Visualization Techniques: Histograms

A Brief Introduction to Histograms

Visual analytics is defined as the “science of analytical reasoning supported by interactive visual interfaces” (Keim, Mansmann & Thomas, 2010, p.5). According to these authors, visual analytic tools allow those making decisions the flexibility and creativity necessary to gain insight into complex problems considering the incredible storage and processing capacity of today’s devices.

There are many visualization techniques used to answer analytical problems. Among them are line plots, histograms, heat-maps, etc. One of my favorites? Histograms. Kraska (2018) state that histograms “require bining the data into buckets and then performing an aggregation per bucket” (p.2153). First introduced by Pearson (1894), histograms are a representation of the distribution of numerical data, an estimate of the probability distribution of a continuous variable.

Building a Histogram

Python and its libraries are especially suitable for visual analytics techniques. Van der Walt et. al. (2014) state that the Numpy and Matplotlib libraries in Python can be used to easily perform operations such as building histograms. We will show an example of using the Matplotlib and Numpy libraries here following the guidance set forward by Kazarinoff (2018).

Import matplotlib, numpy and, if using Jupyter notebooks, make sure to use inline

```
import matplotlib.pyplot as plt
import numpy as np
%matplotlib inline
```

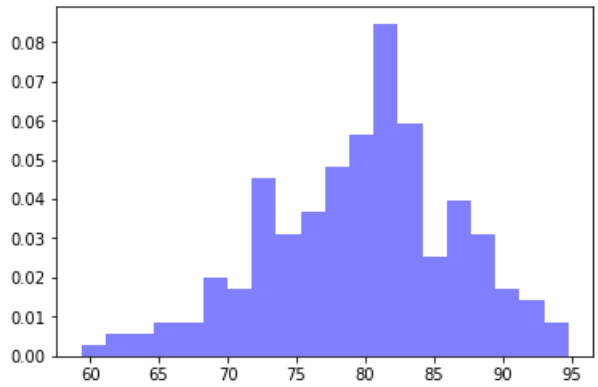
Define a mean and a standard deviation, then use a numpy function to create an array of random numbers with a normal distribution (i.e. 200 numbers).

```
mu = 80
sigma = 7
x = np.random.normal(mu, sigma, size=200)
```

Use plt.hist() to plot the histogram along with some keyword arguments.

```
plt.hist(x, 20,
        density=True,
        histtype='bar',
        facecolor='b',
        alpha=0.5)

plt.show()
```



Some Key Arguments for plt.hist() as shown by Kazarinoff (2018) are found below:

keyword argument	description	example
bins=	list of bin edges	bins=[5, 10, 20, 30]
density=	if true, data is normalized	density=false
histtype=	type of histogram: bar, stacked, step or step-filled	histtype='bar'
color=	bar color	color='b'
edgecolor=	bar edge color	color='k'
alpha=	bar opacity	alpha=0.5

How to Use to Answer Analytical Problems

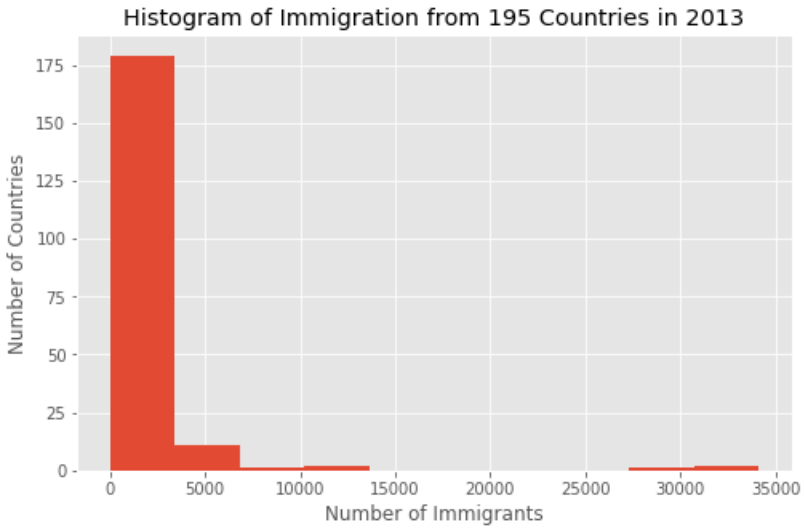
The following answer to an analytical problem using histograms is based on the work by Rajasekharan, Kermani, and Markovic (2018).

Question: What is the frequency distribution of the number (population) of new immigrants from the various countries to Canada in 2013?

```
df_can[2013].plot(kind='hist', figsize=(8, 5))

plt.title('Histogram of Immigration from 195 Countries in 2013') # add a title to the histogram
plt.ylabel('Number of Countries') # add y-label
plt.xlabel('Number of Immigrants') # add x-label

plt.show()
```



This histogram is skewed right. Most countries that have migrated to Canada are providing between 0 and 5 thousand immigrants each. A few countries have provided between 10 and 15 thousand migrants, and a handful of countries have high migration rates of 30 to 35 thousand migrants.

Key Strengths and Weaknesses

Based on the work by Biehler (2005).

Strengths.

- Allow viewers to easily compare data.
- Works well with large ranges of information.
- Provides a more concrete form of consistency as intervals are always equal.
- Great for dealing with large value ranges.

Weaknesses.

- It is extremely difficult and practically impossible to extract the exact amount of "input" in the histogram unless it is a frequency histogram.
- Histograms are often considered inconvenient when comparing multiple categories.

References

Biehler, R. (2005, February). Strength and weaknesses in students' project work in exploratory data analysis. In *Proceedings of the Fourth Congress of the European Society for Research in Mathematics Education, Sant Feliu de Guíxols, Spain–17-21 February* (pp. 580-590).

Kazarinoff, P. (2018, October 8). Plotting histograms with matplotlib and Python. Retrieved from https://pythonforundergradengineers.com/histogram-plots-with-matplotlib-and-python.html

Keim, D. A., Mansmann, F., & Thomas, J. (2010). Visual analytics: how much visualization and how much analytics?. *ACM SIGKDD Explorations Newsletter*, *11*(2), 5-8.

Kraska, T. (2018). Northstar: An interactive data science system. *Proceedings of the VLDB Endowment*, *11*(12), 2150-2164.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, *185*, 71-110.

Rajasekharan, J., Kermani, E.M. & Markovic, S. (2018). Area plots, histograms, and bar plots. Retrieved from https://labs.cognitiveclass.ai/tools/jupyterlab/lab/tree/labs/DV0101EN/DV0101EN-2-2-1-Area-Plots-Histograms-and-Bar-Charts-py-v2.0.ipynb

Van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., ... & Yu, T. (2014). scikit-image: image processing in Python. *PeerJ*, *2*, e453.