Assignment 4: DS 203, Fall 2021

Vinit Awale
18D070067

Question 1

Download, read, and display the data at the following URLs in in python using Google CoLab and pandas, and print the type of each variable. Comment on the difference between python data types (float, int, object, etc.) and the data types taught in class (categorical/nominal, ordinal, numerical (integer, quantized,continuous) etc.)

Answer

Dataset 1

There are 11 columns (features) and 74 rows (observations) in the dataset.

Variable	Python Datatype	Kind of data
District	object	Nominal
Facility Type	object	Nominal
Total No. of Facilities	int64	Numerical (Quantized)
No. of facilities reporting nil performance *	int64	Numerical (Quantized)
Performance - Overall Average **	int64	Numerical (Quantized)
Performance - Maximum	int64	Numerical (Quantized)
Performance - Minimum	int64	Numerical (Quantized)
No. of facilities by performance - 1 to 30	int64	Numerical (Quantized)
No. of facilities by performance - 31 to 150	int64	Numerical (Quantized)
No. of facilities by performance - 151 to 300	int64	Numerical (Quantized)
No. of facilities by performance $\geq$ 300	int64	Numerical (Quantized)

Dataset 2

There are 7 columns (features) and 252 rows (observations) in the dataset.

Variable	Python Datatype	Kind of data
Date	object	Temporal
Open	float64	Numerical (Continuous)
High	float64	Numerical (Continuous)
Low	float64	Numerical (Continuous)
Close	float64	Numerical (Continuous)
Adj Close	float64	Numerical (Continuous)
Volume	int64	Numerical (Quantized)

Dataset 3

There are 23 columns (features) and 190 rows (observations) in the dataset.

Variable	Python Datatype	Kind of data
ID	object	Nominal
Motorway	object	Nominal
SR	object	Numerical (Quantized)
NR	object	Numerical (Quantized)
TR	object	Nominal
VR	object	Nominal
SUR1	object	Nominal
SUR2	object	Nominal
SUR3	object	Nominal
UR	object	Nominal
FR	object	Nominal
OR	object	Numerical (Quantized)
RR	object	Ordinal
BR	object	Ordinal
MR	object	Nominal
CR	object	Nominal
Green frogs	object	Nominal
Brown frogs	object	Nominal
Common toad	object	Nominal
Fire-bellied toad	object	Nominal
Tree frog	object	Nominal
Common newt	object	Nominal
Great crested newt	object	Nominal

The kind of datatype of each variable is mentioned in the link where the data is downloaded.

Question 2:

Assume that you are analyzing work travel habits of people from various localities of Mumbai. Classify the following into types of analyses into exploratory, descriptive, predictive, or prescriptive:

- a. Finding whether people from Bandra and Powai have different distance traveled distributions
  Answer: Descriptive Analysis

- b. Analyzing net savings in carbon footprint if a new train station is added to Bandra versus Powai
  Answer: Predictive Analysis

- c. Modeling distance traveled as a function of income, job type, and residence locality
  Answer: Predictive Analysis

- d. Finding ranges of distance traveled variable in the data
  Answer: Exploratory Analysis

- e. Finding the number of samples that have distance traveled variable missing in the data
  Answer: Exploratory Analysis

- f. Finding whether the distribution of distance traveled by commuters is Gaussian or beta
  Answer: Descriptive Analysis
- g. Plotting histograms of number of people by residence locality variable in your data
  Answer: Exploratory Analysis

Question 3:

For each of the following scenarios, search for datasets related to the problem domain (even if the data is not pertaining to the exact situation) for a few minutes to get a sense of what data is collected around the world. Then exercise your imagination to write down reasonable exploratory, descriptive, predictive, and prescriptive data analyses to be done in case of each of the following hypothetical scenarios. Indicate sources of a few other data sets that you find related to each theme. Feel free to indicate if some of these categories of analyses do not apply to a particular scenario. Some loosely related links are provided:

Scenario 1:

As an advisor to a state government, you want to close the gap between the neonatal mortality in the biggest city versus rest of the state, but you have limited resources to work on only a few hospitals.

