Product_Sales_Analysis

ERICK@Guru

2023-07-029

Product Sales Analysis Project For Data Analyst Professional

The Business goals

1. How many customers were there for each approach?
2. What does the spread of the revenue look like overall? And for each method?
3. Was there any difference in revenue over time for each of the methods?
4. Based on the data, which method would you recommend we continue to use?

The Business Metrics

The Recommendations

Using the validation criteria, the following validation was made:

- week: 6 unique values, without any missing data.
- sales_method: had 5 unique values before validation: Email, Call, Email + Call, em + call, and email, which after validation were Email, Call, and Email + Call.
- customer_id: 15,000 unique values. Needed no cleaning.
- nb_sold: 10 unique values, no cleaning required and no missing values.
- revenue: had 1074 missing values, of which the rows were removed from the data set.
- years_as_customer: had two major values not corresponding: 47 and 63 which were way more than the number of years Pens and Printers has been in existence, 39 years. It made no sense having a customer when the business was not in existence. These rows were dropped.
- nb_site_visits: Needed no cleaning.
- state: Needed no cleaning too. At the end of the validation and cleaning process, the data that remained is 13,924 rows and 8 columns

Data Import, Validation, Cleaning, and Exploration

```
#import Libraries
library(tidyverse)
library(janitor)
library(DataExplorer)
```

```r
sales <- read_csv("product_sales.csv")
head(sales)
```

```
## # A tibble: 6 x 8
##    week sales_method customer_id                nb_sold revenue years_as_customer
##   <dbl> <chr>        <chr>                       <dbl>   <dbl>              <dbl>
## 1     2 Email        2e72d641-95ac-497b-bbf8-~      10    NA                    0
## 2     6 Email + Call 3998a98d-70f5-44f7-942e-~      15   225.                   1
## 3     5 Call         d1de9884-8059-4065-b10f-~      11    52.6                  6
## 4     4 Email        78aa75a4-ffeb-4817-b1d0-~      11    NA                    3
## 5     3 Email        10e6d446-10a5-42e5-8210-~       9    90.5                  0
## 6     6 Call         6489e678-40f2-4fed-a48e-~      13    65.0                 10
## # i 2 more variables: nb_site_visits <dbl>, state <chr>
```

```r
dim(sales)
```

```
## [1] 15000     8
```

The data set contains *15,000 rows/observations and 8 columns/features before the cleaning and validation process.

```r
#check duplicates
sum(duplicated(sales$customer_id))#no duplicates
```

```
## [1] 0
```

```r
#Check mising values
colSums(is.na(sales)) %>% as.data.frame()#1074 missing values in #revenue column
```

```
##                      .
## week                 0
## sales_method         0
## customer_id          0
## nb_sold              0
## revenue           1074
## years_as_customer    0
## nb_site_visits       0
## state                0
```

```r
#Remove Missing values
sales <- na.omit(sales)
```

```r
colSums(is.na(sales))#No Missing values
```

```
##              week      sales_method       customer_id           nb_sold
##                 0                 0                 0                 0
##           revenue years_as_customer    nb_site_visits             state
##                 0                 0                 0                 0
```

```r
names(sales)
```

```
## [1] "week"              "sales_method"      "customer_id"
## [4] "nb_sold"           "revenue"           "years_as_customer"
## [7] "nb_site_visits"    "state"
```

```r
#Remove Column customer_id
sales$customer_id <- NULL
```

```r
# week: 6 unique values, without any missing data.
unique(sales$week)
```

```
## [1] 6 5 3 4 1 2
```

```r
# sales_method: had 5 unique values before validation: Email, Call, Email + Call, em + call, and email,

unique(sales$sales_method)
```

```
## [1] "Email + Call" "Call"         "Email"        "em + call"    "email"
```

