DS 203
Programming for Data Science
Assignment 3

Chirag Garg
Electrical Engineering, Dual Degree
IIT Bombay
Roll No.: 190100042
Mail id: chiraggargb@gmail.com

September 9, 2020

Question 1

The datasets given in the three parts have been displayed in the .ipynb file attched alongwith. The type of each variable has also been printed in the same file.

Difference in python data type and other data type classification

The python data types such as int64, float64, string, etc are useful for analying the given dataset, plotting graphs and creating models to draw meaningful conclusions from them. However, the other type of data classification (categorical/nominal, ordinal, numerical (integer, quantized, continuous) etc.) is useful while data collection, to get a stock of the type of data available and what operations should be performed on a particular variable.

Question 2

The following can be the classification of the analyses mention in the question:

1. Exploratory Data Analysis
2. Descriptive Data Analysis
3. Exploratory Data Analysis
4. Descriptive Data Analysis
5. Predictive Data Analysis
6. Predictive Data Analysis
7. Descriptive Data Analysis

Question 3

Part a

For the prediction of bell-weather stocks, the following fields would be useful to have in the dataset:

- Date
- Stock Name/Stock Code
- Open
- Close
- High
- Low

- Volume Transacted
- Stock Market Exchange
- Market Sector
- Market Cap
- Percentage change
- Percentage Change in Respective Sector Index

The following data analyses should be conducted:

1. Exploratory Data Analysis
   - See if there is any missing field or NaN value for any stock, such as, Opening Price, Closing Price, Percentage Change, High, low, etc.
   - Check if the market sector of each stock is available or not.
   - Check if the coding variable types of respective columns match the expected python variable type.
   - Find out how much data is available for each stock, sector, and economy as a whole.
   - Find out the common time period for which the data is available for each stock, sector, and economy.
   - Check if there are any duplicate rows in the dataset. Remove them before going for the further analyses.
2. Descriptive Data Analysis
   - Plot the candle stick graph for each stock in a sector.
   - Plot a graph between the percentage change in Stock price and the percentage change in price of respective sector index.
   - Draw histogram for the number of stocks available in each sector to know if there are enough stocks to compare with.
3. Predictive Data Analysis
   - Model the future stock price for each stock using the past data available.
   - Predict how much the change in price of the stock is related to the change in prices of respective sector index.
4. Prescriptive Data Analysis
   - Prescribe which stock can be taken as the bell-weather stock for a sector, i.e. , changes in the prices of which stock affect the price change of sector the most.

Part b

For recommending one among two roads for forest clearance, the following fields would be useful in the dataset:

- Year of construction
- Place of construction
- Length of road
- Width of road
- Forest cover destroyed for creation of road
- Reduction in travel time as compared to when the road was absent.
- Carbon footprint

The following data analyses should be conducted:

1. Exploratory Data Analysis
    - Make sure there are no missing values or NaN values in any field such as road length, width, etc.
    - Find out how was the economic growth of the region affected after the construction of road.
    - Find out how was the carbon footprint of the region affected after the construction of road
    - Check if the coding variable types of respective columns match the expected python variable type.
    - Check if there are any duplicate rows in the dataset. Remove them before going for the further analyses.
2. Descriptive Data Analysis
    - Make a graph of carbon footprint versus the length and width of the road.
    - Make histogram showing the number of roads constructed in different years.
    - Analyse the reduction in travel time as a function of carbon footprint generated.
    - Find the correlation between the economic growth of region as a result of construction of road with other factors.
    - Find the correlation between the carbon footprint of region as a result of construction of road with other factors.
3. Predictive Data Analysis

- Model the carbon footprint generated as a function of length and width of the road, and forest cover destroyed for the construction.
- Find out the economic impact of the road on the nearby connecting regions.

4. Prescriptive Data Analysis

    - On the basis of the analysis, prescribe which of the two roads, given their length, width and forest cover destroyed, should be allowed for forest clearance.

Part c

For working upon reducing the gap between the neonatal mortality in the biggest cities versus rest of the state, the following fields would be useful to have in the dataset:

- District Name
- State Name
- City Name
- Year/Time Period
- Literacy level of father
- Literacy level of mother
- Employment Status of father
- Employment status of mother
- Income level of family
- Caste Group of family
- Type of house (Pucca/Kaccha)
- Availability of electricity
- Gender of infant
- Did mother experienced delivery complications (Yes/No)
- Number of tetanus toxoid injections taken by mother

The following data analyses should be conducted:

1. Exploratory Data Analysis:

    - See if there is any missing field such as district name, city name, employment status, etc.

- Check if the Time/Year of birth is available for each infant.
- Check if the coding variable types of respective columns match the expected python variable type.
- Check if there are any duplicate rows in the dataset. Remove them before going for the further analyses.

2. Descriptive Data Analysis:
    - Plot a graph of the number of neonatal mortalities reported versus the year in which they were reported.
    - Plot a histogram of the number of neonatal mortalities reported in a given city versus the year in which they were reported.
    - Create a plot of the number of neonatal mortalities reported in a year versus the city in which they were reported.
3. Predictive Data Analysis:
    - Model the number of neonatal mortalities as a function of family income level, employment status of parents, caste group of family, living conditions of family and gender of infant.
    - Create a model for predicting the number of neonatal mortalities in a given city for future years.
    - Predict a relation between number of neonatal mortalities in a given year for different cities.
4. Prescriptive Data Analysis:
    - Prescribe the measures which can be taken by each district council to reduce the neonatal mortality rate such as ensuring the consumption of an adequate quantity of Tetanus Toxoid (TT) injections by pregnant mothers, targeting vulnerable groups like young, first time and Scheduled Caste mothers, and improving overall household environment by increasing access to improved toilets, electricity, and pucca houses.
    - Prescribing special measures to keep the mortality rate down in rural areas.

This dataset can be useful to work upon.