VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Theory Digital Assignment 1

B.Tech in Computer Science and Engineering **(CSE)**, **Winter** Semester **2020-21**

Name:	Swaranjana Nayak
Registration Number:	19BCE0977
Slot:	B2
Date:	04/04/2021

I. Attributes and their types

Attribute	Data Type
dt	Temporal/Chronological (datetime64[ns])
LandAverageTemperature	Quantitative - Continuous (float64)
LandAverageTemperatureUncertainty	Quantitative - Continuous (float64)
LandMaxTemperature	Quantitative - Continuous (float64)
LandMaxTemperatureUncertainty	Quantitative - Continuous (float64)
LandMinTemperature	Quantitative - Continuous (float64)
LandMinTemperatureUncertainty	Quantitative - Continuous (float64)
LandAndOceanAverageTemperature	Quantitative - Continuous (float64)
LandAndOceanAverageTemperatureUncertainty	Quantitative - Continuous (float64)

```
gt.dtypes

LandAverageTemperature                     float64
LandAverageTemperatureUncertainty          float64
LandMaxTemperature                         float64
LandMaxTemperatureUncertainty              float64
LandMinTemperature                         float64
LandMinTemperatureUncertainty              float64
LandAndOceanAverageTemperature             float64
LandAndOceanAverageTemperatureUncertainty  float64
dtype: object
```

```
[9] gt.index

DatetimeIndex(['1850-01-01', '1850-02-01', '1850-03-01', '1850-04-01',
               '1850-05-01', '1850-06-01', '1850-07-01', '1850-08-01',
               '1850-09-01', '1850-10-01',
               ...
               '2015-03-01', '2015-04-01', '2015-05-01', '2015-06-01',
               '2015-07-01', '2015-08-01', '2015-09-01', '2015-10-01',
               '2015-11-01', '2015-12-01'],
              dtype='datetime64[ns]', name='dt', length=1992, freq=None)
```

II. Plotting and explanation

1. Importing Libraries

Code:

```
# importing libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import copy
%matplotlib inline
```

2. Reading dataset CSV file and NaN values handling

Code:

```
gt = pd.read_csv('GlobalTemperatures.csv', header=0, index_col=0,
parse_dates=True, squeeze=True)
```

```
gt.dropna(inplace = True)
gt.head()
```

Output:

Out[2]:

dt	LandAverageTemperature	LandAverageTemperatureUncertainty	LandMaxTemperature	LandMaxTemperatureUncertainty	LandMinTemperature	LandMinTemp
1850-01-01	0.749	1.105	8.242	1.738	-3.206	
1850-02-01	3.071	1.275	9.970	3.007	-2.291	
1850-03-01	4.954	0.955	10.347	2.401	-1.905	
1850-04-01	7.217	0.665	12.934	1.004	1.018	
1850-05-01	10.004	0.617	15.655	2.406	3.811	

3. **Declaring column arrays**

Code:

```
col = [gt.columns[0], gt.columns[2], gt.columns[4], gt.columns[6]]
col
```

Output:

```
['LandAverageTemperature',
 'LandMaxTemperature',
 'LandMinTemperature',
 'LandAndOceanAverageTemperature']
```

Code:

```
col2 = gt.columns
col2
```

Output:

```
Index(['LandAverageTemperature', 'LandAverageTemperatureUncertainty',
       'LandMaxTemperature', 'LandMaxTemperatureUncertainty',
       'LandMinTemperature', 'LandMinTemperatureUncertainty',
       'LandAndOceanAverageTemperature',
       'LandAndOceanAverageTemperatureUncertainty'],
      dtype='object')
```

4. **Month-wise box plot**

Code:

