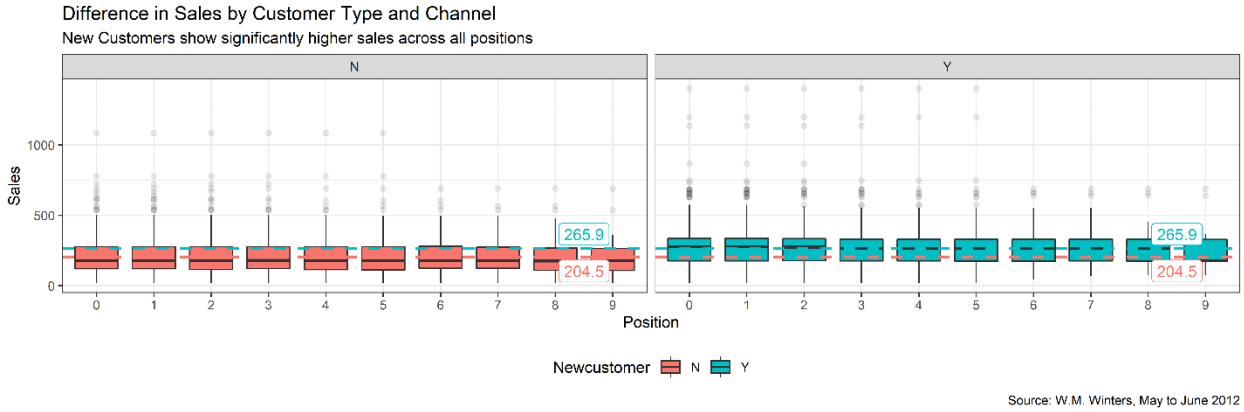

Figure A2: Difference in Sales between Customer Type and along Position

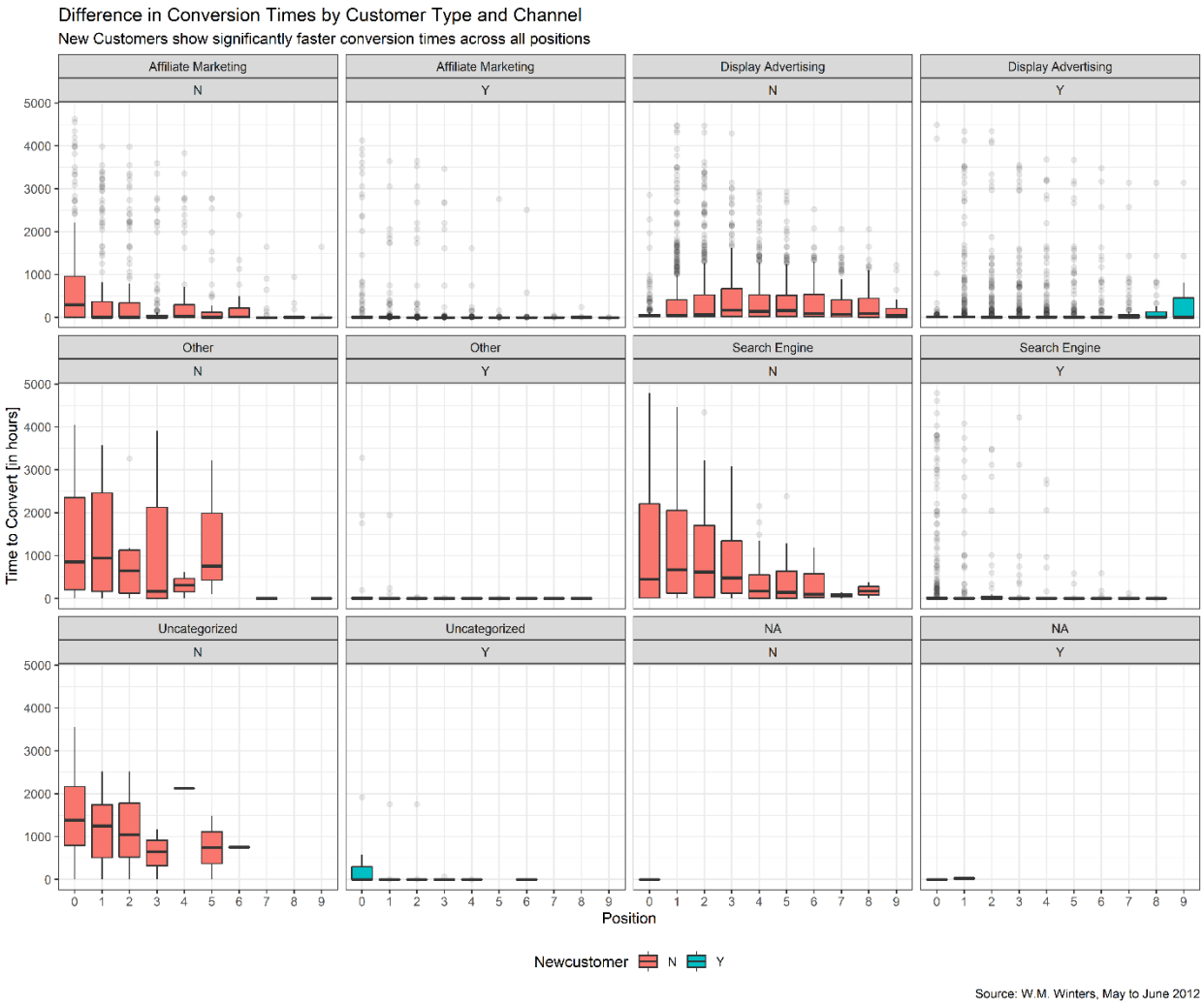

Figure A3: Difference in Conversion Times by Customer Type and Channel

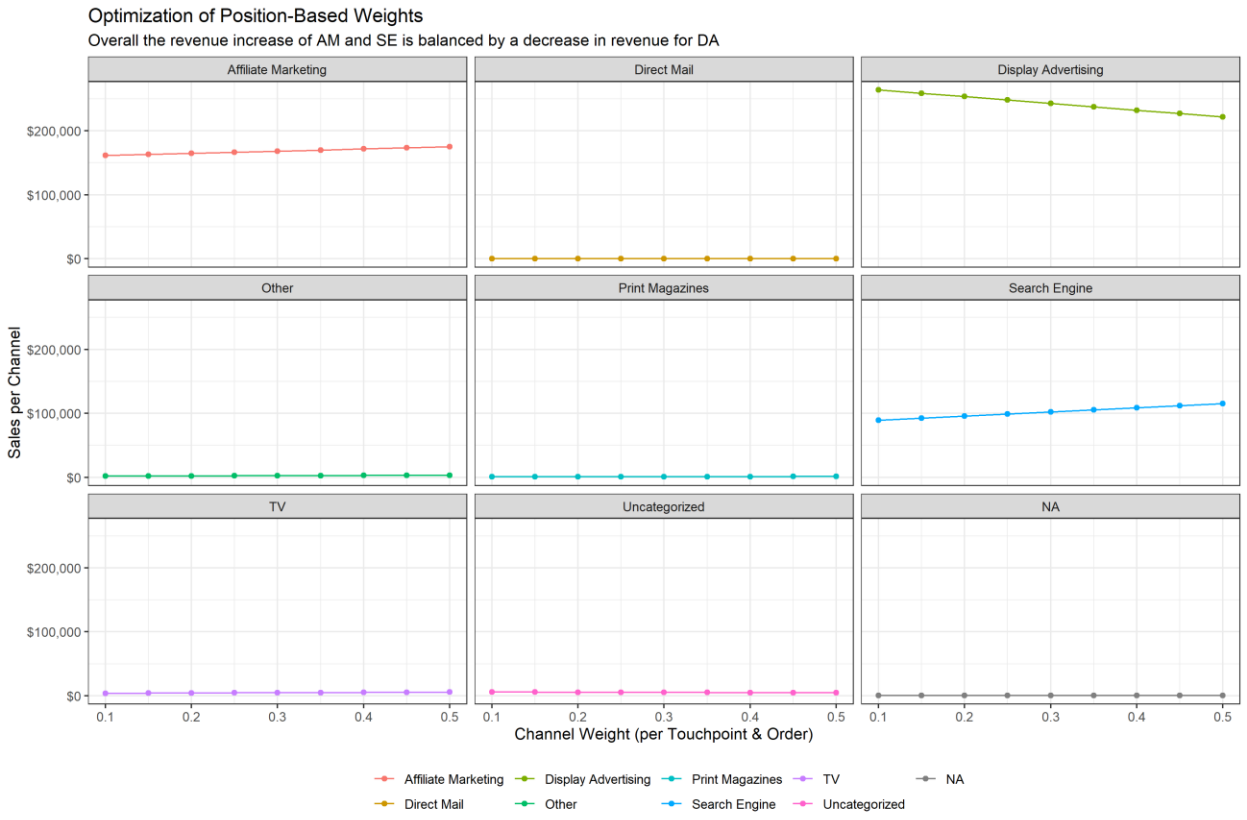

Figure A4: Weight-Optimization (k) for Position-Based Attribution Strategy

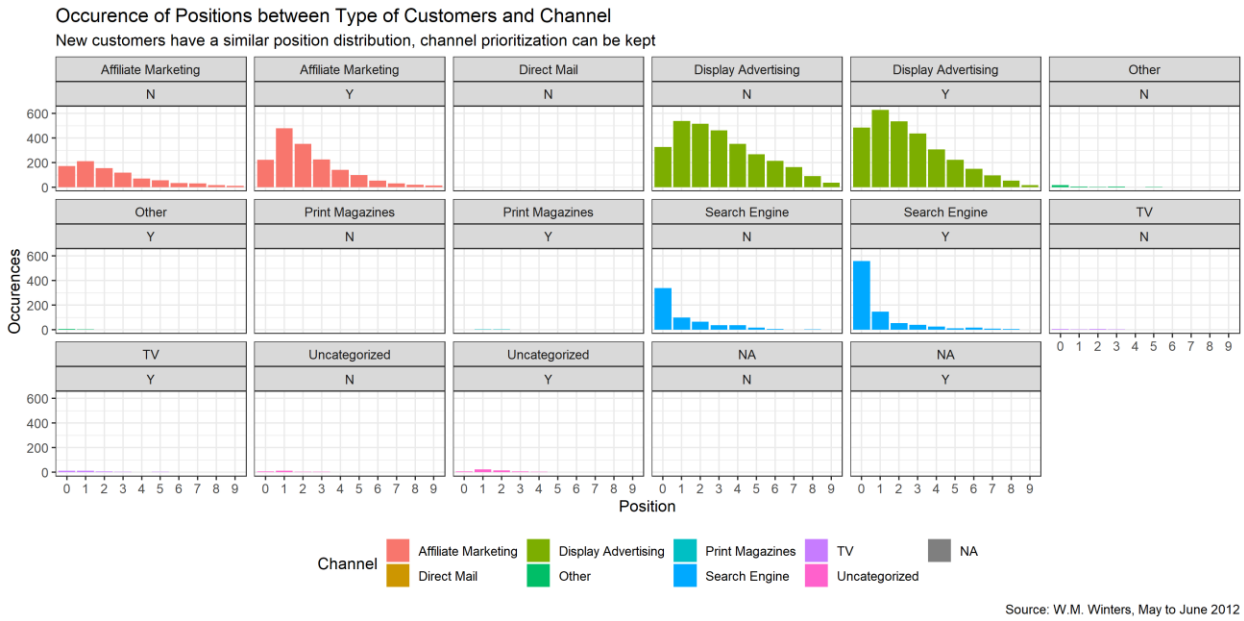

Figure A5: Position Occurrence Comparison between Type of Customer and Channel

Appendix B: Assumptions and Data Analysis

Assumptions and processing heuristics influence the eventual results of all data-based reports. This section will shortly discuss the different data characteristics and point out, which aggregation level were chosen for the remainder of this report.