Answer:

Exploratory analysis

- What data do we have on neonatal mortality in the biggest city and other cities?
- Is the data biased towards urban or rural areas?
- Is the data from a single hospital or from multiple hospitals?
- Is the data biased towards the richer or poorer class?
- Plotting histograms of neonatal mortality by number of days of the child's survival.
- Plotting histograms of neonatal mortality for various cities.

Descriptive analysis

- What is the mean neonatal mortality in the biggest city?
- What is the mean neonatal mortality in the other cities?
- Is it true that the neonatal mortality in the biggest city is lower than the rest of the state?
- What is the range of days when the child's life is at the greatest risk?

Predictive analysis

- What will the neonatal mortality in a city be if a new hospital is added?
- Will the neonatal mortality become lower if the hospital is upgraded?
- Will the rate of neonatal mortality decrease if better family planning measures are adopted?
- Does a decrease in population size affect neonatal mortality?

Prescriptive analysis

- What are the best practices for adding a new hospital?
- What measures should be taken by hospital to reduce neonatal mortality?
- Should the hospital be upgraded?
- Should better family planning measures be adopted?

Datasets

- The World Bank data on neonatal mortality for India.
  `https://data.worldbank.org/indicator/SH.DYN.NMRT?end=2019&locations=IN&start=1969`
- UNICEF data on neonatal mortality for India.
  `https://data.unicef.org/topic/child-survival/neonatal-mortality/`
- Determinants of neonatal mortality in rural India by NIH.
  `https://pubmed.ncbi.nlm.nih.gov/23734339/`

Scenario 2:

As an analyst for a stock market newsletter, you want to recommend bell-weather stocks for different sectors
Answer:

Exploratory analysis

- What data do we have on the stock market?
- Is the data from a single stock exchange or from multiple stock exchanges?
- Is the data biased towards certain sectors?
- Do we have data on the stock market for the last few years?

Descriptive analysis

- Does the price of stock have any relationship with it being ballwether or not?
- What is the mean stock price for a sector in the stock market?
- Is it true that Alphabet (GOOGL), the parent company of Google, is a bellwether stock of the technology sector?
- Is it true that FedEx's (FDX) strong revenues and earnings suggest strong consumer and business shipping activity?
- Does each sector have a bellwether stock?

Predictive analysis

- Which stocks in a sector show a strong correlation with the stock market of the particular sector? Can these stocks be recommended as the ballwether stocks?
- Which stocks will grow if the ballwether stocks (predicted in previous parts) grows?
- Does the stock market of a particular sector have a bellwether stock?
- What will the stock price of Alphabet (GOOGL) be if the company is acquired by a different owner?

Prescriptive analysis

- What stocks should be bought in a particular sector if the predicted ballwether stocks grows?
- What are the best stock which act as an indicator for the behaviour of the ballwether stocks?
- What are the best stocks to buy in a particular sector?

Datasets

- NIFTY-50 Stock Market Data (2000 - 2021)
  `https://www.kaggle.com/rohanrao/nifty50-stock-market-data`
- SP 500 Companies with Financial Information
  `https://datahub.io/core/s-and-p-500-companies`
- Daily News for Stock Market Prediction
  `https://www.kaggle.com/aaron7sun/stocknews`

Scenario 3:

As an intern at the Ministry of Environment, you are under pressure to approve one of the two roads that have been proposed, and you want to recommend the lesser of the two evils.
Answer:

Exploratory analysis

- What data do we have on the road network?
- Is the data biased towards a certain road?
- Is the data from a single construction company or from multiple construction companies?

Descriptive analysis

- What is the mean length of the roads?
- Is the construction cost of any one road much lower than the other?
- Are both the roads of the same type and in similar locations?

Predictive analysis

- Which road is the least expensive?
- Which road will improve the connectivity of the area the most?
- Which of the two roads is the most environment friendly?
- What is the mean length of the roads if the road is built by a different construction company?

Prescriptive analysis

- Which road should be built considering the environmental impact and the construction cost?

Datasets

- Global patterns of current and future road infrastructure
  `https://iopscience.iop.org/article/10.1088/1748-9326/aabd42`