CASE 1 - eLAB Business Case

10 May 2023

Authors:

- Cornelius Kötting
- Jakub Czubaszek
- Tshiamo Kgosidiile
- Ajay Rayudu
- Viktor Vazquez Krenev

Tutor: Hugo Schyns

1. Introduction

Nowadays, having a social media presence has become a vital part of many companies marketing strategies. Social media offers marketers a lot of advantages such as increased visibility, the ability to target a specific consumer segment, sharing engaging short-form content, as well as feedback and insights to optimize market offerings. In recent years, Tik Tok has become one of the biggest and most influential social media platforms with over 1 billion monthly active users. In today's cut-throat market, it is imperative for gastronomy companies to maintain relevance and adapt to innovative technologies and marketing tactics that entice and retain customers. Social media has established itself as an intrinsic part of several firms' marketing strategies. Therefore, comprehending the impact of Tik Tok on a business's performance can furnish valuable insights for enterprises striving to escalate their visibility, engagement, and ultimately boost business growth. Moreover, the COVID-19 pandemic has left a lasting impression on the gastronomy industry, causing a plethora of restaurants to rely on delivery and takeaway services. Consequently, the significance of having a robust social media presence has escalated. By leveraging platforms like Tik Tok, enterprises can display their products and services in a novel and engaging manner, simultaneously reaching out to hitherto untapped customers. Therefore we choose to evaluate and quantify the effect that operating a Tik Tok account has on the performance of companies in the gastronomy industry.

2. Data import & merge

The first step was to collect the data. The process looked as follows: For every company in the dataset, we went onto Tik Tok (https://tiktok.com) to check if they had an account and if so since when. Using this information we created a new categorical variable tt indicating whether the firm operated an account on the social media platform or not (0 = no, 1 = yes). We also created a second variable ttyear that consists of the year when the company started uploading content on Tik Tok. Furthermore, we encountered some instances of a company in the dataset not having a Tik Tok account but a subsidiary that did. In these cases we opted to set *tt* to 1, as these firms were usually holding companies owning one or more restaurant brands, where a Tik Tok account for the holding company would probably have made very little difference to the firms performance. Out of the 106 gastronomy companies, we found that 60 had an active Tik Tok account

(04.05.2023). Examples of gastronomy companies studied include *Benihana Inc*, a chain of Hibachi restaurants and *Chipotle Mexican Grill Inc*, a chain of mexican fast food restaurants. In the code below, we import the original dataset, filter it to only include firms from the gastronomy industry, import the Tik Tok data we collected and merge it all into one unified dataset.

```
data <- read.delim("/Users/cornelius/Downloads/sicdatas.txt", header = TRUE) %>%
  filter(sic == 5812)

tiktok_data <- read_excel("/Users/cornelius/Downloads/Elab_case02.xlsx", col_names = TRUE)

merged_data <- left_join(data, tiktok_data, by = c("gvkey", "conml")) %>%
  mutate(tt = ifelse(is.na(tt), tiktok_data$tt[match(gvkey, tiktok_data$gvkey)], tt))
```

2. Data preparation

Before any meaningful analysis can be conducted on the dataset at hand, the data had to be prepared and cleaned. This process includes checking for missing values, identifying descriptive statistics such as min, max and mean to identify possible outliers and check consistency of the data. Thus, we calculated the above mentioned descriptive statistics for every variable present in the dataset and took corrective action if necessary.

2.1 *tt*

As the dataset is in the long-format, meaning that for every business there is one observation (row) per year. Therefore we realized that we did not need the variable *ttyear*, as this information could be simply expressed using the value of *tt* between the different observations. So if *tt* is 0 in 2019 and 1 in 2020, we know that the company started actively using Tik Tok in 2020. The above described transformation can be seen in the code block below.

```
merged_data <- merged_data %>%
  mutate(
    tt = ifelse(fyear < ttyear & !is.na(ttyear), 0, tt)
  )
merged_data$ttyear<-NULL
```

2.2 Missing values

Before any meaningful analysis can be conducted, we need to check for missing values in the dataset and replace them if necessary, as they can heavily skew analysis results. The table below shows the number of missing values as well as some other statistics per variable in the dataset. As can be seen, data is missing for most variables, which we will adress below.

```
diagnose(merged_data)%>%flextable()
```

variables	types	missing_count	missing_percent	unique_count	unique_rate
gvkey	numeric	0	0.000000	106	0.130221130
datadate	character	0	0.000000	91	0.111793612
fyear	integer	0	0.000000	13	0.015970516
indfmt	character	0	0.000000	1	0.001228501
consol	character	0	0.000000	1	0.001228501
popsrc	character	0	0.000000	1	0.001228501
datafmt	character	0	0.000000	1	0.001228501
curcd	character	0	0.000000	1	0.001228501
curncd	character	0	0.000000	6	0.007371007
act	numeric	38	4.668305	763	0.937346437
at	numeric	35	4.299754	770	0.945945946
ceq	numeric	35	4.299754	778	0.955773956
csho	numeric	63	7.739558	689	0.846437346
dltt	numeric	42	5.159705	655	0.804668305
emp	numeric	110	13.513514	571	0.701474201
lct	numeric	38	4.668305	767	0.942260442
ni	numeric	38	4.668305	774	0.950859951
sale	numeric	38	4.668305	749	0.920147420
xad	numeric	184	22.604423	507	0.622850123
xsga	numeric	53	6.511057	752	0.923832924
costat	character	0	0.000000	2	0.002457002
prcc_f	numeric	125	15.356265	653	0.802211302
busdesc	character	0	0.000000	106	0.130221130
conml	character	0	0.000000	106	0.130221130
sic	integer	0	0.000000	1	0.001228501
weburl	character	0	0.000000	98	0.120393120
tt	numeric	0	0.000000	2	0.002457002

To adress the above issue, the code below, loops through all variables in the dataset, checks for missing values and replaces them with the mean of all company observations if necessary. Furthermore, some missing data could be recovered from the internet. For example, we found up to date employee (*emp*) data in filings on the website of the SEC (https://sec.gov). Additionally, advertising expenditures (*xad*) were missing for some firms. As per the assignment instructions, the disclosure of this data is only mandatory if expenditures are material. Thus, we replaced all missing values with 0, as it is reasonable to assume that the companies that did not report the figures did not incur material advertising expenditure. Moreover, there was a number of firms where the information regarding shares (*csho* & *prcc_f*) was missing completely and could not be found on the internet. These companies where therefore removed from the dataset and will not be considered for further analysis. Lastly, some companies present in the dataset only invest in restaurant companies and therefore a lot of the variables representing firm characteristics such as sales are 0. In order to keep these data points from skewing our analysis, we opted to remove these companies to our dataset. An example of such an investment company is *Grey Fox Holdings Corporation*.

```
#Selecting variables for removal of missing values
var_names <- c("act", "at", "ceq", "csho", "dltt", "emp", "lct", "ni", "sale", "xad", "xsga", "prc
```

```
# Loop through the variable names
for (var in var_names) {
  # Calculate the mean for each group of gvkey
  df_mean <- merged_data %>%
    group_by(gvkey) %>%
    summarize(mean_var = mean(!!sym(var), na.rm = TRUE))

  # Join the mean to the original data frame and replace missing values
  merged_data <- merged_data %>%
    left_join(df_mean, by = "gvkey") %>%
    mutate(!!sym(var) := ifelse(is.na(!!sym(var)), mean_var, !!sym(var))) %>%
    select(-mean_var)
}

#Replacing NA values with 0
merged_data$xad <- ifelse(is.na(merged_data$xad), 0, merged_data$xad)

merged_data[448:452,"emp"]<-1.065
merged_data[463:465,"emp"]<-5
merged_data[487,"emp"]<-30
merged_data[488,"emp"]<-37

#Selecting firms to remove from dataset
firms_to_remove <- c("bowmo", "cordia", "Grey Fox", "DBUB", "Cannae Holdings Inc", "Fat Brands Inc

#Removing above selected firms
merged_data <- merged_data[!grepl(paste(firms_to_remove, collapse = "|"), merged_data$conml, ignor

#Checking if any missing values remain
cat("Number of missing variables:", sum(is.na(merged_data)))
```

```
## Number of missing variables: 0
```

2.3 Creating new variables

After ensuring the completeness and correctness of the dataset at hand, we created new variables for model building. The variables and the corresponding code can be seen below:

- Measures of Firm Performance:
  - *Tobin's q*: Equals the market value of a company divided by its assets' replacement cost.
  - *Return on Sales (ROS)*: Firm's net income (in $ millions) divided by its gross sales (in $ millions).
- Managerial Investment Decisions:
  - *Advertising intensity*: Advertising expenditures / Sales
  - *Marketing investments*: Marketing expenditures / total assets
  - *Investment*: Yes/no (Tiktok presence)

