ENTITY

DSO104 Activities

Starting your First Data Science Portfolio Project!

Google Classroom Extra Activities for the modules from here on out, if done with passion and dedication, can lead to a complete portfolio project before you even reach DSO110! This means you can enter your data science interviews with TWO full real-world projects in your portfolio, the optimal numbers according to our data science mentors who are currently in the workforce. If you choose not to put this amount of effort into these activities then, at the very least, participating in these at least a bit will result in a substantial reusable code base that will give you a *huge* head start when you reach DSO110, the final project course module! This will give you more time to make your final ENTITY portfolio project the best it can be. Having multiple portfolio projects and/or higher quality projects will give you that competitive edge in the job market, which is what we want for ALL our ENTITY graduates!

Here we go, let’s get started!!

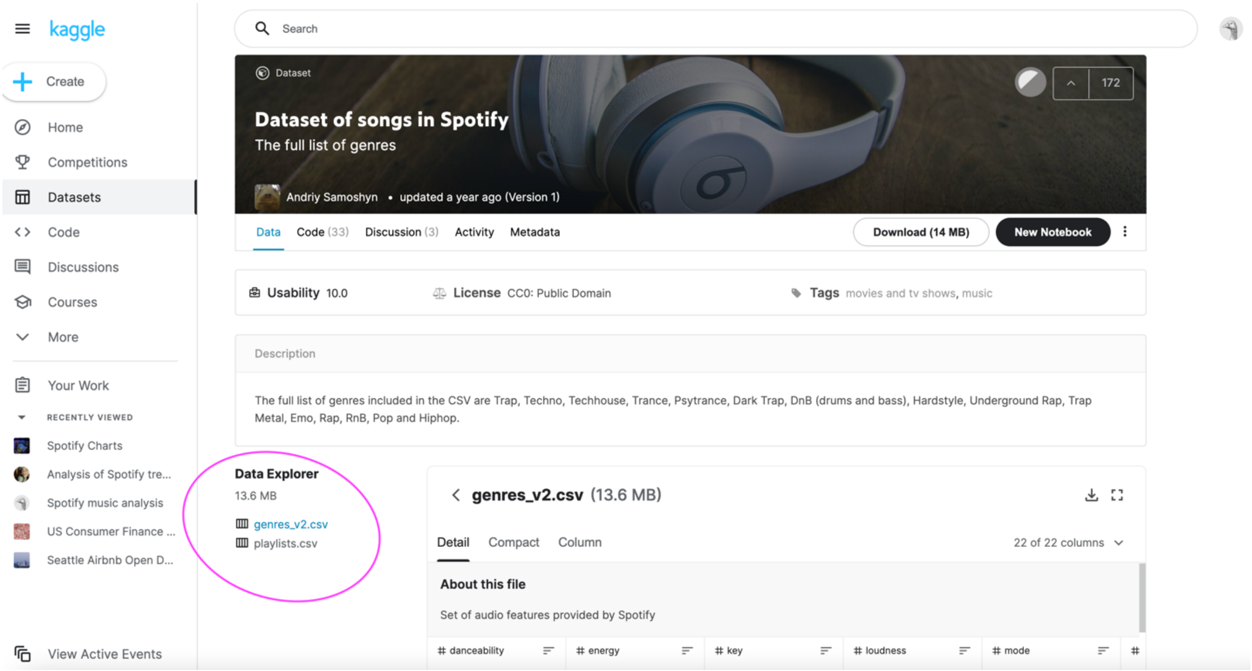
Introduction

On Kaggle.com, search different topics that

(a) interest you
(b) relate to your academic or work history
(c) seem like practical subjects for a data science project

Choose 5 dataset listings from different Kaggle URLs. Here is an example of one dataset listing:

Although there are 2 CSV files, this is considered one dataset listing (see below):



Dataset of songs in Spotify

The full list of genres

Andriy Samoshyn • updated a year ago (Version 1)

Usability 10.0 License CC0: Public Domain Tags movies and tv shows, music

Description

The full list of genres included in the CSV are Trap, Techno, Techhouse, Trance, Psytrance, Dark Trap, DnB (drums and bass), Hardstyle, Underground Rap, Trap Metal, Emo, Rap, RnB, Pop and Hiphop.

