Python and Tableau: The Compete Data Analytics Bootcamp!

A Data Analysts Toolkit

When you become a Data Analyst, there are two things that you should be skilled at

- Python (using Anaconda)
- Data Visualization tool like Tableau

Projects We'll Be Working On

Sales Analysis for Value Inc

Value Inc.

Sales Analysis for Value Inc: Value Inc is a retail store that sells household items all over the world by bulk.

The Sales Manager has:

- No sales reporting he has a brief idea
- Has no idea of the monthly cost, profit and top selling products.
- He wants a dashboard on this and says the data is currently stored in an excel sheet.

Projects We'll Be Working On

Loan Analysis for Blue Bank

BLUE BANK
SIMPLE BANKING

Blue Bank is a bank in USA that has a loan department which is currently understaffed. Using Python and Tableau, they'd like to see:

- Report of borrowers who may have issues paying back the loan in a Tableau Dashboard

Projects We'll Be Working On

Sentiment and Keyword Analysis for BlogMe

BlogMe, a famous blogging business has a dataset of news articles that they need further analysis on.

- Firstly, they’d like keywords to be extracted from headlines of the article.
- Secondly, they would need to determine the sentiment of the news articles.

What is Python?

- Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language.
- Created by Guido van Rossum during 1985- 1990.
- Used to create a variety of different programs and isn’t specialized for any specific problems.
- This versatility, along with its beginner-friendliness, has made it one of the most-used programming languages today.

Python has become one of the most popular programming languages in the world in recent years.

It's used in everything from machine learning to building websites and software testing. It can be used by developers and non-developers alike. Python is commonly used for developing websites and software, task automation, data analysis, and data visualization.

What is Tableau?

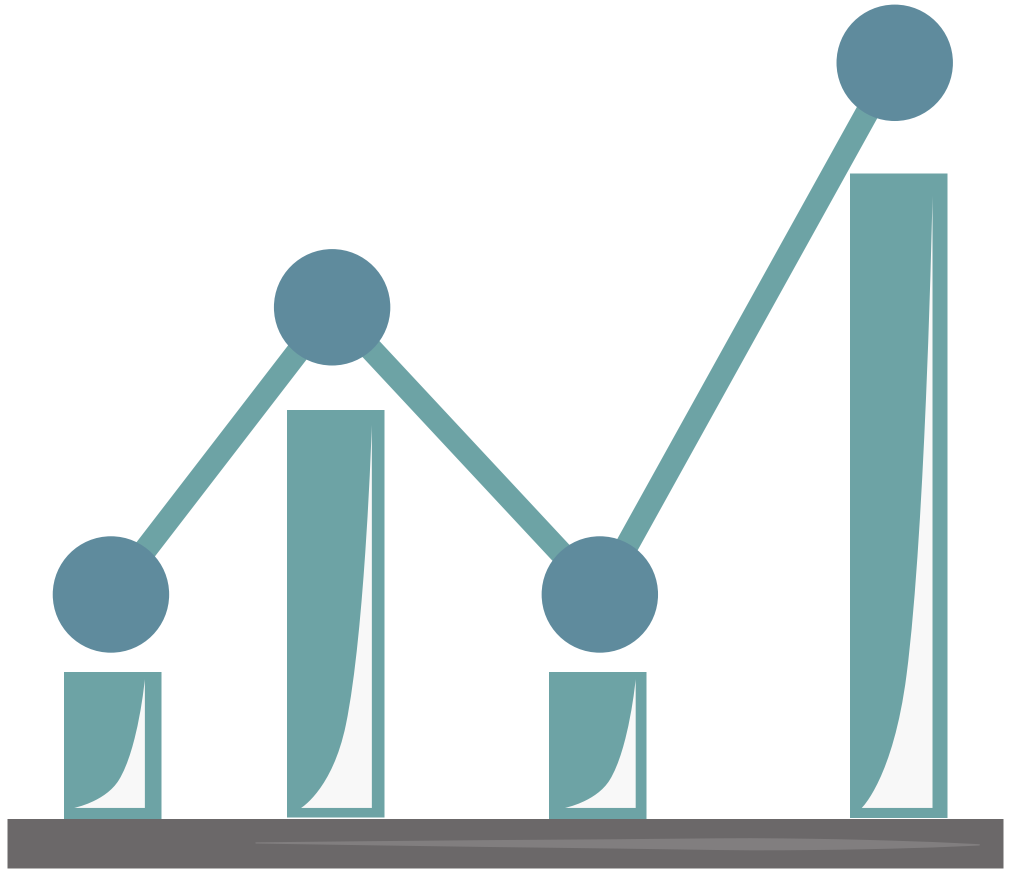
- Tableau is a Business Intelligence tool for visually analyzing the data.
- Users can create and distribute an interactive and shareable dashboard, which depict the trends, variations, and density of the data in the form of graphs and charts.
- Tableau can connect to files, relational and Big Data sources to acquire and process data.
- The software allows data blending and real-time collaboration,

