Technical Report

Motivation for Choosing the Topic

I chose this subject for my project because I recently had an unpleasant experience at a veterinary clinic and wanted to investigate the industry further. I noticed a clear difference between 2022, when I last visited the clinic, and 2025, when I visited again. In the interim, I was living in Poland and took my cat to Polish veterinary clinics. After I noticed the change in care (massively increased prices, large focus on billing and extra procedures, rigid policies) I did some research and learned they were acquired by VetCor.

Dataset Origin and Description

> **`Pet-Related Consumption Data`**

Link to Data

This dataset provides a detailed breakdown of how much money American consumers have spent on various categories of products and services over a selected time period.

> The specific row used for the chart is from:
>
> **Table 2.4.5U. Personal Consumption Expenditures by Type of Product**
> *Veterinary and other services for pets*
>
> This data shows what US consumers spend on their pets over time and was the easiest dataset to find and process.

> **`Vet Clinic Count Data`**

Link to Data

> Getting an accurate count of vet clinics over time was a time-consuming process. I had to navigate several different pages for each year to download the corresponding file. The files often had different naming conventions, and some data was provided as raw text that I had to manually save into a file.
>
> One of the challenges when working with multiple datasets over long periods is that the codes used to reference data can change. In this instance, the codes referring to vet clinics changed in 1998 and 2007.
>
> Before 1998, the Standard Industrial Classification (SIC) system was in use. Under the old system codes starting with `074` corresponded to 'Veterinary Services'.
>
> Starting in 1998 the North American Industry Classification System (NAICS) is used. Under the NAICS system the code `541940` specifically refers to "Veterinary Services."
>
> After 2007, an additional column, `lfo` (Legal Form of Organization) may be added. When this is included it can change the aggregate count and needs to be accounted for.
>
> The following Python code was used to handle this:

```python
vet = None
if 'naics' in df.columns:
    # (1998+)
    vet = df[df['naics'] == '541940'].copy()
    # After 2007
    if year > 2007 and 'lfo' in vet.columns:
        vet = vet[vet['lfo'] == '-']
#before 1998
elif 'sic' in df.columns:
    vet = df[df['sic'].str.startswith('074', na=False)].copy()
```

> **`Consolidation Data`**
>
> The last set of data, for the consolidation graph, was created by searching through reports, articles and other online media to find particular counts on specific dates and was then compiled into a basic CSV file. Finding values before 2010 was difficult and reduced the impact of the argument being made.
>
> The original scope was to identify individual clinics acquired by the 20 largest conglomerates. This proved to be infeasible within the project's timeframe, as it would have required a much more extensive data collection effort (I'm not sure it is even possible at this point).
>
> Once it became clear that per-clinic data was unavailable, I initially struggled with how to best format the aggregated data. I wanted to create a Sankey-like diagram, but representing both ownership changes and clinic counts simultaneously was challenging.
>
> I spent a lot of time trying to get the Sankey diagram to do something it's not really designed for. Plotly's Sankey diagram is very limited in its implementation; I had almost no control over node position or intuitively adding other elements to the graph. For the final version, I added a basic timeline and then manually positioned the nodes to approximate their correct placement on the timeline. I also had to add 'placeholder' rows to get the diagram to render correctly; I feel like this pollutes my data, but I couldn't figure out another solution.

Cleaning/preprocessing

Most of the pre-processing steps were done during the data acquisition phase. The only dataset that really need pre-processing was the `Personal Consumption Expenditures by Type of Product` table. I had to rename a column, drop a column, and remove an empty row.

```python
df = df.rename(columns={'Unnamed: 1': 'Category'})

# remove line column
df.drop('Line', axis=1, inplace=True)

# remove row 0
df = df.drop(df.index[0]).reset_index(drop=True)
```

Visualization choices

The overall color scheme for the graphs was meant to evoke thoughts of vet health care.
- `#FFF`
- `#D84040`
- `#ECDCBF`

The Sankey graph was largely too complex to include in the color scheme, so I used the defaults.

Source for color scheme at Color Hunt: link

Both the bar chart for `Clinic Count` and the line chart for `Pet-Related Spending` were chosen because they were the most obvious choice for the data presented. The Sankey diagram was chose because the idea I had for presenting consolidation was a bunch of lines merging into larger and larger lines, and the Sankey diagram was the closest to that. I think I would have to write something entirely custom in D3.js if I wanted to present my original vision of 30,000 vet clinics merging into 20 conglomerates.

Reflection

I think starting with the data, then making inferences and insights is vastly easier than starting with an assumption (in this case, that the consolidation of vet clinics under massive conglomerates is harming pet owners and pet outcomes) and trying to find the data to support it. More often than not, the data simply does not exist, or it's behind a paywall that is primarily targeting investors (and costs thousands of dollars to access).

This was a good exercise for 'investigating' and working with real-world data, rather than the sanitized, organized, and largely unrealistic datasets from sources like Kaggle. However, the scope of information necessary to really answer the questions I started with was beyond what I could accomplish in the given time; and maybe beyond my skill-set.

If I had had access to all the data I needed at the start of the project, I would have spent the majority of time writing a custom chart in place of the Sankey diagram.

> ### AI Usage
>
> - AI (ChatGPT & Claude) was almost exclusively used to find data.
> - Claude was used to check for grammar, spelling, and overall conformity to the assignment requirements.
> - AI generated code was noted inline