Data Explorer
13.6 MB
genres_v2.csv
playlists.csv

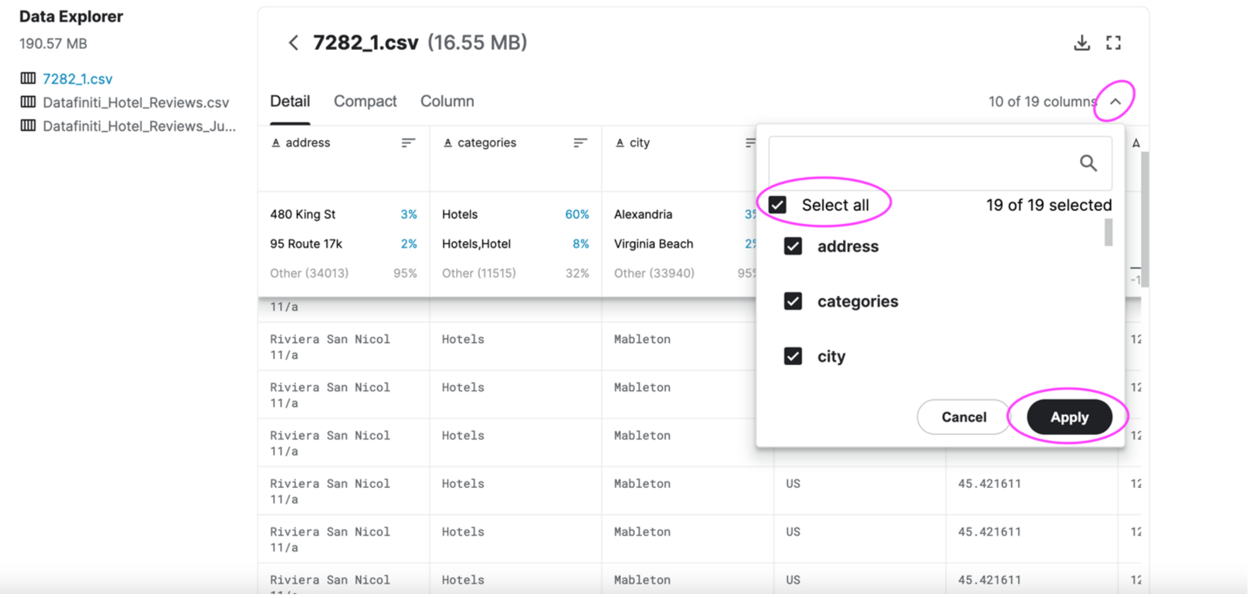
genres_v2.csv (13.6 MB)

About this file

Set of audio features provided by Spotify

# danceability	# energy	# key	# loudness	# mode	#

The 5 can be all one topic or different topics. Make sure each dataset or related datasets (like the Spotify example above) have categorical variables and numeric variables (floats/ints). You can check the datasets contents by *clicking dropdown arrow > checking “Select all” box > clicking “Apply”*.



