UDACITY Google spark

Palestine Launchpad Program

Data Analysis NanoDegree

Our agenda for today

1. Recap about the last session
2. Answer Your Questions on Google form
3. ETL Example with Marwan
4. How to extract data from PCS with Saeeda
5. Project (Two) Rubric
6. Knowledge sharing
7. Q\A

Palestine Launchpad
منصة فلسطين التقنية
Google

Recap

What did we talk about last session

UDACITY

Last Session we covered the following:

Here's a breakdown of the key areas we covered:

- ✓ **Project (two) overview:** We gave an example of project two submission and made clear what is project two about.
- ✓ **Case Study:** Provided detailed examples and innovative ideas for Project Two, which I believe would be highly beneficial for your ongoing work.
- ✓ **Q&A:** We addressed and responded to various inquiries that were raised.
- ✓ **New Concepts:** Scrape Google Play Store

Palestine Launchpad
منصة فلسطين التقنية
Google

Q\A

Answers to your enquieres

UDACITY

1- I didn’t get the differences in quality issues, which function visually and programmatically should be used, should we address only the issues that does exist as problems in our dataset? I mean I chose 4 quality issues but only 2

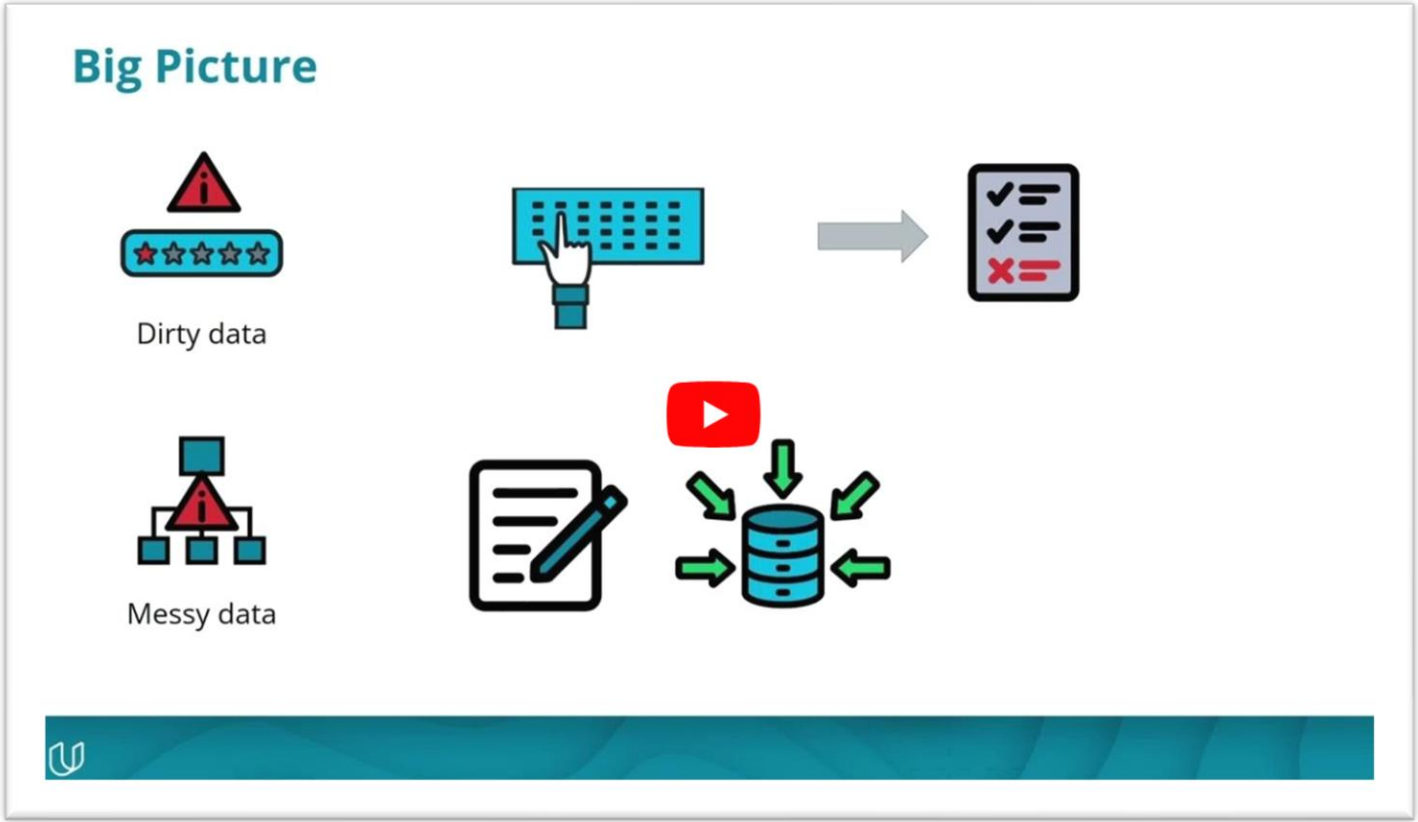
2- what if my datasets that I collected were clean? didn't have actual issues?

