# DS 203: Assignment 4

~ Shubham Lohiya, 18D100020

## Question 1

**Given 3 datasets have been dowloaded and read as Pandas DataFrames using Python. The information regarding each attribute of a data point (columns) has been summarized using the `info` member method of the DataFrame object. The info method displays the datatype of values of a certain attribute. In case of our datasets, `object` type refers to string objects.**
**Also, DateTime data if present in the dataset has been parsed as DateTime objects.**

In [1]:

```
import pandas as pd
```

In [2]:

```
df_facilities = pd.read_excel('facilities.xls')
df_facilities.head()
```

```
WARNING *** OLE2 inconsistency: SSCS size is 0 but SSAT size is non-zero
```

Out[2]:

| | District | Facility Type | Total No. of Facilities # | No. of facilities reporting nil performance * | Performance - Overall Average ** | Performance - Maximum | Performance - Minimum | No. of facilities by performance - 1 to 30 | No. of facilities by performance - 31 to 150 | No facilities performa - 151 to |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Alipurduar | DH | 2 | 1 | 274 | 274 | 274 | 0 | 0 | |
| 1 | Alipurduar | SDH | 2 | 1 | 64 | 64 | 64 | 0 | 1 | |
| 2 | Bankura | DH | 4 | 1 | 302 | 487 | 93 | 0 | 1 | |
| 3 | Bankura | CHC | 23 | 9 | 179 | 607 | 6 | 2 | 4 | |
| 4 | Birbhum | DH | 2 | 0 | 561 | 589 | 534 | 0 | 0 | |

In [3]:

```
df_facilities.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 74 entries, 0 to 73
Data columns (total 11 columns):
 #   Column                                        Non-Null Count  Dtype 
---  ------                                        --------------  ----- 
 0   District                                      74 non-null     object
 1   Facility Type                                 74 non-null     object
 2   Total No. of Facilities #                     74 non-null     int64 
 3   No. of facilities reporting nil performance *  74 non-null     int64 
 4   Performance - Overall Average **              74 non-null     int64 
 5   Performance - Maximum                         74 non-null     int64 
 6   Performance - Minimum                         74 non-null     int64 
 7   No. of facilities by performance - 1 to 30    74 non-null     int64 
 8   No. of facilities by performance - 31 to 150  74 non-null     int64 
 9   No. of facilities by performance - 151 to 300  74 non-null     int64 
 10  No. of facilities by performance - >300       74 non-null     int64 
dtypes: int64(9), object(2)
memory usage: 6.5+ KB
```

In [13]:

```
df_stock = pd.read_csv('stock.csv', parse_dates=['Date'])
df_stock.head()
```

Out[13]:

| | Date | Open | High | Low | Close | Adj Close | Volume |
|---|---|---|---|---|---|---|---|
| 0 | 2020-09-03 | 1709.713989 | 1709.713989 | 1615.060059 | 1641.839966 | 1641.839966 | 3107800 |
| 1 | 2020-09-04 | 1624.260010 | 1645.109985 | 1547.613037 | 1591.040039 | 1591.040039 | 2608600 |
| 2 | 2020-09-08 | 1533.510010 | 1563.864990 | 1528.010010 | 1532.390015 | 1532.390015 | 2610900 |
| 3 | 2020-09-09 | 1557.530029 | 1569.000000 | 1536.051025 | 1556.959961 | 1556.959961 | 1774700 |
| 4 | 2020-09-10 | 1560.640015 | 1584.081055 | 1525.805054 | 1532.020020 | 1532.020020 | 1618600 |

In [14]:

```
df_stock.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 253 entries, 0 to 252
Data columns (total 7 columns):
 #   Column     Non-Null Count  Dtype         
---  ------     --------------  -----         
 0   Date       253 non-null    datetime64[ns]
 1   Open       253 non-null    float64       
 2   High       253 non-null    float64       
 3   Low        253 non-null    float64       
 4   Close      253 non-null    float64       
 5   Adj Close  253 non-null    float64       
 6   Volume     253 non-null    int64         
dtypes: datetime64[ns](1), float64(5), int64(1)
memory usage: 14.0 KB
```

In [20]:

```
df_amphibians = pd.read_csv('amphibians.csv', delimiter=';', skiprows=[0])
df_amphibians.head()
```

Out[20]:

| | ID | Motorway | SR | NR | TR | VR | SUR1 | SUR2 | SUR3 | UR | ... | BR | MR | CR | Green frogs | Brown frogs | Common toad | Fire-bellied toad | Tree frog | Comm n |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | A1 | 600 | 1 | 1 | 4 | 6 | 2 | 10 | 0 | ... | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | |
| 1 | 2 | A1 | 700 | 1 | 5 | 1 | 10 | 6 | 10 | 3 | ... | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | |
| 2 | 3 | A1 | 200 | 1 | 5 | 1 | 10 | 6 | 10 | 3 | ... | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | |
| 3 | 4 | A1 | 300 | 1 | 5 | 0 | 6 | 10 | 2 | 3 | ... | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | |
| 4 | 5 | A1 | 600 | 2 | 1 | 4 | 10 | 2 | 6 | 0 | ... | 5 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | |

**5 rows × 23 columns**

In [16]:

```
df_amphibians.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 189 entries, 0 to 188
Data columns (total 23 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   ID                  189 non-null    int64 
 1   Motorway            189 non-null    object
 2   SR                  189 non-null    int64 
```

```
 3   NR                  189 non-null    int64
 4   TR                  189 non-null    int64
 5   VR                  189 non-null    int64
 6   SUR1                189 non-null    int64
 7   SUR2                189 non-null    int64
 8   SUR3                189 non-null    int64
 9   UR                  189 non-null    int64
 10  FR                  189 non-null    int64
 11  OR                  189 non-null    int64
 12  RR                  189 non-null    int64
 13  BR                  189 non-null    int64
 14  MR                  189 non-null    int64
 15  CR                  189 non-null    int64
 16  Green frogs         189 non-null    int64
 17  Brown frogs         189 non-null    int64
 18  Common toad         189 non-null    int64
 19  Fire-bellied toad   189 non-null    int64
 20  Tree frog           189 non-null    int64
 21  Common newt         189 non-null    int64
 22  Great crested newt  189 non-null    int64
dtypes: int64(22), object(1)
memory usage: 34.1+ KB
```

**Datatypes in python like `int`, `float`, `bool`, `str` exist to serve as the best way to represent and efficiently store data of various types. Different forms of data have different memory and functional requirements, hence efficient handling of such data requires us to have different tailored object classes or containers to store this information. These specialized classes/containers form the python datatypes. For eg, floating point data like decimal numbers are stored as `float` objects in python whereas strings which can be of arbitrary length (forming words, sentences, articles, or even entire books) are stored in specialized `str` objects that are optimized for the same.**
**Since different types of operations are valid on different kinds of data, having datatypes supporting these specialized operations is very beneficial and efficient. For eg, we might want to perform additions or subtractions of floating numbers, whereas in case of strings we have different operations like concatenation and capitalization. In python, datatypes are equipped with support for such common operations on the data they contain.**

**In contrast, statistical data types are defined to segregate the different kinds of data seen in the real world. This data is largely either categorical or numerical. But to have a more finer distinction between different kinds of data and the operations they support, we have the following scales of data: nominal, ordinal, interval and ratio (the latter two being numerical in nature). Numerical datatypes are also classified as either continuous or quantized.**

**So in conclusion, the type of data seen is categorized among the statistical data types, and when we work with such data on our computer in say python programming, we use the programming data types to store this data and perform operations on them. For eg, floating point data can be categorized as continuous, and either ratio or interval. But when we work with this data in python, they are available to us in the form of `float` objects.**

## Question 2

**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**
**This is *descriptive analysis*, as we are trying to answer a question using the data.**

**b. Analyzing net savings in carbon footprint if a new train station is added to Bandra versus Powai**

**This is *prescriptive analysis*, as we are trying to see how to achieve our objective of greatly impacting the carbon footprint through our data analysis.**

**c. Modeling distance traveled as a function of income, job type, and residence locality**
**This is *predictive analysis*, as we are trying to predict traveled distance by modelling it as a function of given variables.**

**d. Finding ranges of distance traveled variable in the data**
**This is *exploratory analysis*, as we are analysing a property of our data.**

**e. Finding the number of samples that have distance traveled variable missing in the data**
**This is *exploratory analysis*, as we are analysing a property of our data.**

**f. Finding whether the distribution of distance traveled by commuters is Gaussian or beta**
**This is *descriptive analysis*, as we are trying to answer a question based on our data.**

**g. Plotting histograms of number of people by residence locality variable in your data**
**This is *exploratory analysis*, as we are trying to analyse a property of our data through visulaization.**

## Question 3

**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:**

**For all questions -> Initial Check: Verify data authenticity, and check if it is representative of the task at hand.**

**a) 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.**

**References:**

- https://pubmed.ncbi.nlm.nih.gov/23734339/
- https://www.sciencedirect.com/science/article/pii/S0140673605710485
- https://academic.oup.com/ije/article/35/3/706/735707?login=true

***Exploratory Analysis***

- **Check out the data attributes available in the dataset which are being considered as factors in determining neonatal mortality, and the data type of each attribute. If the attribute is categorical, discern the number of categories and counts. If the attribute is numerical, analyze the spread, mean, and other aspects of its data**
- **Deal with missing values by methods like dropping incomplete rows, or filling in missing information with an appropriate choice of measure such as mean, median, mode, or weighted average**
- **Evaluate the composition of the dataset, i.e. check what proportions of the dataset correspond to positive and negative cases of neonatal mortality. Evaluate how balanced or unbalanced our dataset is**
- **Do this analysis on complete state's data and also the data subset corresponding to the target city and note differences**

