Global Warming Prediction Report

Pirashanth satkunabalasingam 002244891

CS503B01 201901 - Data Visualization

CS504B01 201901 - Programming Languages for Data Analysis

1. Problem definition

The project aim is to implement a solution for the analyse the data from the dataset and visualize the prediction of the temperature or humidity. When performing the data analysis i came across many scenarios that restrict me to analyse the specific data as user wants. The factors that make restricts such as missing data, enormous number of data of many cities from 16 countries which contain each and every temperature data and humidity data from 1840 - 2013 almost 40 million data. In order to get rid of that issue, Data analyse technique has been used. But in the data prediction part the solution cannon be able to visualize the 100% accurate solution.

2. Summery

The global warming project has been developed using the python. The goal of the project is to analyse data of temperature and humidity from enormous number of data and show the changes using multiple visualization techniques to users also it includes calculations such as mean, standard deviation and variance then data prediction of temperature and humidity is the main purpose. This report describes and evaluates the overview of the project and testing solutions. A lack of commercially available solutions to the problem was identified and this project can get values from the user for the specific input and show multiple number of visualization techniques as user wants it makes more unique. Evaluation of the solutions is analyse the data and visualize, named 'Data analyse and Data visualization', as the optimal solution to the problem also it include high technique such as machine learning for data prediction. This data analyse get multiple input from the user for the country, city,year range and weather type(Temperature or humidity) . Right after the user input the solution will analyse the CSV data file using 'pd.read_csv()' method then solution will remove unwanted data such as missing data using fillna(),dropna() functions. The second part of the solution start analyse and retrive the only information as user wants (Country, city, time period and weather type).

The Third part contains multiple visualization techniques part such as line plot, scatter plot, histogram, pie chart, lifecycle plot/horizontal bar chart, multiple plots. The command prompt will ask the user for the input for the visualization part. User got an option to generate calculations for mean,variable,variance of each year separately which has the own data analysis techniques which contain 2 dimensional array for the action. The another action is 'year wise line plot visualization'. It will show the user line graph of each year changes of the temperature or humidity for the each 12 months within the year range. Next one is 'Scatter visualization' it will visualize the scatter plot of the data. Then 'Histogram visualization, 'Pie chart' , 'Lifecycle visualization' then complete plot visualization which contains previous 3 visualization together to make it easy to look by user.

Before the last one 'mean visualization' will show the changes of the each year by the calculated mean. So the previous calculated method will pass the value of the array_mean and display it in the visualize it.

As last part data prediction which can able to predict the temperature or Humidity using 'LinearRegression' and 'model.predict' using sklearn machine learning library in python. 'Linear regression' is a basic and commonly used type of predictive analysis and the 'model.predict' is classification and regression predictions with a finalized deep learning model

A commercial evaluation suggests that ‘data analyse and visualize' will not only be a possible design solution to the problem but will also be a commercially viable solution to the problem.

3. Context

Dealing with the enormous big data would be hard to handle in the real world scenario but this project makes it more easy to access the data as user wants. As explained in the summery user can able to get the weather data of specific location and time period as user wants.

4. RAID analysis

a. Risk
    i. Estimation and scheduling- The estimated time has been increased right after the sudden growth in the requirement of the project which affect the most risk in the solution
    ii. Sudden growth in requirements- after the testing process with the user the project solution got some new requirements in the project such as multiple data type, future data prediction, calculations of the mean, standard deviation and variance
    iii. Breakdown of specification-During the initial phases of integration and coding, requirements conflict.
b. Assumptions
    i. Technology-The technical context of the project such as platforms and environments. The solution has been decided to use Python and pycharm.
    ii. Facilities- The available data and open source libraries made it successful of the project solution
c. Issues
    i. As a primary issue the occur during the development is data prediction part which consumed more time. But later on it has been solved
d. Dependencies
    i. The ‘Mean visualization’ should be done after the calculation section. Because the mean visualisation depends on the mean calculation.

5. Requirements

a. Functional Requirement

i. User input for the data analyze part
ii. Analyse the data based on user requirement
iii. Retrive the data and store it for the solution use
iv. Show the data of the temperature or humidity to user
v. Calculations of the mean standard deviation and variance for each year
vi. year wise line plot visualization
vii. Mean visualization
viii. Pie chart visualization
ix. Scatter visualization

x. Histogram visualization
xi. Life cycle visualization
xii. Data prediction

b. Non functional requirement

i. Performance of the computer
ii. Scalability
iii. Capacity
iv. Availability
v. Reliability
vi. Recoverability
vii. Maintainability
viii. Manageability
ix. Data Integrity
x. Usability

6. Algorithm

There are multiple algorithm has been used to develop this solution such as data analysis, data visualization, algorithm for calculations and finally data prediction.

a. Data Analysis

Data analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names. This part has been achieved by using Python language and Pandas library. THe main algorithm i used in data analysis is Predictive analytics. Predictive analytics encompasses a variety of statistical techniques from data mining, predictive modelling, and machine learning, that analyze current and historical facts to make predictions about future or otherwise unknown events.
This library has been used to read data from the CSV file using ‘pd.read_csv’. As explained in the summery, right after the reading data it will store the data and clean the unwanted or missing data from the stored data frame right then analysement will take part. So as user required the data will be gathered from the specific country and the city between specific year also user has the option to select weather type such as humidity or temperature.

b. Data Visualization

Data visualization is the graphical representation of information and data. By utilizing different type of visual elements such as charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. The solution has been developed using matplotlib library in python language. It contains multiple solutions of the Data visualization techniques those i mentioned in the summery. The basic way of the creation of the visualization is creating the plot to show the graph using ‘plt.subplots()‘. The next step is mentioning the type of the graph the solution needs such as line,bar,scatter,pie,hist for histogram(plt.hist(temp, num_bins, facecolor='blue', alpha=0.5) the given code line for creating histogram using some inputs from my

code. Also it have an option to mention name of the x , y axis and title even more. By using the sklearn can able to drew and show the future prediction in visualization.

c. Data prediction

The Predictive analytics part has been developed using the sklearn library. From the given data the skylearn library will reshape the array into 2d array then solution will assign the LinearRegression to the model. linear regression is a linear way to deal with modelling the connection between a scalar reaction. The model will predict the x, y data from using the prediction function.

d. calculations

It contains the Mean calculation, Standard deviation, Variance calculation of the each and every values of the temperature/Humidity based on the user limit also solution consist of the mean,variance,Standard deviation of the every year.

