Data Analytics

CS40003

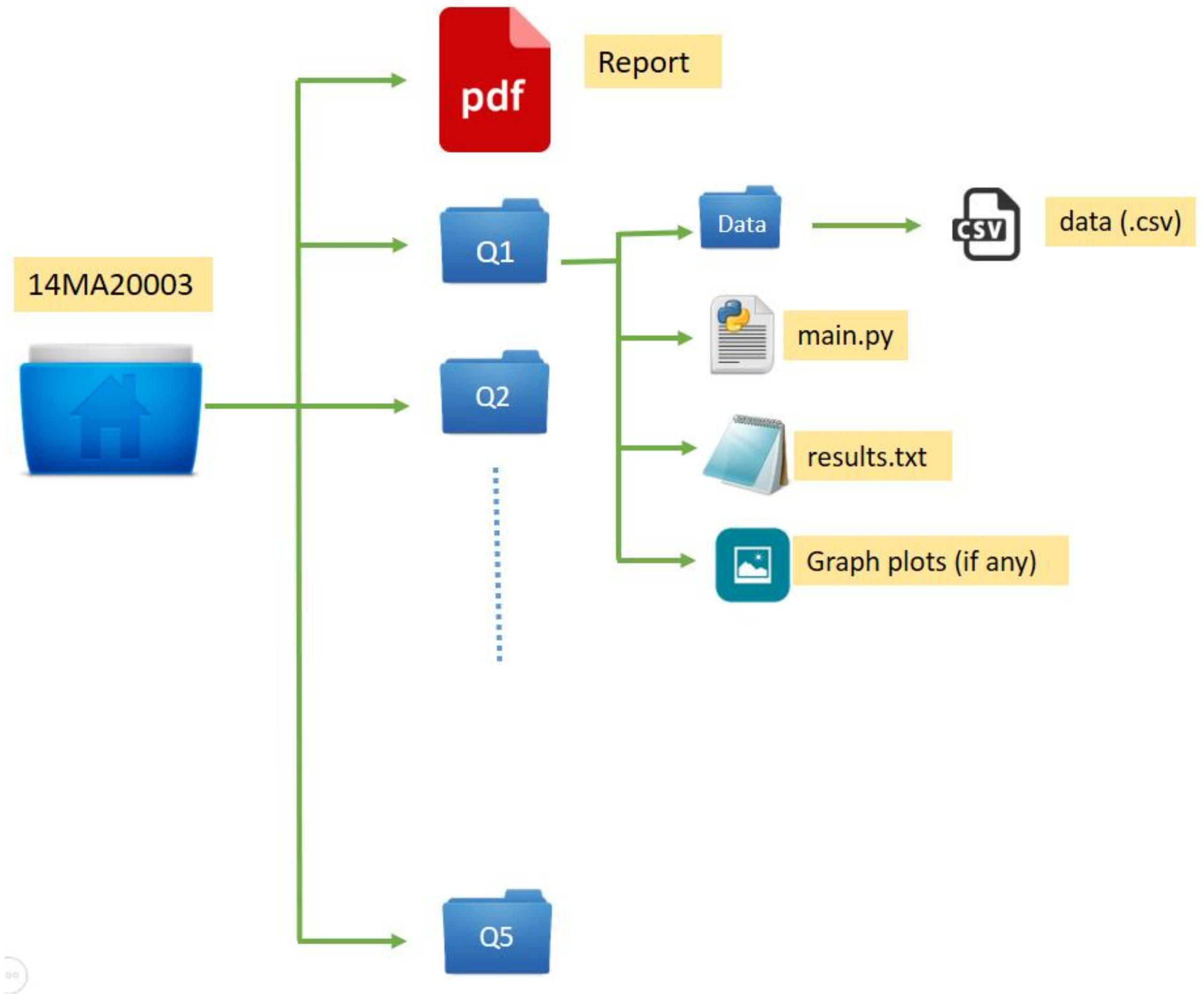
Project Assignment -1

2017-2018

14MA20003

Aakash Singh

Folder Structure :



Topic 1 :

(CAR data with 50 observations)

Tools used : Python-Programming, Spyder, Ubuntu, Python Libraries (pandas, scipy.stats.mstats)

Methodology : Calculate Arithmetic Mean, Geometric Mean and Harmonic mean using inbuilt functions of pandas and scipy.

Reasonable assumptions : none

Topic 2 :

(EARTHQUAKE data with 8086 observations)

Tools used : Python-Programming,, Spyder, Ubuntu, Python Libraries (numpy, pandas, matplotlib.pyplot)

Methodology : -Calulated Lower Quartile, Sample Median, Upper Quartile and Sample Maximum using numpy and pandas inbuilt functions.

-Obtaining Outliers: the range between lower and upper quartiles is multiplied by 1.5 and the values below and above the bigger range are taken as Outliers.

-Boxplot is obtained using matplot

Reasonable assumptions : none

Topic 3 :

(AUTOMOBILE data with 205 observations)

Tools used : Python-Programming,, Spyder, Ubuntu, Python Libraries (numpy, pandas, scipy, matplot)

Methodology :

- Categorization according to NOIR topology manually.
- Cleaning of data (Missing data '?' is replaced by column's average)
- Obtained probability distribution using scipy.gaussian_kde , numpy.linspace and matplot.

Reasonable assumptions : Missing data is assumed a to be mean of remaining data

Topic 4 :

(IRISH data with 50 observations)

Tools used : Python-Programming,, Spyder, Ubuntu, Python Libraries (pandas)

Methodology : Sample Variance and Population Variance is calculates using inbuilt pandas functions and their absolute difference is calculate for comparison.

Reasonable assumptions : none

Topic 5 :

(IRISH data with 50 observations)

Tools used :

Python-Programming,, Spyder, Ubuntu, Python Libraries (pandas, numpy)

Methodology :

-Data Cube is stored as a multidimensional array.

-For the first roll up, set of cities is made.

-For second roll up, set of months are reduced to years.

-Drill downs are going back to saved data frames during roll up operations.

-For slice and dice, python multidimensional array operations are used.

Reasonable assumptions :

none

x