Tableau Features

- Tableau supports powerful data discovery and exploration that enables users to answer important questions in seconds
- No prior programming knowledge is needed; users without relevant experience can start immediately with creating visualizations using Tableau
- It can connect to several data sources that other BI tools do not support. Tableau enables users to create reports by joining and blending different datasets

In this course we are using Tableau Public, the free version of Tableau Desktop. It has most features of Tableau except that in Tableau Public we can only connect to flat files and we can only publish online (We can't save our work locally on our computer- have to save it online)

What Next



We've set up Python

- Download Anaconda on Windows/Mac
- Use Spyder

We've set up Tableau Public

- Download Tableau on Windows/Mac

Make sure you have the following documents

- Project Brief.pdf
- Lecture Slides.pdf

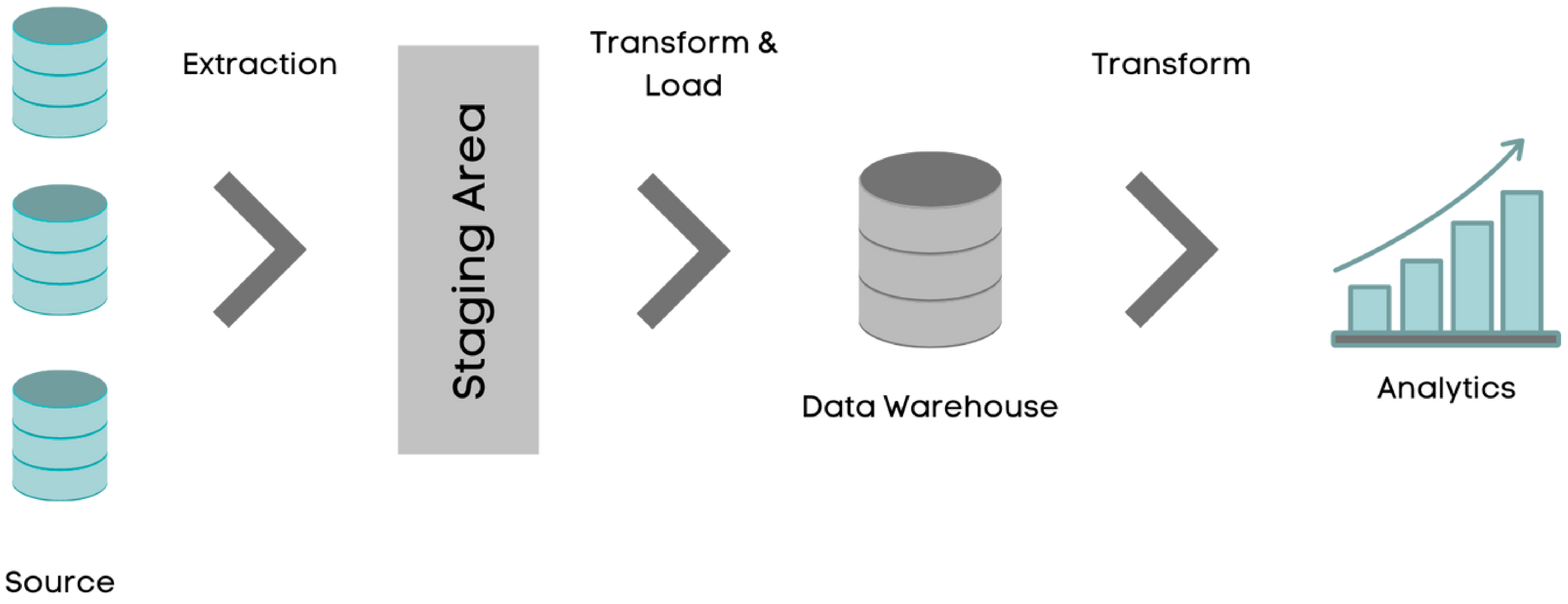
The course is structured in two halves.

- First Half - Learning Python
- Second Half - Learning Tableau

Introduction: The Data Pipeline

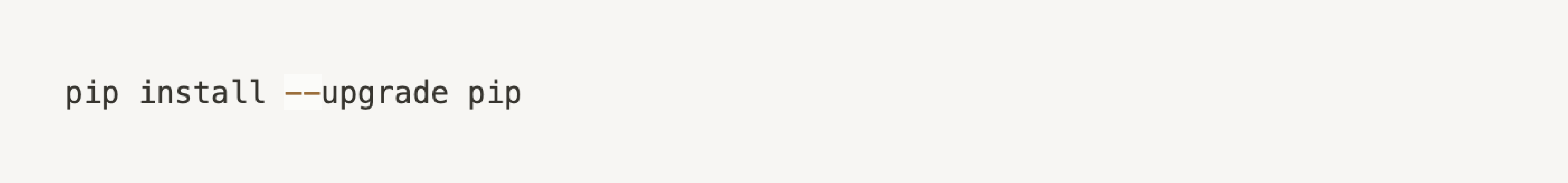
So before we get into the nitty gritty of Python and Tableau, it's really vital for you to understand the Data Pipeline. A data pipeline is a set of actions that ingest raw data from disparate sources and move the data to a destination for storage and analysis.