Data Structure

The given data set includes 10,119 data points, which represent individual touchpoints made per order. Every order is uniquely identified using an ID, showing that 2,170 actual orders are in the data set. This distinction is important, because revenue is only gathered per order, and not per touchpoint. The granular attribution of revenue to channels and positions is the objective of Section II.

Touchpoint Logic

Every touchpoint has an associated position, ranging from zero to nine. This facilitates the analysis of how many touchpoints are needed per different channel, position, or customer type. Additionally, this serves as basis for the classification into Originator, Roster, Assist, and Converter. The following logic applies for the classification:

- Two positions: Always Originator + Converter
- Three positions: Always Originator + Assist + Converter
- Four positions: Originator + Roster + Assist + Converter
- N positions: Originator + (n-3) x Roster + Assist + Converter

Appendix C: Software

The underlying data and graphs were implemented using R and the Integrated Development Environment (IDE) RStudio. All imported functionalities and packages can be seen in the attached script in Appendix D.

Appendix D: R Script

```
# LIBRARIES
library(xlsx) # excel import
library(dplyr) # data preprocessing and data wrangling
library(corrr) # simple correlation analysis
library(scales) # adjustable axes
library(tsibble) # dealing with time-series tibbles
library(ggplot2) # general plotting
library(patchwork) # side-by-side plotting
library(lubridate) # advanced date operations
library(tidyverse) # advanced data wrangling

# LOADING THE DATA -----
sdat <- read.xlsx("attribution_data.xlsx",
                  sheetIndex = 1,
                  as.data.frame = T) %>%
  mutate(ID = seq(1:nrow(.)))

# VARIABLE TRANSFORMATIONS ----
# introduce new channel variables

# TODO: Uncategorized as "Other" ??
sdat <- sdat %>%
  mutate(Channel = case_when(Groupname == "BUZZ AFFILIATE" ~ "Affiliate Marketing",
                             Groupname == "CJ" ~ "Affiliate Marketing",
                             Groupname == "CPM" ~ "Display Advertising",
                             Groupname == "SEARCH GOOGLE NON-BRAND" ~ "Search Engine",
                             Groupname == "SEARCH MSN NON-BRAND" ~ "Search Engine",
                             Groupname == "SEARCH GOOGLE BRAND" ~ "Search Engine",
                             Groupname == "SEARCH MSN BRAND" ~ "Search Engine",
                             Groupname == "SEARCH YAHOO BRAND" ~ "Search Engine",
                             Groupname == "SOCIAL" ~ "Social Media Sites",
                             Groupname == "Uncategorized" ~ "Uncategorized",
                             Groupname == "OTHER" ~ "Other",
                             Groupname == "PRINT - MAGAZINES" ~ "Print Magazines",
                             Groupname == "TV" ~ "TV",
                             Groupname == "DIRECT MAIL"~ "Direct Mail"))


# change to factors
sdat_fact <- sdat %>%
  mutate(Newcustomer = as.factor(Newcustomer),
         Groupname = as.factor(Groupname),
         Brand = as.factor(Brand),
         Positionname = as.factor(Positionname),
         Channel = as.factor(Channel),
         Position = as.factor(Position),
         TimeToConvert = difftime(Orderdatetime, Positiondatetime, unit = "hours"),
         TimeToConvert = as.numeric(TimeToConvert)) %>%
  select(Orderid, Saleamount, Position, TimeToConvert, Channel, Groupname, Positionname, Newcustomer)



# PREPROCESSING -----
# TODO: needs to be mentioned in the beginning because later posiion and channel analyses will be wrong because
no attribution assumptions


# real orders (everything is an order, just every touchpoint is recorded)
sdat %>% distinct(Orderid) %>% count()

# there is never an order with only one touchpoint
sdat_fact %>%
  group_by(Orderid) %>%
  count() %>%
  ggplot(aes(n)) +
  stat_density() +
  labs(title = "Number of Positions per Actual Order",
       subtitle = "No order was finished with only one touchpoint",
       x = "Number of Touchpoints",
       y = "Density") +
  theme_bw()
```

```r
# sales per order
sdat %>% group_by(Newcustomer, Orderid) %>%
  summarise(sales = mean(Saleamount)) %>%
  summarise(total_sales = sum(sales))

# distribution analysis, right tail revenue distribution seems normal
sdat %>%
  pivot_longer(cols = c(Saleamount, Position),
               names_to = "variables",
               values_to = "values") %>%
  ggplot(aes(values)) +
  stat_density() +
  facet_wrap(~variables, scales = "free", ncol = 1)

# relationship analysis
sdat_fact %>%
  ggplot(aes(TimeToConvert, Saleamount)) +
  geom_point(alpha = 0.1) +
  geom_smooth(group = 1) +
  facet_wrap(~Channel + Newcustomer, ncol = 2) +
  theme_bw()

# relationship of position classes and actual positions
sdat_fact %>%
  ggplot(aes(Position)) +
  stat_density() +
  facet_wrap(~Positionname) +
  labs(x = "Position",
       y = "Density",
       title = "Distribution of Positions across Position Classifications",
       subtitle = "Different position classification play a role across all positions",
