## Author: Siyabonga Mahlangu

## Student No: 223055539

## Module:231ISM8X04 (231ISM8X04) LEARNING FROM DATA

## Github url:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import os
for dirname, _, filenames in os.walk('Siyabonga-Mahlangu_-223055539'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
```

```
pop = pd.read_csv('Siyabonga-Mahlangu_-223055539.csv')
```

```
pop.dtypes
```

```
Unnamed: 0             int64
country name          object
current population    object
population 2022       object
area                  object
land area             object
density               object
growth rate           object
world percentage      object
rank                   int64
dtype: object
```

## Checking head and tail

```
pop.head(3)
```

```
   Unnamed: 0   country name current population population 2022  area
\
0           0          India      1,423,118,510   1,417,173,173  3.3M

1           1          China      1,425,820,141   1,425,887,337  9.7M

2           2  United States        339,231,549     338,289,857  9.4M
```

```
  land area density growth rate world percentage  rank
0        3M     481       0.81%            17.85%     1
1      9.4M     151      -0.02%            17.81%     2
2      9.1M      37       0.50%             4.25%     3
```

```
pop.tail(3)
```

```
     Unnamed: 0  country name current population population 2022 area
\
202         202         Nauru             12,780          12,668   21

203         203        Tuvalu             11,396          11,312   26

204         204  Vatican City                518             510  < 1


    land area density growth rate world percentage  rank
202        20     639       0.88%             0.00%   225
203        30     380       0.74%             0.00%   227
204       < 1   1,177       1.57%               NaN   234
```

## Displaying mean for Rank

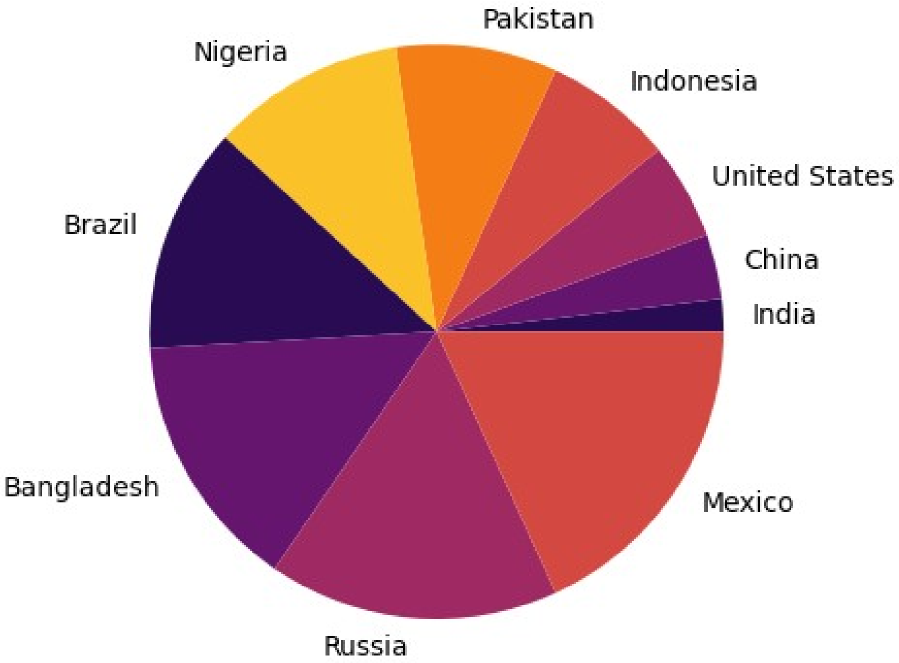
```
pop['rank'].mean()
```