3- Should I get wrong datasets to try to clean them?

1. I didn’t get the differences in quality issues,
2. which function visually and programmatically should be used,
3. should we address only the issues that does exist as problems in our dataset? I mean I chose 4 quality issues but only 2

Assessing Data: Dirty Data Vs Messy Data

1. Dirty data
2. Messy data



- Assessing Data: Dirty Data Vs Messy Data

Hint: Remember that **low-quality** data has **content issues**, and **untidy data** has **structural issues**.

https://vita.had.co.nz/papers/tidy-data.pdf

UDACITY

Messy Data

The following is a table accounting for the number of produce deliveries over a weekend.

What are the variables in this dataset? What object or event are we measuring?

	Friday	Saturday	Sunday
Morning	15	158	10
Afternoon	2	90	20
Evening	55	12	45

What's the issue? How do we fix it?

Messy Data

We're measuring individual deliveries; the variables are Time, Day, Number of Produce.

	Friday	Saturday	Sunday
Morning	15	158	10
Afternoon	2	90	20
Evening	55	12	45

Problem: each column header represents a single value rather than a variable. Row headers are "hiding" the Day variable. The values of the variable, "Number of Produce", is not recorded in a single column.

Fixing Messy Data

We need to reorganize the information to make explicit the event we're observing and the variables associated to this event.

ID	Time	Day	Number
1	Morning	Friday	15
2	Morning	Saturday	158
3	Morning	Sunday	10
4	Afternoon	Friday	2
5	Afternoon	Saturday	9
6	Afternoon	Sunday	20
7	Evening	Friday	55
8	Evening	Saturday	12
9	Evening	Sunday	45

Tabular = Happy ☺

Common causes of messiness are:

- Column headers are values, not variable names
- Variables are stored in both rows and columns
- Multiple variables are stored in one column/entry
- Multiple types of experimental units stored in same table

In general, we want each file to correspond to a dataset, each column to represent a single variable and each row to represent a single observation.

We want to **tabularize** the data. This makes Python happy.

dimensions of data quality:

- **Completeness** is a metric that helps you understand whether your data is sufficient to answer interesting questions or solve your problem. **df.info()**
- **Validity** is a metric helping you understand how well your data conforms to a defined set of rules for data, also known as a schema. **df.dtypes, isnull(), df.colmn.value_counts()**
- **Accuracy** is a metric that helps you understand whether your data accurately represents the reality it aims to depict. **df.describe()**
- **Consistency** is a metric that helps you understand two things: whether your data follows a standard format and whether your data's info matches with information from other data sources. **df.describe()**
- **Uniqueness** is a metric that helps you understand whether there are duplicate or overlapping values in your data. **df.colmn.value_counts()**

How Do we Assess Data Quality?

UDACITY

First Method: Visual Assessment

Visual assessment is **simply opening data** and looking through it in its entirety.

You can visually assess data in Jupyter Notebook via **pandas** using the **.head(), .tail() or .sample()** functions, a text editor, or a spreadsheet application.

UDACITY

Second Method:
Programmatically

Visual Assessment Example

```
import pandas as pd
test_scores = pd.read_csv('test_scores.csv')

test_scores.head(10)
```

	Name	Age	Test A Score
0	Amy Linn	14	95'
1	Marc Fletcher	15	50'
2	Naima Barry	NaN	100'
3	John Carter	14	NaN
4	Dewey Cobb	14	100'
5	Amy Linn	14	85'
6	Dewey Cobb	Fourteen	Sixty six
7	Zeeshan Gibson	120	108'
8	Liem Gibson	14	NaN
9	Marc Fletcher	15	32'

- Accuracy: Incorrect parsing of Liem Gibson's name
- Accuracy: Student's age reported as 120 in row 7
- Validity : Test scores are strings not integers
- Consistency: Age and test score aren't digits in row 6
- Completeness: Missing values in rows 2, 3, 8
- Uniqueness: same student and age with different test scores for same test

