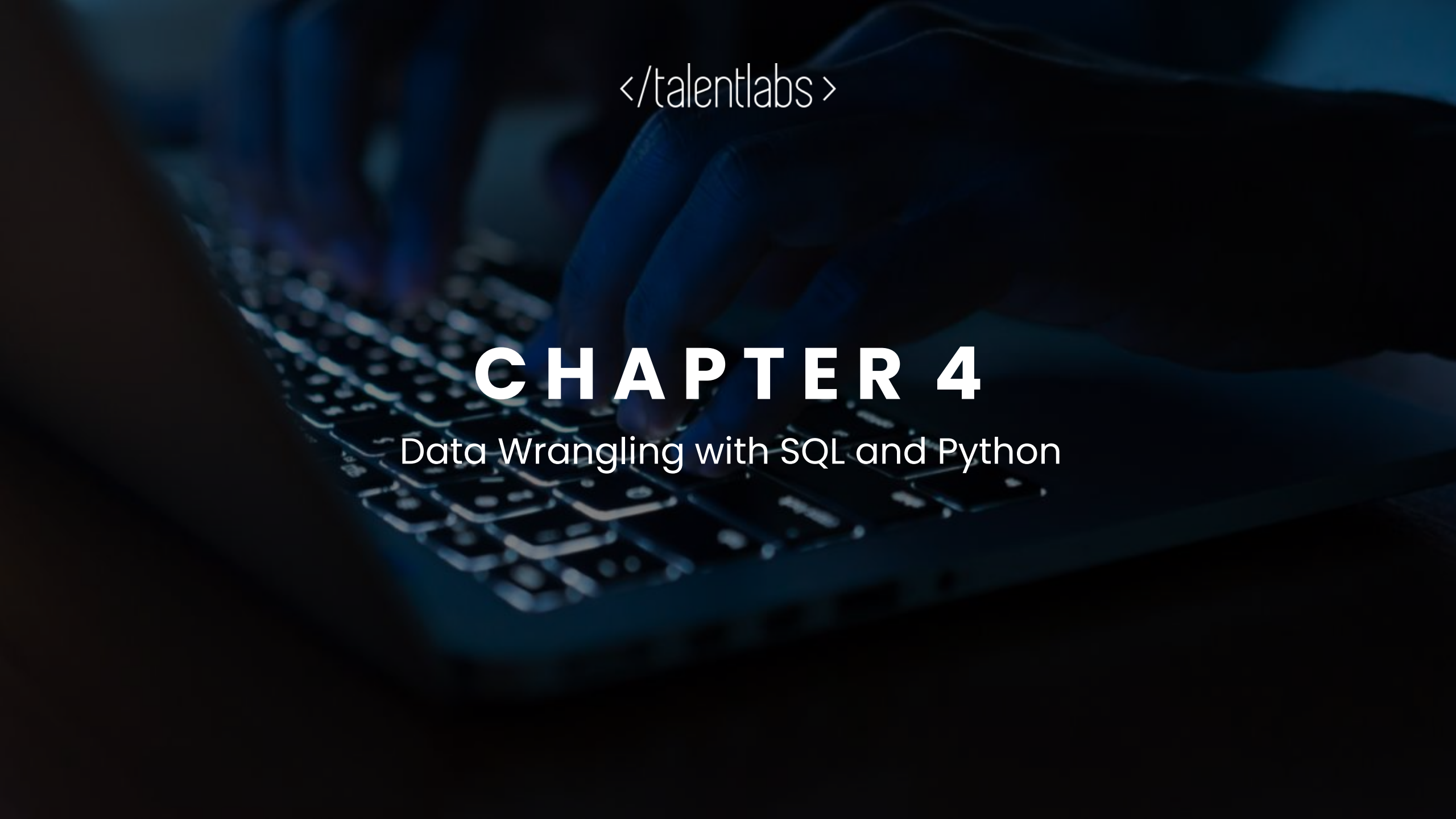
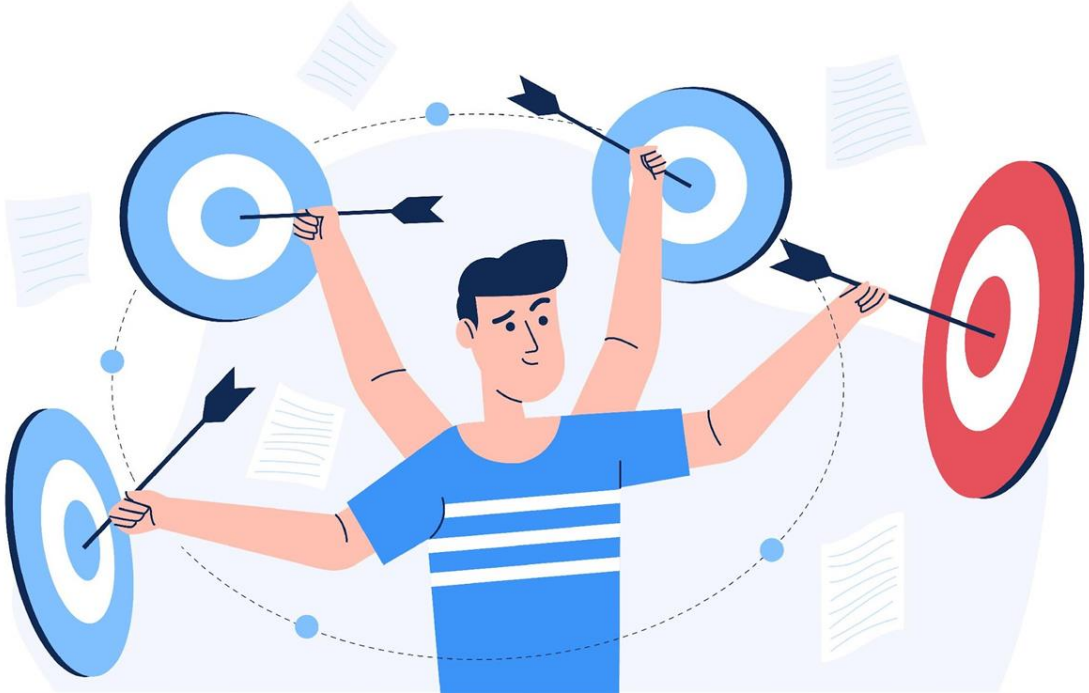