***Descriptive Analysis***

- **Analyze correlations of various attributes with the response variable ( $1$[is neodatal death example]) and see if there are any obvious patterns. This will help to see if a particular attribute has any importance, or if it just acts as noise (and thus can be discarded)**
- **Check the distribution of various cause of neonatal mortality (like Congenital, Sepsis, Asphyxia, Preterm, Diarrhoea, Tetanus, etc.) and try to answer the question: Which are the most common causes, and how does this cause distribution vary between the entire state vs just the target city**
- **Ponder the questions: How do factors like family income, education, and child birthweight affect the chance of neonatal mortality, and how to these factors differ over the entire state vs just the target city**
- **Are some hospitals in the target city more likely to have neonatal deaths? What could be the cause?**

***Predictive Analysis***

- **Model the probability of neonatal mortality as a function of available attributes, such that this model can be used to predict chance of neonatal mortality given the attributes**
- **Compare the models for the entire state vs the target city, and see which factors have a greater effect on the model's output how these differ between the two models**
- **Analyze the outcome of various possible steps that could be taken to reduce the rate of neonatal mortality in the target city and close the gap with the rest of the state**

***Prescriptive Analysis***

- **Make data-driven recommendations on what changes need to taken, which income or education status group needs specific care to avoid neonatal cases**
- **Recommend some hospitals on priority (due to limited funds) that require more funding for better care and sanitation which can help reduce the neonatal mortality rate gap**
- **Get suggestions based on what kind of awareness needs to be spread to avoid such cases in the long run**

**b) As an analyst for a stock market newsletter, you want to recommend bell-weather stocks for different sectors.**

**References:**

- https://www.investopedia.com/terms/b/bellwether-stock.asp
- https://academic.oup.com/rfs/article-abstract/28/11/3153/1636922
- https://www.atlantis-press.com/proceedings/ssmi-19/125925420

***Exploratory Analysis***

- **Check out the stock data available sectorwise, and what kind of attributes we have for a particular stock. Also gather data about the general economy's and every sector's historical performance**
- **Deal with missing values by methods like dropping incomplete rows, or filling in missing information with an appropriate choice of measure such as mean, median, mode, or weighted average**
- **Evaluate the proportion of large-cap, mid-cap, and small-cap companies in the dataset**
- **Check the time-horizon of the data available**

***Descriptive Analysis***

- **Analyze how the performance of particular stocks compares with the overall performance of the sector in belongs to, as well as the overall economy, giving more weight to bigger (higher market-cap) companies. Can use correlation, trend and seasonal analyses**
- **How much do certain companies contribute a particular sector's performance?**

***Predictive Analysis***

- **Model the sector's or the economy's performance as a function of the most important stocks in that sector or the economy**
- **Discern a relation between trends / seasonal patterns of particular stock vs its sector's performance**
- **Will a currently important candidate for Bellwether stock for a sector be consistently important, or is it volatile? Forecasting analyses can be conducted**

***Prescriptive Analysis***

- **Recommend Bellwether stock for each sector and the whole economy that makes up a portfolio reflecting the sector or economy's performance over the desired time horizon in expectation**
- **Prescribe bellwether stock that is expected to be stable and representative of the sector's growth over the desired time horizon**

**c) 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.**

**References:**


- https://www.nbmcw.com/article-report/infrastructure-construction/roads-and-pavements/clearances-required-under-environment-acts-for-highway-projects.html
- https://gshp2.gov.in/sites/default/files/2.pdf
- https://www.tandfonline.com/doi/full/10.1080/14615517.2016.1176403


***Exploratory Analysis***

- **Check out the data available, and discern what kind of regions and geographies do we have the data from, and what kind of attributes are under consideration**
- **Deal with missing values by methods like dropping incomplete rows, or filling in missing information with an appropriate choice of measure such as mean, median, mode, or weighted average**
- **What do the attributes tell the impact of construction of a certain road, or its bad consequences?**
- **Visualize the data for projects to see the outcomes vs cost for initial analysis**

***Descriptive Analysis***

- **How much did the travellers benefit from the construction for the road for past projects? How does it compare to the cost of deforestation and road widening required?**
- **How do benefits and costs compare between the two options presented?**
- **In what ways can road construction affect life in rural areas and forests?**
- **What were the pollution (air and noise) related implications of past projects?**

***Predictive Analysis***

- **Model consequences of both road constructions based on past data available from similar projects and compare benefits vs cost, to check if benefits more than outweigh the costs, and how it differs between the two projects.**
- **Evaluate how robust our analysis is by modelling the uncertainty in our models.**

***Prescriptive Analysis***

- **Prescribe based on the analysis which road construction should be approved**
- **Recommend methods to reduce the bad effects on wildlife and environment based on data analysis from past projects and evaluating contributing factors**
- **Estimate the funds and resources required to complete the project based on analysis of similar projects and our objective of minimizing environmental harm**