- Firm Characteristics:
  - *Firm size*
  - *Financial liquidity*: Current assets / current liabilities
  - *Firm leverage*: Debt / assets
- Industry Characteristics:
  - *Industry dummies*: Dummy variables according to the 4- digit SIC codes

```
#Creating new variables of interest
merged_data$tobinsq <- (merged_data$at+(merged_data$csho*merged_data$prcc_f)-merged_data$ceq)/merg
merged_data$ros<- merged_data$ni/merged_data$sale
merged_data$ai<- merged_data$xad/merged_data$sale
merged_data$mi<-merged_data$xsga/merged_data$at
merged_data$fs<-log(merged_data$emp)
merged_data$fliq<- merged_data$act/merged_data$lct
merged_data$flev<-merged_data$dltt/merged_data$at

#Showing structure of updated dataset
variables_of_interest = c("gvkey", "tobinsq", "ros", "ai", "mi", "fs", "fliq", "flev", "conml", "f
vi <- merged_data[, variables_of_interest]
str(vi) %>% kable()
```

```
## 'data.frame':    701 obs. of  11 variables:
##  $ gvkey  : num  2163 2163 3007 3007 3007 ...
##  $ tobinsq: num  0.95 1.27 2.07 2.43 2.73 ...
##  $ ros    : num  0.00407 0.01135 0.05108 0.0535 0.0574 ...
##  $ ai     : num  0.0208 0.0205 0.029 0.0284 0.0291 ...
##  $ mi     : num  0.1158 0.1399 0.0938 0.1053 0.0988 ...
##  $ fs     : num  1.9 1.93 4.1 4.06 4 ...
##  $ fliq   : num  0.453 0.814 0.546 0.485 0.509 ...
##  $ flev   : num  0.0258 0 0.3385 0.4094 0.5371 ...
##  $ conml  : chr  "Benihana Inc" "Benihana Inc" "Brinker International Inc." "Brinker Internatic
##  $ fyear  : int  2010 2011 2011 2012 2013 2014 2015 2016 2017 2018 ...
##  $ tt     : num  0 0 0 0 0 0 0 0 0 0 ...
```

2.4 Validating new variables

Before the above created variables can be used for further analysis and model construction, validity needs to be ensured. This includes checking for missing values and outliers, as well as checking if the values are consistent with common sense.

```
cat("Number of missing variables:", sum(is.na(vi)))
```

```
## Number of missing variables: 0
```

```
diagnose_numeric(vi)%>%flextable()
```

variables	min	Q1	mean	median	Q3	max	zero	n
gvkey	2,163.000000000	18,839.000000000	77,049.52496434	31,846.00000000	164,255.00000000	279,170.0000000	0	
tobinsq	0.442159002	1.376855861	3.07238544	1.86446366	3.14062310	49.3545615	0	
ros	-5.723936277	-0.009874390	-0.02132042	0.03078996	0.06693068	0.4077962	0	
ai	-0.001000204	0.005515476	0.03034174	0.01757241	0.03438598	0.7997272	105	
mi	0.009752410	0.092807741	0.19407376	0.14007626	0.21348551	2.0195652	0	
fs	-3.611918413	1.119394828	2.26794436	2.13888900	3.52636052	6.2897156	0	
fliq	0.016892978	0.545758023	1.12210893	0.88211325	1.30126728	13.5837838	0	
flev	0.000000000	0.119140640	0.42872303	0.32591876	0.57111654	3.8524893	58	
fyear	2,010.000000000	2,013.000000000	2,015.84450785	2,016.00000000	2,018.00000000	2,022.0000000	0	
tt	0.000000000	0.000000000	0.08701854	0.00000000	0.00000000	1.0000000	640	

As can be seen in the output of the tables above, some variables encompass a significant number of outliers. In order to address this issue, we opted to winsorize these variables. Winsorizing is a data transformation technique that involves replacing outliers with less extreme values to prepare the data for further analysis. We selected a trim level of 5%, meaning that the lower 5% of the data will be replaced by the 5th percent quantile and the upper 5% of data by the 95th percent quantile. 0.05 is commonly used as a trim level as it provides a good compromise between reducing the effect of extreme outliers in the dataset while retaining as much information as possible.

```
 #Selecting variables to winsorize
variables_to_winsorize <- c("ai", "mi", "fs", "fliq", "flev")


trim_level <- 0.05
#Winsorize function
winsorize <- function(x, trim_level) {
  q_lower <- quantile(x, trim_level)
  q_upper <- quantile(x, 1-trim_level)
  ifelse(x < q_lower, q_lower, ifelse(x > q_upper, q_upper, x))
}
#Looping over all variables of interest and attaching the winsorized versions as new variables wit
for(var in variables_to_winsorize){
  vi[paste0(var,".w")] <- winsorize(vi[[var]], trim_level)
}
```

```
diagnose_numeric(vi)%>%flextable()
```

variables	min	Q1	mean	median	Q3	max	zero
gvkey	2,163.000000000	18,839.000000000	77,049.52496434	31,846.00000000	164,255.00000000	279,170.00000000	0
tobinsq	0.442159002	1.376855861	3.07238544	1.86446366	3.14062310	49.35456154	0
ros	-5.723936277	-0.009874390	-0.02132042	0.03078996	0.06693068	0.40779616	0
ai	-0.001000204	0.005515476	0.03034174	0.01757241	0.03438598	0.79972716	105

variables	min	Q1	mean	median	Q3	max	zero
mi	0.009752410	0.092807741	0.19407376	0.14007626	0.21348551	2.01956522	0
fs	-3.611918413	1.119394828	2.26794436	2.13888900	3.52636052	6.28971557	0
fliq	0.016892978	0.545758023	1.12210893	0.88211325	1.30126728	13.58378378	0
flev	0.000000000	0.119140640	0.42872303	0.32591876	0.57111654	3.85248930	58
fyear	2,010.000000000	2,013.000000000	2,015.84450785	2,016.00000000	2,018.00000000	2,022.00000000	0
tt	0.000000000	0.000000000	0.08701854	0.00000000	0.00000000	1.00000000	640
ai.w	0.000000000	0.005515476	0.02079059	0.01757241	0.03438598	0.05857461	106
mi.w	0.052730540	0.092807741	0.18024937	0.14007626	0.21348551	0.55436384	0
fs.w	-0.740238788	1.119394828	2.32623631	2.13888900	3.52636052	5.85507192	0
fliq.w	0.193501455	0.545758023	1.02326494	0.88211325	1.30126728	2.80783465	0
flev.w	0.000000000	0.119140640	0.37439725	0.32591876	0.57111654	1.02284599	58

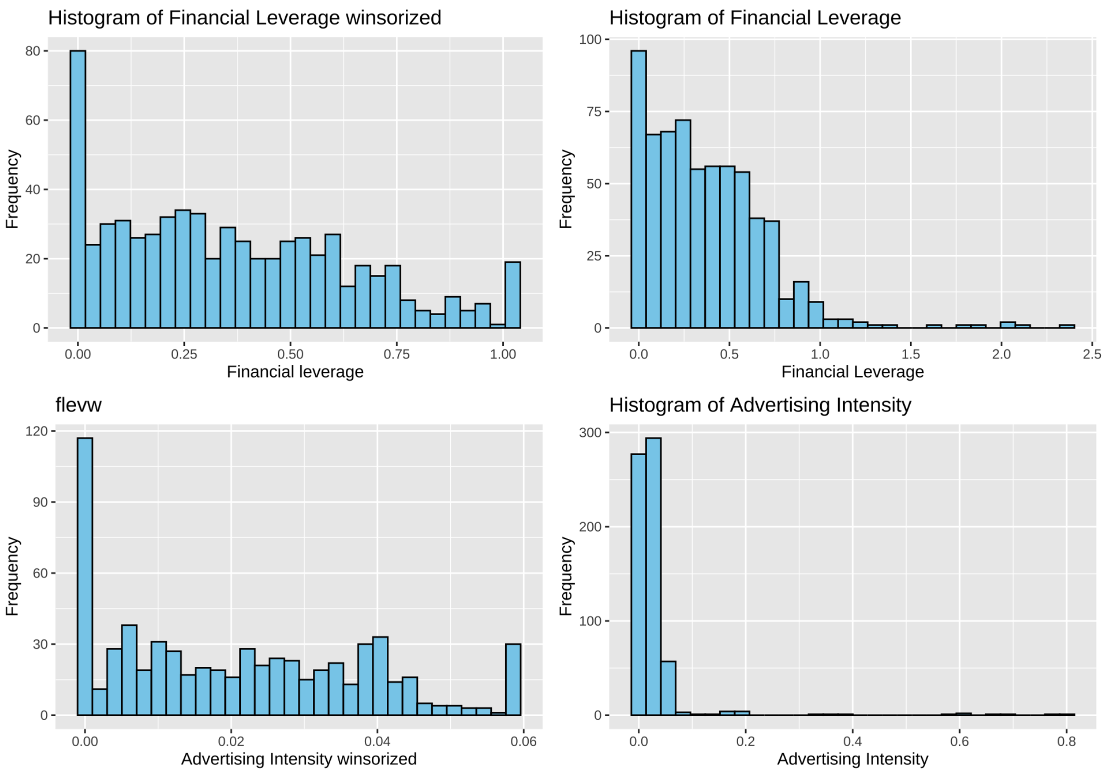
The only exception to this are the variables *tobin's q* and *ros*, where we instead opted to manually remove extreme outliers, as it is generally not considered good practice to winsorize variables that will be later used as the dependent variable in a regression model, as this can lead to information loss, introduce bias into the model and make the interpretation harder, as the coefficients may no longer represent the true effect of the independent variables on the dependent variable.

```
vi <- vi[!grepl("28394", vi$gvkey, ignore.case = TRUE), ]
vi <- vi[!grepl("33134", vi$gvkey, ignore.case = TRUE), ]
vi <- vi[!grepl("22672", vi$gvkey, ignore.case = TRUE), ]
vi <- vi[!grepl("23843", vi$gvkey, ignore.case = TRUE), ]
vi <- vi[!grepl("39756", vi$gvkey, ignore.case = TRUE), ]
vi <- vi[!grepl("160211", vi$gvkey, ignore.case = TRUE), ]
#ros
vi <- vi[!grepl("35967", vi$gvkey, ignore.case = TRUE), ]
vi <- vi[!grepl("160211", vi$gvkey, ignore.case = TRUE), ]
vi <- vi[!grepl("20775", vi$gvkey, ignore.case = TRUE), ]
```

To validate the winsorization of our data, we choose two variables from the dataset and plotted a histogram for the winsorized and non-winsorized version. The results below clearly show that the distribution of the winsorized variables is significantly less skewed when compared to the default variables. These histograms confirm our decision to winsorize variables to prevent extreme outliers from skewing our analysis.

```
plot1 <- ggplot(vi, aes(x = flev.w)) +
  geom_histogram(bins = 30, fill = "skyblue", color = "black") +
  labs(x = "Financial leverage", y = "Frequency") +
  ggtitle("Histogram of Financial Leverage winsorized")
plot2 <- ggplot(vi, aes(x = flev)) +
  geom_histogram(bins = 30, fill = "skyblue", color = "black") +
  labs(x = "Financial Leverage", y = "Frequency") +
  ggtitle("Histogram of Financial Leverage")
plot3 <- ggplot(vi, aes(x = ai.w)) +
  geom_histogram(bins = 30, fill = "skyblue", color = "black") +
  labs(x = "Advertising Intensity winsorized", y = "Frequency") +
  ggtitle("flevw")
```

```
plot4 <- ggplot(vi, aes(x = ai)) +
  geom_histogram(bins = 30, fill = "skyblue", color = "black") +
  labs(x = "Advertising Intensity", y = "Frequency") +
  ggtitle("Histogram of Advertising Intensity")
grid.arrange(plot1, plot2, plot3, plot4, ncol = 2)
```



3. Exploratory Data Analysis

After completing the pre-processing of the dataset, it is now ready for model building. However, for that to be successful we first need to get a better understanding of the data, the different variables it is made of and the relationships between them.