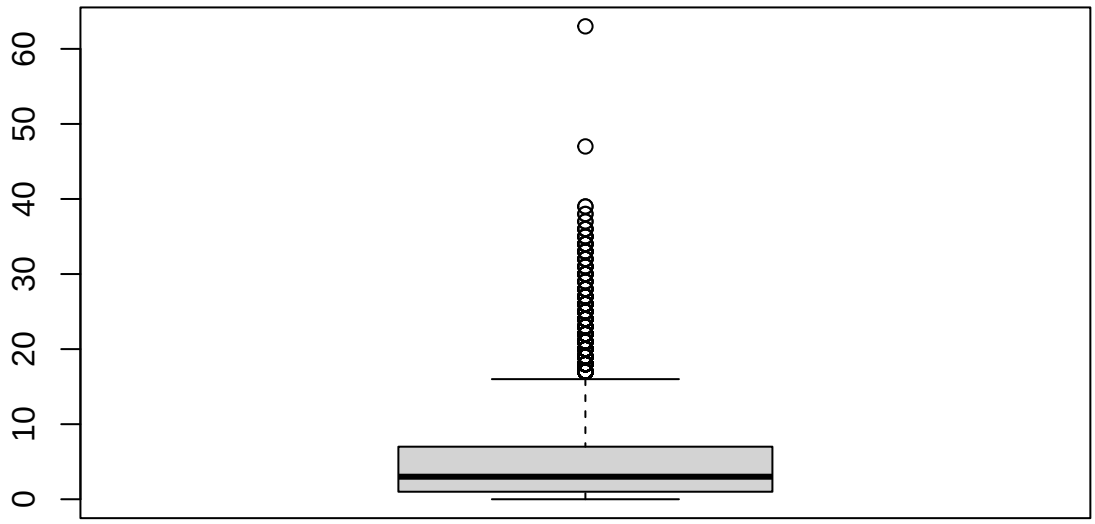
```r
#Sales method now has 3 uniques values as per description
sales <- sales %>%
     mutate(
       sales_method=ifelse(sales_method=="em + call","Email + Call",
                           ifelse(sales_method=="email","Email", sales_method)      )
     )

#Then we Check
unique(sales$sales_method)
```

```
## [1] "Email + Call" "Call"         "Email"
```

```r
# years_as_customer: had two major values not corresponding: 47 and 63 which were way more than the num
summary(sales$years_as_customer)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   3.000   4.979   7.000  63.000
```

```r
boxplot(sales$years_as_customer)
```




```
sales1 <- sales %>%  filter(!years_as_customer>39)#remove outliers(47,63)
```

Our Data is Clean Now

```
sales_clean <- sales1
sales_clean$sales_method <- as.factor(sales_clean$sales_method)
glimpse(sales_clean)
```

```
## Rows: 13,924
## Columns: 7
## $ week              <dbl> 6, 5, 3, 6, 4, 1, 5, 5, 3, 2, 5, 2, 5, 4, 1, 1, 1, 1~
## $ sales_method      <fct> Email + Call, Call, Email, Call, Email, Email, Email~
## $ nb_sold           <dbl> 15, 11, 9, 13, 11, 10, 11, 11, 9, 9, 11, 10, 10, 10,~
## $ revenue           <dbl> 225.47, 52.55, 90.49, 65.01, 113.38, 99.94, 108.34, ~
## $ years_as_customer <dbl> 1, 6, 0, 10, 9, 1, 10, 7, 4, 2, 2, 1, 1, 2, 4, 2, 4,~
## $ nb_site_visits    <dbl> 28, 26, 28, 24, 28, 22, 31, 23, 28, 23, 30, 28, 30, ~
## $ state             <chr> "Kansas", "Wisconsin", "Illinois", "Mississippi", "G~
```

```
#Reorder labels
levels(sales_clean$sales_method) #<- c("Email","Call" ,"Email + Call")
```

```
## [1] "Call"         "Email"        "Email + Call"
```