       caption = "Source: W.M. Winters, May to June 2012") +
  theme_bw() +
  theme(legend.position = "bottom", legend.box = "horizontal") +
  scale_color_discrete(NULL) +
  guides(colour = guide_legend(nrow = 1))

ggsave("01_00_Distribution of Positions across Position Classifications.png", width = 8, height = 3)



# TASK 1.1 ------
# old form including all channel
sdat_fact %>%
  group_by(Positionname, Groupname) %>%
  count() %>%
  arrange(desc(n)) %>%
  ungroup() %>%
  ggplot(aes(Groupname, n, fill = Groupname), colour = "white") +
  geom_bar(stat = "identity") +
  geom_label(aes(label = n), vjust = -0.3, label.size = 0.05, label.r = unit(0.05, "lines")) +
  facet_wrap(~Positionname, nrow = 3) +
  coord_cartesian(ylim = c(0,3500)) +
  labs(x = "Position",
       y = "Channel Touchpoints",
       title = "Touchpoints per Channel and Position",
       caption = "Source: W.M. Winters, May to June 2012") +
  theme_bw() +
  theme(axis.text.x = element_blank(),
        legend.position = "bottom", legend.box = "horizontal",
        text = element_text(size = 8)) +
  scale_color_discrete(NULL) +
  guides(colour = guide_legend(nrow = 1))

ggsave("01_1.1_Touchpoints per Touchpoint Channel and Position.png", width = 12, height = 6)
```

```r
# TASK 1.2

# hyp 1: TTC was calculated correctly, just grouping needed
sdat_fact %>%
  group_by(Channel, Positionname) %>%
  filter(Positionname == "ORIGINATOR" | Positionname == "CONVERTER") %>%
  summarise(mean_ttc = mean(TimeToConvert)) %>%
  ggplot(aes(Channel, mean_ttc, fill = Channel)) +
  geom_col() +
  geom_label(aes(label = number(mean_ttc, accuracy = .1)), vjust = -.5, show.legend = F) +
  facet_wrap(~Positionname, nrow = 1) +
  coord_cartesian(ylim = c(0,1700)) +
  labs(x = "Channels",
       y = "Average Conversion Time [in hours]",
       title = "Conversion Time across Channels",
       subtitle = "Other and uncategorized channels perform the worst, offline channels with good performance")
+
  theme_bw() +
  theme(axis.text.x = element_blank(),
        legend.position = "bottom", legend.box = "horizontal") +
  scale_color_discrete(NULL) +
  guides(colour = guide_legend(nrow = 1))

ggsave("01_1.2_Conversion Time Differences.png", width = 12, height = 6, dpi = 600)



# amount of sales
sdat_fact %>%
  filter(Positionname == "ORIGINATOR"| Positionname == "CONVERTER") %>%
  group_by(Positionname, Channel) %>%
  summarise(sum_sales = sum(Saleamount)) %>%
  arrange(desc(sum_sales)) %>%
  ggplot(aes(Channel, sum_sales, fill = Channel)) +
  geom_col() +
  geom_label(aes(label = dollar(sum_sales, accuracy = 1)), vjust = -.5, show.legend = F) +
  facet_wrap(~Positionname, nrow = 2) +
  coord_cartesian(ylim = c(0,380000)) +
  labs(x = "Channel",
       y = "Aggregated Sales [in USD]",
       title = "Sales across different Channels and Positions",
       caption = "Source: W.M. Winters, May to June 2012") +
  scale_y_continuous(labels = dollar) +
  theme_bw() +
  theme(axis.text.x = element_blank(),
        legend.position = "bottom", legend.box = "horizontal") +
  scale_color_discrete(NULL) +
  guides(colour = guide_legend(nrow = 1))

ggsave("01_1.2_Sales across different Channels and Positions.png", width = 12, height = 5, dpi = 600)



# TASK 1.3
# conversion times

mean_conversion_time <- sdat_fact %>%
  group_by(Newcustomer) %>%
  summarise(mean_conversion_time = mean(TimeToConvert))

sdat_fact %>%
  group_by(Newcustomer, Channel) %>%
  summarise(mean_conversion_time = mean(TimeToConvert)) %>%
  ggplot(aes(Channel, mean_conversion_time, fill = Channel)) +
  geom_col() +
  geom_label(aes(label = number(mean_conversion_time, accuracy = 1)), vjust = -.5, show.legend = F) +
  facet_wrap(~Newcustomer, nrow = 2) +
  coord_cartesian(ylim = c(0,2300)) +
  labs(x = "Channel",
       y = "Average Conversion Time per Channel [in hours]",
       title = "Converstion Times across different Channels and Customer Types [in hours]",
       subtitle = "Lower conversion times for every channel for new customers, NA being the only exception",
       caption = "Source: W.M. Winters, May to June 2012") +
  theme_bw() +
  theme(axis.text.x = element_blank(),
        legend.position = "bottom", legend.box = "horizontal") +
  scale_color_discrete(NULL) +
  guides(colour = guide_legend(nrow = 1))

ggsave("01_1.3_New Customer Converstion Times.png", width = 12, height = 6, dpi = 600)
```

```r
# spending behavior

mean_spending <- sdat_fact %>%
  group_by(Newcustomer) %>%
  summarise(mean_spending = mean(Saleamount))

sdat_fact %>%
  group_by(Newcustomer, Channel) %>%
  summarise(mean_sales = mean(Saleamount)) %>%
  ggplot(aes(Channel, mean_sales, fill = Channel)) +
  geom_col() +
  geom_label(aes(label = dollar(mean_sales, accuracy = 1)), vjust = -.5, show.legend = F) +
  facet_wrap(~Newcustomer, nrow = 2) +
  coord_cartesian(ylim = c(0,700)) +
  labs(x = "Channel",
       y = "Aggregated Sales [in USD]",
       title = "Sales across different Channels and Customer Types",
       subtitle = "Higher Sales for every channel for new customers, NA being the only exception",
       caption = "Source: W.M. Winters, May to June 2012") +
  scale_y_continuous(labels = dollar) +
  theme_bw() +