3.1 Descriptive statistics

The table below shows descriptive statistics such as min, mean, median and max for the numeric variables in our dataset. These values are consistent with common sense.

```
diagnose_numeric(vi)%>%
  select(-"outlier")%>%
  flextable()
```

variables	min	Q1	mean	median	Q3	max	zero
gvkey	2,163.000000000	18,047.000000000	78,407.053763440	31,846.00000000	164,284.00000000	279,170.00000000	0
tobinsq	0.442159002	1.344298989	2.354403903	1.78069828	2.92155156	9.72966586	0
ros	-1.554484804	-0.007297706	0.005461061	0.03034246	0.06418581	0.40779616	0
ai	-0.001000204	0.005445965	0.030845661	0.01897070	0.03475718	0.79972716	101
mi	0.009752410	0.091553272	0.173855164	0.13661289	0.20217146	1.75108696	0
fs	-3.611918413	1.184130234	2.393433578	2.19722458	3.61778507	6.28971557	0
fliq	0.056185845	0.545565445	1.102732178	0.87304816	1.25704379	13.58378378	0

variables	min	Q1	mean	median	Q3	max	zero
flev	0.000000000	0.121545238	0.371993696	0.31718707	0.55915084	2.36101695	55
fyear	2,010.000000000	2,013.000000000	2,015.735791091	2,015.00000000	2,018.00000000	2,022.00000000	0
tt	0.000000000	0.000000000	0.078341014	0.00000000	0.00000000	1.00000000	600
ai.w	0.000000000	0.005445965	0.020971671	0.01897070	0.03475718	0.05857461	102
mi.w	0.052730540	0.091553272	0.167320652	0.13661289	0.20217146	0.55436384	0
fs.w	-0.740238788	1.184130234	2.439613203	2.19722458	3.61778507	5.85507192	0
fliq.w	0.193501455	0.545565445	1.003039416	0.87304816	1.25704379	2.80783465	0
flev.w	0.000000000	0.121545238	0.359598244	0.31718707	0.55915084	1.02284599	55

3.2 Visualizations

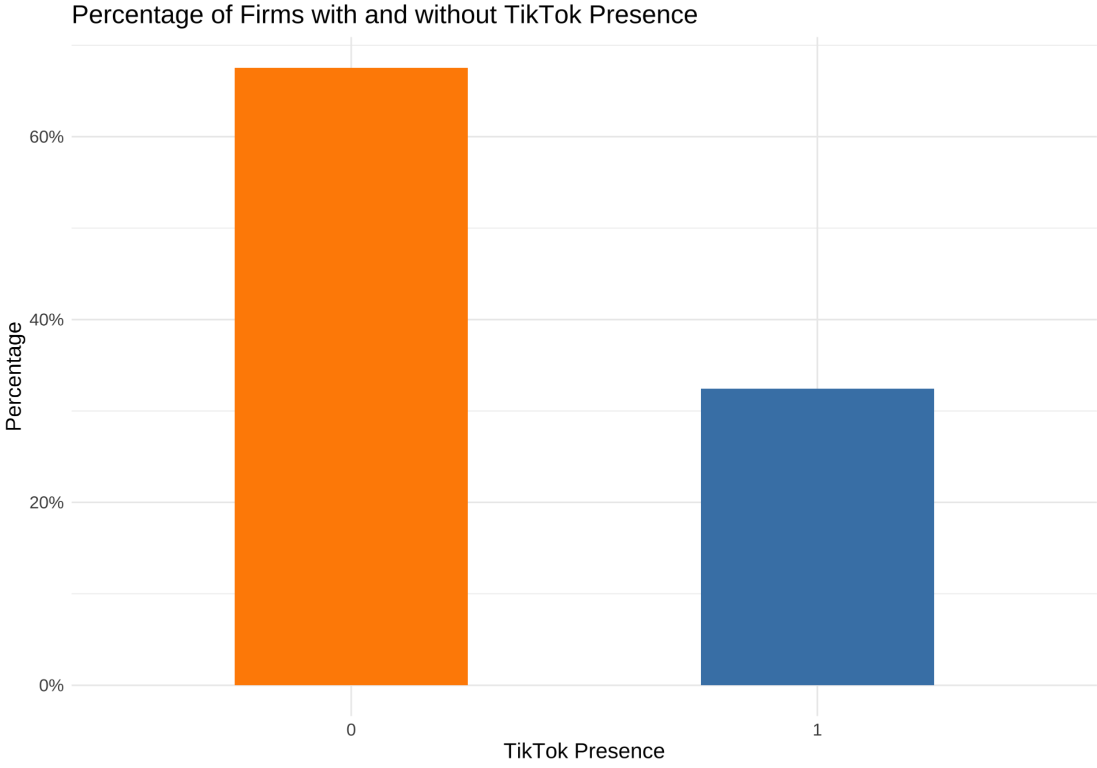
In order to produce meaningful visualizations of the relationships in the dataset, it can be beneficial to aggregate the data from one observation per fiscal year per firm into one observation per firm using the mean of the different variables over the years. Thus, we created *aggregated_data*, that we will be using in the visualizations below.

```
aggregated_data <- vi %>%
  group_by(gvkey) %>%
  summarise(
    tt = as.factor(last(tt)),
    tobinsq = mean(tobinsq, na.rm = TRUE),
    ros = mean(ros, na.rm = TRUE),
    ai.w = mean(ai.w, na.rm = TRUE),
    fs.w = mean(fs.w, na.rm = TRUE),
    fliq.w = mean(fliq.w, na.rm = TRUE),
    flev.w = mean(flev.w, na.rm = TRUE),
    mi.w = mean(mi.w, na.rm = TRUE)
  )
```

3.2.1 Percentage of firms with and without a Tik Tok presence

Before we can start to analyze the effect of operating a Tik Tok account has on a firm, we first need to visualize how many firms already do so. As can be seen in the plot below, the majority of firms in the dataset does not currently operate a Tik Tok account, while approximately 35% do. This makes sense, as Tik Tok is still a relatively new social media platform that primarily attracts younger users and has only enjoyed wide-spread adoption for around three years.

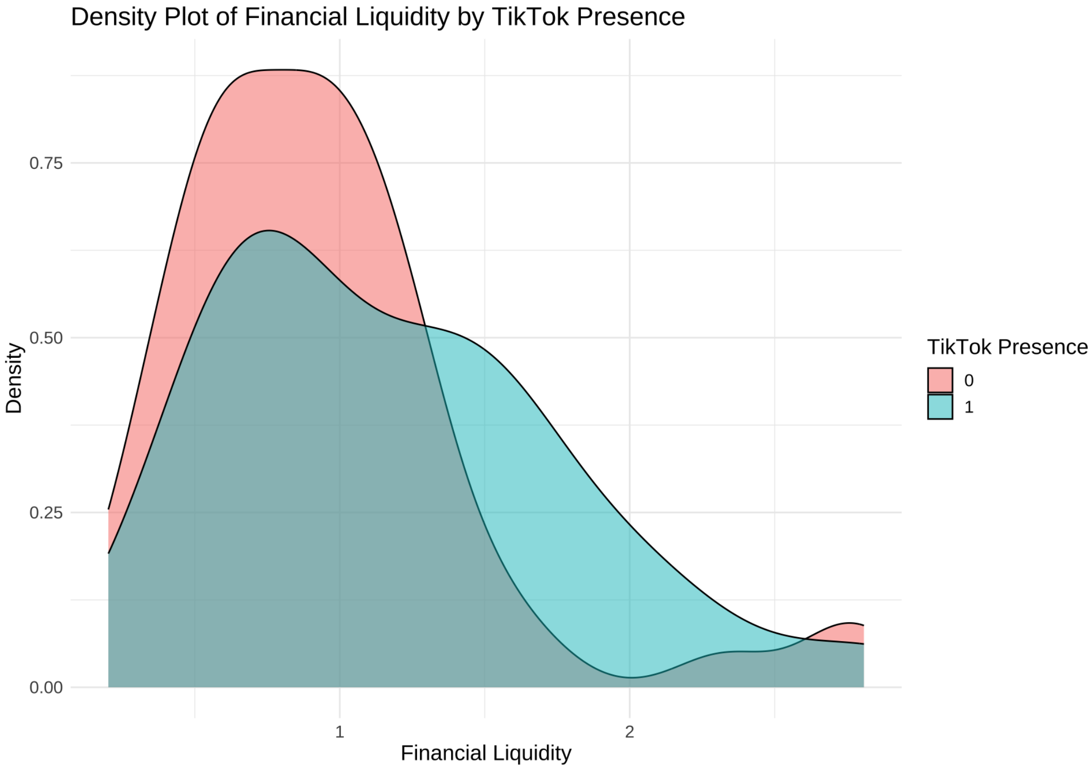
```
ggplot(aggregated_data, aes(x = factor(tt))) +
  geom_bar(aes(y = after_stat(count)/sum(after_stat(count))), width = 0.5, fill = c("darkorange",
  scale_y_continuous(labels = scales::percent) +
  labs(x = "TikTok Presence", y = "Percentage", title = "Percentage of Firms with and without TikT
  theme_minimal() +
  theme(text = element_text(size = 14))
```



3.2.2 Financial Liquidity by TikTok Presence

Another interesting relationship to investigate is the financial liquidity of firms by whether or not they currently have a Tik Tok presence. We can observe notable differences in the financial liquidity between firms that currently operate a Tik Tok account and those that do not. Firms without Tik Tok generally exhibit a higher density at lower liquidity levels. Conversely, firms embracing Tik Tok, have a higher density at higher financial liquidity levels when compared to firms that do not. This could indicate a correlation between the financial liquidity and therefore firm performance and operating an active Tik Tok account.

