Summarizing Categorical Data

DASC 512

Overview

- Tabular Summaries
- Bar Graph
- Pie Chart

Categorical or Qualitative Data

Categorical data is typically non-numerical and describes a category to which the experimental unit belongs

- Think of a dog: breed, sex, eye color, pointy or floppy ears, etc.

Class: Each category for classification of qualitative data

Class Frequency: Number of observations falling into a particular class

Class Relative Frequency: Class Frequency divided by total number of observations in the data set $(0 \leq \hat{p} \leq 1)$

Class Percentage: The Class Relative Frequency multiplied by 100%

Contingency Table

- A contingency table or frequency table is a tabular summary of class frequency

Car Color	Frequency
Black	7705
Silver	6805
Blue	5802
White	4212
Grey	3751
Other	10209

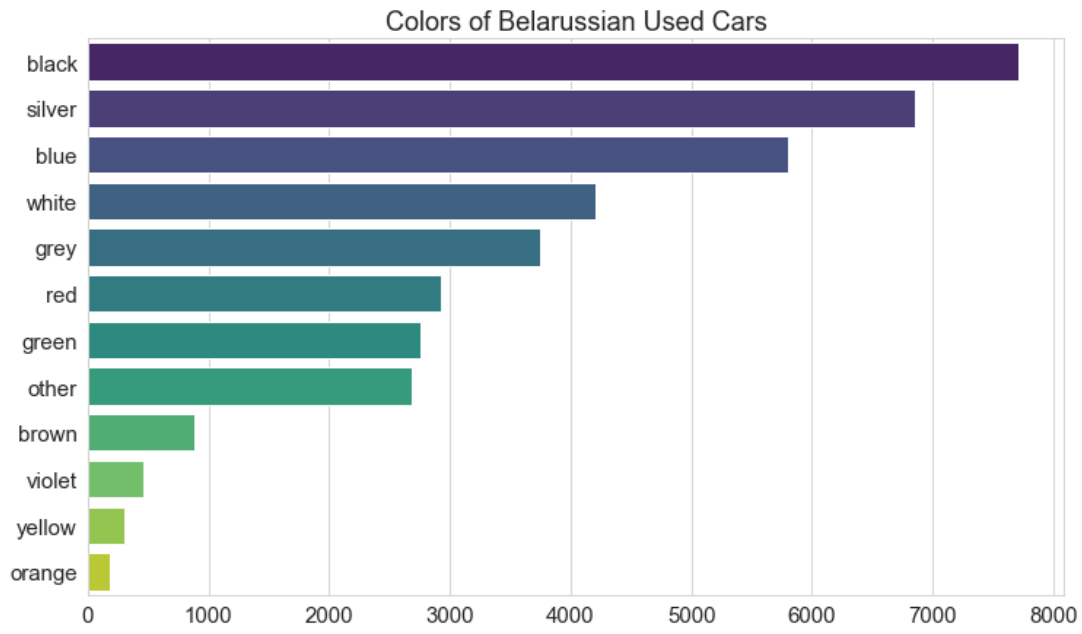
Relative Frequency Table

- A relative frequency table is a tabular summary of relative frequency

Car Color	Relative Frequency
Black	0.20
Silver	0.18
Blue	0.15
White	0.11
Grey	0.10
Other	0.26

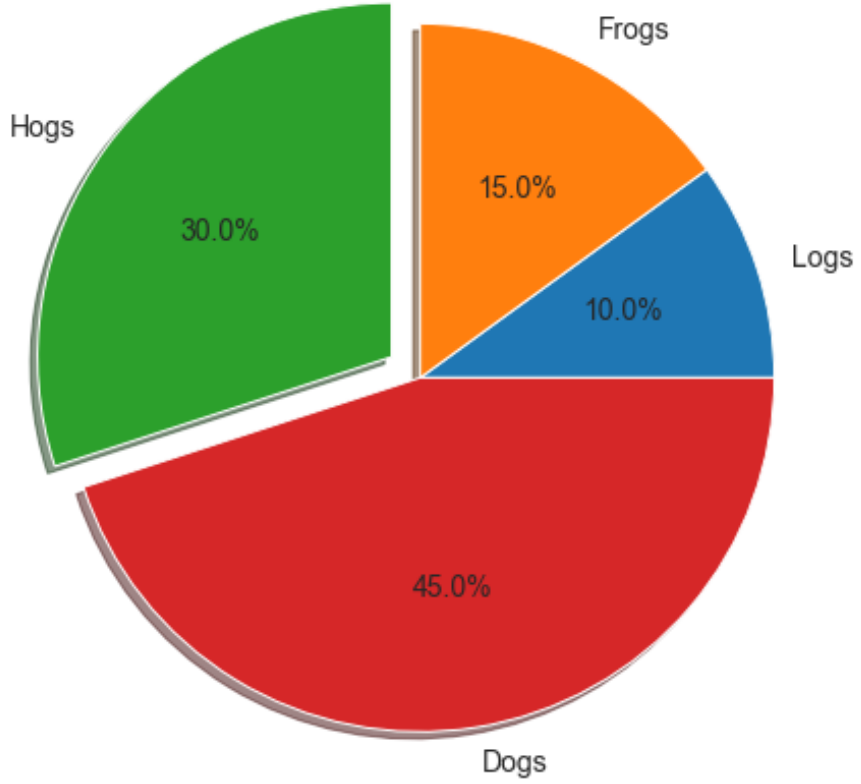
Bar Graph

- Great for comparing class frequency
- Flexible for many situations
- Easily interpreted by anyone



Pie Chart

- Commonly used alternative to bar graphs
- Very rarely is a pie chart the best choice
  - https://kristinhenry.medium.com/in-defense-of-pie-charts-and-why-you-shouldnt-use-them-df2e8ccb5f76
- Pie charts in Python aren't the most flexible



Resources

- Matplotlib Example Gallery
  - https://matplotlib.org/stable/gallery/index.html
- Seaborn Example Gallery
  - https://seaborn.pydata.org/examples/index.html
- Pandas User Guide
  - https://pandas.pydata.org/pandas-docs/stable/user_guide/index.html

Recap

- Tabular Summaries
- Bar Graph
- Pie Chart