- **Mean(m)=average of the values**
- **Variance (S2) = average squared deviation of values from mean**
- **Standard deviation (S) = square root of the variance**

7. Design and architecture

a. Development models

Selecting suitable development models make system extends simply to manage. The development models recognize the requirement for the system and it stays away from risk and makes system extends simply to manage. While considering an appropriate design model for system, the for the development of the system following methodologies were considered

i. Spiral model

Spiral model is a model just like increment model with the more important of the risk analysis. Spiral model specially has 4 stages Planning, Risk Analysis, Engineering and Evaluation. Data visualization of the solution has been developed the same model for the development. All the requirement has been collected and planed for the visualization part then analyzed the all of the risk then developed and tested the system. As the final stage it was evaluated the output with the one student about that section.

Advantages of Spiral model:

- Can analyze and identify the lot of risks and it can be fixed in early stages
- Useful for large and mission-critical system
- It has the proper documentation control and approval
- Extra Functionality can be included at a later time.
- The system could be delivered ahead of schedule in the product life cycle.

Disadvantages of Spiral model:

- It is a costly model to utilize
- To find the risks should be in the higher level of expertise in that area.
- Success of the system mainly depend on the risk analysis
- It needs a bigger scale system

ii. Waterfall model

Waterfall model (also called linear-sequential life cycle model) was the first model which was introduced. It is a simple model to use than other. In the waterfall model, system has each stages to complete. It should complete one stage before start another stage. It will go like stage by stage. Waterfall model has been selected for the development of data analysis. It has several phases like reading, extracting,cleaning, etc. those all stages has been completed using waterfall model.

Advantages of waterfall model:

- By using this model system was easy to develop
- System was easy to manage by the rigidity of the system. Every stages has review and deliverable
- It will work correctly in the small project which has proper requirement
- At a time one stage will be developed

Disadvantages of waterfall model

- Can't have the prototype until finish the project completely
- It has high risk
- It is not a proper model for complex project
- When it is about to complete if the developer has any changes it's hard to make changes after the development

iii. Incremental model

Increment model is the entire requirement of the system divided into multiple smaller system. It is easier to manage the project as multiple component. Each smaller component have the requirement gathering, designing, implementation and testing for the system. For the complete solution system has been divided into three component. First one is data analysis and calculation second one is data visualization and as the final component is data prediction.

Advantages of the Increment model

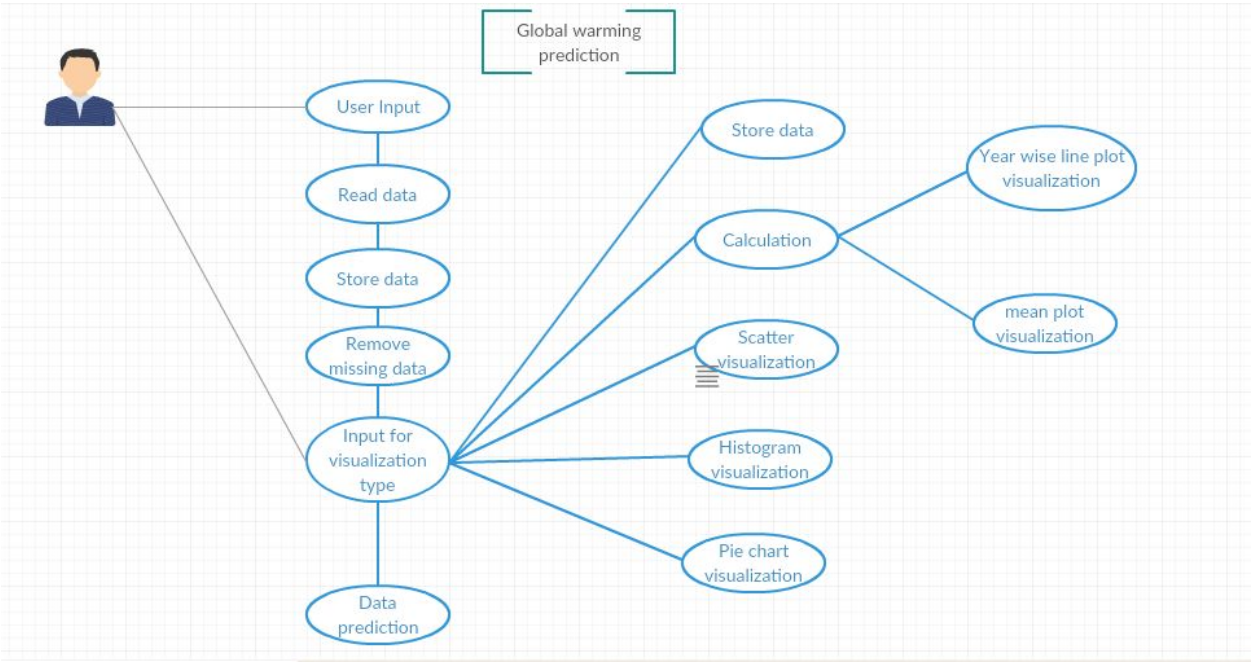
- Produces working project rapidly and early while the software life cycle.
- This model is more adaptable – can change requirement and scope in less cost
- Very easy to test and debug after the smaller updates
- In this model. The client can react to each assembled.
- Brings down starting conveyance cost.

