CERTIFICATION - DATA ANALYST - PRODUCT SALES

datalab

est.1984

Pens and Printers

Six weeks ago we launched a new line of office stationery. Despite the world becoming increasingly digital, there is still demand for notebooks, pens and sticky notes.

We have tested three different sales strategies for this, targeted email and phone calls, as well as combining the two.

Email: Customers in this group received an email when the product line was launched, and a further email three weeks later. This required very little work for the team.

Call: Customers in this group were called by a member of the sales team. On average members of the team were on the phone for around thirty minutes per customer.

Email and call: Customers in this group were first sent the product information email, then called a week later by the sales team to talk about their needs and how this new product may support their work. The email required little work from the team, the call was around ten minutes per customer.

We need to know:

- How many customers were there for each approach?
- What does the spread of the revenue look like overall? And for each method?
- Was there any difference in revenue over time for each of the methods?
- Based on the data, which method would you recommend we continue to use? Some of these methods take more time from the team so they may not be the best for us to use if the results are similar.

other differences between the customers in each group?

Column Name	Details
week	Week sale was made, counted as weeks since product launch
sales_method	Character, which of the three sales methods were used for that customer
customer_id	Character, unique identifier for the customer
nb_sold	Numeric, number of new products sold
revenue	Numeric, revenue from the sales, rounded to 2 decimal places.
years_as_customer	Numeric, number of years customer has been buying from us (company founded in 1984)
nb_site_visits	Numeric, number of times the customer has visited our website in the last 6 months
state	Character, location of the customer i.e. where orders are shipped

We have some steps to take first:

- Check the data into a workbook
- Cleaning the data

```
library(readr)

data <- read_csv("product_sales.csv")


Rows: 15000 Columns: 8
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (3): sales_method, customer_id, state
dbl (5): week, nb_sold, revenue, years_as_customer, nb_site_visits

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
clean_data <- data %>%
  rename_with(~ str_to_lower(.x)) %>%
  mutate(
    sales_method = str_trim(str_to_lower(sales_method)), # trim and lowercase

    # Fix common variants
    sales_method = case_when(
      sales_method %in% c("email", "e-mail") ~ "Email",
      sales_method %in% c("call", "phone call") ~ "Call",
      sales_method %in% c("email and call", "email + call", "email & call", "e-mail and call") ~ "Email
and Mail",
      TRUE ~ NA_character_  # anything else becomes NA
    ),

    # Continue cleaning other fields
    customer_id = as.character(customer_id),
    week = as.integer(week),
    nb_sold = as.integer(nb_sold),
    revenue = round(as.numeric(revenue), 2),
    years_as_customer = as.integer(years_as_customer),
    nb_site_visits = as.integer(nb_site_visits),
    state = str_to_upper(state)
  )
```

Great now that our data is cleaned we can take some closer looks!

```
library(dplyr)

customers_per_method <- clean_data %>%
  group_by(sales_method) %>%
  summarise(num_customers = n_distinct(customer_id)) %>%
  arrange(desc(num_customers))

print(customers_per_method)

# A tibble: 3 × 2
  sales_method   num_customers
  <chr>                  <int>
1 Email                   6922
2 Call                    4781
3 Email and Mail          2203
```

```
library(readr)
library(dplyr)
library(stringr)
library(ggplot2)

# Load data
data <- read_csv("product_sales.csv")

# Clean the data
clean_data <- data %>%
  rename_with(~ str_to_lower(.x)) %>%
  mutate(
    sales_method = str_trim(str_to_lower(sales_method)),
    sales_method = case_when(
      sales_method %in% c("email", "e-mail") ~ "Email",
      sales_method %in% c("call", "phone call") ~ "Call",
      sales_method %in% c("email and call", "email + call", "email & call", "e-mail and call") ~ "Email
and Mail",
      TRUE ~ NA_character_
    ),
    revenue = round(as.numeric(revenue), 2)
  ) %>%
  filter(!is.na(revenue), !is.na(sales_method))  # remove missing values for analysis
```

```
Rows: 15000 Columns: 8
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (3): sales_method, customer_id, state
dbl (5): week, nb_sold, revenue, years_as_customer, nb_site_visits

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

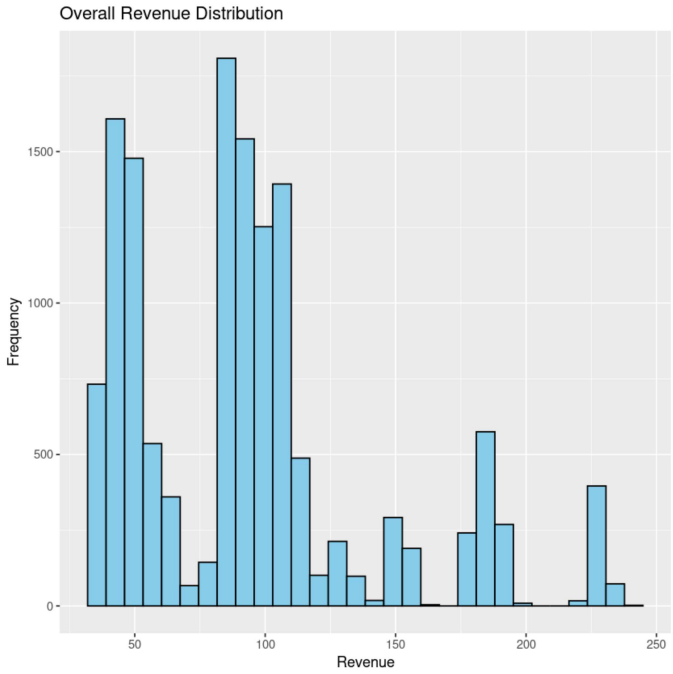
```
# Overall summary
summary(clean_data$revenue)

