# PHASE 3: DEVELOPMENT PART 1

ANALYSIS OBJECTIVES:

The objective is to assess and highlight variations in the daily COVID-19 cases and deaths within the European Union and European Economic Area (EU/EEA) member countries. This analysis aims to compare and contrast the mean values to identify regional trends and disparities, while also examining standard deviations to understand the extent of data variability, severity, and trends of COVID-19 in different EU/EEA countries.

PREPROCESSING OF DATASET AND CLEANING THE DATA:

Cleaning a dataset involves the process of preparing data for analysis by identifying and rectifying inconsistencies, errors, and missing values. This procedure is essential for accurate and reliable data-driven insights. Cleaning ensures that the dataset is structured and consistent, making it ready for further analysis and modelling. This crucial step guarantees that data-driven decisions are based on reliable, high-quality information.

Dataset : <https://www.kaggle.com/datasets/chakradharmattapalli/covid-19-cases>

==============\*\*PYTHON CODE(JUPYTER NOTEBOOK)\*\*===============

# COVID-19 Cases Analysis[¶](#COVID-19-Cases-Analysis)

In [1]:

# Reading the Excel file into a Pandas DataFrame

import pandas as pd

file\_path = r'C:\Users\sankar\Desktop\Covid\_19\_cases4.xlsx'

# Load the Excel file into a Pandas DataFrame

data = pd.read\_excel(file\_path)

print(data)

dateRep day month year cases deaths countriesAndTerritories

0 2021-05-31 31 5 2021 366 5 Austria

1 2021-05-30 30 5 2021 570 6 Austria

2 2021-05-29 29 5 2021 538 11 Austria

3 2021-05-28 28 5 2021 639 4 Austria

4 2021-05-27 27 5 2021 405 19 Austria

... ... ... ... ... ... ... ...

2725 2021-03-06 6 3 2021 3455 17 Sweden

2726 2021-03-05 5 3 2021 4069 12 Sweden

2727 2021-03-04 4 3 2021 4884 14 Sweden

2728 2021-03-03 3 3 2021 4876 19 Sweden

2729 2021-03-02 2 3 2021 6191 19 Sweden

[2730 rows x 7 columns]

In [2]:

# Creating copy of original data

cdata=data.copy()

In [3]:

# Structure of the dataset

cdata.info()

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 2730 entries, 0 to 2729

Data columns (total 7 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 dateRep 2730 non-null datetime64[ns]

1 day 2730 non-null int64

2 month 2730 non-null int64

3 year 2730 non-null int64

4 cases 2730 non-null int64

5 deaths 2730 non-null int64

6 countriesAndTerritories 2730 non-null object

dtypes: datetime64[ns](1), int64(5), object(1)

memory usage: 149.4+ KB

In [4]:

# Summary of numerical variables

summary\_num = cdata.describe()

print(summary\_num)

day month year cases deaths

count 2730.000 2730.000 2730.000 2730.000 2730.000

mean 16.000 4.011 2021.000 3661.011 65.292

std 8.766 0.819 0.000 6490.510 113.957

min 1.000 3.000 2021.000 -2001.000 -3.000

25% 8.000 3.000 2021.000 361.250 2.000

50% 16.000 4.000 2021.000 926.500 14.500

75% 24.000 5.000 2021.000 3916.250 72.000

max 31.000 5.000 2021.000 53843.000 956.000

In [5]:

#Summary of categorical variables

summary\_cate = cdata.describe(include = "O")

print(summary\_cate)

countriesAndTerritories

count 2730

unique 30

top Austria

freq 91

In [6]:

# Removing duplicate records

cdata.drop\_duplicates(keep='first',inplace=True)

In [7]:

# Check for missing values

cdata.isnull()

print('Data columns with null values:\n', cdata.isnull().sum())

Data columns with null values:

dateRep 0

day 0

month 0

year 0

cases 0

deaths 0

countriesAndTerritories 0

dtype: int64

In [8]:

# Calculate Mean Daily Cases

mean\_daily\_cases = cdata['cases'].mean()

print("Mean Daily Cases:", mean\_daily\_cases)

# Calculate Mean Daily Deaths

mean\_daily\_deaths = cdata['deaths'].mean()

print("Mean Daily Deaths:", mean\_daily\_deaths)

# Calculate Standard Deviation of Daily Cases

std\_daily\_cases = cdata['cases'].std()

print("Standard Deviation of Daily Cases:", std\_daily\_cases)