Back to the Business Objectives

The Business goals

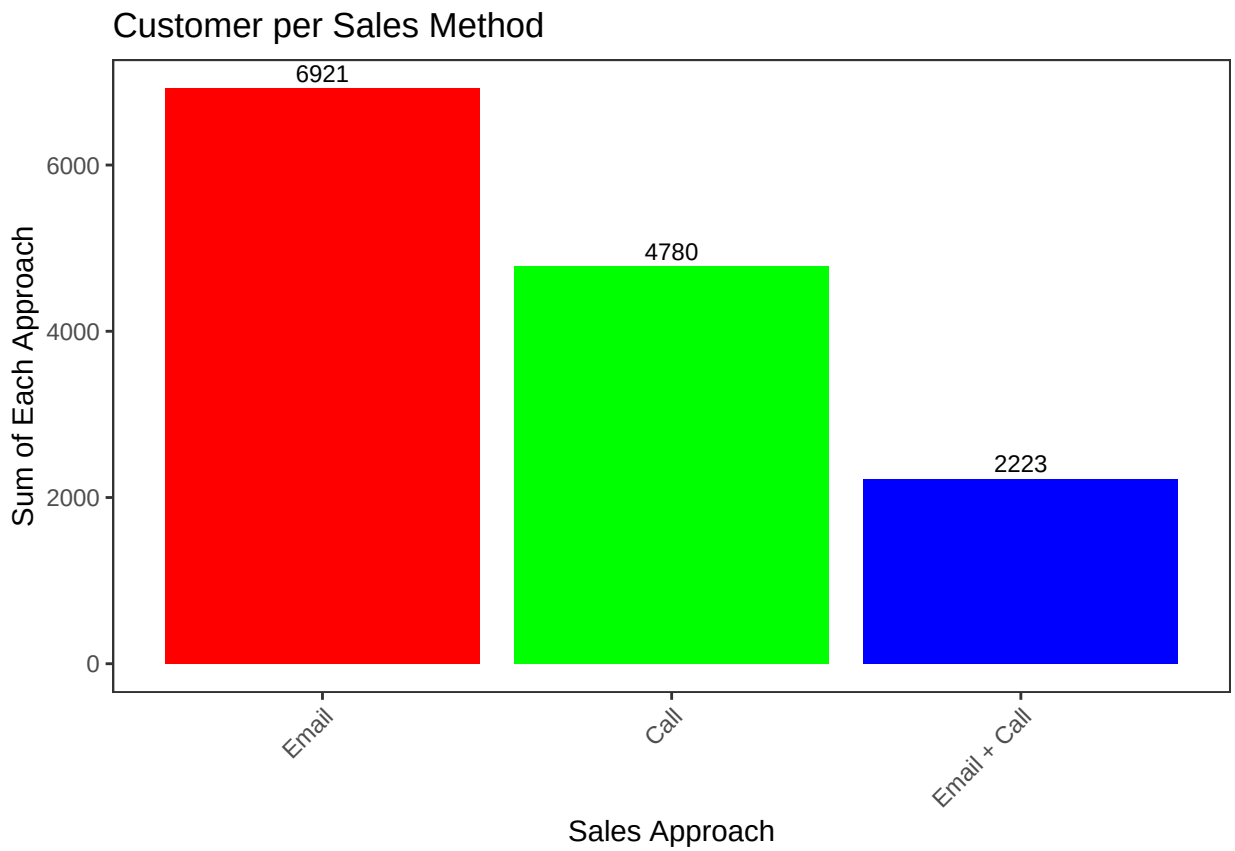
1. How many customers were there for each approach?
2. What does the spread of the revenue look like overall? And for each method?
3. Was there any difference in revenue over time for each of the methods?
4. Based on the data, which method would you recommend we continue to use?

