

Start Assignment

Due Nov 5 by 23:59 **Points** 100 **Submitting** a file upload
Available after Oct 4 at 0:00

A central graphic with the text "Exploratory Data Analysis" in bold black font. The background is a light blue and white geometric pattern of interconnected lines and dots. Surrounding the text are several colorful icons: a group of people in a meeting, a network diagram, a person presenting to a screen, a cloud with circuit lines, a person at a computer, a magnifying glass over a bar chart, a person with a robot, a line graph, a group of people in a meeting, a code editor icon, a person at a computer, a database cylinder, and a person with a robot.

This section will provide the ability for the user to load a CSV file to be analysed. Once loaded, the application will convert it into a Pandas dataframe and display the information listed below.

The web application will provide the following mandatory functionalities:

- Upload CSV file and load data as Pandas dataframe

- Display header called “Overall Information”
- Display filename
- Display number of rows
- Display number of columns
- Display number of duplicated rows
- Display number of rows with missing values

1. Overall Information

Name of Table: csse_covid_19_daily_reports_us_01-01-2021.csv

Number of Rows: 58

Number of Columns: 18

Number of Duplicated Rows: 0

Number of Rows with Missing Values: 58

- Display list of columns and their data type (text, numeric, date)

List of Columns:

Province_State, Country_Region, Last_Update, Lat, Long_, Confirmed, Deaths, Recovered, Acti

Type of Columns:

	type
Province_State	object
Country_Region	object
Last_Update	object
Lat	float64
Long_	float64
Confirmed	int64
Deaths	int64
Recovered	float64
Active	float64
FIPS	float64
Incident_Rate	float64

Select the number of rows to be displayed

- Display slider for selecting the number of rows to be displayed



- Display top N rows (default 5 rows) of dataset

Top Rows of Table

	Province_State	Country_Region	Last_Update	Lat	L
0	Alabama	US	2021-01-02 05:30:44	32.3182	-86
1	Alaska	US	2021-01-02 05:30:44	61.3707	-152
2	American Samoa	US	2021-01-02 05:30:44	-14.2710	-170
3	Arizona	US	2021-01-02 05:30:44	33.7298	-111
4	Arkansas	US	2021-01-02 05:30:44	34.9697	-92
5	California	US	2021-01-02 05:30:44	36.1162	-119
6	Colorado	US	2021-01-02 05:30:44	39.0598	-105
7	Connecticut	US	2021-01-02 05:30:44	41.5978	-72
8	Delaware	US	2021-01-02 05:30:44	39.3185	-75
9	Diamond Princess	US	2021-01-02 05:30:44	NaN	
10	District of Columbia	US	2021-01-02 05:30:44	38.8974	-77

Note: each time the slider is changed the number of rows displayed will be updated

- Display bottom N rows (default 5 rows) of dataset

Bottom Rows of Table

	Province_State	Country_Region	Last_Update	Lat	Lo
8	Delaware	US	2021-01-02 05:30:44	39.3185	-75.
9	Diamond Princess	US	2021-01-02 05:30:44	NaN	
10	District of Columbia	US	2021-01-02 05:30:44	38.8974	-77.
11	Florida	US	2021-01-02 05:30:44	27.7663	-81.
12	Georgia	US	2021-01-02 05:30:44	33.0406	-83.
13	Grand Princess	US	2021-01-02 05:30:44	NaN	
14	Guam	US	2021-01-02 05:30:44	13.4443	144.
15	Hawaii	US	2021-01-02 05:30:44	21.0943	-157.
16	Idaho	US	2021-01-02 05:30:44	44.2405	-114.
17	Illinois	US	2021-01-02 05:30:44	40.3495	-88.
18	Indiana	US	2021-01-02 05:30:44	39.8494	-86.