# Calculate Standard Deviation of Daily Deaths

std\_daily\_deaths = cdata['deaths'].std()

print("Standard Deviation of Daily Deaths:", std\_daily\_deaths)

Mean Daily Cases: 3661.010989010989

Mean Daily Deaths: 65.29194139194139

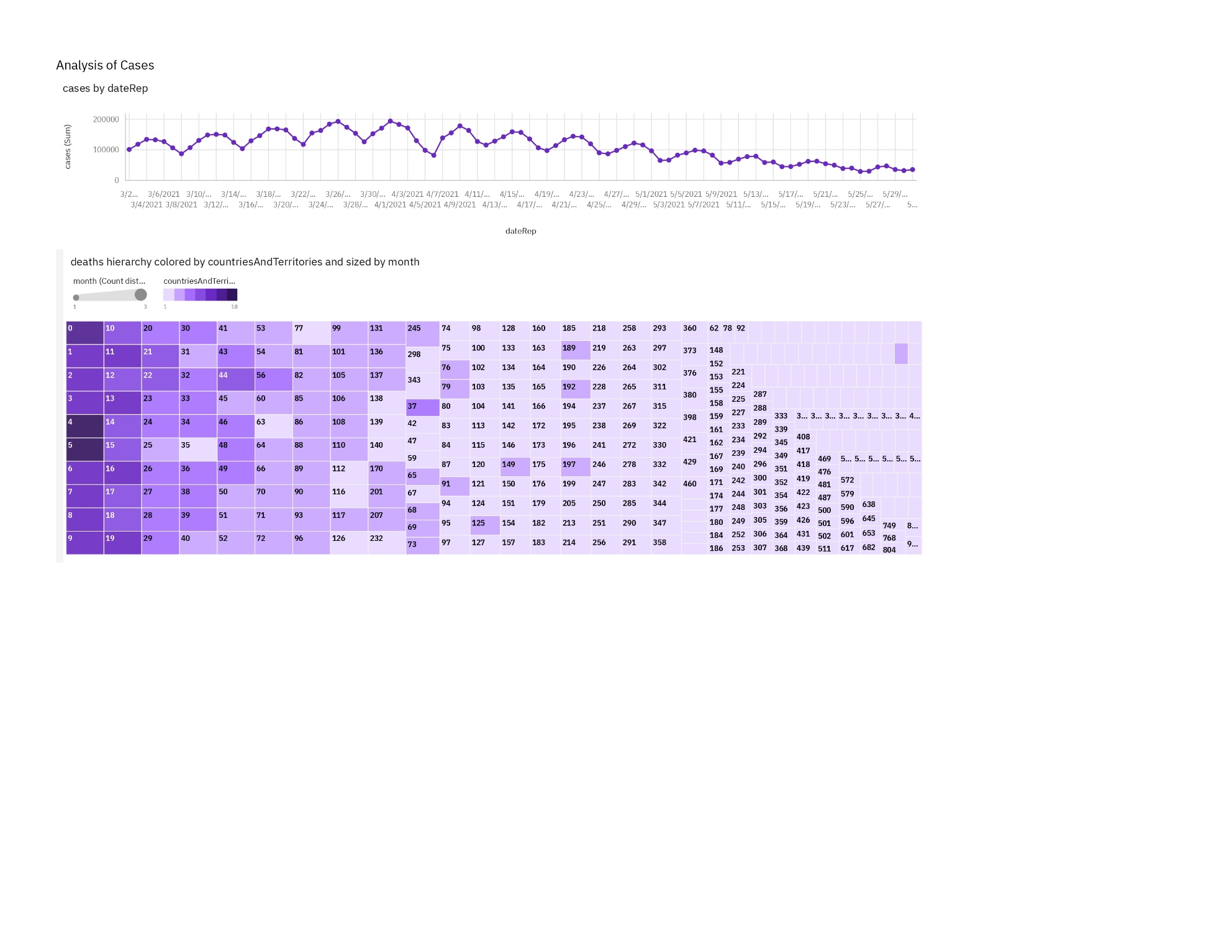
Standard Deviation of Daily Cases: 6490.510073102111