CHAPTER 4

Data Wrangling with SQL and Python

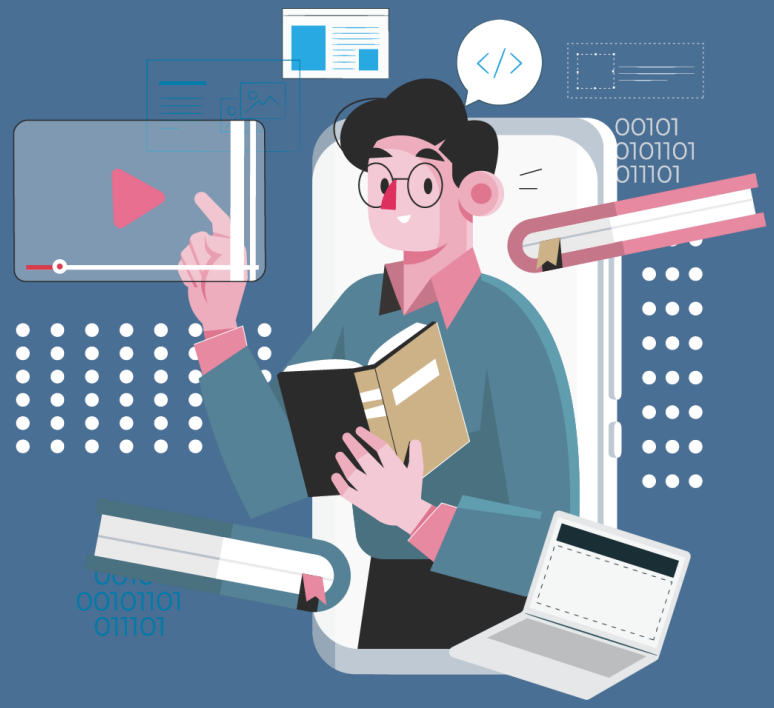
Learning Objectives

- Understand how SQL and Python can be used for data wrangling
- Clean big datasets with BigQuery and Pandas
- Appreciate first-hand the different advantages and disadvantages of using both SQL and Python for data wrangling