The Business Metrics

The Recommendations

1. How many customers were there for each sales method/ approach?

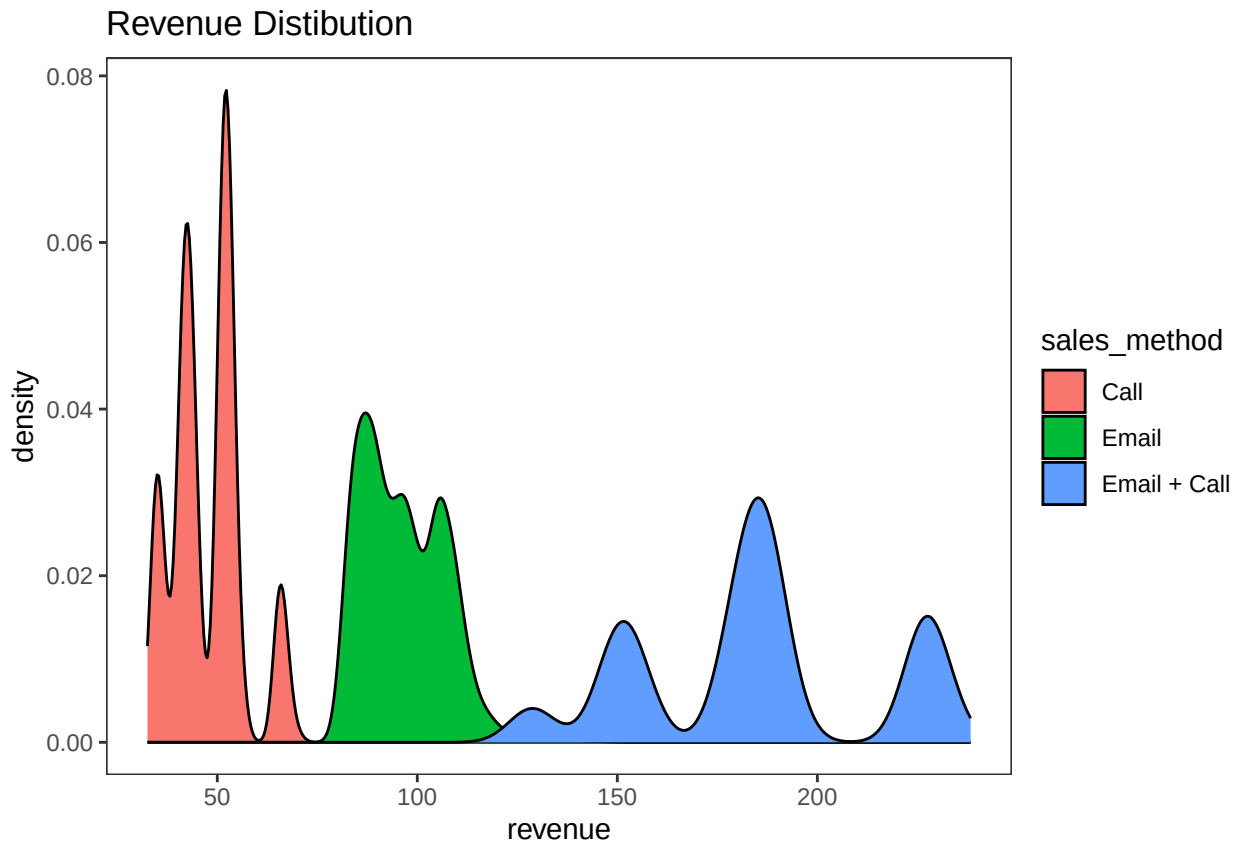
```
theme_set(theme_test())
sales_clean %>%
  ggplot(aes(x=fct_infreq(sales_method)))+
  geom_bar(fill=rainbow(3))+
   geom_text(aes(label=after_stat(count)),
     stat='count',
    position=position_dodge(1.0),
     vjust= -0.4,
     size=3)+
  theme(legend.position = 'bottom')+
   labs(