The Data Cycle



What is Pip

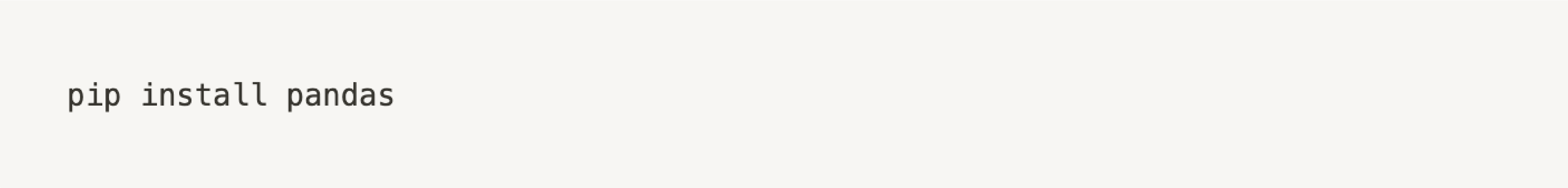
- Pip is a package-management system written in Python used to install, uninstall and manage software packages.
- You need to go to command prompt on windows or terminal in mac and type the following



```
pip install --upgrade pip
```

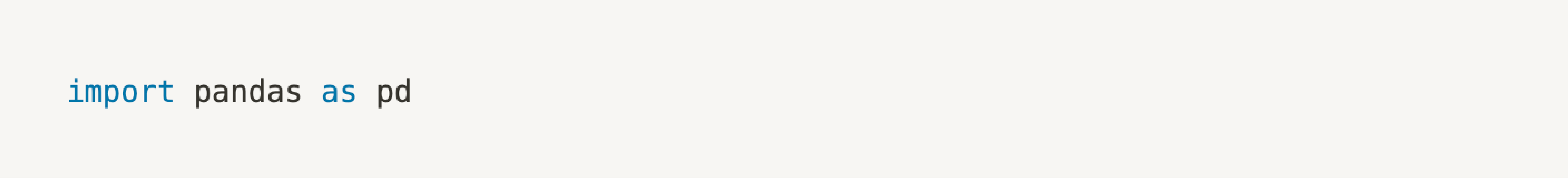
What is Pandas?

- Pandas is a Python library for data analysis. Started by Wes McKinney in 2008 out of a need for a powerful and flexible quantitative analysis. You'll use it a lot and you'll see why.
- You need to go to command prompt on windows or terminal in mac and type the following*