```
104.35121951219512
```

```
piepop = pop[0:10]
```

## Pie Chart

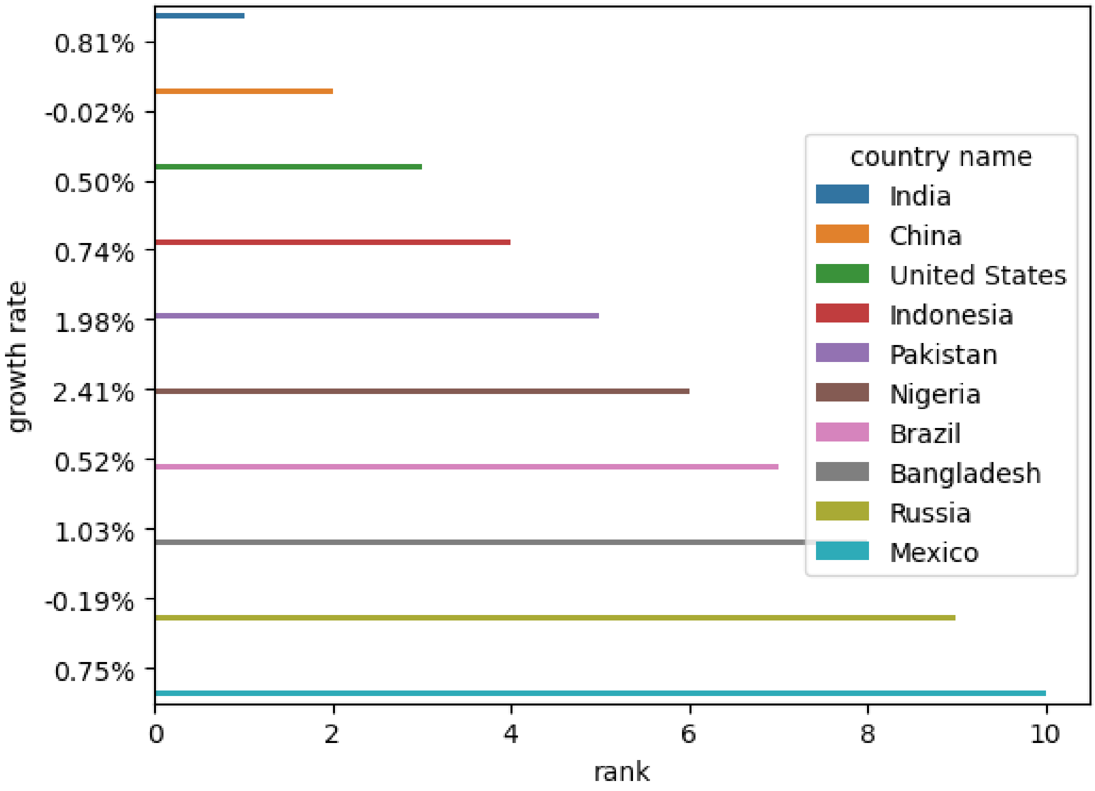
```
plt.pie(piepop['rank'], labels = piepop['country name'], colors =
sns.color_palette('inferno'))
plt.show()
```



## Bar plot showing country population

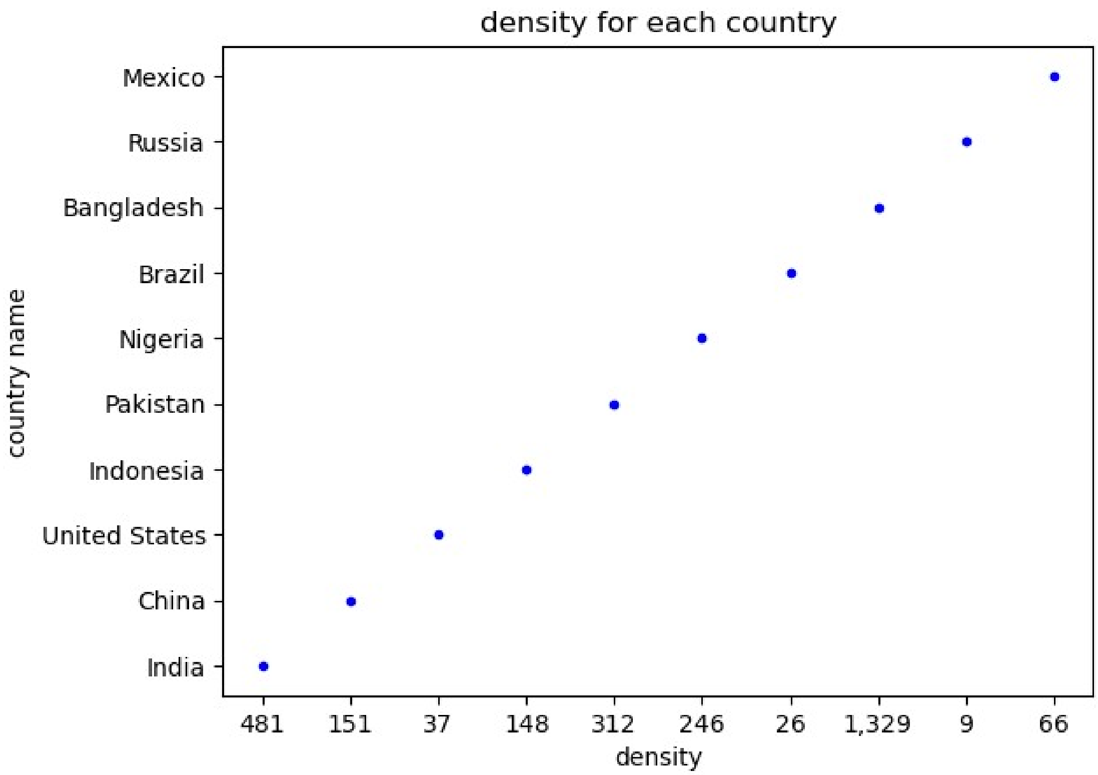
```
sns.barplot(x="rank", y="growth rate", hue="country name", data =
piepop)
```