Data Explorer
190.57 MB

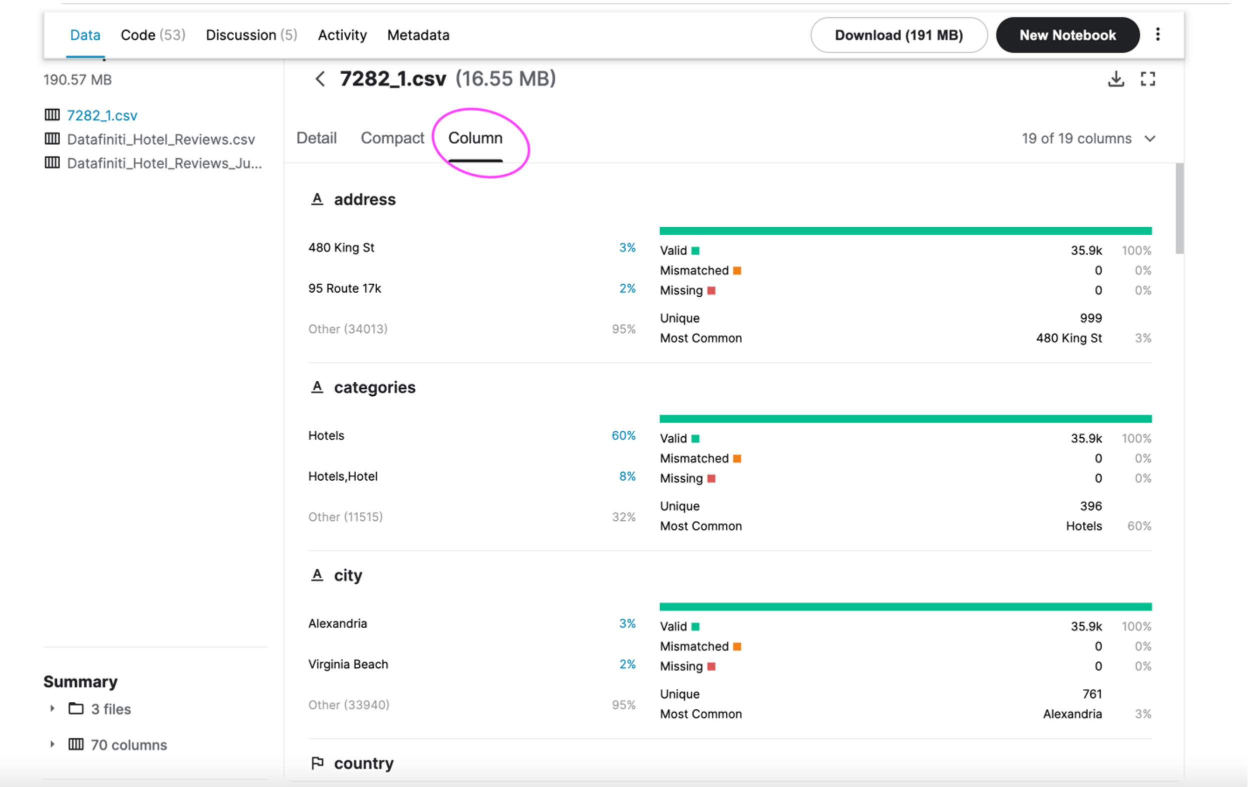
- 7282_1.csv
- Datafiniti_Hotel_Reviews.csv
- Datafiniti_Hotel_Reviews_Ju...

7282_1.csv (16.55 MB)

☑ Select all   19 of 19 selected
☑ address
☑ categories
☑ city

address	categories	city			
480 King St 3%	Hotels 60%	Alexandria 3			
95 Route 17k 2%	Hotels,Hotel 8%	Virginia Beach 2			
Other (34013) 95%	Other (11515) 32%	Other (33940) 95			
11/a					
Riviera San Nicol 11/a	Hotels	Mableton			12
Riviera San Nicol 11/a	Hotels	Mableton			12
Riviera San Nicol 11/a	Hotels	Mableton			12
Riviera San Nicol 11/a	Hotels	Mableton	US	45.421611	12
Riviera San Nicol 11/a	Hotels	Mableton	US	45.421611	12
Riviera San Nicol	Hotels	Mableton	US	45.421611	12

Also, make sure columns of interest do not have too many missing values. You can investigate by clicking “Column” tab.



Data | Code (53) | Discussion (5) | Activity | Metadata | Download (191 MB) | New Notebook

190.57 MB

- 7282_1.csv
- Datafiniti_Hotel_Reviews.csv
- Datafiniti_Hotel_Reviews_Ju...

Summary

- 3 files
- 70 columns

7282_1.csv (16.55 MB)

Detail | Compact | Column

19 of 19 columns

address

480 King St	3%	Valid 🟩	35.9k	100%
		Mismatched 🟧	0	0%
95 Route 17k	2%	Missing 🟥	0	0%
		Unique	999	
Other (34013)	95%	Most Common	480 King St	3%

