7Shifts Data Analysis

## Objective

Find 1-3 interesting things you find about the conversion rates or other trends in the data. Data provided consists of two files: companies and trials.

## Approach

* Load packages needed
* Load data (companies and trials)
* Get summary statistics for data
* Perform data cleaning if needed
* Find any trends or insights and add write-up as needed

## Load packages

packages <- c('ggplot2', 'corrplot','tidyverse','scales','dplyr', 'lubridate','gridExtra','grid')  
for (package in packages) {  
 if (!require(package, character.only=T, quietly=T)) {  
 install.packages(package)  
 library(package, character.only=T)  
 }  
}

## corrplot 0.90 loaded

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✓ tibble 3.1.3 ✓ dplyr 1.0.7  
## ✓ tidyr 1.1.3 ✓ stringr 1.4.0  
## ✓ readr 1.4.0 ✓ forcats 0.5.1  
## ✓ purrr 0.3.4

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

##   
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':  
##   
## discard

## The following object is masked from 'package:readr':  
##   
## col\_factor

##   
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':  
##   
## date, intersect, setdiff, union

##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

## Load data

trials <- read\_csv("trials.csv")

## Warning: Missing column names filled in: 'X1' [1]

##   
## ── Column specification ────────────────────────────────────────────────────────  
## cols(  
## X1 = col\_double(),  
## trial\_date = col\_date(format = ""),  
## customer\_name = col\_character(),  
## utm = col\_character(),  
## display\_price = col\_double(),  
## company\_id = col\_double()  
## )

companies <- read\_csv("companies.csv")

## Warning: Missing column names filled in: 'X1' [1]

##   
## ── Column specification ────────────────────────────────────────────────────────  
## cols(  
## X1 = col\_double(),  
## company\_id = col\_double(),  
## signup\_date = col\_date(format = ""),  
## cancellation\_date = col\_date(format = ""),  
## name = col\_character()  
## )

## Review data

## Summary and preview company information  
summary(companies)

## X1 company\_id signup\_date cancellation\_date   
## Min. : 0.0 Min. : 101.0 Min. :1920-03-02 Min. :2019-07-05   
## 1st Qu.:249.8 1st Qu.: 350.8 1st Qu.:2020-01-16 1st Qu.:2020-01-22   
## Median :499.5 Median : 600.5 Median :2020-08-24 Median :2020-07-04   
## Mean :499.5 Mean : 600.5 Mean :2020-04-16 Mean :2020-07-05   
## 3rd Qu.:749.2 3rd Qu.: 850.2 3rd Qu.:2021-02-18 3rd Qu.:2020-12-24   
## Max. :999.0 Max. :1100.0 Max. :2021-09-06 Max. :2021-06-20   
## NA's :702   
## name   
## Length:1000   
## Class :character   
## Mode :character   
##   
##   
##   
##

glimpse(companies)

## Rows: 1,000  
## Columns: 5  
## $ X1 <dbl> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15…  
## $ company\_id <dbl> 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 11…  
## $ signup\_date <date> 2017-03-18, 2017-06-21, 2017-09-26, 2017-11-05, 201…  
## $ cancellation\_date <date> 2019-10-29, 2019-09-01, 2019-11-23, 2020-07-08, 202…  
## $ name <chr> "Miller-Alvarez", "Bishop-Chapman", "Schmidt, Shah a…

## Summary and preview trial information  
summary(trials)

## X1 trial\_date customer\_name utm   
## Min. : 0 Min. :1920-02-24 Length:10000 Length:10000   
## 1st Qu.:2500 1st Qu.:2018-10-19 Class :character Class :character   
## Median :5000 Median :2019-10-24 Mode :character Mode :character   
## Mean :5000 Mean :2019-09-23   
## 3rd Qu.:7499 3rd Qu.:2020-09-16   
## Max. :9999 Max. :2021-08-30   
##   
## display\_price company\_id   
## Min. : 9.00 Min. : 101.0   
## 1st Qu.:19.00 1st Qu.: 350.8   
## Median :29.00 Median : 600.5   
## Mean :29.27 Mean : 600.5   
## 3rd Qu.:40.00 3rd Qu.: 850.2   
## Max. :49.00 Max. :1100.0   
## NA's :9000

glimpse(trials)