```
<Axes: xlabel='rank', ylabel='growth rate'>
```



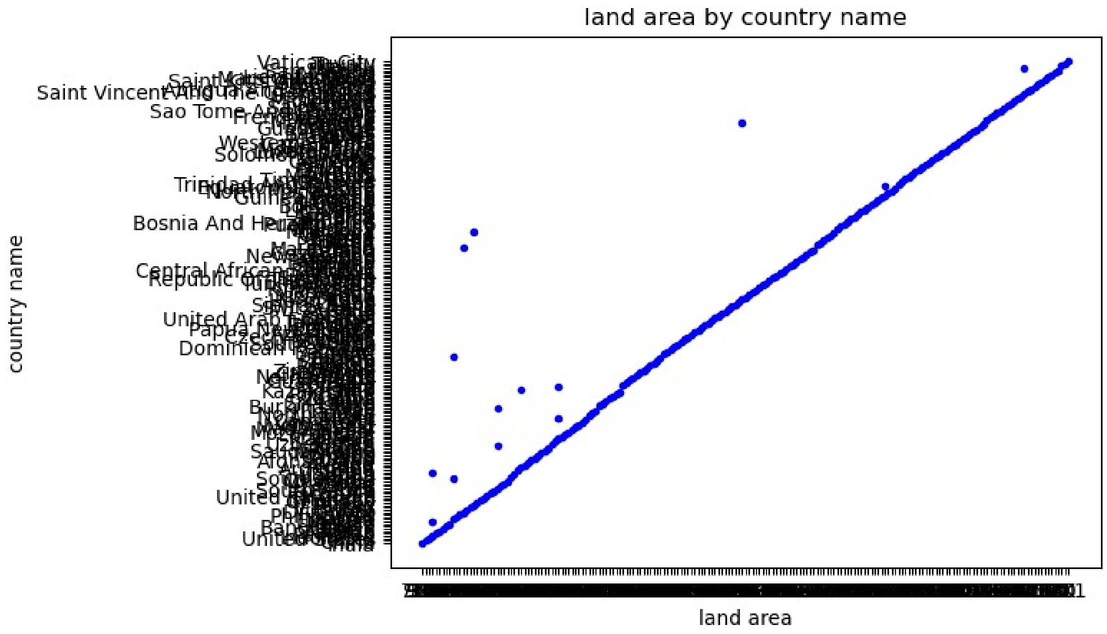
### Using plot to display density per country