Agenda

- Setting up Google BigQuery
- Data Wrangling with SQL
- Verifying and Exporting Data in BigQuery
- Introduction to Pandas
- Inspecting Data with Python
- Data Wrangling with Python
- Chapter Summary & Assignment



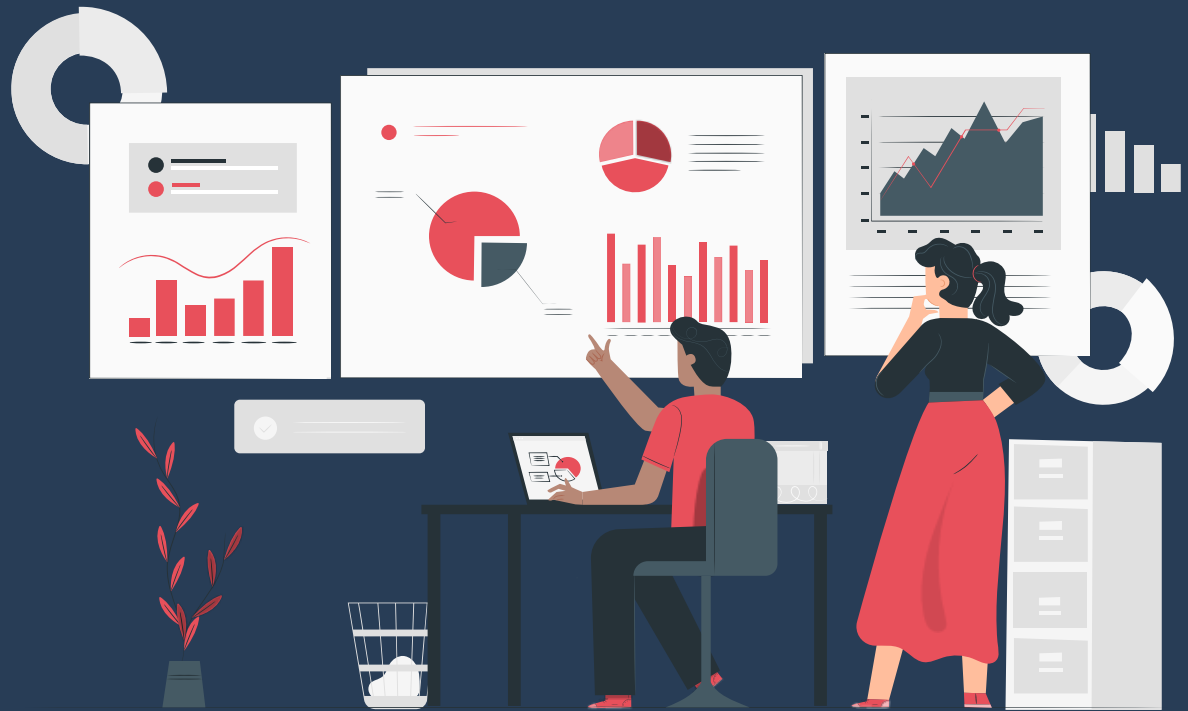
</talentlabs>

Setting up Google BigQuery

What is Google BigQuery

Getting started with BigQuery

Loading data into BigQuery



Google BigQuery

Google BigQuery is a data warehouse that helps manage and analyse your data. It allows you to query terabytes of data in seconds!