## Rows: 10,000  
## Columns: 6  
## $ X1 <dbl> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16…  
## $ trial\_date <date> 1920-02-24, 1921-01-04, 2017-03-11, 2017-06-14, 2017-09…  
## $ customer\_name <chr> "William Anthony", "Mark Wood", "Mary Hampton", "Robin W…  
## $ utm <chr> "organic", "organic", "google", "organic", "organic", "l…  
## $ display\_price <dbl> 21, 43, 13, 39, 12, 44, 45, 31, 47, 42, 47, 20, 32, 47, …  
## $ company\_id <dbl> 396, 790, 101, 102, 103, NA, NA, NA, NA, NA, NA, NA, NA,…

## Drop first columns  
trials[1] <- NULL  
companies[1] <- NULL

## Business information

* Conversion rate is defined as the proportion of prospective customers (trials) that made a purchase and became actual customers (companies) within a given time period.
* Trials can be joined to companies via company\_id. A trial with no company ID did not convert.

# Combine trials and companies  
trials\_company <- trials %>%  
 inner\_join(companies,by = "company\_id")

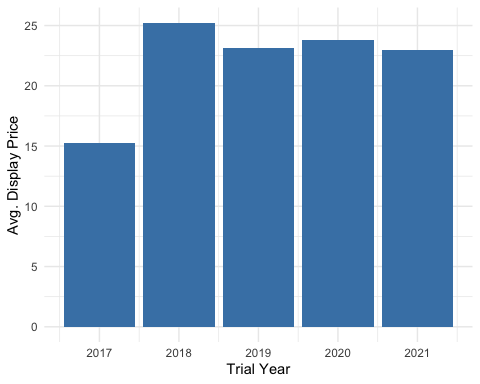
## Create trial year  
trials\_company$trial\_year <- lubridate::year(trials\_company$trial\_date)  
## Unique years  
unique(trials\_company$trial\_year)

## [1] 1920 1921 2017 2018 2019 2020 2021

#filter for years on or after 2017 since there are some odd dates  
trials\_company <- trials\_company %>%  
 filter(trial\_year >= 2017)

## Trends

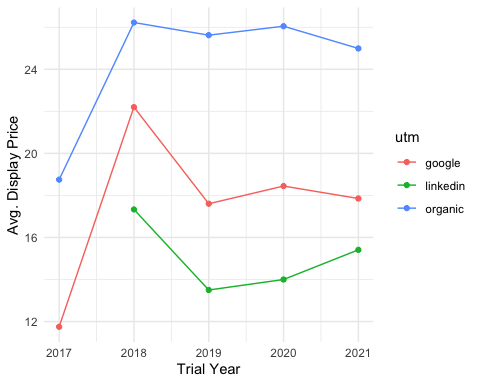
#trial year vs avg display price  
trial\_avg\_display\_price <- trials\_company %>%  
 dplyr::group\_by(trial\_year) %>%  
 dplyr::summarise(avg\_display\_price = mean(display\_price)) %>%  
 dplyr::select(trial\_year, avg\_display\_price)  
   
  
ggplot(data=trial\_avg\_display\_price, aes(x=(trial\_year), y=avg\_display\_price)) +  
 geom\_bar(stat="identity", fill="steelblue")+   
 theme\_minimal()+  
 xlab("Trial Year") +   
 ylab("Avg. Display Price")

 - Looks like the average display price had a significant increase after 2017.