     x='Sales Approach',
     y='Sum of Each Approach',
     title = 'Customer per Sales Method'
   ) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



The **Email** sales method has the vast majority of **6921** customers, followed by **Call** & Combination of **Email and Call, with the respective count of 4780 and 2223.**

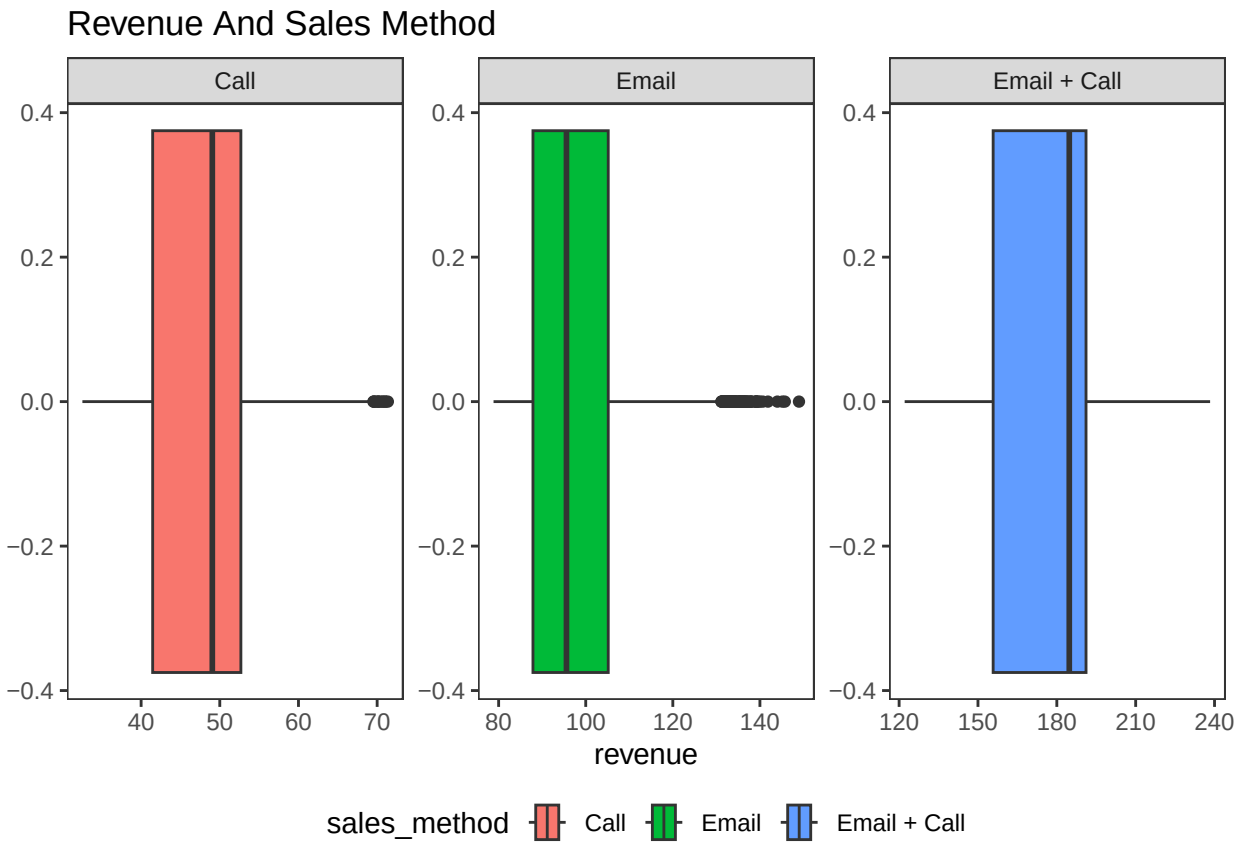
2.What does the spread of the revenue look like overall? And for each method?

