



Palestine Launchpad Program

Data Analysis NanoDegree



Our agenda for today

- 1 Recap about the last session
- 2 Answer Your Questions on Google form
- 3 Project (one) overview
- 4 New Concepts
- 5 Weekly Schedule
- 6 Activity (Study Case)
- 7 Q\A

Recap

What did we talk about last session

Last Session we covered the following:

- ✓ Introduction to the Data Analysis program.
- ✓ Insightful discussions on the Udacity community (Circle).
- ✓ Overview of the Connect Session and its importance.
- ✓ Community guidelines and expectations.
- ✓ A brief exploration of the Data Analysis process and some basic statistical concepts .
- ✓ Plans for gathering your enquiries and suggestions via a Google Form to tailor upcoming sessions to your needs, I'll send you a link on Thursday.
- ✓ Commitment to adhering to our weekly schedule to ensure we cover all material thoroughly before each session.

Q\A

Answers to your enquieres

I recived the following

- ☐ Panda explode
- ☐ Didn't know how to setup Jupiter notebooks on my local machine and ended up using it only online.
- ☐ Slicing, loc and iloc
- ☐ Kafka
- ☐ Can I still working in workspace or I should install the program in my computer? Can I work the first project on the workspace? and submit it if there is an error occur, and I cant find it?
- ☐ the way for submitting projects in details.
- ☐ Yes, I'd like to know more about web scrapping and the best tools and how to use those tools to gather data from the web.
- ☐ I hope we received email before more days.

Project one : Invistigating a dataset

What is needed to make a successful submit

What is this project about?

1. Choose one of three given datasets.
2. Go through the entire data analysis process
3. Start by asking questions and end by sharing findings

Step One - Choose Your Dataset

Dataset	Overview and Notes	Example Questions
TMDb movie data - (cleaned from original data on Kaggle)	<p>This data set contains information about 10,000 movies collected from The Movie Database (TMDb), including user ratings and revenue.</p> <ul style="list-style-type: none">• Certain columns, like 'cast' and 'genres', contain multiple values separated by pipe () characters.• There are some odd characters in the 'cast' column. Don't worry about cleaning them. You can leave them as is.• The final two columns ending with "_adj" show the budget and revenue of the associated movie in terms of 2010 dollars, accounting for inflation over time.	Which genres are most popular from year to year? What kinds of properties are associated with movies that have high revenues?
No-show appointments - (original source on Kaggle)	<p>This dataset collects information from 100k medical appointments in Brazil and is focused on the question of whether or not patients show up for their appointment. A number of characteristics about the patient are included in each row.</p> <ul style="list-style-type: none">• 'ScheduledDay' tells us on what day the patient set up their appointment.• 'Neighborhood' indicates the location of the hospital.• 'Scholarship' indicates whether or not the patient is enrolled in Brazilian welfare program Bolsa Familia.• Be careful about the encoding of the last column: it says 'No' if the patient showed up to their appointment, and 'Yes' if they did not show up.	What factors are important for us to know in order to predict if a patient will show up for their scheduled appointment?
FBI Gun Data - (original source on Github)	<p>The data comes from the FBI's National Instant Criminal Background Check System. The NICS is used by to determine whether a prospective buyer is eligible to buy firearms or explosives. Gun shops call into this system to ensure that each customer does not have a criminal record or isn't otherwise ineligible to make a purchase. The data has been supplemented with state level data from census.gov.</p> <ul style="list-style-type: none">• The NICS data is found in one sheet of an .xlsx file. It contains the number of firearm checks by month, state, and type.• The U.S. census data is found in a .csv file. It contains several variables at the state level. Most variables just have one data point per state (2016), but a few have data for more than one year.	What census data is most associated with high gun per capita? Which states have had the highest growth in gun registrations? What is the overall trend of gun purchases?

Project Rubric: Investigate a Dataset

Code Functionality

Criteria	Submission Requirements
Does the code work?	<ul style="list-style-type: none"> All code is functional and produces no errors when run. The code given is sufficient to reproduce the results described.
Does the project use NumPy and Pandas appropriately?	<ul style="list-style-type: none"> The project uses NumPy arrays and Pandas Series and DataFrames where appropriate rather than Python lists and dictionaries. Where possible, vectorized operations and built-in functions are used instead of loops.
Does the project use good coding practices?	<ul style="list-style-type: none"> The code makes use of at least 1 function to avoid repetitive code. The code contains good comments and meaningful variable names, making it easy to read.

Project Rubric: Investigate a Dataset

Quality of Analysis

Criteria	Submission Requirements
Is a question clearly posed?	The project clearly states one or more questions, then addresses those questions in the rest of the analysis.

Start by stating the questions firstly then one-by-one you addressed them.

Project Rubric: Investigate a Dataset

Start by stating the questions firstly then one-by-one you addressed them.

Keep your report tidy and maintain good structure

1. Introduction

Introduction

Dataset Description

The no-show appointments dataset compiles information regarding medical appointments in Brazil, specifically focusing on patient attendance behavior. This dataset, comprising 110,527 records, aims to explore the factors influencing whether patients show up for their scheduled appointments. The data was sourced from [Kaggle](#), providing a comprehensive view of various aspects related to medical appointments.

The dataset encompasses 14 features for each patient, including:

- **ScheduledDay:** Tells us on what day the patient set up their appointment.
- **Neighborhood:** Indicates the location of the hospital.
- **Scholarship:** Indicates whether or not the patient is enrolled in the Brazilian welfare program [Bolsa Família](#).
- **No-show:** Indicates whether the patients showed up for their appointment or not (No means showed up while Yes means didn't show up).

The list of all features include:

	Feature	Dtype
0	PatientId	float64
1	AppointmentID	int64
2	Gender	object
3	ScheduledDay	object
4	AppointmentDay	object
5	Age	int64
6	Neighbourhood	object
7	Scholarship	int64
8	Hipertension	int64
9	Diabetes	int64
10	Alcoholism	int64
11	Handcap	int64
12	SMS_received	int64
13	No-show	object

Question(s) for Analysis

1. What is the strength and direction of the correlation between age and the date difference between scheduling and attending appointments, as measured by the Spearman correlation coefficient?
2. How does the distribution of patient age vary between attendance and absence at appointments?
3. Is there a relationship between the waiting time and the likelihood of patients attending?
4. Does sending SMS reminders to patients have a significant impact on their attendance for appointments?

[Bravo-EDA/Bravo_EDA.ipynb at main · POS-Cross/Bravo-EDA \(github.com\)](#)

Project Rubric: Investigate a Dataset

Data Wrangling Phase

Criteria	Submission Requirements
Is the data cleaning well documented?	The project documents any changes that were made to clean the data, such as merging multiple files, handling missing values, etc.

How to well-document our cleaning phase

- User sections and subsections
- Each subsection should contain the main idea
- Describe this subsection, describe in your language what exactly did you do, to which variables and most importantly Why:

Case Study

Put all thingd together

QA

Udacity, Students and Session Leads

Next Steps

→
Let's get things started...

First Project May 28, 2024

May

June

July

August

September

Today
May 11, 2024



First Project Due Date

May 28, 2024
- Investigating a Dataset

Start working with the project:

1. Select the dataset
2. Understand the context
3. Finish the cleaning process
4. Document your steps by writing the introduction, and the Data clean sections.
5. Start thinking about the research questions.



Program Period
May 9, 2024 - September 14, 2024

I do not paint things,
I paint only the differences
between things

Henri Matisse
Paris, 1943

“The greatest value of a picture is when it forces us to notice what we never expected to see.”



John Tukey

Thank You!

→
And Good Luck!