# Summary by sales method
clean_data %>%
  group_by(sales_method) %>%
  summarise(
    count = n(),
    min_revenue = min(revenue),
    q1 = quantile(revenue, 0.25),
    median = median(revenue),
    mean = mean(revenue),
    q3 = quantile(revenue, 0.75),
    max_revenue = max(revenue),
    sd = sd(revenue)
  )
```

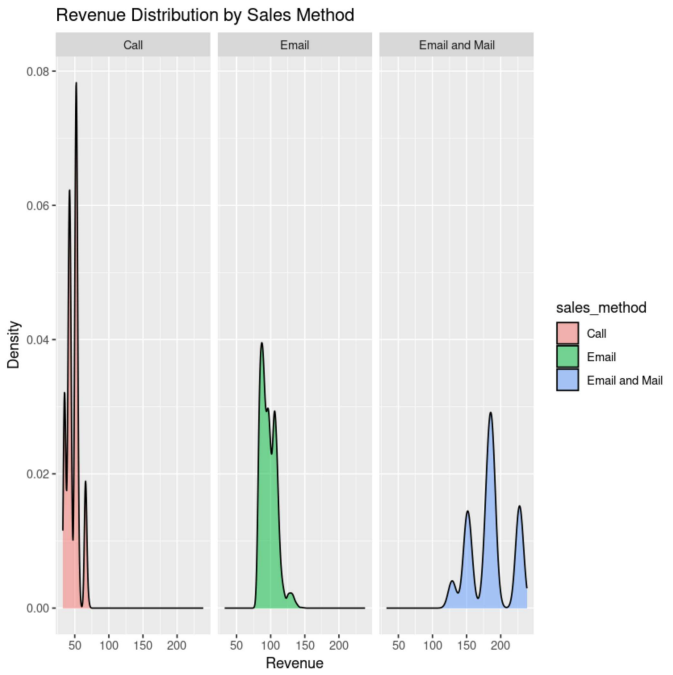
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  32.54   52.46   89.47   93.82  107.23  238.32
```

...	sales_m...	...	...	mi...	...	...	...	...	...	...	ma...	...	...
1	Call	4781	32.54	41.47	49.07	47.5975	52.68	71.36	8.6
2	Email	6922	78.83	87.88	95.58	97.1277	105.17	148.97	11.2
3	Email and Mail	2203	122.11	155.775	184.77	183.7436	191.215	238.32	29.