```
sales_clean %>%
 ggplot(aes(revenue, fill=sales_method))+
 geom_density()+
  ggtitle('Revenue Distibution')
```



Calls is associated with lower Revenues in comparison to other methods. Both **Email** and (**Email+Call**) generates more revenues

```
library(patchwork)
 sales_clean %>%
  ggplot(aes(revenue, fill=sales_method)) +
  geom_boxplot() +
  facet_wrap(~sales_method, scales = "free")+
   theme(legend.position = 'bottom')+
   ggtitle('Revenue And Sales Method')
```



Low revenues were majorly generated by the calls method. This can be clearly seen on the Call graph above, with revenue ranging between 0-70

Email sales_approach generated revenues in the range of 80-120, with larger values start from 130-150.

Both (Email + Call) generate higher revenues ranging from 120-240 as seen from the histogram and boxplot for Email + call.

3.Was there any difference in revenue over time for each of the methods?

```
d <- sales_clean %>%
    group_by(sales_method) %>%
  summarise(Total_Revenue=sum(revenue)) %>%
  arrange(desc(Total_Revenue))

DT::datatable(d)
```

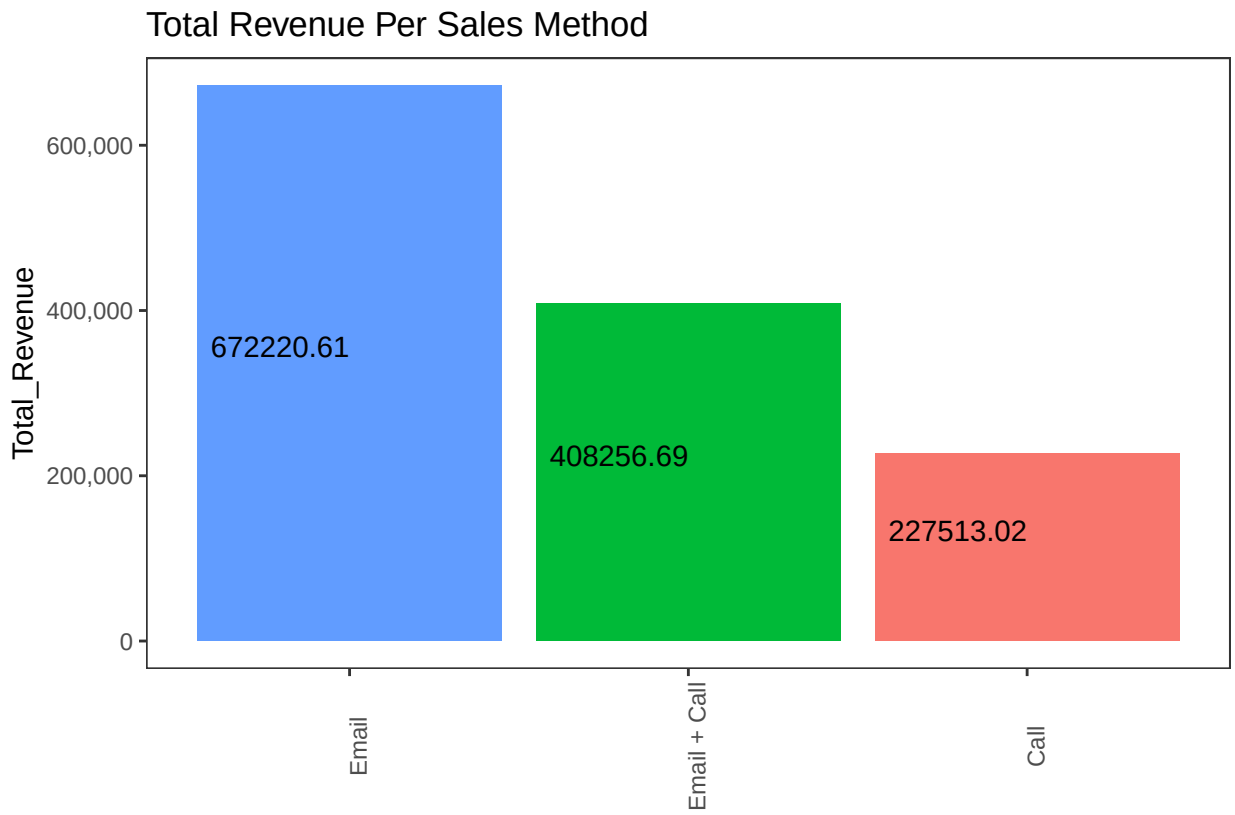
Show 10 entries Search:

	sales_method	Total_Revenue
1	Email	672220.61
2	Email + Call	408256.69
3	Call	227513.02

Showing 1 to 3 of 3 entries Previous 1 Next

```
library(scales)
d %>%
```

```
  ggplot(aes(x=reorder(sales_method, desc(Total_Revenue)),
             y=Total_Revenue, fill=rainbow(3)))+
  geom_col()+
  theme(legend.position = 'none') +
  scale_y_continuous(labels = scales::comma)+
  theme(axis.text.x = element_text(angle = 90))+
  geom_text(
    aes(label=Total_Revenue),
    position = position_stack(0.5),
    hjust=1,
    vjust=-.3
  )+
  xlab(' ')+
   ggtitle('Total Revenue Per Sales Method')
```




```
library(scales)
d1 <- sales_clean %>%
     group_by(week) %>%
       summarize(Total_revenue=sum(revenue)) %>%
       arrange(desc(Total_revenue))