Palestine Launchpad
منصة فلسطين التقنية
Google

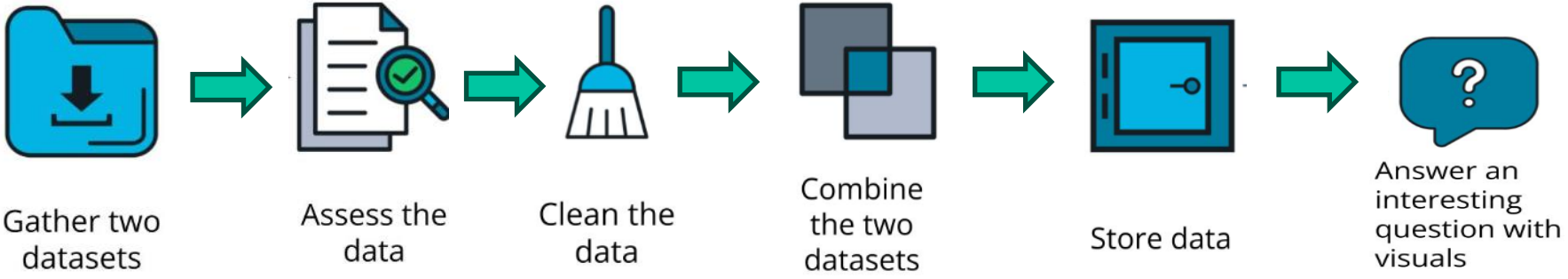
Project Two : Data Wrangling

What is needed to make a successful submit

UDACITY

Project: Real World Data Wrangling with Python

Project: Real World Data Wrangling with Python



awraq-english

Project Rubric

Code Quality and Submission Phase

✓	The student has uploaded a .zip folder containing their Jupyter Notebook for code review, and their datasets (file/link) for running the code.
✓	The project shows thorough documentation of justification of wrangling decisions.
✓	Does the code work?

Project Rubric

Gathering, Assessment, and Cleaning

✓	The project has a proper explanation of the problem statement.
✓	The student has gathered at least two separate datasets using two different data gathering methods.
✓	The student cleans the data issues they identified with the explanation and justifications.
✓	The student assesses the datasets for quality and tidiness.
✓	Remove unnecessary variables and combine datasets

Project Rubric

Data Storage and Answering the Research Question

- Students must update their data store.
- Students must identify next steps for the project.
- Students will define and answer a short research question.

Palestine Launchpad
منصة فلسطين التقنية
Google

New Concepts

Udacity, Students and Session Leads

UDACITY

Skimpy

https://colab.research.google.com/gist/aeturrell/7bf183c559dc1d15ab7e7aaac39ea0ed/skimpy_demo.ipynb

Dataset Sources

🚀 Exploring Data Sources for Data Analysis Projects | by Alaa' Omar | Jun, 2024 | Medium

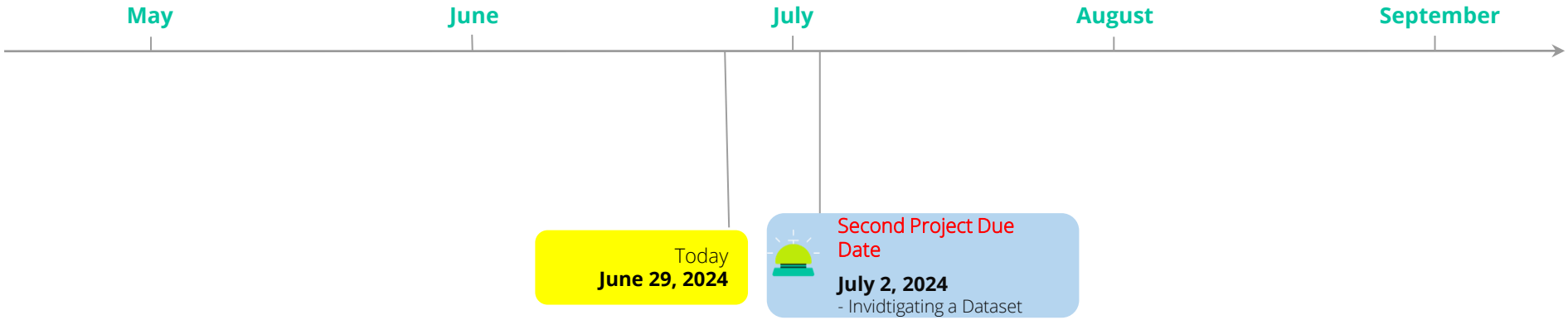
Palestine Launchpad
منصة فلسطين التقنية
Google

Next Steps

Let's get things started...

UDACITY

Second Project July 2, 2024



Start working with the second project:

1. Finish Project
2. Meet Rubric

Palestine Launchpad
منصة فلسطين التقنية
Google

Thank You!

And Good Luck!

UDACITY