Rows: 3 Expand

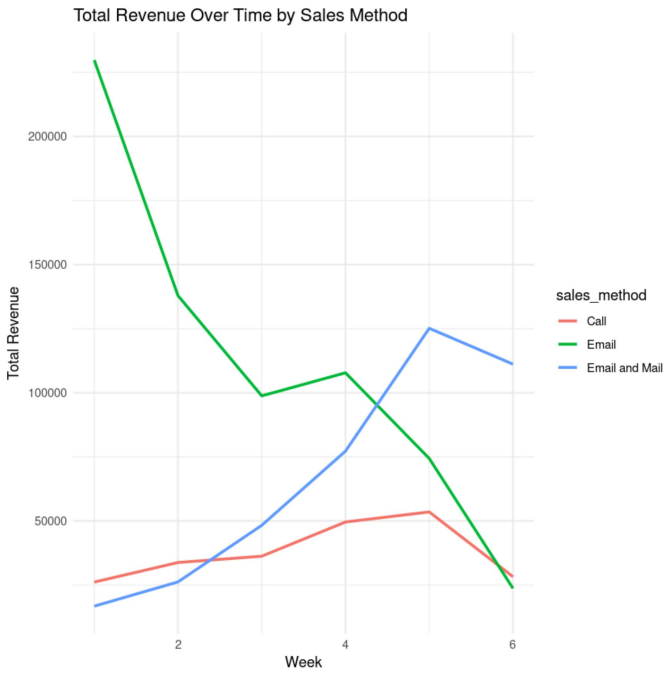
```r
# Histogram - Overall
ggplot(clean_data, aes(x = revenue)) +
  geom_histogram(bins = 30, fill = "skyblue", color = "black") +
  labs(title = "Overall Revenue Distribution", x = "Revenue", y = "Frequency")

# Density Plot by Sales Method
ggplot(clean_data, aes(x = revenue, fill = sales_method)) +
  geom_density(alpha = 0.5) +
  labs(title = "Revenue Distribution by Sales Method", x = "Revenue", y = "Density") +
  facet_wrap(~sales_method)
```



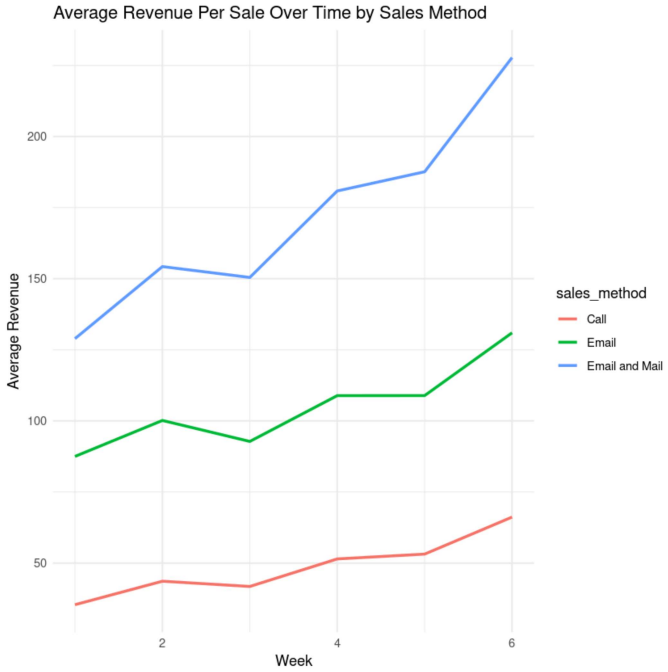


```r
library(dplyr)
library(ggplot2)

# Summarize total revenue per week per method
revenue_by_week <- clean_data %>%
  group_by(week, sales_method) %>%
  summarise(
    total_revenue = sum(revenue, na.rm = TRUE),
    avg_revenue = mean(revenue, na.rm = TRUE),
    .groups = "drop"
  )
```

```r
ggplot(revenue_by_week, aes(x = week, y = total_revenue, color = sales_method)) +
  geom_line(linewidth = 1) +
  labs(title = "Total Revenue Over Time by Sales Method",
       x = "Week",
       y = "Total Revenue") +
  theme_minimal()
```



```
ggplot(revenue_by_week, aes(x = week, y = avg_revenue, color = sales_method)) +
  geom_line(linewidth = 1) +
  labs(title = "Average Revenue Per Sale Over Time by Sales Method",
       x = "Week",
       y = "Average Revenue") +
  theme_minimal()
```



Great these are our major factors ready for analysis!