DT::datatable(d1)
```

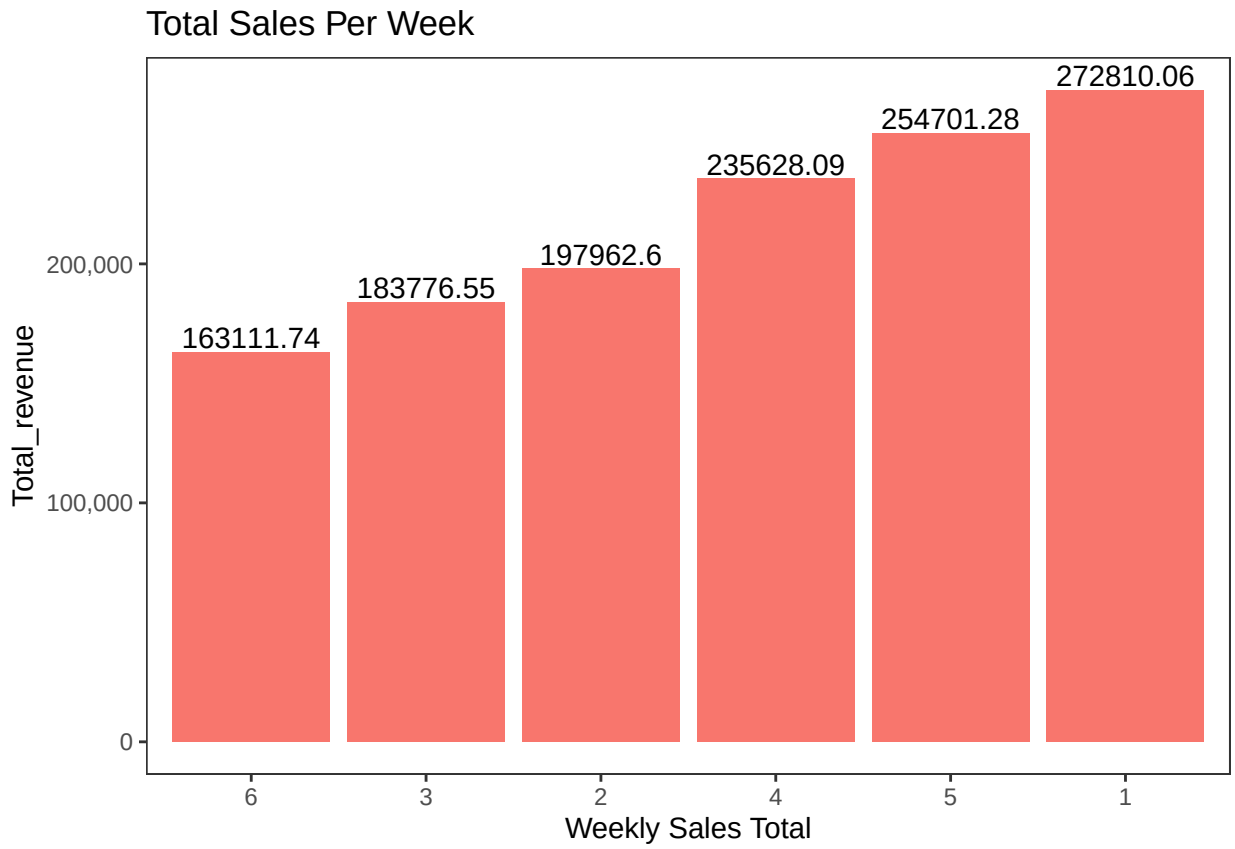
Show 10 entries Search:

	week	Total_revenue
1	1	272810.06
2	5	254701.28
3	4	235628.09
4	2	197962.6
5	3	183776.55
6	6	163111.74

Showing 1 to 6 of 6 entries Previous 1 Next

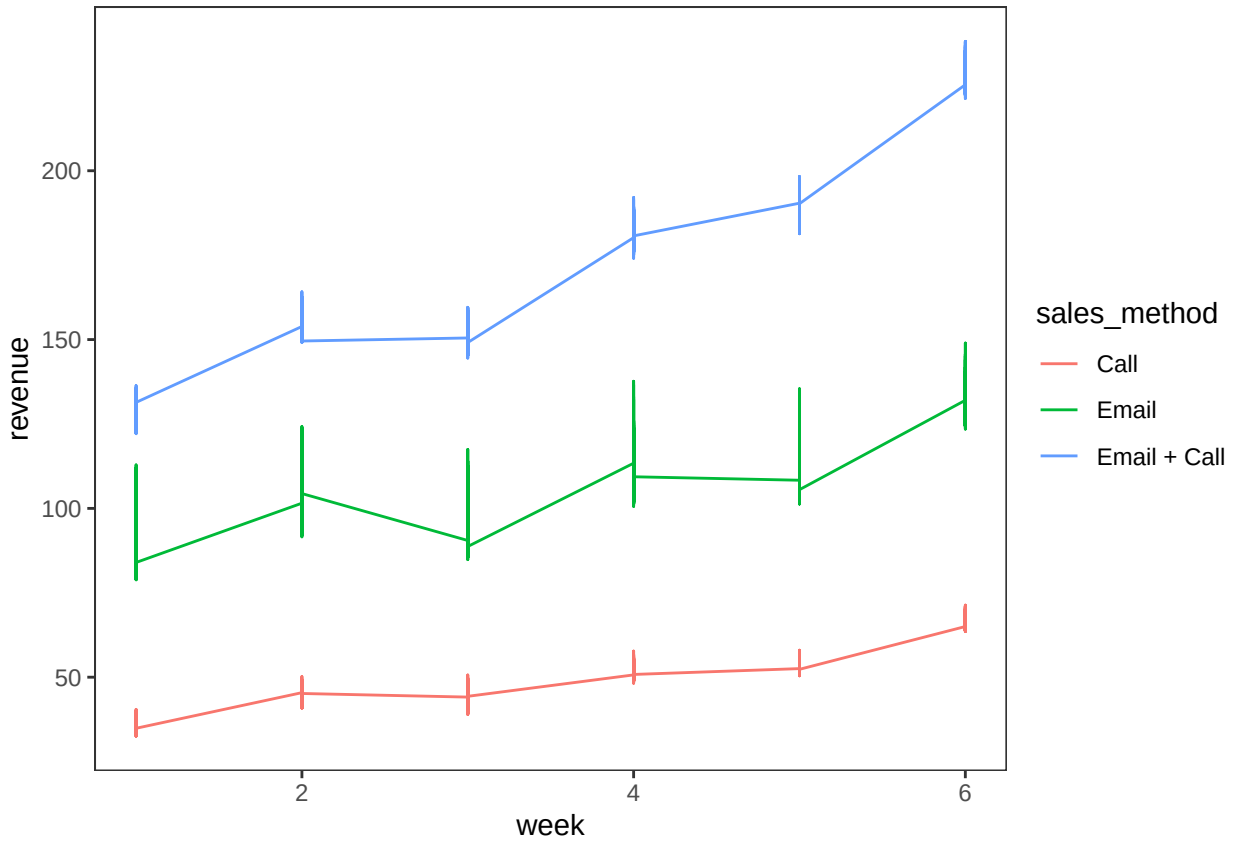
```
d1 %>%
  ggplot(aes(x=fct_infreq(reorder(week, Total_revenue)),
```