Disadvantages of Increment model

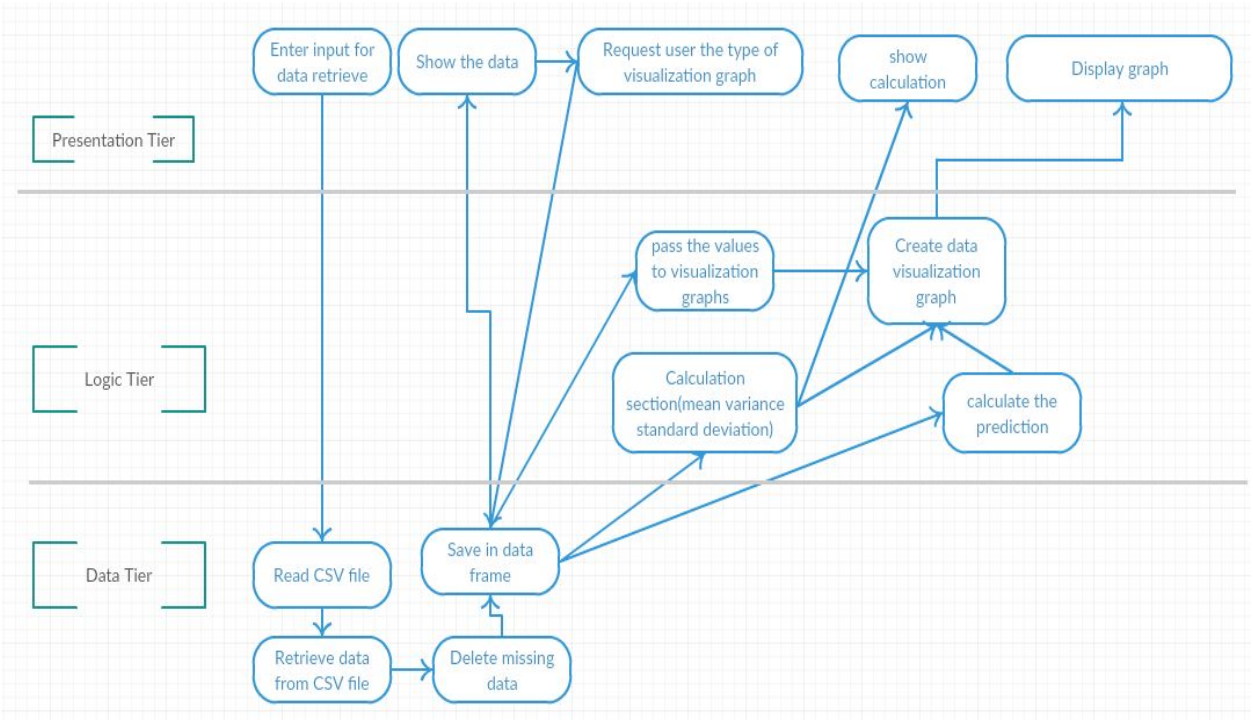
- Should have good plan and design for the system
- Cost of the system is expensive than using waterfall model

- If the system doesn't have proper definition and requirement system will be broken down

b. Use case diagram



c. Architecture Diagram



8. Testing

a. Testing Environment

The a python implementation tested using a HP laptop (HP Hexa-Core 2.0 GHz) with 8 GB ram and Intel system (Dual-core 3.3 GHz) with 8 GB ram.

b. Software Testing

Testing is the technique of evaluating a project or its component(s) to find whether it satisfies the foreordained requirements or not. Testing is executing a project to recognize any crevices, mistakes, or missing essentials in as opposed to the genuine prerequisite. (TUTORIALSPOINT.COM, 2015)

c. Functional Requirements Testing

Testing method for the functional requirement
Incremental methodology, which is the software development methodology mainly used for the system. It works with freedom and flexibility to carry out the entire testing for the development. Table given below will show the summary of the functional requirement results

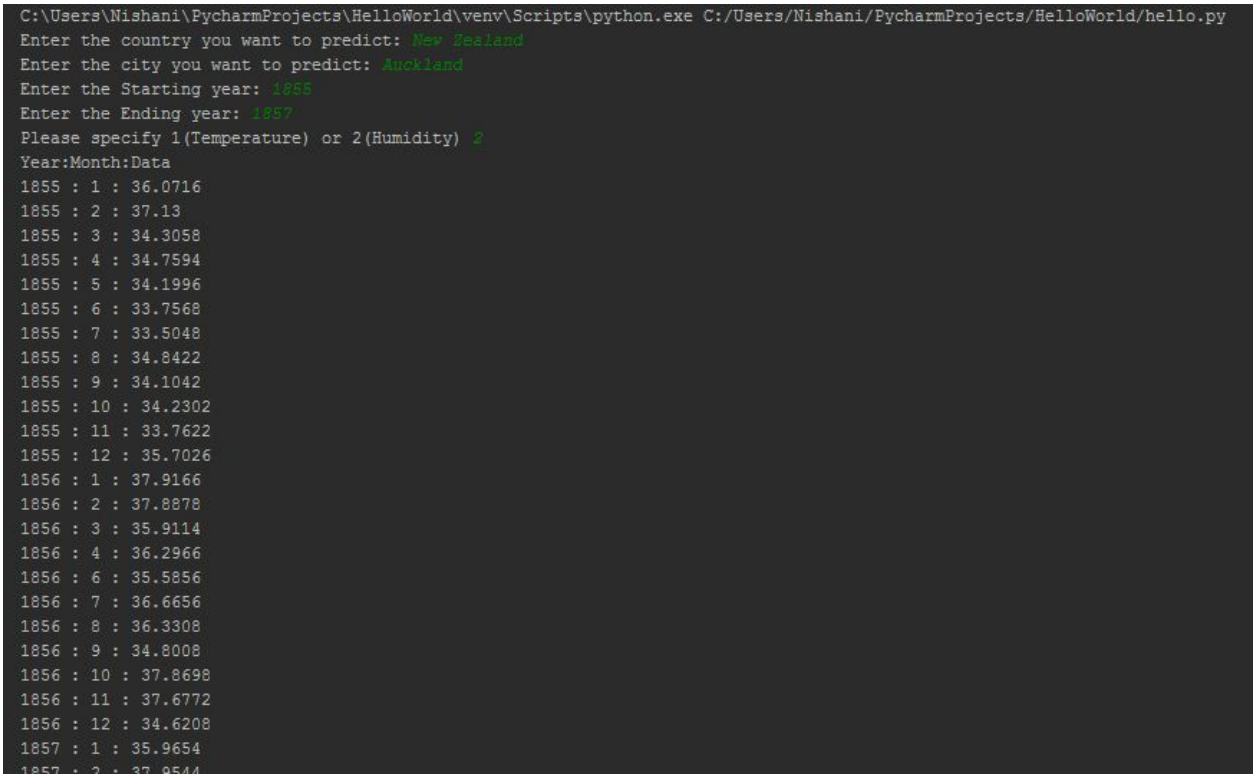
d. Functional requirement Test cases and results.