Google BigQuery – Getting started

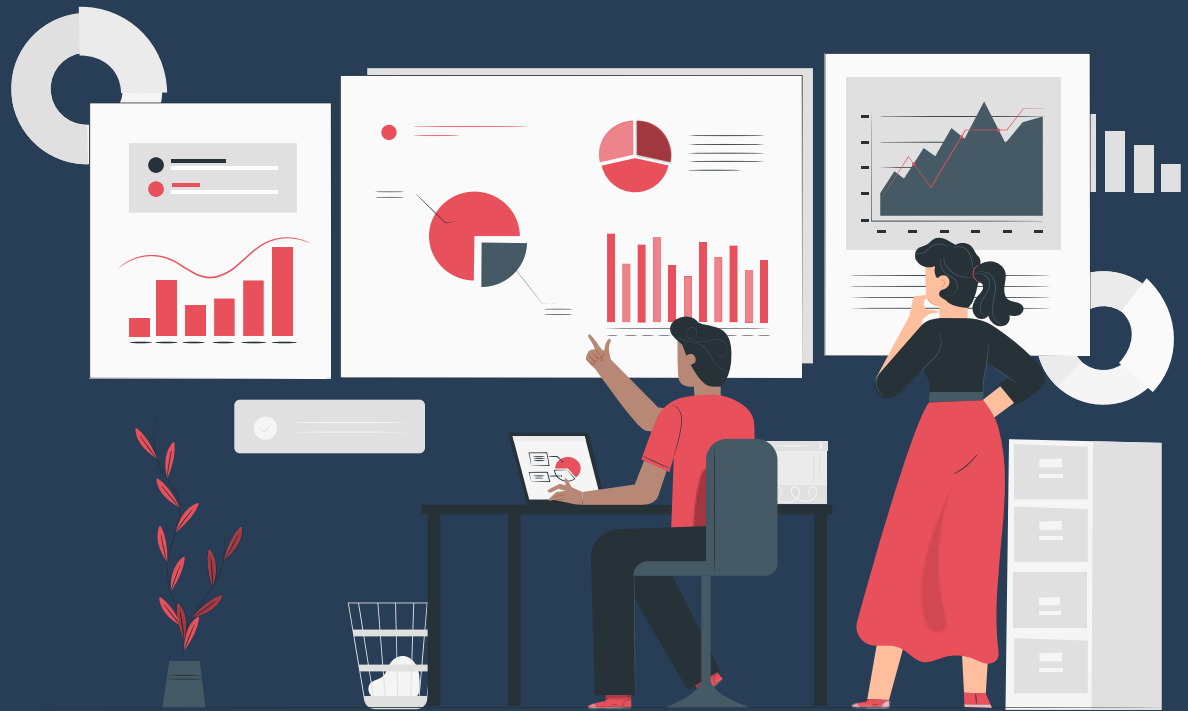
Go to: cloud.google.com/bigquery/docs/introduction#get-started-with-bigquery

Get started for free

</talentlabs>

Data Wrangling with SQL

- Data inspection in BigQuery
- Data cleaning in BigQuery
- Data wrangling in BigQuery



Google BigQuery – Data cleaning

Drop columns:

```
ALTER TABLE tableName
DROP COLUMN columnName;
```

Drop duplicates:

```
SELECT
    DISTINCT columnName
FROM
    dataset.tableName
```

Google BigQuery

</talentlabs>

Google BigQuery – Data cleaning

Check missing values:

```
SELECT
    columnName
FROM
    dataset.tableName
WHERE
    columnName is NULL;
```

Fill Missing Values:

```
UPDATE
    dataset.tableName
SET
    columnName="value"
WHERE
    columnName is NULL;
```

Google BigQuery

</talentlabs>

Google BigQuery – Data cleaning

Remove whitespaces with TRIM():

```
UPDATE
    dataset.tableName
SET
    columnName= TRIM(columnName)
WHERE
    TRUE;
```

Replace errors:

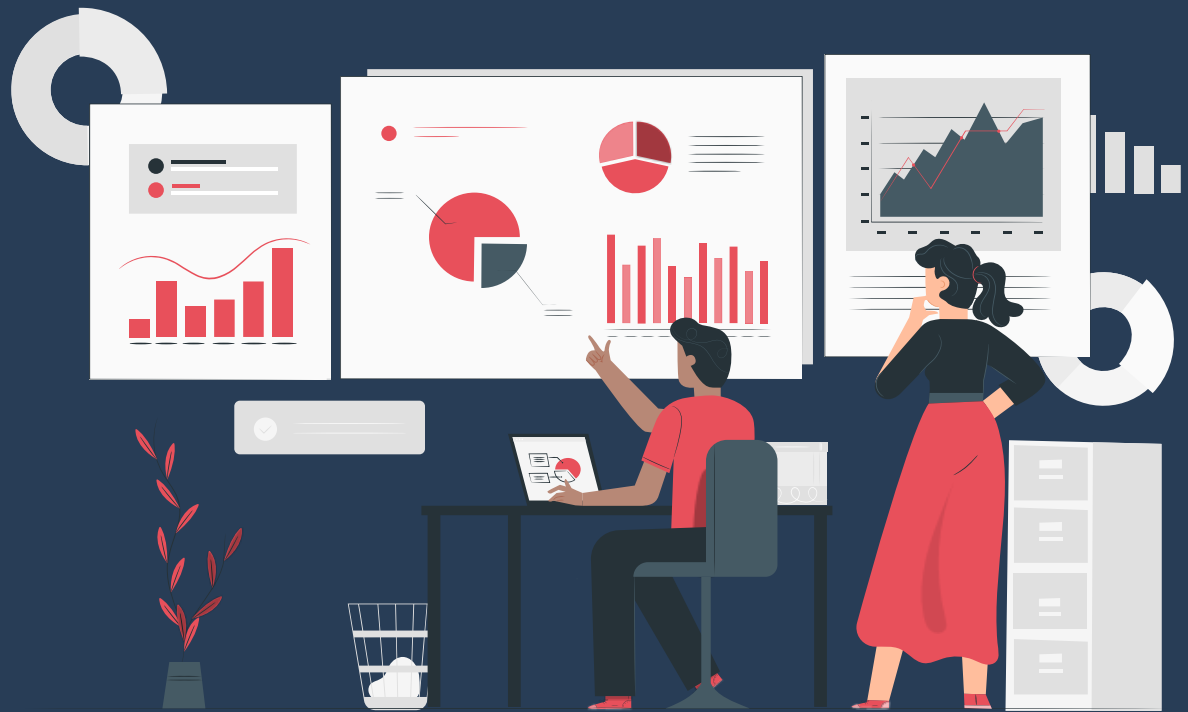
```
UPDATE
    dataset.tableName
SET
    columnName= “new value”
WHERE
    idColumn = id_value
```

Google BigQuery

</talentlabs>

Verifying and Exporting Data in BigQuery

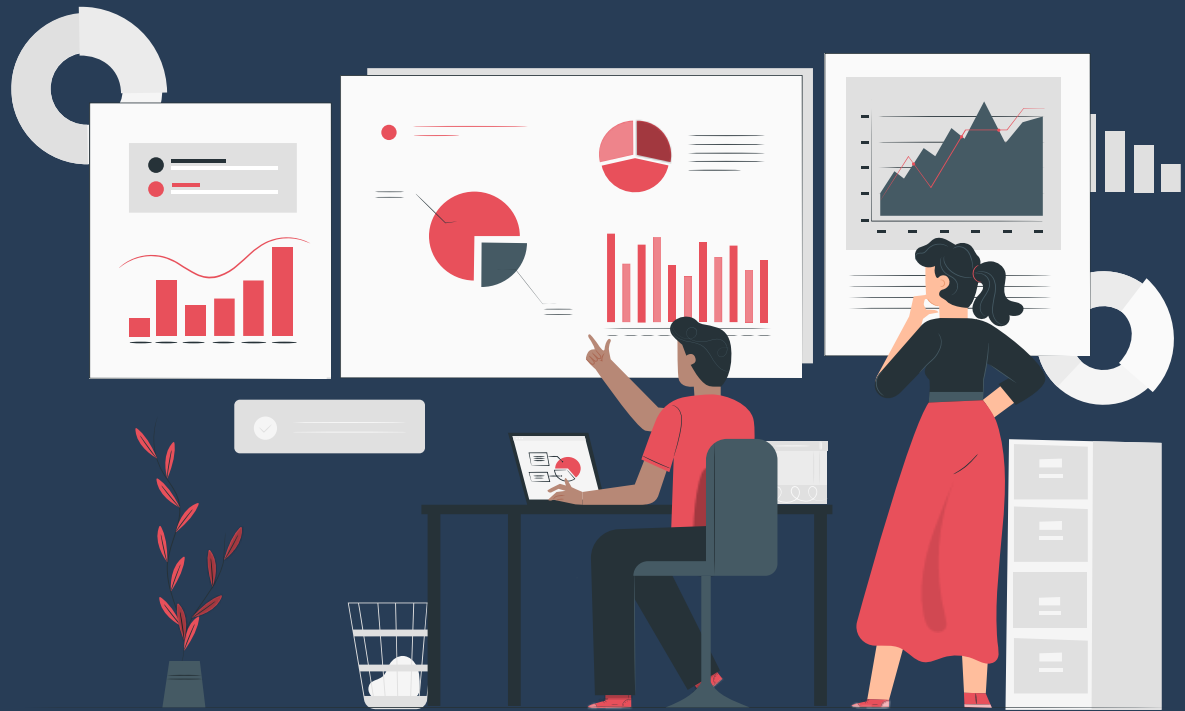
- Query history
- Saving queries
- Exporting data