Note: each time the slider is changed the number of rows displayed will be updated

- Display N randomly sampled rows (default 5 rows) of dataset

Random Sample Rows of Table

	Province_State	Country_Region	Last_Update	Lat	L
46	South Carolina	US	2021-01-02 05:30:44	33.8569	-80
7	Connecticut	US	2021-01-02 05:30:44	41.5978	-72
22	Louisiana	US	2021-01-02 05:30:44	31.1695	-91
31	Nebraska	US	2021-01-02 05:30:44	41.1254	-98
12	Georgia	US	2021-01-02 05:30:44	33.0406	-83
10	District of Columbia	US	2021-01-02 05:30:44	38.8974	-77
32	Nevada	US	2021-01-02 05:30:44	38.3135	-117
55	West Virginia	US	2021-01-02 05:30:44	38.4912	-80
8	Delaware	US	2021-01-02 05:30:44	39.3185	-75
25	Massachusetts	US	2021-01-02 05:30:44	42.2302	-71
38	North Dakota	US	2021-01-02 05:30:44	47.5209	-99

Note: each time the slider is changed a new set of randomly sampled rows are displayed

- Display a multi select box for choosing which text columns will be converted to datetime

Which columns do you want to convert to dates

Choose an option

Note: once selected, the columns will be automatically converted to datetime and the Text Column and Datetime Column sections will be refreshed accordingly

2. Information on numeric columns

This section will provide the ability for the user to get a better understanding of the information contained for each numeric column of the dataset.

The web application will provide the following mandatory functionalities:

- Display name of column as subtitle
- Display number of unique values
- Display number of missing values
- Display number of occurrence of 0 value
- Display number of negative values
- Display the average value
- Display the standard deviation value
- Display the minimum value

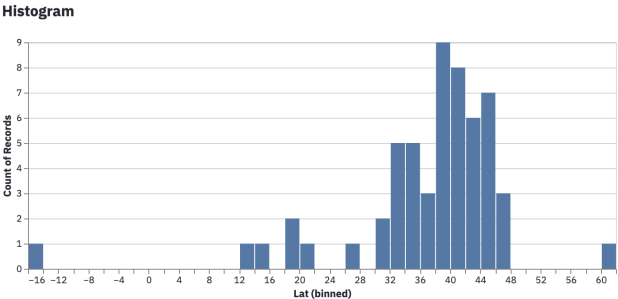
- Display the maximum value
- Display the median value

2. Numeric Column Information

2.0 Field Name: *Lat*

	value
Number of Unique Values	56
Number of Rows with Missing Values	2
Number of Rows with 0	0
Number of Rows with Negative Values	1
Average Value	36.8401
Standard Deviation Value	10.8870
Minimum Value	-14.2710
Maximum Value	61.3707
Median Value	39.0619

- Display a histogram chart with maximum number of bins: 50



- Display a table listing the occurrences and percentage of the top 20 most frequent values

Most Frequent Values

	value	occurrence	percentage
0	32.3182	1	0.0179
1	61.3707	1	0.0179
2	38.3135	1	0.0179
3	43.4525	1	0.0179
4	40.2989	1	0.0179
5	34.8405	1	0.0179
6	42.1657	1	0.0179
7	35.6301	1	0.0179
8	47.5289	1	0.0179
9	15.0979	1	0.0179
10	40.3888	1	0.0179

3. Information on text columns

This section will provide the ability for the user to get a better understanding of the information contained for each text column of the dataset.

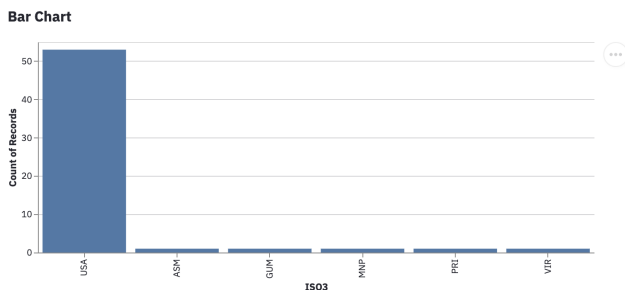
The web application will provide the following mandatory functionalities:

- Display name of column as subtitle
- Display number of unique values
- Display number of missing values
- Display number of rows with empty string
- Display number of rows with only whitespaces
- Display number of rows with only lower case characters
- Display number of rows with only upper case characters
- Display number of rows with only alphabet characters
- Display number of rows with only numbers as characters
- Display the mode value

3.3 Field Name: **ISO3**

	value
Number of Unique Values	6
Number of Rows with Missing Values	0
Number of Empty Rows	0
Number of Rows with Only Whitespace	0
Number of Rows with Only Lowercases	0
Number of Rows with Only Uppercases	58
Number of Rows with Only Alphabet	58
Number of Rows with Only Digits	0
Mode Value	USA

- Display a bar chart showing the number of occurrence for each value



- Display a table listing the frequencies and percentage for each value

Most Frequent Values

	value	occurrence	percentage
0	USA	53	0.9138
1	ASM	1	0.0172
2	GUM	1	0.0172
3	MNP	1	0.0172
4	PRI	1	0.0172
5	VIR	1	0.0172

1. Information on datetime columns

This section will provide the ability for the user to get a better understanding of the information contained for each datetime column of the dataset.