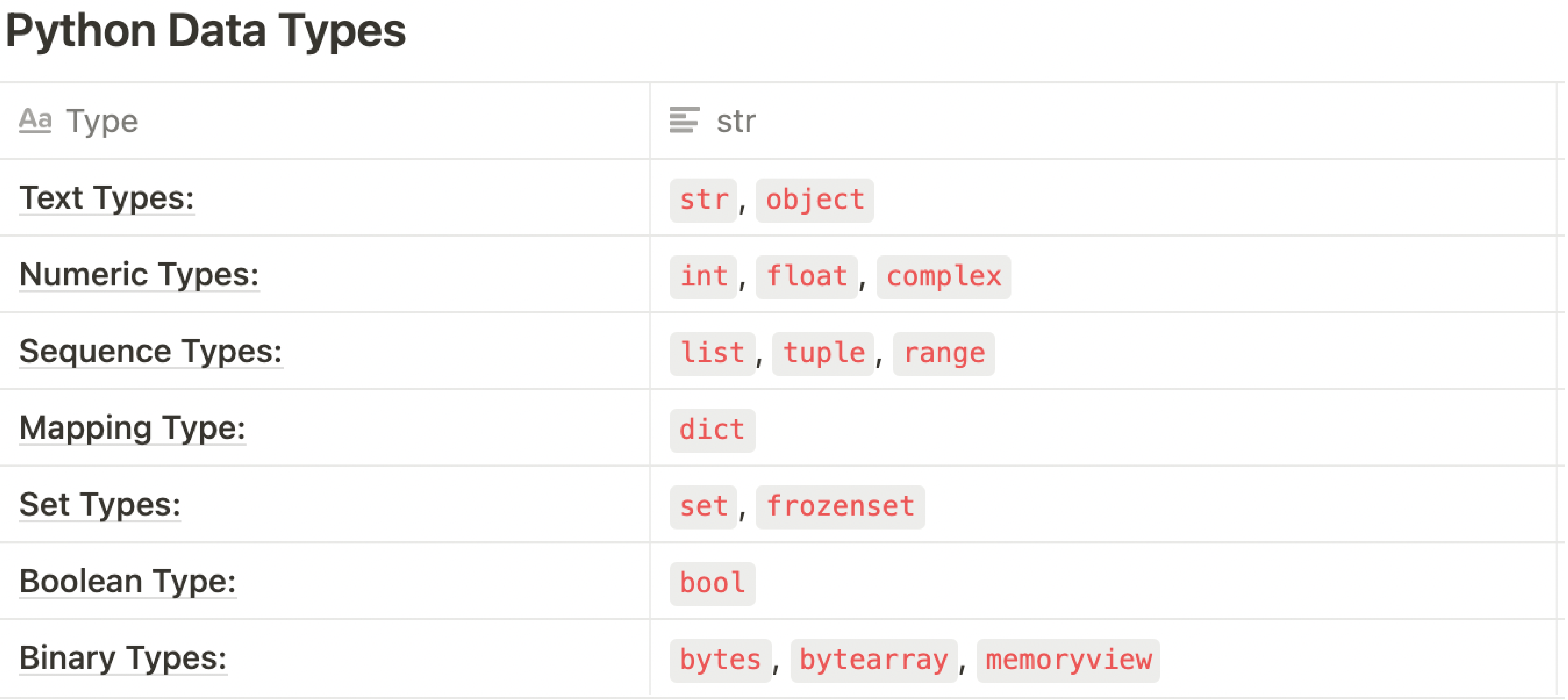


```
pip install pandas
```

Once you install pandas once, you don't need to install it again. Now that this is done, we can go back to Spyder and let's import pandas



```python
import pandas as pd
```



Python Data Types

Type	str
Text Types:	`str`, `object`
Numeric Types:	`int`, `float`, `complex`
Sequence Types:	`list`, `tuple`, `range`
Mapping Type:	`dict`
Set Types:	`set`, `frozenset`
Boolean Type:	`bool`
Binary Types:	`bytes`, `bytearray`, `memoryview`

Investigating Variables

String: These are characters or a mix of characters and numbers
Int: These are whole numbers
Float: These are decimals
List: A collection of items. You can change a list, for example, if I want to change pear to banana I can with python.
Tuple: A collection of items but you can't change the items in a tuple. So if I want to change pear to banana for a tuple, I'm not able to. I'll have to create a new tuple.
Range: This is a range of numbers ex range(10) represents a start point of 0 and an end point of 10. A range like this: range(2,9) then the start point is 2 and the end point is 9.
Dictionary: Dictionaries consist of pairs of keys and their corresponding values.
Set: Sets store unordered values. And unlike Tuples and Lists, Sets can have no duplicate data
Bool: Represents true or false

Sales Analysis for Value Inc

Value Inc.

Sales Analysis for Value Inc: Value Inc is a retail store that sells household items all over the world by bulk.

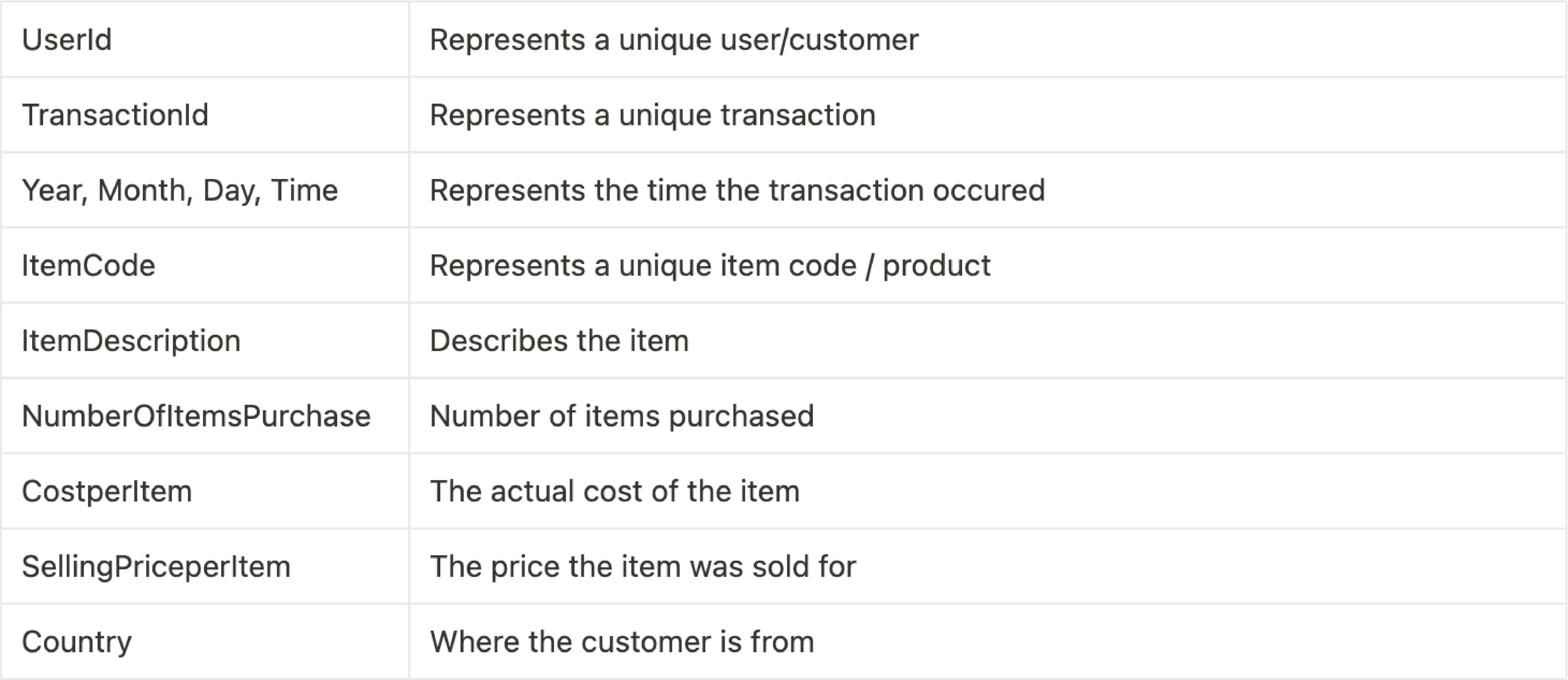
The Sales Manager has:

- No sales reporting but he has a brief idea
- Has no idea of the monthly cost, profit and top selling products.
- He wants a dashboard on this and says the data is currently stored in an excel sheet.

Files to Download

Data Files:	transaction.csv
Logo:	Value Inc. Logo.png

Looking at the Columns



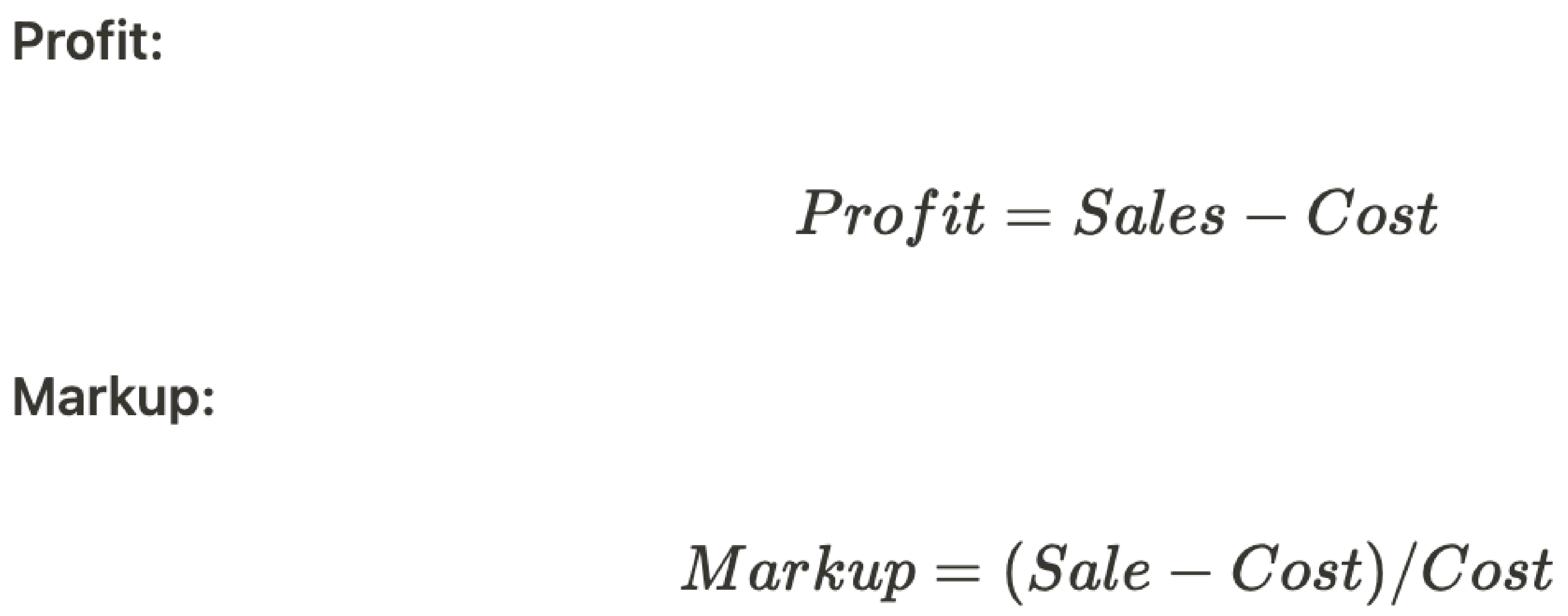
UserId	Represents a unique user/customer
TransactionId	Represents a unique transaction
Year, Month, Day, Time	Represents the time the transaction occured
ItemCode	Represents a unique item code / product
ItemDescription	Describes the item
NumberOfItemsPurchase	Number of items purchased
CostperItem	The actual cost of the item
SellingPriceperItem	The price the item was sold for
Country	Where the customer is from

What is a Series?

- A Pandas Series is like a column in a table. It is a one-dimensional array holding data of any type.