Introduction to Pandas

The Pandas Library

Google Colaboratory



Pandas

Pandas is:

- a fast and easy to use tool for data manipulation and data analysis.
- the most popular Python library for data analysis.
- open-source.

Documentation: https://pandas.pydata.org/docs/

Notebooks

Allow editing and running Python via a web browser.

Popular options:

- Jupyter Notebook (via Anaconda)
- Google Colaboratory
- VS Code with Jupyter Notebook extension

Google Colaboratory

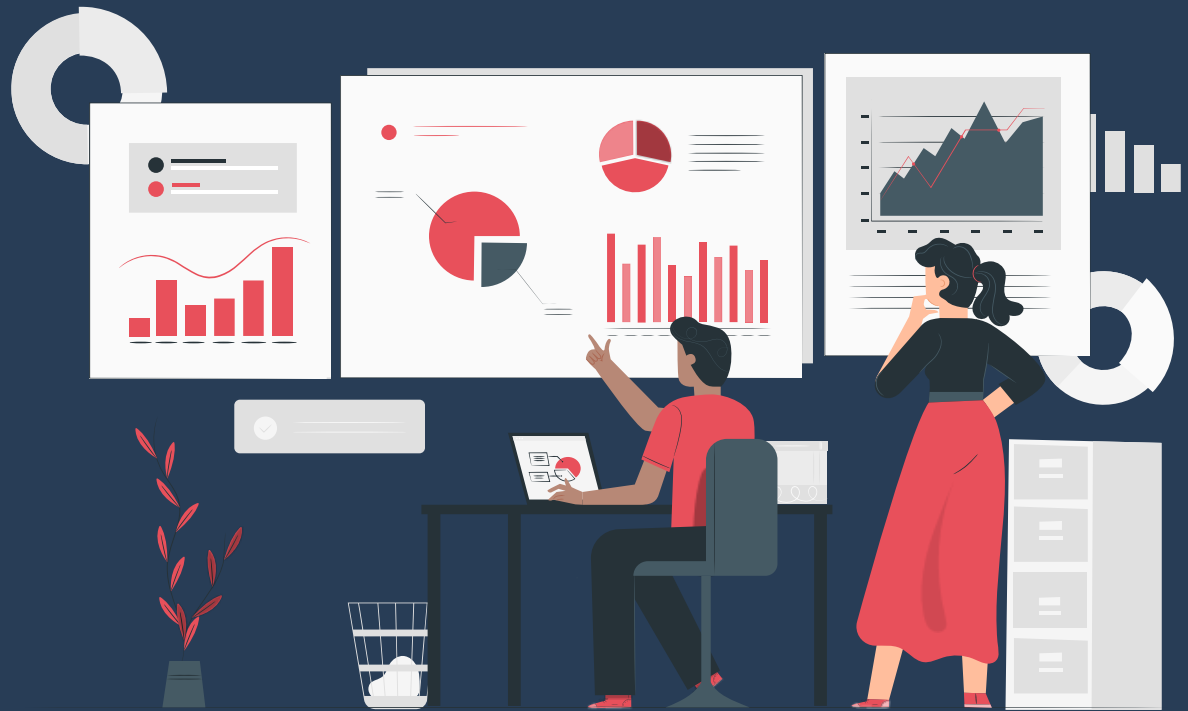
To set up, go to: colab.research.google.com

Next, login with your Google Account.