The web application will provide the following mandatory functionalities:

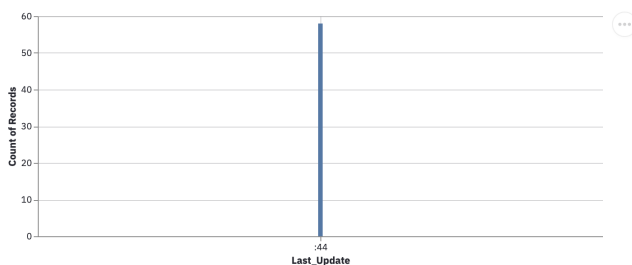
- Display name of column as subtitle
- Display number of unique values
- Display number of missing values
- Display number of occurrence of days falling during weekend (Saturday and Sunday)
- Display number of weekday days (not Saturday or Sunday)
- Display number of cases with future dates (after today)
- Display number of occurrence of 1900-01-01 value
- Display number of occurrence of 1970-01-01 value
- Display the minimum date
- Display the maximum date

4.0 Field Name: *Last_Update*

	value
Number of Unique Values	1
Number of Rows with Missing Values	0
Number of Weekend Dates	58
Number of Weekday Dates	0
Number of Dates in Future	0
Number of Rows with 1900-01-01	0
Number of Rows with 1970-01-01	0
Minimum Value	2021-01-02 05:30:44
Maximum Value	2021-01-02 05:30:44

- Display a bar chart showing the number of occurrence for each value

Bar Chart



- Display a table listing the frequencies and percentage for each value

Most Frequent Values

	value	occurrence	percentage
0	2021-01-02T05:30:44+11:00	58	1

Submission:

You will submit a zip file containing your web application and a final report. You can find the structure template here: [link \(https://drive.google.com/file/d/1Y3gHDBafiZEgjfvuZoSFd7VqKI9J67pR/view?usp=sharing\)](https://drive.google.com/file/d/1Y3gHDBafiZEgjfvuZoSFd7VqKI9J67pR/view?usp=sharing)

The zip file needs to contain the following folder and files:

- README.md
- requirements.txt

- Dockerfile
- docker-compose.yml
- app/
 - streamlit_app.py
- src/
 - data.py
 - datetime.py
 - numeric.py
 - text.py
 - test/
 - test_data.py
 - test_datetime.py
 - test_numeric.py
 - test_text.py

Instructions:

Each group will need to set up a public Github repository. Every team member will need to publish his/or work to this repository using his/her personal account. The team needs to set up an internal process for merging code to the master/main branch which needs to be described in the final report.

Each team member will be assigned a specific section of the application:

- Student A for Overall information of the dataset (responsible of src/data.py, src/test/test_data.py)
- Student B for Information on each numeric column (responsible of src/datetime.py, src/test/test_datetime.py)
- Student C for Information on each text column (responsible of src/numeric.py, src/test/test_numeric.py)
- Student D for Information on each datetime column (responsible of src/text.py, src/test/test_text.py)

The final report needs to cover the following topics:

- Presentation of the web application and link to the group Github repository
- Description of the design (services required, architecture, flow chart,...)
- Instructions to setup and launch the web application
- Description of the defined processes for group collaboration on this project such as meetings, rules, best practises, ... (include screenshots if required)
- Description of the contribution of each team member
- Problems faced and implemented solutions
- Suggested improvements of the code base or web application

All assignments need to be submitted before the due date on Canvas. Penalties will be applied for late submission.

Assessment Criteria:

- Quality and reliability of Python code and Unix commands
- Readability and consistency of coding style
- Level of clarity and relevance for documentation, flowcharts and instructions for installing and running the web application
- Robustness of the web application
- Comprehensiveness of repository structure and level of clarity of documentation of Git workflows (branch management, code review and pull request)
- Level of clarity of explanation of the web application design
- Level of clarity and quality of analysis and visualizations displayed by the web application and written report highlighting individual and teams efforts and problems faced