Day 19: *Histograms with Matplotlib*

After their time in Asia, Dot hopped on a long overnight flight back to the other side of the globe. On a large plane filled with snoozing passengers, Dot headed east back to the Americas. Their destination: Honolulu, Hawaii. They deboarded after many hours, stretching and yawning and ready to bask in the sun. "I'm heading over to the Honolulu Zoo; what about you?" a fellow passenger asked kindly. Dot yelped and shook their head fiercely. No more zoos for them! They had a room in a hotel right on Waikiki Beach. They were excited to spend their time in Hawaii doing nothing except sunbathing. They were grateful for the excitement of the last couple of weeks, but now was the time for pure, unbothered relaxation.

As soon as Dot checked into their hotel, they headed right for the beach with sunscreen, a towel, and a favourite book. They lowered their sunglasses over their eyes and stretched out among the sand and the heat. They were surrounded by palm trees and crystal-clear water, and down the shoreline, they could see the Lē'ahi volcano. Other beachgoers were splashing around or paddling lazily on surfboards in the water. What a beautiful day! After finishing their book and snapping it closed, it was time for Dot to start thinking about the end of their around-the-world trip. First, they needed to plan their way home. Can you help Dot find out which US city has the best connection with Vancouver?

Tutorial

While working with data, we may want to figure out the frequency distribution of a numerical dataset. The frequency distribution refers to how often each value occurs within a dataset. This can be important to know so that we can understand whether the data we are analyzing is normally distributed or skewed.

The best way to visualize the distribution of our dataset is with a histogram. Histograms are a graphical representation of data using bars of different heights. Similar to bar charts, histograms group numbers into buckets. The size of each bar shows how many fall into each range.

Building a histogram with MatPlotLib

Here is how a histogram can be built using Matplotlib:

```
plt.figure()
plt.hist(df['numerical_data'], bins = 40) #Play around with the bin sizes when
plotting your histogram
plt.show()
```

Normal Distribution

A histogram with the following characteristics would have a normal distribution:

- the mean & the median are the same.
- 1 standard deviation from the mean captures 68.2% of the data.
- 2 standard deviations from the mean captures 95.4% of the data.

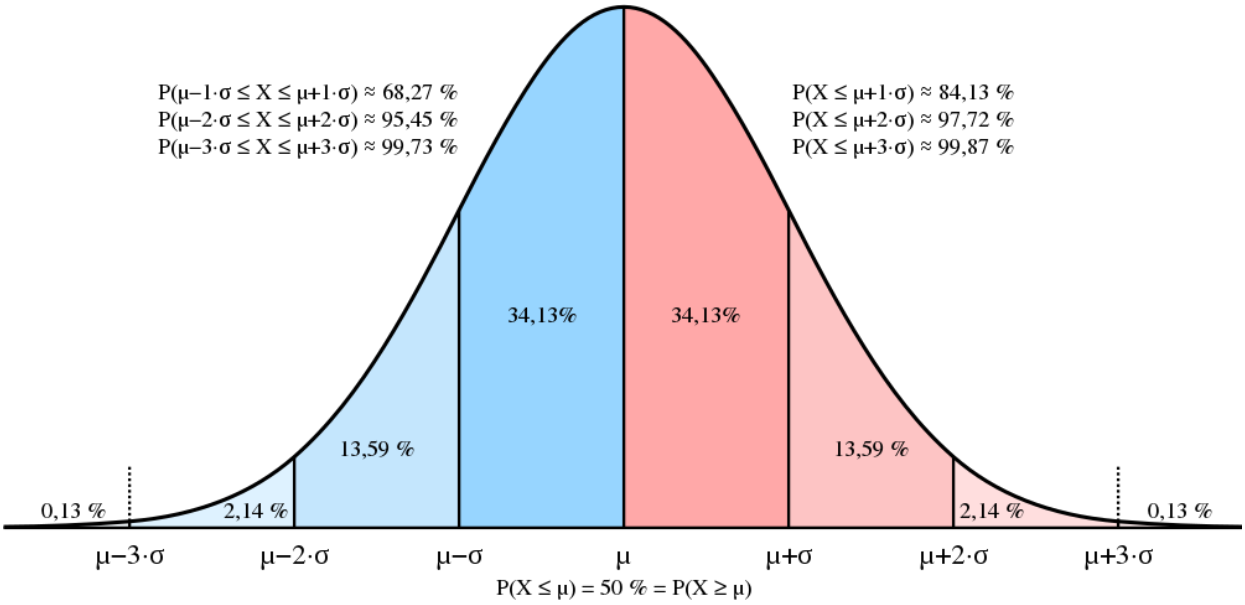


Image Taken from WikiMedia Commons

Skewed Distribution

Within this image, we can see the two types of skewed distribution:

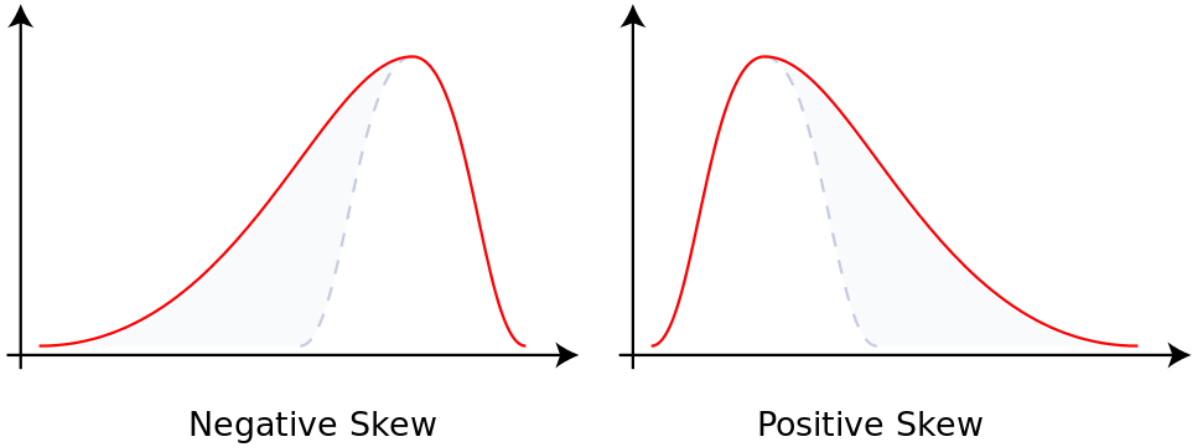


Image Taken from WikiMedia Commons.svg)

Note: Negative skew also refers to **left skew** and positive skew also refers to **right skew.**

Histogram vs. Bar Plot?

Histograms and barplots look nearly identical, and it can be easy to mistake them in some instances. Their differences lie in what they're typically used for. Bar plots tend to measure categorical data, whereas histograms measure the frequency (probability that specific value occurs) in numerical data.

To learn more on interpreting histograms, read this article.

Challenge

Use air traffic data to help Dot by finding out which US city has the best connection with Vancouver.

1. **What is the origin_city_name which was used by the most people for travelling to Vancouver?**
2. **According to our database, how many people travelled from that city?**
3. **Use a histogram to plot the probability distribution of distances for all routes in June 2021.**

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
```

In [6]:
```
df = pd.read_csv('air_traffic_data.csv')
```

In [7]:
```
# SOLUTION

# 1. and 2.
df_van = df[df.DEST_CITY_NAME.str.contains("Vancouver")]
```

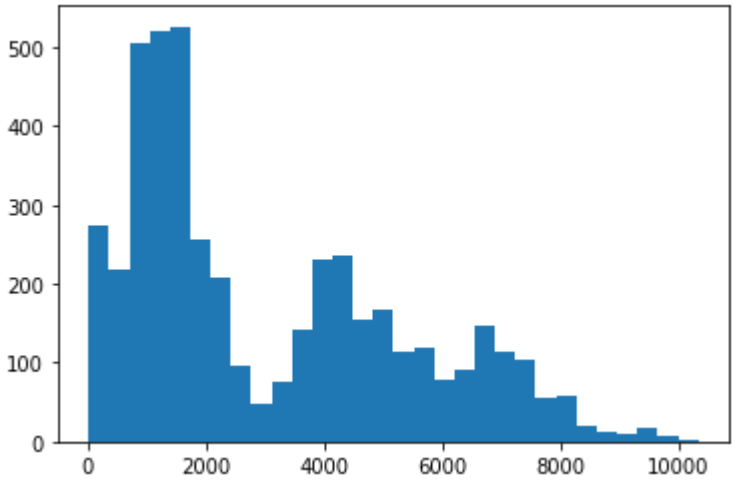
In [8]:
```
df_van.groupby('ORIGIN_CITY_NAME')[['PASSENGERS']].sum().sort_values("PASSENGERS",ascen
```

Out[8]:

ORIGIN_CITY_NAME	PASSENGERS
Seattle, WA	17109.0
San Francisco, CA	15559.0
Los Angeles, CA	11434.0
Dallas/Fort Worth, TX	8934.0
Portland, OR	1806.0

In [9]:
```
# 3.

plt.figure()
plt.hist(df[df['MONTH'] == 6]['DISTANCE'], bins = 30) #Play around with the bin sizes w
plt.show()
```


500
400
300
200
100
0
0
2000
4000
6000
8000
10000