Test case	Scenarios	Input data	Expected result	Actual result	Status	Success rate
1	User input	Country, city,start year end year, weather	Getting input from user	Getting input	pass	100%
2	Retrieve information from user	User input	Read and Store from the CSV file	Storing data successfully	pass	100%
3	Display the data	Read data	Show the data to the user	Showing correct data to the user	pass	100%
4	Getting input from the user	Visualization type	Show the expected output	Successfully showing the output	pass	100%
5	Calculate	Extracted	Calculate	Successful	pass	100%

	the mean standard deviation and variance	data	and show it to the user of each year	ly showing the calculation of each year to the user		
6	Year plot visualization	Extracted data	Show the output yearwise to the user	Successfully showing to the user	Pass	100%
7	Scatter visualization	Extracted data	Display the scatter graph to the user	Successfully display the scatter graph to the user	Pass	100%
8	Pie chart visualization	Extracted data	Display the pie chart to the user	Successfully display the Pie chart graph to the user	Pass	100%
9	Histogram	Extracted data	Display the Histogram chart to the user	Successfully display the Histogram chart to the user	Pass	100%
10	Life cycle graph	Extracted data	Display the Life cycle graph to the user	Successfully display the Life cycle graph to the user	Pass	100%
11	Complete plot visualization	Extracted data	Display the 3 plots together to the user	Successfully display the complete plot visualization to the user	Pass	100%
12	Mean plot	calculation	Display	Successful	Pass	100%

	visualization		the Mean plot visualization graph to the user	ly display the Mean plot visualizationgraph to the user		
13	Data prediction	Extracted data	Display the Extracted visualization graph to the user	Successfully display the predicted visualization graph to the user	Pass	100%

9. Sample screen shots of the program



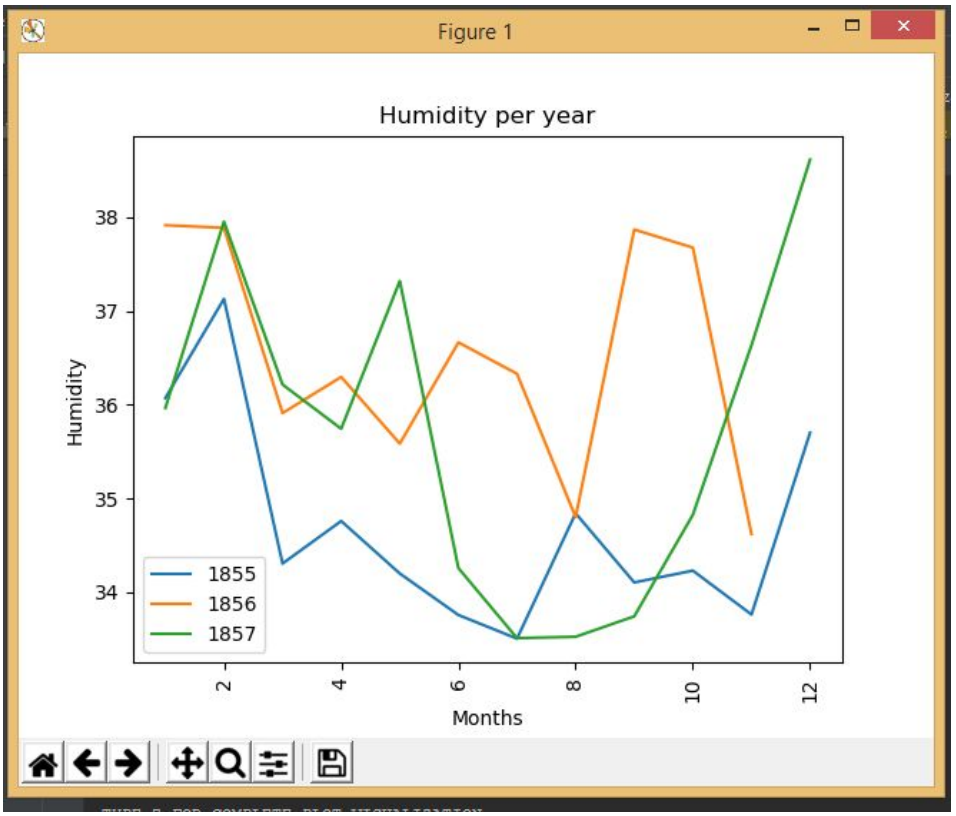
- This image shows the user input and output. The green values are user input. So as user wants solution displayed the set of the auckland, new zealand data between 1855-1857 also user has selected Humidity.

```
Data Processing.....
TYPE 1 FOR MEAN STANDARD DEVIATION AND VARIANCE CALCULATION
TYPE 2 FOR YEAR-WISE LINE PLOT VISUALIZATION
TYPE 3 FOR SCATTER VISUALIZATION
TYPE 4 FOR HISTOGRAM VISUALIZATION
TYPE 5 FOR PIE CHART VISUALIZATION
TYPE 6 FOR LIFE CYCLE VISUALIZATION
TYPE 7 FOR COMPLETE PLOT VISUALIZATION
TYPE 8 FOR MEAN VISUALIZATION
TYPE 9 FOR DATA PREDICTION
TYPE 10 FOR EXIT
Which visualization type do you want?
```