```
plt.plot(piepop['density'],piepop['country name'],'b.')
plt.xlabel('density')
plt.ylabel('country name')
plt.title('density for each country')
plt.show()
```

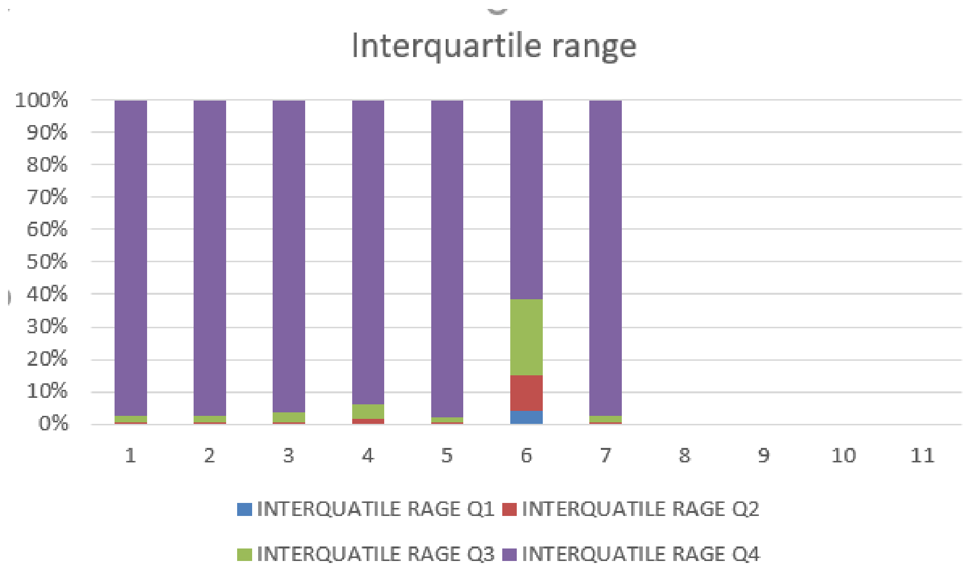


## Population per land area

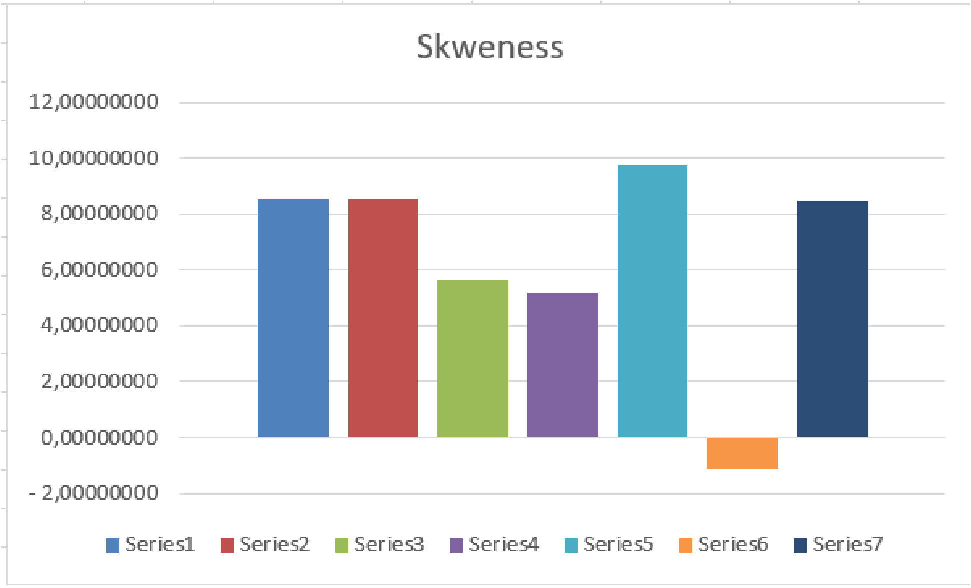
```
plt.plot(pop['land area'],pop['country name'],'b.')
plt.xlabel('land area')
plt.ylabel('country name')
plt.title('land area by country name')