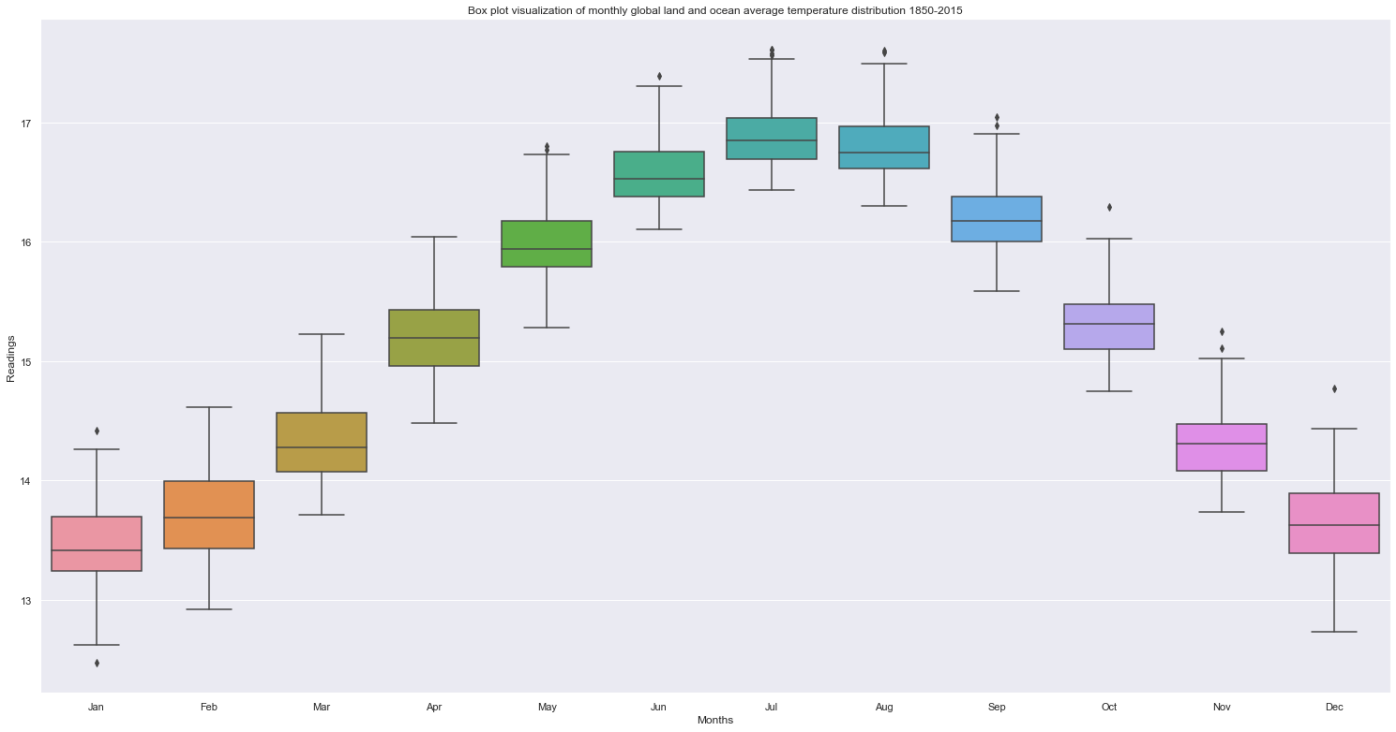
```
groups = gt[col[3]].groupby(pd.Grouper(freq='A'))
LandAndOceanAverageTemperature = pd.DataFrame()
for name, group in groups:
    LandAndOceanAverageTemperature[name.year] = group.values

LandAndOceanAverageTemperature =
LandAndOceanAverageTemperature.transpose()
months = ['Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun', 'Jul', 'Aug',
'Sep', 'Oct', 'Nov', 'Dec']
LandAndOceanAverageTemperature.columns = months

fig = plt.figure(figsize = (20, 10))
ax = fig.add_axes([0, 0, 1, 1])
sns.boxplot(x="variable", y="value",
data=pd.melt(LandAndOceanAverageTemperature), ax = ax)
```

```
ax.set_title('Box plot visualization of monthly global land and
ocean average temperature distribution 1850-2015')
ax.set_xlabel('Months')
ax.set_ylabel('Readings')
plt.savefig('monthlyboxplot.png', bbox_inches = 'tight')
```