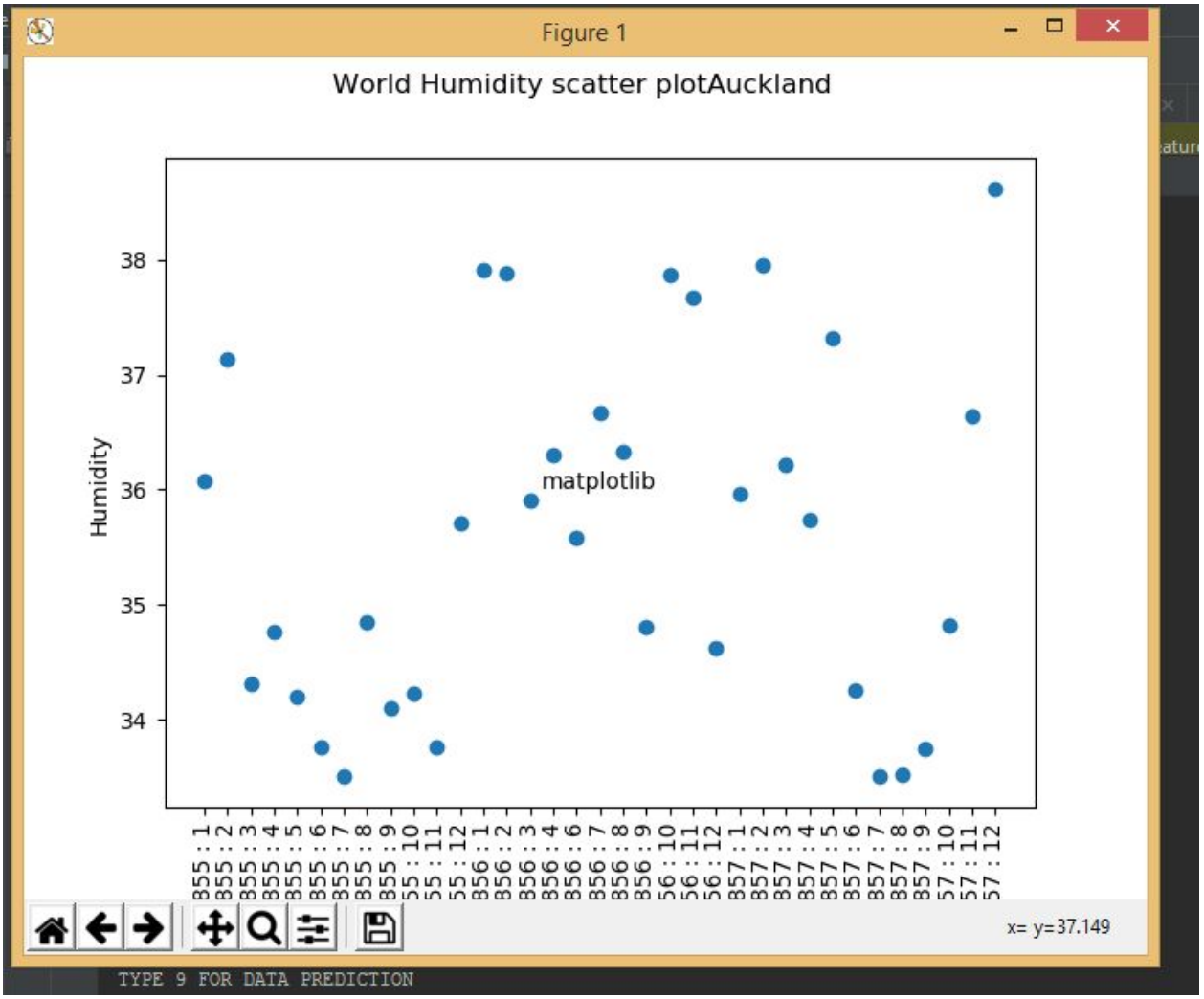
- Second user input for the selection of the feature.

```
Which visualization type do you want?1
--------------------------------------------------
Year of the calculation is   1855
Standard Deviation of temperature is  1.086869821175387
Mean of temperature is  34.697449999999996
Variance of temperature is  1.1812860081818177
--------------------------------------------------
--------------------------------------------------
Year of the calculation is   1856
Standard Deviation of temperature is  1.2203773851484694
Mean of temperature is  36.50572727272727
Variance of temperature is  1.4893209621818158
--------------------------------------------------
--------------------------------------------------
Year of the calculation is   1857
Standard Deviation of temperature is  1.7486142846161457
Mean of temperature is  35.6927
Variance of temperature is  3.057651916363635
--------------------------------------------------
TYPE 1 FOR MEAN STANDARD DEVIATION AND VARIANCE CALCULATION
```

- User has been selected the first feature which will display the year,mean,standard deviation and variance of each year.



The image above is the Line visualization of humidity(user selected humidity) and 12 months of each year. So as you see difference has been shown by different colours and year label below. User can able to see temperature of the each year like this.



This image above mentioning about the scatter visualization of each month and how it is increasing. Label has been rotated into 90 degree manually to avoid unwanted label overlapping.