plt.show()
```



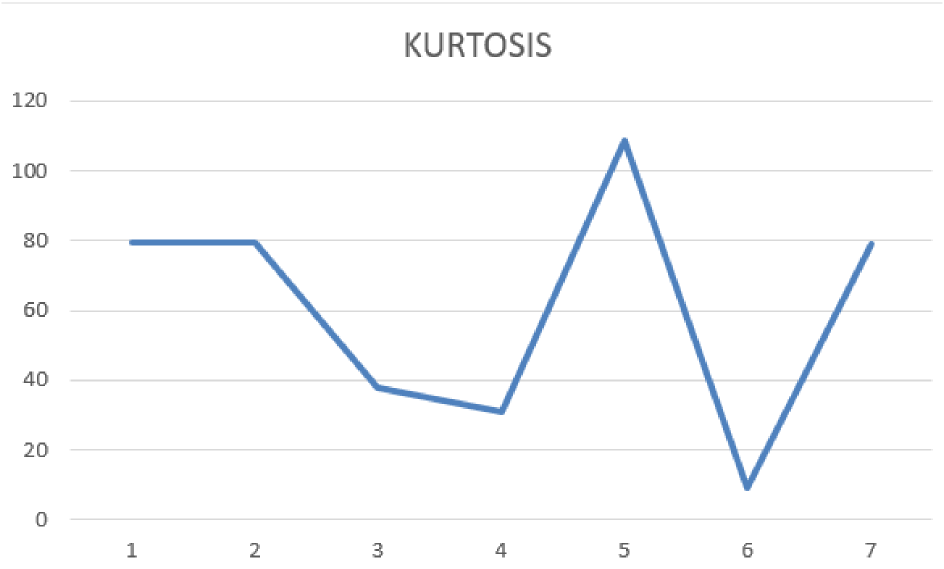
## Interquartile Range for my dataset



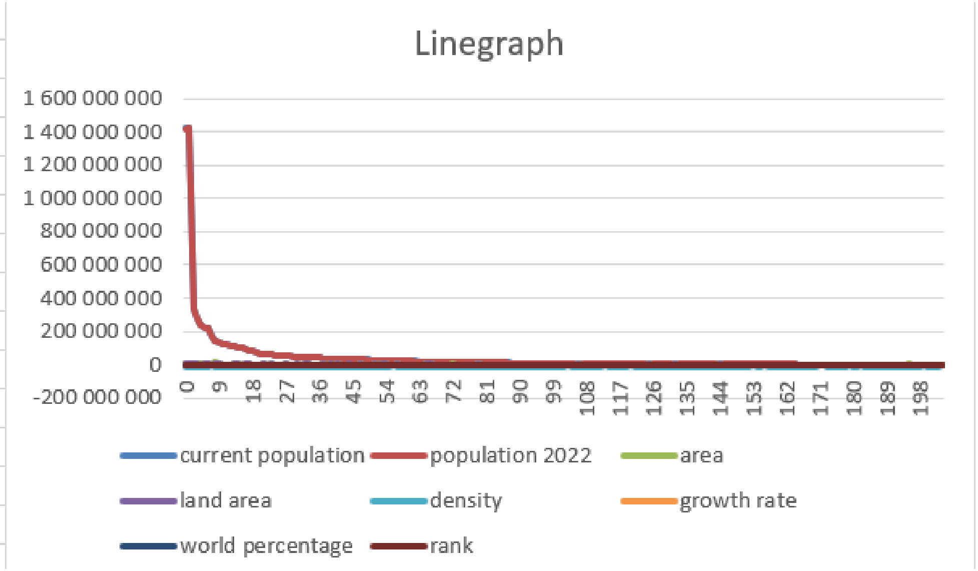
## Skweness Diagram



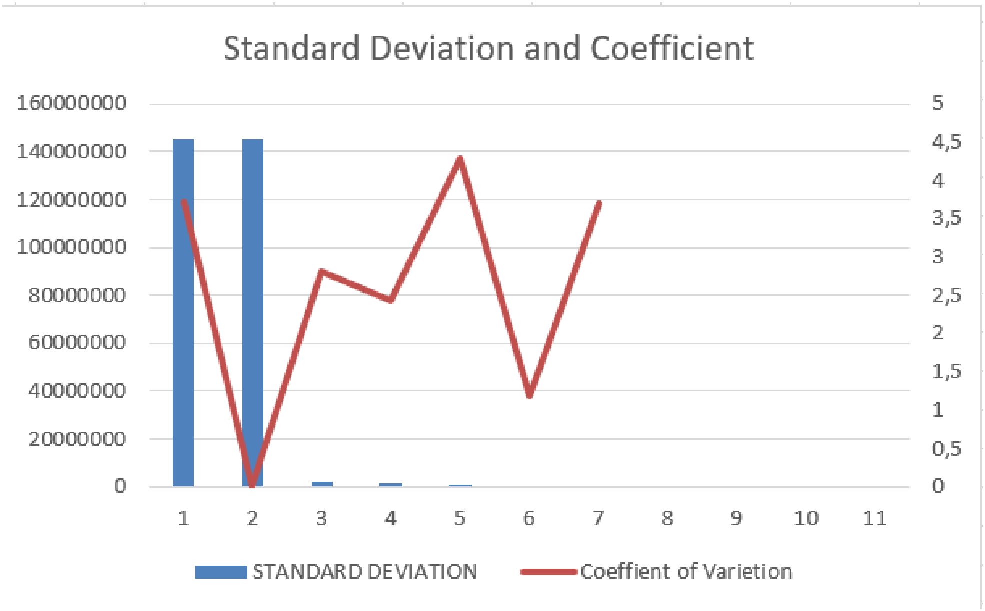
## Kurtosis Diagram



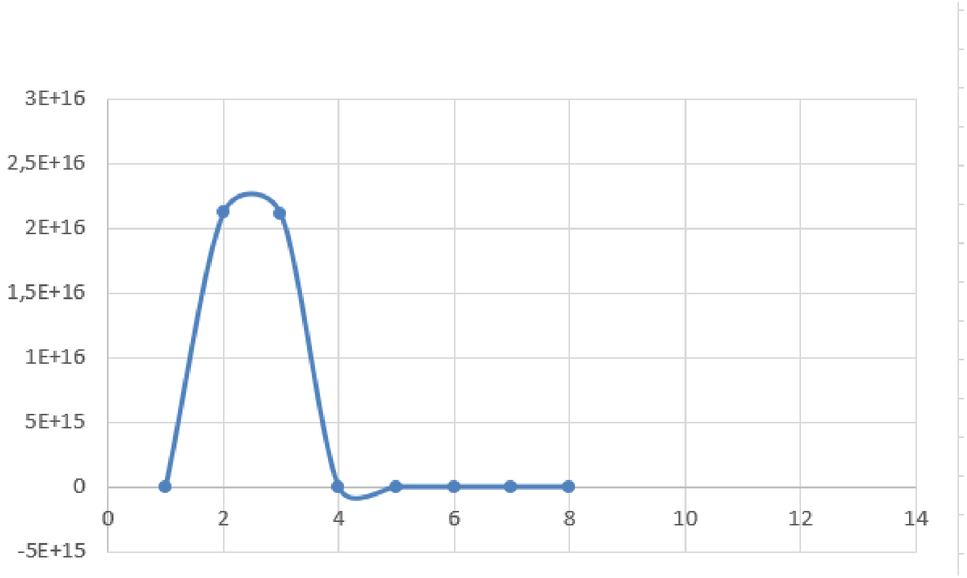
## Line Graph for my Dataset



## Standard Deviation & Coefficient of varietion



## Variance Diagram



## Country with least population per rank