Output:



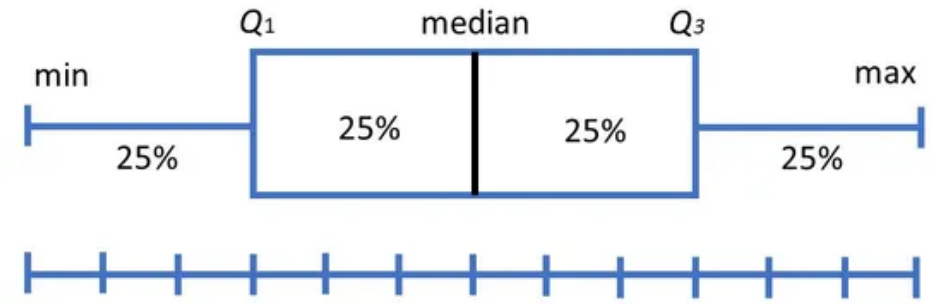
Explanation

- The dataframe LandAndOceanAverageTemperature has 166 rows for years 1850-2015 and 12 columns for each month in the year.

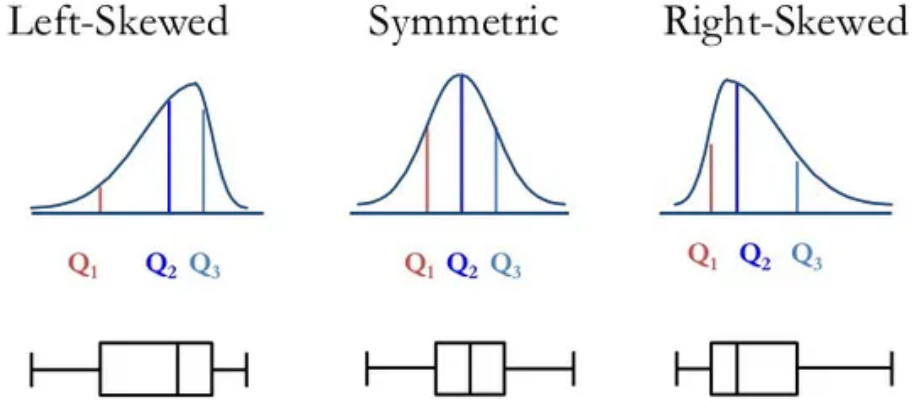
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1850	12.833	13.588	14.043	14.667	15.507	16.353	16.783	16.718	15.886	14.831	13.897	13.300
1851	13.245	13.331	13.897	14.640	15.771	16.496	16.831	16.621	16.058	15.213	14.161	13.638
1852	13.231	13.311	13.736	14.786	15.899	16.619	16.984	16.566	16.038	15.178	13.948	13.782
1853	13.143	13.362	14.033	14.919	15.793	16.455	16.999	16.789	15.942	14.874	13.829	13.324
1854	12.983	13.248	14.089	14.945	15.793	16.286	16.775	16.707	16.098	15.378	14.123	13.467
...	...	...	...	...	...	...	...	...	...	...	...	...
2011	13.928	14.193	14.880	15.832	16.523	17.203	17.568	17.475	16.762	15.873	14.799	14.198
2012	13.859	14.164	14.863	15.881	16.699	17.252	17.450	17.420	16.882	16.019	15.001	14.138
2013	14.117	14.359	14.952	15.749	16.609	17.257	17.503	17.462	16.894	15.905	15.107	14.339
2014	14.136	14.157	15.090	16.038	16.804	17.303	17.508	17.607	16.975	16.029	14.899	14.410
2015	14.255	14.564	15.193	15.962	16.774	17.390	17.611	17.589	17.049	16.290	15.252	14.774