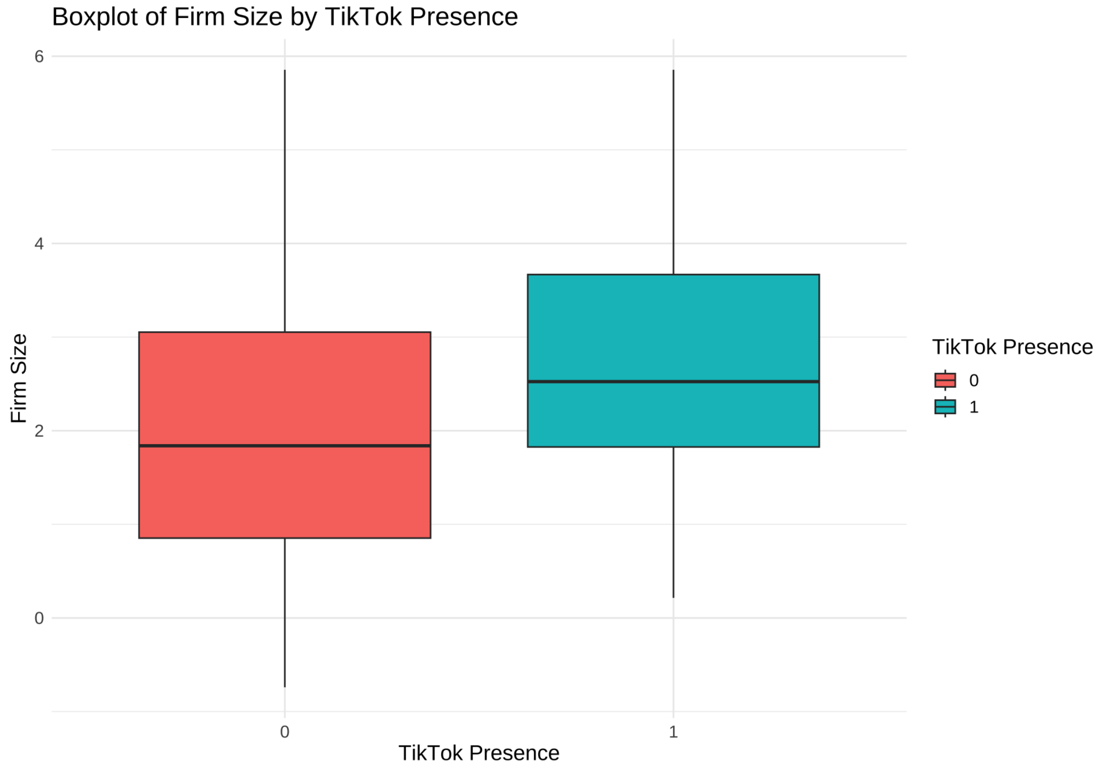
```
ggplot(aggregated_data, aes(x = fliq.w, fill = factor(tt))) +
  geom_density(alpha = 0.5) +
  labs(x = "Financial Liquidity", y = "Density", title = "Density Plot of Financial Liquidity by T
  theme_minimal() +
  theme(text = element_text(size = 14))
```



3.2.3 Firm size by Tik Tok presence

To further investigate and visualize the relationship between a firms performance and whether they embrace Tik Tok or not, the below box plot displays firm size by Tik Tok presence. We can observe that on average, firms that operate a Tik Tok account are generally larger than those that do not. Similarly to the density plot above, this visualization suggests a positive relationship between firm performance and marketing through Tik Tok.

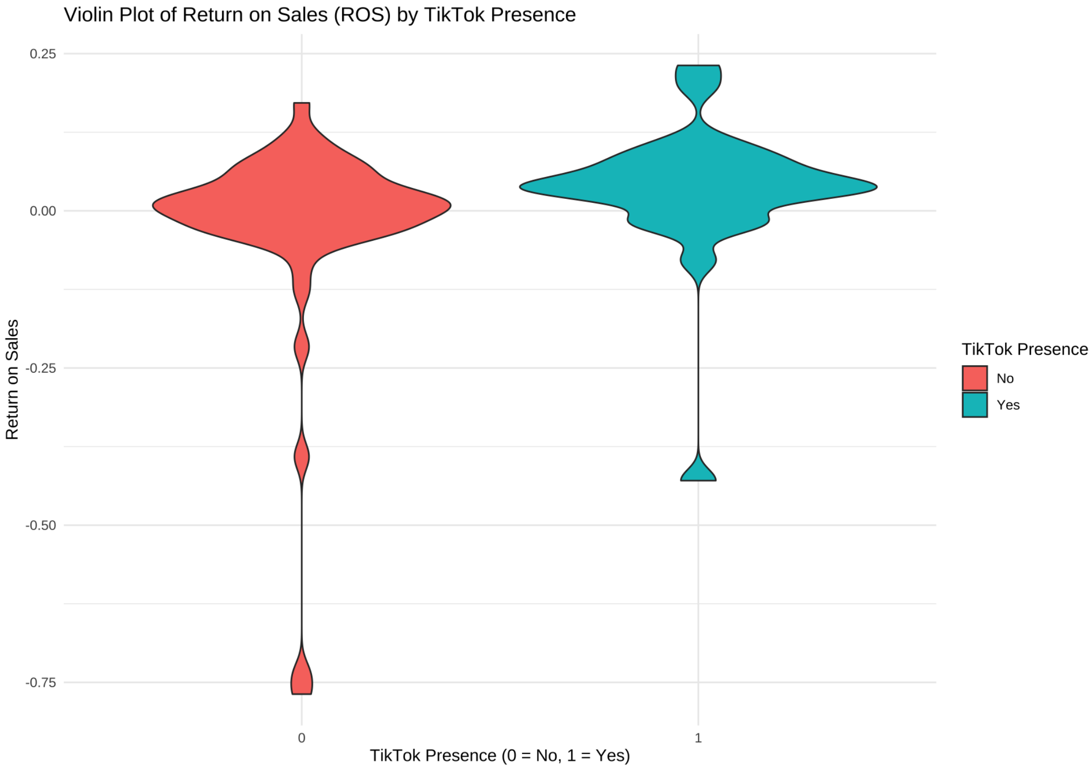
```
ggplot(aggregated_data, aes(x = factor(tt), y = fs.w, fill = factor(tt))) +
  geom_boxplot(outlier.shape = NA) +
  coord_cartesian(ylim = c(min(vi$fs.w), quantile(vi$fs.w, 0.95))) +
  labs(x = "TikTok Presence", y = "Firm Size", title = "Boxplot of Firm Size by TikTok Presence",
  theme_minimal() +
  theme(text = element_text(size = 14))
```



3.2.4 Return on Sales (ROS) by Tik Tok presence

The violin plot below showcases the relationship between Tik Tok marketing and Return on Sales, while also highlighting the density at different ROS levels. Once again, we can observe that firms that currently have a Tik Tok presence have a higher density at a slightly higher Return on Sales. Hence, this visualization also suggest a postive correlation between ROS and marketing using Tik Tok.

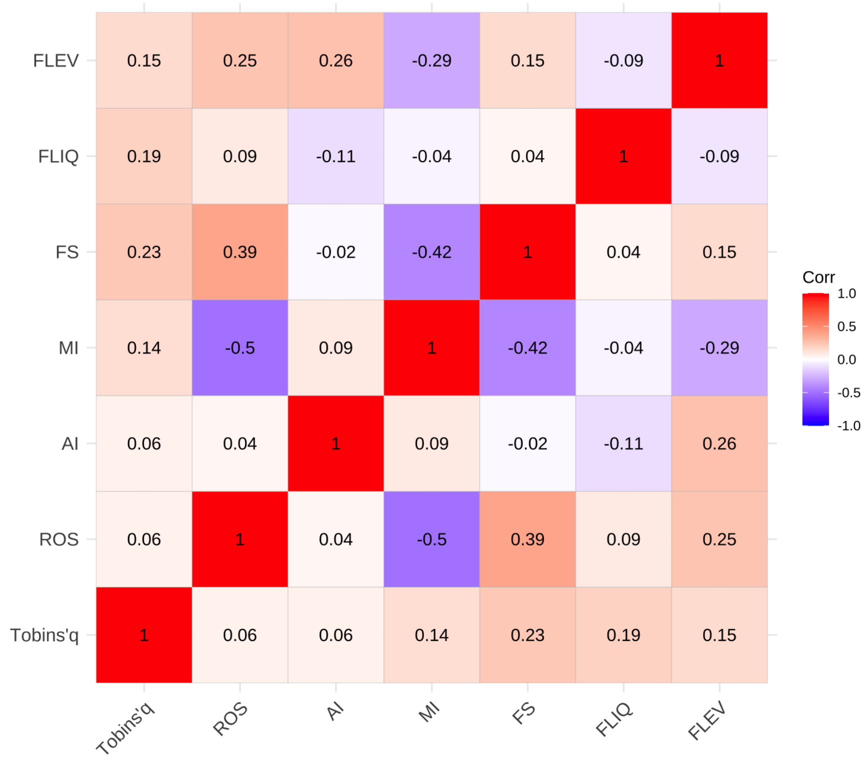
```
ggplot(aggregated_data, aes(x = tt, y = ros, fill = tt)) +
  geom_violin(trim = TRUE) +
  labs(
    title = "Violin Plot of Return on Sales (ROS) by TikTok Presence",
    x = "TikTok Presence (0 = No, 1 = Yes)",
    y = "Return on Sales"
  ) +
  theme_minimal() +
  scale_fill_discrete(name = "TikTok Presence", labels = c("No", "Yes"))
```