Standard Deviation of Daily Deaths: 113.95663405806982

========================\*\*END OF THE CODE\*\*===============================

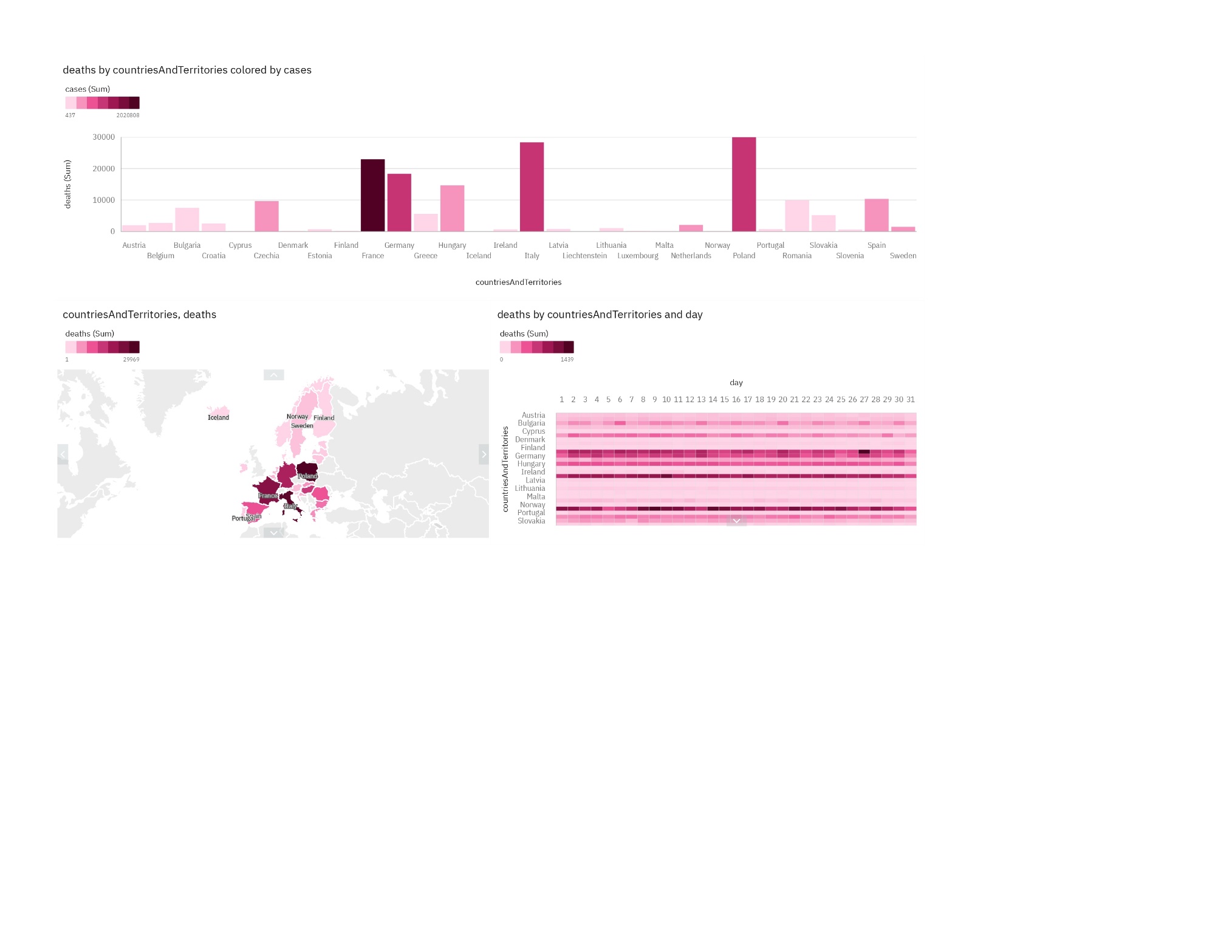
OVERVIEW OF VISUALIZATION OF COVID-19 CASES AND DEATHS IN IBM COGNOS:

ANALYSIS OF CASES:





ANALYSIS OF CASES AND DEATHS:



ANALYSIS OF DEATHS:

CONCLUSION:

In this phase, we initiated the development of our COVID-19 cases analysis project. We outlined our objectives, which involve leveraging IBM Cognos for visualization, providing a powerful platform for data exploration and presentation. The initial focus was on data preprocessing and cleaning to guarantee the data's accuracy and reliability. Also, we had an overview on visualization of the cases and deaths using IBM Cognos with various types of visualizations charts.