but lets grab some other numbers just in case

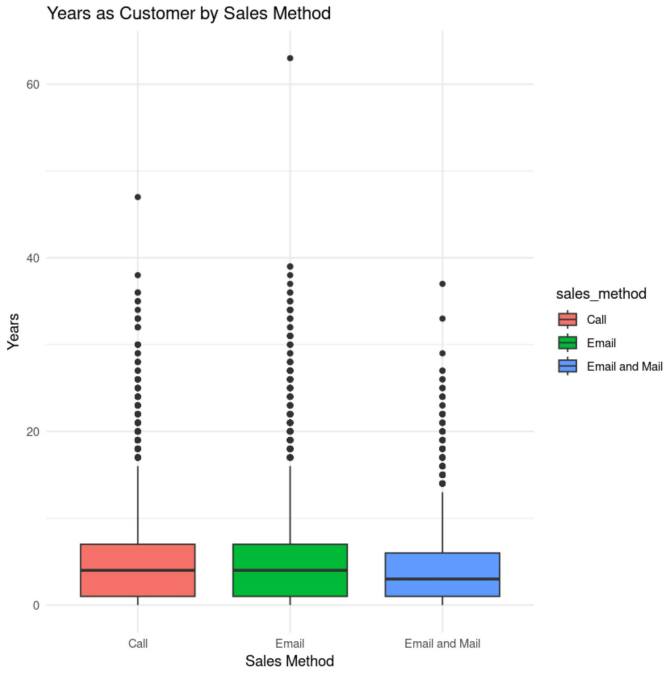
- years as customer
- site visits
- number sold
- revenue

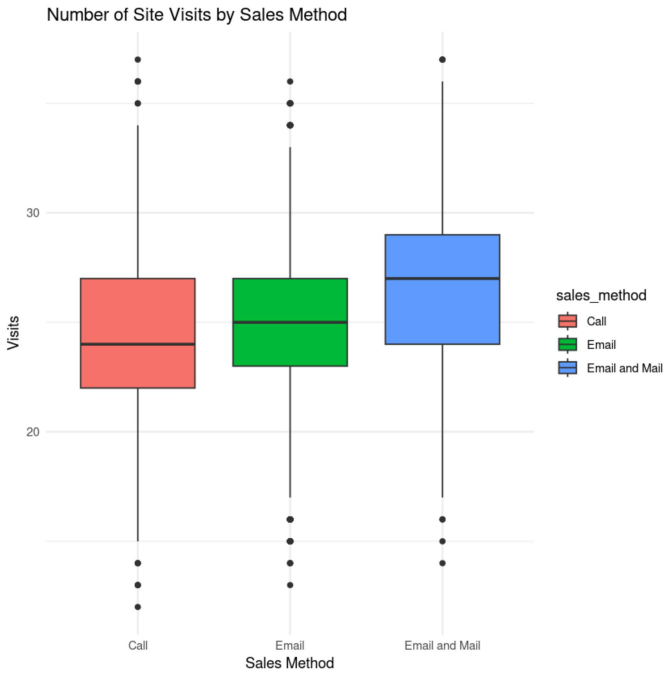
are all pertinent factors.

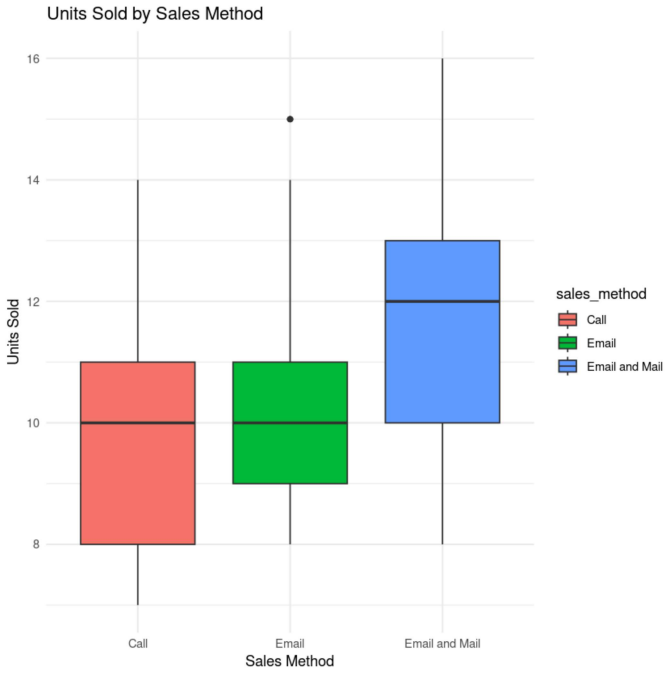
```r
library(dplyr)

# Compare customer features by sales method
customer_summary <- clean_data %>%
  group_by(sales_method) %>%
  summarise(
    avg_years_as_customer = mean(years_as_customer, na.rm = TRUE),
    avg_site_visits = mean(nb_site_visits, na.rm = TRUE),
    avg_units_sold = mean(nb_sold, na.rm = TRUE),
    avg_revenue = mean(revenue, na.rm = TRUE),
    count = n()
  )

print(customer_summary)
```