Profit and Markup

- One of the important metrics in something like sales data is Profit and Markup. The formula is below



Profit:

$$Profit = Sales - Cost$$

Markup:

$$Markup = (Sale - Cost)/Cost$$

Round() Function

- The round() function returns a floating point number that is a rounded version of the specified number, with the specified number of decimals.
- The default number of decimals is 0, meaning that the function will return the nearest integer.
- Syntax: ROUND(variable, digits)
- You can also view other lists of functions here:

🔗 https://www.w3schools.com/python/python_ref_functions.asp

Loc

- Pandas DataFrame.loc attribute accesses a group of rows and columns by label(s) or a boolean array in the given DataFrame.

- See the link below

🔗 https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.loc.html

Split()

- The split() method splits a string into a list.
- You can specify the separator, default separator is any whitespace.
- string.split(separator, maxsplit)

🔗 https://www.w3schools.com/python/ref_string_split.asp

Replace()

- The replace() method replaces a specified phrase with another specified phrase.
- string.replace(oldvalue, newvalue, count)

🔗 https://www.w3schools.com/python/ref_string_replace.asp

Lower()

- The lower() method returns a string where all characters are lower case.
- string.lower()

https://www.w3schools.com/python/ref_string_lower.asp

drop()

- The drop() function is used to drop specified labels from rows or columns.
- Remove rows or columns by specifying label names and corresponding axis, or by specifying directly index or column names. When using a multi-index, labels on different levels can be removed by specifying the level.

🔗 https://www.w3resource.com/pandas/dataframe/dataframe-drop.php

pandas.DataFrame.to_csv()

- By using pandas.DataFrame.to_csv() method you can write/save/export a pandas DataFrame to CSV File.
- By default to_csv() method export DataFrame to a CSV file with comma delimiter and row index as the first column

🔗 https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.to_csv.html

Blue Bank Loan Analysis

Blue Bank is a bank in USA that has a loan department which is currently understaffed. They supply loans to individuals and don’t have much reporting on how risky these borrowers are. Using Python and Tableau, they’d like to see a report of borrowers who may have issues paying back the loan.

Files to Download

Data Files:	loan_data.csv
Logo:	Blue Bank Logo.png

JSON Files

JSON is a lightweight data-interchange format and is plain text written in JavaScript object notation

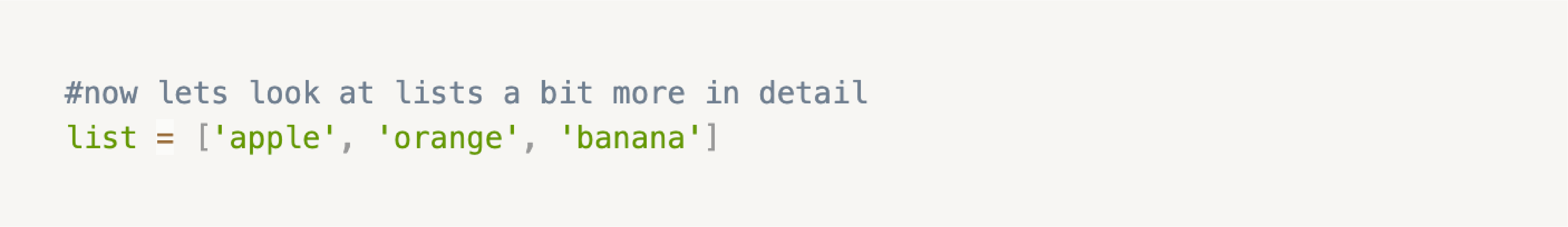
with statement

- with statement in Python is used in exception handling to make the code cleaner and much more readable. It simplifies the management of common resources like file streams.

🔗 https://www.w3schools.com/python/ref_string_lower.asp

Lists

- Lists are used to store multiple items in a single variable.
- Lists are one of 4 built-in data types in Python used to store collections of data, the other 3 are Tuple, Set, and Dictionary, all with different qualities and usage.
- Lists are created using square brackets

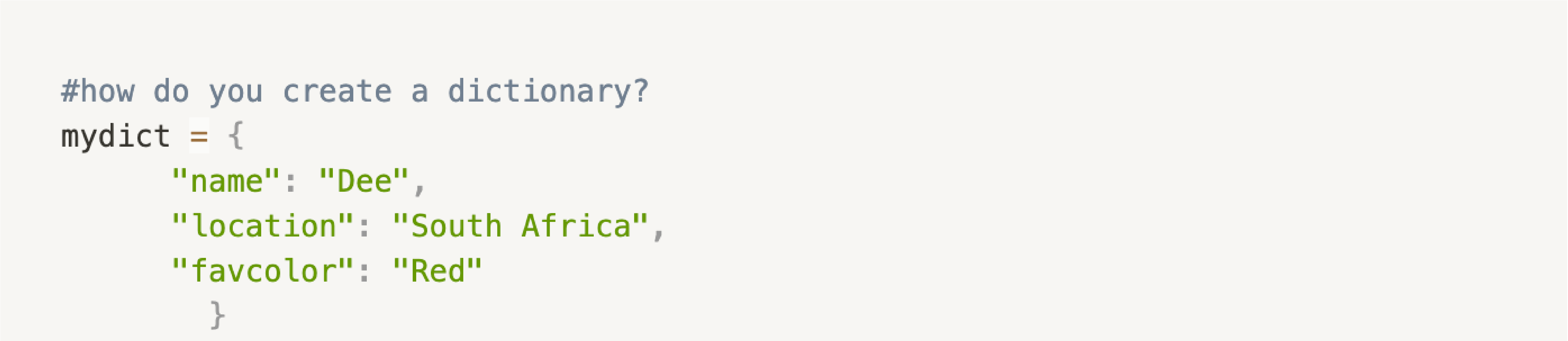


```
#now lets look at lists a bit more in detail
list = ['apple', 'orange', 'banana']
```