```
            y=Total_revenue,  fill='orange')) +
  geom_col() +
  scale_y_continuous(labels=scales::comma)+
  geom_text(aes(label=Total_revenue),
            position = position_stack(1),
           # hjust=1,
            vjust=-.2)+
  theme(legend.position = 'none')+
  xlab('Weekly Sales Total')+
  ggtitle("Total Sales Per Week")
```



4.Based on the data, which method would you recommend we continue to use?

```
sales_clean %>%
  ggplot(aes(x=week, y=revenue, color=sales_method))+
  geom_line()
```




```
# facet_wrap(~sales_method, scales = "free")
```

The Business Metrics

Based on the analysis, I recommend discontinuing the Calls method and focusing only on the Email and Email + Call sales method. This is due to the higher sales and revenue generated by these approaches, as well as the shorter average time required per sale compared to calls approach which is (30 minutes). However, the Calls approach can still be used on condition the customer doesn't have an email address.

The Recommendations

Based on the analysis conducted using the provided data, the following recommendations are proposed:

- Monitor key metrics to track any changes in the sales approach.
- It is recommended to utilize the Email method frequently to inform customers about new products. Additionally, follow-up calls in the second and third week can be made to discuss their needs and how the new product will assist them.
- It is advisable to minimize the usage of the Call method or eliminate it altogether. This approach consumes more time for sales and ultimately generates the lowest revenue, despite having a higher number of sales.

- The sales team should prioritize the Email and Email + Call approaches. As demonstrated in analysis, the Email sales approach yields the highest revenue during the initial three weeks, although it declines as the week progresses. To enhance sales and generate more revenue, a follow-up call should be made in the second or third week.
- To broaden the customer segment, focus on enhancing marketing strategies and improving the conversion rate based on website visits. As indicated in the correlation graph, the longer customer tenure corresponds to lower revenue. To address this, onboard new customers and establish customer retention initiatives to boost sales and revenue from both new and existing customers.
- Ensure accurate data collection to facilitate comprehensive analysis, particularly for revenue, which contains numerous missing values.