```
pop[pop['rank'] == pop['rank'].max()]
```

```
     Unnamed: 0  country name current population population 2022 area  
\
204          204  Vatican City                 518             510  < 1


    land area density growth rate world percentage  rank
204       < 1   1,177       1.57%                NaN   234
```

## Country with most population per rank

```
pop[pop['rank'] == pop['rank'].min()]
```

```
   Unnamed: 0 country name current population population 2022  area  
land area  \
0           0        India      1,423,118,510   1,417,173,173  3.3M  
3M

  density growth rate world percentage  rank
0     481       0.81%           17.85%     1
```

## What I learnt

I believe in the next coming years, Python will overtake Excel becuase on Python you can integrate data extraction, do anaylitics in one environment, so Python is good for companies who work with big data like banks.

Python allow you to work in big dataset and with excel you still can but the formulas and filtering work mostly for small data. Python is amazing becuase tasks are automated so it become easy to replicate a task and on Excel it's difficult.

Excel makes it difficult to test the correctness of data and changing one number can effect hundreds of calculations, during my assignment 1, I struggled a lot with Excel especially when using their calculations but with Python it was flowing and If i did something wrong, system throws an error.

I doubt I will continue to use Excel, Python is it for me, for someone like me who's in the Software Environment, it will be easy to use Python and it saves time when working with large dataset.

If I was given a chance to do this assignment over and over again, I will definitely master the use of python,the calculations behind it.

I have discovered that you can use python also for the following reasons:

```
- Data Cleaning
- Data visualisation
- Statistical Modelling
```

Another why I would choose Python over Excel becuase it has multimedia resouces, as you can see on top that I have attached Images and can also upload video if i had too, which is difficult to do with Excel, and last reason it because on Explonatory text.