Inspecting Data with Python

Previewing data

Data Inspection



Loading and previewing data

Read data:

```
df=pd.read_csv("data_path.csv")
```

Preview data:

```
df.head()
```

View specific column:

```
df["ColumnName"] OR df.ColumnName
```

Index based selection:

```
df.iloc[0,0] OR df.loc[0, "ColumnName"]
```

Data Inspection

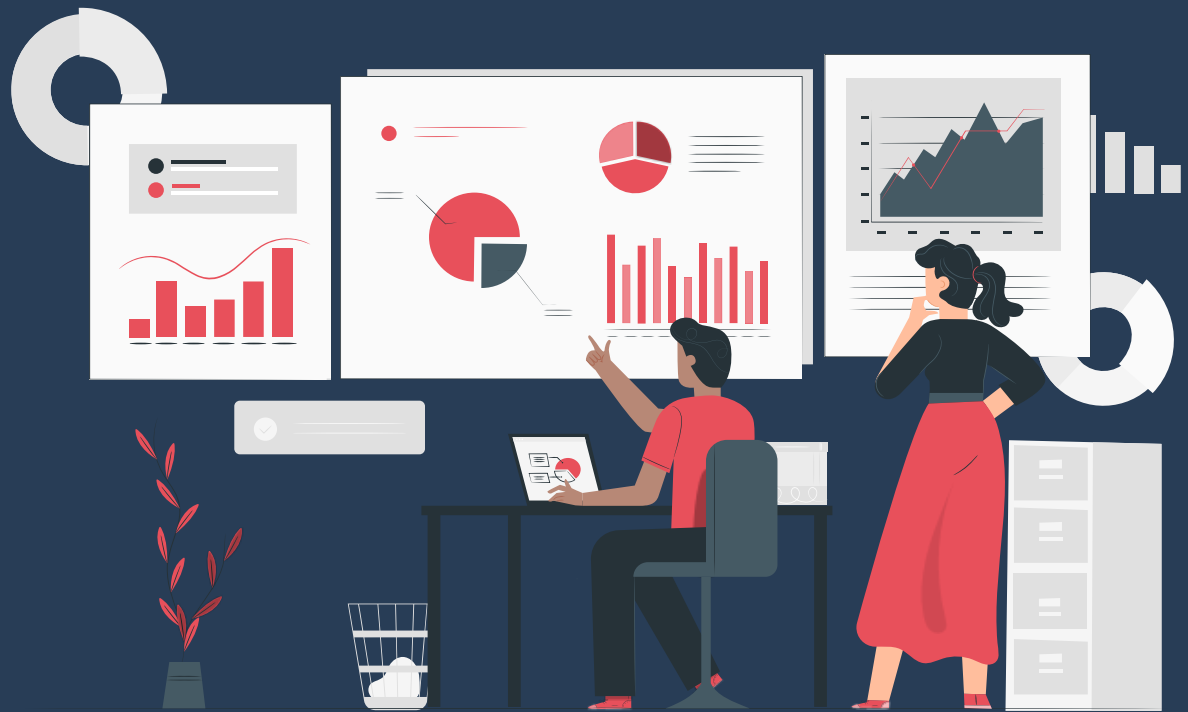
Multiple methods including:

- .head()
- .describe()
- .info()
- .shape
- .value_counts()

Data Wrangling with Python

Cleaning data with Python

Enriching data with Python



Duplicates

Check for duplicates:

```
.duplicated().sum()
```

Remove duplicates:

```
.drop_duplicates(inplace=True)
```

Missing Values

Check for missing values:

.isna().sum()

Drop missing value rows:

.dropna()

Drop missing value columns:

.dropna(axis=1)

Fill missing values

.fillna(0)

Flag missing values

df[“Missing”]=df.ColumnName.isna()

Text formatting

Rename columns, index, values with:

.rename()

Remove leading spaces:

.str.strip()

Remove text from left/right of a string:

.str.lstrip(“text”) / .str.rstrip(“text”)

Conditional Selection Filtering

```
df.loc[df.columnThree > Value0]

df.loc[(df.columnOne == “Value1”) & (df.columnTwo < Value2)]

df.loc[df.columnThree.isin(“Value3”, “Value4”)]
```

Merging Date

Can join data in multiple ways using: join, merge and concatenate.

Check out pandas documentation:

https://pandas.pydata.org/pandas-docs/dev/user_guide/merging.html

Chapter Summary & Assignment