DataCamp Data Analyst Professional Certification

Nora Anzawi

Overview: Pens and Printers

Established in 1984, Pens and Printers provides quality office products to states in the US nationwide.

The primary business goal is to develop a sales strategy for the product launch of a new line of office stationery. This includes:

- Assessing revenue generated across the different sales methods used (email, call, email+ call)
- Difference in customers across various groups (years as customer, state) and if this impacts revenue
- What approach or method should the company use to increase cost-effectiveness regarding revenue and time spent on each sales strategy

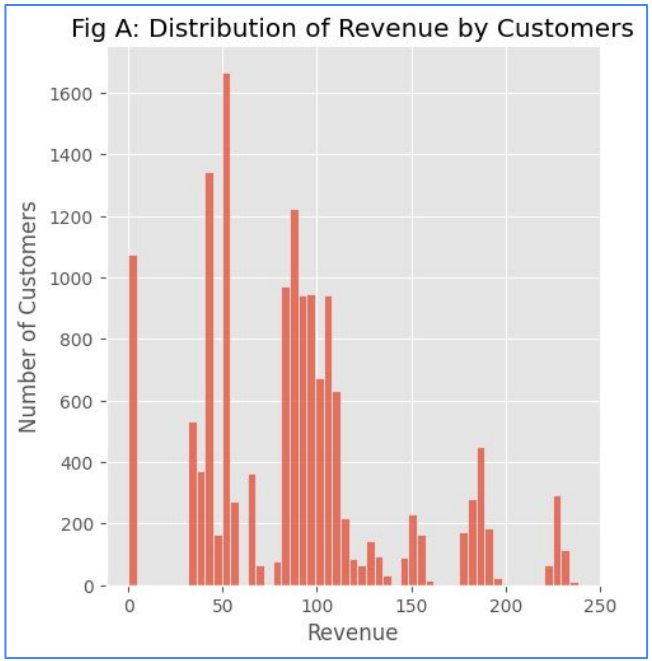
Summary of Work - Data Validation

	type_	nunique	nempty
week	int64	6	0
sales_method	object	5	0
customer_id	object	15000	0
nb_sold	int64	10	0
revenue	float64	6743	1074
years_as_customer	int64	42	0
nb_site_visits	int64	27	0
state	object	50	0

Reviewed dataframe for datatypes and missing values

- All values matched the description
- One column had missing values - `revenue`
  - Out of 15000 observations given, 1074 were NaN. Filled these with 0 and created a categorical column `sales_made`
  - Two observations fell outside of the timeline for `years_as_customer` if company established in 1984
- Cleaned up `sales_method` column
- Created a `sales_times` column to estimate how long it took for each `sales_method`

Summary of Work - Exploratory Data Analysis



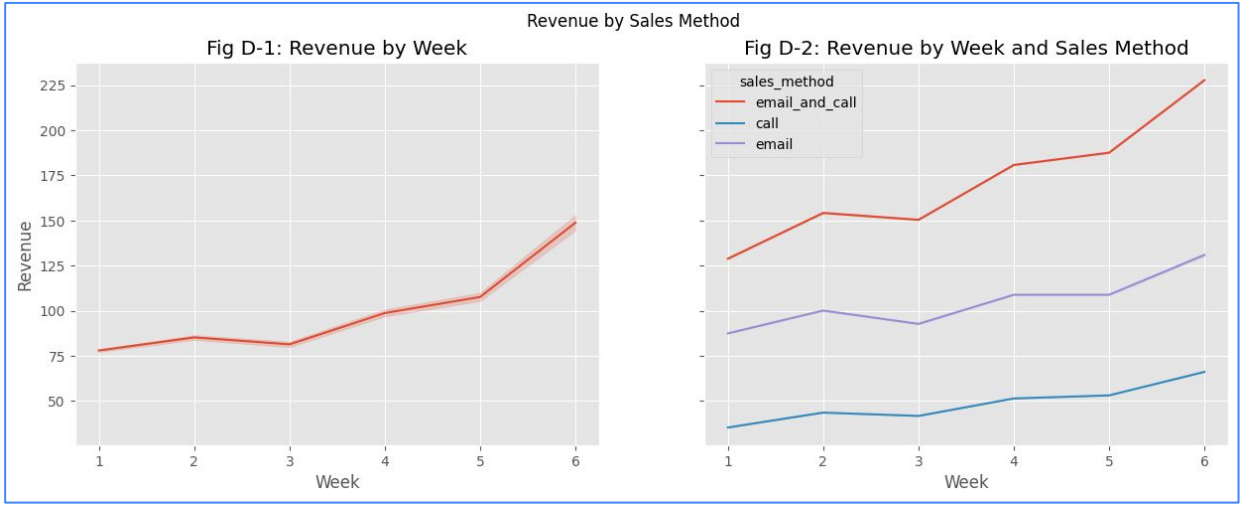
Distribution of Revenue

- Fig A: Distribution of revenue is skewed to the right, although summary stats only show `years_as_customer` being the variable where the median > mean.
- Fig B-1 and B-2: Difference in revenue generated by `sales_method`
  - Violinplots capture the distribution masked for whether `sales_made` was equal to "Yes", filtering out the observations that had no revenue, or 0 as a value.
    - When revenue == 0 remained, distribution of data skewed much more