  theme(axis.text.x = element_blank(),
        legend.position = "bottom", legend.box = "horizontal") +
  scale_color_discrete(NULL) +
  guides(colour = guide_legend(nrow = 1))

ggsave("01_1.3_New Customer Sales.png", width = 12, height = 6, dpi = 600)



# different channels as ORIGINATOR
sdat_fact %>%
  filter(Positionname == "ORIGINATOR") %>%
  group_by(Newcustomer, Channel) %>%
  count() %>%
  ggplot(aes(Channel, n, fill = Channel)) +
  geom_bar(stat = "identity") +
  geom_label(aes(label = n), vjust = -0.5) +
  facet_wrap(~Newcustomer, nrow = 2) +
  coord_cartesian(ylim = c(0,650)) +
  labs(x = "Channel",
       y = "Number of Touchpoints",
       title = "Deep-Dive Originators | Touchpoints per Channel, split by New and Old Customers",
       subtitle = "More New Customers originated from Display Advertising, Search Engines and Affilitate
Marketing",
       caption = "Source: W.M. Winters, May to June 2012") +
  theme_bw() +
  theme(axis.text.x = element_blank(),
        legend.position = "bottom", legend.box = "horizontal") +
  scale_color_discrete(NULL) +
  guides(colour = guide_legend(nrow = 1))

ggsave("01_1.3_New Customers Originators.png", width = 8, height = 6, dpi = 600)


# different channels as CONVERTER

sdat_fact %>%
  filter(Positionname == "CONVERTER" | Positionname == "ORIGINATOR") %>%
  group_by(Newcustomer, Channel, Positionname) %>%
  count() %>%
  ggplot(aes(Channel, n, fill = Channel)) +
  geom_bar(stat = "identity") +
  geom_label(aes(label = n), vjust = -0.5, show.legend = F) +
  facet_wrap(~Newcustomer + Positionname, nrow = 2) +
  coord_cartesian(ylim = c(0,800)) +
  labs(x = "Channel",
       y = "Number of Touchpoints",
       title = "Touchpoints per Channel, split by New and Old Customers",
       subtitle = "Converters: Considerable more New Customers converted via Affiliate Marketing and Display
Advertising \nOriginators: More New Customers originated from Display Advertising, Search Engines and
Affilitate Marketing",
       caption = "Source: W.M. Winters, May to June 2012") +
  theme_bw() +
  theme(axis.text.x = element_blank(),
        legend.position = "bottom", legend.box = "horizontal") +
  scale_color_discrete(NULL) +
  guides(colour = guide_legend(nrow = 1))
```

```r
ggsave("01_1.3_New Customer Converters and Originators.png", width = 12, height = 6, dpi = 600)

## PART II: -----

# TASK 2.1
# FIRST CLICK ATTRIBUTION

attribution_results <- sdat_fact %>%
  filter(Position == 0) %>%
  group_by(Channel) %>%
  summarise(first_click_sales = sum(Saleamount)) %>%
  arrange(desc(first_click_sales)) %>%
  mutate(share_first = (first_click_sales / sum(first_click_sales))*100)

# LAST CLICK ATTRIBUTION

attribution_results <- sdat_fact %>%
  group_by(Orderid) %>%
  arrange(desc(Orderid, Position)) %>%
  slice(1) %>%
  group_by(Channel) %>%
  summarise(last_click_sales = sum(Saleamount)) %>%
  arrange(desc(last_click_sales)) %>%
  mutate(share_last = (last_click_sales / sum(last_click_sales))*100) %>%
  left_join(attribution_results)

# EVEN ATTRIBUTION

attribution_results <- sdat_fact %>%
  group_by(Orderid) %>%
  add_tally() %>%
  mutate(rev_share = Saleamount/n) %>%
  group_by(Channel) %>%
  summarise(even_attribution_sales = sum(rev_share)) %>%
  arrange(desc(even_attribution_sales)) %>%
  mutate(share_even = (even_attribution_sales / sum(even_attribution_sales))*100) %>%
  left_join(attribution_results)


# graph
l1 <- c("Even Attribution", "First Click Attribution", "Last Click Attribution")
names(l1) <- c("even_attribution_sales", "first_click_sales", "last_click_sales")

attribution_results %>%
  pivot_longer(cols = c(even_attribution_sales, last_click_sales, first_click_sales),
               names_to = "attribution",
               values_to = "value") %>%
  mutate(value = round(value),1) %>%
  ggplot(aes(Channel, value, fill = Channel)) +
  geom_col() +
  geom_label(aes(label = dollar(value), accuracy = .1), vjust = -.5, show.legend = F) +
  facet_wrap(~attribution, nrow = 3,
             labeller = labeller(attribution = l1)) +
  coord_cartesian(ylim = c(0,380000)) +
  labs(x = "Channel",
       y = "Sales [in USD]",
       title = "Comparison between Attribution Strategies [in USD]",
       subtitle = "Depending on the strategy, different channels in the customer journey are prioritized",
       caption = "Source: W.M. Winters, May to June 2012") +
  scale_y_continuous(labels = dollar) +
  theme_bw() +
  theme(axis.text.x = element_blank(),
        legend.position = "bottom", legend.box = "horizontal") +
  scale_color_discrete(NULL) +
  guides(colour = guide_legend(nrow = 1))

ggsave("01_2.1_Comparison between Attribution Strategies.png", width = 12, height = 6, dpi = 600)
```

```r
# CASE FOR EVEN ATTRIBUTION MODEL
# check relationship between revenue and channel

sdat_fact %>%
  group_by(Orderid) %>%
  add_tally() %>%
  mutate(rev_share = Saleamount/n) %>%
  group_by(Channel) %>%
  ggplot(aes(Position, rev_share, colour = Channel)) +
  geom_point(alpha = .1) +
  geom_smooth() +
  facet_wrap(~Channel) +
  theme_bw()


# FIXME: For what? Title etc.
sdat_fact %>%
  group_by(Orderid) %>%
  add_tally() %>%
  mutate(rev_share = Saleamount/n) %>%
  group_by(Channel) %>%