3.2.5 Correlation plot

The correlation plot below shows the degree of correlation between all numerical variables in the dataset. There is a low to moderate degree of correlation present in the data at hand, with the highest degree of correlation between the variables Marketing Investments (*MI*) and Return on Sales (*ROS*). The lowest degree of correlation can be observed between Financial Liquidity (*FLIQ*) and Marketing Investments (*MI*). Furthermore, the correlation plot also shows that there is no Multicollinearity present in the dataset, which is important for model building.

```
variables <- cbind(vi$tobinsq, vi$ros, vi$ai.w, vi$mi.w, vi$fs.w, vi$fliq.w, vi$flev.w )
cortable <- round(cor(variables), 2)
colnames(cortable) = c("Tobins'q", "ROS", "AI", "MI", "FS", "FLIQ", "FLEV")
rownames(cortable) = c("Tobins'q", "ROS", "AI", "MI", "FS", "FLIQ", "FLEV")
ggcorrplot(cortable, lab = TRUE)
```



3.3 Exploratory data analysis conclusion

The prior visualizations show that firms that currently have a Tik Tok presence are generally in better financial health, larger by the number of employees and have a higher Return on Sales. To further study these observations and relationships and see whether they are statistically significant or not, we will build multiple linear regression models in the next part of the report.

4. Analysis

After concluding the exploratory data analysis, in this next part we will now construct two regression models to gain some insights on what factors drive a firms performance. For these models we will take a cross-sectional approach, meaning we will utilize the dataset that aggregated the data into one observation per firm, using the mean of the variables over the years.

4.1 Econometric models

To robustly judge a company's performance with a single metric, we are using two econometric models: *Tobin's q* and *Return on Sales (ROS)*. Using these econometric models, multiple linear regression models will be constructed.

4.1 Multiple Linear Regression

Using the above mentioned econometric models, we constructed two multiple linear regression models. The results and interpretations can be seen below:

4.1.1 Tobin's q

Tobins'q is an econometric model that is made up of the market value of a company divided by it's assets' replacement cost. We included all variables of interest in our regression model with the exception of Industry dummies, as our analysis only encompasses firms from a single industry and therefore they all have the same SIC code. The resulting equation for the multiple linear regression model and the model itself can be seen below:

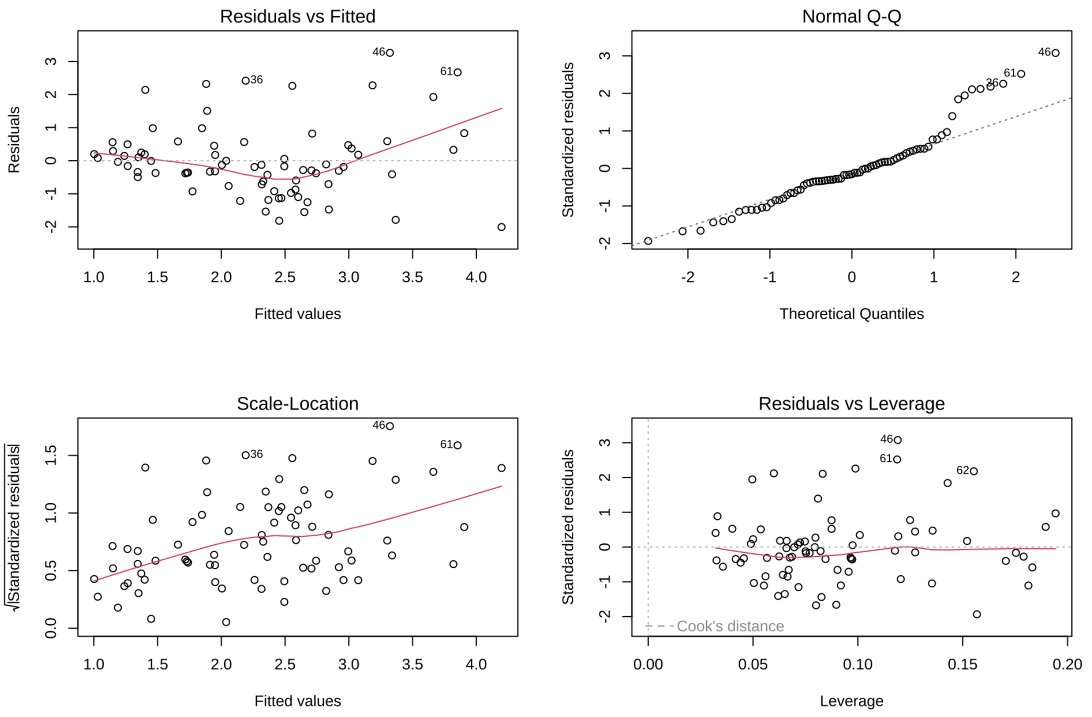
$$Y = -0.58 + 0.58 * tt + 5.51 * mi.w + 1.61 * flev.w + 0.5 * fliq.w + 0.27 * fs.w + 2.99 * ai.w$$

```
tobinsq_model <- lm(tobinsq~tt+mi.w+flev.w+fliq.w+fs.w+ai.w ,data=aggregated_data)
summary(tobinsq_model)
```

```
##
## Call:
## lm(formula = tobinsq ~ tt + mi.w + flev.w + fliq.w + fs.w + ai.w,
##     data = aggregated_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0058 -0.6254 -0.1613  0.4514  3.2618
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.58192    0.55146  -1.055 0.294947
## tt1          0.57847    0.29357   1.971 0.052736 .
## mi.w         5.50551    1.34255   4.101 0.000109 ***
## flev.w       1.60627    0.69805   2.301 0.024368 *
## fliq.w       0.50281    0.23828   2.110 0.038419 *
## fs.w         0.26501    0.08855   2.993 0.003818 **
## ai.w         2.99301    8.72156   0.343 0.732496
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.129 on 70 degrees of freedom
## Multiple R-squared:  0.3265, Adjusted R-squared:  0.2688
## F-statistic: 5.657 on 6 and 70 DF,  p-value: 7.795e-05
```

```
par(mfrow = c(2, 2))
```

```
# Generate the diagnostic plots
plot(tobinsq_model, which=1)
plot(tobinsq_model, which=2)
plot(tobinsq_model, which=3)
plot(tobinsq_model, which=5)
```




```
# Reset the plot layout
par(mfrow = c(1, 1))
```

4.1.2 Tobin's q model evaluation

The results of the *Tobin's q* multiple linear regression model can be seen above. The overall p-value of 7.795e-05 suggests that the model is statistically significant as it is below the significance level of 0.05. Moreover, the multiple R-squared value of 0.3265 suggest that the model fits the data somewhat well, with 32.65% of the variance in *Tobin's q* and therefore the performance of the business being explained by our independent variables. The adjusted R-squared value is naturally lower with 0.2688, as this metric penalizes for the number of predictor variables. When looking at our explanatory variables, interesting results can be observed as well. Only the variables *mi.w*, *flev.w*, *fliq.w* and *fs.w* are statistically significant predictors with p-values ranging from 0.00109 to 0.03818. Furthermore, the underlying assumptions of linear regression models have been met, which can be deduced from the diagnostic plots below showcasing residuals vs fitted values, standardized residuals vs theoretical quantiles, the square root of standardized residuals vs fitted values and Cook's distance test. Our variable *tt*, which indicates whether a firm operates a Tik Tok account, has a p-value of 0.052736 just above the significance level. This suggests, that there is a relationship between *tt* and *Tobin's q*, but it is not strong enough to reject the null hypothesis. However, it is important to keep in mind that statistical significance is not the same as practical significance. Even if an independent variable is not statistically significant, it may still be important in practice if it has a meaningful effect on the dependent variable. Further analysis would be required to investigate this potentially meaningful effect, however this would go beyond the scope of this project.