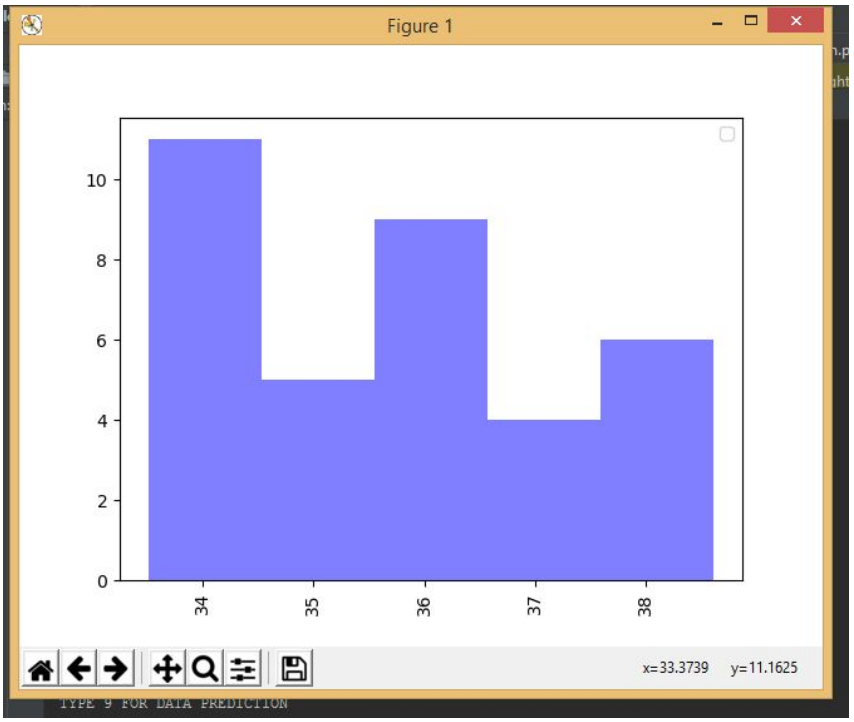
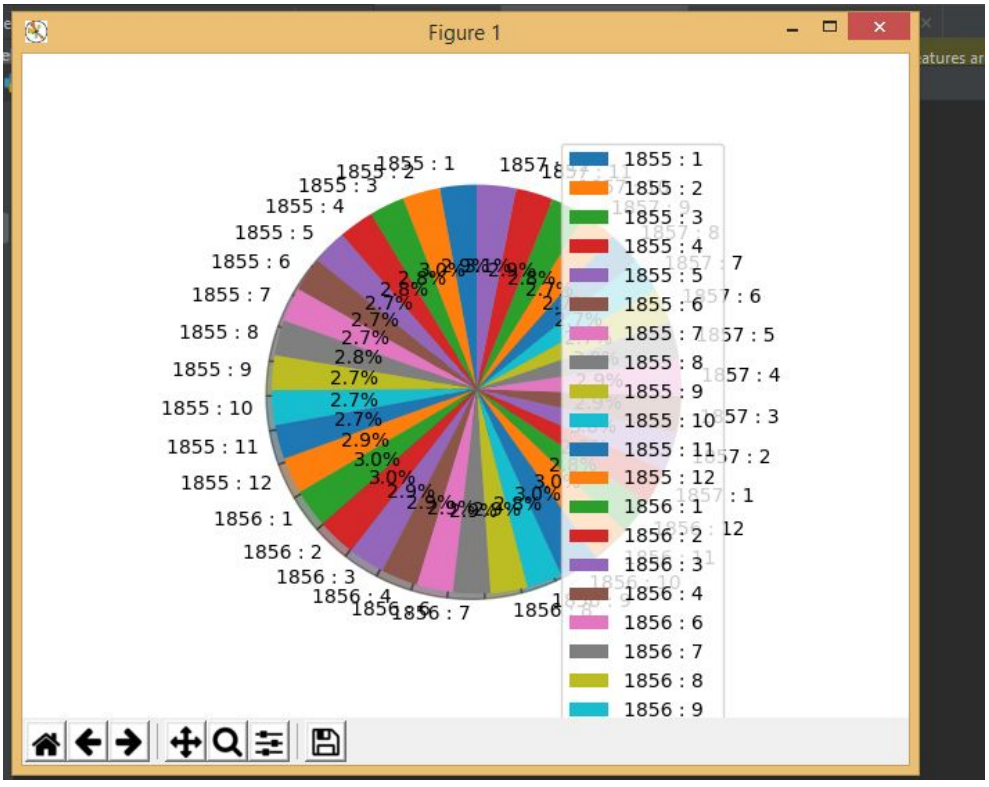
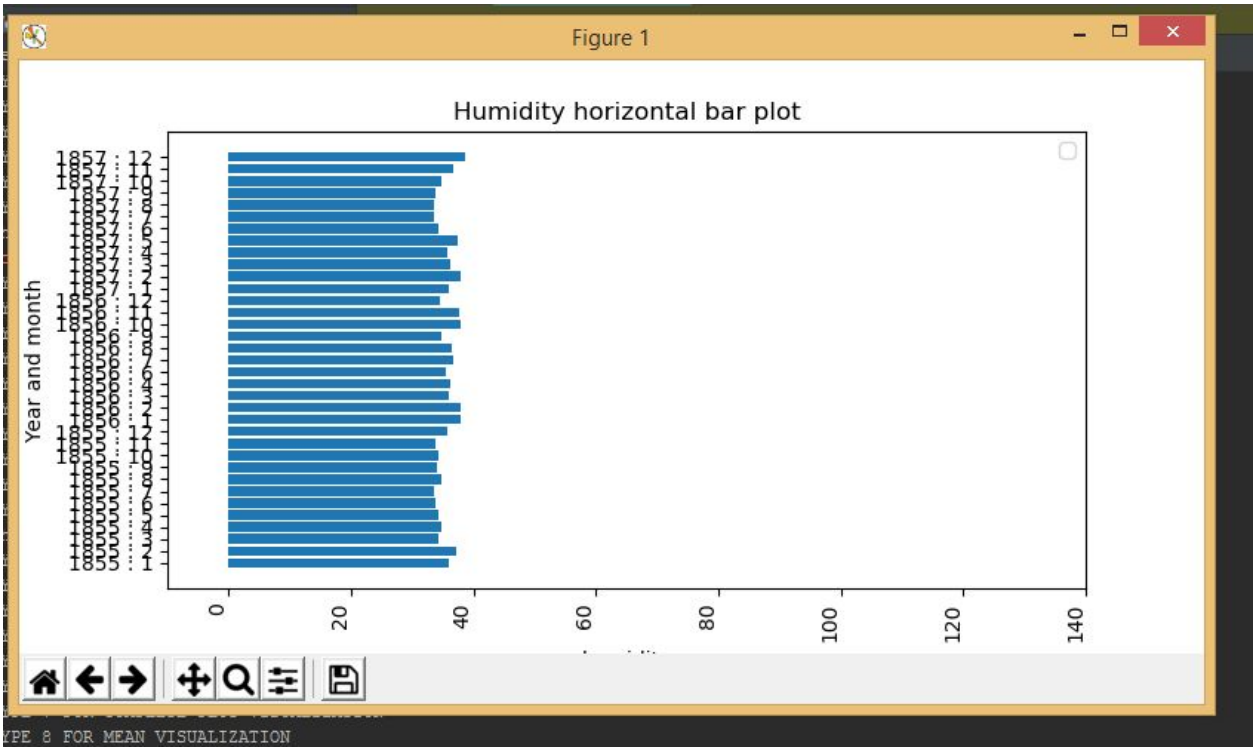