List items are ordered, changeable, and allow duplicate values.

Dictionaries

- Dictionaries are used to store data values in key:value pairs.
- A dictionary is a collection which is ordered, changeable and do not allow duplicates
- Dictionaries are written with curly brackets, and have keys and values



```
#how do you create a dictionary?
mydict = {
      "name": "Dee",
      "location": "South Africa",
      "favcolor": "Red"
        }
```

Dataframe

- Pandas DataFrame is two-dimensional size-mutable, potentially heterogeneous tabular data structure with labeled axes (rows and columns).
- A Data frame is a two-dimensional data structure, i.e., data is aligned in a tabular fashion in rows and columns. Pandas DataFrame consists of three principal components, the data, rows, and columns.

Columns Blue Bank Data

credit.policy: 1 if the customer meets the credit underwriting criteria of Blue Bank, and 0 otherwise.

purpose: The purpose of the loan (takes values "credit_card", "debt_consolidation", "educational", "major_purchase", "small_business", and "all_other").

int.rate: The interest rate of the loan, as a proportion (a rate of 11% would be stored as 0.11). Borrowers judged by Blue Bank to be more risky are assigned higher interest rates.

installment: The monthly installments owed by the borrower if the loan is funded.

log.annual.inc: The natural log of the self-reported annual income of the borrower.

dti: The debt-to-income ratio of the borrower (amount of debt divided by annual income).

- dti>1 then borrower has more debt than income.
- dti<1 then borrower has more income than debt

Columns

fico: The FICO credit score of the borrower.

- 300 - 400: Very Poor
- 401 - 600: Poor
- 601 - 660: Fair
- 661 - 780: Good
- 781 - 850: Excellent

days.with.cr.line: The number of days the borrower has had a credit line.

revol.bal: The borrower's revolving balance (amount unpaid at the end of the credit card billing cycle).

revol.util: The borrower's revolving line utilization rate (the amount of the credit line used relative to total credit available).

inq.last.6mths: The borrower's number of inquiries by creditors in the last 6 months. (If there are a lot of inquiries, that’s an issue)

delinq.2yrs: The number of times the borrower had been 30+ days past due on a payment in the past 2 years.
pub.rec: The borrower's number of derogatory public records (bankruptcy filings, tax liens, or judgments).

Unique()

- unique() method is used to know all type of unique values in a column.

describe()

- Pandas describe() is used to view some basic statistical details like percentile, mean, std etc. of a data frame or a series of numeric values.
- Syntax: DataFrame.describe(percentiles=None, include=None, exclude=None)

🔗 https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.describe.html

Numpy()

- NumPy is a Python library used for working with arrays.
- It also has functions for working in domain of linear algebra, fourier transform, and matrices.

Why Use NumPy?

- In Python we have lists that serve the purpose of arrays, but they are slow to process.
- NumPy aims to provide an array object that is up to 50x faster than traditional Python lists.

https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.describe.html

IF Statement

- Python if Statement is used for decision-making operations.
- It contains a body of code which runs only when the condition given in the if statement is true. If the condition is false, then the optional else statement runs which contains some code for the else condition.

https://www.w3schools.com/python/python_conditions.asp

Python supports the usual logical conditions from mathematics:

- Equals: a == b
- Not Equals: a != b
- Less than: a < b
- Less than or equal to: a <= b
- Greater than: a > b
- Greater than or equal to: a >= b

FICO RANGE

fico >= 300 and < 400: 'Very Poor'

fico >= 400 and ficoscore < 600: 'Poor'

fico >= 601 and ficoscore < 660: 'Fair'

fico >= 660 and ficoscore < 780: 'Good'

fico >=780: 'Excellent'

For Loops

- A for loop is used for iterating over a sequence (that is either a list, a tuple, a dictionary, a set, or a string).
- This is less like the for keyword in other programming languages, and works more like an iterator method as found in other object-orientated programming languages.
- With the for loop we can execute a set of statements, once for each item in a list, tuple, set etc.

🔗 https://www.w3schools.com/python/python_for_loops.asp

Python Try and Except

- When an error occurs, or exception as we call it, Python will normally stop and generate an error message.
- These exceptions can be handled using the try statement
- The try block lets you test a block of code for errors.
- The except block lets you handle the error.