4.1.3 Return on Sales

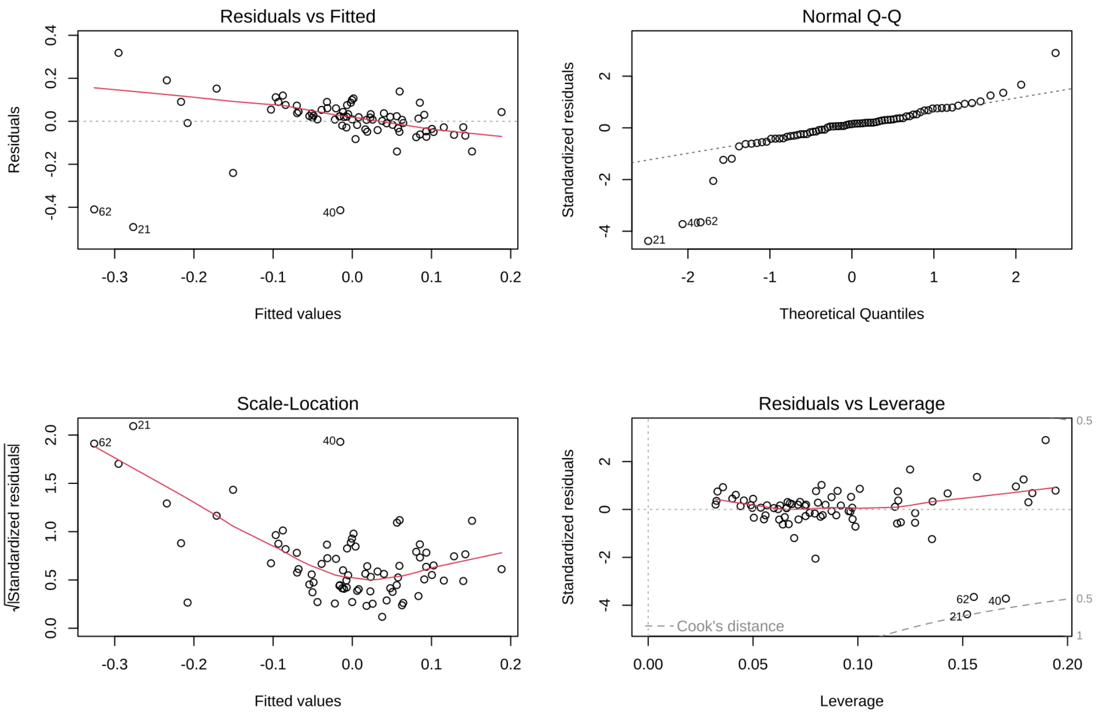
The second econometric model utilized in this analysis, is *Return on Sales*, which is calculated by dividing the firm's net income by its gross sales. The resulting equation for the mutliple linear regression model can be seen below:

$$Y = -0.01 + 0.01 * tt - 0.61 * mi.w + 0.06 * flev.w + 0.03 * fliq.w + 0.02 * fs.w + 0.52 * ai.w$$

```r
ros_model <- lm(ros~tt+mi.w+flev.w+fliq.w+fs.w+ai.w ,data=aggregated_data)
summary(ros_model)
```

```
##
## Call:
## lm(formula = ros ~ tt + mi.w + flev.w + fliq.w + fs.w + ai.w,
##     data = aggregated_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49210 -0.03247  0.01712  0.05284  0.31841
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.014244   0.059597  -0.239   0.8118
## tt1          0.007349   0.031726   0.232   0.8175
## mi.w        -0.610883   0.145092  -4.210 7.46e-05 ***
## flev.w       0.059839   0.075439   0.793   0.4303
## fliq.w       0.033597   0.025751   1.305   0.1963
## fs.w         0.022918   0.009570   2.395   0.0193 *
## ai.w         0.518279   0.942553   0.550   0.5842
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1221 on 70 degrees of freedom
## Multiple R-squared:  0.4278, Adjusted R-squared:  0.3787
## F-statistic: 8.721 on 6 and 70 DF,  p-value: 4.299e-07
```

```r
#Setting plot layout
par(mfrow = c(2, 2))
# Generate the diagnostic plots
plot(ros_model, which=1)
plot(ros_model, which=2)
plot(ros_model, which=3)
plot(ros_model, which=5)
```




```
# Reset the plot layout
par(mfrow = c(1, 1))
```

4.1.4 ROS model evalutaion

The results of the *ROS* multiple linear regression model can be seen above. The overall p-value of 4.299e-07 suggests that the model is statistically significant as it is below the significance level of 0.05. Moreover, the multiple R-squared value of 0.4278 suggest that the model fits the data moderately well, with 42.78% of the variance in *ROS* and therefore the performance of the business being explained by our independent variables. The adjusted R-squared value is naturally lower with 0.3787, as this metric penalizes for the number of predictor variables. When looking at our explanatory variables, interesting results can be observed as well. There are only two statistically significant explanatory variables in this model: Marketing investment (*mi*) and firm size (*fs*) with p-values of 7.46e-05 and 0.0193 respectively. All other variables had p-values significantly over the significance level ranging from 0.1963 to 0.8175. Furthermore, the underlying assumptions of linear regression models have been met, which can be deduced from the diagnostic plots below showcasing residuals vs fitted values, standardized residuals vs theoretical quantiles, the square root of standardized residuals vs fitted values and Cook's distance test.

4.2 Model comparison

Both of the above constructed multiple linear regression models aim to measure the effect investment decisions and firm characteristics have on a businesses performance. To evaluate this performance two different metrics are utilized: *Tobin's q* and *Return on sales*. Both models performed adequately and are statistically significant. The *Tobin's q* model fitted the data less well, as seen by the lower R-squared value of 0.3265. However, all variables except for advertising intensity (*ai*) and Tik Tok (*tt*) were statistically significant predictors for a firms performance measured using *Tobin's q*. Furthermore, the p-value for *tt* was only slightly above the significance level of 0.05 suggesting a correlation between operating a Tik Tok account and a company's performance. The *Return on sales* model on the other hand fitted the data better with an R-squared value of 0.4278. However, the only statistically significant predictors in this model were Marketing investments (*mi*) and firm size (*fs*). The reason could be that while both metrics and therefore models measure a firms performance, *Tobin's q* values a company based on its market value and asset replacement costs. These metrics do not solely depended on marketing and / or investment decisions but also financial metrics such as liquidity and leverage. These financial metrics are not taken into account in the *ROS* model, as this measures the firm's performance purely by comparing net income to sales.

4.3 What type's of firms perform better?

According to the first regression model built, firms with a high ratio of marketing investments, firms with more employees and to a lesser extent firms with healthy financials (leverage and liquidity) tend to perform better. Whether a firm operates a Tik Tok account or not also seems to have an effect on the company's performance, however this effect is not statistically significant. According to the second model constructed (*Return on sales*), large firms with a high ratio of marketing investments perform better. In conclusion, if both models are taken into account, large firms with a high ratio of marketing investments seem to perform better than their peers.

4.4 Based on your regression results, would you recommend firms to invest in marketing?

Yes, we would recommend that firms invest in marketing, as marketing investments are by far the strongest predictor for a business's performance across both multiple linear regression models.

4.5 Tik Tok presence as a competitive advantage

Based on the analysis conducted we can say that having a Tik Tok account does seem to have an effect on a firms performance according to *Tobin's q*. This could also be seen during the exploratory data analysis, where firms that had an active Tik Tok presence were generally larger and in better shape financially. As described earlier, this effect is not statistically significant according to our multiple linear regression models. Therefore, we would definitely recommend that businesses have an active Tik Tok presence. This marketing decision by itself is not a guarantee for success. However, we believe that in conjunction with other marketing and advertising decisions and investments, a Tik Tok account can be beneficial for a company's performance. Especially when considering that Tik Tok is still a relativeley new social media platform that is still expected to grow.

4.6 Opportunities for improvement

The earlier constructed models could be improved in several ways. Firstly, we could choose a different approach to handling missing values and outliers. Additionally, we could also employ techniques like backward elimination, forward selection, and stepwise selection to select the most important features. Moreover, we could also choose a different type of regression model that is better suited for the dataset at hand such as PLS-regression for example. Lastly, we could also collect more data on Tik Tok. For this analysis we choose to only collect data indicating whether a firm has an Tik Tok account or not. Additional data such as the number of followers, views and interactions could give us a more robust estimate of the effect that Tik Tok has on a corporation's performance.

5. Longitudinal Approach

For the analysis conducted in part 4 of this report we took a cross sectional approach. For this last part of the analysis we will now use a longitudinal approach, meaning we will now take the original dataset that has one observation per firm per fiscal year instead of just one observation per firm.

5.1 Firms in our sample

As shown below, we have observations for 77 unique businesses in our dataset.

```
length(unique(vi$gvkey))
```

```
## [1] 77
```

5.2 Multiple linear regression

Just as we did in the previous part of this investigation, we will now analyze the performance of firms using the two econometric models using longitudinal data.