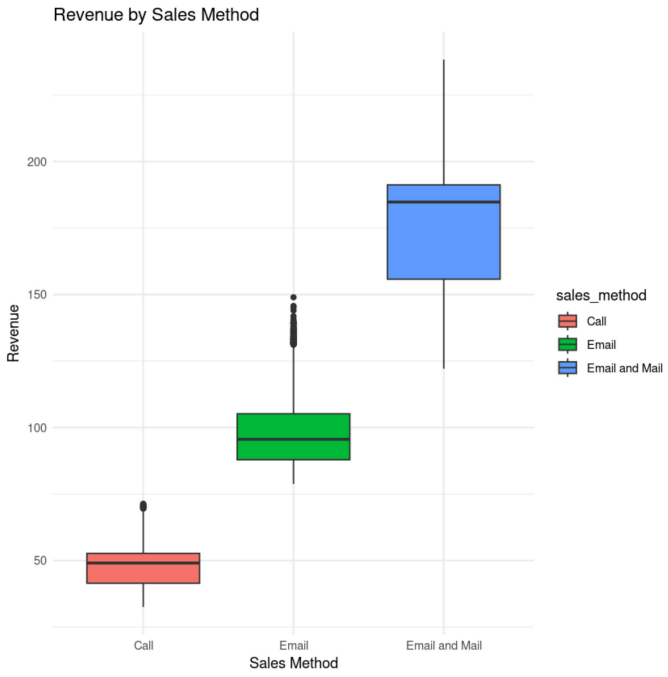
```
# A tibble: 3 × 6
  sales_method  avg_years_as_customer avg_site_visits avg_units_sold avg_revenue
  <chr>                         <dbl>           <dbl>          <dbl>       <dbl>
1 Call                           5.16            24.4           9.50        47.6
2 Email                          5.00            24.7           9.72        97.1
3 Email and Ma…                  4.52            26.7          12.2        184.
# ℹ 1 more variable: count <int>
```

```r
library(ggplot2)

# Years as Customer
ggplot(clean_data, aes(x = sales_method, y = years_as_customer, fill = sales_method)) +
  geom_boxplot() +
  labs(title = "Years as Customer by Sales Method", x = "Sales Method", y = "Years") +
  theme_minimal()

# Site Visits
ggplot(clean_data, aes(x = sales_method, y = nb_site_visits, fill = sales_method)) +
  geom_boxplot() +
  labs(title = "Number of Site Visits by Sales Method", x = "Sales Method", y = "Visits") +
  theme_minimal()

# Units Sold
ggplot(clean_data, aes(x = sales_method, y = nb_sold, fill = sales_method)) +
  geom_boxplot() +
  labs(title = "Units Sold by Sales Method", x = "Sales Method", y = "Units Sold") +
  theme_minimal()

# Revenue
ggplot(clean_data, aes(x = sales_method, y = revenue, fill = sales_method)) +
  geom_boxplot() +
  labs(title = "Revenue by Sales Method", x = "Sales Method", y = "Revenue") +
  theme_minimal()
```




Number of Site Visits by Sales Method
Visits
30
20
Call
Email
Email and Mail
Sales Method
sales_method
Call
Email
Email and Mail


Units Sold by Sales Method
Units Sold
16
14
12
10
8
Call
Email
Email and Mail
Sales Method
sales_method
Call
Email
Email and Mail




```
# ANOVA test for revenue by sales method
anova_result <- aov(revenue ~ sales_method, data = clean_data)
summary(anova_result)

# Tukey's HSD post-hoc test if needed
TukeyHSD(anova_result)
```

```
                Df   Sum Sq  Mean Sq F value Pr(>F)
sales_method     2 28104463 14052232   63117 <2e-16 ***
Residuals    13903  3095350      223
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = revenue ~ sales_method, data = clean_data)

$sales_method
                           diff       lwr       upr p adj
Email-Call             49.53022  48.87252  50.18791     0
Email and Mail-Call   136.14613 135.24553 137.04674     0
Email and Mail-Email   86.61592  85.76037  87.47146     0
```

These graphs give us a full picture of our data, and the anova may not be used but it is helpful to have!