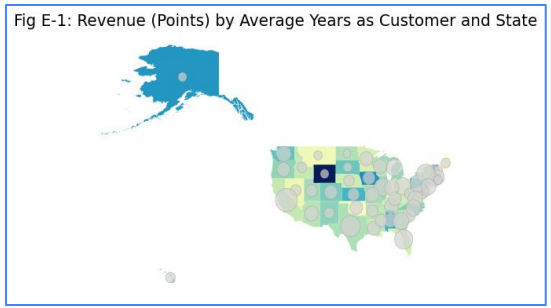
Summary of Work - Exploratory Data Analysis

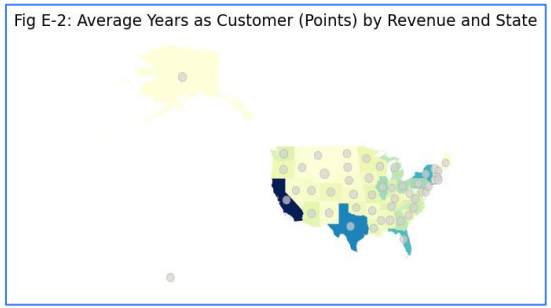
Revenue by Week

- Figures D-1 and D-2: Show revenue distribution by week overall, and by sales method.
- There is a noticeable uptick in revenue distribution overall and by sales method during week 5 and 6, with most profit generated by `sales_method` == `email_and_call`.
- Email the most cost-effective for customers overall and generated most revenue.



Summary of Work - Exploratory Data Analysis





Revenue by Location

- Fig E-1: Revenue (Points) by Average Years as Customer and State areas using natural breaks method. with a larger average customer base have fewer revenue sales where this occurs
- Fig E-2: Average Years as Customer (Points) by Revenue and State using quantiles method. Revenue is concentrated heavily on costal states and there doesn't appear to be a significant difference in average years as customer.
- Figures F extracting data by `sales_method` The location of observations showed little variance regarding the spread of revenue. These visuals are not included in this presentation.

Key Findings

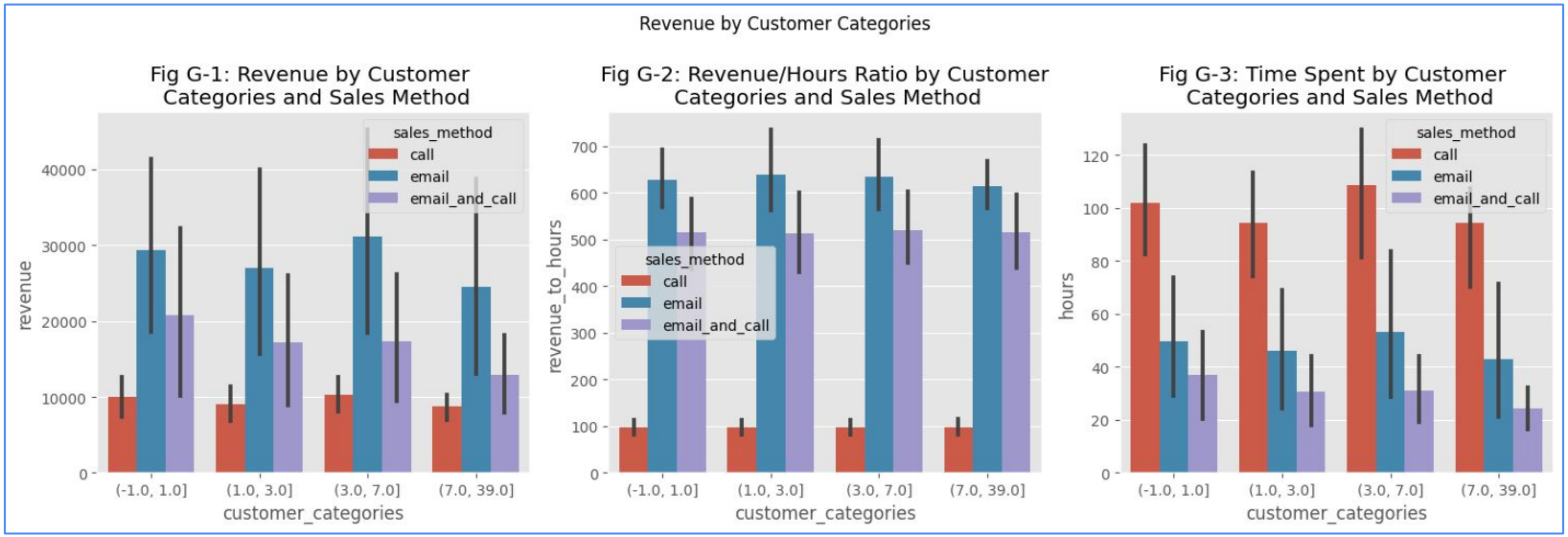
- Interest about revenue and curiosity about the `sales_method` used due some methods using more staff time.