🔗 https://www.w3schools.com/python/python_try_except.asp

Matplotlib

- Matplotlib is a multi-platform data visualization library built on NumPy arrays, and designed to work with the broader SciPy stack
- Most of the Matplotlib utilities lies under the pyplot submodule, and are usually imported under the plt alias:

🔗 https://www.w3schools.com/python/matplotlib_pyplot.asp

Groupby

- Pandas groupby is used for grouping the data according to the categories and apply a function to the categories. It also helps to aggregate data efficiently.
- Pandas dataframe.groupby() function is used to split the data into groups based on some criteria. pandas objects can be split on any of their axes.
- Syntax: DataFrame.groupby(by=None, axis=0, level=None, as_index=True, sort=True, group_keys=True, squeeze=False, **kwargs)

🔗 https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.groupby.html

Size()

- The size of an array is the total number of elements in the array

BlogMe: Sentiment and Keyword Analysis

BlogMe, a famous blogging business has a dataset of news articles that they need further analysis on. Firstly, they'd like keywords to be extracted from headlines of the article and secondly, they would need to determine the sentiment of the news articles.

Files to Download

Data Files:	articles.xlsx BlogMe_sources.xlsx
Logo:	BlogMe Logo.png

Functions

- A function is a block of code which only runs when it is called.
- You can pass data, known as parameters, into a function.
- A function can return data as a result.

https://www.w3schools.com/python/python_functions.asp

Classes

Functions generally represent general calculations/formula in your script.

Classes are similar to functions however Classes (or rather their instances) are for representing things.

Classes are used to define the operations.

If your application needs to keep track of people, then **Person** is probably a class; the instances of this class represent particular people you are tracking (the data).

🔗 https://www.w3schools.com/python/python_classes.asp

Classes

A class is a blueprint for how something should be defined. It doesn’t actually contain any data. So something like a Car class will specify that a car name and car make are necessary for defining a car, but it doesn’t contain the name or make of any specific car.

While the class is the blueprint, an instance is an object that is built from a class and contains real data. An instance of the Car class is not a blueprint anymore. It’s an actual car with a car name, like Ford, that is a F150.

Put another way, a class is like a form or questionnaire. An instance is like a form that has been filled out with information. Just like many people can fill out the same form with their own unique information, many instances can be created from a single class.

Class is a Blueprint, an instance of a class is the actual data.

https://www.w3schools.com/python/python_classes.asp

VADER

- In our project, we will be using VADER sentiment analysis.
- VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. VADER uses a combination of A sentiment lexicon is a list of lexical features (e.g., words) which are generally labelled according to their semantic orientation as either positive or negative.
- VADER has been found to be quite successful when dealing with social media texts, NY Times editorials, movie reviews, and product reviews. This is because VADER not only tells about the Positivity and Negativity score but also tells us about how positive or negative a sentiment is.
- It is fully open-sourced under the MIT License.

https://pypi.org/project/vaderSentiment/

VADER

Advantages of using VADER

- VADER has a lot of advantages over traditional methods of Sentiment Analysis, including:
- It works exceedingly well on social media type text, yet readily generalizes to multiple domains
- It doesn’t require any training data but is constructed from a generalizable, valence-based, human-curated gold standard sentiment lexicon

🔗 https://pypi.org/project/vaderSentiment/

TABLEAU

Tableau Workbook, Worksheets and Dashboards

Tableau uses a workbook and sheet file structure, much like Microsoft Excel. A workbook contains sheets. A sheet can be a worksheet, a dashboard, or a story.

- A worksheet contains a single view along with shelves, cards, legends, and the Data and Analytics panes in its side bar.

- A dashboard is a collection of views from multiple worksheets. The Dashboard and Layout panes are available in its side bar.

Joins and Unions

UNION – the data sources have the same columns and will be stacked on top of one another, creating a longer table

JOIN – the data sources have one or more columns in common that you can combine together, creating a wider table

Groups

You can create a group to combine related members in a field.

For example, if you are working with a view that shows average test scores by major, you might want to group certain majors together to create major categories.

English and History might be combined into a group called Liberal Arts Majors, while Biology and Physics might be grouped as Science Majors. Groups are useful for both correcting data errors as well as answering "what if" type questions

Sets

You can use sets to compare and ask questions about a subset of data.

Sets are custom fields that define a subset of data based on some conditions.

Filters

Filtering is an essential part of analyzing data. This article describes the many ways you can filter data from your view.

It also describes how you can display interactive filters in the view, and format filters in the view.

Calculated Fields

If your underlying data doesn't include all of the fields you need to answer your questions, you can create new fields in Tableau using calculations and then save them as part of your data source.

These fields are called calculated fields.

Parameters

A parameter is a workbook variable such as a number, date, or string that can replace a constant value in a calculation, filter, or reference line.

For example, you may create a calculated field that returns True if Sales is greater than $500,000 and otherwise returns False.

You can replace the constant value of “500000” in the formula with a parameter. Then, using the parameter control, you can dynamically change the threshold in your calculation.