Process book_data vis

2024年4月21日 星期日 上午4:10

Title: Pet Data Insights

Part 1	Part 2
Pet Data CSV	Reddit Data Crawling
2022	2023.4 — 2024.4

Part 1:

Shiny App:

- Main steps:

1. create separate shiny apps ... & merge into one major app. then merge w/ part 2 together to a final integrated app

2. Data Wrangling:
clean & merge csv files to a single df

3. The initial variables we wanted to investigate are vaccination rate, expenditure, adoption rate, microchip rate.

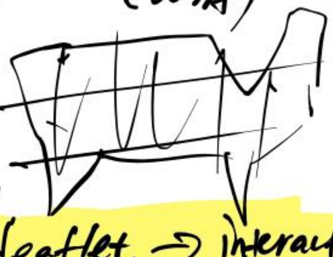
BUT During data wrangling, we decided not to simply visualize these two variables. Instead we divided return ÷ adoption → adoption rate

1. expenditure → distribution of expenditure
⇓
which visualization to use?

? possible options: stacked bar ✓
grouped bar → too many states
tree map (×) → effect not good

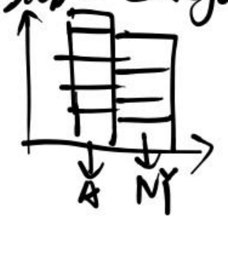
5. Main Plots:

① choropleth geo plots (USA) ⟸ vaccine, microchip, return rates, total score



leaflet → interactivity

② stacked bar plot for "expenditure"
in order to display sub-cost of each sub-category



interactivity rendered by shiny.
The User can select specific state data to display (≥1) or "All"

4Challenges (Solved)   (hawii & Alaska are exclued)

1. Since the data set only has 48 state data "map" function cannot be used directly. ⇒ Manually added geometry, lat, long data to the dataset.

st_
states_data + state_st = states_data_st
(geometry)
polygon

Leaflet choropleth:

- Challenges: adding the ABBR on the Static geomap, solved by manually adding abbrs to the dataset

(Solved)

- tweak the legend for expenditure ⇒ Fail to make it shiny because of the legend

⇓

simple leaflet html plot for displaying dog vs. cat population ratio in 2022

- change the pal multiple times in order to keep the paletee consistent and obvious & * express sentiment & emotions.

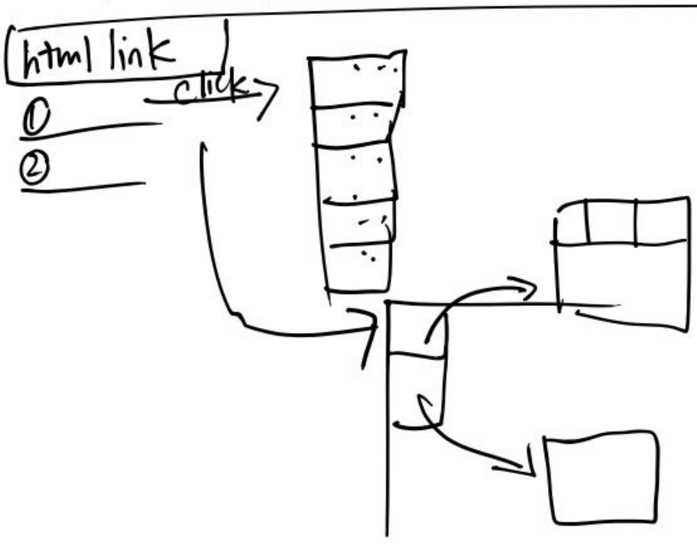
eg. adoption- return rates is just wanna make the stase w/ highest return rates obvious

Adjust the UI height too 600, put the legand at the bottom right

if we finally merge 2 shiny apps to one, how to embed the html leaflet? (Dog vs Cat population)

↓

embed html to the main dashboard title?



Part 2 : Reddit Data Analysis & DV

Step 1: How to get the data (So hard!!!!!)

① : Registered for Reddit API to directly fetch data

⇒ Failed , due to connection issues. ☹

② Searched online for existing code

⇒ Even worse , Recent updates to Reddit's terms imposed stricter data retrieval limits and robust anti-Scraping measures

② Found a script for the real time data scraping

⇒ Success , but not access historical

④ Further efforts : Combined Several online code example.

⇒ Get Some past data , limits 1000 record , within a year.

Final strategy : Created three Separte loops to extract . data by year, month, week within given restrictions .

This is the maxium data collection.

Step 2: Data cleaning

① remove duplicates, special characters / space

② remove basic stop word & name

③ Extracted three detailed. articles from a pet-focused website.

Clean this three. txt, just like what. I did in ②

Select Top 100 most frequency words.

generic pet specific stop_word list

④ clean data again.

Step4 : Data Analysis

1、Topic Model ⇒ using Data 2023/04 - 2024/04

① creat DTM

② using LDA to get topic. Words. (5)

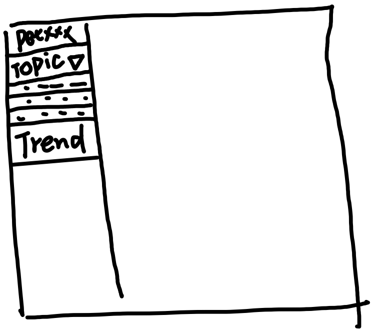
③ Calculate topic word Frequency

④ Calculate. coherent score .

2. Topic Trend Analysis ⇒ using data segmenting into season

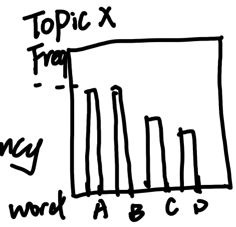
Just do what 1. Topic Model did for each season.

Step 5 ☆ Data visualization ☆ ⇒ Shiny!

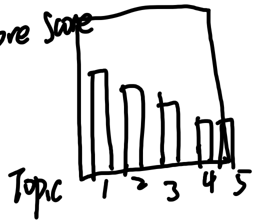


Topic Model

Topic name & word frequency

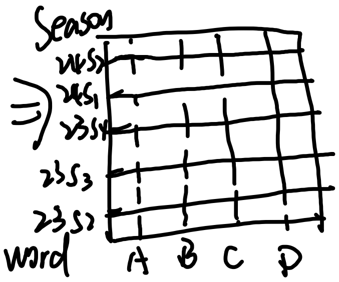


coherent score



word cloud

Trend ⇒



heat