  ggplot(aes(Position, rev_share, fill = Channel)) +
  geom_violin() +
  labs(title = "",
       subtitle = "Low") +
  facet_wrap(~Channel) +
  theme_bw() +

  theme(legend.position = "bottom", legend.box = "horizontal") +
  scale_color_discrete(NULL) +
  guides(colour = guide_legend(nrow = 1))


# TASK 2.3: DEVELOP YOUR OWN ATTRIBUTION MODEL

q <- seq(0.1, 0.5, by = 0.05)


run.summary <- function(q) {
  sdat %>%
    group_by(Orderid) %>%
    add_tally() %>%
    mutate(Position = Position + 1,
           # heuristic based on:
           share = case_when(n == 2 ~ 0.5, # shares equal 50% if only 2 positions (min)
                             Position == 1 ~ q, # otherwise first position = 35%
                             Position == n ~ q, # and last position = 35%
                             TRUE ~ (1-2*q)/(n-2)), # rest shares 30%
           rev_share = Saleamount * share) %>%
    group_by(Channel) %>%
    summarise(position_based_attribution = sum(rev_share)) %>%
    mutate(position_based_attribution = round(position_based_attribution,1)) %>%
    arrange(desc(position_based_attribution)) %>%
    mutate(k = q)
}

# create df with all data
df <- map(q, ~run.summary(.x)) %>% reduce(bind_rows)

df %>%
  ggplot(aes(k, position_based_attribution, color = Channel)) +
  geom_line() +
  geom_point() +
  facet_wrap(~Channel) +
  scale_y_continuous(labels = dollar) +
  labs(x = "Channel Weight (per Touchpoint & Order)",
       y = "Sales per Channel",
       title = "Optimization of Position-Based Weights",
       subtitle = "Overall the revenue increase of AM and SE is balanced by a decrease in revenue for DA") +
  theme_bw() +
  theme(legend.position = "bottom", legend.box = "horizontal") +
  scale_color_discrete(NULL) +
  guides(colour = guide_legend(nrow = 2))

ggsave("01_2.3_Attribution Optimization.png", width = 12, height = 8, dpi = 300)
```

```r
# -------

s1 <- sdat_fact %>%
  group_by(Orderid) %>%
  add_tally() %>%
  mutate(rev_share = Saleamount/n) %>%
  group_by(Channel, Positionname) %>%
  summarise(even_attribution_sales = sum(rev_share)) %>%
  arrange(desc(even_attribution_sales)) %>%
  mutate(even_attribution_sales = round(even_attribution_sales),1) %>%
  ggplot(aes(Channel, even_attribution_sales, fill = Channel)) +
  geom_col() +
  geom_label(aes(label = dollar(even_attribution_sales, accuracy = 1)), vjust = -.2,
             show.legend = F) +
  facet_wrap(~Positionname, nrow = 4) +
  coord_cartesian(ylim = c(0, 160000)) +
  labs(x = "Channel",
       y = "Sales",
       title = "Even Attribution Model | Detailed View",
       subtitle = "Depending on the strategy, different channels in the customer journey are prioritized")+
       # caption = "Source: W.M. Winters, May to June 2012") +
  scale_y_continuous(labels = dollar) +
  theme_bw() +
  theme(legend.position = "none",
        axis.text.x = element_blank())
  # theme(axis.text.x = element_blank(),
  #       legend.position = "bottom", legend.box = "horizontal") +
  # scale_color_discrete(NULL) +
  # guides(colour = guide_legend(nrow = 1))

ggsave("01_2.2_Even Attribution Model - Detailed View.png", width = 8, height = 6, dpi = 600)

k <- 0.3 # efficient weight

# position based attribution
s2 <- sdat %>%
  # filter(!Channel %in% c("Uncategorized", NA, "Other")) %>% # does not work, sometimes negative values
because of the share formula
  group_by(Orderid) %>%
  add_tally() %>%
  mutate(Position = Position + 1,
         # heuristic based on:
         share = case_when(n == 2 ~ 0.5, # shares equal 50% if only 2 positions (min)
                           Position == 1 ~ k, # otherwise first position = 30%
                           Position == n ~ k, # and last position = 30%
                           TRUE ~ (1-2*k)/(n-2)), # rest shares 40%
         rev_share = Saleamount * share)  %>%

  group_by(Channel, Positionname) %>%
  summarise(position_based_attribution = sum(rev_share)) %>%
  mutate(position_based_attribution = round(position_based_attribution,1)) %>%
  arrange(desc(position_based_attribution)) %>%
  ggplot(aes(Channel, position_based_attribution, fill = Channel)) +
  geom_col() +
  geom_label(aes(label = dollar(position_based_attribution, accuracy = 1)), vjust = -.2,
             show.legend = F) +
  coord_cartesian(ylim = c(0, 160000)) +
  # additionally one could include Newcustomer
  facet_wrap(~Positionname, nrow = 4) +

  labs(x = "Channel",
       y = "Sales",
       title = "Position-Based Attribution Strategy | Detailed View",
       subtitle = "Detailed view shows that PB shows a budget allocation in favor of both, converters and
originators",
       caption = "Source: W.M. Winters, May to June 2012") +
  theme_bw()+
  theme(axis.title.y=element_blank(),
        axis.text.y=element_blank(),
        axis.ticks.y=element_blank())



s1 + s2 +
  theme(axis.text.x = element_blank(),
        legend.position = "right", legend.box = "horizontal") +
  scale_color_discrete(NULL) +
  guides(colour = guide_legend(nrow = 1))

ggsave("01_2.3_Attribution Strategy Comparison.png", width = 12, height = 6, dpi = 300)
```

```r
sdat %>%
  group_by(Orderid) %>%
  add_tally() %>%
  mutate(Position = Position + 1,
         # heuristic based on:
         share = case_when(n == 2 ~ 0.5, # shares equal 50% if only 2 positions (min)
                           Position == 1 ~ k, # otherwise first position = 35%
                           Position == n ~ k, # and last position = 35%
                           TRUE ~ (1-2*k)/(n-2)), # rest shares 30%
         rev_share = Saleamount * share)  %>%

  group_by(Channel) %>%
  summarise(position_based_attribution = sum(rev_share)) %>%
  mutate(position_based_attribution = round(position_based_attribution,1)) %>%
  arrange(desc(position_based_attribution)) %>%
  ggplot(aes(Channel, position_based_attribution, fill = Channel)) +
  geom_col() +
  geom_label(aes(label = dollar(position_based_attribution, suffix = "k", scale = 1/1000)), vjust = -.5,
show.legend = F) +
  coord_cartesian(ylim = c(0, 300000)) +
  # additionally one could include Newcustomer
  # facet_wrap(~Positionname, nrow = 4) +
  scale_y_continuous(labels = dollar) +
  labs(x = "Channel",
       y = "Sales",
       title = "Position-Based Attribution Strategy | Detailed View",
       subtitle = "Detailed view shows that PB shows a budget allocation in favor of both, converters and
originators",
       caption = "Source: W.M. Winters, May to June 2012") +
  theme_bw() +
  theme(axis.text.x = element_blank(),
        legend.position = "bottom", legend.box = "horizontal") +
  scale_color_discrete(NULL) +
  guides(colour = guide_legend(nrow = 1))



# TASK 2.4
# Same analysis as before split by type of customer

sdat_fact %>%
  group_by(Newcustomer, Positionname, Channel) %>%
  count() %>%
  ggplot(aes(Channel, n, fill = Channel)) +
  geom_col() +
  facet_wrap(~Positionname + Newcustomer, nrow = 2) +
  labs(x = "Position Classification",
       y = "Number of Touchpoints",
       title = "Difference in Touchpoint Attribution by Type of Customer",
       subtitle = "Dataset includes more new customers than existing customers, which illustrated by the
different channels",
       caption = "Source: W.M. Winters, May to June 2012") +
  theme_bw() +
  theme(axis.text.x = element_blank(),
        legend.position = "bottom", legend.box = "horizontal") +
  scale_color_discrete(NULL) +
  guides(colour = guide_legend(nrow = 1))

sdat_fact %>%
  group_by(Newcustomer, Position, Channel) %>%
  count() %>%
  ggplot(aes(Position, n, fill = Channel)) +
  geom_col() +
  facet_wrap(~Channel + Newcustomer, nrow = 3) +
  labs(x = "Position",
       y = "Occurences",
       title = "Occurence of Positions between Type of Customers and Channel",
       subtitle = "New customers have a similar position distribution, channel prioritization can be kept",
       caption = "Source: W.M. Winters, May to June 2012") +
  theme_bw() +
  theme(legend.position = "bottom", legend.box = "horizontal") +
  scale_color_discrete(NULL) +
  guides(colour = guide_legend(nrow = 1))

ggsave("01_2.4_Position Comparison between Channels and Type of Customers.png", width = 12, height = 6, dpi =
300)
```

```r
# TASK 2.5. BUDGET ALLOCATION

budgetallocation <- sdat %>%
  # filter(!Channel %in% c("Uncategorized", NA, "Other")) %>% # does not work, sometimes negative values
because of the share formula
  group_by(Orderid) %>%
  add_tally() %>%
  mutate(Position = Position + 1,
         # heuristic based on:
         share = case_when(n == 2 ~ 0.5, # shares equal 50% if only 2 positions (min)
                           Position == 1 ~ k, # otherwise first position = 35%
                           Position == n ~ k, # and last position = 35%
                           TRUE ~ (1-2*k)/(n-2)), # rest shares 30%
         rev_share = Saleamount * share)  %>%

  group_by(Channel) %>%
  summarise(position_based_attribution = sum(rev_share)) %>%
  mutate(position_based_attribution = round(position_based_attribution,1)) %>%
  arrange(desc(position_based_attribution)) %>%

  mutate(share = round((position_based_attribution / sum(position_based_attribution))*100,2))

budgetallocation_adj <- budgetallocation %>%
  filter(Channel %in% c("Display Advertising", "Affiliate Marketing", "Search Engine")) %>%
  mutate(share_adj = share + (share / (46.0+31.9+19.4)) * (1.01+0.92+0.47+0.2+0.07+0.01))

# Draw final budget allocation
l2 <- c("(a) With Ineffective Channels", "(b) Without Ineffective Channels", "Even Attribution")
names(l2) <- c("share", "share_adj", "share_even")

budgetallocation %>%
  select(Channel, share) %>%
  left_join(budgetallocation_adj %>%
              select(Channel, share_adj)) %>%
  left_join(attribution_results) %>%
  select(Channel, share, share_adj, share_even) %>%
  pivot_longer(cols = c(share, share_adj, share_even),
               names_to = "attribution",
               values_to = "value") %>%
  mutate(value = value/100) %>%
  ggplot(aes(Channel, value, fill = Channel)) +
  geom_col() +
  geom_label(aes(label = percent(value, accuracy = .1)), vjust = -.5, show.legend = F) +
  facet_wrap(~attribution, nrow = 3,
             labeller = labeller(attribution = l2)) +
  coord_cartesian(ylim = c(0,1)) +
  labs(x = "Channel",
       y = "Ratio [in percent]",
       title = "Comparison between Attribution Strategies [in percent]",