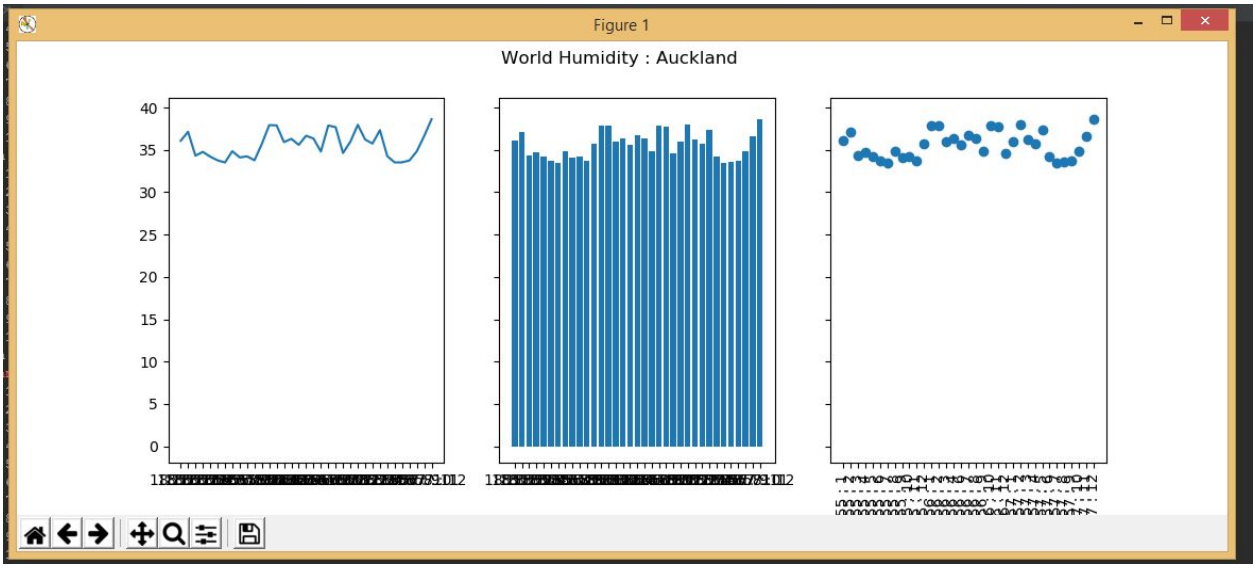
Image above is Histogram visualization image of the Humidity how it change over year.



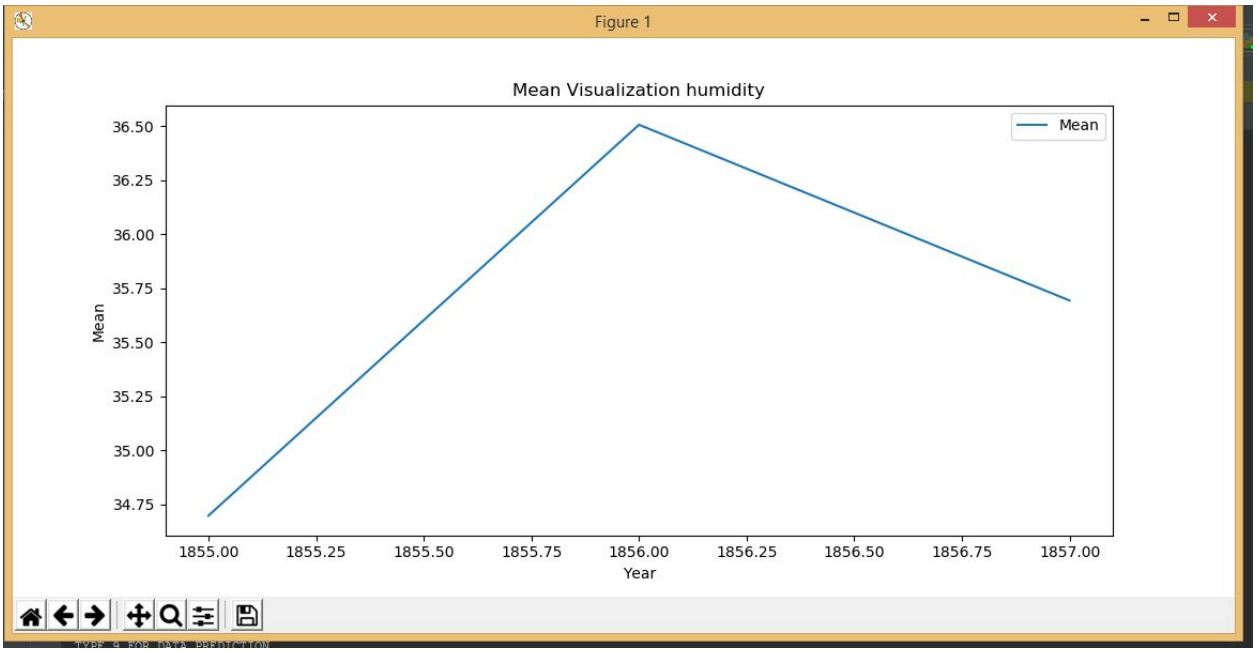
This image above is Pie chart of the every month and it shows how many percentage it occurred out of 100%.