- With this in mind, if cost-efficiency is important to understand our customer base (ex: `state` and `years_as_customer`) my suggestion is to focus on the proportion of `revenue` generated by time spent for each `sales_method`.

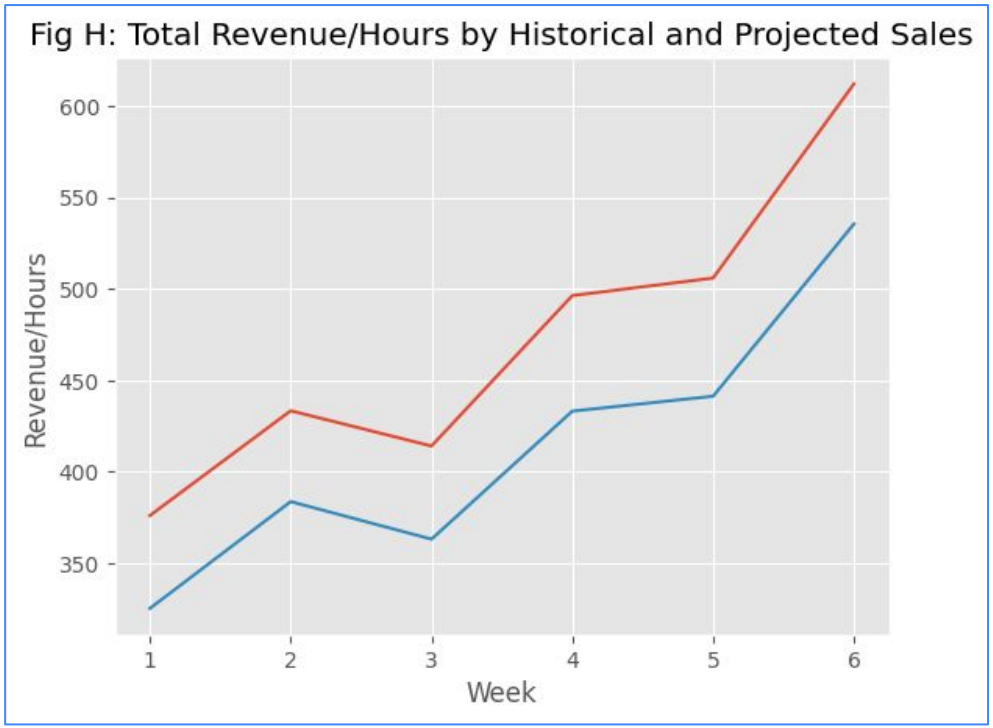
- I suggest this because there appeared to be very little variation about the geographic location of observations recorded, but `years_as_customer` had a fairly disbursed range of customers for a pen and paper store in 2023.

Key Findings

- Since I've hypothesized the `sales_method` most time-consuming and least profitable is `call`, we could (1) analyze what customer categories are already spending the most and (2) assess what the preferred `sales_method` was for each group.
  - The first category of customers under one year (`(-1.0, 1.0]`) generated the most revenue and the most used sales_method was `emails_and_calls`.
  - Customers from 3 to 7 years (`(3.0, 7.0]`) generated the second most revenue and most used sales_method was `call`



Key Findings



- If we replicate the dataset projecting a different trend by `sales method`, we could replace `call and email` for `customer_categories == (-1.0, 1.0]` and `call` for `customer_categories == (3.0, 7.0]` along with the associated deductions on time spent, the savings could be astronomical.

- Figure H: Total Revenue/Hours by Historical vs Projected Sales shows an increase from $414 dollars of revenue/time spent earned per hours to $472 dollars of revenue/time spent earned per hour.

Final Recommendations

Business Focus: Get to know your new customers.

- Likely many differences worth capturing not in this data
- Worthwhile to understand customers that generated no revenue, while the top `sales_method` for them was the most used and revenue-generating:`email`.

Sales Strategy: No calling.

- Ultimately, for business metrics my suggestion is to focus on revenue/time spent, and eliminate calls as a `sales_method`.
- Tracking observations for both email and call for customers with `sales_method` categorized as `email_and_call` would be helpful to further breakdown the revenue/hours ratio.
- Look into site visits by customer age, or years as customer. There might be a correlation between this and nb_site_visits which could aid a marketing campaign