categories

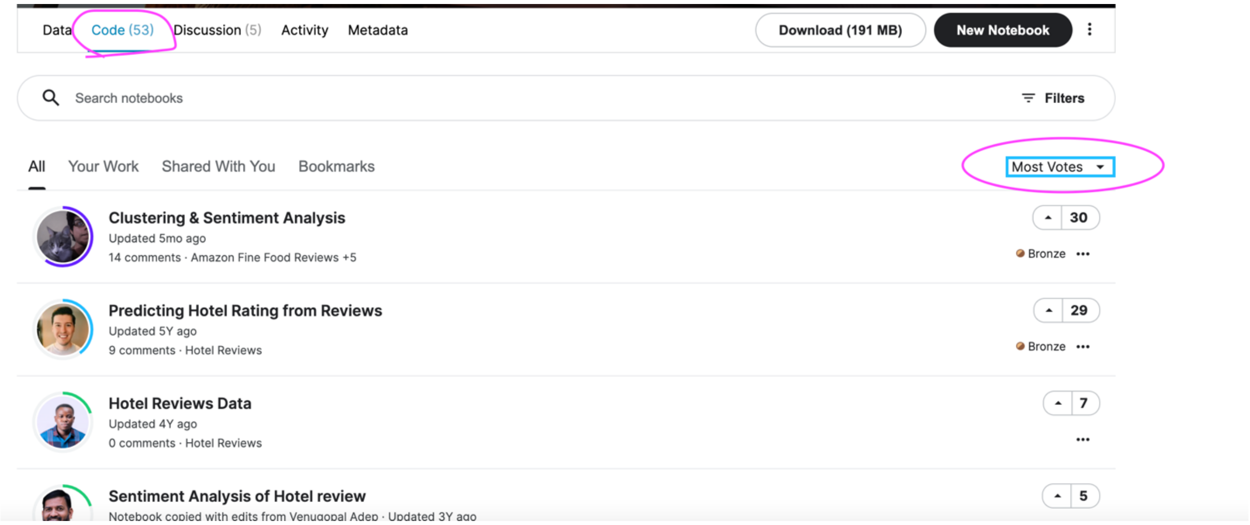
Hotels	60%	Valid 🟩	35.9k	100%
		Mismatched 🟧	0	0%
Hotels,Hotel	8%	Missing 🟥	0	0%
		Unique	396	
Other (11515)	32%	Most Common	Hotels	60%

city

Alexandria	3%	Valid 🟩	35.9k	100%
		Mismatched 🟧	0	0%
Virginia Beach	2%	Missing 🟥	0	0%
		Unique	761	
Other (33940)	95%	Most Common	Alexandria	3%

country

This is not required, but a monumental plus. Check if the datasets have notebooks associated with them. This is the benefit of Kaggle.com and the ethos of open-source platforms! You can get great ideas by “standing on the shoulders of giants” and on many occasions, do not have to reinvent the wheel! You can find associated notebooks in the “Code” tab. Click “Most Votes” to find the highest rated notebooks. This is how data scientists operate in the real world—they help one another, use one another’s code, and pay it forward on these platforms when they can!



Data Code (53) Discussion (5) Activity Metadata Download (191 MB) New Notebook

Search notebooks Filters

All Your Work Shared With You Bookmarks Most Votes

Clustering & Sentiment Analysis
Updated 5mo ago
14 comments · Amazon Fine Food Reviews +5
30 Bronze

Predicting Hotel Rating from Reviews
Updated 5Y ago
9 comments · Hotel Reviews
29 Bronze

Hotel Reviews Data
Updated 4Y ago
0 comments · Hotel Reviews
7

Sentiment Analysis of Hotel review
Notebook copied with edits from Venugopal Adep · Updated 3Y ago
5

This is a time-consuming task, but a task you will inevitably have to do. You will be glad you have done extensive dataset perusing months before you reach DSO110: Final Project.

After you have completed all of the above, you are ready to move on to the activities associated with each lesson module.

