Building a Comprehensive Sales Dataset for 2024

In [ ]:

```
import datetime
import calendar
import random
import numpy
import pandas as pd
import uuid
import os

products = {
    'iPhone': [700, 10],
    'Google Phone': [600, 8],
    'Samsung Galaxy Phone': [650, 3],
    'Alienware Monitor': [400.99, 6],
    'Dell UltraSharp Monitor': [410.99, 9],
    'Samsung Odyssey Monitor': [409.99, 9],
    'LG UltraGear Monitor': [399.99, 11],
    'Flatscreen TV': [300, 7],
    'Macbook Pro Laptop': [1700, 7],
    'Dell Laptop': [1500.99, 6],
    'AA Batteries (4-pack)': [5.84, 30],
    'AAA Batteries (4-pack)': [4.99, 30],
    'USB-C Charging Cable': [11.95, 30],
    'Lightning Charging Cable': [14.95, 30],
    'Galaxy buds Headphones': [120, 26],
    'Bose SoundSport Headphones': [99.99, 19],
    'Apple Airpods Headphones': [150, 22],
    'Amana Washing Machine': [600.00, 1],
    'Amana Dryer': [600.00, 1]
}

columns = ['Order ID', 'Product Name', 'Units Purchased', 'Unit Price', 'Order Date

def generate_random_day(month):
    day_range = calendar.monthrange(2024, month)[1]
    return random.randint(1, day_range)

def generate_random_time(month):
    day = generate_random_day(month)
    if random.random() < 0.5:
        date = datetime.datetime(2024, month, day, 12, 0)
    else:
        date = datetime.datetime(2024, month, day, 20, 0)
    time_offset = numpy.random.normal(loc=0.0, scale=180)
    final_date = date + datetime.timedelta(minutes=time_offset)
    return final_date.strftime("%m/%d/%y %H:%M")

def generate_random_address():
    street_names = ['Main', '2nd', '1st', '4th', '5th', 'Park', '6th', '7th', 'Mapl
    cities = ['San Francisco', 'Boston', 'New York City', 'Austin', 'Dallas', 'Atla
    weights = [9, 4, 5, 2, 3, 3, 2, 0.5, 6, 3]
    zips = ['94016', '02215', '10001', '73301', '75001', '30301', '97035', '04101',
    states = ['CA', 'MA', 'NY', 'TX', 'TX', 'GA', 'OR', 'ME', 'CA', 'WA']

    street = random.choice(street_names)
    index = random.choices(range(len(cities)), weights=weights)[0]

    return f"{random.randint(1, 999)} {street} St, {cities[index]}, {states[index]}

def write_row(order_number, product, order_date, address):
    product_price = products[product][0]
    quantity = numpy.random.geometric(p=1.0-(1.0/product_price), size=1)[0]
    output = [order_number, product, quantity, product_price, order_date, address]
    return output

if __name__ == '__main__':
    order_number = 141234
    os.makedirs(r"C:\2024_Monthly_Sales", exist_ok=True)  # Create folder if doesn'

    for month in range(1, 13):
        if month <= 10:
            orders_amount = int(numpy.random.normal(loc=12000, scale=4000))
        elif month == 11:
            orders_amount = int(numpy.random.normal(loc=20000, scale=3000))
        else:  # month == 12
            orders_amount = int(numpy.random.normal(loc=26000, scale=3000))

        product_list = list(products.keys())
        weights = [products[product][1] for product in products]

        df = pd.DataFrame(columns=columns)
        print(f"Generating data for {calendar.month_name[month]}...")

        i = 0
        while orders_amount > 0:
            address = generate_random_address()
            order_date = generate_random_time(month)

            product_choice = random.choices(product_list, weights=weights)[0]
            df.loc[i] = write_row(order_number, product_choice, order_date, address
            i += 1

            # Add related products for certain items
            if product_choice == 'iPhone':
                if random.random() < 0.15:
                    df.loc[i] = write_row(order_number, "Lightning Charging Cable",
                    i += 1
                if random.random() < 0.05:
                    df.loc[i] = write_row(order_number, "Apple Airpods Headphones",
                    i += 1
                if random.random() < 0.07:
                    df.loc[i] = write_row(order_number, "Galaxy buds Headphones", o
                    i += 1

            elif product_choice in ["Google Phone", "Samsung Galaxy Phone"]:  # Cor
                if random.random() < 0.18:
                    df.loc[i] = write_row(order_number, "USB-C Charging Cable", ord
                    i += 1
                if random.random() < 0.04:
                    df.loc[i] = write_row(order_number, "Bose SoundSport Headphones
                    i += 1
                if random.random() < 0.07:
                    df.loc[i] = write_row(order_number, "Galaxy buds Headphones", o
                    i += 1

            # Sometimes add random extra product
            if random.random() <= 0.02:
                random_product = random.choices(product_list, weights=weights)[0]
                df.loc[i] = write_row(order_number, random_product, order_date, add
                i += 1

            # Sometimes insert bad data
            if random.random() <= 0.002:
                df.loc[i] = columns
                i += 1
            if random.random() <= 0.003:
                df.loc[i] = ["", "", "", "", "", ""]        # Add empty row
                i += 1

            order_number += 1
            orders_amount -= 1

        month_name = calendar.month_name[month]
        file_path = rf"C:\Monthly_Sales\Sales_{month_name}_2024.csv"
        df.to_csv(file_path, index=False)
        print(f"{month_name} Complete")
```