#trial year vs utm vs avg display price  
trial\_utm\_avg\_display\_price <- trials\_company %>%  
 dplyr::group\_by(trial\_year, utm) %>%  
 dplyr::summarise(avg\_display\_price = mean(display\_price)) %>%  
 dplyr::select(trial\_year, utm, avg\_display\_price)

## `summarise()` has grouped output by 'trial\_year'. You can override using the `.groups` argument.

ggplot(trial\_utm\_avg\_display\_price, aes(x=(trial\_year), y=avg\_display\_price, group = utm)) +  
 geom\_line(aes(color=utm)) +  
 geom\_point(aes(color=utm)) +  
 theme\_minimal()+  
 xlab("Trial Year") +   
 ylab("Avg. Display Price")



* Organic average display price has been the highest price across all years.
* Google’s average display price took a drop from 2018 and 2019 and has remained steady since.
* Linkedin’s average display price also took a drop from 2018 to 2019 and has been steadily rising since.

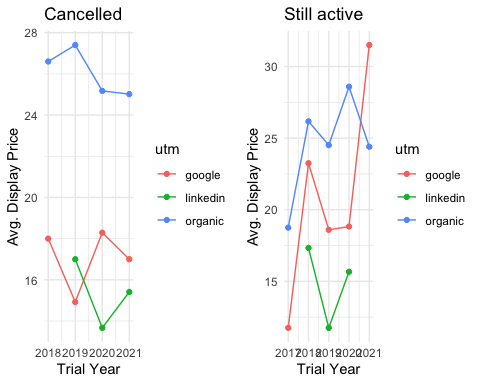
#differences between companies that have cancelled vs not cancelled  
trial\_utm\_avg\_display\_price\_other <- trials\_company %>%  
 dplyr::filter(is.na(cancellation\_date)) %>%  
 dplyr::group\_by(trial\_year, utm) %>%  
 dplyr::summarise(avg\_display\_price = mean(display\_price)) %>%  
 dplyr::select(trial\_year, utm, avg\_display\_price)

## `summarise()` has grouped output by 'trial\_year'. You can override using the `.groups` argument.

plot1 <- ggplot(trial\_utm\_avg\_display\_price\_other, aes(x=(trial\_year), y=avg\_display\_price, group = utm)) +  
 geom\_line(aes(color=utm)) +  
 geom\_point(aes(color=utm)) +  
 theme\_minimal()+  
 ggtitle("Cancelled ")+  
 xlab("Trial Year") +   
 ylab("Avg. Display Price")  
  
  
trial\_utm\_avg\_display\_price\_other1 <- trials\_company %>%  
 dplyr::filter(!is.na(cancellation\_date)) %>%  
 dplyr::group\_by(trial\_year, utm) %>%  
 dplyr::summarise(avg\_display\_price = mean(display\_price)) %>%  
 dplyr::select(trial\_year, utm, avg\_display\_price)

## `summarise()` has grouped output by 'trial\_year'. You can override using the `.groups` argument.

plot2 <- ggplot(trial\_utm\_avg\_display\_price\_other1, aes(x=(trial\_year), y=avg\_display\_price, group = utm)) +  
 geom\_line(aes(color=utm)) +  
 geom\_point(aes(color=utm)) +  
 #geom\_bar(stat="identity",position = "dodge")+   
 theme\_minimal()+  
 ggtitle("Still active")+  
 xlab("Trial Year") +   
 ylab("Avg. Display Price")  
  
  
grid.arrange(plot1, plot2, ncol=2)



* For customers that are still active
  + Linkedin data seems to have stopped after 2020. Might be a data issue. Needs more investigation
  + Google average display prices have been increasing with a dip between 2018 and 2020. Had a sharp rise after 2020
  + Organic average display prices seem to be the most costly from 2017 to 2020. Saw a drop after 2020
* For customers that have cancelled
  + Organic average display prices look to be the most expensive
  + Linkedin prices seem to be the lowest with mild fluctuations
  + Google average display prices have had mild fluctuations