166 rows × 12 columns

- The X-axis has months
- The Y-axis has the distribution of land and ocean average temperatures visualized as a box plot. (which is a quantitative attribute)
- Given a set of data values, a Box Plot shows the five number summary of that set - minimum score, first (lower) quartile, median, third (upper) quartile, maximum score.
- Box plots also show outliers, if any.
- Outliers:
    - **Mark**: Point
    - **Channel:** Unaligned spatial position - how far outside the dataset range the point lies can be seen by where it is with respect to the end of the whiskers.
    - Y position corresponds to the value of the outlier
    - X position corresponds to the Month.
    - **Task performed by outliers** - An outlier is an observation that is numerically distant from the rest of the data. When reviewing a box plot, an outlier is defined as a data point that is located outside the whiskers of the box plot.
- Five number summary:
    - **Mark**: Line
    - **Channel**: Unaligned spatial position for measurement; area for data distribution between first quartile, median and and third quartile; length for data distribution between minimum score and first quartile, and third quartile and maximum score; and color hue (identity channel) to distinguish between the months.



    - The five measurements are marked by lines in order are - minimum score, first quartile, median, third quartile and maximum score.
    - Lengths of marks of minimum and maximum scores are equal.
    - Lengths of marks of first quartile, median and third quartile are equal.
    - The X position of the measurements corresponds to the month.
    - The Y positions of the measurements correspond to their values.
    - The lengths of lines connecting the minimum and maximum scores to first and third quartiles respectively (the whiskers) represent the amount of data values concentrated in that region.
    - The areas between first quartile and median, and median and third quartile represent the amount of data values concentrated in that region.
    - The position of the median mark can be checked to know how skewed the data is.



- **Task performed by the measurements**: We can compare values of two months by comparing their moments of distribution and skew of the data.

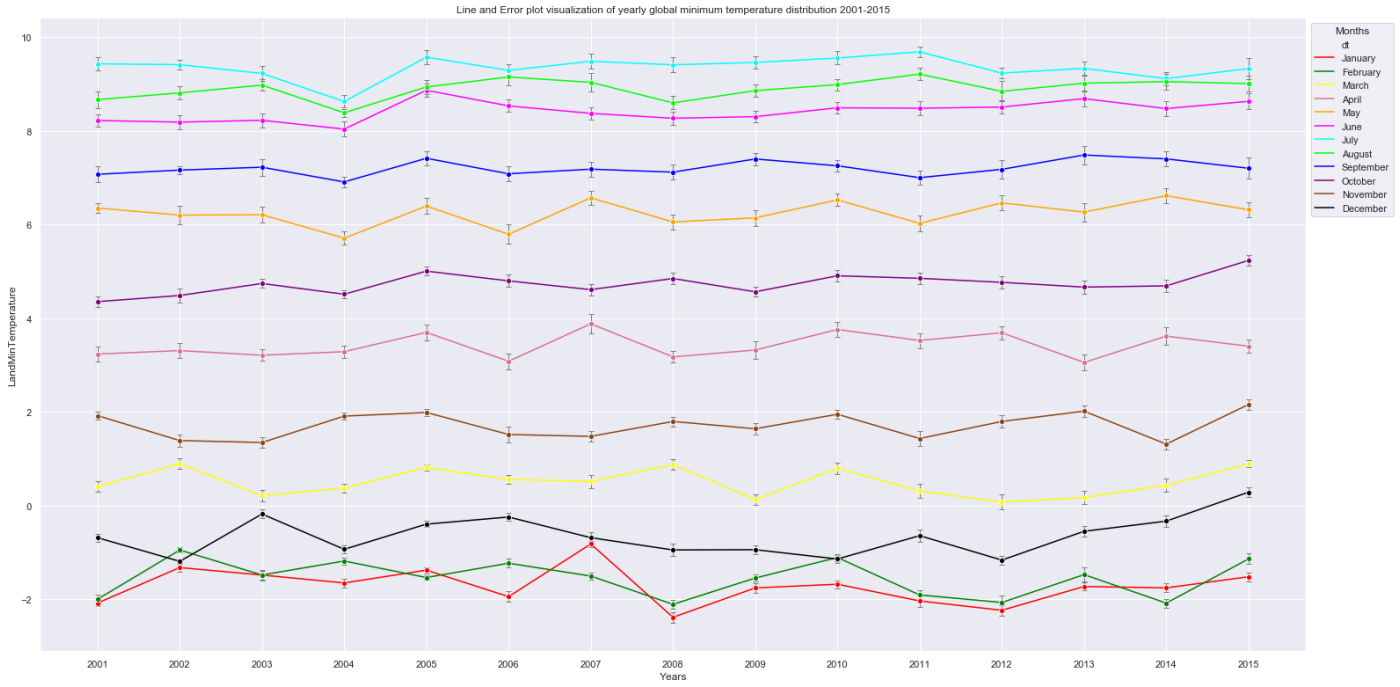
- Result and Interpretation:
  - Maximum average temperature experienced all year long is in July.
  - Minimum average temperature experienced all year long is in January.
  - **Trend**: Temperature increases from January to July and decreases from July to December.