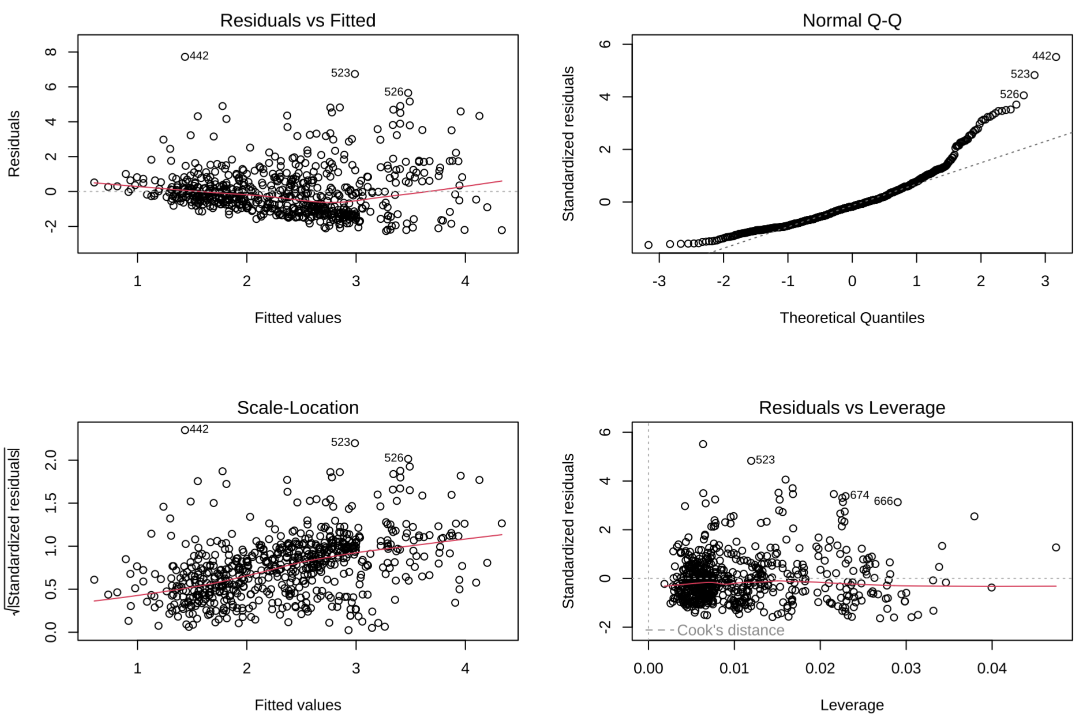
5.2.1 *Tobin's q*

```
tobinsq_model_long <- lm(tobinsq~tt+mi.w+flev.w+fliq.w+fs.w+ai.w ,data=vi)
summary(tobinsq_model_long)
```

```
##
## Call:
## lm(formula = tobinsq ~ tt + mi.w + flev.w + fliq.w + fs.w + ai.w,
##     data = vi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2664 -0.9541 -0.2470  0.5793  7.7231
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.12969    0.21132  -0.614    0.540
## tt           0.31991    0.21173   1.511    0.131
## mi.w         4.69915    0.55420   8.479  < 2e-16 ***
## flev.w       1.14880    0.22305   5.150 3.46e-07 ***
## fliq.w       0.50512    0.08769   5.760 1.30e-08 ***
## fs.w         0.29733    0.03510   8.472  < 2e-16 ***
## ai.w         1.31814    3.40436   0.387    0.699
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.405 on 644 degrees of freedom
```

```
## Multiple R-squared:  0.1982, Adjusted R-squared:  0.1907
## F-statistic: 26.53 on 6 and 644 DF,  p-value: < 2.2e-16
```

```
#Setting plot layout
par(mfrow = c(2, 2))
# Generate the diagnostic plots
plot(tobinsq_model_long, which=1)
plot(tobinsq_model_long, which=2)
plot(tobinsq_model_long, which=3)
plot(tobinsq_model_long, which=5)
```




```
# Reset the plot layout
par(mfrow = c(1, 1))
```

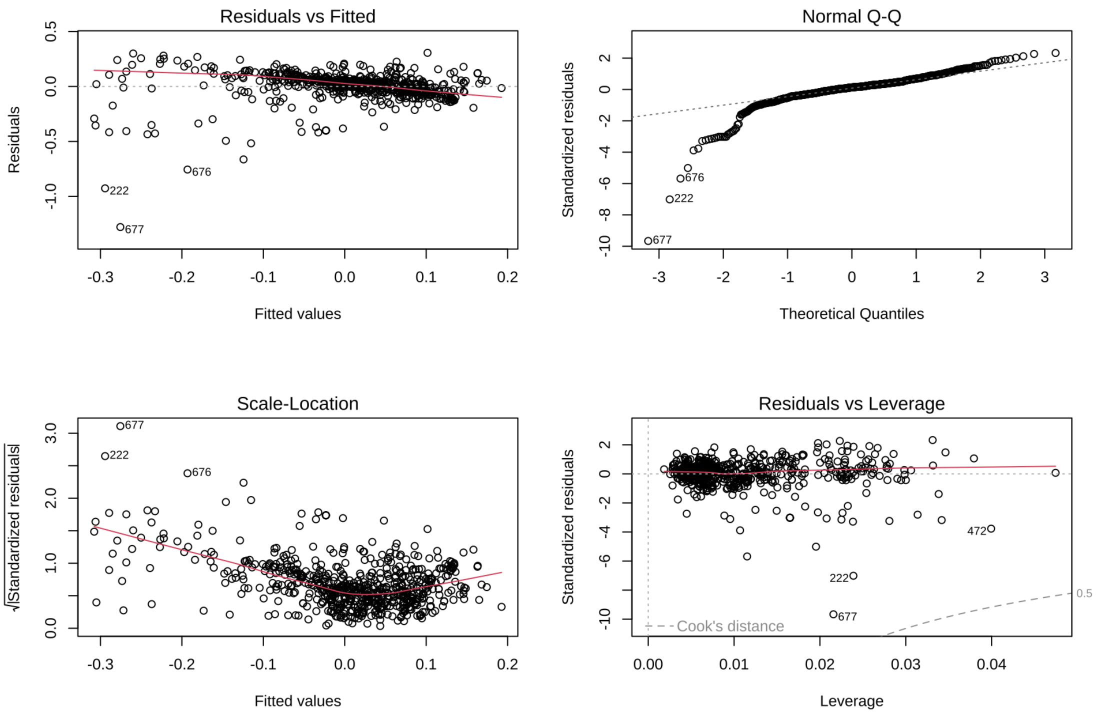
5.2.2 Return on sales

```
ros_model_long <- lm(ros~tt+mi.w+flev.w+fliq.w+fs.w+ai.w ,data=vi)
summary(ros_model_long)
```

```
##
## Call:
## lm(formula = ros ~ tt + mi.w + flev.w + fliq.w + fs.w + ai.w,
##     data = vi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.27891 -0.03757  0.01792  0.05964  0.30615
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.007473   0.020123  -0.371  0.71048
## tt          -0.029461   0.020162  -1.461  0.14444
## mi.w        -0.530340   0.052774 -10.049  < 2e-16 ***
## flev.w       0.063368   0.021240   2.983  0.00296 **
## fliq.w       0.022947   0.008350   2.748  0.00616 **
## fs.w         0.019840   0.003342   5.937 4.76e-09 ***
## ai.w         0.466070   0.324181   1.438  0.15101
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1338 on 644 degrees of freedom
## Multiple R-squared:  0.3093, Adjusted R-squared:  0.3029
## F-statistic: 48.07 on 6 and 644 DF,  p-value: < 2.2e-16
```

```
 #Setting plot layout
par(mfrow = c(2, 2))
# Generate the diagnostic plots
plot(ros_model_long, which=1)
plot(ros_model_long, which=2)
plot(ros_model_long, which=3)
plot(ros_model_long, which=5)
```




```
 # Reset the plot layout
par(mfrow = c(1, 1))
```

5.3 Comparing cross-sectional & longitudinal models

When comparing the results of our newly created multiple linear regression models to those we built in part 4 of our analysis there are some interesting observations. Firstly, the models fit the data less well with lower R-squared values of 0.1982 and 0.3093 respectively. Secondly, in both new models the statistically significant variables are identical: Marketing investments, financial leverage, financial liquidity and firm size, while they differed between models earlier. Furthermore, in both models having a Tik Tok account is not a statistically significant predictor. However it must be noted that a lot of the data present in the dataset is older than three years (2020). During this time Tik Tok was not a popular or widely adopted social media platform. Lastly, it must be said that the above mentioned metrics are not enough to accurately measure a model's performance. More metrics and more robust evaluation technique's such as cross-validation are necessary to better compare the two model pairs.

5.4 Reverse causality

In order to rule out reverse causality and to account for dynamic marketing effects, we will introduce lead and lag variables. As a firms marketing decisions usually take some time to materialize into a measurable effect on the companies performance, a lead variable was created. Furthermore, a lag variable was created to account for dynamic marketing effects in our models. The code can be seen below

```
 #Lead variables
vi <- slide(data = vi, Var = "tobinsq", TimeVar = "fyear", GroupVar = "gvkey",
                   NewVar = 'leadtobinsq', slideBy = +1, keepInvalid = TRUE, reminder = TRUE)
vi$leadtobinsq<-ifelse(is.na(vi$leadtobinsq),vi$tobinsq,vi$leadtobinsq)

vi <- slide(data = vi, Var = "ros", TimeVar = "fyear", GroupVar = "gvkey",
            NewVar = 'leadros', slideBy = +1, keepInvalid = TRUE, reminder = TRUE)
vi$leadros<-ifelse(is.na(vi$leadros),vi$ros,vi$leadros)

##Lag Variables
vi <- slide(data = vi, Var = "mi.w", TimeVar = "fyear", GroupVar = "gvkey",
            NewVar = 'lagmi', slideBy = -1, keepInvalid = TRUE, reminder = TRUE)
vi$lagmi<-ifelse(is.na(vi$lagmi),vi$mi.w,vi$lagmi)

vi <- slide(data = vi, Var = "ai.w", TimeVar = "fyear", GroupVar = "gvkey",
            NewVar = 'lagai', slideBy = -1, keepInvalid = TRUE, reminder = TRUE)
vi$lagai<-ifelse(is.na(vi$lagai),vi$ai.w,vi$lagai)
```

5.5 New Multiple Linear Regression

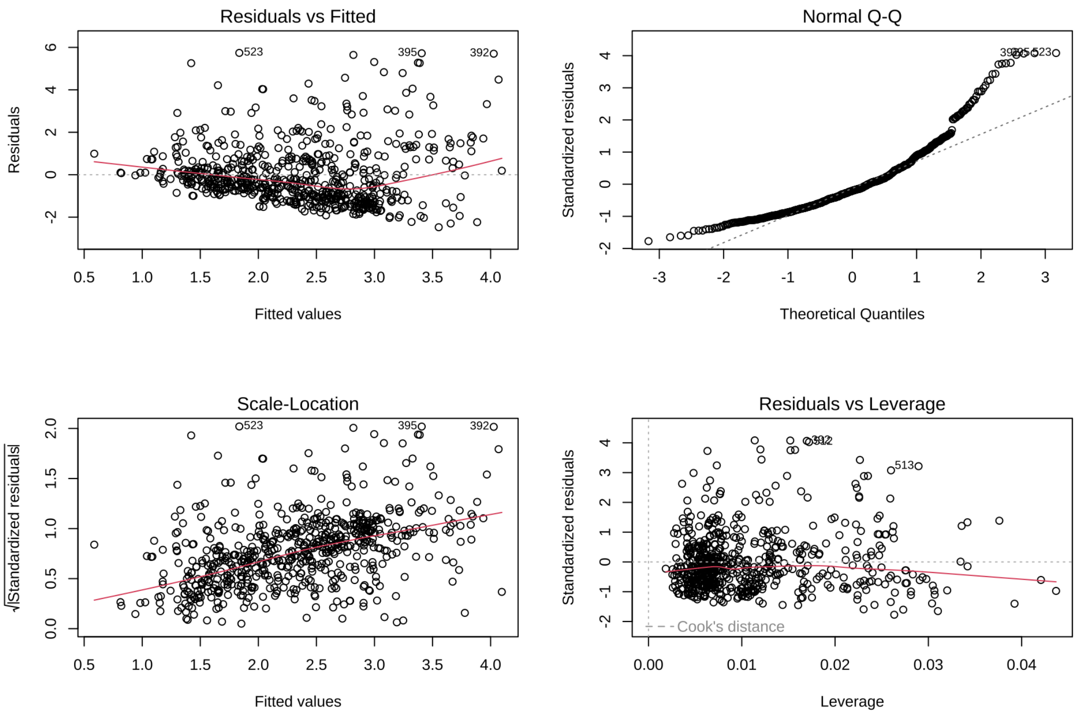
As mentioned above, to account for delays and dynamic effects of marketing two new multiple linear regression models using the same econometric models will be built.