Lesson 1

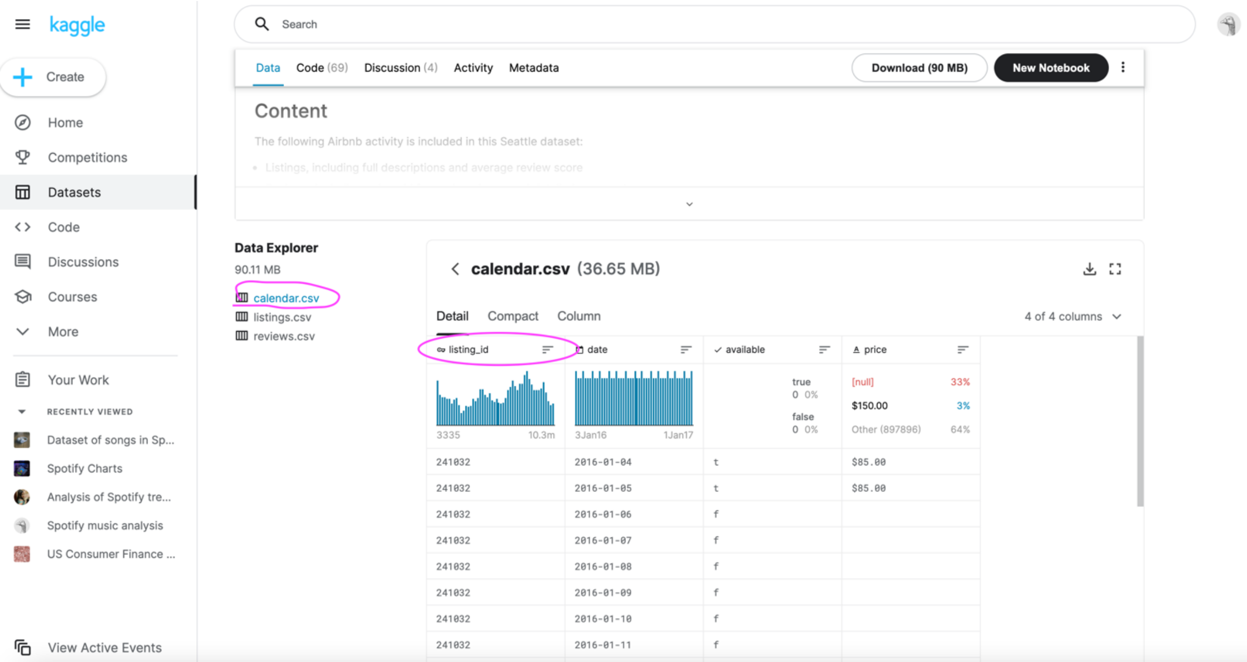
Out of the 5, choose a dataset that has columns that seem unnecessary and/or that could be renamed. Make a subset of that dataframe without those columns and/or rename the columns.

You may do this in Python and/or R.

Lesson 2

Option 1 (if dataset listing has multiple related datasets):

Out of the 5, choose a dataset listing that has multiple datasets and merge them on the appropriate columns. Example:



Content

The following Airbnb activity is included in this Seattle dataset:

- Listings, including full descriptions and average review score

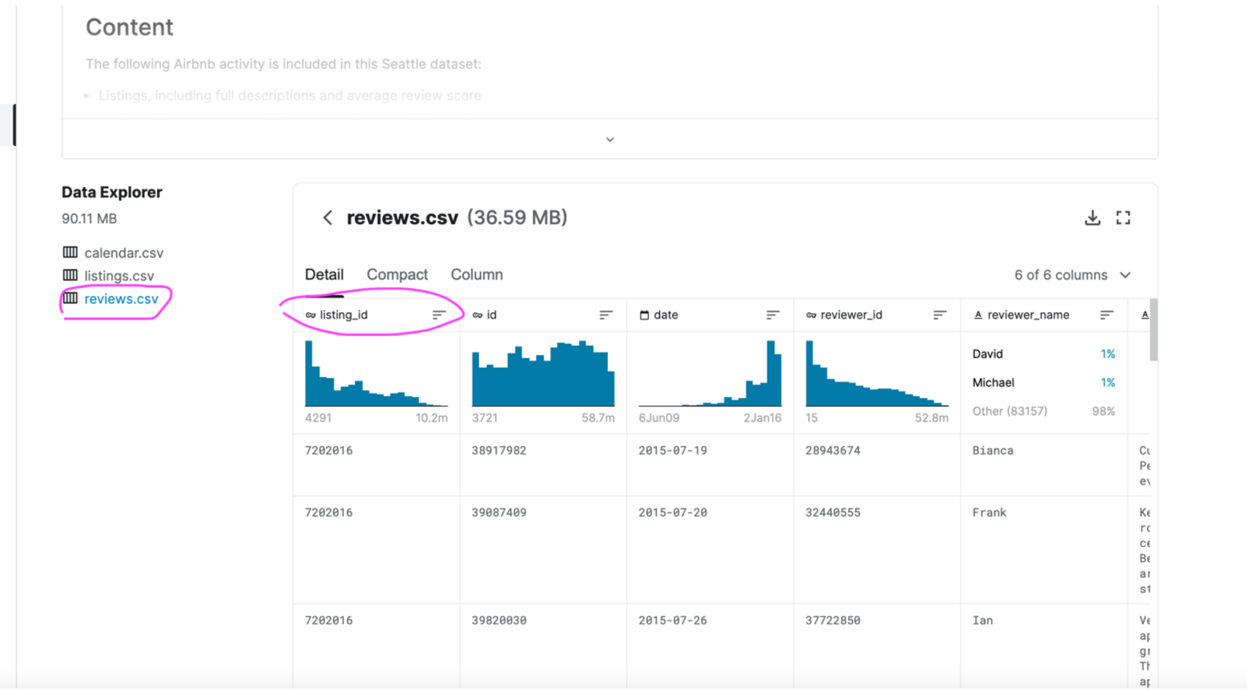
Data Explorer

90.11 MB

- calendar.csv
- listings.csv
- reviews.csv

calendar.csv (36.65 MB)

listing_id	date	available	price
		true 0 0% false 0 0%	[null] 33% $150.00 3% Other (897896) 64%
241032	2016-01-04	t	$85.00
241032	2016-01-05	t	$85.00
241032	2016-01-06	f	
241032	2016-01-07	f	
241032	2016-01-08	f	
241032	2016-01-09	f	
241032	2016-01-10	f	
241032	2016-01-11	f	



Content

The following Airbnb activity is included in this Seattle dataset:

- Listings, including full descriptions and average review score

Data Explorer

90.11 MB

- calendar.csv
- listings.csv
- reviews.csv

reviews.csv (36.59 MB)

listing_id	id	date	reviewer_id	reviewer_name
				David 1% Michael 1% Other (83157) 98%
7202016	38917982	2015-07-19	28943674	Bianca
7202016	39087409	2015-07-20	32440555	Frank
7202016	39820030	2015-07-26	37722850	Ian

The above example with data on Airbnb reviews has two datasets that can be merged on *listing_id*.

Option 2 (if dataset listings only have one dataset file or datasets don't have columns appropriate for merging):

Search Kaggle for another dataset that has a related column and dataset subjects have value for comparison. For instance, a marketing dataset with zip codes can be merged with another dataset that has the zip code populations and average annual income. This is just one example of the many possible scenarios.

Option 3 (if Option 1 and 2 are not viable or becoming a time suck):

Choose a dataset and separate it into two subsets with one categorical column in common. Merge them back into one dataset.

You may do any of these options in Python and/or R.

Lesson 3

Option 1:

Out of the 5 datasets, choose one and find a categorical column (string) where the groups have no inherent order (nominal) and dummy code that column.

Option 2:

Out of the 5 datasets, choose one and find a categorical column (string) where there are only 2 groups (binary), and change the groups to 0 and 1.

Option 3:

Out of the 5 datasets, choose one and find a column that is a continuous variable. Recode that column as a string providing a useful group name. For example, let's say there is an age column.

You could recode as the following:

age < 18 = “minor”
age > 18 = “adult”
age > 55 = “senior”

You may do any of these options in Python and/or R.

Lesson 4

Out of the 5 datasets, choose one column of interest that is a continuous variable and create a histogram. Then, choose one column of interest that is a categorical variable and create a bar chart.

Lesson 5

Out of the 5 datasets, choose one that has 3 or more continuous variables. Make a correlation heatmap.

You may do this in Python and/or R.

Lesson 6

Out of the 5 datasets, choose one and make 2 visualizations with each representing 2 or more variables.

Lesson 7

Peruse the internet and find an infographic that makes an impression on you.

Lesson 8

Watch this tutorial on interactive dashboard in Tableau:

https://www.youtube.com/watch?v=Nr31rv9tsJ8

Lesson 9

Out of the 5 datasets, choose one. Identify columns of interest. Then decide whether they would serve better as independent or dependent variables in an analysis. Finally, use the mind maps provided in the curriculum to choose an appropriate statistical analysis or predictive model. Choose at least 2 statistical analyses or predictive models that can be performed on the dataset.

Lesson 10

Watch this presentation on Storytelling with Data by Cole Knaflic. It does not have to be all in one sitting!

https://www.youtube.com/watch?v=8EMW7io4rSI