5. Line and error plot

Code:

```
fig = plt.figure(figsize = (20, 10))
ax = fig.add_axes([0, 0, 1, 1])
x = gt.index.year[1812:]
y = gt[1812:][col[2]]
dy = gt[1812:][col2[5]]
colors = ['red', 'green', 'yellow', 'palevioletred', 'orange',
'magenta', 'cyan', 'lime', 'blue', 'purple', 'saddlebrown',
'black']
plt.errorbar(x, y, yerr=dy, fmt='none', ecolor='gray',
elinewidth=1, capsize=3)
sns.lineplot(x, y, hue=gt.index.month_name()[1812:], marker="o",
palette=colors)#, s = 70, ax = ax)
ax.set_title('Box plot visualization of yearly global minimum
temperature distribution 2001-2015')
ax.set_xlabel('Years')
ax.set_xticks(x.unique())
plt.legend(title='Months', bbox_to_anchor=(1, 1), loc='upper left')
plt.savefig('scatteranderrorplot.png', bbox_inches = 'tight')
```

Output:



Explanation:

- dt (index of the dataframe) is the key attribute of the dataframe.
- LandMinTemperature, a quantitative attribute, is taken for this plot. It has month-wise minimum temperature readings from 1850 to 2015. For the purpose of this plot, only readings from 2001-2015 are considered.
- We also have LandMinTemperatureUncertainity, a quantitative attribute that gives us the uncertainty in the LandMinTemperature readings.
- The uncertainty is the experimenter's best estimate of how far an experimental quantity might be from the "true value."
- So each reading can be taken as Best Estimate ± Uncertainty, or in our case LandMinTemperature ± LandMinTemperatureUncertainity
- Line Plot:
  - **One key, One value**
  - **Mark**: points for the values; line connection marks between points of the same month.
  - **Channel**: aligned lengths to express quant value, separated and ordered by key attribute into horizontal regions, color hue (identity channel) for distinguishing between the months.
  - There are 12 lines for 12 different months. The vertical position of the points corresponds to the value of LandMinTemperature readings, and horizontal position corresponds to the year.
  - There are 12 dots vertically aligned for each year for the 12 months in that year.
  - The months are color-encoded with different colors.
  - **Task**: find trend – connection marks emphasize ordering of items along key axis by explicitly showing relationship between one item and the next
- Error Plot:
  - **One key, Two values**
  - Key is dt
  - The values are :
    LandMinTemperature + LandMinTemperatureUncertainity and
    LandMinTemperature - LandMinTemperatureUncertainity
  - **Mark**: Lines for the values, and perpendicular line connection marks between LandMinTemperature readings and the error range values.
  - **Channel**: Length of the connection line - represents the error range.
  - **Task**: The error plot shows the error range that we derive from the uncertainty values.
- Results and Interpretations:
  - June, July and August are the hottest months of the year.
  - December, January and February are the coldest months of the year.
  - Error range is more significant when fluctuation from previous year is higher.
  - **Trend**: Temperatures are increasing with years.