5.5.1 *Tobin's q*

```
tobinsq_model_long_ll <- lm(leadtobinsq~tt+lagmi+flev.w+fliq.w+fs.w+lagai ,data=vi)
summary(tobinsq_model_long_ll)
```

```
##
## Call:
## lm(formula = leadtobinsq ~ tt + lagmi + flev.w + fliq.w + fs.w +
##     lagai, data = vi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4750 -0.9869 -0.2878  0.6165  5.7409
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.05800    0.20712   0.280    0.780
## tt           0.16789    0.21335   0.787    0.432
## lagmi        3.72897    0.54922   6.790 2.56e-11 ***
## flev.w       1.15168    0.22372   5.148 3.50e-07 ***
## fliq.w       0.44972    0.08837   5.089 4.73e-07 ***
## fs.w         0.28603    0.03534   8.095 2.87e-15 ***
## lagai        3.33608    3.41228   0.978    0.329
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.415 on 644 degrees of freedom
## Multiple R-squared:  0.1785, Adjusted R-squared:  0.1709
## F-statistic: 23.32 on 6 and 644 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2, 2))
# Generate the diagnostic plots
plot(tobinsq_model_long_ll, which=1)
plot(tobinsq_model_long_ll, which=2)
plot(tobinsq_model_long_ll, which=3)
plot(tobinsq_model_long_ll, which=5)
```




```
# Reset the plot layout
par(mfrow = c(1, 1))
```

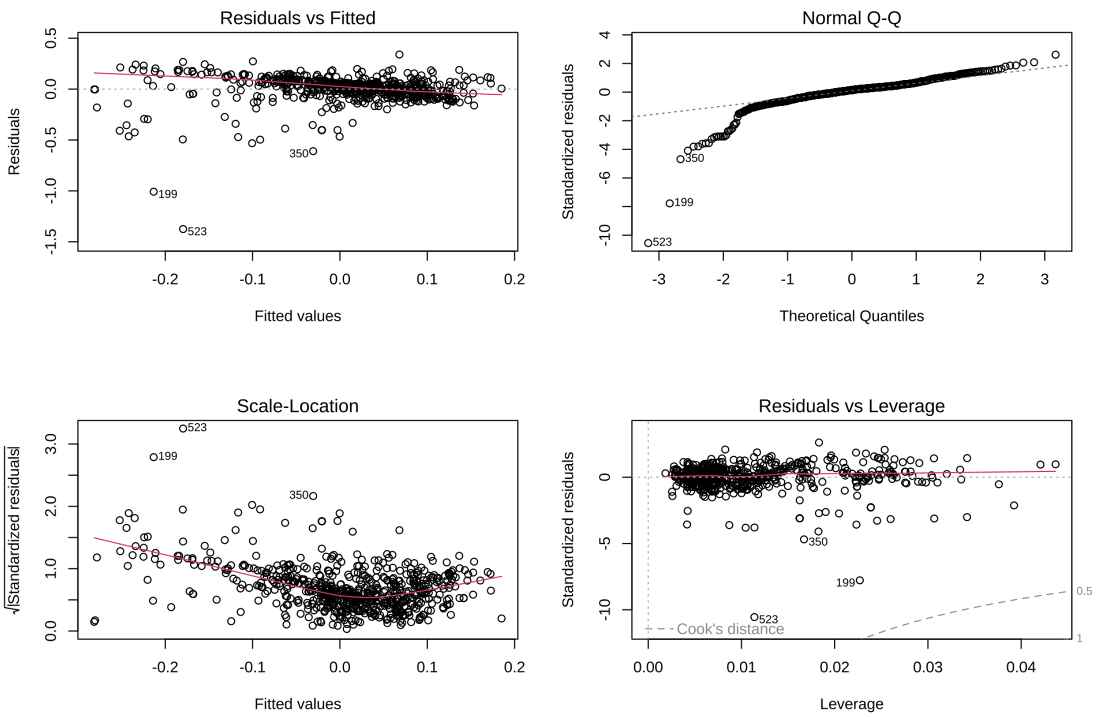
5.5.2 *Return on Sales*

```
ros_model_long_l1 <- lm(leadros~tt+lagmi+flev.w+fliq.w+fs.w+lagai ,data=vi)
summary(ros_model_long_l1)
```

```
##
## Call:
## lm(formula = leadros ~ tt + lagmi + flev.w + fliq.w + fs.w +
##     lagai, data = vi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.37496 -0.03551  0.01855  0.05827  0.33969
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.021936   0.019192  -1.143  0.25349
## tt          -0.024880   0.019769  -1.259  0.20865
## lagmi       -0.458952   0.050892  -9.018  < 2e-16 ***
## flev.w       0.085059   0.020730   4.103 4.60e-05 ***
## fliq.w       0.023028   0.008188   2.812  0.00507 **
## fs.w         0.018564   0.003274   5.670 2.16e-08 ***
## lagai        0.437414   0.316188   1.383  0.16702
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1311 on 644 degrees of freedom
## Multiple R-squared:  0.2863, Adjusted R-squared:  0.2797
## F-statistic: 43.06 on 6 and 644 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2, 2))
# Generate the diagnostic plots
plot(ros_model_long_ll, which=1)
plot(ros_model_long_ll, which=2)
plot(ros_model_long_ll, which=3)
plot(ros_model_long_ll, which=5)
```




```
# Reset the plot layout
par(mfrow = c(1, 1))
```

5.6 Model comparison

Our newly created models fit the un-aggregated data similarly well when compared to those without lead and lag variables. However the model fit is worse when compared to the regression models built using aggregated data with an r-squared value of 0.1785 and 0.2863 respectively. The statistically significant predictors are identical to those in the non-aggregated data models but differ from those in the aggregated models. They are marketing intensity (lag), financial leverage, financial liquidity, and firm size. The new models also meet the underlying assumptions of linear models which can be seen in the diagnostic plots. Having a Tik Tok presence is not statistically significant for a firms performance in both models. Furthermore we also could not observe a relationship between Tobin’s q and having a Tik Tok account like we could in the first model created using the aggregated data. However once again it must be noted that these metrics are not enough to accurately measure and compare the six models with each other. We would need to compare their performance using additional techniques while also keeping the model's use case in mind.

5.7 Firms performance, marketing investments and Tik Tok presence compared

When it comes to which firms perform better, the answer has only changed slightly, as now marketing intensity (lag), financial leverage, financial liquidity, and firm size are all factors that contribute to a comparatively better performance of a business across both econometric models, while it was only marketing investments and firm size, if both models are taken into account, for the aggregated data. However, the first *Tobin's q* model had the same characteristics causing superior performance. When it comes to investing in marketing generally, the new models confirm what has been shown earlier in our analysis, firms that have a larger marketing investment ration tend to perform better than their competitors in the restaurant industry. In all models created, having a Tik Tok presence did not serve as a statistically significant predictor for a corporation's performance. However, we would still recommend that firms operate a Tik Tok account. We arrive at this conclusion because firstly, we could observe a minor correlation between *Tobin's q* and having an active Tik Tok presence or not in the very first model we created in 4.1.2. Furthermore and perhaps more importantly, Tik Tok is a growing social media platform that is becoming increasingly popular both for businesses and individual consumers. A Tik Tok presence should be a part of a carefully constructed marketing mix that encompasses different types of advertising and marketing investments to maximize a firm's performance.

6. Conclusion

In summary, we analyzed the effect that operating a Tik Tok account has on companies in the restaurant industry. To achieve this, we first collected Tik Tok data and integrated it into our existing dataset of gastronomy firms. Next, we pre-processed the data by addressing missing values and outliers present in our data. Afterwards, we introduced new variables that would form the basis for our multiple linear regression models. Then, we conducted exploratory data analysis to get a better understanding of the data and the relationships it contains. Lastly, we constructed six different multiple linear regression models using two econometric models: *Tobin's q* and *Return on sales*. The models differed in whether they were using aggregated data or not and whether they included lead & lag variables or not. While we could observe a relationship between operating a Tik Tok presence and *Tobin's q* in the very first model we built, this correlation was not statistically significant and could not be observed in the remaining five models. As mentioned earlier, this could be due to the fact that a lot of the data that we based this analysis and the models on was pre 2020 when Tik Tok was not widely used or known. Thus, further analysis with different data is required to definitely quantify the effect of Tik Tok. In today's competitive gastronomy market environment, social media has become an important part of the marketing mix. Therefore we would still recommend that firms have an active presence on Tik Tok, while also considering that this is by itself is not a guarantee for success and needs to be carefully integrated with other social media platform presences.