       subtitle = "Posiiton-based strategy (b) dominates other strategies because it concentrates on more
successful channels (AM)",
       caption = "Source: W.M. Winters, May to June 2012") +
  scale_y_continuous(labels = percent) +
  theme_bw() +
  theme(axis.text.x = element_blank(),
        legend.position = "bottom", legend.box = "horizontal") +
  scale_color_discrete(NULL) +
  guides(colour = guide_legend(nrow = 1))

ggsave("01_2.5_Final Comparison in Attribution Strategies.png", width = 12, height = 6)
```

```r
# APPENDIX:

# Appendix 1:

sdat_fact %>%
  group_by(Positionname, Channel) %>%
  count() %>%
  arrange(desc(n)) %>%
  ungroup() %>%
  group_by(Positionname) %>%
  mutate(share = round((n / sum(n)),4)) %>%

  ggplot(aes(Channel, share, fill = Channel), colour = "white") +
  geom_bar(stat = "identity") +
  geom_label(aes(label = percent(share, accuracy = 0.1)), vjust = -0.3, show.legend = F) +
  facet_wrap(~Positionname, nrow = 3) +
  coord_cartesian(ylim = c(0,1)) +
  labs(x = "Position",
       y = "Channel Touchpoints [in percent]",
       title = "Touchpoints per Channel and Position [in percent]",
       subtitle = "Search Engines provide strong support for initial clicks, \nAffiliate Marketing serves a
strong role in converting and rosting (retargeting) \nDisplay Advertisements are strong across all categories",
       caption = "Source: W.M. Winters, May to June 2012") +
  scale_y_continuous(labels = percent) +
  theme_bw() +
  theme(axis.text.x = element_blank(),
        legend.position = "bottom", legend.box = "horizontal") +
  scale_color_discrete(NULL) +
  guides(colour = guide_legend(nrow = 1))


ggsave("01_A1_Touchpoints per Channel and Position in percent.png", width = 12, height = 6)



# Appendix 2:

# conversion time comparison
sdat_fact %>%
  ggplot(aes(Position, TimeToConvert, fill = Newcustomer)) +
  geom_boxplot(outlier.alpha = .1) +

  geom_hline(yintercept = mean_conversion_time[[2,"mean_conversion_time"]],
             linetype = "dashed", colour = "#00BFC4", size = 1) +
  annotate(geom = "label", x = 9.5, y = mean_conversion_time[[2,"mean_conversion_time"]] + 200,
           label = round(mean_conversion_time[[2,"mean_conversion_time"]],1), colour = "#00BFC4") +

  geom_hline(yintercept = mean_conversion_time[[1,"mean_conversion_time"]],
             linetype = "dashed", colour = "#F8766D", size = 1) +
  annotate(geom = "label", x = 9.5, y = mean_conversion_time[[1,"mean_conversion_time"]] + 200,
           label = round(mean_conversion_time[[1,"mean_conversion_time"]],1), colour = "#F8766D") +

  facet_wrap(~ Newcustomer) +
  labs(x = "Position",
       y = "Time to Convert [in hours]",
       title = "Difference in Conversion Times by Customer Type",
       subtitle = "New Customers show significantly faster conversion times across all positions",
       caption = "Source: W.M. Winters, May to June 2012") +
  theme_bw() +
  theme(legend.position = "bottom", legend.box = "horizontal") +
  scale_color_discrete(NULL) +
  guides(colour = guide_legend(nrow = 1))

ggsave("01_A2_Difference in Conversion Times by Customer Type.png", width = 12, height = 4, dpi = 600)
```

```r
# Appendix 3:

# spending  comparison
sdat_fact %>%
  ggplot(aes(Position, Saleamount, fill = Newcustomer)) +
  geom_boxplot(outlier.alpha = .1) +

  geom_hline(yintercept = mean_spending[[2,"mean_spending"]],
             linetype = "dashed", colour = "#00BFC4", size = 1) +
  annotate(geom = "label", x = 9.5, y = mean_spending[[2,"mean_spending"]] + 100,
           label = round(mean_spending[[2,"mean_spending"]],1), colour = "#00BFC4") +

  geom_hline(yintercept = mean_spending[[1,"mean_spending"]],
             linetype = "dashed", colour = "#F8766D", size = 1) +
  annotate(geom = "label", x = 9.5, y = mean_spending[[1,"mean_spending"]] - 100,
           label = round(mean_spending[[1,"mean_spending"]],1), colour = "#F8766D") +

  facet_wrap(~ Newcustomer) +
  labs(x = "Position",
       y = "Sales",
       title = "Difference in Sales by Customer Type and Channel",
       subtitle = "New Customers show significantly higher sales across all positions",
       caption = "Source: W.M. Winters, May to June 2012") +
  theme_bw() +
  theme(legend.position = "bottom", legend.box = "horizontal") +
  scale_color_discrete(NULL) +
  guides(colour = guide_legend(nrow = 1))

ggsave("01_A3_Difference in Sales by Customer Type and Channel.png", width = 12, height = 4, dpi = 600)


# Appendix 4:
# APPENDIX additional difference by channel
sdat_fact %>%
  ggplot(aes(Position, TimeToConvert, fill = Newcustomer)) +
  geom_boxplot(outlier.alpha = .1) +

  facet_wrap(~ Channel + Newcustomer) +
  labs(x = "Position",
       y = "Time to Convert [in hours]",
       title = "Difference in Conversion Times by Customer Type and Channel",
       subtitle = "New Customers show significantly faster conversion times across all positions",
       caption = "Source: W.M. Winters, May to June 2012") +
  theme_bw() +
  theme(legend.position = "bottom", legend.box = "horizontal") +
  scale_color_discrete(NULL) +
  guides(colour = guide_legend(nrow = 1))

ggsave("01_A4_Difference in Conversion Times by Customer Type and Channel.png", width = 12, height = 10, dpi =
600)
```