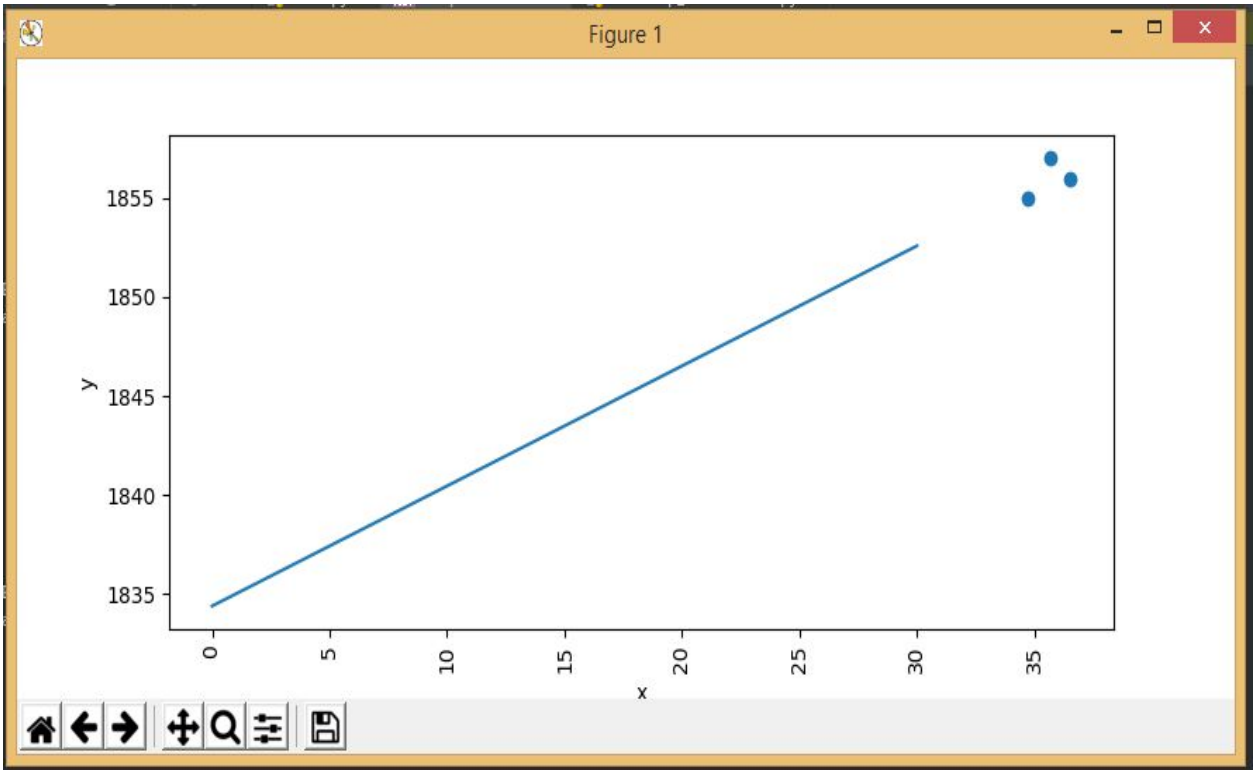
This image is a Humidity horizontal bar/ lifecycle visualization. It shows each year and month as y axis and percentage data as x axis.



This image above shows the line,bar,scatter visualization together. But i could not able to avoid all the label overlapping manually.



This image is line visualization of the mean. it is about How the humidity mean change each year. 1855 to 1856 the mean of year was increasing next year it fell down by next year.



This image is a data prediction image.how the line can increase over the year.

10. References:

a. Websites

i. Introduction to Data Visualization in in Python https://towardsdatascience.com/introduction-to-data-visualization-in-python-89a54c97fbed access-feb-15-2019

ii. Data analysis library -https://pandas.pydata.org/ access: 25-January-2019

iii. Data Analysis with Python and Pandas Tutorial Introduction https://www.youtube.com/watch?v=Iqjy9UqKKuo access - 28-january -2019

iv. Visualization with Matplotlib https://jakevdp.github.io/PythonDataScienceHandbook/04.00-introduction-to-matplotlib.html access feb -20 2019

v. Python Visualization with Matplotlib https://stackabuse.com/python-data-visualization-with-matplotlib/ access-25 feb 2019

vi. Python Matplotlib library -https://matplotlib.org/february-5-2019

vii. Python Visualization with Matplotlib with udemy https://www.udemy.com/data-visualization-with-python-and-matplotlib/ feb-28-2019

viii. Hands-On Introduction To Scikit-learn (sklearn) https://towardsdatascience.com/hands-on-introduction-to-scikit-learn-sklearn-f3df652ff8f2 access march 15 2019

ix. Python scikit library documentation https://scikit-learn.org/stable/ march-20-2019

b. Journal/Books

i. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython-Book by Wes McKinney - Originally published: December 30, 2011

ii. Python Data Science Handbook - Publisher: O'Reilly Media Release Date: November 2016